A Machine Learning Approach for the Diagnosis of Parkinson's Disease via Speech Analysis

Review III for

CSE4020

MACHINE LEARNING

By

Naman Dutta (17BCE2379)

Under Guidence of

Prof. Jaisakthi S M

School of Computer Science Engineering

Vellore Institute of Technology, Vellore – 63014

Review Video Link:

https://drive.google.com/open?id=1wVZFff7cea_RzbnnxL-D7iGyd2xtoWV1

Abstract:

This paper proposes a methodology to detect early signs of Parkinson's disease (PD) through free-speech in uncontrolled background conditions. The early detection mechanism uses signal and speech processing techniques integrated with machine learning algorithms. There is no single test which can be administered for diagnosis of Parkinson's Disease. Instead, doctors must perform a careful clinical analysis of the patient's medical history. Unfortunately, this method of diagnosis is highly inaccurate. Because of these difficulties, I investigate a machine learning approach to accurately diagnose Parkinson's, using a dataset of various speech features. Why speech features? Speech is very predictive and characteristic of Parkinson's disease; almost every Parkinson's patient experiences severe vocal degradation.

Motivation:

Parkinson disease is the most common form of parkinsonism, a group of neurological disorders with Parkinson disease–like movement problems such as rigidity,

slowness, and tremor. More than 6 million individuals worldwide have Parkinson disease.

Objective:

To compare the efficiency of different Machine Learning algorithms based on their accuracy.

Scope:

Diagnosis of Parkinson's Disease.

Introduction

Mental disorders, such as neurodegenerative, psychiatric and advancement diseases have human, social, and money related impacts, both at personal, professional, and social levels. This illness reduces average life expectancy and deteriorates its satisfaction. Contributions to this zone are especially significant, considering their pertinence for the everyday nature of individuals and their families. The costs of mental disorders are higher than those of the mortality figures.

The estimates of the World Health Organization (WHO), known as Disability Adjusted Life Years (DALYs) are of utmost pertinence (World Health Organization, 2004). There is no medication found for Parkinson's Disease even though studies suggest that particular types of treatments are compelling in improving the personal satisfaction of patients when applied at a beginning time, therefore diminishing the aggregate estimated costs of the pathology. Despite this reality, most methods used to identify signs of Parkinson's Disease require a motor manifestation of the symptoms, which just happen at an advanced stage of the disease.

Parkinson’s Disease is the second most prevalent neurodegenerative disorder after Alzheimer’s, affecting more than 10 million people worldwide. Parkinson’s is characterized primarily by the deterioration of motor and cognitive ability. There is no single test which can be administered for diagnosis. Instead, doctors must perform a careful clinical analysis of the patient’s medical history. Unfortunately, this method of diagnosis is highly inaccurate. A study from the National Institute of Neurological Disorders finds that early diagnosis (having symptoms for 5 years or less) is only 53% accurate. This is not much better than random guessing, but an early diagnosis is critical to effective treatment. Because of these difficulties, I investigate a machine learning approach to accurately diagnose Parkinson’s, using a dataset of various speech features (a non-invasive yet characteristic tool) from the University of Oxford.

Why speech features? Speech is very predictive and characteristic of Parkinson’s disease; almost every Parkinson’s patient experiences severe vocal degradation (inability to produce sustained phonations, tremor, hoarseness), so it makes sense to use voice to diagnose the disease. Voice analysis gives the added benefit of being non-invasive, inexpensive, and very easy to extract clinically. Studies give that indications of PD can be detected through speech. [1]Tsanas et al. (2011) presented a methodology that enabled the establishment of the ten best features to distinguish patients with indications of PD. The best accuracy was 98.6%, and, on average, it was marginally lower, with a value of 91.4%, meaning that the algorithm created by [1]Tsanas et al. (2011) was able to distinguish indications of PD in 457 out of 500 patients. For the experiment, the examination utilized a database of 263 samples from 43 individuals to register more than 132 dysphonia measures from sustained vowels, using RF and SVM to construct the statistical classifier.

Literature Survey

S.N.	Paper	Author	Published in	Description
[1]	**Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease**	Athanasios Tsanas ; Max A. Little ; Patrick E. McSharry ; Jennifer Spielman ; Lorraine O. Ramig	09, January 2012	This paper test how precisely these novel algorithms can be utilized to segregate PD subjects from sound controls. Altogether, we register 132 dysphonia measures from supported vowels. At that point, they select four miserly subsets of these dysphonia estimates utilizing four-component determination algorithms, and guide these element subsets to a double classification reaction using two accurate classifiers: arbitrary woodlands and bolster vector machines. We utilize a current database comprising of 263 samples from 43 subjects and exhibit that these new dysphonia measures can beat cutting edge results, arriving at practically 99% by and high classification accuracy utilizing just ten dysphonia highlights.
[2]	**Diagnosis of Parkinson's disease in human using voice signals**	Hamid Karimi Rouzbahani, Mohammad Reza Daliri	1st December 2011	This paper depicted a system to analyze PD in people utilizing voice signals. The proposed approach uses the Praat programming **[3]** to separate highlights from the signing voice. These highlights depend on parameters, for example, the significant recurrence, jitter, sparkle, pitch, HNR, and measures in light of these parameters (for instance, nonlinear proportions of essential recurrence variety). The creators chose the ideal highlights for the classification task by utilizing connection rates, Fisher's Discriminant Ratio, t-test, and ROC bend. The number of excellent highlights was resolved using an element execution bend with a SVM classifier (RBF bit). For the classification procedure, the accompanying ideal highlights were received: average vocal principal recurrence, most excellent vocal essential recurrence, detrended vacillation examination, spread1, spread2 (both nonlinear proportions of central recurrence variety), D2 (nonlinear dynamical multifaceted nature measure) and pitch period entropy. SVM, K-Nearest Neighbor (KNN), and Segregation Function-Based classifiers were tried for their accuracy, blunder rate, affectability, and particularity. The

				examination finished up the best execution (93.82% accuracy) was gotten utilizing the KNN classifier.
[4]	**Characterizing Parkinson's Disease Speech by Acoustic and Phonetic Features**	Jorge Proença, Arlindo Veiga, Sara Candeias, João Lemos	October 2014	Proença et al. (2014) indicated a procedure to recognize acoustic, what's more, phonetic qualities present in the discourse of Portuguese PD patients, utilizing a database with subjects in the early or late phases of the PD. Concerning phonetic attributes, the proposed system depicts the extraction and investigation of highlights, for example, vowel formants and formant measurements, contrasting the first and second formant frequencies of patients of the two gatherings, including the Vowel Space Territory and Vowel Articulatory Index. The paper reasoned that the beginning also, second formant frequencies showed contrasts in focal vowels for patients at later phases of the PD, just as that of the typical formant measurements, only the Vowel Articulatory Index demonstrated to have factual centrality in the correlation of control guys versus guys at later phases of PD. For the acoustic investigation, the methodology depends on the Mel Frequency Cepstral Coefficients and its inferred highlights, such as the Gaussian Mixture Model and openSMILE toolbox attributes, the last-named by the creators "Smile highlights". Measurable outcomes reasoned that probably the most noteworthy highlights for PD discourse separation are made out of unique highlights. Besides, the results exhibited that the discovery of the PD issue can be accomplished through a few acoustic highlights.
[5]	**Early detection of Parkinson's disease through Voice**	Vikas, Sharma, R. K.	2014	Considering the voice parameter investigation, Vikas and Sharma (2014) investigated the contrasts between voice parameters in PD patients, furthermore, solid subjects, utilizing Praat as the product for removing highlights from the voice signal. The creators thought about Mel-Frequency Cepstral Coefficients, pitch, jitter, and shine esteems, just as the subject's glottal heartbeat, between 14 patients and seven healthy subjects between with ages somewhere in the range of 45 and 85. It was discovered that, as to pitch, male patients would, in general, have higher qualities. For Mel-Frequency Cepstral Coefficients and the glottal heartbeat, these

				voice parameters will, in general, have greater varieties and a less comparative example, when contrasted and solid subjects. At last, to jitter and gleam, patients score higher values when contrasted with solid subjects.
[6]	**Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine**	Shahbakhi, M., Far, D.T., Tahami,	2014	Shahbakhi et al. (2014) showed the discovery of PD through voice, utilizing the Genetic Algorithm, and SVM. The creators used an information set made by Max Little, containing 195 supported vowels from 31 male and female subjects, of which 23 are influenced by PD. From the informational collection, 22 direct and non-straight highlights were extricated, of which 14 depend on essential recurrence (pitch), jitter, gleam, and Clamor to-Harmonics proportion (NHR). The ideal removed highlights were chosen utilizing the Genetic Algorithm executed in MATLAB. During the classification procedure, the SVM calculation was applied to the informational collection using the typically separated highlights. They had the option to accomplish the best classification results (94.50%) utilizing four typical highlights: Maximum vocal essential recurrence, most high vocal principal recurrence, jitter (RAP), and sparkle (APQ5).
[7]	**Parkinson-speech analysis: Methods and aims**	Baasch, C., Schmidt, G., Heute, U., Nebel, A., Deuschl, G.,	2016.	All the more as of late, Baasch et al. (2016) portrayed the advancement of an apparatus for acoustical examination of PD discourse, the classification of the seriousness of the discourse disability, to improve achievement paces of language courses, just as streamlining the procedure of classification. This approach comprises of setting up recording meetings, with the preferred position of having a characteristic account condition that recreates a whiz discussion. In a subsequent stage, discourse issues were broke down utilizing attributes, for example, pitch, signal level, and vowel-space territory, all together to order the seriousness of the discourse issue using discourse quality scales, for instance, NTID.
[8]	**Detecting Parkinson's disease from sustained phonation**	Vaiciukynas, E., Verikas, A., Gelzinis, A., Bacauskiene, M.,	2017	Some distributed work utilizing RF to distinguish PD can be found. Vaiciukynas et al. (2017) proposed a system to identify PD from supported phonations using RF. The work comprised of removing events of the phonation of the/a/vowel from spoken Lithuanian

	and speech signals.			short sentences, which were recorded utilizing a cell phone (SP) and an acoustic cardioid (Air conditioning) amplifier. The RF calculation ordered information using 18 highlights, which depended on singular capabilities and choice level combination sets. The calculation's presentation was allotted utilizing the Of-Bag mistake, in which the best qualities acquired using unique skills, for the SP and AC recorded information, were 25.57% and 20.30% accuracy, individually. For the choice level combination sets, the best execution values for SP and AC recorded information was 23.00% and 19.27% accuracy, personally.
[9]	**Classificatio n of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples.**	Zhang, H.-H., Yang, L., Liu, Y., Wang, P., Yin, J., Li, Y., Yan, F.,	2016	Zhang et al. (2016) investigated the chance of consolidating the Multi-Edit Nearest-Neighbor (MENN) and the RF methods when for the classification of PD. In the proposed strategy, the creators first executed the MENN to choose the ideal preparing discourse samples and, at that point, executed the RF to create the classification model. These creators had the option to infer that the blend of the MENN with the RF calculation improved the classification's accuracy.
[10]	**Extração de Conheciment o de Dados, second ed. Edições Sí, .**	Gama, J., Ponce de Leon Carvalho, A., Oliveira, M., Lorena, A.C., Faceli, K.,	2015	The processing power of the NN has put away in the associations between these units in the structure of loads (Gama et al., 2015). The estimations of the loads are inferred from the way toward gaining from a given informational collection. The NN can be made out of multiple layers, and each layer can have n number of neurons (units), where the absolute input esteem u is acquired by a right blend of the jth neuron's input xj with weight wj. The yield of every neuron relates to the response of the complete input esteem that is usually produced by capacity, for example, a sigmoid or hyperbolic tangent, among others.

[11]	**Diagnosing Parkinson by using artificial neural networks and support vector machines.**	Gil, D., Magnus, B.J.,	2009	A case of use of NN in the subject tended to in this paper can be found in work done by Gil and Magnus (2009). The creators removed the absolute most well-known highlights from 195 continued vowels from 31 patients, for example, jitter, gleam, DFA, and pitch-based highlights. The most remarkable accuracy got with the proposed strategy was 93.33%.
[12]	**Automatic Recognition of Parkinson Disease from Sustained Phonation Tests Using ANN and Adaptive Neuro-Fuzzy Classifier, Vol. 1.**	Caglar, M., Cetisli, B., Toprak, I.,	2009	Caglar et al. (2009) utilized an MLP, a Radial Basis Function Network, and an Adaptive Neuro-Fuzzy Classifier to fabricate a classification model for programmed acknowledgment of PD, from a biomedical voice informational collection. The cross-validation procedure was the half holdout strategy; that is, the informational index was partitioned in two equivalent parts — one section for testing and the other for preparing the model. The best accuracy results got were about 94.72% for the testing samples and 95.38% for the prepared samples.

Proposed Methodology

Dataset Description

Source:

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.

Description:

Data Set Characteristics: Multivariate

Number of Instances: 197

Area: Life

Attribute Characteristics: Real

Number of Attributes: 23

Date Donated: 2008-06-26

Associated Tasks: Classification

Data Set 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 one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD. The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.

Performance Metrics

TP = true positive, FP = false positive, TN = true negative, FN = false negative

Accuracy: (TP+TN)/(P+N)

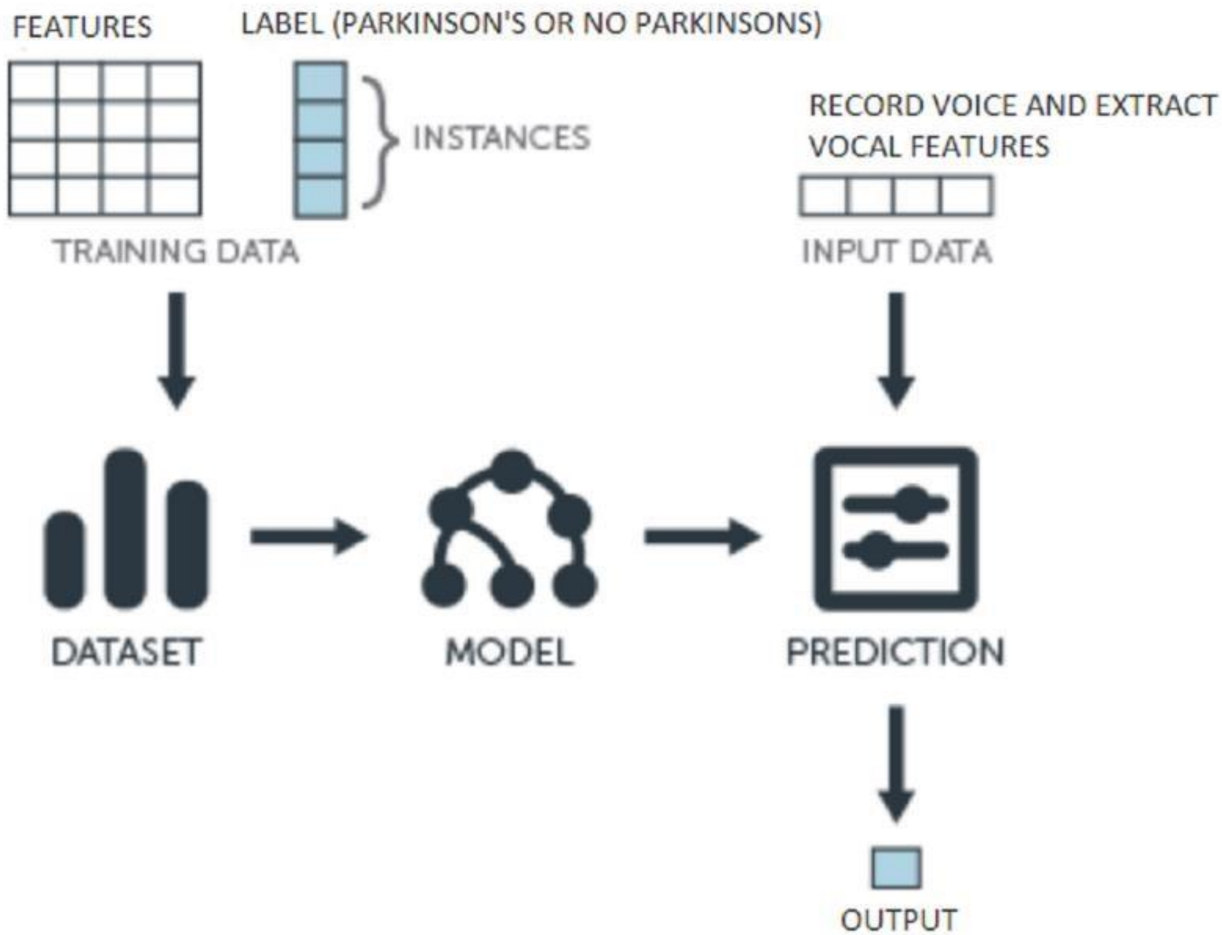


Figure 1: Working Model of Proposed algorithm

Proposed Method

1. <u>KNN – K-Nearest Neighbours</u>

   KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

Architecture Diagram

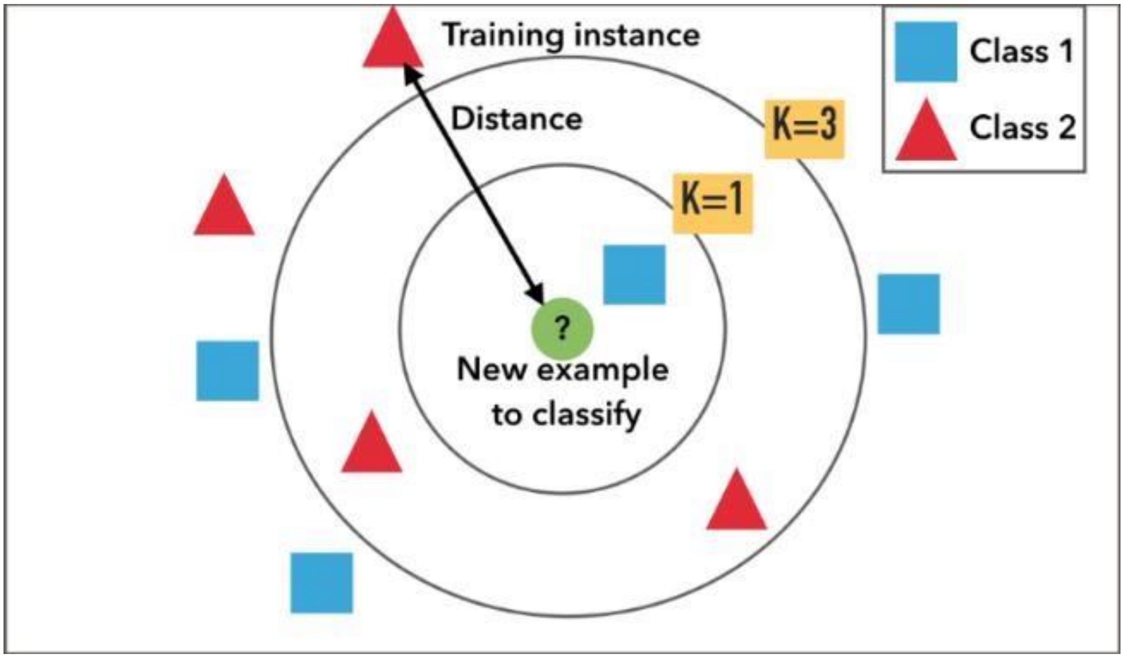


Figure 2 Architecture Diagram of K-Nearest Neighbour

PSEUDOCODE OF KNN (Algorithm)

- Load the data
- Initialise the value of k
- For getting the predicted class, iterate from 1 to total number of training data points.
- Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
- Sort the calculated distances in ascending order based on distance values
- Get top k rows from the sorted array
- Get the most frequent class of these rows
- Return the predicted class

2. Decision Tree Classifier

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

Architecture Diagram

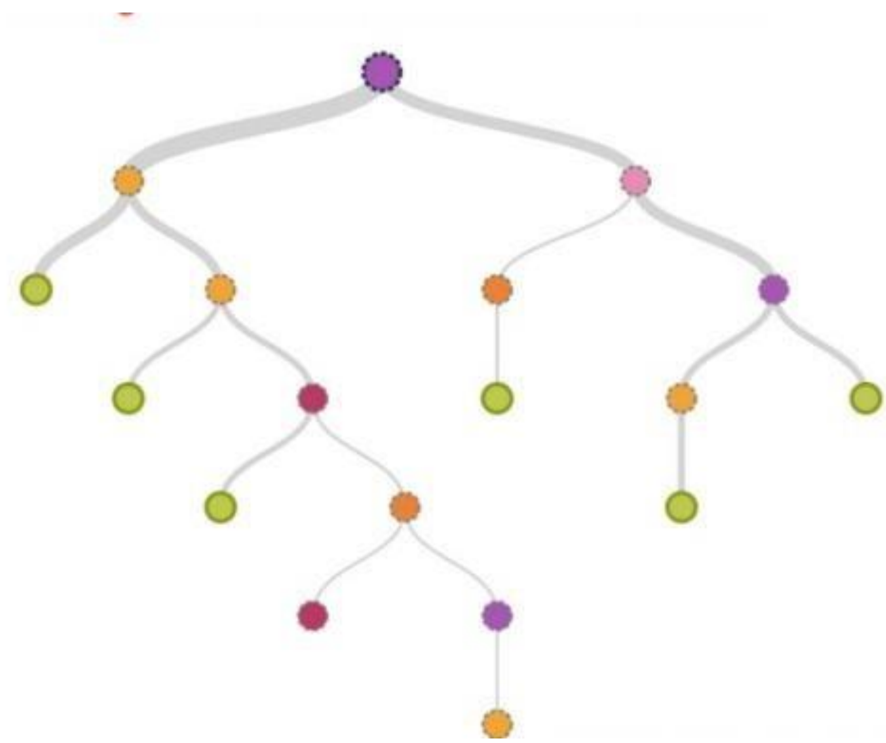
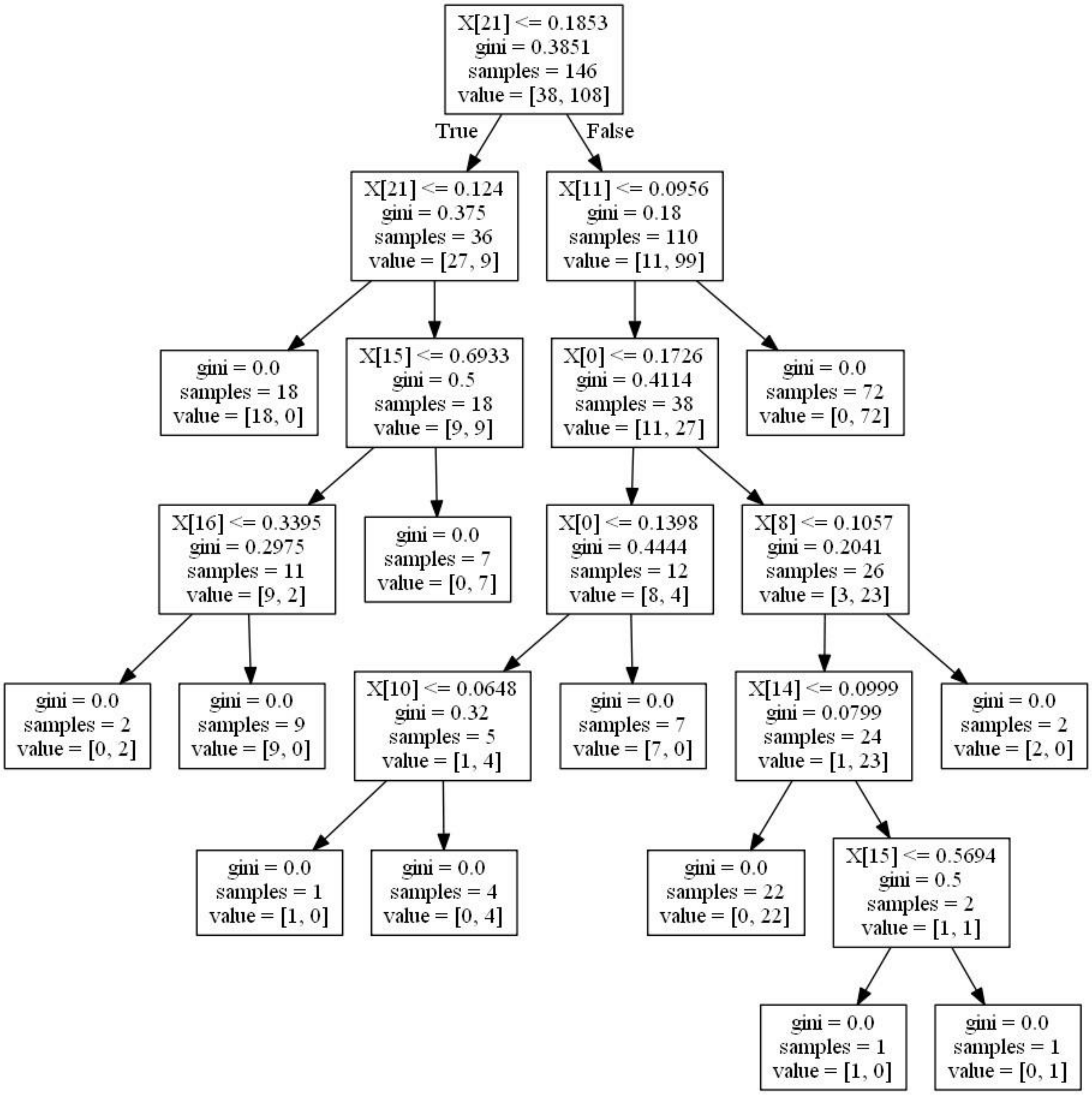


Figure 3: Architecture Diagram of Decision Tree

Modular Diagram:



PSEUDOCODE OF DECISION TREE CLASSIFIER (Algorithm)

- Place the best attribute of the dataset at the root of the tree.
- Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
- Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

3. Naïve Bayes Classifier

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

PSEUDOCODE OF NAÏVE BAYES CLASSIFIER (Algorithm)

- Read the training data set.
- Calculate the mean and standard deviation of the predictor variables in each class.
- Repeat and calculate the probability of f_i using the gauss density equation in each class.
- Until the probability of all predictor variable (f_1,f_2,f_3, …. F_n) has been calculated.
- Calculate the likelihood for each case.
- Get the greatest likelihood.

Experimental Setup

Application used**: Python 3.7.4**

Libraries used:

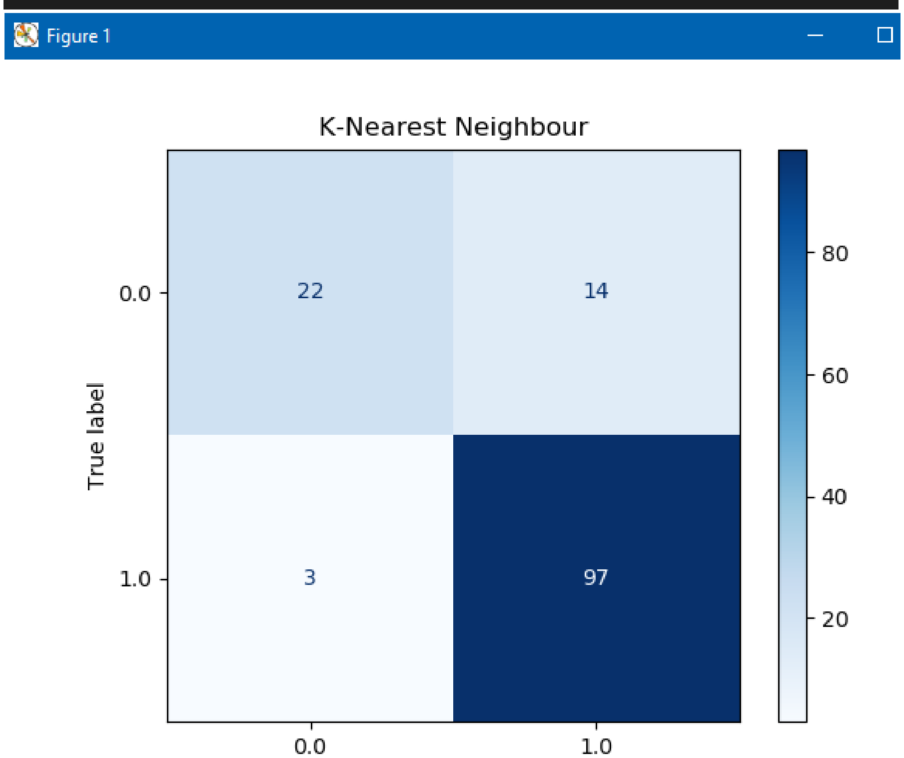
- import pandas
    - For reading the data from the dataset file which is in the format of .csv
- from sklearn.metrics import confusion_matrix
- from sklearn.metrics import plot_confusion_matrix
    - This is used for calculating the confusion matrix and plot them
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.naive_bayes import GaussianNB
    - *models = []*
    - *models.append(('KNN', KNeighborsClassifier()))*
    - *models.append(('DTC', DecisionTreeClassifier()))*
    - *models.append(('NB', GaussianNB()))*
    - *predictions = model.predict(X_validation)*
    - All the above are used for prediction of the output

- from sklearn.metrics import accuracy_score

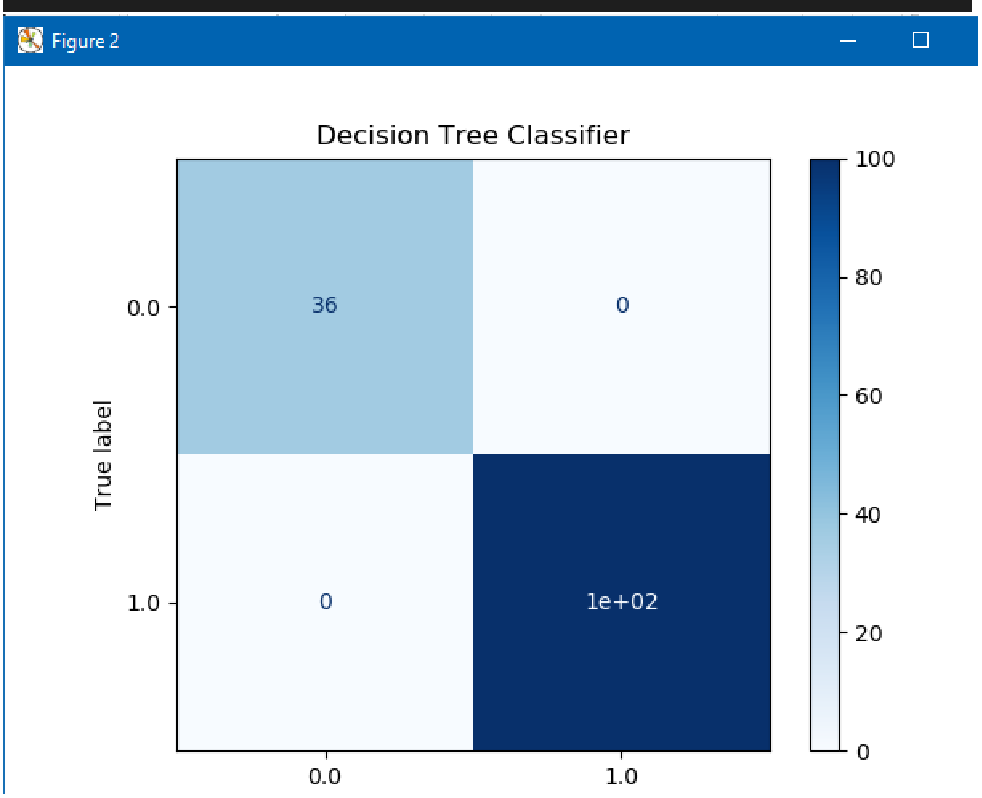
    - This is used for calculating the accuracy of the algorithm model.

Result and Discussion

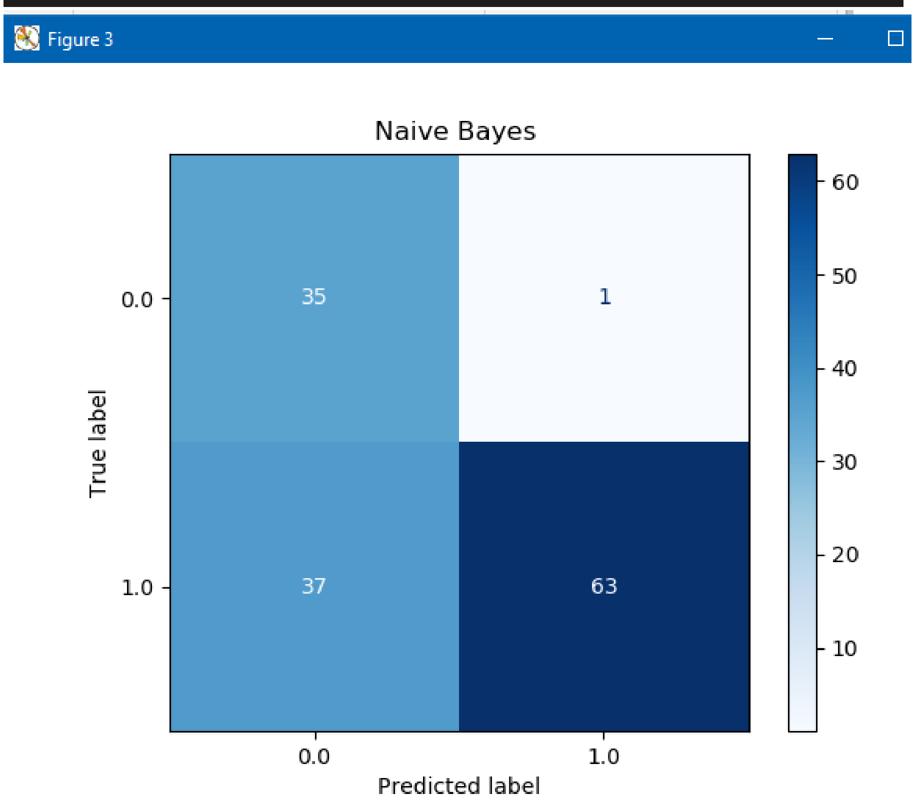
```
Scores for each algorithm:


Name                    |         Accuracy
-----------------       |       ------------
K-Nearest Neighbour     |        86.4406779661017



              precision    recall  f1-score   support

         0.0       0.83      0.42      0.56        12
         1.0       0.87      0.98      0.92        47

    accuracy                           0.86        59
   macro avg       0.85      0.70      0.74        59
weighted avg       0.86      0.86      0.85        59



Confusion Matrix:
 [[ 5  7]
 [ 1 46]]
Normalization Matrix K-Nearest Neighbour

[[22 14]
 [ 3 97]]
```



```
Name                       |          Accuracy
------------------         |        ------------
Decision Tree Classifier         |         86.4406779661017



              precision    recall  f1-score   support

         0.0       0.61      0.92      0.73        12
         1.0       0.98      0.85      0.91        47

    accuracy                           0.86        59
   macro avg       0.79      0.88      0.82        59
weighted avg       0.90      0.86      0.87        59



Confusion Matrix:
 [[11  1]
 [ 7 40]]
Normalization Matrix Decision Tree Classifier

[[ 36   0]
 [  0 100]]
```



```
Name                        |          Accuracy
-----------------           |        ------------
Naive Bayes      |          71.1864406779661



              precision    recall  f1-score   support

         0.0       0.40      0.83      0.54        12
         1.0       0.94      0.68      0.79        47

    accuracy                           0.71        59
   macro avg       0.67      0.76      0.67        59
weighted avg       0.83      0.71      0.74        59



Confusion Matrix:
 [[10  2]
 [15 32]]
Normalization Matrix Naive Bayes

[[35  1]
 [37 63]]
```



**the 0.0 indicates people without Parkinson's disease and 1.0 indicates people with Parkinson's disease*

- The accuracy obtained by using K-Nearest Neighbour (KNN) algorithm is 86.44%.
  F1-score of K-Nearest Neighbour has best score of 0.92 and the worst score of 0.59.
- The accuracy obtained by using Decision Tree Classifier (DTC) algorithm is 89.83%.
  F1-score of Decision Tree Classifier has best score of 0.92 and the worst score of 0.79.
- The accuracy obtained by using Naïve Bayes Classifier (NB) algorithm is 71.18%.
  F1-score of NAÏVE BAYES has best score of 0.79 and the worst score of 0.54.

Conclusion

Ongoing advancements in procedures to distinguish PD through speech analysis have demonstrated promising outcomes, filling in as proof for the developing capability of the utilization of this innovation in the biomedical field, intending to help individuals' lives.

This project demonstrates the use and comparison of K- Nearest Neighbors, Decision Tree Classifier and Naïve Bayes Classifier algorithms and their performance based on their accuracy. The output obtained the accuracy of each of the algorithms was calculated and compared with each other, on comparing them we have come to the conclusion that the KNN algorithm works best with an accuracy of 86.44% compared to the other algorithms.

References

[1] Tsanas, A., Little, M.A., McSharry, P.E., Spielman, J., Ramig, L.O., 2011. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's. Retrieved from http://ieeexplore.ieee.org/iel5/10/4359967/06126094.pdf.

[2] Rouzbahani, H.K., Daliri, M.R., 2011. Diagnosis of Parkinson's disease in human using voice signals. Basic Clin. Neurosci. 2 (3), 12.

[3] Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM, pp. 144–152.

[4] Proença, J., Perdigão, F., Veira, A., Candeias, S., Lemos, J., Januário, C., 2014. Characterizing Parkinson's disease speech by acoustic and phonetic features. In: Computational Processing of the Portuguese Language, Vol. 8775, pp. 24–35. Retrieved from https://www.researchgate.net/publication/312762862_Characterizing_Parkinson's_Disease_Speech_by_Acoustic_and_Phonetic_Features.

[5] Vikas, Sharma, R. K., 2014. Early detection of Parkinson's disease through Voice. In: 2014 International Conference on Advances in Engineering and Technology (ICAET), pp. 1–5. http://dx.doi.org/10.1109/ICAET.2014.7105237.

[6] Shahbakhi, M., Far, D.T., Tahami, E., 2014. Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine. J. Biomed. Sci. Eng. 07 (04), 147–156. http://dx.doi.org/10.4236/jbise.2014.74019.

[7] Baasch, C., Schmidt, G., Heute, U., Nebel, A., Deuschl, G., 2016. Parkinson-speech analysis: Methods and aims. In: Speech Communication; 12. ITG Symposium, pp. 1–5.

[8] Vaiciukynas, E., Verikas, A., Gelzinis, A., Bacauskiene, M., 2017. Detecting Parkinson's disease from sustained phonation and speech signals. PLoS One 12 (10), http://dx.doi.org/10.1371/journal.pone.0185613.

[9] Zhang, H.-H., Yang, L., Liu, Y., Wang, P., Yin, J., Li, Y., Yan, F., 2016. Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples. BioMed. Eng. OnLine 15, http://dx.doi.org/10.1186/s12938-016-0242-6.

[10] Gama, J., Ponce de Leon Carvalho, A., Oliveira, M., Lorena, A.C., Faceli, K., 2015. Extração de Conhecimento de Dados, second ed. Edições Sí, Retrieved from http://www.silabo.pt/Conteudos/8117_PDF.pdf.

[11] Gil, D., Magnus, B.J., 2009. Diagnosing Parkinson by using artificial neural networks and support vector machines. Global J. Comput. Sci. Technol. 9 (4).

[12] Caglar, M., Cetisli, B., Toprak, I., 2009. Automatic Recognition of Parkinson Disease from Sustained Phonation Tests Using ANN and Adaptive Neuro-Fuzzy Classifier, Vol. 1.

VIT Vellore Institute of Technology (Deemed to be University under section 3 of UGC Act, 1956)

Appendix – Your Code here

```
import pandas
import numpy as np

from pandas.plotting import scatter_matrix
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import RepeatedKFold
from sklearn.metrics import classification_report
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
from matplotlib import pyplot as plt

url = "C:/Users/User/Downloads/edu6\Machine learning/Parkinson's Disease/data.csv
"

features = ["MDVP:Fo(Hz)", "MDVP:Fhi(Hz)", "MDVP:Flo(Hz)", "MDVP:Jitter(%)", "MDV
P:Jitter(Abs)", "MDVP:RAP", "MDVP:PPQ", "Jitter:DDP", "MDVP:Shimmer",
            "MDVP:Shimmer(dB)", "Shimmer:APQ3", "Shimmer:APQ5", "MDVP:APQ", "Shim
mer:DDA", "NHR", "HNR", "RPDE", "DFA", "spread1", "spread2", "D2", "PPE", "status
"]
dataset = pandas.read_csv(url, names=features)

array = dataset.values

X = array[:, 0:22]

Y = array[:, 22]


validation_size = 0.3

seed = 7
X_train, X_validation, Y_train, Y_validation = train_test_split(
    X, Y, test_size=validation_size, random_state=seed)

num_folds = 10
num_instances = len(X_train)
seed = 7
```

```
scoring = 'accuracy'
models = []
models.append(('K-Nearest Neighbour', KNeighborsClassifier()))
models.append(('Decision Tree Classifier', DecisionTreeClassifier()))
models.append(('Naive Bayes ', GaussianNB()))

names = []
score = [5]
i = 0
print("Scores for each algorithm:\n\n")
for name, model in models:
    kfold = RepeatedKFold(n_splits=num_instances,
                          n_repeats=num_folds, random_state=seed)
    cv_results = cross_val_score(
        model, X_train, Y_train, cv=kfold, scoring=scoring)
    names.append(name)

    cls = model.fit(X_train, Y_train)
    print("Name\t\t\t|\t  Accuracy\n-----------------\t|\t-----------")
    predictions = model.predict(X_validation)
    print(name, "\t|\t", accuracy_score(Y_validation, predictions)*100)
    print("\n\n")
    print(classification_report(Y_validation, predictions))
    print("\n")
    matrix = confusion_matrix(Y_validation, predictions)
    print("Confusion Matrix:\n", matrix)
    disp = plot_confusion_matrix(cls, X_train, Y_train,
                                 cmap=plt.cm.Blues
                                 )
    disp.ax_.set_title(name)
    print("Normalization Matrix", name, "\n")
    print(disp.confusion_matrix)
    print("\n\n\n")

plt.show()
```