Pós-Graduação em Ciência da Computação

RENAN WILLIAMS MARQUES FERREIRA

A Inteligência Artificial na Identificação de Espécies de *Candida*.

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

Recife
2020

RENAN WILLIAMS MARQUES FERREIRA

A Inteligência Artificial na Identificação de Espécies de *Candida*.

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

Área de Concentração: Inteligência Computacional

Orientador: Leandro Maciel Almeida

Recife
2020

Catalogação na fonte
Bibliotecária Mariana de Souza Alves CRB4-2105

F383i Ferreira, Renan Williams Marques
A Inteligência Artificial na Identificação de Espécies de *Candida* / Renan Williams Marques Ferreira. – 2020.
73f.: il., fig., tab.

Orientador: Leandro Maciel Almeida
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2020.
Inclui referências.

1. Inteligência Computacional. 2. Diagnóstico de *Candidemia*. 3. Narizes Eletrônicos. 4. Aprendizagem de Máquina. I. Almeida, Leandro Maciel. (orientador) II. Título.

006.31 CDD (22. ed.) UFPE-CCEN 2021-01

Renan Williams Marques Ferreira

“A Inteligência Artificial na Identificação de Espécies de Candida”

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

Aprovado em: 30 de janeiro de 2020.

BANCA EXAMINADORA

_____

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

_____

Prof. Dr. Reginaldo Gonçalves de Lima Neto
Centro de Ciências da Saúde/UFPE

_____

Prof. Dr. Leandro Maciel Almeida
Centro de Informática/UFPE
(**Orientador)**

Sem restrições

Dedico este trabalho às mulheres da minha vida: minha mãe e minha irmã.

AGRADECIMENTOS

Começo agradecendo a todos os professores que tive a imensa oportunidade de aprender a partir de seus conhecimentos. Em especial, sou grato ao professor Leandro Almeida por, desde o início desta pesquisa, sempre estar disponível para trocar conhecimentos. Aqui lhe exprimo a minha enorme gratidão.

Agradeço à Maria Conceição, responsável pelas amostras de Candida utilizadas neste estudo. Obrigado por se juntar a mim e ao professor Leandro nesta aventura.

À toda minha família, em particular, meus pais e meus irmãos. Vocês sabem o quanto tivemos que abdicar para que eu chegasse aqui. Mãe e Pai, muito obrigado por tudo que vocês fizeram, e continuam fazendo, para que seus filhos consigam alcançar suas melhores versões como pessoas e como profissionais. Dani e Bruno, obrigado por sempre acreditarem em mim, mesmo quando eu não acreditei, vocês são incríveis.

Aos meus amigos, aqueles de longa data e aqueles que entraram na minha vida através do mestrado. Aos integrantes do grupo Acidaz, vocês me inspiram cada dia a ser melhor, obrigado por me permitirem chamá-los de amigos. A Dinho, Fernando e Victor, vocês estiveram mais próximos durante todo esse processo, sabem o quão importante para mim é fechar mais este ciclo, serei eternamente grato por todo apoio recebido. À Chaina, Kecia e Deborah, vocês foram um dos maiores presentes que o Centro de Informática poderia me dar, obrigado por tudo, inclusive pelos momentos de descontração vividos no banquinho do pátio.

Enfim, agradeço de coração a todo mundo que torceu para que eu chegasse aqui. Essa vitória é nossa :)

RESUMO

Candidemia é a infecção da corrente sanguínea causada por leveduras do gênero *Candida*. Uma das maiores dificuldades no tratamento dessa doença é a demora no seu diagnóstico e na identificação do fungo causador da infecção. Implicando também na demora do tratamento correto e, portanto, tornando tardio a cura do paciente. Neste trabalho utilizamos a tecnologia dos narizes eletrônicos, que é um dispositivo que mimetiza o olfato humano, para avaliarmos seu desempenho na identificação de leveduras clínicas do gênero *Candida* através da amostra de ar obtida. A partir de uma parceria com o Departamento de Micologia da Universidade Federal de Pernammbuco, três bases de dados contendo amostras de *Candida albicans, C. parapsilosis* e *C. krusei* foram criadas. Um total de oito métodos de inteligência artifical foram avaliados. Nas três bases de teste utilizadas, foram obtidos os seguintes resultados em termos de acurácia: 96%, 88% e 90%. Considerando que este estudo tem por objetivo a desenvoltura de um possível novo método para a identificação de leveduras do gênero *Candida*, os resultados mostram que a utilização de um nariz eletrônico com alguns algoritmos de aprendizagem de máquina é promissora nesta área.

Palavras-chaves: Diagnóstico de *Candidemia*. Narizes Eletrônicos. Aprendizagem de Máquina.

ABSTRACT

Candidemia is a bloodstream infection caused by yeasts of the genus *Candida*. One of the biggest difficulties in the treatment of this disease is the delay in its diagnosis and in the identification of the fungus that causes the infection. It also implies the delay of the correct treatment and, therefore, delaying the cure of the patient. In this work we use the technology of electronic noses, which is a device that mimics the human sense of smell, to evaluate its performance in the identification of clinical yeasts of the genus *Candida* through the air sample obtained. From a partnership with the Mycology Department of the Federal University of Pernambuco, three databases containing samples of*Candida albicans, C. parapsilosis* and *C. krusei* were created. A total of eight artificial intelligence methods were evaluated. In the three test bases used, the following results were obtained in terms of accuracy: 96%, 88% and 90%. Considering that this study aims to develop a possible new method for the identification of yeasts of the genus *Candida*, the results show that the use of an electronic nose with some machine learning algorithms is promising in this area.

Keywords: Identification of Candida Species. Electronic Noses. Machine Learning.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CatBoost	*Categorical Boosting*
DT	*Decision Tree*
ENS-MLP	*Ensemble de Multilayer Perceptron*
GC-MS	*Gas Chromatography* com *Mass Spectrometry*
LR	*Logistic Regression*
MLP	*Multilayer Perceptron*
PCA	*Principal Components Analysis*
PCR	*Polymerase Chain Reaction*
RF	*Random Forest*
SVM	*Support Vector Machine*
XGBoost	*Extreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo introdutório será mostrado o contexto ao qual o presente trabalho está inserido e os objetivos que foram traçados.

1.1 PROBLEMA E MOTIVAÇÃO

A candidemia é uma infecção caracterizada pela presença do fungo *Candida* na corrente sanguínea. Infecções por leveduras desse tipo são frequentes nos pacientes críticos e poli-invadidos por instrumentos assistenciais como cateteres e sondas, principalmente em hospitais terciários onde a incidência das infecções fúngicas documentadas causadas pela *Candida* chega a ser 80%. E, devido a grande dificuldade do diagnóstico precoce e preciso, esse é um dos grandes desafios dos profissionais que atuam nessa área (Colombo e Guimarães 2003).

Para uma maior eficácia no tratamento, é de suma importância que o diagnóstico da infecção e a identificação do tipo do fungo que a causou sejam realizados de forma correta e rápida. Três abordagens laboratoriais clássicas que, historicamente, são bastante utilizadas para a realização do diagnóstico de infecções por *Candida*: microbiológica e imunológica (32). O isolamento do organismo no laboratório é uma das etapas principais de alguns métodos laboratoriais, porém além do isolamento ser uma etapa difícil de ser realizada, dependendo da cultura de sangue, a etapa de identificação do organismo ainda pode demorar dias, isso tudo devido a adaptação in vivo para in vitro necessário ao crescimento do fungo (Alexander 2002).

A alta mortalidade de pacientes diagnosticados com candidemia
(Colombo e Guimarães 2003), devido ao longo tempo gasto para identificação do tipo de candida que gerou a infecção, causou a necessidade da criação de outros métodos de identificação, que começaram a ser propostos com o objetivo de diminuir esse tempo mas, também, manter a precisão na identificação. Com isso, a administração terapêutica específica para o fungo identificado pode ser feita de maneira mais rápida e eficaz, aumentando assim as chances de cura dos pacientes.

Sistemas comerciais, como o *CHROMagar Candida system*, e métodos de diagnóstico molecular, como os métodos baseados em *Polymerase Chain Reaction* (PCR) são utilizados na identificação dessas espécies, porém o alto custo e a necessidade de um vasto conhecimento técnico para execução desses métodos limita seu uso (Deorukhkar e Shahriar 2018).

Narizes eletrônicos são dispositivos inspirados na capacidade de humanos e animais em reconhecer odores. Foram desenvolvidos com o objetivo de tornar menos custoso o processo industrial de avaliação de alguns produtos, como perfumes, alguns tipos de comidas e algumas bebidas (Gardner e Bartlett 1994). Antigamente, essas avaliações eram feitas por

humanos e, por isso, demandavam um investimento em treinamento desses profissionais. Porém, apesar de apresentarem um bom desempenho nessa tarefa, os mesmo não poderiam trabalhar por um longo período de tempo (Gardner e Bartlett 1994).

Desde seu desenvolvimento, os narizes eletrônicos têm sido bastante usados. Por exemplo, foram utilizados no estudo de Ping et al. 1997 para criação de um novo método não invasivo do diagnóstico de diabetes, apresentando resultados promissores, onde o método foi capaz de distinguir entre amostras do hálito de pessoas diabéticas e não diabéticas. Já Jonsson et al. 1997 utilizaram essa tecnologia para analisar diferentes amostras de grãos (aveia, centeio, cevada e trigo) com odores diferentes, diferentes níveis de ergosterol e Unidades Formadoras de Colônias (UFCs), com o objetivo de verificar se o dispositivo seria capaz de distinguir as amostras. E, apesar de ter apresentado um ótimo desempenho na distinção de amostras de aveia, na distinção dos outros grãos os resultados não foram promissores. Nos estudos de Olsson et al. 2000 foi verificada a possibilidade de usar um nariz eletrônico ou Espectrometria de Massas acoplada a Cromatografia a Gás para quantificar ergosterol e UFCs de algumas amostras de cevada naturalmente contaminadas. Das 40 amostras avaliadas pelo nariz eletrônico, apenas 3 foram classificadas erroneamente, já quando avaliadas por *Gas Chromatography* com *Mass Spectrometry* (GC-MS) esse número foi de 6 amostras.

1.2 OBJETIVOS

Desenvolver uma solução para a identificação rápida e simples de fungos utilizando um Nariz Eletrônico composto por sensores e métodos avançados de Inteligência Artificial. Analisar o desempenho desta tecnologia e verificar se esse sistema inteligente poderia auxiliar os profissionais que já trabalham na identificação das espécies desses fungos, a terem um diagnóstico mais rápido e tão preciso quanto os métodos já utilizados.

Com base nisto, alguns objetivos específicos foram traçados para chegarmos no objetivo principal, são eles:

- Desenvolver e aprimorar o sensor utilizado;
- Definir as espécies de *Candida* que serão utilizadas;
- Criar as bases de dados;
- Investigar alguns métodos de Inteligência Artifical;

1.3 ESTRUTURA DA DISSERTAÇÃO

Este trabalho está organizado em sete capítulos:

- Capítulo 1: apresenta a motivação deste trabalho, introduz o problema da Candidemia, além dos objetivos que também são apresentados neste capítulo.

- Capítulo 2: uma visão geral sobre a Candidemia e os métodos de identificação de espécies de *Candida* são apresentados.
- Capítulo 3: uma visão histórica e sobre a aplicabilidade e funcionalidade dos Narizes Eletrônicos é apresentada neste capítulo.
- Capítulo 4: todos os algoritmos que foram utilizados neste estudo são brevemente apresentados neste capítulo, além dos métodos de treinamento, validação e teste. As métricas que foram utilizadas também são apresentadas neste capítulo.
- Capítulo 5: todo o processo de coleta dos dados e construção das bases de dados utilizadas é apresentado neste capítulo.
- Capítulo 6: o esxperimentos realizados e seus respectivos resultados são apresentados neste capítulo.
- Capítulo 7: as conclusões acerca desta pesquisa são apresentadas, além de algumas sugestões para trabalhos futuros.

2 CANDIDEMIA

Neste capítulo, será mostrado um pouco sobre a candidemia, algumas espécies de *Candida*, além de alguns métodos que são utilizados para identificação dessas espécies atualmente.

2.1 CANDIDEMIA

Os fungos, assim como os animais, são organismos eucariontes, ou seja, possuem membrana nuclear. A obtenção de nutrientes para sua sobrevivência é realizada também a partir de outros organismos, hospedeiros, sendo que alguns fungos se hospedam em organismos vivos e outros em organismos mortos. Outros fungos, ao infectar um organismo vivo, matam as células vivas de seu hospedeiro para assim obter seus nutrientes (9).

Várias infecções em humanos podem ser causadas por fungos, entre as mais frequentes a dermatológica. Um dos principais fungos causadores dessas infecções é a *Candida*, que possuem várias espécies, mas apenas algumas são capazes de causar doenças em seus hospedeiros (9).

Candida está presente em aproximadamente 71% dos indivíduos considerados saudáveis, localizada em regiões como cavidade oral, trato gastrointestinal e na vagina. E, apesar de ser inofensiva em indivíduos saudáveis, esse fungo pode causar algumas infecções superficiais em hospedeiros com seu sistema imunológico afetado. Porém, em casos mais graves, a *Candida* pode entrar na corrente sanguínea de pacientes que passaram por algum procedimento em que dispositivos médicos foram introduzidos em seu organismo. Causando assim, a candidemia, que é a infecção da corrente sanguínea pelo fungo *Candida*. Essa infecção pode comprometer todos os órgãos internos, incluindo pulmões, rim, coração, fígado, baço e cérebro, trazendo um alto risco de vida ao paciente de acordo com Mavor et al. 2005 e Chandra et al. 2001.

Segundo ANVISA, os fatores reconhecidos de risco para infecção invasiva por *Candida* são:

- Permanência > 4 dias em UTI
- Antibioticoterapia de largo espectro
- Cirurgia abdominal
- Cateterização venosa central
- Nutrição parenteral total
- Imunodepressão
- Índice APACHE II > 10

- Ventilação mecânica > 48h
- Neutropenia
- Quimioterapia citotóxica

As espécies de *Candidas* são um dos patógenos principais responsáveis pelas infecções na corrente sanguínea em pacientes de todo o mundo (Morgan et al. 2005). No estudo de Li et al. 2016, foram identificados 190 pacientes com candidemia. A taxa de mortalidade desses pacientes hospitalizados foi de 27,9% num período de 30 dias após a coleta da amostra de sangue, sendo 16,7% com apenas 7 dias.

Dentre as mais de 200 espécies de *Candida* reconhecidas atualmente, a mais prevalente é a *Candida albicans* correspondendo a 50% de todos os casos detectados. Porém, nas últimas décadas houve um crescimento na identificação das chamadas *Candida* Não-*albicans*, e a identificação das mesmas foi se tornando cada vez mais importante pois algumas dessas espécies apresentam resistência a a terapia antifúngica como fluconazol, assim como a outros antifúngicos. Alguns dos tipos de *Candida* Não-*albicans* são: *Candida glabrata, Candida tropicalis, Candida parapsilosis, Candida krusei* (Quindós 2014).

Algumas características dessas *Candida* Não-*albicans* foram notadas. A *Candida parapsilosis*, por exemplo, vem frequentemente sendo identificada em indivíduos recém nascidos e também em indivíduos jovens adultos. Já a *Candida glabrata*, *Candida tropicalis* e *Candida krusei*, são bastante identificadas em pacientes com mais de 65 anos de idade. Além de que foram constatadas algumas diferenças na incidência geográfica dessas espécies (Quindós 2014).

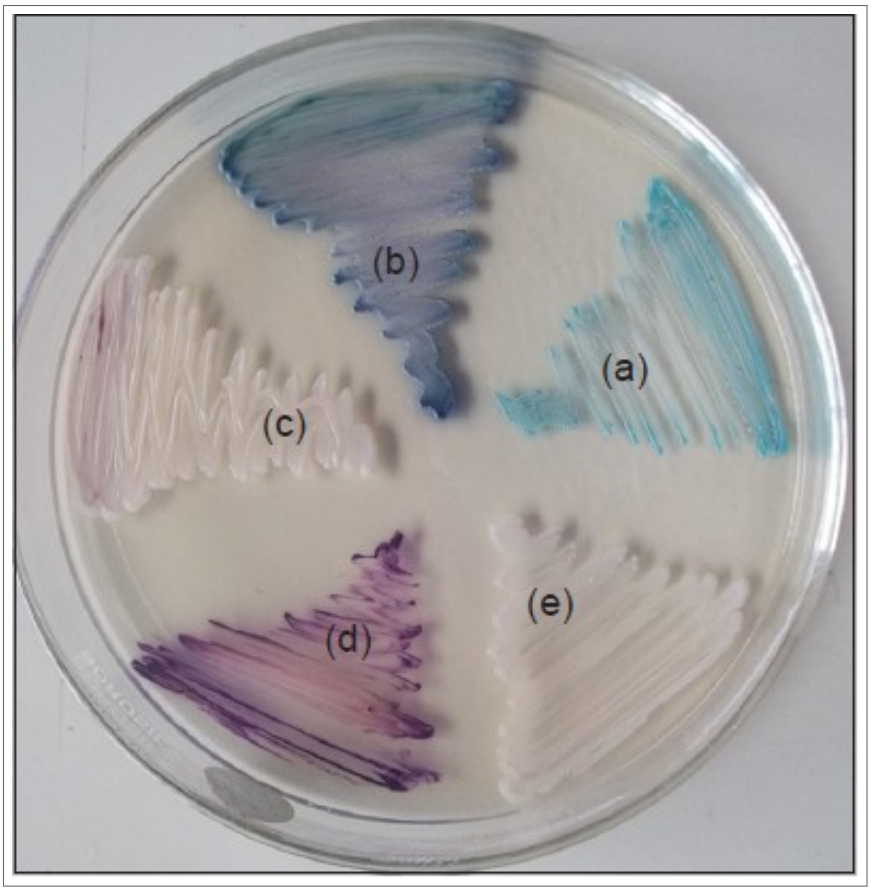
2.2 MÉTODOS DE IDENTIFICAÇÃO DE ESPÉCIES DE *CANDIDA*

Segundo (Raju e Rajappa 2011), a identificação de leveduras, como a *Candida*, pode ser feita utilizando alguns critérios, quando a identificação é baseada em meios de cultura, os critérios são: Morfológico, Fisiológico/Bioquímico.

Tradicionalmente, a primeira tentativa de identificação da espécie de *Candida Albicans* é realizada através do teste de tubo germinativo. E, apesar deste teste ser consideravelmente rápido, o crescimento das colônias em meios de cultura requer um tempo mínimo de 24 horas até 72 horas antes que o método para a identificação possa ocorrer (Sheppard et al. 2008).

A identificação das espécies de *Candida* através de meios cromogênicos como o CHROMagar, após hidrólise pela enzima correspondente, permite que colônias de *Candida albicans* fiquem com uma coloração verde, colônias de *Candida tropicalis* fiquem com uma coloração azul mas rodeada de uma auréola rosa, e colônias de *Candida krusei* fiquem com uma coloração rosa além de uma aparência felpuda (4). Uma identificação por esse método é mostrada na Figura 1.

Figura 1 – Identificação de *Candida albicans* (a), *C. tropicalis* (b)



Fonte: http://www.jlponline.org/text.asp?2014/6/1/28/129087.

Uma das técnicas moleculares existentes utilizadas para identificação dessas espécies é a PCR, que é um ensaio enzimático simples que permite a amplificação de um fragmento de DNA específico a partir de um conjunto complexo de DNA (DOES 2013). Com base na presença de regiões constantes e variáveis no rRNA, é possível projetar uma PCR para a identificação direta de várias espécies de *Candida* (Niesters et al. 1993).

Numa visão geral de métodos utilizados para identificação de espécies de *Candida* promovida por (Deorukhkar e Shahriar 2018), podemos verificar que a maioria dos diagnósticos laboratoriais ainda recorrem a métodos de identificação convencionais, como o teste de tubo germinativo, estudo da morfologia dos isolados e testes de fermentação e assimilação de carboidratos. Uma das conclusões deste mesmo estudo foi que, apesar do ótimo desempenho dos sistemas comerciais e métodos de diagnóstico molecular, o alto custo e a necessidade de um vasto conhecimento técnico para execução desses métodos limita seu uso.

3 NARIZES ELETRÔNICOS

Neste capítulo, será abordado um pouco sobre os Narizes Eletrônicos. Apresentaremos uma breve história sobre o surgimento desses dispositivos e também sobre sua funcionalidade. Além de algumas aplicações para resolução de problemas em diferentes áreas.

3.1 UM POUCO DE HISTÓRIA

A análise de odores para o diagnóstico de doenças é realizada desde 2000 a.C. e vem sendo aprimorada desde então. Detecção de doenças como câncer de pulmão, ovário, mama, bexiga, colorretal, melanoma, além de algumas infecções como infecção pulmonar e intestinal, vêm sendo diagnosticadas através da análise do cheiro do hálito, suor, urina e etc (Bomers e Smulders 2013). Tais doenças podem ser associadas a alterações no metabolismo de seus hospedeiros, acompanhados de diferentes compostos metabólicos, gerando assim um odor diferente (Bomers e Smulders 2013).

Na década de 1980, uma treinadora de cães começou a desconfiar de uma verruga em sua perna após um dos cães tentar mordê-la, o dermatologista consultado diagnosticou a verruga como um melanoma que é o tipo de câncer de pele mais perigoso. Desde então, alguns estudos têm agregado animais como uma técnica para detecção de odores nos diagnósticos de doenças (Bomers e Smulders 2013). Estudos afirmam que a utilização de ratos treinados, para detecção da bactéria Bacilo de Koch, é tão eficiente quanto a técnica de coloração Ziehl-Neelsen, porém com muito mais rapidez já que dois ratos conseguiram analisar cerca de 70 amostras em 32 minutos enquanto que clínicos laboratoriais são recomendados a analisar em média 20 amostras por dia (Mgode et al. 2012). Por mais que o desempenho de animais nessa área seja promissor, o uso dos mesmos não é tão convencional nesses estudos já que os animais necessitam de certo tipo de treinamento, o que demanda tempo e recursos financeiros.

O sentido do cheiro depende totalmente de células sensoriais especializadas localizadas no nariz para perceber compostos voláteis presentes no metabolismo do hospedeiro da doença (Bomers e Smulders 2013). O olfato é muito mais complicado que outros sentidos como, por exemplo, a visão e a audição, em relação aos mecanismos responsáveis pela reação primária a estímulos externos. Estima-se que seres humanos possam detectar de 10.000 a 100.000 substâncias químicas com odores diferentes. Uma explicação mais detalhada sobre como o cérebro humano processa as informações do cheiro é mais explícita em Linda Buck 2004 e 2005 [(Buck 2004), (Buck 2005)]. Uma tentativa de imitar os mecanismos de identificação do odor e os processos responsáveis por esse fenômeno torna-se viável graças à elaboração e aperfeiçoamento de dispositivos que compõem uma matriz de sensores, os chamados Narizes Eletrônicos (Gębicki e Kamysz 2017).

Apesar dos Narizes Eletrônicos terem sido criados na década de 60 do século vinte, o primeiro Nariz Eletrônico comercial formado por uma matriz inteligente de sensores capaz de classificar diferentes odores, só foi apresentado por Persaud e Dodd em 1982 (Persaud e Dodd 1982). Desde então, iniciou-se uma busca do melhoramento desses dispositivos dada a sua grande importância em processos como o diagnóstico de doenças.

3.2 FUNCIONALIDADE E APLICABILIDADE EM ALGUMAS ÁREAS

Narizes Eletrônicos são dispositivos que tentam imitar a estrutura do nariz humano, eles funcionam de uma maneira singular e têm, de certo modo, uma similaridade com o nariz humano (Gutiérrez e Horrillo 2014). Em geral, eles têm dois componentes principais: um sensor e um sistema de reconhecimento padrão.

O sensor é responsável por emitir os sinais, ao computador, ao entrar em contato com os compostos orgânicos voláteis presentes na amostra de ar no qual foi exposto. Uma variedade de sensores está disponível para uso em sistemas de nariz eletrônico. Conduzir polímeros oferece a vantagem de que eles são capazes de responder rapidamente e reversivelmente a temperatura ambiente. Eles são não específicos, mas podem ser altamente sensíveis, respondendo a uma variedade de compostos diferentes. A condutividade do polímero muda quando moléculas são absorvidas na superfície do sensor. Os sensores respondem a presença de álcoois, cetonas, ácidos graxos e ésteres, mas tiveram respostas reduzidas à espécies totalmente oxidadas, como CO2, NO2 e H2O (Magan e Evans 2000).

Já o sistema de reconhecimento de padrão geralmente apresentam dois estágios, um de aprendizado e outro de teste. No primeiro, a saída da matriz de sensores é treinada por um método de reconhecimento de padrão. No segundo estágio uma amostra de ar, onde sua classe pertencente é desconhecida, é testada e a previsão da classe é dada (Gardner e Bartlett 1994).

Zanchettin e colaboradores, em 2005, investigaram Redes Neurais Artificiais Híbridas para o melhoramento do sistema de reconhecimento padrão dos Narizes Eletrônicos, utilizando as que são mais consagradas na literatura para análise de odores, são elas: Multi-Layer Perceptron (MLP), Time Delay Neural Network (TDNN). Também foram utilizadas novas abordagens híbridas, tais como: o neuro-difuso, Feature Weighted Detector (FWD); além da Evolving Fuzzy Neural Network (EFuNN). Para efeito de comparação, foram utilizadas duas bases de dados, uma com cheiros de vinhos e outra com gases derivados do petróleo. Uma comparação entre MLP e TDNN mostrou que TDNN teve um desempenho melhor que MLP. Já entre as abordagens híbridas, EFuNN mostrou um resultado superior ao do FWD. Considerando um teste estatístico não paramétrico, com nível de significância de 5%, os melhores métodos para classificação de odores foram TDNN e EFuNN (Zanchettin e Ludermir 2005).

Dutta e colaboradores, também em 2005, analisaram alguns tipos de sistemas de reconhecimento padrão para classificar três tipos de bactérias: *Methicillin-Resistant Staphylo-*

coccus Aureus (MRSA); *Methicillin-Susceptible Staphylococcus Aureus* (MSSA); *Coagulase-Negative Staphylococci* (C-NS). As técnicas mais convencionais para o agrupamento de dados que foram utilizadas são: *Principal Component Analysis* (PCA), *Fuzzy C Means* (FCM), *Self-Organizing Map* (SOM). Uma nova abordagem também foi proposta, gerada a partir da combinação dessas três técnicas. Além disso, os três tipos de bactérias foram analisados usando três tipos de classificadores supervisionado *Artificial Neural Network*: *MultiLayer Perceptron* (ANN-MLP), *Probabilistic Neural Network* (PNN) e *Radial Basis Function Network* (RBF). Concluíram que o método combinando as três técnicas (PCA, FCM, SOM) pode resolver o problema de extração com dados muito complexos e melhorar o desempenho do Nariz Eletrônico utilizado (Cyranose 320). Por outro lado, utilizando as técnicas de classificadores supervisionados de ANN-MLP, PNN e RBF foram capazes de predizer os três tipos de bactérias com 78%, 96% e 99.69% de precisão, respectivamente (Dutta et al. 2005).

3.3 NARIZES ELETRÔNICOS NA DETECÇÃO DE DOENÇAS

De acordo com Turner e Magan 2004, o desenvolvimento dos narizes eletrônicos podem ter um papel significativo quando se trata de diagnóstico de doenças microbianas. E também afirmam que, com o uso da inteligência artificial, esses dispositivos são bastante úteis no monitoramento da epidemiologia de doenças, assim como podemos ver na revisão feita por Wojnowski et al. 2019.

Espécies microbianas produzem uma variedade de compostos voláteis, alguns dos quais possuem odores característicos, e dependendo do meio de cultura utilizado, assim como a idade da mesma, a quantidade de compostos gerados pode ser influenciada, podendo então esse ser um biomarcador de doenças (Wilson 2015).

Como os narizes eletrônicos são compostos por sensores que são sensíveis a específicos compostos orgânicos voláteis. E, podemos ver nos exemplos citados tanto no capítulo 1 na seção 1.1 e na seção anterior (3.2), os Narizes Eletrônicos apresentaram ótimos desempenhos em diferentes áreas. A aplicabilidade desses dispositivos nesse ambiente de identificação de espécies de *Candida* é factível.

4 ALGORITMOS DE INTELIGÊNCIA ARTIFICIAL UTILIZADOS

Neste capítulo, será abordado uma descrição de todos os algoritmos que foram utilizados nesse domínio de identificação de diferentes espécies de *Candida*.

4.1 INTELIGÊNCIA ARTIFICAL E APRENDIZAGEM DE MÁQUINA

Algumas atividades realizadas com facilidade por humanos, como por exemplo reconhecer uma pessoa através do seu rosto ou através da sua fala, não são tão facilmente realizadas por um programa de computador, pois não é trivial escrever um algoritmo ou pseudocódigo que consiga responder a perguntas como: 'Que características serão utilizadas?', 'O que fazer quando houver alguma mudança na aparência do rosto de uma pessoa ou na mudança de sua voz?'. Humanos conseguem realizar este tipo de tarefa utilizando reconhecimento de padrão, após analisarem vários exemplos de rostos e falas diferentes (Faceli et al. 2011).

Tarefas que exigem reconhecimento de padrão são executadas muitas vezes diariamente, e é de extrema importância ter esse reconhecimento feito de forma correta. Na medicina, por exemplo, um médico precisa analisar os sintomas de um paciente, conjuntamente com os resultados de alguns exames, para realizar o diagnóstico de alguma doença. Para executar tal atividade, foram necessários anos de estudo e prática para que este médico analisasse as informações de seu paciente e fizesse um diagnóstico preciso, possibilitando o tratamento da doença de forma correta (Faceli et al. 2011).

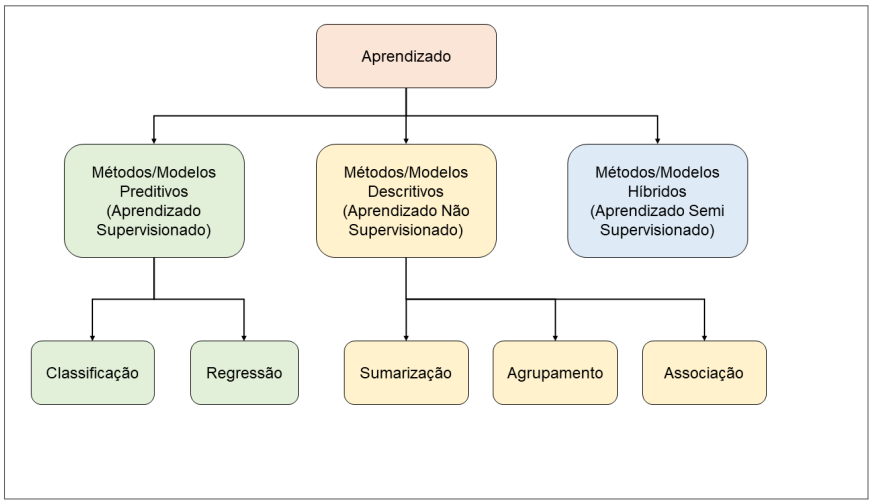
Segundo Andresen 2002, para Jhon McCarthy a Inteligência Artificial era a área responsável pela construção de máquinas inteligentes, principalmente programas de computador, utilizando dos mesmos para o entendimento e exploração da inteligência humana, porém, sem se limitar a fenômenos biologicamente observáveis.

Com o aumento de atividades que necessitam ser resolvidas computacionalmente e também com a enorme quantidade de dados que estão cada vez mais sendo gerados, a necessidade do implemento de métodos inteligentes que fossem capazes de realizar tais tarefas de forma autônoma, reduzindo a intervenção humana, também cresceu. Esses métodos deveriam ser capazes de criar por si próprias, a partir de experiências passadas, uma hipótese, ou função, capaz de resolver o problema que se deseja tratar. A esse processo dá-se o nome de Aprendizagem de Máquina (Faceli et al. 2011).

Sendo uma das áreas da Inteligência Artificial, Aprendizagem de Máquina é a área responsável pelo processo de aprendizado dos métodos inteligentes. Todo esse processo parte do princípio do reconhecimento de um certo padrão, ou algum tipo de comportamento previsível, que é possível de ser identificado através das características fornecidas pelos dados. O procedimento responsável pelo aprendizado automático desses padrões é chamado de modelo (15).

A Figura 2 apresenta uma hierarquia de aprendizado de acordo com os tipos de tarefas de aprendizado. Aprendizado Supervisionado ocorre sobre os dados com classes previamente definidas, possibilitando assim que uma avaliação do modelo utilizado possa ser realizada, pois conhecemos o rótulo desejado de cada um dos exemplos de entrada. Este aprendizado se divide em dois: Classificação e Regressão. Esta divisão é definida de acordo com os valores da variável de interesse, em casos de variável do tipo categórico, temos um problema de classificação, já em casos de variável do tipo numérico, temos um problema de regressão. Aprendizado Não Supervisionado ocorre quando os dados que serão utilizados não possuem classes associadas aos exemplos. Este aprendizado se divide em três: Sumarização, cujo objetivo é encontrar uma descrição simples e compacta dos dados; Associação, que tem por objetivo encontrar padrões frequentes de associação entre os atributos de um conjunto de dados; E, por último, Agrupamento, onde os dados são agrupados de acordo com sua similaridade. Aprendizado Semi Supervisionado ocorre quando alguns exemplos dos dados utilizados apresentam rótulo previamente, mas existem alguns exemplos que não possuem esse rótulo, então técnicas de Aprendizado Supervisionado e técnicas de Aprendizado Não Supervisionado são utilizadas para a resolução deste problema (Faceli et al. 2011).

Figura 2 – Hierarquia de Aprendizado



Fonte: https://www.embarcados.com.br/classificacao-multirrotulo-hierarquica-intro/

Neste trabalho, estamos no ambiente de Aprendizado Supervisionado com um problema de Classificação. E, portanto, foram utilizadas técnicas dessa área para a devida solução do problema. A seguir, na seção 4.2, será apresentada essas técnicas e uma breve descrição de cada uma delas.

4.2 ALGORITMOS

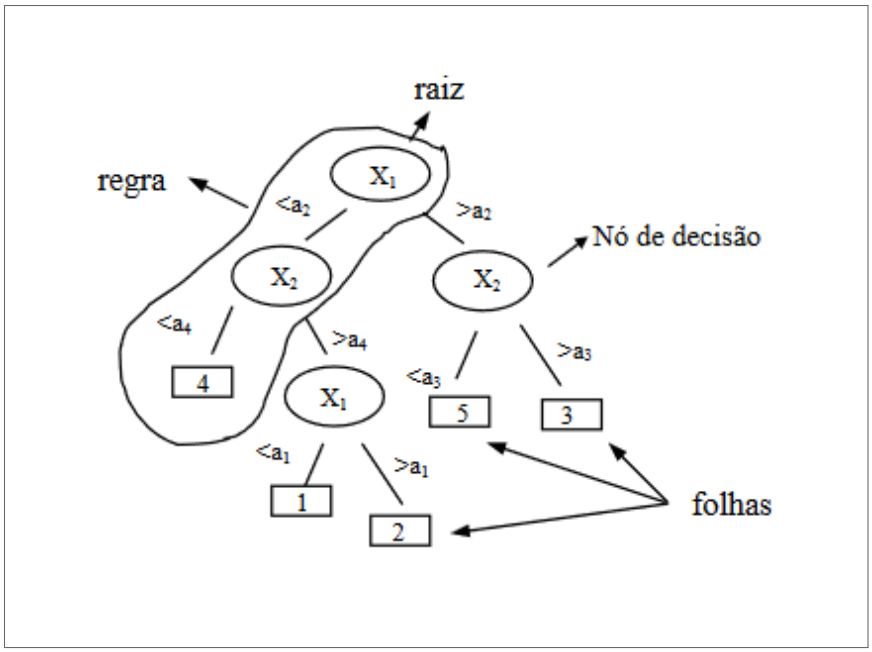
4.2.1 *Decision Tree*

Decision Tree (DT) são modelos estatísticos não paramétricos (conceito mais detalhado em 5.6) que utilizam um treinamento supervisionado para a classificação e previsão de dados. Estes modelos são baseados na estratégia de dividir para conquistar, ou seja, um problema mais difícil é dividido em problemas mais fáceis. Então, a solução desses problemas mais fáceis são combinadas em um formato de árvore e juntas resolvem o problema mais difícil (Faceli et al. 2011).

A divisão dos atributos é feita através do ganho de informação calculado através da entropia de uma variável, ou seja, através do nível de aleatoriedade da variável. A cada nó decisão, o atributo que mais reduz a aleatoriedade da variável alvo, será escolhido para dividir os dados (Faceli et al. 2011).

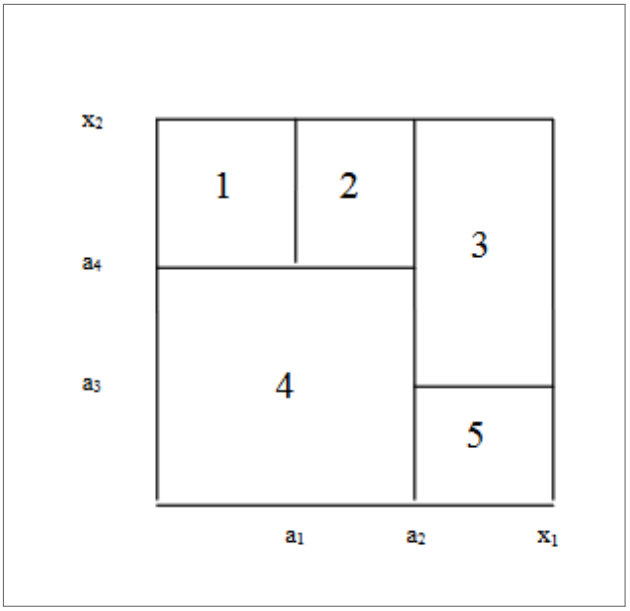
A Figura3 mostra uma representação de uma árvore de decisão e a Figura 4 representa a divisão no espaço definida por seus atributos $X1$ e $X2$. As folhas representam a decisão final para a instância avaliada, por exemplo, num problema de decisão sobre categorização de um novo cliente bancário, o nó folha é a categoria dada ao novo cliente podendo ser: Cliente de Alto Risco, Cliente de Médio Risco, Cliente de Baixo Risco. O nó decisão é a divisão de um atributo que foi selecionado, através do ganho de informação (explicado a seguir), no exemplo de categorização de um novo cliente, um dos nós de decisão poderia ser 'Faixa Salarial do Cliente' com as possíveis respostas: < 2 salários mínimos; $>= 2$ salários mínimos. Já o nó raiz, é o nó que apresenta o maior ganho de informação, neste exemplo poderia ser 'Cliente está negativado?' com possíveis respostas: Sim, Não. O percurso do nó raiz até uma folha, indicada na Figura 3, é chamado de regra de classificação.

Figura 3 – Decision Tree



Fonte: https://www.maxwell.vrac.puc-rio.br/7587/7587_4.PDF

Figura 4 – Divisão do Espaço de Decisão da Decision Tree da Figura 3



Fonte: https://www.maxwell.vrac.puc-rio.br/7587/7587_4.PDF

Algumas vantagens de utilizar *Decision Trees* para a classificação é:

- Flexibilidade: Esses modelos são não paramétricos. Ou seja, independente da distribuição dos dados, a aplicabilidade desses modelos é factível.
- Interpretabilidade: Ao final do treinamento desses modelos, nós temos como identificar quais atributos foram mais importantes para a classificação das instâncias testadas, tornando o entendimento das classificações mais fáceis.

- Eficiência: Esses modelos são algoritmos gulosos (sempre encontrando o ótimo global pois considera todas as features para cálculo do ganho de informação), realizando a seleção dos atributos mais eficientes para sempre otimizar o ganho de informação. Sua complexidade computacional cresce linearmente à medida que a quantidade de exemplos também cresce.

Algumas desvantagens de utilizar *Decision Trees* para a classificação são:

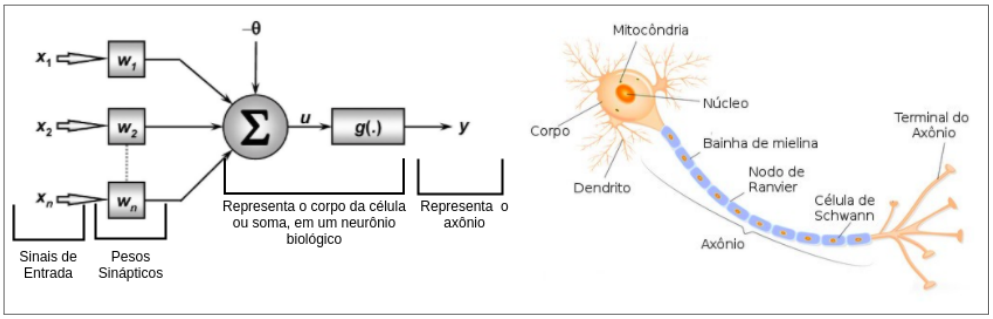
- Valores ausentes: Se o valor do atributo for um valor desconhecido, isso causa problemas em decidir que ramo seguir.
- Instabilidade: Pequenas mudanças no conjunto de treinamento podem produzir grandes variações na árvore final.

4.2.2 *Multilayer Perceptron*

Redes Neurais Artificiais são algoritmos desenvolvidos tomando como inspiração o cérebro humano além de sua estrutura e funcionalidade do sistema nervoso. Os neurônios são células nervosas e a unidade fundamental do sistema nervoso. Eles respondem a estímulos internos e externos possibilitando assim a transmissão de impulsos nervosos a outros neurônios, às células musculares e glandulares (Faceli et al. 2011).

Na Figura 5 é mostrada uma comparação entre um neurônio artificial e um neurônio biológico. Os sinais de entradas e os pesos recebem um valor cada e simulam os dendritos do neurônio biológico. Os valores recebido são ponderados e combinados por uma função matemática, imitando a função do corpo do neurônio biológico. E, por fim, a saída da rede representa o axônio (Faceli et al. 2011).

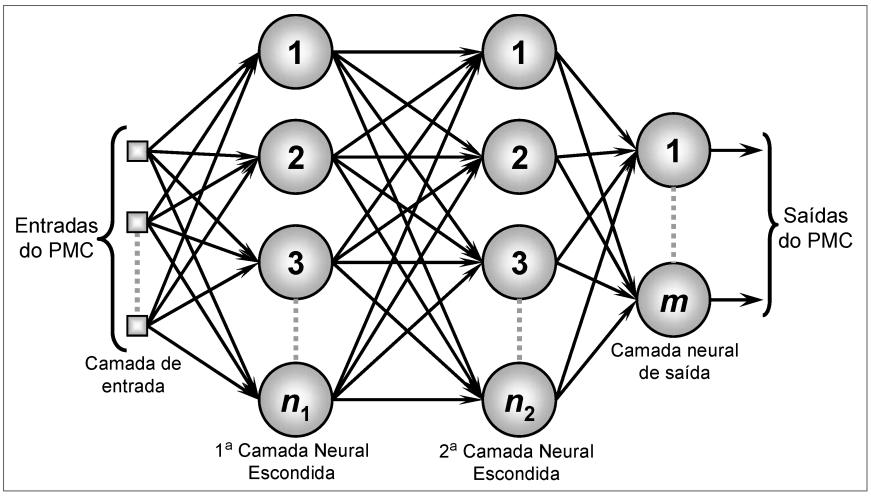
Figura 5 – Comparação entre o neurônio artificial e o neurônio biológico.



Fonte: https://medium.com/@avinicius.adorno/redes-neurais-artificiais-418a34ea1a39

Com objetivo de resolver problemas não linearmente separáveis, as *Multilayer Perceptron* (MLP) foram desenvolvidas e sua estrutura pode ser vista na Figura 5. Como podemos ver, essas redes podem apresentar uma ou mais camadas escondidas além da camada de entrada e saída. E, para resolver problemas não linearmente separáveis, é adicionada uma função não linear entre as camadas escondidas, transformando os sinais da camada anterior para a camada seguinte (Faceli et al. 2011).

Figura 6 – Estrutura de uma Rede Neural Artificial *Multilayer Perceptron*.



Fonte:
https://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9

Essas redes são treinadas com base no algoritmo *back-propagation*, que consiste em duas partes:

- *Forward*: É onde acontece todo o processo inicial de onde as informações da camada de entrada passam por toda a rede neural até chegarem na camada de saída, obtendo seu valor.
- *Backward*: Nessa fase, o valor obtido na camada de saída na fase forward é comparado com o valor esperado (por ser um problema supervisionado, nós sabemos essa informação). Então, o valor do erro obtido a partir dessa comparação é utilizado para a atualização dos pesos das camadas anteriores. Esse ajuste ocorre de trás para frente, ou seja, da camada de saída até a primeira camada escondida.

As atualizações do algoritmo *back-propagation* são feitas até que algum critério de parada seja atendido. Esses critério podem ser, por exemplo, a quantidade de ciclos a serem realizados ou uma taxa máxima de erro permitida (Faceli et al. 2011).

Algumas vantagens de utilizar essas redes neurais são:

- Por ter um pode Generalização e Tolerância a ruídos, esses modelos são muito úteis quando os dados apresentam algumas informações faltantes.

Algumas desvantagens de utilizar essas redes neurais são:

- Por possuir muitos parâmetros, dependendo da quantidade de exemplos utilizados para o treinamento desses modelos, o custo computacional pode ser inviável.
- Outra desvantagem bastante discutida é seu funcionamento interno, pois é composto de muitas fórmulas matemáticas, por isso a interpretabilidade de seus resultados não é trivial.

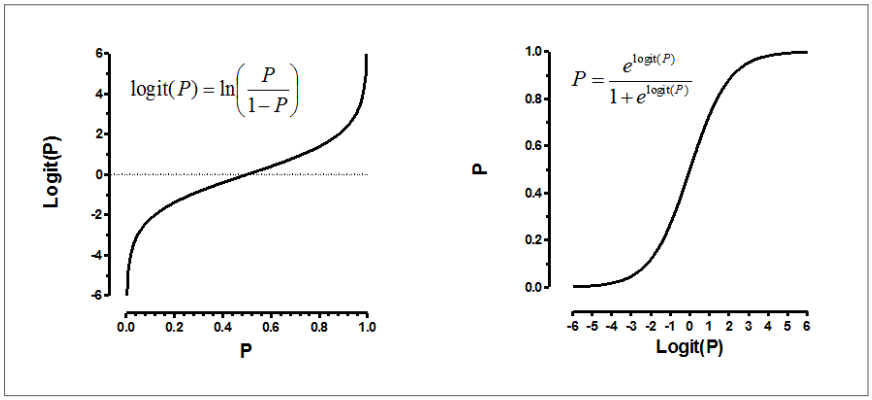
4.2.3 *Logistic Regression*

Logistic Regression (LR) é uma modelo estatístico capaz de modelar a probabilidade de um evento ocorrer, a partir de um conjunto de dados. Semelhante ao modelo de regressão linear, porém nesse caso a distribuição da variável alvo é Bernoulli, ou seja, a variável alvo é binária e só assume dois valores (0 ou 1; Sim ou Não). Porém, pode também ser utilizada em problemas com mais de duas categorias, nesse caso, por exemplo, o algoritmo pode aplicar o esquema Um-Versos-Resto transformando um problema com três categorias em três problemas com duas categorias cada, podendo assim calcular a probabilidade de cada categoria (GONZALEZ 2018).

Ao contrário do modelo de regressão linear, aqui os parâmetros do modelos são estimados a partir do conjunto de dados disponível através do método de estimação de máxima verossimilhança. Este método consiste em maximizar a verossimilhança com relação aos parâmetros de interesse (GONZALEZ 2018).

A Figura 7 apresenta a função logit, função utilizada para o cálculo das probabilidades de cada exemplo pertencer a certa categoria (GONZALEZ 2018).

Figura 7 – Função Logit.



Fonte: https://www.analyticsvidhya.com/blog/2015/10/basics-logistic-regression/

Algumas vantagens da *Logistic Regression*:

- Fornece os resultados em termos de probabilidade.
- Alto grau de confiabilidade por considerar a distribuição da variável alvo.

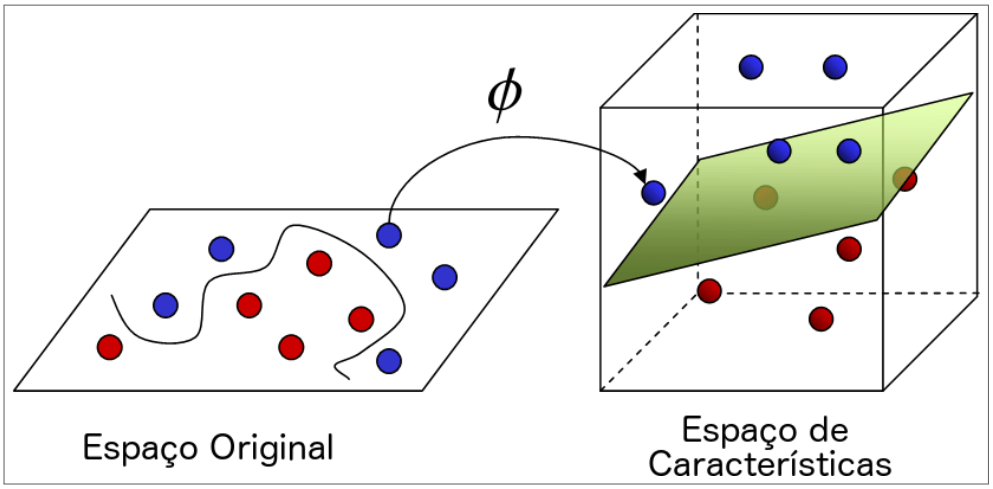
Desvantagens desse métodos:

- Interpretação do modelo também não é tão trivial por apresentar muitos cáculos matemáticos.
- Em problemas com muitos atributos, variáveis independentes, esse método necessita de mais exemplos para realizar uma estimação precisa dos parâmetros do modelo.

4.2.4 *Support Vector Machine*

Os modelos *Support Vector Machine* (SVM) são algoritmos inicialmente desenvolvido para solucionar problemas linearmente separáveis. Dado o espaço dimensional do problema, é traçado um hiperplano de modo que otimize a distância entre os indivíduos de classes diferentes. Porém, mais tarde, essa abordagem foi estendida para definir fronteiras lineares para problemas mais gerais. Ou seja, em casos de problemas não linearmente separáveis, uma função é aplicada nos atributos dos dados disponíveis de modo a tentar transformar esse problema em um problema linearmente separável em outra dimensão. Essa função é chamada de Função Kernel (Faceli et al. 2011).

Figura 8 – Aplicação da Função Kernel Φ para mudança do espaço dimensional e possível separação linear dos dados.



Fonte: https://www.researchgate.net/figure/Figura-215-Classicacao-perfeita-pelo-hiperplano-otimo-do-SVM-com-kernel-nao-linear_fig1_318598388

O hiperplano gerado na Figura 16, após a aplicação da função kernel Φ, foi capaz de realizar a separação linear dos dados. E, apesar de existirem muitos hiperplanos entre as regiões de separação dos indivíduos de bola azul e os indivíduos de bola vermelha, o hiperplano selecionado será aquele que maximize a distância entre as instâncias. Portanto, é de suma importância a escolha de uma função kernel, em casos de problemas não linearmente separáveis.

Algumas vantagens dos *Support Vector Machine*:

- Apresentam uma boa capacidade de generalização.
- Robustos diante de problemas com alta dimensionalidade.

Algumas desvantagens desses modelos:

- Alta sensibilidade a escolha dos parâmetros.
- Complexidade de interpretação dos seus resultados.

4.2.5 *Random Forest*

Random Forest (RF) são algoritmos baseados em Ensembles, que são técnicas que utilizam a expertise de mais de um modelo para a realização da tomada de decisão ao classificar uma instância, com o objetivo de ter uma melhor performance do que utilizando apenas um modelo. A ideia é, por exemplo, obter a opinião de mais de um 'especialista' de modo a tomar uma decisão acerca de um dado problema com mais confiança. Por exemplo, é comum pessoas consultarem mais de um médico para obterem informações mais confiáveis sobre um determinado diagnóstico. Porém, nesse caso, os especialistas são os modelos (Faceli et al. 2011).

Existem ensembles homogêneos, que são um conjunto de classificadores/modelos gerados pelo mesmo algoritmo, e os heterogêneos, que são um conjunto de classificadores/modelos gerados a partir de algoritmos diferentes. *A Random Forest* é um ensemble homogêneo gerado por *Decision Trees* utilizando a técnica de geração de ensemble chamada *bagging* (Faceli et al. 2011) e uma demonstração é apresentada na Figura 9.

O *bagging* é realizado da seguinte forma (Faceli et al. 2011), tendo um conjunto de dados disponíveis T:

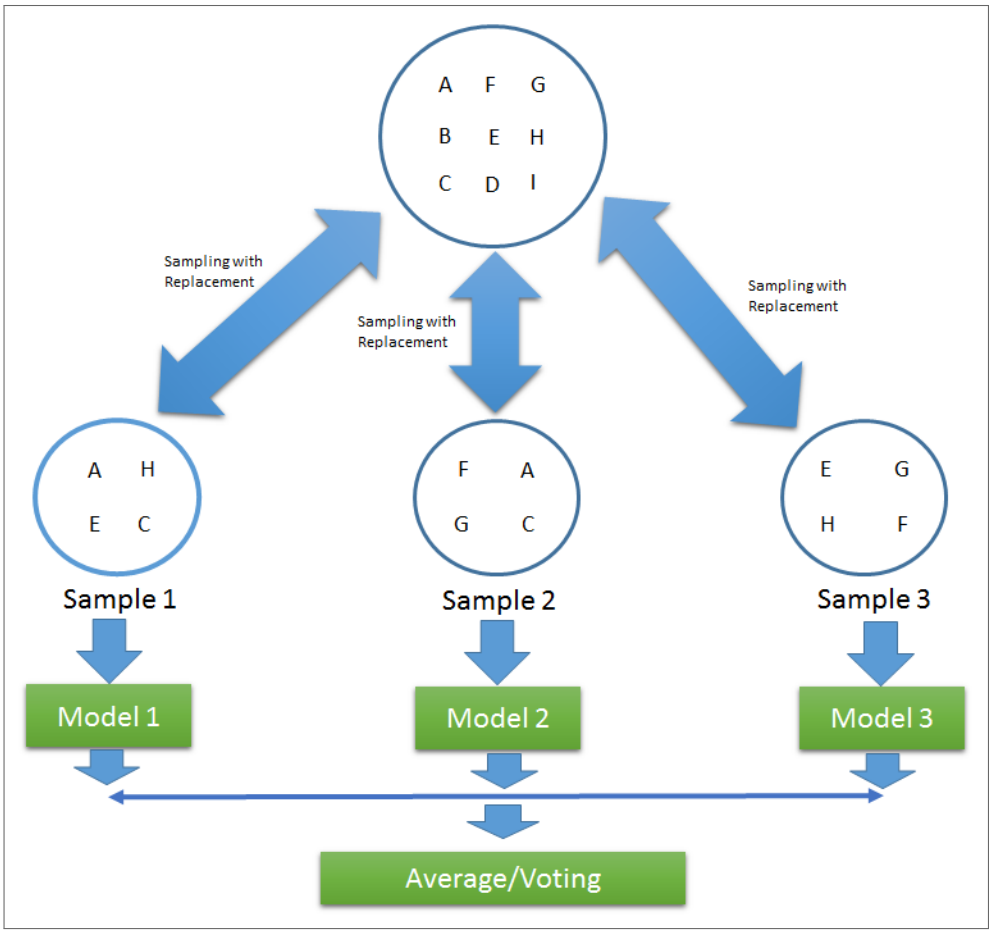
1. Realizar n amostras COM REPOSIÇÃO do conjunto de dados T, sendo n o número de Decision Trees desejado para criação da *Random Forest*.

2. Treinar n Decision Trees nos n diferentes conjuntos selecionados na etapa anterior.

3. Após o treinamento das n Decision Trees, submeter os exemplos não utilizados na fase de treinamento para sua devida classificação as n Decision Trees.

4. Combinar o resultado das n Decision Trees através de um dos métodos de combinação de classificadores.

Figura 9 – Demonstração do método de geração de ensemble Bagging.



Fonte: https://subscription.packtpub.com/book/big_data_and_b usiness_intelligence/9781788830577/3/ch03lvl1sec34/bagging

A combinação dos classificadores normalmente é feita através do voto majoritário, que funciona da seguinte maneira: Se em um problema de classificação categórica temos 3 classes, os classificadores do ensemble gerado irão cada um classificar uma instância de teste baseado em seu treinamento, gerando assim possíveis divergência entre os classificadores na classificação da instância. Portanto, a instância de teste submetida à classificação receberá a classe mais votada entre os classificadores. Por exemplo, temos 3 classes, um ensemble contendo 5 classificadores (C1, C2, C3, C4 e C5), para uma instância de teste os classificadores C1, C2 e C3 atribuíram a classe 1 e os classificadores C4 e C5 atribuíram a classe 2. Pela combinação do voto majoritário, a instância de teste receberá a classe 1 (Faceli et al. 2011).

Algumas vantagens deste algoritmo, além de permitir método efetivo de substituição de valores ausentes, este método apresenta bons resultados em bases de dados desbalanceadas. Já uma desvantagem do mesmo é a dificuldade de interpretação dos resultados.

4.2.6 *Ensemble* de *Multilayer Perceptron*

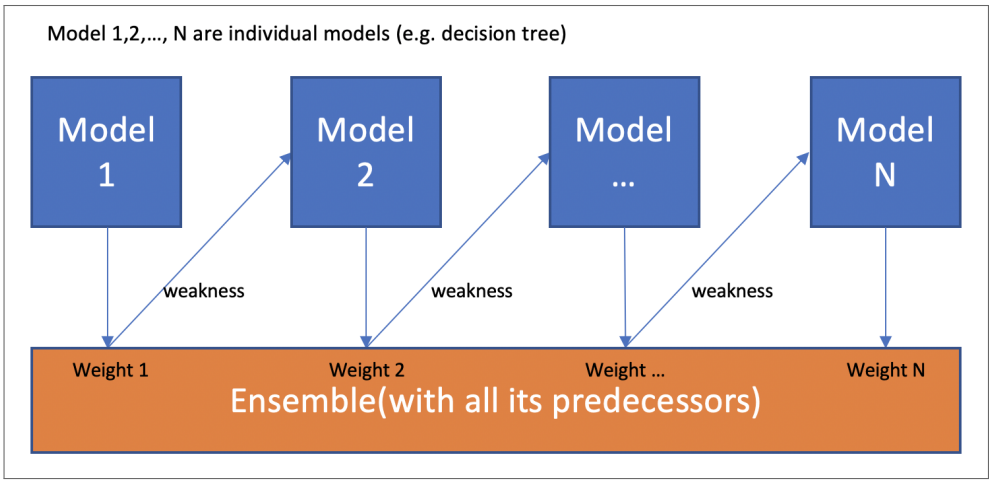
Ensemble de Multilayer Perceptron (ENS-MLP) é mais um algoritmo de ensemble homogêneo. Também foi desenvolvido utilizando a técnica de geração de ensemble *bagging*, explicada na seção 4.2.5, porém ao invés de utilizarmos o algoritmo *Decision Tree* (4.2.1), aplicamos o algoritmo *Multilayer Perceptron* explicado na seção 4.2.2.

4.2.7 *Extreme Gradient Boosting*

Extreme Gradient Boosting (XGBoost) é uma biblioteca de software onde foi implementada o Gradient Boosting Trees com um foco mais em engenharia, promovendo assim a melhora de performance e a rapidez desse algoritmo. Ultimamente sendo bastante utilizados nas competições do *Kaggle* por promover aos seus usuários um poder computacional superior quando comparado com bibliotecas no *Python*, R e *Spark*
(Benchmarking Random Forest Implementations 2015).

Boosting é uma das técnicas de geração de ensembles onde novos modelos são adicionados ao conjunto para corrigir os erros dos modelos já existentes no ensemble, por exemplo, as amostras selecionadas para o Model 2, na Figura 10, são aleatorimente selecionadas com reposição, porém, os exemplos em que o Model 1 errou sua classe, recebem uma probabilidade maior de serem selecionados para serem treinadas/avaliadas pelo Model 2, com o objetivo que o Model 2 consiga acertar os exemplos errados pelo Model 1. Os modelos são adicionados sequencialmente até que nenhuma melhora no desempenho possa mais ser realizada. Já o *Gradient Boosting* é uma abordagem do *Boosting* onde novos modelos são criados para prever os erros dos modelos já existentes no ensemble, e então são adicionados ao ensemble para fazerem a predição final de uma instância de teste. São chamados de *Gradient Boosting* pois utilizam o algoritmo *Gradient Descent* para minimizar a perda quando novos modelos forem adicionados. E, por fim, *Gradient Boosting Tree* é o *Gradient Boosting* utilizando *Decision Trees* como os algoritmos de formação do ensemble (Faceli et al. 2011).

Figura 10 – Demonstração do método de geração de ensemble *Boosting*.



Fonte: https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30

Uma das maiores vantagens do XGBoost é que foi implementado para que a execução do mesmo apresentasse uma rapidez maior do que quando implementado em outras bibliotecas de software. Já uma desvantagem desse método é que os dados utilizados para realização de experimentos com o XGBoost precisam caber na memória do dispositivo que será utilizado. Em casos de bases de dados gigantescas será necessário um dispositivo com mais memória.

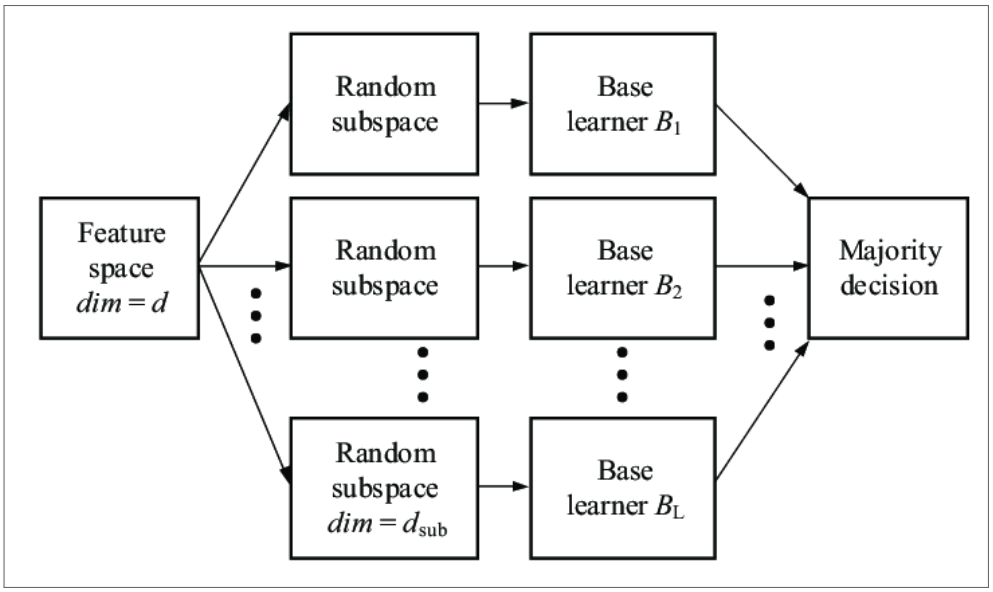
4.2.8 *Categorical Boosting*

Muito similar ao XGBoost, o *Categorical Boosting* (CatBoost) é uma biblioteca de software que foi criada para lidar mais facilmente com atributos categóricos. Ou seja, quando a base de dados apresenta algum atributo que seja categórico, por exemplo o atributo 'Altura' medido em Alto, Médio e Baixo. Essa biblioteca facilita o processamento desses dados transformando esses valores em valores numéricos sem que seja necessário um pré-processamento nesses atributos, apenas o indicativo de quais features são categóricas (CatBoostAI 2017).

A biblioteca do CatBoost, apesar de ser focada também no método de geração de ensemble Boosting, explicado na seção 4.2.7 e ilustrado na Figura 10, fornece a possibilidade de utilizar a técnica Random Subspace como método de geração de ensemble (CatBoostAI 2017).

O *Random Subspace*, também conhecido como Feature Bagging, é exatamente a aplicação do método Bagging porém ao invés de ser retiradas amostras com reposição dos exemplos de treinamento, nesse método as amostras são retiradas com reposição a partir do conjunto de atributos, a Figura 11 apresenta uma demonstração de como o método funciona. E, o CatBoost utiliza também o algoritmo Decision Tree para geração de um ensemble homogêneo (CatBoostAI 2017).

Figura 11 – Demonstração do método de geração de ensemble *Random Subspace*.



Fonte: https://www.researchgate.net/figure/Scheme-of-Ensemble-Learning-Random-subspaces-are-obtained-from-feature-space-dim-and$_fig1_3$17828745

Uma das vantagens desse algoritmo é, quando trabalhado com dados categóricos, a não preocupação com a realização de enconding em atributos categóricos. E uma das desvantagens é que o mesmo não funciona para atributos com valores não existentes na base de dados utilizada.

5 MATERIAIS E MÉTODOS

Neste capítulo será mostrado todo processo até conseguirmos os dados finais utilizados no presente estudo.

5.1 PRIMEIROS PASSOS

Inicialmente vários testes, sem ser com fungos, foram realizados para fins de estabilização do sensor utilizado.

Os primeiros testes com fungos foram realizados no CETENE no dia 19/09/2018, onde tivemos a oportunidade de coletar amostras de fungos de três placas diferentes: uma com a identificação (por outros meios) já realizada; outra sem identificação; e uma última que apresentava quatro tipos de fungos na mesma placa.

Em testes anteriores, foi constatado que é necessário um tempo para a estabilização do sinal do sensor, porém, nesses primeiros testes com fungos no laboratório do CETENE, o tempo de estabilização do sinal foi inferior (10 a 20 minutos) ao tempo levado em outros experimentos (de 1 hora a 1 hora e 30 minutos). Uma possível explicação para este acontecimento, vem do fato que os experimentos feitos anteriormente foram realizados em um laboratório de química e o sensor não estava acoplado a um filtro de carvão, possibilitando assim a entrada de várias substâncias contidas na atmosfera do laboratório para dentro do sensor.

Após a estabilização do sinal emitido pelo sensor, foi iniciada a coleta das amostras dos fungos contidos nas placas de cultura. As primeiras amostras coletadas serviram para construção de um protocolo a ser seguido durante o processo de coleta de dados desses fungos. Esse processo foi necessário pois foi a primeira vez em que estávamos monitorando o sinal do sensor enquanto o mesmo entrava em contato com amostras reais de fungos. Tendo em vista que esse protocolo afeta diretamente o desempenho do sensor em termos de captação dos compostos voláteis presentes nas amostras de ar de cada um dos fungos.

Analisando as primeiras amostras coletadas, o seguinte protocolo foi utilizado:

1. Abertura parcial da placa com fungos para a coleta dos compostos voláteis com filtro acoplado à seringa;
2. Coleta de 10ml do ar contido dentro da placa;
3. Permanência de um filtro acoplado à seringa durante todo o processo de coleta;
4. Monitoramento e coleta do sinal emitido pelo sensor por 1 minuto após a injeção;
5. Purga (limpeza) do sensor com a sucção de ar interno e filtro de carvão ativado na entrada da câmara por aproximadamente 1 minuto;

6. Após isso, uma nova coleta de amostra é realizada.

As seringas utilizadas para coletar as amostras foram de materiais descartáveis, cada uma com capacidade máxima de sucção de 10 mililitros (ml), e uma vez utilizada a mesma era descartada. Os filtros que foram acoplados nas seringas são de PFTE (Polytetrafluoroethylene) / TF (Teflon) 0.22, sendo um único filtro usado na coleta de cinco amostras da mesma placa. Os indivíduos que realizaram a coleta desses fungos fizeram uso de jaleco e luvas de proteção, desde o início dos experimentos até o seu fim.

5.2 DADOS UTILIZADOS NO ESTUDO

Para a realização desse projeto, uma parceria com o departamento de Micologia da Universidade Federal de Pernambuco foi feita, de modo que conseguíssemos a obtenção dos fungos.

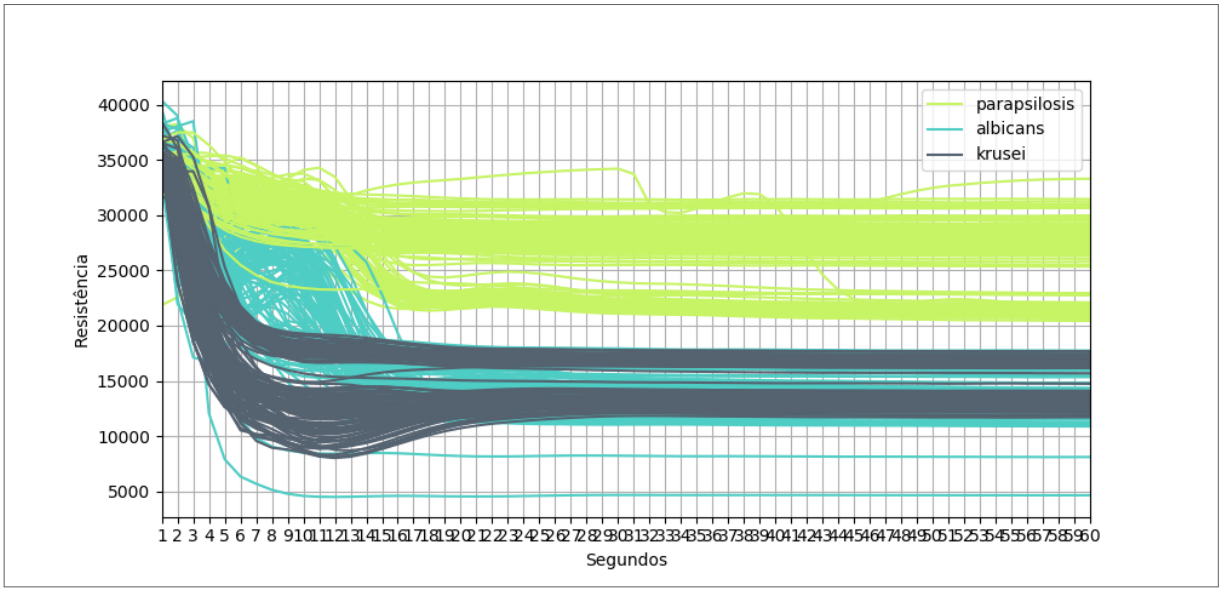
Após análise das primeiras amostras de fungos, algumas decisões foram tomadas: Foram fixados três diferentes tipos de *Candida*, a *albicans*, a *parapsilosis* e a *krusei*. Foram analisadas 10 leveduras de cada uma das espécies para a coleta de dados. Tendo como objetivo verificar o quão bem um sistema inteligente consegue identificar características únicas de cada um desses tipos, de modo que quando receber uma instância nunca vista após o período de treinamento, tenha um poder de generalização relativamente bom.

Três bases de dados foram construídas para a realização desta pesquisa: Base com amostras consideradas Padrões; Base com amostras consideradas Não-Padrões; Base obtida a partir da junção das duas bases com amostras Padrões e Não-Padrões.

Amostras Padrões são amostras que são tomadas como base para identificação de outras amostras. Existe todo um estudo para que uma amostra seja considerada padrão. Para cada espécie de *Candida* existe uma amostra padrão, para que dela outras amostras possam ser identificadas também. Nós utilizamos, nessa base, amostras Padrões reconhecidas mundialmente das espécies *albicans*, *parapisilosis* e *krusei*.

1. Amostras consideradas Padrões;

Figura 12 – Curvas das amostras consideradas Padrões.

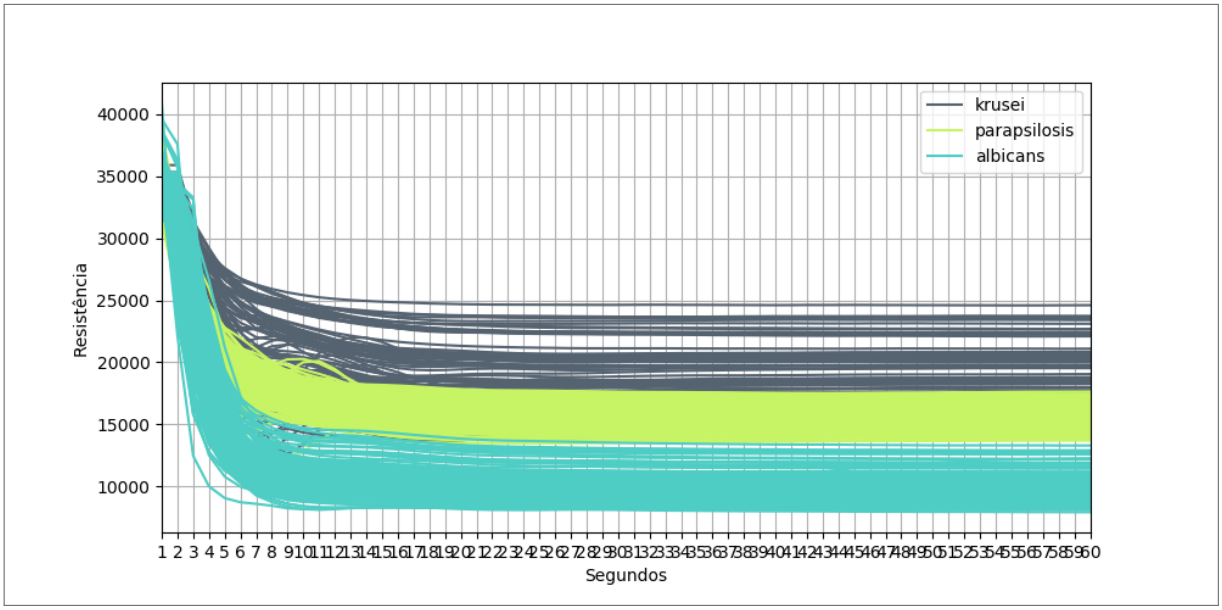


Fonte: Elaborada pelo autor.

Na Figura 12 podem ser vistas as curvas das amostras coletadas na base das amostas Padrões, onde foram obtidas 122 do tipo *albicans*, 122 do tipo *krusei* e 123 do tipo *parapsilosis*.

2. Amostras consideradas Não-Padrões;

Figura 13 – Curvas das amostras Não-Padrões.

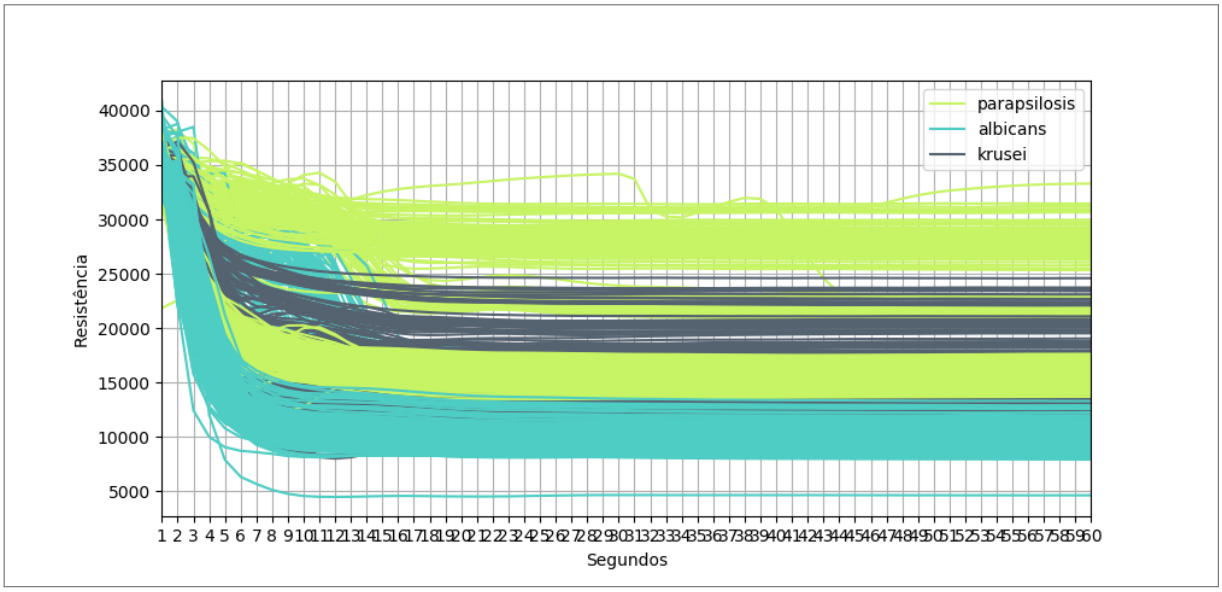


Fonte: Elaborada pelo autor.

Na Figura 13 podem ser vistas as curvas das amostras coletadas na base das amostas Não-Padrões, onde foram obtidas 201 do tipo *albicans*, 120 do tipo *krusei* e 198 do tipo *parapsilosis*.

3. A junção das amostras Padrões com as amostras Não-Padrões.

Figura 14 – Curvas das amostras Padrões e Não-Padrões juntas.



Fonte: Elaborada pelo autor.

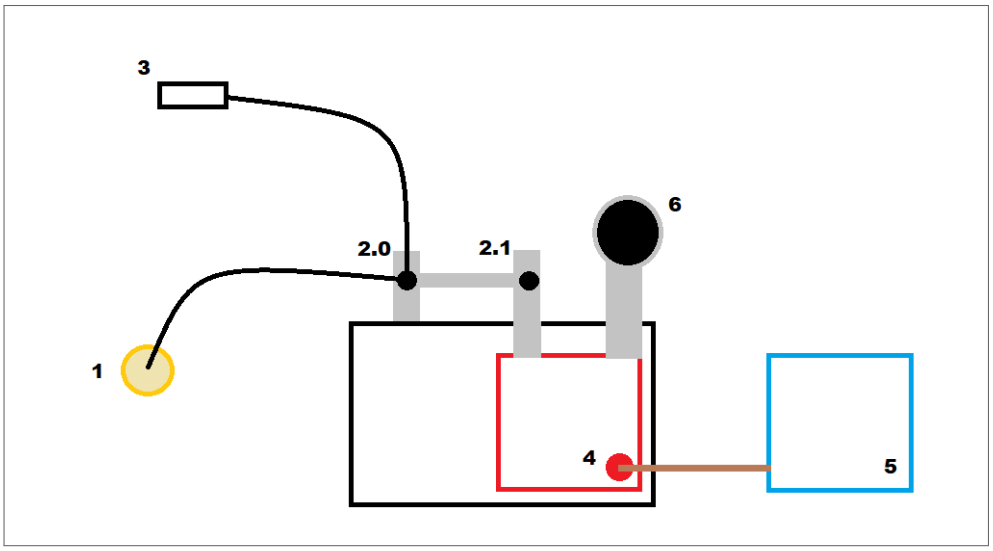
Na Figura 14 podem ser vistas as curvas das amostras coletadas na base das amostas Padrões e Não-Padrões conjuntamente, onde foram obtidas 323 amostras de *Albicans*, 242 amostras de *Krusei* e 321 amostras de *Parapsilosis*.

5.3 NARIZ ELETRÔNICO UTILIZADO NESTE ESTUDO

A Figura 15 mostra um protótipo do Nariz Eletrônico utilizado neste estudo. Uma explicação de cada um dos intens contidos na Figura 15 é apresentada a seguir:

- O item **1** indica a amostra de *Candida*.
- Os itens **2.0** e **2.1** são duas válvulas de abertura/fechadura de passagem da amostra de ar.
- O item **3** é uma seringa que é acoplada a um filtro de teflon.
- O item **4** é o sensor utilizado que na presença de um gás detectável, a condutividade do sensor aumenta dependendo da concentração de gás no ar. Um circuito elétrico simples converte a mudança de condutividade em um sinal de saída que corresponde à concentração de gás.
- O item **5** é o computador utilizado.
- O item **6** é uma bomba conectada a um filtro de carvão ativado.

Figura 15 – Protótipo do Nariz Eletrônico utilizado neste estudo.



Fonte: Elaborado pelo autor.

Inicialmente, a válvula **2.0** é aberta enquanto que a válvula **2.1** permanece fechada, possibilitando que a seringa (**3**), que é acoplada a um filtro de teflon, sugue o ar presente na amostra de *Candida*. Em seguida, a válvula **2.0** é fechada e a **2.1** é aberta, permitindo agora que a seringa possa injetar o ar da amostra de *Candida* para a câmara onde se encontra o sensor (**4**). Durante todo o tempo de exposição do sensor (**4**) ao ar da amostra de *Candida*, sinais são emitidos ao computador (**5**), que capta esses dados para preenchimento da base de dados. Após um minuto de exposição, a bomba (**6**), acoplada a um filtro de carvão ativado, é ligada para realizar a limpeza (por um minuto) do ar contido na câmara onde o sensor está. Após a realização desta limpeza, o processo é realizado novamente para a captura de uma nova amostra de ar. Importante destacar também que todo o processo de coleta de dados foi realizado em um local limpo e com temperatura controlada.

5.4 TREINAMENTO, VALIDAÇÃO E TESTE DOS ALGORITMOS

Para a realização do processo de treinamento, validação e teste dos algoritmos utilizados é comumente separado 3 diferentes conjuntos, não necessáriamente divididos em 70%, 20% e 10%, mas para este estudo essas porcentagens foram utilizadas para divisão do conjunto de dados original. O primeiro 70% que é o conjunto de treinamento, pois é o conjunto que necessita de mais exemplos, pois é nele em que os algoritmos conseguem aprender os padrões baseados nos atributos para conseguir prever uma nova instância nunca vista antes. O conjunto de validação corresponde a 10%, nesse conjunto é onde será realizada a otimização dos parâmetros dos classificadores, ou seja, os parâmetros que retornarem um melhor desempenho no conjunto de validação serão os parâmetros utilizados que os algoritmos irão utilizar para prever as instâncias nunca vistas. O conjunto de teste

corresponde a 20% e essas instâncias não podem participar das fases de treinamento nem de validação, pois é nelas que iremos verificar se o desempenho dos algoritmos obtidos nas fases de treinamento e de validação se repetem na base de teste, para verificarmos se os algoritmos apresentam um bom poder de generalização. As divisões dos três conjuntos é realizada de um modo estratificado, de modo que a distribuição de exemplos das classes do problema se repita em cada um desses conjuntos (Faceli et al. 2011).

5.5 MÉTRICAS PARA AVALIAÇÃO

Para avaliarmos o desempenho de cada um dos classificadores, algumas métricas foram utilizadas. E são baseadas nos valores da matriz de confusão a seguir:

Figura 16 – Matriz de Confusão

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: https://medium.com/@vitorborbarodrigues/

- Acurácia: Métrica que verifica o percentual de acertos dentre todos os exemplos avaliados. Ou seja, se 30 exemplos são submetidos à avaliação, a acurácia de um classificador C1 é a razão do número de exemplos em que C1 acertou a classe, supomos 26, sobre o número total de exemplos, nesse caso a acurácia de C1 é 0.87 ou 87% (Faceli et al. 2011). A partir da matriz de confusão, a acurácia pode ser obtida da seguinte forma:

$$\frac{(VP + VN)}{(VP + VN + FN + FP)}$$

- Precisão: Proporção de exemplos positivos classificados corretamente entre aqueles preditos como positivos por C1 (Faceli et al. 2011). A partir da matriz de confusão a precisão pode ser calculada da seguinte forma:

$$\frac{(VP)}{(VP + FP)}$$

- Recall: Corresponde à taxa de acerto na classe positiva (Faceli et al. 2011). A partir da matriz de confusão pode ser obtida da seguinte forma

$$\frac{(VP)}{(VP + FN)}$$

5.6 TESTES ESTATÍSTICOS APLICADOS

Neste estudo foram aplicados alguns testes estatísticos, alguns paramétricos e outros não-paramétricos. Testes paramétricos são teste que são restritos a quando conhecemos a distribuição dos dados que estamos utilizando, por ter essa restrição, esses testes são considerados mais robustos que os testes não-paramétricos. Já os testes não-paramétricos não necessitam do conhecimento da distribuição dos dados para que possam ser executados, por esse motivo estes testes são menos robustos que os testes paramétricos. Porém, a aplicabilidade de testes não-paramétricos são mais comuns que a dos testes paramétricos, porque nem sempre conseguimos definir a distribuição dos dados que estamos utilizando (REIS e Junior 2007).

Para verificarmos se o vetor de desempenho dos classificadores se distribuem de acordo com uma distribuição normal, aplicamos o teste de Shapiro-Wilk neste vetor. Este teste avalia a hipótese nula de que os dados provém de uma distribuição normal, ou seja, a não rejeição desta hipótese nos confirma que os dados se distribuem de acordo com uma distribuição normal (Razali et al. 2011).

Posteriormente, utilizamos dois testes, um para analisarmos o desempenho dos classificadores em grupo, e caso hovesse a comprovação de que há uma diferença estatisticamente significativa entre o desempenho dos classificadores, nós aplicávamos um teste no vetor do desempenho dos classificadores dois a dois, para identificarmos qual dos classificadores apresentou desempenho estatisticamente diferente.

Em caso de não rejeição da hipótese nula do teste de Shapiro-Wilk, ou seja, os dados seguem uma distribuição normal. Nós podemos utilizar técnicas paramétricas no vetor de desempenho dos classificafores. Para a análise de grupo foi utilizado o teste de ANOVA para medidas repetidas, que testa a hipótese nula de que todos os classificadores apresentam o mesmo desempenho médio (Martinez e Ferreira 2007). Em caso de reijeição desta hipótese, ou seja, os classificadores apresentam desempenhos estatisticamente diferentes, então aplicamos o Teste T pareado.

O Teste T pareado testa a hipótese nula de que dois classificadores apresentam desempenho médio iguais (Kim 2015). Em caso de rejeição desta hipótese, os classificadores apresentam desempenho médio estatisticamente diferentes, e podemos identificar qual dos dois tem o melhor desempenho.

Em caso de rejeição da hipótese nula do teste de Shapiro-Wilk, nós podemos utilizar técnicas não-paramétricas para a realização da comparação do desempenho dos classificadores. O teste de Friedman é equivalente ao teste ANOVA, porém sem a restrição de que os dados pertençam a uma distribuição normal (Zimmerman e Zumbo 1993). Aqui, as mesmas conclusões tomadas no teste de ANOVA com a rejeição da hipótese nula podem ser tomadas, a diferença é que perdemos um pouco de poder ao tomar essa conclusão.

Em caso de rejeição da hipótese nula do teste de Friedman, ou seja, os classificadores apresentam desempenhos estatisticamente diferentes, aplicamos o teste de Wilcoxon.

Novamente, considerando técnicas não-paramétricas, o teste de Wilcoxon é o relativo ao Teste T das técnicas paramétricas (Zimmerman e Zumbo 1993). Portanto as mesmas conclusões após rejeição da hipótese nula, podem ser tomadas.

6 EXPERIMENTOS E RESULTADOS

Com os dados determinados, alguns experimentos foram realizados de modo a alcançar os objetivos traçados inicialmente. Nesse capítulo, os experimentos que foram realizados serão descritos assim como os resultados obtidos por eles.

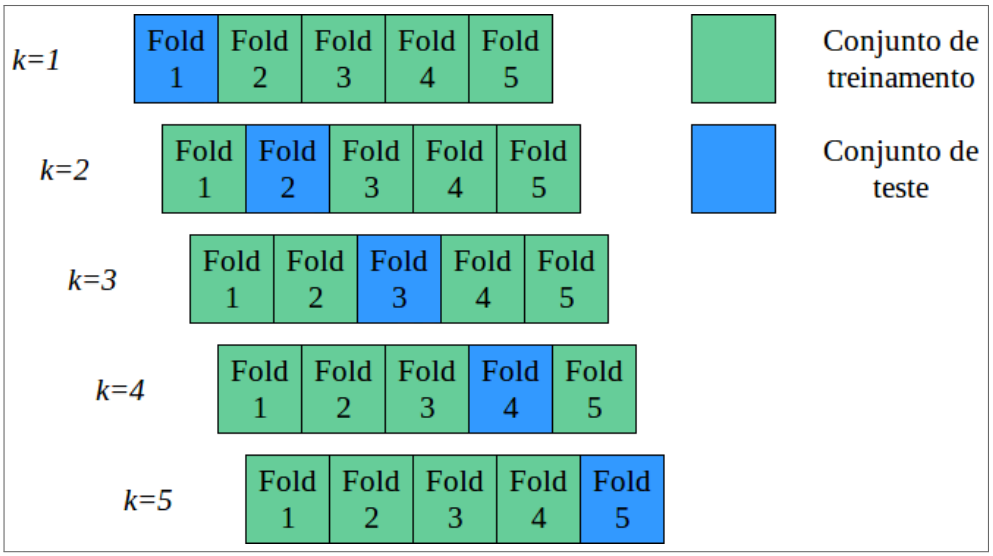
6.1 EXPERIMENTOS

Para aplicação dos algoritmos nas bases de dados, alguns pré-processamentos nas bases foram realizados. Inicialmente, uma transformação nos valores dos atributos foram realizadas. A transformação realizada foi a normalização sendo a medida de localização a média, e a medida de escala utilizada foi o desvio-padrão. Por exemplo, um atributo X_1 terá todos os seus valores $x_1, x_2, x_3, \ldots, x_n$ subtraídos pela média de X_1 e divididos pelo desvio-padrão de X_1 (Faceli et al. 2011).

Após a normalização dos dados, aplicamos uma técnica de redução de dimensionalidade chamada de Principal Components Analysis (*Principal Components Analysis* (PCA)). Que é uma técnica que correlaciona estatisticamente os exemplos, reduzindo a dimensionalidade por meio de funções lineares (Faceli et al. 2011).

Depois de passarem por esses processos de transformação, os dados estão prontos para serem utilizados pelos algoritmos, que realizarão as fases de treinamento, validação e teste a partir destes dados. No processo de treinamento e validação, onde é separado estratificadamente 70% e 10% do conjunto total de dados, respecticamente, os algoritmos passam por um etapa chamada de validação cruzada utilizando o método K-Fold. Este método opera da seguinte forma: O conjunto de treinamento e validação (80%) é variado K vezes. Esse conjunto de 80% é repartido estratificadamente em K conjuntos de tamanhos aproximadamente iguais. Os exemplos contidos em K - 1 conjuntos são utilizados no treinamento dos classificadores que serão avaliados no conjunto restante. Esse processo é repetido K vezes utilizando um conjunto diferente em cada uma das vezes para avaliação dos classificadores. O desempenho final dos classificadores será expressado através da média dos desempenhos observados sobre cada conjunto de avaliação. A Figura 17 demonstra visualmente esse método de validação cruzada.

Figura 17 – Representação da validação cruzada utilizando k-fold.



Fonte: https://www.researchgate.net/figure/Figura-3-Divisao-do-conjunto-de-dados-em-k-5-subconjuntos-folds$_f$ig2$_3$21027902

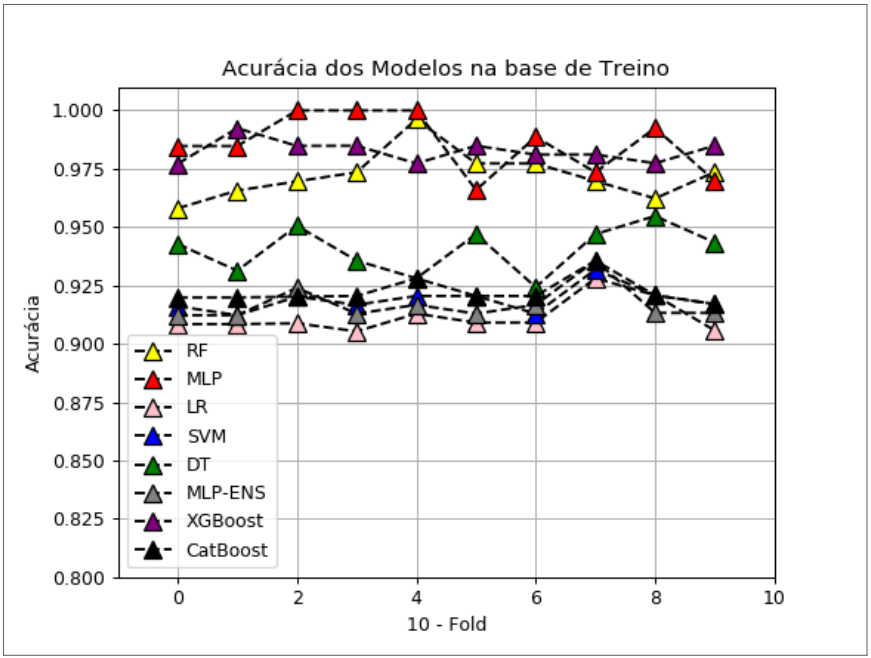
Para garantirmos que estaríamos avaliando na futura base de teste (composta por 20%) os classificadores em sua melhor composição, considerando as bases de treinamento e de validação. Utilizamos um algoritmo chamado Grid Search que é um algoritmo de busca gulosa que tem o objetivo de encontrar, no caso da aplicação deste algoritmo em aprendizagem de máquina, a melhor combinação dos hiperparâmetros de todos os classificadores que retornassem o melhor desempenho na base de validação (Faceli et al. 2011). Para cada combinação de hiperparâmetros possível, os classificadores passam pelo processo de validação cruzada utilizando o K-Fold. Por fim, a melhor combinação dos hiperparâmetros que retornou o melhor desempenho médio será utilizada para prever as instâncias da base de teste.

6.2 RESULTADOS

6.2.1 Desempenho na base com amostras padrões

Na Figura 18 podemos ver o desempenho dos classificadores na própria base de treinamento, podendo assim analisarmos se os mesmos estão sendo capazes de aprender padrões que distiguem as amostras com as características fornecidas.

Figura 18 – AMOSTRAS PADRÕES: Acurácia dos algoritmos na etapa de treinamento.



Fonte: Elaborada pelo autor.

Quadro 1 – AMOSTRAS PADRÕES: Acurácia Média e seu Desvio-Padrão na fase de treinamento.

Classificadores	Acurácia Média	Desvio-Padrão
RF	0.9723	0.0099
MLP	**0.9859**	0.0120
LR	0.9116	0.0068
SVM	0.9188	0.0052
DT	0.9404	0.0096
ENS-MLP	0.9169	0.0070
XGBoost	0.9825	0.0045
CatBoost	0.9222	0.0051

Fonte: Elaborada pelo autor

Como podemos ver no Quadro 2, ao aplicarmos o teste de Shapiro-Wilk para verificação da normalidade dos dados, observamos que o vetor de desempenhos dos classificadores LR, MLP-ENS e CAT não são normalmente distribuídos, considerando um nível de significância de 5%. Com isso, Técnicas Não-Paramétricas foram utilizadas para a verificação

da existência de diferença estatisticamente significativa entre o desempenho dos classificadores.

Quadro 2 – AMOSTRAS PADRÕES: Valores de P obtidos a partir do teste de Shapiro-Wilk quando aplica no vetor de desempenho dos classificadores nas fases de treinamento e validação.

Classificadores	**Fase de Treinamento**	**Fase de Validação**
RF	0.285083	0.802167
MLP	0.219603	0.044511
LR	0.006379	0.027203
SVM	0.079350	0.006662
DT	0.645792	0.506052
ENS-MLP	0.000571	0.092053
XGBoost	0.135795	0.737324
CatBoost	0.000716	0.001535

Fonte: Elaborada pelo autor

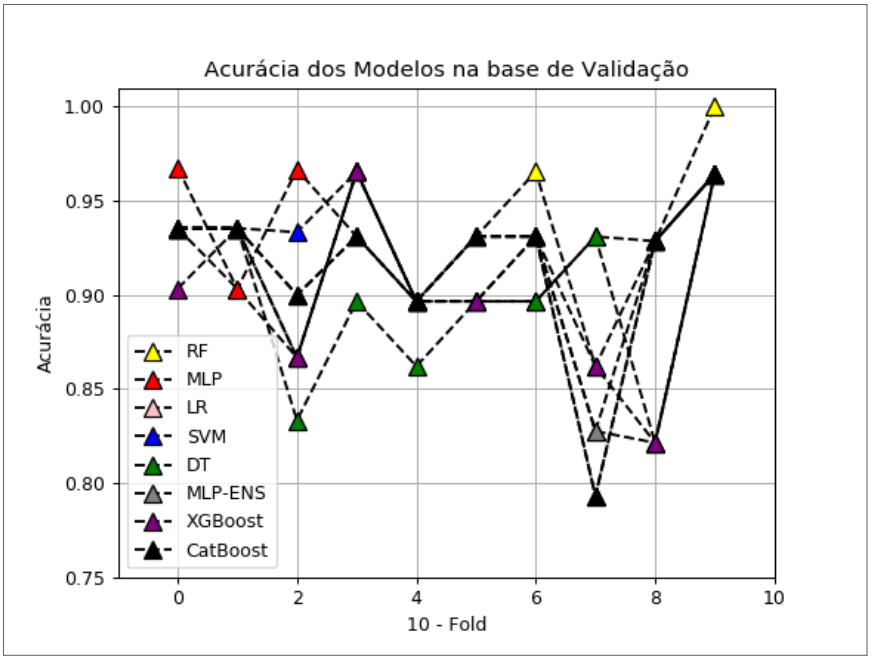
O teste de Friedman para medidas repetidas foi utilizado para analisarmos se existia diferença no grupo. E, ao nível de significância de 5%, com o valor de p de aproximadamente 0.0000, rejeitamos a hipótese de que o desempenho dos classificadores no treinamento tenha sido o mesmo. E, utilizando o teste de Wilcoxon par a par, verificamos no Quadro 3 que o desempenho dos classificadores SVM e ENS-MLP não apresentaram diferença estatisticamente significativa, assim como os classificadores MLP e XGB também. Todos os outros apresentaram desempenho estatisticamente diferente, fazendo assim com que os classificadores MLP e XGB tenham obtido os melhores desempenhos nessa base, com uma média de 99% e 98% de acurácia, respectivamente como podemos ver no Quadro 1.

Quadro 3 – AMOSTRAS PADRÕES: Valores de P obtidos a partir do teste de WILCOXON quando aplicado par a par nos classificadores na fase de treinamento, após a confirmação da diferença do desempenho do grupo pelo teste de Friedman.

Pares Analisados	**Valor de P**
RF vs MLP	0.024745
RF vs LR	0.004948
RF vs SVM	0.005034
RF vs DT	0.005062
RF vs ENS-MLP	0.005034
RF vs XGBoost	0.046710
RF vs CatBoost	0.005034
MLP vs LR	0.005034
MLP vs SVM	0.005062
MLP vs DT	0.005034
MLP vs ENS-MLP	0.005034
MLP vs XGBoost	0.413912
MLP vs CatBoost	0.005034
LR vs SVM	0.007579
LR vs DT	0.005034
LR vs ENS-MLP	0.024277
LR vs XGBoost	0.005062
LR vs CatBoost	0.007526
SVM vs DT	0.005062
SVM vs ENS-MLP	0.256180
SVM vs XGBoost	0.005034
SVM vs CatBoost	0.026857
DT vs ENS-MLP	0.005062
DT vs XGBoost	0.005034
DT vs CatBoost	0.007686
ENS-MLP vs XGBoost	0.005034
ENS-MLP vs CatBoost	0.020655
XGBoost vs CatBoost	0.005034

Fonte: Elaborada pelo autor.

Figura 19 – AMOSTRAS PADRÕES: Acurácia dos algoritmos na etapa de validação.



Fonte: Elaborada pelo autor.

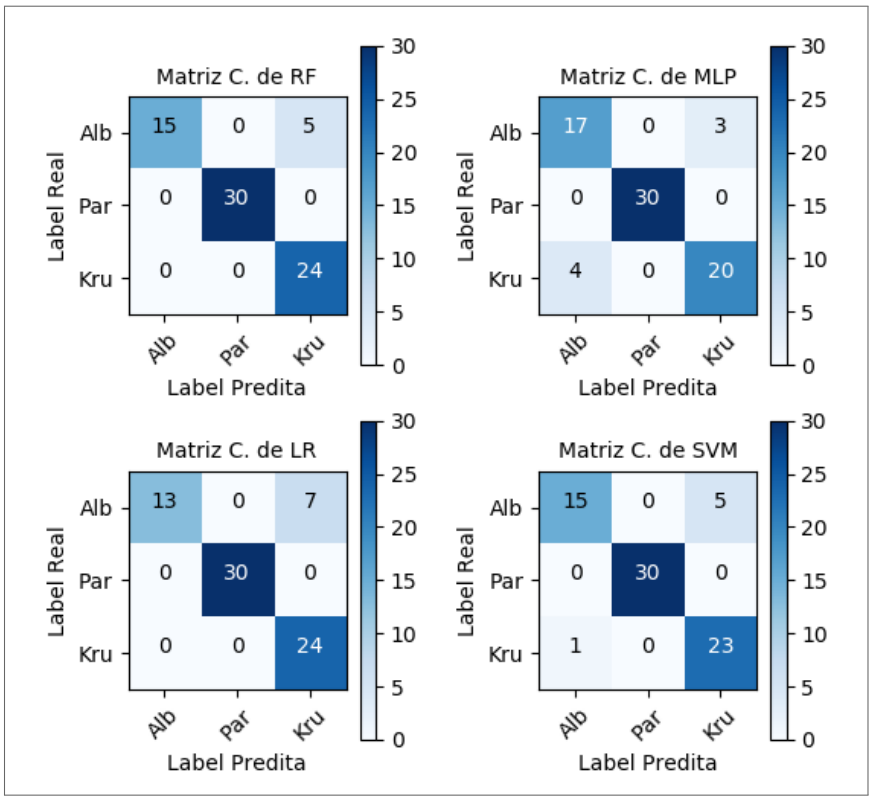
Já quando vamos avaliar o desempenho desses classificadores na base de validação, apresentada na Figura 19, verificamos que visualmente não conseguimos identificar qual classificador obteve um melhor desempenho. Porém, ao aplicarmos o teste de Shapiro-Wilk e verificarmos no Quadro 2 que os únicos classificadores que apresentaram vetores de desempenho normalmente distrbuídos foram RF, DT e XGB. Aplicamos o teste de Friedman para medidas repetidas e confirmamos, com um valor de p de 0.2658, que os classificadores apresentam um mesmo desempenho nessa base. O que pode ser visto também no Quadro 4. Com isso, fazendo com que o algoritmo menos computacionalmente custoso tenha mais vantagem em relação aos outros.

Quadro 4 – AMOSTRAS PADRÕES: Acurácia Média e seu Desvio-Padrão na fase de validação.

Classificadores	**Acurácia Média**	**Desvio-Padrão**
RF	0.9254	0.0420
MLP	**0.9282**	0.0282
LR	0.9073	0.0451
SVM	0.9179	0.0469
DT	0.8972	0.0443
ENS-MLP	0.9182	0.0409
XGBoost	0.9042	0.0438
CatBoost	0.9146	0.0443

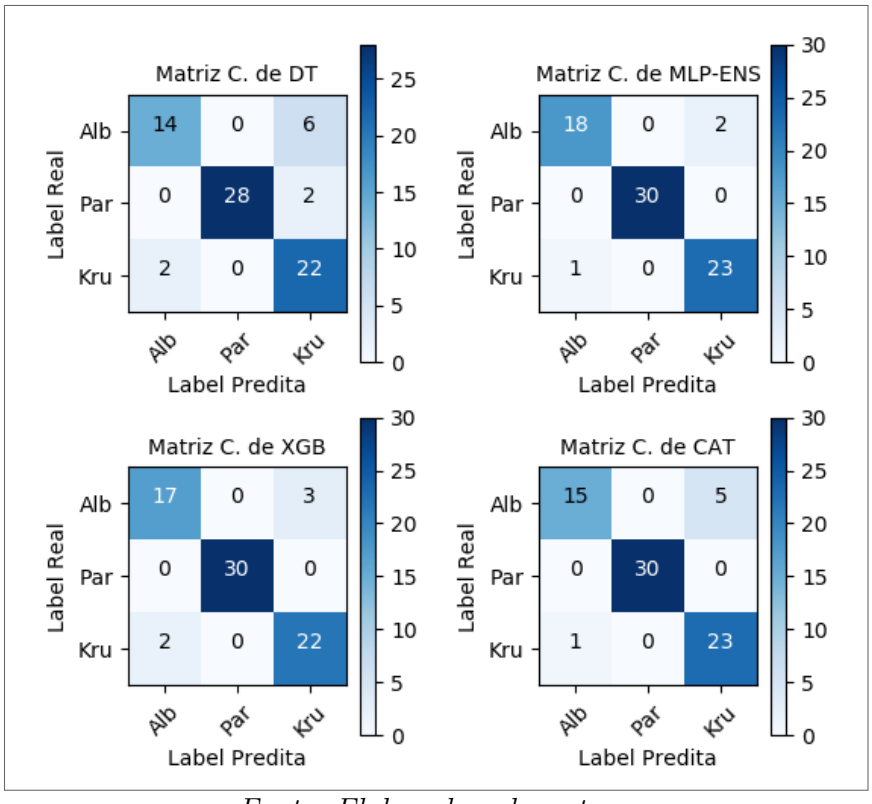
Fonte: Elaborada pelo autor

Figura 20 – AMOSTRAS PADRÕES: Matriz de Confusão dos algoritmos RF, MLP, LR e SVM na etapa de teste.



Fonte: Elaborada pelo autor.

Figura 21 – AMOSTRAS PADRÕES: Matriz de Confusão dos algoritmos DT, MLP-ENS, XGB e CAT na etapa de teste.



Fonte: Elaborada pelo autor.

Quadro 5 – AMOSTRAS PADRÕES: Precisão, Recall e Acurácia dos classificadores na fase de teste.

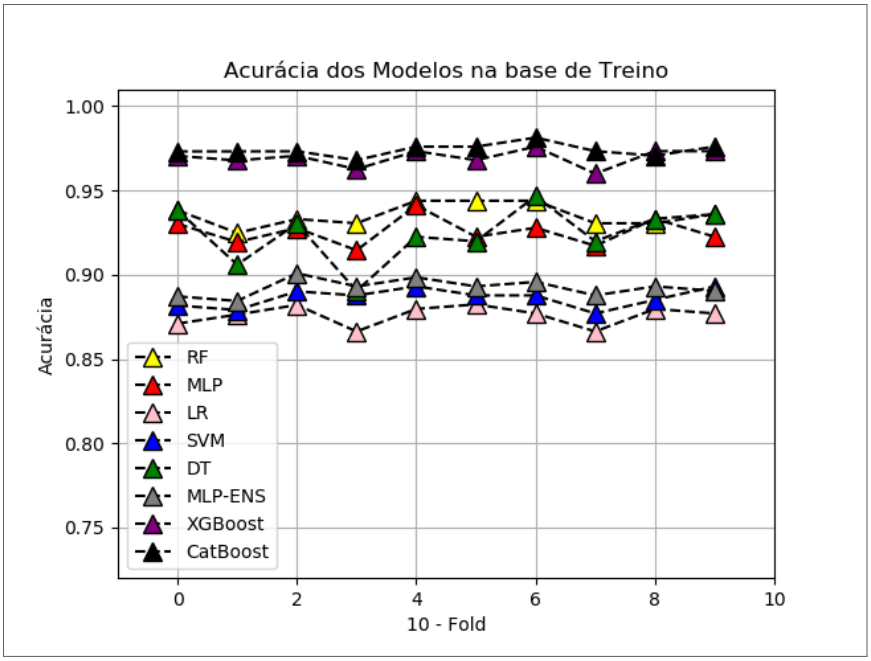
Classificadores	**Precisão**			**Recall**			**Acurácia**
	Alb	**Par**	**Kru**	**Alb**	**Par**	**Kru**	
RF	1.00	1.00	0.83	0.75	1.00	1.00	0.93
MLP	0.81	1.00	0.87	0.85	1.00	0.83	0.91
LR	1.00	1.00	0.77	0.65	1.00	1.00	0.91
SVM	0.94	1.00	0.82	0.75	1.00	0.96	0.92
DT	0.88	1.00	0.73	0.70	0.93	0.92	0.86
ENS-MLP	0.95	1.00	0.92	0.90	1.00	0.96	**0.96**
XGBoost	0.89	1.00	0.88	0.85	1.00	0.92	0.93
CatBoost	0.94	1.00	0.82	0.75	1.00	0.96	0.92

Fonte: Elaborada pelo autor.

Utilizando o algoritmo Grid-Search nas bases anteriores, foi escolhido a melhor combinação de parâmetros, otimizados na base de validação, para verificação de seu desempenho na base de teste, que é composta por 20% da base original. Os resultados podem ser vistos em forma de Matriz de Confusão nas Figuras 20 e 21 e também expressados em forma de Precisão, Recall e Acurácia no Quadro 5. Como podemos ver, o comportamento analisado na base de validação se repete na base de teste, onde não conseguimos identificar uma diferença entre o desempenho dos classificadores. Porém, em termos de acurácia podemos verificar que o classificador ENS-MLP obteve o melhor resultado (96%) na base de teste.

6.2.2 Desempenho na base com amostras consideradas Não-Padrões.

Figura 22 – AMOSTRAS NÃO-PADRÕES: Acurácia dos algoritmos na etapa de treino.



Fonte: Elaborada pelo autor.

O desempenho dos algoritmos na base de treino é ilustrado na Figura 22. E, após a aplicação do teste de Shapiro-Wilk, como podemos verificar no Quadro 7, que todos os vetores apresentaram uma distribuição normal, sendo assim utilizamos técnicas paramétricas para realizar a comparação do desempenho desses algoritmos, tanto em grupo quanto par a par. O teste de ANOVA para medidas repetidas apresentou um valor de p de aproximadamente 0.0000, indicando diferença entre os desempenhos. Já quando utilizamos o Teste T para verificarmos par a par, os únicos classificadores que apresentaram o mesmo desempenho, estatisticamente, foram MLP e DT com um valor de p de 0.7655, como podemos ver no Quadro 8. Sendo assim, o classificador que apresentou o melhor desempenho foi o CAT com uma acurácia média de 97.40%, em segundo lugar o XGB com 96.95%, e com o pior desempenho o classifcador LR apresentou 87.58% de acurácia média, como podemos ver no Quadro 6.

Quadro 6 – AMOSTRAS NÃO-PADRÕES: Acurácia Média e seu Desvio-Padrão na fase de treinamento.

Classificadores	**Acurácia Média**	**Desvio-Padrão**
RF	0.9354	0.0064
MLP	0.9255	0.0076
LR	0.8757	0.0056
SVM	0.8862	0.0052
DT	0.9242	0.0157
ENS-MLP	0.8923	0.0048
XGBoost	0.9694	0.0048
CatBoost	**0.9740**	0.0034

Fonte: Elaborada pelo autor.

Quadro 7 – AMOSTRAS NÃO-PADRÕES: Valores de P obtidos a partir do teste de Shapiro-Wilk quando aplicado no vetor de desempenho dos classificadores nas fases de treinamento e validação.

Classificadores	**Fase de Treinamento**	**Fase de Validação**
RF	0.209439	0.642230
MLP	0.874539	0.395228
LR	0.101611	0.588548
SVM	0.450295	0.729603
DT	0.582718	0.203595
ENS-MLP	0.899946	0.383820
XGBoost	0.321198	0.181923
CatBoost	0.462757	0.022974

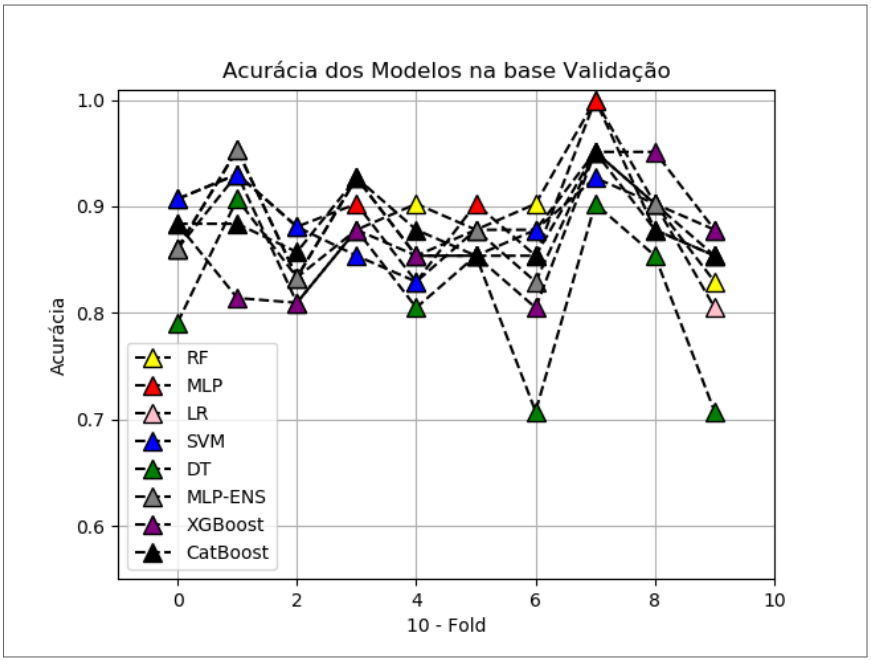
Fonte: Elaborada pelo autor.

Quadro 8 – AMOSTRAS NÃO-PADRÕES: Valores de P obtidos a partir do Teste T para amostras pareadas quando aplicado par a par nos classificadores na fase de treinamento, após a confirmação da diferença do desempenho do grupo pelo teste de ANOVA

Pares Analisados	**Valor de P**
RF vs MLP	0.002112
RF vs LR	0.000000
RF vs SVM	0.000000
RF vs DT	0.035867
RF vs ENS-MLP	0.000000
RF vs XGBoost	0.000000
RF vs CatBoost	0.000000
MLP vs LR	0.000000
MLP vs SVM	0.000000
MLP vs DT	0.765527
MLP vs ENS-MLP	0.000000
MLP vs XGBoost	0.000000
MLP vs CatBoost	0.000000
LR vs SVM	0.000218
LR vs DT	0.000003
LR vs ENS-MLP	0.000005
LR vs XGBoost	0.000000
LR vs CatBoost	0.000000
SVM vs DT	0.000043
SVM vs ENS-MLP	0.000617
SVM vs XGBoost	0.000000
SVM vs CatBoost	0.000000
DT vs ENS-MLP	0.000158
DT vs XGBoost	0.000002
DT vs CatBoost	0.000002
ENS-MLP vs XGBoost	0.000000
ENS-MLP vs CatBoost	0.000000
XGBoost vs CatBoost	0.007452

Fonte: Elaborada pelo autor.

Figura 23 – AMOSTRAS NÃO-PADRÕES: Acurácia dos algoritmos na etapa de validação.



Fonte: Elaborada pelo autor.

Quadro 9 – AMOSTRAS NÃO-PADRÕES: Acurácia Média e seu Desvio-Padrão na fase de validação.

Classificadores	**Acurácia Média**	**Desvio-Padrão**
RF	0.8892	0.0504
MLP	**0.8937**	0.0455
LR	0.8818	0.0476
SVM	0.8840	0.0312
DT	0.8238	0.0681
ENS-MLP	0.8866	0.0430
XGBoost	0.8677	0.0500
CatBoost	0.8819	0.0314

Fonte: Elaborada pelo autor.

Podemos avaliar, na Figura 23, o desempenho dos algoritmos na base de validação. Novamente verificamos no Quadro 7 a normalidade dos vetores de desempenho dos classificadores e apenas o CAT não apresentou distribuição normal. Como estamos analisando o grupo, temo que utilizar técnicas não-paramétricas. O teste de Friedman foi aplicado e, com um valor de p de 0.0167, o desempenho dos classificadores na base de validação são estatisticamente diferentes. Após aplicarmos o teste de Wilcoxon par a par, verificamos a partir do Quadro 10 que o classificador DT apresentou desempenho estatisticamente diferente de todos os outros classificadores, menos com o XGB (valor de p 0.0684). Fazendo assim com que o DT apresente o pior desempenho nessa base de validação, com

uma média de 82% de acurácia, como apresentado no Quadro 9. O desempenho dos outros algoritmos não foram estatisticamente diferentes, sendo assim, o único critério que poderíamos utilizar para escolher um melhor classificador seria a simplicidade computacional do mesmo.

Quadro 10 – AMOSTRAS NÃO-PADRÕES: Valores de P obtidos a partir do teste de WILCOXON quando aplicado par a par nos classificadores na fase de validação, após a confirmação da diferença do desempenho do grupo pelo teste de Friedman.

Pares Analisados	**Valor de P**
RF vs MLP	0.831589
RF vs LR	0.259678
RF vs SVM	0.609523
RF vs DT	0.015067
RF vs ENS-MLP	0.730647
RF vs XGBoost	0.205343
RF vs CatBoost	0.471474
MLP vs LR	0.199381
MLP vs SVM	0.336289
MLP vs DT	0.005005
MLP vs ENS-MLP	0.375193
MLP vs XGBoost	0.201152
MLP vs CatBoost	0.394627
LR vs SVM	0.623639
LR vs DT	0.011514
LR vs ENS-MLP	0.592980
LR vs XGBoost	0.446060
LR vs CatBoost	0.864813
SVM vs DT	0.016324
SVM vs ENS-MLP	1.000000
SVM vs XGBoost	0.607959
SVM vs CatBoost	0.766299
DT vs ENS-MLP	0.007579
DT vs XGBoost	0.068364
DT vs CatBoost	0.010793
ENS-MLP vs XGBoost	0.176296
ENS-MLP vs CatBoost	0.472783
XGBoost vs CatBoost	0.351681

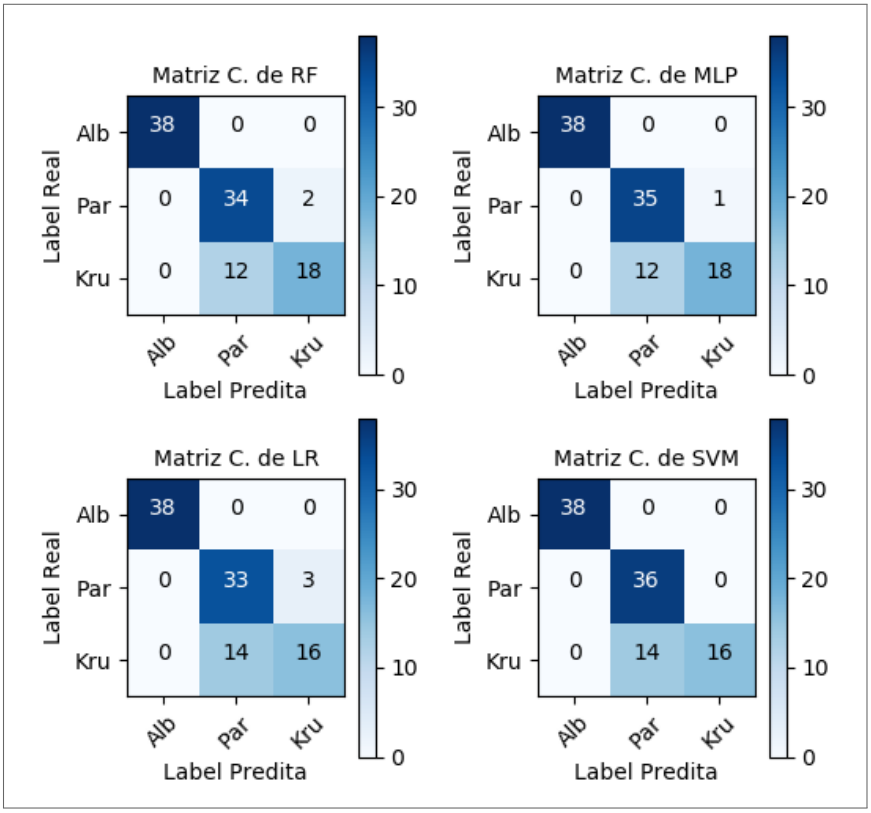
Fonte: Elaborada pelo autor.

Quadro 11 – AMOSTRAS NÃO-PADRÕES: Precisão, Recall e Acurácia dos classificadores na fase de teste.

Classificadores	**Precisão**			**Recall**			**Acurácia**
	Alb	**Par**	**Kru**	**Alb**	**Par**	**Kru**	
RF	1.00	0.74	0.90	1.00	0.94	0.60	0.87
MLP	1.00	0.74	0.95	1.00	0.97	0.60	**0.88**
LR	1.00	0.70	0.84	1.00	0.92	0.53	0.84
SVM	1.00	0.72	1.00	1.00	1.00	0.53	0.87
DT	0.79	0.60	0.65	0.82	0.81	0.37	0.68
ENS-MLP	1.00	0.74	0.95	1.00	0.97	0.60	**0.88**
XGBoost	1.00	0.77	0.91	1.00	0.94	0.67	**0.88**
CatBoost	1.00	0.74	0.95	1.00	0.97	0.60	**0.88**

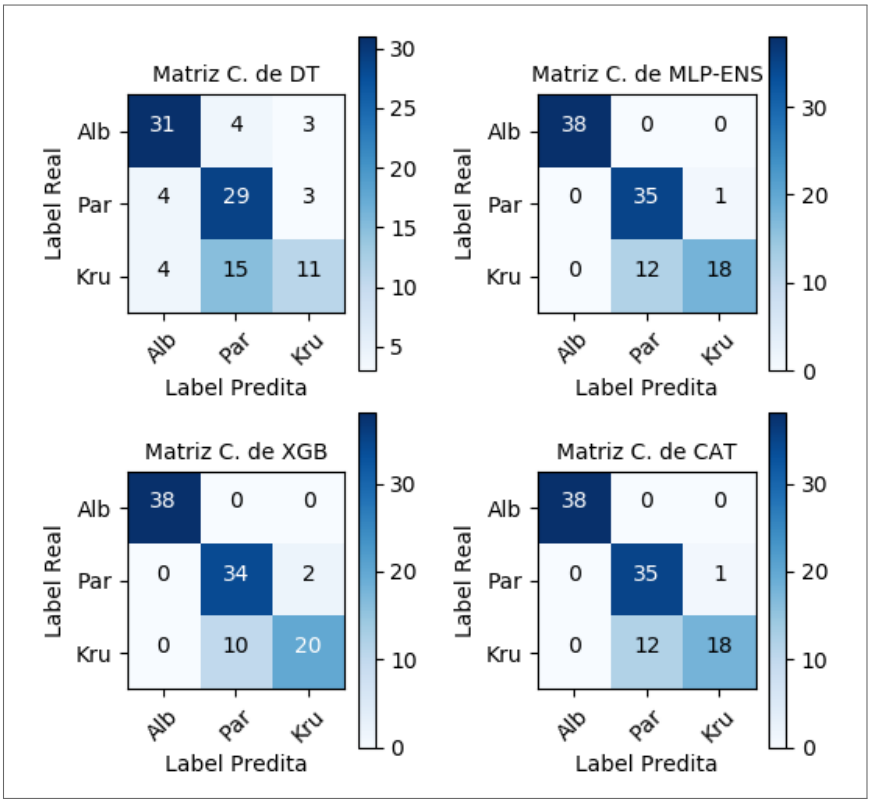
Fonte: Elaborada pelo autor.

Figura 24 – AMOSTRAS NÃO-PADRÕES: Matriz de Confusão dos algoritmos RF, MLP, LR e SVM na etapa de teste.



Fonte: Elaborada pelo autor.

Figura 25 – AMOSTRAS NÃO-PADRÕES: Matriz de Confusão dos algoritmos DT, MLP-ENS, XGB e CAT na etapa de teste.

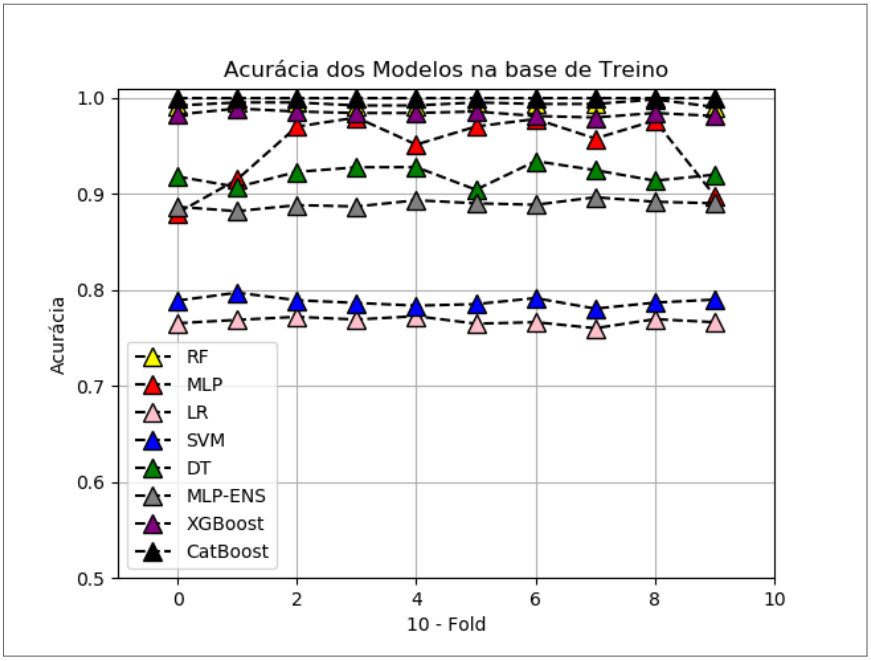


Fonte: Elaborada pelo autor.

Nas Figuras 24 e 25, podemos verificar através das Matrizes de Confusão dos classificadores que o comportamento analisado na base de validação foi replicado na base de teste, DT comprova ter o pior desempenho perante aos outros algoritmos. E, não conseguimos identificar uma diferença no desempenho dos outros classificadores. Este resultado fica mais explícito no Quadro 11, onde são apresentadas as medidas de Precisão, Recall e Acurácia dos classificadores obtidas na fase de teste da base das amostras consideradas Não-Padrões. Porém, podemos verificar também através das matrizes de confusão que todos os classificadores tiveram dificuldades em predizer algumas amostras de *Candida krusei*, predizendo elas como sendo *Candida parapsilosis*.

6.2.3 Desempenho na base com amostras analisadas conjuntamente

Figura 26 – AMOSTRAS JUNTAS: Acurácia dos algoritmos na etapa de treino.



Fonte: Elaborada pelo autor.

Na Figura 26, verificamos o desempenho dos classificadores na base de treino. Aqui, apenas o vetor do algoritmo MLP rejeitou a hipótese de distribuição normal, como podemos ver no Quadro 13. Com um valor de p de aproximadamente 0.0000, o teste de Friedman rejeita a hipótese de semelhança entre o desempenho dos classificadores. E, aplicando o teste de Wilcoxon, verificamos no Quadro 14 que o desempenho de todos os algoritmos apresentam uma diferença estatisticamente significativa. Fazendo com que o algoritmo LR apresente o pior desempenho e o CAT o melhor, com 77% e 100% de acurácia média, respectivamente, como podemos ver no Quadro 12.

Quadro 12 – AMOSTRA JUNTAS: Acurácia Média e seu Desvio-Padrão na fase de treinamento.

Classificadores	Acurácia Média	Desvio-Padrão
RF	0.9938	0.0021
MLP	0.9477	0.0345
LR	0.7675	0.0035
SVM	0.7879	0.0043
DT	0.9201	0.0089
ENS-MLP	0.8895	0.0037
XGBoost	0.9838	0.0026
CatBoost	**1.0000**	0.0000

Fonte: Elaborada pelo autor.

Quadro 13 – AMOSTRA JUNTAS: Valores de P obtidos a partir do teste de Shapiro-Wilk quando aplicado no vetor de desempenho dos classificadores nas fases de treinamento e validação.

Classificadores	**Fase de Treinamento**	**Fase de Validação**
RF	0.495721	0.301207
MLP	0.028047	0.952030
LR	0.735503	0.773344
SVM	0.938141	0.496322
DT	0.814008	0.887462
ENS-MLP	0.983843	0.325505
XGBoost	0.801148	0.400025
CatBoost	1.000000	0.979707

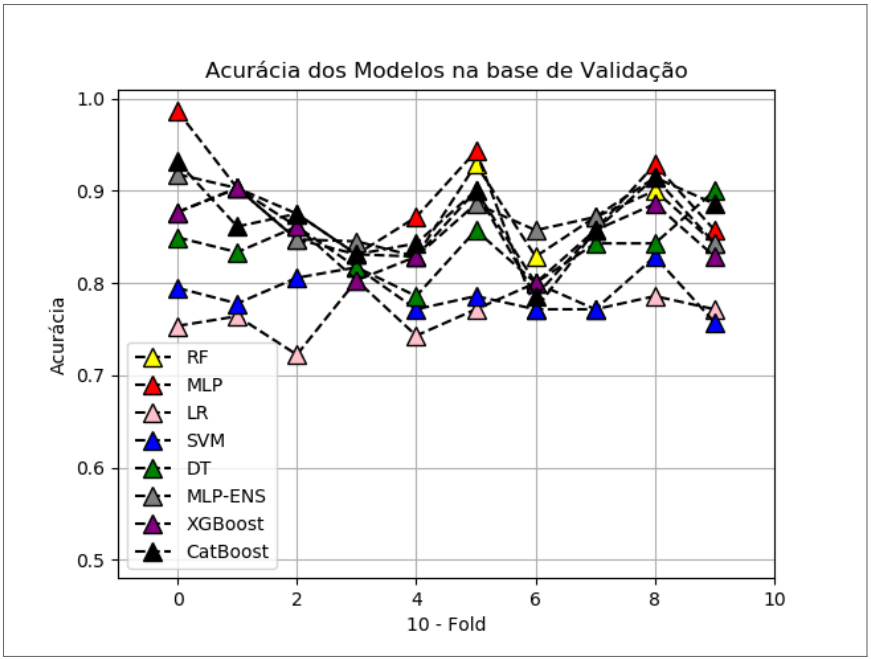
Fonte: Elaborada pelo autor.

Quadro 14 – AMOSTRA JUNTAS: Valores de P obtidos a partir do teste de WILCOXON quando aplicado par a par no desempenho dos classificadores na fase de treinamento, após a confirmação da diferença do desempenho do grupo pelo teste de Friedman.

Pares Analisados	**Valor de P**
RF vs MLP	0.005062
RF vs LR	0.005062
RF vs SVM	0.005062
RF vs DT	0.005062
RF vs ENS-MLP	0.005062
RF vs XGBoost	0.005034
RF vs CatBoost	0.005005
MLP vs LR	0.005062
MLP vs SVM	0.005062
MLP vs DT	0.036658
MLP vs ENS-MLP	0.006910
MLP vs XGBoost	0.005062
MLP vs CatBoost	0.005062
LR vs SVM	0.005062
LR vs DT	0.005062
LR vs ENS-MLP	0.005034
LR vs XGBoost	0.004948
LR vs CatBoost	0.005034
SVM vs DT	0.005062
SVM vs ENS-MLP	0.005062
SVM vs XGBoost	0.005062
SVM vs CatBoost	0.005062
DT vs ENS-MLP	0.005062
DT vs XGBoost	0.005062
DT vs CatBoost	0.005062
ENS-MLP vs XGBoost	0.005005
ENS-MLP vs CatBoost	0.005034
XGBoost vs CatBoost	0.005005

Fonte: Elaborada pelo autor.

Figura 27 – AMOSTRAS JUNTAS: Acurácia dos algoritmos na etapa de validação.



Fonte: Elaborada pelo autor.

Na Figura 27, podemos analisar o desempenho dos classificadores na base de validação. E, após aplicação do teste de Shapiro-Wilk, verificamos no Quadro 13 que todos os vetores mostraram ser normalmente distribuídos. Com isso, utilizamos técnicas paramétricas para comparação dos mesmos. O teste de ANOVA para medidas repetidas apresentou um valor de p de aproximadamente 0.0000, confirmando a diferença estatisticamente significativa entre o desempenho dos classificadores analisados em grupo. Já no Teste T par a par, verificamos no Quadro 16 que os pares que não apresentaram desempenho diferentes foram: RF e MLP, RF e ENS-MLP, RF e CAT, MLP e DT, MLP e ENS-MLP, MLP e XGB, MLP e CAT, LR e SVM, DT e XGB, ENS-MLP e XGB, ENS-MLP e CAT, XGB e CAT. Portanto, como podemos ver no Quadro 15 com acurácias médias de 86.85%, 87.96%, 83.89%, 87.13%, 85.43%, 86.84%, os classificadores RF, MLP, DT, MLP-ENS, XGB e CAT apresentaram, respectivamente, desempenho superior aos classificadores LR e SVM, que apresentaram acurácias médias de 76.85% e 78.80%.

Quadro 15 – AMOSTRA JUNTAS: Acurácia Média e seu Desvio-Padrão na fase de validação.

Classificadores	**Acurácia Média**	**Desvio-Padrão**
RF	0.8685	0.0333
MLP	**0.8795**	0.0587
LR	0.7685	0.0235
SVM	0.7880	0.0217
DT	0.8389	0.0309
ENS-MLP	0.8712	0.0305
XGBoost	0.8543	0.0359
CatBoost	0.8684	0.0406

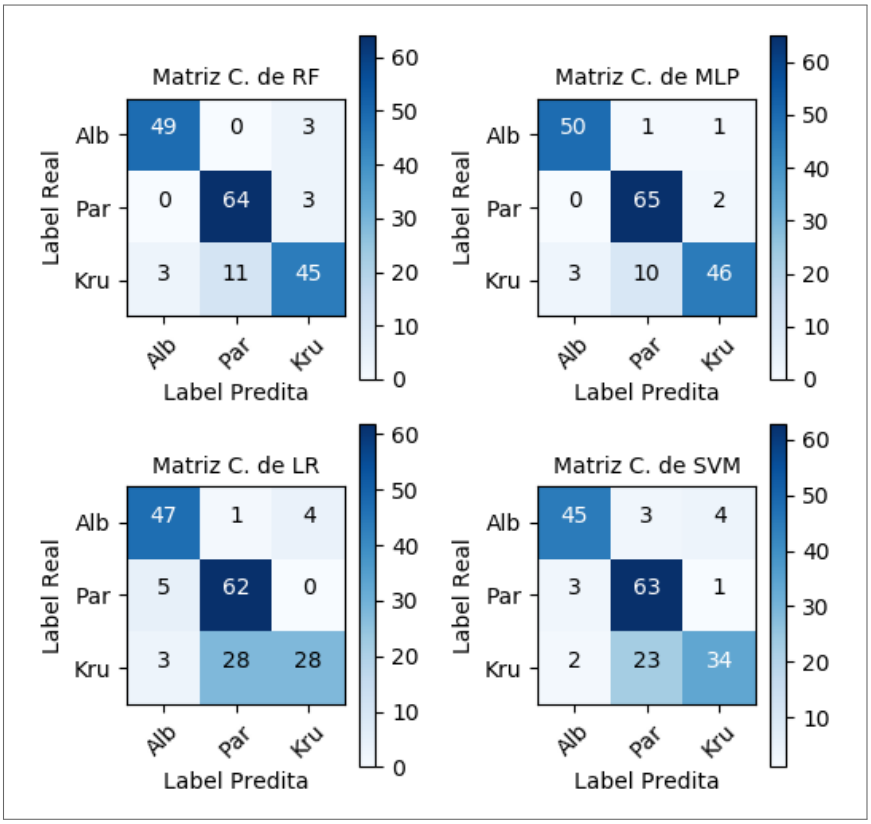
Fonte: Elaborada pelo autor.

Quadro 16 – AMOSTRA JUNTAS: Valores de P obtidos a partir do Teste T quando aplicado par a par no desempenho dos classificadores na fase de validação, após a confirmação da diferença do desempenho do grupo pelo teste ANOVA.

Pares Analisados	**Valor de P**
RF vs MLP	0.456251
RF vs LR	0.000081
RF vs SVM	0.000065
RF vs DT	0.031969
RF vs ENS-MLP	0.732625
RF vs XGBoost	0.003794
RF vs CatBoost	0.991407
MLP vs LR	0.000990
MLP vs SVM	0.000959
MLP vs DT	0.060526
MLP vs ENS-MLP	0.563840
MLP vs XGBoost	0.071634
MLP vs CatBoost	0.271061
LR vs SVM	0.083910
LR vs DT	0.000765
LR vs ENS-MLP	0.000015
LR vs XGBoost	0.000572
LR vs CatBoost	0.000356
SVM vs DT	0.003489
SVM vs ENS-MLP	0.000019
SVM vs XGBoost	0.000557
SVM vs CatBoost	0.000182
DT vs ENS-MLP	0.034065
DT vs XGBoost	0.250935
DT vs CatBoost	0.020861
ENS-MLP vs XGBoost	0.058127
ENS-MLP vs CatBoost	0.796114
XGBoost vs CatBoost	0.174713

Fonte: Elaborada pelo autor.

Figura 28 – AMOSTRAS JUNTAS: Matriz de Confusão dos algoritmos RF, MLP, LR e SVM na etapa de teste.



Fonte: Elaborada pelo autor.

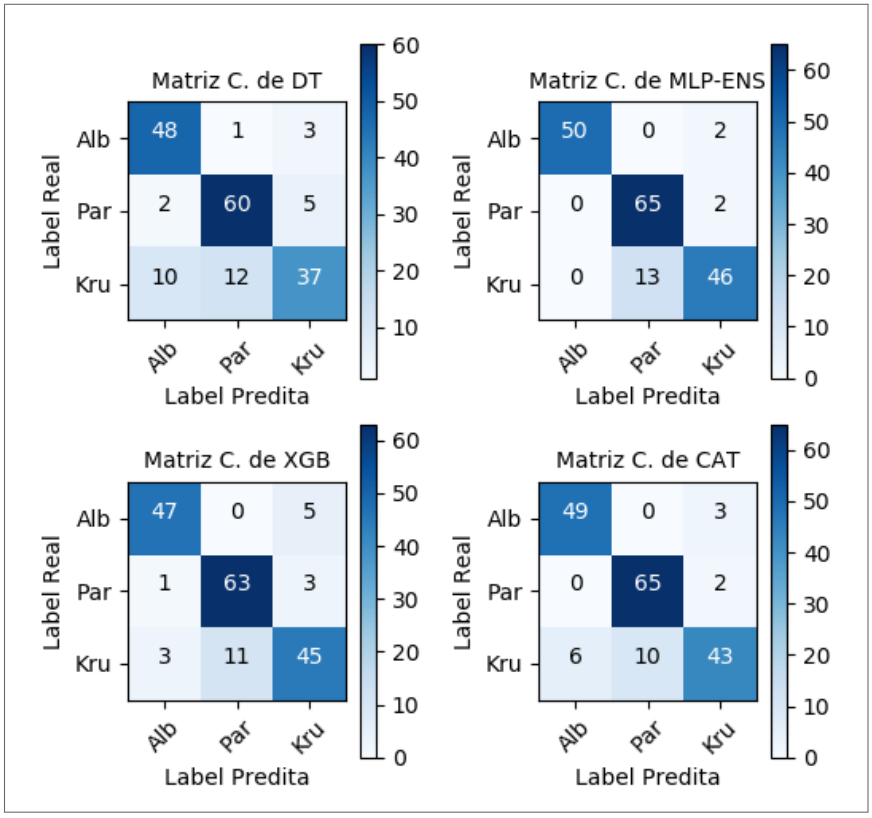
E, novamente, podemos notar atráves das Figuras 28 e 29 além do Quadro 17, que os desempenhos dos classificadores analisados na fase de validação se repetiram na fase de teste na base das amostras analisadas conjuntamente. Como podemos ver, os classificadores também continuam predizendo algumas amostras de *Candida krusei* como sendo *Candida parapsilosis*. Este comportamento já era esperado pois juntamos as duas bases anteriores e, portanto, as amostras da base Não-Padrão estão presentes nesta base também. Destacamos que os classificadores LR e SVM erraram mais amostras que os outros classificadores em termos de amostras de *Candida krusei* sendo predita como *Candida parapsilosis*. Já o classificador DT, neste caso, errou também a predição de algumas amostras de *Candida krusei*, predizendo as mesmas como sendo *Candida albicans*. Por fim, destacamos que, em termos de acurácia na base de teste, o classificador ENS-MLP também obteve o melhor resultado neste cenário, fazendo dele o classificador mais indicado para predizer amostras de *Candida*, quando analisamos apenas as bases de teste.

Quadro 17 – AMOSTRA JUNTAS: Precisão, Recall e Acurácia dos classificadores na fase de teste.

Classificadores	Precisão			Recall			Acurácia
	Alb	Par	Kru	Alb	Par	Kru	
RF	0.94	0.85	0.88	0.94	0.96	0.76	0.89
MLP	0.94	0.86	0.94	0.96	0.97	0.78	0.90
LR	0.85	0.68	0.88	0.90	0.93	0.47	0.77
SVM	0.90	0.71	0.87	0.87	0.94	0.58	0.80
DT	0.80	0.82	0.82	0.92	0.90	0.63	0.81
ENS-MLP	1.00	0.83	0.92	0.96	0.97	0.78	**0.90**
XGBoost	0.92	0.85	0.85	0.90	0.94	0.76	0.87
CatBoost	0.89	0.87	0.90	0.94	0.97	0.73	0.88

Fonte: Elaborada pelo autor.

Figura 29 – AMOSTRAS JUNTAS: Matriz de Confusão dos algoritmos DT, MLP-ENS, XGB e CAT na etapa de teste.



Fonte: Elaborada pelo autor.

7 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho foi desenvolvido com base na área de Inteligência Artificial e Aprendizagem de Máquina aplicada a um domínio totalmente novo, a identificação de espécies de *Candida* que causam a Candidemia (infecção na corrente sanguínea causada pelo fungo *Candida*). Na desenvoltura de um possível novo método de identificação de espécies de *Candida*, aplicamos técnicas de Aprendizagem de Máquina, no âmbito de Aprendizado Supervisionado, são elas: Decision Tree, Multilayer Perceptron, Logistic Regression, Support Vector Machine, Random Forest, Ensemble de Multilayer Perceptron, Extreme Gradient Boosting, Categorical Boosting.

Fizemos uma breve revisão de alguns métodos que são utilizados para a realização da identificação de espécies de *Candida*, verificando que métodos laboratoriais ainda recorrem a métodos convencionais como o teste de tubo germinativo.

Podemos verificar através das análises realizadas no Capítulo 6 na seção 6.2, que a utilização de um dispositivo (Nariz Eletrônico) composto por um sensor capaz de captar os compostos orgânicos voláteis presentes nas amostras de ar de *Candida*, emitindo sinais para um computador, para que algoritmos de Aprendizagem de Máquina realizem futuras análises no âmbito de identificação de espécies de *Candida* é bastante promissor. Pois, em termos de acurácias, obtivemos bons resultados nas três bases (Amostras Padrões; Amostras Não-Padrões; Amostras Juntas) na fase de teste, onde os classificadores nunca tiveram acesso à espécie de *Candida* em que os exemplos pertenciam:

- 96% nas amostras padrões;
- 88% nas amostras não-padrões;
- 90% nas amostras analisadas conjuntamente.

As acurácias apresentadas acima são, respectivamente, dos classificadores: ENS-MLP; MLP, ENS-MLP, XGBoost, CatBoost; ENS-MLP. Mostrando assim que, analisando a base de teste dos três cenários, o classificador que obteve o melhor resultado em termos de acurácia e, portanto, é o mais indicado para realizar a predição de amostras de *Candida albicans*, *Candida parapsilosis* e *Candida krusei*, é o classificador ENS-MLP. Confirmando nossa suposição, pois o mesmo se beneficia da normalização realizado nos dados a partir da média e do desvio padrão, se beneficia também em explorar vários classificadores MLPs, pois é um ensemble. Além de que os pesos iniciais das MLPs, utilizadas para construção deste classificador, são aleatórios, permitindo assim avaliarmos diferentes regiões e combinarmos, através do voto majoritário, a decisão das diferentes MLPs.

Esse método de identificação se mostra vantajoso pois não necessitamos de um alto conhecimento para execução do mesmo, uma vez que o treinamento dos classificadores

tenha sido realizado, apenas precisamos apresentar novas amostras de *Candida* contendo o ar dentro dos meios de cultura e os classificadores serão capazes de identificar a espécie a qual aquela amostra de ar pertence. Além de um baixo custo para a obtenção do sensor utilizado no nariz eletrônico, esse método também se mostra vantajoso pois apresenta fácil mobilidade para outros locais por ser um dispositivo de porte consideravelmente pequeno. Outra vantagem deste método é que, uma vez que os algoritmos de Aprendizagem de Máquina sejam treinados com várias amostras de diferentes espécies de *Candida*, o mesmo será capaz de distinguir entre essas espécies. Diferentemente de alguns métodos que são discriminatórios na identificação de algumas espécies mas não conseguem identificar outras espécies com a mesma facilidade, como o Tubo Germinativo que é discriminatório em identificar *Candida albicans* e o método CHROMagar que é discriminatório em identificar espécies do tipo *Candida krusei*, *Candida albicans* e *Candida tropicalis*.

Alguns trabalhos futuros são:

- Aumentar as bases de dados utilizadas, tanto em quantidade de exemplos das espécies utilizadas quanto em quantidade de outras espécies.
- Melhorar o protocolo de coleta de dados, de modo a deixar o processo com o mínimo de intervenção humana possível.
- Avaliar o desempenho dos classificadores em amostras de *Candida* que passaram pelo processo de crescimento em cultura em menos de 24 horas, com o objetivo de avaliar se os classificadores têm a capacidade de discriminar entre as amostras com menos tempo de cultura. Diminuindo, assim, o tempo do auxílio ao diagnóstico.
- Testar o desempenho de outros classificadores (LightGBM, por exemplo), inclusive avaliar a possibilidade de utilizar *Deep Learning*, pois com o aumento das bases de dados, esta aplicação se torna mais factível.

REFERÊNCIAS

ALEXANDER, B. Diagnosis of fungal infection: new technologies for the mycology laboratory. *Transplant infectious disease: an official journal of the Transplantation Society*, v. 4, p. 32–37, 2002.

ANDRESEN, S. L. John mccarthy: father of ai. *IEEE Intelligent Systems*, IEEE, v. 17, n. 5, p. 84–85, 2002.

ANVISA. Detecção e identificação dos fungos de importância médica.

BAUMGARTNER, C.; FREYDIERE, A.-M.; GILLE, Y. Direct identification and recognition of yeast species from clinical material by using albicans id and chromagar candida plates. *Journal of clinical microbiology*, Am Soc Microbiol, v. 34, n. 2, p. 454–456, 1996.

BENCHMARKING Random Forest Implementations. 2015. <http://datascience.la/benchmarking-random-forest-implementations/>. Accessed: 2020-01.

BIJLAND, L.; BOMERS, M.; SMULDERS, Y. Smelling the diagnosis a review on the use of scent in diagnosing. *Neth. J. Med*, v. 71, n. 2013, p. 300–307, 2013.

BUCK, L. B. The search for odorant receptors. *Cell*, Elsevier, v. 116, p. S117–S120, 2004.

BUCK, L. B. Unraveling the sense of smell (nobel lecture). *Angewandte Chemie International Edition*, Wiley Online Library, v. 44, n. 38, p. 6128–6140, 2005.

CARRIS, L. M.; LITTLE, C. R.; STILES, C. M. Introduction to fungi. *The Plant Health Instructor*, Washington State University, Kansas State University, and Georgia Military . . . , 2012.

CATBOOSTAI. 2017. <https://catboost.ai/>. Accessed: 2020-01.

CHANDRA, J.; KUHN, D. M.; MUKHERJEE, P. K.; HOYER, L. L.; MCCORMICK, T.; GHANNOUM, M. A. Biofilm formation by the fungal pathogencandida albicans: development, architecture, and drug resistance. *Journal of bacteriology*, Am Soc Microbiol, v. 183, n. 18, p. 5385–5394, 2001.

COLOMBO, A. L.; GUIMARÃES, T. Epidemiologia das infecções hematogênicas por candida spp. *Revista da Sociedade Brasileira de Medicina Tropical*, Sociedade Brasileira de Medicina Tropical-SBMT, 2003.

DEORUKHKAR, S.; SHAHRIAR, R. Identification of candida species: Conventional methods in the era of molecular diagnosis. *Remedy Publications LLC. Annals of Microbiology and Immunology*, v. 1, n. 1, p. 1002, 2018.

DOES, W. P. Polymerase chain reaction. *Journal of Investigative Dermatology*, v. 133, 2013.

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

DUTTA, R.; MORGAN, D.; BAKER, N.; GARDNER, J. W.; HINES, E. L. Identification of staphylococcus aureus infections in hospital environment: electronic nose based approach. *Sensors and Actuators B: Chemical*, Elsevier, v. 109, n. 2, p. 355–362, 2005.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. LTC, v. 1, n. 1, 2011.

GARDNER, J. W.; BARTLETT, P. N. A brief history of electronic noses. *Sensors and Actuators B: Chemical*, 1994.

GONZALEZ, L. d. A. Regressão logística e suas aplicações. Universidade Federal do Maranhão, 2018.

GUTIÉRREZ, J.; HORRILLO, M. Advances in artificial olfaction: Sensors and applications. *Talanta*, Elsevier, v. 124, p. 95–105, 2014.

JONSSON, A.; WINQUIST, F.; SCHNÜRER, J.; SUNDGREN, H.; LUNDSTRÖM, I. Electronic nose for microbial quality classification of grains. *International Journal of Food Microbiology*, Elsevier, v. 35, n. 2, p. 187–193, 1997.

KIM, T. K. T test as a parametric statistic. *Korean journal of anesthesiology*, Korean Society of Anesthesiologists, v. 68, n. 6, p. 540, 2015.

LI, Y.; DU, M.; CHEN, L.-a.; LIU, Y.; LIANG, Z. Nosocomial bloodstream infection due to candida spp. in china: species distribution, clinical features, and outcomes. *Mycopathologia*, Springer, v. 181, n. 7-8, p. 485–495, 2016.

MAGAN, N.; EVANS, P. Volatiles as an indicator of fungal activity and differentiation between species, and the potential use of electronic nose technology for early detection of grain spoilage. *Journal of Stored Products Research*, Elsevier, v. 36, n. 4, p. 319–340, 2000.

MARTINEZ, L.; FERREIRA, A. *Análise de Dados com SPSS*. [S.l.]: Escolar editora, 2007.

MAVOR, A.; THEWES, S.; HUBE, B. P. K.; HOYER, L. L.; MCCORMICK, T.; GHANNOUM, M. A. Systemic fungal infections caused by candida species: epidemiology, infection process and virulence attributes. *Current drug targets*, Bentham Science Publishers, v. 6, n. 8, p. 863–874, 2005.

MGODE, G. F.; WEETJENS, B. J.; NAWRATH, T.; COX, C.; JUBITANA, M.; MACHANG'U, R. S.; COHEN-BACRIE, S.; BEDOTTO, M.; DRANCOURT, M.; SCHULZ, S. et al. Diagnosis of tuberculosis by trained african giant pouched rats and confounding impact of pathogens and microflora of the respiratory tract. *Journal of clinical microbiology*, Am Soc Microbiol, v. 50, n. 2, p. 274–280, 2012.

MORGAN, J.; MELTZER, M. I.; PLIKAYTIS, B. D.; SOFAIR, A. N.; HUIE-WHITE, S.; WILCOX, S.; HARRISON, L. H.; SEABERG, E. C.; HAJJEH, R. A.; TEUTSCH, S. M. Excess mortality, hospital stay, and cost due to candidemia: a case-control study using data from population-based candidemia surveillance. *Infection Control & Hospital Epidemiology*, Cambridge University Press, v. 26, n. 6, p. 540–547, 2005.

NIESTERS, H.; GOESSENS, W.; MEIS, J.; QUINT, W. Rapid, polymerase chain reaction-based identification assays for candida species. *Journal of Clinical Microbiology*, Am Soc Microbiol, v. 31, n. 4, p. 904–910, 1993.

OLSSON, J.; BÖRJESSON, T.; LUNDSTEDT, T.; SCHNÜRER, J. Volatiles for mycological quality grading of barley grains: determinations using gas chromatography–mass spectrometry and electronic nose. *International journal of food microbiology*, Elsevier, v. 59, n. 3, p. 167–178, 2000.

PERSAUD, K.; DODD, G. Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. *Nature*, Nature Publishing Group, v. 299, n. 5881, p. 352, 1982.

PFALLER, M. A.; WOLK, D. M.; LOWERY, T. J. T2mr and t2candida: novel technology for the rapid diagnosis of candidemia and invasive candidiasis. *Future microbiology*, Future Medicine, v. 11, n. 1, p. 103–117, 2016.

PING, W.; YI, T.; HAIBAO, X.; FARONG, S. A novel method for diabetes diagnosis based on electronic nose. *Biosensors and Bioelectronics*, Elsevier, v. 12, n. 9-10, p. 1031–1036, 1997.

QUINDÓS, G. Epidemiology of candidaemia and invasive candidiasis. a changing face. *Revista iberoamericana de micologia*, Elsevier, v. 31, n. 1, p. 42–48, 2014.

RAJU, S. B.; RAJAPPA, S. Isolation and identification of candida from the oral cavity. *ISRN dentistry*, Hindawi Publishing Corporation, v. 2011, 2011.

RAZALI, N. M.; WAH, Y. B.; REIS, G. M. others; JUNIOR, J. R. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, v. 2, n. 1, p. 21–33, 2011.

REIS, G. M.; JUNIOR, J. R. Comparação de testes paramétricos e não paramétricos aplicados em delineamentos experimentais. *Simpósio Acadêmico de Engenharia de Produção*, v. 3, p. 2007–3, 2007.

SHEPPARD, D. C.; LOCAS, M.-C.; RESTIERI, C.; LAVERDIERE, M. Utility of the germ tube test for direct identification of candida albicans from positive blood culture bottles. *Journal of clinical microbiology*, Am Soc Microbiol, v. 46, n. 10, p. 3508–3509, 2008.

TURNER, A. P.; MAGAN, N. Electronic noses and disease diagnostics. *Nature Reviews Microbiology*, Nature Publishing Group, v. 2, n. 2, p. 161, 2004.

WASILEWSKI, T.; GĘBICKI, J.; KAMYSZ, W. Bioelectronic nose: Current status and perspectives. *Biosensors and Bioelectronics*, Elsevier, v. 87, p. 480–494, 2017.

WILSON, A. Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath. *Metabolites*, Multidisciplinary Digital Publishing Institute, v. 5, n. 1, p. 140–163, 2015.

WOJNOWSKI, W.; DYMERSKI, T.; GĘBICKI, J.; NAMIEŚNIK, J. Electronic noses in medical diagnostics. *Current medicinal chemistry*, Bentham Science Publishers, v. 26, n. 1, p. 197–215, 2019.

ZANCHETTIN, C.; LUDERMIR, T. B. Sistemas neurais híbridos para reconhecimento de padrões em narizes artificiais. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, SciELO Brasil, v. 16, n. 2, p. 159–172, 2005.

ZIMMERMAN, D. W.; ZUMBO, B. D. Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks. *The Journal of Experimental Education*, Taylor & Francis, v. 62, n. 1, p. 75–86, 1993.