UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Larissa Feliciana da Silva Britto

Identificação Automática de Restrições Alimentares em Receitas Culinárias Através de Técnicas de Aprendizagem de Máquina

Recife

2023

Larissa Feliciana da Silva Britto

Identificação Automática de Restrições Alimentares em Receitas Culinárias Através de Técnicas de Aprendizagem de Máquina

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientadora: Profa. Dra. Teresa Bernarda Ludermir

Coorientador: Prof. Dr. Luciano Demétrio Santos Pacífico

Recife

2023

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

B862i Britto, Larissa Feliciana da Silva
Identificação automática de restrições alimentares em receitas culinárias através de técnicas de aprendizagem de máquina / Larissa Feliciana da Silva Britto. – 2023.
78 f.: il., fig., tab.

Orientadora: Teresa Bernarda Ludermir.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2023.
Inclui referências.

1. Inteligência artificial. 2. Aprendizagem de máquina. I. Ludermir, Teresa Bernarda (orientadora). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2023-84

Larissa Feliciana da Silva Britto

“Identificação Automática de Restrições Alimentares em Receitas Culinárias Através de Técnicas de Aprendizagem de Máquina”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 03/03/2023

BANCA EXAMINADORA

_____

Profa. Dra. Teresa Bernarda Ludermir
Centro de Informática / UFPE
(**Orientadora**)

_____

Prof. Dr. Douglas Veras e Silva
Departamento de Computação / UFRPE

_____

Prof. Dr. João Fausto Lorenzato de Oliveira
Escola Politécnica de Pernambuco / UPE

Dedico este trabalho à minha mãe, Mércia.

AGRADECIMENTOS

Agradeço a minha mãe Mércia, a minha avó Maria e minha tia Ana, por toda dedicação e força, e por serem as grandes responsáveis por todas as minhas conquistas.

Agradeço ao meu irmão e melhor amigo Gustavo, que está sempre ao meu lado, me apoiando incondicionalmente.

Agradeço à minha querida tia Ana pelo seu constante apoio e suporte durante toda minha vida.

Agradeço ao professor Luciano, por me aconselhar, instruir e estimular durante todos esses anos, com toda sua dedicação e paciência.

Agradeço também à professora Teresa, por toda sua confiança, incentivo e apoio.

RESUMO

Com o avanço e popularização da internet e de tecnologias, como os smartphones, a culinária sofreu uma revolução, na qual as receitas se tornaram um dos tópicos mais acessados e comentados da Web. Apesar da facilidade de acesso trazida pela internet, encontrar a receita ideal ainda é uma tarefa árdua, especialmente para pessoas que possuem algum tipo de restrição alimentar. A identificação correta dos alimentos que infrinjam determinada restrição é fundamental para saúde e bem estar do indivíduo que segue uma dieta restritiva. Este trabalho tem como principal objetivo realizar a identificação automática de alimentos que infrinjam restrições alimentares, através da classificação de receitas dietéticas, utilizando técnicas de Aprendizagem de Máquina. Essa classificação pode ser aplicada para facilitar a busca de usuários de sites de receitas que possuem algum tipo de restrição alimentar. Alguns dos principais modelos da literatura de classificação de receitas culinárias são escolhidos para avaliação, no intuito de apurar qual deles seria o mais adequado para execução da tarefa de identificação de dietas e restrições em receitas. Os seguintes classificadores são adotados: Árvore de Decisão (AD), *Bidirectional Encoder Representations from Transformers* (BERT), Floresta Aleatória (FA), K-Vizinhos Mais Próximos (K-NN), Naive Bayes (NB), Perceptron Multicamadas (MLP), Regressão Logística (RL) e Máquinas de Vetores de Suporte (SVM). As listas de ingredientes, modos de preparo, títulos e descrições das receitas são avaliados, individualmente e combinados, com o propósito da seleção dos conjuntos de dados que mais contribuem para o processo de aprendizado dos classificadores. Esses dados se encontram em formato textual, sendo necessário o emprego de técnicas de Processamento de Linguagem Natural para a extração de características dos documentos da base. Experimentos são realizados, nos quais os classificadores selecionados são executados, e seus desempenhos são mensurados e comparados uns com os outros. Os resultados experimentais são avaliados empiricamente, e através de métodos estatísticos (teste de hipóteses de Friedman/Nemenyi). A avaliação aponta o bom desempenho dos modelos adotados na tarefa de classificação de receitas dietéticas, com destaque para os classificadores Regressão Logística, BERT e MLP. Os resultados obtidos indicam ainda que as características mais adequadas para a classificação podem variar de restrição para restrição.

Palavras-chaves: aprendizagem de máquina; BERT; classificação de receitas; classificação de texto; processamento de linguagem natural.

ABSTRACT

With the advancement and popularization of the internet and technologies, such as smartphones, culinary went through a revolution, in which recipes have become one of the most accessed and commented topics on the Web. Despite the ease of access brought by the internet, finding the ideal recipe may be challenging, especially for people who have some kind of dietary restriction. The correct identification of foods that violate a certain restriction is essential for the health and well-being of the individual who follows a restrictive diet. The aim of this work is to execute the automatic identification of foods that violate dietary restrictions, through the classification of dietary recipes, using Machine Learning techniques. This classification can be applied to assist users of recipes websites that have some type of dietary restriction, whether for health reasons, cultural issues or simply personal preferences. For this, some diets associated with dietary restrictions are selected from a dataset of cooking recipes in the English language, obtained through the website Food.com. Some of the main models in cooking recipe classification literature are chosen for evaluation, in order to determine which ones would be the most suitable for performing the target task. The following classifiers are adopted: Decision Tree (DT), Bidirectional Encoder Representations from Transformers (BERT), Random Forest (RF), K-Nearest Neighbors (K-NN), Naive Bayes (NB), Multilayer Perceptron (MLP), Logistic Regression (LR) and Support Vector Machines (SVM). The list of ingredients, preparation steps, titles and descriptions of the recipes are evaluated, individually and combined, with the purpose of selecting the data sets that most contribute to the learning process of the classifiers. These data are in textual format, requiring the use of Natural Language Processing techniques to extract features from the base documents. Experiments are carried out, in which the selected classifiers are executed, and their performances are measured and compared with each other. The experimental results are evaluated empirically, and through statistical methods (Friedman/Nemenyi hypothesis test). The evaluation pointed out the good performances of the adopted models in the dietary recipes classification task, emphasizing Logistic Regression, BERT and MLP classifiers. The obtained results also indicate that the most suitable feature sets for classification may vary from one restriction to another.

Keywords: BERT; machine learning; natural language processing; recipe classification; text classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvore de Decisão
BERT	*Bidirectional Encoder Representations from Transformers*
CD	Diferença Crítica
CNN	Convolutional Neural Network
CT	Classificação de Texto
FA	Floresta Aleatória
GCN	Graph Neural Networks
GRU	Gated Recurrent Unit
K-NN	K-Vizinhos Mais Próximos
LSTM	*Long Short Term Memory*
MLP	Perceptron Multicamadas
NB	Naive Bayes
RL	Regressão Logística
SVM	Máquinas de Vetores de Suporte
TF-IDF	*Term Frequency-Inverse Document Frequency*

LISTA DE SÍMBOLOS

X	Conjunto de Documentos
Y	Classes associadas ao conjunto de documentos X
t	Um termo do vocabulário do conjunto de documentos X
d	Um documento pertencente ao conjunto de documentos X
N	Quantidade de documentos no conjunto de documentos X
$f_{t,d}$	Frequência do termo t no documento alvo d.
df_t	Número de documentos em que um termo t está presente
TP	Verdadeiros Positivos
TN	Verdadeiros Falsos
FP	Falsos Positivos
FN	Falsos Negativos
k	Número de Algoritmos
q_α	Valores Críticos
D	Quantidade de Conjuntos de Dados
s	Segundos

SUMÁRIO

1 INTRODUÇÃO

A alimentação exerce uma função essencial para a saúde humana, sendo a principal fonte de nutrientes fundamentais para o corpo, fornecendo energia para as atividades básicas, crescimento, e todas as funções corporais, como respiração, digestão e termorregulação, mantendo também o sistema imunológico saudável (NATIONS, 2018).

A alimentação costuma ser resumida a uma necessidade biológica indispensável, devido a sua importância nutricional. No entanto, a alimentação representa para os seres humanos, algo muito mais significativo e simbólico que apenas seus benefícios fisiológicos. A comida também é um forte fator cultural, que contribui para a identificação pessoal e social (ZUIN; ZUIN, 2009), além de ser uma ferramenta de aproximação, de conhecimento e de inclusão sociocultural. O ato de comer inclui seleção, escolha, rituais, ideias e significados (FRANZONI, 2016).

O processo de escolha alimentar é dinâmico e complexo, definido por diversos fatores que são construídos desde o nascimento e podem ser modificados durante toda a vida (JOMORI; PROENÇA; CALVO, 2008).

Os fatores que influenciam o processo de escolha do alimento são tradicionalmente divididos em três categorias principais: os fatores relacionados à comida, as características individuais e fatores ambientais. Esses fatores são descritos na Tabela 1.

Alguns dos principais fatores de escolha de alimentos estão relacionados às restrições alimentares. O termo "restrição alimentar" pode ser utilizado para designar uma condição apresentada por alguém que não pode consumir determinado alimento, devido à alergia, intolerância, ou problemas de saúde, como doença celíaca, diabetes, hipertensão, entre outros. Essa condição também é apresentada quando o indivíduo deixa de consumir um alimento ou grupo de alimentos por questões religiosas ou ideológicas, como o vegetarianismo e o veganismo (CATARINA, 2022). As restrições alimentares podem englobar diversos dos fatores citados na Tabela 1. As alergias e doenças que requerem dietas especiais, como a diabetes e hipertensão, podem ser incluídas nos **fatores biológicos e fisiológicos**. Restrições relacionadas a crenças, como a dieta *kosher* (dieta judaica), são englobadas nos **fatores socioculturais**, mais especificamente em **normas e valores culturais**, enquanto restrições relacionadas a convicções e ideais, como a dieta vegana, se encaixam nos **fatores externos ao alimento**, como o **histórico de ética da produção**. Um indivíduo também pode limitar sua alimentação por não gostar de um determinado alimento, um fator diretamente relacionado às **características**

Tabela 1 – Fatores que influenciam a escolha de alimentos.

Categoria	Fator	Exemplos
Relacionados à Comida	Fatores Internos	Características possuídas pelo próprio alimento, como propriedades sensoriais (**sabor**, **cheiro** e **textura**) e perceptivas (**cor**, **tamanho da porção**, **valor nutricional** e de **saúde** e **qualidade**).
	Fatores Externos	Informações sobre o item alimentar (**rótulos nutricionais**, **alegações de saúde**, **embalagem**, **estética**, **histórico de ética da produção**, **marca** e **propaganda**).
Individuais	Estado Pessoal	Características biológicas (**fatores genéticos**, **padrões pessoais de dieta** e **metabolismo**, **condição física/saúde**), necessidades fisiológicas (**fome**, **apetite** e **peso**), componentes psicológicos (**emoção**, **motivação**, **personalidade**), **hábitos** e **experiências pessoais**
	Cognitivos	Os fatores cognitivos estão associados a características como **conhecimento** e **habilidades do indivíduo**, **preferências** e **consequências** (benefícios e riscos do consumo de tal alimento). Características demográficas do consumidor também são incluídas nesse fator (**idade**, **gênero** e **etnia**).
Relacionados à Sociedade	Socioculturais	Os fatores do ambiente que influenciam a escolha alimentar individual incluem a **renda**, **status socioeconômico** e **preço dos alimentos**, **normas e valores culturais**, e **políticas e regulamentações agrícolas e alimentares**

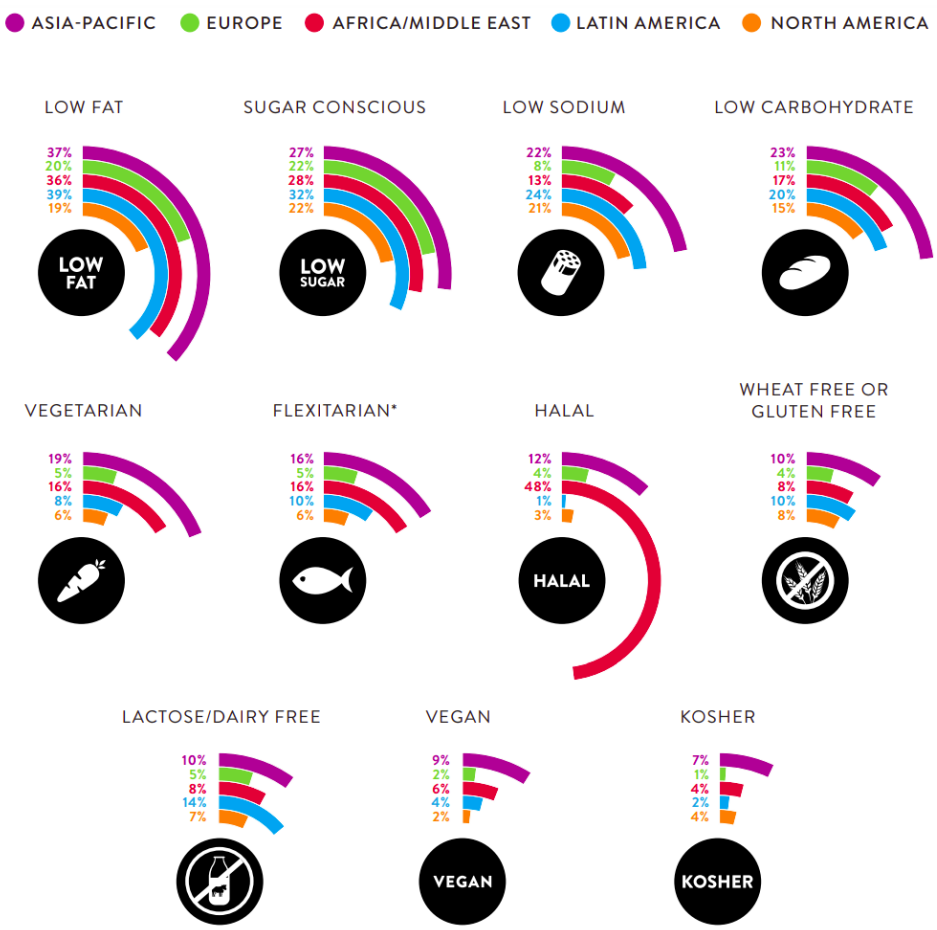
Fonte: (CHEN; ANTONELLI, 2020).

internas do alimento. Uma pesquisa realizada em 2016 (NIELSEN, 2023), com mais de 30 mil participantes, identificou que quase dois terços dos entrevistados globais seguem uma dieta que limita ou proíbe o consumo de alguns alimentos ou ingredientes, e essa taxa pode ser ainda maior em certas regiões, como pode ser visto na Figura 1. A identificação dos alimentos que infrinjam determinada restrição é fundamental para a saúde e o bem estar do indivíduo que segue uma dieta específica.

Além dos fatores citados anteriormente, o conhecimento necessário para o preparo de uma dada refeição também pode ser considerado um importante fator limitante na escolha do prato a ser consumido por um indivíduo em um dado momento. Nesse contexto, as receitas culinárias surgem como um importante fator de difusão dos conhecimentos acerca do preparo correto dos alimentos, assim como ferramentas de extrema utilidade na expansão da capacidade técnica de um indivíduo no que diz respeito à manipulação de alimentos e ingredientes.

Receitas culinárias são textos instrucionais para o preparo de um determinado prato, apresentando todos os detalhes necessários para que um indivíduo que nunca o preparou seja capaz de reproduzi-lo. Indo muito além de um registro instrucional, as receitas são alvo de diversos estudos de múltiplas áreas, por serem fonte de conhecimento sobre a história e a cultura de uma sociedade (SANTOS, 2005; LIMA, 2015).

Figura 1 – Porcentagem de pessoas, em uma pesquisa com mais de 30 mil participantes, que seguem uma dieta especial que limita ou restringe alimentos ou ingredientes específicos.



Fonte: (NIELSEN, 2023).

No passado, as instruções para preparo de um prato eram transmitidas de geração em geração através da oralidade, onde mulheres propagavam seu conhecimento às suas filhas e netas, que só com o tempo e convivência seriam capazes de reproduzir tais preparos. Essa cultura muitas vezes era preservada através da escrita em cadernos, que também atravessavam gerações (CONSIDERA, 2008; LIMA, 2015).

Com o passar do tempo, as receitas culinárias passaram a integrar as páginas de revistas, jornais e livros. No Brasil, os registros culinários em formato de livros começaram a ser publicados no fim do século XIX, se popularizando durante todo século XX e atingindo seu ápice na década de 1990 (CORDEIRO et al., 2020).

Nos últimos anos, o compartilhamento de receitas sofreu uma revolução. A gastronomia se tornou um dos tópicos mais acessados em redes sociais, blogs e sites de compartilhamentos de

vídeos. Canais de gastronomia alcançam milhões de seguidores e bilhões de visualizações [1,2,3], enquanto os repositórios de receitas culinárias estão entre os mais acessados no mundo [4]. A distribuição de receitas on-line apresenta diversas vantagens em relação à publicação por meio de livros e revistas de culinárias. Entre essas vantagens está o grande volume e variedade de receitas, para os mais diversos públicos, disponíveis a qualquer momento em qualquer lugar.

Apesar dos benefícios trazidos por esses sites, encontrar a receita ideal, que atenda restrições, necessidades e preferências dos usuários, ainda é um processo complexo. Esses repositórios geralmente são compostos por milhares de receitas, o que torna inviável a análise manual desses documentos. Para auxiliar os usuários a encontrarem os alimentos mais adequados, Sistemas de Recomendação têm sido aplicados ao problema de busca por receitas (YANG et al., 2016; SCHäFER et al., 2017; MAIA; FERREIRA, 2018). Esses sistemas auxiliam na filtragem dos dados retornados pela busca, exibindo apenas um conjunto reduzido de receitas, que sejam relevantes para o usuário, no contexto especificado pelo mesmo. Como a escolha do alimento a ser consumido leva em consideração diversos fatores, o ideal é que a recomendação gerada seja o mais personalizada possível. Técnicas de Inteligência Artificial têm sido empregadas para identificar e inferir as características das receitas úteis para recomendação.

A inferência de informações sobre receitas culinárias tem sido tradicionalmente modelada como um problema de Classificação de Texto (CT). A CT é uma das tarefas mais populares do Processamento de Linguagem Natural, e tem crescido cada vez mais nos últimos anos com a popularização da internet e disponibilização em massa de dados textuais on-line. Através da CT, é possível, a partir do conhecimento obtido das características de um documento textual, associar este documento a uma ou várias categorias (*classes*) pré-definidas. O principal ponto da CT é criar um modelo de Aprendizagem de Máquina Supervisionada que seja capaz de associar, corretamente, o maior número de documentos às suas respectivas classes.

O enorme volume de documentos presentes nos sites de culinária torna-os de grande interesse para as pesquisas na área de classificação automática de receitas, uma vez que os mesmos representam ricas fontes de dados para o treinamento dos modelos inteligentes. Diversas informações das receitas podem ser inferidas a partir dos modelos de classificação, dependendo apenas dos conjuntos de dados fornecidos como entrada aos mesmos. Assim, o uso desses repositórios on-line pode representar vários benefícios no desenvolvimento de sistemas mais

[1] <https://www.youtube.com/@VillageCookingChannel/>
[2] <https://www.youtube.com/@buzzfeedtasty/>
[3] <https://www.youtube.com/@gordonramsay/>
[4] <https://www.similarweb.com/pt/top-websites/food-and-drink/cooking-and-recipes/>

precisos para o atendimento das necessidades de usuários que possuam algum tipo de restrição alimentar. Entre os benefícios, podemos citar:

- A inferência automática de informações torna os sistemas mais utilizáveis por parte de seus usuários, uma vez que garante o fornecimento de informações mais completas acerca dos alimentos, assim como retira dos usuários a necessidade do preenchimento de formulários longos e complexos no ato de submissão das receitas. Tais informações podem ser fundamentais para auxiliar os usuários no processo de tomada de decisões;
- O preenchimento automático de informações (ou a sugestão aos usuários de como determinados campos poderiam ser corretamente preenchidos) pode evitar a inserção de dados incorretos, assim como diminuir ou eliminar a ocorrência de campos não preenchidos;
- A rotulação (classificação) automática de receitas é parte fundamental na elaboração de Sistemas de Recomendação de Receitas mais precisos. Com o aumento do volume dos dados nos repositórios on-line, esses sistemas de recomendação têm se tornado de grande interesse, tanto por parte de pesquisadores, quanto por parte de empresas (GE; RICCI; MASSIMO, 2015; MAO et al., 2016; ELSWEILER; TRATTNER; HARVEY, 2017; CHEN et al., 2020);
- A inferência automática de informações sobre receitas pode ainda beneficiar a elaboração de outros tipos de sistemas complexos, como sistemas de geração de cardápios, e de dietas completas e personalizadas (MAJUMDER et al., 2019; AZZIMANI et al., 2022).

O restante deste Capítulo está dividido da seguinte forma: na próxima seção (Seção 1.1), os objetivos do trabalho serão apresentados, seguidos das produções bibliográficas realizadas ao longo do desenvolvimento desta dissertação (Seção 1.2), e, por fim, uma visão geral da organização do trabalho como um todo é apresentada (Seção 1.3).

1.1 OBJETIVOS DO TRABALHO

Este trabalho tem como principal objetivo realizar a identificação automática de receitas culinárias que infrinjam dietas restritivas, através da classificação de receitas dietéticas. Essa classificação pode ser aplicada para facilitar a busca de usuários que possuem algum tipo de

restrição alimentar. Para isso, algumas dietas restritivas são selecionadas, sendo elas: a **diabética**, ***gluten-free*** (livre de glúten), ***lactose-free*** (livre de lactose), ***kid-friendly*** (adequada para crianças) e **vegana**.

Através de uma revisão da literatura (Capítulo 2), alguns dos principais modelos da área de Aprendizagem de Máquina Supervisionada (classificadores) foram selecionados, buscando investigar quais dessas técnicas seriam as mais apropriadas para execução da tarefa de identificação automática de receitas dietéticas e com restrições. Os seguintes classificadores são adotados: Árvore de Decisão (AD), *Bidirectional Encoder Representations from Transformers* (BERT), Floresta Aleatória (FA), K-Vizinhos Mais Próximos (K-NN), Naive Bayes (NB), Perceptron Multicamadas (MLP), Regressão Logística (RL) e Máquinas de Vetores de Suporte (SVM).

Com o objetivo de selecionar as informações que mais contribuem para o processo de aprendizado dos classificadores, conjuntos de dados compostos por diferentes componentes textuais das receitas são avaliados. Esses conjuntos de dados são compostos pelas **listas de ingredientes**, **modos de preparo**, **títulos** e **descrições das receitas**, sendo tais conjuntos de dados avaliados individualmente, e através da combinação dos mesmos.

As principais contribuições deste trabalho são:

- Revisão dos trabalhos publicados na tarefa de classificação de receitas culinárias;
- Análise da base de dados Food.com e de suas categorias, com foco nas subcategorias dietéticas;
- Comparação dos principais modelos utilizados na literatura de classificação de receitas culinárias, na tarefa de identificação de receitas para dietas restritivas;
- Avaliação do impacto dos diferentes conjuntos de características das receitas no treinamento dos modelos de classificação;

1.2 PRODUÇÃO BIBLIOGRÁFICA

Nesta seção, as contribuições bibliográficas originadas durante o desenvolvimento deste trabalho são apresentadas. Serão listados, na sequência, os trabalhos publicados pela autora Larissa F. S. Britto no decorrer de seu mestrado acadêmico no Centro de Informática da Universidade Federal de Pernambuco (CIn-UFPE).

1.2.1 Capítulos de livros publicados

1. PACÍFICO, L. D. S.; **BRITTO, L. F. S.**; LUDERMIR, T. B. (2022). *Improved Alternative Average Support Value for Automatic Ingredient Substitute Recommendation in Cooking Recipes*. Publicado em: **Lecture Notes in Computer Science**. Vide (PACIFICO; BRITTO; LUDERMIR, 2022b).

1.2.2 Artigos completos publicados e aceitos para publicação em congressos e conferências

1. PACÍFICO, L. D. S.; **BRITTO, L. F. S.**; LUDERMIR, T. B. (2022). *Improved Alternative Average Support Value for Automatic Ingredient Substitute Recommendation in Cooking Recipes*. Artigo aceito para publicação no **XI Brazilian Conference on Intelligent Systems (BRACIS 2022)**. Vide (PACIFICO; BRITTO; LUDERMIR, 2022c).
2. PACÍFICO, L. D. S.; **BRITTO, L. F. S.**; LUDERMIR, T. B. (2022). *Automatic Recipe Ingredient Substitution Based on Text Mining and Data Clustering Approaches*. Artigo aceito para publicação no **Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2022)**. Vide (PACIFICO; BRITTO; LUDERMIR, 2022a).
3. **BRITTO, L. F. S.**; PACÍFICO, L. D. S.; LUDERMIR, T. B. (2021). *Inferência Automática de Nível Calórico de Receitas Culinárias Através de Técnicas de Aprendizagem de Máquina*. Artigo aceito para publicação no **XVII Simpósio Brasileiro de Sistemas de Informação (Anais Estendidos do SBSI 2021)**. Vide (BRITTO; PACIFICO; LUDERMIR, 2021).
4. PACÍFICO, L. D. S.; **BRITTO, L. F. S.**; LUDERMIR, T. B. (2021). *Geração de Receitas Culinárias para Usuários com Restrições Alimentares pela Substituição Automática de Ingredientes*. Artigo aceito para publicação no **XXXIII Seminário Integrado de Software e Hardware (SEMISH 2021)**. Vide (PACIFICO; BRITTO; LUDERMIR, 2021a).
5. PACÍFICO, L. D. S.; **BRITTO, L. F. S.**; LUDERMIR, T. B. (2021). *Ingredient Substitute Recommendation Based on Collaborative Filtering and Recipe Context for Automatic Allergy-Safe Recipe Generation*. Artigo aceito para publicação no **XXVII Simpósio**

Brasileiro de Multimídia e Web (WebMedia '21). Vide (PACIFICO; BRITTO; LUDERMIR, 2021b).

1.3 ORGANIZAÇÃO DO TRABALHO

O restante deste trabalho está dividido da seguinte forma:

- Capítulo 2 – **Revisão da Literatura**: Nesse capítulo, são apresentados e analisados trabalhos que abordam a classificação de receitas. Esses trabalhos são agrupados de acordo com as informações inferidas, e avaliados quanto às abordagens utilizadas.
- Capítulo 3 – **Metodologia**: A metodologia adotada no presente trabalho é apresentada em detalhes nesse capítulo. São descritos desde a base de dados, aos métodos para extração de características, e ainda os modelos de classificação.
- Capítulo 4 – **Análise Experimental**: Nesse capítulo é descrita a configuração experimental usada para avaliar a metodologia proposta; além disso, os resultados obtidos através dos experimentos são apresentados e discutidos.
- Capítulo 5 – **Conclusões**: Nesse capítulo, as conclusões do trabalho são apresentadas, assim com propostas de linhas para pesquisas futuras.

2 REVISÃO DA LITERATURA

Neste capítulo, uma breve revisão da literatura em classificação de receitas culinárias será realizada. Uma vez que o presente trabalho será conduzido através de uma abordagem de análise de documentos textuais por meio de técnicas de **Processamento de Linguagem Natural** (vide Capítulo 3), pesquisas realizadas em classificação de receitas por meio da **Visão Computacional**, como as realizadas em (WANG et al., 2015; PAN et al., 2017; KAYIKÇI; BAŞOL; DÖRTER, 2019; VIJAYAKUMARI; VUTKUR; VISHWANATH, 2022), não serão incluídas na análise final. Como outro critério de inclusão e exclusão, apenas trabalhos publicados nos idiomas **inglês** e **português** serão discutidos.

Uma síntese dos trabalhos analisados pode ser vista das Tabelas 2 à 4.

Os trabalhos são agrupados nas próximas seções de acordo com as informações que são inferidas pelos autores. Na próxima seção (Seção 2.1), são avaliados os trabalhos que realizam a identificação do tipo de culinária das receitas. Na Seção 2.2, são revisados os trabalhos que que analisam o nível de dificuldade de preparo da receita e o tipo de refeição preparada. Por fim, os trabalhos que realizam a inferência de restrições e alergias são apresentados na Seção 2.3.

2.1 TIPO DE CULINÁRIA

Competições de extração de conhecimentos têm estimulado o esforço de pesquisadores para o desenvolvimento de ferramentas capazes de realizar a inferência de informações em receitas culinárias através da classificação textual. Uma das principais competições nessa área é a *"What's Cooking?"* (KAN, 2015), que, só na plataforma Kaggle, conta com mais de mil competidores. A competição tem como objetivo principal a predição automática do tipo de culinária de uma receita, dada sua lista de ingredientes, isto é, a partir dos ingredientes, um modelo deve ser capaz de identificar a qual país, ou região, aquela receita pertence. A base de dados proposta na competição, composta por receitas extraídas do site culinário Yummly, tem sido amplamente utilizada em trabalhos que aplicam diferentes metodologias para lidar com esse problema, como no trabalho de (HOLSTE; NYAYAPATI; WONG, 2015), em que os classificadores Floresta Aleatória e Regressão Logística, comumente utilizados para problemas de Classificação de Texto, foram adotados. Nesse mesmo trabalho, uma análise exploratória

foi realizada, através da qual foi possível observar os seguintes desafios presentes na base de dados:

- **Desbalanceamento das classes** - A culinária mais popular na base de dados, a italiana, é encontrada em quase 20% das observações, enquanto culinárias como a brasileira e a russa, aparecem em apenas 1% das receitas. Esse aspecto observado no conjunto de dados pode causar um viés nos classificadores, direcionando-os à predição dos novos documentos nas classes majoritárias;

- **Ingredientes de baixíssima frequência** - Outro problema encontrado através dessa análise é a existência de ingredientes com frequência extremamente baixa, como *mahlab* e *pão chinês* (*chinese buns*), que possuem apenas uma ocorrência em toda base de dados, o que pode diminuir a capacidade de generalização e levar ao *overfitting* dos classificadores;

- **Ingredientes de altíssima frequência** - Em contraste ao item anterior, também foi possível observar na base de dados a existência de ingredientes que ocorrem com grande frequência em todas as cozinhas, como *sal* e *água*. Esses ingredientes teriam, em teoria, pouco poder preditivo;

- **Ruídos na lista de ingredientes** - O maior desafio encontrado no entanto, foi a existência de ruídos na lista de ingredientes, como medidas e preparos. Esses ruídos podem fazer com que duas referências a um mesmo ingrediente sejam consideradas distintas, como por exemplo, *ovos grandes* e *ovos médios*, o que poderia levar a perda de acurácia dos classificadores.

Para lidar com os ruídos presentes nas listas de ingredientes, os autores aplicaram uma extensa etapa de pré-processamento e normalização, que inclui a conversão dos documentos para *lowercase* (letras minúsculas), remoção de pontuação e espaços em branco excessivos, e remoção de adjetivos de uma lista pré-definida. Para extração de características dos ingredientes, foi utilizada uma representação *Bag-of-Ingredients*, que é uma variação do *Bag-of-Words* (KOWSARI et al., 2019) para as receitas culinárias. Os autores propõem ainda um sistema de pontuação baseado em *Term Frequency-Inverse Document Frequency* (TF-IDF) (vide Seção 3.2), que calcula um peso para cada ingrediente em cada uma das culinárias, essa abordagem, porém, não demonstrou eficácia, tendo em vista que o modelo que alcançou melhores resulta-

dos é composto apenas pela representação *Bag-of-Ingredients* tradicional com o classificador Regressão Logística, alcançando uma taxa de erro de teste de 22.13%.

Outros trabalhos usaram diferentes abordagens para lidar com o problema de ruídos na lista de ingredientes. Em (KUMAR; KUMAR; SOMAN, 2016), um processo de stemização, que é um método para padronizar diferentes formas de um mesmo termo (KOWSARI et al., 2019), foi aplicado em toda base de dados, na tentativa de remover redundâncias nos ingredientes, como em *ovo* e *ovos*. Outras etapas de processamento também foram aplicadas, como a remoção de pontuação, espaços em branco e *stopwords* (termos que possuem pouca significância (KOWSARI et al., 2019)). A base de dados foi convertida em uma representação de *Bag-of-Words*, e de forma semelhante à (HOLSTE; NYAYAPATI; WONG, 2015), o classificador Floresta Aleatória foi adotado, desta vez sendo comparado com o algoritmo *Extreme Gradient Boosting*, onde este último teve o melhor desempenho, alcançando 80% de acurácia.

A stemização da lista de ingredientes também é realizada em (LI; WANG, 2015), em conjunto com outras etapas de pré-processamento e normalização, como remoção de conteúdos entre parênteses, pontuação, medidas e demais dígitos. Representações de *Bag-of-Words* e TF-IDF são utilizadas, e abordagens para seleção de características, como a medida de Informação Mútua (XU et al., 2007) e a remoção de ingredientes com frequência igual ou inferior a 5, são aplicadas, reduzindo o espaço de características em até 54%. Por fim, diferentes classificadores (Naives Bayes, SVM, Regressão Logística e K-NN) têm suas performances comparadas. Os resultados apontam que a seleção de características foi capaz de promover uma melhora na acurácia dos modelos. Os classificadores Regressão Logística e SVM se destacaram entre os demais, alcançando 78% e 76% de acurácia, respectivamente. Uma análise na matriz de confusão indicou que algumas classes são altamente correlacionadas, e que essa correlação pode reduzir a precisão dos algoritmos.

Ainda como resposta à competição *"What's Cooking?"*, (VERMA; ARORA, 2015) fizeram uso de diferentes abordagens para extração de características, incluindo uma comparação entre *Bag-of-Words*, TF-IDF, e variações na quantidade de n-gramas extraídos. Alguns classificadores, até então não testados nessa tarefa foram avaliados, como o Perceptron e um classificador Passivo-Agressivo, porém nenhum deles foi capaz de superar o desempenho dos classificadores SVM (com unigramas) e Regressão Logística (com bigramas), ambos alcançando 79% de acurácia. O TF-IDF demonstrou sua capacidade para extração de características, tendo um desempenho superior a representação *Bag-of-Words*.

O bom desempenho dos classificadores SVM e Regressão Logística na tarefa de classifi-

cação de receitas foi reforçado em (GHEWARI; RAIYANI, 2015), onde ambos os classificadores alcançaram 81% de acurácia. De forma conflitante aos demais trabalhos, o uso TF-IDF implicou numa piora no desempenho dos classificadores, quando comparado à representação de *Bag-of-Words* binária. Os autores utilizaram a técnica Análise dos Componentes Principais (*Principal Components Analysis*) (KOWSARI et al., 2019) para redução em mais de 60% do espaço de características. O vetor de características foi estendido com a adição de 20 características correspondentes a cada a uma das culinárias, representando o número de ingredientes característicos daquela culinária presentes na receita, abordagem similar ao sistema de pontuação proposto em (HOLSTE; NYAYAPATI; WONG, 2015). A dificuldade na predição de cozinhas correlacionadas, observada em (LI; WANG, 2015) também foi constatada neste trabalho, através da análise da representação *T-Distributed Stochastic Neighbor Embedding* (KOWSARI et al., 2019) do espaço de características.

Enquanto os trabalhos citados anteriormente seguiam uma metodologia baseada em aprendizagem supervisionada, (PAREKH, 2015) gera um modelo que combina tanto técnicas não supervisionadas, quanto supervisionadas. A base de dados é utilizada para treinar um modelo *Word2Vec* (KOWSARI et al., 2019). Os dados são transformados em um espaço gaussiano que em seguida é agrupado pelo algoritmo *K-means* (MACQUEEN, 1965). Por fim, os dados são classificados. A abordagem, no entanto, não teve um bom desempenho, alcançando apenas 76% no melhor caso. Os autores acreditam, porém, que o aumento da base de dados poderia melhorar os resultados do modelo de forma significativa.

(KALAJDZISKI et al., 2018) foi capaz de obter resultados promissores ao comparar as seguintes metodologias para limpeza e redução dos dados:

1. **Correção de Erros de Digitação** - A medida Distância de Levenshtein é aplicada em todos os ingredientes da base de dados para filtrar ingredientes que, por erros de digitação, seriam considerados diferentes pelos classificadores, como por exemplo *groundnut* e *ground nut*;

2. **Remoção de *Outliers*** - Através da análise da frequência, foram removidos ingredientes considerados *outliers*;

3. **Remoção de Ruídos** - A anotação das classes gramaticais (*Part-of-Speech Tagging*) foi executada em toda base de dados para filtrar ruídos na lista de ingredientes, assim como nos preparos e processos, como em *picado*, *grelhado*, *frito*, entre outros.

A correção dos erros de digitação usando a Distância de Levenshtein foi capaz de melhorar a acurácia dos modelos adotados. O melhor desempenho, 81% de acurácia, foi obtido pelo classificador SVM. As outras etapas de pré-processamento (remoção de *outliers* e ruídos), apesar de não melhorarem a acurácia, diminuíram de forma significativa o espaço de características.

Além de abordar a classificação do tipo de culinária, (ROITHER; KURZ; SONNLEITNER, 2022) utiliza a lista de ingredientes para desempenhar também a classificação de alergênicos, e para isso, além da base de dados Yummly, é utilizada também a base *Openfoodfacts*, contendo informações relativas a alimentos e seus alergênicos. Uma etapa de pré-processamento é aplicada, na qual é realizada remoção de *stopwords* e caracteres não alfanuméricos, assim como a lematização (KOWSARI et al., 2019) dos ingredientes. Os dados são balanceados, e é feita a extração e seleção de características através do peso TF-IDF. Por meio dos resultados experimentais, foi possível observar que a seleção de um número de características menor que 2000 resulta em uma queda na acurácia de classificação. Diferentes algoritmos para classificação são comparados, todos eles passando por um processo de otimização dos parâmetros.

Os resultados reforçam, novamente, o bom desempenho dos classificadores Regressão Logística e SVM para essa tarefa. Já na classificação de alergênicos, destacam-se os classificadores Perceptron Multicamadas e Floresta Aleatória. A maior contribuição desse trabalho é a execução de um estudo de caso comparando o desempenho do sistema proposto com o de 10 participantes humanos, na tarefa de detecção de 14 alergênicos em receitas. Os resultados apontam que o sistema obteve melhor performance que um participante médio, sem conhecimento prévio a respeito das alergias. O sistema consegue ainda realizar a identificação de maneira muito mais rápida, levando no máximo 5 segundos, enquanto os participantes levam uma média de 48 segundos por receita. Apesar disso, o sistema tem um desempenho inferior a um participante que recebe informações prévias sobre ingredientes que causam alergias. Por esse motivo os autores não recomendam o uso do sistema para uma tarefa crítica, como identificar alérgenos para alguém com restrições alimentares.

Por fim, alguns problemas e desafios encontrados na base de dados são levantados pelos autores. A base de dados *OpenFoodFacts* é colaborativa, isto é, depende da contribuição dos seus usuários para inserção das informações que compõe a base, de forma que qualquer usuário pode inserir informações sem que haja qualquer tipo de verificação, estando essas informações suscetíveis a erros e a falta de padronização. Mais trabalhos que abordam a classificação de ingredientes alergênicos em receitas serão analisados na Seção 2.3.

Enquanto os trabalhos citados até aqui adotaram a base de dados Yummly da competição

"What's Cooking?", outros autores optaram por propor suas próprias bases de dados para a tarefa de classificação do tipo de culinária.

Em (NAIK; POLAMREDDI, 2015), uma base de dados altamente desbalanceada, composta por receitas do sites Epicurious e Menupan, é proposta. Diferentes conjuntos de dados de entrada são criados a partir dessa base, no intuito de compreender o impacto do desbalanceamento nos dados de treinamento no desempenho dos classificadores. Apesar do classificador treinado com dados desbalanceados alcançar maior acurácia, esse modelo acaba atribuindo a classe majoritária para quase todos os documentos, indiscriminadamente, sendo incapaz de identificar documentos das demais classes.

(SU et al., 2014) coletou receitas de 70 diferentes tipos de culinária do site de receitas Food.com. Apenas receitas das 6 classes majoritárias foram mantidas, e técnicas para *upsampling* e *downsampling* foram aplicadas para balanceamento da base de dados. Os dados são convertidos em uma *Bag-of-Ingredients* e o método *Singular Value Decomposition* (detalhado pelos autores no trabalho) é utilizado para a redução de dimensão da matriz de características. O classificador SVM e técnicas de classificação associativa foram adotados para a predição. Os melhores resultados foram alcançados pelo classificador SVM, sem a redução de dimensionalidade. Os autores observaram a dificuldade de distinção pelos classificadores da culinária chinesa para a japonesa, observação semelhante às feitas por (LI; WANG, 2015). Essa mesma correlação entre cozinhas também é notada por (JAYARAMAN; CHOUDHURY; KUMAR, 2017), que observou o fenômeno de uma dada cozinha A ser classificada como cozinha B, e com frequência semelhante, cozinha B ser classificada como A. Essa relação foi observada principalmente entre as cozinhas francesas e italianas, chinesas e tailandesas, e mexicanas e indianas. Esse problema é tratado por (CELESTIN; PATRICK; SIMON, 2020), onde a classificação de culinárias correlacionadas é explorada. Os classificadores Regressão Logística, SVM e Perceptron Multicamadas são comparados na classificação da culinária de receitas de países asiáticos (chinesa, japonesa, tailandesa, vietnamita e indiana), em uma base de dados extraída do site BigOven. O classificador SVM foi o mais bem sucedido na detecção de todas as cozinhas.

Em (SHARMA; UPADHYAY; BAGLER, 2020), modelos tradicionais de Aprendizagem de Máquina são comparados com os modelos neurais *Long Short Term Memory* (LSTM), BERT e RoBERTa, na classificação do tipo de culinária por meio da base de dados RecipeDB, um largo conjunto de receitas extraídas de diversas fontes (AllRecipes, Epicurious e TarlaDalal). Além da lista de ingredientes, os autores consideram também outras informações das receitas, como os utensílios e modos de preparo. O bom desempenho dos modelos BERT e RoBERTa (que al-

cançaram 69% e 73% de acurácia, respectivamente), aponta para a necessidade e importância do uso de representações contextuais para a tarefa de classificação de receitas.

2.2 NÍVEL DE DIFICULDADE E TIPOS DE REFEIÇÃO

Outra competição popular que estimulou o desenvolvimento de pesquisas na tarefa de classificação de receitas culinárias foi a DEFT (*DÉfi Fouille de Textes*), uma tradicional competição francesa de mineração de dados textuais, que a cada ano aborda uma temática diferente. No ano de 2013, a competição teve como objetivo principal a análise automática de receitas culinárias no idioma francês, onde os seguintes desafios foram propostos aos competidores:

1. Com base no título e no texto da receita, identificar o nível de dificuldade entre 4 níveis: *muito fácil*, *fácil*, *bastante difícil*, *difícil*;
2. Com base no título e no texto da receita, identificar o tipo de refeição preparada: *entrada*, *prato principal*, *sobremesa*;
3. Combinar o texto de uma receita com seu título;
4. A partir do título e do texto da receita, extrair a lista de ingredientes.

Em (CHARTON et al., 2013), é proposta uma solução para os dois primeiros desafios da competição DEFT 2013. A classificação textual das receitas é executada adotando diferentes características, como classes gramaticais, nomes de ingredientes normalizados, números de palavras em uma seção (título, preparos) ou até quantidade de ingredientes. Uma análise na distribuição das classes da base de dados demonstrou um alto nível de desbalanceamento, onde as classes majoritárias, *muito fácil* e *fácil*, estão presentes em 90% das receitas. Já para a classificação do tipo de refeição, apesar da distribuição das classes ser mais balanceada, quase 50% da base de dados é composta pela classe *prato principal*. Os autores comparam diversos classificadores, no intuito de superar os resultados obtidos por outros autores nessa mesma competição. Os melhores resultados são apresentados pelos classificadores Árvore Logística (Uma combinação de Árvores de Decisão com Regressão Logística), que até a publicação do trabalho havia alcançado o primeiro lugar entre os competidores para a classificação do nível de dificuldade. Já para o segundo desafio, o SVM obteve o melhor resultado, aparecendo em segunda posição entre os modelos publicados para a competição.

As mesmas características linguísticas propostas por (CHARTON et al., 2013), também são adotadas por (MOHAMMADI et al., 2020a; MOHAMMADI et al., 2020b), onde são combinadas com características neurais extraídas por meio dos modelos BERT (CamemBERT)(MOHAMMADI et al., 2020a; MOHAMMADI et al., 2020b) e FastText (MOHAMMADI et al., 2020b). Essas características são concatenadas para alimentar modelos neurais como Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU) e LSTM. Os resultados alcançados mostraram que, em geral, o desempenho dos modelos adotados melhorou após a adição das características linguísticas. A metodologia empregada nesse trabalho, foi capaz de superar a proposta por (CHARTON et al., 2013) no primeiro desafio, mostrando a efetividade desses modelos em lidar com dados desbalanceados. O melhor desempenho no conjunto de teste foi obtido pelo modelo que combinava *embeddings* (KOWSARI et al., 2019) CamemBERT com o CNN como camada oculta.

Em (BRITTO; PACÍFICO; LUDERMIR, 2020) é realiza a classificação binária do nível de dificuldade de preparo de pratos, em uma base de dados extraída do site Food.com (no idioma inglês). Diferentes métodos para extração de características e classificadores são comparados, no intuito de selecionar a abordagem mais adequada para resolver esse problema. Além da extração das características, também foi feita a seleção das características mais significativas, de acordo com os pesos dados por cada método de extração. Uma análise dos resultados experimentais apontaram que o classificador Regressão Logística seria a melhor escolha para compor o sistema de inferência de dificuldade, alcançando 95% de acurácia. A seleção de características demonstrou ser uma boa abordagem para diminuir o custo computacional de execução, tendo em vista que a variação no desempenho dos classificadores em relação às quantidades de características avaliadas é baixa. De todos os métodos de extração de características, o modelo *Bag-of-Words* binária obteve o melhor resultado na maioria das métricas, independentemente do classificador selecionado.

2.3 ALERGIAS E RESTRIÇÕES ALIMENTARES

Apesar do grande volume de dados contidos nos repositórios de receitas culinárias, ainda é difícil para determinados grupos de usuários encontrar a receita desejada. A maior parte das receitas contidas nesses repositórios não contempla pessoas que possuem algum tipo de restrição alimentar, como pacientes alergênicos e alguns grupos culturais (ou seja, veganos, vegetarianos, grupos religiosos e assim por diante) (OOI; IIBA; TAKANO, 2015).

(ALEMANY et al., 2016) propõe um sistema multi-agente capaz de detectar automaticamente alergias alimentares em informações nutricionais, e rotular ingredientes com seus potenciais alérgenos. O sistema foi projetado como um conjunto de agentes especialistas, em que cada especialista representa uma técnica de Aprendizado de Máquina que tenta rotular novos ingredientes. O sistema é coordenado por um agente de tomada de decisão que leva em consideração a saída de cada agente. O sistema proposto foi integrado ao receteame.com, um site em espanhol que utiliza um sistema de recomendação social persuasivo para recomendar receitas, levando em consideração as preferências e restrições alimentares de seus usuários. Diferentes classificadores foram testados, e os que alcançaram os melhores desempenhos foram selecionados para executar o papel de agente especialista no sistema (Regressão Logística, Árvore de Decisão e K-NN). Os resultados mostram que a combinação de classificadores permite que o sistema melhore a detecção de alérgenos, e também aumente a precisão e a confiabilidade, partindo da ideia que o voto majoritário pode descartar erros de classificação individuais. Os autores levantam ainda a importância de uma correta classificação nessa tarefa, na qual erros de classificação podem comprometer seriamente a reputação do sistema e, no pior dos casos, causar sérios problemas de saúde aos usuários.

(BRITTO et al., 2020) apresenta uma abordagem *multi-label* para classificação automática de restrições alimentares em receitas culinárias. A abordagem proposta é implementada como uma metodologia de avaliação em três etapas. Primeiramente, é testado o classificador Floresta Aleatória, na tentativa de selecionar o melhor número de estimadores para o modelo. Na segunda etapa é analisada a seleção de características a partir da frequência dos ingredientes. De forma semelhante, a terceira etapa também avalia a seleção de características, mas dessa vez utilizando apenas a frequência dos ingredientes em cada uma das classes. Os resultados experimentais mostraram que a abordagem para seleção de ingredientes baseada na frequência nas classes foi a mais bem sucedida. A seleção de características demonstrou ser uma boa opção para compor um sistema de identificação de restrições alimentares, uma vez que o conjunto de dados avaliado foi reduzido para menos da metade, mantendo bons resultados na classificação, além de reduzir a quantidade de tempo necessária para o treinamento do algoritmo de classificação adotado. Numa avaliação geral, o melhor cenário de avaliação foi obtido quando são utilizados os 900 ingredientes mais frequentes de cada categoria, para um classificador com 100 estimadores.

(BRITTO; PACÍFICO; LUDERMIR, 2021) propõe uma abordagem de classificação textual binária para a construção de uma ferramenta de inferência automática do nível calórico de receitas.

A abordagem é dividida em seis etapas. Na primeira etapa é feita a obtenção da base de dados, extraída por (MAJUMDER et al., 2019) do site Food.com. Os modos de preparo são utilizados para essa tarefa, por, além de conter os ingredientes, possuírem também os preparos realizados sobre o mesmo, partindo da ideia que os preparos realizados sobre os ingredientes podem afetar sua composição nutricional. Na segunda etapa é feita a anotação da base de dados. Através das informações nutricionais, as receitas são categorizadas em duas classes referentes ao seu nível calórico, de acordo com as regras da *U.S. Food and Drug Administration* (U.S. Food and Drug Administration, 2020). Para evitar que o desbalanceamento das receitas impacte na performance dos classificadores, na terceira etapa é feito o balanceamento dos dados. Para transformar os textos brutos em dados numéricos suportados pelos classificadores, na quinta etapa é feita a extração de características por meio da medida TF-IDF. Por último, é feita a classificação, onde os classificadores Regressão Logística e SVM alcançaram os melhores resultados, com 81% de acurácia.

Em (MUPPALA, 2022), Graph Neural Networks (GCN) e Redes LSTM são usadas para identificar, através dos modos de preparo, se uma determinada receita é vegetariana ou não. Os resultados demonstram a capacidade do modelo GCN alcançar resultados melhores de forma mais rápida em comparação ao LSTM.

(KICHERER et al., 2017) propõe uma abordagem *multi-label* e multi-classes para a classificação de diferentes informações das receitas, como nível calórico, alergias e restrições, tipo de refeição, preparos, entre outros. Uma base de dados extraída do site alemão de receitas Chefkoch é adotada, e diversas informações das receitas são utilizadas, como ingredientes, modos de preparo, tempo de preparo, unidades de medidas, quantidades dos ingredientes, além de características linguísticas. Por meio da anotação das classes gramaticais, os autores concluíram que substantivos são mais importantes para a classificação que os verbos, indicando que ingredientes seriam mais descritivos que preparos. A visualização *T-Distributed Stochastic Neighbor Embedding* do espaços de características destaca a dificuldade da tarefa, na qual para algumas categorias, a classificação é comparativamente simples, com grupos bem definidos no gráfico, enquanto outras classes são desafiadoras, não chegando a formar grupos na visualização.

Tabela 2 – Análise de trabalhos que tratam a classificação de receitas culinárias (I).

Trabalho	Idioma	Categorização	Base de Dados	Características	Classificadores
(CHARTON et al., 2013)	Francês	▪ Nível de Dificuldade ▪ Tipo de Refeição	Marmiton (DEFT 2013)	▪ Ingredientes (Matriz Receita-Ingrediente) ▪ Características Linguísticas	▪ Árvore de Decisão ▪ Árvore Logística ▪ Naive Bayes ▪ Rede Bayesiana ▪ Regressão Logística ▪ SVM
(SU et al., 2014)	Inglês	Tipo de Culinária	Food.com	Ingredientes (Matriz Receita-Ingrediente)	Classificador Associativo SVM
(GHEWARI; RAIYANI, 2015)	Inglês	Tipo de Culinária	Yummly (*"What's Cooking?"*)	Ingredientes (Bag-of-Words, TF-IDF)	▪ Árvore de Decisão ▪ Floresta Aleatória ▪ K-NN Naive Bayes ▪ Regressão Logística ▪ SVM
(HOLSTE; NYAYAPATI; WONG, 2015)	Inglês	Tipo de Culinária	Yummly (*"What's Cooking?"*)	Ingredientes (Matriz Receita-Ingrediente, TF-IDF)	▪ Floresta Aleatória ▪ Regressão Logística
(LI; WANG, 2015)	Inglês	Tipo de Culinária	Yummly (*"What's Cooking?"*)	Ingredientes (Bag-of-Words, TF-IDF)	▪ K-NN ▪ Naive Bayes ▪ Regressão Logística ▪ SVM
(NAIK; POLAMREDDI, 2015)	Inglês	Tipo de Culinária	▪ Epicurious ▪ Menupan	Ingredientes (Matriz Receita-Ingrediente)	▪ K-NN ▪ Naive Bayes ▪ Perceptron ▪ Regressão Logística ▪ SVM
(PAREKH, 2015)	Inglês	Tipo de Culinária	Yummly (*"What's Cooking?"*)	Ingredientes (Word2Vec)	Floresta Aleatória

Fonte: A autora (2023).

Tabela 3 – Análise de trabalhos que tratam a classificação de receitas culinárias (II).

Trabalho	Idioma	Categorização	Base de Dados	Características	Classificadores
(VERMA; ARORA, 2015)	Inglês	Tipo de Culinária	Yummly (*"What's Cooking?"*)	Ingredientes (Matriz Receita-Ingrediente)	▪ Classificador Passivo Agressivo ▪ K-NN ▪ Naive Bayes ▪ Perceptron ▪ Regressão Logística ▪ SVM
(ALEMANY et al., 2016)	Espanhol	Alergias	receteame.com	Ingredientes (Matriz Receita-Ingrediente / TF-IDF) Informações Nutricionais	▪ Árvore de Decisão ▪ K-NN ▪ Regressão Linear ▪ Regressão Logística ▪ SVM
(KUMAR; KUMAR; SOMAN, 2016)	Inglês	Tipo de Culinária	Yummly (*"What's Cooking?"*)	Ingredientes (Matriz Receita-Ingrediente)	▪ Floresta Aleatória ▪ XGBoost
(JAYARAMAN; CHOUDHURY; KUMAR, 2017)	Inglês	Tipo de Culinária	▪ Epicurious ▪ Food.com ▪ Yummly	Ingredientes (Matriz Receita-Ingrediente / TF-IDF)	▪ Floresta Aleatória ▪ Naive Bayes ▪ Regressão Logística ▪ SVM
(KICHERER et al., 2017)	Alemão	▪ Dietas ▪ Informações Nutricionais ▪ Tipo de Culinária ▪ Tipo de Refeição e outros	Chefkoch	▪ Ingredientes (Matriz Receita-Ingrediente) ▪ Medidas ▪ Modos de Preparo (Bag-of-Words) ▪ Tempo de Preparo	Regressão Logística
(SAJADMANESH et al., 2017)	Inglês	Tipo de Culinária	▪ BBC Food Data ▪ Yummly	Ingredientes (Matriz Receita-Ingrediente)	▪ DNN ▪ SVM
(KALAJDZISKI et al., 2018)	Inglês	Tipo de Culinária	Yummly (*"What's Cooking?"*)	Ingredientes (Bag-of-Words / TF-IDF)	▪ Naive Bayes ▪ Redes Neurais ▪ SVM
(KICHERER et al., 2018)	Alemão	▪ Dietas ▪ Informações Nutricionais ▪ Tipo de Culinária ▪ Tipo de Refeição e outros	Chefkoch	▪ Ingredientes (Matriz Receita-Ingrediente) ▪ Medidas ▪ Modos de Preparo (Bag-of-Words) ▪ Tempo de Preparo	▪ Árvore de Decisão ▪ Regressão Logística
(BRITTO et al., 2019)	Português	▪ Dieta ▪ Tipo de Refeição e Alergias ▪ Tipo de Refeição	Páginas Brasileiras de Receitas	Ingredientes (TF-IDF)	▪ Floresta Aleatória ▪ Perceptron Multicamadas ▪ Naive Bayes ▪ Regressão Logísitica ▪ SVM

Fonte: A autora (2023).

Tabela 4 – Análise de trabalhos que tratam a classificação de receitas culinárias (III).

Trabalho	Idioma	Categorização	Base de Dados	Características	Classificadores
(BRITTO et al., 2020)	Inglês	Dieta/Restrição	Food.com	Ingredientes (Matriz Receita-Ingrediente)	Floresta Aleatória
(BRITTO; PACÍFICO; LUDERMIR, 2020)	Inglês	Nível de Dificuldade	Food.com	Modos de Preparo // (Bag-of-Words / TF-IDF / Matriz Binária)	▪ ▪ ▪ Floresta Aleatória ▪ Naive Bayes Regressão Logística
(CELESTIN; PATRICK; SIMON, 2020)	Inglês	Tipo de Culinária (Asiática)	BigOven	Ingredientes (TF-IDF)	▪ Perceptron Multicamadas ▪ Regressão Logística ▪ SVM
(MOHAMMADI et al., 2020a)	Francês	▪ Nível de Dificuldade ▪ Tipo de Refeição	Marmiton (DEFT 2013)	Características Linguísticas Custo Ingredientes Modos de Preparo Título	▪ BERT ▪ CNN-BERT ▪ GRU-BERT
(MOHAMMADI et al., 2020b)	Francês	Nível de Dificuldade	Marmiton (DEFT 2013)	▪ Custo ▪ Ingredientes ▪ Modos de Preparo ▪ Título	▪ BERT ▪ CNN-BERT ▪ GRU-BERT ▪ LSTM-BERT
(NIRMAL; CALDERA,)	Inglês	Tipo de Culinária	AllRecipes Epicurious Menupan	Ingredientes (Matriz Receita-Ingrediente)	Floresta Aleatória
(SHARMA; UPADHYAY; BAGLER, 2020)	Inglês	Tipo de Culinária	RecipeDB (AllRecipes, Epicurious e TarlaDalal)	▪ Ingredientes ▪ Modos de Preparo ▪ Utensílios (TF-IDF)	▪ BERT/RoBERTa ▪ Floresta Aleatória ▪ LSTM ▪ Naive Bayes ▪ Regressão Logística ▪ SVM
(BRITTO; PACÍFICO; LUDERMIR, 2021)	Inglês	Nível Calóricos	Food.com	Modos de Preparo (TF-IDF)	▪ Naive Bayes ▪ Regressão Logística ▪ SVM
(MUPPALA, 2022)	Inglês	Vegetariano	Food.com	Modos de Preparo	▪ GCN ▪ LSTM
(ROITHER; KURZ; SONNLEITNER, 2022)	Inglês	▪ Alergias ▪ Tipo de Culinária	▪ Open Food Facts ▪ Yummly (*"What's Cooking?"*)	Ingredientes (Matriz Receita-Ingrediente)	▪ Árvore de Decisão ▪ Floresta Aleatória ▪ K-NN ▪ Perceptron Multicamadas ▪ Regressão Logística ▪ SVM

Fonte: A autora (2023).

2.4 CONSIDERAÇÕES FINAIS

Este capítulo apresentou uma revisão dos trabalhos que abordam a classificação textual de receitas culinárias. Essa tarefa tem sido fortemente incentivada por competições de mineração de dados, como a *"What's Cooking?"* e a DEFT (*DÉfi Fouille de Textes*). Apesar da principal aplicação da classificação de receitas ser a descoberta do tipo de culinária, diversas informações têm sido inferidas através desse processo, como nível de dificuldade, tipo de prato, alergias, informações calóricas, entre outros. Por meio da revisão realizada, foi possível levantar algumas das principais técnicas utilizadas nessa tarefa, desde técnicas de processamento e extração de características dos textos, até os modelos de classificação mais bem-sucedidos. Algumas das principais técnicas levantadas foram adotadas neste trabalho, e serão melhor apresentadas no próximo capítulo (Capítulo 3).

3 METODOLOGIA

Neste capítulo, a metodologia adotada é apresentada. Uma vez que um problema de classificação de receitas é mapeado como um problema de classificação comum, podemos resolvê-lo através da execução de três etapas básicas, que serão descritas nas próximas seções: aquisição e pré-processamento da base de dados (Seção 3.1); extração de características (Seção 3.2); e a etapa de classificação (Seção 3.3).

3.1 BASE DE DADOS

Uma receita é composta basicamente por três partes: o **título** (nome dado à receita); a **lista de ingredientes**, contendo, geralmente, os ingredientes e suas medidas; os **modos de preparo**, com as instruções necessárias para o preparo da receita. Nos sites de receitas, detalhes adicionais, que auxiliam na escolha e preparo dos pratos, também são disponibilizadas, como tempo de preparo, nível de dificuldade, número de porções, informações nutricionais, restrições alimentares, tipo de refeição ou ocasião, entre outros. Esses detalhes usualmente são solicitadas aos próprios usuários no momento do cadastro de uma nova receita.

A base de dados usada neste trabalho foi proposta por (MAJUMDER et al., 2019), e extraída da rede social culinária Food.com[1].

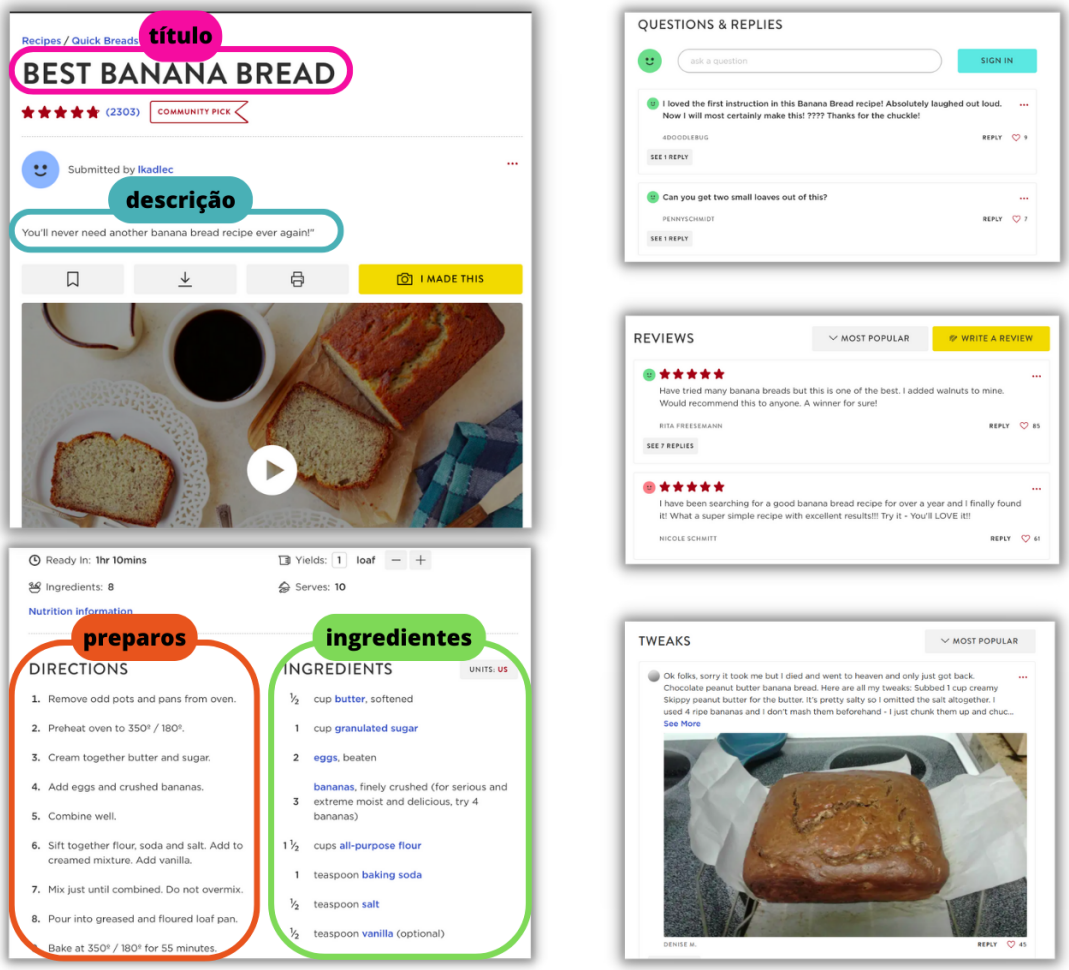
O Food.com é uma plataforma colaborativa em inglês, que permite que usuários compartilhem receitas culinárias, e interajam uns com os outros. A plataforma conta com mais de 500 mil receitas, além de um *feed* de atividades, no qual os usuários compartilham avaliações, sugestões de modificações, perguntas e fotos de seus pratos. Na Figura 2 é possível visualizar a página de uma receita do Food.com.

A base de dados proposta por (MAJUMDER et al., 2019) está publicamente disponível no Kaggle[2]. Ela contém mais de 230 mil receitas culinárias em inglês, e 1 milhão de interações de usuários, publicadas entre os anos de 2000 e 2018. A base é composta por diversos dados textuais e numéricos, que serão descritos a seguir:

- ***name*** – Título da receita (*string*);
- ***id*** – Identificador único da receita (número inteiro);

[1] <www.food.com>
[2] <https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions>

Figura 2 – Receita no site Food.com. Destacados nas imagens estão os componentes das receitas selecionadas para o presente trabalho.



Fonte: A autora (2023).

- ***minutes*** – Tempo de preparo da receita, em minutos (número inteiro);
- ***contributor_id*** – Identificador único do usuário autor da receita (número inteiro);
- ***submitted*** – Data de submissão da receita (*string*);
- ***tags*** – *Tags* informativas (lista de *strings*);
- ***nutrition*** – Informações nutricionais (lista numérica);
- ***n_steps*** – Número de etapas no modo de preparo da receita (número inteiro);
- ***steps*** – Modo de preparo da receita (lista de *strings*);
- ***description*** – Resumo descritivo da receita (*string*);

Figura 3 – Exemplo de entrada da base de dados Food.com.


```
[
  {
    "name": "best banana bread",
    "id": 2886,
    "minutes": 65,
    "contributor_id": 1762,
    "submitted": "1999-09-26",
    "tags": ['time-to-make', 'course', 'main-ingredient', 'cuisine', 'preparation',
             'north-american', 'breads', 'fruit', 'american', 'oven', 'dietary',
             'quick-breads', 'equipment', '4-hours-or-less'],
    "nutrition": [272.8,16, 97, 14, 7, 31, 14],
    "n_steps": 13,
    "steps": ['remove odd pots and pans from oven', 'preheat oven to 350 / 180',
              'cream together butter and sugar', 'add eggs and crushed bananas',
              'combine well', 'sift together flour , soda and salt',
              'add to creamed mixture', 'add vanilla', 'mix just until combined',
              'do not overmix', 'pour into greased and floured loaf pan',
              'bake at 350 / 180 for 55 minutes', 'keeps well , refrigerated'],
    "description": "you'll never need another banana bread recipe ever again!",
    "ingredients": ['butter', 'granulated sugar', 'eggs', 'bananas',
                    'all-purpose flour', 'baking soda', 'salt', 'vanilla'],
    "n_ingredients": 8
  }
]
```


Fonte: (MAJUMDER et al., 2019).

- ***ingredients*** – Ingredientes da receita (lista de *string*);
- ***n_ingredients*** – Número de ingredientes na receita (número inteiro).

Os dados disponíveis para cada receita podem ser visualizadas no exemplo apresentado na Figura 3.

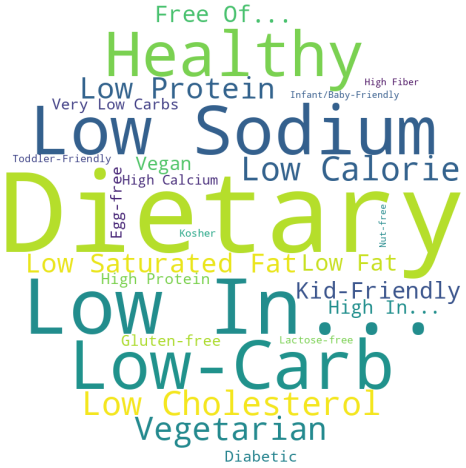
3.1.1 Restrições e Anotação da Base de Dados

Uma dos itens mais importantes da base de dados Food.com são as *tags* informativas, uma categorização feita pelos usuários ao cadastrar uma receita nova, a partir de um conjunto de *tags* pré-definidas disponibilizadas pelo site. Uma receita pode conter nenhuma, uma ou múltiplas *tags*. Essas *tags* abordam diversas propriedades das receitas, e são agrupadas hierarquicamente, sendo divididas em seis grupos principais: *Culinária* (*Cuisine*), *Dietético* (*Dietary*), *Ingrediente Principal* (*Main Ingredient*), *Ocasião* (*Occasion*), *Prato* (*Course*) e *Preparos* (*Preparation*). As *tags* pertencentes a cada um desses grupos são exibidas nas nuvens de palavras na Figura 4.

O principal objetivo deste trabalho é identificar receitas que se adequem às necessidades

Figura 4 – Nuvens de palavras das *tags*, divididas por grupos.

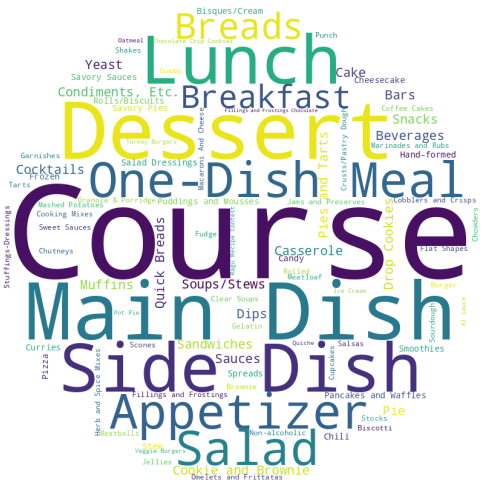
(a) *Culinária*



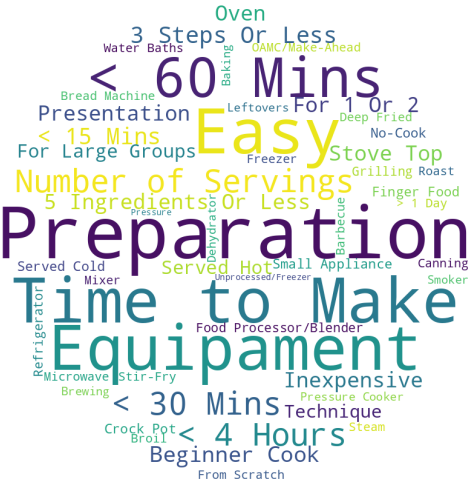
(b) *Dietético*

(c) *Ingrediente Principal*

(d) *Ocasião*



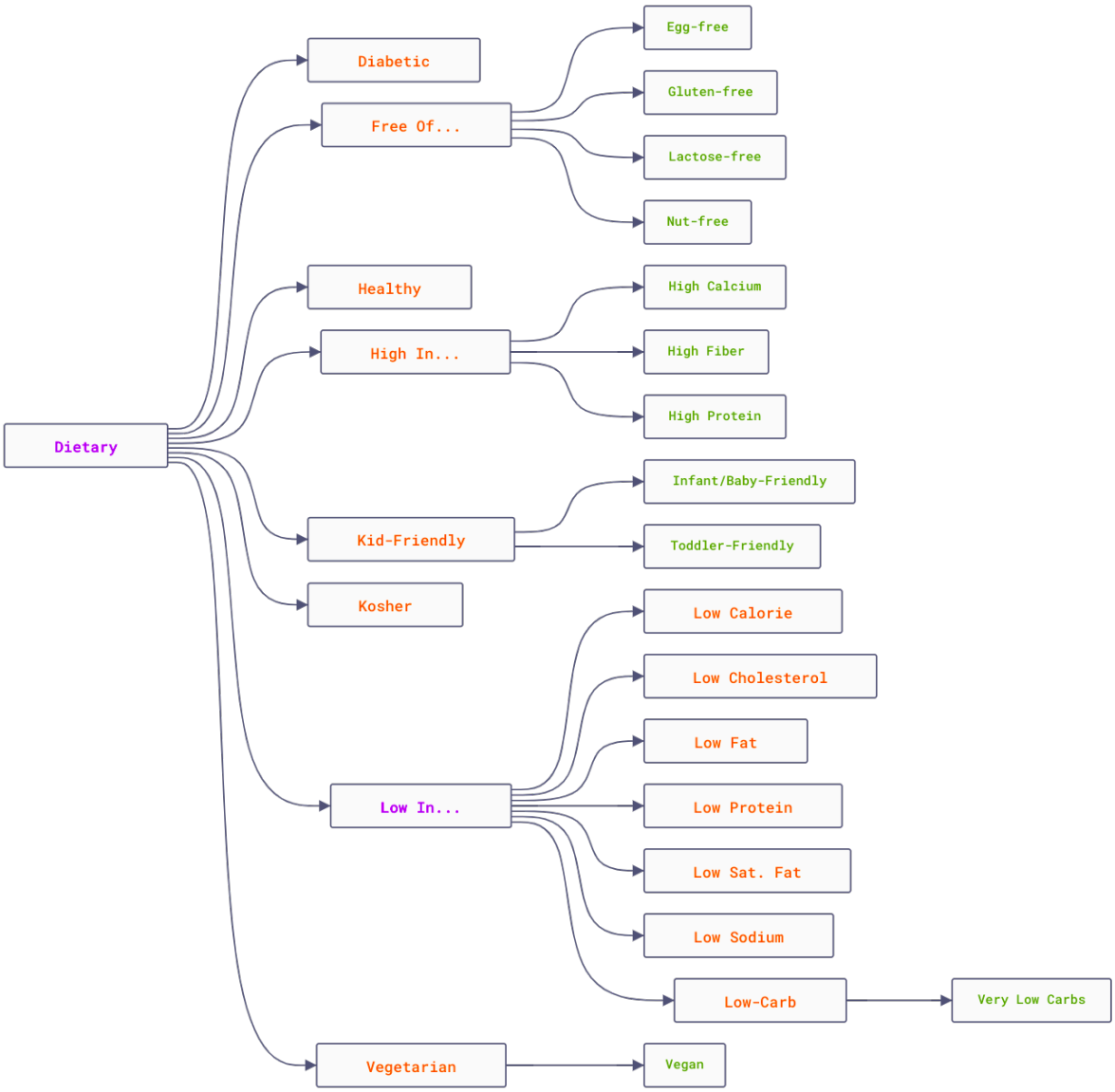
(e) *Prato*



(f) *Preparo*

Fonte: A autora (2023).

Figura 5 – Hierarquia do grupo de *tags* dietéticas.



Fonte: A autora (2023).

de grupos de usuários que possuam uma ou mais restrições alimentares, estando essas receitas rotuladas pelas *tags* do grupo *Dietético* da base de dados. Cinco grupos de dietas foram considerados na abordagem proposta: a diabética, *gluten-free*, *lactose-free*, *kid-friendly* e vegana. Mais detalhes sobre as dietas de interesse podem ser vistos na Tabela 5. A estrutura de relação de todas as *tags* dietéticas pode ser visto na Figura 5.

Para a geração das bases de dados utilizadas no presente trabalho, são escolhidas, dentre as receitas da base de dados Food.com, aquelas que contém a *tag* associada a uma determinada dieta (dentre as citadas na Tabela 5). Esses documentos são então rotulados como

Tabela 5 – Detalhes sobre as dietas selecionadas para a metodologia deste trabalho.

Dietas	Descrição	Quantidade de Receitas
Diabético	Doença crônica cujo tratamento exige uma alimentação especial. Entre as necessidades especiais estão a alimentação equilibrada, rica em carboidratos, fibras, proteínas, gorduras, vitaminas e minerais [3]	6284
Gluten-free (Livre de glúten)	Restrições ao glúten, composto de proteínas presente em alguns cereais	5546
Kid-friendly (Adequado para crianças)	Apropriado para o consumo por crianças	26591
Lactose-free (Livre de lactose)	Restrições à lactose, açúcar presente no leite e em seus derivados	4069
Vegano	Consumo apenas de produtos que não possuam origem animal	9858

Fonte: A autora (2023).

pertencentes à classe ***positiva***. Para a classe ***negativa*** (receitas que infringem a dieta), foram selecionadas aleatoriamente n receitas que não possuíam a *tag* em questão, sendo n o número de receitas na classe positiva, garantindo assim o balanceamento das classes na base de dado. Esse processo é repetido para cada uma das dietas restritivas selecionadas.

3.1.2 Conjuntos de Dados

Apesar de a maioria dos trabalhos que abordam a classificação de receitas se limitarem ao uso da lista de ingredientes, conforme analisado das Tabelas 2 a 4 (vide Capítulo 2), algumas pesquisas alcançaram resultados promissores ao adotar outros componentes das receitas, como os modos de preparo (KICHERER et al., 2018; BRITTO; PACÍFICO; LUDERMIR, 2020; MOHAMMADI et al., 2020a; MOHAMMADI et al., 2020b; SHARMA; UPADHYAY; BAGLER, 2020; BRITTO; PACÍFICO; LUDERMIR, 2021; MUPPALA, 2022). Neste trabalho, expandiremos as metodologias propostas para a tarefa de classificação de receita culinária, com a adição de características extraídas de diferentes elementos textuais das receitas, sendo eles: a **lista de ingredientes** (*ingredients*), o **modo de preparo** (*steps*), o **título** (*name*) e a **descrição da receita** (*description*). Esses elementos podem ser visualizados, em destaque na Figura 2. Os elementos textuais também são agrupados em um conjunto único, formando uma nova base de dados, que também será avaliada nos experimentos, sendo referida, de agora em diante, como conjunto **combinado** (vide Capítulo 4).

3.1.3 Pré-Processamento

Todos os itens selecionados para este trabalho foram originalmente publicados em formato textual (linguagem natural), sendo necessária a padronização desses documentos para um formato que seja mais adequado e significativo, visando garantir o bom desempenho dos classificadores. A base de dados adotada foi previamente pré-processada por (MAJUMDER et al., 2019), onde as seguintes etapas foram executadas:

- Conversão de todo o texto para ***lowercase*** (forma minúscula);
- Remoção de **pontuação** e **caracteres especiais**;
- **Tokenização**.

3.2 EXTRAÇÃO DE CARACTERÍSTICAS

Por meio da extração de características é possível transformar textos brutos em vetores numéricos suportados pelos modelos de classificação. Neste trabalho, a medida *Term Frequency-Inverse Document Frequency* (TF-IDF) (SALTON; MCGILL, 1983) foi aplicada na base de dados para a criação da matriz de características de entrada dos algoritmos de aprendizagem rasos (*Shallow Classifiers*).

O TF-IDF combina a Frequência do Termo (*Term Frequency* - TF), que tenta mensurar quão importante um termo é em determinado documento, com o Inverso da Frequência dos Documentos (*Inverse Document Frequency* - IDF), que mensura a importância do termo em todo o conjunto de dados, tentando diminuir assim a influência de termos que ocorrem com uma grande frequência, mas que possuem pouca relevância. A fórmula do TF-IDF pode ser vista na Equação (3.1).

$$TF-IDF_{t,d} = \frac{f_{t,d}}{\sum_{t_n \in d} f_{t_n,d}} \times \log \frac{N}{df_t} \tag{3.1}$$

onde t representa o termo e d o documento, N é o número total de documentos, df é o número de documentos em que t ocorre, e f retorna sua frequência.

Devido à simplicidade, facilidade de uso e baixo custo computacional, a extração de características com TF-IDF se tornou uma abordagem popular na tarefa de classificação de receitas culinárias. Outra vantagem do TF-IDF é sua capacidade de diminuir a influência de termos

menos significativos, como artigos, preposições e pronomes. No entanto, o TF-IDF é incapaz de capturar conhecimento semântico e sintático dos documentos, o que é fundamental para a compreensão textual. No intuito de superar essa limitação, o presente trabalho também analisará o comportamento de um modelo de linguagem pertencente ao paradigma da Aprendizagem Profunda (*Deep Learning*), o BERT.

3.3 CLASSIFICADORES

Em trabalhos anteriores na literatura de classificação de receitas culinárias, abordagens de Aprendizagem Máquina Rasa foram comumente utilizadas, no entanto, em trabalhos mais recentes, métodos de Aprendizagem Profunda têm obtido resultados promissores. Neste trabalho, comparamos diferentes abordagens para a identificação de receitas dietéticas, incluindo técnicas tradicionais de Aprendizagem de Máquina (Seção 3.3.1) e o modelo pré-treinado de Aprendizagem Profunda BERT (Seção 3.3.2).

O processo de identificação de cada uma das restrições por parte dos algoritmos foi modelado como um problema de **classificação binária**, no qual para cada uma das restrições selecionadas é gerado um classificador, que identifica se aquela receita pode ser consumida por um indivíduo que segue tal dieta ou se a mesma fere suas restrições.

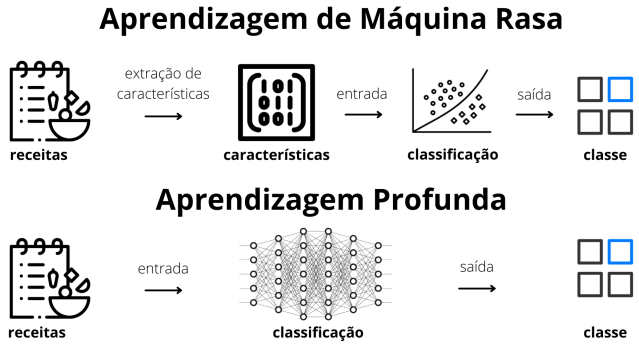
3.3.1 Classificadores Rasos

A Aprendizagem Rasa refere-se às abordagens mais tradicionais de Aprendizagem de Máquina, nas quais, ao alcançar um determinado nível de performance, os modelos tendem a não apresentar melhora com a adição de novos dados, além de possuírem uma alta dependência de um bom conjunto de características previamente extraídas. Uma comparação entre o fluxo de funcionamento de métodos rasos e profundos para Aprendizagem de Máquina é exibida na Figura 6.

3.3.1.1 Árvore de Decisão

Árvores de Decisão (*Decision Tree*) são modelos supervisionados de suporte à tomada de decisão, que classificam padrões usando uma sequência pré-definida de regras. A abordagem mais antiga baseada em árvores foi proposta por (MORGAN; SONQUIST, 1963). Desde então,

Figura 6 – Comparação entre métodos de Aprendizagem Rasa e Profunda.



Fonte: A autora (2023).

Árvores de Decisões têm sido utilizadas com sucesso em diversas tarefas de classificação (SAFAVIAN; LANDGREBE, 1991).

A ideia base desta abordagem é dividir uma decisão complexa em uma união de várias decisões mais simples. Uma árvore de decisão realiza classificação de dados com o auxílio de uma série de perguntas sobre suas características. Cada pergunta é representada por um nó da árvore, e cada um dos nós internos aponta para um nó filho, representando cada possível resposta da pergunta. As questões formam, assim, uma hierarquia, codificada como uma árvore. Um dado de entrada é associado a uma classe seguindo o caminho do nó superior (raiz), até um nó sem filhos (folha), de acordo com as respostas que se aplicam a essa entrada. É então atribuído a esse instância a classe do nó folha que a mesma atinge. As Árvores de Decisão são desenvolvidas pela adição incremental de nós de pergunta, usando exemplos de treinamento rotulados para orientar a escolha dessas questões. Idealmente, uma única e simples pergunta dividiria perfeitamente os exemplos de treinamento em suas classes. Se não existir nenhuma questão que gere tal separação, deve ser escolhida uma questão que separe os exemplos da forma mais clara possível. Várias medidas foram projetadas para avaliar o grau de não homogeneidade, ou impureza, em um conjunto de dados. Para Árvores de Decisão, duas medidas comumente adotadas são a entropia e o índice de Gini (KINGSFORD; SALZBERG, 2008).

Uma das maiores vantagens das Árvores de Decisão é a facilidade de aplicação. A estrutura possui uma natureza visual, tornando as árvores facilmente interpretáveis, compreensíveis e aplicáveis (DJURIS; IBRIC; DJURIC, 2013). Devido a sua interpretabilidade, informações relacionadas à identificação de características importantes e relacionamentos interclasses podem ser usadas para apoiar as decisões e analisar os dados (BROWN; MYLES, 2009).

Apesar de serem algoritmos de rápida execução, as Árvores de Decisão são extremamente

sensíveis a variações, sendo tão instáveis, que pequenas variações podem gerar modelos de árvores completamente diferentes. A criação de regras muito complexas e específicas pode também levar o classificador ao *overfitting* (QUINLAN, 1987; GIOVANELLI et al., 2017; KOWSARI et al., 2019).

3.3.1.2 Floresta Aleatória

Floresta Aleatória (*Random Forest*) (HO, 1995) é um classificador *ensemble*, que usa múltiplos modelos de Árvore de Decisão para melhorar o desempenho que uma única Árvore teria. Devido a sua efetividade, a técnica Floresta Aleatória tem sido amplamente empregada para tarefas de Classificação de Texto (KOWSARI et al., 2019).

O algoritmo Floresta Aleatória integra múltiplas Árvores de Decisão em uma floresta, partindo da ideia de *Ensemble Learning*, uma abordagem que explora múltiplos classificadores com o objetivo de gerar um modelo mais preciso e robusto (SIMSKE, 2019). O algoritmo Floresta Aleatória treina várias Árvores de Decisão através uma abordagem de *Bootstrapping Aggregation*, também conhecido como *Bagging*. Na *Bagging*, várias Árvores de Decisão são treinadas em paralelo com diferentes subconjuntos das características da base de dados de treinamento, garantindo assim que cada árvore na floresta seja única. Para a decisão final, a Floresta Aleatória agrega as decisões das árvores individuais, escolhendo a predição mais votada como resultado (MISRA; LI; HE, 2020).

Apesar de sua alta eficiência e baixo tempo de treinamento quando comparado com técnicas mais complexas, a Floresta Aleatória apresenta as seguintes desvantagens (KOWSARI et al., 2019):

- Bastante lento para criar previsões, quando treinadas;
- O bom desempenho do classificador está diretamente ligado à escolha do número de árvores na floresta. O crescimento do número de árvores pode aumentar de forma significativa a complexidade do tempo de execução;
- As decisões tomadas pelo algoritmo não são de fácil interpretabilidade.

3.3.1.3 K-Vizinhos Mais Próximos

O *K*-Vizinhos Mais Próximos (*K*-Nearest Neighbor – K-NN) é um método não paramétrico, inicialmente proposto por (FIX; HODGES, 1951), e mais tarde expandido por (COVER; HART, 1967). Esta técnica é usada para classificação nas mais diversas aplicações. O K-NN baseia-se na suposição mais básica de todos os modelos de classificação: observações semelhantes, ou seja, com características similares, tendem a ter uma saída correspondente (RICHMAN, 2011).

O processo de aprendizado nesse algoritmo ocorre apenas pelo armazenamento das instâncias rotuladas de treinamento para comparação futura. Já o teste do modelo é realizado comparando um determinado conjunto de teste (padrões novos, ainda não rotulados) ao conjunto de treinamento. Dado um documento de teste *d*, o algoritmo K-NN encontra os *K* vizinhos mais próximos de *d* entre todos os documentos no conjunto de treinamento (KOWSARI et al., 2019). Essa proximidade entre as instâncias pode ser medida por meio de uma métrica de distância, como a Distância Euclidiana. Por fim, o método calcula sua saída com base na média, ponderada ou não, dos *K* vizinhos mais próximos da instância de teste no conjunto de treinamento, isto é, o voto majoritário na vizinhança é usado para definir qual classe será atribuída ao padrão desconhecido.

O K-NN é um método simples, intuitivo, fácil de implementar e que se adapta a qualquer tipo de espaço de características. No entanto, esse classificador é limitado por restrições de armazenamento para problemas de pesquisa com grandes volumes de dados. Além disso, é um desafio encontrar um valor ótimo de *k* e uma função de distância significativa para documentos textuais.

3.3.1.4 Naive Bayes

O Naive Bayes é um classificador probabilístico baseado no teorema de Bayes, formulado por Thomas Bayes no século XVII, que usa as probabilidades conjuntas de termos e categorias para estimar as probabilidades das possíveis categorias de um determinado dado de teste (MUSHTAQ; MELLOUK, 2017). O classificador Naive Bayes assume que as características não interagem entre si, de forma que todas contribuem independente e igualmente para a probabilidade de uma amostra pertencer a uma classe específica (MISRA; LI; HE, 2020). O algoritmo Naive Bayes tem sido amplamente utilizado para as mais variadas tarefas de categorização textual (KOWSARI et al., 2019).

Através do conjunto de características, representados como X na Equação 3.2, o Naive Bayes é capaz de calcular a probabilidade dos dados de entrada pertencerem a uma determinada classe, representada por um Y.

$$p(X|Y) = \frac{p(X|Y)p(Y)}{p(X)} \tag{3.2}$$

onde:

- $p(X|Y)$ é a probabilidade posterior da classe dado o conjunto de características;
- $p(X|Y)$ é a probabilidade do conjunto de características dada a classe;
- $p(Y)$ é a probabilidade *a priori* da classe;
- $p(X)$ é a probabilidade *a priori* do conjunto de características.

Após calcular a probabilidade de cada uma das possíveis categorias, o algoritmo atribui à nova observação a classe mais provável, isso é, a classe com a maior probabilidade $p(X|Y)$.

A técnica Naive Bayes é simples de implementar, computacionalmente rápida, não é sensível a ruídos e funciona bem em conjuntos de dados grandes e com alta dimensionalidade, sendo útil para aplicações em tempo real (MISRA; LI; HE, 2020). Uma das limitações desse método é que a suposição de que as características são independentes uma das outras raramente é amparada por problemas reais.

3.3.1.5 Perceptron Multicamadas

O Perceptron Multicamadas (*Multi-Layer Perceptron*) é uma classe de Redes Neurais Artificiais do tipo *feed-forward*, na qual as informações fluem ao longo das várias camadas da rede para obter uma saída. Assim como as outras redes neurais, o Perceptron Multicamadas se inspira na estrutura biológica do cérebro. Apesar do Perceptron de camada única ser utilizado para tratar problemas lineares, apenas com o uso de múltiplas camadas, problemas complexos e não lineares podem ser solucionados de maneira efetiva.

A estrutura básica do MLP contém três camadas, que são a **camada de entrada**, a **camada escondida** e a **camada de saída**, todas completamente conectadas, isto é, todos os neurônios de uma camada estão conectados com todos os neurônios da próxima camada. A camada de entrada recebe os dados a serem processados pela rede. Por meio de uma função

de ativação, um sinal de saída é calculado e encaminhado para a próxima camada da rede, onde os neurônios coletam os sinais da camada anterior e calculam o próximo sinal, dando diferentes pesos aos neurônios, até atingirem a última camada, onde a saída é computada (EDWARD, 2018). O tamanho do vetor de entrada determina a quantidade de neurônios na camada de entrada, enquanto a quantidade de neurônios na camada de saída é definido pelo número de classes aprendidas. Os Perceptron Multicamadas são treinados por meio de um método chamado retro-propagação (*backpropagation*), no qual a cada iteração de aprendizado (época), o valor de saída do classificador para os dados de treinamento é calculado. Essa saída é comparada com o resultado correto, e uma função de perda é utilizada para calcular o erro do modelo. Esse erro é então propagado para as camadas anteriores, e utilizado para ajustar os pesos de entrada dos neurônios.

O treinamento de redes neurais geralmente exige muito tempo e um alto custo computacional. Apesar disso, após o treinamento, o Perceptron Multicamadas pode ser fácil e rapidamente aplicado para predição da classe de novas amostras. Uma outra vantagem, é que o modelo consegue ter alta eficiência quando treinado, tanto com grandes quanto com pequenos conjuntos de dados (AKKAYA; ÇOLAKOĞLU, 2019).

3.3.1.6 Regressão Logística

A Regressão Logística (*Logistic Regression*) foi um das primeiras técnicas para classificação desenvolvidas, proposta por (COX, 1970). A Regressão Logística, estima a probabilidade associada à ocorrência de determinado evento. Através deste algoritmo é também possível avaliar a influência de cada variável na ocorrência do evento alvo.

Assim como na Regressão Linear, a Regressão Logística tem pesos associados às instâncias de entrada, mas ao contrário da Regressão Linear, o algoritmo realiza uma combinação das características, e aplica a elas uma **função logística**, que pode ser visualizada na Equação 3.3, onde X representa as variáveis de entrada. A função logística é utilizada para modelar a saída binária do modelo, reduzindo-a de uma grande escala, para o intervalo de 0 a 1. O algoritmo usa a função de perda máxima verossimilhança: se a probabilidade calculada for maior que 0.5, a saída do classificador será 0, caso contrário, a saída será 1 (BELYADI; HAGHIGHAT, 2021).

$$f(x) = \frac{L}{1 + e^{-x}} \tag{3.3}$$

A Regressão Logística é de fácil implementação e não exige muitos recursos computacionais. A sua principal limitação é a suposição de linearidade entre a variável dependente e as variáveis independentes. Além disto, assim como o classificador Naive Bayes, a Regressão Logística supõe que as características são independentes uma das outras, mas isso raramente acontece em problemas reais (MISRA; LI; HE, 2020).

3.3.1.7 Máquinas de Vetores de Suporte

As Máquinas de Vetores de Suporte (*Support Vector Machines*) são algoritmos de Aprendizagem de Máquina supervisionada propostos por (CORTES; VAPNIK, 1995), que podem ser usados tanto para classificação, quanto para regressão. O SVM constrói um hiperplano num espaço multidimensional que separa as diferentes classes do conjunto de dados.

O principal objetivo do algoritmo SVM é usar um conjunto de treinamento para encontrar um hiperplano no espaço de características que produza a maior margem entre os objetos que pertencem a diferentes classes. A função de *kernel* desempenha um papel fundamental no classificador. Como os dados podem ser não lineares, a função de *kernel* é usada para transformá-los, fazendo-os mais separáveis linearmente, tornando possível a classificação.

O SVM é um método robusto, sendo pouco suscetível a problemas de *overfitting*, especialmente para documentos textuais, devido ao espaço de alta dimensionalidade. A maior limitação do SVM é a complexidade de tempo e memória. Além disso, a eficiência do modelo está diretamente ligada à escolha de uma função de *kernel* adequada (KOWSARI et al., 2019).

3.3.2 BERT

O *Bidirectional Encoder Representations from Transformers* (DEVLIN et al., 2018) é um modelo de representação de linguagem projetado para pré-treinar representações bidirecionais profundas de textos não rotulados. Nesta seção, serão descritos alguns aspectos do desenvolvimento e aplicação desse modelo.

3.3.2.1 Arquitetura

A arquitetura do classificador BERT é composta por um codificador *Transformer* bidirecional multicamadas, baseado na implementação descrita por (VASWANI et al., 2017). O *Trans-*

former é uma arquitetura de aprendizagem profunda capaz de aprender relações contextuais entre palavras, que tem sido amplamente utilizada em tarefas de Processamento de Linguagem Natural. Essa arquitetura foi originalmente proposta como um modelo sequência a sequência para tradução automática (VASWANI et al., 2017). A base dos modelos *Transformers* são os Mecanismos de Atenção, inspirados em uma propriedade importante da percepção humana, na qual ao processar um largo volume de informações, os humanos tendem a não processar toda a informação de uma só vez, mas ao contrário, focam seletivamente em partes importantes da informação, enquanto ignoram partes que seriam menos relevantes em determinado contexto (NIU; ZHONG; YU, 2021). Os Mecanismos de Atenção executam uma operação entre palavras, descobrindo como cada palavra se relaciona com as outras em uma dada sequência (ROTHMAN, 2021). O modelo base do BERT possui 12 camadas, com o tamanho da camada escondida de 768 e 12 cabeças de Atenção.

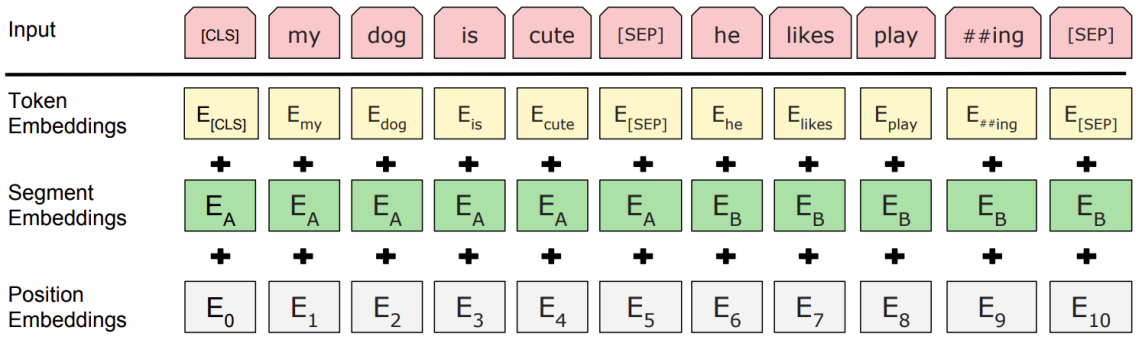
3.3.2.2 Representação

A entrada do modelo BERT é uma sequência de *tokens* (que é uma segmentação significativa de um texto, como uma palavra ou frase (KOWSARI et al., 2019)), que pode ser composta por uma única sentença ou duas sentenças agrupadas. O BERT usa a técnica *WordPiece* para segmentação de palavras. Nessa abordagem, o vocabulário é inicializado com caracteres individuais do idioma. Em seguida, o vocabulário é combinado, e as combinações mais frequentes são adicionadas iterativamente ao vocabulário. Através dessa representação, o vocabulário pode ser composto não só por palavras, mas também por sub-palavras, facilitando a representação de palavras raras e desconhecidas pelo modelo (WU et al., 2016).

Além disso, *embeddings* contendo as posições dos *tokens* e sentença à qual esses *tokens* pertencem também são adicionados à entrada, como pode ser visualizado na Figura 7.

Enquanto as técnicas anteriores com o mesmo propósito treinam o modelo na ordem da sequência das palavras, o BERT usa duas direções, analisando tanto o contexto à esquerda quanto à direita da palavra. Para alcançar esta vantagem, o BERT passa por um complexo processo de pré-treinamento descrito a seguir.

Figura 7 – Representação das entradas do BERT.



Fonte: (DEVLIN et al., 2018).

3.3.2.3 Pré-Treino

O BERT é treinado com o auxílio de duas tarefas: a predição de palavras mascaradas (*masked language models*) e predição da próxima sentença (*next sentence prediction*).

Para ser capaz de capturar o contexto à esquerda e à direita das palavras, o BERT usa modelos de linguagem mascarada, que mascara aleatoriamente alguns dos *tokens* de uma sentença de entrada, com o objetivo de prever o *token* mascarado. Através desse processo, os pesos do modelo são otimizados, buscando reproduzir na saída a mesma sentença de entrada.

Muitas tarefas de Processamento de Linguagem Natural se baseiam na compreensão das relações entre sentenças. Para alcançar essa compreensão, o BERT também é treinado na tarefa de predição da próxima sequência, na qual o classificador busca identificar a partir de um par de sentenças de entrada, se a primeira sentença precede a segunda no documento original.

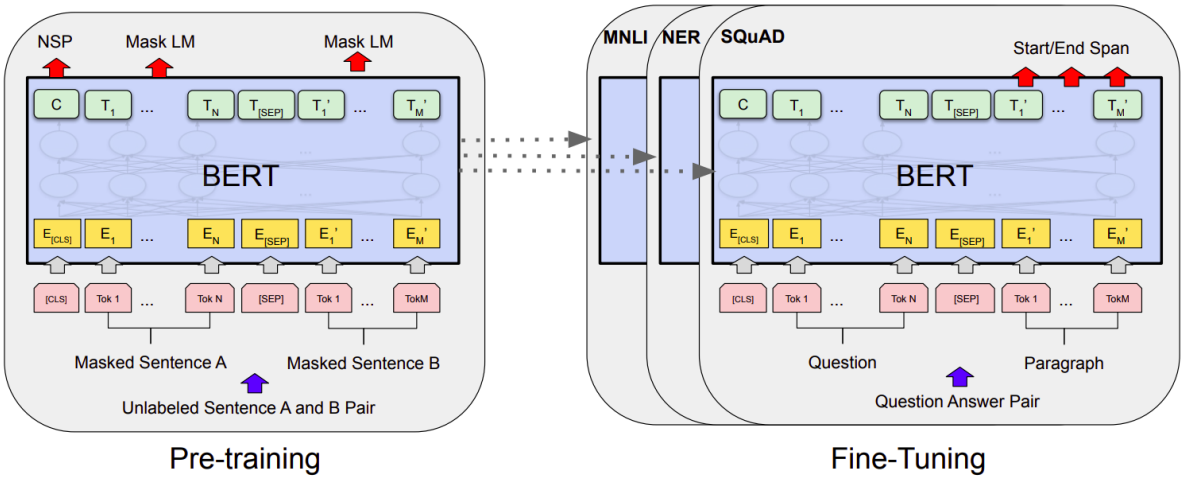
Para realizar todos esses procedimentos, o BERT utiliza os *corpora BooksCorpus* (ZHU et al., 2015) (composto por mais de 11 mil livros de diferentes gêneros) e *English Wikipedia* (contendo textos de páginas do Wikipédia).

Após a etapa de pré-treinamento, o BERT pode ser aplicado a tarefas específicas através do processo de *fine-tuning*.

3.3.2.4 Fine-Tuning

O *fine-tuning* é uma etapa de treinamento de parâmetros com dados rotulados e específicos da tarefa alvo a partir de um *checkpoint* de um modelo que foi previamente treinado. O pro-

Figura 8 – Visão geral do pré-treinamento e *fine-tuning* para BERT. Além das camadas de saída, as mesmas arquiteturas são usadas tanto no pré-treinamento quanto no fine-tuning. Os mesmos parâmetros dos modelos pré-treinados são utilizados para inicializar modelos para diferentes tarefas específicas. Durante o *fine-tuning*, todos os parâmetros são ajustados.



Fonte: (DEVLIN et al., 2018).

cesso de *fine-tuning* é eficiente e econômico, diminuindo a necessidade de grandes quantidades de dados, tempo e recursos computacionais. Esse processo é ilustrado na Figura 8.

3.4 CONSIDERAÇÕES FINAIS

Neste capítulo é apresentada a metodologia adotada para a tarefa de classificação de receitas dietéticas. A base de dados Food.com, rica nas mais diversas informações sobre as receitas culinárias, foi analisada, e algumas das suas principais categorias dietéticas foram selecionadas para a análise experimental proposta neste trabalho (dibética, *gluten-free*, *kid-friendly*, *lactose-free* e vegana). As técnicas para extração de características, assim como os classificadores adotados, são brevemente descritos. No próximo capítulo (Capítulo 4), tanto as configurações dos modelos selecionados, quanto os resultados experimentais obtidos serão apresentados e discutidos.

4 ANÁLISE EXPERIMENTAL

A avaliação experimental deste trabalho tem como objetivo comparar a performance de diversos modelos de classificação na identificação automática de receitas dietéticas. Conjuntos de dados de entrada constituídos por diferentes componentes textuais das receitas também são comparados, visando verificar quais seriam as informações que mais contribuiriam para o bom desempenho dos classificadores. Na próxima seção (Seção 4.1), detalharemos a configuração adotada nos experimentos. Na Seção 4.2, os resultados experimentais são apresentados e, em seguida (Seção 4.3), discutidos.

4.1 CONFIGURAÇÃO EXPERIMENTAL

Nesta seção, os detalhes sobre a configuração dos experimentos serão apresentados. Na sequência (Seção 4.1.1), serão apresentadas as métricas de classificação empregadas na avaliação e comparação dos classificadores. Em seguida, o processo de reamostragem por meio da abordagem de validação cruzada é brevemente descrito (Seção 4.1.2). Por fim, as ferramentas utilizadas para implementação dos classificadores, assim como os hiper-parâmetros selecionados para a configuração desses classificadores são detalhados (Seção 4.1.3).

4.1.1 Métricas

Para conseguir comparar os diferentes classificadores, é necessário mensurar a performance de cada um deles. Para isso, algumas das métricas mais comuns para problemas de classificação são adotadas. Essas métricas são descritas a seguir.

- Acurácia – Proporção de documentos classificados corretamente.

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

- Precisão – Proporção de documentos classificados como positivos que realmente são positivos.

$$Precisão = \frac{TP}{TP + FP} \tag{4.2}$$

- Revocação – Proporção de documentos positivos que foram classificados corretamente.

$$Revocação = \frac{TP}{TP + FN} \tag{4.3}$$

- F-*measure* (F-1) – Balanceia a precisão e revocação em uma única métrica.

$$F - Measure = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação} \tag{4.4}$$

Essas métricas são baseadas na matriz de confusão, e usam os números de verdadeiros positivos (*true positives* - TP), falsos positivos (*false positives* – FP), falsos negativos (*false negatives* – FN) e verdadeiros negativos (*true negative* – TN).

4.1.2 Validação Cruzada

Para evitar resultados obtidos por acaso, nos experimentos é empregada uma abordagem de validação cruzada com 5-*folds*. Nessa abordagem, o conjunto de dados é dividido aleatoriamente em cinco partes (*folds*) balanceadas para formar o conjunto de treinamento e o conjunto de teste. Quatro partes são usadas, a cada iteração da experimentação, para constituir o conjunto de treinamento, e a parte restante compõe o conjunto de teste. Os resultados finais são gerados pelo cálculo da média dos resultados obtidos para cada um dos *folds*.

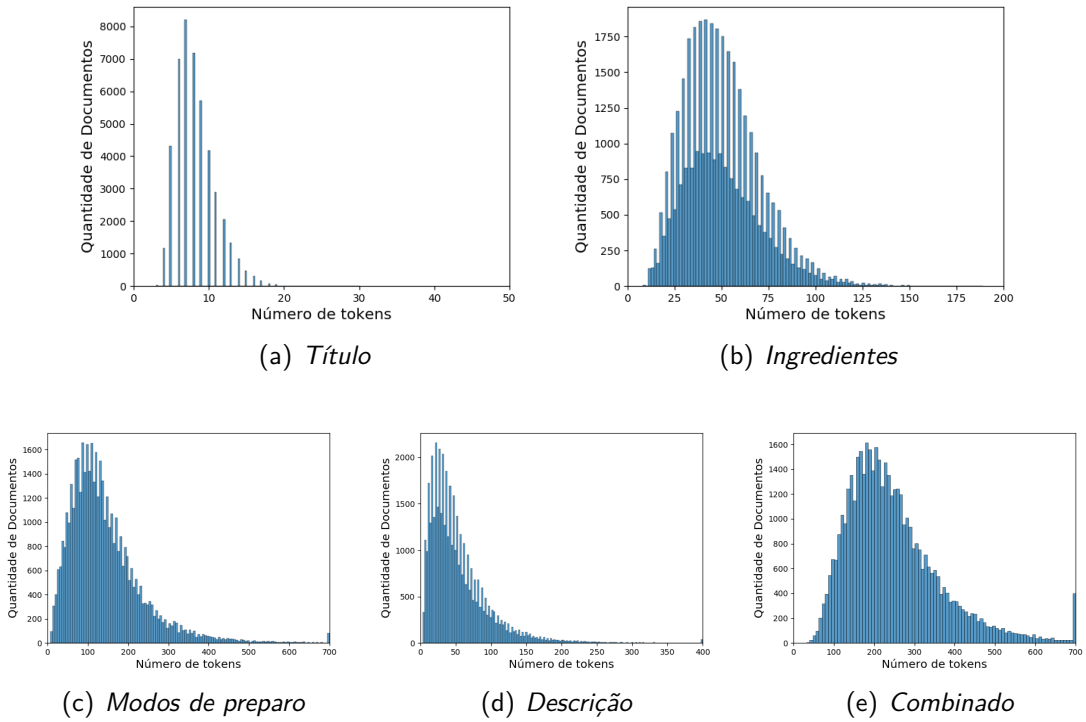
4.1.3 Implementação e Hiper-parâmetros

Os classificadores de Aprendizagem Rasa foram implementados com o auxílio da biblioteca para Aprendizagem de Máquina *Scikit-Learn* (PEDREGOSA et al., 2011), da linguagem de programação Python. Os Hiper-parâmetros utilizados na avaliação foram selecionados do trabalho proposto por (ROITHER; KURZ; SONNLEITNER, 2022).

Já para a implementação e ajuste do modelo BERT, o *framework Transformers* (HuggingFace) (WOLF et al., 2020) foi empregado. A estratégia e hiper-parâmetros usados por (MOHAMMADI et al., 2020b) para o *fine-tuning* do BERT foram selecionados para este trabalho. A Entropia Cruzada é adotada como função de perda. O modelo é ajustado por quatro épocas, e os parâmetros do modelo são salvos em cada uma das épocas. O modelo final escolhido é aquele para o qual o menor erro no conjunto de validação é obtido. Como a base de dados adotada é completamente em inglês, o modelo base do BERT é adotado (*bert-base-uncased model*[1]). O otimizador AdamW é utilizado, com uma taxa de decaimento de peso (*weight_decay*) de 0.01, e uma taxa de aprendizagem (*learning_rate*) de 10^{-5}. O tamanho do lote (*batch_size*) é definido como 16.

[1] <https://huggingface.co/bert-base-uncased>

Figura 9 – Distribuição do tamanho dos documentos.



(a) *Título* (b) *Ingredientes*

(c) *Modos de preparo* (d) *Descrição* (e) *Combinado*

Fonte: A autora (2023).

O BERT funciona com um tamanho fixo de sequência como entrada. Para definir esse parâmetro, analisamos a distribuição da frequência do tamanho dos documentos, essa distribuição é apresentada na Figura 9. Sequências grandes tornam o processo de ajuste caro computacionalmente, por isso, optou-se pela escolha do menor tamanho de sequência que fosse capaz capturar o maior número de documentos completos. Os seguintes tamanhos foram selecionados: *título* - 20 tokens, *ingredientes* - 100 tokens, *descrição* - 200 tokens, *modos de preparo* - 300 tokens e *combinado* - 400 tokens.

Tabela 6 – Hiper-parâmetros adotados para os modelos de classificação.

Modelo	Hiper-parâmetros	Descrição	Valor
Árvore de Decisão	max_depth	Profundidade máxima da árvore.	120
	max_features	O número de características a serem considerados ao procurar a melhor divisão.	auto
	min_samples_leaf	O número mínimo de amostras necessárias em um nó folha.	1
Floresta Aleatória	class_weight	Peso associado com as classes.	balanced
	max_depth	Profundidade máxima das árvores.	75
	max_features	O número de características a serem considerados ao procurar a melhor divisão.	auto
	n_estimators	Número de árvores na floresta.	100
K-NN	n_neighbours	Número de vizinhos a serem considerados.	75
Naive Bayes	alpha	Parâmetro de suavização.	0.01
Perceptron Multicamadas	activation	Função de ativação da camada escondida.	relu
	early_stopping	Encerra o treinamento quando não houver melhora nos resultados obtidos no conjunto de validação.	True
	hidden_layer_sizes	Tamanho das camadas escondidas.	(130,)
	learning_rate	Taxa de aprendizagem do modelo.	constant
	max_iter	Número máximo de iterações	300
Regressão Logística	C	Inverso da regularização.	0.5
	max_iter	Número máximo de iterações.	1000
	multi_class	Define se o problema tratado é binário ou multiclasse.	auto
	solver	Algoritmo a utilizar para optimização.	lbfgs
SVM	C	Inverso da regularização.	10
	kernel	Especifíca o tipo de *kernel* utilizado pelo algoritmo.	rbf
	gamma	Coeficiente do *kernel*.	0.001
BERT	batch_size	Tamanho do batch.	16
	epochs	Quantidade de épocas.	4
	learning_rate	Taxa de aprendizagem do modelo.	10^{-5}

Fonte: A autora (2023).

4.2 RESULTADOS

Os resultados são apresentados da Tabela 7 à 11. O tempo de execução é medido em **segundos** (s).

O tempo de execução dos modelos para os conjuntos de dados combinados podem ser melhor visualizados através dos gráficos exibidos na figura 10.

A avaliação experimental inclui uma análise empírica nos resultados obtidos no conjunto de teste e o tempo médio de execução de cada algoritmo. Além disso, a avaliação também inclui um sistema de *rank* gerado através da aplicação do teste de Friedman (FRIEDMAN, 1937) para os valores médios das métricas de teste. O teste de Friendman é um teste de hipóteses não paramétrico que ranqueia todos os algoritmos para cada conjunto de dados separadamente. Se a hipótese nula de que todos os *ranks* não são significativamente diferentes é rejeitada, o teste de Nemenyi (NEMENYI, 1963) é adotado como um teste *post-hoc*. De acordo com o teste de Nemenyi, o desempenho de dois algoritmos é considerado significativamente diferentes se os *ranks* médios correspondentes diferirem em pelo menos a Diferença Crítica (*Critical Difference* - Diferença Crítica (CD)), cuja a equação pode ser vista a seguir (Equação 4.5):

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6D}} \tag{4.5}$$

onde k representa o número de algoritmos, D representa o número de conjuntos de dados e q_α são valores críticos baseados numa distribuição t-Student dividido por $\sqrt{2}$ (DEMŠAR, 2006).

Para execução da análise estatística foi adotada a biblioteca de Python AutoRank (HERBOLD, 2020). Os testes foram realizados a um nível de significância de 0.05.

Os p-*values* obtidos através do teste são apresentados na Figura 4.2 e os *ranks* gerados são exibidos na Figura 4.2.

–

Fonte: A autora (2023).

Tabela 7 – Resultados experimentais para a identificação de receitas na dieta *diabético*.

Conjunto de Dados	Acurácia	Precisão	Revocação	F-1	T. de Treino (s)	T. de Teste (s)
Árvore de Decisão						
Combinado	0.5878	0.5871	0.5945	0.5907	1.1922	0.2520
Descrição	0.5777	0.5971	0.5112	0.5434	0.4446	0.0750
Ingredientes	**0.6150**	0.6175	**0.6052**	**0.6112**	0.2267	0.0363
Preparo	0.5513	0.5512	0.5532	0.5521	0.7633	0.1529
Título	0.6074	**0.6488**	0.4687	0.5437	**0.1266**	**0.0184**
Floresta Aleatória						
Combinado	**0.7368**	**0.7255**	**0.7618**	**0.7432**	11.2603	0.3695
Descrição	0.6805	0.7137	0.6026	0.6534	5.6377	0.1852
Ingredientes	0.7151	0.7011	0.7497	0.7246	5.4323	0.1248
Preparo	0.6653	0.6574	0.6910	0.6737	9.6427	0.2535
Título	0.6688	0.6973	0.5968	0.6430	**2.4301**	**0.1064**
K-NN						
Combinado	0.6879	**0.6848**	0.6962	0.6904	1.0367	2.2281
Descrição	0.6552	0.6846	0.5764	0.6252	0.3291	1.2255
Ingredientes	**0.7030**	0.6800	**0.7672**	**0.7209**	0.1456	0.9809
Preparo	0.6452	0.6505	0.6278	0.6389	0.6226	1.7186
Título	0.6587	0.6407	0.7233	0.6794	**0.0727**	**0.4754**
Multi-Layer Perceptron						
Combinado	**0.7426**	**0.7338**	**0.7621**	**0.7474**	43.5428	0.2930
Descrição	0.6838	0.6960	0.6528	0.6736	29.9859	0.1144
Ingredientes	0.7208	0.7141	0.7371	0.7253	**5.1452**	0.0576
Preparo	0.6782	0.6801	0.6738	0.6766	26.0082	0.1995
Título	0.6766	0.6787	0.6706	0.6744	10.2459	**0.0269**
Naive Bayes						
Combinado	**0.7175**	**0.6991**	0.7640	**0.7301**	1.0560	0.2430
Descrição	0.6664	0.6706	0.6540	0.6621	0.3332	0.0731
Ingredientes	0.7052	0.6817	**0.7701**	0.7231	0.1483	0.0349
Preparo	0.6554	0.6456	0.6889	0.6665	0.6228	0.1486
Título	0.6537	0.6449	0.6851	0.6642	**0.0724**	**0.0169**
Regressão Logística						
Combinado	**0.7493**	**0.7485**	**0.7510**	**0.7497**	1.3588	0.2824
Descrição	0.6855	0.7019	0.6450	0.6722	0.4851	0.1087
Ingredientes	0.7274	0.7237	0.7357	0.7296	0.2283	0.0379
Preparo	0.6851	0.6816	0.6951	0.6882	0.8624	0.1879
Título	0.6820	0.6910	0.6585	0.6743	**0.1340**	**0.0169**
SVM						
Combinado	0.6819	0.6528	**0.7771**	0.7095	100.5312	23.7973
Descrição	0.5955	**0.8631**	0.2272	0.3595	36.6688	8.8926
Ingredientes	**0.6986**	0.6808	0.7479	**0.7128**	16.8542	3.6163
Preparo	0.6253	0.5986	0.7610	0.6700	66.8807	15.9415
Título	0.6092	0.7750	0.3090	0.4411	**6.4480**	**1.2974**
BERT						
Combinado	**0.7631**	**0.7667**	**0.7562**	**0.7611**	1277.0563	31.9177
Descrição	0.7013	0.7121	0.6760	0.6935	634.6675	15.2131
Ingredientes	0.6720	0.5687	0.5811	0.5743	349.4664	7.9047
Preparo	0.6854	0.6875	0.6805	0.6836	938.4942	23.1413
Título	0.6679	0.6729	0.6566	0.6637	**139.0983**	**2.4562**

Fonte: A autora (2023).

Tabela 8 – Resultados experimentais para a identificação de receitas na dieta *gluten-free*.

Conjunto de Dados	Acurácia	Precisão	Revocação	F-1	T. de Treino (s)	T. de Teste (s)
Árvore de Decisão						
Combinado	0.5849	0.5843	0.5912	0.5877	1.1244	0.2323
Descrição	0.5638	0.5631	0.5840	0.5710	0.4360	0.0722
Ingredientes	**0.6277**	0.6286	**0.6241**	**0.6262**	0.2107	0.0329
Preparo	0.5549	0.5541	0.5602	0.5569	0.7227	0.1400
Título	0.6087	**0.6594**	0.4477	0.5321	**0.1179**	**0.0172**
Floresta Aleatória						
Combinado	0.7284	**0.7174**	0.7537	0.7351	11.3652	0.4212
Descrição	0.6447	0.6717	0.5660	0.6143	6.1241	0.1798
Ingredientes	**0.7310**	0.7033	**0.7993**	**0.7482**	5.0371	0.1231
Preparo	0.6559	0.6454	0.6924	0.6680	9.7852	0.2599
Título	0.6708	0.7162	0.5660	0.6322	**2.4186**	**0.1072**
K-NN						
Combinado	0.6931	**0.6895**	0.7028	0.6961	0.9586	2.1189
Descrição	0.6360	0.6687	0.5400	0.5970	0.3255	1.1456
Ingredientes	**0.7055**	0.6889	**0.7497**	**0.7180**	0.1264	0.9224
Preparo	0.6359	0.6432	0.6103	0.6262	0.6113	1.6610
Título	0.6665	0.6446	0.7429	0.6901	**0.1077**	**0.5732**
Multi-Layer Perceptron						
Combinado	**0.7363**	**0.7285**	**0.7542**	**0.7407**	60.1868	0.2982
Descrição	0.6525	0.6604	0.6291	0.6435	35.8365	0.1326
Ingredientes	0.7291	0.7203	0.7501	0.7346	**6.8195**	0.0671
Preparo	0.6687	0.6585	0.7016	0.6790	29.4204	0.2033
Título	0.6769	0.6796	0.6727	0.6755	13.7467	**0.0294**
Naive Bayes						
Combinado	**0.7147**	**0.7020**	0.7461	0.7233	0.9600	0.2268
Descrição	0.6396	0.6442	0.6237	0.6337	0.3307	0.0723
Ingredientes	0.7143	0.6952	**0.7632**	**0.7276**	0.1346	0.0344
Preparo	0.6518	0.6433	0.6816	0.6619	0.5835	0.1380
Título	0.6699	0.6605	0.6994	0.6793	**0.0699**	**0.0149**
Regressão Logística						
Combinado	**0.7484**	**0.7419**	0.7618	**0.7517**	1.2851	0.2801
Descrição	0.6591	0.6734	0.6181	0.6445	0.5262	0.1245
Ingredientes	0.7312	0.7171	**0.7638**	0.7397	0.2244	0.0332
Preparo	0.6712	0.6608	0.7038	0.6816	0.7640	0.1368
Título	0.6891	0.7052	0.6498	0.6763	**0.1440**	**0.0177**
SVM						
Combinado	0.6868	0.6590	**0.7748**	**0.7121**	88.3131	23.4008
Descrição	0.5562	0.9670	0.1163	0.2075	33.3803	8.5652
Ingredientes	**0.6872**	0.6694	0.7400	0.7029	14.9379	3.6170
Preparo	0.6146	0.5892	0.7580	0.6629	58.6042	15.3365
Título	0.5840	**0.9792**	0.1718	0.2920	**5.8865**	**1.2218**
BERT						
Combinado	**0.7634**	**0.7607**	0.7705	**0.7650**	1122.7957	28.0266
Descrição	0.6675	0.6739	0.6500	0.6615	559.2981	13.4236
Ingredientes	0.7393	0.7242	**0.7739**	0.7476	307.9414	7.1246
Preparo	0.6769	0.6701	0.6974	0.6832	829.6029	20.4584
Título	0.6745	0.7281	0.5893	0.6274	**122.1692**	**2.1233**

Fonte: A autora (2023).

Tabela 9 – Resultados experimentais para a identificação de receitas na dieta *kid-friendly*.

Conjunto de Dados	Acurácia	Precisão	Revocação	F-1	T. de Treino (s)	T. de Teste (s)
Árvore de Decisão						
Combinado	0.6490	0.6472	0.6553	0.6511	4.1873	0.8049
Descrição	0.6083	0.5933	0.6907	0.6379	1.7504	0.2447
Ingredientes	**0.7054**	**0.7027**	0.7124	**0.7074**	0.9046	0.1112
Preparo	0.6258	0.6265	0.6234	0.6249	2.7893	0.4791
Título	0.6485	0.6113	**0.8156**	0.6988	**0.4854**	**0.0529**
Floresta Aleatória						
Combinado	0.7959	0.7943	0.7994	0.7967	70.0950	1.3809
Descrição	0.7309	0.7217	0.7519	0.7364	31.9829	0.6071
Ingredientes	**0.8025**	**0.8087**	0.7930	**0.8007**	30.4311	0.4052
Preparo	0.7470	0.7460	0.7505	0.7480	58.7203	0.9035
Título	0.7261	0.6875	**0.8329**	0.7529	**11.1698**	**0.3290**
K-NN						
Combinado	0.7704	0.7358	**0.8443**	0.7863	3.3455	22.9631
Descrição	0.7216	0.7294	0.7053	0.7169	1.0594	11.9888
Ingredientes	**0.7881**	**0.7729**	0.8167	**0.7941**	0.4584	10.2547
Preparo	0.7416	0.7125	0.8110	0.7584	1.9663	17.8838
Título	0.7203	0.6995	0.7727	0.7343	**0.2292**	**4.5528**
Multi-Layer Perceptron						
Combinado	**0.8303**	**0.8318**	**0.8286**	**0.8300**	231.3739	0.9076
Descrição	0.7558	0.7479	0.7721	0.7597	180.6412	0.3220
Ingredientes	0.8109	0.8111	0.8117	0.8110	**41.1892**	0.1742
Preparo	0.7753	0.7751	0.7759	0.7755	126.5196	0.5663
Título	0.7363	0.7051	0.8132	0.7552	64.1658	**0.0944**
Naive Bayes						
Combinado	**0.7860**	0.7843	**0.7896**	**0.7868**	3.2629	0.7906
Descrição	0.7334	0.7267	0.7483	0.7374	1.0535	0.2417
Ingredientes	0.7846	**0.7847**	0.7853	0.7848	0.6913	0.1450
Preparo	0.7420	0.7467	0.7339	0.7400	1.9693	0.4695
Título	0.7194	0.6989	0.7716	0.7334	**0.2307**	**0.0526**
Regressão Logística						
Combinado	**0.8279**	**0.8239**	**0.8347**	**0.8291**	4.6550	0.8405
Descrição	0.7548	0.7517	0.7612	0.7564	1.8079	0.2945
Ingredientes	0.8116	0.8143	0.8081	0.8110	0.8285	0.1072
Preparo	0.7753	0.7757	0.7755	0.7755	2.9222	0.5210
Título	0.7420	0.7103	0.8190	0.7606	**0.4791**	**0.0487**
SVM						
Combinado	**0.7987**	0.7981	0.8009	**0.7993**	849.1798	209.2173
Descrição	0.7137	0.7539	0.6349	0.6892	406.5087	102.9616
Ingredientes	0.7947	**0.7996**	0.7873	0.7932	152.0769	31.8943
Preparo	0.7483	0.7594	0.7282	0.7432	635.3008	156.6352
Título	0.7046	0.6562	**0.8658**	0.7461	**67.5419**	**13.1479**
BERT						
Combinado	0.7755	0.7469	**0.8493**	**0.7924**	3906.7608	97.2247
Descrição	0.7616	0.7543	0.7766	0.7651	2155.3681	46.6421
Ingredientes	0.7370	0.7299	0.8287	0.7649	1072.1410	23.7631
Preparo	**0.7794**	**0.7679**	0.8027	0.7844	2887.4041	71.0769
Título	0.7286	0.7069	0.7840	0.7430	**430.1238**	**7.1419**

Fonte: A autora (2023).

Tabela 10 – Resultados experimentais para a identificação de receitas na dieta *lactose-free*.

Conjunto de Dados	Acurácia	Precisão	Revocação	F-1	T. de Treino (s)	T. de Teste (s)
Árvore de Decisão						
Combinado	0.5775	0.5774	0.5785	0.5779	0.8102	0.1684
Descrição	0.5543	0.5612	0.5023	0.5294	0.3146	0.0564
Ingredientes	**0.6252**	**0.6229**	**0.6346**	**0.6285**	0.1468	0.0249
Preparo	0.5603	0.5598	0.5645	0.5621	0.5064	0.1018
Título	0.5850	0.6165	0.4524	0.5214	**0.0837**	**0.0132**
Floresta Aleatória						
Combinado	0.7075	0.6834	0.7737	0.7257	7.5121	0.2736
Descrição	0.6353	0.6386	0.6233	0.6308	5.0294	0.1316
Ingredientes	**0.7245**	**0.6855**	**0.8302**	**0.7508**	3.7117	0.0928
Preparo	0.6687	0.6516	0.7250	0.6863	6.4999	0.1877
Título	0.6359	0.6490	0.5915	0.6189	**1.8552**	**0.0768**
K-NN						
Combinado	0.6810	0.6419	0.8194	0.7198	0.7181	1.2054
Descrição	0.6304	0.6159	0.6943	0.6524	0.2431	0.6608
Ingredientes	**0.7057**	**0.6680**	**0.8196**	**0.7359**	0.0957	0.4977
Preparo	0.6432	0.6171	0.7545	0.6789	0.4654	1.0610
Título	0.6159	0.5900	0.7599	0.6641	**0.0478**	**0.2701**
Multi-Layer Perceptron						
Combinado	**0.7220**	**0.7039**	**0.7677**	**0.7341**	33.5110	0.2347
Descrição	0.6443	0.6442	0.6473	0.6448	20.8555	0.1027
Ingredientes	0.7095	0.6936	0.7518	0.7212	**5.0235**	0.0483
Preparo	0.6746	0.6606	0.7186	0.6880	17.1233	0.1647
Título	0.6373	0.6296	0.6726	0.6487	6.3476	**0.0222**
Naive Bayes						
Combinado	0.6907	0.6723	0.7444	0.7065	0.7220	0.1690
Descrição	0.6126	0.6061	0.6429	0.6239	0.2470	0.0531
Ingredientes	**0.6967**	**0.6728**	**0.7663**	**0.7164**	0.1014	0.0230
Preparo	0.6504	0.6382	0.6945	0.6652	0.4326	0.1002
Título	0.6256	0.6154	0.6704	0.6416	**0.0525**	**0.0109**
Regressão Logística						
Combinado	**0.7314**	**0.7030**	0.8017	**0.7491**	1.0221	0.2298
Descrição	0.6530	0.6519	0.6564	0.6541	0.3529	0.0950
Ingredientes	0.7249	0.6926	**0.8095**	0.7464	0.1539	0.0244
Preparo	0.6824	0.6618	0.7459	0.7013	0.5456	0.1055
Título	0.6466	0.6472	0.6446	0.6458	**0.1003**	**0.0108**
SVM						
Combinado	0.6844	0.6316	0.8855	0.7372	47.1222	12.5221
Descrição	0.5892	0.6260	0.5803	0.5535	16.0322	4.3489
Ingredientes	**0.7050**	0.6495	**0.8906**	**0.7511**	7.8053	1.9500
Preparo	0.6468	0.6131	0.7965	0.6927	29.5828	7.8432
Título	0.5423	**0.7723**	0.1212	0.2092	**3.2757**	**0.6915**
BERT						
Combinado	**0.7292**	**0.7071**	**0.7857**	**0.7436**	824.5138	20.7212
Descrição	0.6393	0.6279	0.6862	0.6550	411.6669	9.9512
Ingredientes	0.6084	0.5100	0.6032	0.5330	226.9288	5.2148
Preparo	0.6842	0.6618	0.7542	0.7044	608.9783	14.9484
Título	0.6202	0.6672	0.5781	0.5715	**90.3740**	**1.6712**

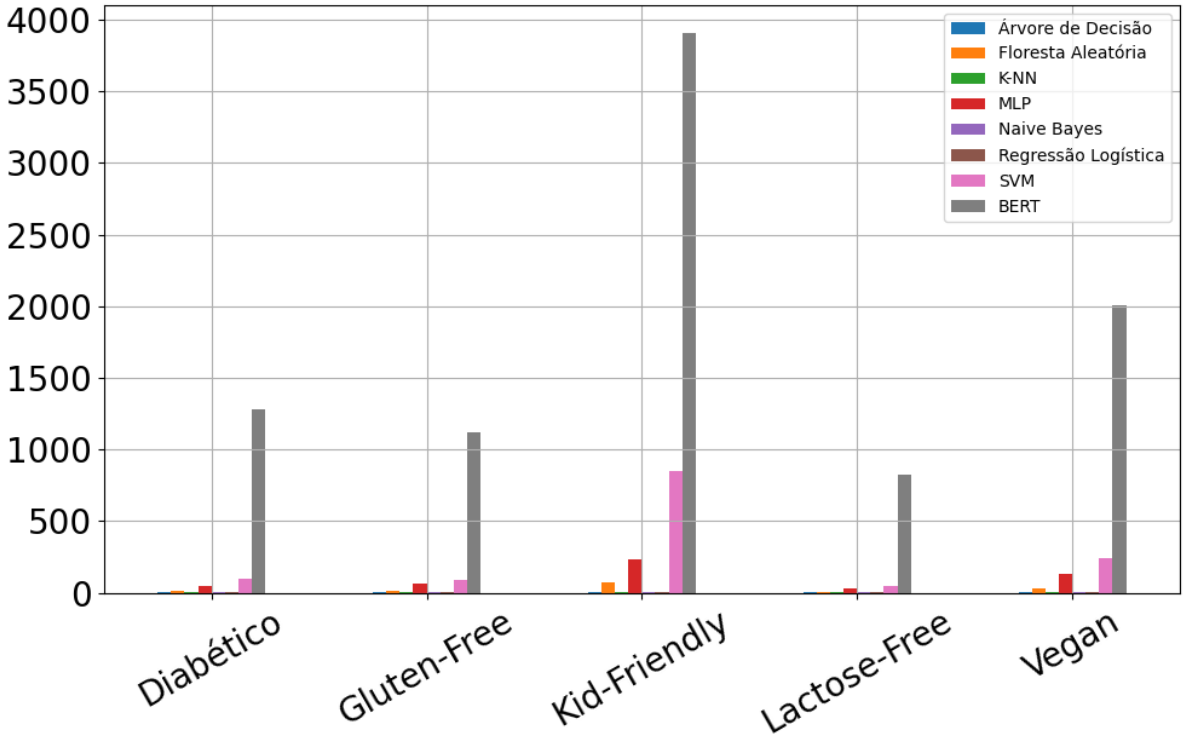
Fonte: A autora (2023).

Tabela 11 – Resultados experimentais para a identificação de receitas na dieta *vegana*.

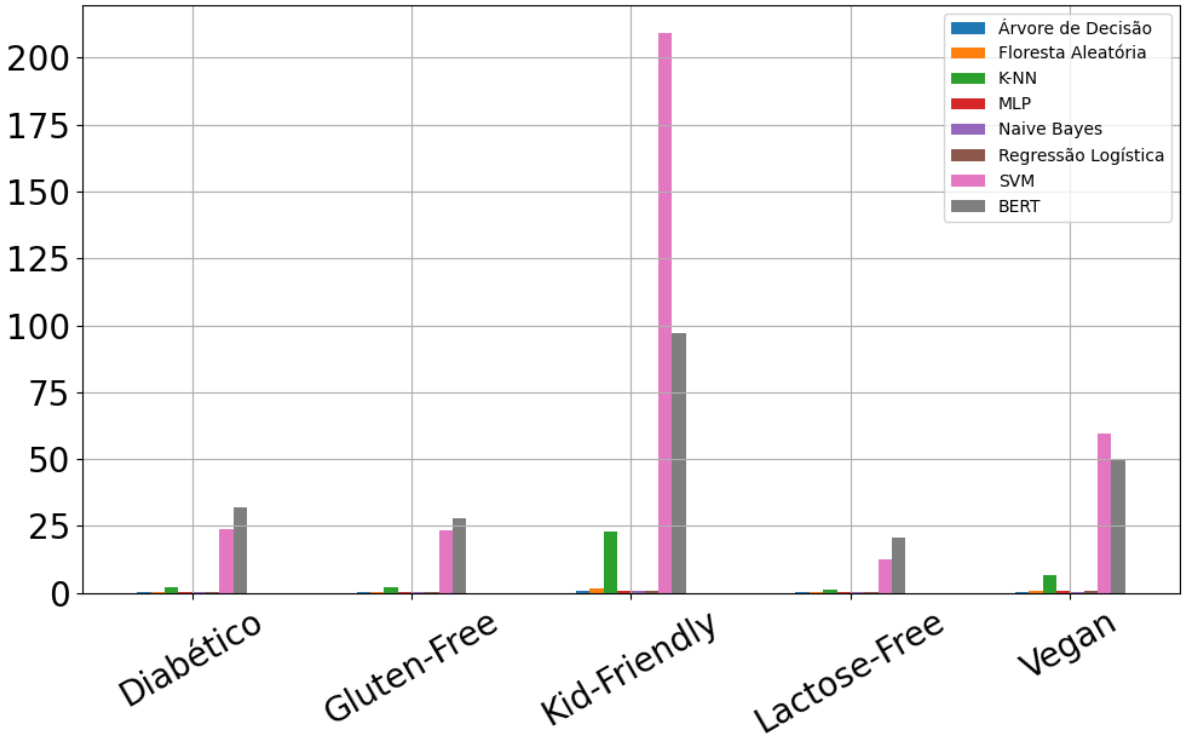
Conjunto de Dados	Acurácia	Precisão	Revocação	F-1	T. de Treino (s)	T. de Teste (s)
Árvore de Decisão						
Combinado	0.6751	0.6735	0.6798	0.6766	2.1315	0.4342
Descrição	0.5998	0.6169	0.5299	0.5693	0.8708	0.1367
Ingredientes	**0.8026**	**0.7949**	**0.8160**	**0.8053**	0.3785	0.0605
Preparo	0.6667	0.6651	0.6716	0.6683	1.3314	0.2544
Título	0.6935	0.7205	0.6822	0.6828	**0.2249**	**0.0295**
Floresta Aleatória						
Combinado	0.8579	0.8321	0.8967	0.8632	25.7718	0.7047
Descrição	0.7299	0.7407	0.7074	0.7237	13.3982	0.3231
Ingredientes	**0.8971**	**0.8720**	**0.9310**	**0.9005**	9.3958	0.2022
Preparo	0.8151	0.7954	0.8485	0.8211	21.0594	0.4460
Título	0.7945	0.8172	0.7587	0.7868	**4.9258**	**0.1777**
K-NN						
Combinado	0.8341	**0.8127**	0.8684	0.8396	1.9360	6.4671
Descrição	0.7151	0.6837	0.8036	0.7382	0.6065	3.5203
Ingredientes	**0.8362**	0.7930	**0.9101**	**0.8475**	0.2676	3.1980
Preparo	0.7969	0.7661	0.8547	0.8079	1.0394	4.9071
Título	0.7686	0.7475	0.8120	0.7782	**0.1224**	**1.4108**
Multi-Layer Perceptron						
Combinado	0.8937	0.8780	0.9146	0.8959	133.8582	0.5056
Descrição	0.7501	0.7508	0.7498	0.7499	90.2833	0.1982
Ingredientes	**0.9061**	**0.8858**	**0.9324**	**0.9085**	**16.9010**	0.1116
Preparo	0.8449	0.8260	0.8741	0.8493	67.2291	0.3243
Título	0.8174	0.8157	0.8201	0.8179	30.2796	**0.0500**
Naive Bayes						
Combinado	0.8403	0.8320	0.8528	0.8422	1.7737	0.4220
Descrição	0.7245	0.7254	0.7226	0.7239	0.6507	0.1306
Ingredientes	**0.8670**	**0.8600**	**0.8769**	**0.8683**	0.2487	0.0614
Preparo	0.8036	0.7913	0.8246	0.8076	1.0535	0.2460
Título	0.7872	0.7848	0.7916	0.7881	**0.1312**	**0.0275**
Regressão Logística						
Combinado	0.8928	0.8677	0.9269	0.8963	2.4669	0.5032
Descrição	0.7529	0.7590	0.7415	0.7501	0.8192	0.1740
Ingredientes	**0.9007**	**0.8700**	**0.9423**	**0.9047**	0.3743	0.0543
Preparo	0.8471	0.8189	0.8913	0.8535	1.3307	0.2857
Título	0.8212	0.8175	0.8271	0.8223	**0.2328**	**0.0244**
SVM						
Combinado	0.8477	0.8048	0.9182	0.8577	241.5664	59.5620
Descrição	0.7123	0.7110	0.7157	0.7131	102.8884	26.7817
Ingredientes	**0.8659**	**0.8156**	**0.9457**	**0.8759**	34.4010	7.6886
Preparo	0.7948	0.7514	0.8811	0.8111	169.3436	42.2841
Título	0.7668	0.8281	0.6734	0.7427	**17.7069**	**3.5917**
BERT						
Combinado	**0.9142**	**0.9048**	0.9258	**0.9152**	2005.9872	50.1983
Descrição	0.7695	0.7691	0.7706	0.7697	994.2128	23.7470
Ingredientes	0.9097	0.8964	**0.9265**	0.9112	547.6022	12.0959
Preparo	0.8665	0.8547	0.8829	0.8686	1471.3744	36.1333
Título	0.8276	0.8223	0.8358	0.8289	**218.7500**	**3.7295**

Fonte: A autora (2023).

Figura 10 – Tempo de execução para os maiores conjuntos de dados (*combinado*)



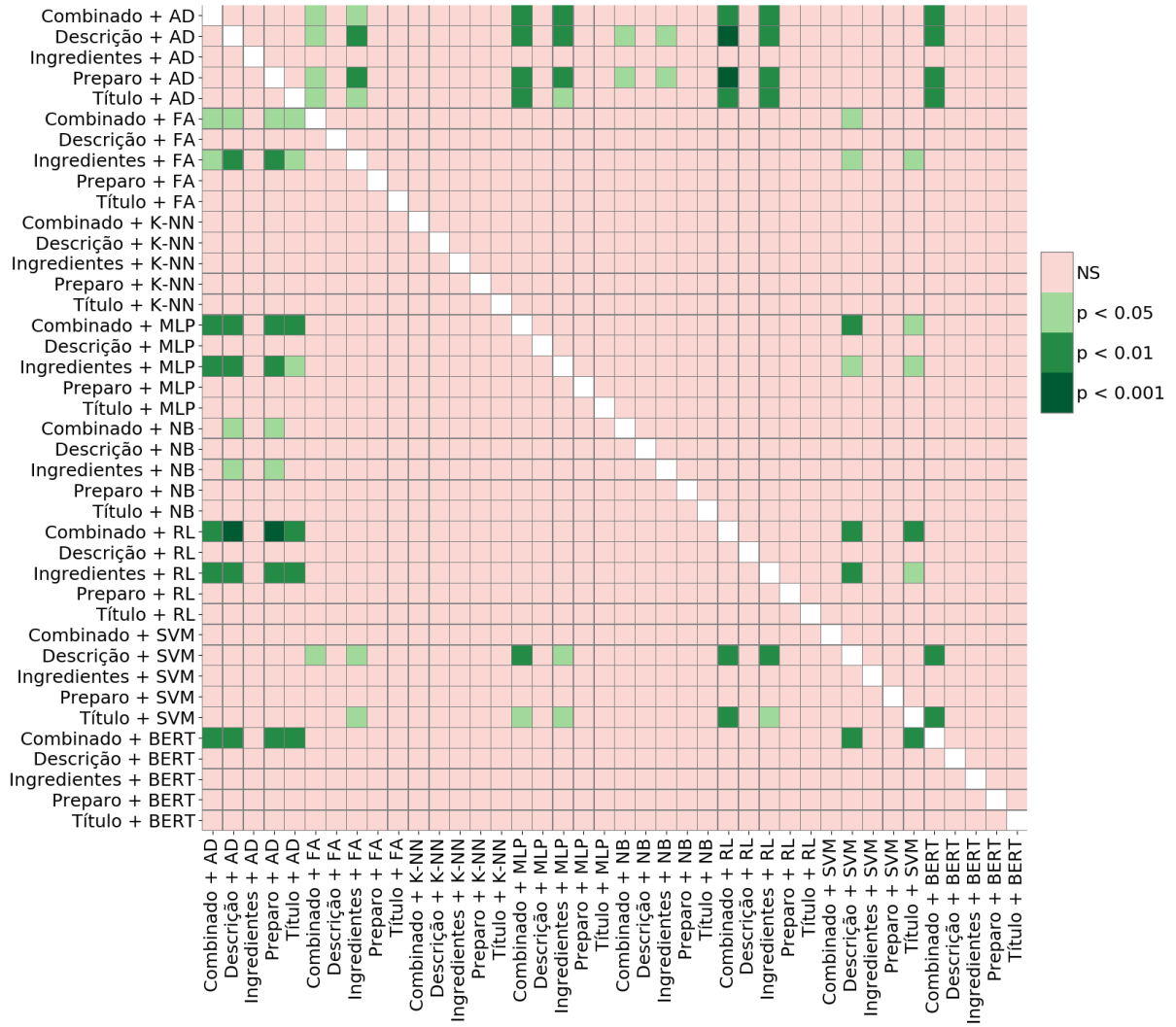
(a) *Tempo de Treino*



(b) *Tempo de Teste*

Fonte: A autora (2023).

Figura 11 – *P-Values* do Teste de Friedman-Nemenyi.



Fonte: A autora (2023).

4.3 DISCUSSÃO

Nas próximas seções serão discutidos os resultados obtidos na Seção 4.2. Na Seção 4.3.1, são analisadas as performances de cada um dos classificadores adotados empiricamente. Em seguida (Seção 4.3.2), o impacto dos diferentes conjuntos de dados adotados como entrada dos classificadores são explorados. Avaliamos o tempo de execução dos classificadores na Seção 4.3.3. Por fim, analisamos estatisticamente os resultados dos modelos testados.

4.3.1 Performance dos Classificadores

Como pode ser verificado através das Tabelas 7 - 11, a maioria dos algoritmos foi capaz de realizar a identificação das receitas dietéticas de forma satisfatória. É possível observar um padrão no desempenho desses classificadores, no qual alguns modelos foram capazes de alcançar bons resultados em todas as restrições, como o BERT, que teve o melhor desempenho nas dietas *diabético*, *gluten-free* (73% de acurácia em ambas as dietas), e *vegano* (com 91% de acurácia); o MLP, com o melhor resultado na dieta *kid-friendly* (com 83% de acurácia); e a Regressão Logística, com o melhor desempenho na dieta *lactose-free* (com 73% de acurácia).

O modelo Árvore de Decisão demonstrou não ser uma das escolhas mais adequadas para a tarefa de identificação de receitas dietéticas, tendo um desempenho consideravelmente inferior aos demais métodos em todos os cenários testados, alcançando, no pior caso (dieta *diabética*), uma diferença de 15 pontos percentuais para o melhor classificador (BERT). Os classificadores K-NN e Naive Bayes também obtiveram resultados relativamente inferiores, porém mais próximos, percentualmente, aos dos classificadores com melhores performances. Considerando a métrica **acurácia**, o classificador Floresta Aleatória não foi o melhor modelo em nenhuma das dietas avaliadas. No entanto, de modo geral, esse algoritmo se destacou entre os classificadores, tendo resultados sempre próximos aos modelos com melhor desempenho.

Enquanto a grande parte dos modelos alcançaram resultados similares entre todas as métricas de classificação, o classificador SVM demonstrou dificuldade no processo de aprendizagem ao apresentar resultados discrepantes entre as métricas adotadas. Esse problema pode ser observado, por exemplo, na classificação da dieta *gluten-free* a partir das características do título das receitas, situação na qual o classificador obteve a maior precisão (98%), porém com apenas 17% de revocação. Embora este seja um problema a ser investigado com maior profundidade, no contexto da aplicação atual (de identificação de receitas que atendam às restrições alimentares dos usuários), uma taxa de revocação relativamente baixa pode implicar que exemplos positivos foram classificados como negativos (alta taxa de FN), o que ainda garantiria que as recomendações realizadas seriam seguras aos usuários, tendo em vista que o sistema tenderia a não recomendar receitas que aparentassem não garantir as condições da dieta em observação.

4.3.2 Conjuntos de Dados

Os conjuntos de dados usados para treinamento influenciam diretamente a performance dos classificadores. Os modelos que adotaram características da lista de ingredientes, individualmente, e de todas as informações combinadas obtiveram os melhores desempenhos, enquanto as características dos títulos e dos textos descritivos, quando usadas isoladamente, mostraram não terem poder discriminatório suficiente para o treinamento dos classificadores, o que resultou em desempenhos ruins. O conjunto de dados mais adequado para classificação também varia de dieta para dieta. A lista de ingredientes é responsável pelos melhores resultados em dietas livres de alérgenos (*lactose-free* e *gluten-free*), ou de certos alimentos restritivos (veganos). Nesses casos, apenas a presença de determinado ingrediente já é o suficiente para infringir a restrição, independente da quantidade ou da forma de preparo. Enquanto para as demais dietas, outros fatores além dos ingredientes são decisivos, como a quantidade e ações realizadas nos ingredientes, como para o caso das dietas diabéticos e *kid-friendly*, o que justificaria o fato dos melhores resultados para esses casos terem sido obtidos do conjunto de dados que combinam todas as informações das receitas.

4.3.3 Tempo de Execução

Assim como esperado, os modelos neurais exigiram um maior tempo de execução na etapa treinamento, com o MLP chegando a quase 4 minutos e o BERT passando de uma hora, para execução no maior conjunto de treinamento (na dieta *kid-friendly*, com todas as informações combinadas). Apesar do tempo de treinamento relativamente alto, o MLP tem um tempo de teste significativamente menor, não passando de um segundo. Enquanto isso, o BERT mantém um alto tempo na etapa de teste, quando comparado com os demais modelos, mas expressivamente menor, comparado ao tempo de treinamento, não ultrapassando dois minutos, no pior caso. O SVM também apresentou elevados custos de tempo de execução, levando mais de 14 minutos no treinamento, e um pouco mais de 3 minutos na etapa de teste, no pior caso. Os demais modelos levaram poucos segundos no processo de treinamento e teste.

Apesar de alguns modelos apresentarem tempos de execução significativamente maiores que outros, todos os classificadores estariam aptos à aplicação em um sistema real, tendo em vista que os tempos de teste foram significativamente menores em relação ao tempo de treino, e considerando também que o tempo apresentado é resultado da inferência em todos

os documentos do conjunto de teste, o que significa que o tempo de predição por documento não passaria de um segundo em nenhum dos modelos.

4.3.4 Teste de Hipóteses

Por meio da aplicação do Teste de Friedman, foi rejeitada a hipótese nula de que todos os modelos apresentaram o mesmo desempenho, isto é, pelo menos um dos modelos testados obteve resultado significativamente diferente. Os *p-values* gerados através do teste são apresentados na Figura 4.2. Através do ranqueamento obtido com o Teste de Friedman-Nemenyi (Figura 4.2), podemos reforçar alguns fatos já observados anteriormente, como o bom desempenho dos modelos Regressão Logística, BERT e MLP, que aparecem, nessa ordem, no topo do *rank* gerado. Podemos observar também como os modelos que usam todas as informações da receita no processo de treinamento obtiveram desempenhos melhores que os demais, seguidos pelos modelos que usaram apenas a lista de ingredientes. Assim, como esperado, os modelos Árvore de Decisão e SVM (com características do *título* e da *descrição*) se encontram nas últimas posições do *rank*, fora da distância crítica (28.7426) dos melhores modelos. Os demais modelos não apresentaram diferenças estatísticas significativas em relação aos modelos mais bem ranqueados.

4.4 CONSIDERAÇÕES FINAIS

Os métodos adotados foram explorados e comparados nesta seção. Todos os classificadores foram capazes de obter resultados expressivos para todas as dietas classificadas, o que demonstra a efetividade da identificação de receitas dietéticas por meio de uma abordagem de Classificação de Texto. Diferentes informações das receitas são avaliadas como entrada para os classificadores. Os resultados experimentais são avaliados empiricamente, e através da execução do teste de hipóteses de Friedman/Nemenyi. Os modelos Regressão Logística, BERT e MLP, se destacaram entre os modelos avaliados, aparecendo, nessa ordem, no topo do *rank* gerado. Porém, os modelos Floresta Aleatória, Naive Bayes e K-NN tiveram resultados estatisticamente próximos. O conjunto de dados contendo todas as informações combinadas, e o que contia apenas a lista de ingredientes, contribuíram positivamente para os modelos de classificação.

Os modelos de Aprendizagem de Rasa, de maneira geral, por suas limitações inerentes,

não teriam a capacidade de capturar informações semânticas e sintáticas das sentenças, como os modelos de Aprendizagem Profunda. Porém, nos experimentos realizados, esses modelos alcançaram resultados estatisticamente próximos ou superiores (como na combinação das características TF-IDF e com a Regressão Logística) ao modelo BERT, que, teoricamente, levaria em consideração tais informações no processo de classificação. Esse fato pode representar um indicativo de que, no domínio de receitas culinárias, a ordem e relação das palavras não sejam tão relevantes quanto em outros domínios da Classificação de Texto, isto é, a presença ou ausência de um ingrediente seria mais relevante do que a ordem em que tais ingredientes são citados no documento.

No próximo capítulo (Capítulo 5, as conclusões finais do trabalho são apresentadas, além de algumas propostas para trabalhos futuros.

5 CONCLUSÕES

A escolha da refeição a ser consumida por um indivíduo com restrições alimentares pode ser um processo complexo e frustrante. A identificação dos pratos que infrinjam determinada restrição é fundamental para saúde e bem estar do indivíduo que segue dieta restritiva. Neste trabalho, executamos a classificação automática de receitas dietéticas por meio de técnicas de Aprendizagem de Máquina e Processamento Linguagem Natural.

Ao longo dos últimos anos, as receitas culinárias têm sido percebidas como fontes ricas de informações a serem explorados por aplicações de Mineração de Dados e Aprendizagem de Máquina, uma vez que os repositórios on-line representados por páginas especializadas concentram bases de documentos cada vez maiores, servindo como inspiração para o desenvolvimento de pesquisas que levem em consideração os aspectos relacionados aos gostos pessoais dos usuários desses sistemas.

Por meio de uma ampla revisão na literatura no tema de Classificação de Receitas, os principais fatos relacionados a essa tarefa foram observados e usados como inspiração para o desenvolvimento do presente trabalho. A revisão da literatura revelou que há um grande interesse dos pesquisadores na classificação do tipo de culinária das receitas, tendo esta aplicação sido impulsionada por competições pioneiras como a *"What's Cooking?"*. A análise exploratória de outras aplicações em receitas também foram sendo desenvolvidas, assim como novas competições, focando na inferência automática de outros aspectos das receitas, como a classificação do tipo de refeição e do nível de dificuldades de execução das mesmas, como a competição DEFT 2013.

Como já esperado, a grande maioria dos trabalhos abordam a classificação de documentos no idioma inglês (viés idiomático). Porém, outros idiomas também são contemplados por pesquisas, mesmo que em menor número de publicações, como os idiomas francês, alemão, espanhol e português. Tradicionalmente, classificadores de Aprendizagem de Máquina Rasa são usados, embora, nos últimos anos, trabalhos que focam no uso de modelos neurais profundos, como BERT, CNN e LSTM, também sejam encontrados.

Os principais classificadores observados na revisão da literatura foram selecionados, de acordo com o critério de efetividade demonstrada pelos mesmos ao longo dos trabalhos avaliados, como forma de garantir a efetividade da metodologia proposta. Através de uma análise empírica, e da aplicação de testes de hipóteses estatísticos, conseguimos explorar e comparar o

desempenho dos classificadores adotados. Os modelos de Regressão Logística, BERT e Perceptron Multicamadas, alcançaram, nessa ordem, os melhores desempenhos globais. Analisando principalmente a métrica da acurácia, o BERT foi capaz de obter o melhor desempenho nas dietas *diabético* (73%), *gluten-free* (73%) e *vegano* (91%). Enquanto isso, o MLP alcançou os melhores resultados na dieta *kid-friendly*, com 83% de acurácia. Já a Regressão Logística, teve o melhor desempenho para a dieta *lactose-free*, com 73% de acurácia. Com exceção da Árvore de Decisão e do SVM, que obtiveram resultados significativamente inferiores, os demais modelos também alcançaram bons resultados, com pouca diferença estatística para os modelos com os melhores desempenhos.

O modelo BERT é, atualmente, um dos mais prestigiados no estado da arte em várias tarefas de Processamento de Linguagem Natural, devido a sua capacidade de assimilar o contexto das palavras em uma sentença. No entanto, essa habilidade não foi tão relevante no contexto dos experimentos realizados, nos quais os modelos de Aprendizagem Rasa foram capazes de obter resultados estatisticamente similares, mesmo que esses modelos não tenham a mesma capacidade de compreensão textual do BERT. Isso pode ser justificado pelo fato de que a presença e ausência de termos (ingredientes, preparos, entre outros), que são facilmente capturadas pela medida TF-IDF, podem representar informações mais significativas que a relação e ordem desses termos na análise de receitas culinárias.

Expandindo as pesquisas realizadas na literatura, que tradicionalmente adotam apenas a lista de ingredientes como dados a serem analisados, diferentes informações textuais das receitas foram exploradas e usadas para o treinamento dos classificadores, e, de forma semelhante, o impacto desses conjuntos de dados no desempenho dos modelos foi verificado. Por meio dos experimentos foi possível constatar como as informações mais significativas para os modelos podem variar de restrição para restrição, onde diferentes tipos de dietas podem ser mais beneficiadas por determinados conjuntos de dados. Enquanto as dietas relacionadas à exclusão de determinados ingredientes são facilmente classificadas apenas através da lista de ingredientes das receitas, para o caso de dietas mais complexas, que envolvem outros fatores, a combinação os componentes textuais (lista de ingredientes, modos de preparo, títulos e descrições) podem ser mais informativos, auxiliando no desempenho dos modelos supervisionados.

Tendo em vista o grande impacto que esse sistema pode ter na saúde e bem estar dos usuários, é necessário, em trabalhos futuros, analisar a confiabilidade dos dados utilizados no treinamento, uma vez que informações inseridas por usuários comuns podem conter erros.

Como trabalhos futuros, aplicaremos a classificação de receitas culinárias para a inferência

de mais informações que podem ser úteis para o processo de escolha por parte dos usuários de sites de receitas, como o custo e tempo de preparo, e informações nutricionais. Também pretendemos combinar o BERT com outros modelos de aprendizagem profunda, como LSTM e CNN. Pretendemos utilizar esses modelos para integrar um sistema de geração de receitas, que identifica receitas que infrinjam determinada dieta e realiza a adaptação da mesma, através da substituição dos ingredientes que causam determinada restrição. A substituição desses ingredientes por equivalentes, que se adequem a tais dietas, pode tornar uma receita compatível com as expectativas e necessidades de usuários que tenham restrições alimentares. Através da substituição de ingredientes, esses usuários podem expandir seu cardápio, adaptando receitas que antes não poderiam consumir. Por fim, os modelos de classificação também serão utilizados para compor um sistema de recomendação de receitas culinárias personalizadas para pessoas que seguem dietas restritivas, sugerindo desde de receitas, a dietas completas, que sejam balanceadas e adequadas às suas restrições e gostos pessoais.

REFERÊNCIAS

AKKAYA, B.; ÇOLAKOĞLU, N. Comparison of multi-class classification algorithms on early diagnosis of heart diseases. In: *y-BIS 2019 Conference Book: Recent Advances n Data Sc ence and Bus ness Analyt cs*. [S.l.: s.n.], 2019. p. 162.

ALEMANY, J.; HERAS, S.; PALANCA, J.; JULIÁN, V. An agent-based application for automatic classification of food allergies and intolerances in recipes. In: SPRINGER. *Advances in Practical Applications of Scalable Multi-agent Systems. The PAAMS Collection: 14th International Conference, PAAMS 2016, Sevilla, Spain, June 1-3, 2016, Proceedings*. [S.l.], 2016. p. 3–12.

AZZIMANI, K.; BIHRI, H.; DAHMI, A.; AZZOUZI, S.; CHARAF, M. E. H. An ai based approach for personalized nutrition and food menu planning. In: IEEE. *2022 IEEE 3rd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*. [S.l.], 2022. p. 1–5.

BELYADI, H.; HAGHIGHAT, A. Chapter 5 - supervised learning. In: *Machine Learning Guide for Oil and Gas Using Python*. [S.l.]: Gulf Professional Publishing, 2021. p. 169–295.

BRITTO, L.; OLIVEIRA, E.; PACÍFICO, L.; LUDERMIR, T. Uma abordagem de análise de textos para a classificaçao de receitas culinárias baseadas em documentos em português brasileiro. In: SBC. *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2019. p. 436–447.

BRITTO, L.; PACÍFICO, L.; LUDERMIR, T. Inferência automática do nível de dificuldade em receitas culinárias usando técnicas de processamento de linguagem natural. In: SBC. *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2020. p. 104–115.

BRITTO, L.; PACÍFICO, L.; OLIVEIRA, E.; LUDERMIR, T. A cooking recipe multi-label classification approach for food restriction identification. In: SBC. *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2020. p. 246–257.

BRITTO, L. F. d. S.; PACÍFICO, L. D. S.; LUDERMIR, T. B. Inferência automática de nível calórico de receitas culinárias através de técnicas de aprendizagem de máquina. In: SBC. *Anais Estendidos do XVII Simpósio Brasileiro de Sistemas de Informação*. [S.l.], 2021. p. 33–36.

BRITTO, L. F. S.; PACIFICO, L. D. S.; LUDERMIR, T. B. Inferência automática de nível calórico de receitas culinárias através de técnicas de aprendizagem de máquina. In: *Anais Estendidos do XVII Simpósio Brasileiro de Sistemas de Informação*. [S.l.]: SBC, 2021. p. 33–36.

BROWN, S.; MYLES, A. 3.17 - Decision Tree Modeling in Classification. In: *Comprehensive Chemometrics*. Oxford: Elsevier, 2009. p. 541–569.

CATARINA, I. F. de S. *Alergias e intolerâncias alimentares: saiba o que são*. 2022. Acessado: 04 de Fevereiro 2023. Disponível em: <https://www.ifsc.edu.br/conteudo-aberto/-/asset_publisher/1UWKZAkiOauK/content/id/10481011>.

CELESTIN, T. K.; PATRICK, T. K.; SIMON, N. B. A comparative study of classification methods: case of application to asian cuisines ingredients. *International Journal of Computer Science Issues (IJCSI)*, International Journal of Computer Science Issues (IJCSI), v. 17, n. 3, p. 19–26, 2020.

CHARTON, E.; MEURS, M.-J.; JEAN-LOUIS, L.; GAGNON, M. Using collaborative tagging for text classification: From text classification to opinion mining. In: MDPI. *Informatics*. [S.l.], 2013. v. 1, n. 1, p. 32–51.

CHEN, M.; JIA, X.; GORBONOS, E.; HOANG, C. T.; YU, X.; LIU, Y. Eating healthier: Exploring nutrition information for healthier recipe recommendation. *Information Processing & Management*, Elsevier, v. 57, n. 6, p. 102051, 2020.

CHEN, P.-J.; ANTONELLI, M. Conceptual models of food choice: influential factors related to foods, individual differences, and society. *Foods*, MDPI, v. 9, n. 12, p. 1898, 2020.

CONSIDERA, E. M. O livro das noivas: receitas culinárias e relações de gênero na sociedade brasileira. *Caderno espaço feminino*, Universidade Federal de Uberlândia, v. 20, n. 2, p. 13–29, 2008.

CORDEIRO, E. P.; FONSECA, L. R. A.; RIBEIRO, R. de C.; SATHLER, M. M. Evolução dos livros de culinária brasileiros e sua relação com o cozinhar na contemporaneidade. *DEMETRA: Alimentação, Nutrição & Saúde*, v. 15, p. 47370, 2020.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, p. 273–297, 1995.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967.

COX, D. R. *The Analysis of Binary Data*. [S.l.]: Methuen, 1970. (Methuen's monographs on applied probability and statistics).

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, JMLR, v. 7, p. 1–30, 2006.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018.

DJURIS, J.; IBRIC, S.; DJURIC, Z. Neural computing in pharmaceutical products and process development. In: *Computer-Aided Applications in Pharmaceutical Technology*. [S.l.]: Elsevier, 2013. p. 91–175.

EDWARD, E. Doutorado em Ciência da Computação, *Comparing methods of text categorization*. 2018.

ELSWEILER, D.; TRATTNER, C.; HARVEY, M. Exploiting food choice biases for healthier recipe recommendation. In: *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. [S.l.: s.n.], 2017. p. 575–584.

FIX, E.; HODGES, J. L. Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, v. 57, p. 238, 1951.

FRANZONI, E. Mestrado em Ciências da Educação, *A gastronomia como elemento cultural, símbolo de identidade e meio de integração*. 2016.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937.

GE, M.; RICCI, F.; MASSIMO, D. Health-aware food recommender system. In: *Proceedings of the 9th ACM Conference on Recommender Systems*. [S.l.: s.n.], 2015. p. 333–334.

GHEWARI, R.; RAIYANI, S. *Predicting cuisine from ingredients*. 2015. Acessado: 02 de Fevereiro 2023. Disponível em: <https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/029.pdf>.

GIOVANELLI, C.; LIU, X.; SIERLA, S.; VYATKIN, V.; ICHISE, R. Towards an aggregator that exploits big data to bid on frequency containment reserve market. In: IEEE. *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*. [S.l.], 2017. p. 7514–7519.

HERBOLD, S. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, v. 5, n. 48, p. 2173, 2020.

HO, T. K. Random decision forests. In: IEEE. *Proceedings of 3rd international conference on document analysis and recognition*. [S.l.], 1995. v. 1, p. 278–282.

HOLSTE, H. H.; NYAYAPATI, M.; WONG, E. What cuisine?-a machine learning strategy for multi-label classification of food recipes. *University of California San Diego*, p. 1–7, 2015.

JAYARAMAN, S.; CHOUDHURY, T.; KUMAR, P. Analysis of classification models based on cuisine prediction using machine learning. In: IEEE. *2017 international conference on smart technologies for smart nation (SmartTechCon)*. [S.l.], 2017. p. 1485–1490.

JOMORI, M. M.; PROENÇA, R. P. d. C.; CALVO, M. C. M. Determinantes de escolha alimentar. *Revista de Nutrição*, SciELO Brasil, v. 21, p. 63–73, 2008.

KALAJDZISKI, S.; RADEVSKI, G.; IVANOSKA, I.; TRIVODALIEV, K.; STOJKOSKA, B. R. Cuisine classification using recipe's ingredients. In: IEEE. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. [S.l.], 2018. p. 1074–1079.

KAN, W. *What's Cooking?* Kaggle, 2015. Disponível em: <https://kaggle.com/competitions/whats-cooking>.

KAYIKÇI, Ş.; BAŞOL, Y.; DÖRTER, E. Classification of turkish cuisine with deep learning on mobile platform. In: IEEE. *2019 4th International Conference on Computer Science and Engineering (UBMK)*. [S.l.], 2019. p. 1–5.

KICHERER, H.; DITTRICH, M.; GREBE, L.; SCHEIBLE, C.; KLINGER, R. What you use, not what you do: automatic classification of recipes. In: SPRINGER. *Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings*. [S.l.], 2017. p. 197–209.

KICHERER, H.; DITTRICH, M.; GREBE, L.; SCHEIBLE, C.; KLINGER, R. What you use, not what you do: Automatic classification and similarity detection of recipes. *Data & Knowledge Engineering*, Elsevier, v. 117, p. 252–263, 2018.

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? *Nature biotechnology*, Nature Publishing Group US New York, v. 26, n. 9, p. 1011–1013, 2008.

KOWSARI, K.; MEIMANDI, K. J.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. Text classification algorithms: A survey. *Information*, MDPI, v. 10, n. 4, p. 150, 2019.

KUMAR, R.; KUMAR, M. A.; SOMAN, K. Cuisine prediction based on ingredients using tree boosting algorithms. *Indian Journal of Science and Technology*, v. 9, n. 45, p. 12, 2016.

LI, B.; WANG, M. *Cuisine Classification from Ingredients*. 2015. Acessado: 02 de Fevereiro 2023. Disponível em: <https://cs229.stanford.edu/proj2015/313_report.pdf>.

LIMA, J. F. O. Receitas culinárias de família como expressão de cultura. *Revista Estação Científica. Centro Universitário Estácio de Juiz de Fora Edição Especial VII Seminário de Pesquisa da Estácio e III Jornada de Científica da UNESA*, v. 2, 2015.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proc. 5th Berkeley Symposium on Math., Stat., and Prob.* [S.l.: s.n.], 1965. p. 281.

MAIA, R.; FERREIRA, J. C. Context-aware food recommendation system. *2018 World Congress on Engineering and Computer Science, WCECS 2018*, International Association of Engineers, p. 349–356, 2018.

MAJUMDER, B. P.; LI, S.; NI, J.; MCAULEY, J. Generating personalized recipes from historical user preferences. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.]: Association for Computational Linguistics, 2019. p. 5976–5982.

MAO, X.; YUAN, S.; XU, W.; WEI, D. Recipe recommendation considering the flavor of regional cuisines. In: IEEE. *2016 International Conference on Progress in Informatics and Computing (PIC)*. [S.l.], 2016. p. 32–36.

MISRA, S.; LI, H.; HE, J. Noninvasive fracture characterization based on the classification of sonic wave travel times. *Machine learning for subsurface characterization*, Gulf Professional Publishing Houston, TX, USA, v. 243, 2020.

MOHAMMADI, E.; NAJI, N.; MARCEAU, L.; QUEUDOT, M.; CHARTON, E.; KOSSEIM, L.; MEURS, M.-J. Classification of rare recipes requires linguistic features as special ingredients. In: SPRINGER. *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13–15, 2020, Proceedings 33*. [S.l.], 2020. p. 426–437.

MOHAMMADI, E.; NAJI, N.; MARCEAU, L.; QUEUDOT, M.; CHARTON, E.; KOSSEIM, L.; MEURS, M.-J. Cooking up a neural-based model for recipe classification. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. [S.l.: s.n.], 2020. p. 5000–5009.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, Taylor & Francis, v. 58, n. 302, p. 415–434, 1963.

MUPPALA, G. K. T. *Recipe Text Classification using Graph Neural Networks Stanford CS224N Custom Project*. 2022. Acessado: 02 de Fevereiro 2023. Disponível em: <https://web.stanford.edu/class/cs224n/reports/custom_117176180.pdf>.

MUSHTAQ, M.-S.; MELLOUK, A. Methodologies for subjective video streaming qoe assessment. *Quality of Experience Paradigm in Multimedia Services*, Elsevier, p. 27–57, 2017.

NAIK, J.; POLAMREDDI, V. *Cuisine classification and recipe generation*. 2015. Acessado: 02 de Fevereiro 2023. Disponível em: <http://cs229.stanford.edu/proj2015/233report.pdf>.

NATIONS, F. *Family Nutrition Guide*. [S.l.]: Food and Agriculture Organization of the United Nations, 2018.

NEMENYI, P. B. *Distribution-free multiple comparisons*. [S.l.]: Princeton University, 1963.

NIELSEN. *What's in our food and on our mind*. 2023. Acessado: 02 de Fevereiro 2023. Disponível em: <https://ilmanagement.files.wordpress.com/2016/09/qui2.pdf>.

NIRMAL, I.; CALDERA, H. *Test Engineering Management*, v. 82, n. 12731, p. 12731–12737.

NIU, Z.; ZHONG, G.; YU, H. A review on the attention mechanism of deep learning. *Neurocomputing*, Elsevier, v. 452, p. 48–62, 2021.

OOI, A.; IIBA, T.; TAKANO, K. Ingredient substitute recommendation for allergy-safe cooking based on food context. In: *PACRIM*. [S.l.: s.n.], 2015. p. 444–449.

PACIFICO, L. D. S.; BRITTO, L. F. S.; LUDERMIR, T. B. Geração de receitas culinárias para usuários com restrições alimentares pela substituição automática de ingredientes. In: *Anais do XLVIII Seminário Integrado de Software e Hardware*. [S.l.]: SBC, 2021. p. 183–190.

PACIFICO, L. D. S.; BRITTO, L. F. S.; LUDERMIR, T. B. Ingredient substitute recommendation based on collaborative filtering and recipe context for automatic allergy-safe recipe generation. In: . [S.l.]: Association for Computing Machinery, 2021. (WebMedia '21), p. 97–104.

PACIFICO, L. D. S.; BRITTO, L. F. S.; LUDERMIR, T. B. Automatic recipe ingredient substitution based on text mining and data clustering approaches. In: *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2022)*. [S.l.]: SBC, 2022. p. 1–8.

PACIFICO, L. D. S.; BRITTO, L. F. S.; LUDERMIR, T. B. Improved alternative average support value for automatic ingredient substitute recommendation in cooking recipes. In: *Intelligent Systems*. [S.l.]: Springer International Publishing, 2022. p. 373–387.

PACIFICO, L. D. S.; BRITTO, L. F. S.; LUDERMIR, T. B. Improved alternative average support value for automatic ingredient substitute recommendation in cooking recipes. In: *Brasilian Conference on Intelligent Systems (BRACIS 2022)*. [S.l.]: SBC, 2022. p. 373–387.

PAN, L.; POUYANFAR, S.; CHEN, H.; QIN, J.; CHEN, S.-C. Deepfood: Automatic multi-class classification of food ingredients using deep learning. In: IEEE. *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)*. [S.l.], 2017. p. 181–189.

PAREKH, S. S. B. K. V. *Identifying Cuisines From Ingredients*. 2015. Acessado: 02 de Fevereiro 2023. Disponível em: <https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/039.pdf>.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR, v. 12, p. 2825–2830, 2011.

QUINLAN, J. R. Simplifying decision trees. *International journal of man-machine studies*, Elsevier, v. 27, n. 3, p. 221–234, 1987.

RICHMAN, J. S. Multivariate neighborhood sample entropy: a method for data reduction and prediction of complex data. In: *Methods in enzymology*. [S.l.]: Elsevier, 2011. v. 487, p. 397–408.

ROITHER, A.; KURZ, M.; SONNLEITNER, E. The chef's choice: System for allergen and style classification in recipes. *Applied Sciences*, MDPI, v. 12, n. 5, p. 2590, 2022.

ROTHMAN, D. *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. [S.l.]: Packt Publishing Ltd, 2021.

SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, v. 21, n. 3, p. 660–674, 1991.

SAJADMANESH, S.; JAFARZADEH, S.; OSSIA, S. A.; RABIEE, H. R.; HADDADI, H.; MEJOVA, Y.; MUSOLESI, M.; CRISTOFARO, E. D.; STRINGHINI, G. Kissing cuisines: Exploring worldwide culinary habits on the web. In: *Proceedings of the 26th international conference on world wide web companion*. [S.l.: s.n.], 2017. p. 1013–1021.

SALTON, G.; MCGILL, M. J. *Introduction to Modern Information Retrieval*. [S.l.]: McGraw-Hill, 1983. (International student edition).

SANTOS, C. R. A. dos. A alimentação e seu lugar na história: os tempos da memória gustativa. *História: questões & debates*, v. 42, n. 1, 2005.

SCHäFER, H.; ELAHI, M.; ELSWEILER, D.; GROH, G.; HARVEY, M.; LUDWIG, B.; RICCI, F.; SAID, A. User nutrition modelling and recommendation: Balancing simplicity and complexity. In: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*. [S.l.]: Association for Computing Machinery, 2017. p. 93–96.

SHARMA, T.; UPADHYAY, U.; BAGLER, G. Classification of cuisines from sequentially structured recipes. In: IEEE. *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*. [S.l.], 2020. p. 105–108.

SIMSKE, S. Chapter 1 - Introduction, overview, and applications. In: *Meta-Analytics*. [S.l.]: Morgan Kaufmann, 2019. p. 1–98.

SU, H.; LIN, T.-W.; LI, C.-T.; SHAN, M.-K.; CHANG, J. Automatic recipe cuisine classification by ingredients. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. [S.l.]: Association for Computing Machinery, 2014. (UbiComp '14 Adjunct), p. 565–570.

U.S. Food and Drug Administration. *How to Understand and Use the Nutrition Facts Label*. 2020. Disponível em: <https://www.fda.gov/food/new-nutrition-facts-label/how-understand-and-use-nutrition-facts-label>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

VERMA, R. S.; ARORA, H. *CSE 255 Assignment 2 Cuisine Prediction/Classification based on ingredients*. 2015. Acessado: 02 de Fevereiro 2023. Disponível em: <https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/028.pdf>.

VIJAYAKUMARI, G.; VUTKUR, P.; VISHWANATH, P. Food classification using transfer learning technique. *Global Transitions Proceedings*, Elsevier, v. 3, n. 1, p. 225–229, 2022.

WANG, X.; KUMAR, D.; THOME, N.; CORD, M.; PRECIOSO, F. Recipe recognition with large multimodal food dataset. In: IEEE. *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. [S.l.], 2015. p. 1–6.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. v.; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; SCAO, T. L.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. M. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2020.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

XU, Y.; JONES, G. J.; LI, J.; WANG, B.; SUN, C. A study on mutual information-based feature selection for text categorization. *Journal of Computational Information Systems*, Binary Information Press, v. 3, n. 3, p. 1007–1012, 2007.

YANG, L.; HSIEH, C.-K.; YANG, H.; DELL, N.; BELONGIE, S.; ESTRIN, D. Yum-me: Personalized healthy meal recommender system. *ACM Trans. Inf. Syst.*, Association for Computing Machinery, v. 36, n. 1, 2016.

ZHU, Y.; KIROS, R.; ZEMEL, R.; SALAKHUTDINOV, R.; URTASUN, R.; TORRALBA, A.; FIDLER, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 19–27.

ZUIN, L. F. S.; ZUIN, P. B. Alimentação é cultura: aspectos históricos e culturais que envolvem a alimentação e o ato de se alimentar:[revisão]. *Nutrire Rev. Soc. Bras. Aliment. Nutr*, p. 225–241, 2009.