Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística

RANAH DUARTE COSTA

CLASSIFICAÇÃO DE ASSINATURAS MANUSCRITAS COM QUANTIFICADORES NÃO PARAMÉTRICOS.

Recife

2020

RANAH DUARTE COSTA

CLASSIFICAÇÃO DE ASSINATURAS MANUSCRITAS COM QUANTIFICADORES NÃO PARAMÉTRICOS.

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientador: Raydonal Ospina Martínez

Recife

2020

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

C837c Costa, Ranah Duarte
Classificação de assinaturas manuscritas com quantificadores não paramétricos / Ranah Duarte Costa. – 2020.
72 f.: il., fig., tab.

Orientador: Raydonal Ospina Martínez.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2020.
Inclui referências.

1. Estatística aplicada. 2. Classificação binária. I. Martínez, Raydonal Ospina (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2020 - 45

RANAH DUARTE COSTA

CLASSIFICAÇÃO DE ASSINATURAS MANUSCRITAS COM QUANTIFICADORES NÃO PARAMÉTRICOS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 20 DE FEVEREIRO DE 2020

BANCA EXAMINADORA

Prof.(º) Raydonal Ospina Martínez
UFPE

Prof.(º) Marcelo Rodrigo Portela Ferreira
UFPB

Prof.(ª) Renata Maria Cardoso Rodrigues de Souza
UFPE

AGRADECIMENTOS

Em primeiro lugar, agradeço aos meus pais que tanto se dedicaram à minha criação e educação, e assim como meu irmão, me incentivaram nesse trajeto.

Ao meu orientador, Raydonal, pelo incentivo, paciência, dedicação e pelas sessões de terapia ao longo desse ano.

A todos os meus amigos e colegas da pós, por estarem comigo nos bons e maus momentos.

A todos os meus professores que muito contribuíram na minha formação.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro.

RESUMO

Essa dissertação tem como objetivo utilizar quantificadores não paramétricos no processo de classificação binária de assinaturas manuscritas. Os dados representam as informações das assinaturas de 100 indivíduos da base de dados MCYT (MCYT Fingerprint subcorpus), sendo que para cada indivíduo apresenta-se réplicas com 25 assinaturas falsas e 25 assinaturas verdadeiras. Aqui, as assinaturas falsas e verdadeiras são rotuladas com zeros e uns no problema de classificação binária, respectivamente. Para o processamento da cada assinatura é extraída a série temporal correspondente a cada coordenada do plano xy. Adicionalmente, para cada uma das séries temporais obtidas, foram calculadas a primeira e a segunda derivada a fim de avaliar a dinâmica em termos de sua velocidade e a aceleração, respectivamente. Também, em cada uma das séries temporais foram extraídos quantificadores de informação não paramétricos a partir da distribuição de padrões (feature extraction), a saber: entropia, complexidade, informação de Fisher e tendência. De posse dos quantificadores extraídos, uma nova base de dados foi construída a fim de avaliar a capacidade dessas informações para separar as assinaturas falsas e verdadeiras. Dessa maneira, foram usados critérios de seleção de variáveis para a classificação, sendo esses: Ganho de informação, análise de variância (ANOVA) e fator de inflação da variância. No que tange aos classificadores, foram utilizados a Regressão Logística, Máquinas de Vetores de Suporte (SVM), Florestas aleatórias (Random Forest), XGBoost (Extreme Gradient Boosting) e regressão regularizada tipo LASSO e Ridge. Neste trabalho, as métricas de avaliação de performance dos classificadores foram a acurácia, sensibilidade, especificidade, área sob a curva ROC (AUC) e taxa de erro de classificação. Os resultados mostram que, entre os quantificadores utilizados, a complexidade, a Informação de Fisher e a estatística de Wallis e Moore foram os quantificadores não paramétricos que conseguem melhorar a performance dos classificadores. Adicionalmente, os classificadores SVM e Florestas aleatórias apresentaram melhor desempenho no grupo de teste quando comparados aos demais segundos as métricas usadas. Por outro lado, a etapa de classificação usando a regressão regularizada tipo LASSO e Ridge e a Regressão Logística não regularizada mostrou que, para esse conjunto de dados, a regressão não regularizada apresenta melhor desempenho.

Palavras-chave: Assinaturas manuscritas. Extração de característica. Classificação binária.

ABSTRACT

This work explores the use of nonparametric quantifiers in the binary classification process of handwritten signatures. We use the MCYT (MCYT Fingerprint subcorpus) database with 100 subjects, where each one contains 25 genuine and 25 skilled forged signatures. Here, false and true signatures are labeled with zeros and ones for the binary classification problem, respectively. We work with the discrete-time sequences position xt in the x-axis and position yt in the y-axis provided in the database. We pre-process each time series and employ time causal information based on nonparametrics quantifiers such as an entropy, complexity, Fisher information, and trend. Also, we evaluate these quantifiers with the time series obtained by applying the first and second order derivatives of each sequence position to evaluate the dynamic behaviour looking their velocity and acceleration, respectively. To assess the ability of nonparametrics quantifiers information to separate false and true signatures, we used criteria selection variables, such as: Information gain, analysis of variance (ANOVA), and variance inflation factor. In the next, we classify the signatures in the MCYT-100 database with nonparametrics quantifiers via Logistic Regression, Support Vector Machines (SVM), Random Forest (Random Forest), regularized regression type Lasso, and Extreme Gradient Boosting (XGBoost). We evaluate the performance of the classifiers by analyzing the accuracy, sensitivity, specificity, area under the ROC curve (AUC), and the Error Rate (ER). The results show that, among the quantifiers used, the Complexity, Fisher Information, and the Wallis and Moore information are the nonparametric quantifiers that improve the performance of the classifiers. Additionally, the SVM and Random Forest classifiers perform better in the test group compared to the others, according to the metrics used. In the classification step, we use LASSO and Ridge regularized regression and the non-regularized Logistic Regression, and the results show that, for this data set, the non-regularized regression presents better performance.

Keywords: Handwritten signatures. Feature extraction. Binary classification.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A biometria é o estudo das características físicas ou padrões comportamentais dos seres vivos. Dessa maneira, a biometria física se refere aos traços físicos das pessoas, como impressões digitais, identificação da íris, veias, geometria das mãos, DNA (Ácido Desoxirribonucléico), etc. Por outro lado, a Biometria comportamental retrata verificação de assinatura, voz, dinâmica de escrita, etc (FRANKE; SOLAR; KÖPPEN, 2012) . A verificação de assinatura, comparada à outras técnicas biométricas de reconhecimento, tem algumas vantagens como um mecanismo de verificação de identidade, pois a captura das assinaturas é mais fácil de ser aceito socialmente e, além disso, forjar uma assinatura mostra-se mais difícil do que uma impressão digital (HOU; YE; WANG, 2004).

A assinatura manuscrita se destaca entre as demais biometrias por ser um meio bastante utilizado como identificação pessoal, visto que é a representação manuscrita do nome de alguém ou uma marca de identificação escrita. A assinatura constitui um dos meios menos invasivos comparada a outras técnicas, como DNA, impressão digital e veias. Além disso, é utilizada há séculos em autorização de transações legais. Sendo assim, é indispensável o uso de técnicas automáticas que auxiliem a verificação dessas (PANSARE; BHATIA, 2012).

Entretanto, as assinaturas manuscritas estão bastantes sujeitas a falsificações, pois a partir de um tempo de treinamento suficiente por parte do falsificador é possível forjar uma imagem e ser difícil distinguir que esta não pertence ao legítimo proprietário da assinatura (SANTOS *et al.*, 2004; ADAMSKI; SAEED, 2014; HOU; YE; WANG, 2004). Desse modo, o problema de reconhecimento de padrões não é considerado uma tarefa trivial no âmbito das assinaturas, mas sim, um problema difícil, pois, além das habilidades do falsificador, as amostras de assinaturas de uma mesma pessoa, apesar de semelhantes, não são idênticas. Isso se deve ao fato que com o passar dos anos a assinatura de um mesmo indivíduo sofre alterações, por condições físicas e estado psicológico ou mental e, assim, podem aumentar a variabilidade das assinaturas de uma mesma pessoa (PLAMONDON; LORETTE, 1989).

Hilton (1992) destaca em seus estudos, que as assinaturas possuem pelo menos três características, forma, movimento e variação. Dentre esses atributos, destaca-se o movimento, pois esse é produzido pelos músculos dos dedos, mão, punho e ombro. Dado que uma pessoa está acostumada a produzir sua assinatura, esses impulsos nervosos são controlados pelo cérebro sem nenhuma atenção particular aos detalhes da escrita. De forma geral, a assinatura verdadeira é considerada um movimento balístico, ou seja, é um movimento rápido, sem um feedback

posicional, predeterminado pelo cérebro e, assim, não podendo ser feito lentamente. Em contrapartida, Nalwa (1997) leva em conta que a assinatura forjada é considerada uma ação deliberada, isto é, uma tentativa de reproduzir uma escrita com o auxilio de um feedback posicional.

A verificação de assinaturas, durante um longo período, foi realizada de forma manual, que, além de demandar tempo, também exigia uma mão-de-obra especializada para tarefa. Nesse sentido, essa classe de verificação está propensa a gerar discordância, visto que a análise feita pelo o humano é subjetiva. Então a verificação de assinaturas é uma tarefa complexa, já que uma verificação manual para uma grande quantidade de documentos é tediosa e facilmente influenciada por fatores físicos e psicológicos do avaliador (XIAO; LEEDHAM, 1999).

Para a obtenção das assinaturas, existem duas formas gerais de registrar e de analisar, o modo *offline* e o modo *online*. O método *online*, o indivíduo assina em um dispositivo digitalizador (mesa digitalizadora, *tablet* ou celular) de forma a converter as informações da mão em dados sequenciais da caligrafia, sendo assim, obtêm-se informações, como a ordem temporal dos traços, pressão exercida na caneta, velocidade da mão, entre outras, a depender da tecnologia do dispositivo utilizado (NGUYEN; BLUMENSTEIN, 2010). SegundoAqili *et al.* (2016), a verificação de assinaturas pelo método *online* têm apresentado melhores resultados ao que se refere à acurácia.

Sob outra perspectiva, no modo *offline* as assinaturas são registradas em um papel e são digitalizadas a partir de dispositivos ópticos, como *scanner* e câmeras digitais, resultando em imagens estáticas das assinaturas (PLAMONDON; SRIHARI, 2000). A verificação de assinatura *online* têm alcançado uma taxa de verificação melhor do que o modo *offline*, isso ocorre devido à possibilidade de, no modo *online*, extrair mais informações da assinatura (QIAO; LIU; TANG, 2007).

A verificação de assinatura trata-se de um problema cuja assinatura a ser testada será classificada como verdadeira ou falsa. Segundo Esmael *et al.* (2015) a abordagem de reconhecimento de padrões consiste em quatro fase; preparação dos dados, recurso de extração de características (*features*), seleção das características e a classificação.

A extração de caracteríscticas das séries temporais desempenha um papel de destaque na verificação de assinatura (RASHIDI; FALLAH; TOWHIDKHAH, 2012). O banco de dados que será utilizado nessa dissertação é o MCYT (MCYT Fingerprint subcorpus) que fornece a série de tempo discreto e a partir dele são extraídos os quantificadores de informação ou características. Para a seleção das características, técnicas de seleção de variáveis são aplicadas

e, dessa maneira, obtêm-se conjuntos de variáveis relevantes, que alimentam modelos, para predizer a autenticidade da assinatura.

No que tange a classificação, os métodos utilizados serão Regressão logística, Máquina de Suporte de vetores, Florestas Aleatórias e *Extreme Gradient Boosting* que serão aplicados nos modelos obtidos através dos métodos de seleção de variáveis e, por fim, a Regressão logística com penalização tipo LASSO e Ridge será avaliada. Essa última, será usada na base de dados que contém todas as características extraídas das séries temporais de assinaturas.

Essa dissertação inicia-se com uma referencial teórico sobre classificação binária. Posteriormente, encontra-se a metodologia utilizada, seguida dos resultados obtidos. Posto disso, a última parte descreve as conclusões seguida das referências bibliográficas.

2 CLASSIFICAÇÃO BINÁRIA

Este capítulo traz uma revisão bibliográfica dos métodos de classificação que foram utilizados nessa dissertação, sendo esses, Regressão logística, Máquinas de Vetores de Suporte, Florestas Aleatórias, *Extreme Gradient Boosting* e regressão logística com penalização tipo LASSO e Ridge. Na literatura, estudos mostram que esses métodos são bastante competitivos tanto na área de regressão, como também, na classificação (FAN *et al.*, 2018; FERNÁNDEZ-DELGADO *et al.*, 2014). Sendo assim, tais métodos foram utilizados para classificar assinaturas manuscritas verdadeiras e falsas. Dessa maneira, antes de introduzi-los, é necessário um breve resumo sobre a classificação binária.

Classificar algo é uma tarefa rotineira para o ser humano, onde usam-se informações prévias para tomar uma decisão, como exemplo, dada informações sobre o estado de saúde de um paciente, um médico consegue classificá-lo em duas classes, isto é, em ter determinada doença ou não, porém é possível que esse tipo de tarefa precise ser aplicado a grandes bancos de dados ou que as informações sobre o que se deseja classificar não seja interpretável pelo humano, como o caso de detecção de SPAM em uma caixa de e-mail (AWAD; ELSEUOFI, 2011). Então, por esses aspectos, é importante que uma tarefa da classificação seja desempenhada por uma máquina.

Na área da computação, o aprendizado de máquina (*machine learning*) é um ramo da inteligência artificial cujos estudos são feitos para que, com auxilio de algoritmos, o computador tome decisão com base em informações inseridas neles. Dessa forma, pode-se dividir o aprendizado de máquina em duas categorias aprendizado supervisionado e não supervisionado.

Na aprendizagem supervisionada, o conjunto de dados que será utilizado está dividido em conjunto de entrada, que se refere as variáveis independentes, e o conjunto de saída, a variável dependente e, assim, utiliza as variáveis de entrada para predizer a variável de saída. A aprendizagem supervisionada é bastante utilizada em classificação e em regressão. Por outro lado, o aprendizado não supervisionado o conjunto de dados não possui os rótulos de entrada e de saída (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007), sendo utilizado em clusterização. Aliado ao aprendizado de máquina, a mineração de dados consiste em extrair de conjunto de dados informações úteis que auxiliem no objetivo do estudo, nesse contexto, a classificação binária é um exemplo de aprendizado supervisionado em que tem-se a junção dessas duas técnicas (HAND, 2006; MAGLOGIANNIS, 2007).

No que se refere à classificação, usa-se algum método que, de maneira automática,

consiga separar as classes em estudo, com base em informações sobre os indivíduos (KUMARI; SRIVASTAVA, 2017). Na literatura, existem alguns métodos usados para fazer essa separação, para mais detalhes Yadav e Shukla (2016). Nessa dissertação, como forma de realizar a classificação, utiliza-se um método chamado *holdout*, esse processo acontece de tal forma que, são feitas amostragens aleatórias simples do banco de dados e esses são divididos em base de treinamento e base de teste, normalmente utiliza-se 70% dos dados para treinamento e 30% para a fase de teste (KOHAVI *et al.*, 1995). Dessa maneira, na fase de treinamento o algoritmo utilizado usa uma a forma de classificação a partir do conhecimento prévio dos dados, ou seja, nessa fase o sistema aprende como classificar esse tipo de dados, estimando coeficientes de um modelo de tal forma que minimize o erro de classificação e, assim, na fase de teste o sistema utiliza o conhecimento adquirido para colocar em prática a classificação supondo que as classes, nessa fase, sejam desconhecidas. No entanto, quando se utiliza o *holdout* uma única vez, dependendo da semente, ou seja, valor inicial para a geração de números aleatórios, pode ocorrer que os grupos de treinamento e teste não sejam representativos e, portanto, não gerem resultados condizentes com a amostra. Sendo assim, Kourou *et al.* (2015) sugere que se faça uma amostragem aleatória pra cada método *holdout* e, assim, melhorar a precisão, pois com essa abordagem é possível obter estimativas das médias e dispersões dos critérios de seleção a serem usados para eleger o melhor classificador para esses dados.

No que tange aos modelos usados para classificação, busca-se um que tenha um bom desempenho tanto na fase de treinamento, como também na fase de teste. Entretanto, é possível que um modelo apresente uma excelente estimação na fase de treinamento, porém na fase de teste seja um modelo com baixo poder de predição. Dessa forma, no âmbito da classificação, esse problema é chamado de *overfitting* (KHOSHGOFTAAR; ALLEN, 2001).

Posto disso, para essa dissertação, o interesse é classificar assinaturas, em duas classes, falsas e verdadeiras, então dada uma certa obervação, Y é de natureza binária, tal que $\{Y = 0\}$ se refere à assinatura cuja classe é falsa e $\{Y = 1\}$ para a classe de assinatura verdadeira, assim, $C = \{0, 1\}$ é o conjunto das classes (BREIMAN, 2017). Suponha que existam p variáveis explicativas para cada observação e essas estão alocadas no vetor $\mathbf{x} = (x_1, x_2, \ldots, x_P)$ de tal forma que esteja definida no espaço multidimensional Ω contendo todos os vetores de $\mathbf{x}$, então um classificador é uma função

$$g : \Omega \to C$$

$$\mathbf{x} \mapsto g(\mathbf{x}) \in C$$

que para cada vetor de $\mathbf{x}$, o classificador $g(\mathbf{x}) \in \{0,1\}$.

2.1 REGRESSÃO LOGÍSTICA

Nos modelos de regressão linear simples ou múltipla, temos que Y é uma variável aleatória dependente, e de natureza contínua e X uma matriz de p colunas que são as variáveis explicativas contendo vetores de $\mathbf{x}$ e, assim, nessa regressão o objetivo é desenvolver um modelo estatístico que possa ser usado para prever os valores esperados da variável Y condicionado as variáveis explicativas conhecidas de X, expresso por, $E(Y|X) = \beta X$, onde β é o vetor de coeficientes da reta de regressão associados a matriz X (HOSMER DAVID W; JOVANOVIC; LEMESHOW, 1989; HOSMER DAVID W; LEMESHOW; STURDIVANT, 2013).

Para o uso desse modelo, é necessário que a variável dependente pertença ao conjunto dos reais. Contudo, existem algumas situações onde a variável dependente é de natureza binária e, assim, é preciso utilizar uma transformação para que a condição seja satisfeita e, diante disso denomina-se regressão logística.

A regressão logística está contida na classe de modelos lineares generalizados (MLG). Posto disso, essa classe de modelos apresenta três componentes: a componente aleatória, onde y_i corresponde ao i-ésimo valor da variável aleátoria Y e é responsável por identificar a distribuição da variável dependente e, no nosso caso, Y possui duas categorias $\{0,1\}$, uma componente sistemática, obtida através do vetor $\eta = (\eta_1, \eta_2, \ldots, \eta_n)$ que está associado, por meio de um modelo linear, ao conjunto de variáveis independentes x, cuja função é especificar uma função linear entre os regressores e, por último, essa função de ligação que representa a relação matemática entre o valor esperado da componente aleatória e a componente sistemática, assim, η é definido como $\eta = g(\cdot)$, em que g é uma função monótona e diferenciável, vale ressaltar que se $g(\cdot)$ for a função identidade, tem-se o modelo de regressão linear (HOSMER DAVID W; LEMESHOW; STURDIVANT, 2013). Usando o modelo linear múltiplo, temos que:

$$y = \beta^T x_i + \varepsilon \tag{2.1}$$

Aqui, $y = (y_1, \ldots, y_n)$ é o vetor $n \times 1$ de respostas, $\beta^T = (\beta_1, \ldots, \beta_p)$ é o vetor de parâmetros de dimensão $1 \times p$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ é o vetor $n \times 1$ contendo os erros e o vetor $x_i = (x_{i1}, \ldots, x_{ij})^T$ é o vetor contendo as observações de cada variável do indivíduo i, sendo esse, um vetor linha pertencendo a matriz X. Quando o modelo contém um intercepto, a primeira coluna de X é um vetor de uns.

Sendo assim, para o caso onde Y é de natureza binária, precisamos aplicar uma transformação na variável resposta para que essa pertença ao conjunto de números reais. Dessa maneira, existem, na literatura, diversas transformações possíveis, como a *logit*, *probit* e *log-log* (DEMÉTRIO, 2001), no presente trabalho, a transformação de interesse é a *logit*, então temos que:

$$g(x_i) = ln\left\{\frac{p_i}{1-p_i}\right\} = \beta^T x_i \tag{2.2}$$

em que,

$$p_i = P(Y_i = 1|x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} = \frac{1}{1+e^{-(\beta^T x_i)}} \tag{2.3}$$

em que Y segue a distribuição de Bernoulli em virtude da natureza ditocômica da váriável(HOSMER DAVID W; JOVANOVIC; LEMESHOW, 1989). Em termos de interpretabilidade, p_i é a probabilidade da ocorrência de $Y = 1$ dado as variáveis explicativas representadas por X.

A razão contida na Eq. (2.2) é conhecida como razão de chances (*odds ratio*) da ocorrência do evento $\{Y = 1\}$ sob o evento $\{Y = 0\}$ dado um x_i. Seja Y um evento de interesse, então a chance do evento Y ocorrer, i.e, $Y = 1$, é a relação entre a probabilidade de ocorrência de $Y = 1$ e a probabilidade de não ocorrência de $Y = 0$. Dessa maneira, supondo que a probabilidade de ocorrência de Y, ou seja, $\mathrm{P}(Y = 1)$ é de $0,8$, então a chance de ocorrer esse evento é de $4:1$ (4 ocorrências para 1 não ocorrência). De maneira similar, se um evento Y tem chance de $0,25$ (1 ocorrência para 4 não ocorrência) de ocorrer, então sua probabilidade de ocorrência é de $0,2$. A chance esta compreendida no intervalo $(0,+\infty)$, ela é usada como uma maneira de interpretabilidade dos parâmetros do modelo (HOSMER DAVID W; LEMESHOW; STURDIVANT, 2013).

Como forma de estimar os coeficientes do modelos logístico, utilizamos o método de máxima verossimilhança com algorítimos numéricos de maximização.

$$\widehat{\beta} = \underset{\beta}{argmax}\left\{\sum_{i=1}^{n} ln(1+\exp(-\beta^T x_i))\right\} \tag{2.4}$$

Para a classificação, temos $\widehat{g(x_i)}$ como estimativa do logaritmo natural da razão de chances, então, desse modo, a exponencial de $\widehat{g(x_i)}$ nos retorna a razão de chances estimada. Assim sendo, definindo $0,5$ como um ponto de corte, conseguimos estimar Y_i através do valor estimado da exponencial de $\widehat{g(x_i)}$, ou seja, se esse valor for menor que $0,5$, então existe menor probabilidade de ocorrer o evento $\{\widehat{Y}_i = 1\}$ sob $\{\widehat{Y}_i = 0\}$ dado x_i, dessa maneira, classificamos $\widehat{Y}_i = 0$, caso contrário, $\widehat{Y}_i = 1$.

2.2 REGRESSÃO REGULARIZADA TIPO LASSO E RIDGE

A regressão linear e logística, além de apresentar as dificuldades para a estimação dos parâmetros, quando há presença de muitas covariáveis, isto é, vetores de x pertencentes a matriz X de variáveis explicativas, podem gerar problemas de multicolinearidade, ou seja, quando as variáveis explicativas do conjunto de dados são altamente correlacionadas, alguns tipos de técnicas de regressão pressupõem que isso pode causar um problema na análise do modelo de regressão (DISATNIK; SIVAN, 2016). Dessa maneira, havendo problema de multicolinearidade, e usando variáveis explicativas desnecessárias, isso pode ocasionar sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado (HAIR *et al.*, 2009).

Posto disso, existem na literatura diversas técnicas para lidar ou eliminar a multicolinearidade, para mais detalhes vide Imon e Khan (2003). Uma técnica usada para controlar o problema de *overfitting* é a regularização, nessa abordagem, se obtêm um processo de estimação que, além de apresentar pouca variabilidade, devido a um parâmetro de ajuste que penaliza coeficientes, também gere altas probabilidades para as estimativas dos $\hat{\beta}$ em que várias de suas componentes sejam próximas a zero. Nessa dissertação, temos o processo de estimação regularizada do tipo Ridge proposta por Hoerl e Kennard (1970) e o LASSO (*Least Absolut Shrinkage and Selection Operator*) proposta por Tibshirani (1996), onde têm-se um processo automático de seleção de variáveis para o modelo já que essa regularização tende a eliminar completamente os pesos das características menos importantes.

Essas regularizações, inicialmente, foram estabelecidos para resolver problemas de regressão, porém é possível ser adotado em problemas de classificação binária utilizando a regressão logística e, na estimação dos coeficientes do modelo, é quando entra o parâmetro de regularização. Engelstad *et al.* (2015) e Meier, Geer e Bühlmann (2008) descrevem o uso da regressão regularizada tipo LASSO como uma forma de classificação binária e Cessie e Houwelingen (1992) e Schaefer, Roi e Wolfe (1984) descrevem o uso da regressão de Ridge no âmbito da regressão logística.

Sendo assim, aqui, usaremos o conceito visto na seção 2.1, porém, nesse contexto, usaremos um modelo de regressão múltipla, sendo assim, estendendo a probabilidade condicional para o caso múltiplo, temos:

$$\mathrm{P}[Y_i = 1|x_i] = \frac{1}{1+\exp(-\beta^T x_i)} \tag{2.5}$$

No contexto múltiplo, o vetor $x_i = (x_{i1}, \ldots, x_{ij})^T$ é o vetor contendo as observações de cada variável do indivíduo i, sendo esse, um vetor linha pertencendo a matriz X.

Para estimar os parâmetros desconhecidos no modelo LASSO e no modelo Ridge é usado a seguinte minimização, respectivamente:

$$\widehat{\beta}_l = \underset{\beta}{argmin} \left\{ -\sum_{i=1}^{n} ln(1+\exp(-\beta^T x_i)) + \lambda \sum_{j=1}^{d} |\beta_j| \right\} \tag{2.6}$$

$$\widehat{\beta}_r = \underset{\beta}{argmin} \left\{ -\sum_{i=1}^{n} ln(1+\exp(-\beta^T x_i)) + \lambda \sum_{j=1}^{d} \beta_j^2 \right\} \tag{2.7}$$

onde, para as Eq. (2.6) e (2.7), $\widehat{\beta}_l$ é o parâmetro estimado usando a regularização LASSO e $\widehat{\beta}_r$ usando a regularização Ridge e $\beta = 1, 2, \vdots, p$, sendo p o número de variáveis.

Os dois métodos de regressão logística regularizada, tentam minimizar a equação de estimação dos parâmetros adicionando um termo de penalidade. As Eq. (2.6) e (2.7) possuem duas somas, a primeira soma, igual nas duas equações, é a log verossimilhança e a segunda soma, onde difere o método LASSO do Ridge, é a penalização associada a um parâmetro introduzido a fim de controlar o processo de estimação, a medida que λ aumenta, mais coeficientes são reduzidos a zero, sendo menos variáveis selecionadas e há mais encolhimento dos coeficientes não nulos. Aqui λ é o parâmetro de regularização que é um número não negativo, vale ressaltar que, se $\lambda = 0$ retorna as estimativas usuais de máxima verossimilhança para o modelo logístico

O grande desafio, nos dois métodos, é encontrar o melhor valor de λ. Sendo assim, para realizar essa tarefa, na literatura, encontra-se diferentes abordagens, nessa dissertação é utilizado a técnica para avaliar o quão bem um modelo pode ser generalizado para um conjunto de dados independente. Hastie, Tibshirani e Friedman (2009), propôs um um método chamado validação cruzada, em que os dados são particionados em k-subconjuntos de tamanhos aproximadamente iguais e um desses subconjuntos se torna o conjunto de validação, sendo assim, esse procedimento é repetido k vezes, sempre com um conjunto de validação diferente e, por fim, o valor ótimo de λ é estimado de tal forma que a probabilidade da log-verossimilhança da validação cruzada seja máximo (GOEMAN; MEIJER; CHATURVEDI, 2018). A validação cruzada pode ser usada para determinar qual abordagem é melhor em um conjunto de dados específico (JAMES *et al.*, 2013).

A partir da estimação dos parâmetros, a classificação segue de maneira similar a Regressão logística apresentada na seção 2.1.

2.3 MÁQUINAS DE VETORES DE SUPORTE

Uma máquina de vetores de suporte (SVM) desenvolvido por Vapnik (1995), é uma técnica de aprendizado de máquina que pode ser usada para classificação e análise de regressão (HEARST *et al.*, 1998). Esse método procura, com base nas informações oferecidas pela matriz de variáveis explicativas X, encontrar um hiperplano como uma superfície de decisão de modo que a margem de separação das classes seja máxima (JAMES *et al.*, 2013). Essa técnica visa a maximização da capacidade de generalização que, no conceito de aprendizado de máquina, é a capacidade da máquina classificar de forma eficiente dado o conjunto de treinamento visando minimizar a probabilidade de classificação errada de padrões que ainda não fora apresentados à máquina (HAYKIN, 2007). Esse método trabalha com os dados que podem ser separados de maneira linear e não linear. Desse modo, quando estamos lidando com dados linearmente separáveis, ou seja, quando é possível separar as classes por pelo menos um hiperplano. Seja X um conjunto de treinamento com n observações $x_i \in X$ com suas respectivas classes $y_i \in Y$ sendo $Y = \{-1, +1\}$, representa as classes linearmente separáveis.

A equação que separa os padrões através de hiperplanos pode ser definida como:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b = 0 \tag{2.8}$$

em que $\mathbf{w}$ é um vetor de pesos ajustáveis, $\mathbf{x}$ é o vetor de entrada que configura os padrões de entrada do conjunto de treinamento, de modo que, $\mathbf{w} \cdot \mathbf{x}$ é o produto escalar entre os vetores $\mathbf{w}$ e $\mathbf{x}$, sendo que $\mathbf{w} \in X$ é o vetor normal ao hiperplano e b é um limiar conhecido como viés, de tal forma que $b \in \mathbb{R}$ e $b/||w||$ corresponde à distância do hiperplano em relação à origem.

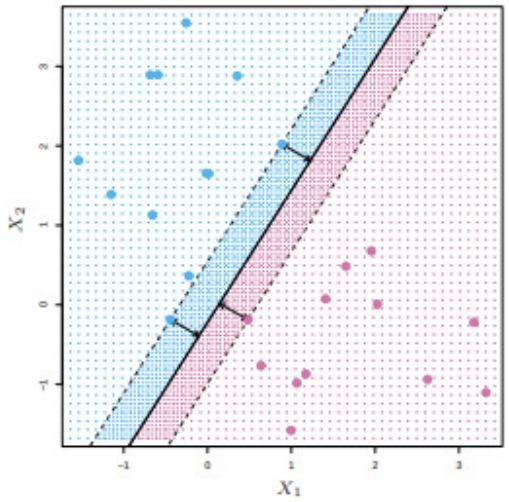
A Eq. (2.8) divide o espaço dos dados X em duas regiões e usa-se uma função sinal $g(x) = sgn(f(x))$ para obter a classificação, conforme, i. e,

$$g(\mathbf{x}) = sgn(f(\mathbf{x})) = \begin{cases} -1 \text{ se} & \mathbf{w} \cdot \mathbf{x} + b < 0 \\ +1 \text{ se} & \mathbf{w} \cdot \mathbf{x} + b \geq 0 \end{cases} \tag{2.9}$$

Sendo assim, a partir de $f(\mathbf{x})$ é possível obter um número finito de hiperplanos tendo em vista o resultado da multiplicação de $\mathbf{w}$ e b por uma constante (MULLER *et al.*, 2001). Então, para obter o melhor hiperplano para a separação deles, calcula-se a distância entre cada observação de treinamento e um hiperplano pré-fixado e, assim, escolhe-se o hiperplano que maximiza a margem de separação entre as diferentes classes dos dados (JAMES *et al.*, 2013).

Para mais detalhes sobre o método SVM ver Muller *et al.* (2001) e (SMOLA; SCHÖLKOPF, 2004).

Na Figura 1 podemos ver duas classes, cujas cores são azul e roxa. O hiperplano de margem máxima é mostrado na linha com cores mais solidas, desse modo, a margem é a distância da linha para qualquer uma das linhas tracejadas, os dois pontos azuis e o ponto roxa que estão nas linhas tracejadas são os vetores de suporte e sua distância até a margem estão indicadas por setas. A parte azul e roxa em destaque, indicam as regiões de decisão por um classificador com base nesse hiperplano.



Fonte: baseado em James *et al.* (2013).
Figura 1 – Método SVM para a escolha de hiperplano que maximiza a separação das classes

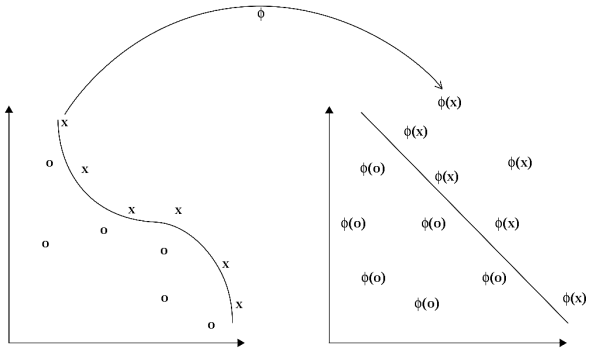
No entanto, o uso do SVM linear na prática é limitado, pois há muitos casos em que não é possível dividir os dados de treinamento por um hiperplano. Sendo assim, como forma de utilizar o SVM, mesmo para dados não lineares, realiza-se uma transformação nos dados originais a fim de construir um espaço de características em altas dimensões através de uma função de mapeamento não linear (por exemplo a distância de *Kernel* (*"kernel trick"*)) e, assim, é possível aplicar a técnica mencionada anteriormente (ÜSTÜN; MELSSEN; BUYDENS, 2006).

Como forma de realizar essa transformação, a distância de *Kernel* são recursos do SVM que auxiliam na transformação do algorítimo, do caso, não linear para o linear, ou seja, através dessa função, o SVM mapeia as classes não separáveis linearmente, no espaço original, para um espaço do *Kernel* onde, assim, essas classes conseguem ser separadas por um hiperplano (THISSEN *et al.*, 2004). Para o uso do SVM deve-se escolher a distância de *Kernel* que traga um melhor desempenho para o classificador. Neste trabalho, consideramos quatro núcleos de

Kernel a serem testados nos dados, a fim de encontrar a que aumente a acurácia do classificador. Sendo assim, na Tabela 0 temos as quatro possíveis distância de *Kernel*, onde γ e d são ajustados com base nos dados.

Kernel	Núcleo
Radial	$K(x_i, x_j) = \exp(-\gamma \|\|x_i - x_j\|\|^2)$
Linear	$K(x_i, x_j) = (x_i^T \bullet x_j)$
Polynomial	$K(x_i, x_j) = (\gamma x_i^T x_j)^d$
Sigmoid	$K(x_i, x_j) = tanh(\gamma x_i^T x_j)$

Tabela 0 – Tipos de distâncias *Kernel*



Fonte: baseado em Shawe-Taylor, Cristianini *et al.* (2004).
Figura 2 – Método SVM não linear

Na Figura 2, o lado esquerdo mostra o espaço de entrada onde as classes não são linearmente separáveis e através da função Φ, função de *Kernel*, cria o espaço das características onde, assim, é possível separar, de forma linear, as classes pelo hiperplano.

Kotsiantis, Zaharakis e Pintelas (2007) aborda em seu estudo que o SVM, apresentando somente variáveis de natureza contínua no banco de dados, tem apresentado melhor desempenho quando comparado à bancos com presença de muitas variáveis categóricas.

2.4 FLORESTA ALEATÓRIA

Floresta aleatória é um método que se baseia em árvores de decisão, onde múltiplas árvores são geradas e ao final elege a classe mais popular, de maneira a obter uma classificação com maior acurácia (BREIMAN, 2001). Essa técnica, além de muito usada em classificação,

também apresenta bom desempenho quando utilizada em regressão e estudo de importância (VERIKAS; GELZINIS; BACAUSKIENE, 2011).

Sendo assim, como forma de entender melhor esse método se faz necessário compreender de maneira mais generalizada, o método de árvore de decisão. Em que é um modelo não-paramétrico, com o objetivo de modelar relações complexas de um problema de classificação ou regressão. Além disso, essa técnica reproduz em certa medida a forma como as pessoas tomam decisões.

A árvore de decisão é um mapa dos resultados de escolhas relacionadas. Sendo assim, para começar elaborar a árvore precisamos determinar o nó raiz, que se divide em possíveis resultados. Na Figura 3, o primeiro círculo apresenta o nó raiz que é de onde se ramifica a árvore. Os nós estão associados as variáveis da matriz X presente no banco de dados e os ramos os possíveis resultados associados à essas variáveis. A medida que as decisões são tomadas a árvore vai ganhando forma até chegar ao nó folha que indica o resultado estimado da classificação.

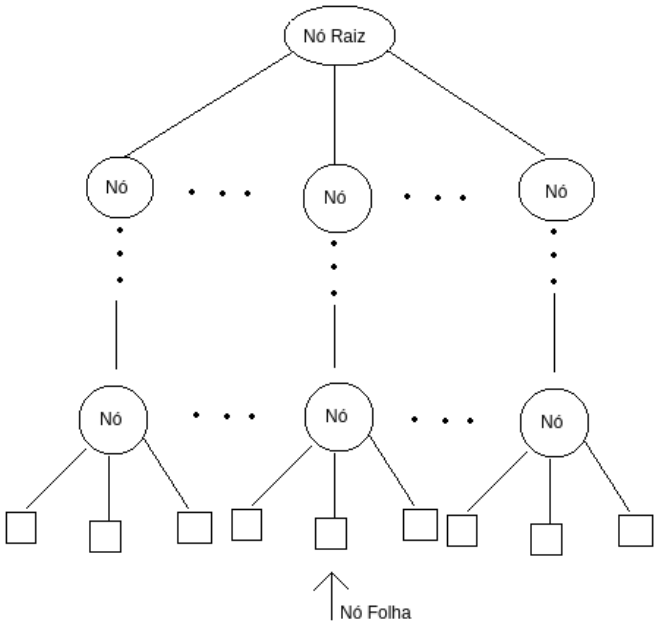


Figura 3 – Exemplo árvore de decisão.

Como mencionado anteriormente, as florestas aleatórias são uma extensão do método de árvore de decisão, onde tem-se um conjunto de árvores aleatórias não correlacionadas. A construção de uma floresta aleatória utiliza a técnica de bootstrap para criar o subconjunto de dados utilizados para a construção do crescimento das árvores, reamostrando de forma aleatória

e com reposição o vetor das classes Y e a matriz X de variáveis explicativas (BREIMAN, 2001). As árvore de decisão geralmente apresentam alta variabilidade e alto viés, dessa maneira, a floresta aleatória tenta encontrar um equilíbrio entre esses dois problemas e, além disso, ele também apresenta bom desempenho com conjuntos de dados de alta dimensão e com problemas de multicolinearidade (BELGIU; DRĂGUŢ, 2016). O resultado da classificação é alcançado a partir de um sistema de votação da classe mais popular entre as árvores que foram criadas (BREIMAN, 2001).

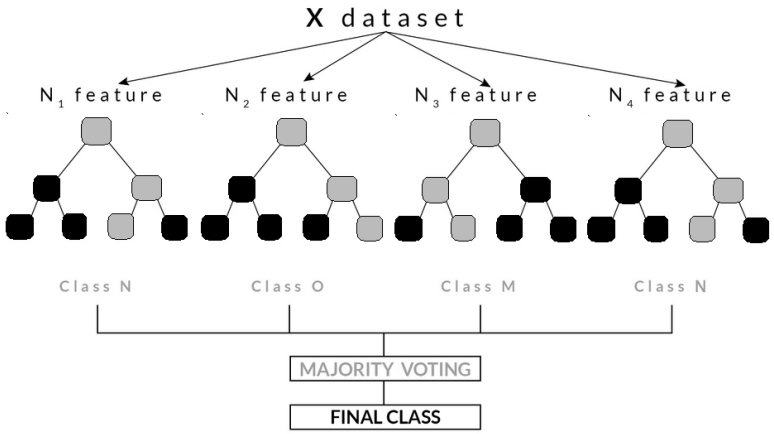


Figura 4 – Exemplo adaptado Florestas aleátorias baseado em Shagufta Tahsildar (2019)

Na Figura 4, podemos observar que cada árvore de decisão previu uma classe, a classe final que será selecionada pela floresta aleatória é a classe N, pois foi a classe presente em duas das quatro árvores de decisões existentes.

2.5 EXTREME GRADIENT BOOSTING

O Extreme Gradient Boosting (XGBoost) é um algoritmo utilizado no aprendizado de máquina que propõe uma melhoria do algoritmo *Gradient boosting* proposto por Friedman (2001). O XGBoost utiliza o conceito de árvore de decisão aliado a estrutura do *Gradient boosting*. Dessa maneira, esse método se basea na técnica conhecida como *Boosting*, que é uma metodologia cujos classificadores são re-amostrados com reposição, diversas vezes, porém, os dados re-amostrados são construídos de maneira que tenham obtido aprendizados com a classificação feita na amostragem anterior, como forma de obter o resultado final, após todas re-amostragens, utiliza-se um método de combinação ponderado pelo desempenho da classificação em cada modelo (GROVER, 2017).

Desenvolvido por Chen e Guestrin (2016), o XGBoost pode ser usado para diversas

finalidades, como exemplo, resolver problemas de regressão, classificação, entre outros. Como forma de entender esse classificador, é necessário entender o *Gradient boosting*, pois o XGBoost é o classificador *Gradient boosting* com técnicas combinadas de otimização de *software* e *hardware*, para que assim, possa reproduzir resultados superiores otimizando tempo e recursos computacionais.

De forma geral, a ideia principal do algoritmo é uma generalização do método *Adboost* proposto por Freund, Schapire *et al.* (1996), onde o objetivo consiste na combinação de uma série de classificadores considerados fracos e, após iterações e estimações, resultar em uma resposta conjunta mais assertiva.

No *Gradient boosting*, o objetivo de cada árvore de decisão construída é minimizar a função perda, isto é, minimizar o gradiente da função objetivo do modelo, em geral, quando se lida com problemas de classificação binária a função perda mais utilizada é a dos mínimos quadrados. Sendo assim, o XGBoost procura otimizar a função objetivo de forma mais robusta, esse difere de outras técnicas, pois usa um algoritmo mais sensível à dispersão na hora de ramificar as árvores (CHEN; GUESTRIN, 2016).

Dado um conjunto de dados de treinamento, com n observações e m váriaveis, sendo $D = \{\mathbf{x}_i, y_i\}$, onde $\mathbf{x}$ e y denotam as variáveis explicativas e as classes, respectivamente. O XGBoost utiliza K árvores para fazer a classificação, pois sendo um modelo iterativo, esse tende a melhorar, a cada passo, a árvore de decisão anterior, de tal forma que pode ser descrito como:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathbb{F}, \tag{2.10}$$

em que $\hat{y}_i$ é a predição da classe e f é uma função no espaço $\mathbb{F}$ de todas as possíveis árvores. A função objetivo, no XGBoost, é composta por dois elementos, sendo eles, a função perda e um termo de regularização, dessa maneira, a função objetivo é dada por:

$$\mathscr{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.11}$$

em que, l é a função perda e $\Omega(\cdot)$ é o termo de regularização que penaliza a complexidade do modelo e, nessa perspectiva, busca evitar que o modelo gere sobreajuste (*overfitting*).

Conforme a iteração é introduzida, a função objetivo é melhorada levando em consideração a árvore anterior, sendo assim, tem-se:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2.12}$$

em que $\hat{y}_i^{(t-1)}$ é a previsão da classe do i-éssimo elemento da amostra na t-éssima iteração e f_t são as funções de aprendizado que possuem a estrutura da árvore de decisão. Então, a cada iteração, o algoritmo tenta melhorar o modelo de árvore de decisão anterior (CHEN; GUESTRIN, 2016).

O algoritmo XGBoost é, atualmente, o mais robusto dentro do âmbito de aprendizado de máquina, devido a seu conjunto de parâmetros, que buscam controlar a complexidade e regularizar suas estimativas, a fim de melhorar a acurácia da classificação (GEORGANOS *et al.*, 2018; DONG *et al.*, 2018).

3 MOTIVAÇÃO

O presente capítulo apresenta a metodologia utilizada neste trabalho, bem como a descrição da base de dados utilizada e a extração de quantificadores de não paramétricos, além disso, os critérios de seleção das variáveis e também as métricas de avaliação do desempenho dos classificadores são apresentados

3.1 BANCO DE DADOS COM ASSINATURAS MANUSCRITAS

O MCYT-100 (MCYT Fingerprint subcorpus) é um banco de dados com assinaturas manuscritas e está disponível gratuitamente. Essa base de dados contém informações sobre 100 pessoas, para cada indivíduo foram capturadas 25 assinaturas verdadeiras. E também 25 falsificações são produzidas para cada usuário (ORTEGA-GARCIA *et al.*, 2003). A obtenção de cada assinaturas é feita de maneira *online* com o auxilio de um *tablet* com caneta eletrônica. A ferramenta usada para a obtenção desses dados é o WACOM ©, modelo INTOUS A6 USB. A resolução do tablet é de 100 linhas/mm e com precisão de ± 0.25 mm. Esse aparelho armazena em função do tempo discreto t, a posição $X(t)$ no eixo x, posição $Y(t)$ no eixo y, além disso, essa ferramenta também permite a captura da pressão aplicada na caneta p_t, posição azimutal γ_t e o ângulo da caneta em relação ao tablet ϕ_t. No presente trabalho, usaremos apenas funções de tempo $X(t)$ e $Y(t)$, pois essas quantidades são oferecidas pelos aparelhos de capturas de assinaturas dos mais simples aos mais sofisticados.

Como parte da coleção de assinaturas, os falsificadores, que geraram as assinaturas falsas, são altamente especializados, com o objetivo de reproduzir o mais próximo da dinâmicas natural das assinaturas verdadeiras (GARCIA-SALICETTI SONIA; HOUMANI; SCHEIDAT, 2009; ROSSO O A; OSPINA; FRERY, 2016). O número total de assinaturas é de 5000, sendo metade verdadeiras e a outra metade falsas.

Como apontado anteriormente, foi usado somente as séries temporais correspondentes às coordenadas x e y de cada assinatura. Nessa perspectiva, deve-se notar que, as séries temporais possuem tamanhos diferentes. Sendo assim, como maneira de facilitar as análises, fez-se um pré-processamento em cada série temporal, de tal forma que, as coordenadas foram redimensionadas para o quadrado unitário $[0,1] \times [0,1]$ e, por fim, considerando esses valores escalonados, o número total de dados em cada série é de M = 5000 pontos, isso foi feito usando o polinômio cúbico de Hermite para suavizar a assinatura. Dessa maneira, cada indivíduo k $(k = 1, \ldots, 100)$ associado à assinatura j $(j = 1, \ldots, 25)$, foram analisadas duas séries temporais

denotadas por $\mathbf{X}_j^{(k;\alpha)} = \{0 \leq \tilde{x}_{j;i}^{(k;\alpha)} \leq 1, i = 1,\ldots,M\}$ e $\mathbf{Y}_j^{(k;\alpha)} = \{0 \leq \tilde{y}_{j;i}^{(k;\alpha)} \leq 1, i = 1,\ldots,M\}$, em que o índice $\alpha = V, F$ se refere à assinatura verdadeira ou falsa, respectivamente, e $\tilde{x}$ e $\tilde{y}$ são os valores da interpolação (ROSSO O A; OSPINA; FRERY, 2016).

Na Figura 5, pode-se ver as assinaturas de três indivíduos, cujas azuis se referem as verdadeiras e a vermelha é um exemplo de assinatura falsa, além disso, é possível observar também a série temporal normalizada das coordenadas x e y.

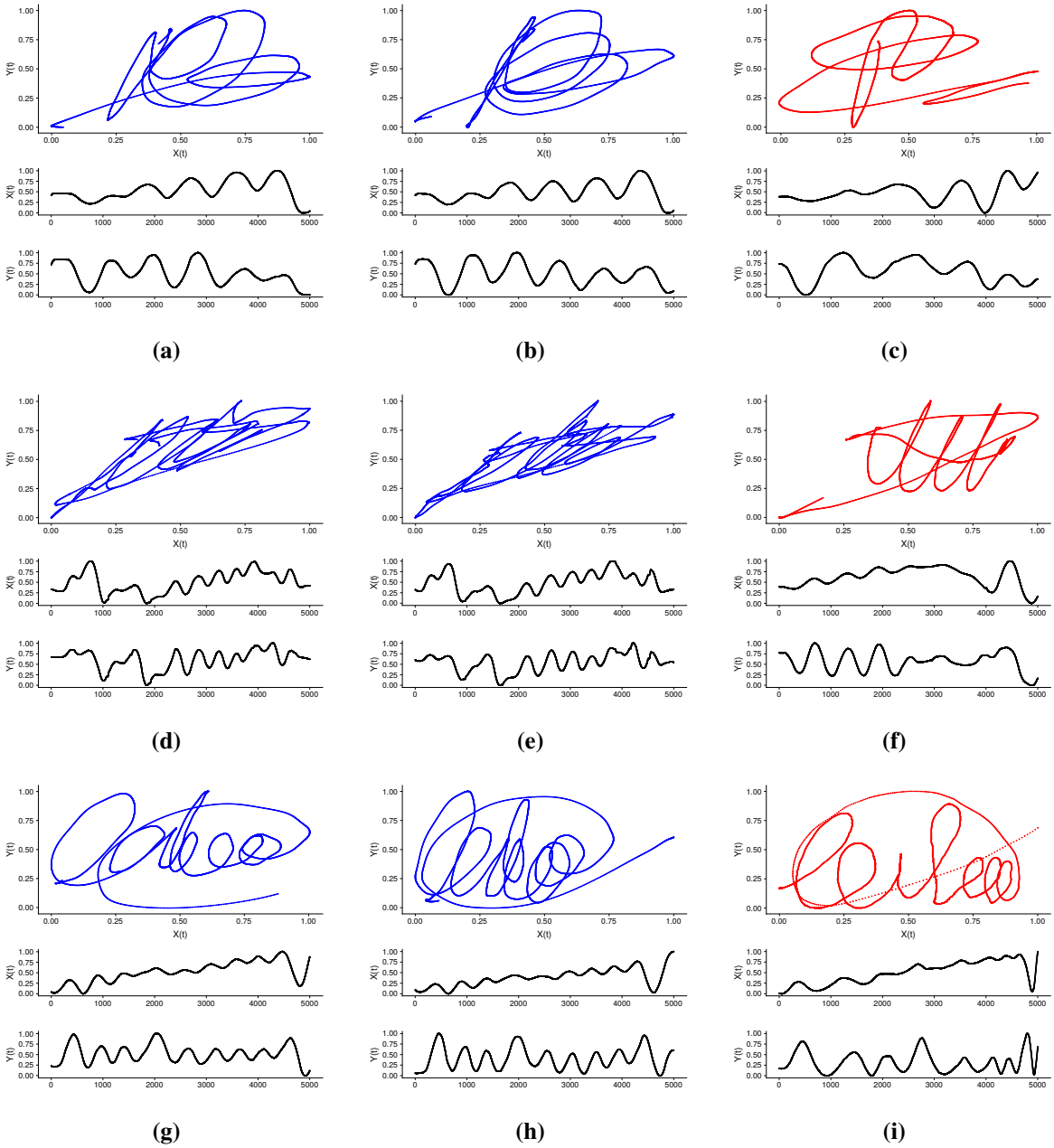


Figura 5 – Assinaturas de três indivíduos do banco MCYT-100. Duas verdadeiras (esquerda, azul) e uma falsa (direita, vermelho)

3.2 QUANTIFICADORES NÃO PARAMÉTRICOS

3.2.1 Avaliação da dinâmica das séries temporais das assinaturas

Na área da física, a dinâmica, ou seja, a velocidade e a aceleração são conceitos bastante difundidos. Dessa maneira, a velocidade está relacionada ao tempo que um corpo leva para percorrer um determinado espaço, enquanto a aceleração é a taxa de variação da velocidade em relação ao tempo, ou seja, é uma maneira de quantificar a variação da velocidade de um determinado objeto. Sendo assim, o uso dessas quantidade no presente trabalho pode trazer a tona novas intuições das séries temporais.

Posto disso, a velocidade e aceleração podem ser obtidas através da derivada da velocidade e da posição original dos dados, respectivamente (ASSUNçãO; TEIXEIRA; FARIA, 2009). Como forma de extrair mais informações sobre a dinâmica das assinaturas, para cada série temporal também será calculado a derivada de primeira e segunda ordem para cada série.

3.2.2 Quantificadores não paramétricos da teoria de informação

A dinâmica de sistemas é a busca por inferir a respeito de um determinado sistema não familiar analisando os registros sobre tal fenômeno ao longo de um tempo, ou seja, dado um sistema observável desse tipo, procura-se o quanto de informação pode ser extraído da dinâmica dele. Dessa maneira, utiliza-se a função de distribuição de probabilidade (FDP), como maneira de avaliação do conteúdo da informação de um sistema. Pode-se definir quantificadores não paramétricos da teoria de informação como medidas que envolvam FDP associadas as séries temporais (ROSSO O A; OSPINA; FRERY, 2016).

3.2.2.1 Entropia de Shannon, Informação de Fisher e Complexidade Estatística

A entropia é uma grandeza com muitas interpretações, sendo frequentemente utilizada na física para mensurar o grau de irreversibilidade de um sistema, geralmente associada ao grau de desordem, volume de espaço de estado e falta de informação. A entropia está presente em diversas áreas, por esse motivo tem-se a importância das diferentes interpretações desse conceito (BRISSAUD, 2005). A entropia de Shannon é considerada a medida de informação fundamental e mais natural (SHANNON, 1948).

Dada uma função de distribuição de probabilidade contínua (FDP), $f(x)$ com $x \in \Omega \subset \mathbb{R}$, $\int_\Omega f(x)dx = 1$, a entropia de Shannon (S) depende da $f(x)$ e denotada por $S[f]$ (SHANNON,

1948):

$$\mathrm{S}[f] = -\int_{\Omega} f(x)\ln[f(x)]dx. \tag{3.1}$$

$S[f]$ é considerado uma medida global, ou seja, a distribuição não é muito sensível à mudanças fortes que ocorrem em uma região pequena em Ω, o espaço de f. Por outro lado, a informação de Fisher $(F[f])$é sensível a pequenas alterações, e sua expressão:

$$F[f] = \int \frac{1}{f(x)}\left[\frac{df(x)}{dx}\right]^2 dx = 4\int \left[\frac{d\psi(x)}{dx}\right]^2, \quad onde \psi(x) = \sqrt{f(x)} \tag{3.2}$$

A informação de Fisher apresenta duas atribuições importantes na teoria da informação, seja a capacidade de estimar um parâmetro, seja de informação da variabilidade uma medida do estado de desordem de um sistema ou fenômeno (FRIEDEN, 2004). A propriedade mais importante dessa quantidade é o limite de Cramér-Rao. De fato, como parte de sua fórmula, a informação de Fisher depende do operador gradiente que influencia de maneira local as distribuições (FRIEDEN, 2004). Quando é relevante a noção de ordem nos cenários, se faz necessário o uso da sensibilidade local. Na Eq. (3.2), é possivel observar que não é conveniente a divisão por $f(x)$ se $f(x) \to 0$ para certos valores do suporte Ω (FRIEDEN, 2004). Como forma de evitar esse problema, pode-se usar as amplitudes de probabilidade reais através de uma expressão alternativa $\psi(x)$. Dessa maneira, não terá mais divisões e F mede o gradiente em $\psi(x)$ (FRIEDEN, 2004).

Seja $P = \{p_i; i = 1,\ldots,N\}$ com $\sum_{i=1}^{N} p_i = 1$ uma distribuição de probabilidade discreta, com N sendo o número de cenários possíveis no estudo, a medida de informação logarítmica de Shannon é dada por:

$$\mathrm{S}[P] = -\sum_{i=1}^{N} p_i \ln[p_i]. \tag{3.3}$$

Essa quantidade mede o grau de incerteza relativa à distribuição P e, desse modo, $p_1,\ldots,p_n$ é chamado entropia da distribuição P (SHANNON, 1948; RÉNYI *et al.*, 1961). Nessa perspectiva, se $\mathrm{S}[P] = S_{min} = 0$ sabemos quais dos possíveis resultados i, sendo a probabilidade associada p_i, irá ocorrer e, nesse caso, nosso conhecimento sobre o processo descrito por uma distribuição de probabilidade é máximo. No entanto, quando trata-se de uma distribuição uniforme nosso conhecimento é mínimo, visto que cada resultado apresenta a mesma probabilidade

de ocorrência, ou seja, $P_e = \{p_i = 1/N; i = 1,\ldots,N\}$ logo, $S[P_e] = S_{max} = \ln N$ e a incerteza é máxima(ROSSO O A; OSPINA; FRERY, 2016). Para o caso discreto, é definido a Entropia de Shannon normalizada, e segue

$$H[P] = \frac{S[P]}{S_{max}} \tag{3.4}$$

para $0 < H < 1$.

No que tange a informação de Fisher, existem algumas expressões discretizadas que podem ser utilizadas para seu cálculo, ainda que não apresentem os mesmo resultados (SÁNCHEZ-MORENO P; YÁNEZ; DEHESA, 2009). No trabalho, adotaremos como medida de informação de Fisher normalizada e discretizada $F[P]$ (OLIVARES; PLASTINO; ROSSO, 2012b), definida por:

$$F[P] = F_0 \sum_{i=1}^{N-1} [\sqrt{p_{i+1}} - \sqrt{p_i}]^2. \tag{3.5}$$

que deriva da Eq. (3.2), para a constante de normalização F_0 segue

$$F_0 = \begin{cases} 1, & \text{Se } p_i^* = 1 \text{ para } i^* = 1 \text{ ou } i^* = N \text{ e } p_i = 0, \forall i \neq i^*, \\ 1/2, & \text{caso contrário.} \end{cases} \tag{3.6}$$

A complexidade está presente em diversas áreas, contudo é difícil encontrar uma forma de defini-la quantitativamente. Isso se da pelo fato que, não há uma definição universal para essa quantidade (MARTIN; PLASTINO; ROSSO, 2006). Nesse sentido, há na literatura algumas maneiras para o cálculo da complexidade. Para o presente trabalho, pode-se entender um sistema complexo quando esse não apresenta padrões considerados simples, então a medida de complexidade pode ser definida por uma quantidade que dependa da medida de informação e uma medida de desequilíbrio, dessa forma(LOPEZ-RUIZ; MANCINI; CALBET, 1995):

$$C[P] = Q_j[P, P_e].H[P] \tag{3.7}$$

em que H é a entropia normalizada de Shannon, vista em Eq. (3.4), e o desequilíbrio Q_j é definido como uma função da divergência de Jensen-Shannon $J[P, P_e]$, usada para quantificar a divergência entre distribuições de probabilidade (GROSSE *et al.*, 2002). De tal forma se P é a probabilidade associada ao sistema em análise e P_e é a distribuição uniforme, então

$$Q_j[P, P_e] = Q_0 J[P, P_e] = Q_0\{S[(P + P_e)/2] - S[P]/2 - S[P_e]/2\}, \tag{3.8}$$

para $0 \leqslant Q_j \leqslant 1$ e Q_0 é uma constante de normalização que pode ser definida como:

$$Q_0 = -2\left\{\frac{N+1}{N}ln(N+1) - ln(2N) + ln(N)\right\}^{-1}, \tag{3.9}$$

vale destacar que, para um valor de H fixo, existe uma parcela de valores possíveis para C, assim, a complexidade pode fornecer informações adicionais sobre mais detalhes da distribuição de probabilidade que não foram captados pelo uso da entropia (ROSSO; MASOLLER, 2009).

Nessa perspectiva, se o sistema estiver em um estado muito ordenado, ou seja, a maior parte dos valores de p_i estiverem estão próximos à zero, então a entropia normalizada de Shannon e a complexidade estatística estarão próximas à zero ($H \approx 0$ e $C \approx 0$), por outro lado, a informação de Fisher normalizada estará próxima à um ($F \approx 1$). Em contraste, se o sistema se encontra em estado muito desordenado, o que ocorre quando os valores de p_i estão em torno de um mesmo valor, tem-se que $H \approx 1$, enquanto $C \approx 0$ e $F \approx 0$. Pode-se afirmar que o comportamento, em geral, da medida discreta da informação de Fisher, apresentada na Eq. (3.5), é oposto ao da entropia de Shannon (ROSSO *et al.*, 2010). A sensibilidade da informação de Fisher discretizada é pautada no caso em que a ordenação dos valores de p_i devem ser considerados no cálculo da Eq. (3.5) (OLIVARES; PLASTINO; ROSSO, 2012a; OLIVARES; PLASTINO; ROSSO, 2012b). Caso a ordenação não seja seriamente levada em conta, ocorrerá um valor da informação de Fisher diferente. Sendo assim, de forma a estabelecer essa ordenação para avaliação, seguiremos a abordagem de Bandt e Pomp (BANDT; POMPE, 2002).

3.2.2.2 Abordagem de Bandt e Pompe para a determinação da FDP

O uso das características mencionadas em 3.2.2.1 pressupõe algum conhecimento prévio sobre o sistema em estudo, efetivamente, é necessário fornecer a distribuição de probabilidade mais adequada associada à série temporal (ROSSO *et al.*, 2013). Sendo assim, algumas metodologias estão disponíveis para determinar a função de distribuição de probabilidade a partir das séries temporais. Nessa dissertação, usaremos a abordagem Bandt e Pompe porposta por Bandt e Pompe (2002), é utilizada para se obter uma distribuição de probabilidade P, sendo uma metodologia simbólica simples e robusta que destaca-se levando em consideração a causalidade da série temporal (BANDT; POMPE, 2002). Um resultado notável de Bandt e Pompe é uma melhoria no desempenho dos quantificadores de informação obtido usando seu algoritmo de geração de P (KOWALSKI *et al.*, 2007). Os dados simbólicos são utilizados para classificar os valores da série e definir a reordenação dos dados inseridos em ordem crescente, equivalendo a

uma reconstrução do espaço de fase com um comprimento padrão D e *lag* τ. Dessa maneira, é possível quantificar a diversidades dos padrões oriundo de uma série temporal (ROSSO *et al.*, 2013).

É importante destacar que não é necessária nenhuma suposição sobre o modelo para que se obtenha a sequência de símbolos apropriada, ou seja, as "partições" necessárias são feitas comparando a ordem dos valores relativos com seus vizinhos. Dentro desse aspecto, é importante destacar que a abordagem leva em consideração a estrutura temporal da série gerada pelo processo (ROSSO *et al.*, 2013).

Para o uso dessa metodologia, para encontrar a FDP P deve-se levar em consideração a causalidade da série temporal comparando seus valores com o vizinho mais próximo da série temporal. Os dados simbólicos são criados reordenando os valores das séries temporais em ordem crescente, ou seja, a uma reconstrução do espaço de fase com dimensão de incorporação D, onde D é a partição do espaço em que contém detalhes relevantes da estrutura ordinal da série temporal apresentada (comprimento do padrão). Tomamos $\mathscr{X}(t) = \{x_t; t = 1, \ldots N\}$ com dimensão de incorporação $D > 1$ ($D \in \mathbb{N}$) e o *lag* τ ($\tau \in \mathbb{N}$). Estamos buscando padrões ordinais de comprimento D que são gerados por:

$$(s) \mapsto (x_{s-(D-1)\tau}, x_{s-(D-2)\tau}, \ldots, x_{s-\tau}, x_s), \tag{3.10}$$

para cada tempo s especificamos um vetor D-dimensional para os tempos $s, s-\tau, \ldots, s-(D-2)\tau, s-(D-1)\tau$, notoriamente quanto maior o valor de D, mais informações sobre o passado são agregadas à esses vetores. O padrão ordinal D associado ao tempo s, são permutações $\pi = (r_0, r_1, \ldots, r_{D-1})$ de $[0, 1, \ldots, D-1]$, dessa forma,

$$x_{s-r_{D-1}\tau} \leqslant x_{s-r_{D-2}\tau} \leqslant \ldots \leqslant x_{s-r_1\tau} \leqslant x_{s-r_0\tau}, \tag{3.11}$$

e para não ter empates, definimos $r_i < r_{i+1}$ se $r_{s-r_i} = r_{s-r_{i-1}}$.

Portanto, para todo $D!$, temos π ordenações possíveis de ordem D e $lag\tau$, a distribuição de probabilidade do padrão ordinal ($P = \{p(\pi)\}$) é dada pela frequência relativa do número de vezes que a sequência com um tipo de padrão foi encontrada na série temporal dividida pelo número total de sequências, ou seja

$$p(\pi_i) = \frac{\#\{s | s \leqslant N-(D-1)\tau; (s) \text{ é do tipo } \pi_i\}}{N-(D-1)\tau}, \tag{3.12}$$

em que # é a cardinalidade do conjunto e $P = \{p(\pi_i), i = 1, \ldots, D!\}$ é obtida a partir da série temporal.

A metodologia de Bandt e Pompe pode ser aplicada para qualquer tipo de série temporal, porém é necessário um suposição de estacionariedade fraca (BANDT; POMPE, 2002). É importante destacar que, para o uso dessa abordagem, a incorporação de D é fundamental na avaliação da distribuição de probabilidade adequada (KOWALSKI *et al.*, 2007). Kowalski *et al.* (2007) propõe que a condição $N \gg D$ seja satisfeita, assim, é conveniente que se trabalhe com $3 \leqslant D \leqslant 7$.

No estudo de sistemas dinâmicos não lineares, essa proposta de associar distribuições de probabilidade a séries temporais proporcionou um avanço significativo no entendimento dos sistemas (ROSSO O A; OSPINA; FRERY, 2016).

Para utilizar esse método de extração de função de distribuição de probabilidade de séries temporais, primeiro divide-se o intervalo $[a,b]$ em infinitos sub-intervalos de tamanhos iguais e que não estejam sobrepostos e, dessa forma, com essa divisão, tem-se a sequência de ordem natural para avaliar o gradiente presente na medida de informação de Fisher.

3.2.3 Estatística não paramétrica de Wallis e Moore

A tendência é uma característica frequentemente encontrada em séries temporais, dessa maneira, quando uma série possui alguma inclinação, dizemos que esta apresenta uma tendência (MORETTIN; TOLOI, 2006). Na literatura, existem alguns testes não paramétricos utilizados para detectar se há tendência em um conjunto de dados. Sendo assim, Wallis e Moore (1941) propôs um teste fundamentado em ordem, cujas principais vantagens são a simplicidade e ausência de suposições sobre a forma da população. Esse teste é baseado em sequências dos sinais obtidos pela diferença entre observações sucessivas, sobretudo testa a aleatoriedade da distribuição dessas sequências por comprimento. Cada ponto cuja série analisada deixa de decrescer e começa a crescer, ou o contrário, é chamado de ponto de virada. Desse modo, o intervalo entre pontos de virada consecutivos são chamados de fase. A hipótese nula consiste na ideia de que a série compreende dados aleatórios, onde cada ordenação é igualmente provável, em que sua média e variância são obtidas de forma simples e a distribuição é assintoticamente normal (STUART, 1952). A estatística de teste de frequência para $N \geq 30$, sendo N o tamanho da série, é dada por:

$$W = \frac{\left| h - \frac{2N-7}{3} \right|}{\sqrt{\frac{16N-29}{90}}} \tag{3.13}$$

em que, h denota o número de fases, sendo a primeira e a última não sendo computadas. Para séries temporais com $N < 30$ uma correção de continuidade $-0,5$ deve ser incluída no denominador.

3.3 CONSISTÊNCIA DAS CARACTERÍSTICAS

O processo de extração de características das séries temporais têm um papel importante, entretanto é necessário verificar se essas quantidades são, de fato, representativas. Por esse aspecto, uma medida para auxiliar sobre a qualidade desses atributos é a consistência, proposta por Lee, Berger e Aviczer (1996). Definida como:

$$d_s(i) = \frac{|\mu_{s;1} - \mu_{s;0}|}{\sqrt{\sigma^2_{s;1} + \sigma^2_{s;0}}}, \quad s = 1, \ldots, T. \tag{3.14}$$

onde, T é o número total de características extraídas das séries temporais, $d_s(i)$ é a consistência da característica s para o indivíduo i, $\mu_{s;1}$ é a média da característica s somente para as assinaturas verdadeiras, enquanto $\mu_{s;0}$ é essa medida para as falsas. $\sigma^2_{s;1}$ e $\sigma^2_{s;0}$ representam a variância na amostra para as assinaturas verdadeiras e falsas, respectivamente. Essa medida resulta em N medidas de consistência para cada características, onde N é o número de indivíduos no estudo, sendo assim, é calculado a média e o desvio padrão entre os N indivíduos da característica s , a fim de consolidar uma medida de consistência para cada característica (ANTAL; SZABÓ, 2017).

No que diz respeito a consistência, uma boa característica deve ter média alta e desvio padrão baixo, dessa maneira (ANTAL; SZABÓ, 2017).

3.4 SELEÇÃO DAS CARACTERÍSTICAS

Durante a condunção do estudo é necessário verificar quais dos quantificadores não paramétricos, entre os extraídas, são realmente relevantes. Nesse sentido, a presença de características irrelevantes prejudica o desempenho dos algoritmos utilizados na classificação, não só na capacidade de predizer corretamente, mas também no desempenho computacional (WITTEN *et al.*, 2016).

Dessa maneira, como forma de utilizar as características que sejam relevantes, foram utilizadas algumas técnicas de seleção que são descritas a seguir.

3.4.1 Ganho de Informação

Esse método é constantemente utilizado no campo de aprendizado de máquina, tem como objetivo principal a seleção de variáveis que melhor conseguem separar as classes (YANG; PEDERSEN, 1997), no presente estudo, separar assinaturas falsas das verdadeiras.

O ganho de informação (G. I), através da entropia, se baseia em uma medida de impureza nos dados. Desse modo, a métrica avalia a redução da entropia causada após separar a classe de acordo com os valores das características. O ganho é dado pela Eq. (3.15)

$$GI(s) = H(\mathrm{s}) - H(\mathrm{s}, \mathrm{Classe}) \tag{3.15}$$

em que s representa a característica, tal que, $H(\mathrm{s})$ é entropia de Shannon para a característica s e $H(\mathrm{s}, \mathrm{Classe})$ é a entropia de Shannon para a característica s após a separação das classes ser considerada (DAI; XU, 2013), essa quantidade está definida na Eq.(3.3). Como critério de avaliação é usado a razão do ganho (R. G), que é relativizar o G. I e é definido como:

$$RG(s) = \frac{GI(s)}{H(s)} \tag{3.16}$$

No que tange a escolha das características a serem utilizados, será considerado apenas as características que obtiverem RG acima da média (ZUBEN; ATTUX, 2001).

3.4.2 Análise de variância

Segundo Hair *et al.* (2009), a análise de variância (ANOVA) é uma técnica estatística que tem como principal objetivo determinar se as amostras de dois ou mais grupos são provenientes de populações com médias iguais em função de uma variável dependente. Dessa maneira, essa metodologia foi utilizada no presente trabalho, de tal forma que somente variáveis que fossem estatisticamente significantes aos níveis de 1% e 5% estariam presentes nestes modelos, para essa análise, foi utilizado o modelo de regressão logístico para o ajuste.

3.4.3 Fator de inflação de variância

O fator de inflação de variância (VIF) é uma medida de tolerância utilizada para medir o grau de multicolinearidade entre variáveis. Nesse sentido, se a correlação é elevada entre

dois ou mais regressores a estimação dos parâmetros pode ser comprometida. Na literatura, é indicado que o VIF da variável não exceda a 10 (O'BRIEN, 2007). Posto disso, para o nosso estudo selecionaremos o conjunto de características que apresentarem o VIF menor que 10, para isso, usaremos o modelo de regressão logístico para o ajuste.

3.5 AVALIAÇÃO DO DESEMPENHO

As métricas de avaliação são parte fundamental em problemas de classificação. O objetivo, em geral, é comparar a classe estimada pelo modelo em relação a classe verdadeira y. Sendo assim, utilizando essas quantidades, é possível avaliar a precisão dos algoritmos de classificação e, assim, escolher o que apresenta melhor desempenho segundo essas. As métricas servirão como critério de escolha entre vários classificadores candidatos.

3.5.1 Acurácia, Sensibilidade e Especificidade

Como forma de entender o conceito de acurácia, especificidade e sensibilidade, é necessário apresentar a matriz de confusão, que é uma tabela representando as frequências de classificação para cada classe. A Tabela 1, mostra a estrutura de uma matriz de confusão em que as linhas representam: Verdadeiros-negativos (VN) que correspondem a classe 0 que foram classificados corretamente, falsos-positivos (FP) cuja a classe era 0, porém foram classificados como 1, falsos-negativos (FN), que, ao contrário de FP, são da classe 1, no entanto foram classificados como 0 e, por fim, verdadeiros-positivos (VP) se refere aos valores da classe 1 que foram classificados corretamente.

Classe estimada ($\hat{y}$)	Classe observada (y)	
	Falsa (0)	Verdaderia (1)
Falsa (0)	VP	FP
Verdaderia (1)	FN	VN

Tabela 1 – Matriz Confusão

A acurácia fornece o acerto médio global entre as classificações, dado por,

$$Acc = \frac{VP + VN}{VN + FP + FN + VP} \tag{3.17}$$

A acurácia distingue erros e acertos levando em consideração toda a informação da matriz de confusão, em contrapartida quando se deseja observar o desempenho relativo de um

classificador em relação a cada classe existente, utiliza-se a sensibilidade e a especificidade. A sensibilidade mede o percentual de verdadeiros-positivos que foram classificados na classe dos positivo, por outro lado, a especificidade considera os verdadeiros-negativos classificados na classe dos negativos (ZHU *et al.*, 2010). Sendo assim, temos respectivamente,

$$Sens = \frac{VP}{VP+FN} \tag{3.18}$$

$$Esp = \frac{VN}{VN+FP}. \tag{3.19}$$

3.5.2 Área sob a curva ROC (AUC)

A área sob a curva ROC (AUC) (HANLEY; MCNEIL, 1982), é uma métrica que é bastante utilizada na área de *machine learning* (FAWCETT, 2006). Essa métrica mostra o quão bom o classificador pode distinguir entre as duas classes. A curva ROC depende dos valores de sensibilidade e especificidade e, a partir desses valores, é possível traçar o gráfico dessa curva. Como forma de facilitar a análise, é possível resumir a informação da curva em um único valor, chamado AUC, que é a área sob a curva ROC. O AUC está limitado no intervalo $[0,1]$ e, assim, é possível utiliza-lo como critério de comparação entre o desempenho de diferentes classificadores, de tal forma que um classificador com melhor desempenho apresenta AUC próximo à 1.

3.5.3 Taxa de erro de classificação

A partir da matriz de confusão, é possível determinar a taxa de erro de classificação (ER) que é uma quantidade difundida na área de *machine learning* e pode ser calculada, a partir da matriz de confusão, como:

$$ER = \frac{FP+FN}{VN+FP+FN+VP}. \tag{3.20}$$

No que se refere essa taxa, quando se trata de classificação desbalanceada, ou seja, quando o número de observações em cada classe é desbalanceado, a taxa de erro de classificação não é uma métrica robusta, por exemplo, supondo que 99% das observações pertença a uma classe. Dessa maneira, pode-se obter com frequência uma taxa de erro baixa, visto que um classificador que sempre retorna a classe majoritária apresentará uma taxa de erro de 1%, porém um modelo que classifica uma nova observação na classe majoritária não conseguirá classificar corretamente quando a classe for a minoritária (BATISTA; PRATI; MONARD, 2004). Entretanto,

no caso dessa dissertação, as classes estão balanceadas, pois apresentam, a mesma quantidade de observações nas duas classes para construção.

3.6 DETALHES COMPUTACIONAIS

Para o desenvolvimento dessa dissertação, a implementação computacional foi executada na linguagem de programação R (R Core Team, 2019) em uma máquina com processador `Intel Core (TM) i3-6006U CPU 2.00GHz` e com `4GB` de memória `RAM`, usando um sistema operacional `Linux` de `64 bits`.

A Tabela 2 apresenta os pacotes e funções que foram utilizados em cada método na etapa de classificação. Após o ajuste do modelo, como forma de obter a predição da variável resposta, foi utilizada a função `predict` disponível no pacote `stats`.

Classificador	Pacote	Função	Sitio
R.L	`stats`	`glm`	*stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html*
SVM	`e1071`	`svm`	*cran.r-project.org/web/packages/e1071/index.html*
F.A	`randonForest`	`randonForest`	*cran.r-project.org/web/packages/randomForest/index.html*
XG	`xgboost`	`xgboost`	*cran.r-project.org/web/packages/xgboost/index.html*
LASSO e ridge	`glmnet`	`cv.glmnet`	*cran.r-project.org/web/packages/glmnet/index.html*

Tabela 2 – Pacotes e funções utilizados para classificação

Para todos os indivíduos do banco de dados tem-se o esquema presente nas Figuras 6 e 7. Desse modo, a primeira se refere a etapa de classificação utilizando critérios de seleção como uma maneira de obter conjuntos de características importantes para as análises, por outro lado, a segunda figura apresenta o esquema em que a regressão logística penalizada é usada como forma de seleção de variáveis relevantes e também na classificação dos dados.

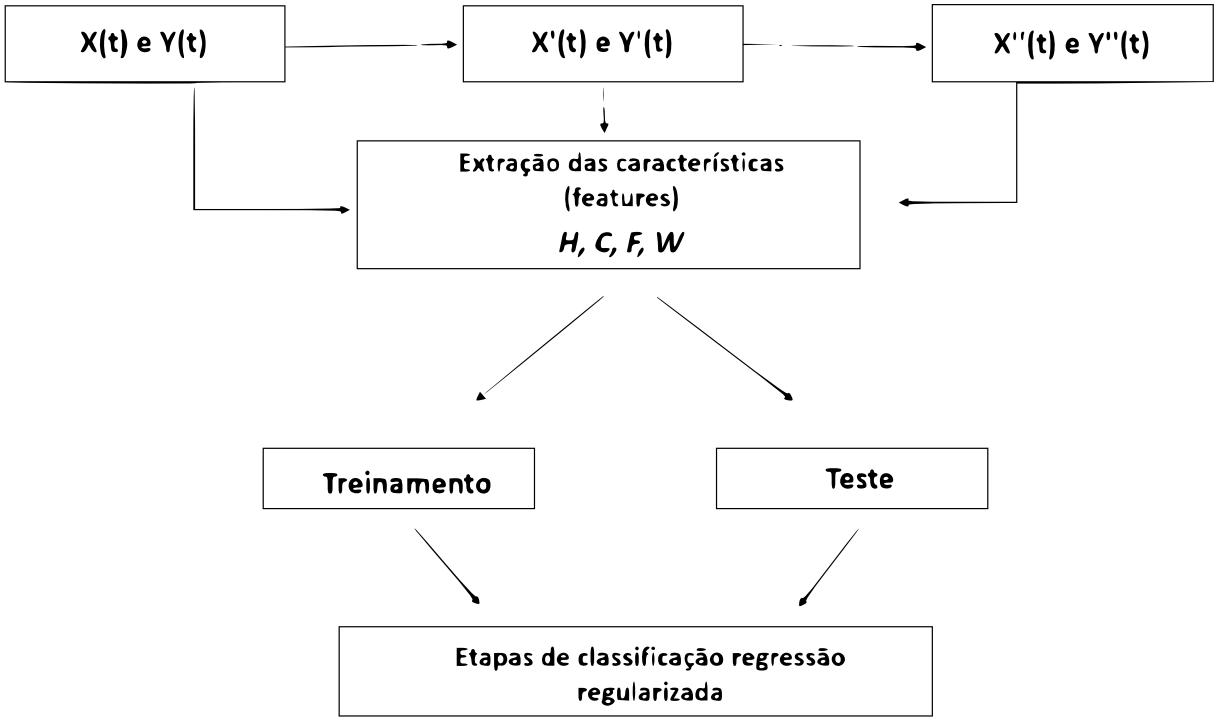


Figura 6 – Esquema da dissertação de acordo com os critérios de seleção

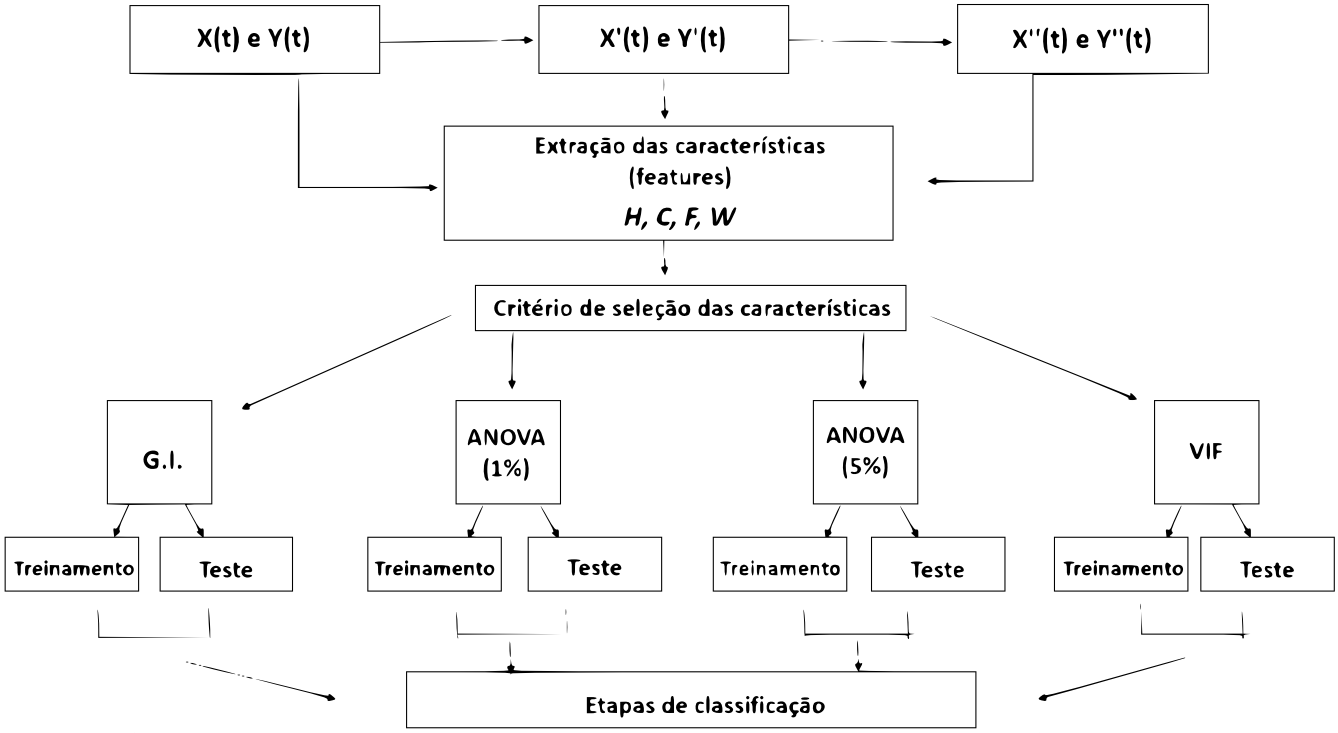


Figura 7 – Esquema da dissertação de acordo com a regressão penalizada

4 RESULTADOS

Neste capítulo apresentamos os resultados computacionais aplicados ao banco de dados utilizando a metodologia descrita nessa dissertação. Primeiramente, apresenta-se uma análise descritiva dos quantificadores de informação, esses mencionados na seção 3.2, posteriormente tem-se a classificação binária para os dados segundo a seleção de características descritas na seção 3.4 e, por fim, a classificação utilizando a regressão regularizada com todas as características extraídas das séries temporais.

4.1 ANÁLISE DESCRITIVA DAS CARACTERÍSTICAS

Para uma melhor compreensão do comportamento dos quantificadores, apresenta-se uma análise descritiva das características extraídas das séries temporais, vale ressaltar que, todas essas foram obtidas através das séries temporais padronizadas para que essas ficassem na mesma escala, sendo assim, para cada característica foi subtraída sua média e dividida pelo desvio padrão da mesma. As Tabelas 3,4, 5 e 6 contém a média ($\hat{\mu}$), desvio padrão ($\hat{\sigma}$) e a mediana ($\hat{m}$) para os quantificadores entropia, complexidade, informação de Fisher e a estatística de Wallis e Moore, respectivamente em sua forma original, primeira derivada e segunda derivada para os eixos x e y provenientes da série temporal de cada assinatura e, dessa forma, cada tabela apresenta as medidas separadas por assinaturas falsas (0) e verdadeiras (1).

É possível observar que os valores de média e mediana em todas as tabelas se diferenciam em relação a magnitude quando se compara em relação à autenticidade das assinaturas, ou seja, assinaturas verdadeiras e falsas apresentam direções opostas, além disso, se tratando do desvio padrão, nota-se que quando se refere à assinaturas verdadeiras, em todas as tabelas, essas apresentam menor variabilidade quando comparadas as falsas.

Características	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{m}_0$	$\hat{m}_1$
$X(t)$	0,3322	-0,3322	1,1723	0,6366	0,1320	-0,4275
$Y(t)$	0,3458	-0,3457	1,1781	0,6108	0,1265	-0,4471
$X'(t)$	0,3337	-0,3337	1,1330	0,7027	0,1985	-0,4236
$Y'(t)$	0,3504	-0,3503	1,1320	0,6879	0,2265	-0,4502
$X''(t)$	0,3161	-0,3161	1,0946	0,7760	0,2628	-0,4131
$Y''(t)$	0,3176	-0,3176	1,0939	0,7758	0,2555	-0,4169

Tabela 3 – Médias, desvios padrão e mediana da Entropia

A Tabela 7 refere-se a medida de consistência onde foi calculado os valores de média das características originais ($\hat{\mu}$), desvio padão das características originais ($\hat{\sigma}$), média

Características	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{m}_0$	$\hat{m}_1$
$X(t)$	0,3313	-0,3313	1,1258	0,7163	0,2203	-0,4210
$Y(t)$	0,3470	-0,3470	1,1319	0,6915	0,2060	-0,4471
$X'(t)$	0,3270	-0,3270	1,0748	0,7944	0,3050	-0,3891
$Y'(t)$	0,3485	-0,3485	1,0731	0,7782	0,3447	-0,4302
$X''(t)$	0,2997	-0,2997	1,0424	0,8566	0,3670	-0,3427
$Y''(t)$	0,3017	-0,3017	1,0426	0,85510	0,3677	-0,3463

Tabela 4 – Médias, desvios padrão e mediana da Complexidade

Características	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{m}_0$	$\hat{m}_1$
$X(t)$	-0.3145	0.3145	1.0976	0.7730	-0.2894	0.3253
$Y(t)$	-0.3336	0.3336	1.1034	0.7484	-0.2820	0.3696
$X'(t)$	-0.3112	0.3112	1.0891	0.7876	-0.2831	0.3139
$Y'(t)$	-0.3284	0.3284	1.0826	0.7825	-0.2929	0.3493
$X''(t)$	-0.2821	0.2821	1.0092	0.9069	-0.4471	0.2863
$Y''(t)$	-0.2826	0.2826	1.0068	0.9093	-0.4445	0.2684

Tabela 5 – Médias, desvios padrão e mediana da Informação de Fisher

Características	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{m}_0$	$\hat{m}_1$
$X(t)$	-0.3252	0.3252	1.2186	0.55097	-0.0393	0.4268
$Y(t)$	-0.3391	0.3391	1.2261	0.5165	-0.0406	0.4502
$X'(t)$	-0.3507	0.3507	1.1952	0.5706	-0.0878	0.4621
$Y'(t)$	-0.3586	0.3586	1.1971	0.5566	-0.1080	0.4731
$X''(t)$	-0.3457	0.3456	1.1670	0.6318	-0.1355	0.4718
$Y''(t)$	-0.3425	0.3424	1.1676	0.6342	-0.1241	0.4709

Tabela 6 – Médias, desvios padrão e mediana da Estatística de Wallis e Moore

da primeira derivada das características ($\hat{\mu}'$), desvio padrão da primeira derivada das variáveis ($\hat{\sigma}'$), média da segunda derivada das variáveis ($\hat{\mu}''$) e o desvio padrão da segunda derivada das características ($\hat{\sigma}''$), para as características obtidas a partir do eixo x e y, de tal forma que H, C, F e W se referem, respectivamente à entropia, complexidade, informação de Fisher e estatística de Wallis e Moore. Segundo Antal e Szabó (2017), uma característica estável deve apresentar média alta e desvio padrão baixo, desse modo, analisando a Tabela 7 pode-se observar que as características não apresentam grande variação e, assim, apresentam valores de média e desvio padrão próximos entre si, porém, em geral, a estatística de Wallis e Moore foi a que apresentou a menor média em todas as formas obtidas, no entanto essa diferença é compensada pelo seu desvio padrão sendo o menor em todas as categorias.

Como forma de visualizar melhor como as características se comportam, a Figura 8 apresenta os gráficos da densidade marginal da entropia e complexidade em sua forma original, primeira derivada e segunda derivada separadas por classe das assinaturas em ambas coordenadas x e y, além disso, pode-se observar também o gráfico de dispersão dessas características referentes

	$\hat{\mu}$	$\hat{\sigma}$	CV	$\hat{\mu}'$	$\hat{\sigma}'$	CV'	$\hat{\mu}''$	$\hat{\sigma}''$	CV''
H_x	0.9365	0.5277	0.5634	1.0759	0.6007	0.5819	1.1688	0.6555	0.5608
H_y	0.9619	0.5152	0.5356	1.1378	0.5918	0.5201	1.1713	0.6559	0.5599
C_x	0.9773	0.5878	0.6014	1.1320	0.6758	0.5969	1.2344	0.7583	0.6143
C_y	1.0038	0.5747	0.5725	1.1978	0.6734	0.5621	1.2386	0.7519	0.6070
F_x	0.9799	0.5943	0.6064	1.0567	0.6132	0.5802	1.2080	0.7622	0.6309
F_y	1.0156	0.5910	0.5819	1.1315	0.6153	0.5437	1.2156	0.7840	0.6449
W_x	0.8914	0.4667	0.5235	0.9195	0.4762	0.5178	0.9866	0.5048	0.5116
W_y	0.9174	0.4530	0.4937	0.9763	0.4696	0.4809	0.9845	0.5187	0.5268

Tabela 7 – Médias, desvios padrão e coeficiente de variação da consistência para as características

ao eixo x e y. Esses mesmos gráficos são mostrados na Figura 9 referendo-se as características da informação de Fisher e estatística de tendência de Wallis e Moore.

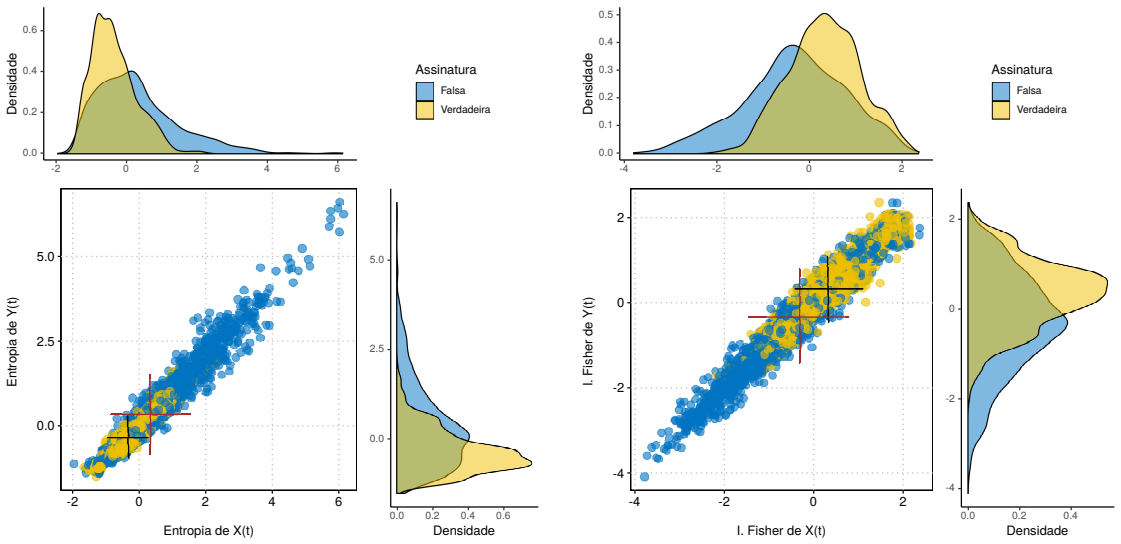
A Figura 10 mostra a correlação entre as características calculadas, ademais, é possível observar a existência de 3 grupos nesse gráfico. O primeiro grupo apresenta a correlação entre as características de entropia e complexidade, sendo assim, pode-se notar uma correlação positiva forte, já que as elipses da figura estão mais achatadas. Por outro lado, o segundo grupo composto pelas características de entropia e complexidade relacionadas com a informação de Fisher e estatística de Wallis e Moore apresenta uma correlação negativa, sendo essa forte na maior parte das formas das características. Por fim, o último grupo que é composto pelas características de informação de Fisher e estatística de Wallis e Moore, nesse grupo, as correlação é positiva, porém é mais fraca ao comparar com o primeiro grupo, visto que as elipses estão mais dispersas.

4.2 SELEÇÃO DAS CARACTERÍSTICAS

Nesta seção são retratados os resultados da seleção das características, foram utilizados quatro métodos para essa seleção: ganho de informação, Anova 1%, Anova 5% e fator de inflação da variância, a partir desses métodos, obteve-se quatro modelos, ou seja, cada modelo é composto pela variável resposta y, classe das assinaturas, e as características x que foram selecionadas a partir desses critérios de seleção mencionados.

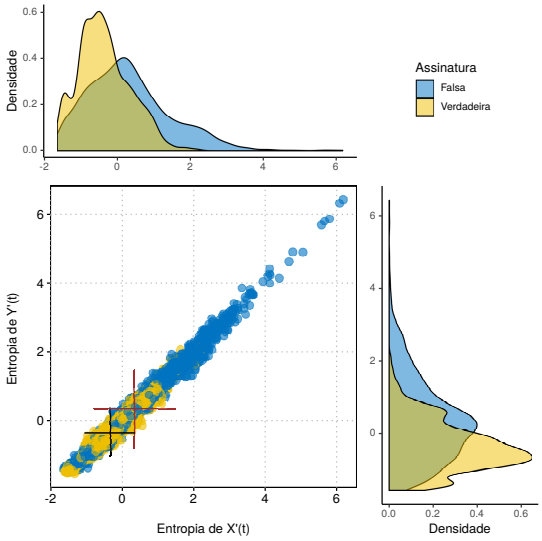
Para o ganho de informação, foi calculado a razão do ganho de informação para todas as características obtidas através das séries temporais, posteriormente, calculamos a média da razão do ganho de informação e, assim, as características selecionadas para compor o conjunto de quantificadores não paramétricos segundo esse critério foram as características cujo a razão do ganho fossem acima da média.

No que se refere ao critério de seleção segundo ANOVA, utilizando o modelo de

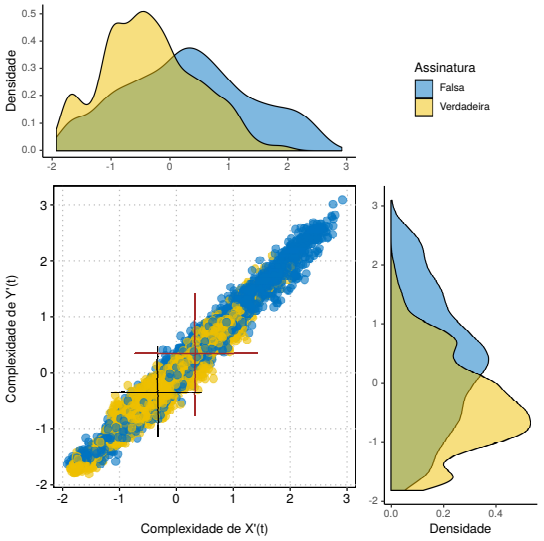


(a) Entropia

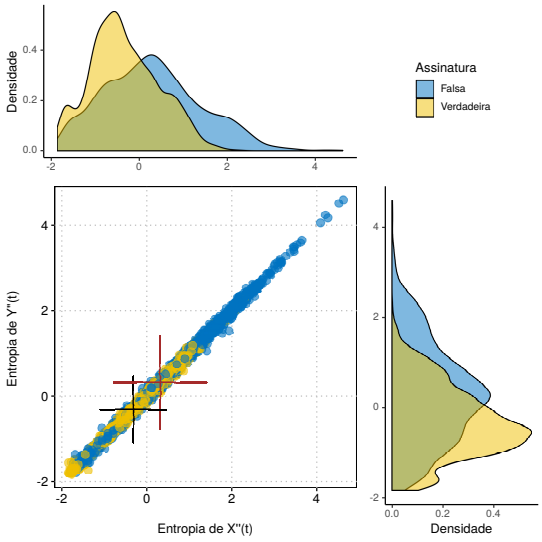
(b) Complexidade



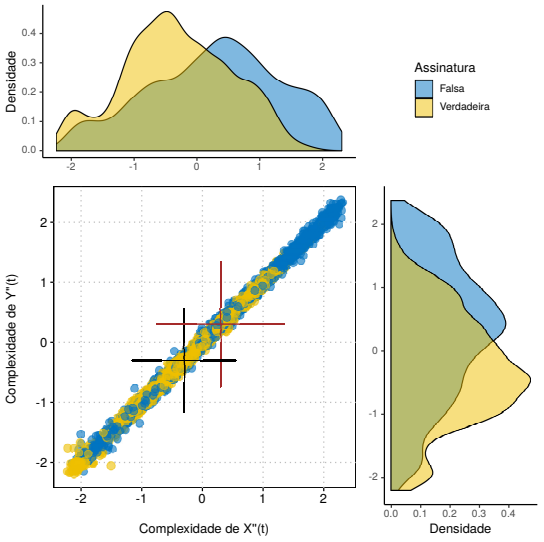
(c) Entropia



(d) Complexidade

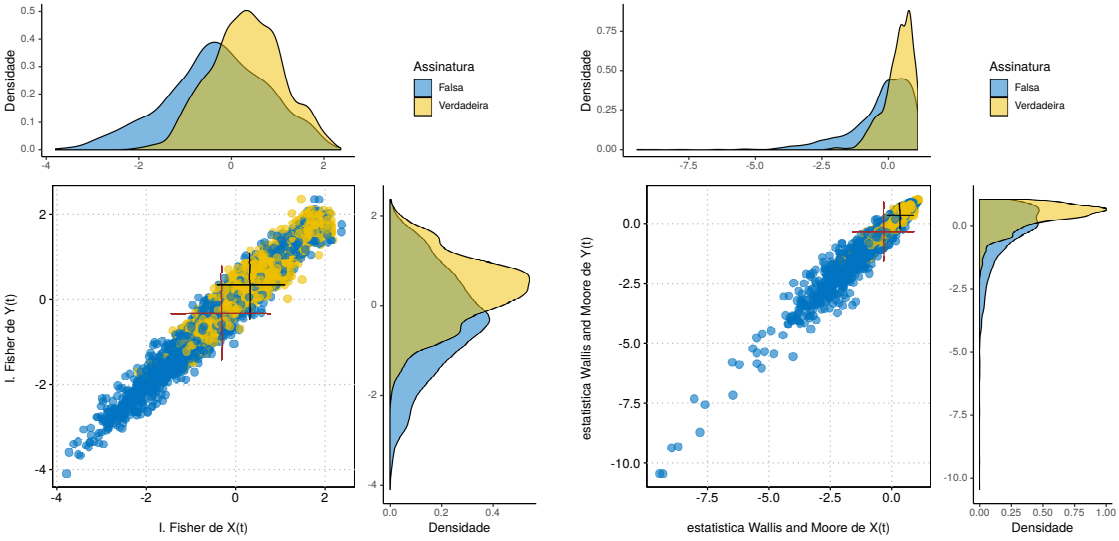


(e) Entropia

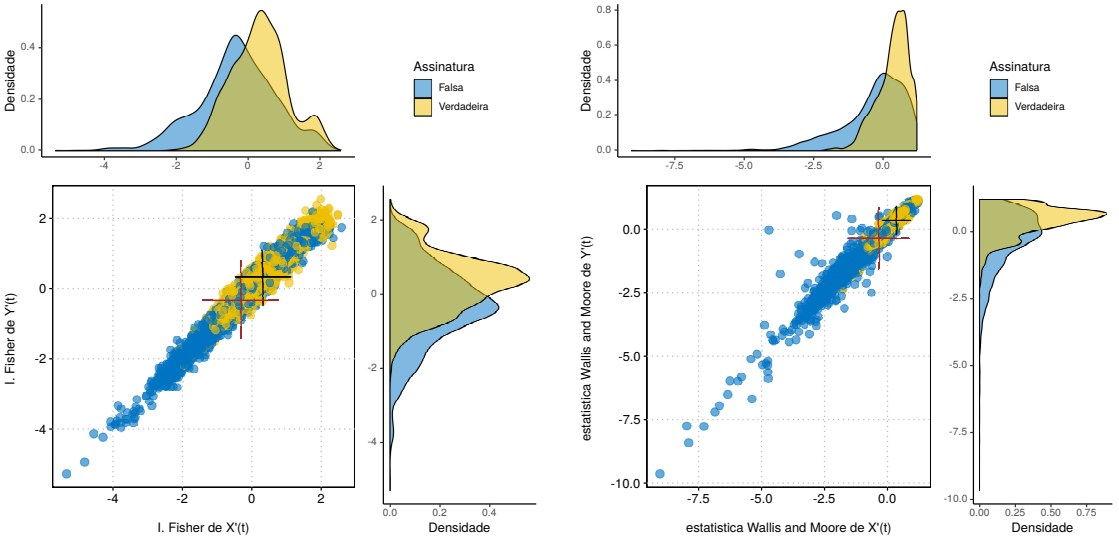


(f) Complexidade

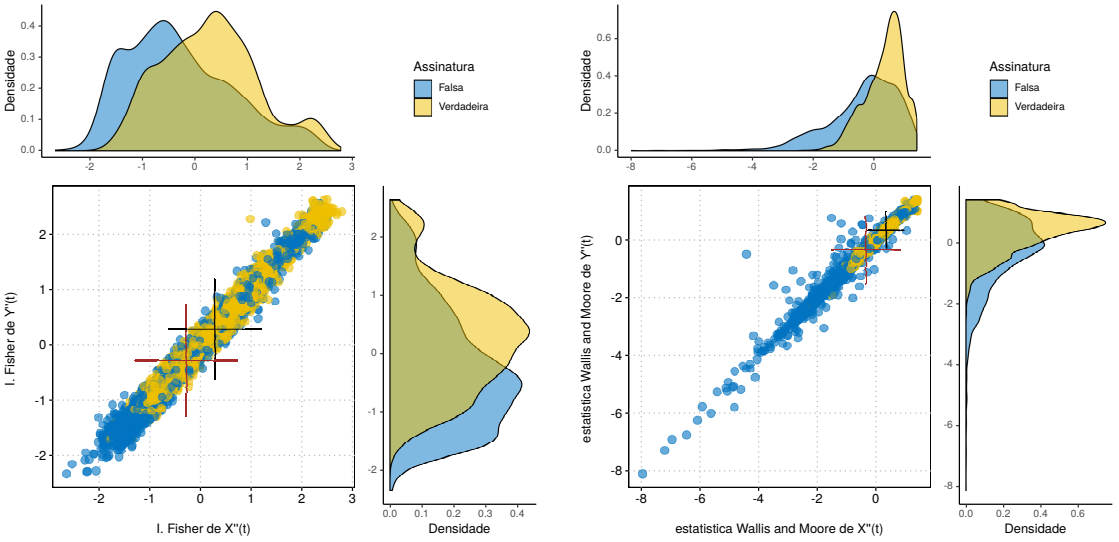
Figura 8 – Diagrama de dispersão com densidades marginais da entropia e complexidade



(a) I. Fisher **(b)** Wallis e Moore



(c) I. Fisher **(d)** Wallis e Moore



(e) I. Fisher **(f)** Wallis e Moore

Figura 9 – Diagrama de dispersão com densidades marginais da informação de Fisher e estatística de tendência de Wallis e Moore

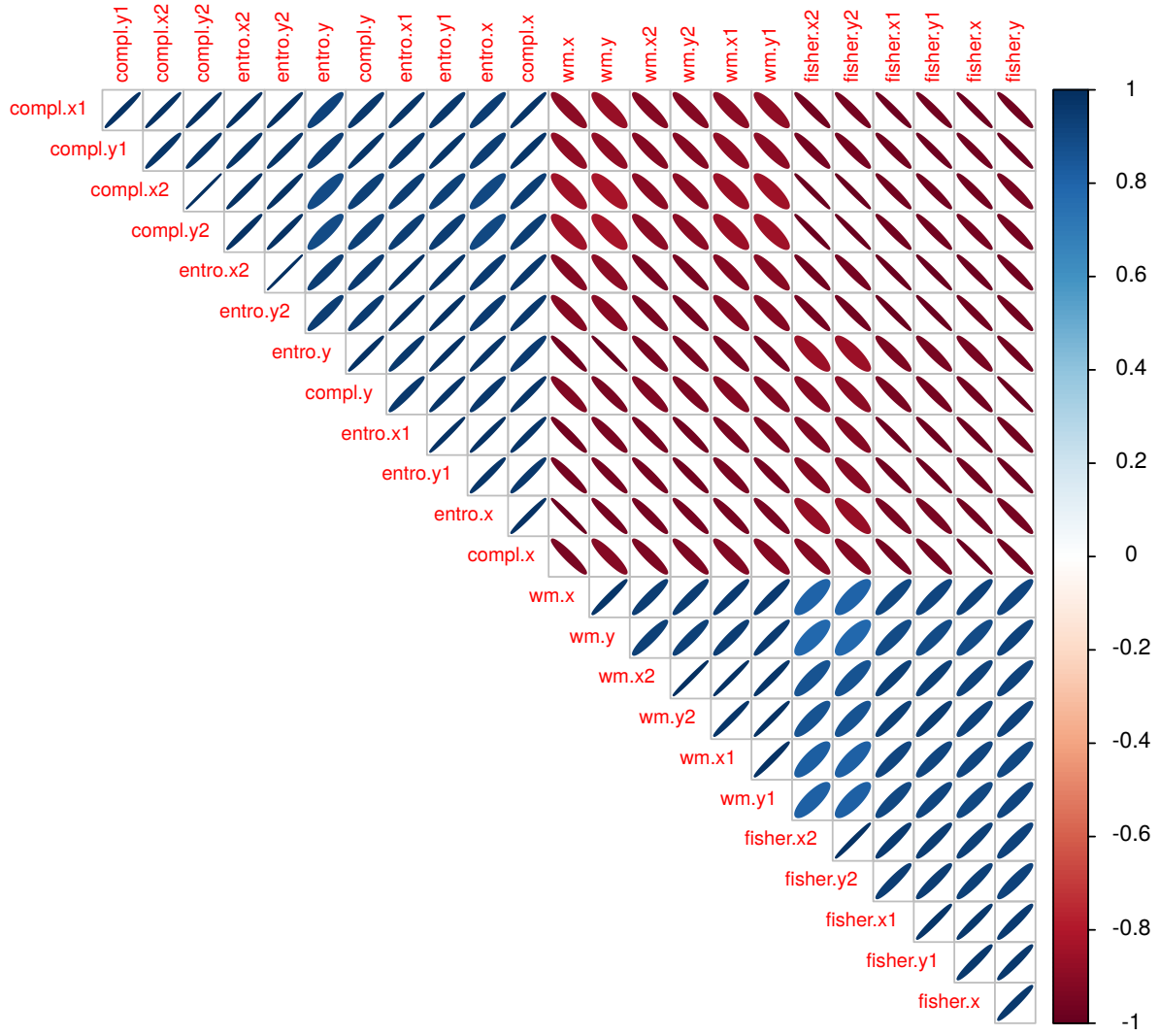


Figura 10 – Correlação das dos quantificadores não paramétricos

regressão logística com função de ligação *logit*, foi estimado os coeficientes desse modelo composto por todos os quantificadores não-paramétricos, então as características que compõe o conjunto segundo esse critério foram as que apresentaram significância ao nível de significância de 1% e 5%.

Para o cálculo do VIF, foi utilizado o modelo de regressão logística com função de ligação *logit* e calculado o VIF para as covariáveis desse modelo, dessa maneira, a característica que apresentou o maior valor do VIF foi retirada da análise e, assim, o modelo logístico foi estimado novamente e recalculado o VIF até que, por fim, as caraterísticas tiveram o VIF menor que 10 foram selecionadas para o conjunto de quantificadores segundo o VIF.

Na Tabela 8, os símbolos x, x', x'', y, y' e y'' se referem, respectivamente, ao valor

Critério de seleção	Características	Nº Características
Ganho de Informação	$H_y, C_y, F_y, W_y, W_{x'}, H_{y'}, C_{y'}, F_{y'}, W_{y'}, W_{x''}, W_{y''}$	11
Anova 5%	$H_x, W_x, C_y, F_y, C_{x'}, C_{y'}, F_{y'}, W_{y'}, F_{x''}, F_{y''}$	10
Anova 1%	H_x,W_x, C_y, F_y,$F_{y'}$, $W_{y'}$,$F_{x''}$	7
VIF	$W_y, F_{y''}$	2

Tabela 8 – Características selecionadas segundo os critérios de seleção.

original, primeira derivada e segunda derivada das características dos eixos x e y. Nessa tabela pode-se notar que as características $C_y, F_y, F_{y'}$ e $W_{y'}$ tiveram mais relevância e estão contidas em três dos critérios de seleção, não sendo relevante somente segundo o VIF, sendo esse, o critério mais conservador, apresentando apenas duas características em sua seleção.

4.3 ETAPA DE CLASSIFICAÇÃO

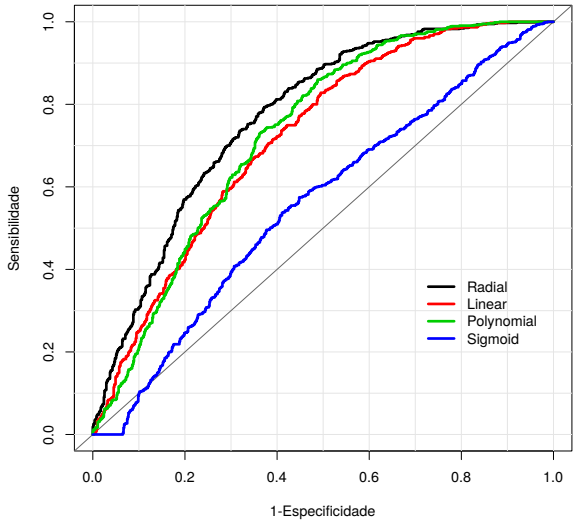
Para a etapa de classificação, utilizando o método de *holdout* descrito na seção 2, o conjunto de caraterísticas selecionadas segundo os critérios foram aplicados nos classificadores. Porém, para o método SVM, é necessário encontrar a função de *Kernel* que melhor se adeque aos dados da amostra. Desse modo, em cada conjunto de característica foram ajustados o método SVM com diferentes funções *Kernel*, sendo assim, como parâmetros dessa função foram utilizados o *default* da pacote *e1071* no *R*. Sendo assim, para as bases de treinamento dos quatro conjuntos de características selecionadas, foram feitas as classificações a fim de selecionar a melhor função e, posteriormente, poder usa-la na iteração do *holdout múltiplo*.

Kernel	Anova 1%	Anova 5%	G. I	VIF
Radial	**0.771**	**0.781**	**0.764**	**0.692**
Linear	0.718	0.725	0.735	0.683
Polynomial	0.728	0.747	0.746	0.688
Sigmoid	0.558	0.550	0.563	0.549

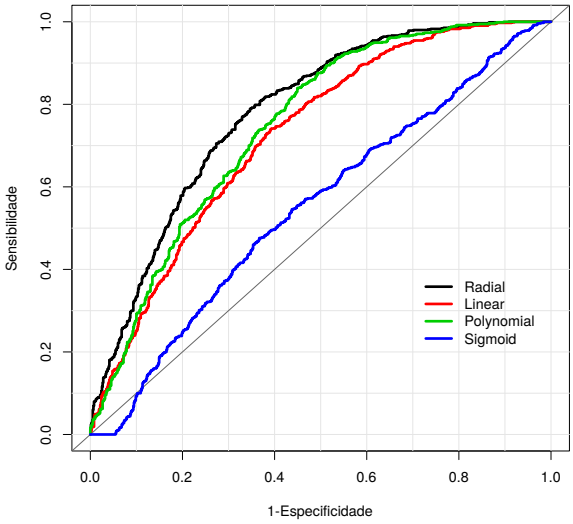
Tabela 9 – AUC para os SVM com diferentes funções kernel

Como é visto na Tabela 9, para todos os conjunto de dados, o classificado SVM que apresentou o maior AUC é a que utiliza a função *Kernel* Radial, na Figura 11 pode-se observar a curva ROC de cada conjunto de característica utilizando diferentes funções Kernel. A Tabela apresenta 9 os valores do AUC para o método SVM. O classificador SVM utilizado na etapa do *holdout múltiplo* é o com a função *Kernel* Radial.

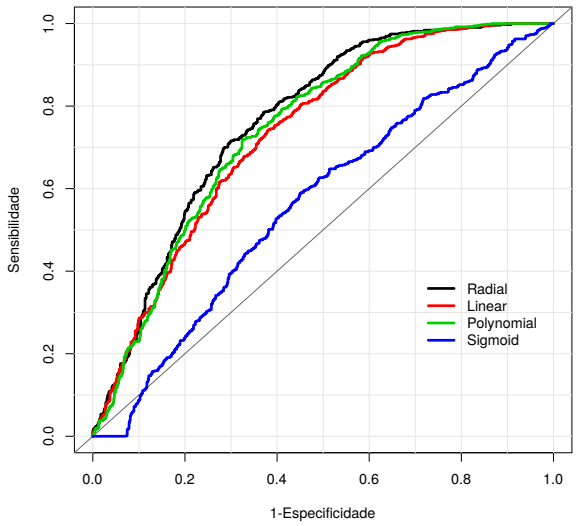
A Figura 12 mostra as médias e desvios padrão das acurácias, tanto no contexto de treinamento (Figura 12a) como da etapa de teste (Figura 12b) para os diferentes métodos



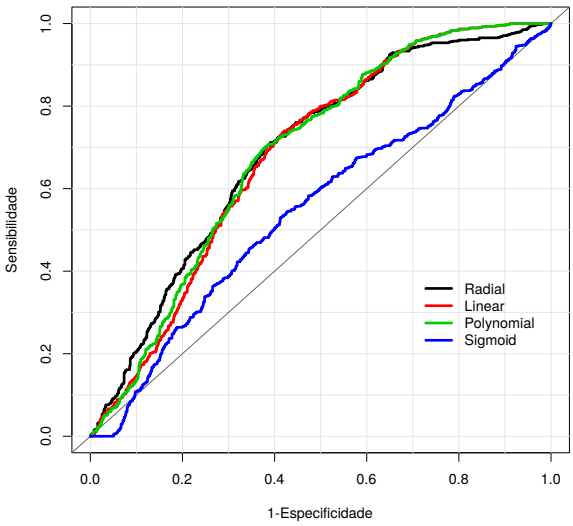
(a) Anova 1%



(b) Anova 5%



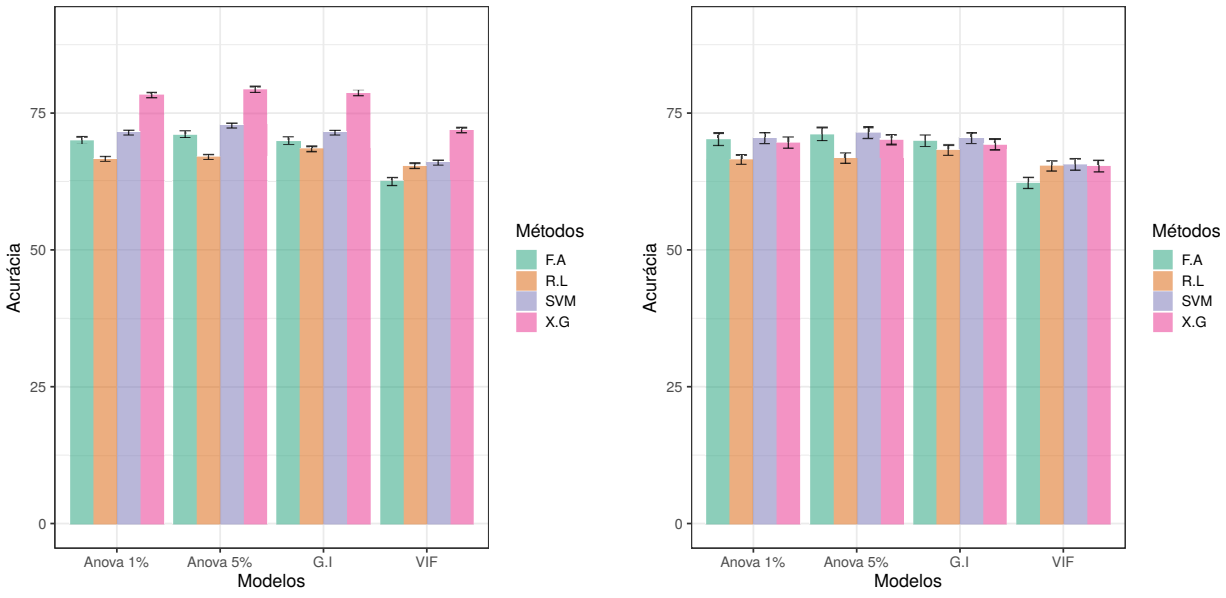
(c) Ganho de Informação



(d) VIF

Figura 11 – Curva ROC da classificação utilizando o SVM com diferentes funções *Kernel* aplicado nos conjuntos de dados segundo o critério de seleção de características

de classificação ajustados para cada conjunto de dados segundo os critérios de seleção. Dessa maneira, observa-se que, na etapa de treinamento, o classificador XGBoost foi o que obteve um melhor ajuste, apresentando para o conjunto de dados Anova 1%, acurácia média de 78.24% de desvio padrão $\pm$0.515%, o conjunto de dados Anova 5% obteve acurácia média de 79.3% e desvio padrão de $\pm$0.541%, o conjunto de características segundo o Ganho de Informação, acurácia média de 78.6% e desvio padrão de $\pm$0.522% e, por fim, o conjunto de características segundo o VIF com acurácia de 71.86% e desvio padrão de $\pm$0.502%. No entanto, no que se refere ao grupo de teste, o classificador XGBoost não se destacou tanto quanto no grupo anterior, de forma geral, não houve grande diferença quanto as acurácias nesse grupo, porém pode-se destacar que os métodos de classificação de florestas aleatória e SVM foram os que apresentaram melhor acurácia e sendo esses valores próximos entre si para os critérios de características Anova 1%, Anova 5% e Ganho de Informação. Somente para o modelo segundo o VIF, o classificador florestas aleatória foi o que apresentou menor acurácia, entretanto, a regressão logística, SVM e XGBoost apresentaram valores de acurácia próximos. A Tabela 10 e 11 mostra a acurácia média para o grupo de treinamento e teste, respectivamente.



(a) Treinamento **(b)** Teste

Figura 12 – Acurácia para cada método de classificação segundo os critérios de seleção de caracteristicas

As matrizes relativas de confusão apresentadas nas Tabelas 12 e 13 mostram os percentuais médios de classificação para as 100 iterações nos grupos de treinamento e teste, respectivamente. Nesse contexto, espera-se que a porcentagem presente na diagonal principal de

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	66.57	66.95	68.44	65.32
F.A	70.02	71.04	69.88	62.42
SVM	71.40	72.72	71.39	65.91
XG	78.25	**79.30**	78.61	71.86

Tabela 10 – Acurácia média (%) para cada método de classificação segundo o critério de seleção das características para o grupo de treinamento

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	66.51	66.75	68.19	65.34
F.A	70.15	**71.12**	69.94	62.23
SVM	70.37	**71.33**	70.36	65.63
XG	69.57	70.13	69.19	65.32

Tabela 11 – Acurácia média (%) para cada o método de classificação segundo o critério de seleção das características para o grupo de teste

cada matriz seja o maior possível, pois, assim, tem-se a porcentagem de indivíduos classificados corretamente. Corroborando com a acurácia, o classificador XGBoost apresentou maior porcentagem de classificação correta na fase de treinamento, já na fase de teste, os classificadores Florestas aleatórias e SVM foram o que apresentaram melhor desempenho no que tange a matriz de confusão.

	Classe	F.A		R.L		SVM		XG	
Critérios	Y \ $\hat{Y}$	0	1	0	1	0	1	0	1
Anova 1%	0	33	17	26	10	31	10	36	14
	1	13	37	24	40	19	40	8	42
Anova 5%	0	33	17	28	11	33	10	37	13
	1	12	38	22	39	17	40	8	42
G. I	0	32	18	28	10	31	10	36	14
	1	12	38	22	40	19	40	8	42
VIF	0	30	20	26	10	29	13	32	18
	1	17	33	24	40	21	37	10	40

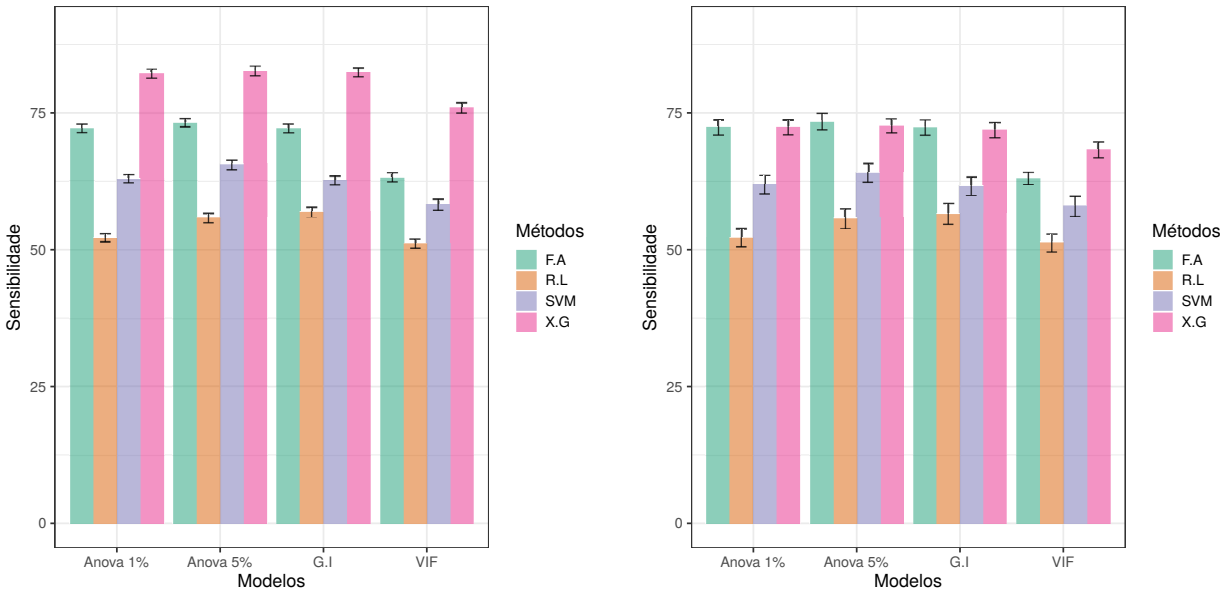
Tabela 12 – Matrizes relativas de confusão (%) para os métodos de classificação (coluna) segundo os critérios de seleção de características (linha) para os dados de treinamento

Como forma de observar o desempenho dos classificadores em relação a cada classe de assinaturas, pode-se observar na Figura 13 os desempenho dos classificadores segundo a sensibilidade, que no contexto dessa dissertação, é a capacidade do método classificar corretamente assinaturas falsas. Logo, a Figura 13a mostra o desempenho do classificador na fase de treinamento e, nesse caso, o método XGBoost se destaca para os todos os critérios de seleção

	Classe	F.A		R.L		SVM		XG	
Critérios	Y \ $\hat{Y}$	0	1	0	1	0	1	0	1
Anova 1%	0	33	17	26	10	31	11	32	18
	1	12	38	24	40	19	39	12	38
Anova 5%	0	33	17	28	11	32	11	32	18
	1	12	38	22	39	18	39	12	38
G. I	0	32	18	28	10	31	10	32	18
	1	12	28	22	40	19	40	12	38
VIF	0	30	20	26	10	29	13	29	21
	1	17	33	24	40	21	37	13	37

Tabela 13 – Matrizes relativas de confusão (%) para os métodos de classificação (coluna) segundo os critérios de seleção de características (linha) para os dados de teste

de características, seguido pelas Florestas Aleatórias, por outro lado, o método de regressão logística teve o pior desempenho nesse grupo. Por conseguinte, na fase de teste, para os critérios Anova 1%, Anova 5% e Ganho de Informação, os classificadores XGBoost e Florestas Aleatórias tiveram desempenho similar, já segundo o VIF, esses dois métodos também foram os com melhor desempenho, porém o XGBoost apresentou maior sensibilidade. A Tabela 14 e 15, mostra os da sensibilidade média segundo os métodos de classificação para os quatro critérios de seleção de características para o grupo de treinamento e teste, nessa ordem.



(a) Treinamento **(b)** Teste

Figura 13 – Sensibilidade para cada método de classificação segundo os critérios de seleção de caracteristicas

De forma análoga, a especificidade é a capacidade de classificar corretamente as

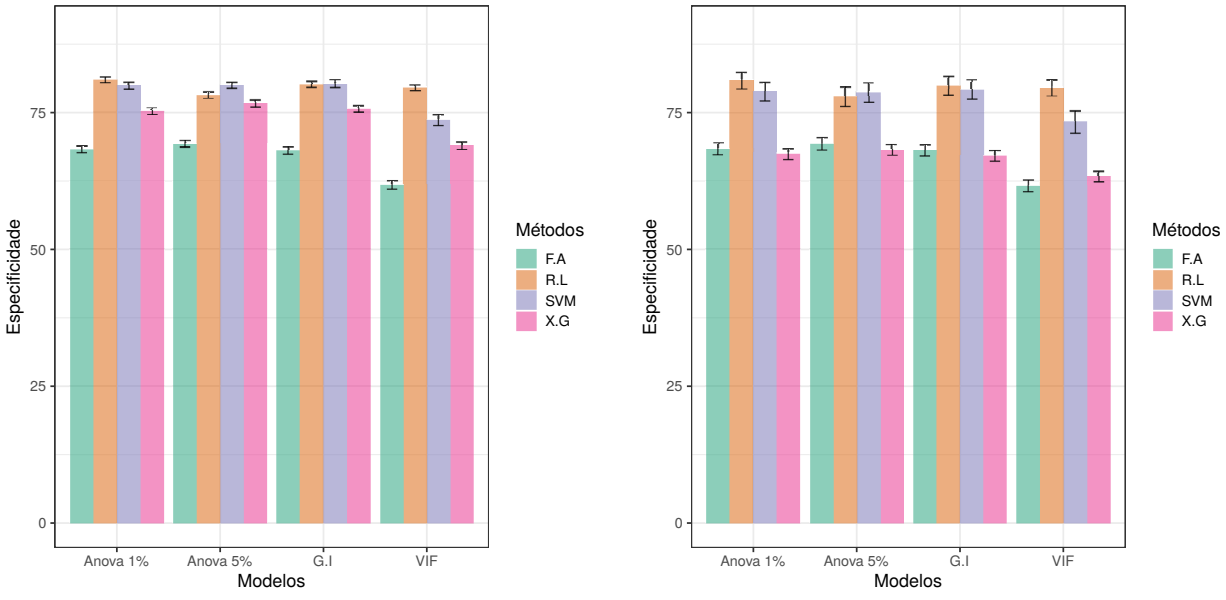
Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	52.17	55.74	56.81	51.15
F.A	72.18	73.21	72.17	63.14
SVM	62.91	65.48	62.59	58.20
XG	82.12	**82.57**	82.34	75.92

Tabela 14 – Sensibilidade média (%) para cada método de classificação segundo o critério de seleção das características para o grupo de treinamento

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	52.21	55.61	56.52	51.21
F.A	72.36	**73.38**	72.30	62.99
SVM	61.93	64.04	61.58	57.94
XG	72.37	72.61	71.91	68.23

Tabela 15 – Sensibilidade média (%) para cada o método de classificação segundo o critério de seleção das características para o grupo de teste

assinaturas verídicas, dessa forma, como mostra na Figura 14, para o grupo de treinamento, nos critérios Anova 1%, Anova 5% e Ganho de Informação, os métodos de regressão logística e SVM apresentaram melhor desempenho. Para o critério segundo o VIF, ainda assim, os classificadores de regressão logística e SVM se destacaram, porém, para esse conjunto de características, a regressão logística apresentou especificidade de 79.47%, enquanto o SVM obteve 73.31%. Na fase de teste, os classificadores tiveram desempenho similar à fase anterior. Na Tabela 16 e 17 é possível ver os valores da especificidade média para cada critério de seleção de características segundo os classificadores.



(a) Treinamento **(b)** Teste

Figura 14 – Especificidade para cada método de classificação segundo os critérios de seleção de caracteristicas

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	**80.98**	78.17	80.08	79.49
F.A	68.24	69.24	68.02	61.78
SVM	79.89	79.95	80.19	73.61
XG	75.23	76.64	75.67	68.93

Tabela 16 – Especificidade média (%) para cada método de classificação segundo o critério de seleção das características para o grupo de treinamento

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	**80.80**	77.89	79.86	79.47
F.A	68.37	69.28	68.06	61.58
SVM	78.80	78.63	79.14	73.31
XG	67.42	68.16	67.09	63.24

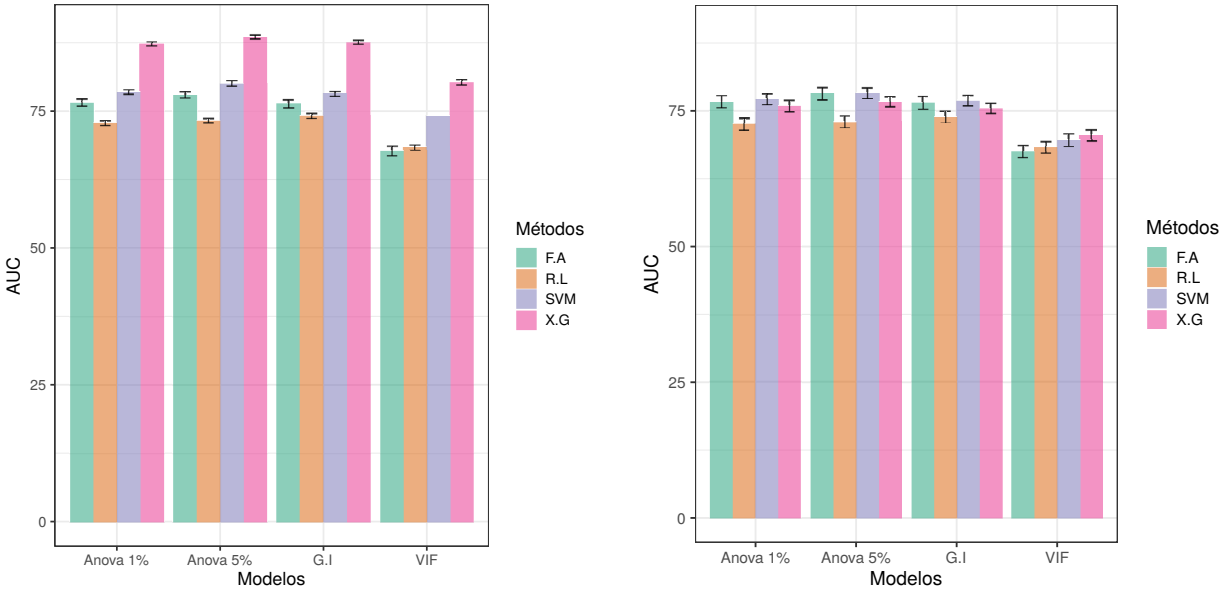
Tabela 17 – Especificidade média (%) para cada o método de classificação segundo o critério de seleção das características para o grupo de teste

No que tange a AUC, como mostra na Figura 15, os métodos de regressão logística e SVM tiveram melhores desempenho segundo essa métrica tanto na fase de treinamento, como também na fase de teste. As Tabelas 18 e 19 apresentam a AUC média para o grupo de treinemento e teste, respectivamente . Como forma de visualizar o comportamento da curva ROC, a Figura 16 é uma exemplificação da curva ROC para os classificadores ajustados para cada conjunto de dados segundo os critérios de seleção. Nesses aspectos, o critério Anova 5% apresentou melhor desempenho para os métodos Florestas aleatórias, SVM e XGBoost, em contrapartida, para o classificador de regressão logística, oconjunto de características segundo o ganho de informação foi o que apresentou maior AUC.

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	72.76	73.22	74.12	68.30
F.A	76.54	77.93	76.34	67.68
SVM	78.45	80.04	78.14	74.08
XG	87.25	**88.50**	87.56	80.27

Tabela 18 – AUC média (%) para cada o método de classificação segundo o critério de seleção das características para o grupo de treinamento

A Figura 17 traz a média da taxa de erro e seus respectivos desvios padrões, paraa essa quantidade, como se trata de a taxa de erros na classificação, busca-se os classificadores que apresentam menor valor segundo essa métrica. Na fase de treinamento, o classificador XGBoost se destacou para todos os conjuntos de dados segundo os critérios de seleção, apresentando menor ER, a média do ER pode ser vista na Tabela 20. Por outro lado, na fase de teste, os



(a) Treinamento **(b)** Teste

Figura 15 – AUC dos critérios de seleção para os diferentes métodos de classificação

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	72.71	73.37	73.67	68.15
F.A	75.90	76.70	75.26	70.63
SVM	77.11	**78.10**	76.40	69.20
XG	75.90	76.70	75.26	70.63

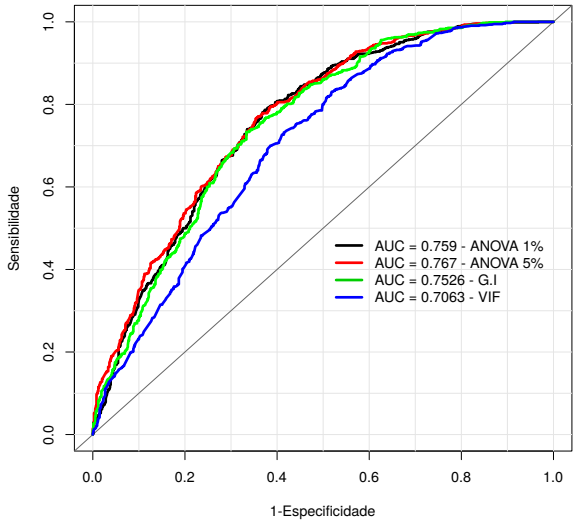
Tabela 19 – AUC média (%) para cada o método de classificação segundo o critério de seleção das características para o grupo de teste

classificadores de Florestas aleatórias, SVM e XGBoost apresentaram médias próximas em relação os diferentes critérios de seleção de características, porém na Tabela 21 nota-se que, entre esses três classificadores, de fato, o SVM apresentou menor taxa de erro nos quatro modelos.

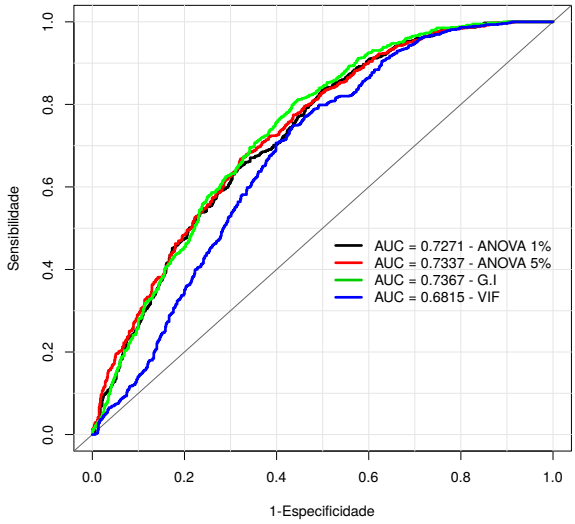
Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	33.42	33.04	31.55	34.67
F.A	29.97	28.95	30.11	37.57
SVM	28.59	27.27	28.60	34.08
XG	21.74	**20.70**	21.38	28.13

Tabela 20 – ER média (%) para cada método de classificação segundo o critério de seleção das características para o grupo de treinamento

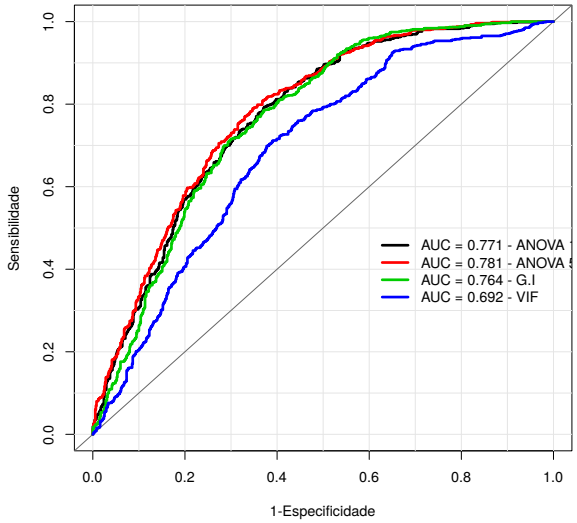
A Tabela 22 traz o tempo médio, em segundos, para a estimação da classificação do conjunto de dados segundo os critérios de seleção de características. Os métodos de florestas aleatórias seguido por SVM são os métodos que mais demandaram tempo para a estimação, sendo o primeiro em torno de 20 segundos e, o segundo, levando 6 segundos. Por outro lado, a regressão logística e o XGBoost foram os métodos mais rápidos com menos de 1 segundo



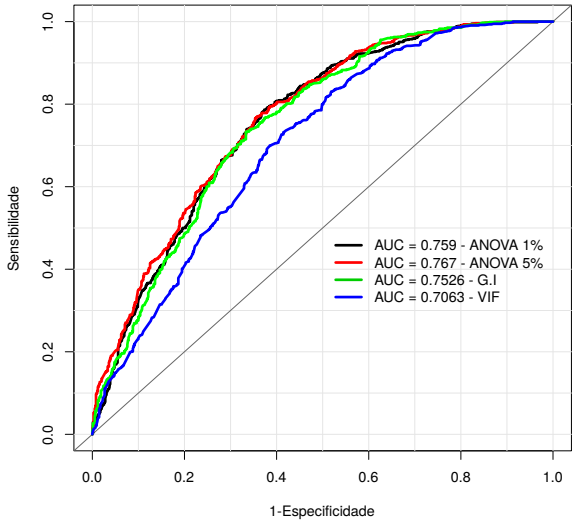
(a) Florestas Aleatórias



(b) Regressão Logstica



(c) SVM

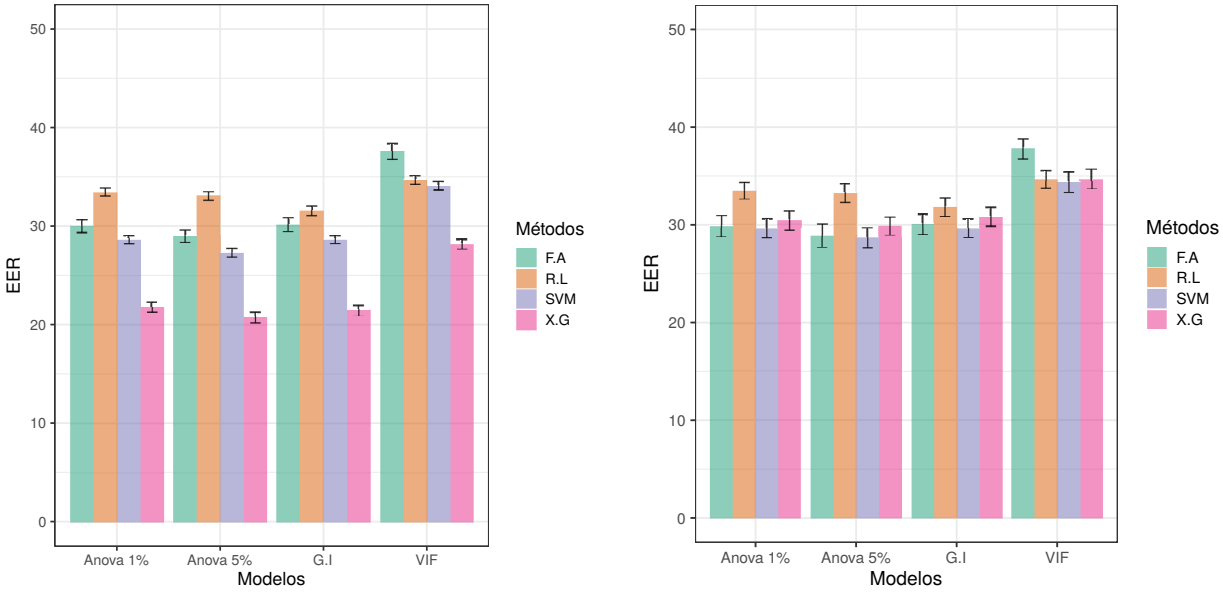


(d) XGBoost

Figura 16 – Curva ROC da classificação utilizando cada método aplicado nos conjuntos de dados segundo o critério de seleção de características no grupo de teste

Métodos	Anova 1%	Anova 5%	G. I	VIF
R.L	33.48	33.24	31.80	34.65
F.A	29.84	**28.87**	30.05	37.76
SVM	29.62	**28.66**	29.63	34.36
XG	30.42	29.86	30.80	34.67

Tabela 21 – ER média (%) para cada o método de classificação segundo o critério de seleção das características para o grupo de teste



(a) Treinamento **(b)** Teste

Figura 17 – ER para cada método de classificação segundo os critérios de seleção de caracteristicas

para obter as estimativas, vale destacar que, a regressão logística leva menos de 1 centésimo de segundo para obter a estimativa da classificação.

Métodos	Anova 1%	Anova 5%	G. I	VIF
F.A	19.6365	21.0733	21.5232	17.1525
R.L	**0.0230**	**0.0261**	0**0.0353**	**0.0159**
SVM	6.0577	6.3253	6.5732	7.1647
XG	**0.2870**	**0.3896**	**0.3710**	**0.2459**

Tabela 22 – Tempo médio de processamento da classificação em segundos

4.4 CLASSIFICAÇÃO COM REGRESSÃO LOGÍSTICA REGULARIZADA

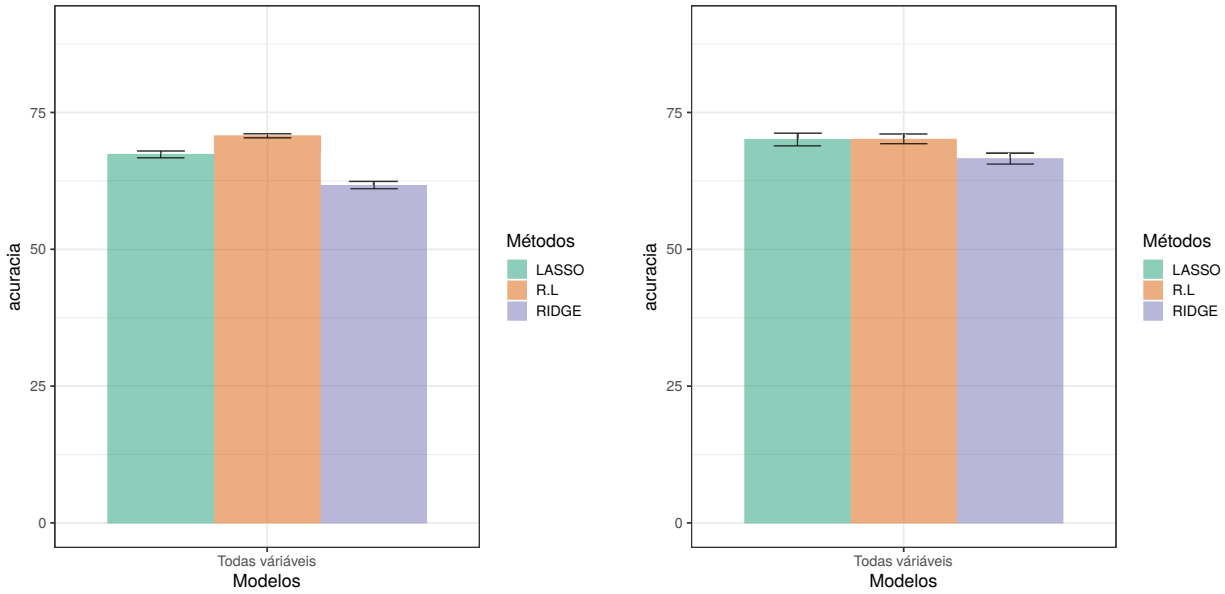
Nessa seção apresenta-se os resultados das classificações utilizando a regressão logística regularizada tipo LASSO e Ridge, bem como a regressão logística sem a penalização. A Figura 10 mostra que existe uma correlação entre as características, podendo ocasionar os problemas discutidos na seção 2.2 quando se utiliza características explicativas desnecessárias, portanto, é interessante observar o desempenho dessas técnicas de classificação utilizando um banco de dados com características que apresentam fortes correlações.

Como mencionado na seção 2.2, a regressão regularizada do tipo LASSO é um processo de seleção automática para as características. Sendo assim, utilizando as 24 características extraídas das séries temporais foi estimado o modelo de regressão logística penalizada

do tipo LASSO. No que refere a esse modelo, nenhuma característica foi penalizada, ou seja, o modelo considerou que todas as características eram importantes para a análise. Dessa maneira, os resultados aqui apresentados foram obtidos utilizando o banco de dados com as 24 características.

No que tange a classificação, após a validação cruzada para encontrar o melhor λ, foi feito o ajuste do modelo e tando para a penalização tipo LASSO, quanto a do tipo Ridge, todas as características do banco de dados foram consideradas importantes para a classificação. A classificação nessa etapa, utilizou o método *holdout* com 100 iterações para calibração.

NA Figura 18 pode-se observar o desempenho dos métodos de classificação segundo a acurácia média, sendo assim, tanto para a fase de treinamento, quanto para a fase de teste, a regressão logística apresentou melhor desempenho, sendo esse de 70,67% (treinamento) e 70.15%. A Tabela 18 apresenta os valores da acurácia média para os modelos na fase de treinamento e teste, respectivamente.



(a) Treinamento **(b)** Teste

Figura 18 – Acurácia para os diferentes métodos de classificação

Métodos	Treinamento	Teste
LASSO	67.26	70.04
R.L	**70.67**	**70.15**
Ridge	61.71	66.54

Tabela 23 – Acurácia média (%) para cada método de classificação utilizando regressão regularizada e regressão logística

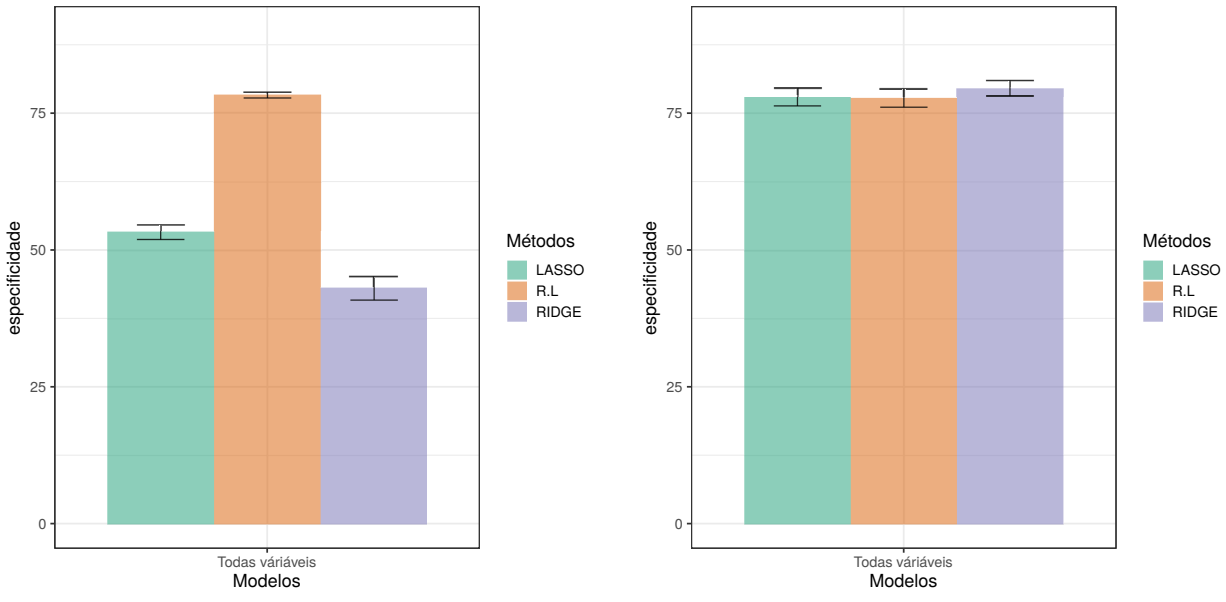
As matrizes relativas de confusão para a etapa de treinamento e teste, respectivamente, apresentam-se nas Tabelas 24 e 25. No treinamento, os 3 classificadores apresentaram erros maiores ao predizer que a assinatura era verdadeira quando essas eram falsificações, porém esse cenário muda na fase de teste, quando o erro foi maior ao classificar a assinatura como falsa quando, na verdade, era verdadeira.

Classe	LASSO		R.L		Ridge	
Y \ $\hat{Y}$	0	1	0	1	0	1
0	41	23	40	29	40	29
1	9	27	10	21	10	21

Tabela 24 – Matrizes relativas de confusão (%) para os métodos de classificação (coluna) segundo os critérios de seleção de características (linha) para os dados de treinamento

Classe	LASSO		R.L		Ridge	
Y \ $\hat{Y}$	0	1	0	1	0	1
0	31	11	28	10	27	10
1	19	39	22	40	23	40

Tabela 25 – Matrizes relativas de confusão (%) para os métodos de classificação (coluna) segundo os critérios de seleção de características (linha) para os dados de teste



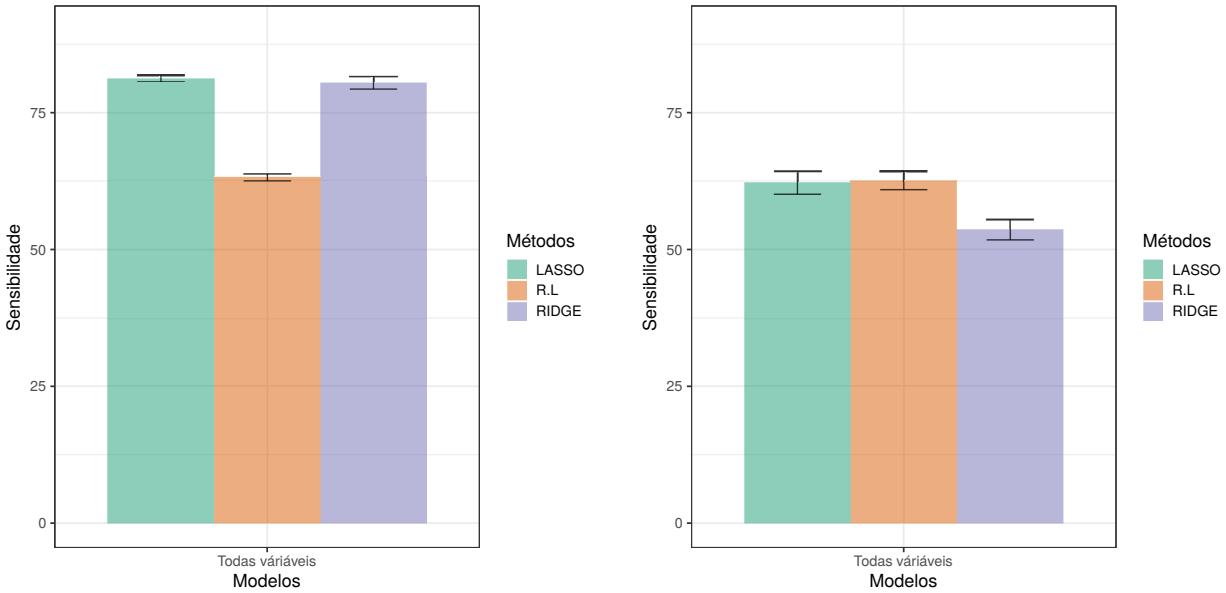
(a) Treinamento **(b)** Teste

Figura 19 – Especificidade para os diferentes métodos de classificação

Métodos	Treinamento	Teste
LASSO	53.27	77.95
R.L	**78.26**	77.77
Ridge	42.98	**79.54**

Tabela 26 – Especificidade média (%) para cada o método de classificação utilizando regressão regularizada e regressão logística

No que se refere a especificidade, na Figura 19 observava-se que a especificidade apresentou valores diferentes para os 3 métodos no treinamento, porém, na etapa de teste, os valores da especificidade foram similares com relação aos 3 métodos. Vale destacar que, a regressão logística não regularizada foi o único método que obteve valores similares tanto no grupo de treinamento, quanto no teste. A Tabela 26, apresenta a especificidade média nos grupos de treinamento e teste. Para a sensibilidade, apresentada na Figura 20, pode-se observar um comportamento similar ao da especificidade, apresentando valores divergentes entre os métodos no grupo de treinamento e teste. Cabe destacar que, novamente, a regressão logística não regularizada foi o método que apresentou comportamento semelhante nas duas etapas. A Tabela 27 apresenta a sensibilidade média nos grupos de teste e treinamento.



(a) Treinamento **(b)** Teste

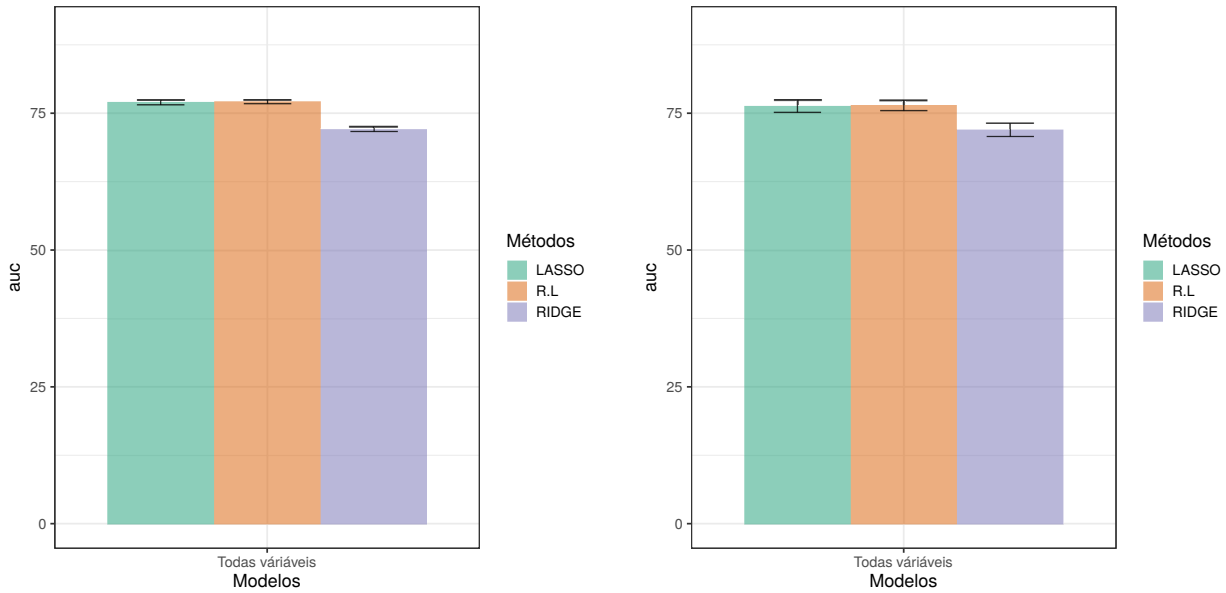
Figura 20 – Sensibilidade para os diferentes métodos de classificação

A Figura 21, apresenta o desempenhos dos classificadores em relação a AUC, dessa maneira, os métodos apresentaram desempenhos semelhantes no que se refere as etapas de teste e treinamento, em destaque, o LASSO e a regressão logística não regularizada, que obtiveram AUC próximo. A Tabela 28 apresenta a AUC média. A Figura 22 apresenta as curva ROC junto

Métodos	Treinamento	Teste
LASSO	81.24	62.12
R.L	**63.08**	**62.54**
Ridge	80.45	53.54

Tabela 27 – Sensibilidade média (%) para cada o método de classificação utilizando regressão regularizada e regressão logística

com o valor do AUC para os 3 métodos de classificação, dessa forma, pode-se observar que o LASSO e e regressão logística não regularizada apresentam curvas semelhantes.



(a) Treinamento **(b)** Teste

Figura 21 – AUC para os diferentes métodos de classificação

Métodos	Treinamento	teste
LASSO	76.90	77.06
R.L	**77.08**	**76.39**
Ridge	72.04	72.66

Tabela 28 – AUC média (%) para cada o método de classificação utilizando regressão regularizada e regressão logística

A Figura 23 apresenta a taxa de erro média da classificação junto com os desvios padrões, vale ressaltar que, busca-se o modelo com menor taxa de erro, dessa forma, a regressão logística não regularizada foi a que apresentou menor taxa nas duas etapas da classificação. Na Tabela 29 apresenta-se a taxa de erro média da classificação para os 3 métodos de classificação.

A Tabela 30 apresenta o tempo médio, em segundo, para estimação da classificação, pode-se notar que a regressão logística não regularizada é a mais rápida. Os métodos do LASSO

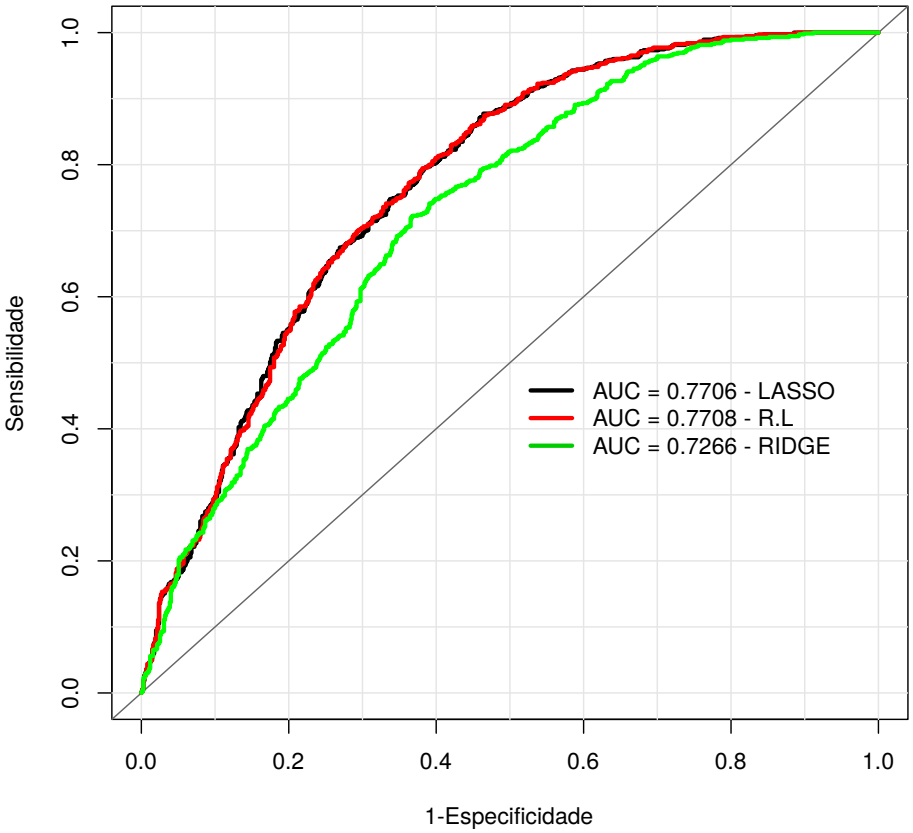


Figura 22 – Curva ROC da classificação utilizando os métodos de classificação aplicado em todo conjuntos de dados

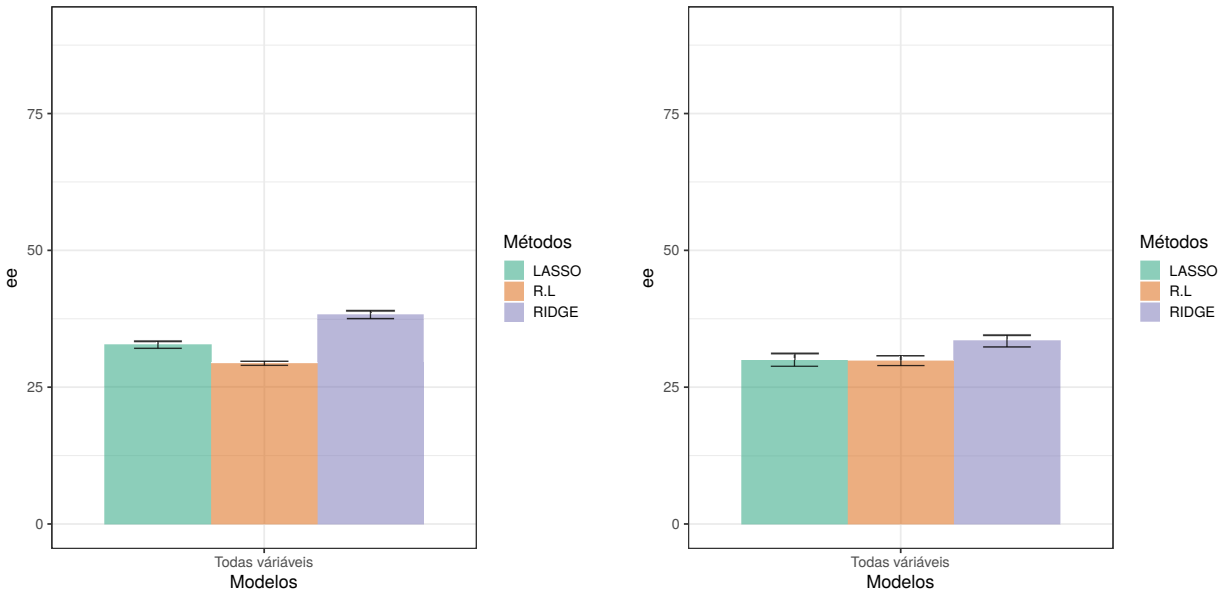
e o Ridge antes do processo de estimação precisam, por meio da validação cruzada, buscam o melhor valor para o parâmetro de penalização λ. O método LASSO foi o que demandou mais tempo, em média 62.26 segundos em cada iteração.

Métodos	Treinamento	Teste
LASSO	32.73	29.96
R.L	**29.32**	**29.84**
Ridge	38.28	33.45

Tabela 29 – ER média (%) para cada o método de classificação utilizando regressão regularizada e regressão logística

Métodos	tempo
LASSO	62.26
R.L	**0.07**
Ridge	3.27

Tabela 30 – Tempo médio de processamento da classificação em segundos



(a) Treinamento **(b)** Teste

Figura 23 – ER para os diferentes métodos de classificação

5 CONSIDERAÇÕES FINAIS

A relevância das assinaturas nos processos de transações legais e, ao mesmo tempo, o risco com as falsificações cada vez mais especializadas mostram a importância em pesquisas que melhore, cada vez mais, o reconhecimento de assinaturas falsas e verdadeiras.

Neste trabalho foi utilizada a proposta *online*, juntamente com o banco de dados MCYT (MCYT Fingerprint subcorpus) para realizar as verificações de assinaturas manuscritas. Sendo assim, antes da classificação das assinaturas, foi feita extrações de quantificadores de informação não paramétricos diretamente das séries temporais como a estatística de tendência de Wallis e Moore e, a partir da distribuição de padrões das séries temporais com a técnica de Bandt e Pompe, foi calculada a entropia, complexidade e informação de Fisher. Além disso, a fim de avaliar a dinâmica das assinaturas, as derivadas de primeira e segunda ordem também foram calculadas e, novamente, calculados os quantificadores.

A partir do banco de dados formados com os quantificadores de informação, técnicas de seleção de características foram utilizadas para, dentre todas as características obtidas, pudessem selecionar as que fossem mais relevantes para a etapa de classificação. Sendo assim, foram obtidos quatro conjuntos de características segundo os critérios de seleção. Vale destacar que entre os quantificadores selecionados a informação de Fisher e estatística de Wallis e Moore foram as características mais relevantes. Pois nos quatro métodos de seleção de características utilizados, esses quantificadores estiveram presente em todos.

De forma geral, apesar de os classificadores apresentarem desempenhos próximos, o XGBoost apresentou melhor desempenho na etapa de treinamento, porém na etapa de teste o SVM e a Florestas Aleatórias apresentaram melhor desempenho na maior parte das métricas usadas para avaliar os classificadores. De acordo com o objetivo dessa dissertação esperamos que o classificador seja mais rigoroso ao classificar corretamente as assinaturas falsas. Dessa maneira, a sensibilidade é uma métrica importante para esse trabalho, sendo assim, o classificador Florestas Aleatórias apresentou melhor desempenho ao que se refere essa métrica entre os demais classificadores.

Na regressão logística regularizada, nenhum quantificador foi penalizado na seleção automática de variáveis, sendo assim, foram considerados as 24 características para a etapa de classificação com a regressão logística penalizada do tipo LASSO e Ridge. Posto isto, também foi avaliado o desempenho da regressão logística não regularizada, a fim de poder comparar se a regularização era uma estratégia mais eficiente nesse âmbito. Nesse sentido, a regressão logística

não regularizada apresentou melhores resultados, em geral, segundo as métricas de avaliação.

Vale ressaltar que, comparando a regressão logística não regularizada, utilizando todas as características, em relação a regressão logística, usando as técnicas de seleção de características, a primeira apresentou melhores resultados na classificação.

Como trabalhos futuros, pretende-se utilizar a classificação de uma classe e criar assinaturas falsas por meio de algoritmos como exemplo, redes neuronais. Além disso, pode-se utilizar função dos dados para encontrar um classificador com mais acurácia e utilização de técnicas de deep learning na verificação de assinaturas.

REFERÊNCIAS

ADAMSKI, M.; SAEED, K. Heuristic techniques for handwritten signature classification. **International Journal of Computing**, v. 5, n. 2, p. 87–92, 2014.

ANTAL, M.; SZABÓ, L. Z. Some remarks on a set of information theory features used for on-line signature verification. In: IEEE. **2017 5th International Symposium on Digital Forensic and Security (ISDFS)**. [S.l.], 2017. p. 1–5.

AQILI, N.; MAAZOUZI, A.; RAJI, M.; JILBAB, A.; CHAOUKI, S.; HAMMOUCH, A. On-line signature verification using point pattern matching algorithm. In: IEEE. **2016 International Conference on Electrical and Information Technologies (ICEIT)**. [S.l.], 2016. p. 410–413.

ASSUNçãO, E.; TEIXEIRA, M. C.; FARIA, F. A. Realimentação da derivada dos estados em sistemas multivariáveis lineares usando lmis. **Revista Controle & Automação**, v. 20, n. 1, 2009.

AWAD, W.; ELSEUOFI, S. Machine learning methods for spam e-mail classification. **International Journal of Computer Science & Information Technology (IJCSIT)**, v. 3, n. 1, p. 173–184, 2011.

BANDT, C.; POMPE, B. Permutation entropy: a natural complexity measure for time series. **Physical review letters**, APS, v. 88, n. 17, p. 174102, 2002.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD explorations newsletter**, ACM, v. 6, n. 1, p. 20–29, 2004.

BELGIU, M.; DRĂGUŢ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier, v. 114, p. 24–31, 2016.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. **Classification and regression trees**. [S.l.]: Routledge, 2017.

BRISSAUD, J.-B. The meanings of entropy. **Entropy**, Molecular Diversity Preservation International, v. 7, n. 1, p. 68–96, 2005.

CESSIE, S. L.; HOUWELINGEN, J. C. V. Ridge estimators in logistic regression. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 41, n. 1, p. 191–201, 1992.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: ACM. **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.], 2016. p. 785–794.

DAI, J.; XU, Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. **Applied Soft Computing**, Elsevier, v. 13, n. 1, p. 211–221, 2013.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. [S.l.]: USP/ESALQ, 2001.

DISATNIK, D.; SIVAN, L. The multicollinearity illusion in moderated regression analysis. **Marketing Letters**, Springer, v. 27, n. 2, p. 403–408, 2016.

DONG, H.; XU, X.; WANG, L.; PU, F. Gaofen-3 polsar image classification via xgboost and polarimetric spatial information. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 18, n. 2, p. 611, 2018.

ENGELSTAD, P. E.; HAMMER, H.; KONGSGÅRD, K. W.; YAZIDI, A.; NORDBOTTEN, N. A.; BAI, A. Automatic security classification with lasso. In: SPRINGER. **International Workshop on Information Security Applications**. [S.l.], 2015. p. 399–410.

ESMAEL, B.; ARNAOUT, A.; FRUHWIRTH, R. K.; THONHAUSER, G. A statistical feature-based approach for operations recognition in drilling time series. **International Journal of Computer Information Systems and Industrial Management Applications**, v. 5, p. 454–461, 2015.

FAN, J.; WANG, X.; WU, L.; ZHOU, H.; ZHANG, F.; YU, X.; LU, X.; XIANG, Y. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in china. **Energy conversion and management**, Elsevier, v. 164, p. 102–111, 2018.

FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FERNÁNDEZ-DELGADO, M.; CERNADAS, E.; BARRO, S.; AMORIM, D. Do we need hundreds of classifiers to solve real world classification problems? **The Journal of Machine Learning Research**, JMLR. org, v. 15, n. 1, p. 3133–3181, 2014.

FRANKE, K.; SOLAR, J. Ruiz del; KÖPPEN, M. Soft-biometrics: Soft-computing for biometric-applications. 2012.

FREUND, Y.; SCHAPIRE, R. E. *et al.* Experiments with a new boosting algorithm. In: CITESEER. **icml**. [S.l.], 1996. v. 96, p. 148–156.

FRIEDEN, B. R. **Science from Fisher information: a unification**. [S.l.]: Cambridge University Press, 2004.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

GARCIA-SALICETTI SONIA; HOUMANI, N. L.-V. B. D. B. A.-F. F. F. J. O.-G. J. V. C.; SCHEIDAT, T. Online handwritten signature verification. In: ____. [S.l.: s.n.], 2009. p. 125–165.

GEORGANOS, S.; GRIPPA, T.; VANHUYSSE, S.; LENNERT, M.; SHIMONI, M.; KALOGIROU, S.; WOLFF, E. Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application. **GIScience & remote sensing**, Taylor & Francis, v. 55, n. 2, p. 221–242, 2018.

GOEMAN, J.; MEIJER, R.; CHATURVEDI, N. L1 and l2 penalized regression models. **Vignette R Package Penalized. URL http://cran. nedmirror. nl/web/packages/penalized/vignettes/-penalized. pdf**, 2018.

GROSSE, I.; BERNAOLA-GALVÁN, P.; CARPENA, P.; ROMÁN-ROLDÁN, R.; OLIVER, J.; STANLEY, H. E. Analysis of symbolic sequences using the jensen-shannon divergence. **Physical Review E**, APS, v. 65, n. 4, p. 041905, 2002.

GROVER, P. Gradient boosting from scratch. **Retrieved from Medium**, 2017.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. [S.l.]: Bookman Editora, 2009.

HAND, D. J. Data mining. **Encyclopedia of Environmetrics**, Wiley Online Library, v. 2, 2006.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer Science & Business Media, 2009.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

HEARST, M. A.; DUMAIS, S. T.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. **IEEE Intelligent Systems and their applications**, IEEE, v. 13, n. 4, p. 18–28, 1998.

HILTON, O. Signatures—review and a new view. **Journal of Forensic Science**, ASTM International, v. 37, n. 1, p. 125–129, 1992.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, Taylor & Francis Group, v. 12, n. 1, p. 55–67, 1970.

HOSMER DAVID W; JOVANOVIC, B.; LEMESHOW, S. Best subsets logistic regression. **Biometrics**, JSTOR, p. 1265–1270, 1989.

HOSMER DAVID W; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. v. 398.

HOU, W.; YE, X.; WANG, K. A survey of off-line signature verification. In: IEEE. **2004 International Conference on Intelligent Mechatronics and Automation, 2004. Proceedings.** [S.l.], 2004. p. 536–541.

IMON, A.; KHAN, M. A. I. A solution to the problem of multicollinearity caused by the presence of multiple high leverage points. **Int. J. Stat. Sci**, v. 2, p. 37–50, 2003.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112.

KHOSHGOFTAAR, T. M.; ALLEN, E. B. Controlling overfitting in classification-tree models of software quality. **Empirical Software Engineering**, Springer, v. 6, n. 1, p. 59–79, 2001.

KOHAVI, R. *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, v. 160, p. 3–24, 2007.

KOUROU, K.; EXARCHOS, T. P.; EXARCHOS, K. P.; KARAMOUZIS, M. V.; FOTIADIS, D. I. Machine learning applications in cancer prognosis and prediction. **Computational and structural biotechnology journal**, Elsevier, v. 13, p. 8–17, 2015.

KOWALSKI, A.; MARTÍN, M.; PLASTINO, A.; ROSSO, O. Bandt–pompe approach to the classical-quantum transition. **Physica D: Nonlinear Phenomena**, Elsevier, v. 233, n. 1, p. 21–31, 2007.

KUMARI, R.; SRIVASTAVA, S. K. Machine learning: A review on binary classification. **International Journal of Computer Applications**, Foundation of Computer Science, v. 160, n. 7, 2017.

LEE, L. L.; BERGER, T.; AVICZER, E. Reliable on-line human signature verification systems. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, IEEE, n. 6, p. 643–647, 1996.

LOPEZ-RUIZ, R.; MANCINI, H. L.; CALBET, X. A statistical measure of complexity. **Physics Letters A**, Elsevier, v. 209, n. 5-6, p. 321–326, 1995.

MAGLOGIANNIS, I. G. **Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies**. [S.l.]: Ios Press, 2007. v. 160.

MARTIN, M.; PLASTINO, A.; ROSSO, O. Generalized statistical complexity measures: Geometrical and analytical properties. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 369, n. 2, p. 439–462, 2006.

MEIER, L.; GEER, S. V. D.; BÜHLMANN, P. The group lasso for logistic regression. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 70, n. 1, p. 53–71, 2008.

MORETTIN, P. A.; TOLOI, C. Análise de séries temporais. In: **Análise de séries temporais**. [S.l.: s.n.], 2006.

MULLER, K.-R.; MIKA, S.; RATSCH, G.; TSUDA, K.; SCHOLKOPF, B. An introduction to kernel-based learning algorithms. **IEEE transactions on neural networks**, IEEE, v. 12, n. 2, p. 181–201, 2001.

NALWA, V. S. Automatic on-line signature verification. **Proceedings of the IEEE**, IEEE, v. 85, n. 2, p. 215–239, 1997.

NGUYEN, V.; BLUMENSTEIN, M. Techniques for static handwriting trajectory recovery: a survey. In: ACM. **Proceedings of the 9th IAPR International Workshop on Document Analysis Systems**. [S.l.], 2010. p. 463–470.

OLIVARES, F.; PLASTINO, A.; ROSSO, O. A. Ambiguities in bandt–pompe's methodology for local entropic quantifiers. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 391, n. 8, p. 2518–2526, 2012.

OLIVARES, F.; PLASTINO, A.; ROSSO, O. A. Contrasting chaos with noise via local versus global information quantifiers. **Physics Letters A**, Elsevier, v. 376, n. 19, p. 1577–1583, 2012.

ORTEGA-GARCIA, J.; FIERREZ-AGUILAR, J.; SIMON, D.; GONZALEZ, J.; FAUNDEZ-ZANUY, M.; ESPINOSA, V.; SATUE, A.; HERNAEZ, I.; IGARZA, J.-J.; VIVARACHO, C. *et al.* Mcyt baseline corpus: a bimodal biometric database. **IEE Proceedings-Vision, Image and Signal Processing**, IET, v. 150, n. 6, p. 395–401, 2003.

O'BRIEN, R. M. A caution regarding rules of thumb for variance inflation factors. **Quality & quantity**, Springer, v. 41, n. 5, p. 673–690, 2007.

PANSARE, A.; BHATIA, S. Handwritten signature verification using neural network. **International Journal of Applied Information Systems**, Citeseer, v. 1, n. 2, p. 44–49, 2012.

PLAMONDON, R.; LORETTE, G. Automatic signature verification and writer identification—the state of the art. **Pattern recognition**, Elsevier, v. 22, n. 2, p. 107–131, 1989.

PLAMONDON, R.; SRIHARI, S. N. Online and off-line handwriting recognition: a comprehensive survey. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 22, n. 1, p. 63–84, 2000.

QIAO, Y.; LIU, J.; TANG, X. Offline signature verification using online handwriting registration. In: IEEE. **2007 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.], 2007. p. 1–8.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>.

RASHIDI, S.; FALLAH, A.; TOWHIDKHAH, F. Feature extraction based dct on dynamic signature verification. **Scientia Iranica**, Elsevier, v. 19, n. 6, p. 1810–1819, 2012.

RÉNYI, A. *et al.* On measures of entropy and information. In: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA. **Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics**. [S.l.], 1961.

ROSSO, O. A.; MASOLLER, C. Detecting and quantifying stochastic and coherence resonances via information-theory complexity measurements. **Physical Review E**, APS, v. 79, n. 4, p. 040106, 2009.

ROSSO, O. A.; MICCO, L. D.; PLASTINO, A.; LARRONDO, H. A. Info-quantifiers' map-characterization revisited. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 389, n. 21, p. 4604–4612, 2010.

ROSSO, O. A.; OLIVARES, F.; ZUNINO, L.; MICCO, L. D.; AQUINO, A. L.; PLASTINO, A.; LARRONDO, H. A. Characterization of chaotic maps using the permutation bandt-pompe probability distribution. **The European Physical Journal B**, Springer, v. 86, n. 4, p. 116, 2013.

ROSSO O A; OSPINA, R.; FRERY, A. C. Classification and verification of handwritten signatures with time causal information theory quantifiers. **PloS one**, Public Library of Science, v. 11, n. 12, p. e0166868, 2016.

SÁNCHEZ-MORENO P; YÁNEZ, R.; DEHESA, J. Discrete densities and fisher information. In: **Proceedings of the 14th International Conference on Difference Equations and Applications. Difference Equations and Applications. Istanbul, Turkey: Bahçesehir University Press**. [S.l.: s.n.], 2009. p. 291–298.

SANTOS, C.; JUSTINO, E. J.; BORTOLOZZI, F.; SABOURIN, R. An off-line signature verification method based on the questioned document expert's approach and a neural network classifier. In: IEEE. **Ninth International Workshop on Frontiers in Handwriting Recognition**. [S.l.], 2004. p. 498–502.

SCHAEFER, R.; ROI, L.; WOLFE, R. A ridge logistic estimator. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 13, n. 1, p. 99–113, 1984.

SHAGUFTA TAHSILDAR. **Random Forest Algorithm In Trading Using Python**. 2019. Disponível em: <https://blog.quantinsti.com/random-forest-algorithm-in-python/>. Acesso em: 22 dez. 2019.

SHANNON, C. E. A mathematical theory of communication. **Bell system technical journal**, Wiley Online Library, v. 27, n. 3, p. 379–423, 1948.

SHAWE-TAYLOR, J.; CRISTIANINI, N. *et al.* **Kernel methods for pattern analysis**. [S.l.]: Cambridge university press, 2004.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and computing**, Springer, v. 14, n. 3, p. 199–222, 2004.

STUART, A. The power of two difference-sign tests. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 47, n. 259, p. 416–424, 1952.

THISSEN, U.; PEPERS, M.; ÜSTÜN, B.; MELSSEN, W.; BUYDENS, L. Comparing support vector machines to pls for spectral regression applications. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 73, n. 2, p. 169–179, 2004.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

ÜSTÜN, B.; MELSSEN, W. J.; BUYDENS, L. M. Facilitating the application of support vector regression by using a universal pearson vii function based kernel. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 81, n. 1, p. 29–40, 2006.

VAPNIK, V. N. The nature of statistical learning theory. Springer-Verlag, 1995.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. **Pattern recognition**, Elsevier, v. 44, n. 2, p. 330–349, 2011.

WALLIS, W. A.; MOORE, G. H. A significance test for time series analysis. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 36, n. 215, p. 401–409, 1941.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016.

XIAO, X.-H.; LEEDHAM, G. Signature verification by neural networks with selective attention. **Applied Intelligence**, Springer, v. 11, n. 2, p. 213–223, 1999.

YADAV, S.; SHUKLA, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: IEEE. **2016 IEEE 6th International conference on advanced computing (IACC)**. [S.l.], 2016. p. 78–83.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: **Icml**. [S.l.: s.n.], 1997. v. 97, n. 412-420, p. 35.

ZHU, W.; ZENG, N.; WANG, N. *et al.* Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. **NESUG proceedings: health care and life sciences, Baltimore, Maryland**, v. 19, p. 67, 2010.

ZUBEN, F. V.; ATTUX, R. **Notas de aula IA004. Tópico 7-Árvores de Decisão [database on the Internet]. Campinas: Unicamp. Faculdade de Engenharia Elétrica e de Computação. Departamento de Engenharia de Computação e Automação Industrial.[Adobe Acrobat document, 44p.]**. 2001.