ANÁLISE COMPARATIVA DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA PARA DIAGNÓSTICO DA DOENÇA DE PARKINSON COM BASE EM CARACTERÍSTICAS VOCAIS

Jorge Antonio Felix Da Silva[1]
Ivina Lorena Oliveira Moura[2]
Alexandre Lima De Oliveira[3]
Antonio Alisson Pessoa Guimaraes[4]

RESUMO

A doença de Parkinson (DP) é uma doença neurodegenerativa crônica que afeta o sistema nervoso central, em particular, as áreas do cérebro responsáveis pelo controle do movimento. Os principais sintomas da doença de Parkinson incluem: tremores, rigidez muscular e bradicinesia. Além desses sintomas motores, a DP pode causar uma série de outros sintomas não motores, como depressão, ansiedade, distúrbios do sono e problemas cognitivos. A DP é uma condição crônica que progride ao longo do tempo, e o tratamento visa melhorar a qualidade de vida dos pacientes e controlar os sintomas da melhor maneira possível. Diante disso, é notório que o diagnóstico precoce é de suma importância para fazer um tratamento eficaz e retardar a progressão da DP. A partir de uma base de dados publica disponível na literatura, com registros de sinais de voz, tem-se como objetivo deste trabalho avaliar o desempenho de diferentes algoritmos de classificação na predição da DP para auxiliar os médicos neurologistas no diagnóstico. O modelo será treinado com base em uma ampla base de dados pública, contendo características extraídas de amostras de voz de 40 pacientes (20 saudáveis e 20 portadores da doença). O almeja-se futuramente criar uma ferramenta que possa ajudar na identificação da doença com base em características vocais, auxiliando no diagnóstico clínico.

Palavras-chave: Doença de Parkinson; Algoritmos de Classificação; Diagnostico Clínico.

Universidade da Integração Internacional da lusofonia Afro-brasileira, IEDS - Instituto de Engenharias e Desenvolvimento Sustentável , Discente, jorgefelix@aluno.unilab.edu.br[1]
Universidade da Integração Internacional da lusofonia Afro-brasileira, IEDS - Instituto de Engenharias e Desenvolvimento Sustentável , Discente, ivinalorena@aluno.unilab.edu.br[2]
Universidade da Integração Internacional da lusofonia Afro-brasileira, IEDS - Instituto de Engenharias e Desenvolvimento Sustentável , Discente, alexandre.computacao@aluno.unilab.edu.br[3]
Universidade da Integração Internacional da lusofonia Afro-brasileira, IEDS - Instituto de Engenharias e Desenvolvimento Sustentável , Docente, alisson@unilab.edu.br[4]

ISSN: 2447-6161

INTRODUÇÃO

A Doença de Parkinson (DP) é uma condição neurodegenerativa crônica que afeta o sistema nervoso central, particularmente as áreas do cérebro responsáveis pelo controle dos movimentos. Seus sintomas incluem tremores, rigidez muscular e bradicinesia. Além dos sintomas motores, a DP também pode se manifestar com sintomas não motores, como depressão, ansiedade, distúrbios do sono e declínio cognitivo. Devido à progressão gradual da doença, a detecção precoce é essencial para uma gestão mais eficaz, ajudando a retardar a evolução dos sintomas.

Este trabalho tem como objetivo realizar um estudo comparativo sobre o desempenho de diferentes algoritmos de aprendizagem de máquina no auxílio ao diagnóstico da Doença de Parkinson. A análise baseia-se em características vocais extraídas de amostras de voz de pacientes com e sem a doença, utilizando um conjunto de dados público composto por gravações de voz de 40 indivíduos. Embora o propósito final seja desenvolver uma ferramenta que, futuramente, possa auxiliar médicos no diagnóstico da DP, este estudo se concentra, primeiramente, em avaliar qual algoritmo classificador oferece o melhor desempenho nesta base de dados.

O estudo envolve diversas etapas, incluindo a seleção das características vocais mais relevantes, o treinamento e a avaliação de modelos com diferentes algoritmos de aprendizagem supervisionada. Foram testados os algoritmos K-Nearest Neighbors (KNN), Random Forest Classifier (RFC), Support Vector Machine (SVM) e Regressão Logística (RL). Cada modelo foi treinado e testado com base nos dados de voz coletados, visando comparar suas respectivas performances em termos de acurácia, precisão, recall e F1-score, com o apoio de uma matriz de confusão para análise mais detalhada dos resultados.

Embora este trabalho ainda não ofereça uma ferramenta pronta para aplicação clínica, ele constitui um passo importante na análise de algoritmos de aprendizagem de máquina que podem, futuramente, ser integrados como um recurso adicional no diagnóstico precoce da Doença de Parkinson. A partir da análise comparativa dos algoritmos, espera-se identificar a abordagem mais promissora para desenvolvimentos futuros, visando contribuir para uma intervenção médica mais eficaz e uma melhor qualidade de vida para os pacientes.

METODOLOGIA

Para desenvolver um sistema auxiliar para a detecção de distúrbios de Parkinson através da análise da voz dos pacientes, adotamos uma metodologia que envolveu várias etapas, desde a seleção e processamento das características vocais mais relevantes até o treinamento e a avaliação do desempenho dos modelos resultantes. Para abordar essa questão, escolhemos utilizar 4 algoritmos de classificação, todos eles têm em comum o fato de serem algoritmos de classificação supervisionada, utilizados para categorizar dados em classes pré-definidas com base em um conjunto de treinamento rotulado. Os algoritmos escolhidos para a execução da pesquisa foram o KNN (K-ésimo vizinho mais próximo), o SVM (Máquina de vetores de suporte), o RFC (Floresta aleatória) e o RL (regressão logística). Os algoritmos foram implementados na linguagem de programação Python na versão mais recente com auxílio das bibliotecas SKLEARN (biblioteca de aprendizado de máquina), NumPy (biblioteca para computação cientifica) e pandas (biblioteca de análise de dados). Todos os valores possíveis de parâmetros usados nos algoritmos, foram os valores padrões do próprio algoritmo, por exemplo no KNN, ele usa o valor padrão para o número de vizinhos mais próximos, que é 5. Isso significa que o algoritmo está usando 5 vizinhos mais próximos para fazer as previsões. O algoritmo K-Nearest Neighbors (KNN), descrito por Cover e Hart (1967), é um método de aprendizado supervisionado utilizado principalmente para tarefas de classificação e regressão. O KNN classifica um novo ponto de dados

ISSN: 2447-6161

com base na maioria dos seus vizinhos mais próximos, sendo o parâmetro K o número de vizinhos considerados. A distância entre os pontos é geralmente medida por métricas como a distância euclidiana. Uma das vantagens do KNN é sua simplicidade, já que não envolve um processo de treinamento explícito, apenas o armazenamento de dados de treinamento. Contudo, isso também implica que ele pode ser computacionalmente custoso para grandes conjuntos de dados, pois cada nova classificação exige o cálculo das distâncias entre o ponto novo e todos os pontos de treinamento. O algoritmo SVM (Support Vector Machine), proposto por Cortes e Vapnik (1995), é uma técnica de aprendizado supervisionado utilizada para classificação e regressão. Ele funciona separando os dados em diferentes classes por meio de um hiperplano de maior margem, ou seja, a linha ou superfície que melhor divide as diferentes classes, maximizando a distância entre os pontos de dados de cada classe. Quando os dados não são linearmente separáveis, o SVM utiliza uma técnica chamada de kernel trick, que transforma o espaço original dos dados em um espaço de maior dimensionalidade, onde eles podem ser separados. O SVM é amplamente utilizado em diversas aplicações, como reconhecimento de padrões e classificação de imagens, devido à sua capacidade de lidar com grandes volumes de dados e proporcionar uma boa generalização. O algoritmo Random Forest Classifier (RFC), introduzido por Breiman (2001), é um método de aprendizado supervisionado baseado em um conjunto de árvores de decisão. O RFC utiliza o conceito de ensemble learning, no qual várias árvores de decisão são construídas a partir de diferentes subconjuntos do conjunto de dados e características, e as previsões finais são feitas com base na votação da maioria das árvores. Esse processo de agregação de várias árvores contribui para uma maior precisão e capacidade de generalização, além de reduzir o risco de overfitting, que é comum em modelos de árvore única. O RFC é amplamente utilizado em tarefas de classificação e regressão, sendo especialmente eficaz em problemas de grande dimensionalidade e dados ruidosos. O algoritmo de Regressão Logística, introduzido por Cox (1958), é um método de aprendizado supervisionado utilizado para tarefas de classificação binária. Ao contrário da regressão linear, que prevê valores contínuos, a regressão logística utiliza a função logística (ou sigmoide) para modelar a probabilidade de uma observação pertencer a uma determinada classe. A saída do modelo é um valor entre 0 e 1, que pode ser interpretado como uma probabilidade. Com base nesse valor, o algoritmo classifica a entrada em uma das duas classes, utilizando um limiar predefinido (geralmente 0,5). A regressão logística é amplamente utilizada em várias áreas, como medicina e ciências sociais, por sua capacidade de lidar com problemas de classificação de forma simples e eficiente. Na etapa de treinamento separamos a base dados em dois conjuntos, um de treinamento e um de teste, respeitando a proporção de 80% e 20% respectivamente, vale destacar que como o conjunto de dados é composto por 20 pacientes saudáveis e 20 portadores da doença de Parkinson, foi escolhido aleatoriamente 4 dentre os saudáveis e 4 dentre os portadores de DP para fazerem parte do conjunto de teste, enquanto o restante dos pacientes foram alocados no conjunto de treinamento dos algoritmos de classificação. Como foram selecionadas as seis variáveis mais correlacionadas com o diagnóstico, o conjunto de dados bruto assumiu o formato de uma matriz, em que as colunas representavam essas variáveis. No entanto, as linhas apresentavam uma particularidade: a cada 26 linhas, correspondia-se a um novo paciente, sendo que cada uma dessas linhas representava um tipo de gravação sonora, incluindo vogais sustentadas, números, palavras e frases curtas. Esse conjunto de dados foi utilizado para treinar algoritmos de classificação. Após as etapas de treinamento e teste, obteve-se um vetor binário, composto por 0’s e 1’s, o qual classifica um mesmo paciente em 26 amostras de voz. Podendo ser classificado como normal ao pronunciar a vogal ‘a’ e ser parkinsoniano ao pronunciar a vogal ‘o’. Para determinar a predição final de um paciente, foi desenvolvido um algoritmo que somava a ocorrência de 0’s e 1’s. Ou seja, para um paciente com 26 gravações sonoras, cada linha possuía uma etiqueta: 0, se o classificador categorizava como saudável, ou 1, se o classificador o classificava como portador da doença. Se o paciente apresentasse mais 1’s

ISSN: 2447-6161

do que 0’s, ele seria classificado como portador da doença de Parkinson; caso contrário, seria classificado como saudável. Contudo, como cada paciente possui 26 etiquetas binárias, poderia ocorrer, em alguns casos, uma quantidade igual de 0’s e 1’s. Esse problema de indeterminação, de fato, ocorreu em dois classificadores diferentes. No KNN, a indeterminação surgiu tanto nos conjuntos de dados de treino quanto nos de teste. Já no SVM, o problema ocorreu apenas no conjunto de teste. Para contornar essa questão, foi analisada qual gravação sonora apresentava o maior número de erros de predição no treinamento, e a coluna correspondente a essa gravação foi removida. Com isso, o número de gravações passou a ser 25 (um número ímpar), eliminando o problema de indeterminação causado pela quantidade igual de etiquetas binárias. Após a submissão dos conjuntos de dados a estes algoritmos citados anteriormente, foi calculado as métricas de desempenho de cada algoritmo tanto no treinamento quanto no teste, e também foi elaborada uma matriz de confusão para cada treinamento e para cada teste para cada algoritmo.

RESULTADOS E DISCUSSÃO

Cada algoritmo teve suas métricas de desempenhos aferidas, seguem abaixo as acurácias de cada algoritmo em cada etapa:

Tabela 1 - Acurácia dos algoritmos.

Fonte: Elaborado pelo autor.

Os algoritmos SVM e RFC apresentaram acurácias semelhantes, mas vale ressaltar que as matrizes de confusão de cada um foram diferentes. Com a finalidade de ter mais um elemento que ajude a entender a eficácia dos algoritmos, segue abaixo uma matriz confusão genérica e logo em seguida a matriz de confusão de cada algoritmo no treinamento e no teste.

Tabela 2 - Matriz de confusão genérica.

Fonte: Elaborado pelo autor.

Verdadeiros positivos são os pacientes que são portadores de DP e o classificador acertou na predição, verdadeiros negativos são os pacientes que são saudáveis e o classificador acertou na predição, falsos positivos são quando os pacientes são saudáveis e o classificador errou na previsão e falsos negativos são quando os pacientes possuem a DP mas o classificador afirmou que o paciente era saudável.

Tabela 3 - Matriz de confusão dos algoritmos no treinamento e no teste.

Fonte: Elaborado pelo autor.

CONCLUSÕES

Através das análises vistas acima, pelos valores da acurácia de cada algoritmo no treino e no teste e as tabelas que mostram as matriz de confusão de cada algoritmo no treino e no teste é possível inferir que para a metodologia usada neste projeto e com a quantidade de variáveis correlacionadas usadas, o algoritmo que apresentou um melhor desempenho na tarefa de prever o diagnóstico do paciente foi o KNN (K-Nearest

ISSN: 2447-6161

Neighbors).

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos ao Professor ANTONIO ALISSON PESSOA GUIMARÃES, que foi fundamental como meu orientador ao longo do desenvolvimento deste projeto de pesquisa. Sua expertise, orientação e apoio constantes foram essenciais para a realização deste trabalho. Durante todo o processo, o Professor demonstrou um compromisso inabalável com minha formação acadêmica e com o progresso da pesquisa, oferecendo contribuições valiosas que elevaram a qualidade dos resultados obtidos.

Agradeço, de forma especial, pela sua disposição em esclarecer dúvidas, por suas sugestões enriquecedoras e pelo constante encorajamento, que foram determinantes para superar os desafios encontrados no decorrer da pesquisa. O tempo e o esforço dedicados à minha orientação foram, sem dúvida, fundamentais para o sucesso deste projeto, e sou profundamente grato por ter tido a oportunidade de aprender sob sua tutela.

Agradeço também à Unilab pelo apoio financeiro ao projeto intitulado ANÁLISE E PREDIÇÃO DA ESCALA HOEHN E YAHR DE AVALIAÇÃO DA DOENÇA DE PARKINSON: UMA ABORDAGEM DE MACHINE LEARNING VIA OS ALGORITMOS REGRESSÃO LOGÍSTICA MULTICLASSE E REDES NEURAIS ARTIFICIAIS, realizado entre 01/10/2023 e 30/09/2024, no âmbito do Programa Institucional de Bolsas de Iniciação Científica (Pibic).

REFERÊNCIAS

COVER, T.; Hart, P. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, v. 13, n. 1, p. 21-27, 1967.

CORTES, C.; Vapnik, V. Support-Vector Networks. Machine Learning, v. 20, n. 3, p. 273-297, 1995.

BREIMAN, L. Random Forests. Machine Learning, v. 45, n. 1, p. 5-32, 2001.

COX, D. R. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society: Series B (Methodological), v. 20, n. 2, p. 215-242, 1958.

ISSN: 2447-6161