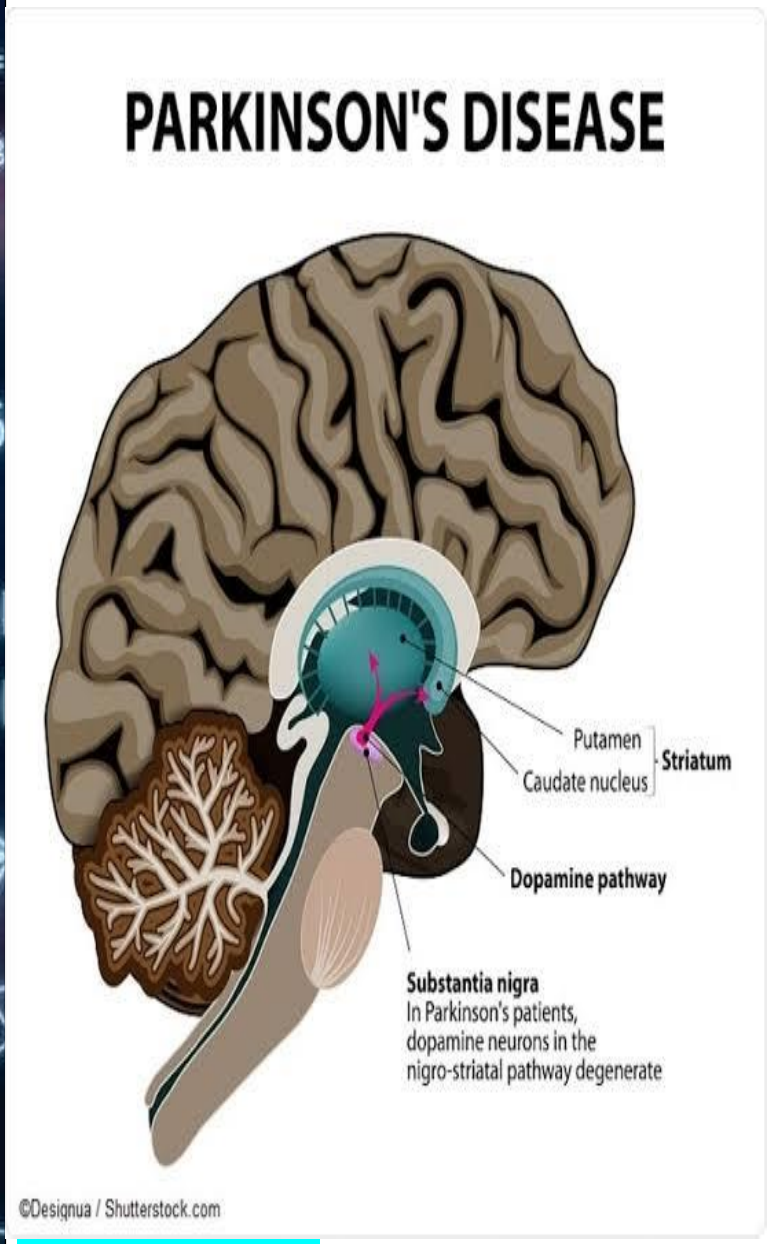
Parkinson's Disease Prediction Using Machine Learning



YERRAM SAI TEJA

Table of Contents

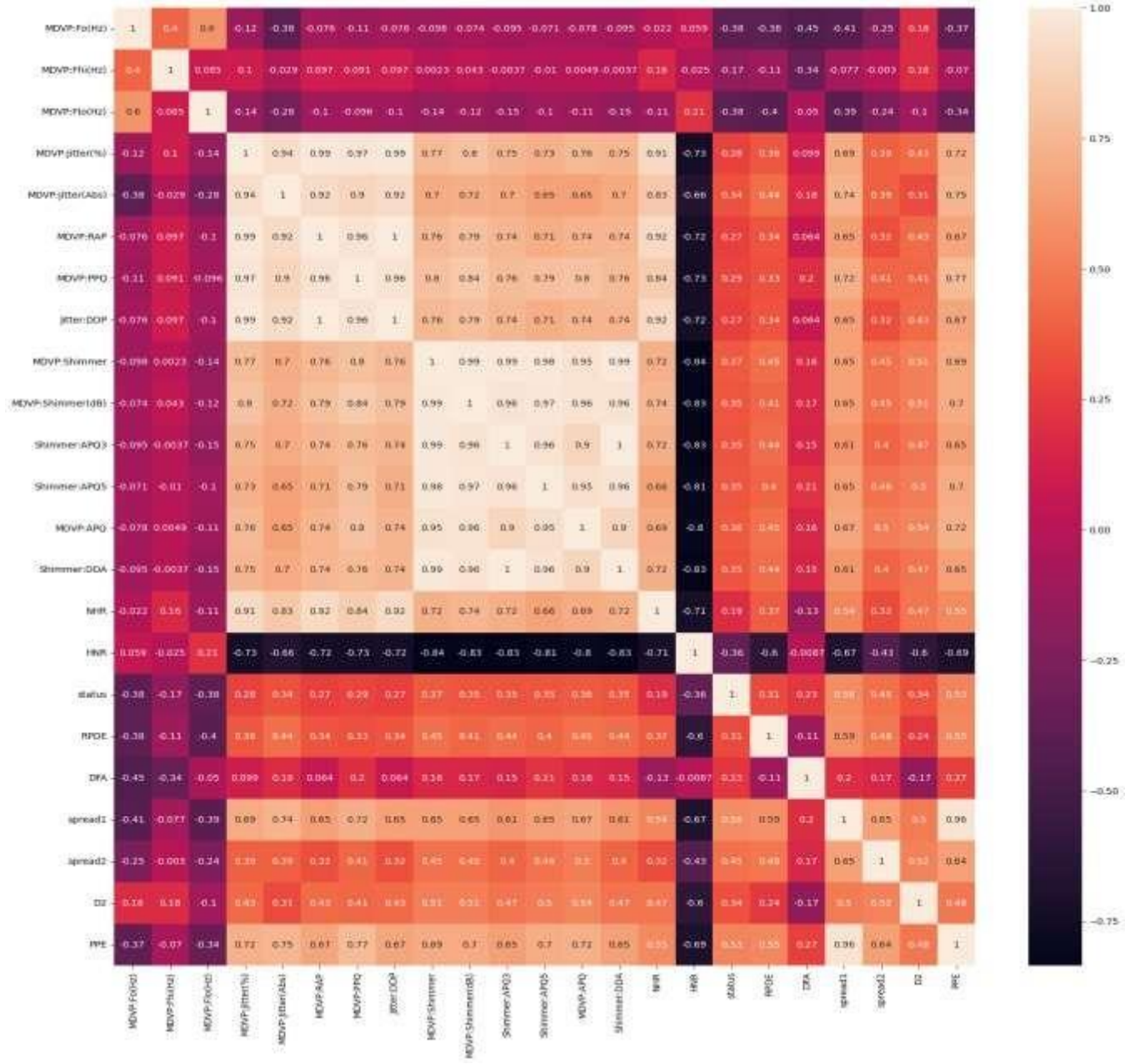
INTRODUCTION

OVERVIEW:

Parkinson's disease is a movement disorder that affects the central nervous system. The symptoms start gradually because of low dopamine levels in the brain. Dopamine is a chemical and is a neurotransmitter responsible for sending signals from the body to the brain. Reduction in the neurons responsible for production of dopamine leads to lowering of dopamine levels resulting in reduced coordination between the brain and the body. Till date there's not any cure for this disease and with the advancement of technology, it is necessary to introduce a quick and reasonable tool to predict the disease.
There are 5 stages of PD. The symptoms for all the stages are mentioned in the picture below.

Fig1: Symptoms associated with PD depending on stage

Machine learning can be used as a powerful tool for the prediction of various kinds of disease. It can help to predict the presence of PD more accurately and cost effectively. One of the main symptoms of PD is loss of speech. The resultant problems may include monotonic speech, slur words, mumbling, breathy or hoarse speech etc. Keeping the following points in mind, we decided to go with a voice dataset containing 25 features. In order to select the best features among these, Pearson method of correlation is used. The heatmap is plotted to find the correlation in the features. Here, we removed the values which are correlated and considered the remaining features.



PURPOSE:

The main aim is to predict the prediction efficiency that would be beneficial for the patients who are suffering from Parkinson and the percentage of the disease will be reduced. Generally, in the first stage, Parkinson's can be cured by the proper treatment. So, it’s important to identify the PD at the early stage for the betterment of the patients.

The main purpose of this research work is to find the best prediction model i.e., the best machine learning technique which will distinguish the Parkinson’s patient from the healthy person. The techniques used in this problem are KNN, Naïve Bayes, and Logistic Regression. The experimental study is performed on the voice dataset of Parkinson’s patients which is downloaded from the Kaggle.

The prediction is evaluated using evaluation metrics like confusion matrix, precision, recall accuracy, and f1-score. The author used feature selection where the important features are taken into consideration to detect Parkinson’s.

LITERATURE SURVEY

Title 1: Detection of Persons with Parkinson's Disease by Acoustic, Vocal, and Prosodic Analysis.
Author: Tobias Bocklet , Elmar N¨oth , Georg Stemmer , Hana Ruzickova , Jan Rusz.
Year: 2011
Description:
"In this paper, the author focused on the automatic discrimination between healthy speakers and speakers within early stages of PD by identifying the speech tasks with the most meaningful acoustic, prosodic, and vocal information to achieve this discrimination. The read speech and monologues contain the most important acoustic, prosodic and vocal information, when it comes to an automatic detection/ of PD speech. The best results was 90.5% recognition rate and 0.97 AUC was achieved with a prosodic system for the detection of PD speakers in early stages."

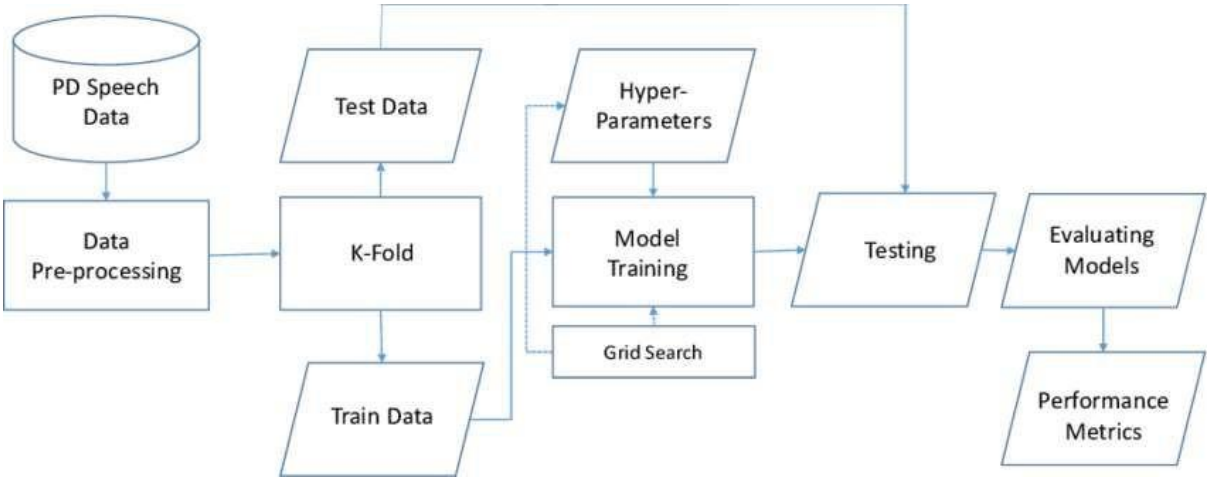
Title 2: Using Machine Learning to Diagnose Parkinson's Disease from Voice Recordings
Author: Akshaya Dinesh, Jennifer He
Year: 2017
Description:
"The author tested the ML model using Microsoft Azure Machine Learning Studio and found the best suited model to be Two-Class Boosted Decision Trees, an ensemble model of boosted regression trees made in a stepwise method. Another conclusion that was derived was that Spread1, spread2, and PPE in the dataset, had the strongest weights in labeling a patient as healthy or not. The results were that we successfully created a predictive model for PD from voice analysis. The best accuracy was found using Boosted Decision Trees and the accuracy achieved was 95%."

Title 3: Parkinson's Disease Diagnosis Using Machine Learning and Voice
Author: Timothy Wroge, Yasin O¨ zkanca,Cenk Demiroglu,Dong Si,David C. Atkins,Reza Hosseini Ghomi
Year: 2019
Description:
"This paper explores the effectiveness of using supervised classification algorithms, such as deep neural networks, to accurately diagnose individuals with the disease. The author‟s peak accuracy was of 85% using AVEC selected feature and Gradient Boosted Decision Tree exceeding the average clinical diagnosis accuracy of non-experts (73.8%) and average accuracy of movement disorder specialists (79.6% without follow-up, 83.9% after follow-up) with pathological postmortem examination as ground truth."

Title 4: Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification
Author: Marek Wodzinski , Andrzej Skalski,Daria Hemmerling,Juan Rafael OrozcoArroyave, Elmar N'oth
Year: 2019
Description:
"In this study, the author presented an approach to Parkinson‟s disease detection using vowels with sustained phonation and a ResNet architecture dedicated originally to image classification." The author converted the frequency based features into images and then performed the ResNet algorithm. The testing was performed using 10 fold cross validation.The accuracy achieved was over 90%.

THEORETICAL ANALYSIS

3.1 BLOCK DIAGRAM



3.2 HARDWARE/ SOFTWARE DESIGN

A.

Python
Python 3

B. Libraries

Numpy
pandas
joblib
sklearn
argparse
pickle
matplotlib, etc.

The non-functional requirements expected from the project are listed below:

- **Correctness:** All the details, reports provided to the system are correct. The system ought to be correct in terms of its functionality, calculations used internally and therefore the navigation ought to be correct.
- **Accuracy:** The system will provide accurate results to the user 96.6% of the time.
- **Reusability:** The components are compatible for changing environment and support upgradability.
- **Maintainability:** The application is easy to maintain.
- **Security:** The system provides security by asking to provide correct credentials to log in to the account.
- **Reliability**: The website will be available and will provide consistent output to users
- **Availability:** The system is functional throughout and data transfer takes place throughout user request

1. SOFTWARE REQUIREMENTS:
    1. Operating System : Windows 10
    2. Framework : Flask
    3. Languages used : Python, Javascript, HTML, CSS
    4. Tools : Visual Studio Code, Jupyter Notebook, Anaconda

2. HARDWARE REQUIREMENTS:
    1. Processor : Intel Core i5, 7th generation CPU @2.70GHz
    2. Hard-Disk : 250 GB or Higher
    3. Ram : 8 GB

EXPERIMENTAL ANALYSIS:

DATA SOURCE

The raw dataset was collected from kaggle.
Details: “This dataset consists of a range of biomedical voice measurements from thirty-one people,out of which twenty-three people have Parkinson's Disease. Every column in the table corresponds to a particular voice measure, and every row corresponds to one of 3315 voice recordings from these people („name‟ column). The main purpose of the data is to distinguish healthy people from those with PD based on the column "status" which is equal to 0 for healthy and 1 for PD.”
The raw dataset has been preprocessed to get a standardized, clean, and normalized dataset which has been utilized for training and testing the model.
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,Shimmer: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

An abundance of data is utilized in an application such as a prediction system. It is essential to organize the data in some standard format for utilization. This is a process of organising data. The designer understands the data available and identifies them as entities, attributes and their inter-linkage. With this analysis a

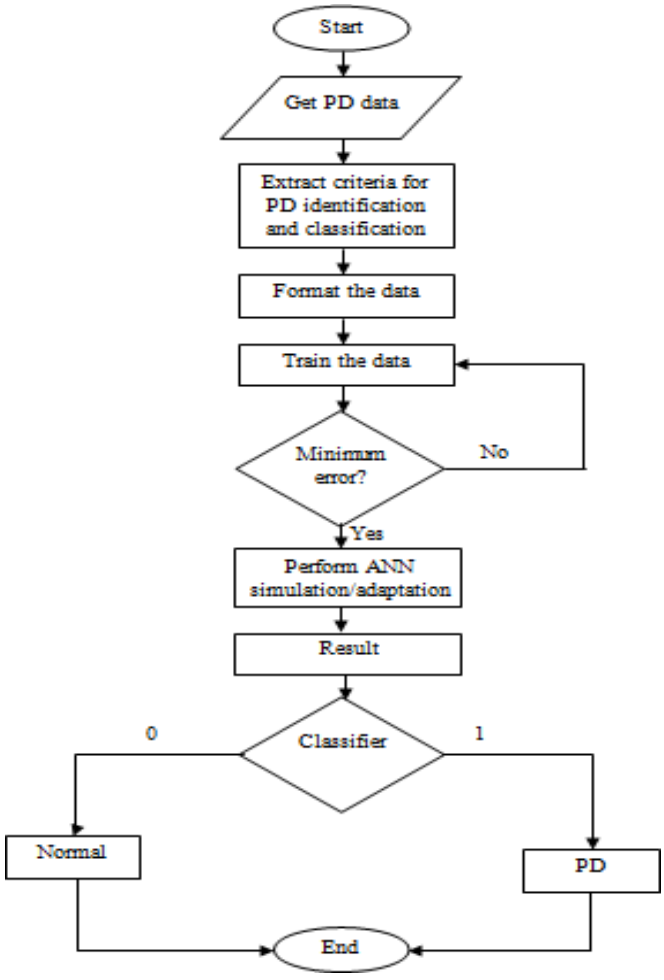
database model can be prepared i.e., the data can be fit into a format. A suitable database management system and tools can be utilized to create, insert, update and manipulate the model and the data itself.

MACHINE LEARNING IMPLEMENTATION ENSEMBLE MODEL:

In order to build a more optimistic model, a combination of various ML algorithms are implemented and the result or prediction of each model in the combination is collected and the final result is calculated. This process of combining multiple base models to make a more efficient model is called an Ensemble Model. The algorithms that we have decided to use in the analysis for choosing the optimistic combination for the proposed system are:

- Random Forest
- SVM
- XGBoost
- KNeighborsClassifier
- Logistic Regression
- Decision tree
- Gaussian Naive Bayes
- Bernoulli Naive Bayes
- voting classifier(Bagging algorithm)

FLOWCHART



RESULT

1. RESULT

- The model is trained successfully with the dataset.
- The application runs smoothly without any glitch.
- The user is able to give the respective inputs in the UI.
- After entering all the details, the result is predicted and displayed on the web-application 86 screen.
- The model operates seamlessly with relevant error and exception handling

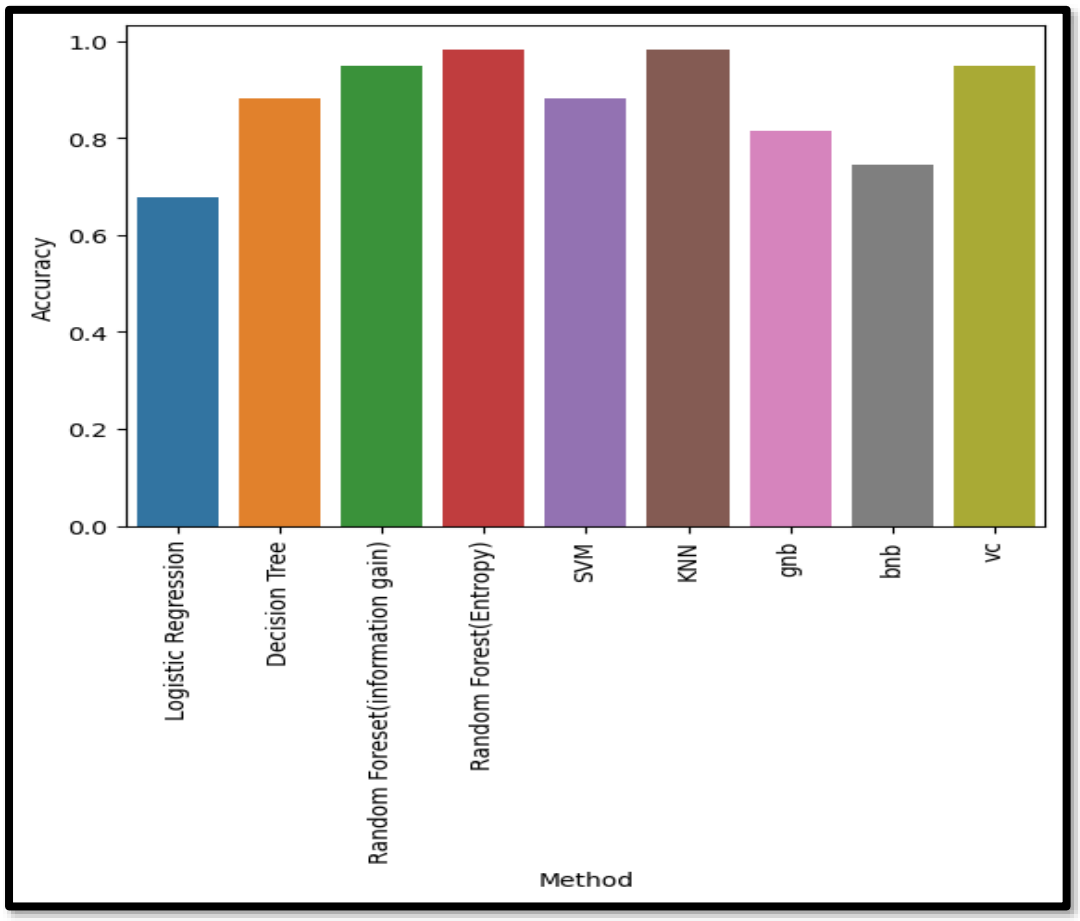
6.2 ANALYSIS

Analysis XGBoost Algorithm

	Method	Accuracy
0	Logistic Regression	0.677966
1	Decision Tree	0.881356
2	Random Foreset(information gain)	0.949153
3	Random Forest(Entropy)	0.983051
4	SVM	0.881356
5	KNN	0.983051
6	gnb	0.813559
7	bnb	0.745763
8	vc	0.949153

Axes(0.125,0.11;0.775x0.77)

MODEL PERFORMANCE FOR EACH MODEL



ADVANTAGES AND DISADVANTAGES

ADVANTAGES:

- Can be accessed offline
- User friendly
- Privacy isn't affected as it's hosted on local server
- Can handle large dataset
- Fast execution
- Accurate results

DISADVANTAGES:

- Non availability of more data
- High cost of production (specific voice features needed for training)
- Data inaccuracy can result in wrong result
- There are innumerable threats that might influence the development of our program. Few of hem are:
  - Having an uneven proportion of healthy and unhealthy person's data could introduce bias in the learning algorithm.
  - Incorrect classification of training dataset could cause hassle within the prediction

APPLICATIONS

- Government initiatives for ML and AI in healthcare
- Health Sector
- Innovation

CONCLUSION

We can conclude that out of three models, xgboost is performing better with higher specificity and accuracy than the other two. Although random forest is having good accuracy and auc but the specificity is low. This means the model is not able to predict the minority classes well which means that this will result in wrongly predicting the cases.

FUTURE WORK

Future Work: The accuracy and performance of the model can be further improved by using a larger volume of data for training. The website can be further extended to be hosted online so that the users don't have to install the software requirements on their system.