MODELAGEM DE REGRESSÃO LOGÍSTICA NA PREDIÇÃO DA DOENÇA DE PARKINSON

Ivina Lorena Oliveira Moura[1]
Jorge Antônio Félix Da Silva[2]
Antonio Alisson Pessoa Guimarães[3]

RESUMO

A Doença de Parkinson (DP) é uma disfunção neurodegenerativa do sistema nervoso central, crônica e progressiva que provoca tremores de repouso (TR), lentidão de movimentos, rigidez muscular, desequilíbrio, além de alterações na fala e escrita. Basicamente, é causada por uma intensa diminuição da produção de dopamina, um neurotransmissor, ou seja, uma substância química que ajuda na transmissão impulsos elétricos entre as células nervosa e devido à sua natureza progressiva, esta doença necessita de um contínuo monitoramento dos sintomas motores. Contudo, o diagnóstico precoce é essencial, pois pode favorecer ao paciente um tratamento eficaz, retardando assim a evolução da DP. Neste contexto, muitos métodos de inteligência computacional baseados em Machine Learning têm sido aplicados para o diagnóstico da doença. Particularmente, este trabalho objetiva propor um classificador computacional, baseado no modelo de Regressão Logística Regularizada (RL-Reg), que auxiliará o médico neurologista no diagnóstico da Doença de Parkinson. A simulação e treinamento do modelo, do qual é supervisionado, levará em consideração uma base de dados pública consolidada na literatura, com extração de características específicas de 40 amostras de voz, tendo-se nesse conjunto pacientes portadores e não portadores de DP. Além disso, a eficiência do método proposto será rigorosamente avaliada em relação à métricas de precisão de classificação. Por fim, promissoramente, vislumbra-se que o simulador seja inserido à sociedade como uma alternativa eficiente às ferramentas existentes no auxílio ao prognóstico de DP.

Palavras-chave: Algoritmo de Predição;; Doença de Parkinson;; Regressão Logística Regularizada;; Machine Learning;.

Universidade da Integração Internacional da Lusofonia Afro-Brasileira, Unidade Acadêmica dos Palmares, Discente, ivinalorena@aluno.unilab.edu.br[1]
Universidade da Integração Internacional da Lusofonia Afro-Brasileira, Unidade Acadêmica dos Palmares, Discente, jorgefelix@aluno.unilab.edu.br[2]
Universidade da Integração Internacional da Lusofonia Afro-Brasileira, Campus das Auroras, Docente, alisson@unilab.edu.br[3]

ISSN: 2447-6161

INTRODUÇÃO

A doença de Parkinson (DP) é uma doença neurológica crônica que afeta o sistema nervoso central, responsável pelo controle do movimento. Os sintomas geralmente começam gradualmente e incluem tremores involuntários, rigidez muscular, lentidão dos movimentos e instabilidade postural. Alterações na fala e na escrita também podem ocorrer. A DP ocorre devido à degeneração de células na substância negra, uma região do cérebro que produz dopamina, um neurotransmissor importante para o controle do movimento e sua falta afeta os movimentos (JUNIOR, 2004; SAKAR et al., 2013).

A pesquisa teve início com uma análise aprofundada do problema de classificar pacientes com DP em duas categorias: pessoas diagnosticadas com a doença de Parkinson e indivíduos sem o diagnóstico da doença. A classificação foi baseada em dados coletados a partir das características dos sinais de voz de cada paciente. Esses dados revelaram padrões que podem ser usados para melhorar o diagnóstico e o tratamento da doença. A escolha do algoritmo utilizado foi de acordo para adequar à tarefa de classificação e sua eficácia na resolução dos problemas envolvidos. A classificação ocorre por meio dos dados binários, logo o algoritmo escolhido foi o de Regressão Logística Regularizada para o desenvolvimento do presente trabalho, pois atende aos critérios e é adequado para solucionar o problema proposto.

Em suma, esta obra visa contribuir de forma notável para a saúde pública, ao mesmo tempo em que busca desenvolver novos métodos que desempenham um papel fundamental na melhoria dos diagnósticos e, consequentemente, no aprimoramento dos tratamentos. Além disso, enfatiza a importância do diagnóstico precoce, um fator que oferece benefícios substanciais aos pacientes, uma vez que a detecção nos estágios iniciais possibilita retardar significativamente a progressão da doença, aumentando a qualidade de vida dos pacientes. Portanto, o presente trabalho tem como objetivo a utilização do algoritmo de predição, Regressão Logística Regularizada a detectar a doença de Parkinson com a finalidade de auxiliar na análise clínica, para uma detecção da DP nos estágios iniciais facilitando, assim, uma implementação de tratamentos precoces e mais eficazes.

METODOLOGIA

Para a realização e desenvolvimento de um sistema auxiliar para a detecção de distúrbios do Parkinson por meio da análise da voz dos pacientes, a metodologia empregada abrangeu desde a seleção e processamento das características vocais mais relevantes, como a etapa de treinamento e a avaliação do desempenho do modelo resultante. O algoritmo escolhido para abordar a questão proposta foi a Regressão Logística Regularizada, uma técnica supervisionada e de classificação binária. Nesse enfoque, o algoritmo assimila conhecimento a partir de um conjunto de dados previamente rotulados, adquirindo a capacidade de antecipar o resultado para novos dados com base nas aprendizagens conquistadas.

Inicialmente, houve um tratamento dos dados públicos disponibilizados em UCI Machine Learning Repository: Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set. Os dados pertencem a 20 pacientes diagnosticados com Parkinson e 20 pessoas saudáveis que recorreram ao Departamento de Neurologia da Faculdade de Medicina de Cerrahpasa, na Universidade de Istambul.

Com a finalidade de orientar a identificação de distúrbios de Parkinson, o processo inicial envolveu cuidadosamente a escolha das características de voz mais relevantes. Esses atributos foram selecionados considerando sua relação com o diagnóstico de Parkinson, usando informações de pacientes que já tinham sido diagnosticados anteriormente.

Na etapa de treinamento, procedeu-se à capacitação do algoritmo de Regressão Logística Regularizada com

ISSN: 2447-6161

um conjunto de 40 pacientes, atentando-se à análise fonética do alfabeto proferido por cada paciente. A partir deste processo de treinamento, foram originadas avaliações parciais relativas a cada letra (Figura 1), as quais foram submetidas à análise de métricas modelares, incluindo precisão, recall e F1-score. A avaliação parcial de cada letra consistia em classificar, com uma avaliação binária, o paciente, ou seja, como Parkinson (1) ou Não-Parkinson (0). Foi elaborado outra figura (Figura 2) que mostra melhor o comportamento do erro de diagnóstico por letra, ou seja, as letras que mais erraram por diagnóstico.

Figura 1 – Comportamento da acurácia, precisão e recall por cada letra do alfabeto



Fonte: Elaboração própria.

Figura 2 – Comportamento do erro de diagnóstico por letra



Fonte: Elaboração própria.

Após a conclusão desta fase, foi instituído um critério no diagnóstico que foi posteriormente incorporado a um banco de dados genérico, concebido sob a orientação do Supervisor, a fim de aprimorar nossa compreensão da estrutura que desejava-se construir e organizar. O banco de dados genérico compreendia as classificações parciais das 26 letras referentes aos 40 pacientes, culminando em uma soma das ocorrências de uns e zeros a fim de determinar a classificação definitiva de cada paciente, o resultado após as operações era mostrado no console (Figura 3).

Figura 3 – O resultado é exibido no console nesse formato



Fonte: Elaboração própria.

Ao término de cada avaliação por letra, ocorria uma soma dos valores de zeros e uns, onde a prevalência determinava o diagnóstico definitivo do paciente. Nesse sentido, considerando a presença de um conjunto de 26 letras, existia a possibilidade de uma situação em que a classificação resultasse em uma divisão equitativa de 13/13, caracterizando um cenário de indeterminação. Para abordar essa possibilidade, foi desenvolvida uma função destinada a minimizar tais indeterminações. Essa função opera excluindo a letra de pior resultado do treinamento com a menor correlação em relação ao diagnóstico final, a partir do conjunto das 10 letras mais fortemente correlacionadas com o diagnóstico. Em outros termos, a última coluna da base de dados era eliminada.

Na fase inicial, abrangendo o treinamento do algoritmo, as ações foram conduzidas por meio do MATLAB. Posteriormente, para a aplicação e execução dos resultados obtidos no treinamento, a abordagem se baseou na linguagem Python, com a utilização da biblioteca Pandas para manipulação de estruturas de dados em forma de data frame e a biblioteca Numpy que é uma função destinada a trabalhar com modelos de matrizes e oferecendo uma vasta gama de operações matemáticas. Nesse ponto, foi onde pôde ser visto a parte prática após o treinamento.

No código desenvolvido, após treinarmos os algoritmos com os dados fornecidos em MATLAB, passamos toda parte do teste e validação do algoritmo para Python. No código elaborado, procedemos à extração dos dados

ISSN: 2447-6161

individuais de cada um dos 40 pacientes a partir de arquivos Excel. Posteriormente, importamos os parâmetros obtidos durante a fase de teste e realizamos uma manipulação matricial a partir dos dados de cada paciente com os dados da matriz de pesos obtidos por cada letra, e em seguida tal manipulação alimentou a função logística que gera como resultado a classificação simulada de Parkinson e Não-Parkinson. O produto dessa manipulação foi armazenado em uma variável denominada "resultados". Vale ressaltar que o foco coincidiu na diagonal dessa matriz resultante, derivada da multiplicação entre os conjuntos de dados de todos os pacientes e as matrizes de pesos exportados do MATLAB. Guardou-se o resultado na variável chamada de “resultados”.

Nessa variável foi aplicado a função sigmoide, que é muito utilizada em algoritmos de classificação binária. Se o valor resultante da função sigmoide for maior que 0.5, o modelo classifica a entrada como 1, caso contrário, a classificação é para 0.

Após a aplicação da função sigmoide para transformar os dados. Em sequência, criamos um terceiro data frame visando a apresentação mais clara dos resultados da função sigmoide ao longo da diagonal. Este novo data frame compreendia duas colunas: "Sigmoide" e "Classificação". A função de classificação atribuí "1" para "Parkinson", caso contrário, "Não-Parkinson" de acordo com a situação. Conforme anteriormente mencionado, quando ocorria uma equitativa classificação de 13/13, um cenário de indeterminação era identificado. Nesse contexto, a última coluna, era removida e, posteriormente, o cálculo original do código era novamente iniciado. Subsequentemente, uma decisão era tomada pelo algoritmo com base na contagem dos valores zero e um. O diagnóstico definitivo era, então, exibido no console.

RESULTADOS E DISCUSSÃO

No processo de concepção e elaboração do projeto, foi empregado o algoritmo de Regressão Logística Regularizada. Este algoritmo extrai e internaliza as informações contidas nos dados guardados, buscando adquirir entendimento para efetuar previsões quando novos dados são inseridos.

Para a parte da avaliação e a precisão do aprendizado do algoritmo, estabelecemos uma tabela na qual a classificação do paciente era inserida na coluna denominada "Diag. Real". Em seguida, comparamos os resultados obtidos pelo algoritmo e os comparamos com a classificação real, registrando esses resultados na coluna "Resultado". Adicionalmente, na mesma tabela.

Dessa forma, a tabela não apenas fornece uma visão geral da concordância entre as previsões do algoritmo e as classificações reais, mas também oferece um entendimento adicional sobre as situações em que o processo de indeterminação influenciou o resultado.

Após a análise, foram identificados os seguintes resultados:

- Matriz de confusão:

	Classe positiva predita (1)	Classe normal predita (0)
Classe positiva real (1)	20	0
Classe normal real (0)	2	18

- Métricas de desempenho:

	Resultado
Acurácia	95.00000
Precisão	90.90909
Recall	100.00000
Score-F	95.23810

ISSN: 2447-6161

- Resultado comparativo (tabela simplificada):

	Diag. Simulado	Diag. Real	Resultado
Paciente 1	1	1	acerto
Paciente 2	1	1	acerto
	...		
Paciente 24	1	0	erro
Paciente 25	0	0	acerto
	...		
Paciente 39	0	0	acerto
Paciente 40	1	0	erro

Na tabela, o algoritmo errou no paciente 24 e 40.

CONCLUSÕES

Neste trabalho, testamos a do algoritmo de Regressão Logística Regularizada como uma ferramenta efetiva na identificação precoce da doença de Parkinson, e a rápida descoberta da doença, não apenas melhora a qualidade de vida dos pacientes, mas também abre possibilidades para tratamentos mais eficientes e personalizados.

Este estudo mostrou como a análise de dados, quando combinada com técnicas de aprendizado de máquina, podem aprimorar a precisão dos diagnósticos. Os resultados mostraram que o algoritmo de estudo, Regressão Logística Regularizada, se mostrou uma abordagem promissora para a detecção de DP em pacientes nos estágios iniciais da doença. A sinergia entre análise de dados e prática clínica tem o potencial de transformar o tratamento e o diagnóstico de doenças neurológicas. No entanto, é fundamental reconhecer que ainda há muito a ser feito para aprimorar o código desenvolvido, bem como para incorporar novos dados ao banco de informações, visando alcançar um desempenho pleno e eficaz do código. Um dos objetivos a serem alcançados é a criação de uma interface, para uma melhor experiência do usuário final.

Por fim, o trabalho obteve resultados promissores que têm o potencial de trazer amplas melhorias para a sociedade afetada pela DP, porém, ainda é o início do que poderá ser alcançado e melhorado.

AGRADECIMENTOS

Expresso minha gratidão ao Orientador, Antonio Alisson Pessoa Guimarães, pelo apoio e pela oportunidade de aprofundar meu conhecimento na área de Machine Learning. Seus direcionamentos e explicações foram fundamentais para definir o rumo da minha pesquisa. Agradeço também ao **Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)** pelo financiamento da pesquisa intitulada Classificação de Distúrbios de Movimento em Pacientes Diagnosticados com Doença de Parkinson e executada entre 01/09/2022 e 31/08/2023, através do Programa Institucional de Bolsas de Iniciação Científica (Pibic) e Tecnológico (Pibiti), da Unilab.

REFERÊNCIAS

ISSN: 2447-6161

1. CASTRO, A. N., FERRARI, D. G., Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações. Editora Saraiva, 2016.
2. SAKAR , B. E. et al Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. IEEE Journal of Biomedical and Health Informatics, v. 17, n. 4, 8828-834, 2013.

ISSN: 2447-6161