MATHEUS ISAC DA SILVA

Avaliação da Sobreamostragem de Dados de Voz na Classificação Automática da Doença de Parkinson

GOIÂNIA
2024

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Matheus Isac da Silva

Título do trabalho: Avaliação da Sobreamostragem de Dados de Voz na Classificação Automática da Doença de Parkinson

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Juliana Paula Felix**, **Usuário Externo**, em 19/12/2024, às 11:39, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Matheus Isac Da Silva**, **Discente**, em 19/12/2024, às 11:40, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5054479** e o código CRC **669ECDFF**.

MATHEUS ISAC DA SILVA

Avaliação da Sobreamostragem de Dados de Voz na Classificação Automática da Doença de Parkinson

Trabalho de conclusão de curso apresentado ao curso de Engenharia de Computação, da Escola de Engenharia Elétrica, Mecânica e de Computação, da Universidade Federal de Goiás (UFG), como requisito para obtenção do título de Engenheiro de Computação.

Universidade Federal de Goiás (UFG)
Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)

Orientadora: Profa. Dra. Juliana Paula Félix

GOIÂNIA
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Silva, Matheus Isac da
Avaliação da Sobreamostragem de Dados de Voz na Classificação Automática da Doença de Parkinson [manuscrito] / Matheus Isac da Silva. - 2024.
xxi, 21 f.: il.

Orientador: Profa. Dra. Juliana Paula Félix.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Engenharia da Computação, Goiânia, 2024.

Inclui siglas, abreviaturas, símbolos, gráfico, tabelas.

1. Doença de Parkinson. 2. Aprendizado de Máquina. 3. Diagnóstico. I. Félix, Juliana Paula, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Ao(s) 19 dia(s) do mês de dezembro do ano de 2024 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado "Avaliação da Sobreamostragem de Dados de Voz na Classificação Automática da Doença de Parkinson", de autoria de Matheus Isac da Silva, do curso de Engenharia de Computação, do(a) Escola de Engenharia Elétrica, Mecânica e de Computação da UFG. Os trabalhos foram instalados pelo(a) Dra. Juliana Paula Félix - FEN/UFG, com a participação dos demais membros da Banca Examinadora: Profa. Dra. Karina Rocha Gomes da Silva e Prof. Dr. Rogerio Lopes Salvini suplente Dr. Gelson da Cruz Junior. Após a apresentação, a banca examinadora realizou a arguição do(a) estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 10,0, tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Juliana Paula Felix**, **Usuário Externo**, em 19/12/2024, às 12:59, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Karina Rocha Gomes Da Silva**, **Professor do Magistério Superior**, em 19/12/2024, às 13:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rogerio Lopes Salvini**, **Professor do Magistério Superior**, em 19/12/2024, às 15:13, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5054873** e o código CRC **6A12F6A3**.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, André Isac e Romilda Romana, por serem meu alicerce e por todo o amor, apoio e ensinamentos que me motivaram a persistir nos momentos mais desafiadores. Aos meus irmãos, pela amizade, incentivo constante e por acreditarem em mim.

À minha orientadora, Dra. Juliana Paula Félix, por toda a orientação, paciência, dedicação e apoio ao longo desta jornada acadêmica. Seu conhecimento e incentivo foram fundamentais para o desenvolvimento deste trabalho e para o meu crescimento como estudante e profissional.

Aos meus amigos e colegas de curso, que compartilharam comigo não apenas os desafios da graduação, mas também momentos de aprendizado, troca de experiências e conquistas. A parceria e o companheirismo de vocês tornaram essa jornada mais leve e especial.

Aos professores, que contribuíram para minha formação ao longo da graduação, por toda a dedicação, o entusiasmo e o compromisso de cada para o meu aprendizado. Seus métodos de ensino, muitas vezes inovadores e desafiadores, criaram um ambiente onde, como bem disse Paulo Freire, *"ensinar não é transferir conhecimento, mas criar as possibilidades para sua própria produção ou sua construção"*.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho, meu sincero agradecimento.

Avaliação da Sobreamostragem de Dados de Voz na Classificação Automática da Doença de Parkinson

Matheus Isac da Silva[1] and Juliana Paula Felix[2,3]

[1]Escola de Engenharia Elétrica, Mecânica e de Computação, Universidade Federal de Goiás
[2]Instituto de Informática, Universidade Federal de Goiás
[3]Faculdade de Enfermagem, Universidade Federal de Goiás

***Resumo*—Este estudo investiga um possível viés na sobreamostragem via janelamento de dados dos sinais vocais. Estudos anteriores indicam que para dados de marcha há um viés quando tratados os dados de forma independentes, além disso há estudos estatísticos que mostram que os dados de um mesmo indivíduo carregam informações semelhantes. Foi utilizada uma abordagem baseada em três bases de dados contendo sinais vocais, sendo duas desbalanceadas e uma balanceada. Os algoritmos *K-Nearest Neighbors (KNN)*, *Support VectorMachine (SVM)*, *Linear Discriminant Analysis (LDA)*, *Naive Bayes* e *Decision Tree (DT)* foram aplicados, com pré-processamento utilizando o *StandardScaler* e análise do comportamento do PCA. A validação cruzada foi feita com *k-fold Cross Validation*, com k=5, em todas as 3 bases, adaptada para cenários com e sem viés nos dados de treinamento. Os modelos avaliados sem considerar o viés apresentaram desempenhos inflacionados, enquanto a abordagem rigorosa mostrou resultados mais modestos. Conclui-se que amostras do mesmo indivíduo em treinamento e teste podem inflar o desempenho dos modelos, sendo crucial aplicar sobreamostragem corretamente para desenvolver modelos confiáveis para o diagnóstico de DP.**

***Palavras-chave*—Doença de Parkinson; Aprendizado de Máquina; Diagnóstico.**

***Abstract*—This study investigates a possible bias in oversampling via data windowing of vocal signals. Previous studies indicate that there is a bias for gait data when the data is treated independently, in addition there are statistical studies that show that data from the same individual carry similar information. An approach based on three databases containing vocal signals was used, two of which were unbalanced and one balanced. The K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Naive Bayes and Decision Tree (DT) algorithms were applied, with pre-processing using StandardScaler and PCA behavior analysis. Cross validation was done with k-fold Cross Validation, with k=5, in all 3 bases, adapted for scenarios with and without bias in the training data. Models evaluated without considering bias showed inflated performances, while the rigorous approach showed more modest results. It is concluded that samples from the same individual in training and testing can inflate the performance of models, and it is crucial to apply oversampling correctly to develop reliable models for diagnosing PD.**

***Index Terms*—Parkinson's disease; Machine Learning; Diagnosis.**

I. Introdução

A Doença de Parkinson (DP) é uma doença neurodegenerativa progressiva que afeta a mobilidade, a fala e a postura, causando tremores, rigidez muscular e bradicinesia [1]. A doença é causada pela morte de neurônios, resultando na diminuição dos níveis de dopamina no cérebro e, por sucessão, dificulta a comunicação entre sinapses, que causa deterioramento das funções motoras [2]. A DP tem uma prevalência de aproximadamente 0,5 a 1 por cento entre aqueles com 65 a 69 anos de idade, aumentando para 1 a 3 por cento entre pessoas com 80 anos ou mais [3], sendo a segunda doença neurodegenerativa mais comum depois da doença de Alzheimer [4]. Espera-se que tanto a prevalência como a incidência da DP aumentem em mais de 30% até 2030, com o envelhecimento da população [3].

A maioria dos indivíduos diagnosticados com DP desenvolve distúrbios de voz e fala durante o curso da doença [5]. Volume vocal reduzido, voz monótona e soprosa ou rouca, e articulação imprecisa são as principais características da fala parkinsoniana [6]. Esses distúrbios de voz e fala, denominados coletivamente de disartria hipocinética, podem estar entre os primeiros sinais da DP [7]. Não há cura para a DP, de forma que os pacientes dependem de detecção precoce e tratamentos personalizados para retardar o progresso da doença e assegurar uma melhor qualidade de vida. Neste sentido, dados acústicos têm sido utilizados para descrever as características vocais de indivíduos com DP [7], e são diversos os trabalhos que propõem o uso de aprendizado de máquina para auxiliar no diagnóstico da DP a partir da classificação de sinais de voz.

Devido à raridade da doença, as bases de dados de voz de pessoas com DP disponíveis publicamente geralmente têm um número limitado de participantes. Como solução, muitos estudos utilizam sobreamostragem, coletando múltiplas amostras de um mesmo indivíduo, aumentando assim a representatividade do conjunto total de amostras e permitindo uma análise mais abrangente da população em estudo. Além disso, a alta dimensionalidade das características vocais, muitas vezes redundantes ou altamente correlacionadas, podendo dificultar a performance dos classificadores e levar a resultados enviesados.

Entretanto, estudos como o de Naranjo *et al.* [8] avalia os dados de voz replicados, indicando que os dados de voz de um mesmo sujeito não são idênticos, porém são mais semelhantes entre si, comparado com sujeitos diferentes. Os autores propõem uma melhoria nos estudos envolvendo a classificação de Doença de Parkinson com dados de voz, no qual eram realizados utilizando diferentes aspectos: métodos lineares versus não lineares, tratamento de todos os indivíduos em conjunto ou divisão por sexo, diferentes algoritmos de seleção

de características ou diferentes tipos de esquemas de validação cruzada, fornecendo diferentes precisões. No entanto, a maioria dessas abordagens consideraram as características extraídas das gravações como se fossem independentes [8].

Recentemente, Chagas *et al.* [9] levantou a hipótese de que, ao realizar experimentos de aprendizado de máquina para classificação de doenças neurodegenerativas, como a Doença de Parkinson, amostras obtidas a partir de um mesmo indivíduo não deveriam ser tratadas de forma independente na modelagem e avaliação dos algoritmos de classificação [9], como é frequentemente observado em estudos encontrados na literatura. Essa hipótese é avaliada pelos autores do estudo utilizando-se sinais de marcha de pessoas com a Doença de Parkinson, Doença de Huntington e Esclerose Lateral Amiotrófica, todas doenças neurodegenerativas que possuem como sintoma comum alterações na marcha. Os autores do estudo avaliam a performance de diferentes classificadores considerando dois cenários de avaliação distintos, sendo o primeiro em que as amostras são todas tratadas de forma independente, e o segundo em que a avaliação do classificador considera que amostras distintas de um mesmo indivíduo devem figurar exclusivamente no conjunto de treinamento ou no conjunto de teste, nunca em ambos os conjuntos no mesmo ciclo de avaliação.

Neste trabalho, investigamos se o comportamento observado para sinais de marcha se repete na análise de sinais de voz de pessoas com DP. Três bases de dados de voz foram utilizadas, todas contendo múltiplas amostras por pessoa, permitindo a análise do impacto da sobreamostragem e do possível viés ao uso repetido de amostras de um mesmo indivíduo em modelos de classificação. Além disso, técnicas de redução de dimensionalidade foram empregadas para buscar um melhor resultado dos classificadores.

No restante deste artigo, apresentamos uma breve revisão da literatura. Na sequência, os materiais e métodos são descritos, incluindo as bases de dados utilizadas e posterior esclarecimento sobre a metodologia utilizada. Finalmente, os resultados são apresentados e discutidos, e as conclusões do estudo são apresentadas.

II. Revisão da Literatura

Esta seção, apresentamos alguns estudos de relevância na classificação de sinais de voz para auxiliar no diagnóstico de DP. Os estudos realizados contaram com diferentes métodos de classificação, separações de variáveis e métodos de avaliação. Na Tabela I é possível ter uma melhor visualização dos estudos com os valores de acurácias obtidas e as metodologias utilizadas.

Little et al. (2009)

Um dos estudos pioneiros neste campo foi realizado por Little *et al.*(2009) [7], que avaliaram medidas de disfonia para discriminar pessoas saudáveis de pessoas com DP. O estudo contou com a participação de 31 pessoas, sendo 23 com DP, das quais foram coletadas um total de 195 amostras vocais, com média de 6 amostras por participante. Diversas características foram geradas a partir de cada amostra de sinal de voz, e métodos de seleção de características foram aplicados. Os autores reportam uma acurácia média de $91,4\%$ obtida com uso da Máquina de Vetores de Suporte (SVM) com *kernel* de base radial gaussiana. A base de dados coletada por Little *et al.* [7] foi disponibilizada publicamente no repositório da *University of California Irvine (UCI)*, sob o nome *"Oxford Parkinson's Disease Detection Dataset"*, e vem servindo de base para novos estudos que envolvem aprendizado de máquina. [7]

Os autores desenvolveram um classificador com base em uma *Support Vector Machine (SVM)* e foi iniciado com a filtragem de alguns atributos utilizando-se uma análise de correlação. A analise consistia em analisar os atributos dois a dois e descartar um atributo sempre que a correlação entre eles era superior a 0,95 [7].

Tsanas et al. (2012)

O estudo tem como objetivo testar o quão preciso novos algoritmos podem ser usados para discriminar indivíduos com DP de indivíduos de controle. Para isso, 132 medidas de disfonia a partir de vogais sustentadas coletados de 43 indivíduos, totalizando 263 amostras, essa base de dados desenvolvida pela *National Center for Voice and Speech* (NCVS) não disponível publicamente.

O desenvolvimento selecionou quatro subconjuntos de medidas de disfonia usando quatro algoritmos de seleção, sendo eles: *Least Absolute Shrinkage and Selection Operator (LASSO)*, *Minimum Redundancy Maximum Relevance (mRMR)*, RELIEF e *Local Learning-Based Feature Selection (LLBFS)*. Os algoritmos utilizados nesse trabalho foram o *Random Forest (RF)* e *Support Vector Machine (SVM)*. A validação realizada foi com *Cross-Validation (CV)* utilizando 90% para treino (237 amostras) e 10% para teste (26 amostras) em cada iteração do CV [10].

Os melhores resultados reportados pelos autores foram com a utilização o selecionador de característica RELIEF e o classificador SVM obtendo 98,60% de acurácia, e também com o RELIEF e o classificador *Random Forest* obtendo 93,50% de acurácia.

Sakar et al. (2019)

Os autores apresentaram um sistema de classificação de Parkinson baseado na voz utilizando a transformada *wavelet Q-factor* ajustável (TQWT). Para isso, o estudo contou com a participação de 252 pessoas, sendo 188 indivíduos com DP e 64 indivíduos de controle, das quais foram coletadas 3 amostras de cada participante, totalizando um total de 756 amostras vocais. A base de dados desenvolvida por esse estudo foi disponibilizada publicamente no *UC Irvine Machine Learning Repository*, sob o nome de "*Parkinson's Disease Classification*". O sistema é avaliado a partir da validação cruzada *leave-one-subject-out*(LOSO). O estudo utilizou *Support Vector Machine*(SVM) com *kernel* linear e *radial bases function*, *Multilayers Perceptron*, *Naive Bayes*, Regressão logística, *Random Forest* e *K Nearest Neighbors* (KNN). Os resultados do treinamento foram [11]:

Table I: Revisão da Literatura.

Referência	Ano	Base utilizada	Seleção de características	Forma de Avaliação	Modelo de AM	Acurácia
Little *et al.*[7]	2009	Base Little *et al.*	-	*bootstrap resampling* com 50 réplicas	SVM	$91,4\% \pm 4,4\%$
Tsanas et al. [10]	2012	NCVS	RELIEF	CV (90% para treino e 10% para teste)	Random Forest	93,50%
			RELIEF	CV (90% para treino e 10% para teste)	SVM	98,60%
Sakar et al. [11]	2019	Base Sakar *et al.*	-	Leave one subject out	Logistic regression	79,00%
			-		SVM	84,00%
			-		Multilayer	78,00%
Aich et al. [12]	2019	Base Little *et al.*	GA	70% para treino e 30% para teste	SVM	97,57%
Solana et al. [13]	2020	Base Sakar *et al.*	Wrappers feature subset selection	-	MLP	86,64%
				-	SVM	94,70%
				-	Random Forest	92,20%
Sharanyaa et al. [14]	2020	Base Little *et al.*	-	-	KNN	90,20%
					Random Forest	87,27%
Ouhmida et al. [15]	2021	Base Little *et al.*	-	85% para treino e 15% para teste	CNN	93,10%
		Base Prez	-	85% para treino e 15% para teste	CNN	88,89%
		Base Little *et al.*	-	85% para treino e 15% para teste	ANN	82,76%
		Base Prez	-	85% para treino e 15% para teste	ANN	72,22%
Rana et al. [16]	2022	Little *et al.*	-	80% dados para treino e 20% para teste	ANN	96,70%
			-	80% dados para treino e 20% para teste	SVM	87,17%
			-	80% dados para treino e 20% para teste	Naive Bayes	74,11%
			-	80% dados para treino e 20% para teste	KNN	87,17%
Govindu e Palwe [17]	2023	Base Little *et al.*	-	75% dados para treino e 25% para teste	Random Forest com Standard Scaler	91,83%
			-	75% dados para treino e 25% para teste	KNN com Standard Scaler	91,83%
			PCA	75% dados para treino e 25% para teste	SVM com Standard Scaler	91,75%
Melo e Gouveia [18]	2023	Base Little *et al.*	Baseado no Little et al, e F0	5-fold CV	Random Forest	$93,80\% \pm 4,4\%$
			F0, D2, RPDE, DFA, PPE	5-fold CV	Random Forest	$92,30\% \pm 4,7\%$
Soliman et al [19]	2024	Base Sakar *et al.*	Select K-Best	-	Ensemble Bagging	92,47%
			-	-	Ensemble Bagging	91,59%

Abreviaturas: ANN: Artificial Neural Network; CNN: Convolutional Neural Network; CV: Cross-Validation; GA: Algoritmos Genéticos; KNN: K-Nearest Neighbors; LDA: Linear Discriminant Analysis; MLP: Multilayer Perceptron; PCA: Principal Component Analysis; SVM: Support Vector Machine.

1) 84% de acurácia com $0,83$ F1-Score e $0,54$ MCC obtido alimentando MFCCs para o classificador SVM-RBF.
2) 85% de acurácia com $0,84$ F1-Score e $0,57$ MCC obtido alimentando recursos de transformada *wavelet* Q-normal com o fator Q relativamente alto (selecionado como 2) para o classificador perceptron multicamadas.
3) 86% de acurácia com $0,84$ F1-Score e $0,59$ MCC obtido alimentando os 50 principais recursos selecionados pelo mRMR para o classificador SVM-RBF.

Os resultados reportados por Sakar *et al.* [11] concluem que as características do TQWT são eficazes na discriminação de pacientes com DP de indivíduos saudáveis e podem ser usadas em sistemas de telediagnóstico de DP baseados em disfonia [11].

Aich et al. (2019)

Neste trabalho, os autores apresentaram uma abordagem de aprendizado de máquina supervisionado, com a utilização da base de dados desenvolvida por Sakar *et al.*, para distinguir dados de voz de pessoas com DP de indivíduos saudáveis. Uma abordagem de seleção de características utilizando Análise de Componentes Principais (PCA) e Algoritmo Genético (GA) foi utilizada. Os resultados são avaliados utilizando-se uma abordagem de separação de 70% para treino e 30% para

teste. Os autores reportam uma acurácia média de 97,57% no conjunto de teste utilizando Máquinas de Vetores de Suporte (SVM) alimentadas por características selecionadas pelo algoritmo genético [12].

Solana et al. (2020)

Solana *et al.* propuseram um método para detecção de doença de Parkinson (DP) com base em recursos vocais. O método envolve a seleção de subconjuntos de recursos, classificação e avaliação de desempenho de detecção. Para a seleção de subconjuntos foi utilizado o algoritmo *Wrappers* com a finalidade de selecionar os melhores recursos para cada classificador [13].

Nesse estudo, realizado na base de dados de Sakar *et al.*, foram utilizados 4 tipos de classificadores, (*K-Nearest Neighbor*, *Multi-Layer Perceptron*, *Support Vector Machine* e *Random Forest*), sendo que o melhor resultado obtido foi com o SVM com uma precisão de $94,70\%$, sensibilidade de $98,40\%$, especificidade de $92,68\%$ [13].

Sharanyaa et al. (2020)

O estudo teve o objetivo de classificar a doença de Parkinson usando características de voz em pacientes com DP e sem DP. Os dados utilizados foram da base desenvolvida por Little *et al.*, dos quais foram realizados pré-processamentos, como a normalização, que é usada para alterar os atributos da coluna que possuem valores numéricos para um grau comum de escala, e a padronização, que é útil quando há valores discrepantes nos dados [14].

Os modelos de aprendizado de máquina utilizados foram o *Naive Bayes*, Regressão Logística, *K-Nearest Neighbors* e *Random Forest*. Os autores reportaram uma maior acurácia com o *K-Nearest Neighbors* com 92,20% e com o *Random Forest* com 87,20%. Concluíram que os modelos não paramétricos forneceram maior precisão de classificação no diagnóstico da doença de Parkinson com base em características de voz, sugerindo sua eficácia para auxiliar no diagnóstico automatizado [14].

Ouhmida et al. (2021)

Neste trabalho, Ouhmida *et al.* realizaram um método para classificação de sinais acústico de pessoas com DP e pessoas saudáveis com duas bases de dados disponíveis pela *UCI Machine Learning repository databases*, sendo as bases de dados desenvolvida por Little *et al.* [7] e por Prez [20]. Os autores utilizaram para classificação da doença de Parkinson as Redes Neurais Artificiais (ANN) e Redes Neurais Convolucionais (CNN) em cada base separadamente [15].

Para a base de dados de Little *et al.*, o CNN obteve uma acurácia de 93,10% (sendo o melhor resultado obtido pelo estudo), enquanto o ANN obteve 82,76%. Para a base de dados de Prez, o CNN obteve uma acurácia de 88,89%, enquanto o ANN obeteve 72,22% de acurácia. Os resultados concluíram que o CNN teve um desempenho melhor do que o ANN para ambas bases de dados no processo de classificação [15].

Rana et al. (2022)

Os autores apresentaram um estudo comparativo entre quatro algoritmos de aprendizado de máquina capazes de classificar DP a partir dos sinais de voz. Para isso o estudo contou com a base dados utilizada por Little *et al.*. O estudo teve a comparação dos classificadores, *Support Vector Machine (SVM)*, Naive Bayes, *K-Nearest Neighbor (KNN)* e *Artificial Neural Network (ANN)*, em relação a valores obtios pelo treinamento e teste como Acurácia. F1-Score, *Matthews Correlation Coefficient (MCC)*, Sensibilidade e Especificidade. [16]

Os autores reportaram que o melhor classificador foi o ANN em relação aos valores de Acurácia, F1-Score, MCC, Sensibilidade e Especificidade que foram, 96,70% 87.01% 70.11% 92.42% 91.25%, respectivamente. Uma igualidade também foi reportado pelos autores dos valores reportados pelos classificadores SVM e KNN, com 87,17% de acurácia. [16]

Govindu e Palwe (2023)

Utilizou a base de dados desenvolvida por Little *et al.* com classificadores com base em *Logistic Regression*, *Radom Forest Classifier*, *Support Vector Machine (SVM)* e *K Nearest Neighbors (KNN)*. Com os classificadores foram utilizados 3 abordagens [17]:

1) base de dados completa (195 gravações e 22 características)
2) utilização do *Principal Component Analysis (PCA)*(195 gravações e 5 atributos)
3) técnica de balanceamento (109 gravações e 22 características).

Na primeira abordagem, o melhor resultado com o *Random Forest* com $91,83\%$. Na segunda abordagem, o melhor resultado foi com *Support Vector Machine (SVM)*, com acurácia de $91,75\%$. [17]

Melo e Gouveia (2023)

Melo e Gouveia desenvolveram a construção de uma ferramenta de auxílio no diagnóstico da DP utilizando sinais de voz associados a técnicas de Aprendizado de Máquina. Os autores fizeram o uso do conjunto de dados elaborado por Little *et al.* (2009), com a base de dados foi iniciado a seleção de atributos, divido em 3 etapas. A primeira etapa consiste numa análise exploratória, observando dispersão, média e presença de instâncias discrepantes. A segunda etapa consiste no uso de testes de hipóteses utilizando a distribuição t com um intervalo de confiança de 99%. Por fim, é feita uma análise de correlação dos atributos [18].

Foram definidas 6 conjunto de dados para o treinamento, com a utilização do classificador *Random Forest (RF)* e a utilização da ferramenta de avaliação *k-fold* com $k = 5$ as bases foram [18]:

- Base 1: Base completa.
- Base 2: Apenas atributos não lineares.
- Base 3: Apenas atributos tradicionais.
- Base 4: Apenas atributos não lineares e F0.
- Base 5: Base de Litte et al.

- Base 6: Base de Litte et al. e F0

Os autores obtiveram uma acurácia de $93,80\%$ com a Base 6 e uma acurácia de 92.30% com a Base 4. [18].

Soliman et al. (2024)

Nesse artigo, Soliman *et al.* [19] tiveram o objetivo de criar e encontrar o melhor modelo que determina se um indivíduo é diagnosticado com DP com base nas características extraídas da fala do indivíduo. Com a utilização da base de dados desenvolvida por Sakar *et al.* os autores realizaram pré-processamento do projeto com uma normalização Min-Max, ajudando a garantir que cada característica contribua igualmente. Foram utilizados técnicas de amostragem para sobreamostragem (SMOTE) e para subamostragem (RandomUserSampler), com o intuito de balancear as quantidades de indivíduos de controle e indivíduos com DP.

Com a utilização do selecionador de características *selectKBest* foram selecionadas as 100 melhores *features* da base de dados. Os classificadores utilizados nesse estudo foram *Logistic Regression (LR)*, *Support Vector Machine (SVM)*, *Decision Tree (DT)*, *Naive Bayes (NB)*, *Ensemble Bagging*, *Ensemble Boosting* e o *Random Forest (RF)*.

Os autores reportaram que após a subamostragem do conjunto de dados e com a seleção de *features*, foi obtido acurácia de 92,47% usando o *Ensemble Bagging* e os dados de subamostragem. Além disso, os autores também reportaram que, após a sobreamostragem dos dados e sem a seleção de *features* obtiveram acurácia de 91,59% usando o *Ensemble Bagging*.

III. Materiais e Métodos

Nesta seção são apresentados detalhadamente as bases de dados utilizadas, os métodos de pré-processamento aplicados, as estratégias de tratamento dos dados, incluindo a análise da sobreamostragem de dados de voz e seu impacto na classificação automática, os modelos de classificadores empregados e a técnica de validação adotada neste estudo. O fluxo completo dessas etapas está ilustrado na Figura 1, que apresenta o fluxograma do processo realizado.

A. Bases de dados

Neste trabalho, foram utilizadas as bases de dados desenvolvida por Little *et al.* [7], Sakar *et al.*(2019) [11] e por Prez (2019) [20]. As três bases não somente contém dados de voz replicados, ou seja, diferentes sinais de voz obtidos de um mesmo indivíduo, mas também possuem semelhanças nas características extraídas disponíveis. As informações da quantidade de indivíduos, separadas por gênero e grupo de controle e grupo de DP, das bases de dados utilizadas são apresentadas na Tabela II. Mais detalhes do processo de coleta, quantidade de características e faixa etária de cada base de dados são listados a seguir.

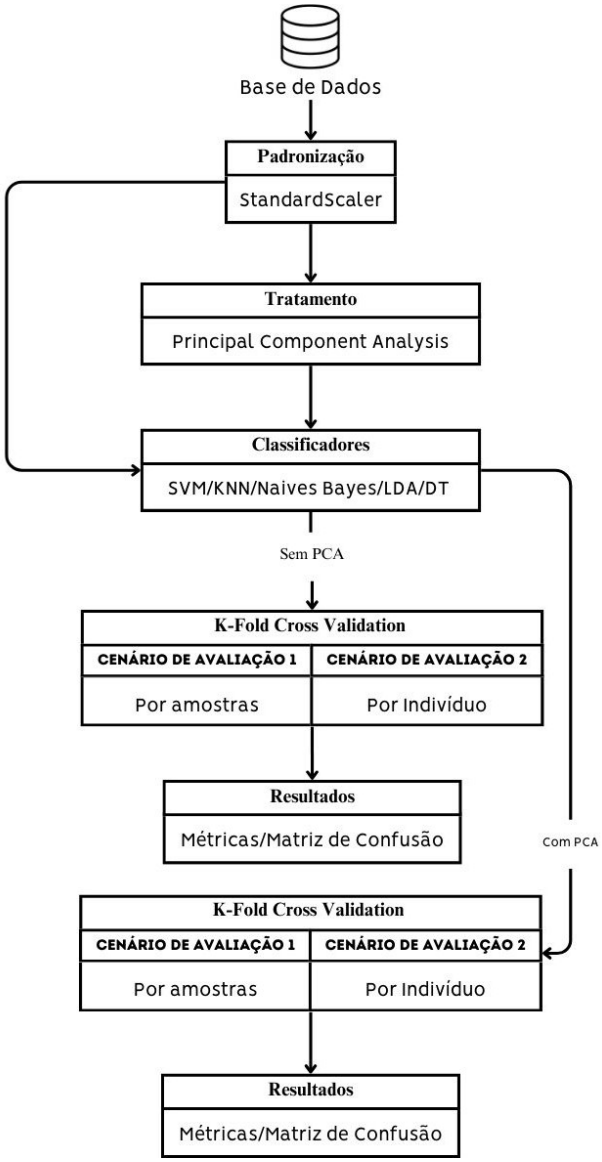


Figure 1: Fluxograma do Processo de Classificação e Avaliação de Dados Vocais.

Table II: Quantidade de amostras por base de dados.

		DP	CO	Total
Little *et al.*	Homens	16	3	19
	Mulheres	7	5	12
	Total	23	8	31
Sakar *et al.*	Homens	81	23	104
	Mulheres	107	41	148
	Total	188	64	252
Prez	Homens	22	27	49
	Mulheres	18	13	31
	Total	40	40	80

Abreviaturas: CO: Controle; DP: Doença de Parkinson.

Base de Little et al.: A base de dados *"Oxford Parkinson's Disease Detection Datase"*, foi desenvolvida por Little *et al.* (2009) [7] e disponibilizada publicamente no *UC Irvine Machine Learning Repository*. Este *dataset* contém informações vocais coletadas de 31 pessoas (faixa etária entre 46 e 85

anos, com média de idade de $65,8 \pm 9,8$), sendo 23 com DP (16 homens e 7 mulheres), e 8 CO (3 homens e 5 mulheres). De cada participante, obteve-se uma média de 6 áudios sustentados (sendo que alguns chegaram a 7 gravações), totalizando em 195 amostras. [7]

Cada fonação foi gravada em uma cabine acústica da *Industrial Acoustics Company (IAC)*, utilizando um microfone *head-mounted (AKG C450)* posicionado a 8 cm dos lábios. Foi colocado um medidor de nível sonoro classe 1 (B&K 2238) a 30cm do alto-falante. As gravações foram feitas pelo hardware*Computerized Speech Laboratory (CSL) 4300B (Kay Elemetrics)*, amostradas em 44,1 kHz com resolução de 16 *bits*. Os dados acústicos disponíveis nessa base são dados já calculados, tratados e preparados de acordo com o apresentado na Tabela III. No total, 22 características (*features*) estão disponíveis.

Base de Sakar et al.: A base de dados *"Parkinson's Disease Classificatio"*, foi desenvolvida por Sakar *et al.* (2019) [11] e disponibilizada publicamente no *UC Irvine Machine Learning Repository*. Os dados disponíveis foram coletados no *Department of Neurology in Cerrahpaşa Faculty of Medicine* da *Istanbul University*. Este *dataset* contém informações vocais coletadas de 252 pessoas (faixa etária entre 33 e 87 anos, com média de idade de $65,1 \pm 10,9$), sendo 188 com DP (81 homens e 107 mulheres), e 64 CO (23 homens e 41 mulheres). De cada participante, obteve-se 3 áudios sustentados, totalizando em 756 amostras. [11]

Durante a coleta de dados, foi usado um microfone ajustado para 44,1 KHz e após exame médio foi coletada de três repetições de fonações sustentada da vogal /a/, totalizando 756 gravações com 754 características da voz, que são descritas na Tabela IV [11].

Base de Prez: A base de dados *"Parkinson Dataset with replicated acoustic feature"*, foi desenvolvida por Prez (2019) [20] e disponibilizada publicamente no *UC Irvine Machine Learning Repository*. Este *dataset* contém informações vocais coletadas de 80 pessoas (faixa etária acima de 50 anos, com média de idade de $67,98 \pm 8,10$), sendo 40 com DP (22 homens e 18 mulheres), e 40 CO (27 homens e 13 mulheres). De cada participante, obteve-se 3 áudios sustentados, totalizando em 240 amostras [20].

Cada indivíduo foi submetido a três gravações de pelo menos 5 segundo, durante o qual deviam sustentar a fonação da vogal /a/. No total, obteve-se 240 gravações com 44 *features*. Estas características acústicas foram divididas em cinco famílias, apresentadas na Tabela V. A gravação foi executada por um computador portátil com placa de som externa (*TASCAM US322*) e microfone de cabeça (*AKG 520*) com padrão cardioide. A gravação digital foi realizada com taxa de amostragem de $44,1KHz$ e resolução de 16 bits/amostra utilizando o *software Audacity*(versão 2.0.5) [20].

B. Pré-processamento

Os dados disponíveis nas bases de dados possuem características com escalas muito diferentes. Considerando que os algoritmos de *Machine Learning*, em geral, não funcionam bem quando os atributos de entrada têm escalas diferentes, realizamos uma técnica de padronização. Existem vários métodos de escalas de padronização [21]. Neste trabalho foi utilizado o *StandardScaler*.

StandardScaler: O *StandardScaler*, sendo um dos requisitos comum para muitos estimadores de *Machine Learning* [22], é um método de escala de padronização dos dados, passando por duas etapas, primeiramente é realizada a subtração do valor médio, fazendo com que os valores padronizados sempre tenham uma média igual a zero, em seguida é realizada a divisão pelo desvio padrão para que a distribuição resultante tenha variância unitária, definida pela Equação 1 [21].

$$Padronizado = \frac{(Amostra - Media)}{DesvioPadrao}. \quad (1)$$

Neste artigo, utilizou-se o *StandardScaler* em todas as análises, com o objetivo de padronizar os dados. Essa padronização é essencial para ajustar as variáveis a uma escala comum, garantindo que nenhuma característica domine as demais devido à diferença em suas magnitudes. Essa abordagem é particularmente relevante, pois melhora a eficiência do treinamento e a convergência, além de contribuir para resultados mais consistentes e comparáveis entre diferentes modelos de aprendizagem de máquina.

C. Principal component analysis (PCA)

Observou-se que as bases de dados analisadas possuem alta dimensionalidade, frequentemente contendo redundâncias, variáveis correlacionadas e ruídos, fatores que podem prejudicar o desempenho dos modelos. Para resolver esses problemas foi aplicado a técnica do *Principal component analysis (PCA)* que decompõe dados de alta dimensão em um componente de subespaço de baixa dimensão e um componente de ruído. Essa decomposição é útil para compressão de dados, bem como para remoção de ruído, tornando-a uma primeira etapa comum para muitas tarefas de processamento de dados. As aplicações dessa transformação tem aplicação que incluem compressão de dados, análise de imagens, visualização, reconhecimento de padrões, regressão e previsão de séries temporais. [23]

Neste artigo, foram reportados os resultados obtidos com e sem a aplicação da transformação, com o objetivo de avaliar o efeito da redução na classificação dos dados a redução da dimensionalidade. A transformação aplicada visa capturar 95% da variância dos dados, simplificando o espaço de características.

D. Classificadores

Os classificadores utilizados foram *K-Nearest Neighbours (KNN)*, *Linear Discriminant Analysis (LDA)* , *Support Vector Machine (SVM)*, *Naive Bayes* e *Decision Tree (DT)*. Esses cinco classificadores são bastante utilizados na literatura,

Table III: Características da Base de Dados do Little *et al.*

Tipos de Features	**Features**	**Significado**
Recursos básicos	MDVP:F0 (Hz)	Média das Frequências da voz
	MDVP:Fhi (Hz)	Pico da Frequência de voz.
	MDVP:Flo (Hz)	Menor valor da Frequência de voz.
Medidas de Variação na Frequência	MDVP:Jitter (%)	Diferença média absoluta entre períodos consecutivos, dividida pelo período médio.
	MDVP:Jitter (Abs)	Diferença média absoluta entre períodos consecutivos, em segundos.
	MDVP:RAP (%)	Perturbação Média Relativa, a diferença média absoluta entre um período e a média dele e dos seus dois vizinhos, dividida pelo período médio.
	MDVP:PPQ (%)	Quociente de Perturbação do Período de cinco pontos, a diferença média absoluta entre um período e a média dele e dos seus quatro vizinhos mais próximos, dividida pelo período médio.
	Jitter:DDP	Diferença média absoluta entre diferenças consecutivas entre períodos consecutivos, dividida pelo período médio.
Medidas de Variação na Amplitude	MDVP:Shimmer (%)	Diferença média absoluta entre as amplitudes de períodos consecutivos, dividida pela amplitude média.
	MDVP:Shimmer (dB)	Logaritmo médio absoluto de base 10 da diferença entre as amplitudes de períodos consecutivos, multiplicado por 20.
	MDVP:APQ	Quociente de Perturbação de Amplitude de onze pontos, a diferença média absoluta entre a amplitude de um período e a média das amplitudes de seus vizinhos, dividida pela amplitude média.
	Shimmer:APQ3	Quociente de Perturbação de Amplitude de três pontos, a diferença média absoluta entre a amplitude de um período e a média das amplitudes de seus vizinhos, dividida pela amplitude média.
	Shimmer:APQ5	Quociente de Perturbação de Amplitude de cinco pontos, a diferença média absoluta entre a amplitude de um período e a média das amplitudes dele e de seus quatro vizinhos mais próximos, dividida pela amplitude média.
	Shimmer:DDA	Diferença média absoluta entre as amplitudes de períodos consecutivos.
Medidas da Proporção entre Ruído e Componentes Tonais na Voz	NHR	Relação ruído-harmônicos.
	HNR	Razão Harmônica-Ruído
Medidas de Complexidade Dinâmica não Linear	RPDE	Medida de entropia de densidade do período de recorrência
	D2	Dimensão de correlação
Expoente de Escala Fractal de Sinal	DFA	Análise de flutuação de tendência
Medidas não Lineares de Variação de Frequência Fundamental	Spread 1	Duas medidas não lineares da frequência fundamental
	Spread 2	Variação de frequência
	PPE	Entropia de período de pitch

MDVP: Programa de Voz Mutidimensional

com destaque para o SVM e KNN, que aparecem com frequência nos trabalhos discutidos na revisão da literatura. As configurações de cada classificador utilizado estão dispostas na Tabela VI.

K-Nearest Neighbours (KNN): O algoritmo de classificação *K-Nearest Neighbours*, ou KNN, é um dos algoritmos de aprendizado de máquina mais simples. Ele consiste na disposição dos dados em um espaço dimensional n, sendo n o número de atributos em questão sendo analisados. Para inferir a classe de um novo dado amostrado, necessitamos dispô-lo neste espaço e contabilizar as classes dos dos "k" (coeficiente a determinar) elementos mais próximos, baseados em uma métrica de distância a ser definida. A classe majoritária determinará a classe do novo dado analisado [24].

Neste trabalho, analisamos o desempenho do KNN, considerando o $k = 5$. A variação do algoritmo considera pesos iguais ao inverso da distância entre os dois pontos, ou seja, os demais vizinhos terão menor influência do que os vizinhos mais próximos de um ponto de consulta [25], e conta com o armazenamento da árvore (*'leaf_size'*=100). A distância utilizada nos algoritmos é a distância euclidiana [26], tal que para dois pontos de coordenadas (x_1, y_1) e (x_2, y_2) no plano cartesiano, pode ser calculada conforme a Equação (2).

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{2}$$

Linear Discriminant Analysis (LDA): O *Linear Discriminant Analysis (LDA)* é um método fundamental de análise de dados originalmente proposto por R. Fisher para discriminar entre diferentes tipos de flores, que consiste em encontrar o hiperplano de projeção, podendo ser usado para classificação, redução de dimensionalidade e para interpretação da importância dos recursos fornecidos, que minimiza a variância interclasse e maximiza a distância entre as médias projetadas das classes. Uma das vantagens da LDA é que a solução pode ser obtida resolvendo um sistema de autovalor generalizado. Isso permite um processamento rápido e massivo de amostras de dados [27].

Em sua configuração há diferentes tipos de solucionadores disponíveis, sendo eles:

- ***Singular Value Decomposition (svd)***: Decomposição de valor singular, sendo o mais recomendado para dados com um grande número de recursos.

Table IV: Características da Base de Dados de Sakar *et al.*

Tipos de Features	Features	Significado	Quantidade de Features
Recursos básicos	Variantes do *Jitter*	São empregados para capturar as instabilidades secundárias no padrão oscilante das pregas vocais e esse subconjunto de características quantifica as mudanças ciclo-ciclo na frequência fundamental.	5
	Variantes do *Shimmer*	São empregados para capturar instabilidades do padrão oscilante das pregas vocais, mas desta vez este subconjunto de recursos quantifica as mudanças ciclo a ciclo na amplitude.	6
	Parâmetros de frequência fundamentais	A frequência de vibração das pregas vocais. Foram utilizados média, mediana, desvio padrão, valores mínimo e máximo.	5
	Parâmetros de Harmonia	Os parâmetros Harmônicos para Ruído e Ruído para Harmônicos, que quantificam a relação entre as informações do sinal e o ruído, foram usados como recursos.	2
	Entropia de densidade do período de recorrência (RPDE)	Fornece informações sobre a capacidade das pregas vocais de sustentar oscilações estáveis das pregas vocais e quantifica a desvios de F_0.	1
	Análise de Flutuação Detendida (DFA)	Quantifica a auto similaridade estocástica do ruído turbulento.	1
	Entropia do período de pitch (PPE)	Mede o controle prejudicado da frequência fundamental F_0 usando escala logarítmica.	1
Características de tempo de Frequência	Parâmetros de Intensidade	A intensidade está relacionada com a potência do sinal de fala em dB. Foram utilizados valores de intensidade média, mínima e máxima	3
	Frequências de formantes	Frequências amplificadas pelo trato vocal, os primeiros quatro formantes foram usados como recursos.	4
	Bandwidth	A faixa de frequência entre as frequências dos formantes, as primeiras quatro larguras de banda foram empregadas como recursos.	4
Coeficientes Cepstrais de Frequência Mel	MFCCs	São empregados para capturar os efeitos BAD no trato vocal separadamente das pregas vocais.	84
Recursos baseados na transformação Wavelet	Recursos de transformada wavelet (WT) relacionados a F_0	Recursos WT quantificam os desvios em F_0	182
Recursos de dobra vocal	Quociente da Glote (GQ)	Fornece informações sobre as durações de abertura e fechamento da glote. É uma medida de periodicidade nos movimentos da glote.	3
	Excitação glótica para ruído (GNE)	Quantifica a extensão do ruído turbulento, causado pelo fechamento incompleto das pregas vocais, no sinal de fala.	6
	Proporção de excitação das pregas vocais (VFER)	Quantifica a quantidade de ruído produzido devido à vibração patológica das pregas vocais usando conceitos de energia não linear e entropia	7
	Decomposição do modo empírico (EMD)	Decompõe um sinal de fala em componentes de sinal elementares usando funções de base adaptativas e valores de energia/entropia obtidos desses componentes são usados para quantificar o ruído.	6
Transformada wavelet de fator Q (TQWT)	TQWT	É um *discrete wavelet transform (DWT)* aprimorado que é fácil de ajustar as características de oscilação da *wavelet*. A função de base *wavelet* do TQWT pode corresponder melhor ao comportamento de oscilação dos sinais de voz.	432

- ***Least Squares QR (lsqr)***: Solução de mínimos quadrados, pode ser combinado com estimativa de encolhimento ou covarância personalizada.
- **Eigenvalue Decomposition (eigen)**: Decomposição de autovalor, pode ser combinado com estimativa de encolhimento ou personalizada.

Neste trabalho foi utilizada o solucionador *Singular Value Decomposition* por ser ideal para alta dimensionalidade de dados, sendo as bases de dados com um número grande de características em comparação com o número de amostras.

***Support Vector Machine (SVM)**:* A Máquina de Vetores de Suporte, ou *Support Vector Machine (SVM)*, é um modelo classificador bastante utilizado na literatura, sendo capaz de fornecer resultados precisos e altamente robustos. O objetivo desse modelo é classificar dados de treinamento separando as classes [17]. Uma das vantagens do SVM é sua versatilidade, possibilitando o uso de diferentes *kernels*, que podem ser especificados para a função de decisão. O *kernel* utilizado neste trabalho foi o "Linear", que define uma fronteira linear a partir de dados linearmente separáveis, separando tais dados com um hiperplano definido pela Equação (3),

$$f(x) = wx + b = 0, \tag{3}$$

onde w é o vetor de pesos perpendicular ao hiperplano de separação, b é um escalar e x é um objeto do conjunto de treinamento [26].

***Naive Bayes (NB)**:* O *Naive Bayes* (NB) é um algoritmo de aprendizagem de classificação simples e eficaz. O NB, tem como base o teorema de Bayes, que determina a probabilidade de um evento ocorrer dependendo das circunstâncias. Este classificador pode ser usado no método estatístico para classificação e método de Aprendizado Supervisionado, sendo facilmente implementado. Ele requer um pequeno conjunto de dados de treinamento prático para julgar uma quantidade

Table V: Características da Base de Dados de Prez.

Tipos de Features	Features	Significado	Quantidade de Features
Recursos básicos	Variantes do Jitter	São empregados para capturar as instabilidades secundárias no padrão oscilante das pregas vocais e esse subconjunto de características quantifica as mudanças ciclo-ciclo na frequência fundamental.	4
	Variantes do Shimmer	São empregados para capturar instabilidades do padrão oscilante das pregas vocais, mas desta vez este subconjunto de recursos quantifica as mudanças ciclo a ciclo na amplitude.	5
	Parâmetros de harmonia	Os parâmetros Harmônicos para Ruído e Ruído para Harmônicos, que quantificam a relação entre as informações do sinal e o ruído, foram usados como recursos.	5
	Entropia de densidade do período de recorrência (RPDE)	Fornece informações sobre a capacidade das pregas vocais de sustentar oscilações estáveis das pregas vocais e quantifica a desvios de F0.	1
	Análise de Flutuação Detendida (DFA)	Quantifica a autosimilaridade estocástica do ruído turbulento.	1
	Entropia do período de pitch (PPE)	Mede o controle prejudicado da frequência fundamental F0 usando escala logarítmica.	1
Recursos de dobra vocal	Excitação glótica para ruído (GNE)	Quantifica a extensão do ruído turbulento, causado pelo fechamento incompleto das pregas vocais, no sinal de fala.	1
Coeficientes Cepstrais de Frequência Mel	MFCCs	São empregados para capturar os efeitos BAD no trato vocal separadamente das pregas vocais.	26

Table VI: Classificadores Utilizados e suas Configurações.

Classificador	Parâmetros
KNN	`n_neighbors=5, weights='distance', leaf_size = 100`
LDA	`solver = 'svd'`
SVM	`kernel='linear'`
Naive Bayes	`--`
DT	`min_samples_split=2, criterion='gini', min_samples_leaf=1, min weight_fraction_leaf=0, max_leaf_nodes= None, max_features=None, random_state=0`

padrão que satisfaça um conjunto específico de equações [28]. Os resultados em sua maioria são bons quando aplicado este classificador.

Para estimar a probabilidade da condição médica, os dados compreendem inúmeras variantes de sinal de fala. O algoritmo *Gaussian Naive Bayes sklearn* é usado para fornecer o módulo classificador para a execução da categorização de *Naives Bayes*.

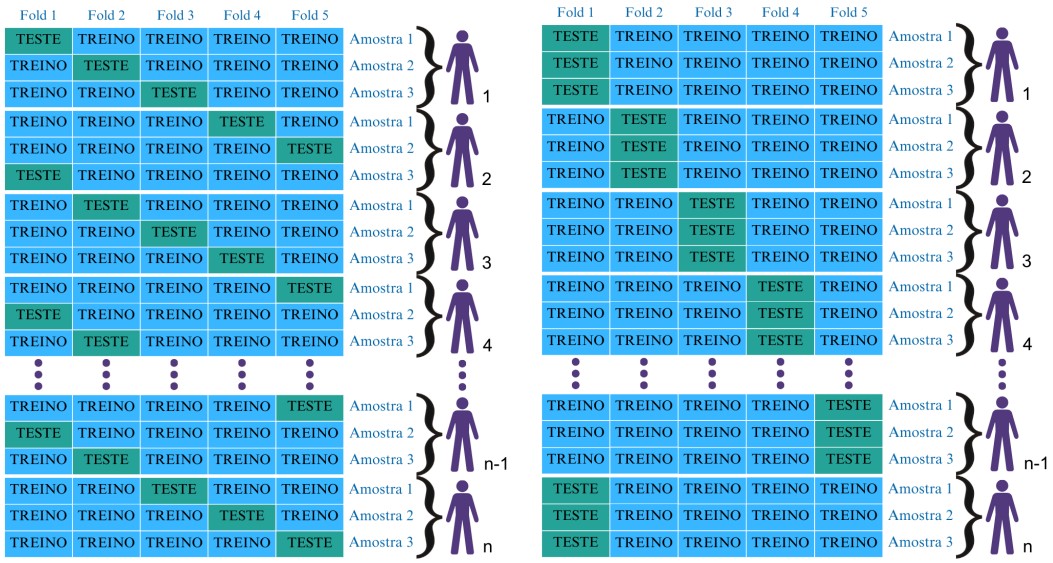
Decision Tree (DT)*: Decision Tree*, ou DT, são algoritmos versáteis de *Machine Learning* que podem executar tarefas de classificação e regressão, e até mesmo tarefas *multioutput* [21]. Nesse trabalho foi utilizada a *Decision Tree* para classificação.

Em sua configuração podemos observar varias configurações que podem ser realizadas e um delas é o parâmetro que foi definida como padrão "gini", que tende a isolar a classe mais frequente em seu próprio ramo da árvore. Outros hyperparâmetros também são definidos, como: *min samples split* (o número mínimo de amostras que um nó deve ter antes de poder ser dividido) configurado por padrão como 2, *min samples leaf* (o número mínimo de amostras que um nó folha deve ter) configurado por padrão como 1, *min weight fraction leaf* (o mesmo que *min samples leaf*, mas expresso como uma fração do número total de instâncias ponderadas) definida como 0, *max leaf nodes* (número máximo de nós folha) foi definida como *None* para que não restringisse a quantidade de nós e *max features* (número máximo de recursos que são avaliados para divisão em cada nó) também foi definido como *None* para que não limitasse a quantidade máxima de recursos [21]. Além destes, foi configurado o *random_state* como 0 para evitar a randomização dos resultados.

E. Avaliação dos métodos

1) ***K-Fold Cross-Validation****:* Para determinar o desempenho de cada classificador, geralmente um modelo é treinado com os dados disponíveis. Em seguida, o desempenho da classificação é avaliado usando dados recém-coletados. Quando não há a disponibilidade de dados novos específicos para a fase de teste, uma parte do conjunto de dados original é separado para a fase de teste. Para superar limitações como o tamanho do banco de dados, o desequilíbrio de dados e a possibilidade de *overfitting*, ao invés de treinar um modelo fixo apenas uma vez, como em uma divisão de treinamento/teste, a abordagem de *cross-validation* (CV) é fortemente recomendada [29].

Na abordagem conhecida como *k-fold cross-validation*, a avaliação do modelo é realizada por k vezes, cada vez usando um particionamento diferente dos dados em conjuntos de treinamento e teste, sendo reportado a média dos resultados obtidos para as k dobras da CV. Neste trabalho, utilizamos o caso especial do *k-fold cross-validation* com $k = 5$, em que o conjunto é divido em 5 partes (*folds*) de tamanhos aproximadamente iguais. Em cada iteração, 4 partes são utilizados para a fase de treino, enquanto a parte restante é reservado para a fase de teste. Esse processo é repetido 5 vezes, de modo que cada parte seja utilizado exatamente uma vez como conjunto de teste, garantindo que todas as amostras sejam utilizadas, em algum momento, na fase de teste. Desse modo podemos ter uma validação do modelo utilizando todo o conjunto de dados.



(a) Cenário de Avaliação 1

(b) Cenário de Avaliação 2

Figure 2: Ilustração dos dois cenários de validação cruzada avaliados neste trabalho, sendo (a) o cenário de avaliação por dados e (b) o cenário de avaliação por indivíduo.

Para avaliar a hipótese do enviesamento dos modelos que realizam a classificação de dados de voz de pessoas com DP, conduzimos os experimentos considerando-se dois cenários do *5-fold*, apresentados com detalhe na Figura 2:

- **Cenário 1:** As amostras disponíveis nas bases de dados são consideradas como amostras independentes. Neste caso, para cada iteração dos *5-fold*, 4 partes são utilizadas no treinamento e 1 parte utilizada para teste, como mostrado na Figura 2a.

- **Cenário 2:** O *5-fold* opera sobre a quantidade total de indivíduos dos quais as amostras foram coletadas. Em cada iteração do *k-fold*, as amostras coletadas de cada indivíduo são destinadas ao mesmo conjunto, de treinamento ou de teste, de forma exclusiva. Neste cenário, garantimos que o modelo a ser testado nunca tenha sido alimentado por amostras de um mesmo indivíduo separado para teste.

A Figura 2 ilustra dois cenários de validação cruzada com *k-fold* (k=5). No Cenário de Avaliação 1, as amostras de um mesmo indivíduo podem estar presentes tanto na fase de treino quanto na fase de teste. Já no Cenário de Avaliação 2, as amostras de cada indivíduo são inteiramente alocadas ou para o treino ou para o teste em cada iteração, garantindo uma separação completa entre as fases.

*2) **Métricas de análise de desempenho**:* Ao calcular a performance de um modelo preditivo, torna-se essencial determinar uma ou mais métricas de avaliação. Na avaliação dos modelos comparados neste artigo, utilizamos as métricas de acurácia, sensibilidade e especificidade, recomendadas quando se trabalha com predição automática de diagnósticos [30]. Essas métricas podem ser obtidas a partir da análise de uma matriz de confusão, como mostrada na Figura 3 sendo que:

		CLASSE DE PREDIÇÃO	
		Doença de Parkinson	Controle
CLASSE ATUAL	Doença de Parkinson	VP	FN
	Controle	FP	VN

Figure 3: Matriz de Confusão.

Para calcular as métricas de avaliação, faz-se necessário definir os seguintes valores:

- **Verdadeiro Positivo (VP):** Acerto do modelo em relação aos dados de pessoas com DP, predições corretas.
- **Verdadeiro Negativo (VN):** Acertos do modelo quanto aos dados de pessoas saudáveis, predições corretas.
- **Falso Positivo (FP):** Erro do modelo que, no contexto desse trabalho, falsamente acusa as amostras como sendo de indivíduos com DP.
- **Falso Negativo (FN):** Erro do modelo que, no contexto desse trabalho, falsamente prediz que a amostra analisada é de um indivíduo saudável.

A matriz de confusão é uma ferramenta fundamental na avaliação do desempenho de modelos de classificação, pois fornece uma visão detalhada da qualidade das predições realizadas, sendo útil tanto para problemas de classificação binária quanto de classificação multiclasse [31]. Sua estrutura permite identificar os acertos e erros do modelo, com a diagonal principal (VP e VN) representando os acertos das predições para o grupo DP e o grupo de controle, respectivamente. Por outro lado, a diagonal secundária (FN e FP) destaca os erros de classificação, com FN representando indivíduos do grupo de controle incorretamente classificados como tendo DP, e

FP indicando indivíduos com DP erroneamente classificados como pertencentes ao grupo de controle.

Com base nas informações da matriz de confusão, podemos agora calcular as métricas de acurácia, sensibilidade e especificidade, utilizadas neste trabalho. Essas métricas fornecem uma visão abrangente e detalhada do desempenho do modelo, permitindo identificar pontos fortes e limitações em diferentes contextos de classificação.

A acurácia determina o quão próximo o valor real está da saída do classificador. Em outras palavras, é a divisão da soma dos acertos em relação à soma do total de amostras analisadas, dada pela Equação (4) [32].

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}. \tag{4}$$

A sensibilidade, ou *recall*, é a porcentagem de acertos dos casos positivos, ou seja, é a porcentagem de amostras com a DP que foram corretamente classificadas, e pode ser calculada pela Equação (5) [30].

$$\text{Sensibilidade} = \frac{VP}{VP + FN}. \tag{5}$$

A especificidade, também chamada de precisão, corresponde à porcentagem de amostras de pessoas saudáveis (verdadeiros negativos) corretamente classificadas, sendo calculada pela Equação (6) [30].

$$\text{Especificidade} = \frac{VN}{VN + FP}. \tag{6}$$

F. Ambiente de Execução

Todos os experimentos foram realizados utilizando um notebook pessoal, da marca *Acer*, modelo Nitro 5. O equipamento roda o Sistema Operacional *Windows* 11, possui um processador *Intel(R) Core(TM)* i7-9750H com uma velocidade de *clock* de 2,6 GHz e conta com 16 GB de memória RAM, operando a 2667 MHz. Esse equipamento foi o suficiente para realizar todos os treinamentos com a linguagem de programação Python 3.10 e a IDE *Visual Studio Code*.

IV. Resultados e Discussão

Nesta seção, são apresentados os resultados obtidos para a avaliação da sobreamostragem de dados de voz na classificação da DP. Os resultados obtidos na avaliação dos dois cenários propostos são apresentados separadamente para cada uma das bases de dados utilizadas. Na sequência, discutimos os resultados observados.

A. Base de Dados Little et al.

Considerando os resultados apresentados na Tabela VII para a base Little *et al.*, observa-se, inicialmente, o desempenho dos classificadores sem a utilização do PCA. No Cenário de Avaliação 1, as acurácias variaram de 70,77% (*Naive Bayes*) a 92,31% (KNN), enquanto no Cenário de Avaliação 2 os valores ficaram entre 67,94% (*Naive Bayes*) e 77,22% (SVM). Embora o *Naive Bayes* tenha apresentado a menor acurácia, ele destacou-se em termos de especificidade, mostrando bom desempenho na identificação de indivíduos de controle, mesmo considerando o maior número de amostras e indivíduos com DP na base. A análise de sensibilidade e especificidade evidencia diferenças no comportamento dos modelos, reforçando que o equilíbrio entre essas métricas é essencial para uma avaliação mais completa.

Comparando os dois cenários, nota-se que os valores de acurácia, sensibilidade e especificidade foram consistentemente maiores no Cenário de Avaliação 1. Isso sugere que, embora esse cenário utilize dados de teste inéditos, a presença de amostras distintas de um mesmo indivíduo no conjunto de treinamento influencia positivamente o desempenho dos modelos ao lidar com amostras do mesmo indivíduo no conjunto de teste, configurando um possível viés.

Com a aplicação do PCA, os classificadores apresentaram um desempenho similar, mas com algumas diferenças notáveis. No Cenário de Avaliação 1, as acurácias variaram de 80,00% (*Naive Bayes*) a 92,82% (KNN), enquanto no Cenário de Avaliação 2 os valores oscilaram entre 72,20% (*Naive Bayes*) e 83,32% (LDA). A análise das métricas indica que a sensibilidade foi consistentemente maior que a especificidade, sugerindo um bom desempenho na identificação de indivíduos com DP, mas com uma maior taxa de falsos positivos, o que compromete a especificidade.

De forma similar aos resultados sem PCA, os valores das métricas no Cenário de Avaliação 1 foram superiores aos do Cenário de Avaliação 2. Isso reforça a influência das informações compartilhadas entre amostras de um mesmo indivíduo, mesmo com a redução de dimensionalidade, evidenciando desafios adicionais para generalização nos modelos ao lidar com indivíduos diferentes.

B. Base de Dados Sakar et al.

Considerando os resultados apresentados na Tabela VIII para a base de dados do Sakar *et al.*, sem a utilização do PCA, observamos que, no Cenário de Avaliação 1, as acurácias médias dos classificadores variaram de 67,47% (LDA) a 85,31% (KNN), enquanto no Cenário de Avaliação 2 as acurácias ficaram entre 60,59% (LDA) e 80,15% (KNN). A análise das métricas de sensibilidade e especificidade indica que a sensibilidade foi consistentemente maior que a especificidade. Isso sugere que os modelos identificaram corretamente indivíduos com DP, mas ao custo de um número considerável de falsos positivos, o que levou a uma especificidade baixa.

Ao comparar os Cenários de Avaliação 1 e 2, para todos os valores de acurácia, sensibilidade e especificidade, os resultados do Cenário de Avaliação 1 foram sempre superiores. Isso sugere que a presença de amostras distintas de um mesmo indivíduo no conjunto de treinamento pode influenciar na capacidade dos modelos em generalizar corretamente para o mesmo indivíduo no conjunto de teste, causando uma redução na acurácia no Cenário de Avaliação 2.

Quando o PCA foi aplicado, os classificadores apresentaram uma leve variação nos resultados, como pode ser observado na Tabela VIII. No Cenário de Avaliação 1, as acurácias médias variaram de 71,82% (*Naive Bayes*) a 85,05% (KNN), enquanto no Cenário de Avaliação 2, as acurácias ficaram entre 72,38%

Table VII: Resultados da base de dados do Little *et al.*

Sem PCA			
Cenário de Avaliação 1			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	92,31%	95,24%	83,00%
LDA	85,64%	93,12%	63,10%
SVM	84,10%	92,49%	58,53%
NaiveBayes	70,77%	64,12%	87,67%
DT	84,62%	90,09%	68,69%
Cenário de Avaliação 2			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	74.58%	88.40%	30.00%
LDA	75.41%	83.02%	51.67%
SVM	77.22%	90.15%	38.33%
NaiveBayes	67.94%	63.05%	86.67%
DT	69.02%	79.57%	33.33%
Com PCA			
Cenário de Avaliação 1			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	92,82%	95,76%	82,32%
LDA	85,13%	95,78%	52,47%
SVM	85,64%	94,34%	57,89%
NaiveBayes	80,00%	87,01%	59,26%
DT	85,13%	89,59%	68,75%
Cenário de Avaliação 2			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	76,62%	89,04%	36,67%
LDA	83,32%	95,38%	46,67%
SVM	80,53%	91,26%	46,67%
NaiveBayes	72,20%	82,90%	41,67%
DT	75,19%	85,86%	36,67%

Table VIII: Resultados da base de dados do *Sakar et al.*

Sem PCA			
Cenário de Avaliação 1			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	85,31%	97,54%	49,38%
LDA	67,47%	69,02%	63,26%
SVM	82,40%	87,16%	69,04%
NaiveBayes	79,89%	84,53%	65,99%
DT	80,56%	87,26%	61,57%
Cenário de Avaliação 2			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	80,15%	94,56%	37,86%
LDA	60,59%	65,23%	46,58%
SVM	78,19%	86,39%	53,89%
NaiveBayes	76,62%	80,72%	64,52%
DT	75,65%	83,75%	51,89%
Com PCA			
Cenário de Avaliação 1			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	85,05%	96,83%	50,53%
LDA	82,81%	91,60%	57,14%
SVM	79,89%	85,73%	62,75%
NaiveBayes	71,82%	87,19%	27,01%
DT	73,67%	81,80%	50,01%
Cenário de Avaliação 2			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	80,55%	94,39%	39,93%
LDA	78,58%	87,99%	50,44%
SVM	77,26%	83,78%	58,00%
NaiveBayes	72,38%	87,49%	27,87%
DT	72,62%	82,66%	43,32%

(*Naive Bayes*) e 80,55% (KNN). Embora a aplicação do PCA tenha levado a uma leve melhora na acurácia no Cenário de Avaliação 2, a sensibilidade ainda foi superior à especificidade, indicando uma boa capacidade de identificar pessoas com DP, mas com a ocorrência de falsos positivos.

Ao comparar os resultados entre os Cenários de Avaliação 1 e 2, os valores de acurácia, sensibilidade e especificidade foram mais elevados no Cenário de Avaliação 1, similar aos resultados sem PCA. Isso reforça a hipótese de que a presença de amostras de um mesmo indivíduo no conjunto de treinamento pode impactar a capacidade de generalização dos modelos. A redução na acurácia no Cenário de Avaliação 2 destaca a importância de uma avaliação mais cuidadosa do efeito das amostras distintas de um mesmo indivíduo durante o treinamento.

C. Base de Dados Prez

Considerando os resultados apresentados na Tabela IX para a base de dados do Prez, sem a utilização do PCA, observa-se que, no Cenário de Avaliação 1, os classificadores retornaram acurácias médias que variam de 71,25% (DT) a 82,50% (*Naive Bayes* e KNN), enquanto no Cenário de Avaliação 2 as acurácias variaram de 65,83% (LDA) a 82,50% (*Naive Bayes*). Ao analisar os valores de sensibilidade e especificidade, observa-se que são bem balanceados, o que torna a base de dados propícia para bons resultados no auxílio aos diagnósticos. Este equilíbrio ocorre devido à estrutura balanceada da base, contendo uma quantidade semelhante de indivíduos com DP e indivíduos de controle.

Ao comparar os Cenários de Avaliação 1 e 2, os resultados de acurácia, sensibilidade e especificidade no Cenário de Avaliação 1 foram geralmente mais elevados do que no Cenário de Avaliação 2. Isso sugere que, embora o Cenário de Avaliação 1 utilize dados de teste nunca vistos pelos modelos, a redução na acurácia no Cenário de Avaliação 2 pode ser atribuída à presença de amostras de um mesmo indivíduo em diferentes conjuntos, o que pode prejudicar a capacidade do modelo de generalizar corretamente quando lida com essas amostras no conjunto de teste.

Com a aplicação do PCA, os resultados mostraram acurácias médias variando de 73,75% (DT) a 80,42% (KNN) no Cenário de Avaliação 1, e de 70,38% (DT) a 76,25% (LDA) no Cenário de Avaliação 2. Assim como no cenário sem PCA, os valores de sensibilidade e especificidade continuaram balanceados, refletindo a boa distribuição de indivíduos com DP e indivíduos de controle na base de dados. Esses resultados indicam que a base de dados, por ser balanceada, ainda favorece os diagnósticos eficientes, mesmo após a redução de dimensionalidade aplicada pelo PCA.

Ao comparar os Cenários de Avaliação 1 e 2 com PCA, observa-se que, novamente, os resultados de acurácia, sensibilidade e especificidade foram superiores no Cenário de

Avaliação 1. Isso reforça a ideia de que a presença de amostras de um mesmo indivíduo no conjunto de treinamento pode afetar a precisão dos modelos ao lidar com amostras desse mesmo indivíduo no conjunto de teste, prejudicando a acurácia no Cenário de Avaliação 2.

Table IX: Resultados da base de dados do Prez.

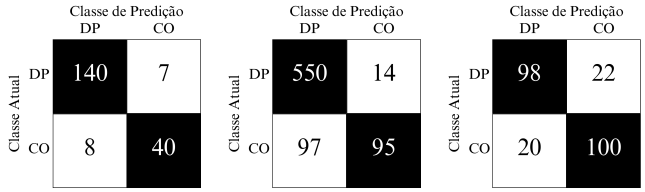
Sem PCA			
Cenário de Avaliação 1			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	82,50%	81,69%	83,39%
LDA	77,08%	76,58%	77,52%
SVM	74,17%	74,98%	73,18%
NaiveBayes	82,50%	79,95%	84,96%
DT	71,25%	69,94%	72,45%
Cenário de Avaliação 2			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	73,33%	73,03%	74,05%
LDA	65,83%	63,41%	69,07%
SVM	70,00%	69,60%	71,48%
NaiveBayes	82,50%	79,75%	85,19%
DT	72,08%	73,58%	70,54%
Com PCA			
Cenário de Avaliação 1			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	80,42%	79,19%	81,69%
LDA	77,08%	75,75%	78,29%
SVM	77,08%	77,45%	76,59%
NaiveBayes	76,25%	77,52%	74,92%
DT	73,75%	67,37%	79,99%
Cenário de Avaliação 2			
Modelo	**Acurácia**	**Sensibilidade**	**Especificidade**
KNN	73,75%	73,98%	73,96%
LDA	76,25%	75,32%	77,66%
SVM	75,83%	71,89%	80,19%
NaiveBayes	72,50%	77,62%	68,37%
DT	70,83%	73,03%	69,21%

D. Discussão

Para aprofundar na análise dos resultados obtidos, construímos as matrizes de confusão detalhando os resultados dos classificadores que mais se destacaram em cada cenário analisado. Foram gerados as matrizes de confusão para o melhor resultado do Cenário de Avaliação 1 e para o melhor resultado do Cenário de Avaliação 2, sem a utilização do PCA e com a utilização do PCA, apresentados nas Figuras 4 e 5, respectivamente.

As matrizes de confusão apresentadas na Figura 4 ilustram os melhores resultados alcançados sem a aplicação do PCA nos diferentes cenários de avaliação para as bases de dados. Observa-se que as bases de dados Little *et al.* e Sakar *et al.* são desbalanceadas, com predominância de amostras de DP, enquanto a base Prez é balanceada, o que impacta diretamente no desempenho dos modelos.

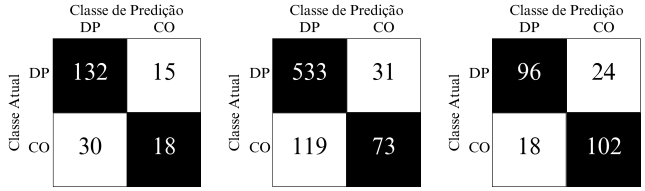
Para o KNN, observam-se ótimos desempenhos nos Cenários de Avaliação 1, nota-se um desempenho robusto em todas as bases de dados com altas taxas de verdadeiro



(a) KNN - Cenário de Avaliação 1 - Base Little *et al.*

(b) KNN - Cenário de Avaliação 1 - Base Sakar *et al.*

(c) KNN - Cenário de Avaliação 1 - Base Prez.



(d) SVM - Cenário de Avaliação 2 - Base Little *et al.*

(e) KNN - Cenário de Avaliação 2 - Base Sakar *et al.*

(f) *Naive Bayes* Confusão Cenário de Avaliação 2 - Base Prez.

Figure 4: Matriz de Confusão dos Melhores Resultados sem PCA.

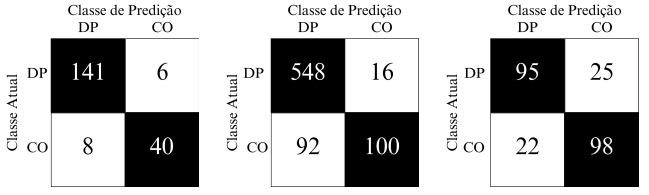
positivos (VP) e poucos falsos negativos (FN), demonstrando boa capacidade do modelo de separar amostras de controle e amostras de DP mesmo em bases desbalanceadas, como as bases Little *et al.* e Sakar *et al.*. No entanto, no Cenário de Avaliação 2, há uma redução geral na capacidade preditiva, especialmente nas bases desbalanceadas, onde se observa um aumento de falsos negativos (FN), o que reflete a dificuldade do modelo em generalizar para novas pessoas. Reforçando a hipótese que as amostras de um mesmo indivíduo carrega informações semelhantes.

O SVM, particularmente no Cenário de Avaliação 2, apresenta uma elevação significativa nos falsos positivos (FP), especialmente em bases desbalanceadas, onde as classes com maior quantidade são preferencialmente previstas. Isso indica que o modelo teve dificuldade em lidar com a separação clara das classes quando desbalanceadas e a variabilidade individual.

Já o *Naive Bayes*, no Cenário de Avaliação 2, mostra um desempenho inferior em relação aos resultados, com taxas mais equilibradas entre falsos positivos (FP) e falsos negativos (FN), porém com menor acurácia geral. Esse comportamento se deve ao balanceamento da base Prez, onde podemos ver o impacto do balanceamento, onde as taxas de acertos são mais consistentes, e os erros, mais equilibrados entres falsos positivos e falsos negativos. Isso reforça a importância de considerar o balanceamento.

Conclui-se, a partir dos resultados no Cenário de Avaliação 2, que os indivíduos podem carregar informações semelhantes entre suas amostras, o que favorece os modelos no Cenário 1, onde as amostras individuais são tratadas de forma independente. No entanto, ao agrupar amostras por pessoa no Cenário de Avaliação 2, os modelos enfrentam maior dificuldade em generalizar, especialmente em bases desbalanceadas. Essa análise ressalta a relevância de considerar as características intrínsecas dos dados e a necessidade de estratégias adequadas para lidar com a redundância e a variabilidade individual.

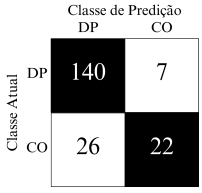
A Figura 5 apresenta as matrizes de confusão dos melhores resultados obtidos após a aplicação do PCA, considerando diferentes classificadores e cenários de avaliação para as bases de dados. A aplicação do PCA sugere a redução da dimensionalidade com intuito de mitigar algumas redundâncias e ruídos presentes nos dados.



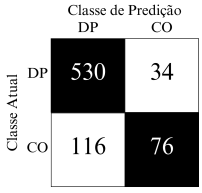
(a) KNN - Cenário de Avaliação 1 - Base Little *et al.*

(b) KNN - Cenário de Avaliação 1 - Base Sakar *et al.*

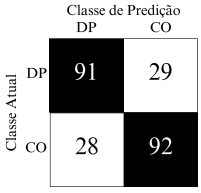
(c) KNN - Cenário de Avaliação 1 - Base Prez.



(d) LDA - Cenário de Avaliação 2 - Base Little *et al.*



(e) KNN - Cenário de Avaliação 2 - Base Sakar *et al.*



(f) LDA - Cenário de Avaliação 2 - Base Prez.

Figure 5: Matriz de Confusão dos Melhores Resultados com PCA.

No Cenário de Avaliação 1, os resultados para a base Little *et al.*, apresentaram uma leve melhora em comparação com a versão sem PCA, indicando que o PCA foi capaz de mitigar algumas redundâncias e ruídos. Na base Sakar *et al.*, apesar de um desempenho sólido do KNN, ainda se observa uma quantidade significativa de falsos negativos (FN), indicando que o desbalanceamento continua impactando negativamente a performance. Já na base balanceada Prez, houve uma diminuição na precisão geral após o PCA, com um aumento nos falsos positivos (aumento de 2) e falsos negativos (aumento de 3). Isso sugere que a técnica não foi tão eficaz para base balanceada, apesar do PCA simplificar o espaço de características.

No Cenário de Avaliação 2, onde a validação foi feita por pessoa, os resultados refletem maior dificuldade em generalizar. O LDA na base Little *et al.* apresentou uma leve melhora em relação ao cenário sem PCA, com uma redução de falsos negativos (FN) e falsos positivos (FP). No entanto, na base Sakar *et al.*, o KNN continuou a apresentar um elevado número de falsos negativos (116 casos), indicando que o impacto do desbalanceamento persiste, mesmo após o PCA. Para a base Prez, o LDA apresentou uma redução de desempenho após a aplicação do PCA, com aumento tanto nos falsos negativos quanto nos falsos positivos, reforçando que a técnica não foi tão eficaz para base balanceada nesse cenário.

Assim como os resultados anteriores sem a utilização do PCA conclui-se que a partir dos resultados no Cenário de Avaliação 2, que os indivíduos podem carregar informações semelhantes entre suas amostras, o que favorece os modelos no Cenário 1, onde as amostras individuais são tratadas de forma independente. No entanto, ao agrupar amostras por pessoa no Cenário de Avaliação 2, os modelos enfrentam maior dificuldade em generalizar, especialmente em bases desbalanceadas. Essa análise ressalta a relevância de considerar as características intrínsecas dos dados e a necessidade de estratégias adequadas para lidar com a redundância e a variabilidade individual.

Os resultados obtidos evidenciam diferenças significativas no desempenho dos classificadores com e sem a aplicação do PCA, destacando os impactos dessa técnica de redução de dimensionalidade sobre as bases de dados analisadas. Dentre os modelos de classificação, o KNN foi o que apresentou desempenho mais consistente e destacado, superando em vários cenários de avaliação os demais modelos em ambas as métricas.

No Cenário de Avaliação 1, sem a utilização do PCA, o KNN na base Little *et al.* alcançou 92,31% de acurácia, 95,24% de sensibilidade e 83,00% de especificidade, superando os outros modelos na mesma base de dados. Na base Sakar *et al.*, o KNN teve 85,31% de acurácia, 97,54% de sensibilidade e 49,38% de especificidade, superando também outros modelos e reforçando sua robustez em identificar padrões mesmo em bases de alta dimensionalidade.

Com a aplicação do PCA, o KNN manteve seu destaque na base Little *et al.*, onde obteve uma leve melhora, atingindo 92,82% de acurácia, 95,76% de sensibilidade e 82,32% no Cenário de Avaliação 1. Esse resultado sugere que o PCA ajudou a eliminar redundâncias e ruídos sem comprometer a capacidade do KNN de capturar padrões importantes. No entanto, na base Prez, que possui menos redundâncias e é mais balanceada, o desempenho do KNN caiu após o PCA, indicando que a redução de dimensionalidade pode ter descartado características importantes para essa base. Esse comportamento ressalta que o impacto do PCA varia conforme a estrutura e as características dos dados.

Já no Cenário de Avaliação 2 os desafios de generalização dos modelos ficaram mais evidentes. O KNN apesar de seu desempenho consistente no Cenário de Avaliação 1, apresentou uma queda considerável no Cenário Avaliação 2, alcançando 74,58% de acurácia sem o PCA e 76,62% com o PCA na base Little *et al.*. Essa redução reflete a dificuldade em generalizar para indivíduos diferentes, um problema amplificado pela semelhança nas informações carregadas pelos indivíduos.

Esses resultados ressaltam que, embora o KNN tenha se destacado em termos de desempenho, ele também é afetado por problemas de generalização, particularmente no Cenário de Avaliação 2. A sobreposição de informações entre indivíduos, característica marcante das bases Little *et al.* e Sakar *et al.*, contribui para taxas elevadas de falsos negativos, indicando que padrões específicos podem ser interpretados como comuns a diferentes classes. Além disso, os resultados reforçam a importância de avaliar cuidadosamente transformações como o PCA, garantindo que elas preservem informações críticas para a diferenciação de padrões entre indivíduos e promovam melhorias reais no desempenho dos modelos em aplicações práticas.

V. Conclusão

Com base no trabalho realizado, podemos concluir que a avaliação de sinais de voz de indivíduos com Doença de Parkinson (DP) requer uma abordagem cuidadosa em relação ao tratamento das amostras, especialmente quando provenientes do mesmo indivíduo. A análise dos resultados obtidos, considerando os dois cenários de avaliação propostos (Cenário de Avaliação 1 e Cenário de Avaliação 2), revelou que tratar as amostras de forma independente introduziu um viés nos resultados, independente da base de dados utilizada, favorecendo a acurácia, sensibilidade e especificidade dos modelos. No entanto, quando as amostras de um mesmo indivíduo foram tratadas de forma dependente, os resultados mostraram uma redução significativa na performance dos classificadores, indicando que este tratamento mais rigoroso é necessário para evitar *overfitting*.

Os resultados obtidos neste trabalho corroboram com os achados de Chagas *et al.* [9], trazendo uma nova analise agora sob o ponto de vista de replicação de sinais de voz para a classificação da DP. Como destacado estatisticamente por Naranjo *et al.* [8], embora os dados de voz de um mesmo indivíduo não sejam idênticas, eles tendem a ser mais semelhantes entre si em comparação com sujeitos diferentes. Essa semelhança nas amostras de um mesmo indivíduo sugere que uma abordagem dependente pode melhorar a capacidade dos modelos de capturar essas variações sutis, levando a uma classificação mais precisa.

A análise realizada também sugere que, em bases de dados desbalanceadas, como no caso das bases utilizadas neste estudo (Little *et al.*, Sakar *et al.*), o uso de técnicas de balanceamento poderia melhorar ainda mais a acurácia, como a sobreamostragem, poderia melhorar ainda mais a acurácia, reduzindo os valores de falso negativos. A sobreamostragem pode ajudar a equilibrar as classes, contribuindo para o aumento de verdadeiros positivos nos modelos e favorecendo a detecção de padrões em amostras menos representadas. Por fim, para trabalhos futuros, propomos a exploração de novas técnicas de pré-processamento, como a seleção de características mais avançadas, e a implementação de métodos de balanceamento de dados nas bases desbalanceadas, de forma a melhorar a qualidade dos resultados obtidos. A combinação dessas estratégias pode aprimorar a capacidade dos modelos de identificar padrões em dados de voz e contribuir para um auxílio de diagnóstico mais preciso e precoce da Doença de Parkinson.

References

[1] K. Prabhavathi and S. Patil, "Tremors and bradykinesia," *Techniques for Assessment of Parkinsonism for Diagnosis and Rehabilitation*, pp. 135–149, 2022.

[2] H. Braak and E. Braak, "Pathoanatomy of parkinson's disease," *Journal of neurology*, vol. 247, pp. II3–II10, 2000.

[3] C. M. Tanner, "Epidemiology of parkinson's disease," *Neurologic clinics*, vol. 10, no. 2, pp. 317–329, 1992.

[4] L. V. Kalia and A. E. Lang, "Parkinson's disease," *The Lancet*, vol. 386, no. 9996, pp. 896–912, 2015.

[5] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with parkinson's disease," *Behavioural neurology*, vol. 11, no. 3, pp. 131–137, 1999.

[6] J. Atarachi and E. Uchida, "A clinical study of parkinsonism," *Re cent Adv Res Nerv Syst 1959; 3: 871*, vol. 882, 1959.

[7] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–1, 2008.

[8] L. Naranjo, C. J. Perez, Y. Campos-Roca, and J. Martin, "Addressing voice recording replications for parkinson's disease detection," *Expert Systems with Applications*, vol. 46, pp. 286–292, 2016.

[9] A. L. de Bastos Chagas, F. Giordana de Farias, J. P. Félix, A. U. da Fonseca, H. A. do Nascimento, and F. Soares, "Avaliando a sobreamostragem de dados temporais de marcha no diagnóstico automático de doenças neurodegenerativas," in *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pp. 567–578, SBC, 2024.

[10] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[11] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and H. Apaydin, "A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform," *Applied Soft Computing*, vol. 74, pp. 255–263, 2019.

[12] S. Aich, H.-C. Kim, K. L. Hui, A. A. Al-Absi, M. Sain, *et al.*, "A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of parkinson's disease," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pp. 1116–1121, IEEE, 2019.

[13] G. Solana-Lavalle, J.-C. Galán-Hernández, and R. Rosas-Romero, "Automatic parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 505–516, 2020.

[14] S. Sharanyaa, P. N. Renjith, and K. Ramesh, "Classification of parkinson's disease using speech attributes with parametric and nonparametric machine learning techniques," in *2020 3rd international conference on intelligent sustainable systems (ICISS)*, pp. 437–442, IEEE, 2020.

[15] A. Ouhmida, O. Terrada, A. Raihani, B. Cherradi, and S. Hamida, "Voice-based deep learning medical diagnosis system for parkinson's disease prediction," in *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, pp. 1–5, IEEE, 2021.

[16] A. Rana, A. Dumka, R. Singh, M. Rashid, N. Ahmad, and M. K. Panda, "An efficient machine learning approach for diagnosing parkinson's disease by utilizing voice features," *Electronics*, vol. 11, no. 22, p. 3782, 2022.

[17] A. Govindu and S. Palwe, "Early detection of parkinson's disease using machine learning," *Procedia Computer Science*, vol. 218, pp. 249–261, 2023.

[18] M. Melo and T. Gouveia, "Classificação de sinais de voz para auxilio no diagnóstico da doença de parkinson," *Revista Brasileira de Computação Aplicada*, vol. 15, no. 2, pp. 88–104, 2023.

[19] H. A. Soliman, M. S. Elshourbagi, M. A. Dahy, M. M. Osman, B. A. Kamal, M. Abd Elaziz, and N. El-rashidy, "Classifying parkinson's disease using speech features," in *2024 Intelligent Methods, Systems, and Applications (IMSA)*, pp. 544–549, IEEE, 2024.

[20] C. Prez, "Parkinson dataset with replicated acoustic features," *UCI Machine Learning Repository*, 2019.

[21] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.

[22] "Standard scaler — scikit-learn.org." https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. [Accessed 30-11-2024].

[23] T. Minka, "Automatic choice of dimensionality for pca," *Advances in neural information processing systems*, vol. 13, 2000.

[24] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, pp. 1–37, 2008.

[25] "KNeighborsClassifier — scikit-learn.org." https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. [Accessed 18-09-2024].

[26] K. Faceli, A. C. Lorena, J. Gama, T. A. d. Almeida, and A. C. P. d. L. F. d. Carvalho, "Inteligência artificial: uma abordagem de aprendizado de máquina," 2021.

[27] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," *Robust data mining*, pp. 27–33, 2013.

[28] A. Bhatia and R. Sulekh, "Predictive model for parkinson's disease through naive bayes classification," *International Journal of Computer Science & Communication*, vol. 9, no. 1, pp. 194–202, 2017.

[29] R. O. Duda, P. E. Hart, *et al.*, *Pattern classification*. John Wiley & Sons, 2006.

[30] “Diagnostic tests. 1: Sensitivity and specificity - PubMed — pubmed.ncbi.nlm.nih.gov.” https://pubmed.ncbi.nlm.nih.gov/8019315/. [Accessed 18-09-2024].
[31] C. Room, “Confusion matrix,” *Mach. Learn*, vol. 6, p. 27, 2019.
[32] A. Gunawardana and G. Shani, “A survey of accuracy evaluation metrics of recommendation tasks.,” *Journal of Machine Learning Research*, vol. 10, no. 12, 2009.

Matheus Isac da Silva é estudante de graduação em Engenharia da Computação pela Universidade Federal de Goiás, com foco em pesquisa nas áreas de Aprendizado de Máquina e Ciência de Dados. Possui experiência em técnicas de modelagem preditiva e análise de desempenho de algoritmos de classificação, com aplicações voltadas para diagnóstico assistido de doenças neurodegenerativas e outras áreas interdisciplinares.

Juliana Paula Felix é Doutora em Ciência da Computação (2023) pela Universidade Federal de Goiás (UFG), Mestre (2018) e bacharel em Ciência da Computação (2015) pela mesma universidade. Desenvolve pesquisas em diversas áreas da computação, com interesse especial em pesquisas aplicadas à saúde. Tem interesse nas áreas de ciência de dados, aprendizado de máquina, processamento de sinais e imagens, visualização da informação e otimização.