PROJECT ON

DETECTING PARKINSON'S DISEASE- using Modern ML algorithms

A Project Report Submitted in partial fulfilment on the completion of the project.

BY:

Sunny Chowdhary Muppala, B.E Computer Science

Sathyabama Institute of Science and Technology, Chennai- 119

Submitted to:

Mr K. Rajasekharam
Scientist "G"
DASQA, DRDL
Hyderabad.

2021

*****

ACKNOWLEDGEMENT

I thank the **Director, Defence Research & Development Laboratories and Director, DASQA** for providing the facilities to work throughout my project.

I express my sincere and deep gratitude to **Mr K. Rajasekharam, Technology Director, DASQSA** for providing an opportunity to work with them.

I am very much indebted to their constant support, encouragement and guidance. I am very grateful to them for their friendliness and co-operation.

I also thank **Mr Y. Harinath, Scientist "E", DASQA** who helped me throughout the project.

I also thank the **staff of DASQA, DRDL,** Kanchanbagh for providing a helping atmosphere throughout my stay.

Finally**,** I thank all of them who directly helped me during the project.

INDEX

1. ABSTRACT

Biomarkers derived from human voice can offer insight into neurological disorders, such as Parkinson's disease (PD), because of their underlying cognitive and neuromuscular function. PD is a progressive neurodegenerative disorder that affects about one million people in India, with approximately sixty thousand new clinical diagnoses made each year. Historically, PD has been difficult to quantify and doctors have tended to focus on some symptoms while ignoring others, relying primarily on subjective rating scales. Due to the decrease in motor control that is the hallmark of the disease, voice can be used as a means to detect and diagnose PD. With advancements in technology and the prevalence of audio collecting devices in daily lives, reliable models that can translate this audio data into a diagnostic tool for healthcare professionals would potentially provide diagnoses that are cheaper and more accurate. We provide evidence to validate this concept here using a voice dataset collected from people with and without PD. This paper explores the effectiveness of using supervised classification algorithms, such as Xtreme Gradient Boosting (xgboost),which is a new algorithm for Machine Learning developed with speed and efficiency in mind to accurately diagnose individuals with the disease. Our peak accuracy of 92.3077% provided by the machine learning model with RMSE of 0.277350 which is quite good. Also on computing the k-fold cross validation the model gives Average Test RMSE of 0.281145.

2. INTRODUCTION

2.1 ORGANIZATION PROFILE

Soon after attaining India's independence, a body called Defence Science and Organization was established in 1948 for the development of Defence Science. In 1958 a major Reorganisation was implemented by which the present DRDO (Defence Research & Development Organization) was formed. Amalgamating the Defence Science Organization with some of the existing technical development establishment did the information of DRDO. DRDO has gone through a phased program build-up of infrastructures and expansions of various Defence Science & Technology fields. Now the number of laboratories and small cells attached to DRDO cover practically all the Scientific and Technological disciples of Defence interest. Among these laboratories, DRDL is one of the laboratories. The DRDL plays a vital role in the technical development of national security. The DRDL is one of the biggest laboratories in the country under DRDO. DRDL is a large research and development laboratory engaged in the development of missiles. The whole infrastructure of DRDL consists of large manpower, sophisticated machines and a modern computer system with an advanced management system.

Functions of DRDL:

To design, develop weapons and equipment based on the operational requirements defined by the services and to help the indigenous production. The weapon system should be reliable, safe and cost-effective, should meet the time schedule without comprising quality, and aim at continuous quality improvement through the involvement of all members.

- To render scientific advice to the service headquarters.
- To carry out basic and applied research to solve the problems of the service.
- To evaluate and conduct the technical trail of new weapons and equipment which are designed and developed in the country.
- To render technical support to civil trade for the development.
- The development of the missile system is computer and software-intensive.

Computers are being used for the simulation modelling of the subsystem and checkout, launching, control and guidance of the missile. The majority of the software is developed in the house of DRDL.

The DRDL, as its name suggests, is one of the very few organizations in the country with the sole purpose of designing and developing models and types of equipment to quench the ever raising Defence requirements at a pace with the ever-changing world. Established in 1962, with its location in kanchanbagh, Hyderabad, with the staff of the famous and renowned scientist and sophisticated equipment, it has conquered the upper limits of the world with its technology and research.

DRDL is dealing with real-time systems, object-oriented technology, networking and the internet, multimedia, image processing, artificial neural networks, artificial intelligence and graphical user interfaces. With the use of sophisticated technology and maintenance of its manpower, it has reached the upper limits and still chasing them in an ever-changing world.

DASQA (Directorate of Avionics Software Quality Assurance):

DASQA (Directorate of Avionics Software Quality Assurance) is a special body in DRDL that deals with all aspects of the above-mentioned services. DASQA is well equipped with a sophisticated and advanced computer system as per the requirements of the projects it handles. The lab consists of its official internet applications and D-net of the DRDL, Hyderabad. The state of the art technology has been utilized whenever the need felt for producing high performance.

The organization is at its best in producing the most accurate machines and making the brilliance of its manpower as its backbone. Because of the endless flow of data and the need for the maintenance of secrecy, they opt for complicated design and encoding. At this venture, they had gone much far in the way to reduce the noise and to increase performance.

2.2 Background

Parkinson's disease is a progressive disorder of the central nervous system, characterized by the progressive degeneration of the structure and function of the nervous system. They are incurable and debilitating conditions that cause problems with mental functioning of an individual.

Parkinson's disease affect millions of people worldwide. It affects more than 1 million people in India per year. An estimated 930,000 people in the United States could be living with Parkinson's disease by 2020.

Parkinson's disease (PD) is a neuropathological disorder which deteriorates the motor functions of the human body. It is the second most common neurological disease seen after Alzheimer's disease and it is estimated that more than one million people are suffering from PD in North America alone. In 1817, PD was termed as shaking palsy by Dr. James Parkinson. Various studies have shown that this number will rise in an ageing population as it is commonly seen in the people whose age is over 60.

Parkinson's disease is characterized by the degeneration of certain brain cell clusters that are responsible for producing the neurotransmitters that include dopamine, serotonin and acetylcholine. The loss of dopamine's result in the symptoms like anxiety, depression, weight loss and visual problems. The other symptoms that can be seen in the people with Parkinson's disease are poor balance, voice impairment and tremor. Various research studies have shown that 90% of people who suffer from PD have speech and vocal problems which include dysphonia, monotone and hypophonia. Thus, the degradation of voice is considered to be as the initial symptom of Parkinson's disease.

The cause and cure of PD are yet unknown but the availability of various drug therapies offers the significant mitigation of symptoms especially at its earlier stages, thus improving the life quality of patients and also reduces the estimated cost of the Pathology. The analysis of voice measurement is simple and non-invasive. Thus, to track the progression of PD the measurement of voice can be used. For assessing the progression of PD, various vocal tests have been devised which include sustained phonations and running speech texts. The telemonitoring and telediagnosis systems have been widely used as these systems are based on speech signals which are economical and easy to use. Hence, in this paper, there is an attempt to explore a better machine learning based model for an early detection of PD from the voice samples of the subject.

Purpose of the Project

Early detection of a Parkinson's disease could be useful for the identification of people who can participate in trials of its agents, or ultimately to try and halt disease progression once effective disease-modifying interventions have been identified.

Problem Statement

The goal of this project is to build a model to accurately predict the presence of Parkinson's disease in an individual.

Objective

This paper aims to build a model using an XGBClassifier that accurately predicts the presence of Parkinson's disease in an individual using the mentioned dataset. We will use the python libraries; scikit-learn, numpy, pandas, and xgboost. We'll load the data, explore the data, get the features and labels, then split the dataset, build an XGBClassifier, and then calculate the accuracy and the RMSE of our model. And perform k-fold cross validation to make our model more robust. At last we will calculate the feature importance on our predictive modeling problem.

Structure of thesis

The outline of this thesis is shown as follows -

In Chapter 3, Literature Reviews of some of the previous related studies are provided. In Chapter 4, Information on Machine Learning Algorithms used.
In Chapter 5, Information on Dataset used
In Chapter 6, Software requirements will be provided.
In Chapter 7, Methodology will be discussed.
In Chapter 7, Results will be provided.
In Chapter 8, The conclusion and recommendation will be provided.

3. LITERATURE REVIEW

3.1 Literature Reviews of some of the previous related studies

- Richa Mathur et al [23] suggested a method for predicting the PD. They used a weka tool for implementing the algorithms to perform preprocessing of data, classification and the result analysis on the given dataset. They used k-NN along with Adaboost.M1, bagging, and MLP. It was observed that k-NN + Adaboost.M1 yielded the best classification accuracy of 91.28%.
- A.Yasar et al [24] used artificial neural networksfor the detection of Parkinson's disease. The dataset was taken from UCI machine learning repository. Using the MATLAB tool, 45 properties were chosen as input values and one output for the classification. Their proposed model was able to distinguish the healthy subjects from the PD subjects with an accuracy of 94.93%.
- Max A. little et al [15] suggested a novel technique for the classification of subjects into Parkinson diseased and control subjects by detecting dysphonia. In their work, pitch period entropy (PPE) a new robust measure of dysphonia was introduced. The data was collected from 31 people (23 were PD patients and 8 were healthy subjects) which comprised of 195 sustained vowel phonations. Their methodology consisted of three stages; feature calculation, preprocessing and selection of features and finally the classification. For the classification purpose, they used linear kernel support vector machine (SVM). Their proposed model achieved an accuracy of 91.4%.
- To separate the healthy subjects from PD subjects, Ipsita Bhattacharya et al [20] used a tool for data mining known as weka. They used SVM, a supervised machine learning algorithm for the classification purpose. Prior to classification, the data preprocessing was done on the dataset. Different kernel values were used to get the best possible accuracy by applying libSVM. The linear kernel SVM produced the best accuracy of 65.2174%, whereas the RBF kernel and polykernel SVM achieved the accuracy of 60.8696%

4. ML ALGORITHM

4.1 What is Machine Learning?

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision- making processes based on data inputs.

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are: -

- **Supervised learning** which trains algorithms based on example input and output data that is labeled by humans.
- **Unsupervised learning** provides the algorithm with no labeled data in order to allow it to find structure within its input data.

4.2.Algorithm used- XGBoost?

XGBoost is a new Machine Learning algorithm designed with speed and performance in mind. XGBoost stands for eXtreme Gradient Boosting and is based on decision trees.

How XGBoost Algorithm works?

- XGBoost builds really short and simple decision trees iteratively.
- XGBoost starts by creating a first simple tree which has poor performance by itself.
- It then builds another tree which is trained to predict what the first tree was not able to, and is itself a weak learner too.
- The algorithm goes on by sequentially building more weak learners, each one correcting and reduce the errors of the previous tree until a stopping condition is reached.

Here's a popular graphic from the XGBoost website as an example (Fig. 3.1):

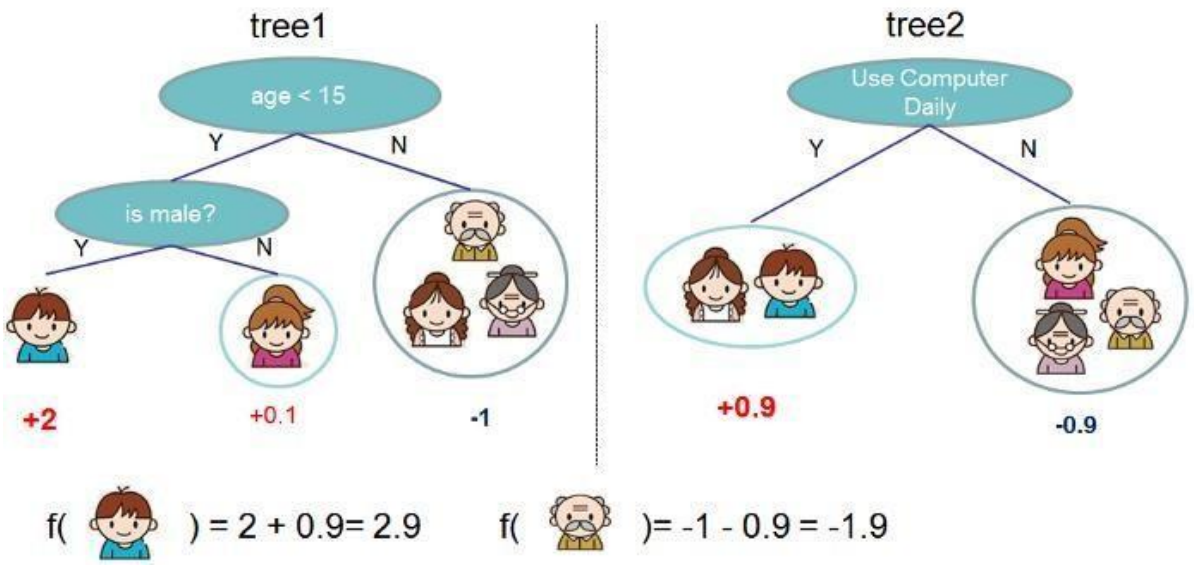

Fig. 3.1: XGBoost example

3.1.1. Advantages of XGBoost:

- **Better Speed and performance**: XGBoost is comparatively faster and it has shown better performance over other algorithms on a variety of machine learning benchmark datasets.

- **Regularization**: Standard GBM implementation has no regularization like XGBoost, therefore it also helps to reduce overfitting.

- **Parallel Processing:** XGBoost utilizes the power of parallel processing and that is why it is much faster than GBM. It uses multiple CPU cores to execute the model.

- **Handling Missing Values:** XGBoost has an in-built capability to handle missing values

5. DATASET

5.1 About the Dataset:

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

From existing metadata, we get some important information, such as:

- Title: Parkinson's Disease Data Set
- Abstract: Oxford Parkinson's Disease Detection Dataset
- Data Set Characteristics: Multivariate
- Number of Instances: 197
- Number of Attributes: 23

5.2 Dataset Information:

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to the "status" column which is set to 0 for healthy and 1 for PD.

5.3 Attribute Information:

- Matrix column entries (attributes):
- name - ASCII subject name and recording number
- 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,Jitter:DDP - Several
- measures of variation in fundamental frequency
- MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP :APQ,Shim mer:DDA - Several measures of variation in amplitude
- NHR,HNR - Two measures of ratio of noise to tonal components in the voice
- status - Health status of the subject (one) - Parkinson's, (zero) - healthy
- RPDE,D2 - Two nonlinear dynamical complexity measures
- DFA - Signal fractal scaling exponent
- spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

This dataset has been downloaded from UCI ML repository.

6. SOFTWARE REQUIREMENTS

6.1 Software Used:

1. Anaconda Navigator

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system *conda*. The Anaconda distribution is used by over 15 million users and includes more than 1500 popular data-science packages suitable for Windows, Linux, and macOS.

We can download this from the official website of Anaconda Navigator. We will need Jupiter Notebook, which we can directly use through Anaconda Navigator

2. Python 3.7 (64-bit)

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. It is high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed.

1. Open a browser window and navigate to the Download page for Windows at python.org.

2. Underneath the heading at the top that says Python Releases for Windows, click on the link for the Latest Python 3 Release - Python 3.x.x. (As of this writing, the latest in Python 3.7.x)

3. Scroll to the bottom and select Windows x86-64 executable installer for 64-bit. We strongly recommend 64-bit because TensorFlow doesn’t support 32-bit.

4. Once you have chosen and downloaded an installer, simply run it by double-clicking on the downloaded file.

5. Then just click Install Now. That should be all there is to it. A few minutes later you should have a working Python 3 installation on your system.

6.2 Libraries Used:

1. NumPy:

To install this package with conda run the following:

conda install -c conda-forge numpy

To install this package with python, run the following:

pip install numpy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

2. Sklearn:

To install this package with conda run one of the following:

conda install -c anaconda scikit-learn

To install this package with python, run the following:

pip install scikit-learn

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

3. Pandas:

To install this package with conda run one of the following:

conda install -c anaconda pandas

To install this package with python, run the following:

pip install pandas

Pandas offer data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

4. Matplotlib:

To install this package with conda run one of the following:

conda install -c conda-forge matplotlib
conda install -c conda-forge/label/testing matplotlib
conda install -c conda-forge/label/testing/gcc7
matplotlib conda install -c conda-forge/label/gcc7
matplotlib
conda install -c conda-forge/label/broken
matplotlib conda install -c conda-forge/label/rc
matplotlib
conda install -c conda-forge/label/cf201901 matplotlib

To install this package with python, run the following:

pip install matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

5. Seaborn:

Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

To install this package with conda run one of the following:

conda install seaborn

To install this package with python, run the following:

pip install seaborn

7. METHODOLOGY

Workflow Diagram:

The methodology for building a model to detect the Parkinson's disease at its early stage using the machine learning algorithms is presented in figure5.3. It consists of the following steps:

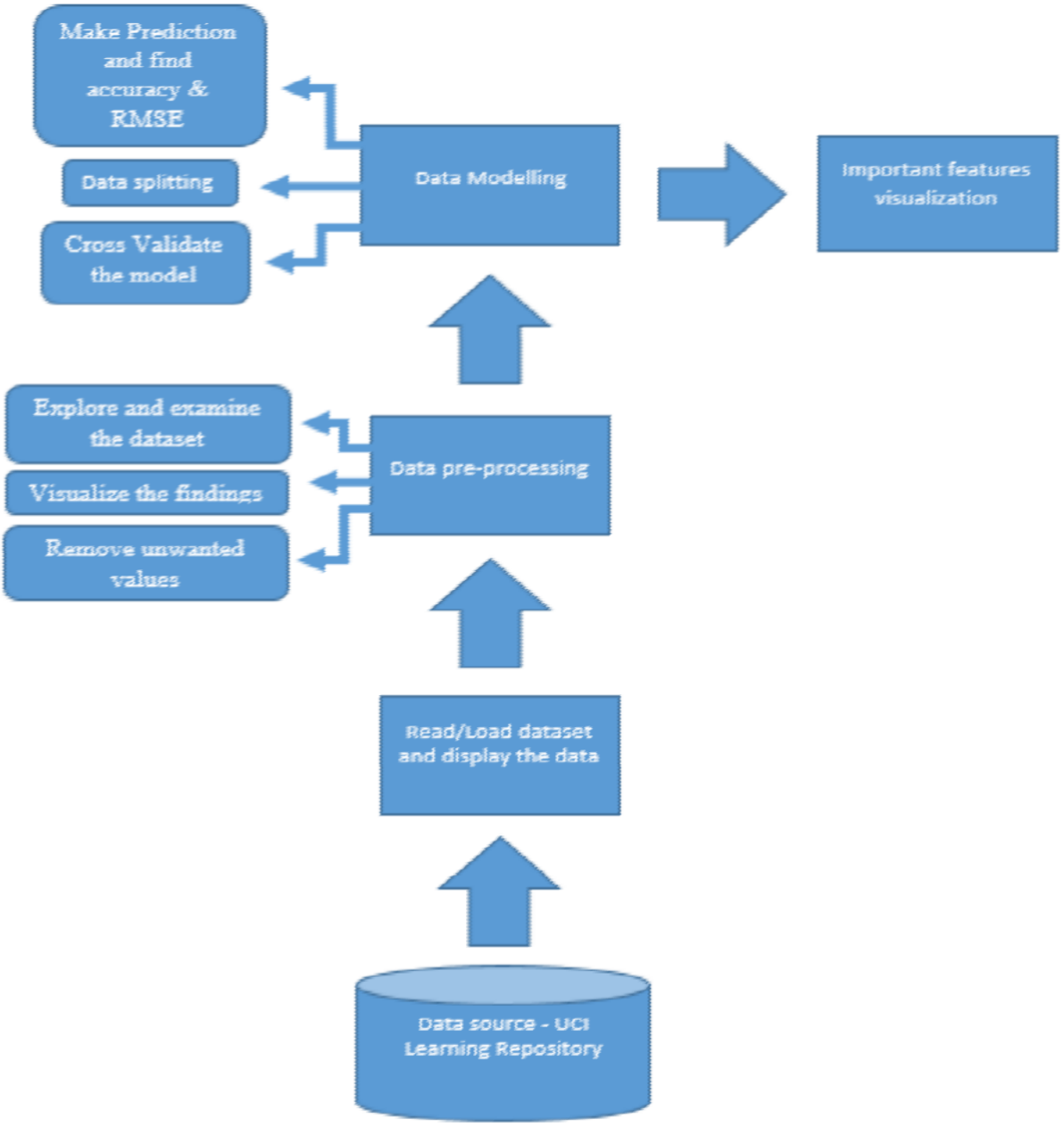

Fig. 5.3: Workflow Diagram

.

Explanation of the different stages depicted in the workflow diagram (Fig. 2) are as follows:

1. **Acquiring the dataset**

- We have download the voice sample dataset from UCI Learning Repository and stored in our PC.

2. **Read/Load the dataset**

Now we will load the dataset from our PC and display it in our code using Pandas data-frame.

3. **Data Pre-processing**

Here we will explore and examine our dataset (using .info, .corr, .describe methods), remove unwanted values and then visualize our findings using heat-map and bar-chart.

4. **Data Modelling**

Within this step we will split our data into training and testing part using sklearn library and we will do our prediction using the XGBoost algorithm and find out the Accuracy and Root Mean Square Error of the model's prediction after that we will again use k-fold cross validation in order to make our model more robust and again find out the mean of the Accuracy and mean Root Mean Square Error of the result of validation.

5. **Important feature visualization using XGboost**

Finally, we will visualize and find out the feature that has the highest importance among all the features

8. CONCLUSION

Currently, the Parkinson's disease research area is of much significance and its detection at the early stage can make the patient's life better. The recent developments in the methodologies through speech analysis have produced significant results. In our work, the problem of identification of Parkinson's disease is coped through a machine learning approach. The main aim of this work is to show the PD diagnosis by analysing the voice signals. From many years, speech processing has an incredible potential in the detection of PD as voice measurements are non- invasive. In our project we have used XGBoost machine learning algorithm. An accuracy of 92.3077% was provided by the machine learning model with RMSE of 0.277350 which is quite good. Also on computing the k-fold cross validation to test the model, the model gives Mean Test RMSE of 0.281145 which is very close to the RMSE given by our model initially. Finally using the XGBoost's plot_importance() function we have found out that the feature MDVP.Fo(Hz) (i.e. Average vocal fundamental frequency) has the highest importance score among all the features. Thus the proposed model is a reliable model to detect Parkinson's disease due to its efficient accuracy rates.

Though the model works efficiently, this is limited by the richness of the dataset with which it is being trained. The selected dataset, has only 197 instances, hence in future if we use a dataset with more no of samples it would help the model generalize even better.

9. REFERENCES

1. https://www.kaggle.com/prashant111/xgboost-k-fold-cv-feature-importance
2. https://randerson112358.medium.com/ai-in-health-2e9f84906bed
3. https://www.datacamp.com/community/tutorials/xgboost-in-python
4. https://codeburst.io/using-python-to-detect-early-onset-parki nsons-disease- b89651b0ed3
5. Sakar, C. O., & Kursun, O. (2010). Telediagnosis ofParkinson's disease using measurements of dysphonia.Journal of medical systems, 34(4), 591-599.
6. Rahn III, D. A., Chou, M., Jiang, J. J., & Zhang, Y. (2007).Phonatory impairment in Parkinson's disease: evidence
7. https://www.datainsightonline.com/post/xgboost_predicting-parkinson-diseases
8. D.J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for
9. Parkinson disease." Archives of neurology, vol. 56, no.1, pp. 3999.
10. R. Das, "A Comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications, vol. 37, no. 2, pp. 1568-1572, 2010.

BIBLIOGRAPHY

Master Muppala Sunny Chowdhary graduating from the beloved university Sathyabama Institute of Science and Technology in the stream of Computer Science and Engineering has excelled in this stream. His research interest is Etymology, Android Applications in real life, Neural Networking, Data Analysis.

Linkedin: https://www.linkedin.com/in/muppalasunnychowdhary/
GitHub: https://github.com/MuppalaSunnyChowdhary
Google Scholar: https://scholar.google.com/citations?hl=en&user=F3srbCEAAAAJ
Medium: https://muppalasunnychowdhary.medium.com/
Portfolio: https://muppalasunnychowdhary.github.io/