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

RENANN RODRIGUES DA SILVA

CLASSIFICAÇÃO BACTERIANA BASEADA EM PROTEÍNAS RIBOSSOMAIS ORIUNDAS DE DADOS GENÔMICOS

PONTA GROSSA

2021

RENANN RODRIGUES DA SILVA

CLASSIFICAÇÃO BACTERIANA BASEADA EM PROTEÍNAS RIBOSSOMAIS ORIUNDAS DE DADOS GENÔMICOS

Dissertação apresentado para obtenção do título de Mestre em Computação Aplicada na Universidade Estadual de Ponta Grossa, Área de concentração: Modelagem Computacional Aplicada.

Orientador: Prof. Dr. Rafael Mazer Etto

PONTA GROSSA

2021

Silva, Renann Rodrigues da
S586 Classificação bacteriana baseada em proteínas ribossomais oriundas de dados Genomicos / Renann Rodrigues da Silva. Ponta Grossa, 2021.
64 f.

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

Orientador: Prof. Dr. Rafael Mazer Etto.

1. Proteínas ribossomais. 2. Espectrometria de massa. 3. Pesos moleculares. 4. Aprendizado de máquina. I. Etto, Rafael Mazer. II. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. III.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

Renann Rodrigues da Silva

CLASSIFICAÇÃO BACTERIANA BASEADA EM PROTEÍNAS RIBOSSOMAIS ORIUNDAS DE DADOS GENOMICOS

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. Rafael Mazer Etto - UEPG

Prof. Dr. José Carlos Ferreira da Rocha - UEPG

Profa. Dra. Maria Isabel Stets - IBMP

Ponta Grossa, 04 de dezembro de 2020.

Documento assinado eletronicamente por **Rafael Mazer Etto**, **Professor(a)**, em 04/12/2020, às 13:25, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Maria Isabel Stets**, **Usuário Externo**, em 04/12/2020, às 13:26, 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 07/12/2020, às 08:36, 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 **0352865** e o código CRC **6D9C9FDE**.

The two most important days in your life are the day you were born
and the day you find out why.
(Mark Twain)

AGRADECIMENTOS

Agradeço aos meus pais, Rozilda e Rinaldo por sempre me apoiarem em todas as minhas decisões sejam acadêmicas ou profissionais, por sempre valorizarem meus esforços em busca dos meus objetivos e por todo apoio incondicional que me deram.

Ao meu orientador Professor Doutor Rafael Mazer Etto, que me aceitou como seu orientando, pela sua ajuda incondicional em todos os momentos deste trabalho, acreditando e confiado em meu potencial.

A professora Carol, que sempre esteve disposta a me auxiliar durante o desenvolvimento desta dissertação, e em outros trabalhos.

Ao professor José Carlos, por todo o auxílio durante essa pesquisa.

Ao professor Daniel Potma por toda ajuda, conselhos, incentivos e cafés.

A Universidade Estadual de Ponta Grossa, e ao programa de pós-graduação de Computação Aplicada.

A Capes pela bolsa de estudos concedida.

Ao meu grande amigo Fabio, por todo apoio e conselho que me deu nesses anos todos. Aos meus amigos, com os quais convivi por muitas horas no laboratório e/ou em sala de aula compartilhando diversos momentos, Bruno, Giancarlo, Douglas, Rodrigo (Digão), Luiz, Alisson, Alessandro e David.

A todos do Laboratório de Biologia Molecular Microbiana (LABMOM) pelo acolhimento.

A todos que participaram direta e indiretamente desse trabalho. Muito obrigado!

RESUMO

Nas áreas da saúde e agricultura, a identificação bacteriana é essencial para compreender a composição da comunidade microbiana e a sua ecologia. As técnicas de identificação de microrganismos buscam maior precisão, rapidez e menos custo. Uma técnica que vem sendo estudada e amplamente utilizada para a identificação de microrganismos é a identificação através de espectros de massa. Por meio de picos referentes às mais abundantes massas moleculares registradas no espectro, é possível gerar um perfil para reconhecimento de um microrganismo. Outra forma de identificar um espectro de massa, é por meio de picos que são esperados que apareçam no espectro, modelo cujo qual este trabalho fez uso. Para presumir os picos esperados no espectro, foram calculados os pesos moleculares estimados de proteínas ribossomais. Essas proteínas são denominadas *house keeping*, isto é, são onipresentes e responsáveis pelo funcionamento celular básico. Além de apresentarem em grande abundância no conteúdo procariótico, as proteínas ribossomais são altamente conservadas, não possuindo alteração em sua fisiologia para diferentes meios ou estágios celulares. Os pesos estimados formaram uma base de dados presumida, contendo todas as informações obtidas do repositório do NCBI e foram utilizados somente dados notados como completo, a base de dados criada recebeu o nome de Puchuy e possui 14689 registro. Esta base de dados presumida foi gerada para taxonomia a nível de Domínio, Filo, Classe, Ordem, Família, Gênero e Espécie, e posteriormente submetida à um aprendizado de máquina. Dessa forma, foi possível obter modelos classificatórios de microrganismos baseado em valores de proteínas ribossomais. Foram gerados modelos para cada nível taxonômico, os quais foram utilizados somente os que possuíram melhor desempenho para cada nível. Ainda foi adicionado um algoritmo de agrupamento para o auxílio da classificação. Com os modelos gerados pelo aprendizado de máquina, foi desenvolvido um software, capaz classificar os microrganismos a nível de Filo, Classe, Ordem, Família, Gênero e Espécie. Por fim, foram comparados diferentes classificadores para cada nível taxonômico, com e sem a utilização de um método de agrupamento.

Palavras-chave: Proteínas Ribossomais; Espectrometria de Massa; Pesos moleculares; Aprendizado de Máquina.

ABSTRACT

In health and agriculture, bacterial identification is essential to understand the composition of the microbial community and its ecology. Microorganism identification techniques seek greater accuracy, speed, and less cost. One technique that has been studied and widely used for the identification of microorganisms is the identification through mass spectra. Through peaks referring to the most abundant molecular masses recorded in the spectrum, it is possible to generate a profile for the recognition of a microorganism. Another way to identify a mass spectrum is through peaks that are expected to appear in the spectrum, the model which this work has made use of. To assume the expected peaks in the spectrum, estimated molecular weights of ribosomal proteins were calculated. These proteins are called housekeeping, that is, they are ubiquitous and responsible for the basic cellular functioning. In addition to their abundant prokaryotic content, ribosomal proteins are highly conserved and do not change their physiology for different cell media or stages. The estimated weights formed a presumed database containing all information obtained from the NCBI repository and only data noted as complete were used, the database created was named Puchuy and has 14689 records. This presumed database was generated for taxonomy at Domain, Phylum, Class, Order, Family, Genus, and Species level, and then subjected to machine learning. Thus, it was possible to obtain classification models of microorganisms based on ribosomal protein values. Models were generated for each taxonomic level, which was used only for those that had better performance for each level. A clustering algorithm was also added to aid classification. With the models generated by the machine learning, the software was developed, able to classify the microorganisms in the Phylum, Class, Order, Family, Genus and Species level. Finally, different classifiers were compared for each taxonomic level, with and without the use of a clustering method.

Keywords: Ribosomal Proteins. Mass spectrometry. Molecular Weights. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AM	Aprendizado de Maquina
EM	Espectrometria de Massa
FN	*Falso Negativo*
FP	*Falso Positivo*
MALDI	*Matrix Assisted Laser Dessorption Ionization*
NCBI	*National Center for Biotechnology*
SVM	*Support Vector Machines*
TOF	*Time of flight*
VN	*Valor Negativo*
VP	*Valor Positivo*

SUMÁRIO

1 INTRODUÇÃO

A Bioinformática, também conhecida como Biologia Computacional, consiste na aplicação da informática para resolver problemas biológicos (GIBAS; JAMBECK; FENTON, 2001). Na agricultura a bioinformática desempenha um papel fundamental na área microbiológica, sendo uma ferramenta essencial para a identificação bacteriana a partir de dados moleculares. Identificar molecularmente bactérias que estão ligadas ao crescimento vegetal é de suma importância para a seleção de novas estirpes, que poderão ser utilizadas como biofertilizantes, visando uma agricultura mais sustentável e de alto rendimento (YADAV; SARKAR, 2019).

Segundo os autores Gans, Wolinsky e Dunbar (2005) um grama de solo pode conter 8,3 milhões de espécies bacterianas. E de acordo com Labuschagne (2003) ter o conhecimento de quais microorganismos estão presentes no solo é importante para direcionar práticas de manejo que auxiliem no aumento da produtividade, controle de patógenos, redução do uso de agroquímicos etc.

Com o objetivo de classificar taxonomicamente as bacterias de forma rápida e precisa, a Espectrometria de Massa (EM), vem sendo utilizada cada vez mais nos laboratórios de pesquisas. A EM, mais especificamente do tipo MALDI-TOF (dessorção/ionização a laser assistida por matriz e analisador de tempo-de-voo, do inglês *Matrix-assisted Dessortion/Ionization time-of-flight*), gera espectros de massas das amostras analisadas, e é amplamente utilizada devido ao baixo custo, especialmente se comparado com a técnica de sequenciamento (GARCÍA *et al.*, 2012).). Os espectros de massas gerados pelo MALDI-TOF geram informações que podem ser consideradas como “impressões digitais” de uma bactéria.

Uma maneira de identificar bactérias é utilizando algoritmos capazes de analisar dados extraídos de amostras e compará-los com dados contidos em um banco de dados (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017). O aprendizado de máquina se destaca quando se trata de análises de dados de EM. Assim como mostra os trabalhos de BRUYNE *et al.* (2010) entre outros, que a predição de um modelo de aprendizado de máquina, pode aumentar a capacidade classificatória de diferentes amostras analisadas na EM.

Quando utiliza-se dados biológicos, deve-se levar em consideração algumas características importantes como: dados faltantes, dados desbalanceados, dimensionalidade dos dados, entre outros fatores que podem prejudicar a classificação. De acordo com Wei *et al.* (2017) uma das dificuldades que envolvem dados biológicos é que estes dados são volumosos e desbalanceados.

O desbalanceamento de um conjunto de dados é prejudicial para o aprendizado de modelos de classificação, como destaca o autor Barella (2016), os classificadores treinados

em conjuntos de dados desbalanceados tendem a gerar modelos acurados para classificação de classes que possuem mais representantes e baixo desempenho nas classes que possuem menos representantes. De acordo com Galar *et al.* (2011) uma maneira de abordar o problema do desbalanceamento de um conjunto é facilitar a escolha de uma função de classificação.

Ao utilizar dados de proteínas, mais especificamente com proteínas oriundas de bactérias, os dados faltantes sobre uma determinada proteína pode ser tratado de duas formas:

1. Considerar que os dados realmente não existem, isto é, a bactéria em questão não possui a proteína;
2. Considerar que os dados para aquela proteína não foram sequenciados ou publicados.

Entretanto, somente a análise dos genomas das bactérias possibilitaria considerar uma proteína como ausente na estirpe. Segundo Eckel-Passow *et al.* (2009) existem duas principais estratégias para tratar o problema da classificação com dados incompletos, o descarte das instâncias ou imputação de dados, ou seja, excluir o representante do conjunto de treinamento ou completar registros que possuem dados faltantes.

Fazer uso de proteínas Ribossomais, as quais são proteínas essenciais para o funcionamento celular, foi a estratégia adotada pelo autor Tomachewski *et al.* (2018), uma vez que essas proteínas são conhecidas como biomarcadores confiáveis. Essas proteínas não sofrem grandes alterações em sua sequência de aminoácidos, e devido a isso, são ditas altamente conservadas (TERAMOTO *et al.*, 2007). Classificar bactérias através de proteínas ribossomais diminui a dimensionalidade da base de dados, tendo impacto direto no processo de classificação. Entretanto, no trabalho de Tomachewski *et al.* (2018) não foi possível discriminar se a ausência de proteínas ribossomais de uma determinada espécie em sua base de dados era devido a falta de registro ou a ausência do gene no genoma da bactéria.

Considerando isto, para esse trabalho foi criado uma base de dados de espectro de massa virtual, contendo apenas dados de proteínas ribossomais de bactérias que possuem genoma completamente sequenciado. A base de dados de espectro de massa virtual foi obtida através da calculadora de peptídeos conforme descrita no trabalho de Tomachewski *et al.* (2018). Logo este trabalho tem como objetivo selecionar e avaliar a eficácia de vários modelo de classificação taxonômica de bactérias baseado em valores de massa/carga de proteínas ribossomais oriundas de dados genômicos de bactérias completamente sequenciadas.

Este trabalho está dividido nas seguintes seções. Seção 2, objetivos, onde são apresentados os objetivos gerais e específicos para o trabalho. Seção 3, revisão da literatura, onde são

apresentadas as definições relacionados ao trabalho. Seção 4, metodologia, na qual se descreve o desenvolvimento e validação do procedimento de análise de um classificador mais preciso para as bactérias. Seção 5 apresenta os resultados e discussões obtidas. Seção 6 apresenta as conclusões.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Selecionar o melhor modelo de classificação taxonômica de bactérias baseados em valores de massa/carga de proteínas ribossomais oriundas de dados genômicos.

2.2 OBJETIVOS ESPECÍFICOS

- Construir um banco de dados de valores de massa/carga de proteínas ribossomais oriundas de bactérias com genoma completo;
- Avaliar modelo de *ensemble* baseado em agrupamento de proteínas ribossomais para classificação bacteriana em diferentes níveis taxonômicos;
- Comparar o desempenho de diferentes classificadores na composição do *ensemble*.

3 REVISÃO DA LITERATURA

3.1 IDENTIFICAÇÃO TAXONÔMICA DE BACTÉRIAS

De acordo com Gillis *et al.* (2001) e Kampfer e Glaeser (2011) a taxonomia bacteriana é a base para gerar hipóteses filogenéticas e evolutivas e compreendem as áreas interrelacionadas de classificação, nomenclatura e identificação.

A taxonomia bacteriana era tradicionalmente baseada em análises fenotípicas que abrangem a fisionomia de um organismo, metabolismo, bem como suas enzimas (MADIGAN *et al.*, 2009). No entanto, segundo os autores Ludwig e Klenk (2005) esta abordagem nem sempre possibilitava uma ampla visão das relações genéticas e filogenéticas dos organismos.

Similaridades de sequências gênicas indicam uma origem em comum, visto que *táxons* (unidade taxonómica, essencialmente associada a um sistema de classificação científica) com sequências similares tendem a ser filogeneticamente relacionadas. A partir da década de 70, tornou-se possível a utilização dos métodos genotípicos, na reconstrução da filogenia de grupos bacterianos (MADIGAN *et al.*, 2009).

Estudos com genes ribossomais, a partir de 1980, passaram a ser empregados como método de identificação bacteriana. Observou-se que as relações filogenéticas podiam ser determinadas comparando partes imutáveis do código genético. Regiões conhecidas como codificantes, como: 5S, 16S e 23S rRNA e espaços intergênicos foram utilizados para definir filogenia. Logo, o gene 16S rRNA tornou-se o mais amplamente utilizado como marcador molecular taxonômico em bactérias, podendo também, ser utilizado em Archaea (CLARRIDGE, 2004).

3.1.1 Morfologia, Fisiologia e Testes Bioquímicos

De acordo com, Duarte, Careli e Silva (2011) é possível classificar as bactérias por suas características fenotípicas ou genotípicas. Na classificação fenotípica dados morfológicos, bioquímicos e fisiológicos são utilizados. Na classificação genotípica relações filogenéticas dos micro-organismos podem ser melhor determinadas, por meio da comparação de sequências de proteínas ou DNA/RNA (DUARTE; CARELI; SILVA, 2011).

Na classificação fenotípica, testes como o da atividade da enzima oxidase, que utilizam aceptor de elétrons artificial, são utilizados para a diferenciação de microrganismos com base na presença ou ausência de enzimas respiratórias. A falta da enzima citocromo C diferencia as enterobactérias de outras gram-negativas (BROOKS *et al.*, 2014).

Os autores Brooks *et al.* (2014) destacam que os testes bioquímicos mais comuns para

a diferenciação de bactérias são:

1. Quebra de carboidratos;
2. Produção de catalase;
3. Utilização de citrato;
4. Coagulase;
5. Descarboxilases e deaminases.
6. Sulfeto de hidrogênio;
7. Indol;
8. Redução do nitrato;
9. Quebra do O-ortonitrofenil-β-d-galactopiranosídeo (ONPG);
10. Produção de oxidase;
11. Produção de proteinase;
12. Produção de urease;
13. Teste de Voges-Proskauer.

Como pode ser observado, existem vários testes bioquímicos que podem determinar a presença ou ausência de funções metabólicas, características essas, que podem ser utilizadas para agrupar táxons específicos. Entretanto, algumas incongruências baseadas em análises fenotípicas só puderam ser resolvidas com os avanços de técnicas moleculares, baseadas nas informações contidas nas sequencias de DNA e proteínas.

3.1.2 Sequenciamento 16S rRNA

Tido como excelente marcador molecular, o gene 16S rRNA teve seus estudos iniciados por Atlas e Bartha (1997). A utilização desse gene causou uma revolução no campo da ecologia microbiana, possibilitando investigar e determinar posições filogenéticas de comunidades bacterianas (LUDWIG *et al.*, 1997). Em virtude das unidades funcionais, porções dos genes rRNA são bem conservadas e suas sequencias podem ser utilizadas para medir distancias filogenéticas (SILVEIRA, 2004).

O gene 16S rRNA esta presente em todas as bactérias, visto que, este gene codifica o RNA que compõe a subunidade menor do ribossomo,envolvido na síntese de proteínas. Esse gene apresenta regiões conservadas e varíaveis e serve como um bom marcador molecular podendo aproximar grupos distantes filogenetciamente ou diferenciar táxons muito próximos. Entretanto, o sequencimento do gene 16S rRNA é um processo custoso, que exige qualificação técnica e não discrimina diferentes estirpes (SILVEIRA, 2004).

3.1.3 Espectrometria de Massa do Tipo MALDI-TOF

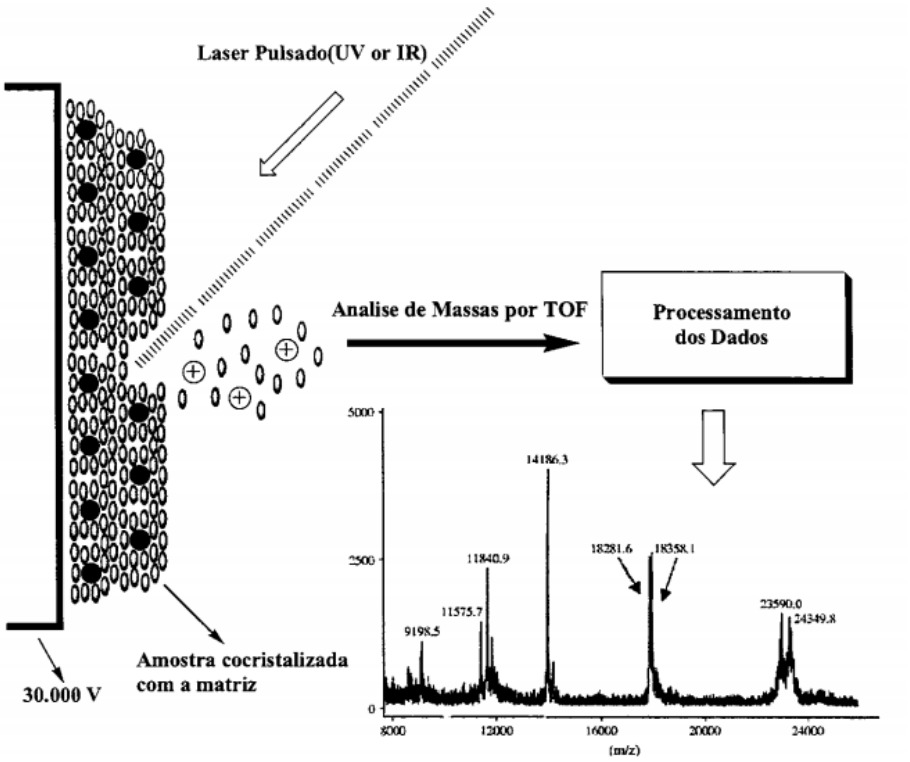
Segundo Meurer *et al.* (2003) a Espectrometria de Massa (EM) é, nos dias de hoje, reconhecidamente uma das técnicas instrumentais mais úteis e poderosas em investigações científicas e com ampla aplicação em várias áreas das ciências.

Willard *et al.* (1988) destaca que a EM é uma técnica que consiste em caracterizar as moléculas pela medida da relação massa/carga (m/z) de seus íons.

A EM é decorrente de uma variedade de técnicas de formação de íons, dentre as quais pode-se citar a ionização e dessorção por laser assistida por matriz "*Matrix Assisted Laser Dessorption Ionization*" (MALDI) e de suas técnicas de análise de razões de massa/carga (m/z) onde destaca-se os analisadores de tempo de vôo "*Time of flight*" (TOF).

A técnica de MALDI leva os íons para fase gasosa, utilizando um feixe de laser (UV ou IR). O laser é absorvido seletivamente pela matriz (ácido nicotínico devido a alta absorção no IR) contendo a amostra, a qual é ionizada e volatilizada. Logo após os íons são dirigidos para o analisador de massas através de um alto campo elétrico aplicado no anteparo da amostra, como pode ser observado na Figura 1.

Figura 1: Ionização de moléculas por irradiação a laser e separação dos íons por m/z



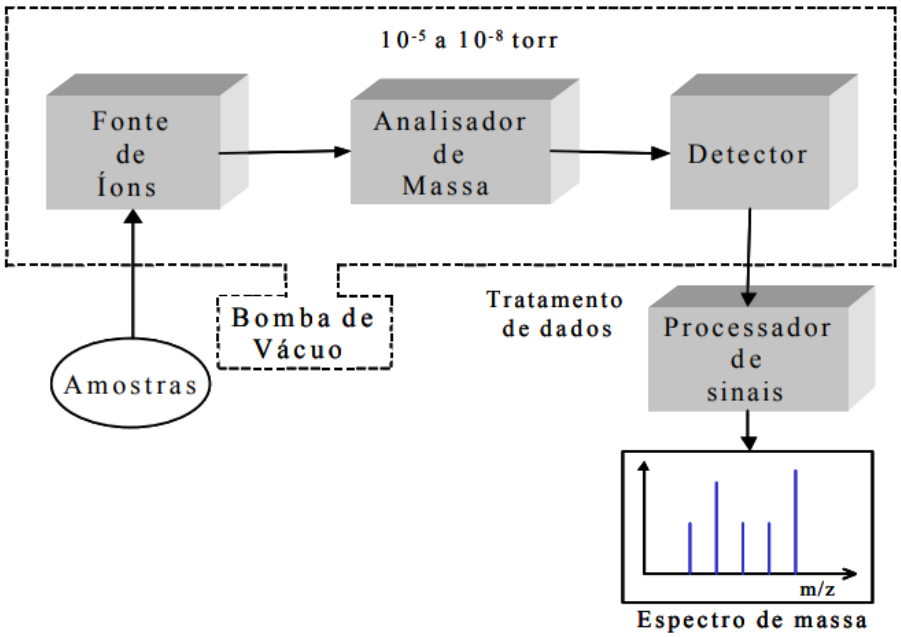
Fonte: Meurer *et al.* (2003).

De acordo com Rodriguez (2003) na Espectrometria de Massa são necessários três componentes:

1. Uma fonte de íons;
2. Um analisador de massa;
3. Um detector.

Na fonte de íons, os componentes de uma amostra são convertidos em íons, pela ação de um agente ionizante, os íons positivos ou negativos são imediatamente acelerados em direção ao analisador de massa. A função do analisador de massa é separar tais íons de acordo com a sua relação massa-carga (m/z). Os espectrômetros de massa podem ser classificados em várias categorias dependendo da natureza do analisador de massa. Finalmente um detector recebe os íons que foram separados pelo analisador, transformando a corrente de íons em sinais elétricos que são processados, armazenados na memória de um computador, conforme indica a Figura 2.

Figura 2: Componentes de um espectrômetro de massa.



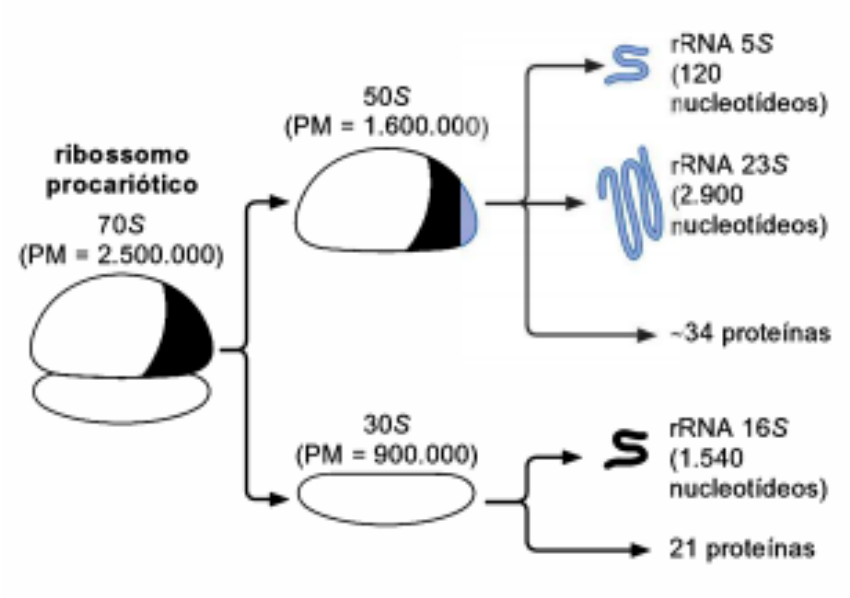
Fonte: Rodriguez (2003).

De acordo com García *et al.* (2012) nos últimos anos, a EM do tipo MALDI-TOF vem sendo utilizada para identificação de bactérias de importância clínica de modo rápido e preciso. MALDI-TOF é uma ferramenta promissora, rápida e bastante barata para a identificação bacteriana, baseada na geração de espectros de massa a partir de células inteiras (HSIEH *et al.*, 2008).

3.2 PROTEÍNAS RIBOSSOMAIS

Ribossomos são estruturas celulares encontradas em procariotos e eucariotos que promovem a síntese proteica. O ribossomo procariótico (70S) é formado por duas subunidades, conforme mostra a Figura 3, que são conhecidas como subunidade maior (50S) e subunidade menor (30S). A subunidade maior contém o centro da peptidiltransferase, e responsável pela formação das ligações peptídicas. A subunidade menor contém o centro de decodificação, no qual os tRNAs carregados leem ou "decodificam" os códons do mRNA (WATSON *et al.*, 2015).

Figura 3: Composição do ribossomo procariótico.



Fonte: Watson *et al.* (2015).

O ribossomo é formado por um conjunto de RNAs e proteínas. As proteínas ribossomais são conhecidas como *housekeeping* (EISENBERG; LEVANON, 2003), uma vez que são expressas constitutivamente e essenciais para o funcionamento celular. Essas proteínas podem ser encontradas nas subunidades menor e maior dos ribossomos. Essas proteínas não sofrem grandes alterações em sua sequência de aminoácidos e por isso, são ditas proteínas altamente conservadas (TERAMOTO *et al.*, 2007). Devido a essas características, as proteínas ribossomais podem ser consideradas como biomarcadores confiáveis para a identificação bacteriana.

Biomarcadores foram utilizados para identificação de *Bacillus subtilis* e *Pseudomonas syringae* por Demirev *et al.* (1999). Os autores realizaram testes com espectros de massa com mistura de microrganismos, espectros de organismos em diferentes estágios de crescimento, e espectros provenientes de outros laboratórios e concluíram que a técnica foi eficiente para identificação.

3.3 CENTRO NACIONAL DE INFORMAÇÕES SOBRE BIOTECNOLOGIA

O Centro Nacional de Informações sobre Biotecnologia do inglês *National Center for Biotechnology Information* (NCBI) abriga uma série de bancos de dados. As principais bases de dados incluem o GenBank para sequencias de DNA e o PubMed.

Segundo Lu e Salzberg (2018) o repositório do GenBank contém a maioria dos dados de sequência do genoma enviados por laboratórios ao redor do mundo.

O NCBI possui um projeto denominado RefSeq o qual analisa e filtra as sequencias do genoma do Genbank para criar um banco de dados mais curado. Esta base de dados também abrange os genomas bacterianos e virais, com mais de 37.000 organismos bacterianos e mais de 7.500 organismos virais representados (LU; SALZBERG, 2018).

3.4 BIOINFORMÁTICA E CLASSIFICAÇÃO

Segundo Orozco *et al.* (2013) a bioinformática é uma ciência que adquire, armazena, organiza, processa, gerencia e distribui grandes quantidades de dados e informações de caráter biológico usando infraestruturas de computação e software sofisticados. O sucesso de aplicações em bioinformática depende da capacidade de analisar dados a fim de adquirir conhecimento sobre um objeto de interesse (LIU *et al.*, 2016).

Nas ultimas décadas o Aprendizado de Maquina (AM) tem ganhado força como um recurso mais sofisticado para solução de problemas. Um algoritmo de AM oferece uma solução por meio de uma hipótese ou função com base em uma experiencia fornecida previamente. De acordo com Faceli *et al.* (2011) espera-se também que um algoritmo de aprendizado de maquina possa analisar informações, de forma a otimizar o modelo de resultados, assim como reconhecer e organizar um conhecimento novo.

Baseado nisto, os algoritmos AM para classificação tem sido utilizados em tarefas como a identificação de bactérias (TOMACHEWSKI *et al.*, 2018), predição de estrutura das proteínas e suas funções (ELBASHIR; JIANXIN; WU, 2012) e reconhecimento molecular baseado em estruturas especificas de proteínas (ELBASHIR; JIANXIN; BINBIN, 2011).

Uma vez que, são usados dados biológicos, existem uma série de problemas que podem dificultar no aprendizado de classificadores. Os autores Tan, Steinbach e Kumar (2006), Faceli *et al.* (2011) e Hardin e McCool (2015) destacam:

- Dados com ruído;
- Dados desbalanceados;
- Amostra pequena;
- Falta de representatividade.

Segundo Wei *et al.* (2017) uma das dificuldades que envolvem trabalhar com dados biológicos, são: dados desbalanceados e dados volumosos. O desbalanceamento prejudica à aprendizagem de um modelo de classificação, porque de acordo com Barella (2016), classificadores

que são treinados com dados desbalanceados criam a tendência a gerar modelos acurados na identificação de instâncias das classes que possuem maior quantidade de representantes (classe majoritárias) e baixo desempenho nas classes que possuem menor quantidade de representantes (classe minoritárias).

3.5 CLASSIFICADORES

Segundo Martins, Guimarães e Fonseca (2002), um classificador é um sistema que visa identificar objetos semelhantes como pertencentes a uma mesma classe. Da-se o nome de classe cada tipo de padrão.

O objetivo da classificação é desenvolver modelos e algoritmos que venham a ser aplicados para identificar categorias de um objeto a partir de uma coleção de atributos ou propriedades. Segundo Faceli *et al.* (2011) um classificador é, em termos matemáticos, uma função *f: X* $\rightarrow$ *Y* que concebe um modelo preditivo, cujo qual, permite determinar valores da variável categórica Y através dos valores das variáveis x_1 ... x_n pertencentes a X. Logo, os valores de Y são denotados por y_1 ... y_m, e são rótulos de possíveis hipóteses de classificação.

Segundo os autores Faceli *et al.* (2011) e Carvalho *et al.* (2011) os classificadores apresentam aprendizagem automática, isto é, inferem funções de classificação a partir de reconhecimento de um conjunto de dados, o qual possuem atributos descritivos e rotulo de classe dos objetos observados.

Os autores Woods, Kegelmeyer e Bowyer (1997) afirmam que, a principal etapa no desenvolvimento de um classificador dá-se pela utilização de algoritmos de aprendizagem de máquina, o qual é responsável por aproximar uma função de classificação que classifique corretamente os objetos do domínio. Com o proposito de alcançar este objetivo, algoritmos de aprendizagem de máquina aplicam procedimentos que combinam métodos de inferência, busca e otimização para deduzir uma função de classificação que generalize o comportamento observado no conjunto de dados (CARVALHO *et al.*, 2011).

A aprendizagem automática de um classificador pode ser separada em tipos:

Aprendizagem supervisionada: Como os autores Lorena, Gama e Faceli (2000) explicam, a aprendizagem supervisionada é realizada com a ajuda de especialistas da área, que realizam a extração dos padrões nos dados, avaliam os resultados e testes a partir da entrada. Na aprendizagem supervisionada as instancias do conjunto de treinamento possuem rotulo, através disso, novas ocorrências serão classificadas baseadas no conjunto de treinamento.

Aprendizagem não-supervisionada: De acordo com os autores, Lorena, Gama e Faceli

(2000) o método de aprendizagem não-supervisionada utiliza algoritmos inteligentes para encontrar relações e padrões. Nesse método não existe rotulo predefinido para nenhuma instancia. Um conjunto de dados é dado com a finalidade de estabelecer a existência de classes ou *clusters*.

Aprendizado semi-supervisionado: Esse método utiliza parte das duas técnicas de classificação, supervisionada e não supervisionada, onde informações são classificadas e algoritmos inteligentes ajustam o modelo a partir das informações disponíveis para alguns dados (CHAPELLE; SCHÖLKOPF; ZIEN, 2006), isto é, uma parcela dos dados de treinamento possui rótulo.

3.6 AGRUPAMENTO

De acordo com os autores Rokach e Maimon (2005) o agrupamento de objetos é tão antigo quanto a necessidade humana de descrever as características dos homens e objetos e identificá-los com um tipo. Deste modo, abrange várias disciplinas científicas: da matemática e estatística à biologia e genética, cada uma das quais utiliza termos diferentes para descrever as topologias formadas. Independente da origem da natureza dos dados, o problema se torna o mesmo: formar categorias e designar indivíduos para os grupos adequados.

O agrupamento é considerado um problema de aprendizado não supervisionado, isto é, trata de encontrar uma estrutura em uma coleção de dados não rotulados. Um *cluster* é, portanto, uma coleção de objetos que são "semelhantes" entre si e "diferentes" dos objetos pertencentes a outros *clusters* (MADHULATHA, 2012b).

Segundo o autor Madhulatha (2012b) os algoritmos de agrupamento de dados podem ser divididos em hierárquicos ou particionais. Algoritmos hierárquicos encontram *clusters* sucessivos usando *clusters* estabelecidos anteriormente, enquanto algoritmos particionais determinam todos os *clusters* ao mesmo tempo. Os algoritmos hierárquicos podem ser aglomerativos (de baixo para cima) ou divisivos (de cima para baixo). Os algoritmos aglomerativos começam com cada elemento como um *cluster* separado e os mesclam em *clusters* sucessivamente maiores. Os algoritmos de divisão começam com todo o conjunto e passam a dividi-lo em *clusters* sucessivamente menores.

Na literatura podem ser encontrados um grande diversidade de algoritmos de agrupamento . A escolha do algoritmo de agrupamento depende tanto do tipo de dados disponíveis quanto do objetivo e aplicativo específicos. Agrupamento pode ser usado como uma ferramenta descritiva ou exploratória (MADHULATHA, 2012a) . Em geral, os principais métodos de agrupamento podem ser classificados nas seguintes categorias:

1. Hierárquico;
2. Densidade;
3. K-means;
4. K-Medoids;
5. Markov Clustering Algorithm (MCL);
6. Non-negative matrix factorization (NMF);
7. Singular Value Decomposition (SVD).

Ainda, segundo Madhulatha (2012a), alguns algoritmos de agrupamento integram as ideias de vários métodos de agrupamento, o que torna difícil classificar um determinado algoritmo como pertencendo exclusivamente a apenas uma categoria de método de agrupamento.

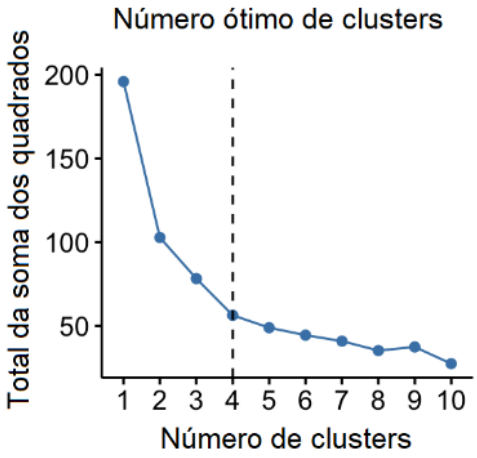
3.6.1 Método do Cotovelo

O método do cotovelo (*elbow*) é uma regra para determinar qual número de clusters deve ser escolhido. O método do cotovelo diz que deve-se escolher um número de clusters o qual a adição de outro cluster não adicione informações suficientes. Mais precisamente, se for representado graficamente a porcentagem de variância explicada pelos clusters em relação ao número de clusters, os primeiros clusters adicionarão muitas informações, mas em algum momento o ganho marginal cairá. (MADHULATHA, 2012b)

Segundo os autores Bholowalia e Kumar (2014) o método do cotovelo é um recurso que analisa a porcentagem de variação explicada em função do número de clusters. Este método consiste da ideia de que se deve escolher um número de clusters e que a partir deste número os próximos clusters adicionados não impactarão significativamente na modelagem.

No método do cotovelo, os primeiros clusters adicionarão muita informação, mas em algum momento o ganho marginal cairá drasticamente e dará um ângulo no gráfico. O número de clusters, notado como ’k’ é escolhido neste momento, daí o "Método de cotovelo". Este método começa com k = 2 e continua aumentando a cada etapa em 1, calculando seus clusters e o custo que acompanha o treinamento. Em algum valor para k, o custo cairá drasticamente e, depois disso, atinge um platô (BHOLOWALIA; KUMAR, 2014). Este é o valor k desejado, como pode ser visto na Figura 4.

Figura 4: Método do cotovelo.



Fonte: (KASSAMBARA, 2017)

3.7 TRABALHOS CORRELATOS

Alguns autores empregaram o método de aprendizagem de maquinas em dados de espectros do tipo MALDI-TOF, para realizar a identificação de bactérias, essa estrategia vem se mostrando eficaz. No trabalho de Bruyne *et al.* (2011) os autores utilizaram classificadores do tipo maquina de vetores de suporte (SVM, do inglês *Support Vector Machine*) e *random forest* para identificar bactérias do gênero *Leuconostoc* e *Fructobacillus*. Para tal, os autores elaboraram um protocolo experimental para gerar espectros de massa do tipo MALDI-TOF. Entretanto, não foram especificados os picos usados para o processo de identificação. Os autores avaliaram o desempenho dos classificadores utilizando validação cruzada, com dez partições, acurácia e F-score3 . A acurácia 98,4% para o SVM e 94,1% para a *random forests*. O F-escore foi de 96,8% para o SVM e 89,7% *random forests*. Entretanto, foi observado pelos autores que esses resultados apresentam viés de seleção em devido a bases de dados utilizadas apresentarem desbalanceamento.

Ainda utilizando dados extraídos pelo espectro de massa do tipo MALDI-TOF, os autores Rossel e Arbizu (2018) apresentaram em seu trabalho um modelo para a identificação de espécies de crustáceos, para tanto, utilizam o classificador baseado em arvore de decisão Random Forest.

O classificador SVM, juntamente com dados do tipo MALDI-TOF, também foi utilizado pelos autores Lee *et al.* (2017), com o proposito de classificar especies bacterianas dos grupos *Mycobacterium abscessus* e *Mycobacterium fortuitum*, cujo o primeiro grupo continha 3 espécies de bactérias e 309 espectros e o segundo 5 espécies de bactérias e 285 espectros.

O classificador teve seu desempenho mensurado através da acurácia. O grupo *M. abscessus* obteve-se uma acurácia de 91,61% e o grupo *M. fortuitum* 92,25%.

Através do algoritmo k-NN os autores Villanueva *et al.* (2004) obtiveram um modelo capaz de diferenciar com precisão 96,4% das amostras entre normal e doente, utilizando dados de espectro de massa extraídos de amostra de soro de pacientes com tumores cerebrais. Em torno de 400 pesos moleculares foram reunidos dos espectros, dentre eles 274 mostraram-se suficientes para diferenciação da doença.

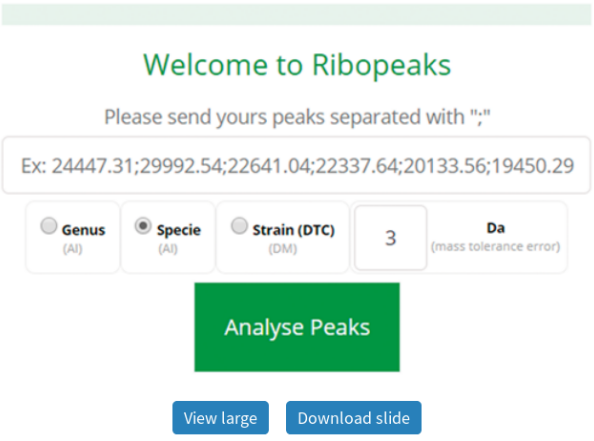
3.7.1 Ribopeaks

O *Ribopeaks* é um software para a identificação de bactérias a partir da análise da massa molecular das proteínas ribossomais geradas por MALDI-TOF (TOMACHEWSKI *et al.*, 2018). O sistema apresentado por Tomachewski *et al.* (2018) utiliza um classificador probabilístico do tipo "Bayes Ingênuo" (*naive bayes*) (RUSSEL; NORVIG, 2004) para codificar a incerteza no relacionamento entre as proteínas e as hipóteses de classificação, também computando a probabilidade posterior de cada hipótese.

Juntamente com o desenvolvimento do software *Ribopeaks* os autores criaram uma base de dados, denominada base de dados PUKYU, que possui 57 tipos diferentes de proteínas ribossomais, com um total de 28.505 registros, referentes a 6.936 espécies e 1.949 gêneros diferentes.

O *Ribopeaks* foi testado com dados extraídos de Ziegler *et al.* (2015). Esses dados não tinham sido usados para realizar o treinamento e, posteriormente, gerar o modelo de classificação taxonômico. O Ribopeaks obteve um resultado de 87,93% de acurácia em nível de espécie e 90,51% em o nível de gênero. A Figura 5 mostra a interface do Ribopeaks.

Figura 5: Interface do software Ribopeaks.



Fonte: Tomachewski *et al.* (2018).

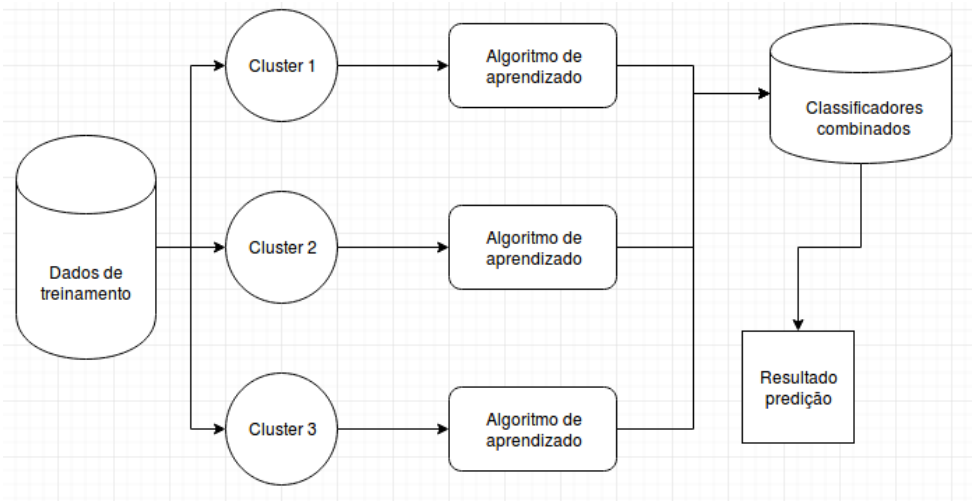
Atualmente o *Ribopeaks* trabalha com uma base de dados onde há dados faltantes, não permitindo afirmar se a falta de registro da proteína no banco de dados é devido a sua ausência no genoma ou ao desconhecimento da sua existência. Um banco construído a partir de genomas completos trará os dados faltantes como informação, isto é, a falta da proteína para aquela espécie será uma informação para aquela espécie e não uma dúvida, sendo mais um fator que contribuirá para a classificação.

3.8 AGRUPAMENTO EM MODELOS DE SISTEMA DE MÚLTIPLOS CLASSIFICADORES

Modelos ensembles podem ser pensados como uma maneira de compensar os algoritmos de aprendizado pobres (POLIKAR, 2006), pois predizem mais de um único modelo. Particionar dados de entrada pode ser relevante no desempenho de um modelo de classificação, a clusterização é uma técnica utilizada para particionar esses dados de entrada (KURLAND; KRIKON, 2011).

Segundo Ghaemi *et al.* (2009) o uso de agrupamento (clusterização) em modelos ensemble é uma técnica de meta-aprendizagem, pois gera partições de dados tornando possível trabalhar localmente, o que gera mais informação do que um trabalho de maneira global. A Figura 6 exibe um modelo ensemble utilizando agrupamento.

Figura 6: Ensemble com clusterização.



Legenda: Diagrama exibindo o processo de classificação.
Fonte: Topchy *et al.* (2004).

3.9 MÉTRICAS PARA CLASSIFICADORES

Uma matriz quadrada $M_{k \times k}$ onde as linhas indicam as classes verdadeiras e as colunas as classes preditas pelo classificador, é dita uma matriz de confusão. De acordo com Faceli *et al.* (2011) uma das métricas utilizadas para mensurar o desempenho de um classificador é através de uma matriz de confusão. O elemento localizado na posição m_{ij} da matriz de confusão, corresponde ao numero de vezes que o classificador selecionou a hipótese Y = j para o caso que possui o rotulo Y = i na base de teste. As instancias da diagonal principal da matriz correspondem as instancias classificadas corretamente, conforme indicado na Figura 7.

Figura 7: Matriz de confusão.

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

Legenda: Tabela de combinações
Fonte: (SANTOS *et al.*, 2018)

A matriz da Figura 6 exibe um problema binário. Onde classes é denotada como positiva (+) e a outra como negativa (-). É indicado com VP (Valor Positivo) o número de casos verdadeiros positivos, instâncias da classe positiva classificados corretamente e VN (Valor Negativo) representa o número de casos verdadeiros negativos, isto é, instâncias que pertencem a classe negativa e foram classificados de maneira correta. FP (Falso Positivo) são os falsos positivos, o qual informa o número de instâncias que pertencem a classe negativa, as quais foram classificados como pertencentes a classe positiva. Sendo assim, FN (Falso Negativo) indica os casos falsos negativos, instâncias que pertencem a classe positiva, porem erroneamente classificados como negativos. Logo, a soma de todos os valores presentes na matriz de confusão retorna o número total de instâncias utilizadas na etapa de teste do mesmo (N) (FACELI *et al.*, 2011).

Segundo os autores Tan, Steinbach e Kumar (2006) a medida de desempenho mais comumente utilizada para avaliar classificadores é a acurácia. Essa medida pode ser obtida através da matriz de confusão e indica a taxa de acerto do classificador, é calculada como a soma dos elementos da diagonal principal divida por N.

$$\textit{Acurácia} = \frac{\textit{total de acertos}}{\textit{total de elementos da amostra}} = \frac{VP + VN}{(VP + FN) + (VN + FP)} \quad (1)$$

Outras medidas que podem ser extraídas da matriz de confusão é o *Recall* (sensibilidade) que corresponde a taxa de acerto da classe positiva, cuja a formula esta descrita na equação 2. O *Recall* (sensibilidade) é proporção de padrões de classe positiva identificada de forma correta.

$$Recall = \frac{\textit{verdadeiro positivo}}{\textit{total de positivos da amostra}} = \frac{VP}{VP + FN} \quad (2)$$

E a precisão, cuja qual, é a porcentagem de padrões classificados como pertencentes a classe positiva de maneira correta e possui a formula descrita conforme a equação 3:

$$\textit{Precisão} = \frac{\textit{predição positiva correta}}{\textit{predição positiva}} = \frac{VP}{VP + FP} \quad (3)$$

Para calcular o F_1-score são utiliza as métricas de precisão e *recall* (sensibilidade). O F_1-score é média harmônica entre precisão e sensibilidade e esta descrito na equação 4.

$$F_1 = 2 \cdot \frac{\textit{precisão * recall}}{\textit{precisão + recall}} \quad (4)$$

Segundo os autores Anderson e Burnham (2004), é denominado *Overfitting* a produção de um modelo muito próximo ou exatamente a um conjunto de dados específico. De acordo com Reunanen (2003) o *overfitting* não possui métricas especificas para validação, porém a definições como sendo a diferença entre o desempenho no conjunto de treinamento e o conjunto de teste, foi criada por alguns autores.

$$\textit{Overfitting} = \frac{Desempenho_{teste}}{Desempenho_{treinamento}} \quad (5)$$

4 METODOLOGIA

A execução foi organizada em três etapas. Na primeira etapa foi realizado a obtenção dos dados genômicos do repositório (NCBI, *National Center for Biotechnology*), os quais foram priorizados e obtidos apenas genomas completos, juntamente com os dados genômicos foram obtidos os dados de classificação taxonômica de cada bactéria, nos níveis de: Domínio, Filo, Classe, Ordem, Família, Gênero e Especie. Após a obtenção dos genomas, foram extraídas as proteínas ribossomais para etapa seguinte. Na segunda etapa foi criado uma base de dados de massa/carga (m/z) com as proteínas ribossomais dos dados obtidos na etapa anterior, criando assim, um banco de dados de espectro de massa virtual, essa base de dados foi utilizada como entrada para a terceira etapa. Por fim, na terceira etapa foi realizado o treinamento do *ensemble* baseado em agrupamento e criado a rotina de atualização automática de todos os passos. Por fim, foi realizado a comparação dos classificadores utilizando agrupamento e não utilizando agrupamento, para constatar se houve impacto no desempenho da classificação.

4.1 ETAPA 1: AQUISIÇÃO DOS DADOS GENÔMICOS E TAXONOMIA

Para essa etapa, foi implementado um *script* para realizar o download dos genomas do repositório de genomas do NCBI. Nessa etapa somente os genomas com o atributo “completo” foram obtidos. A obtenção dos dados foi realizada usando o formato de arquivos “.fasta” contendo todas as sequencias de aminoácidos das proteínas que aquela espécie tem em seu genoma. No arquivo os aminoácidos são representados por letras: G (Glicina ou Glicocola), A (Alanina), L (Leucina), V (Valina), I (Isoleucina), P (Prolina), F (Fenilalanina), S (Serina), T (Treonina), C (Cisteina), Y (Tirosina), N (Asparagina), Q (Glutamina), D (Aspartato ou Ácido aspártico), E (Glutamato ou Ácido glutâmico), R (Arginina), K (Lisina), H (Histidina), W (Triptofano), M (Metionina).

O arquivo FASTA foi escolhido por fornecer várias opções de saída, que podem ser úteis para destacar semelhanças e diferenças em sequências alinhadas (PEARSON, 1990), além de garantir acesso a sequencias de proteínas de forma rápida.

Juntamente com a obtenção dos dados genômicos de cada bactéria foi realizado a obtenção da taxonomia completa, isto é, Domínio, Filo, Classe, Ordem, Família, Gênero e Espécie, para cada genoma obtido. As informações de taxonomia foram obtidas do NCBI através de busca pelo número de identificação da bactéria contida do sumário do repositório de genomas do NCBI. Registro de bactérias que apresentaram as nomenclaturas: “uncultured”, “unidentified” e “candidatus” foram descartadas da base de dados.

Ao fim da primeira etapa, foi obtido um total de 14.689 genomas completos juntamente com suas respectivas classificações taxonômicas. Logo após, foi realizado um pre-processamento da base, pois continham genomas cuja nomenclatura das proteínas encontrava-se fora de padrões. Em alguns caso necessitando de verificação manual de alguns registros.

4.2 ETAPA 2: CRIAÇÃO DA BASE massa/carga (m/z)

Para a construção da base de dados de massa/carga foi usado uma versão adaptada para linguagem *python* da calculadora de peptídeos proposta por Tomachewski *et al.* (2018) que considera as modificações pós-traducionais. A etapa 1 resultou em uma coleção de arquivos, um para cada especie de bactéria. Para criar uma base de dados a partir desses arquivos foi implementado um *script* em linguagem *Python*, o qual utilizou a biblioteca de bioinformática Biopython (COCK *et al.*, 2009) para a manipulação dos arquivos.

Biopython são ferramentas de código aberto disponíveis gratuitamente, para todos os principais sistemas operacionais (COCK *et al.*, 2009). A *Open Bioinformatics Foundation* (OBF, www.open-bio.org) disponibiliza em seu site o código dessas ferramentas. O Biopython permite a manipulação de arquivos FASTA, conexões com banco de dados e classificadores, entre outras funcionalidades uteis para a bioinformática de forma mais ágil e centralizada. A extração das proteínas ribossomais foi feita através da ferramenta de leitura de arquivos do BioPython.

A base de dados contendo informações de massa/carga foi construída somente com as proteínas ribossomais contidas no arquivo do genoma, ou seja, foram usadas somente as proteínas que continham a descrição "*ribosomal protein*'. Para esse fim, foi criado um *script* para a extração das proteínas, visto que, se trata de um arquivo de dados genômicos, todas as proteínas notadas dessa bactérias estão presente no arquivo.

Para cada bactéria foram identificadas e extraídas as 60 proteínas ribossomais, as quais estavam notadas da seguinte forma dentro do genoma: L1, L2, L3, L4, L5, L6, L7A, L10, L11, L7/L12, L13, L14, L15, L18, L22, L23, L24, L29, L30, S2, S3, S4, S5, S7, S8, S9, S10, S11, S12, S13, S14, S15, S17, S19, L7ae, L9, L16, L17, L19, L20, L21, L25, L27, L28, L31, L32, L33, L34, L35, L36, S1, S6, S16, S18, S20, S21, S22, S31e, THX, YCF65. As demais proteínas que não se encaixam nessas descrições foram descartadas.

Ao termino da etapa 1 e 2 foi obtido uma base de dados de massa/carga (m/z) de proteínas ribossomais de bactérias com 14.689 registros os quais possuem a distribuição conforme a Tabela 1 exibe.

Tabela 1: Distribuição taxonômica da base de dados

Filo	Classe	Ordem	Família	Gênero	Espécie
35	72	156	343	1163	3253

Fonte: O autor.

No apêndice A deste trabalho encontra-se uma tabela contendo a distribuição de classes dos 35 filos. Através de uma rápida análise, pode-se observar que existe desbalanceamento, onde 56,30% das instâncias pertencem ao filo Proteobacteria, 22,85% ao filo Firmicutes, 9,83% ao filo Actinobacteria e 3,62% ao filo Bacteroidetes e 7,37% representam os demais 31 filos. Além disso, existem 18 filos que possuem menos de 10 representantes.

4.3 ETAPA 3: CLASSIFICAÇÃO, ANÁLISE E ATUALIZAÇÃO

Nesta etapa foi construído os modelos para classificação utilizando a biblioteca de python *Scikit-learn* (PEDREGOSA *et al.*, 2011). Para tanto, antes do treinamento propriamente dito, foi realizado um pré-processamento da base de dados, tendo em vista que trata-se de um aprendizado supervisionado, alguns rótulos de classes tiveram que ser padronizados. Para cada classificador foram realizados a validação cruzada com dez *folds*.

Os classificadores base do *ensemble* utilizados foram implementados na linguagem Python com o uso da biblioteca *scikit-learn* (PEDREGOSA *et al.*, 2011) os quais serão:

- SVM (*Support Vector Machines*);
- *Gaussian Naive Bayes*;
- *Multinomial Naive Bayes*;
- *AdaBoost Classifier*;
- *Multi-layer Perceptron Classifier*;
- *Random Forest*;
- *Decision Tree*.

Os classificadores anteriormente listados foram treinados da seguinte forma: Para cada nível taxonômico (Filo, Classe, Ordem, Família, Gênero e Espécie) foi treinado um modelo de cada classificador listado. Logo apos foi realizado o *ensemble* baseado em agrupamento, o qual foi aplicado *K-means* para realização desse agrupamento, que visa agrupar os níveis taxonômicos similares auxiliando na classificação.

Para a seleção do melhor classificador para cada etapa da classificação foram utilizados as métricas descritas na seção 4.3.1. Para a escolha do numero de *clusters* criado pelo agrupador, foi utilizado o método do cotovelo, descrito na seção 3.6.1.

4.3.1 Métricas Para Comparações

A fim de avaliar a eficácia de diferentes modelos de classificador para a construção do *ensemble* baseado em agrupamento, para identificação de bactérias usando dados de espectros MALDI-TOF, foram utilizado as métricas descritas na seção 3.9, as quais são:

1. Acurácia, medida de desempenho a qual mostra a proporção de classificações corretas, tanto de casos positivos quanto negativos.

$$\textit{Acurácia} = \frac{\textit{total de acertos}}{\textit{total de elementos da amostra}} = \frac{VP + VN}{(VP + FN) + (VN + FP)}$$

2. *Recall* (sensibilidade), medida de proporção de padrões da classe positiva identificada corretamente.

$$Recall = \frac{\textit{verdadeiro positivo}}{\textit{total de positivos da amostra}} = \frac{VP}{VP + FN}$$

3. Precisão, porcentagens classificadas como pertencentes a classe positiva as quais de fato pertencem a classe positiva.

$$\textit{Precisão} = \frac{\textit{predição positiva correta}}{\textit{predição positiva}} = \frac{VP}{VP + FP}$$

4. F_{medida} ($f_1 score$), média harmônica entre precisão e sensibilidade.

$$F_1 = 2 \cdot \frac{\textit{precisão * recall}}{\textit{precisão + recall}}$$

5. *Overfitting*, diferença entre o desempenho no conjunto de treinamento e o conjunto de teste.

$$\textit{Overfitting} = \frac{Desempenho_{teste}}{Desempenho_{treinamento}}$$

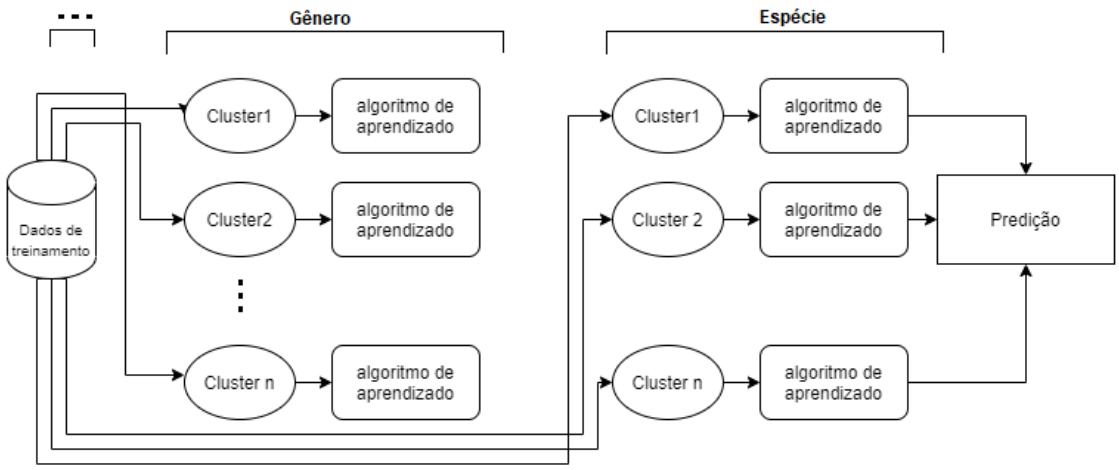
Essas métricas permitem avaliar qual classificador teve melhor desempenho para a classificação bacteriana.

4.3.2 Criação do *Ensemble* Baseado em Agrupamento

O *ensemble* utilizado nessa etapa foi implementado baseado em agrupamento, onde para cada nível taxonômico foi feito um novo agrupamento de dados e uma classificação. O inicio deu-se pelo nível taxonômico Filo onde foram clusterizados os dados de entrada e os cinco melhores resultados passaram para a classificação que por sua vez retornou os cinco melhores resultados.

Os resultados obtidos com o agrupamento e classificação do Filo serviram de entrada para o nível taxonômico Classe, que por sua vez serviu para repetir o processo do Filo passando os cinco melhores resultados para o nível taxonômico Ordem. Esse processo foi repetido até chegar ao nível taxonômico de espécie, conforme ilustrado na Figura 8.

Figura 8: Ensemble com clusterização proposto.



Legenda: Diagrama exibindo o processo de classificação para os níveis taxonômicos de Gênero e Espécie.
Fonte: O autor.

O número de melhores resultados entre cada clusterização e classificação pode ser alterado, não necessariamente sendo fixo em cinco. Por fim, uma rotina de atualização da base de dados foi elaborada para que em um período determinado a base seja atualizada.

5 RESULTADOS E DISCUSSÃO

Nesta seção estão relatados os resultados obtidos com a execução dos experimentos. Em cada caso foram realizados a validação cruzada com dez *folds*. Como foi utilizado agrupamento, o método do cotovelo (*elbow*) foi aplicado para determinar o número de clusters que seria criado pelo agrupador.

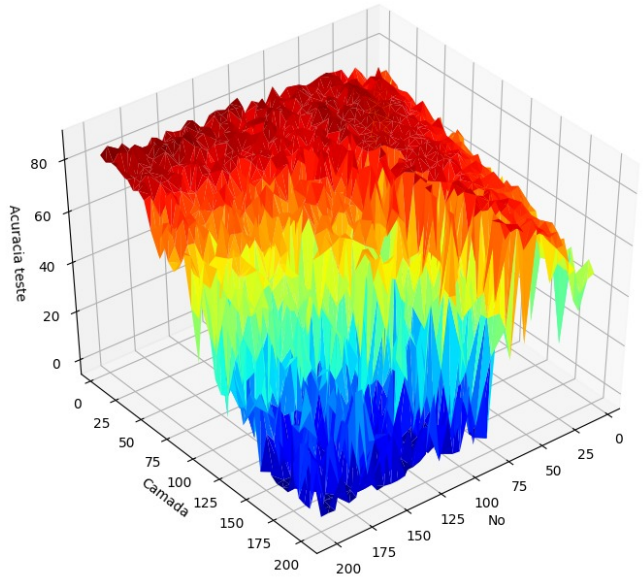
Os classificadores, em sua maioria, foram utilizados sem alterações em seus parâmetros, ou seja, usando as configurações padrões que a biblioteca disponibiliza, que no geral, não afetam significativamente o desempenho do classificador. Porém, o classificador Multilayer Perceptron possui impacto significativo em seu desempenho de acordo com os parâmetros, para esse classificador foi necessário realizar testes a fim de determinar quais parâmetros seriam utilizados.

A base de dados gerada para este trabalho, conforme descrita na seção 4.2, foi registrada junto ao INPI - (Instituto Nacional da Propriedade Industrial). Essa base de dados, por se tratar de dados oriundos de dados genômicos completos, elimina a incerteza sobre os dados faltantes, uma vez que, caso a proteína esteja faltando para uma determinada bactéria isso significa que essa bactéria não possui essa proteína, ou seja, essa proteína não é expressa para aquela bactéria. Ainda, essa base de dados pode gerar estudos posteriores, onde a falta da proteína para uma determinada bactéria pode ser vista como informação na hora de sua classificação.

5.1 PARÂMETROS PARA MULTI-LAYER PERCEPTRON

O Multi-layer perceptron (MLP) é um algoritmo de aprendizado supervisionado que aprende uma função treinando em um conjunto de dados, onde há um número de dimensões de entrada e o número de dimensões para saída. Dado um conjunto de dados e um objetivo, o MLP pode aproximar uma função não linear para classificação ou regressão. Diferente da regressão logística, pois entre a camada de entrada e a de saída, pode haver uma ou mais camadas não lineares, chamadas camadas ocultas. (PEDREGOSA *et al.*, 2011).

Figura 9: Acurácia multi-layer perceptron por parâmetros



Legenda: Gráfico com a variação da acurácia da rede neural Multi-layer Perceptron de acordo com a variação dos parâmetros Camada e Nó
Fonte: O autor.

É possível alterar os parâmetros para o MLP, os quais causam impacto em seu desempenho na classificação. Para este trabalho, foram realizados teste a fim de determinar quais os parâmetros seriam utilizados para o classificador. Como pode ser visto na Figura 9, a Acurácia do MLP é afetada através da variação dos parâmetros: Número de camadas e Número de nós por camada.

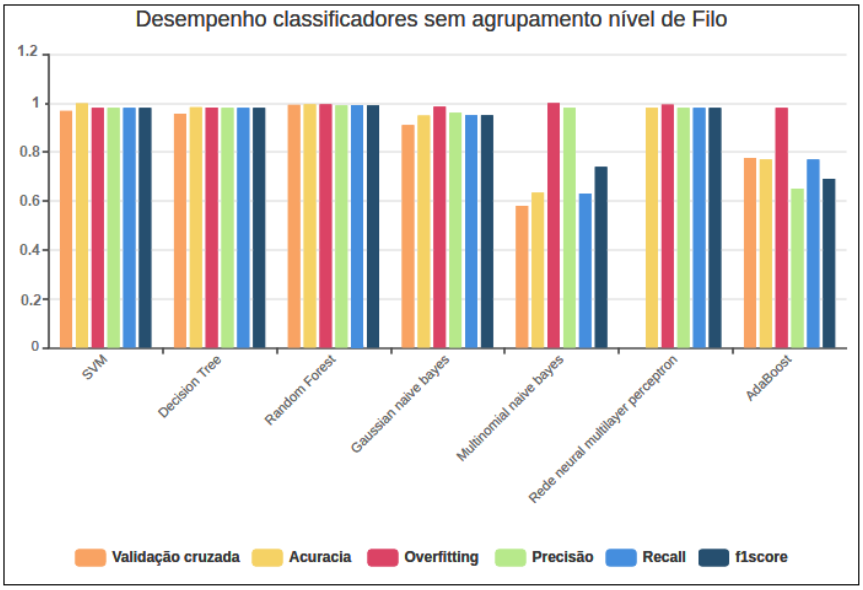
A medida que o número de camadas e de nós por camada aumenta, a acurácia do MLP cai drasticamente. O tempo de processamento para o MLP também é afetado de acordo com o numero de camadas e de nós atribuídos para a rede, quanto mais nós por camadas e quanto mais camadas, maior o tempo de execução. Considerando isso, para este trabalho os parâmetros para o MLP foram ajustados de tal forma que ofereçam o melhor desempenho possível tanto em acurácia do classificador quanto em tempo de execução, portanto foram atribuído para a rede 5 camadas e 195 nós.

5.2 CLASSIFICADORES SEM AGRUPAMENTO

Nesta seção encontra-se o resultado de classificador para cada nível taxonômico, sem o uso do agrupamento. O classificador SVM não possui resultados em validação cruzada devido ao tempo utilizado por esse classificador para gerar o modelo.

Em alguns testes conforme o número de classes da base de treinamento aumentava, alguns classificadores não conseguiram finalizar a analise e gerar um resultado em tempo hábil para apresentação desse projeto.

Figura 10: Desempenho classificadores para nível taxonômico de Filo.

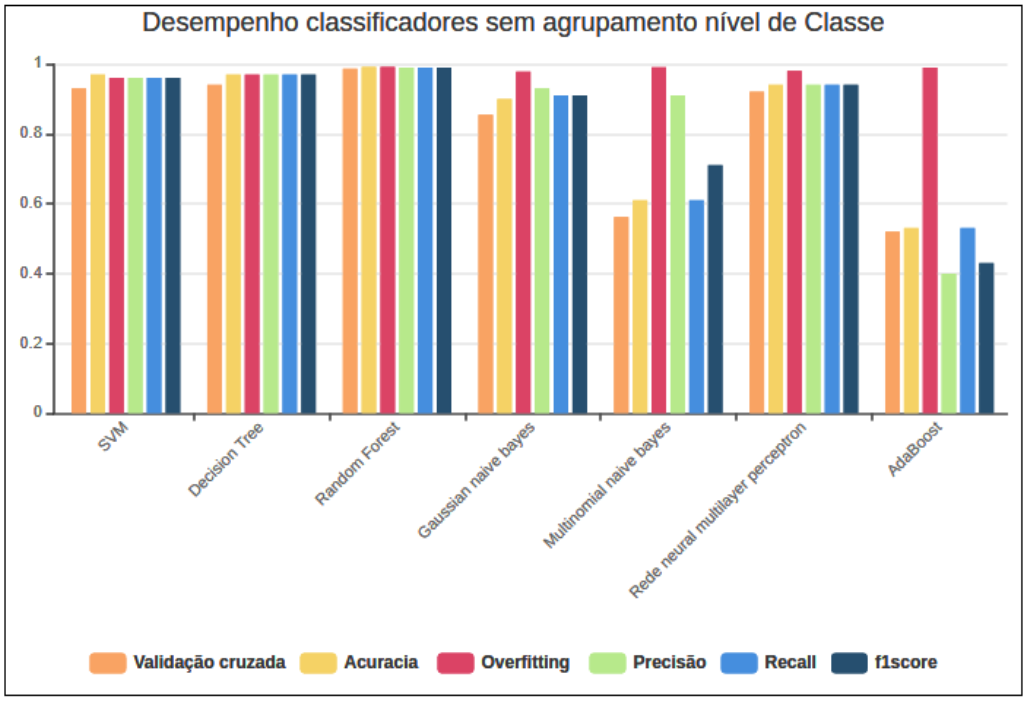


Fonte: O autor.

Como pode ser visto na Figura 10, o classificador que possui o melhor desempenho para a classificação de bactérias ao nível taxonômico de Filo foi o algoritmo de *Random Forest*. O nível taxonômico de Filo possui 35 classes, os classificadores em sua maioria tiveram resultados similares.

Quando trata-se do nível taxonômico de Classe, o numero de classes para classificação salta para 72. Os classificadores, SVM, *Decision Tree*, *Random Forest*, *Gaussian Naive Bayes*, *Multilayer Perceptron* não tiveram seu resultado impactado pelo elevação do numero de classes, mantiveram acurácia e *overfitting* na casa dos 90%. *Random Forest* manteve desempenho na casa dos 99% conforme pode ser observado na Figura 11.

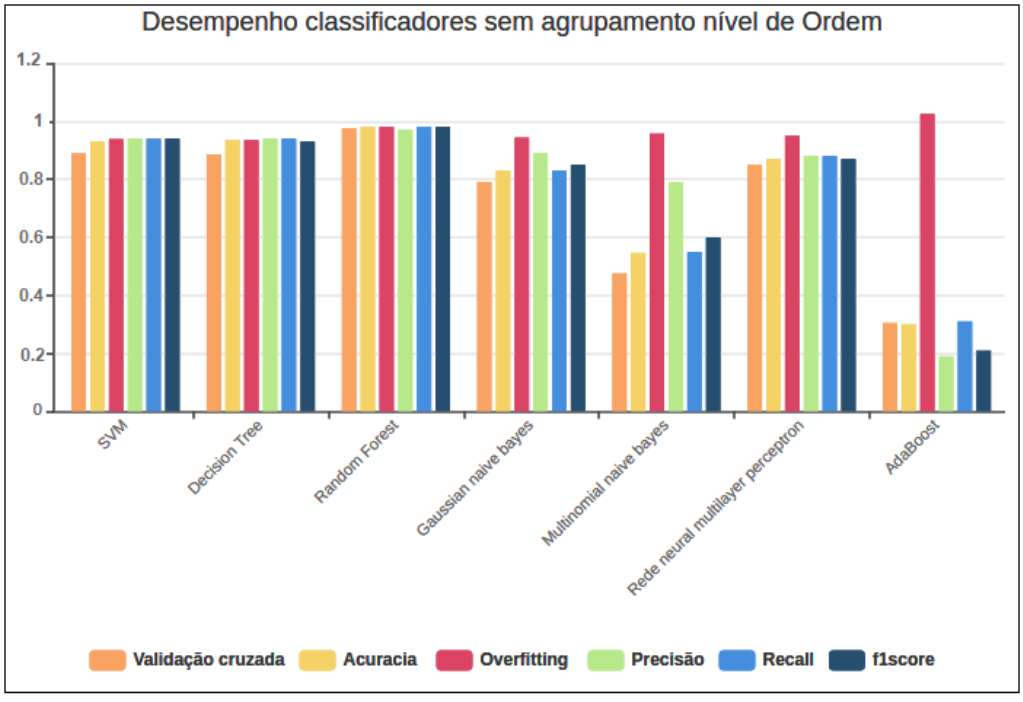
Figura 11: Desempenho classificadores para nível taxonômico de Classe



Fonte: O autor.

No nível taxonômico de Ordem possui 156 classes para serem classificadas, os resultados de cada classificador pode ser visto na Figura 12.

Figura 12: Desempenho classificadores para nível taxonômico de Ordem

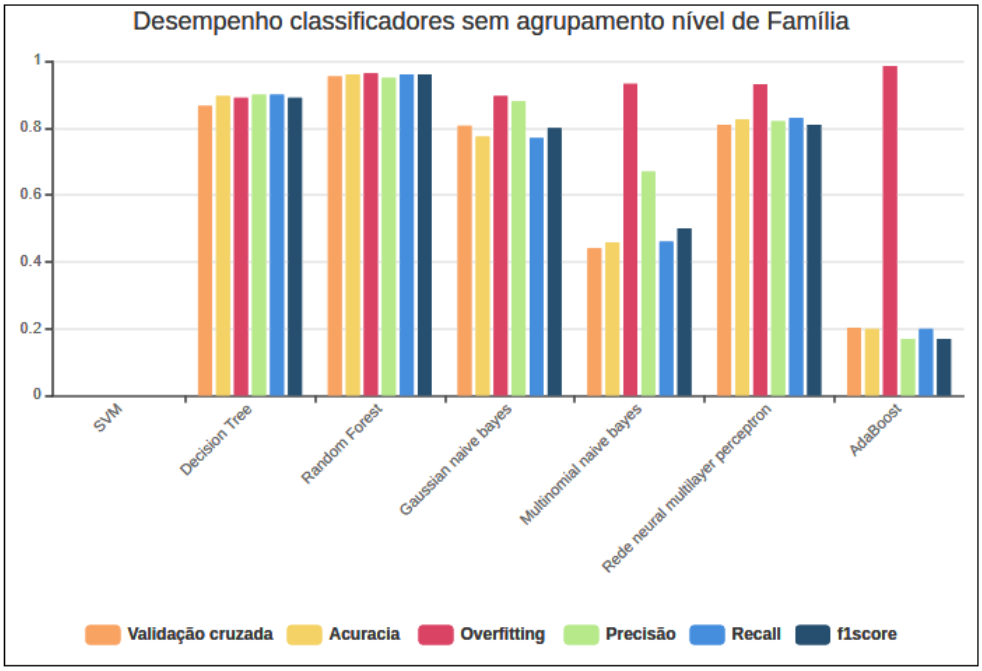


Fonte: O autor.

O nível taxonômico de Família possui 343 classes para serem classificadas. Os classifi-

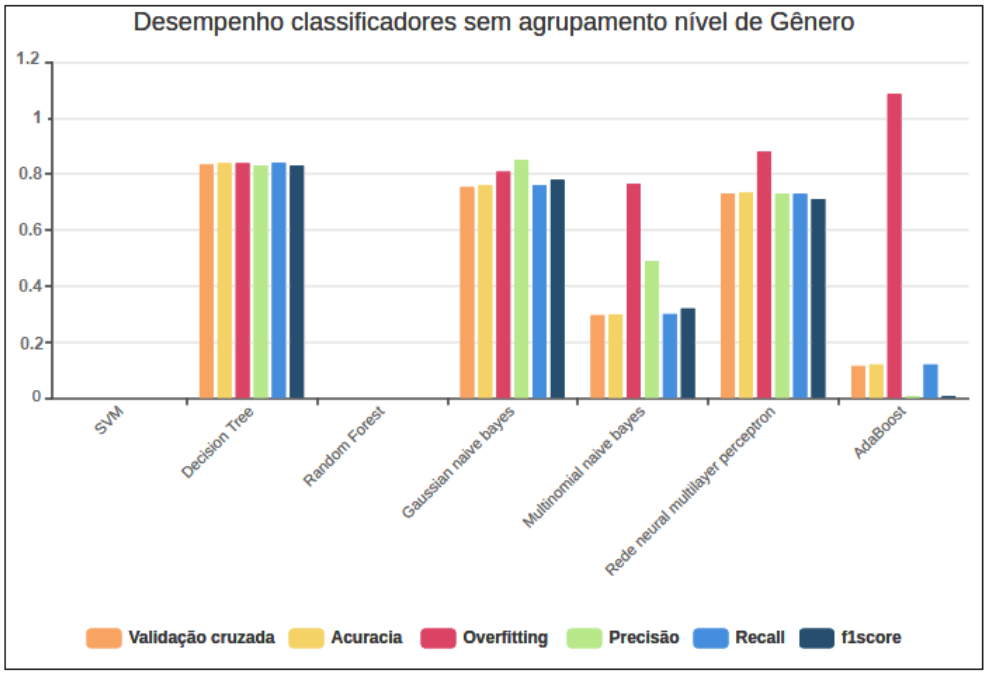
cadores obtiveram resultados conforme a Figura 13 . *Random Forest* obteve o melhor resultado para este nível. O algoritmo SVM não obteve resultados para esse nível devido ao tempo de processamento.

Figura 13: Desempenho classificadores para nível taxonômico de Família



Fonte: O autor.

Figura 14: Desempenho classificadores para nível taxonômico de Gênero

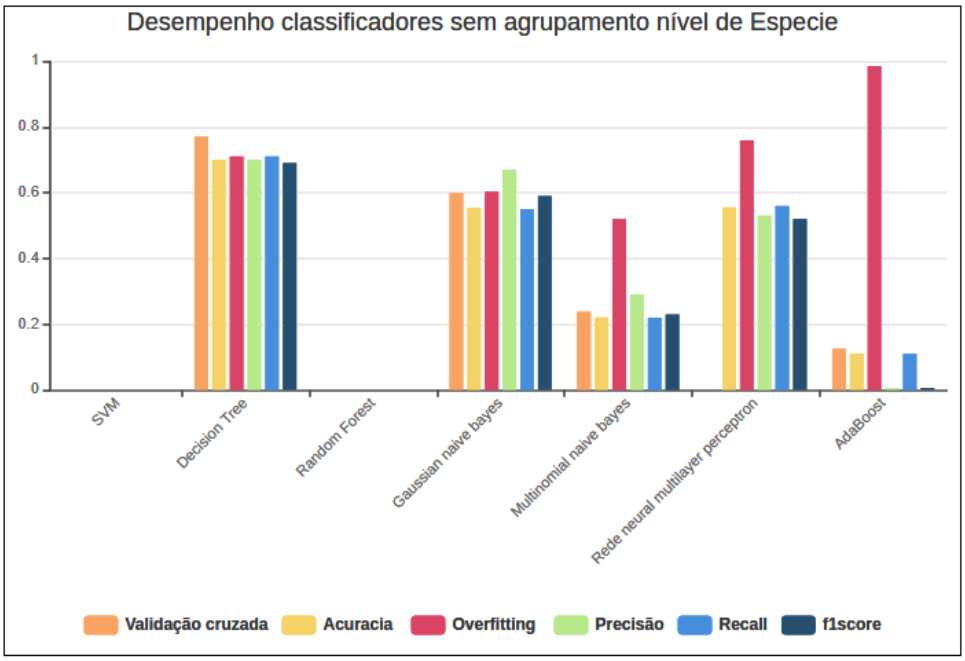


Fonte: O autor.

De acordo com a Figura 14 e 15, pode ser observado que o melhor resultado ficou

com *Decision tree*, pois SVM e *Random Forest* não representaram resultados para este nível taxonômico. SVM não obteve resultados ha tempo, e *Random forest* não executou por limitações de *Hardware*.

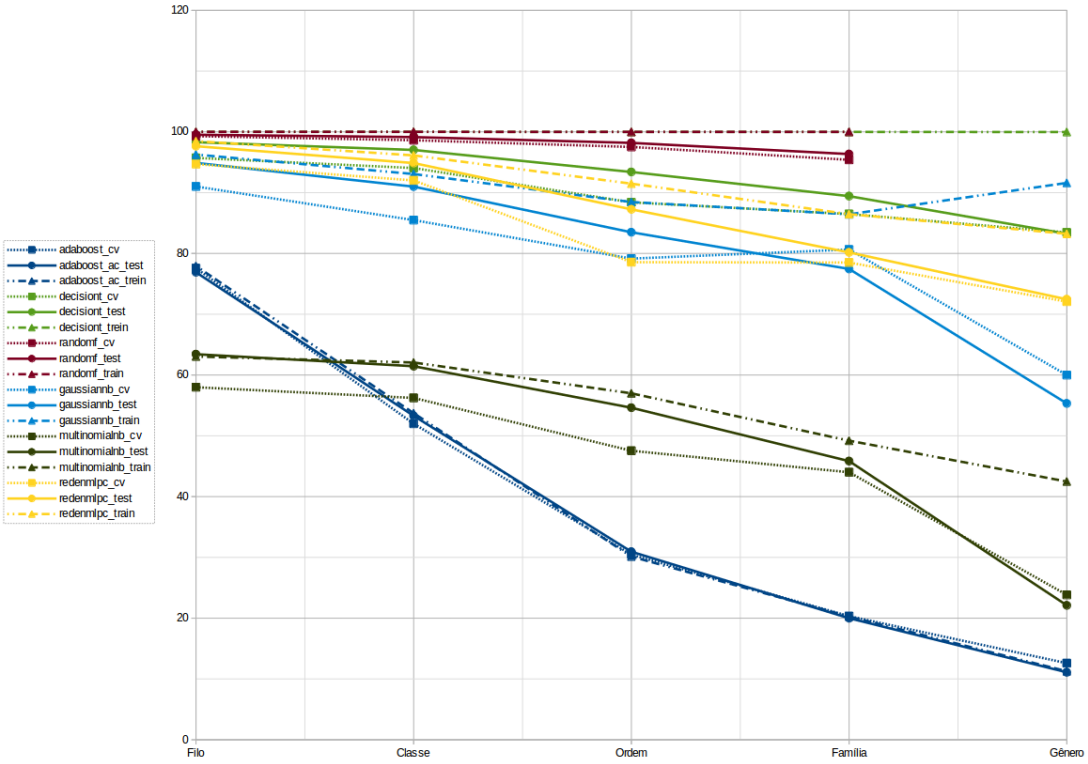
Figura 15: Desempenho classificadores para nível taxonômico de Especie



Fonte: O autor.

A Figura 16 exibe um gráfico de desempenho de todos os classificadores para cada nível taxonômico até gênero. É possível observar a queda de alguns classificadores ao ser aumentado o numero de classes a serem preditas. O classificador *Adaboost* é o mais afetado com o aumento de classes. As linhas que possuem pontilhado indicam o desempenho que o classificador atingiu na validação cruzada, as linhas com tracejado representam o desempenho do classificador no treinamento, por fim a linha representa o desempenho do classificador no teste.

Figura 16: Queda de desempenho dos classificadores *Adaboost* e *Multinomial Naive Bayes* com o aumento de número de grupo (classe) nos diferentes níveis taxonômicos

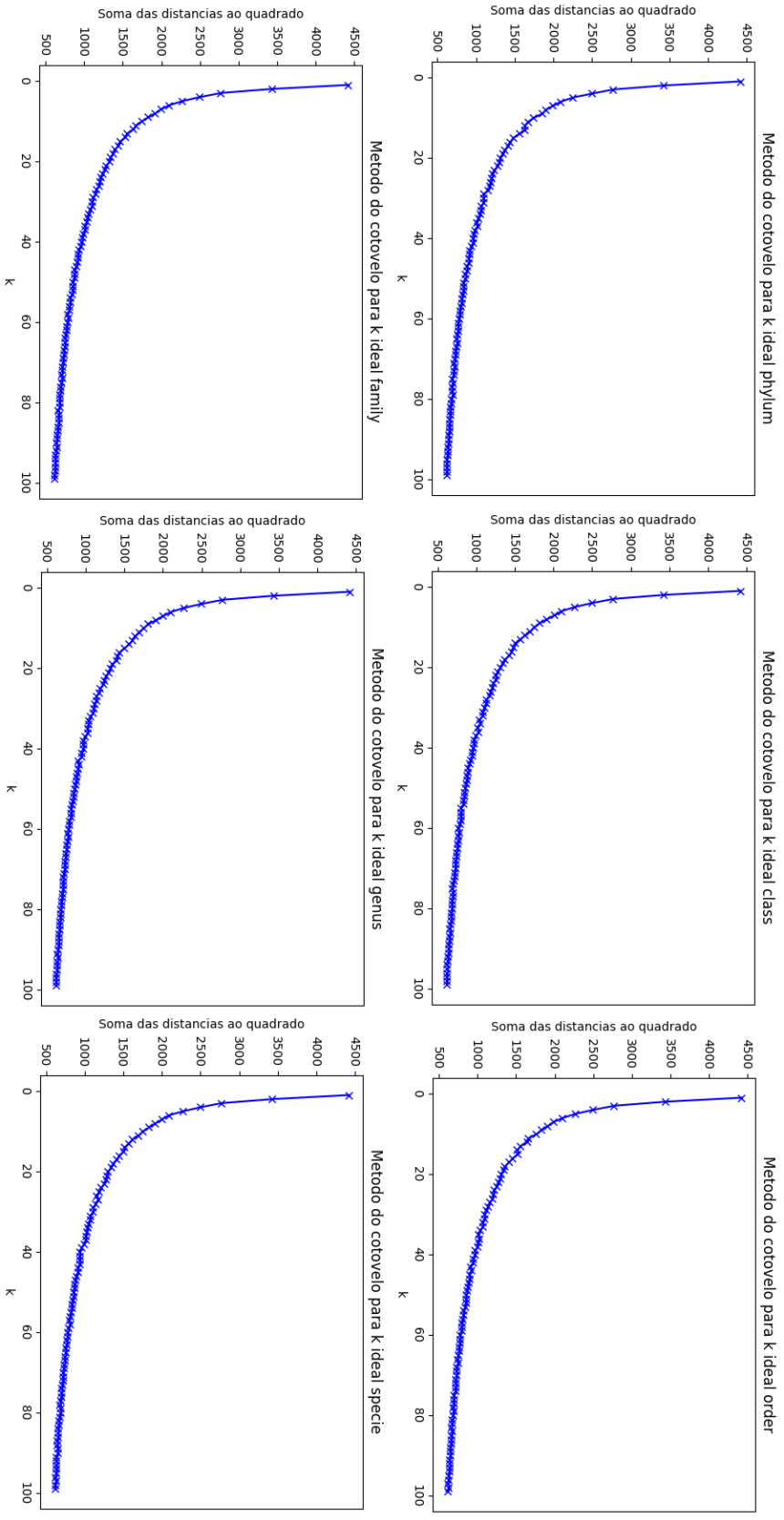


Legenda: Queda do desempenho dos classificadores de acordo com o nível taxonômico.
Fonte: O autor.

5.3 MÉTODO DE AGRUPAMENTO

Para determinar o número de *clusters* que o algoritmo de agrupamento *K-means* iria utilizar para cada nível taxonômico, foi aplicado o método do cotovelo (*elbow*), descrito na seção 3.6.1. Através do método do cotovelo foi obtido resultados conforme mostra a Figura 17. Para todos os níveis taxonômicos o método do cotovelo teve resultados similares, o qual foi estabelecido um valor prefixado de clusters igual a quinze, para todos os níveis taxonômicos.

Figura 17: Método de *elbow* aplicado na base de dados



Legenda: Resultado do teste de *elbow* para cada nível taxonômico.

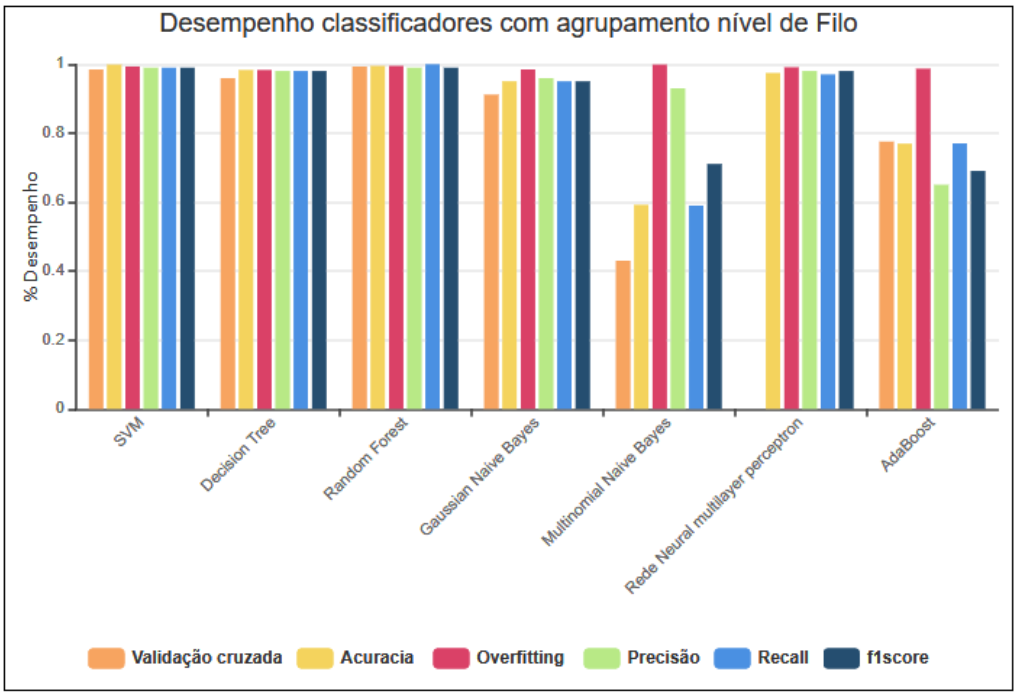
Fonte: O autor.

5.4 CLASSIFICADORES UTILIZANDO AGRUPAMENTO

Após ser identificado o número de *cluster* que seriam criados para cada nível taxonômico, conforme foi descrito na seção 5.3, foi aplicado o agrupador (algoritmo *K-Means*) conforme especificado na seção 4.3.2. Os resultados de cada classificador utilizando o agrupador para o nível taxonômico de Filo estão apresentados na Figura 18.

Utilizando o método de agrupamento, para o nível taxonômico de Filo, houve perda de desempenho somente no classificador *Multinomial Naive Bayes*, os demais classificadores obtiveram melhoras em seus desempenhos.

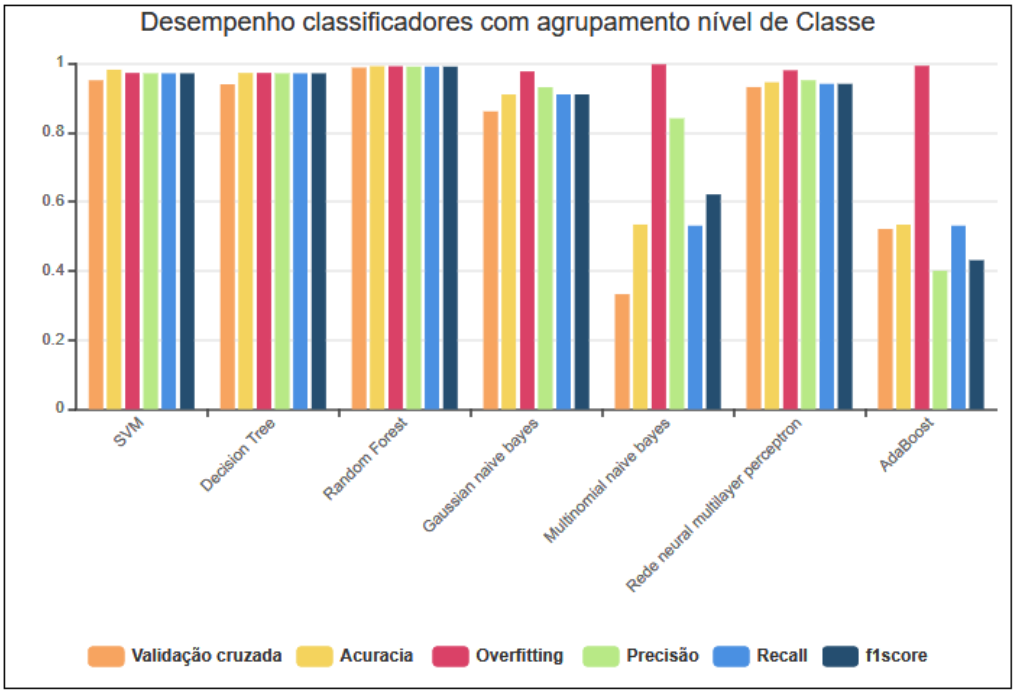
Figura 18: Desempenho classificadores com agrupamento nível de Filo



Fonte: O autor.

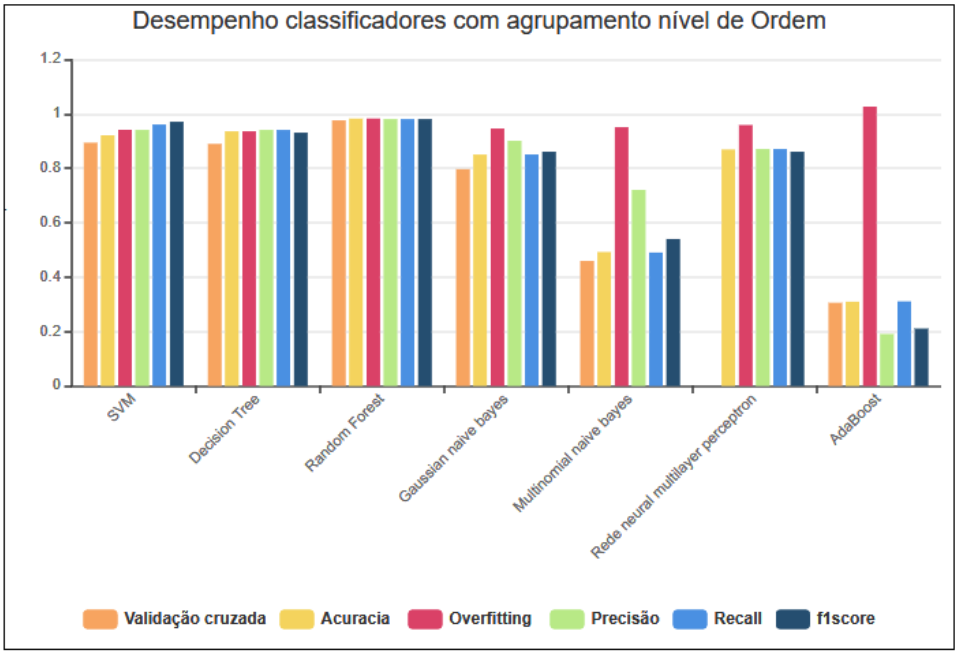
As Figuras 19, 20, 21, 22 e 23 mostram os desempenhos dos classificadores para os níveis taxonômicos Classe, Ordem, Família, Gênero e Especie respectivamente.

Figura 19: Desempenho classificadores com agrupamento nível de Classe



Fonte: O autor.

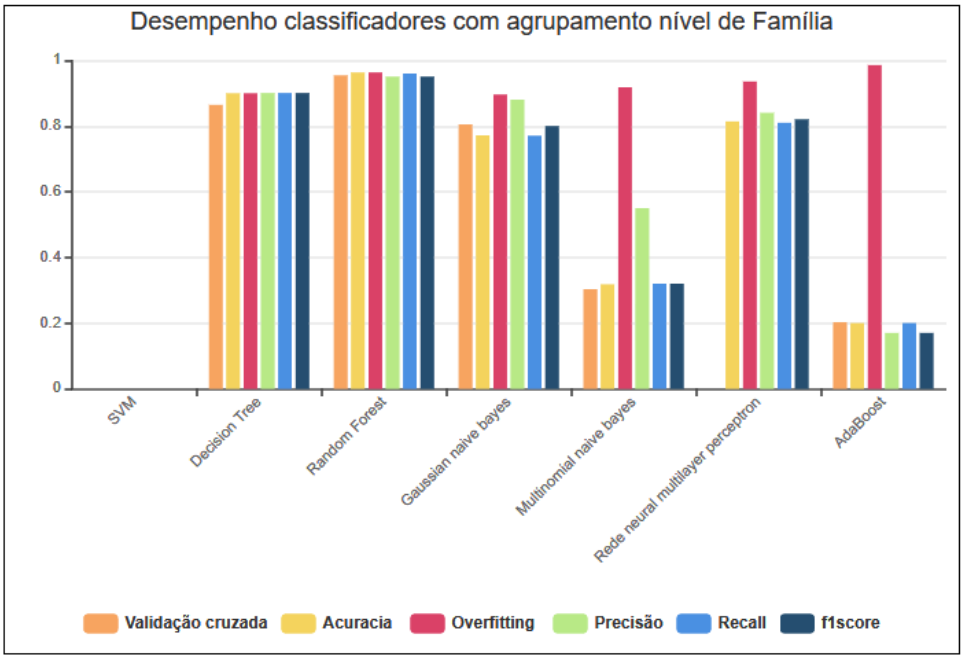
Figura 20: Desempenho classificadores com agrupamento nível de Ordem



Fonte: O autor.

Na Figura 20 é possível observar que o classificador *AdaBoost* obteve um overfitting superior a 1 (100%) indicando que a acurácia do treinamento foi inferior a acurácia de teste.

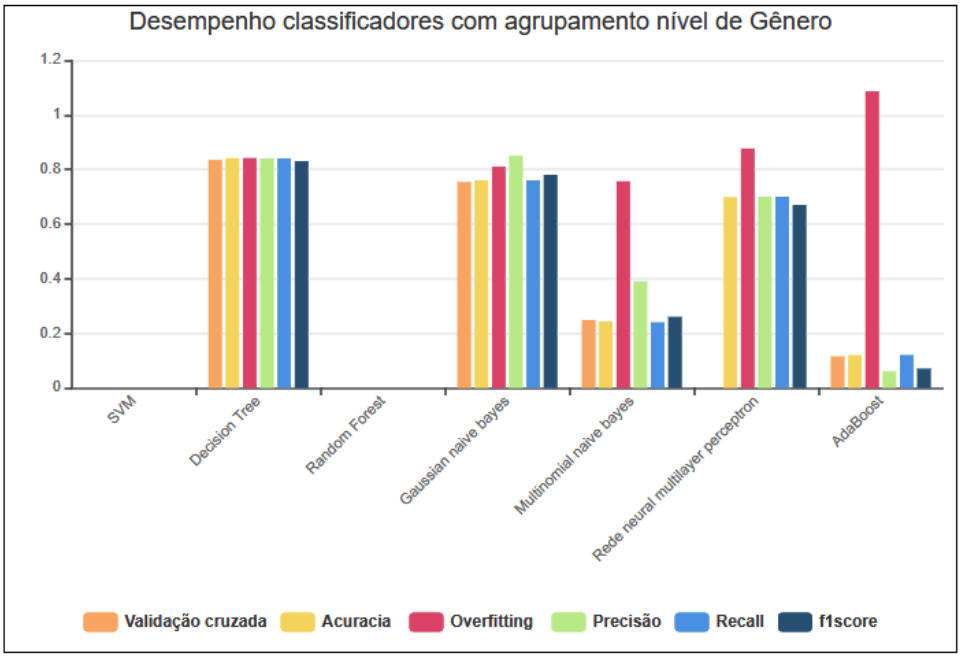
Figura 21: Desempenho classificadores com agrupamento nível de Família



Fonte: O autor.

A Figura 21 exibe os resultados para o nível de Família, o qual pode ser observado uma queda no desempenho dos classificadores *Multinomial Naive Bayes* e *Adaboost*.

Figura 22: Desempenho classificadores com agrupamento nível de Gênero

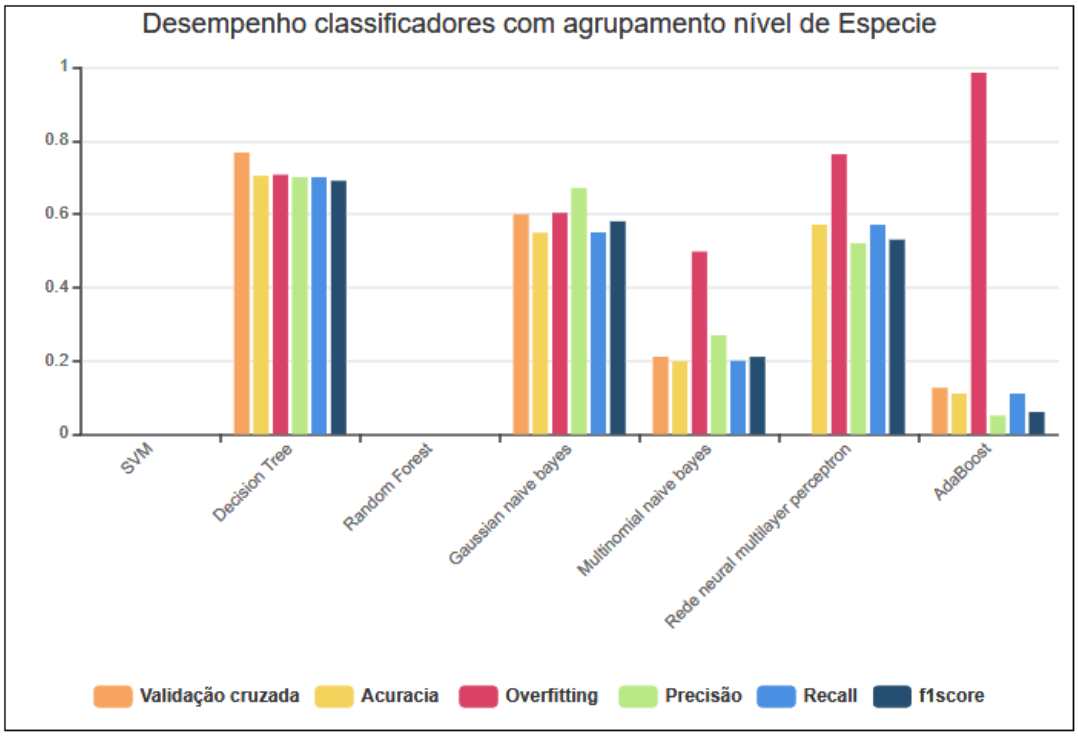


Fonte: O autor.

Para o nível de especie (Figura 23), a base de dados possui 3253 grupos taxonômicos (classes) diferentes, o qual torna o classificador SVM computacionalmente custoso. Devido a

isso, o modelo de classificação para esse classificador não pode ser criado em tempo para esse projeto.

Figura 23: Desempenho classificadores com agrupamento nível de Especie



Fonte: O autor.

Conforme descrito nesta seção, os resultados obtidos foram, de maneira geral, positivo. A utilização de um algoritmo de agrupamento, no caso deste trabalho *K-Means*, mostrou uma melhora no desempenho de alguns classificadores quando comparado com a metodologia sem agrupamento (item 5.2). SVM, *Decision Tree*, *Random Forest* e rede neural *multilayer perceptron* obtiveram ganho em seus desempenhos, porém *Adaboost* e *Multinomial Naive Bayes* tiveram seus desempenhos afetados com o agrupamento.

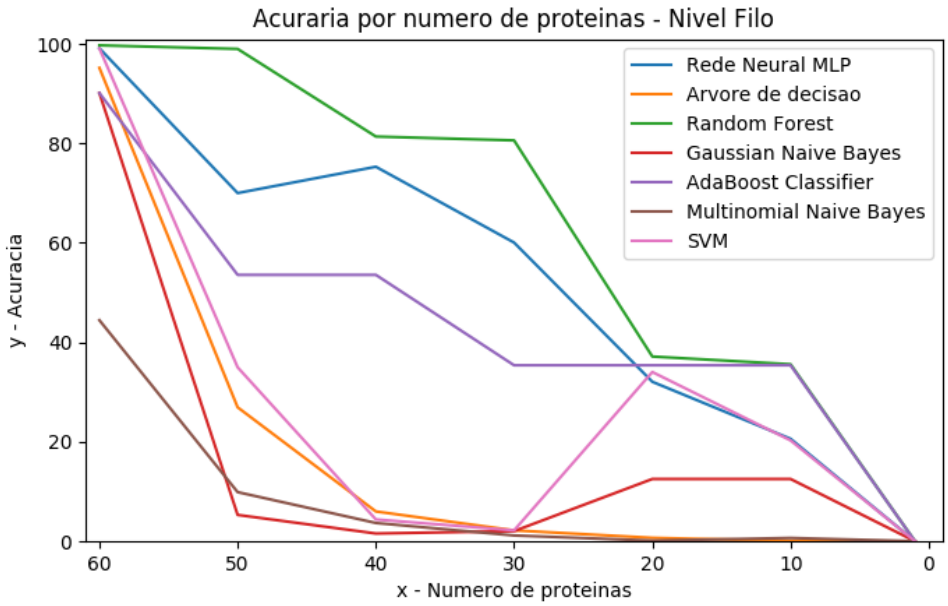
5.5 CLASSIFICAÇÃO COM BASE EM NÚMERO DE PROTEÍNAS

Tendo em vista que os classificadores Árvore de decisão (*Decision tree*) e Floresta randômica (*Random Forest*), que compõe o *ensemble*, são baseados em arvores de decisão para realizar a criação do modelo. Alguns testes com números variados de proteínas foram feitos a fim de verificar o impacto que estes classificadores sofreriam e, para expor os classificadores a uma situação real de classificação.

Pode-se observar na Figura 24 que o classificador Arvore de decisão teve seu desempenho afetado com base no numero de proteínas de entrada no momento da classificação. É

possível observar que o classificador menos afetado pela diminuição das proteínas é o Random Forest.

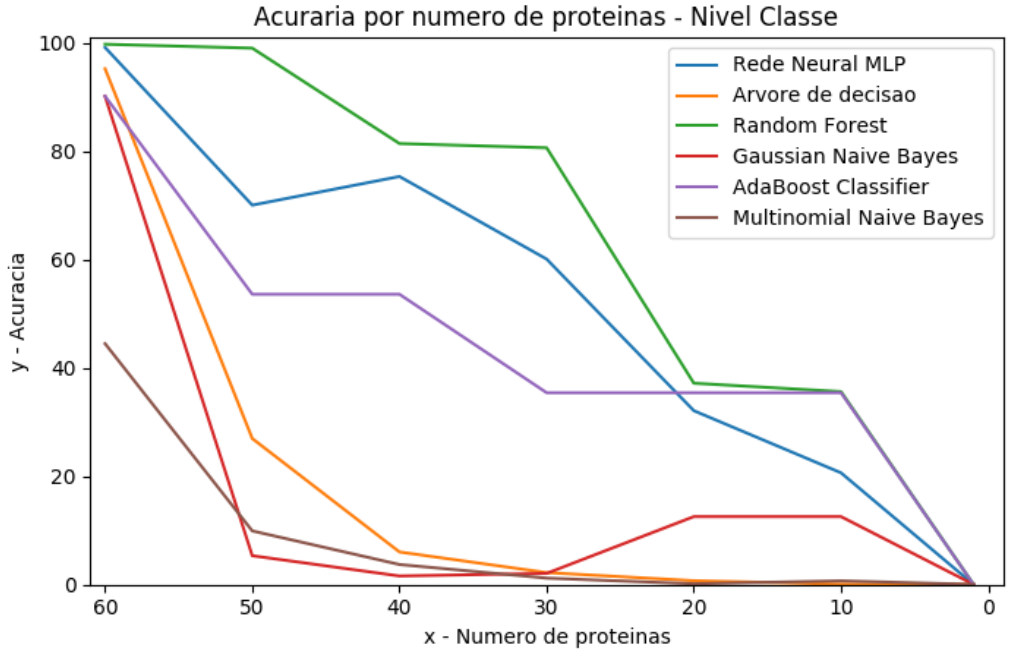
Figura 24: Desempenho classificadores por número de proteínas (com agrupamento) - Filo



Fonte: O autor.

As Figuras 25, 26, 27, 28 exibem o desempenho dos classificadores para os níveis de Classe, Ordem, Família e Gênero de acordo com o número de proteínas submetido ao classificador.

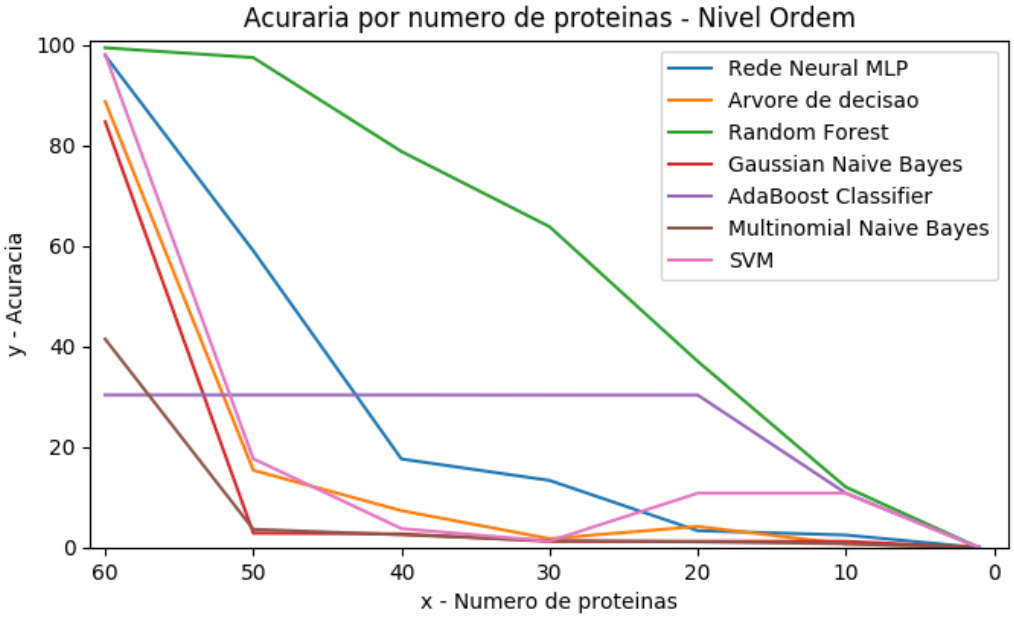
Figura 25: Desempenho classificadores por número de proteínas (com agrupamento)- Classe



Fonte: O autor.

Para o nível taxonômico de ordem, os classificadores tiveram impacto negativo em seu desempenho ao diminuir 10 proteínas, conforme pode ser observado na Figura 26. O classificador *Guassian Naive Bayes* foi o mais impactado nesse teste.

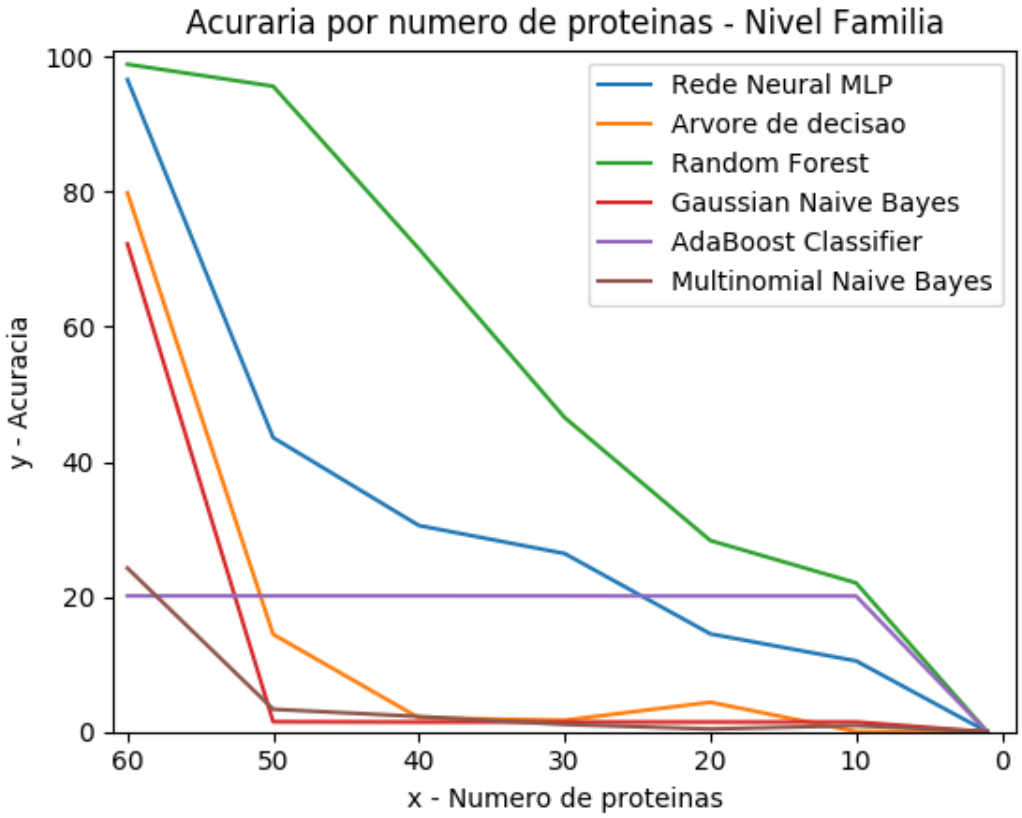
Figura 26: Desempenho classificadores por número de proteínas (com agrupamento) - Ordem



Fonte: O autor.

O classificador *Random Forest* possui modelos somente ate o nível taxonômico de Família. Como pode ser visto na Figura 27 este classificador é o menos impactado pelo numero de proteínas.

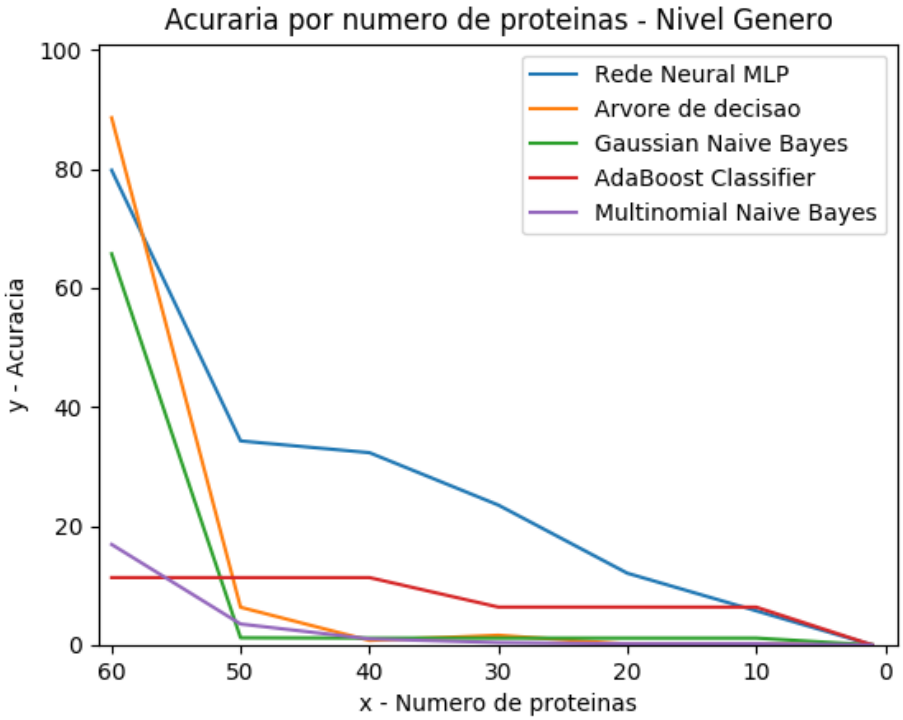
Figura 27: Desempenho classificadores por número de proteínas (com agrupamento) - Família



Fonte: O autor.

Na Figura 28 o desempenho do classificador *Random Forest* e SVM não aparecem, pois não possuem modelos para esse nível de taxonomia. O classificador que obteve menor queda com base no numero de proteínas foi o Adaboost, porém este classificador obteve a pior acurácia dentre os classificadores testados.

Figura 28: Desempenho classificadores por número de proteínas (com agrupamento) - Gênero

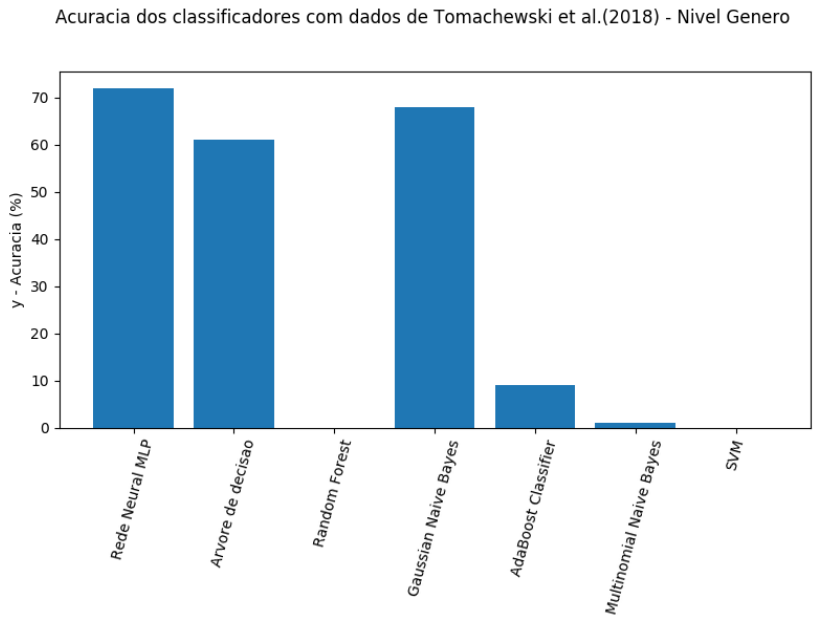


Fonte: O autor.

5.6 CLASSIFICAÇÃO EM OUTRAS BASES DE DADOS

Com o intuito de testar o modelo de classificação baseado em agrupamento, foram submetidos outras bases de dados de espectro de massa. A base desenvolvida por (TOMACHEWSKI *et al.*, 2018) em seu trabalho é composta de 1.949 gêneros, 28.505 registros e 60 proteínas os quais possuem dados faltantes. A Figura 29 mostra o desempenho dos classificadores, sem o uso do agrupamento, utilizando a base de Tomachewski.

Figura 29: Desempenho classificadores sem agrupamento com dados com faltantes

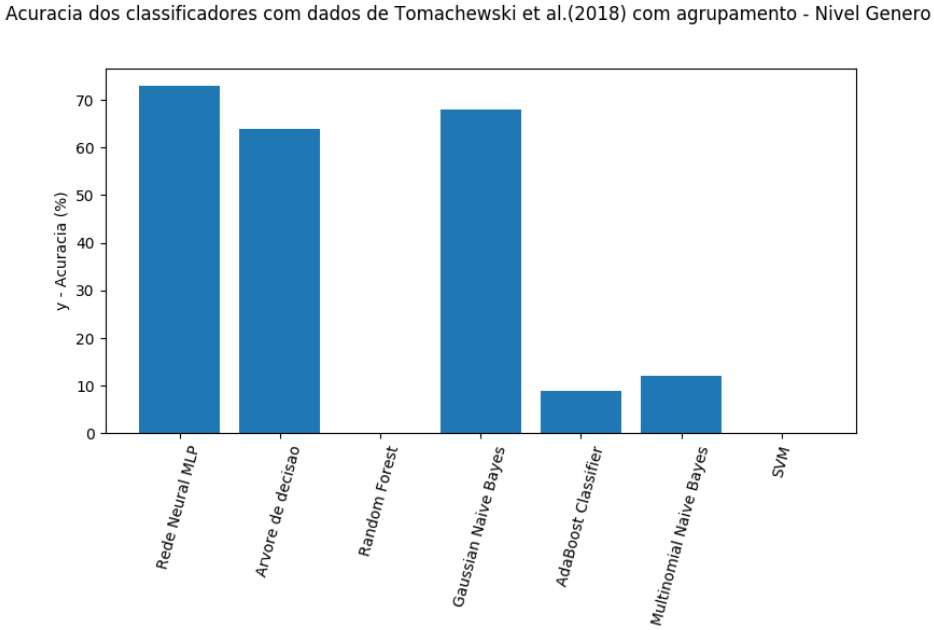


Fonte: O autor.

O classificador arvore de decisão obteve desempenho inferior a rede neural, devido ao fato da base de dados possuir dados faltantes sendo assim o desempenho desse classificador é afetado negativamente. Os classificadores SVM e *Random Forest* não possuem modelos para classificação a nível de gênero, por isso não apresentam resultados para este teste.

Utilizando o agrupamento os classificadores obtiveram ganho em seus resultados, como pode ser visto na Figura 30.

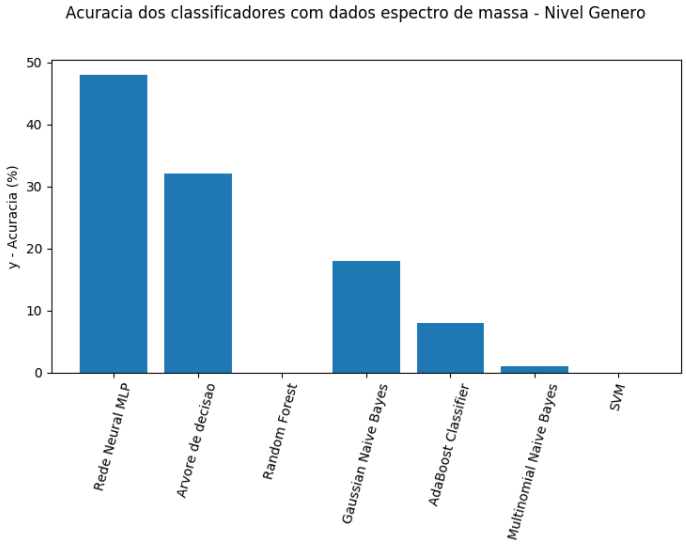
Figura 30: Desempenho classificadores com agrupamento com dados com faltantes



Fonte: O autor.

Por fim, foi utilizado uma base de dados de espectro de massa o qual possui 110 registros e 104 proteínas que também possui dados faltantes. A Figura 31 mostra o desempenho dos classificadores sem agrupamento utilizando esta base de dados. A rede neural *multilayer perceptron* obteve o melhor resultado, o classificador Arvore de decisão obteve desempenho inferior a rede neural. Os classificadores *Random forest* e SVM não possuem modelos para classificação deste nível taxonômico.

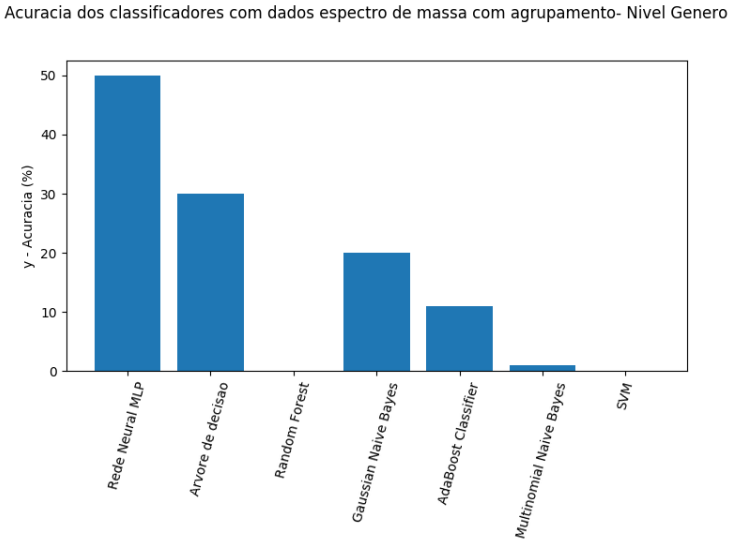
Figura 31: Desempenho classificadores sem agrupamento com dados faltantes



Fonte: O autor.

A Figura 32 exibe os resultados utilizando o agrupamento proposto, pode-se observar que a Rede neural *Multilayer Perceptron* obteve melhora em sua acurácia, o classificador arvore de decisão teve redução em seu desempenho.

Figura 32: Desempenho classificadores com agrupamento com dados com faltantes



Fonte: O autor.

6 CONCLUSÃO

Este trabalho avaliou diferentes classificadores e o efeito do agrupamento utilizando uma base de dados de massa/carga (m/z) de proteínas ribossomais criada a partir de dados genômicos completos, obtidos de um repositório público.

A revisão bibliográfica mostrou que a utilização de aprendizado de máquina para a classificação de bactérias tem obtidos resultados promissores. Contudo, dificuldades relacionadas aos dados biológicos, como dados faltantes e desbalanceamento são recorrentes. Entretanto, a criação de uma base de dados com dados de genomas e a utilização de um ensemble baseado em agrupamento é promissora para a classificação taxonômica.

O algoritmo SVM foi treinado utilizando os computadores de alto desempenho disponibilizados pela Universidade Estadual de Ponta Grossa e, como pode ser visto, o agrupamento aumentou o desempenho do classificador. Entretanto, devido a ser um método computacionalmente custoso não foi possível demonstrar resultados para todos os nível taxonômicos até o presente momento deste trabalho.

Os algoritmos que utilizam arvores de decisão como *Decision Tree* e *Random forest* obtiveram bons resultados nos teste realizados por este trabalho, porém para um caso real foi observado que os mesmos não tiveram bons resultados, devido ao fato que em classificações reais de bactérias, nem todas as proteínas são informadas. Caso alguma proteína chave para a árvore de decisão não seja informada o desempenho do classificador é prejudicado. Portanto pode-se concluir que em uma situação de classificação onde haja dados faltantes esse classificador seria o menos indicado.

Os resultados obtidos foram positivos, alcançando uma melhora de 9% no melhor caso, revelando que a metodologia elaborada no trabalho mostrou-se como uma alternativa na classificação de bactérias utilizando proteínas ribossomais oriundas de dados gnômicos.

Por fim, este trabalho resultou em um software capaz de construir um banco de dados de proteínas ribossomais oriundas de dados genômicos completos, treinar os classificadores e atualizar sua base de dados automaticamente.

6.1 TRABALHOS FUTUROS

Os conhecimentos obtidos por meio do desenvolvimento desta dissertação podem ser consideravelmente ampliados através da utilização de uma técnica eficaz de minimização do desbalanceamento da base de dados, e da ampliação da base de dados. A disponibilidade de uma base de dados menos desbalanceada viabiliza novas metodologias as quais podem reduzir o custo computacional na criação de novos modelos mais precisos.

Como trabalhos futuros propõe-se: Testar outros algoritmos de classificação e agrupamento na composição do ensemble. Gerar um modelo de classificação utilizando outro tipo de rede neural.

7 PUBLICAÇÕES RESULTANTES DA PESQUISA

Essa pesquisa gerou um registro de patente da base de dados.

Patente: Programa de Computador.

Número do registro: BR512019002529-6

Data de registro: 26/08/2019

Título: "Puchuy - Banco de Dados de Massa Molecular de Proteínas Ribossomais Baseado Em Genomas Bacterianos"

Instituição de registro: INPI - Instituto Nacional da Propriedade Industrial.

REFERÊNCIAS

ANDERSON, D.; BURNHAM, K. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, v. 63, 2004.

ATLAS, R.; BARTHA, R. Microbial evolution and biodiversity: the origins of life. *Microbial ecology: fundamentals and applications. 4th. ed. Menlo Park: Book News*, p. 37–39, 1997.

BARELLA, V. H. *Técnicas para o problema de dados desbalanceados em classificação hierárquica*. Tese (Doutorado) — Universidade de São Paulo, 2016.

BHOLOWALIA, P.; KUMAR, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, Citeseer, v. 105, n. 9, 2014.

BROOKS, G. F. *et al. Microbiologia Médica de Jawetz, Melnick & Adelberg-26*. [S.l.]: AMGH Editora, 2014.

BRUYNE, K. D. *et al.* Bacterial species identification from maldi-tof mass spectra through data analysis and machine learning. *Systematic and applied microbiology*, Elsevier, v. 34, n. 1, p. 20–29, 2011.

CARVALHO, A. *et al.* Inteligência artificial–uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, 2011.

CHAPELLE, O.; SCHÖLKOPF, B.; ZIEN, A. Semi-supervised learning mit press. *Massachusetts, USA*, 2006.

CLARRIDGE, J. E. Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews*, Am Soc Microbiol, v. 17, n. 4, p. 840–862, 2004.

COCK, P. J. *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, Oxford University Press, v. 25, n. 11, p. 1422–1423, 2009.

DEMIREV, P. A. *et al.* Microorganism identification by mass spectrometry and protein database searches. *Analytical chemistry*, ACS Publications, v. 71, n. 14, p. 2732–2738, 1999.

DUARTE, E. R.; CARELI, R. T.; SILVA, K. L. da. Classificação e evolução de microorganismos. *Microbiologia Básica para Ciências Agrárias*, p. 25, 2011.

ECKEL-PASSOW, J. *et al.* An insight into high-resolution mass-spectrometry data. *Biostatistics*, Oxford University Press, v. 10, n. 3, p. 481–500, 2009.

EISENBERG, E.; LEVANON, E. Y. Human housekeeping genes are compact. *TRENDS in Genetics*, Elsevier, v. 19, n. 7, p. 362–365, 2003.

ELBASHIR, M. K.; JIANXIN, W.; BINBIN, L. Multiple logistic regression model for beta-turns prediction. *Journal of Convergence Information Technology*, v. 6, n. 10, p. 173–180, 2011.

ELBASHIR, M. K.; JIANXIN, W.; WU, F. A hybrid approach of support vector machines with logistic regression for β-turn prediction. In: IEEE. *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. [S.l.], 2012. p. 587–593.

FACELI, K. *et al.* Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, 2011.

GALAR, M. *et al.* A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 42, n. 4, p. 463–484, 2011.

GANS, J.; WOLINSKY, M.; DUNBAR, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, American Association for the Advancement of Science, v. 309, n. 5739, p. 1387–1390, 2005.

GARCÍA, P. *et al.* Identificación bacteriana basada en el espectro de masas de proteínas: Una nueva mirada a la microbiología del siglo xxi. *Revista chilena de infectología*, Sociedad Chilena de Infectología, v. 29, n. 3, p. 263–272, 2012.

GHAEMI, R. *et al.* A survey: clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, v. 50, p. 636–645, 2009.

GIBAS, C.; JAMBECK, P.; FENTON, J. *Developing bioinformatics computer skills*. [S.l.]: "O'Reilly Media, Inc.", 2001.

GILLIS, M. *et al.* Polyphasic taxonomy. In: *Bergey's Manual® of systematic bacteriology*. [S.l.]: Springer, 2001. p. 43–48.

HARDIN, B.; MCCOOL, D. *BIM and construction management: proven tools, methods, and workflows*. [S.l.]: John Wiley & Sons, 2015.

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.

KAMPFER, P.; GLAESER, S. P. Prokaryotic taxonomy in the sequencing era and the role of mlsa in classification. *Microbiology Australia*, CSIRO PUBLISHING, v. 32, n. 2, p. 66–70, 2011.

KASSAMBARA, A. Determining the optimal number of clusters: 3 must know methods. *Available onli ne: https://www. datanovia. com/en/lessons/determiningthe-optimal-number-of-clusters-3-must-know-methods/.(accessed on 31 April 2018)*, 2017.

KURLAND, O.; KRIKON, E. The opposite of smoothing: a language model approach to ranking query-specific document clusters. *Journal of Artificial Intelligence Research*, v. 41, p. 367–395, 2011.

LABUSCHAGNE, N. Plant growth promoting rhizobacteria as biofertilizers. 2003.

LEE, J. *et al.* Svm classification model of similar bacteria species using negative marker: Based on matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. In: IEEE. *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. [S.l.], 2017. p. 145–150.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, JMLR. org, v. 18, n. 1, p. 559–563, 2017.

LIU, L. *et al.* An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, v. 5, n. 1, p. 1608, 2016.

LORENA, A. C.; GAMA, J.; FACELI, K. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. [S.l.]: Grupo Gen-LTC, 2000.

LU, J.; SALZBERG, S. Removing contaminants from metagenomic databases. *bioRxiv*, Cold Spring Harbor Laboratory, p. 261859, 2018.

LUDWIG, W. *et al.* Detection and in situ identification of representatives of a widely distributed new bacterial phylum. *FEMS Microbiology Letters*, Blackwell Publishing Ltd Oxford, UK, v. 153, n. 1, p. 181–190, 1997.

LUDWIG, W.; KLENK, H.-P. Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics. In: *Bergey's manual® of systematic bacteriology*. [S.l.]: Springer, 2005. p. 49–66.

MADHULATHA, T. S. Graph partitioning advance clustering technique. *arXiv preprint arXiv:1203.2002*, 2012.

MADHULATHA, T. S. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

MADIGAN, M. T. *et al. Brock: Biologia de los microorganismos.* [S.l.]: Pearson Educación., 2009.

MARTINS, M. P.; GUIMARÃES, L. N. F.; FONSECA, L. M. G. Classificador de texturas por redes neurais. In: *Anais do II Congresso Brasileiro de Computação, Itajaí-SC*. [S.l.: s.n.], 2002.

MEURER, E. C. *et al. Técnicas Modernas em Espectrometria de Massas: Aplicações Analíticas e no Estudo de Reações Íon/Molécula na Fase Gasosa*. Dissertação (Mestrado) — Universidade Estadual de Campinas, 2003.

OROZCO, A. *et al.* A review of bioinformatics training applied to research in molecular medicine, agriculture and biodiversity in costa rica and central america. *Briefings in bioinformatics*, Oxford University Press, v. 14, n. 5, p. 661–670, 2013.

PEARSON, W. R. [5] rapid and sensitive sequence comparison with fastp and fasta. Elsevier, 1990.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011.

POLIKAR, R. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, IEEE, v. 6, n. 3, p. 21–45, 2006.

REUNANEN, J. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, v. 3, n. Mar, p. 1371–1382, 2003.

RODRIGUEZ, R. M. *Estudo da Emissão de Íons Estáveis e Metaestáveis (LiF) nLi+ Induzida por Fragmentos de Fissão do 252 Cf*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, 2003.

ROKACH, L.; MAIMON, O. Clustering methods. In: *Data mining and knowledge discovery handbook*. [S.l.]: Springer, 2005. p. 321–352.

ROSSEL, S.; ARBIZU, P. M. Automatic specimen identification of harpacticoids (crustacea: Copepoda) using random forest and maldi-tof mass spectra, including a post hoc test for false positive discovery. *Methods in Ecology and Evolution*, Wiley Online Library, v. 9, n. 6, p. 1421–1434, 2018.

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. Universidade Estadual de Ponta Grossa, 2018.

SILVEIRA, É. L. d. Identificação de comunidades bacterianas de solo por seqüenciamento do gene 16s rrna. Universidade Estadual Paulista (UNESP), 2004.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introduction to data mining, pearson education. *Inc., New Delhi*, 2006.

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.

TOPCHY, A. P. *et al.* Analysis of consensus partition in cluster ensemble. In: IEEE. *Fourth IEEE International Conference on Data Mining (ICDM'04)*. [S.l.], 2004. p. 225–232.

VILLANUEVA, J. *et al.* Serum peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry. *Analytical chemistry*, ACS Publications, v. 76, n. 6, p. 1560–1570, 2004.

WATSON, J. D. *et al. Biologia molecular do gene*. [S.l.]: Artmed Editora, 2015.

WEI, L. *et al.* A novel hierarchical selective ensemble classifier with bioinformatics application. *Artificial intelligence in medicine*, Elsevier, v. 83, p. 82–90, 2017.

WILLARD, H. H. *et al.* Instrumental methods of analysis. 1988.

WOODS, K.; KEGELMEYER, W. P.; BOWYER, K. Combination of multiple classifiers using local accuracy estimates. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 19, n. 4, p. 405–410, 1997.

YADAV, K. K.; SARKAR, S. Biofertilizers, impact on soil fertility and crop productivity under sustainable agriculture. *Environ Ecol*, v. 37, n. 1, p. 89–93, 2019.

ZIEGLER, D. *et al.* Ribosomal protein biomarkers provide root nodule bacterial identification by maldi-tof ms. *Applied microbiology and biotechnology*, Springer, v. 99, n. 13, p. 5547–5562, 2015.

APÊNDICE A - LISTA DE FILOS

. .

Tabela 2: Quantidade de representantes por Filo

Quantidade	Filo
8271	Proteobacteria
3357	Firmicutes
1444	Actinobacteria
533	Bacteroidetes
321	Tenericutes
166	Chlamydiae
128	Cyanobacteria
110	Spirochaetes
61	Fusobacteria
52	Planctomycetes
41	Chloroflexi
40	Deinococcus-Thermus
37	Verrucomicrobia
36	Thermotogae
16	Aquificae
16	Chlorobi
11	Acidobacteria
9	Nitrospirae
5	Deferribacteres
5	Synergistetes
4	Elusimicrobia
4	Thermodesulfobacteria
4	unclassified
3	Gemmatimonadetes
2	Dictyoglomi
2	Fibrobacteres
2	Ignavibacteriae
1	Armatimonadetes
1	Balneolaeota
1	Caldiserica
1	Calditrichaeota
1	Chrysiogenetes
1	Coprothermobacterota
1	Kiritimatiellaeota
1	Lentisphaerae