Pós-Graduação em Ciência da Computação

PAULO MELLO DA SILVA

UMA ABORDAGEM DE *ENSEMBLE REGRESSION* PARA O DIAGNÓSTICO DE PROBLEMAS EDUCACIONAIS

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

Recife
2019

PAULO MELLO DA SILVA

UMA ABORDAGEM DE *ENSEMBLE REGRESSION* PARA O DIAGNÓSTICO DE PROBLEMAS EDUCACIONAIS

Tese apresentada ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciências da Computação.

Área de concentração: Mineração de Dados
Orientador: Profº. Dr.Fernando da Fonseca de Souza

Recife
2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S586a Silva, Paulo Mello da
Uma abordagem de *ensemble regression* para o disgnóstico de problemas educacionais / Paulo Mello da Silva. – 2019.
166 f.: il., fig., tab.

Orientador: Fernando da Fonseca de Souza.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências.

1. Mineração de dados. 2. Diagnóstico educacional. I. Souza, Fernando da Fonseca de (orientador). II. Título.

006.312 CDD (23. ed.) UFPE- MEI 2019-172

Paulo Mello da Silva

“Uma Abordagem de Ensemble Regression para o Diagnóstico de Problemas Educacionais”

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 12/09/2019.

_____

Orientador: Prof. Dr. Fernando Fonseca de Souza

BANCA EXAMINADORA

_____

Profa. Dra. Renata Maria Rodrigues Cardoso de Souza
Centro de Informática/UFPE

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática/UFPE

_____

Prof. Dr. Alex Sandro Gomes
Centro de Informática/UFPE

_____

Prof. Dr. Sérgio Paulino Abranches
Departamento de Fundamentos Sócio Filosóficos da Educação/UFPE

_____

Prof. Dr. José Adson Oliveira Guedes da Cunha
Departamento de Ciências Exatas/UFPB

Dedico esse trabalho:

A Deus

A minha Mãe Ivoneide Pereira da Silva

A minha Esposa Roberta Andrade de Araújo Fagundes

As minhas Filhas Rayssa Fagundes Mello e

Sarah Fagundes Mello

AGRADECIMENTOS

Esta conquista teve a participação e colaboração de várias pessoas. Todas elas contribuíram de alguma forma para a construção desta Tese. Nesse momento gostaria de agradecer a todas as pessoas que direta ou indiretamente contribuíram para que essa relização fosse possível:

Agradeço, a Deus, por me dar forças, para continuar, e nunca desistir dos meus sonhos e objetivos, apesar de todos os obstáculos enfrentados ao longo do Doutorado. Ao meu orientador Fernando Fonseca pela confiança em mim depositada.

A minha esposa Roberta Andrade de Araújo Fagundes, minha parceira, e companheira, que mesmo nos momentos dificieis sempre acreditou, torceu, e compartilhou seus conhecimentos para o sucesso do trabalho.

A meus familiares, em especial a minha mãe Ivoneide Pereira da Silva, que esteve presente, me apoiando, com sua fé, carinho e amor, e apoio a realização desse sonho. As Minhas filhas Rayssa Fagundes Mello e Sarah Fagundes Mello, por entenderem a razão de meu esforço e sempre me apoiando do geitinho delas com carinho amor e muita paciência.

Aos amigos, Maximiliano Carneiro da Cunha (Max), e Sérgio Paiva, pelo companheirismo e conhecimento, que me ajudou a continuar trilhando com perseverança e nunca desistir dos meus objetivos.

E a todos aqueles que, direta ou indiretamente, colaboraram para que este trabalho chegasse a atingir seus objetivos.

“Mesmo desacreditado e ingnorado por todos, não posso desistir, pois para mim vencer é nunca desistir. ”
(EINSTEIN, 1955)

RESUMO

Atualmente, os dados mostram situações alarmantes em relação a problemas da educação, como: niveis baixos de aprendizagem, evasão, reprovação, baixo desempenho em leitura e escrita, entre outros. Nas instituições educacionais, esses problemas são um grande obstáculo na busca pela qualidade na educação. Nesse contexto, é essencial identificar, antecipadamente, quais fatores estão associados a esses problemas. Para isso, utiliza-se técnicas de Mineração de Dados Educacionais (EDM). Essas técnicas são capazes de obter informações e organizar tais informações em conhecimento útil. A EDM requer adaptações de métodos existentes e o desenvolvimento de novas tecnologias. Essa diversidade nos dados representa um potencial para implementação de recursos críticos para auxiliar na melhoria da educação. Partindo dessa necessidade, este trabalho utilizou as teorias do desempenho escolar proposta por Andrade e Soares (2008), e as teorias da evasão proposta por Spady (1970), Vincent Tinto (1975, 1987, 1993), para propor uma abordagem baseada em EDM. Assim, essa abordagem determina a relação dos fatores associados com os problemas educacionais, como também, utiliza-se modelos combinados de regressão (*Ensemble Regression* - ER) para predição da evasão e do desempenho escolar. Mesmo existindo evidências na literatura do uso de diversas técnicas aplicadas a EDM, esses modelos ER reduzem o erro de predição e/ou a variância dos modelos individuais, alcançando melhor acurácia. A metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) foi utilizada e aplicada nesse contexto. Para avaliar a predição dos modelos propostos, um ambiente experimental utilizando bases de dados educacionais reais foi utilizado e o desempenho foi avaliado por meio do erro médio absoluto. Por fim, foi proposta uma abordagem utilizando o diagrama de causa e efeito educacional com base nas teorias: evasão, desempenho escolar e nos resultados dos modelos de predição propostos para o diagnóstico dos problemas educacionais. Assim, essa abordagem serve como uma ferramenta de conhecimento e suporte aos agentes educacionais para a tomada de decisão e o desenvolvimento de estratégias de intervenção educacionais mais eficientes.

Palavras-chave: Mineração de Dados Educacioanis. EDM. Predição. Diagnóstico Educacional.

ABSTRACT

Currently, data show alarming situations related to educational problems, such as: low levels of learning, evasion, failure, low performance in reading and writing, amongst others. In educational institutions, such problems are a big obstacle in the search for the quality of education. In this context, it's essential to identify, beforehand, which factors are associated to those problems. Therefore, educational data mining (EDM) techniques will be used. These techniques are able to obtain and organize useful knowledge. EDM requires adaptations of existing methods and development of new technologies. Such a diversity in data represents a potential for the implementation of critical resources to help improving education. Based on this necessity, this work has used the theories of school performance proposed by Andrade e Soares (2008), and the theories of evasion proposed by Spady (1970) and Vincent Tinto (1975, 1987, 1993), to propose an approach based on EDM. Thus, this approach determines the relationship between factors associated with educational problems, as well as the use of combined regression models (Ensemble Regression -ER) to predict dropout and academic performance. Even though there is evidence in the literature of the use of various techniques applied to EDM, models ER reduce the error and/or variance of the individual models achieving better accuracy. The methodology CRISP-DM (Cross Industry Standard Process for Data Mining) was used and applied in this context. To evaluate the prediction of the proposed models, an experimental environment using real educational databases was used and the performance was assessed through absolute mean error. Finally, an approach was proposed using the cause diagram and the effect educational based on the theories: evasion, academic performance, and in the results of the proposed prediction models for the diagnostic of educational problems. Therefore, this approach serves as a knowledge and support tool for educational agents for decision-making and the development of more effective educational intervention strategies.

Keywords: Educational Data Mining (EDM). Prediction. Educational Diagnostic.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

CRISP-DM	*Cross Industry Standard Process for Data Mining*
DEP-DM	*Diagnosis Educational Problems for Data Mining*
EDM	*Educational Data Mining*
ENADE	Exame Nacional de Desempenho de Estudantes
ENEM	Exame Nacional do Ensino Médio
FGV	Fundaçãom Getulio Vargas
IA	*Artifiicial Inteligence*
IDEB	Indice de Desenvolvimento da Educação Básica
ISEI	*International Socio-Economic Index of Occupational Status*
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira
MEC	Ministério da Educação e Cultura
SAEB	Sistema de Avaliação da Educação Básica

SUMÁRIO

1 INTRODUÇÃO

Este capitulo apresenta a motivação e justificativa do desenvolvimento deste trabalho, como também o método de pesquisa e a estruturação da tese.

1.1 JUSTIFICATIVA E MOTIVAÇÃO

Atualmente, quando o conhecimento se apresenta cada vez mais como variável central para o aumento da produtividade e competitividade, a relevância da educação para o desenvolvimento dos países passa a ser ainda maior. Afinal, em tempos de avanços impressionantes das novas tecnologias, da inteligência artificial e dos processos de automação nos mais diversos setores, especialistas já antecipam que a capacidade dos países em responder a essas demandas por meio de seus sistemas de educação será determinante para o desenvolvimento das nações (TRAJINTENBERG, 2018).

No Brasil, os problemas da educação básica e superior são importantes para a economia do país. Após avanços relevantes nas últimas décadas, os jovens, em sua grande maioria, estão frequentando escolas e universidades. No entanto, da alfabetização ao ensino médio, poucos aprendem em níveis adequados. Os problemas da educação brasileira podem ser classificados em geral como internos e externos. Os externos são: (i) investimentos públicos insuficientes para atender com qualidade as necessidades educacionais; (ii) baixa participação dos pais na vida escolar dos filhos e nos assuntos da escola; (iii) baixa renda familiar; (iv) pouca ou nenhuma escolarização dos país; (v) trabalho informal; e (iv) discriminação por cor ou gênero. Os internos são: (i) elevados índices de reprovação; (ii) professores sem formação mínima; (iii) altas taxas de abandono de alunos devido ao fracasso escolar, ou problemas financeiros; (iv) carência na infraestrutura das escolas; (v) distorção idade-série; (vi) violência em sala de aula; e (vii) baixos níveis de aprendizado.

Um estudo anual do movimento “Todos pela Educação” afirma que no Brasil em 2018, 48,5 milhões de estudantes se matricularam na educação básica. No entanto, os números publicados pelo estudo mostram que 2,46 milhões de crianças e jovens com faixa etária de 6 a 14 anos e 1,7 milhões de jovens entre 15 e 17 anos estão fora da escola. Além disso, o estudo apontou que no geral, os estudantes matriculados em instituições públicas ou privadas apresentam desempenho inferior a metas

estabelecidas pelo PNE – Plano Nacional de Educação (TODOS PELA EDUCAÇÃO, 2018).

Ainda segundo o estudo, o Nível Socioeconômico (NSE) afeta de forma efetiva as chances de aprendizado dos estudantes serem atingidas. Principalmente, aqueles estudantes que vivem em comunidades ou em situações vulneráveis. Para compreender melhor o impacto do índice NSE na educação brasileira: Em 2015, alunos do 9º ano com baixo NSE apresentaram 7,5% de aprendizado adequado em Língua Portuguesa. Já aqueles estudantes com NSE considerado alto atingiram 71,6% em Matemática, esse índice despenca mais, atingindo 2,5% para NSE baixo e 58,2% para NSE alto.

De acordo com dados do Qedu (2016), para o Ensino Médio, a educação brasileira apresenta uma distorção idade-série média de 28%. Isso significa que, de 100 alunos, cerca de 28 estão com até 2 anos a mais do que o esperado para a série em que está matriculado. Essa distorção idade-série, considerada como um fator de desmotivação para o estudo, acarreta taxas significativas de abandono dos estudos. Cerca de 8,6% dos alunos do 1º ano do Ensino Médio saem da escola e 17, 3% reprovam a série (QEDU, 2016).

No Ensino Superior, as realidades são outras, mas os problemas são muito semelhantes aos da educação básica. De acordo com a Pesquisa Nacional por Amostra de Domicílios (PNAD/IBGE), no primeiro semestre de 2017 o número de jovens brasileiros na faixa etária de 18 a 24 anos, que não trabalham e não estudam aumentou para 28%. O fenômeno é complexo e se deve a inúmeros fatores que passam pela crise econômica; pelas escolhas e trajetórias individuais; e pelos contextos em que as pessoas estão inseridas.

Os estudantes que entram na educação superior, e não permanecem, passam a fazer parte do número elevado de estudantes que evadem das instituições de educação superior. Dados do Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (INEP) indicam que em 2010, 11,4% dos estudantes abandonaram o curso para o qual foram admitidos. Em 2014, esse número aumentou chegando a 49%. Estudos realizados ao longo dos últimos anos mostram que as causas da evasão são diversas. Elas são atribuídas a dificuldades financeiras tanto para pagar mensalidades como para manutenção dos estudantes, materiais escolares, moradia e alimentação, dentre outros (AMARAL, 2008); à diversificação e à qualidade dos sistemas, somadas

às características dos cursos e à falta de condições de permanência (CARVALHO, 2006); a insatisfação dos estudantes com professores; a tecnologias inadequadas e à falta de habilidade para o uso de tecnologias (BITTENCOURT; MERCADO, 2014); e, como afirmam Morosin et al (2011), os fatores econômicos não são os únicos responsáveis pelo abandono dos cursos de graduação, os aspectos ligados à vida pessoal, como falsas expectativas e insatisfações por parte dos estudantes em relação aos cursos e às instituições, aliadas as questões associadas ao desempenho, apontando que a decisão de sair ou permanecer na educação superior é tomada pelo próprio estudante (PRATA et al., 2017).

Com relação aos obstáculos acima apresentados, indiscutivelmente os maiores problemas da educação brasileira, tanto na educação fundamental quanto na educação superior são relacionados aos seguintes aspectos: (i) desempenho; e (ii) evasão. Tal constatação surge não só pela gravidade dos números apresentados, mas também ao considerar que os baixos índices de aprendizagem contribuem com os desafios de acesso e permanência (TODOS PELA EDUCAÇÃO, 2018).

Os números apresentados anteriormente referem-se a indicadores oficiais constituídos a partir de avaliações padronizadas realizadas pelo Governo Federal por intermédio do Ministério da Educação (MEC), INEP, e de organismos internacionais. Essas avaliações buscam, ao longo da trajetória escolar, aferir o desempenho acadêmico dos alunos. As principais avaliações realizadas no Brasil são: Avaliação Nacional da Alfabetização (ANA), Sistema de Avaliação da Educação Básica (SAEB), Exame Nacional do Ensino Médio (ENEM), e o Exame Nacional de Desempenho dos Estudantes (ENADE). Os resultados dessas avaliações representam um diagnóstico da educação brasileira, gerando dados que se tornam indicadores de acompanhamento dos alunos ao longo do seu percurso acadêmico. Os indicadores gerados expressam apenas aspectos quantitativos como por exemplo: proporção de alunos com o aprendizado adequado; nível de proficiência dos alunos; rendimento dos alunos concluintes, dentre outros (QEDU, 2018).

Entretanto, a tarefa de utilização e extração de conhecimento dessas bases apresenta desafios técnicos e metodológicos para sua utilização efetiva. Mesmo com essas dificuldades, a relevância dessas informações propiciadas por esses dados, faz com que a utilização dos mesmos represente um poderoso subsídio para elaboração de pesquisas acadêmicas e o aprimoramento das políticas públicas (SILVA,2007).

Em particular, o exame mais detalhado desses tipos de dados oferece uma oportunidade para que os educadores identifiquem rapidamente os problemas, no sentido de apoiar sua metodologia de ensino ao longo do curso. Educadores podem determinar os indicadores que mostram a satisfação do aluno e engajamento no curso, como também monitorar o progresso da aprendizagem (CARD et al. 1999).

Os gestores podem se valer de dados educacionais para traçar metas de desempenho institucionais, identificar fragilidades e utilizar os recursos disponíveis de forma mais precisa e fundamentada em informações concretas, ou seja, avaliar o desempenho da instituição educacional por eles administrada.

Diante desse cenário, diagnosticar de forma eficiente os fatores (causas) e relacioná-los aos problemas (efeitos), a partir de dados educacionais, torna-se um desafio para esta pesquisa. Outro desafio, está em aplicar abordagens\técnicas inteligentes para identificar, analisar e mensurar as causas dos problemas a partir das informações presentes nas bases de dados educacionais.

Portanto, se faz necessária, dentro do contexto de Mineração de Dados Educacionais (MDE, do inglês, Educational Data Mining, ou EDM), a aplicação de técnicas de Aprendizagem de Máquina (AM). A AM é muitas vezes confundida com a Mineração de Dados (MD), já que ambas compartilham conceitos e muitas vezes são usadas juntas. É comum ver essas áreas correlacionadas no desenvolvimento de modelos que facilitem a descoberta de novos padrões em conjuntos de dados, visto que a área de AM fornece muitos de seus conceitos e técnicas para a área de MD (HAN e KAMBER, 2011).

Dentre as técnicas oriundas de AM utilizadas na descoberta de conhecimento em conjuntos de dados, destacam-se os modelos de regressão, os quais analisam os relacionamentos entre as variáveis. Mais especificamente, os algoritmos de regressão estimam o valor de uma variável numérica dependente (y) que faz uso de uma ou mais variáveis independentes (x) (MONTGOMERY et al. 2012).

Ou seja, a função pode usar uma variável ou variáveis múltiplas para explicar a previsão da variável de saída. Além disso, esse procedimento de mapeamento exige que uma função de perda seja minimizada sobre a distribuição conjunta de todos os valores (y e x) (ÖZÇIFT, 2014).

Esse tipo de técnica, quando empregada ao contexto de EDM, poderá estimar os relacionamentos existentes entre as variáveis de causa (fatores) com as variáveis

de efeito (aprendizagem, desempenho e evasão) formando um modelo matemático que descreve a curva de aproximação dos dados.

Uma vez que os problemas científicos podem ser modelados como problemas de regressão, existe uma necessidade contínua de algoritmos precisos em cada cenário estudado (ÖZÇIFT, 2014). Dessa forma, a abordagem de conjuntos ou aprendizagem de combinação de modelos (do inglês, ensemble learning), vem sendo amplamente utilizada para aumentar a acurácia de modelos preditivos.

Diferentes modelos combinados podem ser gerados para diferentes modelos preditivos, sendo considerados, nesta tese, problemas de regressão. Uma metodologia de desenvolvimento facilitaria a sua aplicação, pois na própria literatura não existe uma definição da melhor forma de desenvolver a abordagem de conjunto em EDM. Portanto, uma das motivações para o desenvolvimento desta tese é a formação de uma abordagem para construção de conjuntos no contexto de EDM. Outra motivação é a possibilidade de utilizar as informações oriundas das avaliações padronizadas para desenvolver indicadores e instrumentos de previsões futuras, com a capacidade de auxiliar na compreensão implícita de informações sobre tendências futuras da aprendizagem e do desempenho dos estudantes.

1.2 QUESTÕES DE PESQUISA

Nesse sentido, emerge a questão central de pesquisa desta tese, que busca responder: É possível desenvolver um modelo preditivo no contexto da EDM para o diagnóstico dos problemas educacionais? As questões abaixo norteiam a questão central desta tese:

- Quais fatores melhor descrevem as causas dos problemas educacionais?
- Quais métodos e técnicas computacionais vêm sendo utilizados para predição dos problemas educacionais?
- Como diagnosticar de forma eficiente as causas dos problemas educacionais a partir da combinação de modelos de regressão?
- Como estabelecer uma relação entre os resultados encontrados e os fatores apontados na literatura, no que diz respeito aos problemas educacionais?
- Quais as relações de causa e efeito diagnosticadas nos dados educacionais analisados? e

- De que forma o conhecimento obtido a partir do diagnóstico poderá proporcionar intervenções por parte de professores e gestores nas causas dos problemas educacionais?

1.3 OBJETIVOS

Nesta seção são apresentados os objetivos geral e específicos que este trabalho procurou alcançar.

1.3.1 Geral

O objetivo geral desta tese é desenvolver uma modelagem preditiva para o diagnóstico do problema educacional por meio dos modelos teóricos do problema e métodos e técnicas de Mineração de Dados Educacionais.

1.3.2 Específicos

- Determinar as principais teorias e abordagens relacionadas aos fatores relacionados ao desesempenho educacional e a evasão, bem como, as principais abordagens/técnicas de EDM aplicadas a predição desses problemas no contexto educacional;
- Aplicar técnicas de EDM para extrair, analisar e mensurar as informações educacionais presentes em bases de dados educacionais;
- Construir modelos combinando *Ensemble* e Métodos de Regressão para obtenção de resultados que indiquem os fatores mais representativos do problema educacional;
- Desenvolver um ambiente experimental para avaliação do modelo proposto, utilizando bases de dados reais;
- Identificar as relações de causa e efeito existentes entre os fatores identificados e o problema a partir das abordagens teóricas sobre o problema no contexto educacional; e
- Propor mecanismos para o diagnóstico do problema educacional, a partir da identificação visual e entendimento das relações de causa e efeito dos fatores com o problema, como subsísdios para a melhoria da prática docente e promoção de estratégias de intervenção para minimizar a ocorrência do problema.

1.4 MÉTODO DE PESQUISA

O método utilizado nesta tese foi fundamentado no processo CRISP-DM (*Cross Industry Standard Process for Data Mining*). O processo CRISP-DM é descrito de forma hierárquica, composto por seis fases: entendimento do negócio/domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação (CRISP-DM, 1999).

A fase de **entendimento do domínio** trata da compreensão dos objetivos do projeto de mineração a partir da perspectiva do negócio. A partir desse conhecimento surgirá um problema de mineração de dados. Essa etapa é considerada uma das mais importantes do processo, a partir da qual é desenvolvido um plano preliminar do projeto de mineração em busca de atingir os objetivos (SHEARER, 2000).

A fase de **entendimento dos dados** começa com uma coleta inicial dos dados, visando estabelecer uma familiaridade com esses dados, identificar possíveis problemas de qualidade dos dados, descobrir idéias iniciais ou para detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas. As atividades realizadas nessa etapa são importantes para evitar possíveis problemas na etapa seguinte (preparação dos dados) (IBM,2013).

A fase de **preparação dos dados** é a etapa que abrange todas as atividades necessárias para construir o conjunto de dados final, a partir dos dados brutos iniciais. Esses dados preparados alimentarão a ferramenta de modelagem na etapa seguinte. Tarefas de preparação de dados são sucetíveis de serem excutadas repetidas vezes sem qualquer ordem pré-estabelecida. Essas tarefas incluem a seleção das tabelas (databases relacionais), registros, atributos, assim como a transformação e a limpeza dos dados (CHAPMAN et al., 2000; SHEARER, 2000).

Após a etapa de pré-processamento inicia-se a etapa de **modelagem**, na qual de acordo com o problema de mineração, várias técnicas podem ser usadas. A modelagem é geralmente executada em várias interações, nas quais os analistas de dados executam vários modelos, usando as configurações padrão e vão ajustando os parâmetros para valores otimizados. É comum também retornar para a fase de preparação de dados para manipulações exigidas pelo modelo (SHEARER, 2010; IBM, 2013).

Antes de implementar definitivamente o modelo construído é necessário avalia-lo, para ter a certeza de que ele atinge em sua completude e adequadamente os objetivos planejados. A etapa de **avaliação** garante que a organização possa utilizar os resultados obtidos e os conhecimentos descobertos. É de fundamental importância determinar se há algum problema importante do negócio que não foi considerado (CHAPMAN et al., 2000; IBM, 2013).

Por fim, a ultima etapa é a **implantação**, na qual os novos conhecimentos descobertos são usados para proporcionar melhorias na organização. Nessa etapa, todo o conhecimento adquirido deve ser organizado e apresentado de uma forma que o cliente possa usá-lo efetivamente dentro dos processos de tomada de decisão. Esse pode ser um processo simples como um relatório, um indicador ou complexo com a implementação de um *software* ou de um processo de mineração de dados aplicado a toda organização (SHEARER, 2000; CHAPMAN et al., 2000).

Todos os instrumentos utilizados em cada fase descrita acima foram detalhados no Capitulo 5 desta tese.

1.5 ORGANIZAÇÃO E ESTRUTURA DA TESE

Esta tese está dividida em sete capítulos. Além deste, o Capitulo 2 descreve as abordagens conceituais sobre os problemas educacionais que dão suporte ao trabalho, descrevendo as principais teorias e modelos conceituais sobre os problemas educacionais abordados nesta tese, existentes na literatura.

O Capitulo 3 descreve a revisão bibliográfica relacionada as áreas de Mineração de Dados e Mineração de Dados Educacionais, mostrando as principais abordagens e técnicas utilizadas no âmbito da educação.

O Capitulo 4 descreve, o mapeamento sistemático da literatura sobre as principais abordagens/técnicas de EDM para predição de problemas nos diversos contextos educacionais, buscando identificar lacunas e oportunidades que justifiquem o desenvolvimento desta tese.

No Capitulo 5 são descritas a proposta de método utilizada, levando em consideração a caracterização do problema, entendimento sobre os dados, os participantes envolvidos, o processo de mineração de dados e por fim a apresentação dos resultados.

No Capitulo 6 são apresentados os resultados alcançados em todas as fases do método proposto.

O Capítulo 7 discorre sobre as considerações finais, contribuições, limitações, trabalhos futuros e publicações cientificas.

Por fim, são apresentadas as referências utilizadas neste trabalho.

2 ABORDAGENS CONCEITUAIS DOS PROBLEMAS EDUCACIONAIS

Neste capítulo, é iniciada a fundamentação conceitual, na qual são apresentados os conceitos fundamentais. Este capítulo apresenta algumas teorias e modelos existentes sobre os problemas educacionais elencados nesta tese: o desempenho educacional e a evasão escolar, como também, aspectos relevantes, que fornecem a base para posicionar o trabalho no campo da pesquisa.

2.1 ABORDAGENS TEÓRICAS SOBRE O DESEMPENHO EDUCACIONAL

O desempenho educacional pode ser entendido como a capacidade que os alunos têm de expressar a sua aprendizagem e seu conhecimento adquirido no processo ensino-aprendizagem (PERRENOUD, 2003). Esse infere nas habilidades acdêmicas dos alunos e tem caráter avaliativo na medida em que os estudantes devem demonstrar em suas respostas em testes e provas, por exemplo, o que aprenderam nas aulas. Segundo Miranda et al (2013), o desempenho educacional é consequência de diversos fatores tais como características do corpo docente e dos próprios estudantes, estrutura da instituição de ensino e organização do tempo.

O bom desempenho educacional, segundo Borkowiski (1992), se refere ao fato do aluno realizar as tarefas e atividades escolares de forma eficaz, atingindo seu objetivo final que é o aprendizado. Contudo, é sabido que muitos alunos das diferentes etapas de ensino apresentam um desempenho escolar muito aquém do ideal. Quando se reflete sobre o desempenho escolar, há que se levar em consideração diferentes fatores que se correlacionam e influenciam diretamente o desempenho educacional. Dentre os fatores mais relevantes estão as características relacionadas a escola (tamanho das salas de aula, experiência do professor), a família (expectativas dos pais em relação ao trabalho da escola, participação dos pais nas atividades escolares dos filhos) e ao aluno (adaptação ao ambiente escolar, relações com o trabalho, motivação do aluno, por exemplo).

2.1.1 Estruturas Sociais que Influenciam o Desempenho do Estudante

As discussões sobre as diferenças sociais, embora concebidas no arcabouço da sociologia clássica, se consolidaram nos estudos sociológicos americanos a partir dos anos de 1950. Isso se deu em virtude da necessidade de aferir empiricamente sobre

características populacionais, ocupacionais e socioeconômicas da sociedade contemporânea, o que também ocorreu no Brasil algumas décadas depois (ALVES; SOARES, 2012).

No campo educacional, diversos autores vêm demonstrando interesse em discutir sobre os fatores que influenciam no desempenho cognitivo dos estudantes e nas suas relações com o ambiente interno e externo da escola que, sobremaneira, interferem nos processos de ensino e aprendizagem. Assim, as investigações sobre a influência do ambiente familiar e das variáveis contextuais no aprendizado escolar, das décadas de 1950 e 1960, indicam que as famílias podem influenciar positivamente no desempenho dos estudantes, na motivação para os estudos, no desenvolvimento de competências interpessoais e na boa convivência entre alunos, professores e colegas (MARTURANO, 2006, *apud* FERREIRA e BARREIRA, 2010).

Sobre o contexto educacional, Brooke e Soares (2008) também destacam o pioneirismo americano nos estudos sobre a eficácia escolar e a grande preocupação do país com a qualidade e condições de oportunidades educacionais, desde 1960. Entretanto, Soares (2004) aponta como essencial que as intervenções reforcem a importância social da escola para o aprendizado dos conteúdos cognitivos necessários à formação da consciência ativa dos estudantes na sociedade.

Segundo Brooke e Soares (2008), o marco inicial do objeto de pesquisa sobre o desempenho do estudante é o Relatório Coleman (1966) que diz respeito ao Estudo da Igualdade de Oportunidade Educacional. Este estudo foi realizado em resposta as disposições da Lei de Direitos Civis de 1964 dos Estados Unidos da América e teve como objetivo estudar as causas para as diferenças de desempenho entre as escolas norte-americanas. A pesquisa indagou sobre cor, raça, religião e nacionalidade e ainda aspectos do nível socioeconômico. Foram avaliados cerca de 570 mil estudantes e 60 mil professores e as instalações de aproximadamente quatro mil escolas.

Ainda, segundo os autores, o Relatório Coleman define que as diferenças de infraestrutura, a localização das escolas, e a qualidade dos professores não justificavam a diferença de desempenho dos estudantes e que o principal motivo se dava pelas variáveis socioeconômicas (fatores contextuais) e não pelas diferenças entre as escolas (fatores individuais). Pesquisas realizadas na Inglaterra relatadas no Relatório Plowden de 1967 e na França por meio do estudo longitudinal de 1962 a

1972, se mostraram compatíveis com o resultado do Relatório Coleman (BONAMINO e FRANCO 1999).

Os estudos de Bonamino e Franco (1999), Soares (2004), Brooke e Soares (2008), Soares (2007), e Alves e Soares (2013), dentre outros, possibilitaram a compreensão de que o fator socioeconômico, desde o Relatório Coleman, é reconhecido como o que causa maior influência nos resultados de desempenho de estudantes e não se resolve em curto prazo, mas são necessárias políticas públicas de desenvolvimento econômico para diminuir as diferenças entre os estudantes, e consequentemente, obter melhora no seu desempenho. Os estudos apontam ainda que a escola contribui com políticas de qualidade e melhoria dos resultados educacionais, mas que não surtem efeitos em curto prazo; nas condições socioeconômicas, as atitudes positivas de professores geram bons resultados, mas, sozinhas não alcançam os objetivos esperados; a família por sua vez, influencia por meio de hábitos de estudos e incentivos do estímulo e da manutenção de expectativas educacionais (SOARES, 2004).

Os fatores que melhoram o desempenho dos estudantes, segundo Andrade (2007), são: comparação do aluno com os colegas; recursos culturais de que o aluno dispõe em casa, como acesso a computadores com *Internet* livros, revistas de informação geral, jornais e se o aluno gosta de estudar a disiciplina além de o aluno gosta de fazer o dever de casa.

Albenaz (2002) ressalta que os fatores são considerados eficazes quando proporcionam a equidade, reduzindo a diferença do nível socioeconômico dos estudantes, ou seja, a escola eficaz e equânime é aquela que apresenta um desempenho elevado dos estudantes, independentemente do nível social (ALBERNAZ, 2002).

Clifton e Cook (2013) afirmam que o desenvolvimento educacional é influenciado por vários fatores, os quais incluem as características individuais, a família, o bairro onde vivem e as escolas. Muitas variáveis (fatores) que determinam o baixo desempenho estão fora do controle direto das escolas, contudo, a escola pode ser parte da solução desses problemas. As escolas que atendem comunidades menos favorecidas mostram que a alta qualidade na educação pode ajudar a transformar a vida dos estudantes e compensar as diferenças. Por isso, pode-se afirmar que as escolas podem reduzir as desigualdades por meio da educação, mesmo perante a

existência de uma pobreza social e da pouca escolaridade das famílias. Torna-se fundamental que a escola ofereça ao estudante oportunidades em igualdade para o sucesso do desempenho, independentemente da sua condição social ou familiar, além das características pessoais de cada estudante (raça, sexo, trajetória escolar, saúde) que são determinadas pela sua trajetória de vida (SOARES, 2007).

2.1.2 Modelo Conceitual e Explicativo dos Fatores Associados ao Desempenho dos Estudantes

Para o entendimento completo do desempenho do aluno é necessária uma abordagem multidisciplinar que agregue conhecimentos pelo menos da psicologia, da educação, da sociologia, da economia e inclusive da ciência política, em muitos momentos subsidiados pela coleta e análise de dados por meio de técnicas estatísticas apropriadas (SOARES, 2007).

Dessa forma, apesar das pesquisas apontarem o fator socioeconômico como o mais influente no desempenho do estudante, não se pode afirmar que esse é um fator isolado, mas que a combinação de fatores, tanto internos como externos, podem levar o estudante ao sucesso ou fracasso escolar. São vários os fatores que interferem na aprendizagem dos estudantes: o estudante, sua família, a escola e a rede, ou sistema a que está associado, e finalmente a sociedade em geral (ANDRADE e SOARES, 2008). Os autores apresentam, ainda, o modelo conceitual de inter-relações entre os fatores associados da aprendizagem na Figura 1.

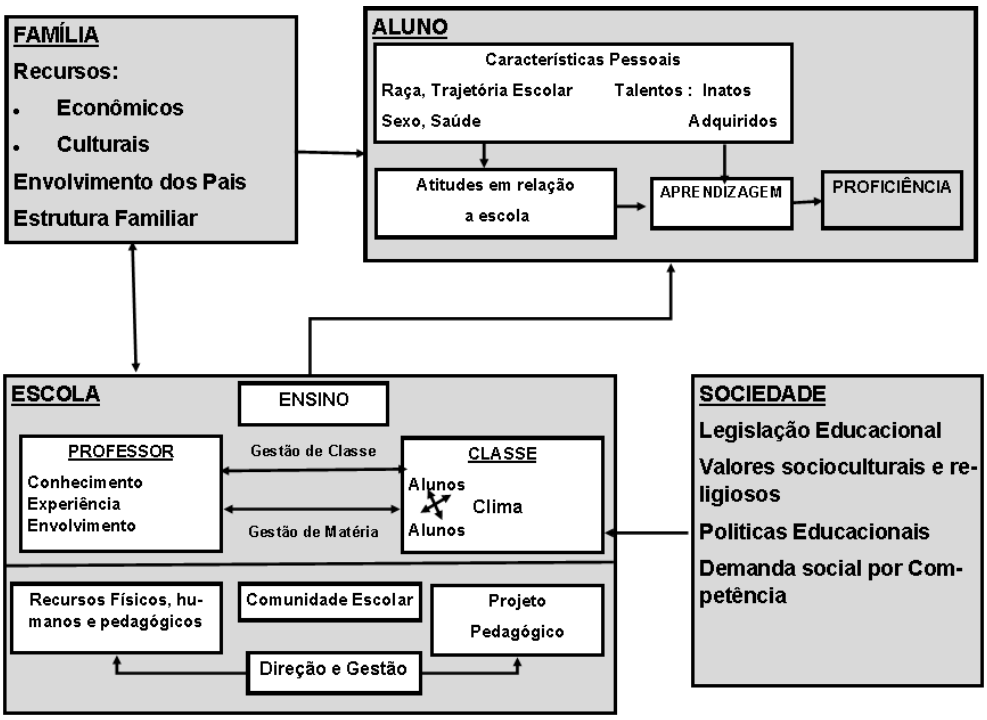


Figura 1. Modelo Conceitual dos fatores associados ao desempenho dos alunos
Fonte: Andrade e Soares (2008)

O modelo explicativo da Figura 1 mostra a interdependência entre as diversas dimensões que interferem na proficiência do estudante quando considera que a sociedade é a dimensão que influencia todas as demais dimensões, pois inclui toda a Legislação Educacional, as demandas sociais por competências, os valores socioculturais e religiosos e as Políticas Educacionais. Alinhado à dimensão social, tem-se a Escola com suas diretrizes, suas políticas de formação de professores e suas rotinas administrativas. Ela por sua vez, interfere diretamente no desenvolvimento dos estudantes, os quais também, são afetadas pelos demais estudantes que a frequentam. E, finalmente, os estudantes que têm o desempenho cognitivo influenciado por todas as dimensões, cujas características individuais, sociais e econômicas têm uma relativa importância (PALERMO et al., 2013).

O trabalho de Palermo et al. (2013) buscou compreender os diferentes fatores que influenciam o funcionamento das escolas da rede pública de ensino brasileira, bem como a sua qualidade, a partir da identificação de variáveis que impactam os resultados escolares, em particular o desempenho dos alunos medido pelos sistemas de avaliação utilizados pelo INEP/MEC. Os fatores que influenciam a proficiência são múltiplos e complexos, contemplando dinâmicas que atuam em diferentes níveis, desde o mais elementar, das características socioeconômicas e culturais dos

indivíduos e de suas famílias, até as dinâmicas que ocorrem nas salas de aula, entre professores e alunos, e as características estruturais da escola. Alunos e seus responsáveis, diretores e professores das instituições de ensino são alguns dos atores relevantes que agem sobre o processo latente gerador de maiores ou menores rendimentos escolares. Mais especificamente, o objetivo do estudo foi analisar os fatores que influenciam o desempenho escolar dos alunos do 5º ano do ensino fundamental, nas escolas públicas municipais da cidade do Rio de Janeiro. Para tanto, foram utilizados modelos hierárquicos com três níveis, os quais permitiram avaliar os efeitos de variáveis socioeconômicas e culturais e dos contextos familiares dos alunos, as práticas pedagógicas e estilos de ensinar dos docentes e, ainda, de políticas educacionais, aspectos da gestão e características das escolas. O estudo teve como fonte principal as informações provenientes da Prova Brasil 2007 (INEP, 2018).

As características individuais ou estruturais referem-se aos fatores de gênero e raça que não mudam de um ano para outro, são considerados fixos. Ainda, como características individuais têm-se os fatores de fluxo, os quais são questões relacionadas ao processo escolar presente e passado do estudante. O fator família está relacionado com a questão de residência (bens), escolaridade dos pais, relacionamento destes com a escola, além de influências diárias, como incentivo à leitura. Os fatores intraescolares são mutáveis de um ano para o outro, pois o corpo docente e a direção podem mudar, ou se aprimorar, utilizando os recursos de forma diferenciada (MACEDO, 2000).

No estudo sobre os fatores associados ao desempenho escolar de estudantes da quinta série do ensino fundamental, Castro (2010) expõe que os órgãos ligados a educação nas esferas federal, estadual e municipal no Brasil, implantam sistemas de avaliação não classificatória em larga escala para aferir os saberes dos estudantes ao final de determinadas séries. Os indicadores dessas avaliações apontaram que o aprendizado é influenciado por características individuais, ambientais, socioeconômicos, aliadas aos fatores de idade, etnia, classe social e condições de moradia.

O autor expõe com base nos resultados da sua pesquisa, assim como Macedo (2000) que o sexo feminino se beneficia mais em Língua Portuguesa e o masculino em Matemática. A partir do efeito do coeficiente de valor adicionado, ser menina para

o rendimento em português é quase o dobro do efeito de ser menino no rendimento de matemática. Outra conclusão diz respeito ao melhor desempenho dos estudantes provenientes de camadas sociais economicamente favorecidas. Castro (2010) ainda expõe que os estudantes brasileiros apresentam baixo rendimento nas avaliações do Saeb e Pisa, em comparação com países mais desenvolvidos. Isto ocorre em decorrência do alto percentual de estudantes sem habilidades mínimas compatíveis com sua escolaridade. Para a autora, a melhoria do resultado nas avaliações junta-se com a melhoria das condições de vida da população, além disso, o processo educacional deve ser capaz de reconhecer e respeitar as diferenças de forma a adaptar-se à realidade dos estudantes. Contudo, não pode ser deixado de lado o conjunto mínimo de habilidades para a escolaridade.

Soares (2005) afirma que “a proficiência escolar é um atributo que tem gênero, cor e é distribuído de forma desigual entre as regiões do país e entre as redes de ensino”. O estudo de Soares, assim como o de Macedo (2000), conclui que o sexo feminino tem tendência a obter rendimento inferior em matemática em relação ao sexo masculino, assim como os negros em relação aos brancos e os estudantes de escolas públicas em relação a escolas privadas. É possível diminuir a desigualdade de proficiência por meio de um ambiente intraescolar equânime que dê igualdade de condições aos estudantes, independentemente da sua classe social.

Ainda, de acordo com Soares et al. (2012), o fator social coletivo impacta mais no desempenho do estudante que fatores individuais, e que, o estudante que convive com colegas de alta condição social é privilegiado em relação aos outros, pois desfruta de vantagens criadas pelo contexto e pela convivência com esses estudantes de melhor condição social e cognitiva.

Andrade e Laros (2007) afirmam que o desempenho escolar é resultado de múltiplas interações aliadas aos fatores pessoais, de ensino, instalações e ambiente, o que exigem a utilização de instrumentos de modelagem complexos e comparativos, mas nem sempre possíveis e fidedignos à realidade. Em se tratando de pesquisa sobre o desempenho escolar, devem-se primeiramente considerar o conhecimento prévio e o nível socioeconômico familiar dos estudantes uma vez que tais fatores influenciam no seu desempenho.

Dentro das interações com ambiente tem-se o clima escolar que é a percepção que os estudantes têm sobre o ambiente que os rodeia. Segundo Cornejo e Redondo

(2001) e Brito e Costa (2010), clima escolar é a forma que o indivíduo percebe coletivamente a “atmosfera” e traz significativas influências sobre o comportamento dos grupos. “Uma escola na qual as relações entre os diferentes grupos membros da comunidade educacional são positivas favorecem um bom clima de trabalho e serão obtidos resultados favoráveis no processo pedagógico”, afirmam (Brito e Costa, 2010).

O estudo de Cornejo e Redondo (2001), sobre clima escolar na região metropolitana de Santiago no Chile, define algumas variáveis para o estudo, dentre elas está o conceito de ambiente escolar, o qual se traduz na percepção dos atores educativos sobre as relações interpessoais estabelecidas na escola e ainda mede a percepção dos estudantes sobre as relações estabelecidas com seus professores a respeito de vários contextos ambientais interrelacionados. Passador e Calhado (2012) citam que a existência de fatores intangíveis capazes de influenciar o desempenho das escolas, e consequentemente a qualidade da educação. Segundo Ala-Harja et al (2000), a ênfase nos resultados constitui elemento central das recentes reformas do setor público. A avaliação é uma ferramenta que visa oferecer informações quanto aos resultados obtidos por organizações e programas.

No Brasil, pesquisadores como Castro (2010), Rocha e Perosa (2008) e Andrade e Laros (2007), entre outros, apontam a avaliação como um processo que pode trazer benefícios para a compreensão da realidade atual, enquanto elemento que se insere no ambiente social e escolar.

A literatura brasileira apresenta os fatores que proporcionam equidade dentro das escolas, ou seja, equidade intraescolar. Mesmo que estas recebam estudantes, com acentuada desigualdade socioeconômica, a equidade pode ser trabalhada sob as perspectivas destes fatores que compreendem cinco grupos, os recursos escolares; a organização e gestão da escola; o clima acadêmico; a formação e salário dos docentes; e o enfoque pedagógico. Com relação aos recursos escolares, diversos estudos realizados no Brasil apontam que as diferenças no desempenho dos estudantes estão relacionadas a falta de equipamentos, recursos financeiros, aos aspectos de liderança do diretor e a características associadas à eficácia escolar. O clima escolar se relaciona com o absenteísmo dos docentes. Há uma estreita relação de desempenho dos estudantes com a frequência dos docentes, e ainda, a ênfase em passar e corrigir tarefas de casa. Para a formação e salário dos docentes não se encontrou grande relação entre o resultado dos estudantes com essa variável. Já para

a ênfase pedagógica nota-se efeito positivo após o movimento de renovação do ensino de Matemática (FRANCO et al. 2007).

A sociedade brasileira espera que a escola de educação básica garanta aos estudantes aprendizado das competências necessárias para uma inserção crítica e produtiva na sociedade (ALVES e SOARES 2012). Sabe-se, desde pesquisas realizadas na década de 70, que o contexto escolar possui influência significativa no comportamento e no desempenho cognitivo dos estudantes (BROOKOVER et al., 1979). Por meio das pesquisas conclui-se que retirando a diferença provocada pelos fatores individuais que não podem variar ao longo do tempo (gênero, raça), a escola é capaz, empregando ações que promovam a equidade, diminuir as diferenças de proficiência entre os estudantes, mesmo que as condições socioeconômicas familiares sejam acentuadas.

Não se pode negar que nas três últimas décadas, desde a promulgação da Constituição Federal de 1988, a universalização da Educação Básica é presente no país. Contudo, “educação para todos” não significa qualidade na educação. A conquista das classes populares pelo direito à educação é notável, mesmo que seja no nível fundamental de ensino. O Brasil atende, em todo território nacional, a população em idade escolar, contudo, o desafio agora é o baixo desempenho escolar por parte de alguns seguimentos da sociedade (SOARES, 2005).

A escola pública brasileira atende ao universo da população em idade escolar e é realidade presente em todo o território nacional. No entanto, é importante refletir sobre o fato de o sistema educacional brasileiro ter substituído a exclusão da escola pelo fracasso escolar para alguns seguimentos da sociedade (SOARES, 2013).

Segundo Machado (2014), a renda familiar é a primeira característica que deve ser considerada em estudos da influência da família no desempenho do aluno.

Para Alves e Soares (2012), o nível socioeconômico é um construto teórico que sintetiza as características dos indivíduos em ralação a sua renda, ocupação e escolaridade, permitindo a criação de estratos ou classe de indivíduos semelhantes em relação a estas características. Ainda, segundo os pesquisadores, não há uma definição única na literatura sobre o que é construto (nesta pesquisa construto será entendido com o que representam atores, práticas e processos capazes de afetar o desempenho dos estudantes).

Alves e Soares (2012) destacam que não há consenso sobre quais dimensões devem ser consideradas nas pesquisas, e que as decisões do que devem fazer parte da construção de indicadores dependem da justificativa teórica e da disponibilidade de dados. Mas destacam o modelo do PISA (*Programme for International Student Assessment)*, o qual inspira as avaliações de vários países do mundo, o qual desenvolveu um indicador que usa dados de ocupação dos pais dos alunos (GAZENBOOM; DE GRAAF; TRAIMAN, 1992; GAZENBOOM, 2010). Esse indicador, chamado de ISEI (*International Socio-Economic Index of Occupational Status*), é estimado a partir de informações sobre a ocupação dos pais dos alunos, a qual é atribuída um escore escalonado de acordo com a educação e a renda da família medida de forma indireta (posse de bens). O ISEI é utilizado para contextualizar os resultados comparativos entre países, exatamente para evitar análises enviesadas sem o controle do NSE (Nível Sócio Econômico) dos alunos que fazem o teste e as diferenças locais. Outra referência interessante sobre o uso da ocupação dos pais no índice de NSE vem da Austrália, onde eleé utilizado para contextualizar a comparação dos resultados educacionais entre as diferentes regiões do país (MARKS et al., 2000).

De acordo com Franco et al (2003), o apoio familiar é um dos principais fatores do processo de aprendizagem do aluno. Estudos realizados sobre o grau de instrução dos pais, e a sua influencia no desempenho escolar dos filhos, constatam que filhos de pais analfabetos ou que não terminaram o ensino fundamental têm maior de chance ter baixo desempenho escolar quando comparados a filhos de pais com curso superior completo. Segundo os pesquisadores, a explicação para a essa influência está no estímulo que as crianças recebem dentro de casa. Não só o grau de instrução, mas também o incentivo à leitura, bem como a disponibilidade de itens de leitura, os quais interferem no desempenho (FRANCO et al, 2003).

O contexto socioeconômico no qual o estudante está inserido é a variável explicativa com maior coeficiente de fidedignidade, alcançando 0,87 de correlação entre o nível socioeconômico e o desempenho escolar (ANDRADE, LAROS, 2007). Mesmo que os fatores socioeconômicos tenham influência no desempenho dos estudantes, “aumentar os níveis de proficiência e diminuir o impacto da posição social no sucesso escolar devem ser os principais objetivos de qualquer sistema educacional” (SOARES, 2007).

Buchman e Hannum (2001) consideram a educação como fator decisivo tanto na reprodução das desigualdades existentes quanto na possibilidade de mobilidade social. As autoras afirmam que as diferenças de desempenho educacional entre os indivíduos podem ser explicadas pela interação entre os chamados fatores da oferta e da demanda. Os fatores da oferta são as oportunidades educacionais disponiveis, enquanto os fatores de demanda dizem respeito às decisões familiares quanto a educação, processo diretamente ligado às características sócio-econômicas e estruturais da família.

As autoras afirmam ainda que os determinantes do desempenho escolar, e consequentemente as desigualdades educacionais entre os indivíduos, dependem da ação conjunta de variáveis micro, como a escolaridade dos pais, a renda familiar e a composição do domicilio, e macro, como os insumos físicos disponíveis na escola, as características dos professores e, em uma esfera mais geral, as politicas públicas voltadas para educação.

Para Hanushek (2002), a importância da análise da relação entre fatores escolares e desempenho é dada principalmente pelo fato de que estes fatores são mais propensos à elaboração de políticas públicas. A atuação governamental no processo educacional, sua contribuição para que os resultados deste processo sejam satisfatórios em termos de qualidade e equidade, é mais facilmente viabilizada na provisão e regulação do sistema de ensino que na mudança das características familiares, principalmente no que tange à escolaridade dos pais.

A mensuração dos efeitos dos fatores de oferta e demanda sobre o desempenho educacional permite reconhecer as causas da estratificação educacional e identificar as características passíveis à ação de políticas públicas eficientes. É recorrente na literatura a utilização da função de produção educacional para este fim, pois permite relacionar os fatores listados anteriormente a uma variável resposta indicadora de desempenho ou eficiência educacional.

Hanushek (2002) propõe que a análise estatística dos determinantes do desempenho escolar seja feita em termos marginais. Assim, é possível delinear com maior clareza os fatores que afetam o processo de aprendizado, entendido como o ganho de conhecimento ou habilidade entre dois pontos, e não a proficiência adquirida pelo aluno até então. Este enfoque, conhecido como valor adicionado, possibilita visualizar o papel dos determinantes em dado horizonte temporal, conforme sua

capacidade de atuar no processo corrente. Caso a análise fosse feita levando-se em conta a nota em um teste apenas, esta refletiria a atuação dos insumos de forma atemporal, o que dificultaria a compreensão de como e em que ponto da formação educacional se dá sua intervenção.

2.1.3 Abordagens de Mensuração do Desempenho Escolar

Os resultados educacionais podem ser analisados por meio de variáveis associadas a dois tipos básicos de enfoque na educação formal: quantidade e qualidade. O enfoque da quantidade está pautado nos trabalhos cujo objetivo é analisar a quantidade de anos de estudos acumulados pelos indivíduos e variáveis relacionadas, como matrícula, a repetência, a evasão e o desempenho escolar (aprovação e reprovação). Já o enfoque da qualidade está normalmente relacionado ao estudo dos efeitos familiares e escolares sobre o rendimento dos alunos obtidos em avaliações padronizadas. É importante atentar para a definição da questão da qualidade, pois o uso de testes padronizados constitui apenas uma das possíveis nuances analíticas assim classificadas. Entre outras abordagens sobre qualidade na educação, deve-se ressaltar aquelas que se referem a aspectos não mensuráveis do processo de aprendizado, como por exemplo a qualidade dos insumos educacionais ofertados. Nos tópicos a seguir serão apresentadas as principais abordagens quantitativas e qualitativas aplicadas a dados educacionais (MACEDO, 2004).

Dentre as abordagens mais utilizadas para mensuração dos resultados educacionais no Brasil, a partir de dados de avaliações em larga escala, destacam-se: o Indice de Desenvolvimento da Educação Básica – IDEB (INEP, 2007); a função de produção educacional (TODD e WOLPIN, 2003); e a teoria de resposta ao item – TRI (ANDRADE e KLEIN, 1999).

O IDEB é calculado a partir dos dados sobre aprovação escolar, obtidos no Censo Escolar realizado todos os anos, e médias de desempenho nas avaliações do INEP. Ele sintetiza em um único indicador dois conceitos importantes para aferir a qualidade do ensino no país: (i) Fluxo - representa a taxa de aprovação dos alunos; e (ii) Aprendizado - correspondente ao resultado dos estudantes no Sistema de Avaliação da Educação Básica – SAEB (INEP,2007).

A função de produção educacional, utiliza os mesmos conceitos da função de produção na economia. Ela examina a relação de produtividade entre os insumos e o

produto final, neste caso os insumos que afetam o desempenho escolar e a proficiência do aluno, respectivamente. Na literatura acerca dos determinantes do desempenho escolar, os pesquisadores utilizam essa analogia com o objetivo de entender a tecnologia de combinar os insumos escolares e familiares de forma que o resultado educacional seja melhorado (TODD e WOLPIN, 2003).

A TRI vem ao longo dos anos sendo umas das abordagens mais aplicadas em várias áreas de conhecimento, em particular na avaliação educacional. Essa abordagem propõe modelos para os traços latentes, ou seja, características do individuo que não podem ser observadas diretamente. Esse tipo de variável deve ser inferido a partir da observação de variáveis secundárias que estejam relacionadas a ela. Ela sugere formas de represntar a relação entre a probabilidade de um individuo dar uma certa resposta a um item e seus traços latentes, proficiências ou habilidades na área de conhecimento avaliada (ANDRADE e CLEIN, 1999).

2.2 ABORDAGENS TEÓRICAS SOBRE A EVASÃO ESCOLAR

A evasão escolar é um grande problema relacionado à educação e atinge todos os níveis de ensino. Evasão escolar é o abandono da escola antes da conclusão de uma série ou de um determinado nível. Trata-se de uma verdadeira ameaça à realidade educacional de muitos países do mundo, tendo o Brasil como um dos campeões desta situação negativa e vergonhosa (BISSOLI, 2010).

Várias formas de interpretação não permitem definir exatamente “evasão e abandono escolar”. A diversidade de conceituação atrapalha a quantificação precisa dos casos, dificultando o estudo das causas e dos princípios que podem levar a alternativas claras e objetivas para superação desse problema que perdura até hoje. São fundamentais a compressão das relações entre os motivos de ingresso e a trajetória dos permanecentes, dos desistentes e egressos desse público, dentre muitas outras questões.

Evasão, segundo Riffel e Malacarne (2010), é o ato de evadir-se, fugir, abandonar, sair, desistir, não permanecer em algum lugar. Quando se trata de evasão escolar, entende-se a fuga ou abandono da escola em função da realização de outra atividade. A diferença entre evasão e abandono escolar foi utilizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP (GIPPS, 1998). Nesse caso, “abandono” significa a situação na qual o aluno se desliga da escola, mas

retorna no ano seguinte, enquanto na “evasão” o aluno sai da escola e não volta mais para o sistema escolar. Já o Índice de Desenvolvimento da Educação Básica/Ideb (2012) aponta o abandono como o afastamento do aluno do sistema de ensino e desistência das atividades escolares, sem solicitar transferência. Steinbach (2012) e Pelissari (2012) adotam o termo abandono escolar, pois consideram “evasão” um “ato solitário”, levando a responsabilizar o aluno e os motivos externos pelo seu afastamento. Ferreira (2013) chama de fracasso das relações sociais que se expressam na realidade desumana que vivencia o aluno em seu cotidiano. Machado (2009) diz que tratar da evasão é tratar do fracasso escolar; o que pressupõe um sujeito que não logrou êxito em sua trajetória na escola (MACHADO, 2009).

Ferreira (2016) vai além, quando afirma que o fracasso escolar e a consequente evasão denotam o próprio fracasso das relações sociais que se expressam na realidade desumana que se vivencia no cotidiano, no qual a distância formada pela teoria e a prática desafia a inteligência do indivíduo. Evasão e abandono não têm uma origem definida e por isso não terão um fim por si só. O problema não é a falta de vinculação às políticas públicas, a desestruturação familiar ou ainda as dificuldades de aprendizagem dos educandos, e sim a soma de vários fatores.

Conforme o pensamento de Digiácomo (2005), a evasão escolar é um problema crônico em todo o Brasil, sendo muitas vezes passivamente assimilada e tolerada por escolas e sistemas de ensino, os quais chegam ao exercício de expedientes maquiadores ao admitirem a matrícula de um número mais elevado de alunos por turma do que o adequado, já contando com a ‘desistência’ de muitos ao longo do período letivo. Em que pese a propaganda oficial sempre alardear um número expressivo de matrículas a cada início de ano letivo, em alguns casos chegando próximo aos 100% (cem por cento) do total de crianças e adolescentes em idade escolar, de antemão já se sabe que destes, uma significativa parcela não irá concluir seus estudos naquele período, em prejuízo direto à sua formação e, é claro, à sua vida, na medida em que os coloca em posição de desvantagem face aos demais que não apresentam defasagem idade-série.

O que chama a atenção é o número de alunos que abandona a escola básica, mas isso também atinge todos os níveis de ensino. É fenômeno que causa prejuízos no campo educativo. Pelo insucesso escolar e pelos baixos rendimentos, constitui uma preocupação constante, pois para o Ministério da Educação (MEC) “o maior

desafio dessa escola é garantir condições para que o aluno possa aprender" (DOURADOS, 2005).

Verifica-se, na atualidade, vários fatores que podem influenciar no agravamento do fenômeno da evasão escolar. Entretanto, fatores intrínsecos e extrínsecos à escola, como drogas, sucessivas reprovações, prostituição, falta de incentivo da família e da escola, necessidade de trabalhar, excesso de conteúdos escolar, alcoolismo, vandalismo, falta de formação de valores e preparo para o mundo do trabalho influenciam diretamente nas atitudes dos alunos que se afastam da escola. Esses obstáculos, considerados, na maioria das vezes, intransponíveis para milhares de jovens, engrossam o desemprego ou os contingentes de mão de obra barata. Em pesquisa feita pela Fundação Getúlio Vargas (FGV), Neri (2009) afirma que o mercado de trabalho é um ator importante na tomada de decisão desse jovem que teima em continuar seus estudos para que possa ser absorvido por ele, ou desiste e torna-se uma mão de obra desqualificada para garantir sua sobrevivência. As escolas não ficam isoladas desse contexto.

2.2.1 Causas e Consequências da Evasão e do Abandono Escolar

Evasão e abandono escolar têm sido associados a situações tão diversas quanto a retenção e a repetência do aluno na escola. Sabe-se ainda que implica uma ampla abordagem da qualidade e a da quantidade. Enguita et al. (2010) acrescentam que a qualidade do sistema educacional de um país é, além de um indicador dos níveis de desenvolvimento e bem-estar social, um indicador de como será o futuro dessa nação. Pesquisas elaboradas por Lucas (1998), Barro (1991) e Mankiw, Romer e Weil (1992) associam níveis educacionais a um maior crescimento econômico. A escassez de informações teóricas e empíricas sobre o problema, bem como as dificuldades para construir indicadores adequados à sua investigação, dificultam ainda mais o seu entendimento e suas definições, e a própria forma de condução do ensino.

Segundo o pensamento de Dore e Lüscher (2001, p. 775), várias situações corroboram para a retenção e repetência do aluno na escola: a saída do aluno da instituição e do sistema de ensino, a não conclusão de um determinado nível de escolaridade, o abandono da escola e o posterior retorno.

Viadero (2001) e Finn (1989) afirmam que a evasão pode ser representada por aqueles indivíduos que nunca ingressaram em um determinado nível. Outro aspecto

considerado relevante nessas situações concerne ao nível escolar no qual estas ocorrem, pois, o abandono da escola fundamental ou de nível médio (Montmarte, Mahseredjian, Houle, 2001) é significativamente diferente daquele que ocorre na educação de adultos ou na educação superior.

Pode-se observar que existem três dimensões conceituais indispensáveis à investigação da evasão escolar: (i) níveis de escolaridade no qual esta ocorre, como a educação fundamental, a educação média ou a superior; (ii) tipos de evasão, como a descontinuidade, o retorno, a não conclusão definitiva, dentre outras; e (iii) razões que motivam a evasão, como, por exemplo, a escolha de outra escola, um trabalho, o desinteresse pela continuidade de estudos, problemas na escola, problemas pessoais ou problemas sociais (JORDAN, LARA, MCPARTLAND, 1996).

Para Rumberger (1995), a chave da compreensão e solução da evasão é encontrar as causas do problema, mas essas causas de forma análoga a outros processos do desempenho escolar têm influência de um conjunto de fatores, como o estudante, a família, a escola e a comunidade em que vive. Revisando diversas pesquisas sobre as causas que levam à evasão, esse autor consegue identificar como problema duas perspectivas: uma individual, a qual envolve o estudante e as circunstâncias de seu percurso escolar; e outra institucional, a qual leva em conta a família, a escola, a comunidade e os grupos de amigos. Ainda podem ser verificadas diferentes teorias que abordam a evasão escolar. Algumas citam a existência de dois tipos principais de engajamento: o escolar (acadêmico ou aprendizagem) e o social (relacionamento com os colegas, com os professores e com os demais membros da comunidade escolar). Essas duas formas são determinantes para a decisão de evadir ou permanecer na escola (Rumberger, 1995). Nesse sentido, Ferreira (2013) afirma que os motivos que levam à evasão podem ser classificados ainda de acordo com os seus fatores determinantes: (i) escola (não atrativa, autoritária, com professores despreparados, insuficiente, com ausência de motivação); (ii) aluno (desinteressado, indisciplinado, com problema de saúde, gravidez); (iii) pais ou responsáveis (não cumpridores do pátrio poder, desinteressados em relação ao destino dos filhos); e (iv) social (trabalho com incompatibilidade de horário para os estudos, agressão entre os alunos, violência em relação a gangues, entre outros).

Lopes (2010) ressalta que, para a amenização de alguns problemas referentes à evasão, é necessária uma ação firme dos poderes públicos, principalmente em

relação aos gestores escolares, os quais precisam assegurar um bom ensino e aprendizagem. Desempenho ruim também é um fator de evasão; oposto a isso, há alunos que evadem por não se sentirem “desafiados e estimulados”. Em um apanhado geral da literatura sobre abandono escolar, em 203 estudos no assunto, chegam-se a algumas conclusões relevantes: notas baixas no início do processo educativo é um forte aspecto de previsão de futuro abandono; desempenho inadequado frequente costuma implicar reprovação; faltas, atos delinquentes e abuso de substâncias ilegais são fortes preditores de abandono. Essa superação poderá acontecer em um ambiente familiar estável, e o acesso a recursos sociais e financeiros influencia de forma significante a probabilidade de o estudante completar seus estudos (RUMBERGER e LIMA, 2008).

2.2.2 A Evasão no Ensino Superior

A evasão estudantil no ensino superior pode ser considerada um fenômeno dos mais graves e complexos, envolvendo a educação, o qual acontece tanto nas instituições públicas quanto nas instituições privadas e em diversos países do mundo. Trata-se de um tema em evolução e estudo e que contribui com conseqüências significativas para o processo educacional.

A evasão é na verdade um fenômeno social complexo, definido como interrupção no ciclo de estudos em qualquer nível de ensino (GAIOSO, 2005). Esta definição é amplamente utilizada em vários trabalhos realizados. Porém, outros conceitos diferentes a respeito desse tema também foram encontrados em estudos variados. Baggi e Lopes (2011) definem a evasão como a saída do aluno da instituição antes da conclusão de seu curso. Já Polydoro (2000) chama a atenção para a distinção entre dois conceitos: a evasão do curso – que consiste no abandono do curso sem a sua conclusão e a evasão do sistema – que reflete o abandono do aluno do sistema universitário. Cardoso (2008) refere-se a conceitos similares, porém a partir de diferentes nomenclaturas: a evasão aparente, que trata da mobilidade do aluno de um curso para o outro e a evasão real que se refere à desistência do aluno do Ensino Superior. A evasão real coincide com a evasão do Sistema, citada por Polydoro (2000), quando o aluno desiste de estudar.

O Relatório da Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras (BRASIL, 1997) traz três maneiras de conceituar a evasão de acordo com nível em que ela ocorre:

1. **Evasão do curso** - quando o estudante se desliga do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional;
2. **Evasão da instituição -** quando o estudante se desliga da instituição na qual está matriculado; e
3. **Evasão do sistema** - quando o estudante abandona de forma definitiva ou temporária o ensino superior.

Silva et al. (2007) definem a evasão como: evasão anual média e a evasão total. A evasão anual média corresponde à porcentagem de alunos que, não tendo se formado, não realizaram matrícula no ano seguinte, demonstrando assim as perdas médias anuais em determinado curso, instituição ou conjunto de instituições. Já o outro tipo, a evasão total, corresponde à quantidade de alunos que em um determinado período, não tendo se matriculado em um determinado curso, IES ou sistema de ensino, não obtiveram diploma. Desta forma não se sabe exatamente o que motivou o aluno a não prosseguir seus estudos, se o mesmo interrompeu o curso, trancando-o, ou se mudou de faculdade ou mesmo se deixou o ensino superior.

Ainda no trabalho de Silva et al. (2007), a evasão é analisada por diferentes formas, levando - se em conta os diferentes tipos de instituição, o conjunto de todas as instituições de um país, em determinado curso, em determinada região geográfica, por categoria administrativa (público/privado), por forma de organização acadêmica (universidades, centros universitários, faculdades, institutos superiores) e por áreas de conhecimento.

Segundo Cislaghi (2008), os primeiros estudos sobre o fenômeno da evasão de estudantes de ensino superior surgiram nos EUA, na década de 50, devido à grande escassez de recursos humanos gerada pela Segunda Guerra Mundial. No Quadro 1, é apresentada uma síntese da evolução dos esforços feitos nas últimas seis décadas sobre o estudo do assunto. Nele é possível verificar que, na última década, as análises dos índices de evasão e permanência ganharam importância para a tomada de decisões nas Instituições de Ensino Superior (IES), principalmente no que diz respeito a manutenção e alocação de receitas.

Quadro 1. Sintese Hisórica dos estudos sobre evasão

Década	Motivo	Educação Superior e os avanços na permanência de estudantes
1950	Expansão	Após as grandes Guerras Mundiais, ocorre uma expansão no número de IES e no contigente de estudantes.
1960	Prevenção da Evasão	Surgem situações problemáticas nas IES provocadas pelo grande contigente de estudantes, pela diversidade que os caracteriza e pela inquietação social causada por vários fatores socioculturais; São realizados os primeiros esforços para controlar a evasão com estudos que não se limitem às abordagens estatisticas descritivas.
1970	Construção de teorias	É criada uma base de conhecimento e propostas as primeiras estruturas teórico conceituais que vão impulsionar o avanço sistemático da compreensão dos processos relacionados ao fenômeno da evasão.
1980	Administração de Matriculas	Crescem os esforços das IES para atrair e manter estudantes; O tema permanência se consolida na área do ensino superior.
1990	Abertura de horizontes	Avançam muitos estudos empiricos para validação das teorias e modelos sobre permanência e evasão; Emerge com força a tendência de considerar o processo de aprendizagem como importante para a permanência de estudantes.
2000	Tendências	índices de permanência passam a ser considerados como indicadores importantes e a serem utilizados por orgãos oficiais para alocação entre as IES, em especial as do setor público; O ensino a distância aparece como elemento novo dentro e fora das IES; Cresce a importância da formação superior para os profissionais que disputam uma colocação no mercado de trabalho cada vez mais exigente.

Fonte: Adaptado de Cislaghi (2008)

Como pode ser observado no Quadro 1, de acordo com estudos de Berger e Lyon (2005), nos anos 50 se iniciou o desenvolvimento de estudos científicos sobre a evasão e a permanência de estudantes no ensino superior nos EUA, cujo objetivo era contribuir para o desenvolvimento de estratégias institucionais que contribuíssem na redução de ocorrências do fenômeno da evasão universitária. Ocorreu nessa época uma grande expansão do número de IES e consequentemente de alunos nos cursos superiores. A sociedade americana se industrializava no período pós-guerra e havia assim grande necessidade de mão de obra com formação superior, mais preparada e com habilidades desenvolvidas para atender a sociedade em desenvolvimento e em busca dos benefícios da tecnologia, para assim reparar os prejuízos advindos do grande conflito ocorrido.

Ainda segundo os autores, a próxima década se inicia com um enorme contingente de estudantes nas diversas universidades e desta forma os problemas de convivência não tardaram a surgir, devido a ampla diversidade cultural, étnica e sócio-econômica que envolvia esses estudantes. Também a economia do país em pleno crescimento exigia novas configurações nos currículos dos cursos e um desempenho acadêmico cada vez mais exigente, porém sem a devida valorização da produção intelectual dos alunos. Este fato provocava grande insatisfação nos estudantes, aliado a um contexto de intensa movimentação, insatisfações políticas, rebeliões, ativismos e movimentos pelos direitos civis. Esta insatisfação vem corroborar com o aumento

dos níveis de evasão universitária.

De acordo com Spady (1970), anteriormente à década de 1970 os estudos sobre evasão se agrupavam em algumas categorias, como: *estudos teóricos*, os quais traziam algumas recomendações para a prevenção da evasão, com objetivo de evitá-la; *censitários,* focados em índices puramente quantitativos; *autópsias,* as quais traziam a visão dos estudantes sobre a razão do abandono do curso ou da instituição; *estudos de caso*, os quais demonstravam trajetórias individuais e as abordagens descritivas, nas quais eram caracterizadas de uma forma geral as experiências de estudantes evadidos; e por fim os *estudos preditivos*, cujo objetivo era prever alternativas de sucesso estudantil. O autor conclui afirmando que até então inexistiam estudos sobre a evasão, pelo menos não do tipo analítico-exploratório, com condições de sintetizar o conhecimento científico sobre o tema e colaborar na compreensão das razões que fazem os estudantes concluírem ou abandonarem o ensino superior iniciado. Este tipo de estudo somente apareceria na década seguinte.

Desta forma na década de 1970, surgem os estudos de Vincent Tinto, por meio dos quais se estabeleceu uma base sólida de conhecimentos que contribuíram de uma forma mais sistemática e científica para a compreensão da problemática da evasão, além de impulsionar a realização de outros estudos baseados em seu conteúdo, os quais surgem nas décadas posteriores, como Pascarella (1980), Bean (1982), Astin (1997), Cabrera, Castaneda, Nora e Hengstler (1992), além de novos estudos do próprio Tinto (1987, 1993), os quais revisam e complementam sua teoria. Somam-se a esses trabalhos a experiência acumulada pelas IES, sobre o tema em questão, o que amplia esta área de estudo.

A partir da década de 2000 muitos estudos surgem com o intuito de ampliar a compreensão do processo de evasão, como aqueles publicados por Miller (2007), Janusik e Wolvin (2007) e também um estudo realizado por Grayson e Grayson (2003). Todos esses pesquisadores tinham como objetivo buscar alternativas adequadas para reduzir o processo de evasão.

Em síntese, nota-se diversas definições de evasão, provenientes de vários estudos e diferentes pesquisadores do tema, porém todas convergem para um mesmo ponto, ou seja, todas buscam entender como ocorre o processo de evasão nas diferentes instituições e quais motivos levam os estudantes a desistirem dos seus objetivos iniciais de graduarem-se e assim conseguir melhor qualificação e

possivelmente maiores possibilidades no mercado de trabalho. Alcançar este entendimento poderá trazer grande contribuição aos estudos sobre evasão e sugestões de alternativas de como as IES devem proceder para reter seus estudantes, já que a evasão é um dos grandes problemas que afligem as instituições de ensino em geral.

Os estudos teóricos sobre a evasão ganharam destaque a partir da década de 70, especialmente quando Tinto (1975) desenvolveu um modelo que serviu como inspiração e base para outros modelos teóricos sobre o assunto. Desde então, vários autores têm empregado esforços na construção e aprimoramento de várias teorias que levam em conta diversas variáveis que contemplam características pessoais do discente, características da IES e o ambiente externo. A seguir são apresentados os principais modelos de evasão propostos na literatura sobre evasão.

2.2.3 Modelo do Abandono Escolar

Spady (1970, 1971) propôs um modelo baseado em evidências empíricas, com objetivo de compreender o fenômeno da evasão. Ele partiu da teoria do suicídio de Durkheim (1951), a qual diz que a possibilidade de um indivíduo cometer suicídio depende do seu nível de integração na sociedade. Em seu modelo, Spady faz um comparativo da possibilidade de suicídio com a possibilidade de evasão.

Na vida acadêmica, se houver adequada congruência, entre o estudante e a instituição, as dificuldades e os desafios serão ultrapassados e as possibilidades de persistir são maiores. A interação entre fatores como, objetivos, interesses e personalidade do estudante influenciam o ambiente universitários e são influenciados por este, impactando em seu nível de satisfação. A Figura 2 apresenta o modelo sociológico explicativo da evasão escolar proposto por Spady (1971).

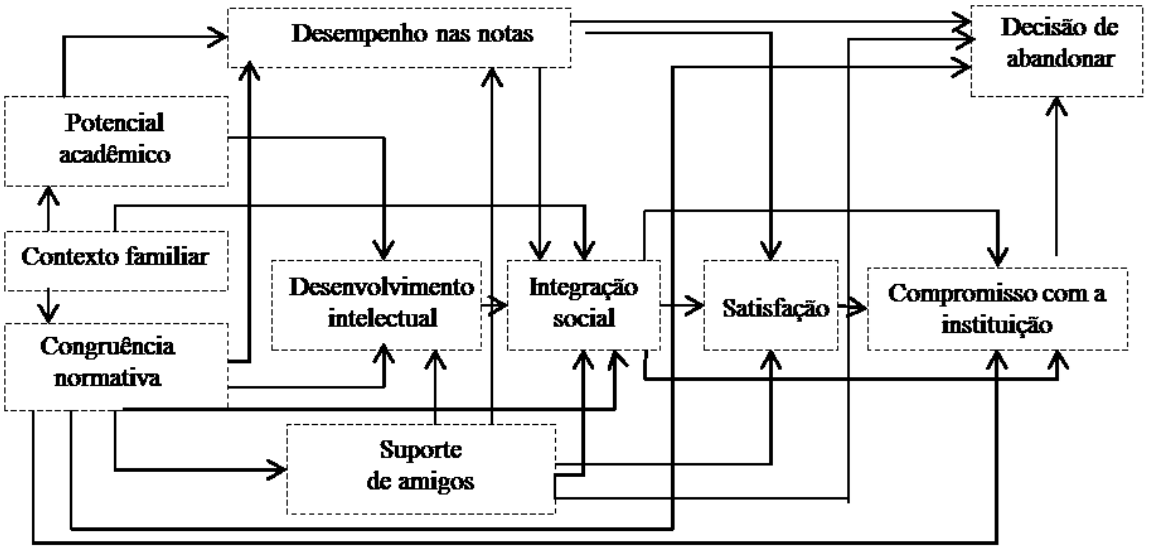


Figura 2. Modelo sociológico explicativo do abandono escolar
Fonte: Spady (1971)

O autor realizou também um estudo longitudinal com 683 estudantes calouros da Universidade de Chicago, com o objetivo de testar seu modelo teórico para explicação da evasão de estudantes no período de graduação. Para ele, cinco variáveis desempenham um papel importante no processo de desistência do aluno no ensino superior. Além das questões referentes ao contexto familiar, são elas: potencialidade acadêmica, congruência normativa, avaliações de desempenho, desenvolvimento intelectual e suporte das amizades. Spady correlaciona a interação dos estudantes com seus valores, interesses, habilidades, características e atitudes com o ambiente acadêmico, representado pelos professores, gestores e colegas. Cislaghi et al. (2008) explicam que a congruência normativa consiste na forma de como os objetivos, os interesses e a personalidade do estudante se articulam com os subsistemas institucionais. O grau de congruência implica o suporte de amizade, o desenvolvimento intelectual e o desempenho acadêmico refletido pelas notas. A interação desses fatores influencia o nível de integração social do aluno com o ambiente acadêmico e o nível de satisfação com suas experiências como universitário. Quanto maior o nível de satisfação do estudante com o ambiente acadêmico, maior é o seu comprometimento com a instituição e menores serão as possibilidades de abandono do curso.

Por fim Spady (1971) sugere que a evasão é o resultado da falta de integração dos estudantes com o ambiente acadêmico e assinala ainda que o meio familiar é uma

das principais fontes que expõe os estudantes a influências e expectativas. A integração plena ao ambiente universitário requer respostas efetivas a diversas demandas dos sistemas acadêmicos e sociais da educação superior. Quanto mais satisfeito o aluno se encontrar, maior será o seu nível de compromisso com o universo acadêmico e isto contribuirá na sua decisão de ficar ou abandonar o curso que escolheu.

2.2.4 Modelo de Integração dos Estudantes

O modelo proposto por Tinto (1975) pode ser entendido como institucionalmente orientado, cujos conceitos centrais são a integração acadêmica e social à instituição, nos quais a decisão do estudante de permanecer ou deixar a universidade é o resultado do nível de conexão desenvolvido com a mesma. Os indivíduos entram na universidade com uma variedade de atributos e características (sexo, raça, habilidades específicas), experiências pré-universitárias (desempenho acadêmico anterior, talentos acadêmico e social) e *background* familiar (atributos de status social, renda, valores e expectativas), cada um deles tendo um impacto direto ou indireto sobre o seu desempenho na universidade.

Esses antecedentes influenciam no desenvolvimento de expectativas e comprometimentos educacionais que o indivíduo traz para o ambiente universitário, os quais podem ser divididos em comprometimento com metas educacionais ou de carreira, mais especificamente a meta de graduação universitária, e comprometimento com a instituição particular em que o estudante se matriculou (TINTO, 1975).

Ainda segundo o autor esses comprometimentos são afetados majoritariamente pelo grau de integração acadêmica e social do estudante no ambiente universitário. Neste sentido, o processo de evasão deve ser visto como longitudinal, no qual as experiências do estudante modificam seus comprometimentos iniciais com a instituição e com o objetivo de graduar-se, de forma a levá-lo ao longo do processo a permanecer ou optar por uma dentre várias formas de evasão.

De acordo com Pascarella e Chapman (1990), as variáveis de comprometimento exercem um efeito direto e intenso na persistência do estudante; enquanto que variáveis como idade, sexo ou nível sócio-econômico tendem para um efeito indireto, em combinação com integração social e acadêmica, na predição da persistência. Tinto (1993) argumentou contra modelos de evasão que percebem a saída da universidade

como o reflexo de alguma deficiência ou fragilidade no indivíduo, e reforçou a importância dos aspectos subjetivos do construto integração à universidade. Na verdade, o autor entende que o indivíduo pode revelar comportamentos e desempenhos excelentes e ainda assim decidir evadir por razões outras que fraquezas ou deficiências. Enfim, o comportamento de evasão não precisa ser, necessariamente, percebido de forma negativa. Os componentes do modelo de Tinto (1993) são apresentados na Figura 3 e serão examinados separadamente em seguida.

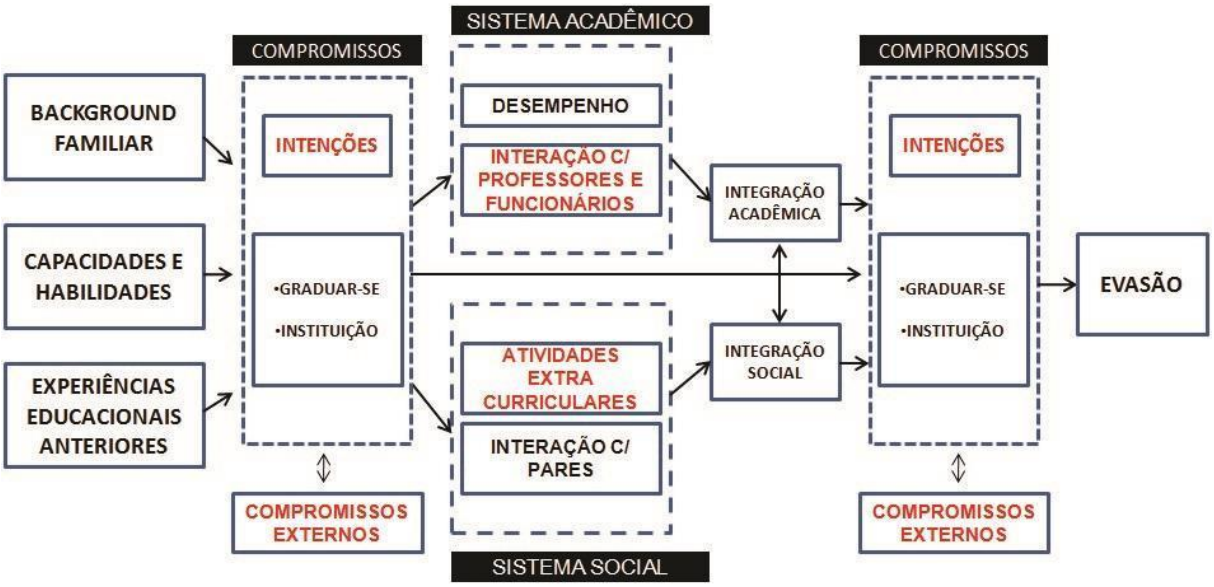


Figura 3. Modelo de integração dos estudantes
Fonte: Tinto (1993)

a) Background Familiar

Em termos gerais, a condição sócio-econômica da família parece estar inversamente relacionada com o comportamento de desistência. Especificamente, quanto mais restritas as condições sócio-econômicas da família, mais altas são as taxas de desistência, mesmo levando em conta as capacidades intelectuais (TINTO, 1993). Existem outros fatores do background familiares também importantes. A qualidade das relações dentro da família e o interesse e expectativas dos pais com relação à educação de seus filhos estão entre eles. Os estudantes que persistem no ensino superior tendem a vir de famílias com pais mais abertos, democráticos, e menos conflituosos nas relações com os filhos. Estes estudantes parecem não apenas receber conselhos, elogios e expressões de interesse em suas experiências

escolares, mas seus pais expressam expectativas elevadas sobre o nível de educação formal de seus filhos. Deste modo, parece que o nível de expectativas dos pais tem tanta influência no comportamento de persistência dos filhos quanto as expectativas destes para com eles próprios (TINTO, 1993).

b) Capacidades e Habilidades

A habilidade cognitiva do estudante tem se mostrado mais importante para a persistência na universidade do que o status social da família (TINTO, 1975, 1993). As medidas de habilidade obtidas por meio de testes de inteligência e as obtidas por meio de grade de notas na escola correspondem a diferentes aspectos da competência do sujeito. A grade de notas tende a ser melhor preditor do sucesso do indivíduo na universidade somente se as características e as exigências desta não forem muito distantes das que o sujeito possuía na escola secundária (TINTO, 1993).

Além das habilidades cognitivas, diferenças significativas de personalidade e de atitude têm sido observadas entre os estudantes que persistem e os desistentes. Os alunos que evadem tendem à impulsividade, a demonstrar pouco compromisso emocional com sua educação e incapacidade de tirar proveito de suas experiências passadas. Tais alunos parecem apresentar uma falta de flexibilidade para lidar com mudanças e tendem a ser mais instáveis, ansiosos e preocupados com o sucesso acadêmico do que os demais estudantes (TINTO, 1993).

c) Rendimento Acadêmico Anterior

Embora, de modo geral, as experiências em ambientes educacionais anteriores não tenham sido diretamente relacionadas com a desistência da universidade, o desempenho na escola, medido por meio da média de notas ou colocação em classe, tem se mostrado um preditor importante do desempenho futuro do estudante. Por outro lado, as características da escola secundária, tais como a qualificação do corpo docente e recursos didáticos, têm impacto no desenvolvimento acadêmico e intelectual, e, por conseguinte, afetarão o desempenho e a persistência na universidade. As características da escola, sua cultura e valores, também influenciam direta ou indiretamente as aspirações, expectativas e motivações do indivíduo com relação à universidade e ao objetivo da graduação (TINTO, 1993).

d) O compromisso com a Meta de Graduação

A pesquisa indica que, depois das habilidades intelectuais, o compromisso que

o sujeito assume com a meta de concluir o ensino superior é o maior preditor da permanência deste na universidade. Se for medida em termos de planos e expectativas educacionais, ou expectativas com a carreira, pode-se dizer que quanto mais elevado o nível de aspiração, maior a probabilidade de permanência. O compromisso com as metas educacionais ou de carreira reflete um processo multidimensional de interações entre o indivíduo, sua família e as experiências prévias de escola (TINTO, 1993).

e) Integração no Ambiente Universitário

A forma como o indivíduo se integra ao sistema social e acadêmico da universidade é um fator diretamente relacionado à sua permanência na instituição. Quanto maior o grau de integração no sistema universitário, maior será o compromisso com a instituição e com as metas de graduação, consequentemente, menor será a probabilidade de desistência. A integração no sistema universitário é subdividida em integração acadêmica e integração social (TINTO, 1993).

Na universidade a integração acadêmica pode ser medida em termos do desempenho sob forma de notas e do desenvolvimento intelectual durante os anos de estudos. O primeiro aspecto está relacionado aos padrões explícitos do sistema acadêmico, já o segundo diz respeito à identificação e à avaliação do indivíduo em relação aos valores deste sistema. As notas são uma recompensa explícita e extrínseca para o indivíduo. O desenvolvimento intelectual representa uma forma mais intrínseca de recompensa, um aspecto integrado ao desenvolvimento pessoal e acadêmico do estudante. Neste sentido, é importante fazer a distinção entre os estudantes que fracassam nos estudos e aqueles que voluntariamente deixam a universidade. Estes últimos, com frequência, apresentam melhor desempenho em vários aspectos (como habilidades cognitivas e notas) do que os estudantes que persistem, e a sua evasão pode ser compreendida como uma percepção de incompatibilidade dos seus valores e objetivos com os do sistema acadêmico (TINTO, 1993).

As notas são o fator singular mais importante no que tange à predição da persistência na universidade. Noutra perspectiva, os alunos que persistem também tendem a valorizar a educação de nível superior mais como um processo de ganho de conhecimento e de apreciação de idéias do que como uma etapa na escalada da carreira (TINTO, 1993).

Integrar-se ao ambiente social, por sua vez refere-se ao envolvimento do estudante no meio social de sua universidade, por meio de associação ao grupo de iguais, atividades extracurriculares, interação com o corpo docente, com os colegas e com as diversas pessoas inseridas no contexto universitário. O sucesso e gratificação em cada uma destas atividades podem ser vistos como uma recompensa que influi na avaliação geral do indivíduo sobre os custos e benefícios de estar na universidade, e também modificam a sua experiência educativa e o seu comprometimento institucional.

A relação do estudante com o corpo docente influi não apenas na integração social e no compromisso com a instituição, mas também na integração acadêmica. Neste sentido, a integração com o corpo docente e a coordenação pode influir tanto no desenvolvimento intelectual do aluno quanto em seu desempenho sob a forma de notas.

O ambiente social, parece não exigir uma congruência entre os valores do estudante e as características do ambiente prevalente na instituição, podendo ocorrer por meio de relações congruentes com somente alguma parte do sistema social da universidade. Mesmo assim, estudantes com valores, interesses e atitudes mais “convencionais” têm maior probabilidade de estabelecer relações mais próximas e consistentes com o ambiente social da instituição como um todo (TINTO, 1993).

Portanto, a integração social por meio de atividades extracurriculares está relacionada à persistência no ensino superior. A participação nestas atividades resulta, na perspectiva do estudante, numa experiência de maior articulação entre o sistema social e acadêmico da universidade, e parece ajudar na redução de possíveis percepções de incompatibilidade entre os dois sistemas. A seguir será analisado o modelo de Bean (1982) relacionado à mesma temática.

2.2.5 Modelo de Evasão de Bean

O modelo de Bean (1982) adquiriu proeminência no campo educacional por ser um modelo causal da evasão universitária, baseado em estudos de *turnover* nas organizações de trabalho, denominado “modelo de evasão”. O autor incluiu em seu modelo variáveis ausentes no modelo de Tinto (1975), tais como as intenções de evasão e oportunidades de transferência.

Assim como Tinto (1975), Bean (1980, 1982) considera que as intenções de permanência e evasão são moldadas a partir das crenças e atitudes prévias que os indivíduos trazem para a universidade. Este modelo é composto por variáveis intencionais, atitudinais, organizacionais, pessoais e ainda algumas que dizem respeito ao ambiente. São dez determinantes: intenção de abandonar a universidade, valores práticos, certeza da escolha da universidade, lealdade institucional, desempenho em notas, satisfação com o curso, metas educacionais, certeza de escolha vocacional, oportunidades de transferência e aprovação da instituição pela família.

A intenção de abandonar a universidade está relacionada à probabilidade do sujeito permanecer na instituição. É considerada a variável mais importante para a predição do comportamento de evasão. As três variáveis de atitude (valores práticos, certeza da escolha da universidade e lealdade institucional) estão diretamente associadas à intenção de abandono. Os valores práticos referem-se à crença do sujeito nas oportunidades de emprego que determinada graduação universitária lhe proporcionará. A variável lealdade institucional significa o quanto é importante para o indivíduo graduar-se em determinada instituição e não em outra. Esta lealdade e a importância da certeza da escolha da universidade possuem relação indireta com o processo de desistência por meio da influência na intenção de abandono (BEAN, 1980, 1982).

Os fatores organizacionais incluem o desempenho sob forma de notas e a satisfação com o curso. Assim como no modelo de Tinto (1993), o desempenho sob forma de notas refere-se à contribuição da média de notas na universidade no processo de evasão. Nos estudos de Bean (1982), a satisfação com o curso apresentou relação direta e significativa com a intenção de abandono.

Os fatores pessoais referem-se às metas educacionais e a certeza sobre a escolha vocacional. As metas educacionais significam a importância da graduação universitária. O grau de certeza do estudante sobre sua escolha profissional não apresenta relação direta com o comportamento de desistência, e contribui para o modelo por meio de sua influência sobre a certeza de escolha da universidade e sobre os valores práticos (BEAN, 1982).

O autor denominou de fatores ambientais a oportunidade de transferência e a aprovação da família. O primeiro avalia as alternativas que o sujeito tem de

transferência para outra instituição, bem como a facilidade ou dificuldade deste processo. A aprovação da família quanto à escolha da instituição universitária revelou ter um papel proeminente na atitude dos estudantes e está fortemente relacionada com todos os demais fatores, com efeitos diretos e indiretos sobre a persistência (BEAN, 1982).

Em síntese as experiências pessoais do estudante, suas características e suas interações com o ambiente universitário influenciam seu nível de satisfação e consequentemente seu comprometimento com a instituição. Este comprometimento seria assim o fator determinante na decisão de permanecer ou de desistir do curso escolhido.

A seguir são abordadas outras teorias referentes à evasão, decorrentes dos estudos já apresentados, na sua maioria, a partir dos conceitos da teoria de Tinto (1975).

2.2.6 Outras Teorias e Modelos aplicadas a Evasão

De acordo com Astin (1984, 1997), quanto mais envolvido for um estudante com a faculdade, maior a probabilidade da sua retenção. Para ele a entrada na Universidade representa mudanças ligadas ao conhecimento, às capacidades vocacionais, aos valores, às atitudes, às crenças e aos comportamentos do estudante. Um aluno que se dedica ao estudo e participa das atividades em grupo, interagindo com colegas, professores e funcionários é um aluno altamente envolvido. Por sua vez um estudante que participa pouco do Campus, não se envolve em atividades extracurriculares, negligencia os estudos e tem contatos pouco frequentes com seus pares e com os demais membros da comunidade acadêmica é considerado um estudante pouco envolvido.

Astin (1984, 1997) afirma que o envolvimento dos estudantes na Universidade depende significativamente da atmosfera institucional criada por eles próprios. Este envolvimento, resultante da adequada integração dos estudantes, vai impactar sobremaneira no desenvolvimento do ensino superior. Por fim esta teoria aponta o relacionamento dos estudantes com seus pares e professores como uma forte influência na maioria das Universidades.

Quando o estudante chega à universidade é imprescindível que ele consiga se relacionar com a comunidade acadêmica, para que ele ali permaneça até a conclusão de seu curso (ASTIN, 1997; TINTO, 1993).

O modelo de Pascarella (1980) é baseado no modelo de Tinto (1975), dando ênfase nas interações vivenciadas no ambiente universitário. Para Pascarella (1980), a interação que ocorre entre os acadêmicos e a instituição na qual estudam acontece por meio de três conjuntos de variáveis independentes que interagem entre si: 1. O nível de contato formalizado entre professores e alunos; 2. As experiências universitárias proporcionadas pelo convívio dentro e fora da sala de aula; e 3. Os resultados educacionais (desempenho formal, intelectual e notas). Segundo este autor apenas as variáveis referentes aos resultados educacionais mostram efeito direto sobre a decisão de abandonar o curso.

No modelo de Cabrera *et al*. (1992), os autores testam um modelo alternativo, que também visa explicar a persistência acadêmica dos estudantes universitários, a partir da comparação dos modelos de Tinto e de Bean. Os resultados apontam para algumas variáveis que têm influencia na permanência acadêmica: primeiramente o compromisso institucional, em seguida o apoio da família e dos amigos; depois a integração acadêmica, a integração social e por fim as questões financeiras.

Este modelo inclui todas as variáveis do modelo de Tinto, ou seja, considera que as experiências dos estudantes influenciadas pela integração acadêmica e pela integração social contribuem para reforçar seus objetivos educacionais e seu comprometimento com a instituição. A partir daí resulta a maior ou menor intenção de persistir em seu objetivo inicial de conclusão do curso. Percebe-se também a inserção de uma nova variável, a qual é a econômica, como elemento influenciador no processo de evasão, porém esta última variável não encontra grande apoio no meio acadêmico e nos estudos posteriores.

Silva et al. (2007) alertaram que muitos preditores da evasão têm sido minimizados e a falta de recursos financeiros do aluno sobrevalorizada como a principal causa para a interrupção de seus estudos. Os autores consideram importante uma maior compreensão de questões de ordem acadêmica, como as expectativas do aluno em relação ao curso ou à instituição, que podem encorajá-lo ou desestimulá-lo a priorizar a conclusão do seu curso.

No modelo proposto por Nora *et al*. (2005), os autores buscaram trazer contribuições para o estudo da evasão nos anos seguintes ao primeiro, visto que os estudos sobre evasão, na sua maioria, se concentram em pesquisas durante o primeiro ano do curso universitário. Assim concluem que a evasão pode acontecer em qualquer momento do curso e orientam as instituições a desenvolverem seus próprios instrumentos para acompanhar a interação dos estudantes com a instituição, pois mesmo estudantes bem-sucedidos em suas experiências universitárias iniciais podem posteriormente encontrar motivos para evadir.

2.2.7 Considerações sobre as Teorias e Modelos Aplicados à Evasão

As teorias apresentadas são de grande relevância para o estudo e entendimento da evasão universitária, sendo mesmo alguns modelos complementares uns aos outros. O modelo de Spady (1980) foi pioneiro e centrou-se no exame da congruência entre o estudante e a instituição. O Modelo Integrado de Permanência de Cabrera, Castaneda, Nora e Hengstler (1992) por sua vez é um modelo que explica a persistência acadêmica dos estudantes universitários a partir da articulação dos modelos de Tinto e de Bean, com a inserção da variável econômica, como elemento influenciador no processo de evasão.

Nota-se que grande parte das teorias abordadas parte das considerações do modelo de integração do estudante de Tinto (1975), como é o caso também do Modelo de Desgaste do Estudante de Pascarella (1980), o qual dá ênfase às interações vivenciadas no ambiente universitário. As ideias de Astin (1997) coincidem com as de Tinto (1975) no tocante à integração universitária, quando ressalta que quando o estudante chega à universidade é imprescindível que ele consiga se relacionar com a comunidade acadêmica, para que ele ali permaneça até a conclusão de seu curso (ASTIN, 1997; TINTO, 1993).

As teorias de Tinto (1975) e também a de Bean (1980, 1982), mesmo com menor ênfase têm recebido atenção considerável na literatura. Embora vários teóricos tenham avançado seus estudos para explicar o processo de persistência da faculdade, os modelos dos dois autores fornecem um quadro teórico mais abrangente sobre as decisões de saída da universidade. Esses autores buscaram desenvolver modelos integradores de variáveis institucionais e não institucionais na tentativa de obter uma melhor compreensão dos processos de ajustamento à universidade. Para

eles, as decisões de evasão são o resultado da interação das expectativas, características e habilidades do estudante com a estrutura, as normas e a comunidade universitária (TINTO, 1975; BEAN 1980, 1982).

Pode-se dizer após a análise realizada que os dois modelos, de Tinto (1975) e de Bean (1980, 1982), têm vários pontos em comum. Ambos os modelos consideram a persistência como o resultado de um conjunto complexo de interações no ambiente acadêmico e levam em conta as características pré-universitárias como fatores a serem considerados no processo de ajuste do estudante à instituição. Além disso, os dois modelos argumentam que a persistência é afetada pela integração bem-sucedida entre o aluno e a instituição.

Algumas diferenças podem ser apontadas entre as duas teorias, pois ao contrário do modelo de integração do estudante de Tinto (1975), o modelo de evasão estudantil de Bean (1980) enfatiza os fatores externos à instituição em afetar atitudes e decisões. O modelo de integração considera o desempenho acadêmico como um indicador de integração acadêmica, porém o modelo de evasão de estudantes considera o desempenho, mas refere este aspecto como uma variável de resultado vinculada a processos sócio-psicológicos. A principal contribuição do modelo de evasão de estudante é explicitar o papel dos fatores externos sobre o processo de persistência universitária (TINTO, 1975; BEAN 1980, 1982).

Analisando as duas teorias, a de Tinto (1975), e a de Bean (1980, 1982), Cabrera, Castañeda, Nora e Hengstler (1992) concluíram que os dois modelos estão corretos em dizer que a persistência na universidade é produto de um complexo conjunto de interações entre fatores pessoais e institucionais, e também em presumir que a intenção de persistência é o resultado de uma relação bem-sucedida entre os estudantes e sua universidade. Assim, os resultados indicaram, segundo a análise desses autores, que as duas teorias não eram mutuamente exclusivas, mas sim complementares e adequadas para a compreensão de evasão. Quanto à questão de qual modelo retrata uma melhor representação do processo de persistência da universidade, a resposta dos autores defende a idéia que as duas são importantes, depende do critério específico em estudo.

Os teóricos citados trouxeram grande contribuição aos estudos sobre evasão, pois definiram alguns modelos de evasão que procuram facilitar a compreensão de como acontece tal fenômeno. É a partir do trabalho destes autores, principalmente

das idéias de Tinto (1975), as quais se configuram em uma moderna pesquisa sobre evasão universitária. Pouca ou mesmo nenhuma pesquisa relevante sobre evasão universitária foi realizada antes do trabalho de Tinto (1975) e grande parte das pesquisas realizadas sobre evasão contém referências às idéias deste autor (Grayson e Grayson, 2003).

Após análise de diversas teorias constata-se que os estudos seminais de Tinto (1975, 1987, 1993) lideram o campo de pesquisa sobre evasão. O modelo teórico proposto e revisto por ele permaneceu como referencial teórico de maior credibilidade e relevância no campo da evasão universitária, sendo foco constante de diversos estudos que buscavam testar empiricamente seus pressupostos.

O modelo de integração do estudante (TINTO, 1975) parece ser mais robusto do que o modelo de evasão de estudantes (BEAN, 1982), bem como em relação aos demais modelos, quando julgado em termos do número de hipóteses validadas. Segundo Cislaghi (2008), o modelo de Tinto (1993) revisto por ele mesmo é um dos mais aperfeiçoados, chegando a obter aprovação empírica relevante em cerca de 70% das proposições que o compõem, e está presente na maioria dos estudos realizados sobre evasão. Sua hegemonia é notadamente reconhecida, sendo o modelo escolhido como base teórica do presente trabalho. O Quadro 2 apresenta a síntese das teorias e modelos sobre evasão e permanência de estudantes.

Quadro 2. Sintese das teorias e modelos sobre evasão

Autores	Denominação	Abordagem	Elementos/ Variáveis	Preditor/ Indicador
Spady (1970)	Modelo do processo de abandono	Sociológica	Contexto familar; congruencia normativa; suporte de amigos; integração social; desempenho acadêmico	Desempenho acadêmico
Tinto (1975, 1993, 1997)	Teoria da integração do estudante	Sociológica	Integração social; integração acadêmica; compromisso com o objetivo; compromisso com a instituição; qualidade do esforço do estudante; compromissos externos	Desempenho em notas; ajustamento na instituição; aprovação e encorajamento por familias e amigos
Bean (1980)	Teoria de desgaste do estudante	Psicológica	Fatores pré-ingresso; fatores ambientais; resultados acadêmicos; resultados psicológicos.	Desempenho em notas; ajustamento na instituição; aprovação e encorajamento por familias e amigos

Pascarella (1980)	Modelo do desgaste	Psicológica	Contato informal com professores; outras experiências universitárias; resultados educacionais.	Resultados Educacionais
Astin (1985)	Teoria do envolvimento do estudante	Psicológica	Oportunidade para envolvimento; envolvimento do estudante.	Desempenho em notas
Cabrera et al. (1992)	Modelo integrado de permanência	Sociológica	Capacidade de pagamento; Desempenho de notas; Compromisso com a instituição; compromisso com o objetivo	Desempenho em notas
Nora et al. (2005)	Modelo do comprometimento estudante-instituição após o primeiro ano	Sociológica	Fatores pré-universitários; Experiências acadêmicas e sociais; resultados cognitivos e não cognitivos; compromissos iniciais e finais	Compromisso com o objetivo; compromisso com a instituição

Fonte: Cislaghi (2008)

O Quadro 2 apresenta o estudo de Cislaghi (2008), no qual é possível observar as abordagens, as variáveis e os indicadores de cada teoria. Pode-se ainda observar que entre as teorias e os modelos, há uma predominância da abordagem sociológica e os indicadores mais citados estão relacionados ao desempenho dos estudantes.

2.3 CONSIDERAÇÕES FINAIS DO CAPITULO

Neste capítulo foram tratadas as principais abordagens teóricas sobre o desempenho e evasão escolar existentes na literatura. É importante observar que grande parte dos trabalhos envolve o desenvolvimento de modelos teóricos para a identificação e mensuração dos fatores associados, por meio de observação, experimentos, ou aplicação de questionários contextuais.

Em ambientes educacionais, é necessário o conhecimento não só do quantitativo de alunos que se matriculam, como também dos que concluem a série ou o curso, reprovam ou evadem. É importante identificar os fatores intra ou extra educacionais que os levam ao sucesso ou fracasso, e que os impede de atingir seus objetivos, além do estabelecimento de relacionamentos entre eles com o ambiente escolar ou na universidade

As abordagens teóricas, aqui abordadas, sintetizam aspectos individuais, familiares, comportamentais, socioeconômicos e relacionados ao ambiente educacional. O conhecimento dos fatores associados e suas relações com o desempenho e a evasão escolar, torna-se imprescindível para os professores no que diz respeito à adoção de novas estratégias de ensino - aprendizagem mais

estimulantes, além de prover uma melhor adaptação do estudante ao contexto educacional.

Para os gestores educacionais o conhecimento dessas relações, por meio de indicadores de desempenho e de evasão, permitirá identificar os pontos fortes e fracos em suas instituições e nas politicas públicas atualmente aplicadas na educação. Servirá ainda como meio para proposição de novas politicas que objetivem diminuir as desigualdades educacionais, obter a equidade e melhorar a qualidade do sitema educacional.

O próximo capítulo aborda a descrição sobre as principais técnicas que podem contribuir para identificar e mensurar os fatores associados ao desempenho e à evasão, por meio de dados educacionais.

3 MINERAÇÃO DE DADOS EDUCACIONAIS (EDM)

Neste capítulo serão apresentadas as principais tarefas de Mineração de Dados Educacionais, como também, os conceitos da área necessários para a fundamentação desta tese.

3.1 DESCOBERTA DO CONHECIMENTO

Os avanços mais recentes nos sistemas de informação, nos dispositivos de coleta e nas tecnologias de armazenamento permitem que empresas e organizações armazenem grandes quantidades de dados. Esses dados representam oportunidades para descoberta de conhecimento relevante que possibilita melhores decisões e planejamento mais adequado.

Por meio de uma ampla variedade de campos, os dados estão sendo coletados e acumulados em um ritmo acelerado. Há uma necessidade crescente de geração de novas teorias e ferramentas computacionais para ajudar as pessoas a extraírem informações úteis (conhecimento) dos volumes de dados digitais de rápido crescimento. Essas teorias e ferramentas são os assuntos do campo da descoberta de conhecimento em bases de dados (KDD - do inglês *Knowledge Discovery in Databases)* (FAYYAD *et al.*, 1996).

De acordo com Côrtes, Porcaro e Lifschitz (2002), a Mineração de dados constitui-se como uma etapa de um processo maior de descoberta de informações. O KDD representa uma metodologia de descoberta de informações a partir de uma sequência definida de passos, sendo eles: Limpeza de dados; Integração dos dados; Seleção dos dados; Transformação dos dados; Mineração dos dados; Avaliação dos Padrões e Apresentação e assimilação do conhecimento (HAN; KAMBER, PEI, 2012).

Alguns autores consideram a mineração de dados o próprio KDD ou como seu sinônimo, como é o caso de Han *et al.* (2011) e Wang (2005), embora a maioria considere a mineração como sendo parte do processo de KDD. Fayyad *et al.* (1996), em seu texto clássico sobre o tema, estabeleceram bem os limites e diferenças de cada área. Para eles, o KDD refere-se a todo o processo de descoberta de conhecimento útil em dados e a mineração refere-se a uma determinada etapa nesse processo. A mineração de dados é a aplicação de algoritmos específicos para extrair padrões de dados. Esse processo consiste de uma série de passos para

transformação, do pré-processamento dos dados até o pós-processamento dos resultados da mineração. A Figura 4, mostra, de maneira simplificada, o processo (TAN *et al.*, 2009).

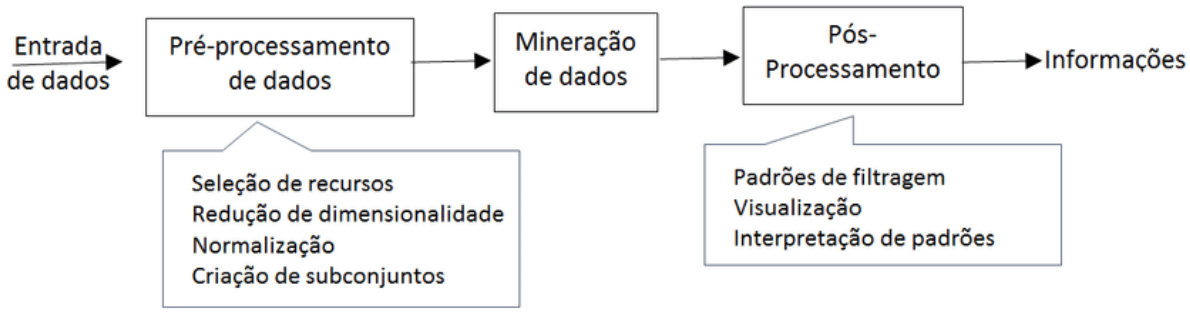


Figura 4. Processo de descoberta de conhecimento
Fonte: Tan et al. (2009)

Para Han *et al.* (2011) e Maimon e Rokach (2010), o KDD é uma análise exploratória, automática e a modelagem de grandes repositórios de dados. É o processo organizado de identificação de padrões válidos, novos, úteis e compreensíveis em conjuntos de dados grandes e complexos.

3.2 MINERAÇÃO DE DADOS

A evolução e o crescimento das fontes de geração e dispositivos de coleta e armazenamento de dados têm permitido que as organizações acumulem vasta quantidade de dados. A extração de informação útil desses dados tem provado ser extremamente desafiadora. Geralmente, as ferramentas e técnicas tradicionais de análises de dados não podem ser utilizadas em razão do tamanho do conjunto de dados ser muito grande.

As principais definições de mineração de dados sempre a associam a uma busca de informação relevante em grandes quantidades de dados. Tan *et al.* (2009) definem a mineração de dados como o processo de descoberta automática de informações úteis em grandes depósitos de dados. Para Ham *et al.* (2011), a mineração de dados é o processo de descobrir padrões interessantes em enormes quantidades de dados.

A mineração de dados pode ser considerada como a principal etapa de um processo de KDD, cujo papel é incluir as tarefas de seleção, preparação e exploração das informações, e a análise e interpretação dos resultados, assimilando o conhecimento extraído do processo. Os padrões citados devem ser novos,

compreensíveis e úteis, devendo trazer algum novo benefício que possa ser compreendido (COSTA, 2012).

A mineração de dados pode ser realizada em qualquer tipo de dados, desde que os dados sejam significativos para um aplicativo de destino, tais como bancos de dados, dados de *data warehouses*, dados transacionais e outros tipos avançados.

Como um domínio altamente orientado para aplicações, a mineração de dados incorporou muitas técnicas de outros domínios, como estatística, aprendizagem de máquina, reconhecimento de padrões, sistemas de banco de dados, recuperação de informação, visualização, algoritmos, computação de alto desempenho e muitos domínios de aplicação (MAIMON e ROKACH, 2010). A natureza interdisciplinar de pesquisa e desenvolvimento de mineração de dados contribui significativamente para o sucesso de mineração de dados e suas amplas aplicações (HAN *et al.*, 2011).

Outro ponto forte tem sido sua ênfase na colaboração com pesquisadores de outras áreas. O desafio de analisar novos tipos de dados não pode ser executado somente aplicando-se técnicas de análise de dados sem a participação daqueles que entendem os dados e o domínio no qual eles estão inseridos. Muitas vezes, a habilidade na construção de equipes multidisciplinares tem sido responsável pelo sucesso de projetos de mineração de dados, como a criação de algoritmos novos e inovadores (TAN et al., 2009). A Figura 5 ilustra a multidisciplinaridade da mineração de dados.

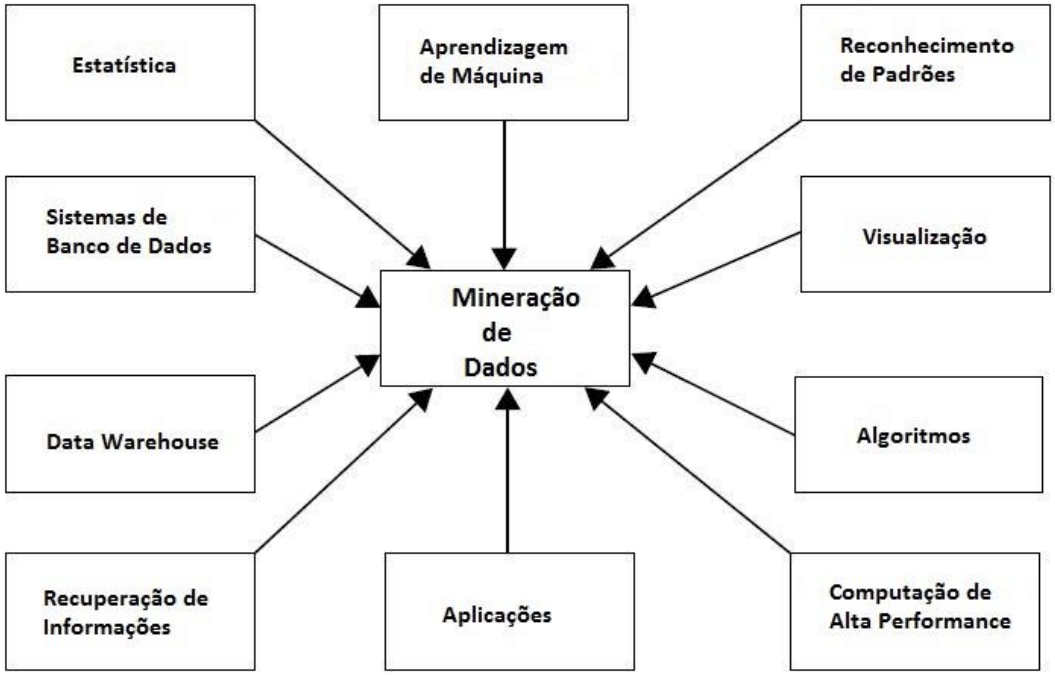


Figura 5. Multidisciplinaridade da Mineração de Dados
Fonte: Han et al. (2011)

Pesquisadores têm sido estimulados a desenvolver novas técnicas de mineração de dados. Isso envolve a investigação de novos tipos de conhecimento, mineração no espaço multidimensional, integração de métodos de outras disciplinas e a consideração das relações semânticas entre objetos de dados. Além disso, as metodologias de mineração devem considerar questões como a incerteza de dados, o ruído e incompletude (HAN et al., 2011).

3.2.1 Aprendizagem de Máquina (AM)

Nas últimas duas décadas, o aprendizado de máquina, notoriamente conhecido pelo termo (ML, do inglês – *Machine learning*), tornou-se um dos pilares da tecnologia da informação, e, consequentemente, traz uma atuação muito importante no dia a dia das pessoas (HOFFMAN *et a*l.,2008). O crescente volume de dados disponíveis só aumenta a necessidade de estudar e aplicar técnicas para análise de dados que empreguem uma inteligência cada vez maior, sendo esse um ingrediente imprescindível para o progresso tecnológico e científico na atualidade (SHALEV et al., 2014).

Segundo Han et al., (2011), a aprendizagem de máquina investiga como computadores podem aprender (ou melhorar o seu desempenho) com base nos dados analisados. A principal área de pesquisa é desenvolver programas de computador

para aprenderem automaticamente a reconhecer padrões complexos e tomar decisões inteligentes baseadas nos dados (HAN *et al.*, 2011).

Os seres humanos estão cercados por dispositivos que utilizam tecnologias baseadas em aprendizado de máquina, direta ou indiretamente. Motores de busca aprendem para trazer os melhores resultados às consultas, *software* aprende com transações de cartão de crédito para encontrar fraudes, sistemas embarcados em câmeras digitais aprendem para reconhecer faces e carros são equipados com sistemas de prevenção de acidentes que são construídos utilizando algoritmos de aprendizado de máquina (SHALEV et al., 2014).

Para Lantz (2013), a aprendizagem de máquina é o campo de estudo interessado no desenvolvimento de algoritmos de computador para transformar dados em ação inteligente. O crescimento de dados exigiu poder de computação adicional, que, por sua vez, impulsionou o desenvolvimento de métodos estatísticos para a análise de grandes conjuntos de dados. Isso criou um ciclo evolutivo, permitindo que os dados, ainda maiores e mais interessantes, possam ser coletados e analisados.

Apesar da grande dimensão, inerente a certos problemas, ser uma das principais motivações de utilizar algoritmos de aprendizado de máquina, ela traz uma dificuldade para a avaliação dos resultados obtidos com os modelos gerados por esses algoritmos, já que muitas vezes a intuição não é suficiente para avaliar certos aspectos. Para tanto, técnicas para avaliação dos algoritmos e modelos gerados são muito estudadas e necessárias para permitir entender o grau de confiança que se pode empregar sobre os resultados obtidos.

As tarefas de aprendizagem podem ser classificadas em aprendizado supervisionado, não supervisionado, semissupervisionado e aprendizado ativo (HAN *et al.*, 2011):

- **Aprendizado Supervisionado**

É a tarefa de aprendizagem de máquina para inferir uma função a partir de dados de treinamento (conjunto de exemplos, observações, medidas, entre outros) previamente rotulados. Na aprendizagem supervisionada, cada exemplo é um par constituído por um objeto de entrada (normalmente um vetor) e um valor de saída desejada (também chamado de sinal de supervisão). Um algoritmo de aprendizagem supervisionada analisa os dados de treinamento e produz uma função inferida, a qual pode ser usada para mapear novos exemplos. Em um cenário ideal, a tarefa irá

permitir ao algoritmo determinar corretamente os rótulos de classe para instâncias invisíveis ou para as seguintes (CORTES *et al.*, 2012). O aprendizado supervisionado está associado aos modelos preditivos, e as suas tarefas mais comuns são a classificação e a regressão (CLEMMENSEN *et al.*, 2011).

- **Aprendizado Não-Supervisionado**

Nas tarefas não-supervisionadas, os dados não precisam de uma pré-categorização ou rótulos. O problema de aprendizado não supervisionado é o de tentar encontrar a estrutura oculta em dados sem rótulo. Uma vez que os exemplos fornecidos são sem rótulos, não há nenhum sinal de erro ou recompensa para avaliar uma solução em potencial (KUMAR *et al.*, 2012). O aprendizado não supervisionado está associado aos modelos descritivos de mineração de dados, e sua tarefa mais comum é a *cluster*ização (agrupamento) (HAN *et al.*, 2011).

- **Aprendizado Semi-Supervisionado**

É uma classe de técnicas de aprendizado de máquina que fazem uso de ambos os exemplos (rotulados e não rotulados) para aprender um modelo. Numa abordagem, os exemplos rotulados são usados para aprender os modelos da classe, e exemplos não rotulados são usados para refinar as fronteiras entre as classes (HAN *et al.*, 2011).

- **Aprendizado Ativo**

É uma abordagem de aprendizado de máquina, a qual permite aos usuários desempenhar um papel ativo no processo de aprendizagem. Uma abordagem de aprendizagem ativa pode solicitar a um usuário (por exemplo, um especialista de domínio) para rotular um exemplo, que pode ser a partir de um conjunto de exemplos não rotulados ou sintetizados pelo programa de aprendizagem. O objetivo é otimizar a qualidade do modelo por meio da aquisição de conhecimento ativo dos usuários humanos, dada a restrição de quantos exemplos que podem ser solicitados os rótulos (HAN *et al.*, 2011).

3.2.2 Modelos de Mineração de Dados

Em geral, o aprendizado de máquina juntamente com a Mineração de Dados busca resolver problemas do mundo real, os quais apresentam relevância e possuam bases de dados contendo informações que possibilitem alcançar a solução. Além disso, é necessário que o volume de dados disponível seja adequado.

Para a resolução desses problemas, é necessário abstrair os processos por meio da construção de modelos que sejam válidos para a zona de operação com que se deseja trabalhar, a qual consiste de um conjunto de entradas e saídas possíveis e relevantes. Logo, para um mesmo processo real, pode-se construir diversos modelos para se obter as soluções de um conjunto de problemas de interesse.

Já em se tratando da construção de um modelo, é importante conhecer o processo no qual os problemas se encontram com um nível de profundidade que permita essa modelagem. Outro aspecto importante é conhecer os dados a que se tem acesso, para poder então se determinar quais tipos de problemas podem de fato ser solucionados.

Visando garantir que os modelos descrevam o fenômeno de estudo com nível adequado de precisão, cada um deles deve ser avaliado e validado. Esse processo de avaliação ocorre conforme critério inerente ao problema em questão, buscando atender às necessidades envolvidas para atingir a solução.

Basicamente, existem dois tipos de modelos de mineração de dados: descritivos e preditivos. Os modelos descritivos, geralmente, aplicam funções de aprendizagem não supervisionada para produzir padrões que explicam ou generalizam a estrutura intrínseca, as relações e inter-relação dos dados extraídos (PENG *et al.*, 2008).

Nos modelos descritivos, o objetivo é derivar padrões (correlações, tendências, grupos, entre outros) que resumam os relacionamentos entre os dados. As tarefas descritivas são, muitas vezes, exploratórias em sua natureza e, frequentemente, requerem técnicas de pós-processamento para validar e explicar os resultados (TAN *et al.*, 2009).

Os modelos preditivos frequentemente aplicam funções de aprendizado supervisionado para estimar valores desconhecidos ou futuros de variáveis dependentes em função das características das variáveis independentes relacionadas (HAND *et al.*, 2001).

Modelos preditivos têm o objetivo específico de permitir predizer os valores desconhecidos de variáveis de interesse a partir de valores conhecidos de outras variáveis. O formato da previsão pode ser pensado como um mapeamento de aprendizagem a partir de um conjunto de entrada como um vetor de medições e uma saída como um escalar (HAND *et al.*, 2001).

3.2.3 Tarefas de Mineração de Dados

A mineração de dados é geralmente classificada de acordo com a sua capacidade de realizar determinadas tarefas. Normalmente, a implementação de um modelo também é feita por uma tarefa preditiva ou descritiva. Por exemplo, o agrupamento (ou *clustering*) e as regras de associação produzem modelos descritivos, enquanto a classificação e a regressão geram modelos preditivos (PEÑA-AYALA, 2014). As tarefas mais comuns da mineração de dados são descritas a seguir.

- **Agrupamento (*Clustering*):** É a tarefa com o objetivo de agrupar um conjunto de dados de tal forma que os dados no mesmo grupo (denominado *cluster*) são mais semelhantes entre si do que aos de outros grupos (*clusters*). Essa tarefa difere da classificação, pois não necessita que os dados sejam previamente categorizados. O próprio processo de agrupamento pode gerar rótulos nos dados após os *clusters* formados e cada conjunto formado pode ser visto como uma classe de objetos, a partir da qual podem ser derivadas regras (HAN *et al.*, 2011). Como exemplos da tarefa, na biologia, a semelhança de dados genéticos é usada em *cluster* para inferir estruturas populacionais. Na educação, a análise de agrupamento pode ser usada para identificar grupos de escolas ou alunos com propriedades semelhantes (TAN *et al.*, 2009).
- **Análise das regras de associação:** É usada para descobrir padrões que descrevem características altamente associadas entre os dados, buscando encontrar relações entre variáveis. Os padrões descobertos são normalmente representados na forma de regras de implicação ou subconjunto de características (HAN *et al.*, 2011). Aplicações úteis da análise de associação incluem identificar preferências de consumidores por determinados produtos comprados juntos, descoberta de genes que possuam funcionalidade associada, entre outros (TAN *et al.*, 2009).
- **Detecção de anomalias:** É a tarefa de identificar observações cujas características sejam significativamente diferentes do resto dos dados (*outliers*) que podem ser interessantes ou erros de dados que requerem mais investigação. Aplicações de detecção de anomalias incluem detecção de fraudes, intromissões na rede ou padrões incomuns em doenças (TAN *et al.*, 2009).

- **Classificação:** É a tarefa de organizar objetos em uma entre diversas categorias pré-definidas. Nessa tarefa, o modelo analisa o conjunto de dados fornecidos, na qual cada dado já contém o rótulo, indicando a qual categoria ele pertence, a fim de "aprender" como classificar novos dados. Por exemplo, um programa de e-mail pode tentar classificar um e-mail como "legítimo" ou como "spam", usando a classificação baseada em e-mails anteriormente recebidos e rotulados (HAN *et al.*, 2011).
- **Regressão:** É uma metodologia estatística que é mais frequentemente usada para previsão numérica, embora outros métodos existam com essa finalidade (WITTEN *et al.*, 2011). A regressão também engloba a identificação de tendências de distribuição com base na informação disponível. O objetivo é tentar encontrar uma função que modele os dados com o menor erro possível. Um exemplo é usar um modelo de regressão para estimar as vendas de um determinado produto em um período, a partir de dados de vendas anteriores (HAN *et al.*, 2011).

A principal diferença entre ambos os modelos preditivos é que, enquanto a classificação prevê rótulos categóricos (discretos, não ordenados) para os dados, a regressão estabelece modelos de funções com valores contínuos.

Todas essas tarefas apresentadas podem usar o mesmo banco de dados de maneiras diferentes e exigem o desenvolvimento de inúmeras técnicas de mineração de dados. Devido à diversidade de aplicações, novas tarefas de mineração continuam a emergir, tornando a mineração de dados um campo dinâmico e de rápido crescimento (HAN *et al.*, 2011).

Com relação as tarefas e técnicas mais utilizadas especificamente para a mineração de dados provenientes de aplicações educacionais, o estudo de Peña-Ayala (2014) que analisou 242 trabalhos entre 2010 a 2013, apontou que a classificação foi a tarefa mais usada nos estudos com 42,15% dos trabalhos seguida pelo agrupamento (26,86%), regressão (15,29%) e regras de associação (6,61%). Mais recentemente, Saa et al. (2019) analisaram 218 artigos entre 2009 e 2018, no âmbito da mineração de dados para analisar as principais técnicas preditivas mais comumente aplicadas no contexto educacional. O autor agrupou as técnicas preditivas em sete grupos de algoritmos, juntamente com a sua frequência de uso nos trabalhos de pesquisa analisados. Como pode ser observado no Quadro 3, as categorias de

algoritmos de mineração de dados mais comumente utilizados são: Árvore de Decisão; Classificadores Naïve Bayes e as Redes Neurais Artificiais.

Quadro 3. Técnicas preditivas mais comumente usadas

Algoritmo	Frequencia	Porcentagem (%)
Árvore de decisão	35	24,8
Classificadores Naïve Bayes	14	9,9
Redes Neurais Artificiais	13	9,2
Regressão	12	8,5
Maquina de vetor de suporte (SVM)	9	6,4
K-vizinhos mais próximo	8	5,7
K-Means	3	2,1
Outros algoritmos	47	33,3

Fonte: Saa et al. (2019)

Ainda segundo os autores, os resultados apresentados estão em consonância ao estudo anterior realizado por Peña-Ayala (2014), no qual mostrou que a tarefa de classificação e suas técnicas são as mais utilizadas em estudos que aplicam a mineração de dados no contexto educacional. Tendo em vista esses resultados, os autores sugerem pesquisas adicionais para se referir a outros algoritmos de mineração de dados que possam acrescentar conclusões mais significativas e razoáveis ao contexto educacional (SAA et al., 2019).

Após a definição das principais tarefas de mineração de dados e as técnicas de predição que vêm sendo utilizadas na literatura, deve-se então, escolher o método, a técnica e/ou algoritmo para realizar o processo de mineração. Para cada tarefa, várias opções podem ser testadas ou combinadas na busca de resultados mais apropriados para o problema tratado.

Nesta tese, o estudo busca diagnosticar as causas dos problemas educacionais baseado em modelos preditivos. Para isso, a técnica de regressão foi utilizada para o desenvolvimento desses modelos a fim de analisar os problemas educacionais.

Em Elbadrawy et al. (2015), os autores utilizaram métodos de regressão para prever com precisão a nota de um aluno em uma determinada atividade do curso. O objetivo foi identificar os alunos que estão em risco de reprovação. A partir das informações acadêmicas dos alunos, os modelos propostos foram capazes de estimar 80% de acerto na predição.

No trabalho de Fu et al (2012), foram utilizados modelos de regressão para estimar a relação entre a personalidade de uma pessoa e o seu desempenho escolar. Foi utilizada uma abordagem baseada em Support Vector Regression (SVR), para

encontrar as correlações a partir dos dados amostrais fornecidos, como por exemplo dados do perfil do estudante, comportamentais e biológicos. Com resultado o desempenho previsto obteve 80% de precisão, mostrando que existem correlações entre o desempenho do estudante e suas características de personalidade. O estudo mostrou ainda que o SVR é um método correto para explorar correlações de personalidade.

Mesmo apresentando resultados satisfatórios quando da sua aplicação em contextos educacionais, a regressão ainda é pouco aplicada em projetos de mineração de dados. Diferentemente da técnica de classificação, ainda dominante no cenário da EDM, a regressão mostra-se uma técnica relevante na literatura de aprendizagem de máquina e que ao longo do tempo tem sido aplicada a problemas educacionais para estimativa numérica e predição.

3.2.4 Modelos de Regressão

Na regressão, a obtenção de um modelo baseia-se em um conjunto de exemplos que descrevem uma função não conhecida. Esse modelo é utilizado para predizer o valor do atributo-meta de novos exemplos. O objetivo da regressão é encontrar uma relação entre um conjunto de atributos de entrada (variáveis de entrada ou variáveis preditoras) e um atributo meta contínuo (WEISS,1997).

Em problemas de regressão, dado um exemplo z = (x; y), em que x é um vetor de n variáveis preditoras, x = (x_1...x_n), também denominadas de variáveis independentes, o objetivo é predizer o valor da variável resposta, ou dependente, y. Em outras palavras, consiste na obtenção de uma equação que explique a variação do y pela variação do x. Para isso, verifica-se a existência de uma relação funcional entre essas variáveis, por meio de um gráfico denominado diagrama de dispersão.

O gráfico pode revelar o comportamento entre y e x de várias formas como relacionamento linear, quadrático, exponencial, logarítmico, entre outros. Portanto, para definir o modelo para explicar o problema de regressão, deve-se verificar qual curva mais se aproxima dos pontos representados no diagrama de dispersão, como mostra a Figura 6.

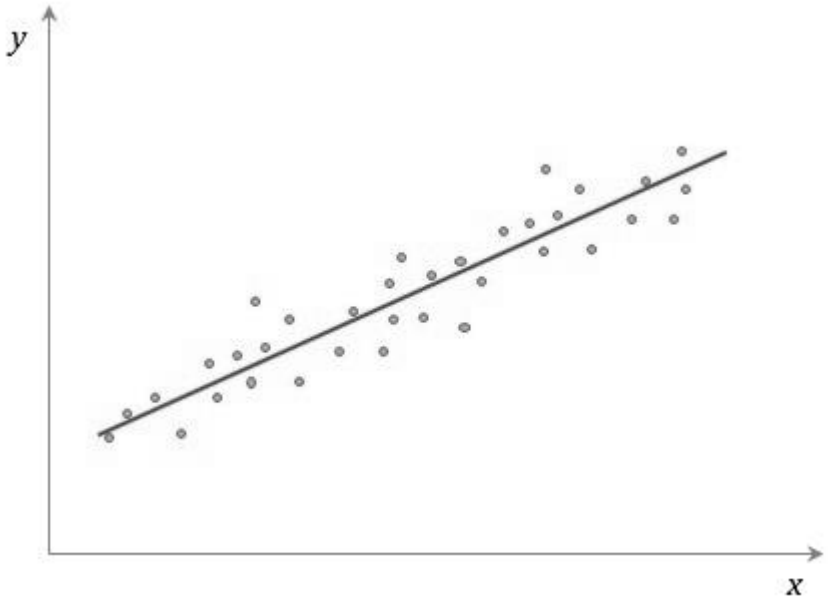


Figura 6. Exemplo de relação linear entre y e x através do diagrama de dispersão
Fonte: Dosualdo (2003)

No entanto, a curva do modelo matemático gerado pode não se ajustar perfeitamente aos pontos. Assim, o objetivo da regressão é obter um modelo que melhor se ajuste aos valores observados de y em função de x. Conforme ilustrado na Figura 6, o relacionamento entre as variáveis possui forma linear. Esse grau de associação entre y e x pode ser medido pelo coeficiente de correlação. A medida de correlação mais comum que reflete o grau de relacionamento linear entre duas variáveis é o coeficiente de correlação de Pearson (r) (STANTON, 2001). O coeficiente r pode assumir valores entre -1 e 1, o que significa r=1 uma correlação perfeita negativa. Quanto mais próximo de 0 o r, torna-se mais fraca a correlação. Porém como este índice avalia apenas a relação linear entre as variáveis, outras medidas são necessárias quando existe uma dependência não linear nos dados (PEARSON, 1994).

Os modelos paramétricos assumem que a forma do relacionamento funcional entre y e x é conhecida, reduzindo o problema para a estimação de um conjunto de parâmetros. Como visto, a regressão linear pode ser usada quando a relação entre as variáveis explicativas e a variável resposta pode ser aproximada por uma reta. Por outro lado, técnicas não paramétricas de regressão fazem apenas algumas suposições sobre a função, a qual será estimada a partir dos dados.

3.2.4.1 Modelos Paramétricos

Os modelos paramétricos caracterizam –se por serem um método utilizado para obter a relação funcional das variáveis. São baseados na obtenção de uma equação estimada, de forma que a distância entre os pontos no diagrama de dispersão e os pontos da curva do modelo matemático sejam os menores possíveis. Esse método é conhecido como Método dos Mínimos Quadrados (MMQ). Este consiste em fazer com que a soma dos erros quadráticos seja a menor possível (MONTEGOMERY, 2012). A Equação 3.1 apresenta o modelo de Regressão Linear (RL), definida para esta situação.

$$\gamma = \beta_0 + \beta_1 x_1 + \varepsilon_1 \tag{3.1}$$

Onde β_0 representa o intercepto da reta com o eixo dos y, $\beta_i = (\beta_i, \ldots, \beta_p)$ é um vetor de parâmetros que representa a variação de y em função da variação de x_i, $x_i = (x_i,\ldots,x_p)$ um vetor de variáveis explicativas, e $\varepsilon = (\varepsilon_0, \ldots, \varepsilon_p)$ é um vetor de erro aleatório. Portanto, os parâmetros β é estimado, minimizando uma função baseada no MMQ que é dada pela Equação 3.2 a seguir.

$$\sum_{i=1}^{n} \varepsilon_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.2}$$

Onde n é o tamanho da amostra, y_i é a variável resposta real e $\hat{y}_i$ a variável resposta estimada. A reta ajustada pelo MMQ é fortemente influenciada pelos *outliers*, pois o método minimiza a soma dos resíduos ao quadrado.

Na estimação da predição, utilizando uma regressão linear múltipla por exemplo, o ajuste da reta em uma distribuição de pontos é uma forma de relacionamento paramétrico entre as variáveis. A Equação 3.3 apresenta o modelo de Regressão Linear Multipla (RLM).

$$\gamma_i = \alpha + \beta^1 x^1{}_i + \ldots + \beta_p x_{pi} + \varepsilon_i (i = 1,2,\ldots,n) \tag{3.3}$$

Onde γ_i é a variável dependente que tem uma distribuição específica de média $\mu_i = E(\gamma_i)$ e variância constante. O α é a representação do intercepto da reta com eixo do Y. Os $(\beta_1,\ldots,\beta_p)$ são os coeficientes de variação de x em relação a y. Os

x_i representa o vetor de variáveis idependentes .Por fim, o ε_i representa um vetor de erro aleatório. Onde o ε_i é calculado utilizando a Equação 3.4

$$\varepsilon_i = y_i - (\alpha + \beta x_i) \tag{3.4}$$

Dessa forma, para ajustar um modelo que possa trazer boas estimativas dos parâmetros β na presença de *outliers*, ou quando y não segue uma distribuição normal, outros métodos podem ser utilizados para obter a relação linear entre as variáveis. Como por exemplo, tem-se a Regressão Linear Robusta (RLR), (MONTEGOMERY, 2012). Este modelo de regressão utiliza M-estimadores, no qual o vetor β é estimado minimizando uma função critério baseada em ρ, conforme apresentado na Equação 3.5.

$$\sum_{i=1}^{n} p\left(\frac{y_i - \beta x_i}{\sigma}\right) \tag{3.5}$$

Onde n é o tamanho da amostra, y_i é a variável resposta real, x_i a variável explicativa, β o parâmetro, σ é um estimador robusto e ρ uma função particular.

Um outro modelo robusto aos inconvenientes causados pela sensibilidade a *outliers* em RL, é a Regressão Quantilica (RQ) (KROENKER et al., 1978). A RQ captura mudanças de localização e inclinação das curvas permitindo distinguir diferenças de importância e de relação entre as variáveis sobre a mediana e sobre os quantis da variável dependente. Possui o vetor de parâmetros e os resíduos associados ao θ-ésimo quantil, $\theta \in (1,0)$. O estimador β_0 é encontrado a partir da minimização da função objetivo:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\,(y_i - x_i\beta_0) \tag{3.6}$$

Onde ρ_0 é a função de *check* definida por:

$$\rho_0(\varepsilon) = \begin{cases} \theta\varepsilon & e, \varepsilon \geq 0 \\ (\theta - 1)\varepsilon & , e\varepsilon < 0 \end{cases} \tag{3.7}$$

A Regressão Ridge (RD), proposta por Hoerl (1970), prevê a obtenção de uma variância menor que a dos mínimos quadradados adicionando uma pequena quantidade positiva, ou seja, viciando o estimador de forma:

$$b(k) = W^T W + kI)^{-1} W^T y \tag{3.8}$$

A esse tipo de estimador atribui-se o nome estimador “Ridge”. Para obtê-lo deve-se encontrar um valor de k. Este é uma matriz diagonal com elementos, $(k_{1,} \ldots, k_p)$, $k_i \geq 0$ para $\forall i$. São várias as propostas para obter o estimador “Ridge” por meio de diferentes quantidades positivas adicionadas na diagnal da matriz $W^T W$, sendo mais usual esses valores serem todos iguais.

Ainda fazem parte dos modelos paramétricos os modelos não – lineares, nos quais o vetor de parâmetros pode ter uma relação não linear nos preditores. Sendo decrito na Equação 3.9.

$$y_{\mathrm{i}} = f\,(x_{\mathrm{i}}\,, y) + \varepsilon_{\mathrm{i}} \tag{3.9}$$

Onde $f\,(x_{\mathrm{i}}\,, y)$ é uma função não-linear e ε_{i} o erro aleatório associado. Esses modelos não serão abordados nesse trabalho.

3.2.4.2 Modelos Não-Paramétricos

Os modelos não-paramétricos de regressão caracterizam-se por não permitir saber de forma antecipada, a forma da função que está sendo aplicada. Na abordagem paramética discutida anteriormente, a maneira mais comum de estimar a função de regressão é por meio dos modelos linear e não-linear, permitindo previsões dos valores de y. No entanto, pode-se especificar formas funcionais para a função de maneira não apropriada. Esse problema, busca ser amenizado na abordagem não-paramétrica.

A abordagem não-paramétrica é flexível, e objetiva explorar o aprendizado sobre a função sem restringí-la a modelos estabelecidos *a priori*. A idéia é permitir que os dados ditem o formato da curva de regressão no processo de estimação. Esse formato se dará por meio de um suavizador, que representa uma ferramenta para resumir a tendência da medida da resposta que seja menos variável que o próprio y coletado da amostra de dados. Na literatura existem diferentes tipos de suavizadores, como os

baseados em *kernels, splines* entre outros (EFROMOVICH, 1999).

A Regressão de *Kernel* (RK) é um método no qual o valor resposta para um ponto de teste é estimado usando uma média ponderada das amostras de treinamento adjacentes (MONTGOMERY, 2012). Os pesos são normalmente obtidos aplicando-se uma função *Kernel* baseada na distância em cada uma das amostras. A idéia é que ao estimar o suavizador de Kernel, é desejável dar maior peso às observações do conjunto de treinamento que estão próximas ao ponto de consulta $\mathcal{X}_q$.

Seja $\mathcal{X}_q$ um ponto de consulta. Considere d(x_q, x_i) a distância entre uma observação x_i do conjunto de treinamentos e o ponto x_q. Para ponderar a vizinhança de x_q, é necessária uma função *kernel* *K*(d((x_q, x_i)), que atribui maior peso às observações próximas ao foco x_q e, em seguida, cai simetricamente e suavemente quando d cresce. A função de Kernel gaussiano é definida pela Equação 3.10.

$$K_g\left(d\left(x_q, x_\mathrm{i}\right)\right) = \frac{1}{\sqrt{2\pi}}\, e^{\left(\frac{-d\left(x_q, x_\mathrm{i}\right)}{2h}\right)} \tag{3.10}$$

Onde d(x_q, x_i) é a distância Euclidiana do quadrado entre x_q e a localização de interesse x_i . Nesta função *kernel,* a largura de banda *h* é o desvio padrão de uma distribuição normal centrada em x_i . A predição usando um estimador de *kernel* é dada pela Equação 3.11.

$$\hat{\mathrm{y}}_q = \sum_{i=1}^{n} \mathcal{W}_\mathrm{i} Y_\mathrm{i} \tag{3.11}$$

Onde n é o tamanho da amostra, $\mathrm{y_i}$ é a resposta real e $\mathcal{W}_\mathrm{i}$ é obtido, como mostrado na Equação 3.12.

$$\mathcal{W}_\mathrm{i} = \frac{K(d\left(x_{q,}\, x_i\right))}{\sum_{i=0}^{n} K(d\left(x_{q,}\, x_\mathrm{i}\right))} \tag{3.12}$$

O *Support Vector Regression* (SVR) é um método não-paramétrico, desenvolvido para resolver problemas lineares e não-lineares (DUCKER et al., 1997). Esse método é capaz de construir aproximações de *spline* de dados independente do número de dimensões de entrada em relação a complexidade do treinamento e com apenas complexidade linear, ao contrário da exponencial em um método

convencional.

A formulação do problema SVR fornece como alternativa trabalhar em um espaço de alta dimensionalidade. Assim, pode-se realizar um mapeamento não-linear dos dados de entrada para um espaço de deimensões maiores. Para o caso da regressão linear, a aproximação num ponto de consulta x_q é formada conforme a Equação 3.13.

$$\hat{y}_q = \left(x_i, x_q\right) + b \; com \; b \in \mathcal{R} \tag{3.13}$$

Onde $\left(x_i, x_q\right)$ denota o produto escalar. Para o caso da aplicação de SVR como regressão não-linear a equação consiste na seguinte forma, conforme a equação 3.14.

$$\hat{y}_q = (\emptyset\left(x_i\right), \emptyset(\, x_q)) + b \; com \; b \in \mathcal{R} \tag{3.14}$$

Onde $\emptyset$ consiste numa função não linear que mapeia o espaço de entrada para um espaço de características dimensional superior (infinito). Nesse caso, o algoritmo SVR envolve o uso de multiplicadores Langrageanos, os quais dependem exclusivamente de produtos de pontos de $\emptyset$. Isso pode ser feito por meio de funções do *Kernel*, definidas como:

$$K = \left(x_i, x_q\right) = (\emptyset\left(x_i\right), \emptyset\left(x_q\right)) \tag{3.15}$$

Dentre as funções *kernel* mais utilizadas no algoritmo SVR destacam-se a linear, polinomial, sigmoidal e a gaussiana (DUCKER et al., 1997)

3.2.4.3 Modelos Combinados (*Ensemble)*

Em AM, o aprendizado de conjuntos consiste de métodos eficientes que integram vários modelos básicos para gerar a saída final. Esse método criou grande popularidade devido ao seu excelente desempenho de generalização. Os modelos *ensemble* geralmente resultam em uma melhor acurácia do que os indivíduos que os compõem. As razões por trás de desenvolver modelos são três (DIETTERICH, 2000). Em primeiro lugar, os dados de treinamento podem não fornecer informações suficientes para selecionar o melhor modelo único. Portanto, a integração de tais modelos com desempenho equivalente pode ser uma escolha melhor. Em segundo

lugar, os conjuntos podem compensar a imperfeição no processo de busca individual. Em terceiro lugar, na prática real, pode não haver uma verdadeira função alvo. Conjuntos podem fornecer uma aproximação relativamente boa e, portanto, resultam em um melhor desempenho e generalização (DIETTERICH, 2000).

Um modelo *ensemble* é uma técnica resultante da integração de dois ou mais algoritmos de tipos semelhantes ou diferentes com o objetivo de criar um sistema mais robusto que incorpore o resultado de todas as técnicas (POLIKAR, 2006).

É um paradigma de aprendizado no qual propostas alternativas de solução para um problema, denominadas *componentes*, têm suas saídas individuais combinadas na obtenção de uma solução final. Esta abordagem tem sido amplamente utilizada na última década, tanto para problemas de regressão quanto para problemas de classificação de padrões, uma vez que os *ensembles* são comprovadamente capazes de aumentar a capacidade de generalização e, consequentemente, o desempenho geral do sistema. Intuitivamente, a combinação de múltiplos componentes é vantajosa, uma vez que componentes diferentes podem implicitamente representar aspectos distintos e, ao mesmo tempo, relevantes para a solução de um dado problema (COELHO *et al*., 2009).

Os modelos *ensemble* podem ser formados por técnicas do mesmo tipo e são conhecidos como homogêneos e quando utilizam diferentes modelos de técnicas de AM, são ditos heterogêneos.

Dentre os modelos *ensemble* existentes desatacam-se o *Bagging* (*Bootstrap Aggregation*) e o *Stacking Generalization* ou, simplesmente, *Stacking* (BREIMAN, 1996; WOLPERT, 1992).

O *Bagging* tem como objetivo gerar um conjunto de dados por amostragem *bootstrap* dos dados originais. Segundo Maddala (2003), o método *Bootstrap* é uma técnica de reamostragem, com o propósito de reduzir desvios e prover desvios padrão mais confiáveis. O conjunto de dados gera um conjunto de modelos utilizando um algoritmo de AM simples para construção de modelos de regressão. Por fim, generaliza o modelo por meio de uma função de fusão (BREIMAN, 1996). A Equação 3.16 apresenta a função de fusão desse modelo combinado.

$$fbagging(x) = \frac{1}{M} \sum_{\mathrm{i}=1}^{\mathrm{n}} \hat{f}\,(x) \qquad (3.16)$$

Onde $fbagging(x)$ representa a previsão do modelo combinado para o

instante x, e $\hat{f}(x)$ é a predição dada pelo i-*ésimo* regressor construído sob i – *ésima* amostra de *bootstrap* dos dados de treinamento.

O método *stacking* tem como objetivo realizar a previsão usando algoritmos diferentes, e as previsões desses algoritmos serão usadas por um meta-preditor para fazer a combinação dos diferentes modelos (WOLPERT, 1992). O meta-preditor é outro algoritmo de aprendizado de máquina usado para combinar as saídas dos modelos básicos e gerar a previsão do modelo. A Equação 3.17 apresenta a fusão do modelo *stacking*.

$$fstacking(x) = \sum_{\mathrm{i}=1}^{\mathrm{K}} [\, f\, x_{\mathrm{i}} - \sum_{i=1}^{M} \alpha_{\mathrm{i}} * \hat{f}_{\mathrm{i}}(x_{\mathrm{i}})] \qquad (3.17)$$

Onde $f$stacking (x) é a predição do modelo combinado para x, K é o tamanho do combinado de dados treinamento, *M* é a quantidade de regressores do modelo, α é um coeficiente que minimiza o erro e $\hat{f}_{\mathrm{i}}(x_{\mathrm{i}})$ é a predição dada pelo i-*regressor*.

3.2.4.4 Avaliação dos Modelos de Regressão

O processo de aprendizado para problemas de regressão, de forma geral, consiste em induzir uma função $\hat{f}$: X $\rightarrow \mathcal{R}$ em que $\hat{f}$ é denominado regressor, modelo ou hipótese. Deseja-se $f(x) = f(x), \forall x \in X$, em que $f(x)$ representa a função real desconhecida. Porém, na prática, $\hat{f}$ não é idêntica a f, mas uma função que minimiza a diferença entre as duas. Dito isso, busca-se modelos de regressão que traduzam uma função de indução minimizando o erro de generalização

O erro de generalização dos modelos contruídos para problemas de regressão pode ser calculado usando diferentes medidas. Uma dessas medidas é o *Mean Absolute Error* (MAE), ou erro médio absoluto, no qual os erros são tratados igualmente de acordo com a sua magnitude. O MAE é uma medida útil amplamente utilizada em avaliações de modelos (DRAXLER et al., 2014).

Portanto, neste trabalho a métrica MAE será utilizada para avaliar os modelos preditivos. A Equação 3.18 descreve o MAE, onde $\hat{y} = \{\hat{y}_1, \ldots, \hat{y}n\}$ são os valores reais da variável resposta.

$$MAE = \frac{1}{n}\sum_{i=0}^{n} |\hat{y} - y_1| \qquad (3.18)$$

3.2.4.5 Métodos de Seleção de Variáveis

Na descrição de processos reais, um fenômeno é associado a um elevado volume de variáveis. Para fins de utilização dos modelos de regressão para a tomada de decisão, é desejável uma redução da quantidade de variáveis de entrada, visto que variáveis não significativas aumentam a variabilidade do modelo e diminuem a qualidade da predição. Assim modelos com grande quantidade de variáveis podem apresentar aderência satisfatória aos dados modelados, porém podem conduzir a predições insatisfatórias devido ao ruído inserido por algumas variáveis. Atributos irrelevantes, com pouco impacto sobre o modelo, também diminuem a exatidão da predição (GUYSON e ELISEEF, 2003; GAUCHI; CHAGNON, 2001; ALLEN, 1974; THANASSOULIS, 1996).

A redução do numéro de variáveis em um modelo de regressão é denominada seleção de variáveis aplicada à modelagem de sistemas. Os métodos devem ser utilizados adequadamente, sendo avaliados e ajustados de forma iterativa, buscando uma solução mais adequada (GUYSON; ELISEEFF, 2003; ANDERSEN; BRO, 2010). Dentre os métodos para seleção mais referenciados pela literatura destacam –se os métodos *Foward Selection, Backward Elimination, e Stepwise Regression.*

O *Foward Selection* inicia sem variáveis na equação, incorporando ao modelo variáveis estatisticamente significativas, uma a uma (medidas por meio de teste estatístico). *Backward Elimination,* ao contrário, conta com todas as variáveis do modelo, removendo sucessivamente as variáveis menos significativas.

O Stepwise Regression resulta de uma combinação dos métodos *Backward e Forward.* A metodologia *Stepwise Regression* é uma das mais utilizadas. Essa família de métodos apresenta vantagens como simplicidade e ampla difusão. Todavia, deve-se ter ciência que os procedimentos podem indicar um modelo que, não

necessariamente, é tido como modelo ótimo. Tal colocação é corroborada por Mantel (1970), segundo o qual um modelo excelente pode passar despercebido devido à restrição de adicionar ou remover apenas uma variável por vez.

3.3 CONTEXTUALIZAÇÃO SOBRE MINERAÇÃO DE DADOS EDUCACIONAIS

3.3.1 Etapas da Mineração de Dados Educacionais

Um dos objetivos do sistema educacional é conseguir, quando possível, um maior nível de qualidade no ensino. As instituições educacionais acumulam uma enorme quantidade de dados sobre alunos, educadores, materiais didáticos, curriculos e processos de aprendizagem como um todo. A coleta dos dados educacionais dentro do âmbito institucional depende de regulamentação legal e das necessidades e políticas institucionais específicas. Analisar dados é um trabalho oneroso, e fazê-lo sem ferramentas poderosas de Mineração de Dados, pode torná-lo ainda mais difícil. A mineração de dados envolve técnicas para encontrar e descrever padrões estruturais em dados, como uma ferramenta para ajudar a explicar os dados e fazer previsões a partir deles (CLEMMENSEN et al., 2011).

O processo de aplicação de técnicas e métodos de Mineração de Dados em dados educacionais pertence à área científica específica. A Mineração de Dados Educacionais (EDM) pode ser definida como uma disciplina preocupada com o desenvolvimento de métodos para explorar os tipos únicos de dados dentro do contexto educacional (ROMERO e VENTURA, 2013). Com a utilização de técnicas e métodos de mineração de dados no domínio educacional, é possível descobrir conhecimentos e padrões ocultos que podem ser utilizados para a tomada de decisões relacionadas à melhoria do processo educacional de ensino/aprendizagem (MILINKOVICK e VUJOVIC, 2019).

O conhecimento obtido por meio da EDM, pode ser usado para oferecer sugestões aos gestores acadêmicos para aprimorar seu processo de tomada de decisões, melhorar o desempenho acadêmico dos alunos, diminuir as taxas de reprovação e evasão, compreender melhor o comportamento dos alunos e auxiliar os docentes a melhorar suas estratégias de ensino e para a construção de modelos de regressão para prever do aluno em termos de notas e porcentagem (MILINKOVICK e VUJOVIC, 2019). Os métodos de EDM também podem ser usados para categorizar os alunos que precisam de apoio, para analisar a aprendizagem dos alunos e agrupá-

los de acordo com seus pontos fortes e fracos para atividades relacionadas à aprendizagem.

Assim, é possível compreender de forma mais eficaz e adequada, os alunos, como eles aprendem, o papel do contexto no qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. Por exemplo, é possível identificar em que situação um método de aprendizagem proporciona melhores benefícios educacionais ao aluno. Também é possível verificar se o aluno está desmotivado ou confuso e, assim, permitir aos professores melhorar seus métodos de ensino para oferecer melhores condições de aprendizagem (BAKER *et al*., 2011).

O processo de EDM converte dados brutos de sistemas educacionais em conhecimento que pode ser usado por desenvolvedores de *software educacional*, professores, pesquisadores educacionais entre outros. Esse processo não difere de outras áreas de aplicação da mineração de dados porque ele se baseia nos mesmos passos do processo de mineração de dados em geral, conforme mostra a Figura 7 (GARCIA *et al*., 2011):

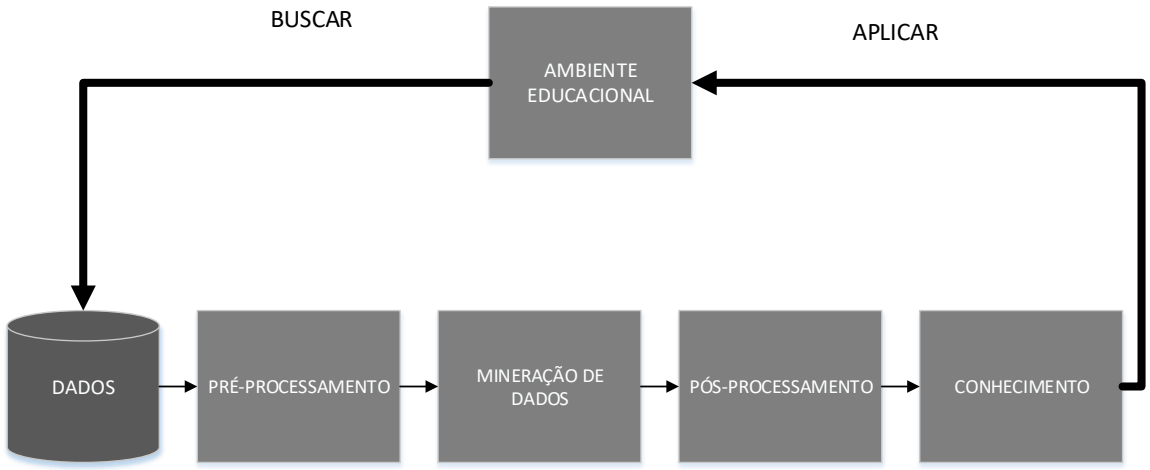


Figura 7. Etapas da Mineração de Dados Educacionais
Fonte: Garcia *et al*. (2011) apud Ramos (2016)

Para que o processo de conhecimento proposto no ciclo seja relevante, é necessário que sua utilização não seja limitada apenas aos atores educacionais, mas também retroalimentar o sistema de ensino, contribuindo para a promoção de melhorias no processo. As etapas da EDM precisam acontecer de maneira efetiva e consistente, possibilitando que os resultados intermediários e finais sejam adequados. A descrição de cada componente e as etapas do processo serão mostradas a seguir:

- **Ambiente Educacional -** constitui-se em sala de aula tradicional, ensino baseado em computador, ou educação baseada na web e sistemas de informação que lhes dêem suporte (gestão de aprendizagem, tutor inteligente, ou sistemas hipermídios). Diferentes tipos de dados podem ser coletados para resolver diferentes problemas educacionais (ROMERO e VENTURA, 2010). Esses dados podem vir de diferentes fontes, incluindo dados administrativos, observações de campo, avaliações padronizadas, censitários, questionários, e medições recolhidas a partir de experimentos (ROMERO e VENTURA, 2013);
- **Dados -** São as fontes de informação geradas pelos alunos, professores, instituições de ensino, ambientes de e-*learning* que podem fornecer rápidas e importantes compreensões a respeito do desempenho, evasão e retenção, infraestrutura, comportamentos e os níveis de motivação e participação dos alunos. Essas compreensões podem sugerir mudanças significativas nas metodologias de ensino/aprendizagem (ROMERO et al., 2008);
- **Pré-Processamento -** Os dados obtidos com o ambiente educacional têm que primeiro ser pré-processados para serem transformados em um formato apropriado para a mineração. Dentre as tarefas a serem executadas estão: a limpeza, seleção de atributos, transformação, normalização, integração, entre outros (RAMOS, 2016). Em contextos educacionais, o pré-processamento de dados é uma tarefa importante e complicada, às vezes demandando mais da metade do tempo total gasto na resolução do problema de mineração de dados (RAMOS, 2016);
- **Mineração de Dados -** É a etapa central que identifica o processo como um todo. No contexto educacional, técnicas clássicas de mineração de dados, como classificação, agrupamento e as técnicas de análise de associação, vêm sendo aplicadas com sucesso no domínio da educação (PENA AYALA, 2014).

No entanto, os sistemas educacionais têm características especiais que requerem um tratamento diferente do problema de mineração clássico. Por exemplo, os métodos de MD hierárquicos, a combinação de modelos e a modelagem

longitudinal de dados tem que ser usados em EDM. Por consequência, são necessárias algumas técnicas específicas de MD para lidar com os problemas de aprendizagem e outros relacionados ao aluno. No entanto, a EDM, ainda, é uma área de pesquisa emergente, e a cada dia, novas contribuições a esta área vem sendo desenvolvidas, o que irá resultar em um melhor entendimento dos desafios específicos da educação, ajudando os pesquisadores envolvidos na área a adotar e desenvolver novas técnicas (ROMERO e VENTURA, 2013):

- **Pós – Processamento -** É a etapa final na qual os resultados obtidos ou modelo são interpretados e usados para tomar decisões sobre o ambiente educacional. Os modelos obtidos pelos algoritmos de EDM têm que ser compreensíveis e úteis para o processo de decisão. Outra forma de compreensão dos modelos é a utilização de técnicas de visualização, as quais são muito úteis para mostrar os resultados, e por fim os sistemas de recomendação são muito úteis na apresentação dos resultados e informações, explicações, recomendações e comentários para usuários leigos em EDM. Desta forma os resultados dos modelos podem ser mostrados na forma de lista de sugestões ou conclusões sobre os resultados e como aplicá-los (ROMERO e VENTURA, 2013); e
- **Conhecimento -** quando extraído deve facilitar e melhorar a aprendizagem como um todo, não apenas transformando dados em conhecimento, mas também filtrando o conhecimento extraído para a tomada de decisão (ROMERO *et al*., 2008).

A EDM surge como um paradigma orientado para projetar modelos, tarefas, métodos e algoritmos para explorar dados de contextos educativos. Também busca descobrir padrões e fazer previsões que caracterizam os comportamentos dos alunos e suas realizações, o conteúdo de conhecimento de domínio, avaliações, funcionalidades e aplicações educacionais. A fonte de informação é armazenada em repositórios gerados e gerenciados por modalidades tradicionais de ensino, ensino aberto, e na educação a distância (PEÑA-AYALA, 2014).

3.3.2 Áreas Relacionadas a EDM

A EDM é uma área interdisciplinar, a qual inclui áreas como a recuperação da informação, sistemas de recomendação, visualização de dados, análise de redes

sociais (SNA), psicopedagogia, psicologia cognitiva, psicometria, dentre outras, em uma visão mais ampla.

Segundo Romero e Ventura (2013), a EDM pode ser visualizada como a combinação de três principais áreas: Ciência da Computação, Educação e Estatistica. Na interseção dessas três áreas surge a relação com algumas subáreas estreitamente relacionadas com EDM, como a área de educação, *E-learning*, Mineração de Dados, Aprendizagem de Máquina (AM) e *Learning Analytics* (LA). Na Figura 8, a seguir, são explicitadas as áreas multidisciplinares que estão relacionadas a EDM.

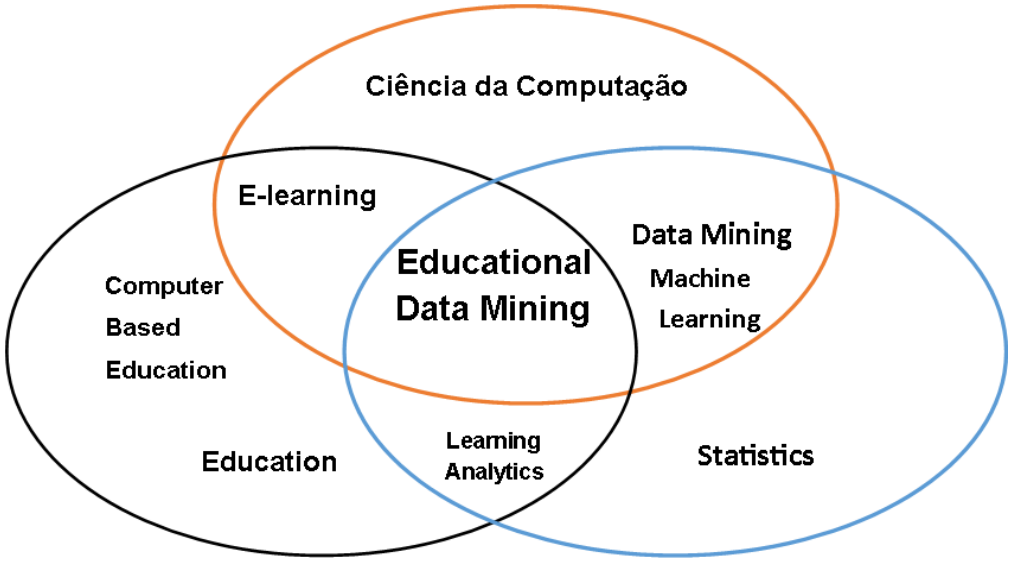


Figura 8. Principais áreas relacionadas a EDM
Fonte: Romero e Ventura (2013)

De todas as áreas interdisciplinares mencionadas, a que mais se relaciona com a EDM é a *Learning Analytics* (LA), tendo como técnicas mais utilizadas a inferência estatistica, a visualização de dados, análise de redes sociais, análise de sentimento e análise de influencia (RODRIGUES, 2016).

Embora boa parte dos trabalhos de EDM utilize as tarefas clássicas da mineração de dados, os sistemas educacionais possuem características diferenciadas que exigem o desenvolvimento de novas abordagens de mineração de dados (ROMERO e VENTURA, 2010). Pesquisadores da área de EDM, não só utilizam as tarefas tradicionais de mineração de dados, mas também propõem desenvolver e alocar métodos e técnicas extraídas de uma variedade de áreas relacionadas a EDM, como por exemplo, a estatistica, aprendizagem de máquina e psicometria.

Assim, diversas abordagens, com a utilização destas tarefas, vêm sendo

desenvolvidas no âmbito educacional. O Quadro mostra as abordagens listadas por Romero e Ventura (2013).

Quadro 4. Abordagens de EDM

Aplicação	Descrição
Predição do desempenho do estudante	Estimação de desempenho de alunos baseado em variáveis comportamentais
Investigação científica	Testes e comprovações de teorias de aprendizagem, com o objetivo de formular novas hipóteses científicas e assim por diante
Fornecendo feedback para apoio aos professores	Fornecimento de *feedback* para apoiar educadores no processo de decisão sobre a melhor forma de aprendizagem dos alunos e permitir tomar ações apropriadas, proativas e de reparação.
Aprendizagem personalizada/adaptativa para alunos	Para adaptar-se automaticamente a diversos perfis de aprendizagem, formas de navegação e conteúdo.
Recomendação para estudantes	Para fazer recomendações aos estudantes com relação a suas atividades ou tarefas, links para visitas, atividades a serem feitas e cursos a serem realizados.
Criação de alertas para estudantes	Para monitorar o progresso da aprendizagem dos alunos, para detecção em tempo real do comportamento indesejável dos alunos, tais com baixa motivação, uso idevido, abandono, e assim por diante.
Modelagem do usuário/estudante	Para o desenvolvimento de modelos cognitivos dos estudantes, representando suas competências e seus conhecimentos
Modelagem do domínio	Para descrever o domínio de instrução em termos de conceitos, habilidades, itens e aprender suas interações.
Agrupamento de perfis de alunos	Para criar grupos de estudantes de acordo com suas características pessoais, dados de aprendizagem.
Construindo material didático	Para ajudar a instrutores e desenvolvedores para realizar o processo de construção/desenvolvimento de material didático e aprendizagem de conteúdos automaticamente.
Estimação de parâmetros	Para inferir parâmetros de modelos probabilísticos a partir de determinados dados para prever a probabilidade de eventos de interesse.

Fonte: Rodrigues (2016)

De acordo com Rodrigues (2016), um trabalho na área de EDM, nem sempre é classificado de acordo com as abordagens mencionadas no Quadro 4. Geralmente os trabalhos perpassam por mais de uma abordagem educacional. Além disso cada

abordagem educacional é desenvolvida por uma ou mais tarefas de MD.

3.3.3 Métodos em EDM

Existem muitos métodos utilizados em EDM, os quais são originalmente da área de MD. Entretanto, de acordo com Baker (2010), muitas vezes, esses métodos precisam ser modificados por causa da necessidade de considerar a hierarquia (em diversos níveis) da informação. Alem disso, existe uma falta de independência estatistica nos tipos de dados encontrados ao coletar informações em ambientes educacionais (BAKER *et al*., 2011).

3.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este capitulo discutiu as diferentes tarefas que vêm sendo utilizadas em pesquisas na área de Mineração de Dados Educacionais. Percebe-se, por meio deste capítulo que existe um grande número de pesquisas no contexto de EDM, as quais utilizam tarefas de classificação, seguidamente por trabalhos que utilizam tarefas de agrupamento. Essa constatação reforça, ainda mais, a necessidade do desenvolvimento de mais pesquisas com a aplicação de técnicas de regressão para a resolução de problemas de previsão dentro do contexto de EDM.

No próximo capítulo serão apresentados os principais trabalhos e técnicas, identificadas por meio de um mapeamento sistemático da literatura, o qual buscou identificar um conjunto de trabalhos que utilizaram abordagens quantitativas para a predição de problemas educacionais.

4 MAPEAMENTO SISTEMÁTICO DA LITERATURA

Neste capítulo, são apresentadas as etapas e resultados do mapeamento sistemático realizado nesta tese. O objetivo é identificar trabalhos relacionados e lacunas de pesquisa que tenham relação com a aplicação de abordagens/técnicas para predição de problemas no âmbito educacional.

4.1 MÉTODO DO MAPEAMENTO SISTEMÁTICO DA LITERATURA (MSL)

O estudo do MSL é projetado para mostrar uma visão ampla dos estudos primários existentes sobre um tema de pesquisa específico, o qual busca identificar as evidências disponíveis a respeito do tema escolhido. De acordo com Kitchenham e Charters (2007), um MSL é um estudo secundário que tem como objetivo identificar e classificar as pesquisas a um tema amplo. Os resultados de um MSL ajudam a identificar lacunas na área, e são capazes de sugerir pesquisas futuras para posicionar adequadamente novas atividades de pesquisa (KITCHEHAM e CHATERS 2007).

O processo de um MSL envolve três fases principais: Planejamento do MSL, Condução do MSL e Publicação dos Resultados (KITCHEHAM e CHATERS 2007). Essas fases, bem como suas atividades são conduzidas de modo interativo. A Figura 9 apresenta as fases e as atividades do processo de MSL aplicados nesta Tese.

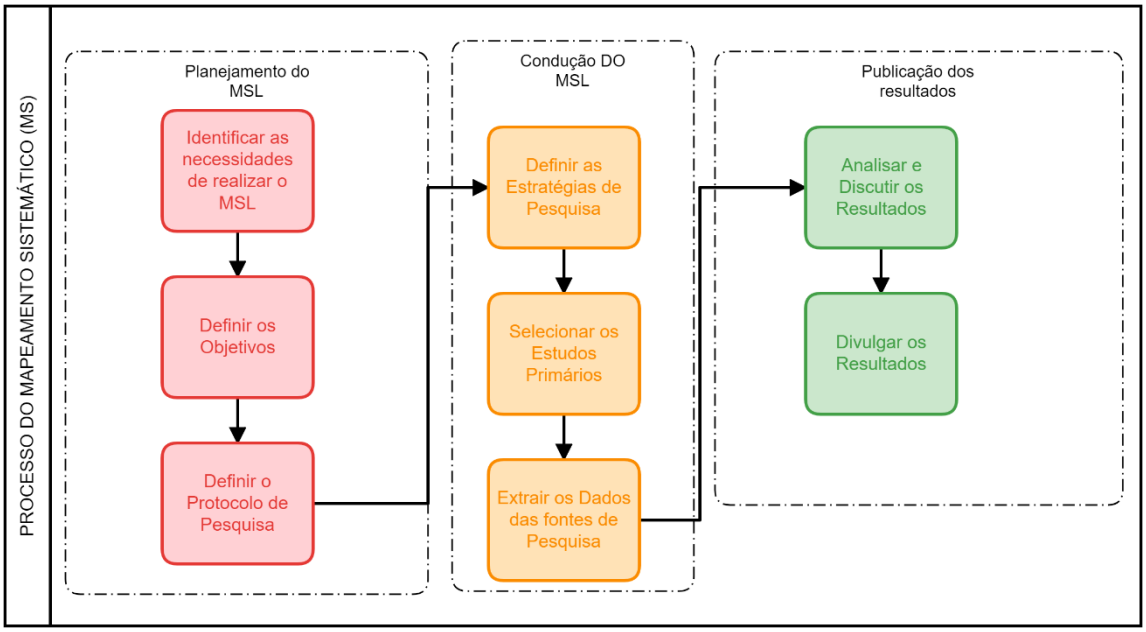


Figura 9. Processo de MSL aplicado nesta Tese
Fonte: Kitcheham e Chaters (2007)

4.2 PLANEJAMENTO DO MSL

Esta etapa tem como objetivo identificar a real necessidade, ou seja, a motivação para a execução de um MSL. É importante antes de iniciar o processo de planejamento do MSL, identificar se já existem estudos secundários no mesmo tema. A seguir busca-se a necessidade da realização deste MSL tendo em vista trabalhos relacionados, bem como descrever o protocolo de pesquisa definido, as etapas e procedimentos realizados no processo (KITCHENHAM e CHATERS,2007).

4.2.1 Necessidade da Pesquisa

Ao longo do tempo, a Educação vem vivenciando diversos problemas que dificultam a eficiência e a eficácia de seus processos. Dentre os problemas mais conhecidos na atualidade estão: a evasão/retenção e os níveis baixos de aprendizagem/desempenho escolar. Esses problemas são apontados como responsáveis pela queda na qualidade dos sistemas educacionais nas diversas modalidades de ensino em todo mundo, principalmente nos países em desenvolvimento como o Brasil. Muitas vezes, a ocorrência desses problemas está associada a fatores internos e externos ao contexto educacional. Identificar, analisar e prever a ocorrência desses problemas, torna-se importante para as instituições educacionais e governos subsidiarem estratégias eficientes para combatê-los. Diante desse cenário, este estudo busca identificar na literatura a natureza das pesquisas realizadas no contexto da predição educacional, os tipos e características dos dados utilizados, as abordagens/técnicas de predição utilizadas para prever a ocorrência dos problemas e os fatores associados, ou seja, as causas dos problemas educacionais.

4.2.2 Objetivos do Mapeamento

Esta pesquisa caracteriza-se pela condução de um MSL, com o objetivo de identificar, catalogar e analisar o maior número possível de estudos primários relevantes e reconhecidos, sobre a aplicação de métodos para predição de problemas educacionais, com o intuito de identificar lacunas de pesquisa que subsidiem e reforcem a questão central de pesquisa desta tese.

4.2.3 Questões de pesquisa

Uma das etapas essenciais do MSL é a definição das questões de pesquisa que conduzem a busca dos documentos relevantes, permitindo, posteriormente, a seleção dos documentos por meio das palavras chave (*keywording*) e resumos *(abstract*) para extração dos dados (PETERSEN, 2007). Objetivando responder à questão central desta tese e conforme os objetivos propostos no Capitulo 1, este mapeamento elencou a seguinte questão principal: **"Quais métodos utilizados na literatura são aplicados à predição de problemas educacionais"?**

Foram elaboradas quatro questões secundárias que serviram como base para mapear sistematicamente os trabalhos na literatura:

Q1: Qual a natureza das pesquisas que vêm sendo desenvolvidas relacionadas à predição em contextos educacionais?

Q2: Quais são os tipos e caracteristicas dos dados utilizados para construção dos modelos de predição educacionais?

Q3: Que abordagens/técnicas vêm sendo, frequentemente, utilizadas para predição no contexto educacional? e

Q4: Quais fatores estão associados aos problemas de predição educacional?

Estas questões serviram para o direcionamento e elaboração da *string* de busca utilizada na pesquisa para a seleção dos estudos primários.

4.2.4 Protocolo de Pesquisa

Após identificar a necessidade da realização do MSL, foi definido o protocolo do mapeamento que é o elemento essencial para sua execução. O protocolo definido neste estudo específica as questões de pesquisa, a estratégia que foi utilizada para conduzir o MSL, os critérios para a seleção dos estudos e a forma como os dados serão extraídos dos estudos selecionados. Vale destacar que a qualidade do protocolo impacta diretamente na qualidade do MSL.

4.3 CONDUÇÃO DO MSL

Após a definição das atividades a serem realizadas no protocolo de pesquisa, tem inicio a fase de condução do MSL. Nesta fase, os estudos primários são identificados utilizando-se a estratégia de busca definida no protocolo. Uma vez identificados, os estudos precisam ser selecionados por meio de aplicação de critérios de seleção (critérios de inclusão e exclusão). Os critérios de seleção devem especificar as principais características e/ou conteúdos que os estudos devem ter para serem incluídos ou excluídos**.** Após a atividade de seleção, os dados contidos nos estudos devem ser extraídos. A extração objetiva coletar os dados que sejam necessários para responder as questões de pesquisa e para facilitar posteriormente as análises e sínteses dos resultados.

4.3.1 Estratégia de Pesquisa

A estratégia de pesquisa usada, nesse MSL, inclui a definição de: (i) seleção dos estudos primários - a forma como as buscas serão realizadas, os locais onde os estudos serão procurados, a *string* de busca a ser usada; (ii) critérios de seleção - garantir a qualidade dos resultados obtidos, estabelecer as características relevantes que os estudos devem conter (critérios de inclusão) e características que levam à exclusão de estudos que não obedecem aos critérios definidos (critérios de exclusão); e (iii) extração dos dados - definir os critérios de classificação dos estudos com relação as questões de pesquisa (PETERSEN *et al*, 2015).

4.3.2 Seleção dos Estudos Primários

Para que um MSL proveja uma visão ampla do tema é necessário realizar buscas automáticas. Assim, para esta tese, foi realizada uma busca automática dos estudos primários no indexador *Elseiver Scopus (http://www.scopus.com*), uma vez que os indexadores indexam os artigos de várias bibliotecas digitais de editoras tais como *IEEE Xplorer, Digital Library, ACM Digital Library, Science Direct* e *Springer Link* (RODRIGUES, 2016). Foi realizada uma primeira interação por meio de uma *string* de busca, apresentada no Quadro 5.

Quadro 5. Primeira String de busca

TITLE-ABS-KEY-AUTH: ("Academic Performance") AND ("Student Performance") AND ("Dropout") AND AND ("Predictive Models") OR ("Prediction Models") OR ("Predictive Analysis") OR ("Predicting Dropout") OR ("Predicting Performance") AND (EXCLUDE (LANGUAGE , "Spanish") OR EXCLUDE (LANGUAGE , "Turkish"))

Fonte: Elaborado pelo Autor (2019)

Esta primeira interação teve como resultado 145 trabalhos primários nos últimos dez anos (2010 a 2019). A Figura 10 apresenta o gráfico com os trabalhos distribuídos por ordem cronológica de publicações nos últimos dez anos. Assim, 36,2% foram artigos publicados em conferências, 52,5% foram artigos publicados em periódicos, 6,4% foram artigos em revisões, 3,5% foram livros, e 1,4% foram capítulo de livros. Portanto percebe-se como esta temática está crescendo no decorrer dos últimos dez anos.

Figura 10. Publicações na área nos ultimos dez anos
Fonte: Elaborada pelo Autor (2019)

A Figura 11 apresenta os principais periódicos e conferências nos quais os estudos foram publicados.

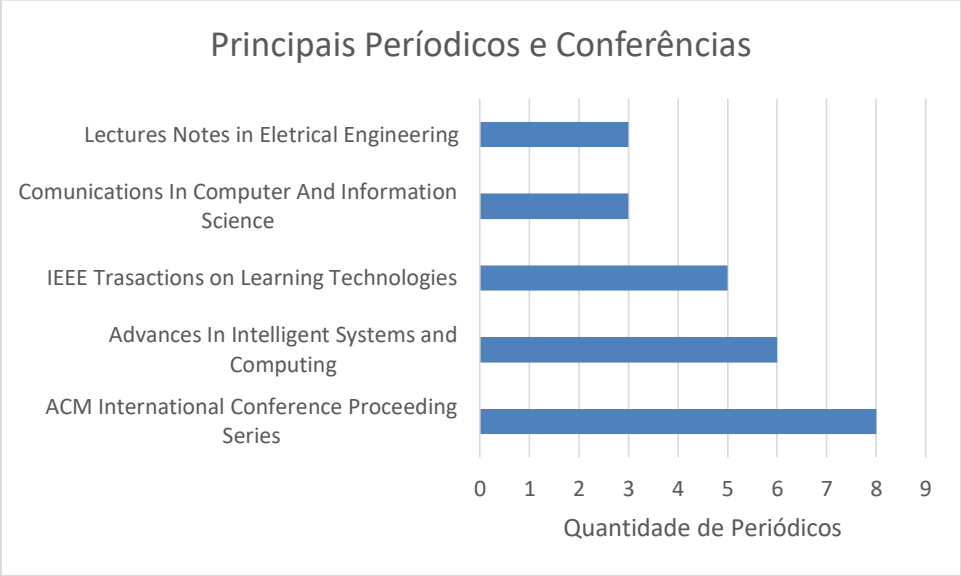


Figura 11. Principais Periódicos e Conferências
Fonte: Elaborado pelo Autor (2019)

O principal periódico foi o *Advances in Intelligent Systems and Computing*, com seis publicações na área de predição educacional, seguido pelo IEEE Trasactions on Learning Tecnologies com cinco publicações. Os demais periódicos obtiveram menos de quatro trabalhos publicados na área.

No intuito de especificar ainda mais a busca, limitando apenas a trabalhos que utilizaram abordagens/técnicas para predição dos fenômenos educacionais, foi necessário desenvolver uma segunda *string* de busca adicionando os termos: Educational Data Mining, Learning Analytics, Machine Learning e Statistic. O Quadro 6 mostra a segunda *string* de busca

Quadro 6. Segunda String de busca

TITLE-ABS-KEY-AUTH: ("Academic Performance") AND ("Student Performance") AND ("Dropout") AND ("Predictive Models") OR ("Prediction Models") OR ("Predictive Analysis") OR ("Predicting Dropout") OR ("Predicting Performance") AND ("Educational Data Mining")AND ("Learning Analytics") AND ("Machine Learning") OR ("Statistic") AND (EXCLUDE (LANGUAGE,"Spanish")

Fonte: Elaborado pelo Autor (2019)

Para especificar a quantidade de trabalhos relevantes no contexto educacional, foi feito um refinamento adicionando os termos *Educational Data Mining (EDM)*, *Learning Analytics*, *Machine Learning*, *Statistic*, na segunda *string* apresentada no Quadro 6, retornando apenas 50 trabalhos primários. Desse total, 14 trabalhos utilizaram abordagens/técnicas de EDM, 9 utilizaram abordagens/técnicas de Aprendizagem de Máquina, 5 utilizaram abordagens/técnicas de *Learning Analytics* e

4 utilizaram abordagens baseadas na modelagem preditiva.

Os 18 trabalhos restantes utilizaram outras abordagens/técnicas aplicadas à previsão educacional. A Figura 12 apresenta a classificação dos trabalhos por área de aplicação.

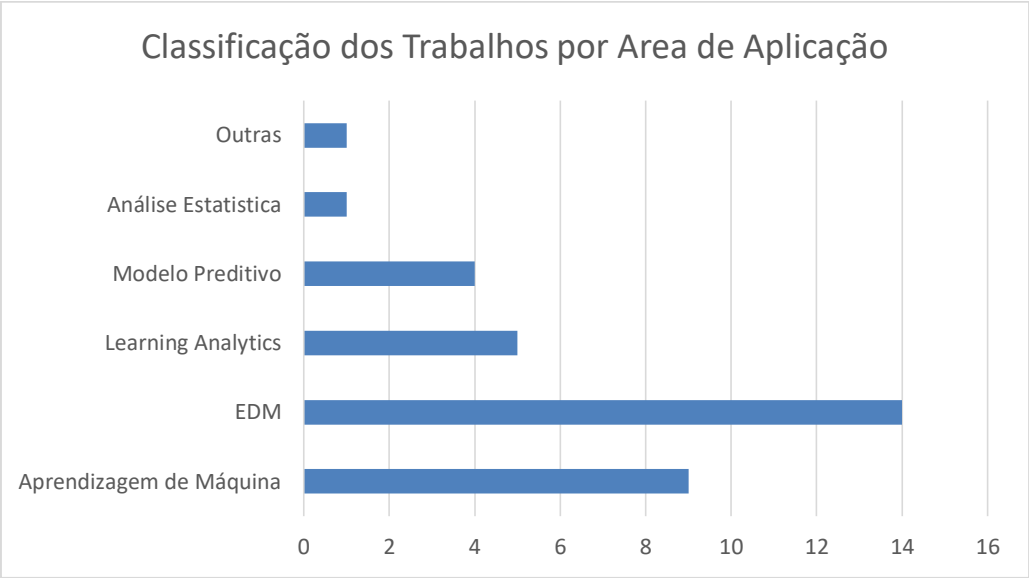


Figura 12. Classificação dos trabalhos por área de aplicação
Fonte: Elaborada pelo Autor (2019)

Considerando os trabalhos selecionados na segunda interação, a Figura 13 apresenta a cronologia dos trabalhos que utilizaram abordagens/técnicas para predição no contexto educacional. De todos os trabalhos retornados, 44% foram publicados no ano de 2019, 36% em 2018, 14% em 2017, 4% em 2016, e 2% em 2015. Os demais anos não apresentaram pesquisas na área educacional que utilizassem abordagens/técnicas de predição.

Figura 13. Cronologia dos trabalhos na área educacional que aplicaram abordagens/ técnicas de predição.
Fonte: Elaborada pelo Autor (2019)

Diante do exposto, para o andamento das próximas etapas do MSL, foram considerados os resultados obtidos na segunda *string* (50 artigos).

4.3.3 Critérios de Seleção

A definição de critérios de seleção é fundamental para garantir a qualidade nos resultados obtidos em um MSL. Os Critérios de Inclusão (CI) estabelecem as características que um estudo deve conter, para ser considerado relevante. Já os Critérios de Exclusão (CE) evidenciam as características que levam à exclusão de estudos que não obedecem aos critérios definidos. Para a inclusão de um trabalho nesta pesquisa foi determinada a sua relevância em relação às questões de pesquisa, por meio da análise do título, palavras chave e resumo. Sendo assim foram definidos os seguintes critérios, conforme Quadro 7 a seguir.

Quadro 7. Critérios de Inclusão e Exclusão

Critérios de Inclusão	
CI1	Artigos publicados em periódicos e conferências
CI2	Artigos publicados entre os anos de 2010 e 2019
CI3	Artigos Escritos em Português ou Inglês
CI4	Artigos que mencionem abordagens, técnicas e processos aplicados a predição dos fenômenos educacionais
CI5	Artigos que respondam à pergunta de pesquisa primária /secundárias
Critérios de Exclusão	
CE1	Artigo duplicado
CE2	Não apresenta acesso gratuito ao texto completo
CE3	Não ter relação com abordagens/ técncicas preditivas no contexto educacional

Fonte: Elaborado pelo Autor (2019)

Tendo em vista a seleção dos trabalhos mais relavantes, este MSL focou especificamente em artigos publicados em periódicos e conferências. A seleção teve como base os trabalhos selecionados na segunda *string* definida no Quadro 6.

Os critérios de inclusão apresentado, pelos itens I1, I2, I3, I4 no Quadro 7, considerou, por uma questão de abrangência, os trabalhos escritos em língua inglesa e portuguesa. Por fim, delimitou-se que os artigos deveriam mencionar alguma abordagem/técnica aplicada à predição no contexto educacional.

Com relação aos critérios de exclusão apresentados pelos itens E1, E2, E3 no Quadro 7, o primeiro critério de exclusão trata da exclusão dos artigos duplicados que retornaram pela busca realizada pelo engenho *Scopus*. O próximo critério de exclusão indica a impossibilidade de acesso gratuito a textos completos. Por fim, o terceiro critério de exclusão trata da aderência dos trabalhos às abordagens/técnicas de predição no contexto educacional. Para este fim, foi realizada uma primeira leitura considerando análise dos resumos (*abstract*) e as palavras chave (*Keywords*), buscando identificar se os trabalhos apresentavam aderência com as abordagens/ técnicas para predição no contexto educacional.

Ao fim dos processos de inclusão e exclusão, foram identificados 15 trabalhos que não atendiam aos critérios de inclusão, sendo excluídos do estudo.

4.3.4 Extração dos Dados

Após a aplicação dos critérios de inclusão e exclusão, foram selecionados 35 artigos nos quais uma leitura completa foi realizada. Nesta etapa de extração dos dados, a qualidade dos estudos selecionados foi analisada de acordo com as seguintes características: a) estrutura adequada; b) definição de abordagens, métodos e técnicas utilizadas; c) fundamentação teórica e referências relevantes; e d) reflexões acerca da temática.

O Quadro 8 apresenta os artigos selecionados neste MSL. Ainda nesse quadro, a primeira coluna utiliza o identificador (ID) para referenciar esses artigos. Já a segunda coluna refere-se ao tipo de estudo: artigos que apresentam estudos de caso (EC) e artigos que apresentam relatos de experimentos (RE), ambos nos diversos contextos educacionais.

ID	TIPO	REFERÊNCIAS
A01	RE	GKONTIZIS, A.F.; KOTSIANTIS, S.; TSONI, R.; VERYKIOS, V.S. An Effective LA Approach to Predict Student Achievement. ACM International Conference Proceeding Series, P. 76-81, 2018.
A02	RE	GARDNER, J.; BROOKS. C. Coenrollment Networks and their Relationship to Grades in Undergraduate Education. ACM International Conference Proceeding Series, P. 295-304, 2018.
A03	RE	HLOSTA, M.; ZDRAHAL, Z.; ZENDULKA. J. Ouroboros: Early Identification of at-Risk Students Without Models based on Legacy Data. ACM International Conference Proceeding Series. P. 6-15, 2017.
A04	EC	ISMAIL, A.; AIMAN M. S.; Predicting the Performance of School: Case Study in Sultanate of Oman. International Conference on Information and Computer Technologies. P. 18 - 21, 2018.
A05	RE	LINGJUN H.; RICHARD A. L.; ANDREW J. B.; JUANJUAN F.; JEANNE S. Predictive Analytics Machinery for STEM Student Success Studies. Journal Applied Artificial Intelligence Vol. 32, no. 4. P. 361-387, 2018.
A06	RE	MUSHTAQ H.; WENHAO Z.; WU Z.; SYED M.R.A; SADAQUAT. A. Using Machine Learning to Predict Student difficulties from Learning Session Data. Springer Science+Business Media B.V. P. 381-407, 2017.
A07	EC	CARMEM L.; ANA L. M.; JOSÉ A. C. L. Learning Analytics to Identify Dropout Factors of Computer Science Studies through Bayesian Networks. Journal Behaviour & Information Technology. Vol 37, p. 993-1007, 2018.
A08	RE	PEDRO M.M.M.; PEDRO J. M.M.; CARLOS A. H.; IRIA E. A.; CARLOS D. K. Analysing the Predictive Power for Anticipating Assignment Grades in a Massive Open Online Course. Journal Behaviour & Information Technology. Vol 37, p. 1021-1036, 2018.
A09	RE	MUSHTAQ H.; WENHAO Z.; WU Z.; SYED M.R.A; SADAQUAT. A. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Journal Computational Intelligence and Neuroscience. Vol 2018, P. 21, 2018.
A10	RE	SAMAN R.; BART R.; SHAKEEL A. K. The role of demographics in online learning: A decision tree based approach. Journal Computers & Education. Vol 137, p 32-47, 2019.
A11	EC	ANA, G.M.; RUBEN, O.V.; JOAQUIM, O.M.; FERNANDO, A.E. Predicting Students' Performance in a Virtual Experience for Project Management Learning. In Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019), p. 665-673, 2019
A12	EC	PATRICIA, M. B.; SONIA, S. G,; MILLARAY, C.S.; HORACIO,M.V. MONICA,B.S. Selection of Determinant Attributes for the results of the SIMCE Matemática 2015 of 8º degree, Region de La Araucanía Chile, using Genetic Algorithms and Support Vector Machines. IEEE Procedings, 2017.
A13	EC	V.I.MIGUEIS.; ANA, F.; PAULO, J.V. G.; ANDRE, S. Early segmentation of students according to their academic performance: A predictive modelling approach. Journal Decision Support Systems. Vol 115, p. 36-51, 2018.
A14	EC	ANNE, S.H.; MICHAEL, S. Early detection of university students with potential difficulties. Journal Decision Support Systems. Vol 101, p. 1-11, 2017.
A15	RE	EVANDRO, B.C.; BALDOINO, F.; MARCELO A. S.; FABRISIA,F.A.; JOILSON, R. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Journal Computers in Human Behavior. Vol 73, p. 247-256, 2017.

A16	RE	SUMYEA, H.; JIUYONG, L.; LIN, L.; ESMAIEL, E.; SHANE D.; DUNCAN, J. M.; QI, L. Predicting academic performance by considering student heterogeneity. Journal Knowledge-Based Systems. Vol 161, p. 134-146, 2018.
A17	EC	ADERIBIGBE, I. A.; ETINOSA, N. O. Data mining approach to predicting the performance of first year student in a university using the admission requirements. Journal Education and Information Technologies. Vol 24, p. 1527-1543, 2019.
A18	EC	ANAT, C. Analysis of student activity in web-supported courses as a tool for predicting dropout. Journal Education Tech Research. Vol 65, p. 1285-1304, 2017.
A19	EC	JOSÉ, A. R. V. PEDRO, J. M. M; CARLOS, D. K.; Improving the prediction of learning outcomes in educational platforms including higher level interaction indicators. Wiley Expert Systems. Vol 35, p. 1-11, 2018.
A20	RE	YURI, N.; VICENTE, G. D.; CARLOS M.; CLAUDIO C. G.; RUBEN G. C. Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions. Journal IEEE Access. Vol 7, p. 75008- 75017, 2018.
A21	RE	AKIN O. Forward stage-wise ensemble regression algorithm to improve base regressors prediction ability: an empirical study. Wiley Expert Systems. Vol 31, no. 1, 2014.
A22	RE	AGORITSA, P.; GEORGE, K. Feature Extraction for Next-Term Prediction of Poor Student Performance. Journal IEEE Transactions On Learning Technologies. Vol 12, no. 2, 2019.
A23	RE	CARLOS, M. V.; ALBERTO C.; CRISTOBAL, R.; AMIN, Y. M. N.; HABIB, M. F.; SEBASTIAN, V. Early dropout prediction using data mining: a case study with high school students. Wiley Expert Systems. Vol. 33, no.1, 2015.
A24	RE	DAVID, B.; M. ELENA, R. G.; MONTSE, S. An Early Feedback Prediction System for Learners At-Risk Within a First-Year Higher Education Course. Journal IEEE Transactions On Learning Technologies. Vol. 12, no.2, 2019.
A25	RE	ALBERTO, C.; JOHN, D. L. Interpretable Multiview Early Warning System Adapted to Underrepresented Student Populations. Journal IEEE Transactions On Learning Technologies. Vol. 12, no. 2, 2019.
A26	RE	DAVID, M. O.; DU Q. H.; MARK R.; MARTIN, D.; DAMYON, W. A Quest for a One-Size-Fits-All Neural Network: Early Prediction of Students at Risk in Online Courses. Journal IEEE Transactions On Learning Technologies. Vol. 12, no. 2, 2019.
A27	EC	SARA, S.; SHARIFULLAH, K.; MUHAMMAD A. A. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. International Journal of Electrical Engineering Education. Vol 54 (2), p. 105-118, 2017.
A28	EC	AUGUSTO, S.; CARLOS, G.; ROSA, A.; KARIM, P.; MAXIMILIANO, M. Centralized student performance prediction in large courses based on lowcost variables in an institutional context. Journal the Internet and Higher Education. Vol 37, p. 76-89, 2018.
A29	RE	OLUGBENGA, W.A.; THOMAS, C. Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education. Vol 10, no.1, p. 61-75, 2018.
A30	EC	EVANDRO, B.C.; BALDOINO, F.; MARCELO A. S.; FABRISIA,F.A.; JOILSON, R. Evaluating the effectiveness of educational data mining techniques for early EDUARDO, F.; MARISTELA, H.; MARCIO, V.; VINICIUS, B. ROMMEL, C. GUSTAVO, V. E. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research. Vol 94, p. 335-343, 2019.
A31	RE	RAHEELA, A.; AGATHE M.; SYED, A. A.; NAJIMI, G. H. Analyzing undergraduate students' performance using educational data mining. Journal Computers & Education. Vol 113, p. 177-194, 2017.
A32	RE	EMMA, H.; MARIA, M.; ANDREW, P. Contrasting prediction methods for early warning systems at undergraduate level . Journal the Internet and Higher Education. Vol 37, p. 66-75, 2018.
A33	EC	IYAD, S.; MAHA, A.; REDA, A.; MICK, R. Prediction Model of School Readiness. Journal of Information & Knowledge Management.Vol. 16, no. 3, 2017.
A34	EC	MARCO, S. M.; EDISON, L. A. Using Decision Trees For Predicting Academic Performance Based On Socio-Economic Factors. International Conference on Computational Science and Computational Intelligence, IEEE, 2017.
A35	EC	HALIL B.; MURALI, M.; SULEYMAN, U. Delineating Factors that Influence Student Performance in a Data Structures Course. Journal IEEE Access. Vol 7, 2018.

Quadro 8. Conjunto de artigos selecionados para revisão
Fonte: Elaborado pelo Autor (2019)

As questões secundárias propostas neste MSL foram consideradas na análise dos artigos apresentados no Quadro 8, no qual criou-se uma sistematização dos principais conteúdos associados às questões descritas na Subseção 4.3.1 (QI, Q2, Q3, Q4), para auxiliar a responder à questão central proposta neste estudo.

4.4 PUBLICAÇÃO DOS RESULTADOS

A última fase do processo do MSL é a escrita dos resultados, os quais serão divulgados aos potenciais interessados (KITCHENHAM e CHATERS, 2007). Os resultados deste MSL serão divulgados como parte desta tese.

4.4.1 Análise e Discussão dos Resultados

Nesta seção serão apresentados os resultados do estudo aprofundado dos 35 artigos selecionados. O MSL foi conduzido no período de janeiro a março de 2019. O processo foi iniciado com a leitura dos artigos selecionados. Foi realizada a análise dos objetivos e das metodologias dos artigos, selecionando as informações julgadas relevantes para responder às questões de pesquisa.

4.4.2 Natureza das Pesquisas

Esta questão de pesquisa focou na identificação da natureza das principais pesquisas desenvolvidas para a predição de problemas educacionais. Para respondê-la foi utilizada a ideia de categorizar estudos em termos de facetas, proposto por Wieringa et al. (2006). Assim, três facetas principais foram criadas. A primeira estruturou o tópico com base na natureza das abordagens de predição, encontradas na literatura, em termos de: desempenho, evasão e comportamento. A segunda considerou a contribuição do estudo, com a proposição ou aplicação de um método, uma técnica e uma ferramenta. Essas categorias foram derivadas da análise dos objetivos propostos nos trabalhos. No entanto, a terceira faceta que reflete a abordagem de pesquisa utilizada nos artigos, é geral e independente de uma área de foco específica. O Quadro 9 apresenta a natureza das abordagens para predição de problemas classificadas com base em facetas.

Quadro 9. Natureza das abordagens\técnicas de predição

Categoria	Descrição	Trabalhos
Predição do Desempenho	Os trabalhos contribuem com a utilização de conhecimentos e/ou técnicas preditivas relacionadas a previsão do desempenho ou engajamento de alunos.	A02, A04, A05, A06, A08, A09, A10, A11, A12, A13, A17, A21, A22, A25, A29, A30, A31, A34, A35
Predição da Evasão	Os trabalhos contribuem com a utilização de conhecimento, métodos e técnicas preditivas para prever com antecedência os alunos em risco e o combate a evasão escolar.	A03, A07, A14, A15, A20, A23, A24, A27, A32
Predição do Comportamento	Os trabalhos contribuem com a predição a partir da análise do comportamento de alunos e interações com o sistema. Neste ultimo caso, alunos da Educação a Distância (EAD).	A01, A16, A18, A19, A26, A28, A33,

Fonte: Wieringa et al. (2006)

Conforme apresentado no Quadro 9, boa parte dos trabalhos está focada no desempenho dos alunos e na identificação de alunos propensos ao abandono escolar (evasão). Uma característica importante observada nos trabalhos foi o predomínio das investigações no nível de ensino superior (graduação) em detrimento ao ensino fundamental e médio. Foi observada a utilização de múltiplas informações para obtenção do perfil do aluno e das características dos fatores associados ao problema do estudo em questão.

4.4.3 Tipos e Caracteristicas dos Dados

A segunda questão de pesquisa trata dos tipos e caracteristicas dos dados utilizados para construção de modelos para predição educacional. Para alcançar os objetivos propostos, os trabalhos utilizaram diferentes tipos de dados: dados demográficos (idade, gênero, classe social, entre outros); dados contidos em questionários (individuais ou em larga-escala); dados contidos em ambientes de aprendizagem (metadados dos objetos de aprendizagem, publicações em fóruns de discussão, entre outros); dados coletados de entrevistas; dados do histórico escolar (notas, disciplinas cursadas, créditos obtidos, total de reprovações, entre outros); dados de navegação dos alunos (páginas acessadas, tempo *online*, frequência de acesso entre outros); coletados nos ambientes de aprendizagem utilizados (*Moodle, Blackboards*, entre outros.); dados coletados em mídias sociais e informações do perfil, comentários realizados, e assim por diante .

A Figura 14 apresenta a origem dos principais tipos de dados utilizados nos trabalhos, para construção dos modelos preditivos, é possível destacar a predominância de estudos que utilizaram dados relacionados ao ambiente de ensino-aprendizagem (22,4%). Seguido pelos dados demográficos (20,9%), e pelos dados presentes nos históricos escolares (19,4%), juntamente com os dados de navegação (19,4%). Os demais dados aparecem em uma quantidade significativa quando comparados aos demais tipos de dados entre (4,5 e 9%).

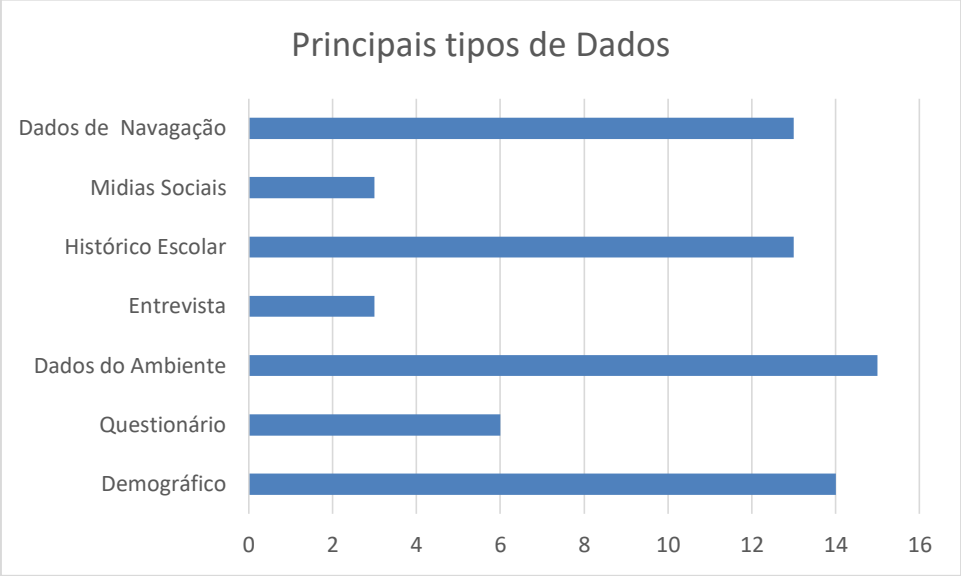


Figura 14. Principais tipos de dados utilizados para predição educacional
Fonte: Elaborada pelo Autor (2019)

4.4.4 Abordagem/Técnicas de Predição

A terceira questão de pesquisa, desenvolvida neste trabalho, diz respeito às abordagens/técnicas que vêm sendo frequentemente utilizadas para predição no domínio educacional. Os resultados encontrados nessa questão de pesquisa mostram que a grande maioria dos trabalhos que se propõem a utilização de abordagens para predição utilizou a Mineração de de Dados Educacionais (EDM). Dos trinta e cinco trabalhos analisados, quatorze utilizaram EDM, para prever o desempenho, a evasão e o comportamento dos alunos, conforme o Quadro 8. Os estudos são: (A04; A10; A11; A12; A13; A14; A17; A20; A22; A23; A25; A26; A27; A30). Esse é um dos métodos mais utilizados na literatura para a descoberta de conhecimento em dados educacionais (PEÑA-AYALA, 2014). A Figura 15 apresenta as principais abordagens/técncicas aplicadas à predição edcuacional, identificadas nos estudos analisados neste MSL.

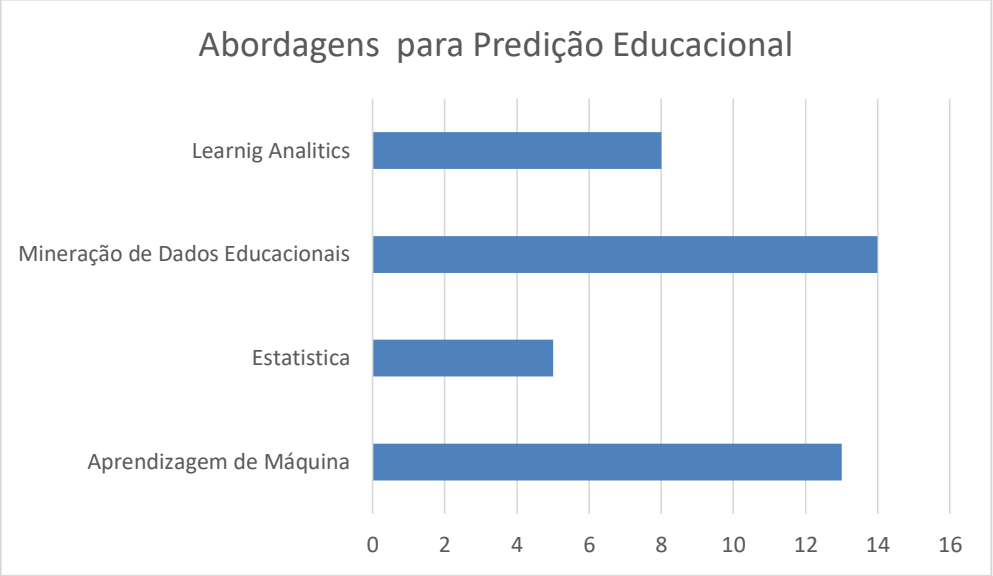


Figura 15. Abordagens\Técnicas de predição educacional
Fonte: Elaborada pelo Autor (2019)

Outras abordagens também foram utilizadas pelos trabalhos (Quadro 6) como: **Aprendizagem de Máquina** (A01; A02; A03; A05; A06; A09; A14; A17; A21; A22; A24; A25), **Learning Analytics** (A01; A03; A07; A08; A19; A24; A32) e **Estatistica** (A07; A17; A21; A35).

É importante esclarecer que alguns trabalhos utilizam mais de uma abordagem para predição a partir da combinação dos modelos preditivos, objetivando compará-los para determinar qual o que apresenta melhores resultados para a sua aplicação no problema de predição. Nesse sentido se destaca o trabalho de Akin (2014), o qual combina abordagens de Aprendizagem de Máquina e Estatisticas, como por exemplo a combinação de técnicas de Ensemble com Regressão Linear, para predição objetivando a redução dos erros de previsão nos modelos. Desta forma, com a combinação de abordagens e técnicas busca a eficiência dos modelos de predição.

Complementando a terceira questão de pesquisa, as técnicas de predição mais utilizadas foram a Classificação e a Regressão. A primeira técnica obteve maior destaque nos trabalhos analisados, como por exemplo o trabalho de Gkontzis e Kontsiantis (2018) o qual aplicou técnicas de Regressão Linerar e técnicas de Classificação a partir dos algoritmos *Random Forest, Neural Network, e Support Vector Machine* –SVM. O foco desse estudo está em prever o desempenho de estudantes a partir das tarefas que o aluno realiza no ambiente de aprendizagem. Obtendo o algotimo *Random Forest* com o melhor desempenho para a previsão.

A Figura 16, apresenta as principais técnicas aplicadas a predição do desempenho, evasão e comportamento do estudante.

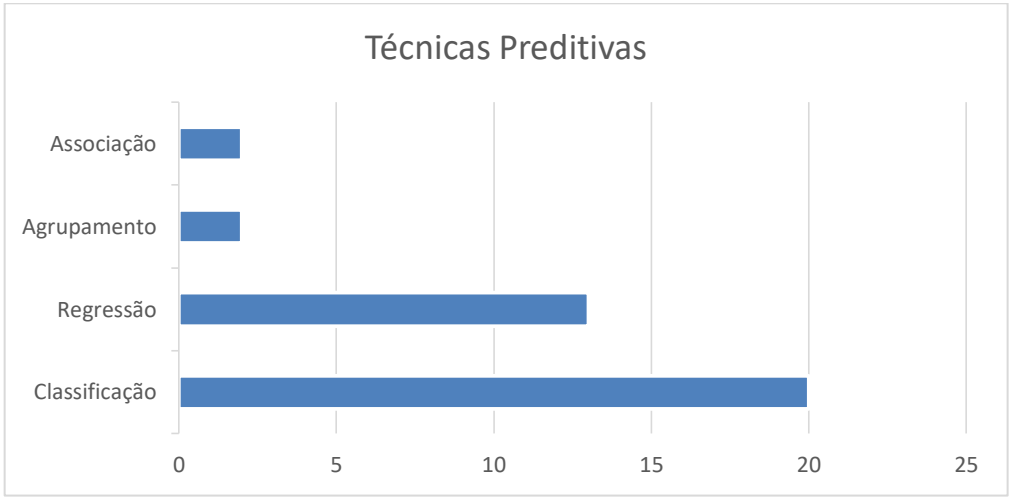


Figura 16. Técnicas aplicadas a predição
Fonte: Elaborada pelo Autor (2019)

A maioria dos trabalhos que utilizaram as técnicas de classificação teve como foco o desenvolvimento de modelos de previsão, especificamente para prever precocemente o desempenho e a evasão dos alunos em risco, bem como o comportamento dos alunos quanto à utilização de ambientes virtuais de aprendizagem. Todos os trabalhos que utilizaram a Regressão como técnica principal ou combinada com técnicas de aprendizagem de máquina, por exemplo, tiveram como objetivo a identificação das variáveis mais significativas para a contrução dos modelos de predição do desempenho e da evasão dos estudantes. Contudo, o estudo não identificou trabalhos, que fizeram uso de abordagens/ técnicas de Regressão para prever características comportamentais dos estudantes.

Além das técnicas de classificação e regressão, outras duas técnicas foram utilizadas: a análise de agrupamento e regras de associação, ambas com 2 trabalhos encontrados na literatura. Os trabalhos de análise de agrupamento (MUSHTAK et al., 2017; GKONTIZ et al., 2018) tiveram como objetivo identificar padrões ou perfis de comportamento dos estudantes em ambientes de aprendizagem. Os trabalhos de regras de associação (SULEIMAN et al., 2017; SANDOVAL et al., 2018) tiveram como foco identificar as relações entre as variáveis relacionadas ao desempenho e ao comportamento dos alunos.

4.4.5 Fatores Associados

A quarta questão de pesquisa está relacionada com a identificação dos fatores associados aos problemas educacionais. Os fatores associados são variáveis internas e externas ao contexto educacional, os quais influenciam de forma positiva ou negativa o andamento da vida acadêmica dos estudantes. Além disso, podem estar relacionados a aspectos pessoais, cognitivos, familiares ou relacionado à própria instituição de ensino (SOARES, 2004). Também podem estar relacionados, a aspectos sociais, intelectuais e comportamentais (TINTO, 1997). Esses aspectos, de forma individual ou em conjunto, podem ser a causa dos principais problemas educacionais enfrentados pela educação na atualidade, como por exemplo, o baixo nível de aprendizagem/desempenho e a retenção/evasão. Nos trabalhos analisados, os fatores associados a instituição de ensino, sócio-econômicos e tecnológicos foram os mais identificados. Este último, no entanto, é um fator presente principalmente nas modalidades de Educação a Distância (EaD). Conforme o Quadro 6, os seguintes trabalhos tratam dos aspectos relacionados a instituição de ensino (A04; A07; A17; A29; A31), **sócio econômicos** (A02; A12; A13; A14; A29) e **fatores tecnológicos** (A02; A03; A06; A09; A29). Além desses, outros fatores também foram identificados, tais como **demográficos** (A10; A34) **pessoais** (A14; A29; A35) e **familiares** (A12). A Figura 17 apresenta os principais fatores associados identificados nos trabalhos.

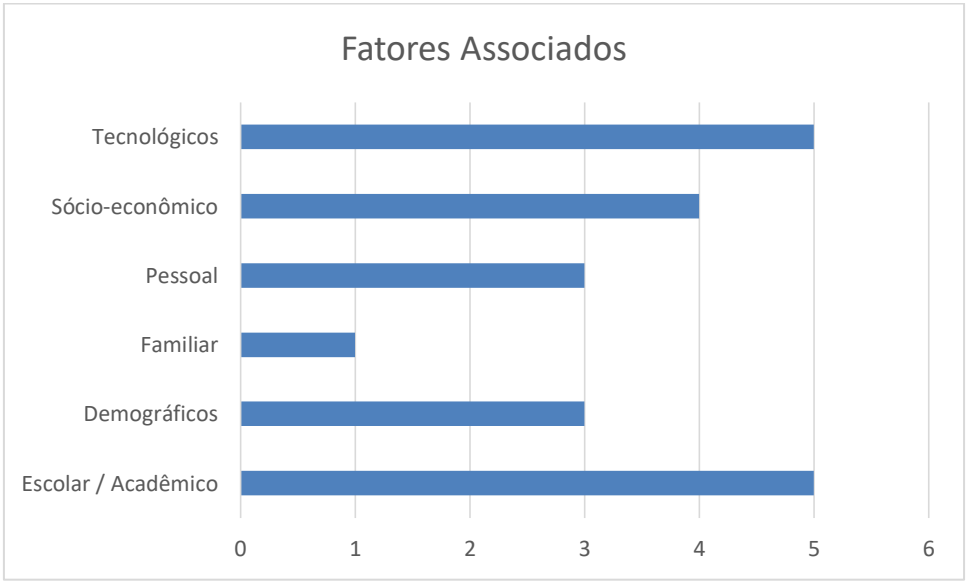


Figura 17. Fatores associados identificados nos trabalhos
Fonte: Elaborada pelo Autor (2019)

A seção seguinte descreve as principais considerações obtidas com este mapeamento sistemático, bem como lacunas de pesquisas que serão abordadas nesta tese.

4.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

O MSL teve como objetivo identificar na literatura os estudos que utilizaram métodos para a predição de problemas educacionais nos últimos dez anos (2010-2019). Foram encontrados no total 141 trabalhos em uma primeira interação realizada. Buscando refinar a seleção dos estudos, foram inseridos alguns termos relacionados, para contemplar as principais abordagens/técnicas utilizadas para predição, tais como: *Educational Data Mining* (EDM), *Machine Learning* (ML), *Learning Analytics* (LA), e Estatistica. Retornando 50 estudos primários, dos quais, 14 aplicaram abordagens\técnicas de EDM, 9 aplicaram abordagens/técnicas de ML, 5 aplicaram abordagens de LA, 4 aplicaram abordagens/técnicas de estatística. Os 18 trabalhos restantes aplicaram outras abordagens/técnicas para a predição dos problemas educacionais. É importante esclarecer que o crescimento exponencial dos estudos nessa área, principalmente nos últimos 5 anos, sendo 2018 e 2019 os anos com o maior número de trabalhos publicados em periódicos e conferências na área.

A maioria dos estudos foca na predição do desempenho escolar/acadêmico, seguido pelos estudos que tratam da predição da evasão/retenção, e pela predição do comportamento dos alunos nas suas interações com os ambientes educacionais. Este último como destaque os estudos que tratam dos ambientes de Educação a Distância (EaD). Além disso, observou-se o grande número de trabalhos que tratam especificamente da predição no contexto do Ensino Superior, abordando principalmente o problema da evasão/retenção. Em contrapartida, poucos trabalhos abordam problemas educacionais relacionados ao Ensino Fundamental.

Os dados utilizados para a construção dos modelos preditivos são de diversos tipos, possuindo diversas origens e características. Grande parte dos dados tem sua origem nos Ambientes Virtuais de Aprendizagem (AVA) e Learning Manangemant Systems (LMS), e dizem respeito ao comportamento do estudante em suas interações com os ambientes, como por exemplo, acesso as atividades, interação com professores e turores, entre outros. Outros tipos de dados significativos encontrados nos estudos são: os dados demográficos que dizem respeito a características como

sexo, idade, cor, por exemplo; os dados acadêmicos, oriundos dos históricos escolares, como notas, reprovações, retenções, entre outros. Os dados de navegação que estão relacionados ao conhecimento tecnológico do estudante no uso das ferramentas de ensino/aprendizagem. Os outros tipos de dados referem-se a questionários que são aplicados a grupos de estudantes/professores, bem como entrevistas. Por fim os dados coletados de mídias sociais como *Facebook* e *Instagram,* entre outras. Sobre o perfil comportamental dos estudantes e suas relações sociais. Como se pode ver, a origem, tipo e caracteristicas dos dados são os mais diversos, contudo, não há uma padronização dos métodos de coleta e análise desses dados, limitando-se estudos a adaptações de outros métodos desenvolvidos.

As abordagens/técnicas de EDM são utilizadas para a predição de problemas educacionais. Dentre as tarefas mais utilizadas, a Classificação se destaca na maioria dos trabalhos, seguida pela tarefa de Regressão. Nesse sentido, pode-se destacar a utilização da combinação de modelos preditivos, ou seja, a combinação de mais de uma técnica, objetivando-se a redução de erros, acurácia e eficiência dos modelos. Destacam-se nos estudos a combinação de técnicas de ML com técnicas de classificação e regressão.

Dos principais fatores associados aos problemas educacionais destacam-se os fatores escolares/acadêmicos, tecnológicos, socioeconômicos, pessoais e demográficos. Observa-se que os estudos não fazem a transposição dos fatores encontrados com os modelos teóricos existentes na literatura relacionada aos problemas educacionais. A literatura trata de problemas como desempenho e evasão, e no decorrer dos anos desenvolveu modelos teóricos que possibilitam o diagnóstico das causas dos problemas educacionais. No entanto, foi observado que os estudos buscam confirmar os fatores associados por meio de análises quantitativas, não relacionando aos fatores já desenvolvidos e consolidados pelos modelos teóricos existentes na literatura. Observou-se ainda que os estudos buscam apresentar resultados por meio de instrumentos de visualização como, por exemplo, gráficos, *softwares* e *dashboards*, entre outros. Contudo, esses métodos não estabelecem de forma concreta a relação de causa e efeito para um diagnóstico dos problemas educacionais. No próximo capitulo será apresentada a abordagem proposta por este trabalho bem como as tarefas descritas em cada uma das fases para análise dos dados educacionais.

5 ABORDAGEM PROPOSTA (DEP-DM)

Em função da natureza dos objetivos propostos nesta tese direcionarem para o uso e aplicação de técnicas de EDM, como forma de determinação dos modelos preditivos, para o diagnóstico dos problemas educacionais, e para adequação ao contexto de dados educacionais, foi proposta uma abordagem específica, com base na metodologia CRISP-DM, denominada, DEP-DM (*Diagnosis of Educational Problems using Data Mining)*. A Figura 18 apresenta a abordagem DEP-DM, a qual consiste em 5 (cinco) fases. A sequência de fases não é obrigatória, podendo ocorrer a transição para diferentes fases dependendo do resultado de cada fase. Os fluxos indicam as mais importantes e mais frequentes dependências entre as fases.

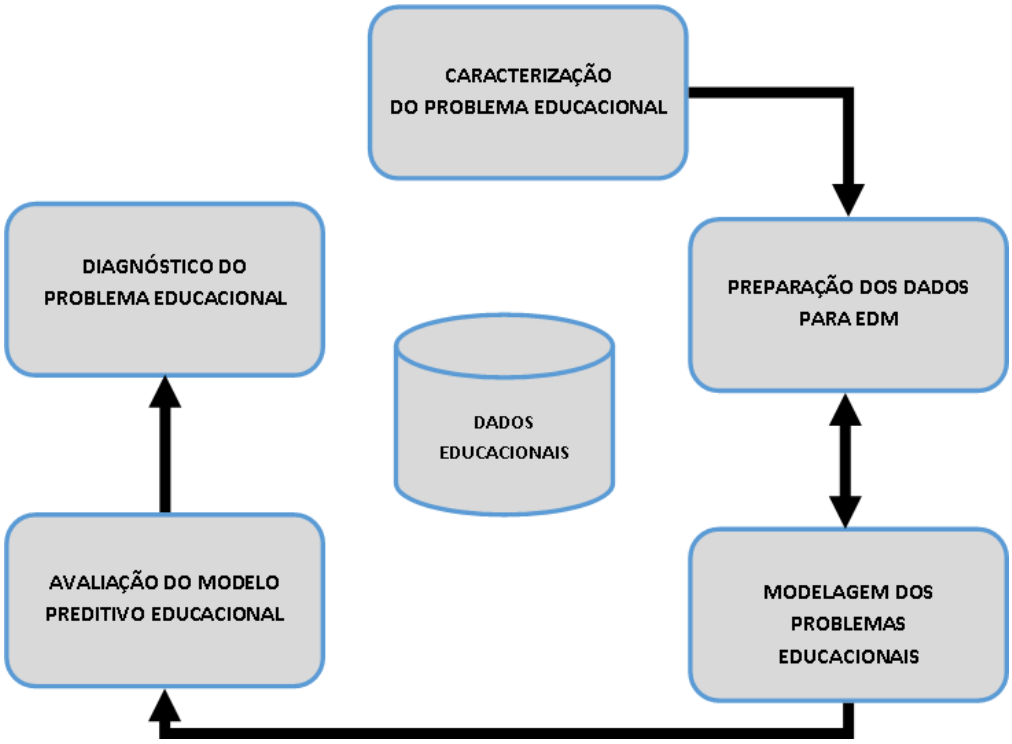


Figura 18. Abordagem DEP-DM
Fonte: Elaborada pelo Autor (2019)

As próximas seções detalham cada uma das fases exibidas na Figura 20, as quais são estruturadas em várias tarefas gerais e específicas, desenvolvidas para determinadas situações.

5.1 CARACTERIZAÇÃO DO PROBLEMA EDUCACIONAL

Esta é uma das fases mais importantes do processo. Nela é realizada a caracterização e o entendimento do contexto educacional com o objetivo de identificar e descrever os principais problemas educacionais. Ou seja, identificar as dimensões dos fatores associados aos problemas e suas principais causas. Também permite identificar as principais abordagens/técncicas que vêm sendo utilizadas para prever a ocorrência dos problemas. Por fim, nesta fase é possivel analisar os fatores e suas dimensões e os indicadores que serão medidos dependendo do contexto educacional.

Nesta tese, a caracterização do problema educacional e dos fatores causadores foi realizada a partir de uma revisão da literatura, a qual teve como foco: identificar os problemas educacionais, identificar os contextos educacionais nos quais eles ocorrem e os fatores associados aos problemas educacionais. Também foi realizado um mapeamento da literatura que teve como objetivo identificar os métodos, abordagens\técnicas para predição de problemas educacionais. A Figura 19 mostra as atividades realizadas na revisão da literatura para caracterização do problema educacional.

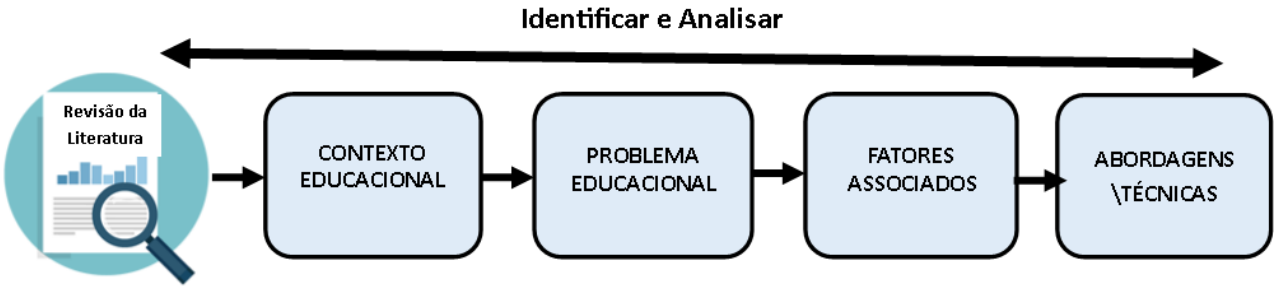


Figura 19. Atividades realizadas na revisão da litertura para caracterização dos problemas educacionais**.**
Fonte: Elaborada pelo autor (2019)

A revisão da literatura foi realizada a partir do entendimento das principais problemáticas presentes na educação. Para este estudo, dentre as diversas problemáticas existentes, destacaram-se o baixo desempenho acadêmico e a evasão. A escolha dessas problemáticas deve-se às evidências analisadas na literatura sobre sua ocorrência nas diversas modalidades de ensino. As evidências dessas ocorrências são apresentadas nos capítulos 2 e 4 desta tese.

A partir das evidências apresentadas na revisão da literatura sobre o baixo desempenho e a evasão, e buscando atender aos objetivos desta tese, foram

utilizadas as bases de dados do INEP. Essas bases são disponibilizadas publicamente, sendo compostas por um conjunto de avaliações externas em larga escala (provas e questionários) coletados periódicamente ou por informações censitárias anualmente.

Nesta pesquisa, os dados utilizados são oriundos do Sistema de Avaliação da Educação Básica (SAEB), do Censo da Educação Superior e Indicadores de Fluxo da Educação Superior. As amostras de dados do SAEB referem-se aos anos de 2013, 2015 e 2017. O conjunto de dados do Censo da Educação Superior refere-se ao ano de 2013 e 2015 para os dados de fluxo da educação superior.

Para esta pesquisa, os dados do SAEB representam os resultados das proficiências (desempenho) nas avaliações de Língua Portuguesa e Matemática, realizadas pelos estudantes do 5º ano das escolas públicas do Estado de Pernambuco. A escolha especifica desses dados deve-se ao baixo desempenho que o estado vem apresentando nas últimas avaliações, ficando abaixo da média nacional de desempenho (INEP, 2018). Além dos dados de proficiciência dos estudantes, a base de dados do Saeb conta ainda com informações das escolas, professores e diretores, além das informações dos questionários contextuais.

Os dados do Censo da Educação Superior reúnem informações de todas as instituições superiores brasileiras (públicas e privadas), sendo coletados a partir do preenchimento dos questionários, por parte das Instituições de Ensino Superior (IES) e por importação dos dados do Sistema e-MEC. Os dados representam os aspectos socioeconômicos, institucionais e indicadores de fluxo dos estudantes. Nesta tese, os dados censitários utilizados referem-se a informações dos alunos dos cursos superiores da área de computação de todo o Brasil. Essa escolha deve-se a esses cursos serem um dos responsáveis pelos altos índices de evasão e retenção.

Foi identificada na literatura, a utilização de várias abordagens e técnicas para o desenvolvimento de modelos preditivos no contexto educacional, dentre as quais destacam-se a Mineração de Dados Educacionais (EDM), e as técnicas e algoritmos de Regressão e Aprendizagem de Máquina. Estes métodos apresentam-se como os mais indicados, tendo em vista o atendimento dos objetivos propostos por esta tese.

Quanto à identificação dos fatores associados ao desempenho, levou-se em consideração as dimensões propostas pelo modelo conceitual dos fatores inter-relacionados à aprendizagem proposto por Andrade e Soares (2008) que estão

relacionados a familia, aluno, escola e sociedade. Para o problema da evasão foram considerados os fatores propostos no modelo explicativo de Spady (1971) e no modelo de integração proposto por Tinto (1993), os quais determinam os fatores que influenciam a decisão do aluno em evadir, sendo relacionados às dimensões dos fatores familia, aluno, instituição e sociedade.

5.2 PREPARAÇÃO DOS DADOS PARA EDM

Nesta fase, foram realizadas as atividades de junção, extração, transformação e limpeza das variáveis para construção da tabela de análise. Para a junção das bases foi realizado um estudo da arquitetura das bases de dados. Existem nas referidas bases 86 (oitenta e seis) variáveis socioeconômicas o que representa um número considerável e significativo. Após a aplicação dos métodos de seleção de variáveis o número de variáveis reduziu para 52 (cinquenta e duas) em ambas as bases. Um fator importante a ser considerado foi a significância teórica das variáveis, como por exemplo Q001 (sexo), Q002 (Raça auto atribuída) e PROFICENCIA_LP_SAEB (Proficiência em língua portuguesa do SAEB) dentre outras. Após o entendimento da estrutura das bases de dados e a significância das variáveis, foi realizada a construção de uma tabela de análise. Essa tabela foi construída a partir da junção das bases (aluno, escola, professor e diretor) para cada ano de edição do SAEB, ou seja, três anos não consecutivos (2013, 2015 e 2017).

Quanto ao Censo da Educação Superior, foi feita uma junção nas bases para que passasse a existir apenas um referente a cada ano, com as variáveis que seriam utilizadas. As duas bases foram mescladas em função da variável que representava o código único de identificação de cada curso, pois essa variável estava presente em ambas as bases. A tabela de análise do censo foi constituída com dados dos anos de (2013 e 2015).

Após a junção das bases foi realizada a seleção das variáveis. Para este fim, nas Bases do SAEB, foram excluídos os alunos que não participaram da avaliação. A caracterização do desempenho (variável dependente) foi definida pelas proficiências dos alunos em Lingua Portuguêsa e Matemática.

Já os fatores associados (variáveis independentes) foram definidos pelas respostas dos questionários contextuais preenchidos pelos estudantes e escola. As variáveis redundantes ou irrelevantes foram excluídas, e nos registros com valores

em branco foi realizada a inserção de valores utilizando a mediana dos valores das colunas, tendo como objetivo obter o mínimo de perda de instâncias significativas.

Nos dados do Censo, foi observado que haviam alguns atributos que constavam com valores em branco, sendo difícil aplicar o processo de substituição pela mediana por conter muitos *missing values.* Estes atributos então foram removidos, bem como atributos que foram observados tendo mais de 50% dos dados faltando.

Algumas dificuldades foram encontradas durante a seleção das variáveis, tais como: mudança da estrutura das bases do SAEB, volume grande de dados, existência de muitos dados em branco (*missing values*) em muitas variáveis, principalmente do Censo e dos indicadores de fluxo, além das poucas informações sobre o perfil socioeconômico dos alunos e divergências nas informações

Para análise do grau de correlação entre as variáveis, foram aplicadas as técnicas de Correlação de Pearson (LEHMAN, 2013) e o Método Stepwise (GUYSON; ELISEEFF, 2003). O objetivo foi identificar as variáveis que mais se correlacionavam com os fatores associados ao desempenho e à evasão.

Por fim, foi realizada a transformação dos dados categóricos em numéricos. Em EDM, os algoritmos necessitam que os dados sejam transformados nas formas de atributos categorizados (classificação) e numéricos (regressão), especialmente o atributo alvo de interesse (TAN, STEIBACH e KUMAR, 2009).

Neste sentido, para aplicação das téncicas de Regressão, as variáveis categóricas encontradas em ambas as bases que possuiam três ou mais opções de resposta foram dicotomizadas. As demais variáveis que tinham duas categorias foram transformadas para numéricas.

As variáveis transformadas foram utilizadas na fase de modelagem dos problemas educacionais, descritas a seguir, para a construção dos modelos de predição e diagnóstico dos problemas educacionais.

5.3 MODELAGEM DOS PROBLEMAS EDUCACIONAIS

Após a análise correlacional das bases de dados educacionais, foram elencadas as variáveis mais significativas para o estudo, tanto no cenário do desempenho, quanto no cenário da evasão. Foram considerados dois casos de estudo em cada base de dados: (i) as variáveis de maior correlação em relação às proficiências em Portugues e Matemática; e (ii) variáveis com maior correlação em relação às taxas de evasão.

Após essas definições das variáveis preditoras, foram aplicadas as técnicas mais apropriadas, dependendo dos objetivos identificados para o processo de EDM. Considerando a abordagem de combinação de modelos aplicada nesta tese, foram consideradas as fases de gerar e integrar modelos. Dessa forma, a metodologia de EDM é adaptada para compor essas subfases. Apesar de poder ser realizada a fase de poda no processo de combinação de modelos, essa etapa não foi considerada para implementação nesta tese. Tendo em vista, que não houve sobrecarga de dados, o que levaria a má precisão em dados não conhecidos, portanto não sendo necessário neste caso a implementação do processo de poda. Assim, a fase geral da modelagem comporta duas subfases para desenvolver o modelo. Com essa metodologia busca-se uma forma padronizada de construir modelos combinados para o contexto da EDM. As subfases são: (i) a geração que consiste em criar um conjunro de elementos candidatos a fazerem parte do modelo combinado; e (ii) na integração é definida uma estratégia para obter a predição do modelo combinado, com base nas previsões dos modelos base.

A função que combina modelos considerada nesta tese foi a *Bagging* ou ensacamento de modelos. A implementação da modelagem foi realizada utilizando a plataforma R, a qual é recomendada para modelagem e análises estatísticas.

O modelo *Bagging* desenvolvido será utilizado como a técnica de predição aplicada aos problemas de regressão no contexto de EDM abordados nesta tese. Os algoritmos gerados no processo *Bagging* são apresentados na Quadro 10. Buscou-se a implementação de regressões paramétricas e não-paramétricas, as quais satisfizessem a necessidade da diversidade de modelos para o desenvolvimento de modelos combinados.

O primeiro modelo considerado foi a Regressão Linear. O modelo do MMQ desenvolve uma equação que cria um relacionamento linear entre as variáveis preditoras e explicativa. No entanto, esse método apresenta alta sensibilidade a *outliers* presentes nos dados. Devido a essa fragilidade do modelo linear, outros modelos foram considerados a compor o conjunto, como a Regressão Robusta. Outro modelo que resolve o problema de sensibilidade aos *outliers* é a Regressão Quantílica. Esta técnica distingui-se dos demais tipos de regressão pois permite diferentes inclinações da curva de regressão, capturando mudanças de localização.

Ou seja, considera diferenças de relação sobre os quantis da variável dependente.

Outra técnica considerada é a SVR, a qual possui vantagem sobre outras técnicas pois fornece como alternativa trabalhar em um espaço de alta dimensionalidade. Assim, permite realizar um mapeamento não linear dos dados de entrada para um espaço de dimensão maior, na qual a regressão linear torna-se possível. Ela também apresenta um bom desempenho em problemas não lineares. Neste caso a SVR torna-se um bom candidato a metapreditor do modelo *Bagging*, porque mesmo que os demais modelos base tenham um alto erro na predição devido ao comportamento dos dados, a SVR pode aumentar a precisão preditiva por se tratar de uma técnica não paramétrica, adequando-se melhor aos dados.

Quadro 10. Modelos de Regressão gerados

Regressão	**Modelos de Nivel 0**	***Package***
Linear (RL)	Canônica	-
Robusta (RLR)	Canônica	*rlm {MASS}*
Quantilica (RQ)	Tau = 0,5	quantreg
Regressão	**Modelo de Nível 1**	***Package***
Support Vector Regression (SVR)	Kernel = radial basis, C=1, y = 0,25, $\epsilon = 0{,}1$	*e1071*

Fonte: Elaborado pelo Autor (2019)

Como se pode observar no Quadro 10, alguns algoritmos sofreram variações de parâmetros para que o modelo construído fosse melhor adaptado ao contexto do problema. Para o SVR, as variações consistiram do *kernel*, sendo gerado um modelo com *kernel* gaussiano. Já a RQ sofreu variações em relação ao quantil (tau), sendo considerado o quantil 0,5. Esse quantil foi utilizado porque apresenta melhores resultados nas análises realizadas e considera a mediana dos dados.

O método *Bagging* aplicado nesta tese foi desenvolvido por Breiman (1996). A sua aplicação pode ser comprovada pelas pesquisas realizadas por Shahid et al., (2015); e Palermo et al. (2006) que apresentam a aplicação desse método mostrando a problemas de regressão. O *Bagging* combinado com modelos de *Regressão* pode gerar diferentes modelos para melhorar a estabilidade, a precisão e o valor preditivo. Sendo o modelo preditivo representado pela equação D= {(y_n.x_n), com n=1..., N}, em que y_n é a variável resposta e x_n é o conjunto de variáveis explicativas.

O processo inicia-se com a definição da base de treino D_{treino} e da base de teste D_{teste}. Ainda dado um conjunto K de técnicas, é invocada a *k-ésima* técnica sobre D_{treino} para criar o modelo, M_k^{treino}, com k =1,... , K. Esses são chamados de modelos de nível 0, ou de base. Então para cada instância x_n em D_{teste}, z_{kn} denota as predições do modelo M_k^{treino} em x_n. No fim do processo, a base de dados formada com a saída das K técnicas é $D_f = \{(y_n, z_{1n}, \ldots z_{Kn}), n = 1, \ldots N\}$. Estes são os dados do nível 1, os quais serão utilizados como base para a predição do modelo do nível 1. Este modelo consiste na aplicação de alguma técnica que derive da base D_f um modelo M_f para y como uma função de (z_1, ..., z_k). As etapas da construção do *Bagging* são apresentadas na Figura 20.

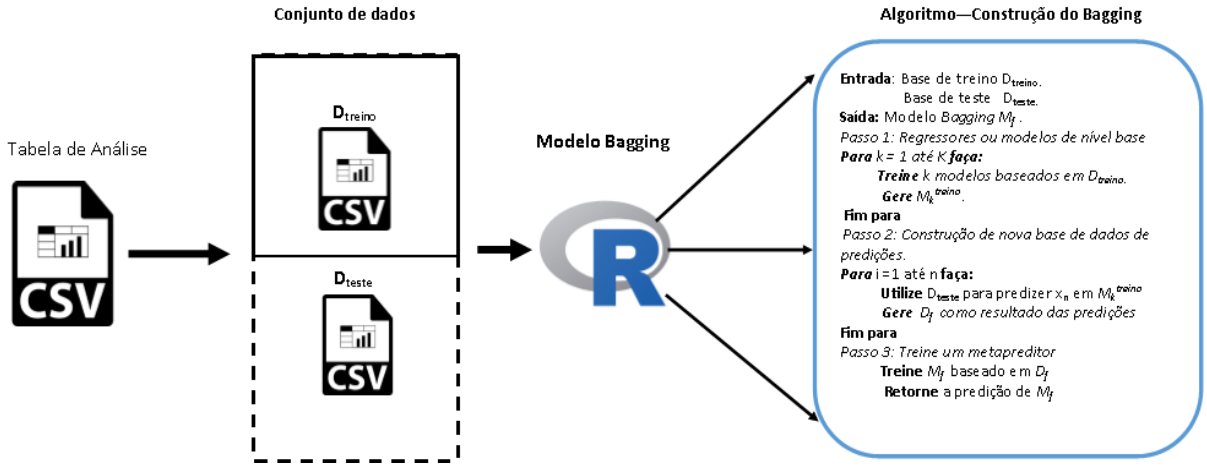


Figura 20. Etapas para a construção do Modelo *Bagging*
Fonte: Elaborada pelo Autor (2019)

A Figura 21 representa os modelos gerados por meio da implementação do *Bagging*. Os {M_1, ..., M_k}, com *k* =4, identificam os quatro modelos de *Ensemble* utilizando a estrutura do *Bagging* caracterizando-se como modelos de base, ou de nível treinados. O M_f identifica o metapreditor, ou modelo de nível 1, no qual constam dados treinados pelos modelos do nível anterior:

- RL, modelo M_1;
- RLR, modelo M_2;
- RQ, com tau = 0,5 modelo M_3; e
- SVR, com Kernel = radial basis. Modelo M_f.

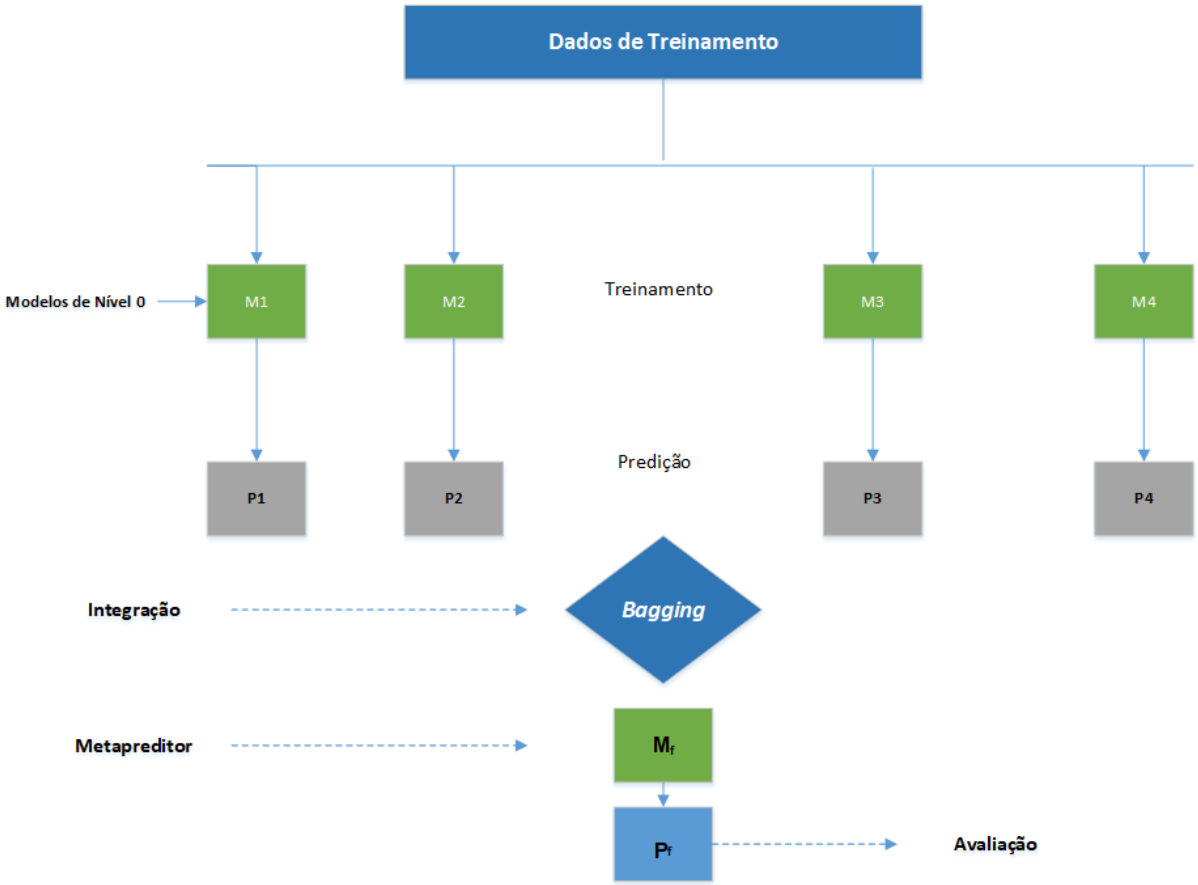


Figura 21. Representação e Integração de modelos baseados em *Bagging*
Fonte: Elaborada pelo Autor (2019)

A proposta desta tese consiste na aplicação de modelos combinados a partir da abordagem *Bagging* no contexto da EDM. A sua aplicação em diferentes cenários proporciona uma melhor avaliação do desempenho do modelo proposto. Pretende-se que o referido modelo reduza o erro de predição associado a diferentes problemas. Com os modelos de base gerados reduz o viés associado aos dados, e alia-se isso a um modelo não-paramétrico como Metapreditor (SVR).

5.4 AVALIAÇÃO DO MODELO PREDITIVO EDUCACIONAL

A avaliação é o processo de verificação de como os modelos desenvolvidos com as técnicas de mineração são executados nos dados reais. É importante validar os modelos de mineração entendendo suas qualidades e características antes de implantá-los em um ambiente de produção (CHAPMAN, *et al.*, 2000).

O desempenho será mensurado em termos do MAE, já definido no Capítulo 3, como denotado na Equação 2.16. Outra forma de medição de desempenho é por meio do ganho relativo (GR). O GR é aplicado para mensurar o ganho em relação à

minimização do erro de predição, dado em por centagem.

O cálculo é mostrado na Equação 5.1. Ainda são abordadas avaliações por meio de testes estatísticos e gráficos.

$$GR = 100\left(\frac{MAE_a - MAE_b}{MAE_a}\right) \quad (5.1)$$

A partir das amostras geradas após a execução das simulações pode-se calcular o desvio padrão do erro, realizar testes estatísticos, gráficos *bosxplots*, para assim também avaliar o desempenho dos modelos propostos nas bases de dados educacionais.

O desempenho do Modelo *Ensemble Regression* (MER) foi testado em dados reais de um repositório público. O objetivo desta análise consiste no estudo do desempenho do MER em bases de dados reais, as quais não são criadas de forma controlada para experimentação. Nesse contexto, foram selecionados dois conjuntos de dados, a saber: SAEB, e Censo da Educação Superior. As características desses conjuntos de dados serão descritas a seguir.

5.5 DIAGNÓSTICO DO PROBLEMA EDUCACIONAL

Esta fase direciona a utilização de métodos e ferramentas para o diagnóstico do problema educacional (efeito) a partir dos fatores (causas) já consolidados na teoria e os identificados, a partir dos resultados apresentados pelos modelos preditivos aplicados nos contextos educacionais estudados nesta tese. O objetivo é estabeler uma relação de causa e efeito entre os fatores identificados com os problemas, além de facilitar a visualização e o conhecimento das princiapis causas relacionadas aos problemas educacionais. Além disso, deve servir como um método que auxilie no processo de implementação de modelos preditivos, quando aplicados para fins da predição educacional.

O diagnóstico educacional é um processo de caráter instrumental, científico e integral que permite um estudo prévio e sistemático, por meio da coleta de informações, do estado real e potencial do problema e de todos os elementos que possam influenciar direta ou indiretamente dos resultados desejados (BIRINGUES, 2010). O diagnóstico proposto neste trabalho busca reunir evidências que permitam aos gestores, professores e demais agentes educacionais ajustar suas politicas e métodos de ensino para atender às necessidades educacionais individuais ou grupais

de seus alunos. Neste sentido, o processo de diagnóstico educacional segue algumas etapas para sua execução, conforme apresentado na Figura 22.

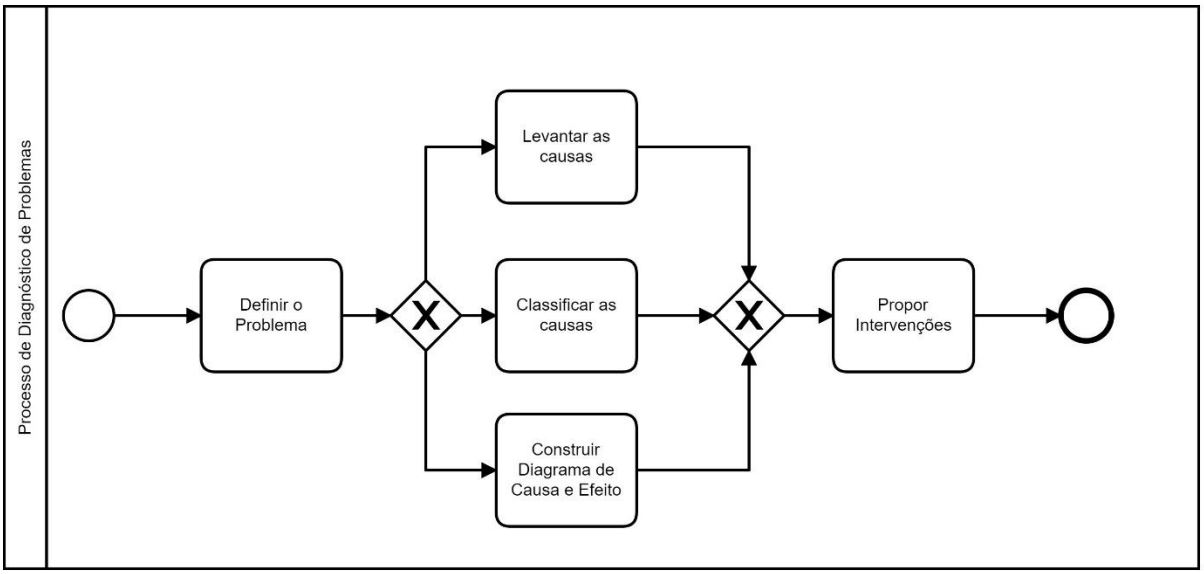


Figura 22. Processo de Diagnóstico do Problema Educacional
Fonte: Elaborada pelo Autor, (2019)

O processo apresentado na Figura 22 define um conjunto de etapas a serem executadas para a realização do diagnóstico educacional. A primeira etapa consiste na definição do problema, definidos nesta tese como problemas relacionados ao desempenho dos estudantes e à evasão no ensino superior.

Os fatores internos e externos ao ambiente educacional são apontados como os principais fatores associados aos problemas educacionais. Assim, quatro dimensões de fatores relacionados destacam-se tanto com o desempenho quanto com a evasão família, aluno, instituição e sociedade.

Observou-se em alguns casos a manutenção de alguns fatores e a inclusão de outros, como por exemplo: fatores comportamentais, tecnológicos, socioeconômicos, demográficos, familiares e institucionais. A partir do conhecimento das dimensões dos fatores (causas) primárias, foi realizada a sua categorização.

Como forma de visualizar a execução desta etapa foi desenvolvido um diagrama de causa e efeito. O diagrama simplifica processos considerados complexos, dividindo-os em processos mais simples e, portanto, mais controláveis (TURBINO, 2000). Essa ferramenta é um método bastante efetivo na busca das raízes do problema (SLACK, 2009).

Ela é utilizada para expor a relação existente entre os resultados de um processo

e as causas que tecnicamente possam afetar esse resultado.

Esse diagrama leva em conta todos os aspectos que podem ter levado a ocorrência do problema dessa forma. Ao utilizá-lo, as chances que algum detalhe seja esquecido diminuem consideravelmente (ISHIKAWA, 1993). A Figura 23 apresenta o diagrama de causa e efeito utilizado como ferrmenta da qualidade.

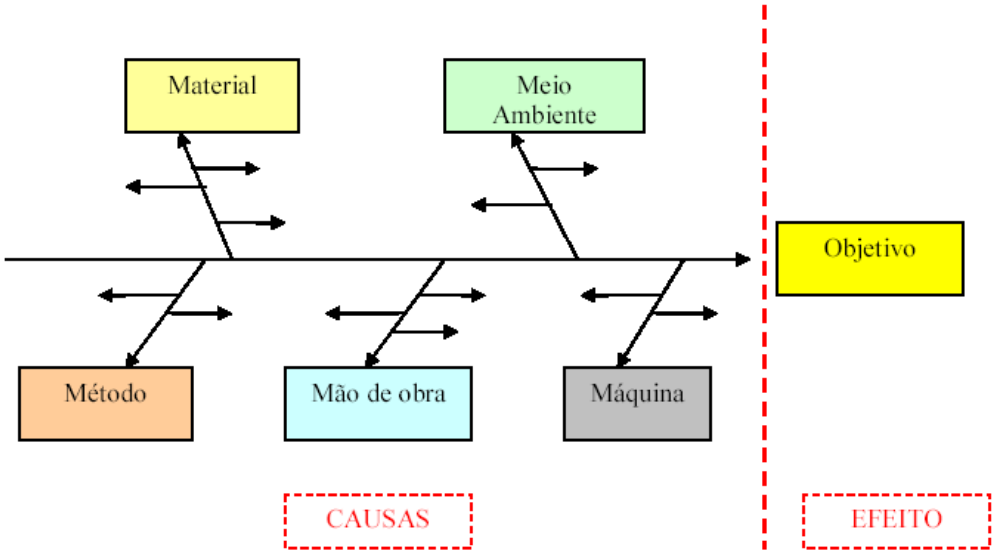


Figura 23. Diagrama de causa e efeito
Fonte: Ishikawa (1993)

No diagrama de causa e efeito os resultados de um processo apresentado podem ser indesejados ou não (efeito) e os diversos fatores responsáveis (causas) que podem contribuir para que tais efeitos ocorram no decorrer de sua execução. Sua relação com a imagem de espinha de peixe se dá devido ao fato que se pode considerar suas espinhas as causas dos problemas levantados, contribuíndo para a descoberta de seu efeito (PEINADO, 2007). É possível aplicar o diagrama de causa e efeito em diversos contextos e de diferentes maneiras, entre elas destaca-se a utilização para:

- Visualizar as causas principais e secundárias de um problema (efeito);
- Ampliar a visão das possíveis causas de um problema, enxergando de maneira mais sistêmica e abrangente;
- Identificar soluções, levantando os recursos disponíveis; e
- Gerar melhorias nos processos.

Neste sentido, o diagrama de causa e efeito foi adaptado ao contexto educacional, de forma a atender aos objetivos propostos nesta tese.

5.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram apresentadas as fases/etapas do DEP-DM, com suas respectivas metas, técnicas e ferramentas que compõem o método de pesquisa desta tese. Na fase de caracaterização do problema educacional foi realizada a revisão da literatura sobre os principais problemas educacionais nas diversas modalidades de ensino, destacando-se o problema do desempenho e da evasão. Na fase de preparação dos dados para EDM, foi realizada a escolha das variáveis, que representam os fatores associados aos problemas educacionais selecionados nesta tese. Em seguida, foi realizada a extração das variáveis, e seu tratamento (limpeza, e transformação) dos dados para construção da base de análise para aplicação das técnicas de EDM. Após o pré-processamento das váriaveis foi executada a fase de modelagem, na qual foram aplicados os modelos combinados de *Ensemble Regression* para a construção dos modelos preditivos educacionais para o diagnóstico dos problemas. Estes modelos passaram por uma fase de avaliação por meio de métricas como a do erro médio absoluto MAE, ganho relativo e teste de hipótese.

Por fim, foi realizado o desenvolvimento de um processo para o diagnóstico de problemas, com base no diagrama de causa e efeito, como forma de materializar os resultados dos modelos preditivos, além de diagnosticar os problemas educacionais tratados nesta tese. Esse conhecimento será útil para gestores e professores implementarem intervenções eficazes para reduzir a ocorrência desses problemas em suas instituições educacionais. Além de servir como um guia de boas práticas para a resolução de problemas semelhantes ou outros tipos de problemas nas mais diversas áreas de conhecimento.

No próximo capítulo, será apresentada uma análise dos resultados alcançados com o desenvolvimento e a aplicação do modelo proposto nesta tese.

6 ANÁLISE DOS RESULTADOS

Neste capítulo são apresentados os resultados dos experimentos baseados no Capitulo 5. Além disso, os métodos propostos e as teorias educacionais foram necessários para atingir os resultados e obter conclusões. Os resultados ratificam a aplicação do modelo combinado MER em dados reais. Desta forma, é apresentada uma aplicação utilizando dados educacionais para investigar o modelo na predição – por meio da EDM. Esta tem como objetivo, conforme já explicitado nesta tese a descoberta de informações que auxiliem no contexto educacional, como por exemplo: a melhoria das condições de infraestrutura escolar, o processo de ensino, a previsão do desempenho dos alunos, além de outros fatores que influenciam a aprendizagem, dentre os quais podem ser citados o desempenho e a evasão escolar.

Para isto, a EDM consiste na aplicação de técnicas de Mineração de Dados em contextos educacionais diversos. Ao mesmo tempo, essa diversidade representa potencial de implementar resoluções de problemas nos diferentes setores da educação (RIGO et al., 2012).

Para alcançar os objetivos propostos nesta tese, uma abordagem denominada DEP-DM (*Diagnosis of Educational Problems using Data Mining*) foi proposta. A DEP-DM é baseada na abordagem CRISP-DM aplicada ao contexto de problemas educacionais. No entanto a DEP-DM, diferentemente da CRISP-DM, incorporou algumas fases em uma única no intuito de tornar mais claro os processos realizados. Portanto, a DEP-DM é constituída por cinco etapas (Caracterização do Problema Educacional, Preparação dos Dados para EDM, Modelagem do Problema Educacional, Avaliação do Modelo Preditivo Educacional e Diagnóstico do Problema Educacional) e foi aplicada a dois problemas no contexto educacional: desempenho e evasão. A descrição dos resultados da aplicação de cada fase do DEP-DM proposta nesta tese é descrita nas seções a seguir.

6.1 DIAGNÓSTICO DO DESEMPENHO ESCOLAR

O diagnóstico do desempenho é um mecanismo que busca conhecer e medir o desempenho, estabelecendo uma comparação entre o desempenho esperado e o apresentado (LOTTA, 2002). Para a avaliação diagnóstica na educação, existem muitos métodos e técnicas para serem aplicados a seu favor e de acordo com as variáveis, dimensões e indicadores que se destinam a diagnosticar com um objetivo

específico e em um determinado nível. O valor da avaliação diagnóstica do desempenho não está apenas na correta seleção dos indicadores predeterminados, mas também no uso que dela é feito.

6.1.1 Caracterização do Problema Educacional

Esta fase foi realizada com o objetivo de compreender os fatores associados aos problemas educacionais (desempenho e evasão) a partir de um estudo relacionado aos cenários educacionais, EDM e modelos de regressão. No cenário educacional foi realizado o entendimento dos problemas, suas principais causas, as formas de ocorrência e as principias técnicas aplicadas para identificação e mensuração. Este entendimento foi feito baseado nas teorias relacionadas aos problemas educacionais: evasão e desempenho, como também, por meio da realização de um mapeamento sistemático da literatura (Capítulo 4).

Os dados utilizados neste trabalho são provenientes de bases de dados abertos educacionais fornecidas pelo INEP (INEP, 2018). O objetivo foi realizar o entendimento dos dados e explicar os fatores associados a problemas educacionais: o desempenho e a evasão. Neste contexto, foram utilizadas as seguintes bases de dados:

- Inicialmente, os dados analisados fazem parte da base de dados do Sistema de Avaliação da Educação Básica (SAEB) referente aos anos de 2013, 2015 e 2017; e
- Posteriormente, os dados analisados faziam parte do Censo e Indicadores de Fluxo da Educação Superior referente aos anos de 2013, 2014 e 2015.

As subseções a seguir apresentam os cenários abordados neste estudo para aplicação da abordagem de diagnóstico proposta. Os cenários estão relacionados ao SAEB, e ao Censo e Indicadores de Fluxo da Educação Superior.

6.1.2 Preparação dos Dados para EDM

O SAEB é uma avaliação realizada periodicamente pelo INEP desde 1990 e tem como objetivo, no âmbito da Educação Básica, avaliar a qualidade, a equidade e a eficiência da educação praticada no país e em seus diversos níveis governamentais. O SAEB é constituído, essencialmente, por provas de Lingua Portugesa e Matemática, visando avaliar o desempenho escolar, como também, questionários que buscam analisar

fatores no âmbito escolar, financeiro e social do aluno. Os questionários contextuais devem subsidiar a obtenção de informações acerca dos fatores supracitados que interferem na qualidade da educação e no desempenho escolar. Segundo Rico (1998), o SAEB considera quatro eixos: eficiência no ensino; (medida por meio de provas de avaliação do desempenho); contexto (que engloba nível socioeconômico, perfil e autonomia das escolas); processo (que envolve planejamento e projeto pedagógico); e insumos (que incluem infraestrutura, instalações e equipamentos). Segundo o Inep (2018), as variáveis presentes nos questionários contextuais classificam-se em seis dimensões da qualidade educacional. O Quadro 11, apresenta a Matiz de referência dos questionários contextuais

Quadro 11. Matriz de referência dos questionários contextuais

Dimensão	Temas envolvidos	Tópicos a serem medidos
Atendimento Escolar	Acesso	Proximidade com a residência
	Infraestrutura	Condições de funcionamento da escola; Espaços internos e externos a escola; insumos e recursos.
Ensino e Aprendizagem	Currículo	Previsto; diversificado; e ministrado.
	Práticas pedagógicas	Apoio pedagógico; forma de atuação do professor; relações interpessoais; reprovação; dever de casa.
Investimento	Mecanismos e programas de financiamento público	Controle social dos gastos; Autonomia e verba da unidade escolar.
	Arrecadação de recursos pela escola	Acompanhamento das inicitaivas escolares e de arrecadação
Profissionais da Educação	Formação Profissional	Formação inicial e continuada
	Condições de Trabalho	Recursos: infraestrutura, materiais didáticos; Organização do trabalho; Volume de trabalho; e Aspectos sociais.
	Condições de emprego	Contrato; Remuneração; e Carreira.
Gestão	Planejamento e gestão da escola.	Organização da rede; Gestão pedagógica; Condições da gestão.
	Participação na escola e na rede	Mecanismos de Autoavaliação da escola; Controle social; Locais de tomada de decisão.
Equidade	Contexto socioeconômico, cultural e espacial	Condições socioeconômicas; Recursos para aprendizagem; Expectativas educacionais da família; Envolvimento da Familia na Escola; Recursos Culturais.
	Intersetorialidade	Politicas sociais; Integração de politicas sociais.
	Inclusão	Desigualdades geracionais; étnico-raciais; gênero; Discriminação violência; Etnia e Imigração.

Fonte: Elaborado pelo Autor (2019)

Os indicadores educacionais disponibilizados pelo SAEB atribuem valor estatístico à qualidade do ensino não atendendo somente ao desempenho dos alunos, mas, também, ao contexto econômico social em que as escolas estão inseridas. Eles

consideram informações como o acesso, a permanência e a aprendizagem dos alunos. Os dados relacionados podem ser obtidos em diferentes granularidades, como nível nacional, estadual ou municipal, escola, turma e aluno. Para este cenário de aplicação foi considerado o nível aluno. As variáveis relacionadas ao aluno são classificadas de acordo com as dimensões da qualidade apresentadas na Matriz de Referência elaborada pelo Inep (2018). O Quadro 12 apresenta a definição dos construtos a serem avaliados nos questionários tendo como base os resultados dos modelos teóricos propostos por Andrade e Soares (2008).

Quadro 12. Construtos avaliados nos questionários contextuais

Construto	Descrição	Dimensão
Caracteristicas sociodemográficas	Caracteristicas geracionais; étnico raciais; gênero; local, espaço, ambiente; recursos para aprendizagem, renda.	Aluno
Capital Social	Expectativas educacionais das famílias; Envolvimento da família com a escola; Políticas sociais nas áreas de saúde, trabalho, cultura, assistência, segurança	Familia/Sociedade
Capital Cultural	Recursos culturais disponíveis em casa; Língua falada em casa; participação em eventos culturais.	Sociedade
Motivação e autoestima	Recursos: infraestrutura, Características das relações estabelecidas no ambiente de aprendizagem; materiais didáticos; Apoio pedagógico;	Escola
Práticas de estudo	Forma de atuação do professor; preparação das aulas; uso do tempo; Avaliação em sala (objetos, tipos, posturas); Reprovação (motivos, programas e ações para evitá-la). Dever de casa; Apoio físico e humano.	Aluno / Escola
Trajetória escolar	Evasão; reprovação; Literacia, numeracia e atividades científicas prévias (anteriores à escolarização)	Aluno

Fonte: Elaborado pelo Autor (2019)

Este estudo utilizou dados secundários coletados pelo INEP, relativos às edições do SAEB realizadas em 2013 (subseção 6.2), 2015 e 2017. Foram analisadas as respostas dos estudantes do 5º ano do ensino fundamental das escolas públicas do Estado de Pernambuco. Foram consideradas escolas das diferentes dependências

administrativas (federais, estaduais e municipais) e localizadas tanto nas cidades quanto no meio rural.

A base de dados de 2015 contém informações sobre 2.071.581 de estudantes que participaram da avaliação em todo Brasil. Já a base de 2017 possui informações de 2.193.137 alunos. Para o estado de Pernambuco as bases de dados possuem informações de 105.456 em 2015, e 113.667 estudantes em 2017. Os instrumentos utilizados foram os questionários de contexto do SAEB. O questionário respondido pelos alunos é composto por 47 questões relacionadas, de modo geral, com o perfil do aluno, estrutura e itens presentes nas casas, escolaridade dos pais ou responsáveis, hábitos de leitura e incentivo dos pais aos estudos, trajetória e práticas escolares (INEP, 2018). De posse dessas bases foi considerado como indicador de desempenho o nível de proficiência dos alunos em Lingua Portuguesa (LP) e Matemática (MT). Neste estudo os indicadores de desempenho foram posicionados em quatro níveis qualitativos de aprendizado. O aprendizado adequado engloba os níveis proficiente e avançado, conforme mostrados no Quadro 13.

Quadro 13. Niveis de Proficiência em LP e MT

Nível	Descrição	Proficiências
Insuficiente	Alunos que apresentaram pouqissímo aprendizado. É necessário a recuperação de conteúdos.	Menor que 125
Básico	Alunos que precisam melhorar. Sugere-se atividades de reforço.	Maior ou igual a 125 e Menor que 200
Proficiente	Alunos neste nível encontram-se preparados para continuar os estudos. Recomendam-se atividades de aprofundamento.	Maior ou igual a 200 e Menor que 275
Avançado	Aprendizado além da expectativa. Recomenda-se para os alunos neste nível atividades desafiadoras	Maior ou igual a 275 e Menor que 325

Fonte: INEP (2018)

De acordo com as proficiências apresentadas no Quadro 13, os alunos pernambucanos apresentaram nas avaliações de LP e MT resultados muito abaixo do esperado (INEP, 2018). A Figura 24 apresenta os níveis de proficiência dos alunos pernambucanos nas últimas edições do SAEB.

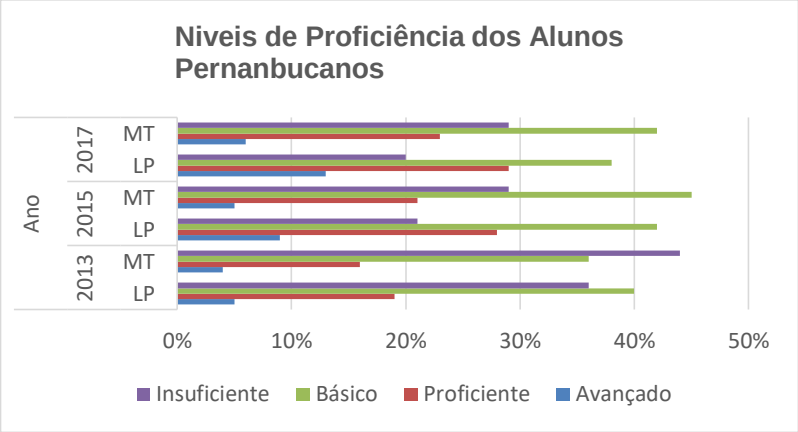


Figura 24. Níveis de Proficiência dos Alunos Pernambucanos no SAEB

Fonte: INEP (2018)

Conforme apresentado na Figura 24, os alunos pernambucanos em sua grande maioria concentram-se no nível básico de aprendizagem em LP e MT, apresentando uma sensível melhora do desempenho em LP no ano de 2015 em relação a 2013 e uma queda em 2017. Entretanto, o nível básico de aprendizagem em MT apresenta um aumento significativo no ano de 2015 em relação a 2013, tendo uma pequena queda em 2017. Com relação aos níveis de proficiente e avançado nos anos de 2015 e 2017 registram uma evolução significativa em LP e MT quando comparadas a 2013. No entanto, apesar da Educação Básica em Pernambuco ter apresentado evolução ao longo dos anos, está longe do ideal, apresentando níveis insuficientes de aprendizado muito significativos principalmente em MT, o que deixa o estado distante da média regional que é de 227.8 pontos, contra 209 pontos obtida pelo estado na ultima edição em 2017. E aquém da média nacional que é 224 pontos.

Objetivando identificar os fatores associados à aprendizagem que vêm afetando os estudantes da educação básica do Estado de Pernambuco, as bases de dados do SAEB foram divididas por ano/disciplina, sendo os estudantes que realizaram a avaliação nos anos de 2015 e 2017, classificados pelos níveis de proficiência em LP e MT. Sendo considerados os alunos cujas proficiências em LP e MT, foram inferiores à média regional e à nacional definidas pelo Inep (2018), para o 5º ano do ensino fundamental nos respectivos anos e disciplinas. Assim foram identificadas as diversas variáveis que compõem as dimensões de fatores associados à aprendizagem, conforme descrito no trabalho de Andrade e Soares (2008).

Após o entendimento das características dos dados, foi realizada uma análise das variáveis e transformações destas, de acordo com as necessidades identificadas. Inicialmente foi realizada a seleção dos dados do estado de Pernambuco, pois consiste no âmbito estudado. Foram excluídas as instâncias que possuíam valor 0 (zero) para o desempenho, ou seja, foram excluídos da base os alunos que não participaram da avaliação. A caracterização do desempenho (variável dependente) foi definida pelas proficiências dos alunos nas disciplinas de LP e MT. Já os fatores associados ao desempenho (variáveis independentes) foram definidos pelas respostas do questionário contextual preenchido pelo estudante que realizou a avaliação. As variáveis redundantes ou irrelevantes foram excluídas e analisou-se a base de dados em busca de registros com valores não preenchidos (*missing values*). Para as variáveis nessa situação, os registros foram preenchidos utilizando a mediana entre os atributos. Os dados do questionário contextual, para efeitos da aplicação das técnicas de regressão, foram transformados em dados numéricos e em alguns casos dicotomizados. Também foi utilizado o método Stepwise para a seleção automática das variáveis. Assim, a Tabela 1 mostra a quantidade de instâncias antes e depois do pré-processamento realizado.

Tabela 1. Dimensão do Conjunto de Dados

Base SAEB	Antes do pré-processamento		Depois do pré-processamento	
	Nº variáveis	Nº instâncias	Nº de variáveis	Nº de instâncias
2015	86	105.456	52	85.036
2017	86	113.667	52	94.554

Fonte: Elaborada pelo Autor (2019)

O método Stepwise foi executado no software Rstudio (R, 2019). O modelo inicialmente incorpora todas as variáveis e depois, por etapas, retira ou não, variáveis do modelo.

A descrição das variáveis selecionadas pelo método Stepwise são apresentadas no Quadro 14. Com a seleção dessas variáveis foi possível a construção dos modelos para a identificação dos subconjuntos útil de preditores para a construção dos cenários. Assim foram elaborados quatro cenários para o estudo: os cenários (1) LP SAEB 2015 e (2) LPSAEB 2017 estão relacionados a disciplina de Língua Portuguesa, os cenários (3) MTSAEB2015 e (4) MTSAEB2017 estão relacionados à disciplina de Matemática.

Quadro 14. Descrição das Variáveis Selecionadas

Cenário 1	Variáveis	Descrição
LPSAEB 2015	BANHEIRO	existência e quantidade de banheiros na residência
	IN_REPROVACAO	existência quantidade de reprovações do aluno
	FZ_DEVERLP	o aluno faz dever de casa de Lingua Portuguesa
	IN_ABANDONO	existência quantidade de abandonos da escola pelo aluno
	TRB_FORA	o aluno trabalha fora de casa
Cenário 2	**Variáveis**	**Descrição**
LPSAEB 2017	BANHEIRO	existência e quantidade de banheiros na residência
	QUARTOS	existência quantidade de quartos na residência
	IDADE	Idade do aluno
	LER_LIVROS	frequencia de leitura de livros
	LER_QUADRINHO	frequencia de leitura de quadrinhos
	TRB_FORA	o aluno trabalha fora de casa
	IN_REPROVACAO	existência quantidade de reprovações do aluno
	IN_ABANDONO	existência quantidade de abandonos da escola pelo aluno
	FZ_DEVERLP	o aluno faz dever de casa de Lingua Portuguesa
Cenário 3	**Variáveis**	**Descrição**
MTSAEB 2015	BANHEIRO	existência e quantidade de banheiros na residência
	GDUPLEX	existência e quantidade de geladeiras duplex
	DOMESTICA	existência e quantidade de empregadas domesticas
	ID_AREA	area de residência do estudante: Capital ou Interior
	MQLAVAR	existência e quantidade de máquinas de lavar
	INC_DEVER	frequencia com que os pais incentivam a fazer o dever de MT
	FZ_DEVERMT	o aluno faz dever de casa de Matemática
Cenário 4	**Variáveis**	**Descrição**
MTSAEB 2017	ID_LOCALIZACAO	a residencia do aluno é na área urbana ou rural
	INC_LER	frequencia com que os pais incentivam a leitura sobre MT
	FREQ_REUNIAO	frequencia com que os pais vão as reuniões na escola
	ESCOL_MAE	nível de escolaridade da mãe
	NUM_PESSOAS	quantidade de pessoas morando na residência
	ALFAB_MAE	a mãe sabe ler e escrever

Fonte: Elaborado pelo Autor (2019)

O Quadro 15 apresenta as variáveis associadas com maior influência positiva ou negativa no desempenho em LP e MT, além dos seus respectivos coeficientes. Além disso, para a correlação dessas variáveis com o desempenho escolar, foi tomado como base o modelo teórico proposto por Andrade e Soares (2008). Esse modelo define quatro dimensões para os fatores associados ao desempenho como: aluno, família, escola e sociedade, os quais englobam os construtos: sóciodemográficos, capital social, capital cultural, motivação, práticas de estudos e a trajetória escolar. Ainda relacionado ao aluno, os itens funcionais e estruturais da sua residência tais como quantidade de trabalhadores domésticos, quantidade de quartos, quantidade de geladeiras e quantidade de banheiros, também têm participação significativa no seu desempenho. Por fim, como fatores relacionados à família tem-se o incentivo dos pais à leitura, se os pais vão às reuniões da escola, o nível de escolaridade da mãe e se a

mãe sabe ler e escrever, os quais apresentam significante influência no desempenho escolar.

Quadro 15. Variáveis associadas ao Desempenho Escolar

Variável	Coeficiente	Construto	Dimensão
Qt. De pessoas que moram na residência	-0,97	Sociodemográfico	Aluno
Área de residência do estudante	0,95	Sóciodemográfico	Aluno
Pais incentivam a fazer o dever de MT	0,93	Capital Social	Familia
Qt. De quartos na residência	0,90	Sóciodemográfico	Aluno
Qt de trabalhadores domésticos (a)	-0,90	Sóciodemográfico	Aluno
Pais Incentivam a ler	-0,86	Capital Social	Familia
O aluno faz o dever de MT	0,80	Praticas de Estudo	Escola
Qt. De geladeiras na residência	0,81	Sociodemográfico	Aluno
Os pais vão as reuniões da escola	0,82	Capital Social	Familia
Qt. De banheiros na residência	0,76	Sociodemográfico	Aluno
Nível de escolaridade da mãe	0,77	Capital Social	Familia
Mãe alfabetizada	0,78	Capital Social	Familia

Fonte: Elaborado pelo Autor (2019)

Dos fatores associados ao desempenho escolar identificados, os que exercem influência mais significativa são: a quantidade de pessoas que moram na residência, a área de residência do estudante (urbana ou rural) e o incentivo dos país à realização de tarefas de Matemática. Com relação à classificação os fatores identificados pertencem aos construtos sociodemográficos, e de capital social, estando relacionados às dimensões do aluno e da família (ANDRADE e SOARES, 2008). Como fator relacionado às atitudes do aluno em relação à escola e ao construto das práticas de estudo, destaca-se a realização de tarefas de MT, a qual se mostrou um item importante dentro da dimensão escola.

6.1.3 Modelagem dos Problemas Educacionais

Nesta fase foram definidas as técnicas de modelagem de dados, especificamente um conjunto de algoritmos de previsão de acordo com as variáveis selecionadas pelo método Stepwise, presentes no Quadro 15. Os métodos paramétricos de Regressão Linear (RL), Regressão Linear Robusta (RLR), Regressão Quantílica (RQ) e não paramétricos, *Support Vector Regression* (SVR), foram utilizados para verificar o desempenho dos alunos em relação aos fatores associados à aprendizagem. Os modelos foram descricos na Seção 3.4.2. Todos os experimentos foram desenvolvidos no ambiente *open source* do R Studio, seguindo os passos do Algoritmo 1.

Algoritmo 1 Composição Experimental

1: **Definir** $n = 30$
2: **Para** todo i igual $1 \leq i \leq n$ **faça**:
3: **Particionar** aleatoriamente a base dados em treino e teste (75-25%).
4: **Construir** os modelos de regressão a partir da base de treino (Equações 1, 2, 3, 4, 5 e 6.)
5: **Estimar** a variável resposta de novos exemplos a partir da base de teste.
6: **Calcular** o erro de predição (MAE) para cada modelo construído.
7: **Salvar** os valores de erro de cada modelo.
8: **Fim Para**
9: **Gerar** as análises com base na amostra do MAE para cada modelo (boxplot, média, desvio padrão e teste estatístico.

Fonte: Elaborado pelo Autor (2019)

6.1.4 Avaliação do Modelo Preditivo Educacional

Nesta fase os modelos desenvolvidos serão avaliados de acordo com os critérios de EDM definidos. Um dos índices de desempenho mais utilizados para o cálculo da previsão baseia-se no erro de previsão. Os resultados foram analisados por meio do erro absoluto médio (ver Equação 2.6).

Foram construídos quatro Cenários de aplicação do modelo de acordo com as variáveis explicativas selecionadas pelo método estatístico *Stepwise* (mostradas no Quadro 15). A primeira subseção agrupa os resultados para a predição da proficiência em LP (Cenário 1 e 2) e a segunda apresenta os resultados para a predição da proficiência em MT (Cenário 3 e 4).

A Tabela 2 mostra os valores da média e desvio padrão obtidos das amostras de cada modelo de regressão construído. O menor valor médio de erro obtido é do *SVR* tanto para o Cenário 1 quanto para o Cenário 2. A Figura 25 mostra o boxplot

após as execuções das técnicas de regressão utilizando as variáveis explicativas definidas para este grupo. Em relação à variância amostral, pode-se ver que entre os métodos há alta variância, mas sem muita diferença entre uma técnica e outra (também observado nos valores de desvio padrão da Tabela 2).

Tabela 2. Valor do erro médio das execuções (desvio padrão) – Cenários 1 e 2.

Cenário/Modelo	RL	RLR	RQ	SVR
1 – LP 2015	19,3823 (**0,0739**)	19,3404 (0,0746)	19,3033 (0,0769)	**19,2826** (0,0791)
2 – LP 2017	20,0586 (**0,0716**)	19,9834 (0,0747)	19,9204 (0,0860)	**19.8688** (0,0888)

Fonte: Elaborada pelo Autor (2019)

Conforme apresentado na Figura 25, as técnicas paramétricas (RL, RLR e RQ) obtiveram uma mediana amostral maior nos dois cenários estudados. O modelo SVR apresenta a menor mediana amostral para o valor do erro médio absoluto. Já o modelo RL possui maior mediana tanto em (a) quanto em (b). A presença de maior dispersão nos dados pode contribuir para esse pior desempenho, enquanto pode ser visto na análise do modelo RLR que a baixa sensibilidade a *outliers* fez com que se obtivesse um desempenho um pouco melhor. Ainda, a análise dos resíduos do modelo RL mostra uma não normalidade, não satisfazendo uma das suposições de sua aplicação. Ou seja, não há uma garantia da explicação do modelo, portanto a regressão linear múltipla não seria a mais indicada.

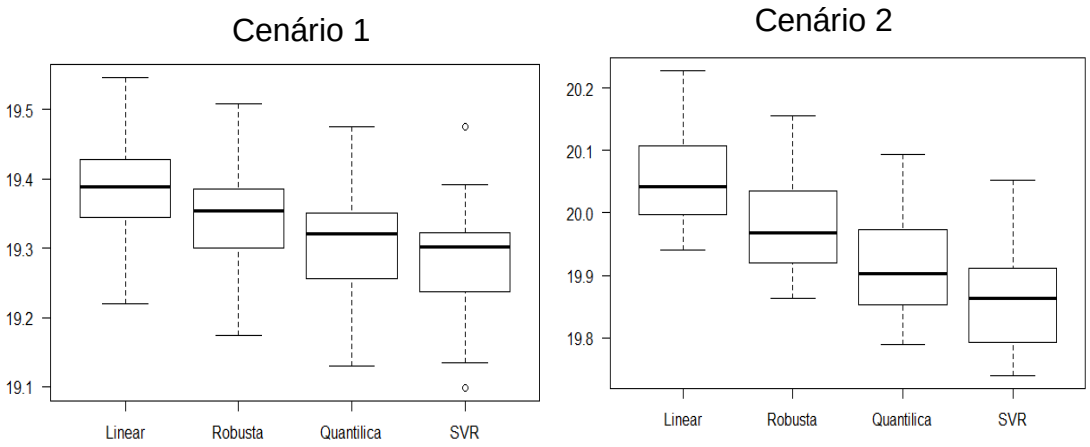


Figura 25. Boxplot para amostra dos erros gerados para Cenários 1 e 2
Fonte: Elaborada pelo Autor (2019)

O teste de hipóteses proposto por Wilcoxon (1945) foi realizado para verificar se o modelo SVR apresenta menor erro que os demais, ou seja, teste unilateral para amostras pareadas. Os resultados foram analisados por meio dos valores de p-valor

(9,313x10-10) para estes dois Cenários, a um nível de significância de 5%. Eles indicaram que, estatisticamente, há evidências de que a SVR tem uma média de amostra menor do que outras técnicas.

A Tabela 3 apresenta os valores de média e desvio padrão obtidos das amostras para os Cenários 3 e 4. O menor valor médio de erro obtido é do modelo RQ em ambos os Cenários. A Figura 26 mostra o *boxplot* após as execuções das técnicas de regressão utilizando as variáveis explicativas definidas para estes grupos. Pode ser observado *outliers* nos *boxplots* dos modelos. Isto indicia a presenta de valores discrepantes nos dados. Dessa forma, pode-se verificar o desempenho da RL, a qual possui sensibilidade a *outliers*, e apresenta a maior mediana amostral em (a) e (b). A RLR apresentou um erro menor comparada a RL, visto seu desempenho em resposta aos *outliers*.

Tabela 3. Valor do erro médio das execuções (desvio padrão) – Cenários 3 e 4.

Cenário/Modelo	RL	RLR	RQ	SVR
3 –MT 2015	15,8781 (**0,0458**)	15,8285 (0,0464)	15,7795 (0,0478)	**15,8038** (0,0461)
4 – MT 2017	17,6607 (**0,0558**)	17,5758 (0,0562)	**17,5026** (0,0560)	17,5175 (0,0563)

Fonte: Elaborada pelo Autor (2019)

No entanto, a menor mediana do MAE é apresentada pela RQ, pois além de ser robusta em resposta aos *outliers*, utiliza em sua estimativa diferentes quantis dos dados. Neste estudo utiliza-se a mediana dos dados (quantil 0.5), sendo a mais representativa faixa para estes cenários. Apesar da SVR ser de natureza não-paramétrica e nos resultados anteriores apresentar melhor desempenho não conseguiu superar a RQ para esse grupo de experimentos.

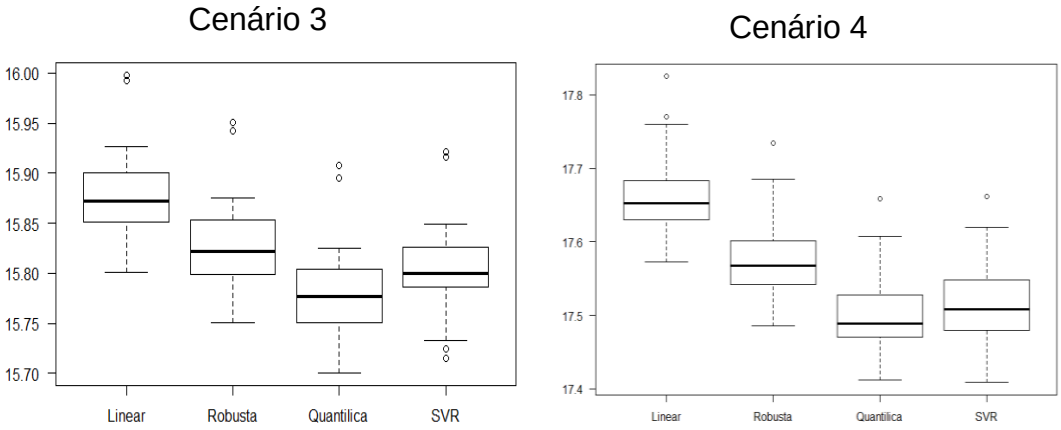


Figura 26. Boxplot para amostra dos erros gerados para Cenários 3 e 4.
Fonte: Elaborada pelo Autor (2019)

Foram analisados os resultados do teste de hipóteses de Wilcoxon unilateral pareado, por meio dos valores de p-valor ($9{,}313x10^{-10}$), para este cenário. A um nível de significância de 5%, os testes indicaram que, estatisticamente, há evidências de que a RQ possui menor média amostral do que outras técnicas para esses cenários. Portanto, pode-se ver que a aplicação da RQ proporcionou um melhor resultado na previsão do desempenho dos alunos nos cenários 3 e 4.

6.1.5 Diagnóstico do Problema Educacional

Esta etapa tem como objetivo apresentar o diagnóstico do problema educacional estudado. O processo de diagnóstico baseia-se no modelo teórico do desempenho escolar proposto por Soares (2004), na abordagem de causa e efeito proposta por Ishikawa (1960;1993) e nos construtos relacionados às carcterísticas contextuais do aluno propostos pelo Inep (2018). O objetivo é, a partir dos fatores identificados e dos resultados apresentados, relacioná-los às causas propostas na teoria, determinando a sua relação de causa e efeito. Além disso, propor ações para possíveis intervenções dos principais envolvidos com a educação (Governo, Instituições Educacionais, Professores, Alunos e a Sociedade).

O processo inicia-se com a identificação das causas que exercem influência sobre o problema do desempenho educacional. Neste caso, conforme Soares (2004), os fatores (causas) pertencem a quatro grupos: Aluno, Familia, Escola e Sociedade. A Figura 27 apresenta o diagrama de causa e efeito para problemas educacionais, no qual pode-se visualizar essa relação.

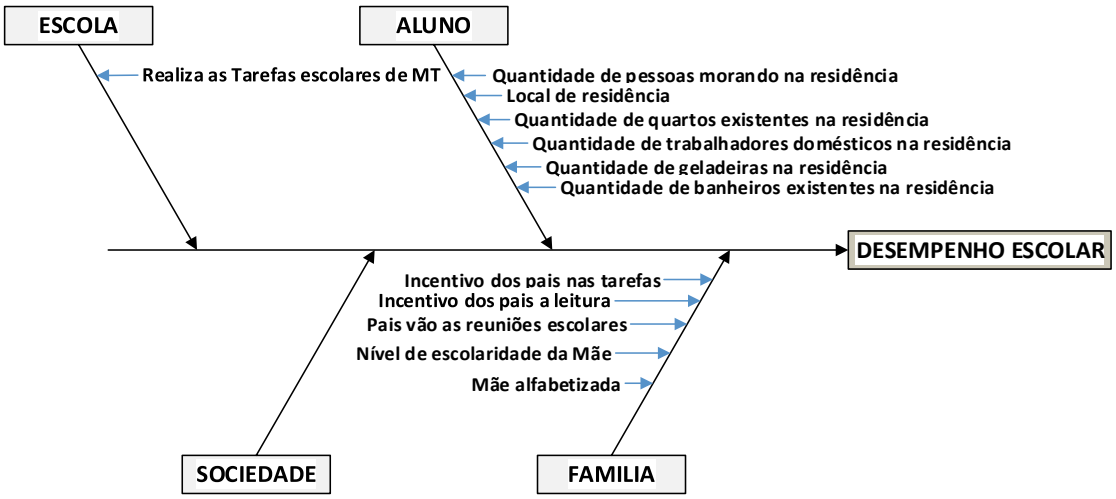


Figura 27. Diagram de Causa e Efeito do Desempenho Escolar.
Fonte: Elaborada pelo Autor (2019)

O diagrama de causa e efeito, apresentado pela Figura 27, destaca as principais causas relacionadas ao fracasso escolar, identificadas neste estudo. Conforme o modelo teórico proposto por Soares (2004), o qual define as quatro dimensões de causas dos problemas educacionais como sendo relacionadas ao aluno, à família, à escola/instituição e à sociedade. Contudo, as principais causas que afetam o desempenho do aluno nas disciplinas de LP e MT estão relacionadas às características de infraestrutura de sua residência como localização, quantidade de comôdos, equipamentos de uso residencial como geladeira, além da existência de trabalhador doméstico. Essas características são relacionadas a aspectos socioeconômicos como apontado por DUARTE (2013) e SOARES (2005). Outros fatores de destaque são os relacionados à família como por exemplo o envolvimento dos pais nas atividades escolares dos filhos, no incentivo à execução das tarefas escolares e à leitura, participação nas reuniões escolares para acompanhar o percurso educacional do filho na escola, além de aspectos relacionados ao nível de alfabetização dos país, em especial o da mãe como apontado pelos resultados (SOARES, 2005).

6.1.6 Discussões

Após a análise desses resultados, percebe-se que a aplicação da SVR proporcionou um melhor resultado na predição do desempenho dos alunos nos cenários que buscam a estimação do valor proficiência em Língua Portuguesa. A vantagem de utilizar uma técnica não paramétrica é que não há suposição da relação entre a variável resposta e a variável explicativa. Assim, formando curvas nos ajustes dos dados (e não uma suposição de parâmetros, como nas regressões paramétricas utilizadas), esse tipo de técnica proporciona melhor desempenho. Além disso, SVR pode obter melhores resultados porque procura o valor ideal no espaço de pesquisa.

Já para os cenários que consideram a proficiência em Matemática como variável resposta, identifica-se o modelo RQ como o que reduz o erro de predição. A SVR nestes grupos de experimentos não conseguiu ser superior, apresar de apresentar resultados mais significativos que as demais técnicas paramétricas. Pode-se perceber nos *boxplots* dos modelos a presença de *outliers*. Ao considerar a mediana dos dados, a RQ quantílica estima sobre a faixa de dados mais significativas, sendo mais robustas a *outliers*.

Estas características da RQ reduziram o erro de predição para este grupo.

Com esses resultados, pode-se observar que técnicas de regressão não-paramétrica podem ser aplicadas em cenários educacionais. O poder de ajustes e flexibilidade aos dados justifica a aplicabilidade desse tipo de modelagem quando os métodos paramétricos são insuficientes. A técnica não-paramétrica utilizada traz uma modelagem de regressão mais flexível, buscando uma resposta que melhor corresponda aos dados. Essas técnicas trazem ganhos para a área educacional em direção à predição de fatores educacionais com mais precisão e menor erro. Contudo, percebe-se que regressão paramétrica pode apresentar resultados superiores em relação às não-paramétricas, tendo em vista as características dos dados utilizados e do modelo utilizado. A regressão quantílica tem a vantagem de estimar a mediana (em vez da média em RL), portanto, seu resultado vai ser mais robusto em resposta à existência de *outliers* nos dados.

Neste estudo utiliza-se um conjunto de variáveis explicativas, as quais foram selecionadas por cenários de aplicação. Estes cenários correspondem a predizer o valor da proficiência dos alunos em Língua Portuguesa e Matemática utilizando dados do SAEB dos anos de 2015 e 2017. Enquanto a maioria dos trabalhos encontrados aplica tarefas de classificação para os dados educacionais, neste trabalho foi proposta a estimação do valor da proficiência investigando os resultados de diferentes modelos de regressão (paramétrico e não-paramétrico) para encontrar o que reduza o erro de predição.

Além disso, foi possível propor uma abordagem para diagnosticar o problema do desempenho educacional a partir das relações de causa e efeito entre os fatores associados e o desempenho escolar. Foram desenvolvidos um diagrama de causa e efeito e um relatório de diagnóstico, nos quais foi possível visualizar os fatores associados e suas relações com o problema de forma sistemática. Além disso, o relatório proposto serve como instrumento para que os agentes educacionais possam intervir, adotando as ações estabelecidas a curto, médio e longo prazos. Estes instrumentos buscam disponibilizar à comunidade educacional uma forma de visualizar os problemas e suas causas além, do conhecimento das ações a serem tomadas para melhorar a qualidade da educação em nosso país.

6.2 MODELOS ENSEMBLE PARA O DIAGNÓSTICO DO DESEMPENHO EDUCACIONAL

Seguindo a metodologia proposta no Capitulo 5, as próximas subseções mostram a aplicação de suas etapas para o cenário de EDM, considerando o diagnóstico da evasão escolar dos estudantes do 5º ano do ensino fundamental que realizaram o SAEB no ano de 2013. Os algoritmos utilizados para o desenvolvimento do MER foram apresentados no Capitulo 3.

6.2.1 Caracterização do Problema Educacional

O sucesso escolar nem sempre é alcançado por todos os alunos ao longo dos ciclos escolares. Tarefas repetitivas, ausência de relação entre os conteúdos escolares e vivências dos alunos, condições de trabalho docente precárias, avaliações maçantes e rígidas têm saturado o ambiente escolar prejudicando o processo de ensino aprendizagem (ZAMBOM e ROSE, 2012). Tais dificuldades enfrentadas pelos estudantes ao longo do percurso escolar podem ser analisadas por meio do fenômeno do fracasso escolar que se manifesta pelos problemas na aprendizagem, problemas de comportamento, baixo desempenho escolar, reprovações, evasões e abandonos (DAZZANO, CUNHA, LUTTIGARDS, ELIAS, 2016).

O desempenho escolar pode ser entendido como a capacidade que o aluno tem de expressar sua aprendizagem e seu conhecimento adquirido no processo de ensino-aprendizagem (PERRENOUD, 2003). Este infere nas habilidades acadêmicas dos alunos e tem caráter avaliativo na medida em que os estudantes devem demonstrar em suas respostas em testes e provas, por exemplo, o que aprenderam nas aulas. Segundo D´Abreu e Maturano (2010), o baixo desempenho escolar ocorre quando o aluno apresenta em notas ou tarefas, um resultado abaixo do nível esperado para sua idade, habilidade e potencial de um indivíduo.

6.2.2 Preparação dos dados para EDM

Para este contexto de aplicação, foi utilizada a base de dados do SAEB realizado no ano de 2013. Essa base foi considerada para verificar os fatores associados que influenciam no fracasso educacional dos estudantes do 5º ano das escolas públicas do Estado de Pernambuco. O exame é composto de alguns instrumentos. Nesta pesquisa foi utilizado o questionário contextual do aluno, considerando informações que caracterizam o perfil dos alunos.

Os conjuntos de dados devem ser preparados adequadamente. Assim, foram realizadas as seguintes atividades:

- Filtrar os dados para atender o foco principal desse estudo;
- Verificar valores ausentes ou em branco;
- Normalização de dados; e
- Seleção de variáveis.

O processo de filtragem de dados tem como objetivo extrair da base original os dados de interesse para este estudo. Neste caso, serão considerados os dados referentes ao Estado de Pernambuco. A base de dados do SAEB relacionada ao estado de Pernambuco possui dimensão inicial de 105.451 instâncias com 214 atributos referentes a informações dos alunos e das escolas publicas e privadas do Estado.

Visando identificar de forma mais ampla os fatores associados à aprendizagem que afetam de forma expressiva o desempenho dos alunos pernambucanos, foi realizada a junção da base de dados aluno e escola. O objeto deste estudo restringiu-se a análise apenas dos alunos e escolas públicas (estaduais e municipais) do Estado de Pernambuco. A razão dessa restrição deve-se a esses alunos e escolas apresentarem os maiores índices de fracasso escolar conforme dados do INEP (2018). Os dados mostram que apenas 19% dos alunos dessas escolas atingiram o aprendizado esperado em Lingua Portuguesa (LP) e 16% em Matemática (MT) no referido ano. Os números ainda mostram que 37% dos alunos das escolas estaduais apresentam níveis de pouco aprendizado em MT e LP.

A partir desse cenário buscou-se identificar as diversos fatores associados que compõem as dimensões propostas no modelo teórico proposto por Soares (2004) que são: Aluno, Familia, Escola e Sociedade.

Foi realizada uma análise das variáveis e transformações destas de acordo com os objetivos e das necessidades identificadas neste estudo. O processo foi iniciado com a seleção dos dados, sendo excluídas as instâncias que possuíam o valor 0 (zero) para o desempenho, ou seja, foram excluídas da base as informações dos alunos que não participaram da avaliação. Além disso foram excluídas as escolas que não responderam o questionário contextual. A caracterização do desempenho (variável dependente) foi definida pelas proficiências dos alunos nas disciplinas de LP e MT. Já os fatores associados ao desempenho (variáveis independentes) foram definidos

pelas informações presentes nos questionários contextuais respondidos pelos alunos e pelas escolas. As variáveis redundantes ou irrelevantes foram excluídas e analisou-se a base de dados em busca de registros com valores não preenchidos (*missing values*). Para as variáveis nessa situação, os registros foram preenchidos utilizando a mediana entre os atributos. Os dados do questionário contextual, para efeitos da aplicação das técnicas de regressão, foram transformados em dados numéricos e em alguns casos dicotomizados.

Para a seleção das variáveis foram utilizados dois métodos: Stepwise e a Correlação de Pearson. Assim, a Tabela 4 mostra a quantidade de instâncias antes e depois do pré-processamento realizado.

Tabela 4. Dimensão do Conjunto de Dados

Base SAEB	Antes do pré-processamento		Depois do pré-processamento	
	Nº variáveis	Nº instâncias	Nº de variáveis	Nº de instâncias
2013	214	105.451	147	48.609

Fonte: Elaborada pelo Autor (2019)

A descrição das variáveis selecionadas pelos métodos Stepwise e Correlação de Pearson são apresentadas no Quadro 16, apartir da seleção das variáveis relacionadas aos cenários de LP e MT, foi possível a construção dos modelos preditivos.

Quadro 16. Variáveis Selecionadas pelo Método Stepwise e Pearson

Cenário	Variáveis	Descrição
LP SAEB 2013 Método Stepwise	LEITUMAE	Mãe costuma a ler
	INCENTDEVER	Pais incentivam a fazer o dever LP
	PCONVESC	Pais conversam sobre a escola
	PATIO	Estado de conservação do pátio da escola
	SEGDIURNO	Escola possui segurança diurno
	POANTFURTO	Escola possui Sistema Antifurto
	PROTEQUIP	Escola possui Mecanismos de Proteção dos equipamentos
	SLESTUDO	Escola possui sala de estudo
	MIDILAZER	Escola possui Mídias (vídeo ou DVD) lazer
Cenário	**Variáveis**	**Descrição**
LP SAEB 2013 Método Correlção Pearson	TAXA_PARTICIPACAO 5EF	Indicador de participação na avaliação
	SLESTUDO	Escola possui sala de estudo
Cenário	**Variáveis**	**Descrição**

MT SAEB 2013 Método Stepwise	LEITUMAE	Mãe costuma a ler
	PC_FORMAÇÃO DOCENTE INICIAL	Indicador de adequação da formação docente
	SILUMI	Quantidade de salas iluminadas
	PATIO	Estado de conservação do pátio da escola
	SEGDIURNO	Escola possui segurança diurno
	POANTFURTO	Escola possui Sistema Antifurto
	PROTEQUIP	Escola possui Mecanismos de Proteção dos equipamentos
	SLESTUDO	Escola possui sala de estudo
	MIDILAZER	Escola possui Mídias (vídeo ou DVD) lazer
Cenário	**Variáveis**	**Descrição**
MT SAEB 2013 Correlação Pearson	TAXA_PARTICIPACAO 5EF	Indicador de participação na avaliação

Fonte: Elaborado pelo Autor (2019)

No Quadro 16, foi possível visualizar as variáveis associadas com maior influência positiva ou negativa no desempenho em LP e MT. Além disso, para a correlação dessas variáveis com o desempenho escolar, foi tomado como base o modelo teórico proposto por Soares (2004). Tal modelo classifica os fatores associados ao desempenho como pertencendo a quatro grupos: ao aluno, à família, à escola e à sociedade. Estes englobam os construtos: sóciodemográficos, capital social, capital cultural, motivação, praticas de estudos e a trajetória escolar.

Dos fatores associados ao desempenho escolar identificados, os que exercem influência mais significativa no desempenho em LP e MT são: a participação do aluno na avaliação, a existência de sala de estudo na escola, o hábito de leitura da mãe, a formação docente inicial, a existência e estado de conservação do pátio da escola, a existência de segurança durante o período diurno na escola, a existência de sistema antifurto na escola, a mecanismos de proteção dos equipamentos, a existência de mídias como vídeos ou DVD, para o lazer dos alunos.

6.2.3 Modelagem dos Problemas Educacionais

Nesta etapa, utilizou-se os seguintes algoritmos: RL, RR, RD, SVR, *Bagging* (B-RL, B-RR, B- RD, ML- SVR) e o modelo proposto nesse estudo PM *(PredictiveModeling*) (PM1, PM2, PM3 e LM4). Os modelos propostos foram combinados da seguinte forma:

- PM1: ensemble Bagging com Regressão Linear;
- PM2: ensemble Bagging com Regressão Robusta;
- PM3: ensemble Bagging com Regressão Ridge; e
- LM4 (Modelo da Literatura), representa o modelo proposto por (Nascimento, 2018) para a construção do modelo baseado em *Stacking* como meta-preditor a RD e as regressões para compor o *ensemble* são: *linear regression*, regressão lasso, *Bagging*, *boosting*, *Randon forest*, *Support Vector Machine*, *Knearest neighbors*.

6.2.4 Avaliação do Modelo Preditivo Educacional

Nesta etapa foram avaliados os resultados obtidos verificando sua relação com os objetivos propostos nesta tese para o diagnóstico e predição do desempenho educacional. Para avaliar com precisão os modelos preditivos, utilizou-se o MAE, além de gráficos e testes estatísticos. Com a amostra de 100 interações, pode-se calcular o desvio padrão (SD) do erro e realizar testes estatísticos. Neste estudo foi utilizado o teste de Komolgorov- Smirnov e teste de hipóteses, além de gráficos *boxplots* para avaliar também o desempenho dos modelos por meio do ganho relativo (RG).

A Tabela 5 mostra os valores médios do MAE e do desvio padrão (SD) para os modelos propostos após as 500 interações. Vale a pena destacar o PM1 e PM2 obtiveram o menor valor médio para as duas métricas no cenário utilizando o Método Stepwise. Enquanto no cenário da Correlação de Pearson, os modelos PM1, PM2 e LM4 apresentaram resultados bem similares, enquanto o PM3 obteve maior erro médio em ambas as técnicas de escolha de variáveis utilizados. Pode-se verificar ainda que os modelos PM1 e PM2 propostos obtiveram erros médios inferiores ao modelo da literatura (LM4) usando *Stepwise*.

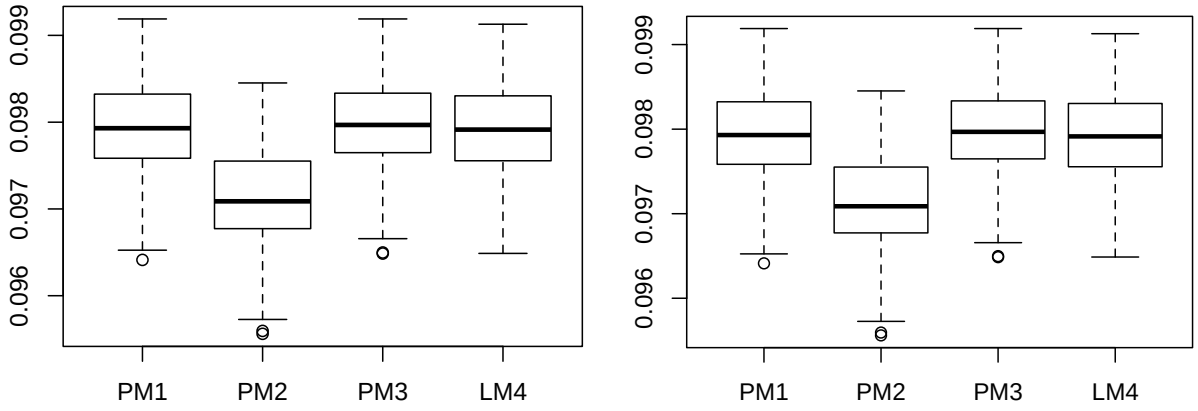
Tabela 5. Média e Desvio Padrão - Conjunto de dados LP SAEB 2013

Técnica escolha das Variáveis	PM1	PM2	PM3	LM4
Stepwise	**$9{,}86 \times 10^{-2}$** ($5{,}47 \times 10^{-4}$)	**$9{,}83 \times 10^{-2}$** ($5{,}49 \times 10^{-4}$)	$1{,}04 \times 10^{-1}$ ($5{,}47 \times 10^{-4}$)	$1{,}08 \times 10^{-1}$ ($6{,}52 \times 10^{-4}$)
Correlação	**$9{,}85 \times 10^{-2}$** ($6{,}03 \times 10^{-4}$)	**$9{,}84 \times 10^{-2}$** ($6{,}03 \times 10^{-4}$)	$1{,}04 \times 10^{-1}$ ($5{,}89 \times 10^{-4}$)	**$9{,}84 \times 10^{-2}$** ($6{,}01 \times 10^{-4}$)

Fonte: Elaborada pelo Autor (2019)

A análise gráfica na Figura 30 apresenta os *boxplots* gerados pelas 500 iterações. Com métrica MAE destaca-se nos gráficos da Figura 28 (a) e (b) que não houve diferença significativa na mediana dos erros entre o PM1, PM2 e LM4 para as

variáveis selecionadas com Correlação/*Setpwise*. Também foi identificada a presença de *outliers* para os modelos utilizando a técnica de *Stepwise* para escolha das variáveis independentes para construção do modelo. Além disso, o modelo PM2 proposto apresenta menor erro de previsão.



a) Stepwise (LP) b) Correlação (LP)

Figura 28. Boxplots dos modelos de conjunto (LP)
Fonte: Elaborada pelo Autor (2019)

A Tabela 6 apresenta o RG do PM2 em relação aos outros modelos utilizados neste trabalho. Ela mostra que PM2 é mais eficiente que os demais modelos propostos (PM1 e PM3), como também, o modelo da literatura (LM4), ratificando os valores médios obtidos na Tabela 5.

Tabela 6. Resultados do RG

Técnica	PM2 x PM1	PM2 x PM3	PM2 x LM4
Stepwise	0.305%	5.79%	9.86%
Correlação	0.1%	5.69%	0%

Fonte: Elaborada pelo Autor (2019)

Portanto, destaca-se que o modelo proposto baseado em *Ensemble Bagging* (PM2) obteve resultado superior ao modelo da literatura baseado em *Ensemble Stacking* (LM4) para o problema do desempenho da proficiência de LP do aluno.

A Tabela 7 mostra os valores médios do MAE e do desvio padrão (SD) para os modelos propostos para MT, após as 500 interações. Vale a pena destacar que o PM1, PM2 e PM3 obtiveram o menor desvio padrão no cenário onde a técnica de seleção de variáveis utiliza o método *Stepwise*. Já no cenário da Correlação de Pearson, os modelos apresentaram resultados bem similares. Pode-se destacar que

os modelos PM1, PM2 e PM3 propostos apresentam menor desvio padrão em relação ao modelo da literatura (LM4) quando utiliza-se a técnica de *Stepwise* para escolha das variáveis independentes para construção do modelo.

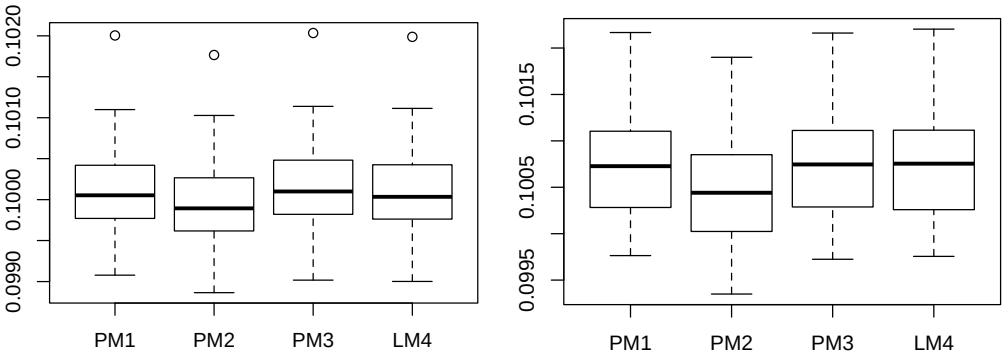
Tabela 7. Média e Desvio Padrão - Conjunto de dados MT SAEB 2013

Técnica escolha das Variáveis	PM1	PM2	PM3	ML4
Stepwise	**$1{,}00 \times 10^{-1}$ $(5{,}43 \times 10^{-4})$**	**$1{,}00 \times 10^{-1}$ $(5{,}40 \times 10^{-4})$**	**$1{,}00 \times 10^{-1}$ $(5{,}25 \times 10^{-4})$**	$1{,}08 \times 10^{-1}$ $(6{,}25 \times 10^{-4})$
Correlação	$1{,}00 \times 10^{-1}$ $(5{,}53 \times 10^{-4})$	$1{,}00 \times 10^{-1}$ $(5{,}53 \times 10^{-4})$	$1{,}01 \times 10^{-1}$ $(5{,}37 \times 10^{-4})$	$9{,}87 \times 10^{-2}$ $(5{,}38 \times 10^{-4})$

Fonte: Elaborada pelo Autor (2019)

A análise gráfica na Figura 29 apresenta os *boxplots* gerados pelas 500 iterações. Com a métrica MAE observa-se nos gráficos (a) e (b) que não houve diferença significativa na mediana dos erros entre o PM1, PM3 e LM4 para as variáveis selecionada com Correlação/*Stepwise*. Também foi identificada a presença de *outliers* para os modelos utilizando a técnica de *Stepwise* para escolha das variáveis independentes para construção do modelo.

Além disso, o modelo PM2 proposto apresenta menor mediana em relação ao erro de previsão dos demais modelos apresentados.



a) Stepwise (MT) b) Correlação (MT)

Figura 29. Boxplot dos modelos de conjunto (MT)
Fonte: Elaborada pelo Autor (2019)

A Tabela 8 apresenta o RG do PM2 em relação aos outros modelos utilizados neste trabalho, comprovando que PM2 é mais eficiente que os demais modelos propostos (PM1 e PM3), como também o modelo da literatura (LM4). Isto ratifica os valores médios obtidos na Tabela 7.

Tabela 8. Resultados do RG

Técnica	PM2 x PM1	PM2 x PM3	PM2 x LM4
Stepwise	0%	0%	8%
Correlação	0%	1%	1.3%

Fonte: Elaborada pelo Autor (2019)

Portanto, destaca-se que o modelo proposto baseado em *Ensemble Bagging* (PM2) obteve resultado superior ao modelo da literatura baseado em *Ensemble Stacking (*LM4) para o problema do desempenho da proficiência de MT do aluno. Ele apresenta uma estimação mais robusta por meio de menores erros de predição para o modelo desenvolvido em relação ao modelo da literatura, tornando-se importante por obter um desempenho em relação a dados que agreguem o modelo final.

Realizou-se o teste de *Kolmogorov-Smirnov* para verificar se o vetor de erros (MAE: MT e LP) referentes às 500 interações seguia uma distribuição normal. Essa hipótese foi rejeitada. Dessa forma foi utilizado o teste de *Wilcoxon* para realizar o teste de hipótese com 5% de significância. A hipótese alternativa elaborada é que o PM2 apresenta melhor desempenho em relação ao PM1, PM3 e LM4.

Assim, foi comprovado estatisticamente, com um nível de confiança de 95%, que o PM2 para a proficiência de MT apresenta menores erros de predição em relação a todos os modelos utilizados para o cenário das variáveis selecionadas com *Stepwise/Correlação*. Utilizando técnica de Correlação o PM2 obteve erros menores que o PM1, PM3 e o LM4 já que os valores de *p-value* obtidos foram*:* 0.0224, 0.009564 e 0.01819, respectivamente.

Também, foi comprovado estatisticamente com um nível de confiança de 95%, que o PM2 para a proficiência de LP apresenta menores erros de predição em relação a todos os modelos utilizados para o cenário das variáveis selecionadas com *Stepwise/Correlação*. Utilizando técnica de Correlação o PM2 obteve erros menores que o PM1, PM3 e o LM4 já que os valores de *p-value* obtidos foram: 2.23×10^{-16}, 2.2×10^{-16} e 1.11×10^{-15}, respectivamente.

6.2.5 Diagnóstico do Problema Educacional

Esta etapa apresenta o diagnóstico do problema educacional no contexto estudado. O processo de diagnóstico baseia-se no modelo teórico do desempenho escolar

proposto por Soares (2004). O objetivo é a partir dos fatores identificados e dos resultados apresentados pelos modelos preditivos, estabelecer uma relação de causa e efeito entre os fatores associados (causas) com o problema educacional (efeito). Para permitir a visualização dessas relações será utilizado o diagrama de causa e efeito proposto por Ishikawa (1993). Dessa forma pretende-se com esse conhecimento contribuir para as intervenções a serem feitas pelos agentes educacionais, governo e sociedade. A Figura 30 apresenta o diagrama de causa e efeito para o problema do fracasso educacional.

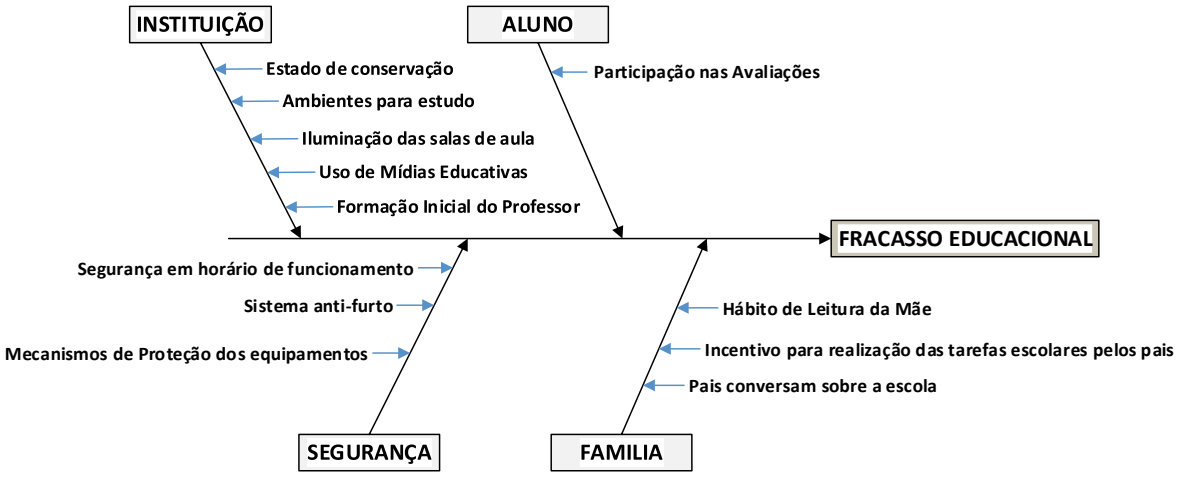


Figura 30. Diagrama de Causa e Efeito do Fracasso Educacional
Fonte: Elaborada pelo Autor (2019)

O diagrama de causa e efeito, apresentado na Figura 30, destaca as principais causas relacionadas ao fracasso escolar, identificadas neste estudo, conforme as dimensões do modelo teórico proposto por Soares (2004). Contudo, devido as mudanças sociais e educacionais nos últimos anos, vêm surgindo novas dimensões de problemas que prejudicam os estudantes interferindo diretamente nas questões relacionadas ao ensino- aprendizagem, bem como ao comportamento e à vida social dos estudantes. Dentre essas novas dimensões, pode-se destacar aspectos relacionados à tecnologia, à saúde, às finanças e à segurança.

Os resultados mostram que as causas relacionadas à instituição/escola têm uma contribuição mais significativa em relação às demais pelo quantitativo de fatores associados. Dentre as causas relacionadas, destacam-se os fatores relacionados à infraestrutura da escola (iluminação das salas, ambientes para o estudo), ao ensino/ aprendizagem (existência e uso de mídias educativas e de laser no cotidiano escolar) e relacionados à formação inicial do professor (poucos professores com formação

especifica para as disciplinas que lecionam). As demais causas também têm sua contribuição significativa ao problema mencionado. Com relação ao aluno, o fato do mesmo não comparecer às avaliações escolares, relacionadas à família, à formação educacional e cultural dos pais aparece como um fator preocupante, além do conhecimento dos pais sobre o andamento escolar dos filhos. Por fim, um fator associado que está em evidência na educação brasileira é a violência. Os dados deste estudo comprovam que fatores associados à violência contribuem para o fracasso escolar dos estudantes. Pode-se destacar neste estudo a existência de segurança, seja eletrônica ou policial nas imediações da escola durante o seu período de funcionamento. Exemplos são a proteção dos bens (equipamentos) da escola de roubo e depredação e a utilização pela escola de sistemas de segurança para repelir ações de roubo e furto dentro ou fora de suas dependências.

Como sugestões de ações de intervenção: (i) adoção de ações dos governos estaduais e municipais para a melhoria da infraestrutura das escolas, bem como da segurança pública; e (ii) a instituição escolar criar estratégias para aproximação dos pais à escola.

6.2.6 Discussões

Este trabalho utilizou modelos de *Ensemble Bagging* para a predição da evasão escolar em Instituições de Ensino Superior no Brasil. Além disso foram utilizados o método Stepwise e a Correlação de Pearson para a seleção das variáveis que serviram com base para a construção dos modelos preditivos. A vantagem de combinar regressores com o método Ensemble Bagging, em uma única previsão, é otimizar os resultados das estimativas.

Os experimentos mostraram que o modelo PM2, o qual combina *Ensemble Bagging* com Regressão Robusta, obteve os melhores resultados em relação ao modelo LM4 que utiliza o método *Ensemble Stacking*, como também em relação ao *Ensemble Bagging* com Regressão Linear (PM1) e Ensemble Bagging com Regressão Ridge (PM3). As métricas com o Método Stepwise e a Correlação alcançaram 95% na previsão do erro em ambos os cenários. Destaca-se ainda, a possibilidade de diagnóstico das causas do fracasso escolar a partir dos fatores associados, relacionando-os aos modelos teóricos da literatura para compreender as causas que influenciam diretamente o referido fracasso escolar.

Os resultados comprovam que os modelos que combinam Ensemble *Bagging* com métodos de Regressão apresentam menor erro quadrático médio em relação aos modelos de regressão simples, ou seja, o *bagging*, pode melhorar o desempenho de preditores instáveis que são basicamente preditores com alta variância. De acordo com os resultados apresentados, os modelos Ensemble Regression trazem ganhos para a área educacional em direção à predição dos problemas educacionais com mais precisão e menor erro.

Nesse estudo utiliza-se um conjunto de variáveis explicativas, as quais foram selecionadas por cenários de aplicação. Estes cenários correspondem a predizer o valor da proficiência dos alunos em Língua Portuguesa e Matemática utilizando dados do SAEB do ano de 2013.

Além disso, foi possível propor uma abordagem para diagnosticar o problema do desempenho educacional a partir das relações de causa e efeito entre os fatores associados e o desempenho escolar. Foi desenvolvido, um diagrama de causa e efeito, onde foi possível visualizar os fatores associados e suas relações com o problema de forma sistemática. A utilização dessa abordagem visa minimizar a falta de compreensão dos resultados relacionados a predição no contexto educacional, servindo como um instrumento para que os agentes educacionais possam intervir nos problemas, adotando as ações pontuais.

6.3 MODELOS ENSEMBLE PARA O DIAGNÓSTICO DA EVASÃO ESCOLAR

Seguindo a metodologia proposta no Capitulo 5, as próximas subseções mostram a aplicação de suas etapas para o cenário de EDM, considerando o diagnóstico da evasão no ensino superior no ano de 2013. Os algoritmos utilizados para o desenvolvimento do MER foram apresentados no Capitulo 3.

6.3.1 Caracterização do Problema Educacional

A evasão escolar está relacionada à perda de estudantes que iniciam, mas não concluem seus cursos. Este é um fenômeno complexo, associado a não concretização de expectativas e reflexo de múltiplas causas que precisam ser compreendidas nos contextos socioeconômico, político e cultural, no sistema educacional e nas instituições de ensino. A evasão escolar significa desistência por qualquer motivo, exceto conclusão ou diplomação, e é caracterizada por ser um processo de exclusão

determinado por fatores e variáveis internas e externas às instituições de ensino, sendo esse fenômeno percebido tanto em instituições públicas de ensino quanto em instituições privadas.

No Brasil, de acordo com os dados do Censo da Educação Superior, as taxas de evasão no ensino superior brasileiro apresentam índices alarmantes. Os dados mostram que 49% dos discentes que ingressaram no ensino superior em 2010 abandonaram os cursos dentro de um intervalo aglomerado de cinco anos. Nas instituições privadas a evasão chegou a 53%, e nas instituições públicas alcançou 47% nas municipais, 38% nas estaduais e 43% nas federais (INEP, 2018).

Diante do cenário exposto, este estudo objetivou propor modelos que combinam *Ensemble* com Regressores para predição da evasão escolar em Instituições de Ensino Superior Brasileiras. Foram consideradas as informações presentes no Censo da Educação Superior e nos Indicadores de Fluxo do Ensino Superior. Essas informaçãoes estão releacionadas a aspectos demográficos, acadêmicos e socioeconômicos. Quando combinadas, essas informações servirão como base para o desenvolvimento dos modelos preditivos.

6.3.2 Preparação dos Dados para EDM

Os indicadores educacionais utilizados pelo estudo são oriundos da Base de Dados do Censo da Educação Superior e dos Inidicadores de Fluxo da Educação Superior, ambas disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teixeira (INEP, 2018), referentes ao ano de 2013.

Para a extração dos dados dessas bases foram levados em consideração os fatores propostos na literatura como fatores associados à evasão, como por exemplo, atributos demográficos, acadêmicos e socioeconômicos. Cabe salientar uma limitação que está relacionada aos dados disponibilizados pelo Censo, a qual é a ausência de outras informações que poderiam complementar ainda mais os atributos extraídos, como por exemplo aspectos relacionados ao aluno, à infraestrutura das instituições, e ao ensino dentre outras.

Depois de realizar o entendimento dos dados a serem trabalhados, foi realizada a junção das duas bases de dados (Censo e Indicadores de Fluxo). Verificou-se a inexistência de dados na variável referente ao abandono escolar. As instâncias que tinham mais de 70% dos valores faltando *missing values* foram excluídas. Para as

demais variáveis que tinham poucos *missing values* foram preenchidas por sua mediana. Também foi utilizado o método Stepwise para a seleção de variáveis. A Tabela 9 apresenta a relação da quantidade de instâncias antes e depois do pré-processmento realizado.

Tabela 9. Dimensão da Base de Dados

Antes do Pré-processamento		Após o Pré-processamento	
Nº de variáveis	Nº de instâncias	Nº de variáveis	Nº de instâncias
50	133.528	14	22972

Fonte: Elaborada pelo Autor (2019)

A descrição das 14 variáveis selecionadas pelo método Stepwise é apresentada no Quadro 17. Além do método Stepwise, utilizou-se a correlação de Pearson para obter as correlações entre as variáveis. A correlação de Pearson foi utilizada para a construção de um outro cenário. Assim, foram elaborados dois cenários de aplicação para o estudo, a saber: um utilizando as variáveis selecionadas pelo método Stepwise e o outro com as 4 variáveis de maior correlação com a variável da evasão escolar nas instituições de ensino superior no Brasil.

Quadro 17. Variáveis Selecionadas para o Estudo da Evasão Escolar

Variáveis	Descrição
CAS	A situação do aluno no curso (ativo, trancado, desvinculado do curso, transferido para outro curso da mesma IES, formado, falecido)
IABP	Informa se o aluno recebe alguma remuneração para a permanência na instituição de ensino superior.
QI	Número de novos alunos
QP	Número de alunos que permanecem no curso
TAP	Indicador de permanência acumulada
TCA	Indicador de conclusão acumulada
INC	Estuda em período noturno
CCRA	Cor/Raça do aluno (branca, preta, parda, amarela, indígena, não declarado, não informado)
IABT	Informa se o aluno recebe remuneração por atividade desenvolvida dentro da instituição de ensino superior
ICE	Informa se o aluno faz atividade extracurricular de estágio não obrigatório
ICEX	Informa se o aluno participa de atividade extracurricular de extensão
QCC	Número de concluintes no curso

IAE	Informa se o aluno participa de alguma atividade extracurricular (estágio, extensão, monitoria e pesquisa)
QC	Número de concluintes
TDA (Y)	Taxa de abandono escolar

Fonte: Elaborado pelo Autor (2019)

Analisando o Quadro 17, foram selecionadas as variáveis de maior correlação que são: TAP (-0,6254), TCA (-02522), INC (0.1566) e QP (-0.1308). As variáveis selecionadas estão relacionadas a aspectos do fluxo do aluno na instituição, como permanência e conclusão do curso. Outra variável significativa foi a INC relacionada aos estudantes que estudam no período noturno, a qual está relacionada às atividades exercidas pelos estudantes fora da instituição educacional, como por exemplo o trabalho. As maiores correlações foram as variáveis TAP e TCA. Este fato pode ser justificado, tendo em vista que esses indicadores estão relacionados diretamente à variável TDA. Assim, quanto maior a permanência do aluno, menor o abandono escolar.

6.3.3 Modelagem do Problema Educacional

Nesta etapa utilizou-se os seguintes algoritmos: RL, RR, RD, SVR, *Bagging* (B-RL, B-RR, B- RD, ML- SVR) e o modelo proposto neste estudo PM *(PredictiveModeling*) (PM1, PM2, PM3 e LM4). Os modelos propostos foram combinados da seguinte forma:

- PM1- ensemble Bagging com Regressão Linear;
- PM2- ensemble Bagging com Regressão Robusta;
- PM3- ensemble Bagging com Regressão Ridge; e
- LM4 (Modelo da Literatura) - representa o modelo aplicado por Nascimento (2018) para a construção do modelo baseado em *Stacking* como meta-preditor a RD. As regressões para compor o *ensemble* são: *linear regression*, regressão lasso, *Bagging*, *boosting*, *Randon forest*, *Support Vector Machine* e *Knearest neighbors*.

6.3.4 Avaliação do Modelo Preditivo Educacional

Nesta etapa foram avaliados os resultados obtidos verificando sua relação com os objetivos propostos nesta tese para o diagnóstico e predição da evasão educacional. Para avaliar com precisão os modelos preditivos, utilizou-se o MAE, além de gráficos e testes estatísticos. Com a amostra de 500 interações, foi possível calcular o desvio

padrão (SD) do erro e realizar testes estatísticos. Neste estudo foram utilizados o teste de Komolgorov- Smirnov e teste de hipóteses, além de gráficos *boxplots* para avaliar também o desempenho dos modelos por meio do ganho relativo (RG).

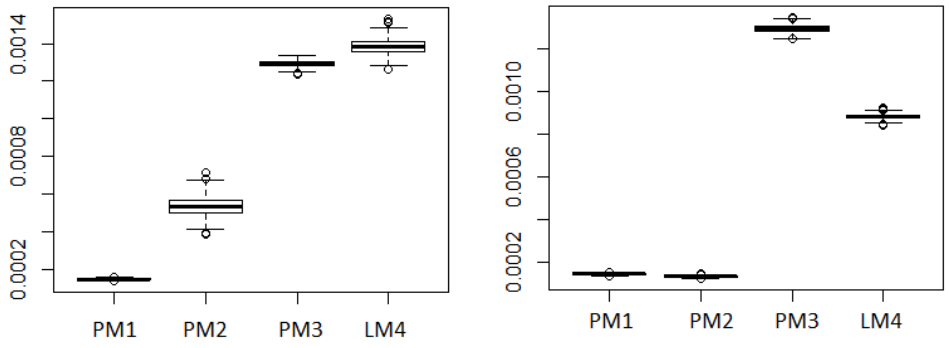
A Tabela 10 mostra os valores médios do MAE e MSE do desvio padrão (SD) para os modelos propostos após as 500 interações. Vale a pena destacar que o PM1 obteve um menor valor médio para as duas métricas no cenário utilizando o Método Stepwise, enquanto que no cenário da Correlação de Pearson, o modelo PM1 e o modelo PM2 apresentaram resultados bem similares. Pode-se verificar ainda que PM1, obteve nos dois cenários valores médios menores que LM4.

Tabela 10. Média e Desvio Padrão – Conjunto de Dados da Evasão

Técnica	Métricas	PM1	PM2	PM3	LM4
Stepwise	MSE	$3.14x10^{-3}$ ($2.75x10^{-4}$)	$1.39x10^{-2}$ ($5.31x10^{-3}$)	$6.54x10^{-3}$ ($2.33x10^{-3}$)	$1.31x10^{-2}$ ($4.57x10^{-3}$)
	MAE	$1.47x10^{-4}$ ($7.89x10^{-3}$)	$5.3x10^{-4}$ ($3.46x10^{-2}$)	$1.29x10^{-3}$ ($1.20x10^{-2}$)	$1.4x10^{-3}$ ($2.69x10^{-2}$)
Pearson	MSE	$2.266x10^{-2}$ ($3.3x10^{-4}$)	$9.29x10^{-3}$ ($2.6x10^{-4}$)	$8.054x10^{-3}$ ($1.04x10^{-4}$)	$4.43x10^{-3}$ ($1.13x10^{-3}$)
	MAE	$1.21x10^{-1}$ ($1x10^{-3}$)	$7.38x10^{-2}$ ($1.15x10^{-3}$)	$7.25x10^{-2}$ ($5.6x10^{-4}$)	$9x10^{-4}$ ($9.33x10^{-3}$)

Fonte: Elaborada pelo Autor (2019)

Para a representação gráfica e análise dos dois cenários, a Figura 31 mostra os *boxplots* gerados pelas 500 interações. Observa-se nos gráficos da Figura 31 que não houve diferença significativa na mediana dos erros entre o PM1 e o PM2 para as variáveis selecionada com correlação. Também foi identificada a presença de *outliers* no PM2 para o cenário Stepwise. Além disso, PM2 apresenta maior variabilidade que os outros modelos para este cenário.



(a) Stepwise MAE **(b) Cor MAE**

Figura 31. Boxplot dos Modelos de Conjunto
Fonte: Elaborada pelo Autor (2019)

Realizou-se o teste de *Kolmogorov-Smirnov* para verificar se o vetor de erros das 500 interações seguia uma distribuição normal. Esta hipótese foi rejeitada. Dessa forma foi utilizado o teste de Wilcoxon para realizar o teste de hipótese com 5% de significância. A hipótese alternativa elaborada é que o PM1 apresenta melhor desempenho em relação ao PM2, PM3 e PM4.

Foi possível comprovar estatisticamente com um nível de confiança de 95% que o PM1 apresenta menores erros de predição em relação a todos os modelos utilizados para o cenário das variáveis selecionadas com *Stepwise*, pois o valor de *p-value* obtido foi de $2{,}47x10^{-7}$. Utilizando o método de correlação o PM1 obteve erros menores que o PM3 e o LM4 já que o valor de *p-value* obtido foi $2{,}47x10^{-7}$. Enquanto que para relação do PM1 ser menor que o PM2, e analisando as métricas MSE e MAE os valores do *p-value* foram de $7{,}916x10^{-2}$ e de $7{,}916x10^{-2}$ respectivamente.

A Tabela 11 apresenta o RG do PM1 em relação aos outros modelos utilizados neste trabalho. Pode-se verificar que o ganho obtido foi muito significativo, monstrando que PM1 é mais eficiente do que os demais modelos, e do que os valores médios obtidos na Tabela 10.

Tabela 11. Resultados do RG

Técnica	PM1x LM4	PM2 x LM4	PM3 x LM4
Stepwise	89,325%	61,464%	65,167%
Correlação	83,431%	84,730%	46,590%

Fonte: Elaborada pelo Autor (2019)

Portanto pode-se ver que o PM1 obteve melhor desempenho na predição da evasão escolar que o LM4 para os dois cenários testados. Além disso, propõe-se a utilização de outros modelos combinados utilizando *Bagging* que também obtiveram bons resultados para o problema da evasão. Assim, comprova-se que os modelos propostos baseados em *Ensemble Bagging* obtiveram resultados superiores ao modelo da literatura (Ensemble Stacking) para o problema da evasão escolar em Instituições de Ensino Superior no Brasil. Pois tais modelos apresentam uma estimação mais robusta por meio de menores erros de predição para o modelo desenvolvido em relação ao modelo da literatura, tornando-se importante por obter um desempenho em relação a dados que agreguem o modelo final.

6.3.5 Diagnóstico do Problema Educacional

Esta etapa apresenta o diagnóstico do problema educacional no contexto estudado. O processo de diagnóstico baseia-se no modelo teórico de Spady (1971) e Tinto (1975, 1993), a partir dos resultados da predição e dos fatores identificados, estabelecendo uma relação de causa e efeito entre eles. Além disso foram propostas soluções para os problemas elencados. A Figura 32 apresenta o diagrama de causa e efeito da evasão no ensino superior.

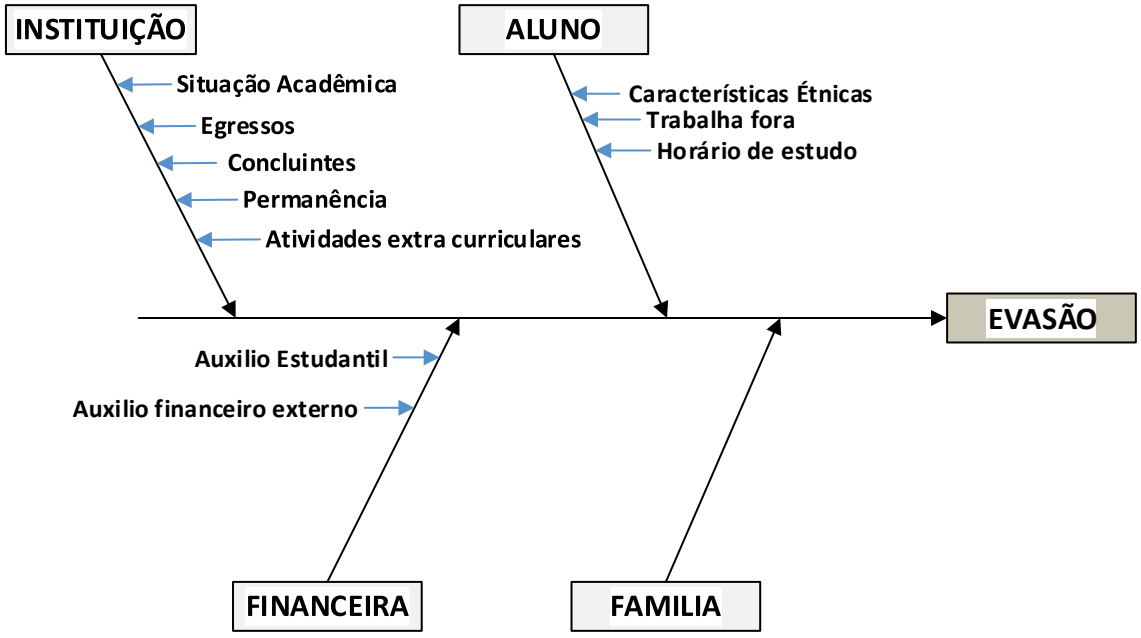


Figura 32. Diagrama de Causa e Efeito da Evasão Escolar
Fonte: Elaborada pelo Autor (2019)

O diagrama de causa e efeito apresenta as principais causas relacionadas à evasão, sendo identificadas neste estudo as causas relacionadas ao aluno, à instituição, além da financeira, como os principais fatores associados ao problema da evasão dos estudantes das Instituições de Ensino Superior brasileiras no ano de 2013. Destacam-se os fatores relacionados à instituição de ensino com o maior número de causas que influenciam a ocorrência da evasão, dentre os quais os fatores associados à permanência e à conclusão do curso. Esses fatores, de acordo com Spady (1970) e Tinto (1993), estão relacionados ao desempenho acadêmico do estudante ao longo de seu percurso dentro da instituição de ensino e a sua adaptação à instituição, fatores estes que podem levá-lo à decisão de abandonar o curso ou a instituição. Outros fatores associados ao estudante e a aspectos financeiros, também se apresentaram significantes, destacando-se os auxílios que o estudante recebe da instituição ou

externos a ela como por exemplo bolsas de pesquisa, estágios e programas governamentais, bem como os aspectos relacionados ao próprio estudante como sociodemográficos (étnico raciais), atividades externas (trabalho) e o comprometimento com a instituição (horário de estudo). Este último obtendo maior destaque em relação aos fatores associados ao aluno.

6.3.6 Discussão

Este trabalho utilizou modelos de *Ensemble Bagging* para a predição da evasão escolar em Instituições de Ensino Superior no Brasil. Além disso foram utilizados o método Stepwise e a Correlação de Pearson para a seleção das variáveis que serviram com base para a construção dos modelos preditivos. A vantagem de combinar regressores com o método Ensemble Bagging, em uma única previsão, é proporcionar a otimização dos resultados das estimativas. Os experimentos mostraram que o modelo PM1, o qual combina *Ensemble Bagging* com Regressão Linear, obteve os melhores resultados em relação ao modelo LM4 que utiliza o método *Ensemble Stacking*, como também em relação ao *Ensemble Bagging* com Regressão Robusta (PM2) e Ensemble Bagging com Regressão Ridge (PM3). As métricas com o Método Stepwise alcançaram 96% na previsão do erro e com o Método de Correlação Pearson, 95%. Destaca-se ainda a possibilidade de diagnóstico das causas da evasão a partir dos fatores associados, relacionando-os aos modelos teóricos da literatura para compreender os motivos que levam os estudantes à decisão de evadir. Os fatores que possuem influência considerável na evasão neste estudo foram: a) indicador de permanência acumulada (TAP); b) indicador de conclusão acumulada (TCA); c) estuda no Período Noturno (INC); e d) Número de alunos que permanecem no curso (QP).

Esses indicadores descrevem a movimentação dos discentes dentro das instituições de ensino desde sua entrada até a conclusão.

6.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capitulo foram utilizadas as técnicas de predição para problemas educacionais, sendo desenvolvido o modelo proposto utilizando as técnicas de RL, RR, RQ, SVM no caso de estudo relacionado ao problema de desempenho educacional. Nos casos relacionados ao fracasso escolar e evasão, foi adicionada a técnica de *Bagging* para

o desenvolvimento do modelo combinado, denominado Ensemble Regression. Ao avaliar o modelo proposto com as demais técnicas de regressão chegou-se à conclusão de que os objetivos propostos neste estudo para dados educacionais foram atingidos por meio dos menores erros de predição alcançados pelo modelo proposto em alguns casos de estudo. A combinação de regressores com o método Ensemble Bagging em uma única previsão, proporciona otimizar os resultados das estimativas. Essa visão holística na qual os modelos são combinados em um todo único e integrado permite resultados mais abrangentes e precisos. Nesta tese, buscou-se propor um modelo preditivo que proporcionasse um melhor ajuste aos dados, com o objetivo de diminuir o erro de estimação dos modelos de regressão utilizados nas variáveis de estudo, para os contextos de EDM estudados. Nestes contextos os modelos propostos conseguiram reduzir o erro para os casos estudados ao estimar o desempenho, o fracasso escolar e a evasão escolar no ensino superior. De forma geral, os modelos propostos nos estudos obtiveram resultados satisfatórios, pois extraíram o melhor de cada uma das técnicas de regressão em um grupo específico de dados nos casos estudados.

Nesse sentido, considerando o caso de estudo relacionado ao desempenho escolar, os resultados mostraram-se satisfatórios, nos quais a aplicação com SVR proporcionou um melhor resultado na predição do desempenho dos alunos nos cenários que buscam a estimação do valor da proficiência em Lingua Portuguesa. A vantagem da utilização da técnica não paramétrica é a suposição da relação entre a variável resposta e a explicativa. Assim, formando curvas de ajuste dos dados e não a suposição de parâmetros, como nas regressões paramétricas utilizadas. Já para os cenários que consideram a proficiência em Matemática, o modelo RQ apresentou-se mais eficiente na redução do erro de predição.

Neste caso, a SVR não conseguiu ser superior. Apesar de apresentar resultados mais significativos que as demais técnicas paramétricas, percebe-se nos *boxplots* dos modelos a presença de *outliers*. Ao considerar a mediana dos dados, a RQ estima sobre a faixa de dados mais significativas, sendo mais robustas a *outliers.* Estas características da RQ minimizaram o erro de predição para este grupo.

O Modelo proposto que utiliza SVR obteve melhores resultados nos cenários 1 e 2 em relação aos modelos RL, RLR, RQ, obtendo métricas de predição do desempenho de 79% e 88% respectivamente. Nos cenários 3 e 4 o modelo proposto

que utiliza RQ obteve os melhores resultados em relação aos demais modelos, obtendo métricas de predição de 47% e 56%, respectivamente.

Os resultados ainda apresentaram os fatores que interferem no desempenho escolar. Conforme o diagrama de causa e efeito, as causas do problema do desempenho estão associadas a questões de infraestrutura da residência do aluno como a quantidade de banheiros e quartos, causas relacionadas a aspectos socioeconômicos como possuir geladeira na residência, de localização da residência se fica na zona urbana ou rural, causas relacionadas à família, as quais referem-se à participação e ao incentivo dos pais na realização das atividades escolares dos filhos, bem como a participação nas reuniões escolares. Por fim, causas associadas ao nível de instrução dos pais, como se a mãe é alfabetizada, e se tem o hábito de leitura.

Considerando o segundo caso de estudo, o qual trata do fracasso escolar, os resultados mostraram-se satisfatórios quando se compara o Modelo *Ensemble Regression* com técnicas não-paramétricas. Procurou-se prever o fracasso escolar a partir de dados do SAEB de 2013. Para os dois casos estudados (LP e MT) foi estatisticamente comprovado que o modelo PM2 (*Ensemblle Bagging* + Reressão Robusta) por meio do ganho relativo (RG) que o modelo é mais eficiente que os demais modelos propostos (P1e PM3), como também o modelo proposto pela literatuta (LM4). Portanto, o modelo *Ensemble Bagging* (PM2) destaca-se obtendo resultado superior ao modelo da literatura baseado em *Ensemble Stacking* (LM4) para o problema de fracasso nas disciplinas de Lingua Portuguesa (LP) e Matemática (MT). Para ratificar foi realizado um teste de hipótese, confirmando o melhor desempenho do modelo PM2. Assim, foi comprovado estatisticamente com nível de confiança de 95% em ambos os cenários (LP e MT) que o modelo apresenta menores erros de predição em relação a todos os modelos utilizados para o cenário das variáveis selecionadas com *Stepwise/Correlação*. Utilizando técnica de Correlação, o PM2 obteve erros menores que o PM1, PM3 e o LM4 já que os valores obtidos de *p-value* foram 0.0224, 0.009564 e 0.01819, respectivamente.

Os resultados apresentados no diagrama de causa e efeito, aplicado ao fracasso escolar, mostram que as causas relacionadas à instituição/escola têm uma contribuição mais significativa em relação às demais pelo quantitativo de fatores associados. Dentre as causas relacionadas destacam-se a infraestrutura da escola (iluminação das salas, ambientes para o estudo), o ensino/aprendizagem (existência

e uso de mídias educativas e de laser no cotidiano escolar) e a formação inicial do professor (poucos professores com formação específica para as disciplinas que lecionam). As demais causas também têm sua contribuição significativa ao problema mencionado. Com relação ao aluno, o fato do mesmo não comparecer às avaliações escolares, problemas relacionados à família, à formação educacional e cultural dos país aparecem como fatores preocupantes, além do conhecimento dos pais sobre o andamento escolar dos filhos. Por fim, um fator associado que está em evidência na educação brasileira é a violência. Os dados deste estudo comprovam que fatores associados à violência contribuem para o fracasso escolar dos estudantes. Pode-se destacar neste estudo a existência de segurança, seja eletrônica ou policial, nas imediações da escola durante o seu período de funcionamento. Além da proteção dos bens (equipamentos) da escola de roubo e depredação e a utilização pela escola de sistemas de segurança para repelir ações de roubo e furto dentro ou fora de suas dependências.

Considerando o terceiro caso de estudo que trata da evasão no ensino superior, os resultados mostraram-se satisfatórios. Os experimentos realizados neste caso mostraram que o Modelo PM1, o qual combina Ensemble Bagging com Regressão Linear (LR), obteve melhores resultados em relação ao modelo da literatura LM4 que usa *Ensemble Stacking*, como também em relação aos modelos PM2 e PM3. As métricas com o método Stepwise alcançaram 96% na previsão e com o método de Correlação Pearson, 95%. Além disso, propõe-se a utilização de outros modelos combinados utilizando *Bagging* que também obtiveram bons resultados para o problema da evasão.

Assim, comprova-se que os modelos propostos baseados em Esemble Bagging obtiveram resultados superiores ao modelo da literatura (Ensemble Stacking) para o problema da evasão escolar em instituicões de ensino superior no Brasil.

Os resultados apresentados no diagnóstico de causa e efeito destacam as causas relacionadas à instituição de ensino como maiores reponsáveis pela ocorrência da evasão, dentre as quais estão os fatores associados à permanência e à conclusão do curso. Esses fatores, de acordo com Spady (1970) e Tinto (1993) estão relacionados ao desempenho acdêmico do estudante ao longo de seu percurso dentro da instituição de ensino e a sua adaptação à instituição, fatores estes que podem levá-lo à decisão de abandonar o curso ou a instituição. Outros fatores associados ao

estudante e a aspectos financeiros também se apresentaram significantes, destacando-se os auxílios que o estudante recebe da instituição, ou externos a ela, como por exemplo bolsas de pesquisa, estágios, programas governamentais e os aspectos relacionados ao próprio estudante como sociodemográficos (étnico raciais), atividades externas (trabalho) e comprometimento com a instituição (horário de estudo), este último obtendo maior destaque em relação aos demais fatores associados ao aluno.

7 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as considerações finais deste trabalho, contribuições, limitações e trabalhos futuros. Também estão listadas as principais produções científicas realizadas durante o desenvolvimento desta tese.

7.1 CONCLUSÃO

O primeiro passo para o desenvolvimento desta pesquisa foi a identificação de lacunas com foco na construção de modelos preditivos voltados aos problemas educacionais. Inicialmente, foi desenvolvido um mapeamento sistemático da literatura, no qual buscou-se identificar, catologar e analisar artigos (Capitulo 4) e seus respectivos métodos/modelos/abordagens/técnicas para mensurar os problemas educacionais. Com isso, observou-se que a literatura sobre a predição de problemas nos diversos contextos educacionais vem aumentando ao longo dos anos, embora a maior concetração desses trabalhos seja nos útimos cinco anos. Tal fato ratifica que o desenvolvimento desta pesquisa possui relevância científica/ acadêmica na atualidade. A maioria dos estudos identificados apresenta os seguintes contextos: predição do desempenho escolar/acadêmico, predição da evasão/retenção, e a predição do comportamento dos alunos. Por outro lado, observou-se o grande número de trabalhos relacionados à predição dos problemas educacionais no ensino superior. Contudo, os trabalhos encontrados não utilizam modelos teóricos e nem identificam as casuas e os efeitos relacionados aos problemas educacionais.

Observa-se que os problemas educacionais mais mencionados nos trabalhos identificados foram o desempenho e a evasão.

Nenhum estudo foi encontrado na literatura pesquisada com foco no estabelecimento das relações de causa e efeito entre os fatores identificados (causas) e os problemas estudados (efeitos).

A utilização de modelos preditivos para o diagnóstico de problemas educacionais, baseados nas relações de causa e efeito, foi uma das principais lacunas de pesquisa abordadas nesta tese. O tema se mostrou desafiador especialmente considerando o grande volume de dados gerados nas bases públicas do INEP, e da aplicação dos modelos teóricos de Andrade e Soares (2008), Spady (1970) e Tinto (1975;1993) para subsidiar o diagnóstico dos problemas educacionais, a partir dos resultados gerados pelos modelos preditivos propostos.

Diante destes achados, esta tese apresenta o diagnóstico dos problemas educacionais, a partir do desenvolvimento de modelos preditivos baseados em *Ensemble Regression*. Esses modelos fundamentam-se nos modelos teóricos de Andrade e Soares (2008), Spady (1971) e Tinto (1975; 1993). Além disso, foi utilizada a identificação das relações de causa e efeito proposta por Ishikawa (1960;1993) para as dimensões aluno, família, escola/instituição e sociedade. Para isso, utilizou-se dados educacionais fornecidos pelo INEP.

Neste sentido, os dados educacionais podem ser analisados por meio de variáveis associadas a dois tipos: quantidade e qualidade. O enfoque da quantidade está pautado nos trabalhos cujo objetivo é analisar a quantidade de anos de estudos acumulados pelos indivíduos e variáveis relacionadas, como matrícula, a repetência, a evasão e o desempenho escolar (aprovação e reprovação). Já o enfoque da qualidade está normalmente relacionado ao estudo dos efeitos familiares e escolares sobre o rendimento dos alunos obtido em avaliações padronizadas (MACEDO, 2004).

Assim, para analisar de forma efetiva os dados educacionais dentro dos preceitos quantitativos e qualitativos, o primeiro passo foi o uso das técnicas de mineração de dados educacionais, a qual se mostrou uma abordagem satisfatória, contribuindo para o entendimento e mensuração dos problemas em estudo. Além disso, com a utilização de modelos preditivos foi possível identificar as principais causas dos problemas, como por exemplo, alunos em situações de risco de desempenho ou de retenção. A possibilidade de prever os problemas antes da sua ocorrência faz da modelagem preditiva uma abordagem eficaz para análise de dados educacionais.

Para a construção dos modelos preditivos deste trabalho foram utilizadas variáveis referentes aos apectos sociodemográficos, socioeconômicos, de infraestrutura escolar e indicadores educacionais. A identificação das variáveis foi fundamentada no modelo teórico dos fatores associados ao desempenho dos alunos proposto por Andrade e Soares (2008), no modelo sociológico explicativo do abandono escolar de Spady (1971) e no modelo de integração do estudante de Tinto (1993). Além disso, para seleção das variáveis foram utilizados os métodos de correlação de Pearson, Stepwise e Postos de Spearman. Essas variáveis foram utilizadas para explicar o desempenho e a evasão escolar.

Posteriormente, foram utilizados quatro algoritmos de regressão (Regressão Linear, Regressão Robusta, Regressão Quantilica e Support Vetor Regression –SVR) e o método *Bagging* de *Ensemble Regression.* Buscou-se realizar um comparativo entre os modelos de regressão que melhor se ajustassem ao problema em estudo. O comparativo foi realizado utilizando as principais métricas encontradas na literatura como a MAE (Mean Absolute Error) e GR (Ganho Relativo). Para ratificar o desempenho dos modelos propostos, foi utilizado teste de hipótese.

De acordo com as métricas utilizadas nos casos estudados, no caso 1- diagnóstico do desempenho - o modelo SVR obteve melhores desempenhos nos cenários apresentados em relação aos modelos RL, RR e RQ, obtendo métricas de predição do desempenho de 79% e 88%, respectivamente. Nos cenários 3 e 4, o modelo que utiliza RQ obteve métricas de predição de 47% e 56%, respectivamente. No caso 2 – Diagnóstico do fracasso escolar – o modelo *Ensemble Bagging* Regressão Robusta (PM2) apresenta menores erros de predição em relação aos demais modelos PM1, PM3 e o modelo da literatura LM4, obtendo uma métrica de predição de 95% em ambos os cenários (LP e MT) nos quais foi aplicado. Por fim, o caso 3 – Diagnóstico da Evasão – o modelo que combina *Ensemble Bagging* (PM1) obteve melhores resultados em relação ao modelo da literatura LM4 que utiliza *Ensemble Stacking,* como também em relação ao PM2 e ao PM3. As métricas com o método Stepwise alcançaram 96% na previsão e com a Correlação de Pearson, 95%. Estes resultados vêm a contribuir com os trabalhos desenvolvidos por Carmem (2018), Aderibigibe (2018), Akin (2014) e Halil (2018), os quais apresentaram modelagem preditiva por meio de regressão como sendo uma técnica adequada para a mineração de dados educacionais.

Como forma de apresentar o diagnóstico dos problemas educacionais baseados nos modelos preditivos propostos, foi utilizado o diagrama de causa e efeito proposto por Ishikawa (1993), no contexto dos problemas educacionais evasão e desempenho, com o objetivo de identificar as causas de forma sistemática, proporcionando intervenções pontuais aos problemas analisados.

Destacam-se nesta solução a simplicidade de construção, aplicação e visualização dos resultados dos modelos preditivos, possibilitando aos agentes educacionais (gestores e professores) o conhecimento das principais causas, para um diagnóstico mais preciso do problema em questão, facilitando o acompanhamento e

a tomada de decisão por parte dos agentes educacionais envolvidos.

7.2 CONTRIBUIÇÕES

Durante o desenvolvimento desta tese, diversas contribuições surgiram conforme a execução da abordagem proposta, destacando as seguintes:

- O desenvolvimento de uma abordagem baseada no CRISP-DM, aplicada ao diagnóstico de problemas educacionais, denominada DEP-DM. A abordagem proposta utiliza os modelos teóricos Andrade e Soares (2008), Spady (1970) e Tinto (1975;1993) e o método causa e efeito de Ishikawa (1993);
- A construção de modelos utilizando técnicas de regressão paramétricas e não paramétricas baseada na abordagem DEP-DM para prever o desempenho da proficiência de LP e MT dos estudantes do 5º ano do Ensino Fundamental das Escolas Púbicas de Pernambuco;
- O Desenvolvimento de modelo *Ensemble Regression* baseado no método *Bagging* para o diagnóstico de problemas educacionais;
- A aplicação da Abordagem DEP-DM utilizando os modelos de *Ensemble Regression* propostos aos problemas de evasão e desempenho escolar (proficiência de LP e MT); e
- Aplicação do método de causa e efeito de Ishikawa (1993), no contexto da abordagem DEP-DM para o diagnóstico das principais causas dos problemas de desempenho (proficiência de LP e MT) e evasão escolar.

7.3 LIMITAÇÕES

Uma das principais dificuldades, durante o desenvolvimento desta tese, foi a extração e pré-processamento das bases de dados utilizadas na fase de modelagem. A grande quantidade de dados e a complexidade da arquitetura dos bancos de dados do INEP tornaram a extração dos dados educacionais numa das fases que exigiram maior dedicação de tempo e esforço durante a aplicação da metodologia DEP-DM.

Além da fase de extração dos dados educacionais das bases do INEP, outra etapa que demandou bastante esforço foi a identificação das variáveis representativas dos fenômenos em estudo nesta tese. Neste sentido o processo empregado nesta tese permitiu selecionar um conjunto satisfatório de variáveis relacionadas ao

comportamento dos estudantes e aos fatores relacionados à família, a aspectos socioeconômicos e aspectos relacionados à instituição educacional.

A abordagem DEP-DM não foi testada e avaliada por professores e gestores educacionais, para que os mesmos pudessem apontar pontos de melhoria. Outra limitação desta tese foi a inexistência de experimento temporal com alunos, tanto da educação fundamental quanto do ensino superior, pertencentes a escolas e universidades do Estado de Pernambuco. Desta forma, tem-se consciência da necessidade, em trabalhos futuros, do levantamento de dados reais *in loco* nas escolas e universidades para que se possa explorar mais os problemas educacionais elencados neste estudo, e para permitir a real percepção da problemática educacional que não consta muitas vezes nos dados oriundos dos questionários do INEP. Pode-se considerar esta como sendo a principal limitação desta tese.

7.4 TRABALHOS FUTUROS

Este trabalho possibilitou o desenvolvimento de uma abordagem para o diagnóstico de problemas educacionais. Como possibilidades de trabalhos futuros, os quais podem ser explorados a partir deste estudo, tem-se:

- Aplicação da abordagem DEP-DM utilizando o modelo *Ensemble Regression*, em outros contextos educacionais, como em ambientes de ensino a distância, para identificar as causas de problemas;
- Realizar a implementação, da análise de resíduos, Correlação de Postos, e teste de Friedman, em dados educacionais;
- Desenvolvimento de um software utilizando a abordagem DEP-DM para o processo de diagnóstico de problemas educacionais;
- Utilização da abordagem DEP-DM em dados de sistemas corporativos (SIGA e ATRIO); e
- Construção de modelo *Ensemble Regression*, como: *boosting* e *stacking* utilizando a abordagem DEP-DM.

7.5 PRODUÇÃO CIENTÍFICA DESENVOLVIDA

Durante o doutorado, foram desenvolvidos artigos científicos, publicados em conferências, com temáticas alinhadas a esta proposta, no sentido de promover um aperfeiçoamento do pesquisador, além da apropriação de conteúdos e ferramentas

necessários ao bom andamento desta tese.

7.5.1 Artigos Aceitos em Conferências

- SILVA, Paulo Mello; LIMA, Marilia N.C.A; SOARES, Wedson L.; SILVA, Iago R.R; FAGUNDES, Roberta A. de A.; SOUZA, Fernando F. **Ensemble Regression Models Applied to Dropout in Higher Education**. Brasilian Conference on Intelligent Systems (BRACIS), 2019, Salvador, Brasil.
- SILVA, Paulo Mello; NASCIMENTO, Rafaella L.S.; LIMA, Marilia N.C.A; FAGUNDES, Roberta A. de A.; SOUZA, Fernando F. **Modelos de Regressão Aplicados a Predição do Desempenho Escolar de Estudantes do Ensino Fundamental.** Simpósio Brasileiro de Informática na Educação (SBIE), 2019, Brasilia.

7.5.2 Artigos em fase de submissão para periódicos

- SILVA, Paulo Mello; Roberta A. de A.; SOUZA, Fernando F. **Mapeamento Sistemático sobre Abordagens para Predição do Desempenho Educacional.**
- SILVA, Paulo Mello; Roberta A. de A.; SOUZA, Fernando F. **Ensemble Regression Models Applied to Dignostic Performance in Higher Education.**

REFERÊNCIAS

A.I., M. T. A Political Economy Perspective. *NEXT GPT*, 15 February 2018. Disponivel em: <http://ssrn.com/abstact=3126215>.

AVNIMELECH, R.; INTRATOR, N. Boosting regression estimators. *Neural computation,* v. 11, p. 499–520, 03 1999.

BAKER, R. S. J. Data mining for education. *Internacional Encyclopedia of Education. ELSEIVER*, Amsterdam, 2010. 112-118.

BAKER, S. J. D.; CARVALHO, M. J. B. D.; ISOTANI. Mineração de Dados Educacionais Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação,* 2011.

BISSOLI, S. C. A. Evasão Escolar: O caso do Colégio Estadual Antonio Francisco Lisboa. *Repositório SEAP*, 2010.

BREIMAN, L. Bagging predictors. *Machine Learning*, Kluwer Academic Publishers, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Heuristics of instability and stabilization in model selection. *Annals of Statistics*. The Institute of Mathematical Statistics, v. 24, n. 6, p. 2350–2383, 12 1996.

BUHLMANN, P. *Bagging, Boosting and Ensemble Methods.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. 985–1022 p.

C., R.; C, V. *Data Mining in Education. WIRES Data Mining Knowl Discov*, February 2013. 12-27.

CAMERON, A. C.; TRIVEDI, P. K. Regression analysis of count data. *Cambridge University Press,* 1998.

CAMPOS, H. Estatistica Experimental não-paramétrica. Esalq, Piracicaba-SP, 1987. 230.

CHAPMAN, P. E. A. CRISP-DM 1.0 Step - by - *Step data mining guide*. [S.l.]: [s.n.], 2000.

CURRAL, J. Statistics Packages: A General Overview. Universidade de Glasgow, Glasgow, 1994.

DA SILVA, L. A.; PERES, S. M.; BOSCARIOLLI, C. Introdução a Mineração de Dados com Aplicações em R. 1ª. ed. Rio de Janeiro: ELSEVIER, 2016.

EINSTEIN, A. Ideas and Opnions. Broadway Books, 1955.

F.M., A. et al. Analysis of student performance using edm methods. 5*th International Multi -Topic ICT Conference* (IMTIC). [S.l.]: IEEE. 2018. p. 1-7.

FAVEIRO, L. E. A. Análise de dados: modelagem multivariada para a tomada de decisão. [S.l.]: Campus, 2009.

GUYSON, I.; ELISSEEFF, A. An Introduction to Variable and Features Selection. *Journal of Machine Learning Research*, 2003.

H, D. et al. Support Vector Regression Machines. I*n Advances in Neural Information Processing Systems,* 1997. 155-161.

HAIR, J. F. E. A. Análise Multivariada de Dados. [S.l.]: Bookman Editora, 2009.

HOFFMAN, R. Análise de Regressão: Uma introdução à econometria, 2015.

INEP. Site do INEP. Disponivel em: <http://www.inep.gov.br>.

J., C.; M., P.; T., K. Second Chance for high school dropouts? a regression discontinuity analysis of postsecondary educational returns to the ged. J*ournal of Labor Economics,* 2017. 273-304.

JACOB, J. et al. Educational Data Mining Techiniques and Their Application. *International Conference on Green Computing and Internet Things* (ICGCIoT). Noida: [s.n.]. 2015. p. 1344-1348.

K., M. Nonlinear Multiobjective Optimization. *Springer Science e Business Media*, 2012.

KITCHENHAM, B. Guidelines for performing Systematic Literature Reviews, 2007.

KOEDINGER, K. R. E. A. A data repository for the EDM community: The PSLC Data Shop. In: *Handbook of educational data mining*. [S.l.]: [s.n.], 2010. p. 43.

KURDI, M. M.; KHAFAGI, H. -; ELZEIN, I. Mining Educational Data to Analyse Students Behavior and Performance. JCCO - Joint International Conference; ICT - Educational and Training ; International Confrence on Computing in Arabic. Hamamet - Tunisia: [s.n.]. 2018. p. 1-5.

LEHMAN, A. et al. JMP for Basic Univariate and Multivariate Statistics: *Methods for Researches and Social Scientists*. [S.l.]: SAS Institute, 2013.

LEMAN, F. Todos Pela Educação. Disponivel em: <http://www.todospelaeducacao.org.br>. Acesso em: 5 Maio 2018.

M.V., C. et al. Dropout Prediction using data mining: a case study with high school students. Expert Systems, 2016. 107-124.

MD HASSIBUR, R.; MD RABIUL, I. Predict Students academic performance and evaluate the impact of diferent attributes on the performace using data mining techiniques. 2nd International Conference on Eletrical e Eletronic Engineering (ICEEE). [S.l.]: IEEE. 2017. p. 1-4.

MILIKOVIC, S.; VUJOVIC, V. Students Sucess Predictive Models Based on Select Imput Parameters Set. 18TH International Symposium Infoteh - Jahorina (INFOTEH). East Sarajevo, Bosnia Herzegovina: [s.n.]. 2019. p. 1-6.

MONTGOMERY, D. C. Introduction to Statistical Quality Control. [S.l.]: John Wiley e Sons, 2007.

MORAES, R. E. Dia dia Educação, 20 Maio 2010. Disponivel em: <http//www.diadiaeducacao.pr.gov.br/portals/pde/arquivos/748-4.pdf>.

PASSOS, E.; GOLDSCHMIDT, R. Data Mining: Um Guia Prático. Rio de Janeiro: Campus, 2005.

PEINADO, J.; GRAEMIL, A. R. Administração da Produção: Operações Industriais e de Serviços. Curitiba: Unicep, 2007.

PEÑA-AYALA, A. Educational Data Mining: A survey and data mining based analysis of recent works. *Expert Systems with appliations*, 2014. 1432-1462.

PETERSEN, K. E. A. Systematic Mapping Studies in Software Engineering, 2007.

PIMENTEL, P. E.; OMAR, N. Descobrindo Conhecimentos em Dados de Avaliação da Aprendizagem com Técnicas de Mineração de Dados, 2006.

QEDU. Disponivel em: <http://www.qedu.org.br>.

QUEIROZ, D. L. Anped, 2010. Disponivel em: <http://www.anped.org.br/reunoes/25/luciledomingosqueiroz13.rtf>.

RIGO, S. J. E. A. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. Revista Brasileira de Informática na Educação, 2014.

RODRIGO, R. E. A. Mapeamento Sistemático sobre abordagens de Mensuração de Autorregulação da Aprendizagem. RENOTE- Revistas Novas Tecnologias na Educação, 2016.

RODRIGUES, R. A Literatura Brasileira sobre Mineração de Dados Educacionais. Workshop de Mineração de Dados Educacionais. Dourados: [s.n.]. 2014.

ROMERO, C. Handbook of Educational Data Mining. [S.l.]: CRC Press, 2010.

ROMERO, C.; VENTURA, C. Educational Data Mining: A Review of the state of the art. Trans. Syst. man Cybern. Part C Appl. Rev, 2010. 601-618.

SHETH, J.; PATEL, B. Best pratices for adaptation of Data Mining techiniques in Education Sector. *National Journal of System and Information Technology*, 2010. 186.

TAN, P. N.; STEINBACH, M.; V., K. Introdução ao Data Mining: Mineração de Dados. Rio de Janeiro: Ciência Moderna, 2009.

VARGAS, F. G. Fundação Getulio Vargas. FGV, 2010. Disponivel em: <http://www.fgv.br/cps/tpemotivos>.

WAZLAWICK, S. Metodologia de Pesquisa para Ciência da Computação. Rio de Janeiro: Elseiver, 2008.

WEISS, S. M.; INDURKHYA, N. Rule - based Machine Learning Methods. *Journal of Artificial Inteligence Research*, 1995. 383-403.

WITTEN, I.; E FRANK, M. A. H. Mineração de Dados: Ferramentas e Técnicas Práticas de Aprendizado de Máquina. Elseiver, 2011.