UNIVERSIDADE ESTADUAL DE PONTA GROSSA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

GABRIEL LUCAS FEDACZ

APRENDIZAGEM DE CLASSIFICADORES PARA IDENTIFICAÇÃO DE BACTÉRIAS RELAÇÃO ENTRE AS MEDIDAS DE COMPLEXIDADE DE DADOS E O DESEMPENHO DOS CLASSIFICADORES

PONTA GROSSA

2020

GABRIEL LUCAS FEDACZ

APRENDIZAGEM DE CLASSIFICADORES PARA IDENTIFICAÇÃO DE BACTÉRIAS RELAÇÃO ENTRE AS MEDIDAS DE COMPLEXIDADE DE DADOS E O DESEMPENHO DOS CLASSIFICADORES

Dissertação submetida ao Programa de Pós Graduação em Computação Aplicada, curso de Mestrado em Computação Aplicada - Área de concentração Computação para Tecnologias em Agricultura - da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de Mestre.

Orientador: Prof Dr. José Carlos Ferreira da Rocha
Coorientador: Prof Dr. Rafael Mazer Etto

PONTA GROSSA

2020

Fedacz, Gabriel Lucas

F292 Aprendizagem de Classificadores para Identificação de Bactérias: relação entre as medidas de complexidade de dados e o desempenho dos classificadores / Gabriel Lucas Fedacz. Ponta Grossa, 2020.

71 f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.

Orientador: Prof. Dr. José Carlos Ferreira da Rocha.
Coorientador: Prof. Dr. Rafael Mazer Etto.

1. Complexidade de dados. 2. Espectrometria de massa. 3. Classificação de bactérias. 4. Desbalanceamento de dados. I. Rocha, José Carlos Ferreira da. II. Etto, Rafael Mazer. III. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. IV.T.

CDD: 004

Ficha catalográfica elaborada por Maria Luzia Fernandes Bertholino dos Santos- CRB9/986

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
Av. General Carlos Cavalcanti, 4748 - Bairro Uvaranas - CEP 84030-900 - Ponta Grossa - PR - https://uepg.br

TERMO

TERMO DE APROVAÇÃO

Gabriel Lucas Fedacz

APRENDIZAGEM DE CLASSIFICADORES PARA IDENTIFICAÇÃO DE BACTÉRIAS:

Relação entre as medidas de complexidade de dados e o desempenho dos classificadores

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Prof. Dr. José Carlos Ferreira da Rocha - UEPG

Prof(a). Dr(a). Maria Berenice Reynaud Steffens - UFPR

Prof. Dr. Alceu de Souza Britto Junior - UEPG

Ponta Grossa, 28 de julho de 2020.

Documento assinado eletronicamente por **Alceu de Souza Britto Junior**, **Professor(a)**, em 28/07/2020, às 12:56, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Jose Carlos Ferreira da Rocha**, **Coordenador(a) do Programa de Pós-Graduação em Computação Aplicada - Mestrado**, em 28/07/2020, às 16:56, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Rafael Mazer Etto**, **Professor(a)**, em 29/07/2020, às 10:19, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Maria Berenice Reynaud Steffens**, **Usuário Externo**, em 03/08/2020, às 15:58, conforme art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.uepg.br/autenticidade informando o código verificador **0249193** e o código CRC **8482B249**.

20.000021289-2 0249193v5

Dedico este trabalho aos meus pais Rosangela e Laertes, por serem meu alento e fontes de minha dedicação, ao meu irmão Guilherme que sempre me apoiou e a todos os amigos que nunca deixaram de me incentivar.

AGRADECIMENTOS

Agradeço a Deus, por ter iluminado meu caminho, aos meus pais Rosangela e Laertes, heróis que me deram apoio e incentivo nos momentos mais difíceis.

Aos meus avós maternos (*in memorian*) e paternos que são minha fonte de inspiração e sempre me incentivaram a alcançar meus objetivos.

Ao meu orientador Prof. Dr. José Carlos F. da Rocha, que me acolheu como seu orientando, agradeço por todo conhecimento que recebi, também por todo apoio e paciência em todos os momentos deste trabalho.

Ao meu coorientador Prof. Dr. Rafael M. Etto e ao Me. Douglas Tomachewski, por toda contribuição, e em especial pela disponibilização dos dados que tornaram este trabalho possível.

À Universidade Estadual de Ponta Grossa e ao programa de Pós-Graduação em Computação Aplicada por me aceitarem como discente e me oferecerem todo amparo e infraestrutura.

Aos professores do programa de Pós-Graduação em Computação agradeço pelo compartilhamento de conhecimentos e dedicação à docência do programa.

Aos colegas de classe e amigos do Lab14, por toda a ajuda e apoio durante este período tão importante da minha formação acadêmica.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financeiro, o qual foi de extrema importância.

Por fim, agradeço a todos os amigos e amigas que direta ou indiretamente fizeram parte da minha formação, muito obrigado!

RESUMO

No meio agrícola, algumas bactérias têm sido utilizadas na promoção do biocontrole e crescimento vegetal. Isto tem motivado o desenvolvimento de ferramentas de software para detectar automaticamente sua presença em amostras coletadas do solo. Uma maneira de proceder tal identificação é o desenvolvimento de classificadores que utilizam padrões de espectros de massa obtido por MALDI/TOF para verificar a frequência de determinados conjuntos de proteínas ribossomais na amostra. A seleção de uma função de classificação adequada para o problema alvo tem grande influência sobre o desempenho do classificador e isto tem incentivado o uso de escores, denominados medidas de complexidade de dados. Tais escores descrevem certas características da base dados e podem fornecer suporte à escolha da função de classificação. Durante o processo de geração dos dados a partir de espectros de massa, é comum a ocorrência do desbalanceamento de classes, o que afeta adversamente as medidas de complexidade de dados. Considerando o exposto, este trabalho aplica um protocolo experimental para verificar a influência do desbalanceamento dos dados sobre o desempenho dos classificadores e nas medidas de complexidade. Os modelos classificadores utilizados nos experimentos foram a regressão logística e o QDA, os quais foram treinados para a identificação de bactérias dos gêneros *Bacillus* e *Rhizobium*. O desempenho dos classificadores apresentou relação exponencial com o balanceamento dos dados. Foram propostos dois índices de complexidade de dados, L2B e N3B que foram submetidas aos testes junto aos índices encontrados na literatura. Os resultados mostram que as medidas F3, *Density*, N3B e L2B estão relacionados ao desempenho dos classificadores treinados com dados desbalanceados. Tais medidas foram avaliadas quanto a capacidade em predizer a acurácia balanceada dos modelos. Na identificação de bactérias do gênero *Bacillus*, a medida de melhor relação com o desempenho em ambos os modelos foi a medida N3B. No caso da identificação do gênero *Rhizobium*, a medida de melhor associação ao modelo logístico foi L2B e N3B no modelo quadrático.

Palavras-chave: Complexidade de dados, Espectrometria de Massa, Classificação de Bactérias, Desbalanceamento de dados.

ABSTRACT

In the agricultural environment, some bacteria have been used as active in biocontrol and plant growth. This has motivated the development of software tools to automatically detect their presence in soil samples. One way to proceed with this identification is the development of classifiers that use MALDI / TOF mass spectra patterns to check the frequency of certain ribosomal proteins in the sample. The selection of a classification function that fits the target problem has a great influence on the classifier's performance, this has encouraged the use of scores, called data complexity measures. Such scores describe certain characteristics of the database and may provide support for choosing the classification function. During the process of generating data from mass spectrometry, it is common for data to be unbalanced, which adversely affects the data complexity measures. Considering the above, this work applies an experimental protocol to verify the influence of unbalanced data on the performance of classifiers and on complexity measures. The classifying models used in the experiments were logistic regression and QDA, which were trained to identify bacteria of the genera *Bacillus* and *Rhizobium*. The performance of the classifiers showed a strong to moderate relationship with the unbalanced data problem. Two data complexity indexes, L2B and N3B, have been proposed and submitted to tests along with the indexes found in the literature. The results show that the measures F3, *Density*, N3B and L2B are related to the performance of the classifiers trained with unbalanced data. Such measures were evaluated for their ability to predict the balanced accuracy of the models. When identifying bacteria of the genera *Bacillus*, the measure of best relation to the performance of the models was the N3B measure. In the case of the identification of the genera *Rhizobium*, the measure of best association with the logistic model was L2B and N3B for the quadratic model.

Keywords: Data Complexity, Mass Spectrometry, Bacterial Classification, Imbalanced Datasets.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As bactérias são organismos unicelulares que prosperam em diversos ambientes e podem ter grande influência no cultivo de produtos agrícolas (AMYES, 2013). Dependendo da interação com a cultura elas podem ser benéficas, inofensivas ou patogênicas para determinadas plantas. Como exemplo Banerjee, Gorthi e Chattopadhyay (2018) observam que algumas espécies de bactérias do gênero *Bacillus* são amplamente utilizadas como agentes de biocontrole. Pongslip (2012) e Masson-Boivin e Sachs (2018) relatam que bactérias do gênero *Rhizobium* convertem nitrogênio atmosférico em íons amônio que podem ser incorporados para gerar moléculas nitrogenadas ou sais (nitrato e nitrito) que agem como fertilizantes para os vegetais.

A importância destes organismos para a agricultura têm motivado o desenvolvimento de ferramentas de software para detectar automaticamente sua presença em amostras retiradas do solo. Em particular, o desenvolvimento de classificadores que analisam padrões associados à proteínas ribossomais em espectros de massa do tipo MALDI/TOF[1] tem se mostrado uma abordagem efetiva para esta tarefa. Neste contexto, destacam-se os trabalhos de Bou *et al.* (2011), Tamura, Hotta e Sato (2013), Schumann e Maier (2014) e Tomachewski *et al.* (2018).

Como observado por Norvig e Russell (2004), o treinamento de classificadores é um problema de aprendizagem supervisionado que envolve a indução de uma função de classificação a partir da análise de um conjunto de dados. Mesmo quando a forma[2] da função da classificação é escolhida *a priori* ainda resta a questão da estimação de seus parâmetros. A seleção de uma função de classificação adequada para o problema alvo tem grande influência sobre a eficácia do classificador. Segundo Ho e Basu (2002), Sotoca, Sánchez e Mollineda (2005) e Lorena *et al.* (2019) as dificuldades associadas à tal escolha têm fomentado a pesquisa por escores, chamados de medidas (ou índices) de complexidade de dados, que descrevam características da base de dados.

Ho e Basu (2002), e Kolaczyk e Csárdi (2014) propuseram uma série de medidas de complexidade que avaliam fatores como:

- a sobreposição dos valores dos atributos em classes diferentes;
- a separabilidade entre as classes - diz respeito à capacidade de determinados classificadores discriminarem instâncias de classes diferentes;

[1] Espectrometria de massa por dessorção/ionização a laser assistida por matriz com analisador por tempo de voo, do inglês *Matrix-Assisted Laser Desorption/Ionization Time-of-Flight*

[2] Polinomial, modelo probabilístico, representação lógica, etc.

- as características de cada classe em termos de estrutura, geometria e densidade.

Segundo Anwar, Jones e Ganesh (2014) os índices de complexidade de dados são afetados adversamente pelo desbalanceamento de classes. Isto é, quando o número de casos de uma classe é muito menor em relação a outra, alguns índices podem não quantificar corretamente o quão complexa é uma base de dados no que tange ao fator avaliado. Este fato estabelece uma contingência para o problema abordado neste trabalho pois, como destacado por Xu *et al.* (2016) e Tomachewski *et al.* (2018), o desbalanceamento de dados é uma condição comum em bases de dados referentes a espectros MALDI-TOF de gêneros de bactérias.

Considerando o exposto, este trabalho objetiva investigar o relacionamento entre os índices de complexidade de dados e o desempenho de classificadores para identificação de gêneros de bactérias em bases com desbalanceamento entre as classes. Adicionalmente, este trabalho também avalia duas medidas de complexidade de dados, L2B e N3B, que estendem duas medidas já existentes, L2 e N3. As medidas propostas exploram um escore que quantifica o desempenho preditivo de classificadores, chamado de acurácia balanceada, para avaliar a separabilidade das classes. Tal escore penaliza classificadores com baixo desempenho em termos de sensibilidade e especificidade.

Para atingir tais objetivos executou-se um conjunto de experimentos que abordaram a identificação de bactérias dos gêneros *Bacillus* e *Rhizobium*. Os conjuntos de dados usados nos experimentos foram produzidos a partir da base de dados PUKYU, oriunda dos trabalhos de Tomachewski (2017) e Tomachewski *et al.* (2018). Basicamente, os registros da base PUKYU foram utilizados na produção de dois conjuntos de dados binários, referentes aos gêneros *Bacillus* e *Rhizobium* que foram então usados para gerar bases de dados com diferentes proporções de balanceamento das classes. Na sequência, para cada uma das bases produzidas calculou-se as medidas de complexidade de dados e realizou-se o treinamento e validação dos modelos classificadores.

Os resultados dos testes indicaram uma relação linear entre o desbalanceamento e a exponencial do desempenho dos classificadores em termos de sensibilidade e acurácia balanceada. Dentre as medidas de complexidade testadas, as que tiveram relação ao desempenho dos classificadores foram F3 que apresentou associação linear e *Density* que apresentou associação exponencial. As medidas propostas a partir da adaptação de L2 e N3, também apresentaram uma associação linear ao desempenho dos modelos. Os testes com as bases de dados do gênero *Bacillus* mostraram que o índice mais relevante para a predição da acurácia balanceada do classificador logístico foi o N3B, seguido por Bal e L2B. Para o classificador QDA[3] o índice

[3] Análise Discriminante Quadrática, do inglês *Quadratic Discriminant Analysis*

de complexidade que teve maior grau de associação com a acurácia balanceada foi o N3B. Em seguida vieram a medida F3 e o nível de balanceamento. Os resultados obtidos com dados do gênero *Rhizobium* indicaram que a medida de complexidade mais relacionada ao desempenho do classificador logístico foi o L2B. Seguido pelo balanceamento e N3B. Para o classificador QDA o nível de balanceamento, seguido por N3B e então L2B.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Utilizar informações extraídas de espectros de massa obtidos por MALDI-TOF, para investigar a relação entre as medidas de complexidade de dados e o desempenho de classificadores na identificação de gêneros de bactérias em conjuntos de dados desbalanceados.

1.1.2 Objetivos Específicos

- verificar existência de dependência linear/não linear entre o desempenho dos classificadores e o desbalanceamento da base de dados;
- identificar os índices de complexidade relevantes para predição do desempenho dos modelos classificadores treinados com dados desbalanceados;
- avaliar a importância de medidas de complexidade de dados que exploram a acurácia balanceada para mensurar a separabilidade entre as classes na predição do desempenho dos modelos classificadores.

1.2 ORGANIZAÇÃO DO TRABALHO

O Capítulo 2 apresenta a revisão bibliográfica sobre identificação automática de bactérias, aprendizagem de classificadores e medidas de complexidade de dados. O Capítulo 3 descreve os materiais e métodos aplicados no procedimento experimental. O Capítulo 4 apresenta os resultados e a discussão. As conclusões são apresentadas no Capítulo 5.

2 REVISÃO BIBLIOGRÁFICA

Autores como Tomachewski *et al.* (2018) e Sauer *et al.* (2008) tem abordado o uso de classificadores na identificação automática de bactérias. No entanto, segundo Ali, Lee e Chung (2017) o desenvolvimento de sistemas de classificação automática depende de diversos fatores, entre eles: (a) da seleção de um modelo de classificação adequado e (b) das características da base de dados (NORVIG; RUSSELL, 2004). Neste sentido, Sotoca, Sánchez e Mollineda (2005) e Zubek e Plewczynski (2016) destacam que as medidas de complexidade de dados propostas por Ho e Basu (2002) fornecem uma maneira de quantificar um conjunto de características da base de treinamento e que têm se mostrado relacionadas à dificuldade do problema de classificação.

Outro fator que pode dificultar a aprendizagem de classificadores é o desbalanceamento classes. No que tange ao problema alvo deste trabalho, o desenvolvimento de classificadores para identificação de bactérias com dados de massa obtidos a partir de espectros do tipo MALDI-TOF, o desbalanceamento de classes tem se mostrado prejudicial para o desempenho preditivo do modelo de classificação em classes que eram minoritárias na base de treinamento (SCHUMANN; MAIER, 2014; TOMACHEWSKI *et al.*, 2018; SANTOS *et al.*, 2018). Além disso, segundo Anwar, Jones e Ganesh (2014), determinados índices de complexidade de dados são afetados adversamente pelo desbalanceamento das classes.

A fim de fornecer subsídios para investigar o efeito do desbalanceamento das classes sobre a complexidade dos conjuntos de dados usados no treinamento de classificadores para identificação de bactérias a partir de observações extraídos de espectros de massa, este capítulo apresenta uma descrição dos principais termos e conceitos relacionados ao treinamento de classificadores, à complexidade de dados e à identificação de bactérias. Desta forma, a Seção 2.1 descreve os principais conceitos ligados à classificação de dados e ao treinamento de classificadores. Em particular, são descritos os modelos classificadores utilizados nos experimentos. A Seção 2.2 descreve os conceitos de complexidade de dados. A seção 2.3 mostra como os dados obtidos por espectrometria de massa do tipo MALDI-TOF tem sido usados no desenvolvimento de classificadores para identificação automática de bactérias.

2.1 CLASSIFICAÇÃO DE DADOS

De acordo com Aggarwal (2014) o problema da classificação automática de objetos consiste em analisar as características de um alvo de interesse e determinar a que categoria ele pertence. O algoritmo responsável por esta tarefa, o classificador, é definido da seguinte maneira

por Norvig e Russell (2004) :

> Definição: Dado um vetor $\mathbf{x} = (x_1, \ldots, x_n)$, $n \in \mathbb{N}^+$, um classificador é uma função $f : \mathbf{X} \rightarrow \mathbf{C}$, $\mathbf{C} = \{c_1 \ldots, c_t\}$, que retorna um rótulo $c^* \in \mathbf{C}$, $t \in \mathbb{N}$ e $t > 1$.

Nesta definição, $\mathbf{X}$ é um conjunto $\{X_1 \ldots X_n\}$ de variáveis aleatórias que representam os atributos (características) que são usados para distinguir os objetos de um determinado domínio de acordo com a sua classe (categoria). Os elementos de $\mathbf{C}$ são rótulos que simbolizam as categorias e $\mathbf{x}$ é um vetor que representa uma n-upla $(x_1 \ldots, x_n)$, tal que x_i é o valor do atributo X_i. O espaço amostral de X_i é denotado por Ω_i. O objetivo do classificador f é determinar qual rótulo de classe $c^* \in \mathbf{C}$ é consistente com os valores observados em $\mathbf{x}$.

A aprendizagem automática de classificadores tem o objetivo de induzir a estrutura (forma) e estimar os parâmetros da função f a partir da inspeção de um conjunto de dados com amostras extraídas do domínio da aplicação (população alvo) (MITCHELL *et al.*, 1997). Na aprendizagem supervisionada, f é estimada com o auxílio de algoritmos que aplicam algum método de inferência indutiva sobre uma base dados $\mathbf{T}$ (NORVIG; RUSSELL, 2004). As entradas de $\mathbf{T}$ registram casos que foram previamente rotulados e o objetivo da inferência é estimar/gerar um modelo (função) que permita relacionar as entradas do classificador a seus respectivos rótulos de classe. Em outros termos, dados $\mathbf{X} = \{X_1 \ldots X_n\}$ e $\mathbf{C}$, a aprendizagem supervisionada objetiva induzir f a partir de uma base de dados $\mathbf{T}$. $\mathbf{T}$ possui $m \in \mathbf{N}^+$ casos tal que cada entrada de $\mathbf{T} \in \Omega_1 \cdots \times \Omega_n$.

Quando a estrutura do classificador (forma da função) é fixada *a priori*, o processo de aprendizagem supervisionada de classificadores também é chamado de treinamento. De acordo com Norvig e Russell (2004), se a base possui apenas dois rótulos de classe, o problema de classificação é dito binário ou dicotômico. Caso contrário, o problema é dito multi-classe ou politômico.

2.1.1 Análise do desempenho da aprendizagem de classificadores

Posto que, uma vez treinado, o classificador f será usado para rotular instâncias que não foram observadas durante o processo de aprendizagem, é necessário verificar se o seu desempenho atende aos requisitos da aplicação antes de utilizá-lo. Uma forma usual de abordar esta tarefa é testar o classificador em uma base de dados diferente daquela usada no treinamento (WONG, 2015). Muito frequentemente, a base de dados $\mathbf{T}$ é particionada em dois subconjuntos: a base de treinamento $\mathbf{T}_T$ e a base de teste $\mathbf{T}_t$. O procedimento de teste então registra a correção (ou erro) dos resultados obtidos pelo classificador ao rotular as instâncias em $\mathbf{T}_t$.

Seja um problema de classificação binária, em que uma classe é denominada *positiva* (+) e a outra *negativa* (-), o resultado produzido por um classificador na categorização de uma instância de teste pode ser descrito como (NORVIG; RUSSELL, 2004):

- Verdadeiro Positivo (VP): quando uma instância da classe alvo (+) é categorizada corretamente;
- Verdadeiro Negativo (VN): quando uma instância pertencente à classe complementar (-) é categorizada corretamente;
- Falso Negativo (FN): quando uma instância da classe alvo (+) é categorizada incorretamente;
- Falso Positivo (FP): quando uma instância da classe complementar (-) é categorizada de forma incorreta.

A partir destes conceitos é possível definir uma matriz, chamada matriz de confusão. Em problemas de classificação binária a matriz de confusão é uma matriz 2×2 cujas linhas indicam a classificação correta (indicada na base de treinamento, esperada) e as colunas indicam a predição do classificador (saída do classificador). Cada célula da matriz armazena o número de resultados que se encaixa nas condições indicadas pelos rótulos de linha e de coluna associado. Assim, o número total de previsões corretas feitas pelo classificador é VP+VN e o total número de previsões incorretas é FP+FN. A Tabela 1 mostra a estrutura de uma matriz de confusão.

Tabela 1: Matriz de confusão para um problema binário

		Classe Predita	
		+	-
Classe Real	+	VP	FN
	-	FP	VN

As seguintes medidas de desempenho podem ser computadas a partir da matriz de confusão (TAN, 2018):

- acurácia - é a proporção de predições corretas de um classificador sem considerar se os casos são positivos ou negativos. Pode ser visto como uma estimativa da probabilidade de se obter uma classificação correta em relação ao número de casos em $\mathbf{T}$ (Equação 1). Esta medida é suscetível ao desbalanceamento dos dados, portanto, sob tais condições

pode fornecer um indicador que superestima o desempenho do classificador (BEKKAR; DJEMAA; ALITOUCHE, 2013). Neste caso N representa o total de predições;

$$acuracia = \frac{VP + VN}{N} \tag{1}$$

- especificidade - mede o sucesso do classificador em predizer corretamente casos pertencentes à classe (-); é computada pela Equação 2:

$$especificidade = \frac{VN}{VN + FP} \tag{2}$$

- sensibilidade - também conhecida como a proporção dos verdadeiros positivos, a sensibilidade é uma medida que expressa o sucesso do modelo em detectar os casos que pertencem à classe (+). É encontrada com o auxilio da Equação 3:

$$sensibilidade = \frac{VP}{VP + FN} \tag{3}$$

O processo que avalia o desempenho de um classificador também pode ser efetuado por meio de um procedimento de validação cruzada (ARLOT; CELISSE *et al.*, 2010). A validação cruzada consiste em dividir aleatoriamente os elementos da base $\mathbf{T}$ em k subconjuntos (*folds*) de mesmo tamanho. Na sequência, um procedimento iterativo, seleciona um um dos subconjuntos para compor a base de teste e usa os $k - 1$ demais subconjuntos para o treinamento do classificador (WONG, 2015). Este processo é executado k vezes, garantindo que todos os subconjuntos sejam usados tanto no treino quanto no teste do modelo. Após a última repetição, o desempenho médio do classificador é computado como a média dos desempenhos obtidos nas bases de testes de cada iteração.

2.1.2 Avaliação do desempenho de classificadores em bases de dados desbalanceadas

Seja um problema de classificação binário em que $\mathbf{C} = \{c_1, c_2\}$. Sejam também m_1 e m_2 o número de casos de $\mathbf{T}$ rotulados como c_1 e c_2, respectivamente. Se $m_1 \ll m_2$ ou $m_1 \gg m_2$, $\mathbf{T}$ é dita desbalanceada, ou mais especificamente, diz-se que a base de dados apresenta desbalanceamento entre as classes (HE; GARCIA, 2009). A classe que contem a maioria dos casos, é denominada classe majoritária e a classe que possui menos ocorrências é dita minoritária.

De acordo com Haixiang *et al.* (2017), a maioria dos algoritmos de aprendizagem assume que a base de dados usada na indução de um classificador é balanceada. Ao violar esta suposição, uma base de dados desbalanceada pode direcionar o procedimento de aprendizagem

para um modelo que não é capaz de discriminar a classe minoritária (HE; GARCIA, 2009; ANWAR, 2012). Como em muitas situações, a classe minoritária é o alvo de interesse, tais classificadores apresentam baixa sensibilidade e alta especificidade. Em situações de desbalanceamento extremo, o desempenho do classificador na classe minoritária pode ser próximo a zero, sem que isto seja refletido na acurácia (KOTSIANTIS *et al.*, 2006).

Com o intuito de evitar uma superestimativa do desempenho de classificadores em bases desbalanceadas tem-se proposto o emprego de outras medidas de desempenho (BEKKAR; DJEMAA; ALITOUCHE, 2013). Uma destas medidas é a acurácia balanceada (*acbal*), que é definida conforme a Equação 4:

$$acbal = \frac{sensibilidade + especificidade}{2} \tag{4}$$

Segundo Patterson e Zhang (2007), ao avaliar o desempenho preditivo do classificador com a média aritmética entre a sensibilidade e a especifidade, a acurácia balanceada penaliza modelos com baixo desempenho em qualquer uma das classes.

2.1.3 Seleção de Atributos

A seleção dos atributos que servirão de entrada para uma função de classificação é um procedimento que, usualmente, ocorre antes do treinamento, como uma fase de pré-processamento. Esta etapa é importante porque o número de atributos utilizados no modelo e a contribuição de cada um deles na discriminação das classes pode influir no desempenho o classificador. Especificamente, se o treinamento de um classificador ocorre sobre uma base com muitos atributos, os seguintes fatores devem ser considerados : (a) possibilidade de sobreajuste do modelo e (b) a maldição da dimensionalidade (HAWKINS, 2004).

O sobreajuste (*overfit, em inglês*) de um classificador fica caracterizado quando o desempenho do modelo na base de testes é significativamente inferior àquela registrada na base de treinamento (SUBRAMANIAN; SIMON, 2013). Isto ocorre, principalmente porque o algoritmo de aprendizagem pode estimar um conjunto de parâmetros que descreve o comportamento da base de treinamento, mas não demonstra capacidade de identificar corretamente registros que não foram previamente apresentados ao modelo (HAWKINS, 2004). Um dos fatores que contribuem para o sobreajuste é o número de casos em $\mathbf{T}$ ser relativamente inferior em comparação ao número de atributos da base. Segundo Subramanian e Simon (2013), o *overfit* pode ser calculado como a divisão da acurácia do modelo na base de teste, pela acurácia do mesmo na base de treinamento, Equação 5. Neste caso, valores próximos a zero indicam a presença de sobreajuste.

$$\textit{overfit} = \frac{acuraciaBaseTeste}{acuraciaBaseTreino} \tag{5}$$

A maldição de dimensionalidade (*curse of dimensionality* em inglês) se refere ao fenômeno de que um espaço com muitas dimensões (ou variáveis) necessita de muitas instâncias para realizar uma determinada tarefa de classificação. Isto fica evidente quando se analisa o conceito de complexidade da amostra. Uma medida de complexidade da amostra define a quantidade de registros que é necessária para o treinamento de um modelo tal que a função aprendida tenha um erro inferior a um limiar ϵ, com probabilidade superior a $1 - \delta$ (KAKADE *et al.*, 2003). Onde, δ e ϵ são valores pré-especificados. O ponto a ser notado aqui é que, conforme observado por Bellman (1957), o número de instâncias exigido para gerar um modelo com um determinado nível de desempenho (acurácia ou acurácia balanceada), cresce exponencialmente em relação ao número das variáveis de entrada.

Os métodos de seleção de atributos provêm uma abordagem para mitigar as dificuldades enumeradas acima. Para tanto executam procedimentos que procuram identificar as variáveis que contribuem efetivamente para a tarefa de classificação proposta e remover aquelas que são irrelevantes (LI *et al.*, 2018). Neste trabalho, foram utilizados dois métodos de seleção de atributos, o teste de Mann-Whitney e o teste *Relief*. O teste de Mann Whitney (Wilcoxon rank-sum test) é um teste não paramétrico do tipo filtro utilizado para verificar se duas amostras pertencem ou não a mesma população (PÉREZ *et al.*, 2015). Na seleção de atributos para classificação binária, ele é empregado da seguinte forma:

- seja $\mathbf{X}'_i$ o conjunto de amostras da variável X_i que pertence à classe c e seja $\mathbf{X}"_i$ o conjunto de amostras pertence à classe $\overline{c}$;
- Aplica-se o teste de Mann-Whitney sobre $\mathbf{X}'_i$ e $\mathbf{X}"_i$;
- se $\mathbf{X}'_i$ e $\mathbf{X}"_i$ diferirem estatisticamente (*p-value* $< \alpha$[1]), o atributo é definido como relevante.

O teste *Relief* é um teste do tipo filtro e foi originalmente projetado para aplicação em problemas de classificação binária com atributos discretos ou numéricos. Utilizando uma medida de distância, este método gera um score pertencente ao intervalo [0,1] que representa a relevância para cada atributo em relação a classe (URBANOWICZ *et al.*, 2018). Este escore pode ser utilizado para criar um sistema de *rank*. Desta forma podem ser selecionados os atributos de melhor pontuação.

[1] nível de significância

2.1.4 Classificadores Discriminativos

Este trabalho considera classificação binária, categorizando uma bactéria como pertencente ou não a um determinado gênero. Dada esta restrição, este trabalho também assume, que na base de dados utilizada, os atributos descritores são contínuos e não possuem uma distribuição gaussiana (TOMACHEWSKI *et al.*, 2017).

Adicionalmente, os algoritmos de aprendizagem empregados neste trabalho, são baseados em modelos estatísticos, como afirma Mitchell *et al.* (1997), os modelos estatísticos são utilizados em tarefas onde todos os atributos têm valores contínuos ou ordinais. Tais modelos, comumente assumem que os valores dos atributos contidos na base T são independentes e identicamente distribuídos. A seguir são apresentados os métodos de classificação utilizados neste trabalho.

2.1.4.1 Regressão Logística

A regressão logística é um procedimento estatístico que tem como objetivo estimar, a partir de um conjunto de instâncias, os parâmetros de modelo probabilístico que relaciona uma variável de resposta $Y \in \{0, 1\}$ (variável de classe/decisão) a um vetor de entrada $\mathbf{x} \in \Omega_1 \times \Omega_2 \cdots \times \Omega_n$ (HAIR *et al.*, 2009). O modelo obtido pode ser visto como um classificador, chamado de classificador logístico, cuja função calcula as probabilidades posteriores de cada rótulo de classe, dada observação $\mathbf{x}$. Assim, dada a entrada $\mathbf{x}$, o classificador logístico computa $P(Y = 0|\mathbf{x})$ e $P(Y = 1|\mathbf{x})$ e aplica a regra de decisão bayesiana para selecionar a hipótese mais provável (DUDA; HART; STORK, 2012).

Na prática, é mais frequente que o classificador logístico utilize uma transformação chamada *logit* para computar o logaritmo das razões entre as probabilidades de cada hipótese conforme a Equação 6:

$$logit(\mathbf{x}) = ln\frac{P(Y = 1|\mathbf{x})}{1 - P(Y = 1|\mathbf{x})} = \beta_o + \beta_1 X_1 + ... + \beta_n X_n \tag{6}$$

Nesta equação, $\beta_i \in \mathbb{R}$, $i = 1 : n$, é o i-ésimo parâmetro do modelo, o coeficiente do atributo X_i. No caso de um classificador binário, a seleção da hipótese $Y = 1$ ($C = +$) acontece quando $logit(\mathbf{x}) > 0$; caso contrário, a hipótese resultante é $Y = 0$ ($C = -$).

2.1.4.2 Análise Discriminante Quadrática

A análise discriminante quadrática (QDA, do inglês *Quadratic Discriminant Analysis*) é um método utilizado para encontrar uma fronteira não linear entre as classes (VENABLES; RIPLEY, 2013). A função de classificação da QDA é dada pela equação:

$$P(Y = k|\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \left|\sum_k\right|^{1/2}} exp(-\frac{1}{2}(\mathbf{x} - \mu_k)^t \Sigma_k^{-1}(\mathbf{x} - \mu_k)) \tag{7}$$

Nesta expressão, n representa o numero de atributos, Σ_k a matriz de covariância da classe $k \in \{0, 1\}$ e μ_k são as médias das classes. Considerando a entrada $\mathbf{x}$, a regra de classificação do QDA consiste em encontrar a classe k que maximiza a função descrita na Equação 7.

2.2 COMPLEXIDADE DE DADOS

Ho e Basu (2002) e Kolaczyk e Csárdi (2014) propõem várias medidas de complexidade de dados para o problema de classificação binária. Tais medidas podem ser vistas como escores que procuram auxiliar os pesquisadores e os desenvolvedores de software envolvidos na confecção de classificadores a compreender a relação entre os conjuntos de dados e o desempenho de diferentes funções de classificação (REYES; OCHOA; TRINIDAD, 2005). Primeiro, porque fornecem indicadores que explicitam determinadas particularidades da estrutura e da distribuição subjacente a um conjunto de dados de treinamento (SOTOCA; SÁNCHEZ; MOLLINEDA, 2005). Segundo, porque proveem uma forma de evidência que tem se mostrado relacionada à complexidade de um dado problema de aprendizagem (CANO, 2013). E, finalmente, porque segundo Anwar (2012), fornecem uma visão geral sobre o espaço do problema.

Ho e Basu (2002) classificaram as medidas de complexidade em três diferentes categorias tal que as medidas dentro de cada grupo procuram capturar diferentes aspectos dos dados. Os grupos enumerados por aqueles autores são os seguintes:

- Medidas baseadas na sobreposição dos atributos;
- Medidas de separabilidade de classes;
- Medidas de topologia, geometria e densidade de variedades.

As medidas baseadas na sobreposição dos atributos estimam a eficácia dos atributos da base de dados considerando a separação ou sobreposição das classes. Seu objetivo é avaliar a

eficácia de cada atributo em separar as classes, e os efeitos de tais atributos sobre a composição da base. As medidas de separabilidade de classes estão relacionadas aos limites das classes, essas medidas identificam se duas classes são separáveis examinando a existência e a forma dos limites de cada classe no espaço de dados. As medidas de topologia, geometria e densidade de variedades oferecem uma caracterização indireta sobre a separabilidade de classes. A forma, a posição e a interconexão dos atributos são utilizados para identificar quão bem duas classes são separadas.

A Tabela 2 lista um conjunto de medidas de complexidade de dados enumeradas por Cano (2013), Anwar, Jones e Ganesh (2014) e Zubek e Plewczynski (2016). Segundo aqueles autores, diferentes experimentos sugerem que aquelas medidas se mostraram associadas ao desempenho dos algoritmos de aprendizagem de classificadores.

Tabela 2: Medidas de Complexidade de Dados

Grupos	ID	Detalhes, intervalo e direção da complexidade
Medidas de sobreposição de atributos	F1	Máxima discriminante de Fisher, $[0 \Leftarrow \infty]$
	F2	Volume de Sobreposição, $[0 \Rightarrow 1]$
	F3	Eficiência Máxima de Atributos, $[0 \Leftarrow 1]$
Medidas de separabilidade de classes	L1	Distância ao Separador Linear, $[0 \Rightarrow 1]$
	L2	Erro do Separador Linear, $[0 \Rightarrow 1]$
	N1	Pontos no limite da Classe, $[0 \Rightarrow 1]$
	N2	Relação entre média intra/inter classe, $[0 \Rightarrow 1]$
	N3	Taxa de Erro do Classificador 1-NN, $[0 \Rightarrow 1]$
Medidas de topologia, geometria e densidade	T1	Fração de Subconjuntos associados a cada classe, $[0 \Rightarrow 1]$
	Density	Densidade Média da Rede, $[0 \Leftarrow 1]$
	ClsCoef	Coeficiente de *Clustering*, $[0 \Leftarrow 1]$

2.2.1 Máxima função discriminante de Fisher (F1)

A função discriminante de Fisher tem como objetivo medir a separação entre dois conjuntos de dados univariados (CANO, 2013). Mais precisamente, esta função pode ser utilizada para estimar a separação ou sobreposição das instâncias de um atributo X_i condicional à variável de classe. O discriminante de Fisher é dado pela expressão:

$$f_i = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{8}$$

em que $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ representam respectivamente as médias e variâncias de X_i em cada classe.

Os autores Ho e Basu (2002) utilizam o escore f_i para definir uma medida de separa-

bilidade da base de treinamento em relação a um espaço d-dimensional. Para tanto, propõem o índice de complexidade de dados F1, o qual é computado como o máximo f_i, $i : 1..d$, do conjunto de atributos preditores $\mathbf{X} = \{X_1 \dots X_d\}$ conforme a Equação 9:

$$F1 = \max f_i \ : i = 1..d \tag{9}$$

Assim, $F1$ está relacionado ao atributo em que o ocorre a máxima separação entre as classes conforme medida pelo discriminante de Fisher. O domínio de F1 se encontra no intervalo $[0, +\infty]$ tal que baixos valores de F1 representam forte sobreposição e baixa separabilidade.

2.2.2 Volume de Sobreposição (*Overlap*) (F2)

De acordo com Ho e Basu (2002) esta medida é definida como o volume d-dimensional da sobreposição das faixas dos valores dos atributos entre as duas classes e é calculado de acordo com a Equação 10:

$$F2 = \prod_{i=1}^{d} \frac{MIN(max(X_i, c_1), max(X_i, c_2)) - MAX(min(X_i, c_1), min(X_i, c_2))}{MAX(max(X_i, c_1), max(X_i, c_2)) - MIN(min(X_i, c_1), min(X_i, c_2))} \tag{10}$$

Nesta equação, $max(X_i, c_j)$ e $min(X_i, c_j)$ são os valores máximo e mínimo de cada atributo X_i na classe c_j, $j = 1, 2$. Baixos valores de F2 representam baixo volume de sobreposição. O alcance desta função é $[0, 1]$.

2.2.3 Eficiência Máxima de Atributos (F3)

Segundo Cano (2013) a eficiência de um atributo é definida pela fração de todos os pontos separáveis pelo próprio atributo. O índice F3 verifica atributos individuais e mensura o quanto cada um contribui para a separação das duas classes. Se for detectada sobreposição (*overlap*), é considerado que a classe não é completamente separável por este atributo.

A eficiência máxima em um problema d-dimensional por atributo é definida por:

$$F3 = \sum_{i=1}^{d} \frac{|MIN(max(X_i, c_1), max(X_i, c_2)), MAX(min(X_i, c_1), min(X_i, c_2))|}{d} \tag{11}$$

onde, $max(X_i, c_j)$ e $min(X_i, c_j)$ são os valores máximo e mínimo de cada atributo

X_i na classe c_j, onde $i = 1 : n$ e $j = 1, 2$ para problemas binários. Baixos valores de F3 representam alta sobreposição. Os valores desta função estão no intervalo $[0, 1]$.

2.2.4 Separação Linear (L1 e L2)

O índice de complexidade L1 é uma medida de linearidade calculada através da minimização do erro de uma função objetivo, geralmente relacionada a um separador linear (HO; BASU, 2002). Para tratar ambos casos separáveis e não separáveis, Smith (1968) propõe a minimização da seguinte formula:

$$\begin{aligned} &\text{minimizar} && a^t t \\ &\text{sujeito a} && Z^t w + t \geq b \\ & && t \geq 0 \end{aligned}$$

Onde a e b são vetores constantes (definidos como 1). w é um vetor de pesos, t é o vetor de erro e Z é a matriz onde cada coluna z é definida como uma entrada do vetor x. A classe c (com valores c_1 e c_2) é expressa na próxima equação:

$$\begin{cases} z = +x \text{ se } c = c_1 \\ z = -x \text{ se } c = c_2 \end{cases} \tag{12}$$

O valor $a't$ da função objetivo visto anteriormente é utilizado como medida (L1). O valor desta medida é zero caso o problema seja linearmente separável. Altos valores de L1 indicam que os dados não são linearmente separáveis. O domínio da medida L1 está entre $[0, 1]$.

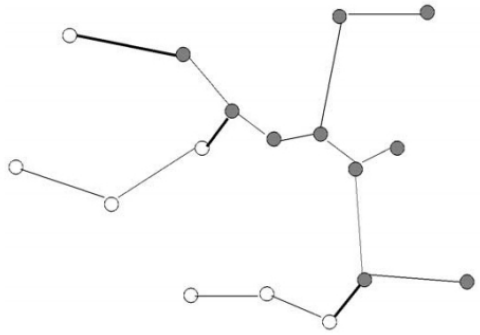
Considerando a descrição da medida L1, o índice L2 computa o erro de classificação para o separador linear utilizado anteriormente. Neste caso é computado o erro com base na acurácia, referente ao treinamento com a base de dados original. O domínio da medida L2 está no intervalo $[0, 1]$ altos valores indicam que a base não é linearmente separável.

2.2.5 Pontos no Limite da Classe (N1)

O escore N1 é calculado por meio da construção de uma árvore geradora miníma (MST, do inglês *minimum spanning tree*) sobre todo o conjunto de dados (HO; BASU, 2002). A MST conecta cada ponto da base ao seu vizinho mais próximo. Em seguida, o número de pontos que se conectam a pontos da classe oposta são contados. Estes são definidos pontos próximos ao limite da classe. A Figura 1, ilustra a aplicação de uma MST, onde os pontos claros e escuros representam os os casos pertencentes às classes c_1 e c_2, respectivamente. As arestas representam

qual o vizinho mais próximo de cada ponto. Neste caso, as arestas destacadas evidenciam quais são os pontos próximos ao limite da classe.

Figura 1: Árvore de abrangência mínima conectando os pontos de duas classes.



Fonte: Adaptado de Ho e Basu (2002)

A fração do número de tais pontos em relação a todos os pontos do conjunto de dados define a medida N1. O domínio desta medida é [0,1] e altos valores indicam que a maioria dos pontos estão próximos ao limite da classe.

2.2.6 Relação entre as médias das distâncias intra e inter classes dos vizinhos mais próximos (N2)

De acordo com Anwar, Jones e Ganesh (2014), quando os dados das classes estão muito sobrepostos, alguns registros podem estar próximos ao limite que separa as classes. Para identificar esta situação, é proposto o índice N2 (HO; BASU, 2002). Neste caso é computada a distância euclidiana de cada registro ao seu vizinho mais próximo pertencente a mesma classe ($intraDist$), e o mesmo é feito considerando apenas os vizinhos pertencentes a classe oposta ($interDist$). A razão entre as somatórias destes valores é a medida N2, como indicada na Equação 13:

$$N2 = \frac{\sum_{i=0}^{N} intraDist(x_i)}{\sum_{i=0}^{N} interDist(x_i)} \tag{13}$$

Nesta equação, para cada instância x_i são calculadas as suas respectivas distâncias $intraDist(x_i)$ e $interDist(x_i)$.

O domínio da métrica N2 está no intervalo [0, 1]. Valores baixos sugerem que as duas classes estão bem separadas, também sugerem que os registros de mesma classe estão bem agrupados.

2.2.7 Taxa de Erro do Classificador 1-NN (N3)

De acordo com Tan (2018) algoritmo de classificação baseado no vizinho mais próximo (do inglês *Nearest Neighbor* -NN) é uma técnica vastamente utilizada para reconhecer padrões. Este algoritmo funciona descobrindo o vizinho mais próximo a uma determinada instância. Dado isto, a medida N3 corresponde ao erro computado sobre a acurácia de classificação de um modelo 1-NN(CANO, 2013).

Esta medida mostra quão próximos estão os exemplos de diferentes classes. Os valores de N3 estão no intervalo $[0, 1]$. Valores baixos indicam que as fronteiras das classes estão relativamente distantes, mostrando que os dados são separáveis.

2.2.8 Fração de Subconjuntos associados a pontos por classe (T1)

Esta medida de complexidade foi proposta por Frank e Hubert (1996) e descreve a variedade das classes através do conceito de agrupamento de dados. Um subconjunto pode ser visto como uma esfera centrada sobre um registro x da base, este subconjunto cresce conforme encontra amostras próximas e que pertencem a mesma classe de x. Quando é encontrado um registro de outra classe, o subconjunto deixa de incluir amostras. A medida T1 computa o número de subconjuntos necessários para representar cada classe. Este valor é normalizado pelo número total de registros.

O domínio dos valores de T1 é o intervalo $[0, 1]$. Valores de baixos de T1 indicam que as instâncias que compõe a base são agrupadas e as fronteiras das classes são bem definidas.

2.2.9 Densidade Média da Rede (*Density*)

Dentre as medidas de complexidade que são derivadas da representação da base através de grafos, destaca-se a densidade média da rede (ou *Density*) (MORAIS; PRATI, 2013). Neste índice, cada registro da base de dados é representado como um nó ou vértice, e os nós são conectados se a sua distância no espaço de busca corresponder a algum critério, neste caso utilizada a distância euclidiana. Em seguida, as arestas que conectam exemplos de classes contrastantes são removidas.

A densidade média da rede é representada pela divisão do número de arestas do grafo pelo número máximo de arestas possíveis entre os pares de pontos (ROSEDAHL; ASHBY, 2019). A medida *Density* é dada pela Equação 14:

$$Density = \frac{N}{(n(n-1))/2}. \tag{14}$$

Nesta expressão, n representa o número de nós (ou registros da base) e N o número de arestas do grafo. O domínio dos valores de *Density* é o intervalo $[0, 1]$. Valores altos para este índice indicam que indivíduos da mesma classe formam grafos de maior densidade, evidenciando menor complexidade.

2.2.10 Coeficiente de *Clustering* (*ClsCoef*)

O Coeficiente de *Clustering* (ou *ClsCoef*) é um escore que mede a densidade das redes locais, obtidas a partir de uma base de dados representada em forma de grafo (ROSEDAHL; ASHBY, 2019). Primeiro para cada nó, é definido um subconjunto com os nós vizinhos que são diretamente conectados. O índice *ClsCoef* se refere a média entre a densidade (ver Seção 3.2.9) destes subconjuntos. Valores altos indicam que a base de dados é mais simples, sugerindo maior separabilidade entre os dados (GARCIA; CARVALHO; LORENA, 2015).

2.3 IDENTIFICAÇÃO DE BACTÉRIAS COM ESPECTROMETRIA DE MASSAS MALDI-TOF

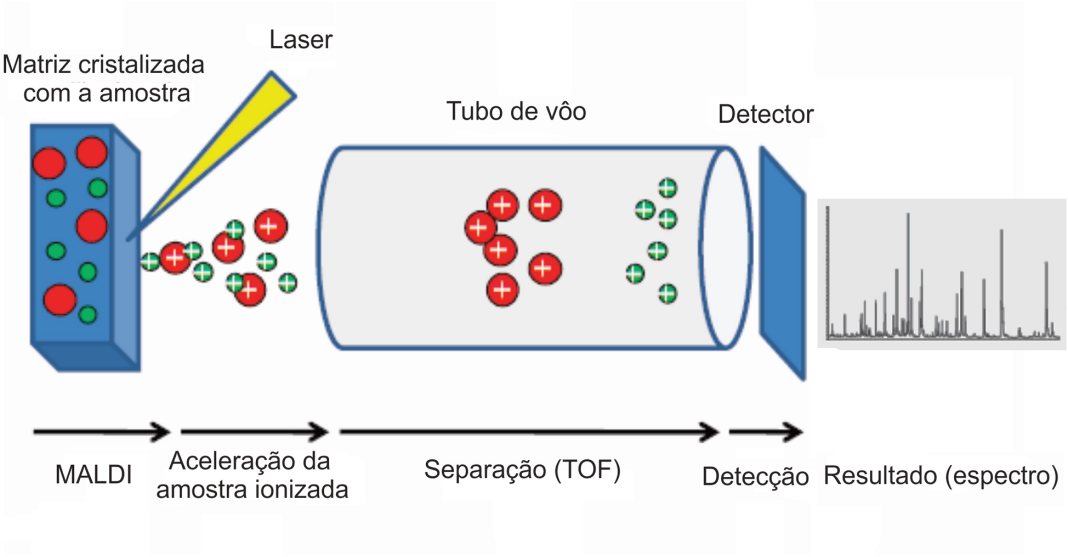
A espectrometria de massa (EM) é uma técnica que permite a identificação e quantificação da relação entre a massa e a carga elétrica de uma molécula (BIEMANN, 1963). Os valores referentes a medidas de espectrometria de massa são então expressos em unidades denotadas por m/z (razão massa / carga). Segundo Pasternak (2012) esta técnica tem sido utilizada para identificar proteínas bacterianas e tem sido aplicado na identificação de microrganismos de diferentes gêneros e espécies.

De acordo com Wieser *et al.* (2012), um dos métodos de EM que tem sido utilizados na análise de bactérias é o MALDI-TOF. MALDI (do inglês, *Matrix Assisted Lazer Desorption Ionization*) é a técnica utilizada para ionização da amostra e o TOF (do inglês, *Time of flight*) é o analisador de massa, responsável pela separação das espécies ionizadas em função da sua razão m/z, este método foi proposto por Tanaka *et al.* (1988). Basicamente, o método MALDI-TOF faz uso de um espectrômetro de massa que é responsável pela separação das espécies ionizadas em função da razão m/z. Após a ionização, um campo elétrico atrai as partículas eletricamente carregadas para o detector. A tensão usada para atrair os íons é constante e a velocidade com que cada partícula se move em direção ao detector depende da sua relação *m/z*. O tempo gasto para cada partícula atingir o detector é chamado de tempo de voo e também depende da razão

m/z. De forma simples, moléculas com maior relação *m/z* são mais lentas, enquanto moléculas de menor *m/z* chegam mais rapidamente ao sensor (WIESER *et al.*, 2012; GOULART; RESENDE, 2013). A disposição dos dados obtidos com a EM MALDI-TOF em um espectro de massas gera o espectro de MALDI-TOF, onde o eixo das abcissas (X) evidencia a relação m/z e o eixo das ordenadas (Y) representa a intensidade de cada pico (GROSS, 2006).

A Figura 2 ilustra o processo de obtenção de um espectro de MALDI-TOF de uma amostra arbitrária.

Figura 2: Exemplo do processo de obtenção de espectro por MALDI-TOF.



Fonte: Adaptado de Croxatto, Prod'hom e Greub (2012)

A espectrometria MALDI-TOF tem sido usada com êxito nas tarefas de identificação de proteínas e identificação de taxonomia de microrganismos, entre outras aplicações (SCHUMANN; MAIER, 2014). Neste contexto, a espectrometria MALDI-TOF tem fornecido uma forma de identificar a presença de proteínas ribossomais em amostras extraídas de bactérias. Segundo Tamura, Hotta e Sato (2013) as proteínas ribossomais contribuem para o próprio funcionamento celular e se caracterizam por serem abundantes e altamente conservadas. Podendo ser facilmente detectadas, de forma independente do ciclo celular.

Suas características incluem alta presença no organismo e alta capacidade de conservação. Ou seja, estas proteínas apresentam a mesma fisiologia independente do crescimento ou ciclo celular.

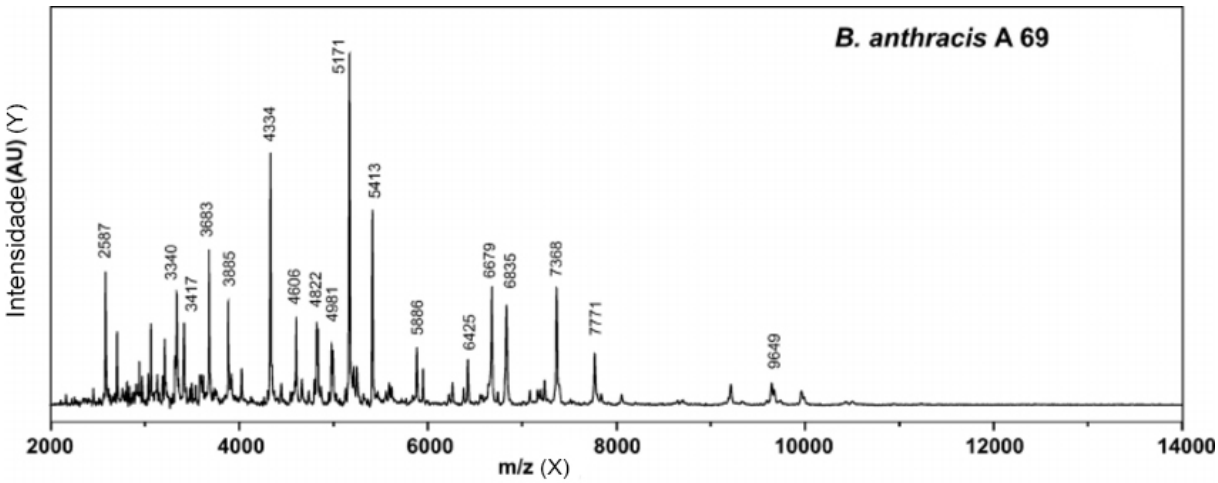
Tais características têm permitido que as proteínas ribossomais sejam biomarcadores confiáveis para identificação de bactérias (TERAMOTO *et al.*, 2007). Um biomarcador (ou marcador biológico) permite que as características de um processo biológico sejam mensuradas.

Estes marcadores, podem fornecer informações para a identificação da espécie ou gênero de uma determinada bactéria (TAMURA; HOTTA; SATO, 2013).

Para evidenciar os padrões dos biomarcadores em um espectro, é necessário determinar os valores *m/z*, relacionados a cada pico do espectro. Em seguida são selecionados aqueles valores no eixo X que estão relacionados às proteínas ribossomais. Os valores selecionados são então usados para compor vetor $(x_1, \ldots, x_n)$, que informa os pesos moleculares de cada proteína detectada na amostra analisada (TERAMOTO *et al.*, 2007). Como diferentes espécies e gêneros de bactérias podem possuir diferentes conjuntos de proteínas ribossomais, o espectro de massa pode fornecer evidências para sua identificação.

A Figura 3, adaptada do trabalho de Lasch *et al.* (2009), representa o espectro de massa para uma bactéria da espécie *Bacillus anthracis*. Neste exemplo foram destacados 18 picos.

Figura 3: Exemplo de Espectro de Massa com indicação de picos para a estirpe *Bacillus anthracis* A 69



Fonte: Adaptado de Lasch *et al.* (2009)

De acordo com Lay Junior (2001) identificar bactérias utilizando EM é uma tarefa promissora e deve receber cada vez mais atenção nos laboratórios, pois, quando a identificação por EM do tipo MALDI-TOF é comparada a métodos tradicionais, tal como sequenciamento de DNA, mostra uma economia na relação tempo/custo. De acordo com Hsieh *et al.* (2008) gerar um espectro MALDI-TOF é um processo que pode levar minutos, já o tempo de um teste tradicional pode levar até dias para ficar pronto. Em particular, Tomachewski *et al.* (2018) e Santos *et al.* (2018) obtiveram modelos classificadores com taxas de acerto superiores a 98% na identificação de gêneros, tais autores analisavam dados referentes a picos do espectro de massa associados às proteínas ribossomais.

3 MATERIAIS E MÉTODOS

Este capítulo descreve as etapas efetuadas para avaliar a relação entre o desempenho dos classificadores, as medidas de complexidade e o desbalanceamento das bases de dados. A fim de obter informações para esta avaliação, a performance dos classificadores foi medida de acordo com a sua acurácia balanceada, sensibilidade e sobre-ajuste na classificação de bactérias, para diferentes níveis de balanceamento na base de treinamento. Adicionalmente, foram computados os índices de complexidade para cada base de treinamento.

Assim, os experimentos realizados neste trabalho agrupam dois conjuntos de testes. O primeiro experimento testa o desempenho dos classificadores logístico e QDA nas tarefas de:

- categorizar bactérias como pertencentes ou não ao gênero *Bacillus*;
- categorizar bactérias como pertencentes ou não ao gênero *Rhizobium*.

Os atributos de entrada para cada classificador foram as abcissas de picos em espectros de massa MALDI-TOF relacionados a um conjunto de proteínas ribossomais. As bases de dados usadas no treinamento dos classificadores foram sintetizadas a partir da base de dados PUKYU e possuem um nível de balanceamento que varia entre 0,05% a 15% no que se refere a razão entre o número de casos da classe minoritária e o tamanho do conjunto de dados. O segundo experimento computou uma coleção de índices de complexidade de dados sobre os conjuntos de dados processados durante o primeiro experimento.

Como visto nas Seções 2.2.4 e 2.2.7, as medidas de complexidade L2 e N3 computam escores que estão relacionados à acurácia (erro) dos classificadores linear e 1-NN, respectivamente. Contudo, como apontam Bekkar, Djemaa e Alitouche (2013), a acurácia não é adequada para identificar o desempenho de modelos treinados com dados desbalanceados. Considerando isto, foram testadas duas medidas de complexidade, L2B e N3B, que quantificam a separabilidade das classes como o erro relativo a acurácia balanceada.

A seguir são detalhados os procedimentos de geração das bases de dados usadas durante os experimentos, Seções 3.1 e 3.2. Na sequência, o protocolo experimental deste trabalho é apresentado, Seção 3.3. A última seção (3.4) resume a forma de análise dos resultados.

3.1 A BASE DE DADOS PUKYU

A base de dados empregada na produção dos conjuntos de dados dicotômicos utilizados na geração das bases de dados sintéticas é intitulada PUKYU (TOMACHEWSKI *et al.*, 2018). Esta base foi produzida a partir de dados obtidos no repositório do NCBI (Centro Nacional de Informação Biotecnológica) na data de 13/06/2016. Os atributos da base PUKYU representam picos em espectros de massa MALDI-TOF que são associados à determinadas proteínas ribossomais. Esta base também contêm um atributo que define a taxonomia do organismo.

O conjunto de dados da PUKYU contém 60 atributos que se referem as massas moleculares das seguintes proteínas: L1, L2, L3, L4, L5, L6, L7, L7a, L7ae, L7/L12, L9, L10, L11, L12, L13, L14, L15, L16, L17, L18, L19, L20, L21, L22, L23, L24, L25, L27, L28, L29, L30, L3 1, L32, L33, L34, L35, L36, S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, S22 e S31e. Nesta base, os pesos moleculares são dados contínuos e as taxonomias dados nominais.

A base PUKYU se refere a um conjunto multi-classe, possui 30.033 registros que representam 7.491 espécies de 2.166 gêneros. Uma das dificuldades que pode surgir ao utilizar classificadores em bases MALDI-TOF é a ausência de dados (HURLBERT; JETZ, 2007). Tomachewski *et al.* (2018) evidencia que 35.14% dos campos da PUKYU não estão preenchidos. A Tabela 3 apresenta um exemplo de registro encontrado na base PUKYU. Naquela Figura, os valores representam a m/z das proteínas em questão e NA indica um dado faltante.

Tabela 3: Exemplo de conjunto de dados (massa/carga) para proteínas de L1-S31e obtidos da base PUKYU.

L1	L2	L3	L4	L5	L6	L7	L7a
24519.08	30023.85	23602.07	22280.5	NA	19655.38	NA	NA
L7ae	L7/L12	L9	L10	L11	L12	L13	L14
NA	12580.37	16453.98	NA	15268.86	NA	16333.43	13135.35
L15	L16	L17	L18	L19	L20	L21	L22
15596.84	16257.05	13616.57	13097.06	13244.54	13582.09	11333.08	12834.92
L23	L24	L25	L27	L28	L29	L30	L31
10605.39	11536.47	NA	10314.78	6846.05	NA	6422.49	9722.93
L32	L33	L34	L35	L36	S1	S2	S3
6353.38	5887.8	6043.32	7566.09	4305.39	NA	28963.2	23968.63
S4	S5	S6	S7	S8	S9	S10	S11
22882.2	17653.37	11724.28	17663.34	14700.13	14698.78	11445.3	13765.8
S12	S13	S14	S15	S16	S17	S18	S19
15203.71	13587.67	10409.24	10477.08	10103.68	10043.66	9220.82	10484.07
S20	S21	S22	S31e	Taxonomia			
8890.26	8004.43	NA	NA	*Staphylococcus aureus*			

A fim de contornar o problema da falta de dados, a base de dados PUKYU foi pré-processada em duas etapas. Inicialmente, foram removidos os registros que possuíam dados de três picos ou menos (com pelo menos 57 registros faltantes). Isto resultou em uma base com 28.505 registros, e 31.89% de casos incompletos. Em seguida os dados faltantes que sobraram foram imputados utilizando o algoritmo kNN (*k Nearest Neighbors*) (LU *et al.*, 2011) conforme descrito por Santos *et al.* (2018). A base de dados obtida com o procedimento de remoção de instâncias e imputação de dados foi denominada PUKYU-i.

Assim como a base PUKYU (TOMACHEWSKI *et al.*, 2018), a base PUKYU-i também apresenta desbalanceamento entre as classes. Este fato pode ser observado na Tabela 4, que indica alguns casos de desbalanceamento presentes no conjunto de dados. Nesta tabela são indicadas as proporções do número de casos de algumas taxonomias em relação ao número total de casos da PUKYU. *Classe 1* indica a categoria alvo, *Classe 2* representa a categoria complementar.

Tabela 4: Exemplos de desbalanceamento entre as classes, obtidos da base PUKYU.

Gênero	Classe 1	Classe 2	Balanceamento
Mycobacterium	1936	28097	6,446%
Bacillus	552	29481	1,837%
Rhizobium	112	29921	0,372%
Lactococcus	33	30000	0,109%

3.2 GERAÇÃO DAS BASES DE DADOS BINÁRIAS E SELEÇÃO DE ATRIBUTOS

Com o intuito de treinar os classificadores dicotômicos citados no início do capítulo, a base de dados PUKYU-i foi utilizada na geração de dois conjuntos de dados para classificação binária. As instâncias contidas no primeiro e no segundo conjunto de dados, chamados de *Bacillus* e *Rhizobium*, respectivamente, foram rotuladas com o propósito de distinguir aquelas que fazem parte dos gêneros *Bacillus* e *Rhizobium* das demais.

O procedimento de geração das bases de dados binárias *Bacillus* e *Rhizobium* é apresentado na listagem do Algoritmo 1. Além dos atributos descritores enumerados na seção 3.1 cada um destes conjuntos de dados possui um atributo binário chamado *Classe* que indica quais casos pertencem à taxonomia de interesse. Assim, nos conjuntos de dados usados no treinamento de classificadores para a base *Bacillus*, $Classe = 1$ (verdadeiro) indica que o registro pertence ao gênero *Bacillus* ao passo que $Classe = 0$ (falso) especifica o contrário. A valoração do atributo *Classe* ocorre de maneira análoga para os conjuntos deados usados nos experimentos sobre o gênero *Rhizobium*.

A escolha do gênero *Bacillus* se deve ao fato de algumas de suas espécies estarem associadas ao biocontrole e também ao crescimento vegetal como é o exemplo das *Bacillus insolitus* e *Bacillus cereus* (RADHAKRISHNAN; HASHEM; ABD_ALLAH, 2017). A seleção do gênero *Rhizobium* foi motivada porque especies desse gênero formam nódulos em leguminosas e, ao mesmo tempo que proliferam nessas estruturas especializadas, fixam nitrogênio em compostos orgânicos que são utilizados pelas plantas (PONGSLIP, 2012).

Algoritmo 1: GERADOR DE BASES DE DADOS BINÁRIAS

Entrada: r, um rótulo de classe; A, atributo alvo na base PUKYU-i

Saída: Base de dados dicotômica em relação ao rótulo R

```
1  início
2      Inicialize uma tabela D_r com os mesmos atributos descritores de PUKYU-i e
         um atributo adicional denominado Classe
3      para cada caso d ∈ PUKYU-i faça
4          adicione um novo caso d' em D_r
5          copie os dados dos atributos descritores de d para d'
6          se d.A = r então
7              d'.Classe = 1
8          senão
9              d'.Classe = 0
10         fim
11     fim
12 fim
13 retorna D_r
```

A execução do Algoritmo 1 com os pares de parâmetros ("Bacillus", Gênero), ("Rhizobium", Gênero) resultou, nos conjuntos de dados *Bacillus* e *Rhizobium* . A Tabela 5 informa o número de casos de cada classe e o nível de balanceamento (em %) nestas bases.

Tabela 5: Casos por classe nas bases binarias produzidas

Base	Classe 1	Classe 2	Balanceamento
Bacillus	552	29.481	1.84%
Rhizobium	112	29.921	0.37%

Para identificar quais proteínas eram relevantes para classificar as taxonomias alvo em cada base, os conjuntos de dados *Bacillus* e *Rhizobium* foram sujeitas a uma série de procedimentos de seleção de atributos. Inicialmente, cada uma das bases foi particionada em dois conjuntos de dados $\mathbf{D}_V$ e $\mathbf{D}_F$, com os casos categorizados como *Verdadeiro* e *Falso*, respectivamente. Em seguida, o teste de independência de Mann-Whitney foi aplicado sobre os casos de $\mathbf{D}_V$ e $\mathbf{D}_F$, separadamente, e foram removidos aqueles cujo *valor-p* era maior que 0,05. A Tabela 6 lista as proteínas que foram selecionadas após a primeira etapa:

Tabela 6: Proteínas excluídas com base no teste de Mann-Whitney

Base	Proteínas excluídas
Bacillus	L4, L7, L7a, L7ae, L9, L12, L15, L19, L20, S1, S10, S22, S31e
Rhizobium	L3, L4, L7a, L7ae, L12, L25, L27, L33, L35, S1, S7, S12, S22, S31e

Em seguida, o método *Relief* foi executado sobre os resultados da primeira etapa (URBANOWICZ *et al.*, 2018). Isto permitiu organizar um *rank* dos atributos de cada base em termos de relevância quanto a predição da classe alvo. Na sequência, o procedimento iterativo descrito no Algoritmo 2 foi executado e os atributos com menor índice *Relief* foram removidos iterativamente. O modelo f utilizado para testar a remoção dos atributos foi a Regressão Logística.

Algoritmo 2: SELEÇÃO DE ATRIBUTOS COM RELIEF

Entrada: $\mathbf{D}_1$, uma base de dados
v, um vetor com os atributos ordenados de acordo com o rank *Relief*
f, um classificador
Saída: $\mathbf{D}_1$

```
1  início
2      treine o classificador f em D_1
3      calcule a acurácia balanceada, acbal_0, de f em D_1
4      defina n como o número de atributos de D_1
5      para i de n até 1 faça
6          copie o conjunto de dados D_1 para D_2
7          remova o atributo v[i] do conjunto D_2 treine o classificador f em D_2
8          calcule a acurácia balanceada, acbal, de f em D_2
9          se acbal ≥ acbal_0 então
10             atribua D_2 a D_1
11         senão
12             pare
13         fim
14     fim
15 fim
16 retorna D_1
```

Ao final da execução do Algoritmo 2, as seguintes proteínas foram removidas:

Tabela 7: Proteínas excluídas com base no teste com o *rank Relief*

Base	Proteínas excluídas
Bacillus	L2, L30, L35, L7/L12, L10, L14, L16, L18, L27, S6, S20
Rhizobium	L2, L7, L7/L12, L10, L24, L29, L31, L32, L34, S4, S8, S10, S11, S13, S14, S19

Por fim, concluindo o processo de pré-processamento dos dados, as seguintes proteínas foram escolhidas para compor as respectivas bases:

- Base *Bacillus* : L1, L3, L5, L6, L11, L13, L17, L21, L22, L23, L24, L25, L28, L29, L31, L32, L33, L34, L36, S2, S3, S4, S5, S7, S8, S9, S11, S12, S13, S14, S15, S16, S17, S18, S19, S21;
- Base *Rhizobium*: L1, L5, L6, L9, L11, L13, L14, L15, L16, L17, L18, L19, L20, L21, L22, L23, L28, L30, L36, S2, S3, S5, S6, S9, S15, S16, S17, S18, S20, S21;

3.3 PROTOCOLO EXPERIMENTAL

Os experimentos descritos nesta seção foram executados sobre coleções de bases de dados sintetizadas a partir dos conjuntos de dados *Bacillus*, *Rhizobium*. As bases de dados sintéticas foram geradas usando um procedimento de síntese de dados baseado em regressão do tipo CART (do inglês, *Classification And Regression Tree*) (REITER, 2005; DRECHSLER; REITER, 2011). O Algoritmo 3 descreve o procedimento de geração de dados sintéticos. Na sua realização foi utilizada a função *syn.cart* do pacote *synthpop* da linguagem R.

Algoritmo 3: GERADOR DE BASES DE DADOS SINTETICAS

Entrada: $\mathbf{D}$, uma base de dados
b, número de bases sintéticas desejado
r, um vetor com os valores de níveis de balanceamento desejados
m, número de amostras a compor as bases sintéticas
Saída: $b * r$ bases de dados sintéticas

início
  Seja $\mathbf{D}_v \subset \mathbf{D}$, o subconjunto de casos rotulados como *Verdadeiro*
  Seja $\mathbf{D}_f \subset \mathbf{D}$, o subconjunto de casos rotulados como *Falso*
  para $i = 1 : b$ **faça**
    para *cada* $j = 1 : |r|$ **faça**
      (a) computar o número de amostras sintéticas de cada classe:
        $m_v = \lceil m * r[j]/100 \rceil$
        $m_f = m - m_v$
      (b) gerar amostras sintéticas a partir de $\mathbf{D}_v$ e $\mathbf{D}_f$:
        $\mathbf{D}_{v_s} = syn.cart(\mathbf{D}_v, m_v)$
        $\mathbf{D}_{f_s} = syn.cart(\mathbf{D}_f, m_f)$
      (c) construir da base sintética dicotômica:
        $\mathbf{D}_s = \mathbf{D}_{v_s} \cup \mathbf{D}_{f_s}$
      (d) armazenar $\mathbf{D}_s^i$
    fim
  fim
fim

Os níveis de balanceamento escolhidos para compor as bases sintéticas utilizadas nos testes estão no intervalo de 0.5% a 15%. Assumidos estes limites, foram geradas bases com os seguintes níveis de balanceamento: 0.5%, 1.0%, 1.5%, 2.0%, 2.5%, 3%, 3.5%, . . . , 15%. Para cada proporção indicada, foram geradas 20 bases contendo 5000 registros cada. Logo, para cada taxonomia de interesse foram produzidas 600 bases sintéticas. Ao final desta etapa as bases sintéticas foram identificadas como $\mathbf{B}_j^t$ e $\mathbf{R}_j^t$, onde t representa o nível de balanceamento (%) e j representa o número da base. As letras $\mathbf{B}$ e $\mathbf{R}$ indicam que as bases estão associadas ao problema de reconhecimento de instâncias do gênero *Bacillus* e *Rhizobium* respectivamente.

Na próxima etapa, as bases foram submetidas ao procedimento descrito no Algoritmo 4 cujo objetivo era o treinamento e a computação dos escores de desempenho dos classificadores. Para cada base, foram calculadas a acurácia, sensibilidade, especificidade, acurácia balanceada e *overfit*. Na validação dos modelos, foi aplicado o método Validação Cruzada Estratificada, o

qual gera uma distribuição uniforme dos registros, certificando que cada subconjunto, receba ao menos uma instância de cada classe. Segundo Krstajic *et al.* (2014) este método é indicado para a validação de classificadores em bases desbalanceadas, pois garante a presença da classe minoritária tanto na base treino ($\mathbf{T}_{treino}$) como na base teste ($\mathbf{T}_{teste}$).

Algoritmo 4: VALIDAÇÃO DOS MODELOS CLASSIFICADORES

Entrada: $\mathbf{D}$, uma base de dados
k, tamanho da validação cruzada
f um classificador
Saída: Resultados dos escores médios de desempenho

1 **início**
2 $\mathbf{kFolds}$ = Particiona $\mathbf{D}$ em k partes;
3 Cria uma tabela $\mathbf{arq.csv}$ de resultados r_{ac}, r_{acbal}, r_{sens}, r_{esp} e $r_{overfit}$
4 **para** *cada* $\mathbf{T_k}$ *de* $\mathbf{kFolds}$ **faça**
5 (a) $\mathbf{T_k}$ é utilizada como base de teste ($\mathbf{T}_{teste}$) e os demais registros como base de treinamento ($\mathbf{T}_{treino}$);
6 (b) Treinar f com $\mathbf{T}_{treino}$;
7 (c) realizar a predição e calcular os escores de desempenho de $\mathbf{T_k}$ e armazenar os resultados na tabela $\mathbf{arq.csv}$.
8 **fim**
9 computa as médias aritméticas dos escores salvos em $\mathbf{arq.csv}$
10 **fim**
11 **retorna** *uma lista contendo os resultados médios da execução*

Em seguida as bases sintéticas foram submetidas a um procedimento que determinava suas medidas de complexidade, Algoritmo 5. Para cada uma das bases $\mathbf{B}_j^t$ e $\mathbf{R}_j^t$, foram computados os resultados para os índices F1, F2, F3, L1, L2, N1, N2, N3, T1, *ClsCoef*, *Density*, L2B e N3B. A acurácia dos classificadores não foi integrada aos resultados. Isto porque, segundo Bekkar, Djemaa e Alitouche (2013), em bases desbalanceadas, a acurácia pode ser enviesada pela taxa de acertos do modelo na classificação de instâncias da classe majoritária.

Algoritmo 5: EXECUÇÃO DOS ESCORES DE COMPLEXIDADE

Entrada: $\mathbf{D}$, uma base de dados
$\mathbf{C}$, uma lista de medidas de complexidade
Saída: Resultados das medidas de complexidade

```
início
    Cria uma tabela complex.csv de resultados cada medida em C
    para cada caso i de C faça
        (a) compute o score i na base D;
        (b) armazenar o resultado em uma lista.
    fim
    armazena a lista de resultados dos scores em complex.csv
fim
retorna Retorna uma lista contendo os resultados dos escores de complexidade
```

3.4 ANÁLISE DOS RESULTADOS

Os resultados da execução dos experimentos foram utilizados para avaliar a relação do desbalanceamento de classes com o desempenho dos classificadores e as métricas de complexidade de dados. Inicialmente, o teste de independência de Hoeffding, com $\alpha = 0,05$ e a análise de correlação de Pearson foram empregados para detectar dependências entre o desempenho dos classificadores e o desbalanceamento de classes (HOEFFDING, 1948; SEBER; LEE, 2012). Estes testes também foram utilizados para verificar a existência de dependência entre o desbalanceamento de classes e as medidas de complexidade. As medidas de desempenho examinadas foram a acurácia balanceada, a sensibilidade e o sobreajuste.

A força da dependência linear, medida pela correlação de Pearson (R), foi categorizada de acordo com a metodologia proposta por Benesty *et al.* (2009) e Mukaka (2012). Segundo aqueles autores, o valor absoluto da correlação linear pode ser categorizada como segue:

(a) = 0 - sem relação linear;

(b) > 0,30 - relação fraca;

(c) > 0,50 - relação moderada;

(d) > 0,70 - relação forte;

(e) > 0,90 - relação muito forte;

(f) = 1 - relação linear perfeita.

A análise de variância (ANOVA), com $\alpha = 0,05$, foi empregado para testar se o desempenho dos classificadores e os índices de complexidade variam conforme diferentes condições no desbalanceamento de classes (KIM, 2014). Este teste é indicado para situações onde existem mais de dois grupos a serem comparados. Os dados utilizados neste teste foram agrupados em dez faixas, conforme o nível de balanceamento das bases. Os níveis de balanceamento que compõe as bases utilizadas nos testes consistem no intervalo 0.5% a 15%, com variação de 0.5% totalizando trinta níveis de balanceamento.

Quando o teste ANOVA identificou uma diferença estatística entre os grupos, o teste t de Student foi utilizado para verificar em quais faixas tais diferenças estatísticas ocorriam. Assim, o teste t de Student, pareado e unicaudal, com $\alpha = 0,05$, foi aplicado para verificar como a variação na condição de balanceamento afetou o desempenho dos classificadores (em termos de acurácia balanceada e sensibilidade) e os índices de complexidade a cada faixa. O mesmo teste foi aplicado para identificar qual classificador obteve melhor desempenho em cada caso.

3.4.1 Análise da importância dos índices de complexidade

A última etapa da análise dos resultados teve o objetivo de ordenar os índices de complexidade de dados e a taxa de desbalanceamento em termos de sua influência sobre o desempenho dos classificadores. O primeiro passo dessa etapa trata da identificação dos índices de complexidade que não satisfaziam um critério de relevância em termos de dependência linear, C_0, definido como:

> Critério de relevância em termos de dependência linear (C_0): Uma medida de complexidade $\mathcal{M}$ de dados é relevante para a predição do desempenho dos modelos classificadores quando atende as duas condições abaixo:
>
> a) os dados sugerem que há uma associação entre $\mathcal{M}$ e o desempenho do classificador em termos de acurácia balanceada ou sensibilidade;
>
> b) um aumento da complexidade de dados, conforme quantificada por $\mathcal{M}$, está associado a um decremento da acurácia balanceada (queda do desempenho do classificador).

Neste trabalho, o grau de associação entre uma medida de complexidade e o desempenho de um classificador seria estimado pela correlação linear (logarítmica ou exponencial) entre

a referida medida e a acurácia balanceada. Além disso, foi definido que um atributo atendia ao primeiro requisito do critério de relevância quando a correlação era moderada, forte ou muito forte.

Seja $\mathbf{V}_0$ um conjunto cujos elementos simbolizam os índices de complexidade que atenderam ao critério C_0 e taxa de desbalanceamento. No segundo passo da análise de importância, o procedimento de seleção de atributos descrito por Kuhn *et al.* (2008) e Kuhn (2015) foi empregado para remover os elementos de $\mathbf{V}_0$ que eram altamente correlacionados. No terceiro passo, o escore denominado importância da variável, apresentado por Gevrey, Dimopoulos e Lek (2003) e Hapfelmeier *et al.* (2014) foi utilizado para ranquear os índices de complexidade de dados em termos de sua capacidade em predizer o desempenho dos classificadores.

4 RESULTADOS E DISCUSSÃO

4.1 DESBALANCEAMENTO DE DADOS E O DESEMPENHO DE CLASSIFICADORES

Os resultados dos experimentos que avaliam a relação entre o nível de balanceamento da base e o desempenho dos classificadores estão em acordo com a literatura. A revisão da bibliografia da área sugere que o desbalanceamento de classes é um dos fatores que podem fazer com que um algoritmo de aprendizagem produza um classificador com baixa capacidade preditiva no que se refere a classe minoritária (DELYON, 1996; ELRAHMAN; ABRAHAM, 2013; LU; CHEUNG; TANG, 2019). Nos experimentos realizados, este fenômeno foi observado no que diz respeito à acurácia balanceada dos classificadores treinados para a identificação de amostras de ambos os gêneros, *Bacillus* e *Rhizobium*.

A Tabela 8 lista o resultado dos testes estatísticos de Pearson e Hoeffding para medir, respectivamente, a correlação linear e a dependência probabilística entre: (a) o nível de balanceamento de classes e a acurácia balanceada, e; (b) o nível de balanceamento de classes e a sensibilidade do classificador. A notação exp índica que a medida de desempenho foi submetida a uma transformação exponencial antes do cômputo da correlação.

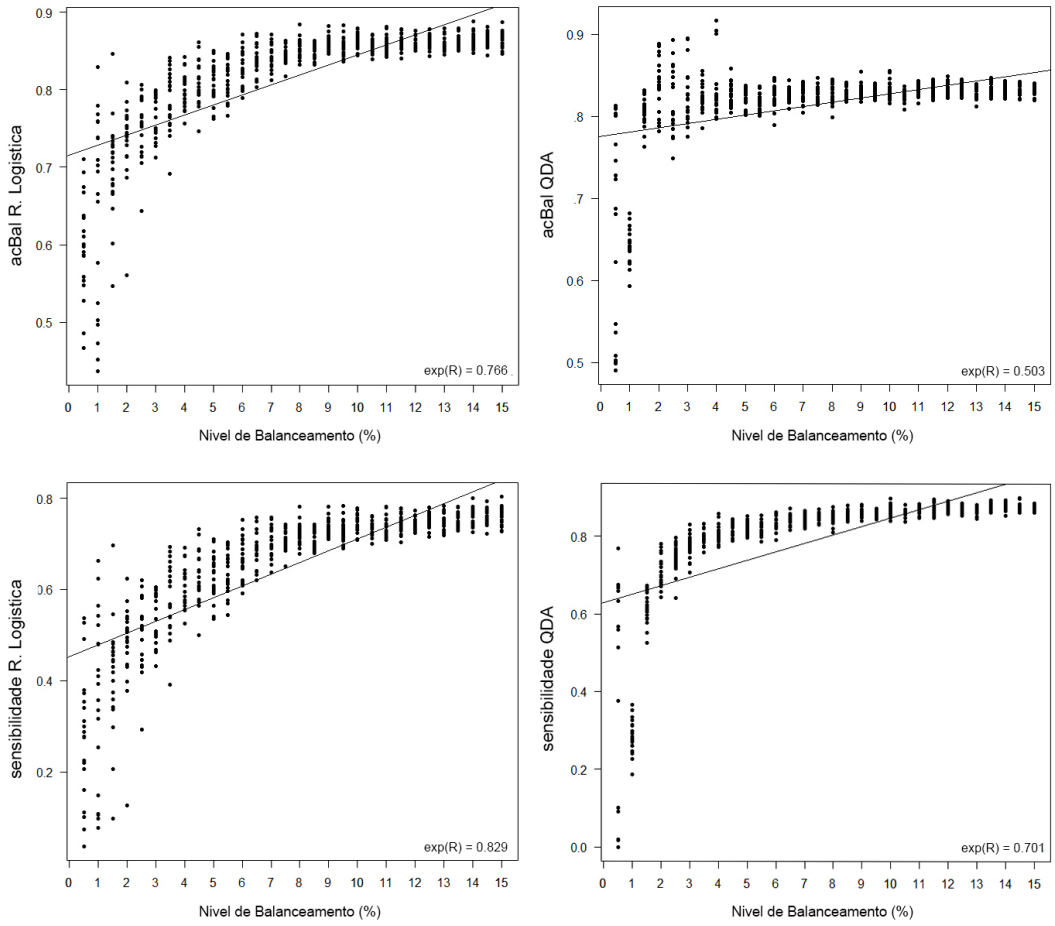
Tabela 8: Valores da correlação de Pearson e independência de Hoeffding entre desempenho dos classificadores e o nível de balanceamento dos dados

Classificador	base de dados	Acurácia Balanceada		Sensibilidade	
		Pearson	Hoeffding	Pearson	Hoeffding
QDA	*Bacillus*	0.503 (exp)	$< 1e^{-8}$	0.701 (exp)	$< 1e^{-8}$
	Rhizobium	0.731 (exp)	$< 1e^{-8}$	0.755 (exp)	$< 1e^{-8}$
Reg. Logística	*Bacillus*	0.766 (exp)	$< 1e^{-8}$	0.829 (exp)	$< 1e^{-8}$
	Rhizobium	0.767 (exp)	$< 1e^{-8}$	0.809 (exp)	$< 1e^{-8}$

Inicialmente, deve ser notado que os resultados do teste de hipóteses (Hoefdding) indicam a existência de dependência entre o o nível de balanceamento e ambas as medidas de desempenho avaliadas. O coeficiente de Pearson aponta para uma correlação linear forte entre o nível de balanceamento e a sensibilidade dos classificadores em ambas as bases. No que diz respeito à acurácia balanceada, o desempenho da regressão logística teve uma correlação forte com o nível de balanceamento das bases de dados dos gêneros *Bacillus* e *Rhizobium*. Por outro lado, os valores de *acbal* registrados para o classificador QDA apresentaram uma correlação moderada com o desbalanceamento de classes quando do processamento das bases de dados relativas à detecção da classe *Bacillus*.

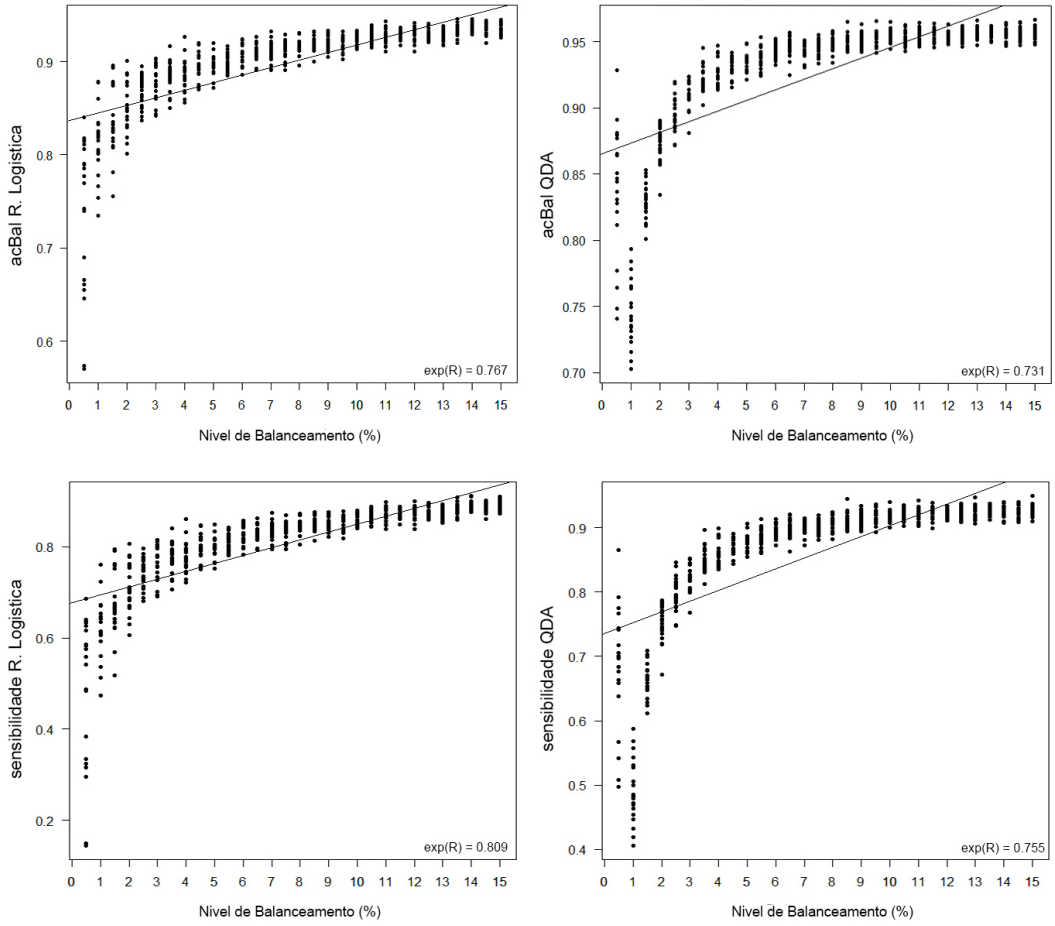
As Figuras 4 e 5 evidenciam a dependência entre o nível de balanceamento dos dados e o desempenho dos classificadores para as bases *Bacillus* e *Rhizobium*, respectivamente. No conjunto de bases de dados referentes ao gênero *Bacillus*, a acurácia balanceada do classificador logístico ficou entre [43,73 ; 88,78 %] e a do QDA entre [49,09 ; 91,67 %]. Para a base *Rhizobium*, a acurácia balanceada do modelo logístico resultou em valores entre [57,07 , 94,54 %], já o modelo QDA atingiu resultados no intervalo [70,31 , 96,69 %].

Figura 4: Relação entre o nível de balanceamento e o desempenho dos classificadores para as bases *Bacillus*



Fonte: O Autor

Figura 5: Relação entre o nível de balanceamento e o desempenho dos classificadores para as bases *Rhizobium*



Fonte: O Autor

4.1.1 Comparação do desempenho dos classificadores e análise de variância

A Tabela 9 organiza os resultados obtidos nos testes de classificação para o gênero *Bacillus* em grupos definidos por intervalos do nível de balanceamento. Naquela tabela, a primeira coluna informa a proporção do número de casos da classe minoritária em relação ao tamanho da base. As demais colunas, informam a média e o desvio padrão da acurácia balanceada, sensibilidade e *overfit* obtidos pelos classificadores treinados durante os experimentos.

Tabela 9: Valores médios de desempenho dos classificadores para a base *Bacillus*

	Regressão Logística						QDA					
Bal (%)	*acbal*		sensibilidade		*overfit*		*acbal*		sensibilidade		*overfit*	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
[0,5-1,5]	0,648	0,097	0,352	0,159	0,997	0,001	0,701	0,106	0,428	0,226	0,996	0,002
[2-3]	0,749	0,041	0,505	0,083	0,997	0,001	0,828	0,035	0,741	0,045	0,995	0,001
[3,5-4,5]	0,798	0,031	0,605	0,062	0,997	0,001	0,826	0,023	0,807	0,022	0,995	0,001
[5-6]	0,816	0,025	0,644	0,050	0,997	0,001	0,821	0,012	0,826	0,018	0,995	0,001
[6,5-7,5]	0,843	0,016	0,701	0,032	0,997	0,001	0,827	0,009	0,846	0,013	0,995	0,001
[8-9]	0,854	0,011	0,725	0,022	0,996	0,001	0,828	0,008	0,854	0,013	0,996	0,001
[9,5-10,5]	0,857	0,011	0,734	0,021	0,996	0,001	0,829	0,008	0,863	0,010	0,996	0,001
[11-12]	0,861	0,009	0,744	0,018	0,996	0,001	0,832	0,007	0,870	0,011	0,996	0,001
[12,5-13,5]	0,862	0,009	0,752	0,017	0,997	0,001	0,832	0,007	0,872	0,010	0,996	0,001
[14-15]	0,864	0,009	0,758	0,017	0,996	0,001	0,832	0,006	0,875	0,008	0,996	0,001

Como pode ser notado, o classificador QDA alcançou um desempenho superior à regressão logística sempre que o nível de balanceamento era inferior a seis por cento. Esta observação foi corroborada pelo teste t de Student ($\alpha = 0.05$) pareado para a diferença entre as médias da acurácia balanceada do classificador logístico e do classificador QDA em cada uma das quatro primeiras linhas da tabela. Para as bases com nível de balanceamento superior a 6, o desempenho do modelo logístico foi superior. Nos referidos intervalos os níveis de *overfit* foram 0,996 em média para ambos os modelos.

Na base *Bacillus*, o teste ANOVA detectou que houve uma diferença estatística nas médias de *acbal* e de sensibilidade entre os grupos definidos pelas faixas de balanceamento (ou linhas) da Tabela 9. O teste t de Student foi usado para identificar em quais faixas ocorrem diferenças significativas na média da *acbal* e sensibilidade. O teste foi aplicado assumindo dados pareados e $\alpha = 0.05$. Dentre os resultados do classificador logístico, houve diferença significativa nas médias de sensibilidade em todos os níveis de balanceamento. Para a *acbal* houve diferença somente para as bases com até nove por cento de nível de balanceamento. Após este valor, o desempenho médio do classificador não apresentou alteração significativa. Para o modelo QDA, o incremento total em *acbal* e sensibilidade é identificado como significativo, neste caso, o teste aponta que as médias da acurácia balanceada não diferem consideravelmente quando o nível de balanceamento das bases é superior a três por cento.

A Tabela 10 lista os resultados obtidos nos testes de classificação para o gênero *Rhizobium*. O desempenho dos classificadores em termos de acurácia balanceada foi comparado via teste t de Student ($\alpha = 0.05$). Considerando os resultados médios do desempenho dos classificadores, em todos os grupos de balanceamento é constatado que o classificador QDA

teve desempenho superior ao classificador logístico. Nos referidos intervalos, não foi detectado *overfit*, que foi 0,996 em média para ambos os modelos.

Tabela 10: Valores médios de desempenho dos classificadores para a base *Rhizobium*

	Regressão Logística						QDA					
Bal (%)	*acbal*		sensibilidade		*overfit*		*acbal*		sensibilidade		*overfit*	
	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP	Média	DP
[0,5-1,5]	0,795	0,069	0,594	0,159	0,997	0,001	0,803	0,106	0,612	0,226	0,996	0,002
[2-3]	0,865	0,024	0,738	0,083	0,996	0,001	0,893	0,035	0,791	0,045	0,996	0,001
[3,5-4,5]	0,887	0,016	0,783	0,062	0,996	0,001	0,926	0,023	0,859	0,022	0,996	0,001
[5-6]	0,900	0,011	0,810	0,050	0,997	0,001	0,938	0,012	0,884	0,018	0,996	0,001
[6,5-7,5]	0,911	0,010	0,833	0,032	0,996	0,001	0,945	0,009	0,900	0,013	0,995	0,001
[8-9]	0,918	0,007	0,849	0,022	0,997	0,001	0,950	0,008	0,911	0,013	0,995	0,001
[9,5-10,5]	0,922	0,008	0,858	0,021	0,997	0,001	0,953	0,008	0,917	0,010	0,995	0,001
[11-12]	0,928	0,007	0,871	0,018	0,997	0,001	0,955	0,007	0,920	0,011	0,996	0,001
[12,5-13,5]	0,930	0,006	0,877	0,017	0,997	0,001	0,955	0,007	0,923	0,010	0,995	0,001
[14-15]	0,935	0,005	0,889	0,017	0,996	0,001	0,956	0,006	0,925	0,008	0,996	0,001

A variância observada no desempenho dos classificadores para a base *Rhizobium*, é avaliada pelo teste ANOVA. Ao comparar todas as faixas de balanceamento, o mesmo confirma que as métricas de desempenho acurácia balanceada e sensibilidade são afetadas pelo desbalanceamento dos dados. Para avaliar a variância das medias dos resultados de forma pareada, foi aplicado o teste t, com $\alpha = 0.05$. Este teste identificou que o incremento no desempenho do modelo logístico em termos de *acbal* e sensibilidade é significativo. Para o modelo QDA, o incremento total no desempenho também é identificado como significativo. Porém, o teste aponta que as médias da acurácia balanceada e sensibilidade deste classificador não diferem quando o nível de balanceamento das bases é superior à doze por cento.

4.1.2 Especificidade dos classificadores

Analisando os dados da Tabela 11 pode-se sugerir que a especificidade não possui uma relação forte com o desempenho dos classificadores. Em alguns casos obteve-se uma correlação fraca e em outros moderada. Em vista disto, esta medida foi descartada em analises posteriores.

Tabela 11: Valores da correlação de Pearson entre *acbal* e o balanceamento dos dados

Classificador	Base de dados	*acbal*
QDA	*Bacillus*	-0.451
	Rhizobium	-0.542
Reg. Logística	*Bacillus*	0.442
	Rhizobium	-0.687

4.2 RELAÇÃO ENTRE AS MEDIDAS DE COMPLEXIDADE, O DESEMPENHO DE CLASSIFICADORES E O DESBALANCEAMENTO DOS DADOS

4.2.1 Medidas de sobreposição de espaços

A Tabela 12 contém os resultados de correlação linear das medidas de sobreposição F1, F2 e F3 com o desempenho dos classificadores (colunas 3 a 6). A última coluna da tabela lista a correlação entre o balanceamento de dados e os índices de complexidade.

Tabela 12: Valores da correlação de Pearson entre as medidas F1, F2 e F3 e o desempenho dos classificadores e o nível de balanceamento dos dados

Medida de		Regressão Logística		QDA		Bal. (%)
complexidade	Base	*acbal*	sens.	*acbal*	sens.	
F1	*Bacillus*	-0,760	-0,804	-0,447	-0,660	-0,972
	Rhizobium	-0,846	-0,860	-0,832	-0,837	-0,950
F2	*Bacillus*	0,218	0,245	0,102	0,174	0,422
	Rhizobium	0,371	0,378	0,376	0,381	0,463
F3	*Bacillus*	**0,695**	**0,692**	**0,560**	**0,704**	**0,721**
	Rhizobium	**0,791**	**0,805**	**0,772**	**0,776**	**0,903**

Os resultados das duas primeiras linhas das colunas 3 a 6 indicam que as correlação entre a medida F1 e os escores de desempenho são fortes ou moderadas na maioria dos testes. Contudo, o fato destas correlações serem negativas indica que a redução da sobreposição dos dados, conforme medida por F1, está relacionada a uma queda no desempenho dos classificadores. Isto é, quanto mais separados os dados da classe (menor complexidade dos dados), menor o desempenho do classificador. Portanto, a medida F1 não atende ao critério de relevância C_0.

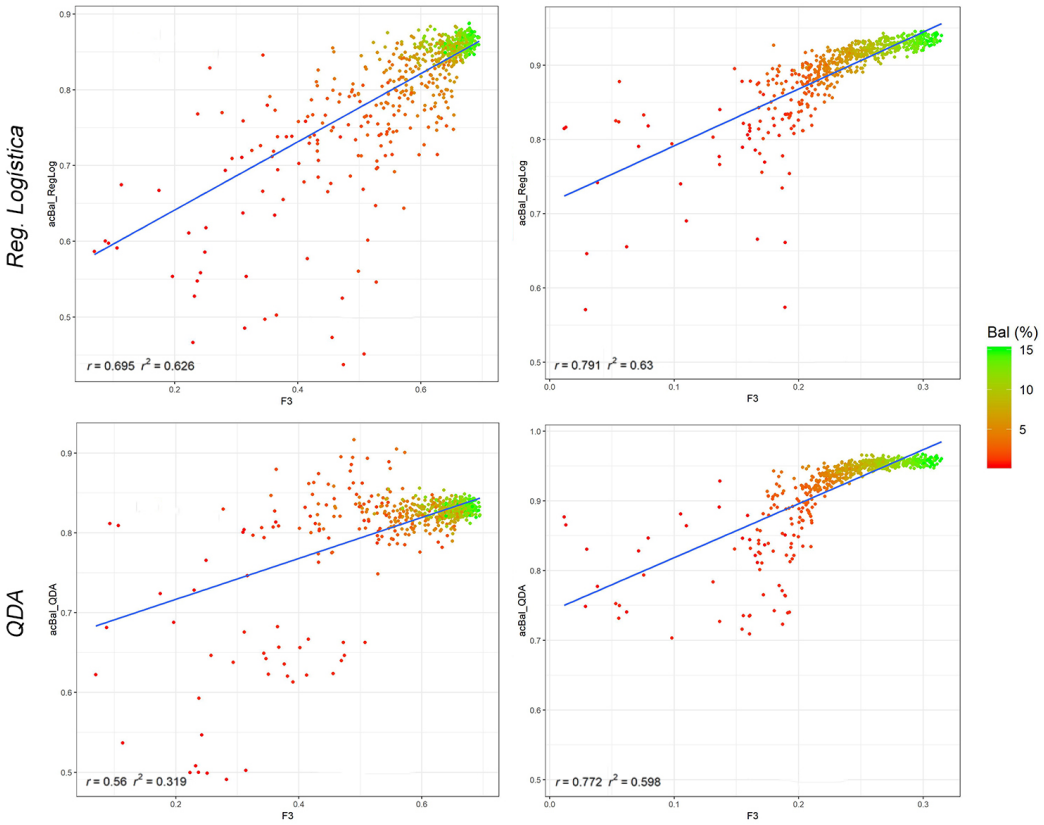
Os resultados relativos à medida F2 mostram uma correlação fraca entre aquele índice e os escores de desempenho dos classificadores. O mesmo acontece entre F2 e o nível de balanceamento. Assim, o índice de complexidade F2 também não atendeu ao C_0.

Os dados experimentais obtidos para a medida de complexidade a F3 atenderam ao Critério C_0 para os classificadores logístico e QDA tanto na identificação de instâncias nas bases de dados *Bacillus* quanto nas bases de dados *Rhizobium*. Ou seja, as correlações positivas listadas na quinta e sexta linhas da Tabela 12 mostram que, nos testes realizados, o crescimento de F3 (redução da sobreposição) foi associada a um aumento no desempenho dos classificadores.

A Figura 6 evidencia a relação entre o índice F3, o nível de balanceamento dos dados e o desempenho dos classificadores para as bases *Bacillus* (Figura 6a) e *Rhizobium* (Figura 6b).

Nesta figura, cada ponto representa o resultado obtido pelo classificador em uma base de dados e sua coloração indica o nível de balanceamento. A abcissa de cada gráfico se refere ao valor de F3 e a ordenada diz respeito à acurácia balanceada.

Figura 6: Resultado da relação entre o nível de balanceamento, o desempenho dos classificadores e o índice F3 para as bases *Bacillus* e *Rhizobium*

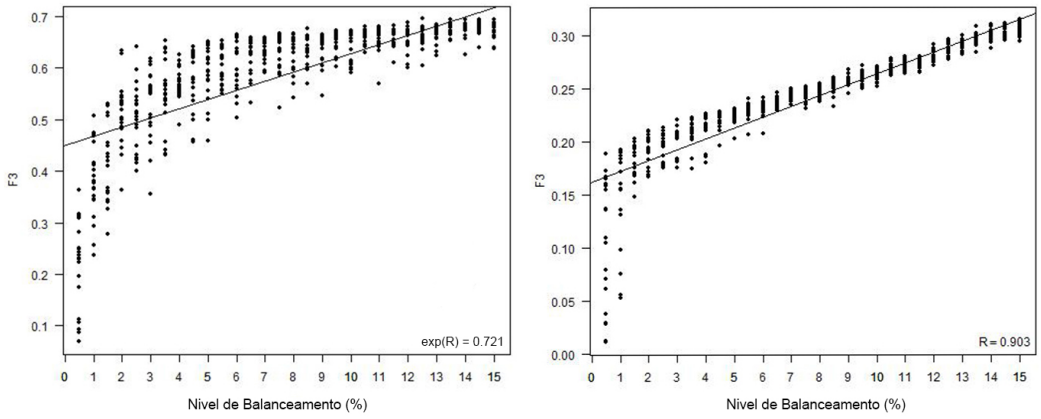


(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

O índice F3 também apresentou uma associação linear com o nível de balanceamento. Neste caso para ambos os gêneros foi detectada uma relação forte. Esta associação é destacada nos gráficos da Figura 7 para os gêneros *Bacillus* e *Rhizobium*.

Figura 7: Resultado da relação entre o nível de balanceamento e o índice F3 para as bases *Bacillus* e *Rhizobium*



(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

Quando aplicado o teste ANOVA, é identificada diferença entre os grupos de resultados de F3. O teste t de Student, com $\alpha = 0,05$, permite inferir um crescimento significativo na média de F3 quando se comparou conjuntos de dados com diferentes níveis de balanceamento. Isto é, segundo os resultados do teste, a média de F3 era maior em conjuntos de bases de dados em que o nível de balanceamento era maior. Este resultado foi obtido para ambas as taxonomias. Neste caso, o teste t revelou que o valor de F3 para a base *Bacillus* não diferiu estatisticamente quando o nível de balanceamento era superior a nove por cento. As médias dos dados utilizados na realização do teste estão listados na Tabela 13.

Tabela 13: Valores médios do índice F3 de acordo com o nível de balanceamento dos dados.

Nível de	*Bacillus (a)*		*Rhizobium (b)*	
Balanceamento (%)	Média	DP	Média	DP
[0,5-1,5]	0,3410	0,1170	0,1448	0,0546
[2-3]	0,5129	0,0648	0,1978	0,0133
[3,5-4,5]	0,5619	0,0573	0,2137	0,0115
[5-6]	0,5953	0,0443	0,2275	0,0076
[6,5-7,5]	0,6282	0,0301	0,2414	0,0059
[8-9]	0,6324	0,0307	0,2520	0,0071
[9,5-10,5]	0,6501	0,0234	0,2644	0,0064
[11-12]	0,6561	0,0224	0,2761	0,0066
[12,5-13,5]	0,6423	0,1128	0,2912	0,0065
[14-15]	0,6530	0,1027	0,3027	0,0058

4.2.2 Medidas de Separabilidade das classes

A Tabela 14 lista os valores dos testes da correlação entre as medidas L1, L2, N1, N2 e N3, o desempenho dos classificadores e o nível de balanceamento dos dados, que foram extraídas dos resultados dos experimentos.

Tabela 14: Valores da correlação de Pearson entre as medidas L1, L2, N1, N2 e N3 e o desempenho dos classificadores e o nível de balanceamento dos dados

		Reg. Logística		QDA		Bal. (%)
ID	Base	*acbal*	sens.	*acbal*	sens.	
L1	*Bacillus*	0,770	0,807	0,483	0,710	0,966
	Rhizobium	0,769	0,785	0,796	0,804	0,921
L2	*Bacillus*	0,673	0,717	0,394	0,581	0,918
	Rhizobium	0,733	0,752	0,760	0,769	0,952
N1	*Bacillus*	0,783	0,815	0,505	0,724	0,936
	Rhizobium	0,737	0,752	0,751	0,757	0,858
N2	*Bacillus*	0,725	0,741	0,541	0,729	0,819
	Rhizobium	0,719	0,732	0,732	0,737	0,826
N3	*Bacillus*	0,721	0,745	0,478	0,677	0,822
	Rhizobium	0,679	0,694	0,694	0,701	0,801

No caso da identificação de bactérias do gênero *Rhizobium*, o procedimento experimental mostrou que apesar do desempenho dos classificadores ter apresentado uma correlação forte ou moderada com todos os índices de separabilidade a interpretação das correlações mostra que, segundo estas medidas, o incremento da complexidade de dados está associada a uma melhora no desempenho dos classificadores. Logo tais medidas não atenderam ao critério C_0. O mesmo relacionamento foi observado na maioria dos testes realizados com as bases de dados do gênero *Bacillus*.

4.2.3 Medidas de Topologia, Geometria e Densidade

A Tabela 15 apresenta os resultados de correlação linear das medidas T1, *Density* e *ClsCoef* com o desempenho dos classificadores e o nível de balanceamento dos dados.

Tabela 15: Valores da correlação de Pearson entre as medidas T1, *ClsCoef* e *Density* e o desempenho dos classificadores e o nível de balanceamento dos dados

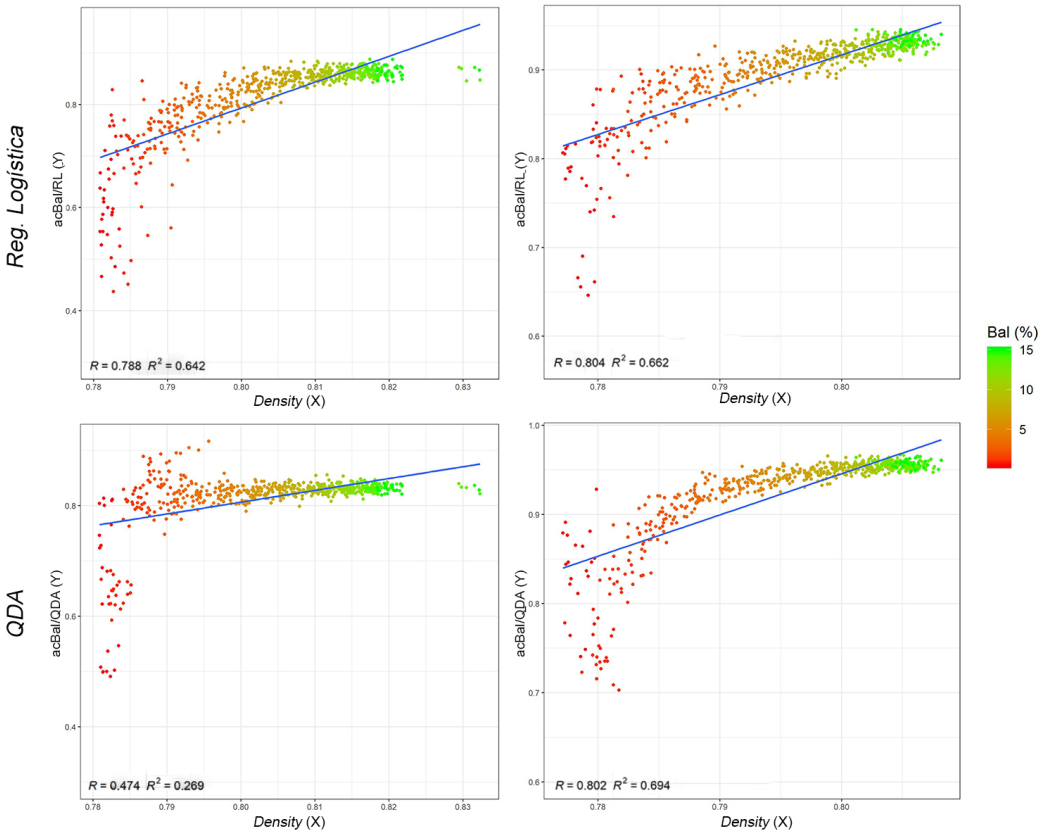
		Reg. Logística		QDA		Bal. (%)
ID	Base	*acbal*	sens.	*acbal*	sens.	
T1	*Bacillus*	0,785	0,814	0,515	0,738	0,915
	Rhizobium	0,793	0,806	0,813	0,817	0,866
ClsCoef	*Bacillus*	-0,308	-0,357	-0,131	-0,235	-0,683
	Rhizobium	-0,632	-0,652	-0,590	-0,598	-0,915
Density	*Bacillus*	**0,788**	**0,830**	(exp) **0,544**	**0,693**	**0,980**
	Rhizobium	**0,804**	**0,820**	(exp) **0,802**	**0,809**	**0,965**

Os resultados relativos à medida de complexidade T1 mostram que a mesma não atendeu à C_0 pois, as correlações obtidas (moderadas a fortes) mostram que a acurácia balanceada cresceu com a complexidade dos dados. O índice *ClsCoef* também não atendeu à C_0. No caso da identificação do gênero *Bacillus* todas as correlações sinalizam que a dependência entre a *acbal* e *ClsCoef* foi fraca ou não detectável. Por outro lado, nos testes realizados com dados do gênero *Rhizobium* detectou-se uma correlação negativa e moderada entre o desempenho dos classificadores e o índice de complexidade. No entanto, aquelas correlações apontam que há um crescimento da *acbal* dos classificadores em bases de dados que são mais complexas conforme *ClsCoef*.

Por sua vez, os resultados obtidos para a medida de complexidade *Density* atenderam ao Critério C_0, apresentando associação linear moderada a forte com os modelos testados. Assim, as correlações positivas listadas na quinta e sexta linhas da Tabela 15 mostram que, nos testes realizados, o crescimento de *Density* foi associado a um aumento no desempenho dos classificadores.

A Figura 8 destaca o relacionamento entre a medida *Density*, o nível de balanceamento dos dados e o desempenho dos classificadores para as bases *Bacillus* e *Rhizobium*, respectivamente. Mais uma vez, o balanceamento dos dados é indicado através de uma legenda colorida. Em termos de acurácia balanceada, o crescimento do valor de *Density* mostrou uma correlação forte com o desempenho do classificador logístico nas bases *Bacillus* e *Rhizobium* (Figura 8a e 8b). Para o classificador quadrático, o valor de correlação indicado, sugere relação exponencial moderada para a base *Bacillus* e forte para a base *Rhizobium*.

Figura 8: Resultado da relação entre o nível de balanceamento, o desempenho dos classificadores e o índice *Density* para as bases *Bacillus* e *Rhizobium*

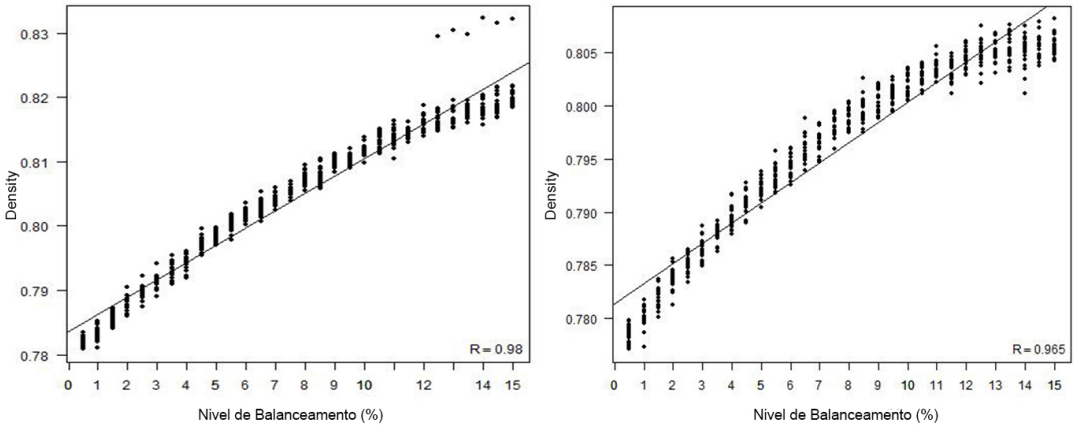


(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

Ao considerar a associação entre a medida *Density* e o nível de balanceamento para ambos os conjuntos de base de dados, foi identificada associação linear. Esta conclusão, segue dos valores registrados na última coluna da Tabela 15, onde indicam uma correlação muito forte entre *Density* e o nível de balanceamento (BENESTY *et al.*, 2009). Esta relação pode ser visualizada nos gráficos da Figura 9.

Figura 9: Resultado da relação entre o nível de balanceamento e o índice *Density* para as bases *Bacillus* e *Rhizobium*



(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

A variação dos valores de *Density* é apresentada na Tabela 16. Considerando um $\alpha = 0,05$ o teste ANOVA identifica diferenças significativas entre os resultados deste índice de acordo com as faixas de balanceamento. Isto para ambas as taxonomias. Considerando o mesmo α, o teste t de Student aplicado par a par nas linhas da tabela, permite inferir que o acréscimo do número de casos na classe minoritária levou a diferenças significativas no valor de *Density*. Neste caso, o teste aponta que as médias do valor de *Density* não diferem apenas na base *Rhizobium*, quando o nível de balanceamento é superior a doze por cento.

Tabela 16: Valores médios do índice *Density* de acordo com o nível de balanceamento dos dados.

Nível de	*Bacillus (a)*		*Rhizobium (b)*	
Balanceamento (%)	Média	DP	Média	DP
[0,5-1,5]	0,7835	0,0019	0,7802	0,0017
[2-3]	0,7897	0,0018	0,7852	0,0016
[3,5-4,5]	0,7949	0,0021	0,7895	0,0015
[5-6]	0,8001	0,0016	0,7934	0,0013
[6,5-7,5]	0,8042	0,0015	0,7968	0,0014
[8-9]	0,8084	0,0015	0,7995	0,0012
[9,5-10,5]	0,8117	0,0013	0,8018	0,0012
[11-12]	0,8147	0,0015	0,8039	0,0011
[12,5-13,5]	0,8177	0,0030	0,8052	0,0011
[14-15]	0,8218	0,0032	0,8055	0,0012

4.2.4 Medidas de separabilidade baseadas em acurácia balanceada

A Tabela 17 lista os dados de correlação entre os índices de complexidade L2B e N3B e o desempenho dos classificadores. Também são listadas as correlações entre estas medidas e o nível de balanceamento.

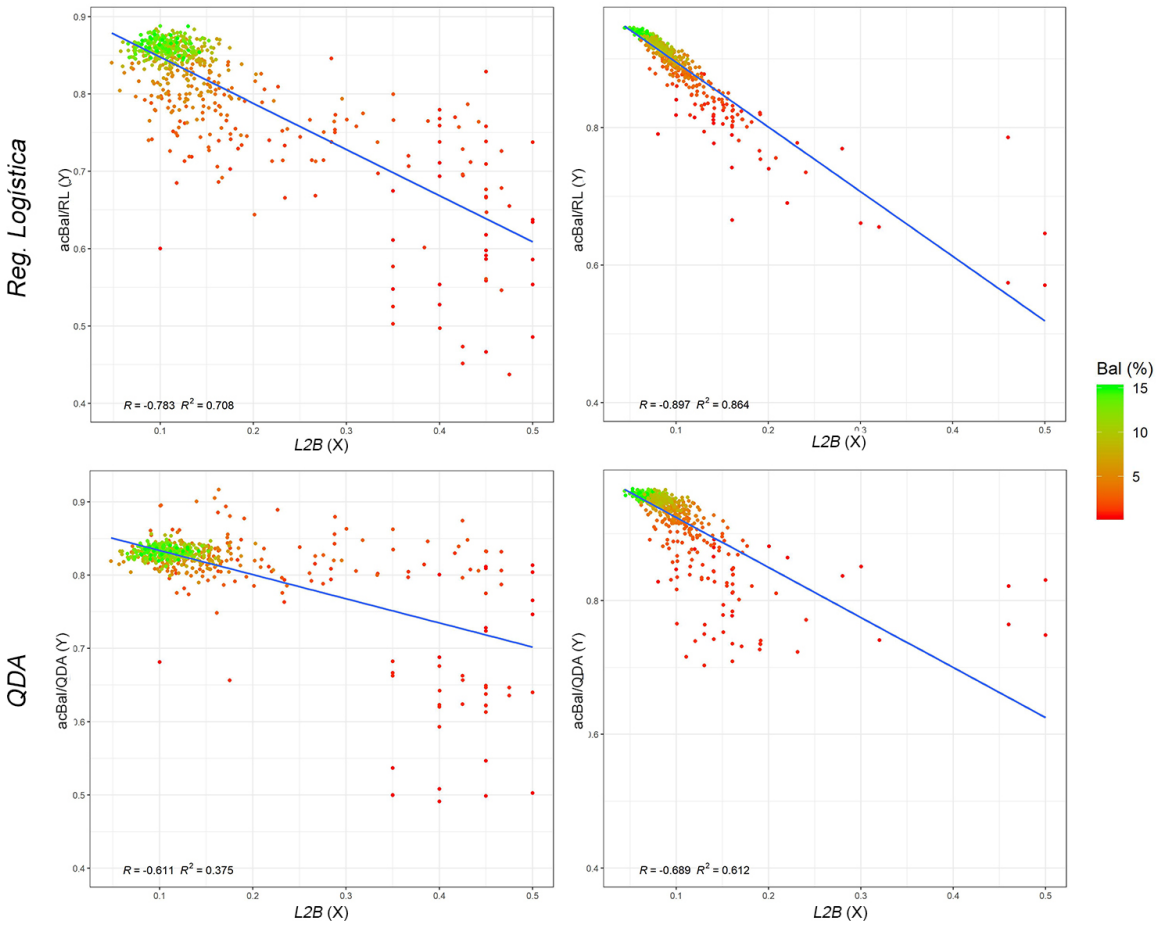
Tabela 17: Valores da correlação de Pearson entre as medidas L2B e N3B e o desempenho dos classificadores e o nível de balanceamento dos dados

		Reg. Logística		QDA		Bal. (%)
ID	Base	*acbal*	sens.	*acbal*	sens.	
L2B	*Bacillus*	**-0,783**	**-0,769**	**-0,611**	**-0,768**	**-0,648**
	Rhizobium	**-0,897**	**-0,895**	**-0,689**	**-0,703**	**-0,615**
N3B	*Bacillus*	**-0,833**	**-0,825**	**-0,642**	**-0,739**	**-0,655**
	Rhizobium	**-0,806**	**-0,805**	**-0,701**	**-0,70**	**-0,558**

As correlações negativas registradas mostram que a dependência linear entre a *acbal* e sensibilidade do classificador logístico e as medidas L2B e N3B pode ser categorizada como forte. No que diz respeito ao classificador QDA, as correlações entre *acbal* e sensibilidade e tais índices de complexidade sugerem uma associação linear entre moderada e forte. Estes resultados indicam que, a redução da complexidade de dados determinada pelos escores apresentados está associada a melhora no desempenho dos classificadores.

Logo, os escores L2B e N3B estão em acordo com o critério de relevância C_0. Este fato é destacado na Figura 10, que evidencia a relação entre a medida L2B, o nivel de balanceamento dos dados e o desempenho dos classificadores para as bases *Bacillus* (a) e *Rhizobium* (b). A Figura 11 evidencia os resultados da mesma relação para o escore N3B.

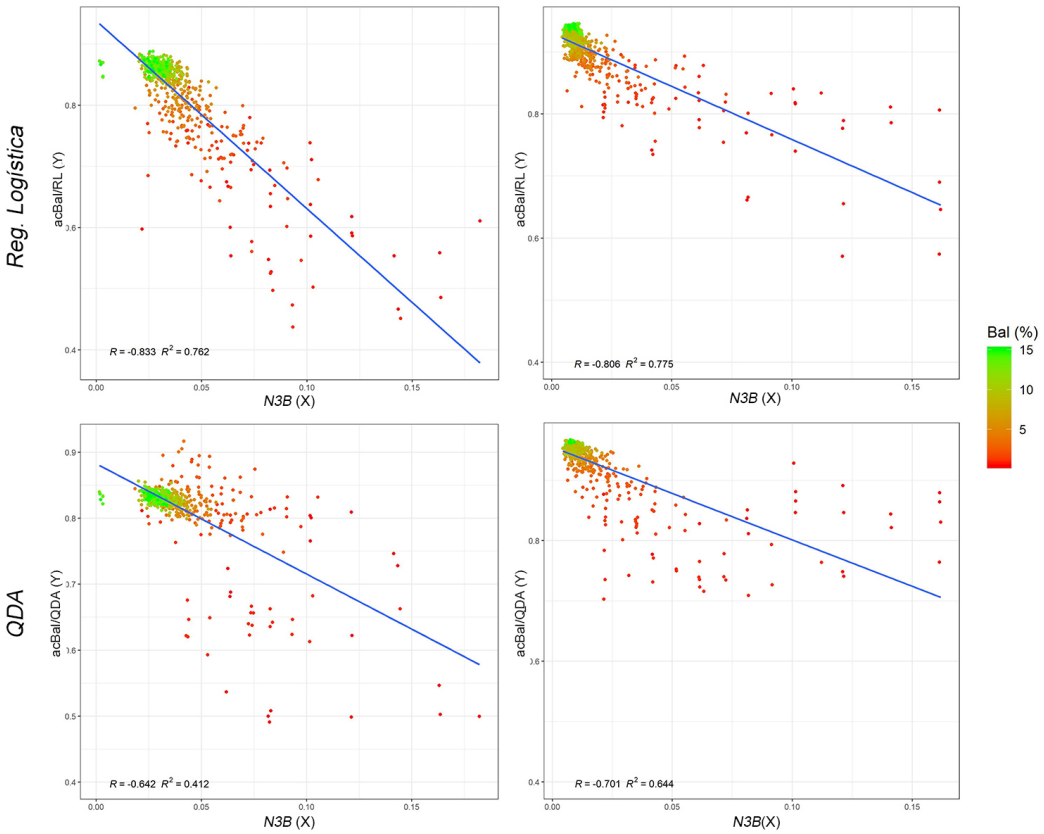
Figura 10: Resultado da relação entre nível de balanceamento, o desempenho dos classificadores e o índice L2B para as bases *Bacillus* e *Rhizobium*



(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

Figura 11: Resultado da relação entre nível de balanceamento, o desempenho dos classificadores e o índice N3B para as bases *Bacillus* e *Rhizobium*

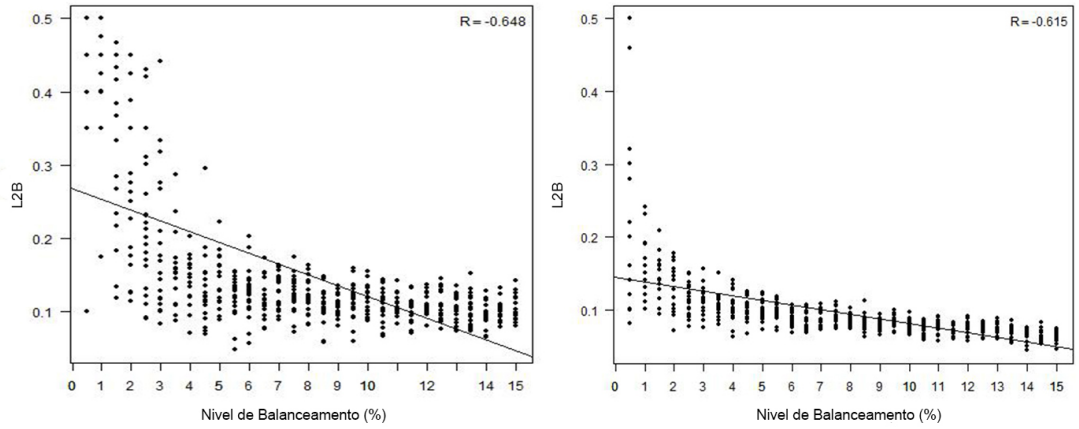


(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

Adicionalmente, ambas as medidas tiveram uma correlação moderada com o nível de balanceamento das bases. Este fenômeno pode ser visualizado nas Figuras 12 e 13.

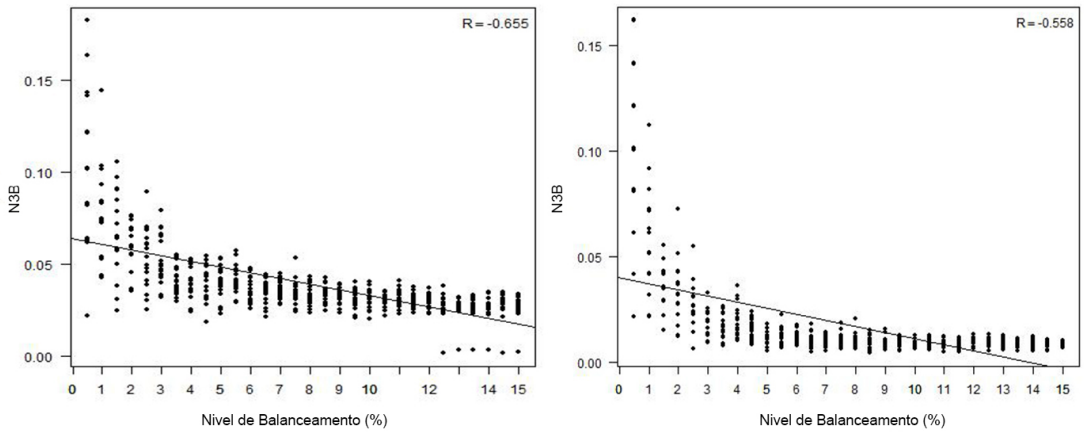
Figura 12: Resultado da relação entre nível de balanceamento e o índice L2B para as bases *Bacillus* e *Rhizobium*



(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

Figura 13: Resultado da relação entre nível de balanceamento e o índice N3B para as bases *Bacillus* e *Rhizobium*



(a) Base Bacillus *(b) Base Rhizobium*

Fonte: O Autor

A variação dos resultados dos índices L2B e N3B pode ser verificada nas Tabelas 18 e 19, respectivamente. Para ambos os casos, o teste ANOVA identificou diferenças estatísticas entre as faixas de resultados indicadas. Utilizando o nível de significância $\alpha = 0,05$, o teste estatístico t de Student detectou uma diferença significativa nos escores conforme os níveis de balanceamento. Neste caso, para os resultados da base *Rhizobium*, o teste t aponta que as médias de N3B não diferem consideravelmente quando o nível de balanceamento é superior a nove por cento. O mesmo é visto na base *Bacillus*, quando o nível de balanceamento é superior a doze por cento. Nos resultados de L2B com a base *Bacillus* as medias deixam de variar

estatisticamente quando o nível de balanceamento é superior a doze por cento. As diferenças observadas e os valores de correlação confirmam que o aumento de casos na classe minoritária implicou na redução do complemento da *acbal* dos índices N3B e L2B, resultando em redução na complexidade das bases *Bacillus* e *Rhizobium*.

Tabela 18: Valores médios do índice L2B de acordo com o nível de balanceamento dos dados.

Nível de	*Bacillus (a)*		*Rhizobium (b)*	
Balanceamento (%)	Média	DP	Média	DP
[0,5-1,5]	0,3859	0,1012	0,1796	0,0937
[2-3]	0,2207	0,1036	0,1175	0,0253
[3,5-4,5]	0,1444	0,0441	0,1041	0,0163
[5-6]	0,1284	0,0324	0,0949	0,0121
[6,5-7,5]	0,1226	0,0248	0,0876	0,0111
[8-9]	0,1106	0,0214	0,0824	0,0087
[9,5-10,5]	0,1116	0,0230	0,0789	0,0089
[11-12]	0,1065	0,0157	0,0738	0,0082
[12,5-13,5]	0,1030	0,0194	0,0710	0,0078
[14-15]	0,1010	0,0164	0,0643	0,0074

Tabela 19: Valores médios do índice N3B de acordo com o nível de balanceamento dos dados.

Nível de	*Bacillus (a)*		*Rhizobium (b)*	
Balanceamento (%)	Média	DP	Média	DP
[0,5-1,5]	0,0813	0,0347	0,0678	0,0417
[2-3]	0,0539	0,0138	0,0254	0,0125
[3,5-4,5]	0,0409	0,0080	0,0168	0,0058
[5-6]	0,0388	0,0076	0,0121	0,0035
[6,5-7,5]	0,0354	0,0059	0,0106	0,0028
[8-9]	0,0328	0,0051	0,0096	0,0029
[9,5-10,5]	0,0308	0,0047	0,0089	0,0018
[11-12]	0,0301	0,0040	0,0087	0,0020
[12,5-13,5]	0,0268	0,0064	0,0089	0,0018
[14-15]	0,0268	0,0065	0,0087	0,0017

4.3 DISCUSSÃO DOS RESULTADOS

Nos experimentos realizados neste trabalho o desempenho dos modelos classificadores foi próximo ao registrado por Santos *et al.* (2018) que utilizou classificadores dicotômicos e obteve uma acurácia superior a 99% e acurácia balanceada acima de 94% ao identificar instâncias do gênero *Bacillus*. Neste contexto, o trabalho de Tomachewski (2017) e Tomachewski et al. (2018) utiliza um modelo de classificação politômico e avalia apenas a acurácia, sendo

superior a 98% na tarefa de identificar gêneros de bactérias. No presente trabalho, a acurácia balanceada obtida ao identificar indivíduos do gênero *Bacillus*, foi em média de 86.4% pelo modelo logístico e 83.2% pelo quadrático (Tabela 9). No caso da identificação de indivíduos do gênero *Rhizobium*, a acurácia balanceada obtida foi em média 93.5% para o modelo logístico e 95.6% pelo quadrático (Tabela 10).

Como destacado na Seção 4.1, os resultados obtidos pelos classificadores na detecção dos gêneros de bactérias *Bacillus* e *Rhizobium* em bases desbalanceadas sugerem que há uma relação de dependência entre a acurácia balanceada dos classificadores e o desbalanceamento de dados. As correlações obtidas pelo ajuste exponencial indicam que tal dependência pode ser categorizada como moderada ou forte. Este resultado evidencia o fato de que o incremento da *acbal* decorreu, principalmente, do aumento da sensibilidade dos classificadores (ver Tabelas 8 e 11). Isto sugere que, como observado na literatura, por Anwar, Jones e Ganesh (2014) e Lorena *et al.* (2019), o desbalanceamento de classes é um dos fatores que afetam o desempenho dos classificadores. Em particular, deve ser notado que Reyes, Ochoa e Trinidad (2005) e Rosedahl e Ashby (2019) indicam que o desempenho de classificadores em bases com dados desbalanceados também é afetado por características das bases tais como sobreposição e separabilidade nos dados.

Sejam as medidas de complexidade que atenderam ao critério de relevância: F3, *Density*, L2B, N3B, e o nível de balanceamento da base de dados. A fim de avaliar a relevância destes fatores para estimar o desempenho dos classificadores QDA e Logístico, inicialmente, foram removidos aqueles índices que possuíam alta correlação entre si. Para tal, foi aplicado o procedimento de seleção de variáveis baseado em correlação, proposto por Kuhn *et al.* (2008), onde foram computadas as correlações entre pares de índices de complexidade. As Tabelas 20 e 21 listam a correlações encontradas. Desta forma, para a base *Bacillus*, foi removido o índice *Density* e na base *Rhizobium* foram removidos *Density* e F3.

Tabela 20: Correlações entre os índices de complexidade selecionados para a base *Bacillus*

	F3	N3B	L2B	*Density*
Bal	0.658	-0.655	-0.647	0.979
F3		-0.565	-0.696	-0.663
N3B			0.721	-0.705
L2B				-0.702

Tabela 21: Correlações entre os índices de complexidade selecionados para a base *Rhizobium*

	F3	N3B	L2B	*Density*
Bal	0.902	-0.558	-0.615	0.965
F3		-0.663	-0.662	0.912
N3B			0.752	-0.651
L2B				-0.660

Na sequência, um procedimento de regressão foi aplicado para estimar a relação linear entre o desempenho dos classificadores testados, medido em termos de *acbal*, e os valores de F3, L2B, N3B e nível de balanceamento da base de dados para *Bacillus* e L2B, N3B e nível de balanceamento para a base *Rhizobium*. Assim, sejam as acurácias balanceadas dos classificadores logístico e quadrático denotadas conforme a Tabela 22. O procedimento de regressão ajustou os seguintes modelos lineares listados nas Equações (1) a (4).

Tabela 22: Notação da variável "acurácia balanceada"na análise de regressão

Taxonomia	Classificador	Notação da *acbal*
Bacillus	R.L	$acbal_{B,L}$
	QDA	$acbal_{B,Q}$
Rhizobium	R.L	$acbal_{R,L}$
	QDA	$acbal_{R,Q}$

$$acbal_{B,L} = a_1 \cdot F3 + a_2 \cdot L2B + a_3 \cdot N3B + a_4 \cdot \%bal + a5 \quad (1)$$

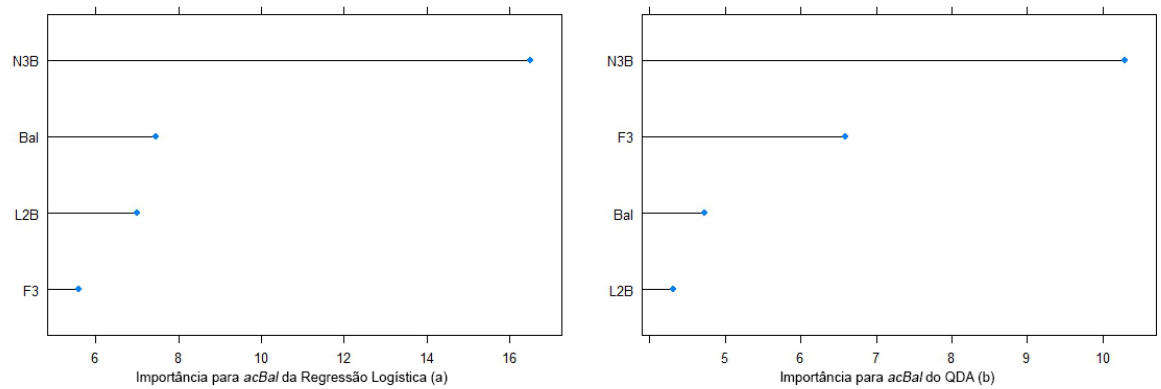
$$acbal_{B,Q} = b_1 \cdot F3 + b_2 \cdot L2B + b_3 \cdot N3B + b_4 \cdot \%bal + b5 \quad (2)$$

$$acbal_{R,L} = c_1 \cdot L2B + c_2 \cdot N3B + c_3 \cdot \%bal + c4 \quad (3)$$

$$acbal_{R,Q} = d_1 \cdot L2B + d_2 \cdot N3B + d_3 \cdot \%bal + d4 \quad (4)$$

Após o ajuste destas funções a importância das covariadas para cada um dos modelos de regressão listados acima foi computada conforme o método proposto por Gevrey, Dimopoulos e Lek (2003) e Hapfelmeier *et al.* (2014). A Figura 14 exibe os resultados deste teste para a base *Bacillus*. O mesmo teste, executado sobre a base *Rhizobium* é visto na Figura 15. Como pode ser notado nos resultados para a base *Bacillus* (Figura 14), o escore mais relevante para predição da *acbal* do classificador linear foi o N3B, seguido pelo balanceamento e L2B. Para o classificador QDA o índice de melhor associação à acurácia balanceada foi o N3B, seguido por F3 e o balanceamento.

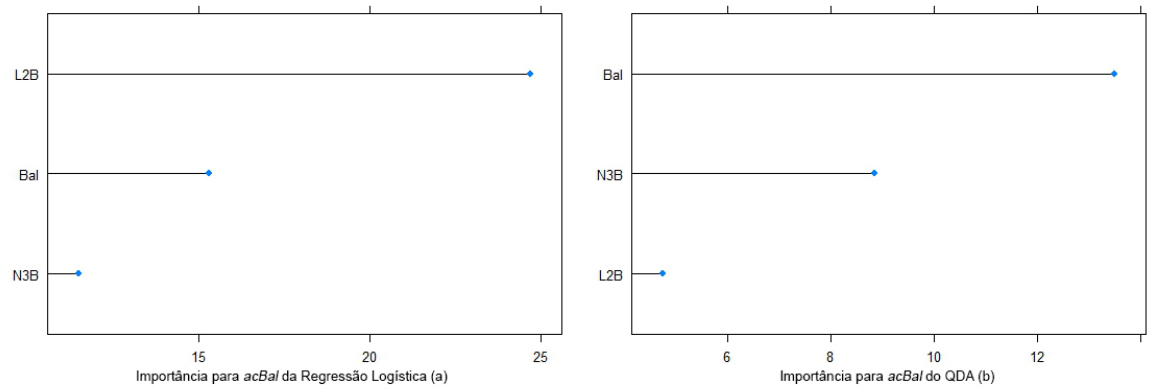
Figura 14: Resultado da importância de cada índice em relação a *acbal* dos modelos logístico e quadrático para a base *Bacillus*



Fonte: O Autor

Para a base *Rhizobium* os resultados da medida de importância (Figura 15) indicam que a medida de complexidade de maior escore de associação com a *acbal* do classificador logístico foi o L2B, seguido pelo balanceamento e N3B. Para o classificador QDA foi o balanceamento, seguido pelo N3B e então L2B.

Figura 15: Resultado da importância de cada índice em relação a *acbal* dos modelos logístico e quadrático para a base *Rhizobium*



Fonte: O Autor

5 CONCLUSÕES

Este trabalho apresentou uma investigação sobre o efeito do desbalanceamento de classes na complexidade de dados e no aprendizado de classificadores dicotômicos treinados para identificar bactérias a partir de dados provenientes de espectros de massa obtidas por MALDI-TOF. A investigação considerou o emprego dos classificadores logístico e quadrático na identificação de bactérias como pertencentes ou não aos gêneros *Bacillus* e *Rhizobium*. Primeiramente foram realizados experimentos para detectar relações de dependência linear ou não linear entre a acurácia balanceada, sensibilidade e especificidade dos classificadores e o desbalanceamento das bases de dados. Estes experimentos mostraram que a sensibilidade dos modelos possui associação exponencial forte com o desbalanceamento de classes em ambos os classificadores testados. O mesmo foi observado em relação à acurácia balanceada, exceto para o classificador quadrático na base *Bacillus*, onde foi detectada dependência exponencial moderada.

Os índices de complexidade foram qualificados conforme um critério de relevância, denominado C_0, que avaliou sua relação com desbalanceamento e à acurácia balanceada dos modelos. Dentre os índices testados, aqueles que atenderam ao critério foram F3 e *Density*. Os mesmos tiveram relação linear ao desbalanceamento dos dados e ao desempenho dos classificadores. No caso do índice *Density*, detectou-se uma relação logarítmica com desempenho do classificador quadrático. Foram propostos dois índices de complexidade de dados, que adaptam as medidas de separabilidade L2 e N3 para dados desbalanceados. As medidas apresentadas, L2B e N3B também atenderam ao critério C_0. Os resultados indicaram a existência de relação linear moderada entre tais índices e o desbalanceamento de classes. Os resultados também mostram que L2B e N3B possuem relação linear forte com o desempenho do classificador logístico e entre moderada a forte com o quadrático.

A análise de importância de descritores, apresentada por Hapfelmeier *et al.* (2014), foi aplicada a fim de avaliar a associação das medidas de complexidade F3, *Density*, L2B e N3B com desempenho dos classificadores para identificação de bactérias. No caso da identificação de bactérias do gênero *Bacillus*, esta análise mostrou que a medida de complexidade N3B obteve maior associação a acurácia balanceada de ambos os modelos. Ao identificar bactérias do gênero *Rhizobium*, o índice de melhor associação ao modelo logístico foi L2B, e N3B ao quadrático.

Como trabalhos futuros, considera-se aplicar os índices que são sensíveis ao desbalanceamento em esquemas de seleção dinâmica de modelos para sistemas com múltiplos classificadores. Adicionalmente, pretende-se estender o estudo de complexidade de dados aqui apresentado sobre outros gêneros de bactérias. A principal motivação para tal estudo é a coleta de

evidências que facilitem a seleção de funções de classificação para identificação de bactérias.

REFERÊNCIAS

AGGARWAL, C. C. *Data classification: algorithms and applications*. [S.l.]: CRC press, 2014.

ALI, R.; LEE, S.; CHUNG, T. C. Accurate multi-criteria decision making methodology for recommending machine learning algorithm. *Expert Systems with Applications*, Elsevier, v. 71, p. 257–278, 2017.

AMYES, S. G. *Bacteria: a very short introduction*. [S.l.]: OUP Oxford, 2013.

ANWAR, M. N. *Complexity measurement for dealing with class imbalance problems in classification modelling: a thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy*. Tese (Doutorado) — Massey University, 2012.

ANWAR, N.; JONES, G.; GANESH, S. Measurement of data complexity for classification problems with unbalanced data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 7, n. 3, p. 194–211, 2014.

ARLOT, S.; CELISSE, A. *et al.* A survey of cross-validation procedures for model selection. *Statistics surveys*, The author, under a Creative Commons Attribution License, v. 4, p. 40–79, 2010.

BANERJEE, G.; GORTHI, S.; CHATTOPADHYAY, P. Beneficial effects of bio-controlling agent bacillus cereus ib311 on the agricultural crop production and its biomass optimization through response surface methodology. *Anais da Academia Brasileira de Ciências*, SciELO Brasil, v. 90, n. 2, p. 2149–2159, 2018.

BEKKAR, M.; DJEMAA, H. K.; ALITOUCHE, T. A. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, v. 3, n. 10, 2013.

BELLMAN, R. E. 1957. dynamic programming. *Princeton UniversityPress. BellmanDynamic programming1957*, p. 151, 1957.

BENESTY, J. *et al.* Pearson correlation coefficient. In: *Noise reduction in speech processing*. [S.l.]: Springer, 2009. p. 1–4.

BIEMANN, K. Mass spectrometry. *Annual Review of Biochemistry*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 32, n. 1, p. 755–780, 1963.

BOU, G. *et al.* Bacterial identification methods in the microbiology laboratory. *Enfermedades infecciosas y microbiología clínica*, v. 29, n. 8, p. 601–608, 2011.

CANO, J.-R. Analysis of data complexity measures for classification. *Expert Systems with Applications*, Elsevier, v. 40, n. 12, p. 4820–4831, 2013.

CROXATTO, A.; PROD'HOM, G.; GREUB, G. Applications of maldi-tof mass spectrometry in clinical diagnostic microbiology. *FEMS microbiology reviews*, Blackwell Publishing Ltd Oxford, UK, v. 36, n. 2, p. 380–407, 2012.

DELYON, B. General results on the convergence of stochastic algorithms. *IEEE Transactions on Automatic Control*, IEEE, v. 41, n. 9, p. 1245–1255, 1996.

DRECHSLER, J.; REITER, J. P. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, Elsevier, v. 55, n. 12, p. 3232–3243, 2011.

DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. [S.l.]: John Wiley & Sons, 2012.

ELRAHMAN, S. M. A.; ABRAHAM, A. A review of class imbalance problem. *Journal of Network and Innovative Computing*, v. 1, n. 2013, p. 332–340, 2013.

FRANK, L.; HUBERT, E. Pretopological approach for supervised learning. In: IEEE. *Proceedings of 13th International Conference on Pattern Recognition*. [S.l.], 1996. v. 4, p. 256–260.

GARCIA, L. P.; CARVALHO, A. C. de; LORENA, A. C. Effect of label noise in the complexity of classification problems. *Neurocomputing*, Elsevier, v. 160, p. 108–119, 2015.

GEVREY, M.; DIMOPOULOS, I.; LEK, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, Elsevier, v. 160, n. 3, p. 249–264, 2003.

GOULART, V.; RESENDE, R. Maldi-tof: uma ferramenta revolucionária para as análises clínicas e pesquisa do câncer. *Nanocell News*, v. 1, 11 2013.

GROSS, J. H. *Mass spectrometry: a textbook*. [S.l.]: Springer Science & Business Media, 2006.

HAIR, J. F. *et al. Análise multivariada de dados*. [S.l.]: Bookman Editora, 2009.

HAIXIANG, G. *et al.* Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, Elsevier, v. 73, p. 220–239, 2017.

HAPFELMEIER, A. *et al.* A new variable importance measure for random forests with missing data. *Statistics and Computing*, Springer, v. 24, n. 1, p. 21–34, 2014.

HAWKINS, D. M. The problem of overfitting. *Journal of chemical information and computer sciences*, ACS Publications, v. 44, n. 1, p. 1–12, 2004.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, Ieee, v. 21, n. 9, p. 1263–1284, 2009.

HO, T. K.; BASU, M. Complexity measures of supervised classification problems. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 24, n. 3, p. 289–300, 2002.

HOEFFDING, W. A non-parametric test of independence. *The annals of mathematical statistics*, JSTOR, p. 546–557, 1948.

HSIEH, S.-Y. *et al.* Highly efficient classification and identification of human pathogenic bacteria by maldi-tof ms. *Molecular & cellular proteomics*, ASBMB, v. 7, n. 2, p. 448–456, 2008.

HURLBERT, A. H.; JETZ, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 104, n. 33, p. 13384–13389, 2007.

JUNIOR, J. O. L. Maldi-tof mass spectrometry of bacteria. *Mass spectrometry reviews*, Wiley Online Library, v. 20, n. 4, p. 172–194, 2001.

KAKADE, S. M. *et al. On the sample complexity of reinforcement learning*. Tese (Doutorado) — University of London London, England, 2003.

KIM, H.-Y. Analysis of variance (anova) comparing means of more than two groups. *Restorative dentistry & endodontics*, v. 39, n. 1, p. 74–77, 2014.

KOLACZYK, E. D.; CSÁRDI, G. *Statistical analysis of network data with R*. [S.l.]: Springer, 2014.

KOTSIANTIS, S. *et al.* Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, v. 30, n. 1, p. 25–36, 2006.

KRSTAJIC, D. *et al.* Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, BioMed Central, v. 6, n. 1, p. 1–15, 2014.

KUHN, M. Caret: classification and regression training. *Astrophysics Source Code Library*, 2015.

KUHN, M. *et al.* Building predictive models in r using the caret package. *Journal of statistical software*, v. 28, n. 5, p. 1–26, 2008.

LASCH, P. *et al.* Identification of bacillus anthracis by using matrix-assisted laser desorption ionization-time of flight mass spectrometry and artificial neural networks. *Appl. Environ. Microbiol.*, Am Soc Microbiol, v. 75, n. 22, p. 7229–7242, 2009.

LI, J. *et al.* Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, ACM, v. 50, n. 6, p. 94, 2018.

LORENA, A. C. *et al.* How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 52, n. 5, p. 1–34, 2019.

LU, L. *et al.* Disease status determination: Exploring imputation and selection techniques. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, JSTOR, p. 197–201, 2011.

LU, Y.; CHEUNG, Y.-m.; TANG, Y. Y. Bayes imbalance impact index: A measure of class imbalanced data set for classification problem. *IEEE transactions on neural networks and learning systems*, IEEE, 2019.

MASSON-BOIVIN, C.; SACHS, J. L. Symbiotic nitrogen fixation by rhizobia—the roots of a success story. *Current opinion in plant biology*, Elsevier, v. 44, p. 7–15, 2018.

MITCHELL, T. M. *et al.* Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, n. 37, p. 870–877, 1997.

MORAIS, G.; PRATI, R. C. Complex network measures for data set characterization. In: IEEE. *2013 Brazilian Conference on Intelligent Systems*. [S.l.], 2013. p. 12–18.

MUKAKA, M. M. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, Medical Association of Malawi, v. 24, n. 3, p. 69–71, 2012.

NORVIG, P.; RUSSELL, S. Inteligência artificial. *Editora Campus*, v. 20, 2004.

PASTERNAK, J. Novas metodologias de identificação de micro-organismos: Maldi-tof. *Einstein*, v. 10, p. 118–119, 2012.

PATTERSON, G.; ZHANG, M. Fitness functions in genetic programming for classification with unbalanced data. In: SPRINGER. *Australasian Joint Conference on Artificial Intelligence*. [S.l.], 2007. p. 769–775.

PÉREZ, N. P. *et al.* Improving the mann–whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artificial intelligence in medicine*, Elsevier, v. 63, n. 1, p. 19–31, 2015.

PONGSLIP, N. *Phenotypic and genotypic diversity of rhizobia*. [S.l.]: Bentham Science Publishers, 2012.

RADHAKRISHNAN, R.; HASHEM, A.; ABD_ALLAH, E. F. Bacillus: a biological tool for crop improvement through bio-molecular changes in adverse environments. *Frontiers in physiology*, Frontiers, v. 8, p. 667, 2017.

REITER, J. P. Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, Statistics Sweden (SCB), v. 21, n. 3, p. 441, 2005.

REYES, E. H.; OCHOA, J. A. C.; TRINIDAD, J. F. M. Classifier selection based on data complexity measures. In: SPRINGER. *Iberoamerican Congress on Pattern Recognition*. [S.l.], 2005. p. 586–592.

ROSEDAHL, L. A.; ASHBY, F. G. A difficulty predictor for perceptual category learning. *Journal of Vision*, The Association for Research in Vision and Ophthalmology, v. 19, n. 6, p. 20–20, 2019.

SANTOS, F. d. *et al.* Algoritmo knn na imputação de dados de espectros de massa do tipo maldi-tof: uma análise da influência da imputação com knn sobre o desempenho de classificadores logísticos para identificação de bactérias. Dissertação (Mestrado em Computação Aplicada), Universidade Estadual de Ponta Grossa, Ponta Grossa,, p. 82, 2018.

SAUER, S. *et al.* Classification and identification of bacteria by mass spectrometry and computational analysis. *PloS one*, Public Library of Science, v. 3, n. 7, p. e2843, 2008.

SCHUMANN, P.; MAIER, T. Maldi-tof mass spectrometry applied to classification and identification of bacteria. In: *Methods in microbiology*. [S.l.]: Elsevier, 2014. v. 41, p. 275–306.

SEBER, G. A.; LEE, A. J. *Linear regression analysis*. [S.l.]: John Wiley & Sons, 2012.

SMITH, F. W. Pattern classifier design by linear programming. *IEEE transactions on computers*, IEEE, v. 100, n. 4, p. 367–372, 1968.

SOTOCA, J.; SÁNCHEZ, J.; MOLLINEDA, R. A review of data complexity measures and their applicability to pattern classification problems. *Actas del III Taller Nacional de Mineria de Datos y Aprendizaje. TAMIDA*, p. 77–83, 2005.

SUBRAMANIAN, J.; SIMON, R. Overfitting in prediction models–is it a problem only in high dimensions? *Contemporary clinical trials*, Elsevier, v. 36, n. 2, p. 636–641, 2013.

TAMURA, H.; HOTTA, Y.; SATO, H. Novel accurate bacterial discrimination by maldi-time-of-flight ms based on ribosomal proteins coding in s10-spc-alpha operon at strain level s10-germs. *Journal of the American Society for Mass Spectrometry*, Springer, v. 24, n. 8, p. 1185–1193, 2013.

TAN, P.-N. *Introduction to data mining*. [S.l.]: Pearson Education India, 2018.

TANAKA, K. *et al.* Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid communications in mass spectrometry*, Wiley Online Library, v. 2, n. 8, p. 151–153, 1988.

TERAMOTO, K. *et al.* Phylogenetic classification of pseudomonas putida strains by maldi-ms using ribosomal subunit proteins as biomarkers. *Analytical chemistry*, ACS Publications, v. 79, n. 22, p. 8712–8719, 2007.

TOMACHEWSKI, D. *et al.* Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins. *Bioinformatics*, Oxford University Press, v. 34, n. 17, p. 3058–3060, 2018.

TOMACHEWSKI, D. *et al.* Utilização de aprendizado de máquina para classificação de bactérias através de proteínas ribossomais. Universidade Estadual de Ponta Grossa, 2017.

URBANOWICZ, R. J. *et al.* Relief-based feature selection: introduction and review. *Journal of biomedical informatics*, Elsevier, 2018.

VENABLES, W. N.; RIPLEY, B. D. *Modern applied statistics with S-PLUS*. [S.l.]: Springer Science & Business Media, 2013.

WIESER, A. *et al.* Maldi-tof ms in microbiological diagnostics—identification of microorganisms and beyond (mini review). *Applied microbiology and biotechnology*, Springer, v. 93, n. 3, p. 965–974, 2012.

WONG, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, Elsevier, v. 48, n. 9, p. 2839–2846, 2015.

XU, G. *et al.* Automatic annotation and visualization tool for mass spectrometry based glycomics. *Rapid Communications in Mass Spectrometry*, Wiley Online Library, v. 30, n. 23, p. 2471–2479, 2016.

ZUBEK, J.; PLEWCZYNSKI, D. M. Complexity curve: a graphical measure of data complexity and classifier performance. *PeerJ Computer Science*, PeerJ Inc., v. 2, p. e76, 2016.