**PROJECT REPORT ON**

**“PREDICTION OF**

**PARKINSON’S DISEASE PROGRESSION”**

**A report submitted in partial fulfillment of the requirements for**

### INTEGRATED PROFESSIONAL CORE COURSE-

### FUNDAMENTALS OF MACHINE LEARNING (AD2001-1)

**Submitted by**

**Name: Anagha Tantry USN: NNM22AD007**

**Name: Anusha Nayak USN: NNM22AD012**

**Name: Nidhi Shetty USN: NNM22AD036**

**Name: Swaroop S USN: NNM22AD058**

**Under the Guidance of**

**Dr. Ankitha K**

**Assistant Professor Gd-III**

### DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

A blue text on a black background

Description automatically generated

A blue text on a black background

Description automatically generated

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE**

**CERTIFICATE**

This is to certify that the Integrated Professional Core Course project entitled “Prediction of Parkinson’s Disease Progression” report has been submitted by Anagha Tantry (NNM22AD007), Anusha Nayak (NNM22AD012), Nidhi Shetty (NNM22AD036) and Swaroop S (NNM22AD058) of II year B. Tech, bonafide students of NMAM Institute of Technology, Nitte, has completed the project work during the academic year 2023-2024 fulfilling the partial requirements of Integrated Professional Core Course lab in Artificial Intelligence & Data Science at NMAM Institute of Technology, Nitte.

**Project Guide Course Coordinator HOD**

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| Chapter No | Contents | Page No. |
|  | Abstract | 4 |
| 1. | Introduction | 5-6 |
| 2. | Literature Survey | 7-9 |
| 3. | Problem Statement | 10 |
| 4. | Dataset description | 11-12 |
| 5. | Methodology | 12-13 |
| 6. | Implementation | 13-18 |
| 7. | Results | 18-22 |
| 8. | Conclusion | 23 |
| 9. | Reference | 24 |

Abstract

Parkinson’s disease (PD) is a disabling brain disorder that affects movements, cognition, sleep, and other normal functions. Unfortunately, there is no current cure and the disease worsens over time. Early prediction of its progression can significantly help in providing required treatments. Traditional diagnosis relies on clinical assessment, which can be challenging, especially in the early stages. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and tool for diagnosis. This project aims to develop a machine learning model for the prediction of Parkinson's Disease using voice recordings, providing a non-invasive and efficient screening tool. The prediction of Parkinson's Disease (PD) using machine learning methods involves applying algorithms to analyze data, particularly voice recordings, for accurate diagnosis. This approach aims to enhance the diagnostic process of PD, which is challenging due to the lack of definitive laboratory tests for early detection. By utilizing machine learning algorithms on voice signal features, researchers can develop predictive models that assist in identifying PD based on characteristic vocal features.

**Chapter 1**

Introduction

1.1 Description of the problem being addressed:

Parkinson’s Disease (PD) is a degenerative neurological disorder marked by decreased dopamine levels in the brain. It manifests itself through a deterioration of movement, including the presence of tremors and stiffness. There is commonly a marked effect on speech, including dysarthria (difficulty articulating sounds), hypophonia (lowered volume), and monotone (reduced pitch range). Additionally, cognitive impairments and changes in mood can occur, and risk of dementia is increased. Traditional diagnosis of Parkinson’s Disease involves a clinician taking a neurological history of the patient and observing motor skills in various situations. Since there is no definitive laboratory test to diagnose PD, diagnosis is often difficult, particularly in the early stages when motor effects are not yet severe. Monitoring progression of the disease over time requires repeated clinic visits by the patient. An effective screening process, particularly one that doesn’t require a clinic visit, would be beneficial. Since PD patients exhibit characteristic vocal features, voice recordings are a useful and non-invasive tool for diagnosis. If machine learning algorithms could be applied to a voice recording dataset to accurately diagnosis PD, this would be an effective screening step prior to an appointment with a clinician.

1.2 Objectives of the project:

1. Survey the existing models and recreate the dataset: Review existing literature and models related to the prediction of Parkinson's Disease progression using machine learning. Collect the suitable dataset ensuring it adequately represents the characteristics of Parkinson's Disease progression, including voice recordings and clinical data.
2. Develop a model to predict Parkinson's Disease progression: Utilize the collected dataset to train machine learning models for predicting the progression of Parkinson's Disease. Experiment with various feature engineering techniques to extract relevant features from voice recordings and clinical data. Implement and fine-tune machine learning algorithms such as decision trees, support vector machines, or k nearest neighbour to develop an accurate prediction model.
3. Compare the proposed model with existing models and suggest an accurate model: Evaluate the performance of the developed model using appropriate metrics such as accuracy, precision, recall, and F1-score. Compare the performance of each proposed model with existing models identified during the survey phase. Identify strengths and weaknesses of each model and provide recommendations for an accurate prediction model for Parkinson's Disease progression.

1.3 Importance of the project in the context of machine learning:

The project holds significant importance in the field of machine learning as it showcases the application of advanced computational techniques in healthcare. By utilizing machine learning algorithms, specifically tailored for analyzing voice data, it aims to address a critical need for early and accurate diagnosis of Parkinson's Disease. This not only demonstrates the potential of machine learning in improving medical diagnostics but also highlights the importance of interdisciplinary approaches in tackling complex healthcare challenges. A successful model would showcase the potential of machine learning for early disease detection using readily available data like voice recordings. This approach could revolutionize PD screening by offering a convenient and cost-effective method.

**Chapter 2**

Literature Survey

2.1 Introduction:

Parkinson's disease is a neurological disorder effect and is a second most common neurodegenerative disorder after Alzheimer's [2]. Parkinson disease (PD), a neurodegenerative disorder of CNS system, also produces movement disorder PD has no cure but several treatments are known to provide relief from the symptoms. Typically, PD occurs in people over the age of 60, among which 1% of people are affected. It is called as young onset PD when it is seen in the people before age 50 [3]. According to estimates, PD affected 6.2 million people and about 117400 deaths globally in 2015.Early prediction of motor symptoms of PD can get a proper treatment at right time to a patient.

2.2 Related Work:

For analysis purposes, voice dataset of PD is collected from UCI machine learning library. By implementing multiple predictive models to disease datasets, multiple accuracies and results of different classifiers are acquired. C4.5, SVM, and ANN give better results than other machine learning algorithms. After comparing the results of these classifiers, best results are chosen for the final decision explored a predictive model for PD that is based on decision tree algorithm. They introduced PD, a second most common neurodegenerative disease with its symptoms, and risk factors associated with it. Various applications of data mining are used for classification purposes that are decision tree, attribute selection measures, ID3 and decision stumps. Their dataset (have 197 instances) is taken from UCI repository and built up from the data of 31 people. For performance analysis, two parameters accuracy and classification error are used. For validation, 10-fold cross-validation technique is used that gives the unbiased outcome. They found that decision tree algorithm performs best and gives the best accuracy and less classification error than other algorithms in their experimental results. Sriram et al. [4] proposed a method for diagnosis of PD using its voice dataset. This voice dataset is built up from the voice of 31 people among which 23 people are affected by PD. This dataset contains 5875 instances and 26 attributes. In their experiment for statistical analysis, classification, evaluation, visualization, and unsupervised methods Weka V3.4.10 and Orange V2.0b software are used. They achieved the best accuracy 90.2% from Random Forest algorithm.

2.3 Problem Definition:

The huge amount of data known as Big Data is generated everywhere, and this data can be used to perform analysis and make future predictions. According to some research, most of the healthcare data is in the unstructured form that can be stored in the centralized repository to make useful interpretation out of it. To improve the quality of patient care at low cost, this unstructured data can be analyzed further by merging it with structured datasets. The problem is to classify and discover data pattern to predict future disease, so that doctors can detect and diagnose the disease at an early stage.

2.4 Proposed model:

We observed that Waikato Environment for Knowledge Analysis (WEKA) to implement data mining algorithms for preprocessing, classification, clustering, and analysis of results. This environment includes java libraries that implement algorithms and provide the best environment to researchers for classifying datasets

2.5 Data Collection:

The dataset used in this paper is taken from UCI machine learning library [5]. The dataset consists of 195 instances and 24 attributes. This feature set consists of name, Fo(Hz), Fhi(Hz), Flo(Hz), Jitter(%), Jitter(Abs), RAP, PPQ, Jitter: DDP, Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, APQ, Shimmer: DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, and PPE. We store these datasets in .CSV format and then convert it into. ARFF format for further analysis. The 360 R. Mathur et al. dataset is divided into two classes according to its “status” column which is set to 0 for healthy subjects and 1 for those PD.

2.6 Data preprocessing:

The poor data quality in the medical dataset is one of the big challenges that are faced by the knowledge discovery process. This process decreases the number of attributes into a better subset which can increase accuracy, and also it brings a reduction in training time. It is done using Filters and Wrappers. WEKA provides “Attribute Selection” filter to choose an attribute evaluation method. We use “cfs Subset Eval” attribute evaluator and “Best First Search” method which considers the individual predictive ability of each feature to evaluate the worth of an attribute from that, a new feature data subset is prepared which contains 11 features.

2.7 Data Mining:

In this proposed framework, we observed that they have used different classification techniques for analyzing PD patient’s record. To evaluate performance, we apply 10-fold cross-validation technique which splits the original set into training sample to train the model and a test set to evaluate results. An approach known as “Information Retrieval Metrics” can be used to evaluate experimental results in terms of precision, recall, f-measure, and accuracy with the use of following formulas.

2.8 Experimental results:

Various techniques are used in the analysis and prediction of PD. Methods that are based on analytics can give an appropriate prediction for a particular disease by grouping people with similar symptoms. In our experiment, obtained accuracies using SMO, KNN, Random Forest, AdaBoost.M1, Bagging, MLP, and DT algorithms are 86.67%, 90.76%, 89.23%, 88.20%, 89.23%, and 89.74% respectively.

2.9 Conclusion and future scope:

The research area for Parkinson’s Disease is significant, early-stage detection of it can improve patient’s health. Big Data analytics plays a huge role in the healthcare industry, as these data are scattered everywhere, big, and complex in nature. In this paper, we observed early-stage prediction of Parkinson disease for that we presented a methodology of data mining using Weka tool for classifying disease dataset. We observed various MLAs for classifying our experimental data that indicate the combined effect of ANN algorithms with KNN and Decision tree algorithm which is better as compared than other algorithms. The system detects the maximum accuracy of the multi-classifier, and their result predicts the disease at its early stage. According to the study, females are more likely to have Parkinson’s than males. The study was done to detect and severity of Parkinson’s disease using a single model for each purpose. The study can be further extended by using other models and comparing the results to find the most optimized and efficient models for detection of disease and to determine the severity of the disease in the patient.

**Chapter 3**

Problem Statement

There is a lack of precise methods to predict the progression of Parkinson's disease, leading to challenges in providing optimal care and planning interventions for patients.

Solution Overview**:**

Our solution involves implementing machine learning algorithms with maximum accuracy on the patient’s data to develop a predictive model that can forecast the progression of Parkinson's disease based on various clinical markers derived from voice recordings. The goal is to apply algorithms such as K Nearest Neighbor, Decision Tree and Support Vector Machine algorithm and analyze the evaluation metrics and predict the status of the patient based on the unseen data given i.e. 1 if the patient is diagnosed with Parkinson’s Disease (PD) and 0 if the patient is healthy.

**Chapter 4**

Dataset Description

4.1 Description of the dataset used in the project:

The dataset being used is composed of a range of biomedical voice measurements from

195 people, 147 diagnosed with Parkinson's disease (PD) and 48 found healthy. The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to 195 voice recording. There are 24 columns where each column in the table denotes a particular voice measure. The main aim of the data is to discriminate healthy people from those with Parkinson’s disease, according to "status" column which is set to 0 for healthy and 1 for PD. [6]

4.2 Source of the dataset:

The dataset used in this project was sourced from Kaggle, a renowned platform for data science enthusiasts and practitioners [7]. 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 in 2008-06-26. The original study published the feature extraction methods for general voice disorders. [8].

4.3 Features included in the dataset:

Data Set Characteristics: Multivariate

Number of Instances: 197

Area: Life

Attribute Characteristics: Real

Number of Attributes: 24

Date Donated: 2008-06-26

Associated Tasks: Classification

Missing Values? N/A

Size of the dataset: 195 X 24

Column 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

4.4 Preprocessing steps applied to the dataset:

* Importing valid dataset
* Handling missing values
* Data Cleaning by replacing missing or inconsistent values with appropriate ones.
* Data Transformation such as normalization to prepare the data for machine learning algorithms.
* Ensuring consistency by scaling features to have similar ranges, aiding model performance.
* Train-Test Split: Dividing the pre-processed data into training and testing sets to assess model performance accurately.

**Chapter 5**

Methodology

5.1 Description of the preprocessing:

The preprocessing of the data for the Parkinson's disease machine learning project involved several steps. First, the data was imported from the relevant sources and inspected for missing values and inconsistencies. Any missing or inconsistent data was cleaned and replaced with appropriate values. The data was then transformed and normalized to ensure that it was in a suitable format for the machine learning algorithms to be used. Finally, the data was split into training and testing sets in the ratio of 7:3 to allow for evaluation of the machine learning models.

5.2 Description of the machine learning techniques/algorithms/models used:

SVM (Support Vector Machine): SVM is a supervised learning algorithm used mainly for classification tasks. It finds the optimal boundary (hyperplane) that maximizes the margin between different classes. For non-linearly separable data, it uses kernels like linear, polynomial, and RBF to project the data into a higher-dimensional space for separation.

Decision Tree: A decision tree is a flowchart-like structure where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. It's used for both classification and regression, offering a simple, interpretable way of making decisions based on data features.

KNN (K-Nearest Neighbors): KNN is a simple, non-parametric algorithm used for classification and regression. It classifies data points based on the 'K' closest training examples in the feature space. The object is assigned to the class most common among its K nearest neighbors (measured by a distance function).

5.3 Justification for the chosen techniques/algorithms/models:

The choice of machine learning algorithms for this project was based on their suitability for the task at hand and their performance on similar problems. KNN was chosen because of its simplicity and effectiveness for classification tasks. Decision Trees were chosen because of their ability to handle both categorical and numerical data, and their interpretability. SVM was chosen because of its ability to handle high-dimensional data and its strong performance on classification tasks. These algorithms were chosen because they have been shown to perform well on similar problems, and because they provide a diverse set of approaches to the problem of predicting Parkinson's disease.

**Chapter 6**

Implementation

6.1 Details of how the methodology was implemented.

6.1.1. Data preprocessing:

* The dataset was first cleaned and pre-processed to remove any irrelevant or missing data.
* The data was then transformed and normalized using zscore from scipy library to ensure that it was in a suitable format for the machine learning algorithms to be used.
* The data was split into training and testing sets using train\_test\_split in the ratio of 7:3 for the evaluation of the machine learning models.

6.1.2. Machine learning algorithms:

* K-Nearest Neighbours (KNN) with k=3 was implemented using the scikit-learn library in Python.
* Decision Trees with a depth of 3 was implemented using the scikit-learn library in Python.
* Support Vector Machines (SVM) with a radial basis function kernel was implemented using the scikit-learn library in Python.

6.1.3. Model evaluation:

* Training and testing the model on different subsets of the data: This involves dividing the data into a training set and a testing set, training the model on the training set, and evaluating its performance on the testing set. This method can provide an estimate of the model's performance on new data. The model is applied on the new unseen data and predictive system is built to predict whether a person has PD or is healthy.
* Using metrics such as accuracy, precision, recall, and F1 score: These metrics can provide an estimate of the model's performance on the data and can help identify any biases or errors in the model.

6.1.4. Model interpretation:

* The k-Nearest Neighbors (kNN) algorithm was chosen for its simplicity and ease of understanding. Interpretation of kNN models revolves around understanding the influence of neighboring data points on the classification decision. The "k" parameter determines the number of nearest neighbors considered for classification. Feature importance in kNN models is implicit and can be inferred from the relative positions and densities of data points in the feature space. Features that contribute significantly to the classification decision are those that exhibit clear patterns or clusters in the data distribution.
* The SVM Trees algorithm was chosen because SVM involves understanding the decision boundary, feature weights, and the relationship between the hyperplane and the classes. SVMs find the hyperplane that maximally separates the classes, and the weights of the hyperplane can be used to interpret the importance of each feature. The accuracy and sensitivity provided by the model makes it ideal to applied for the Parkinson prediction model.
* The Decision Tree algorithm was selected due to its inherent interpretability and simplicity. Decision trees partition the feature space into regions based on feature thresholds, making decisions at each node based on these thresholds. This hierarchical structure allows for easy interpretation of the decision-making process. Each node in the tree represents a feature and a corresponding threshold, with branches leading to subsequent nodes or leaves representing the final classification decision. Decision trees offer insights into feature importance by evaluating the frequency and significance of features in the decision-making process. The precision provided by the decision tree makes it ideal for the Parkinson’s disease prediction model

6.2 Tools and libraries used for implementation:

6.2.1 Libraries:

1. Pandas: Pandas is a powerful data manipulation and analysis library for Python. It provides data structures and functions for handling structured data and performing data analysis tasks such as data cleaning, manipulation, and aggregation.
2. NumPy: NumPy is a fundamental package for scientific computing with Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.
3. matplotlib.pyplot: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. pyplot is a sub-module of Matplotlib that provides a matlab-like interface for creating plots and visualizations.
4. Seaborn: Seaborn is a statistical data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

6.2.2 Tools:

1. scikit-learn (sklearn): Scikit-learn is a machine learning library for Python that provides simple and efficient tools for data mining and data analysis. It includes various algorithms for classification, regression, clustering, dimensionality reduction, and model selection.
2. scipy.stats: Scipy is a scientific computing library for Python that builds on NumPy. The stats module in Scipy provides a large number of probability distributions and statistical functions for analyzing data and performing statistical tests.

6.2.3 Machine Learning Models:

1. KNeighborsClassifier: This is a classification algorithm from scikit-learn (sklearn) that implements the k-nearest neighbors algorithm. It is a non-parametric method used for classification and regression tasks.
2. DecisionTreeClassifier: This is a classification algorithm from scikit-learn (sklearn) that implements a decision tree classifier. Decision trees are a popular method for classification and regression tasks, known for their simplicity and interpretability.
3. svm: This import likely refers to support vector machine (SVM) algorithms provided by scikit-learn (sklearn). SVMs are supervised learning models used for classification and regression tasks. Scikit-learn provides various SVM algorithms, including SVC (Support Vector Classification) for classification tasks.

6.3 Code snippets or pseudocode:

#Train-Test-Split:

from sklearn.model\_selection import train\_test\_split

from scipy.stats import zscore

pdata = pdata.drop(['name'], axis=1)

X = pdata.drop(['status'], axis=1)

y = pdata[['status']]

# Standardize features using Z-score normalization

XScaled = X.apply(zscore)

print(XScaled.describe())

# Split data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(XScaled, y, test\_size=0.30, random\_state=1)

#KNN algorithm:

from sklearn.neighbors import KNeighborsClassifier

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn import metrics

# Creating KNN classifier with 3 neighbors and 'distance' weights

NN = KNeighborsClassifier(n\_neighbors=3, weights='distance')

# Training the model with training data

NN.fit(X\_train, y\_train.values.ravel())

# Predicting labels for test data

y\_predict = NN.predict(X\_test)

predicted\_labels = NN.predict(X\_test)

# Calculating and printing the score of the KNN model

NN\_score = NN.score(X\_test, y\_test)

print('The KNN model score is', NN\_score)

#SVM

from sklearn import svm

svmmodel = svm.SVC(gamma=0.1, C=10)

svmmodel.fit(X\_train , y\_train.values.ravel())

svmmodel\_score1 = svmmodel.score(X\_train, y\_train)

print("Training model score:",svmmodel\_score1)

svmmodel\_score = svmmodel.score(X\_test, y\_test)

print('The SVM Model score is',svmmodel\_score)

#Decision Tree algorithm:

from sklearn.tree import DecisionTreeClassifier

import seaborn as sns

import matplotlib.pyplot as plt

# Initialize Decision Tree Classifier

dTreeR = DecisionTreeClassifier(criterion='entropy', max\_depth=3, random\_state=1)

# Train the Decision Tree Classifier

dTreeR.fit(X\_train, y\_train)

print('Training score:', dTreeR.score(X\_train, y\_train))

print('Testing score:', dTreeR.score(X\_test, y\_test))

# Visualize the decision tree

from sklearn.tree import plot\_tree

fn = list(X\_train)

cn = ['No', 'Yes']

plt.figure(figsize=(15, 10))

plot\_tree(dTreeR, feature\_names=fn, class\_names=cn, filled=True)

plt.savefig('tree.png')

plt.show()

# Make predictions on the testing set

ytree\_pred1= dTreeR.predict(X\_test)

**Chapter 7**

Results

7.1 Evaluation metrics are used to assess the performance of the model(s):

Evaluation metrics, including accuracy, precision, recall, specificity, and F1 score, are crucial for assessing the performance of classification models. They are calculated based on the information provided by a confusion matrix, which breaks down a model's predictions compared to the actual ground truth across different classes. Accuracy measures overall correctness, precision focuses on the correctness of positive predictions, recall assesses the model's ability to identify all positive instances, specificity evaluates the model's ability to correctly identify negative instances, and F1 score provides a balance between precision and recall. These metrics enable stakeholders to understand the strengths and weaknesses of the model and guide optimization efforts for improved performance in classification tasks.

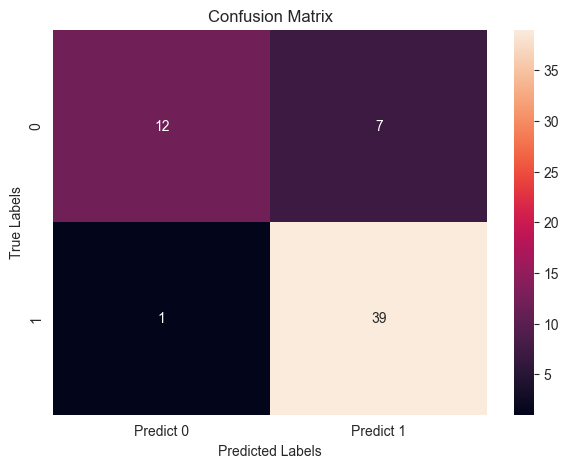


Fig 7.1: Confusion Matrix for KNN

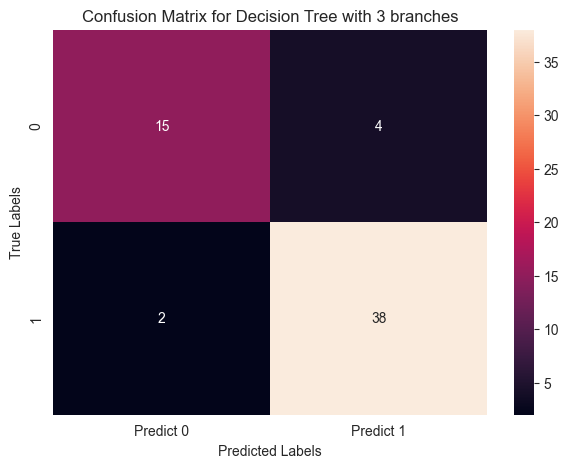


Fig 7.2: Confusion Matrix for Decision Tree

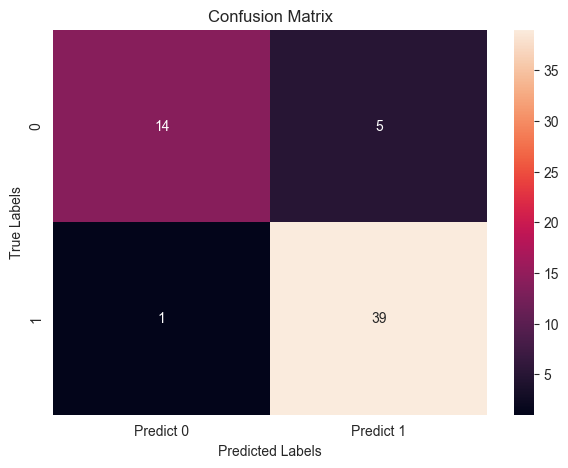


Fig 7.3: Confusion Matrix for SVM

|  |  |  |  |
| --- | --- | --- | --- |
| Model | K Nearest Neighbor Method | Decision Tree | Support Vector Machine |
| Accuracy | 86.44 | 89.83 | 89.83 |
| Sensitivity | 97.5 | 95.0 | 97.5 |
| Specificity | 63.15 | 78.94 | 73.68 |
| False Positive Rate | 36.84 | 21.05 | 26.31 |
| Precision | 84.78 | 90.47 | 86.63 |

Table 7.1: Obtained evaluation Metrix of Machine learning algorithms

Table 7.1 describes that:

1. K Nearest Neighbour (KNN) Method:

* Accuracy: 86.44% - This indicates the overall correctness of the model's predictions.
* Sensitivity (True Positive Rate): 97.5% - The proportion of actual positive cases correctly identified by the model.
* Specificity (True Negative Rate): 63.15% - The proportion of actual negative cases correctly identified by the model
* False Positive Rate: 36.84% - The proportion of actual negative cases incorrectly classified as positive.
* Precision: 84.78% - The proportion of true positive cases among all cases classified as positive.

2. Decision Tree:

* Accuracy: 89.83% - Higher accuracy compared to KNN.
* Sensitivity: 95.0% - Slightly lower sensitivity compared to KNN.
* Specificity: 78.94% - Higher specificity compared to KNN.
* False Positive Rate: 21.05% - Lower false positive rate compared to KNN.
* Precision: 90.47% - Slightly higher precision compared to KNN.

3. Support Vector Machine (SVM):

* Accuracy: 89.83% - Same accuracy as Decision Tree.
* Sensitivity: 97.5% - Same sensitivity as KNN.
* Specificity: 73.68% - Lower specificity compared to Decision Tree.
* False Positive Rate: 26.31% - Higher false positive rate compared to Decision Tree.
* Precision: 86.63% - Slightly lower precision compared to Decision Tree.

Explanation:

KNN has high sensitivity but low specificity, meaning it tends to correctly identify positive cases but struggles with negative cases.

Decision Tree performs well overall with balanced sensitivity and specificity, making it a good choice for this dataset.

SVM performs similarly to Decision Tree in terms of accuracy but has lower specificity and slightly lower precision.

In summary, while all three models achieve relatively high accuracy, Decision Tree might be preferred due to its balanced performance in terms of sensitivity, specificity, and precision.

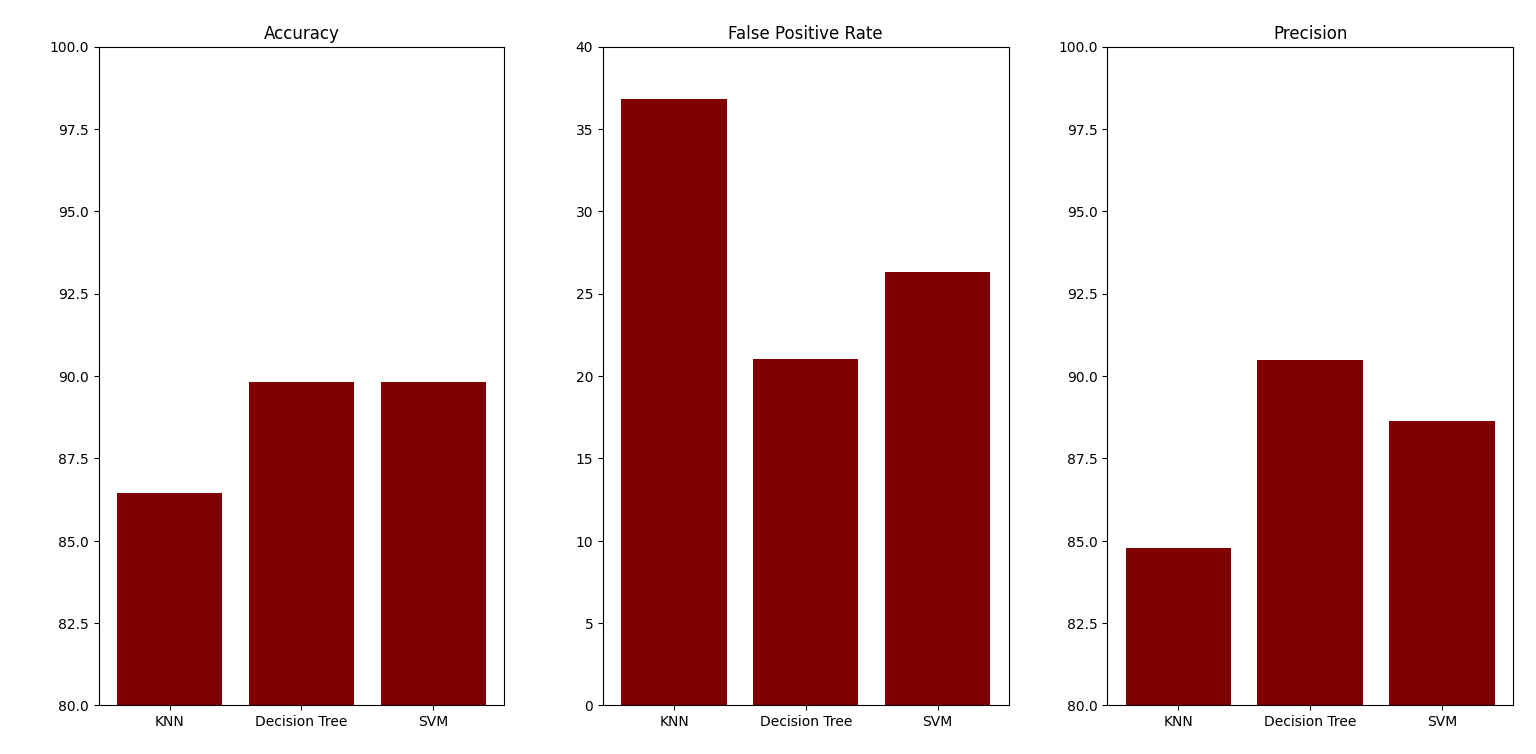


Fig 7.4: Accuracy, False positive rate and Precision

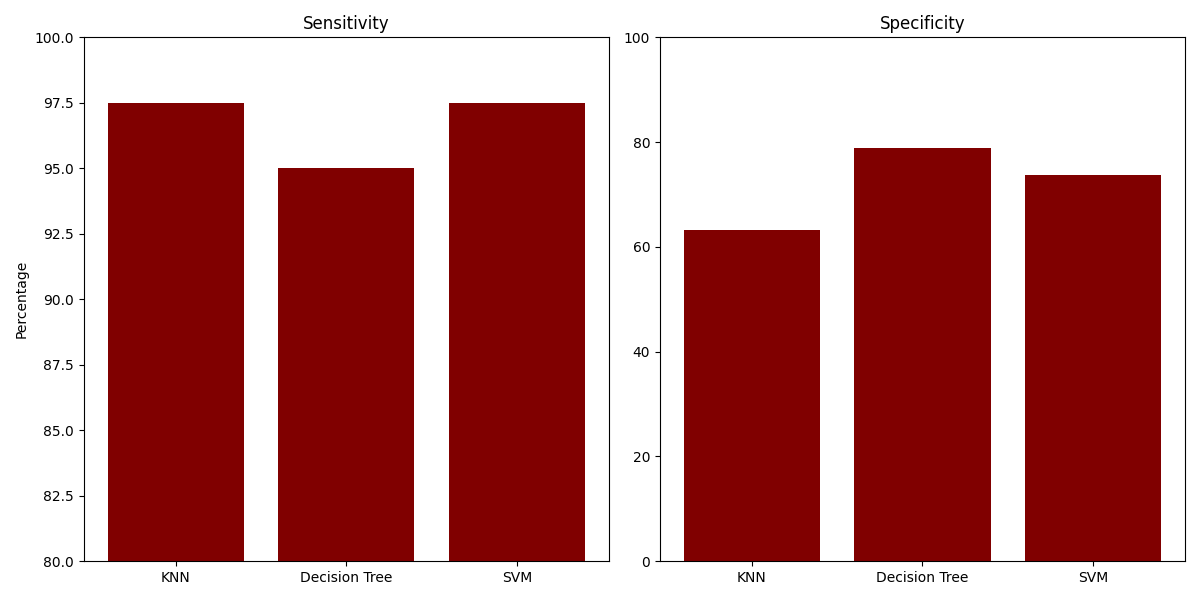
****

Fig 7.5: Sensitivity and specificity

**Chapter 8**

Conclusion

In conclusion, the study demonstrates the potential of machine learning models in predicting Parkinson's Disease progression from voice recordings. Among the models evaluated, the Decision Tree algorithm stands out for its balanced performance in accuracy, sensitivity, specificity, and precision. This indicates its effectiveness in capturing the distinctions in voice features associated with Parkinson's Disease. The Decision Tree model offers a promising avenue for non-invasive and efficient screening, thus useful for early diagnosis and treatment planning. These findings underscore the significance of interdisciplinary approaches in healthcare and highlight the role of advanced computational techniques in improving medical diagnostics and patient care.

**References:**

1. <https://www.mygreatlearning.com/>
2. <https://en.wikipedia.org/wiki/Parkinson%27s_disease>
3. Azad C et al (2014) Design and analysis of data mining based prediction model for Parkinson’s disease
4. Saloni et al (2015) Detection of Parkinson disease using clinical voice data mining 4. Shamli N et al (2016) Parkinson’s Brain disease prediction using big data analytics
5. Sriram TVS et al (2013) Intelligent Parkinson disease prediction using machine learning algorithms.
6. 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection'
7. <https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set>
8. Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease'