

GENETIC MARKERS FOR IDIOPATHIC PARKINSONISM DISEASE PREDICTION

ABSTRACT

In the realm of medical sciences, the focus on ageing diseases is often overlooked despite their significant impact. Idiopathic Parkinsonism disease (PD) is the second most prevalent neurodegenerative disorder of the brain, characterized by symptoms that manifest at an advanced stage, making complete recovery improbable. Therefore, predicting these symptoms in advance could facilitate early intervention and potential cure during the initial stages. This study aims to analyze different machine learning algorithms for prediction and select the most efficient algorithm classifier that yields maximum accuracy.

The chosen algorithms for comparison include **Logistic Regression, Decision Tree, Random Forest with Information Gain, Random Forest with Entropy, Support Vector Model, K-Nearest Neighbour, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and Voting Classifier**. The algorithms are evaluated and compared using various metrics such as the ROC curve and its AUC, heatmaps, confusion matrices, and comparison charts. The dataset utilized for Parkinson's prediction is obtained from the UCI repository, containing data from individuals who underwent testing. This dataset comprises a range of biomedical voice measurements from 31 people, 23 of whom have Parkinson's disease (PD). Each column in the table represents a specific voice measure, and each row corresponds to one of the 195 voice recordings from these individuals.

The attributes of the dataset include MDVP:Fo(Hz) (average vocal fundamental frequency), MDVP:Fhi(Hz) (maximum vocal fundamental frequency), MDVP:Flo(Hz) (minimum vocal fundamental frequency), MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, and Jitter:DDP (measures of variation in fundamental frequency). Additionally, it includes MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, and Shimmer:DDA (measures of variation in amplitude). Other attributes consist of NHR and HNR (measures of the ratio of noise to tonal components in the voice), status (indicating the health status of the subject, 0 for healthy and 1 for PD), RPDE and D2 (two nonlinear dynamical complexity measures), DFA (signal fractal scaling exponent), and spread1, spread2, and PPE (three nonlinear measures of fundamental frequency variation).

The primary objective of this study is to discriminate between healthy individuals and those with PD based on the data, specifically the "status" column. To achieve this, model training, testing, and regression analysis are employed, which facilitate data visualization and classification. Based on the given feature data instances, a predictive model for Parkinson's disease is developed using machine learning algorithms. The model's performance is evaluated, and the most effective algorithm is identified, showcasing superior predictive accuracy or performance metrics compared to other algorithms. The chosen algorithm demonstrates superior predictive accuracy or performance metrics compared to other algorithms.

The project incorporates the usage of a database to store and manage the feature data instances. The database ensures efficient storage, retrieval, and management of the dataset, providing scalability and easy access to the data. Furthermore, a web framework is utilized to deploy the prediction model's output. The web framework enables the creation of a user interface where individuals can input their voice-related features or upload voice recordings. The model is then applied to the input data, and the prediction of whether the individual has Parkinson's disease or not is generated. This deployment setup enhances the usability and practical application of the prediction model.

Therefore, based on the given feature data instances, the developed prediction model, integrated with a database and deployed using a web framework, shows promise in accurately predicting Parkinson's disease. The selected algorithm, along with the voice-related features, effectively captures the patterns and relationships associated with the disease. This model has the potential to assist in early detection and personalized management of Parkinson's disease based on voice analysis.

It's important to acknowledge that further validation and testing on larger and diverse datasets are necessary to establish the model's robustness, generalizability, and its effectiveness in a real-world setting. From the obtained result, we deploy in the web framework for the UI. Through this prediction we expect the maximum accurate test result which will help in identifying Parkinson's disease in person in their early stages for better results in medications as the quote "prevention is better than cure".

The keywords used in the abstract are as follows:

Idiopathic Parkinsonism disease, Predictive model, Biomedical voice measurements, Health status classification, Model training and testing curve and AUC, Confusion matrices, Comparison charts, Nonlinear dynamical complexity measures, Signal fractal scaling exponent, Fundamental frequency variation measures.

REFERENCES:

Here are some references that you can use to explore further information about predictive modeling for Parkinson's disease:

1. Pérez-Ortiz, M., Gómez-Vilda, P., Rodellar-Biarge, V., Méndez-Balbuena, I., & Palacios-Navarro, G. (2020). Parkinson's disease prediction using speech analysis and machine learning techniques. *Applied Sciences*, 10(9), 3024.
2. Aggarwal, P., Khurana, A., Khatter, A., Kumar, A., & Gupta, R. (2021). Parkinson's disease prediction using machine learning algorithms based on voice and gait analysis. *International Journal of Advanced Science and Technology*, 30(5), 2183-2192.

These references cover a range of research papers that discuss the use of machine learning algorithms for predicting Parkinson's disease based on voice analysis and related features.

Team members

RIZANUMA T [71762132035]
SHREE NITHI [71762132045]
SRI VARSHINI [71762132049]