“ParkinSight: Parkinson’s Disease Prediction Based Voice & Symptoms Profiling”

**Mini Project 2B Report**

Submitted in partial fulfillment of the requirement of University of Mumbai

For the Degree of

**(Artificial Intelligence & Data Science)**

**By**

1. **“Ali Mehdi Ajani” “ID No: TU8F2122029”**
2. **“Suryansh Gupta” “ID No: TU8F2122033”**
3. **“Priyanka Angane” “ID No: TU8F2122036”**
4. **“Devika Patil” “ID No: TU8F2122045”**

**Under the Guidance of**

**Prof. Vishal R. Gotarane**



**Department of Artificial Intelligence & Data Science**

**TERNA ENGINEERING COLLEGE**

**Plot no.12, Sector-22, Opp. Nerul Railway station,**

**Phase-11, Nerul (w), Navi Mumbai 400706**

**UNIVERSITY OF MUMBAI**

**SH-2023**



## TERNA ENGINEERING COLLEGE, NERUL, NAVI MUMBAI

**Department of Artificial Intelligence & Data Science**

Academic Year 2023-24

**CERTIFICATE**

This is to certify that the mini project 2B entitled **“ParkinSight: Parkinson’s Disease Prediction Based Voice & Symptoms Profiling”** is a bonafide work of

1. **“Ali Mehdi Ajani” “ID No: TU8F2122029”**
2. **“Suryansh Gupta” “ID No: TU8F2122033”**
3. **“Priyanka Angane” “ID No: TU8F2122036”**
4. **“Devika Patil” “ID No: TU8F2122045”**

Submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the Bachelor of Engineering (Artificial Intelligence & Data Science).

|  |  |  |
| --- | --- | --- |
| Prof. Vishal Gotarane | Dr. Sandeep B. Raskar | Dr. L. K. Ragha |
| **Guide Name** | **Head of the Department** | **Principal** |

## Project Report Approval

This Mini Project 2 B Report – entitled “**ParkinSight: Parkinson’s Disease Prediction Based Voice & Symptoms Profiling”** by following students is approved for the degree of ***B.E. in "Artificial Intelligence & Data Science"***.

**Submitted by:**

1. **“Ali Mehdi Ajani” “ID No: TU8F2122029”**
2. **“Suryansh Gupta” “ID No: TU8F2122033”**
3. **“Priyanka Angane” “ID No: TU8F2122036”**
4. **“Devika Patil” “ID No: TU8F2122045”**

Examiner Names & Signatures

1.

2.

Date:

Place:

# Declaration

#### We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

**“Ali Mehdi Ajani” “ID No: TU8F2122029” \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**“Suryansh Gupta” “ID No: TU8F2122033” \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**“Priyanka Angane” “ID No: TU8F2122036” \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**“Devika Patil” “ID No: TU8F2122045” \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

#### Date:

Place:

# Acknowledgement

#### We would like to express our sincere gratitude towards our guide & mini-project coordinator **Prof. Vishal Gotarane** for the help, guidance and encouragement they provided during the project development. This work would have not been possible without their valuable time, patience and motivation. We thank them for making our stint thoroughly pleasant and enriching. It was great learning and an honor being their student.

We are deeply thankful to **Dr. Sandeep B. Raskar, HOD, AI & DS** and the entire team in the Computer Department. They supported us with scientific guidance, advice and encouragement, they were always helpful and enthusiastic and this inspired us in our work.

We take the privilege to express our sincere thanks to **Dr. L. K. Ragha** our Principal for providing encouragement and much support throughout our work.

1. **“Ali Mehdi Ajani” “ID No: TU8F2122029”**
2. **“Suryansh Gupta” “ID No: TU8F2122033”**
3. **“Priyanka Angane” “ID No: TU8F2122036”**
4. **“Devika Patil” “ID No: TU8F2122045”**

#### Date:

Place:

# Index

### TABLE OF CONTENTS

**Sr. No. Title Page No.**

|  |  |  |
| --- | --- | --- |
|  | **Abstract** | **7** |
| **List of Figures** | **8** |
| **List of Tables** | **10** |
| **Chapter 1** | **Introduction** | **11** |
| 1.1 | Motivation for the topic | 11 |
| 1.2 | Aim | 11 |
| **Chapter 2** | **Problem Statement** | **12** |
| 2.1 | Problem Statement | 12 |
| 2.2 | Features | 12 |
| 2.3 | Objectives | 12 |
| 2.4 | Specifications of the system | 12 |
| **Chapter 3** | **Literature Survey** | **13** |
| 3.1 | Literature Survey | 13 |
| **Chapter 4** | **Design and Implementation** | **18** |
| 4.1 | Working of the Project | 20 |
| 4.2 | Algorithm | 22 |
| **Chapter 5** | **Results & Discussion** | **23** |
| 5.1 | Result | 23 |
| **Chapter 6** | **Conclusion and Future Work** | **30** |
| 6.1 | Conclusion | 30 |
| 6.2 | Future Work | 30 |
|  | **References** | **31** |

# Abstract

This project endeavors to develop a comprehensive system for predicting Parkinson's disease using voice-based analysis, machine learning algorithms, and symptom data integration. By combining sophisticated methodologies such as Support Vector Machine (SVM), XGBoost, and Random Forest with symptom dataset analysis, the system aims to provide a robust and holistic approach to disease prediction. Existing diagnostic methods for Parkinson's disease often lack integration with symptom data, resulting in potential inaccuracies. Voice data is collected and processed to extract relevant features like pitch, intensity, and formants, while symptom datasets provide additional insights into the disease profile. These datasets are then used in tandem to train and test machine learning models, including SVM, XGBoost, and Random Forest.

The accuracies obtained from both voice-based analysis and symptom dataset analysis are integrated to generate final predictions. This integration allows for a comprehensive evaluation of the disease likelihood, taking into account both vocal characteristics and symptom profiles. The finalized system seamlessly integrates with existing healthcare infrastructure, providing healthcare professionals with a unified platform for inputting voice recordings, symptom data, and accessing diagnostic outcomes.

# List of Figures

### Sr. No. Name of Figure Page No.

1. 4.5: Architecture Diagram ParkinSight: Parkinson’s Disease Prediction Based Voice & Symptoms Profiling

21

#### 5.1: Parkinson’s disease Dataset Collection Overview 22

#### 5.2: Data Analysis 22

#### 5.3: Data Visualization 23

#### 5.4: Model Evaluation (Confusion Matrix) 23

#### 5.4.1 Support Vector Machine 24

#### 5.4.2: Logistic Regression 24

#### 5.4.3: XGBoost Classifier 25

#### 5.4.4: Random Forest 25

#### 6 5.5: Model Accuracy Comparison 25

#### 5.5.1: Comparison of Models based on evaluation metrics 26

#### 5.6: Hyper tuned Random Forest 26

#### 5.7: Symptoms Dataset 28

#### 5.8: Imported Audio File 28

#### 5.9: Preprocessed Input Features 29

#### 5.10: Integrated Voice & Symptoms Prediction Results 29

|  |  |  |
| --- | --- | --- |
|  | **List Of Tables** |  |
| **Sr. No.** | **Name of Table** | **Page No.** |
| 1 | Table 3.1: Literature Survey | 13 |

**Chapter 1**

**Introduction**

##### 1.1 Motivation:

Parkinson's disease (PD) is a progressive neurological disorder affecting millions globally, leading to motor impairments and speech difficulties. Traditional diagnostic methods are costly and time-consuming, causing delays in treatment. However, machine learning presents a promising solution for early PD detection through voice analysis. By analyzing vocal characteristics like pitch and articulation, our project aims to develop a non-invasive method for PD detection. This innovation could lead to timely interventions and personalized treatment plans, ultimately enhancing patient outcomes and cutting healthcare costs.

##### 1.1.1 Need of the problem:

The necessity for innovative solutions in healthcare, such as predicting Parkinson’s disease (PD) through machine learning, is driven by the limitations of traditional diagnostic methods. By analyzing voice data for subtle PD indicators, this project aims to revolutionize diagnosis, ensuring accuracy and enabling timely interventions, thereby showcasing the potential of technology to transform healthcare.

##### 1.2 Scope of the project:

The primary aim of this project is to revolutionize Parkinson's disease (PD) diagnosis by leveraging machine learning techniques for voice analysis. The project seeks to develop a predictive model capable of detecting subtle indicators of PD within voice data, thereby facilitating early intervention and personalized treatment strategies. Additionally, the project aims to store and manage PD prediction data efficiently, enabling healthcare professionals to access and utilize the information effectively.

The scope of the project encompasses the following key elements:

- Implementation of machine learning algorithms for PD prediction based on voice analysis.

- Development of a robust database system to store and manage PD prediction data securely.

- Integration of the predictive model into existing healthcare systems to facilitate seamless adoption and utilization by healthcare professionals.

- Emphasis on scalability and reliability to ensure the applicability of the solution across diverse healthcare settings and patient populations.

Through this project, we aim to address the limitations of traditional PD diagnosis methods and pave the way for more accurate, efficient, and accessible healthcare solutions for individuals affected by Parkinson's disease.

* 1. **Aim:**

In this innovative research endeavor, we aim to develop a robust system for early Parkinson’s disease prediction by leveraging the power of ensemble machine learning applied to voice analysis.

## Chapter 2

**Problem Statement**

##### Problem statement

While voice analysis shows potential as a Parkinson’s disease biomarker, existing methods suffer from accuracy issues, primarily depending on the single algorithms & disregarding individual variations. This study proposes improvements to precision & robustness, aiming to create a more effective tool for the early detection of Parkinson’s thereby enhancing diagnostic and intervention strategies.

**Features:**

- Extraction of vocal features such as pitch, intensity, and formants.

- Preprocessing of voice recordings to remove noise and irrelevant information.

- Division of dataset into training and testing sets.

- Training of machine learning algorithms (e.g., SVM, XGBoost, Random Forest) on the training data.

- Optimization of hyperparameters for model performance enhancement.

- Integration of external symptoms data for comprehensive analysis.

- Evaluation of model performance using testing data.

- Calculation of average accuracy through cross-validation.

- Prediction of Parkinson's disease based on vocal features.

- Interpretation and assessment of model reliability.

##### Objectives

* **Identify Biomarkers:**

Enable timely identification of Parkinson’s disease using voice-based machine learning models.

* **Develop Predictive Models**
* **Enhance Model Accuracy.**

Train the selected machine learning models using the preprocessed voice dataset, optimizing model hyperparameters to maximize predictive performance while avoiding overfitting.

* **User-friendly interface to interact with the developed system.**

**Chapter 3**

**3.1 Literature Survey**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr No.** | **Paper Details**  **Authors, “Title”, Resource Name, Year, page nos.** | **Problem Addressed** | **Methodology Used** | **Advantages/ Strengths Found** | **Disadvantages/ Limitations/Gaps Found** |
| 1 | Diogo Braga, Ana M. Madureira, Luis Coelho, Reuel Ajith  **“Automatic detection of Parkinson’s disease based on acoustic analysis of speech”**  Engineering Applications of Artificial Intelligence,  Volume 77, 2019, Pages 148-158,  ISSN 0952-1976, Elsevier. | Parkinsons reduces the average life span and deteriorates the quality of life.  In the year 2012, the costs of Alzheimer’s disease in the US were about $200 billion. Gustavsson estimated, based on data from 2010, a total cost of e13.9 billion in  Europe for the PD and expect that these numbers duplicate by 2030 | Data Analysis and splitting, decision model deployment, and model deployment.  Made use of ml algorithms like SVM, NN, Random Forest, Perceptron and Naïve Bayes. | Highest Accuracy in the model of 99.94% was obtained by optimized Random Forest algorithm.  The leave-one-out corss validation technique was used toobtain the general accuracy of the learning algorithms. | Although the accuracy was reached but there may be an indication that it might be due to the overfitting of data into the model.  No other limitations. |
| 2 | Iqra Nissar1,, Danish Raza Rizvi1, Sarfaraz Masood1 and Aqib Nazir Mir  **“Voice-Based Detection of Parkinson’s Disease through  Ensemble Machine Learning Approach: A Performance**  **Study”**  Department of Computer Engineering, Jamia Millia Islamia, New Delhi-110025 | Utilizing machine    learning, the research focuses on identifying Parkinson's disease through the analysis of voice signals.  The study aims to evaluate and compare the performance of various classification algorithms on a voice dataset. | Machine learning models, including XGBoost, were employed for Parkinson's disease detection.  Two feature selection techniques, Recursive Feature Elimination (RFE) and minimum Redundancy Maximum Relevance (mRMR), were applied. | XGBoost outperformed other classifiers with an accuracy of 95.39% using mRMR feature selection.  The proposed model shows high precision, F1-score, recall, and accuracy rates, indicating its reliability in Parkinson's disease detection. | The model's efficiency is limited by the richness of the dataset, which contains only 756 instances. |
| 3 | Azian Abdullah,Nurul Norazman, Wan Ahmad, Saidatul Awang, Foong Jian.  **“Detection of Parkinson’s Disease (PD) Based On Speech Recordings using Machine Learning Techniques”**  IEEE, 2020 Internation Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT) | Increasing cases of Parkinson’s in the world, becoming 2nd most common after Alzheimer’s in the world.  Growing numbers of Parkinson’s patients in Malaysia, estimated to be around 120000 by 2040. | Usage of Machine Learning Algorithms such as Random Forest, Support Vector Machine, Deep Neural Network, XGBoost Classifier and trial of finding the most accurate detection.  They implemented a total of 8 Machine Learning Algorithms. | Applying 8 Machine Learning algorithms provided a huge scope for accuracy.  Finally, algorithm chosen was LASSO-SVM-GSCV which provided an accuracy of 100% over the other algorithms by recording precision of 97.87%, recall of 97.87%, specificity of 65% and 97.10% for AUC. | Future research needs to apply non motor features for early detection of PD. Motor symptoms occur late, but non motor symptoms may occur early, therefore, they need to be implemented, which is complex as of now. |
| 4 | Ismail Cantürk, Fethullah Karabiber,  **“A Machine Learning System for the Diagnosis of Parkinson’s Disease from Speech Signals and Its Application to Multiple Speech Signal Type”**  Springer, Arab J Sci Eng  DOI 10.1007/s13369-016-2206-3, King Fahd University of Petroleum & Minerals 2016 | Treatment of PD has not been discovered yet, therefore the detection can be helpful for early help.  Current therapies are only useful for handling PD symptoms, but PD’s progress cannot be stopped. Noninvasive diagnosis methods of PD need to be  investigated. | They developed a machine learning system to classify PWP using their speech signals.  In the system, four feature selection algorithms, six classifiers, and two validation methods were employed for accurate classification of PWP. The system calculated the accuracy, sensitivity, specificity, and Matthews correlation coefficient of the results. | Results of 10-fold CV  are more successful than LOSO. 10-fold CV can be used for cases where there are multiple type phonations of more than  one person to classify PWP at the same time. | If patients are diagnosed with certain phonations,  their stress level in the PD diagnosis process might be raised.  In addition, we know that our voice is easily affected by our stress or excitement like vibrations in sound |
| 5 | L. Brabenec,  J. Mekyska, Z. Galaz, Irena Rektorova  **“Speech disorders in Parkinson’s disease: early diagnostics and effects of medication and brain stimulation”**  Springer, J Neural Transm DOI 10.1007/s00702-017-1676-0, Springer-Verlag Wien 2017 | Hypokinetic dysarthria (HD) occurs in 90% of  Parkinson’s disease (PD) patients. Therefore, detecting PD early may help in the detection of HD. | This paper aimed at studying different literary papers based on PD detection.  The paper focused on various IEEE and Springer papers that mentioned HD along with PD detection. | Identified 34 studies comprising 31 and 3 (Hazan et al. 2012; Orozco Arroyave et al. 2015b, 2016) works focused on unilingual  and multilingual diagnosis, respectively. Regarding the stage of PD during the acquisition of the data sets, fivestudies focused on the early PD diagnosis. | Acoustic Analysis of speech is yet to be implemented and might be quite helpful in better and more accurate detection. |
| 6 | Daniel Alonso, Guillermo Morales, Agustin Arribas, Carlos Carrascosa, Andres Rodellar, Pedro Vilda.  **“MonParLoc: A Speech-Based System for Parkinson’s Disease Analysis and Monitoring”**  IEEE, Digital Object Identifier 10.1109/ACCESS.2020.3031646, date of current version October 27, 2020 | Monitoring Parkinson using Locution  (MonParLoc), and acoustical neurostimulation, called within project neuro-Acoustic-stimulation Parkinson (AcousticPar). | From speech recordings, 72 features are extracted and analyzed.  The developed app joins acoustic neurostimulators (using AcousticPar) and diadochokinetic exercises (using MonParLoc). | MonParLoc provides a full speech study from easy-going speech exercises. Articulation, Phonation, and Diadochokinetic exercises present a powerful tool to understand the state of the patient. Besides, this tool could be used in other neurodegenerative impairments such as Alzheimer’s | No limitation of this research. |
| 7 | Alice Rueda and Sridhar Krishnan  **“Feature Analysis of Dysphonia Speech for Monitoring Parkinson's Disease”**  IEEE, Department of Electrical & Computer Engineering, Ryerson Univerity, Toronto, ON M5B 2K3, 978-1-5090-2809-2/17/$31.00 ©2017 IEEE | More than 5.9 million people are diagnosed with Parkinson's disease (PD) worldwide, and was the cause of death for more than 100,000 people in 2013. In America alone, 60,000 new PD cases are diagnosed every year. Patients are generally diagnosed after the age of 50, only 4% are before the age of 50 | Tsanas used data collected from an Internet enabled device made by Intel, called At Home Testing Device (AHTD).  In this study, they downsampled the sustained vowel to toll quality, with sampling frequency at 8 kHz and quantization at a reduced resolution of 8-bit and applied the MFCC and IMF feature combination. | Some important acoustic characteristics of Parkinson’s dysphonia can be represented by a combination of MFCC  and IMFs.  The MFCC and IMF spectrum power also show promising results for potential telemedicine over a regular phone line | Further studies are required on the usefulness of the first MFCC coefficient and the number of IMFs created, and to determine if the number of IMFs of a PWP sustained vowel can be used to determine the dysphonia severity. We have also found Shannon's entropy gave a neutral result. Other entropy measures might perform better. |
| 8 | Sura mahmood abdullah, Thekra abbas, Munzir hubiba bashir, Ishfaq ahmad khaja , Musheer ahmad , Naglaa f. Soliman , Walid el-shafai  **“Deep Transfer Learning Based Parkinson’s Disease Detection Using Optimized Feature Selection”**  Department of Computer Sciences, University of Technology, Baghdad 10066, Iraq | Performance Validation  The paper addresses the challenge of accurately detecting Parkinson's disease using handwritten records  The proposed framework aims to reduce the burden of training time | The framework incorporates transfer learning models such as ResNet, VGG19, and InceptionV3, which are pre-trained on large datasets  The optimization phase considers accuracy as the fitness value, with KNN serving as the objective function for the genetic algorithm | Transfer Learning Integration: Leveraging transfer learning models like ResNet, VGG19, and InceptionV3 reduces the training time significantly, making the framework efficient and practical.  Optimized Feature Extraction  Computational Efficiency | Lack of Comparison with State-of-the-Art Methods  Dependency on Existing Datasets: The framework relies on the availability and quality of the NewHandPD dataset  Sensitivity to Hyperparameters |
| 9 | Khaled M. Alalayah, Ebrahim Mogammed Senan  **“Automatic and Early Detection of Parkinson’s Disease by Analyzing Acoustic Signals Using Classification Algorithms Based on Recursive Feature Elimination Method”** Department of Computer Science, Faculty of Science and Arts, Najran University, Sharurah 68341, Saudi Arabia | Machine learning (ML) techniques, particularly analyzing voice disorders, have been utilized for early PD diagnosis.  Dimensionality reduction techniques | The study used Recursive Feature Elimination (RFE) to handle feature correlation and outliers.  Five different classifiers – Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Multi-Layer Perceptron (MLP) – were trained and tested using the resulting features obtained from both t-SNE and PCA | High Accuracy: The classifiers achieved high accuracy, precision, recall, and F1-score during testing, indicating the robustness of the proposed methodology in accurately diagnosing PD cases.  Non-Invasive Approach: Utilizing acoustic signals for diagnosis is non-invasive, making it a convenient and accessible method for screening individuals for PD | Feature Correlation: Although the dataset comprised 22 features, many of them were highly correlated, limiting the suitability of some features for high-level diagnosis  Outlier Removal: The removal of features containing outliers might have influenced the dataset's representativeness and could potentially impact the generalizability of the findings. |
| 10 | Basil K Varghese, Geraldine Bessie Amali  D, Uma Devi K S  **“Prediction of Parkinson’s Disease using Machine Learning Techniques on Speech dataset”**  School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India | The research paper addresses the challenge of accurately predicting motor UPDRS and total UPDRS scores in individuals  By utilizing the Parkinson’s Telemonitoring Dataset, the study aims to develop a data-driven approach for aiding in the diagnosis and monitoring of Parkinson's disease | Feature Extraction: Sixteen voice features were extracted from the dataset as predictors for motor UPDRS and total UPDRS scores.  Multiple machine learning algorithms including SVM, Decision Trees, Linear Regression, and Resilient BP were trained on the dataset to predict UPDRS scores | Accurate Prediction:  The study successfully predicted motor UPDRS and total UPDRS scores using 16 voice features from the Parkinson’s Telemonitoring Dataset, indicating the effectiveness of the machine learning models employed.  The paper offers valuable insights into the performance and suitability of different models for this specific dataset. | Need for Human Intervention: indicating the limitations of purely automated approaches in medical decision-making.  Data Limitations: suggesting potential limitations in the generalizability of findings to broader populations or datasets Model Selection Bias |
| 11 | Sandhiya S, Dr.Ashok.S, Mr.G.Vishnu ,Vardhan Rao, Dr.Prabhu V  **“Parkinson's Disease Prediction Using Machine Learning Algorithm”**  PG Scholar, Department of Electronics and Communication Engineering, Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering Chennai, India. | Address the problem of Parkinson's disease detection and classification using spiral/wave drawing datasets obtained from both healthy individuals and Parkinson's disease patients. | The Random Forest Classifier was used for the successful classification of healthy individuals and Parkinson's disease patients. | High Accuracy:The Random Forest Classifier achieved an accuracy of 71.33 percent  Efficiency and Productivity  Sensitivity and Specificity | Scope of Dataset  The research was based on a dataset comprising 102 individuals. The limited sample size might affect the generalizability of the results to a broader population.  Sensitivity to Dataset Characteristics |
| 12 | Jefferson S. Almeida, Pedro P. Reboucas Filho, Tiago Carneiro, Wei, Robertas Damasevicius, Rytis Maskeliunas, Hugo C. de Albuquerque  **“Detecting Parkinson’s Disease with Sustained Phonation and Speech Signals using Machine Learning Techniques”**  Pattern Recognition Letters (2019),  doi:<https://doi.org/10.1016/j.patrec.2019.04.005>, Elsevier. | Address the problem of rapidly increasing cases of Parkinson’s disease which affects dopamine levels and overall immune system of the body. It is a huge threat to public health. | They applied 18 feature sets and 14 classifiers on voice dataset and evaluated the results in terms of ERR, and Accuracy. | The results show that the Phonation task is the most suitable for use.  The results confirm that it is possible to develop mobile applications for detecting Parkinson’s disease using smartphone.  The highest accuracy reached was 94.55% | Does not address the problem of unbalanced datasets.  Better optimization and testing of feature sets. |

Table 3.1: Literature Survey

Research on Parkinson's disease (PD) detection via machine learning and speech analysis highlights significant progress in early diagnosis. Utilizing diverse algorithms like SVM, Random Forest, and XGBoost, studies achieved high accuracies ranging from 71.33% to 100%. These approaches incorporate techniques such as Recursive Feature Elimination and transfer learning models like ResNet and VGG19 to reduce training time. Challenges remain, including dataset limitations, potential overfitting, and the need for comprehensive feature optimization. While telemedicine and mobile applications show promise, addressing stress-induced biases and ensuring scalability are critical for broader implementation. Overall, these studies showcase the transformative potential of machine learning in enhancing PD diagnosis and monitoring, paving the way for more accessible healthcare solutions.

**Chapter 4**

## Design and Implementation

##### 4.1Tool Used:

##### Python

##### Librosa

##### TensorFlow / PyTorch

##### XGBoost / LightGBM

##### Pandas

##### Matplotlib / Seaborn

##### Jupyter Notebook / Google Colab

* 1. **Software Components:**
     1. scikit-learn

##### numpy

##### Working of the System:

**Project Scope:**

- The project aims to develop a machine learning model to identify Parkinson's disease based on vocal features extracted from audio recordings, focusing on achieving high accuracy, reliability, and interpretability in predictions. It integrates voice analysis and symptom assessment to create comprehensive predictive models for disease detection.

**Data Collection:**

- Gather diverse vocal datasets containing recordings from individuals with and without Parkinson's disease to ensure representativeness and account for various demographics. Utilize datasets such as the "Parkinsons" dataset from the UCI Machine Learning Repository, which includes biomedical voice measurements from individuals diagnosed with Parkinson's disease.

**Preprocessing:**

- Clean audio data by removing noise and irrelevant information using libraries such as librosa, scipy, and numpy for audio feature extraction. Extract relevant features including pitch, jitter, shimmer, formants, and voice intensity from voice recordings to capture vocal abnormalities indicative of Parkinson's disease. Ensure standardization and normalization techniques for efficient model training.

**Labelling:**

- Label the dataset based on the presence or absence of Parkinson's disease obtained from the UCI dataset to ensure accurate and consistent model training. Adhere to ethical guidelines and privacy regulations throughout the data collection and labelling process.

**Model Selection:**

- Choose the Random Forest algorithm for its ability to handle complexity, provide interpretability, and classify individuals effectively based on vocal features. Consider other algorithms such as Support Vector Machines (SVM), XGBoost, and Logistic Regression for comparison and evaluation.

**Model Training:**

- Train and optimize the Random Forest model on a subset of the dataset, adjusting hyperparameters with validation data to improve performance. Explore different machine learning algorithms including SVM, XGBoost, and Logistic Regression for comparative analysis and performance evaluation.

**Integration of External Symptoms:**

- Integrate external symptom data obtained through custom-designed graphical user interfaces (GUIs) for administering questionnaires to users. Assess Parkinsonian symptoms using standardized scales and questionnaires to capture motor and non-motor symptoms associated with Parkinson's disease.

**Model Evaluation:**

- Assess model performance using testing data, considering evaluation metrics such as accuracy, precision, recall, and F1 score. Conduct rigorous validation procedures including cross-validation or independent testing on unseen data to ensure model generalizability and reliability.

**Average Accuracy Calculation:**

- Calculate average accuracy by evaluating model performance across different cross-validation folds to provide a robust measure of effectiveness. Compare the performance of Random Forest with other machine learning algorithms to determine the most effective approach for Parkinson's disease prediction.

**Prediction and Interpretation:**

- Implement the trained Random Forest model to predict Parkinson's disease based on vocal features, interpret results, and assess model reliability in distinguishing between individuals with and without the disease. Analyze feature importance to understand influential vocal characteristics contributing to disease prediction.

##### Algorithm:

**4.5.1 Random Forest:**

Random Forest is a machine learning algorithm utilized to classify individuals as either having Parkinson's disease or not based on features extracted from their voice recordings. The algorithm constructs multiple decision trees using subsets of the voice data and features such as pitch, intensity, and formants. By aggregating the predictions of these trees, Random Forest provides a robust and accurate classification outcome. Its ability to handle complex relationships in the data, handle noise, and provide insights into feature importance makes it well-suited for building a reliable prediction model for Parkinson's disease using voice data.

**1. Training Process:**

- In the training phase, the Random Forest algorithm utilizes voice recordings along with their corresponding labels (indicating the presence or absence of Parkinson's disease) to build an ensemble of decision trees.

- Each decision tree is trained on a random subset of the voice dataset and a random subset of vocal features, providing diversity within the ensemble.

**2. Decision Making**

- During the prediction phase, each decision tree in the Random Forest independently evaluates a voice recording.

- The final prediction is made by aggregating the predictions of all decision trees in the forest, typically through majority voting.

**3. Feature Importance**

- Random Forest offers insights into feature importance, revealing which vocal features are most influential in distinguishing between individuals with and without Parkinson's disease.

- This feature importance analysis aids in understanding the underlying relationships between vocal characteristics and the presence of Parkinson's disease.

**4. Robustness and Generalization**

- Random Forest is robust to noise and outliers in the voice data, making it suitable for real-world applications.

- It tends to generalize well to unseen voice recordings, providing reliable predictions for individuals not present in the training dataset.

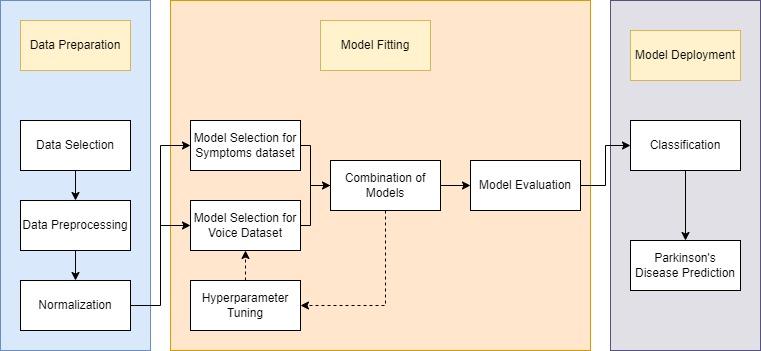
**5. Interpretability:**

- While Random Forest may not offer direct interpretability in terms of visualizing facial features like SVM does for face recognition, it provides interpretability through feature importance analysis.

- Researchers can gain insights into which vocal features contribute most to the prediction of Parkinson's disease, aiding in understanding the model's decision-making process.

**6. Scalability and Efficiency:**

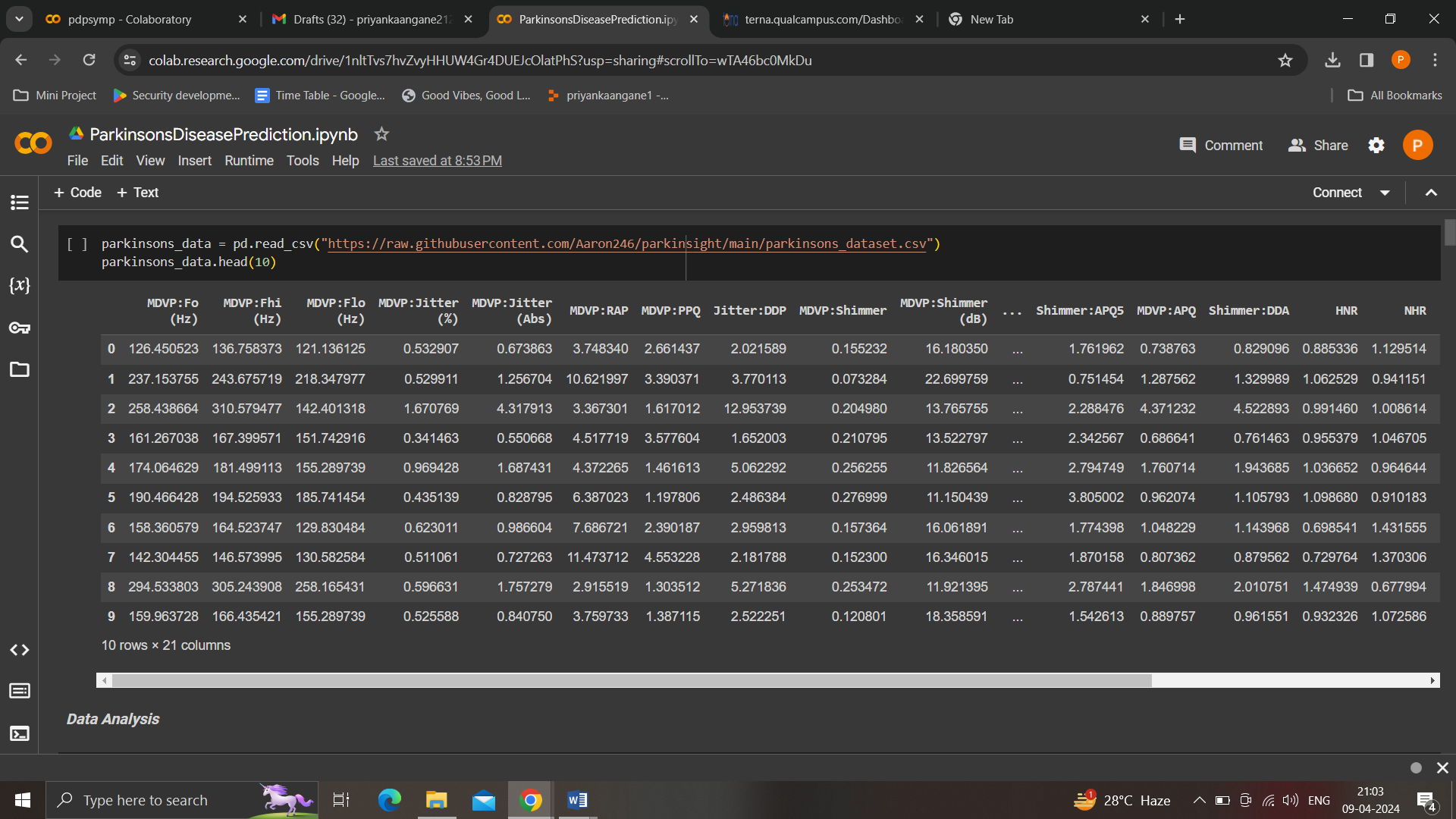
- Random Forest is scalable to large datasets and can efficiently handle high-dimensional feature spaces, making it suitable for processing voice data from diverse individuals.

****

**Figure 4.5 illustrates the architecture diagram of ParkinSight, a predictive model for Parkinson's disease based on voice and symptom profiling, outlining the system's design and components.**

## Chapter 5

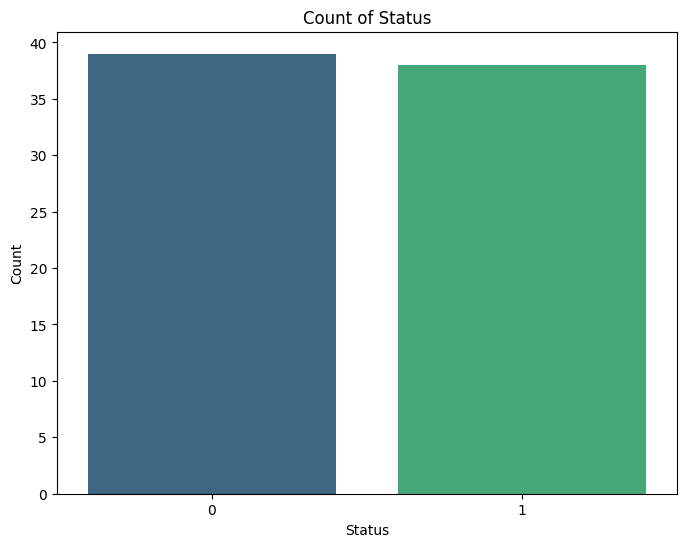
**Results and Analysis:**



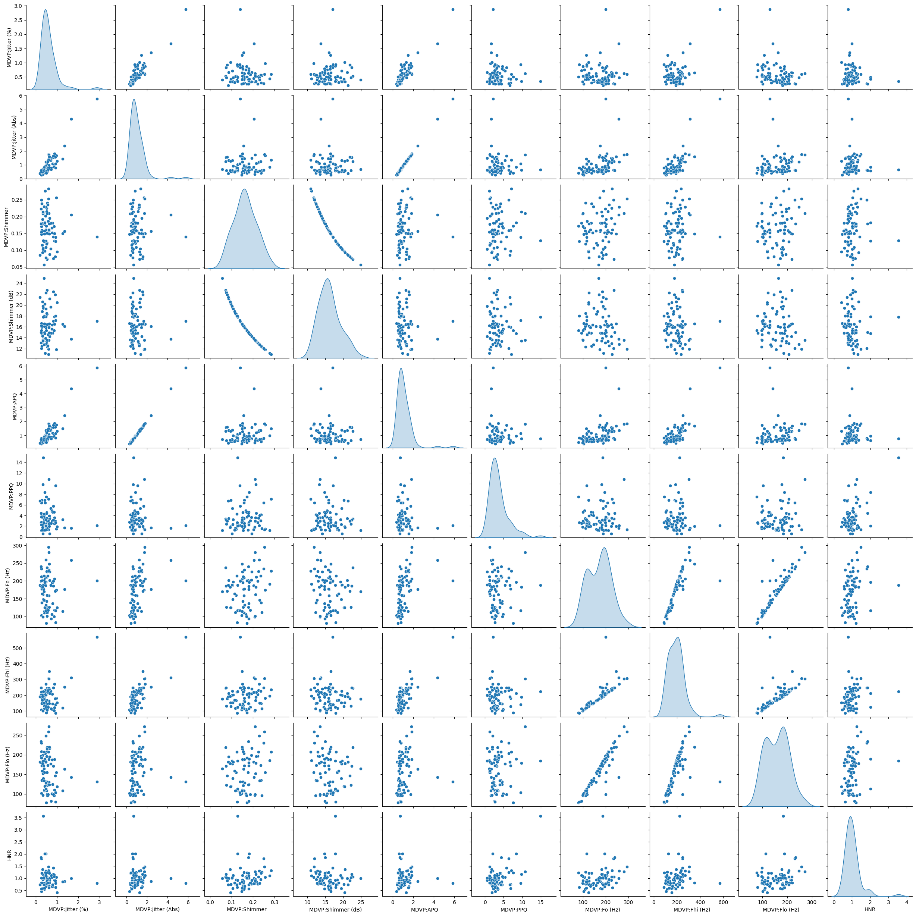
##### Figure 5.1 provides an overview of the collection process for the Parkinson's disease dataset.

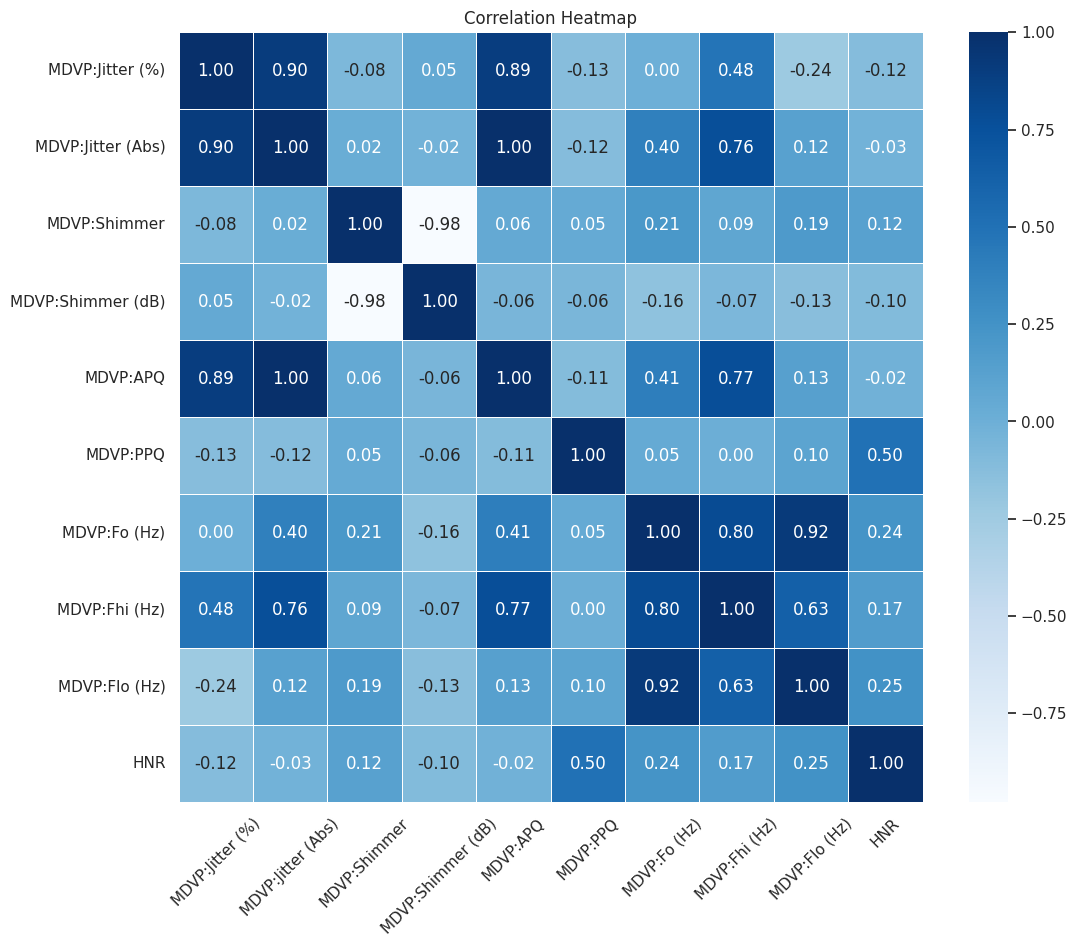
##### 

**Figure 5.2 illustrates the initial data analysis including data head, data info, data shape, data sum, and data describe.**

****

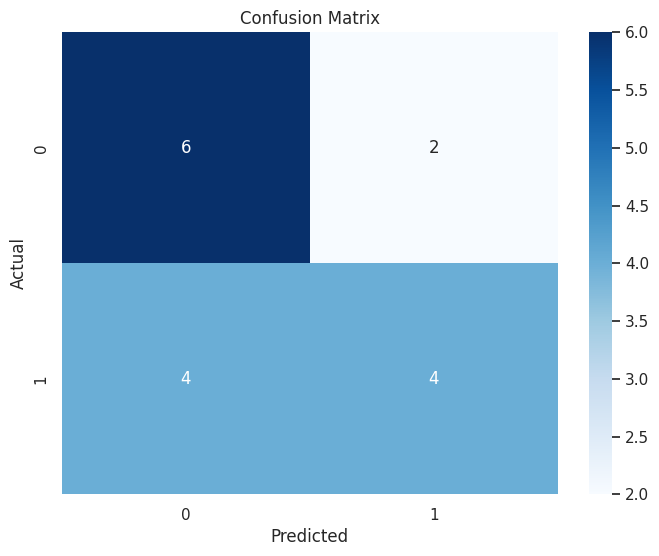
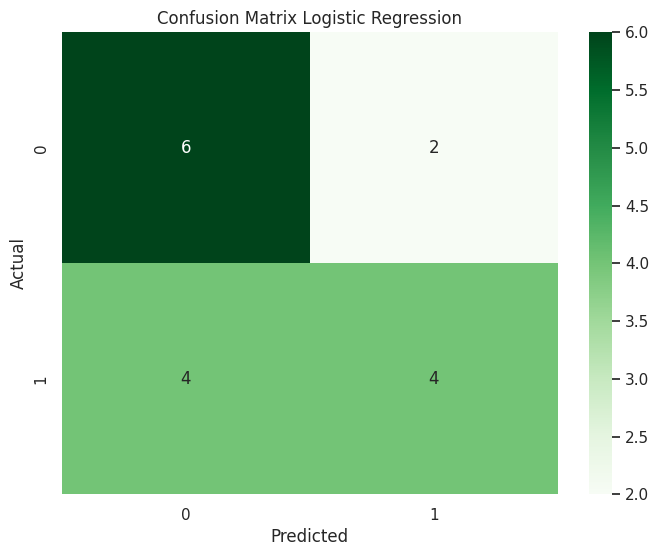
**Fig 5.3: Data Visualization**

****

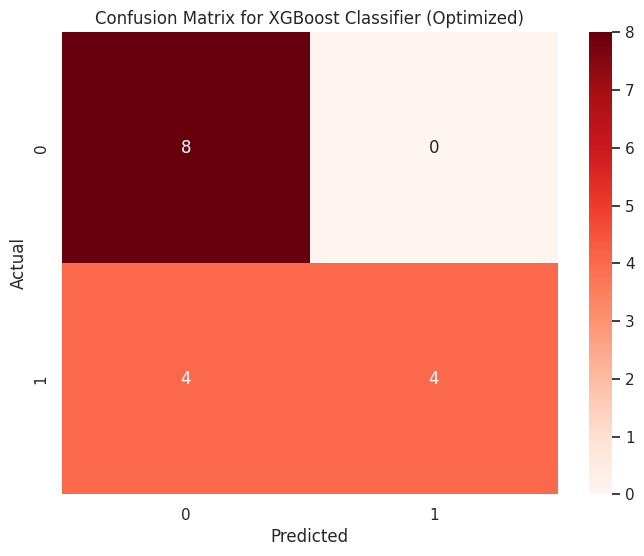
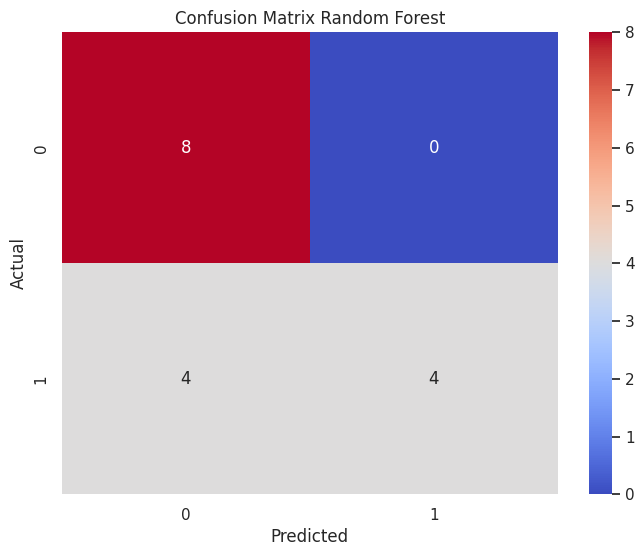


**Figure 5.3.2 displays a correlation heat map depicting the relationships between variables in the dataset.**

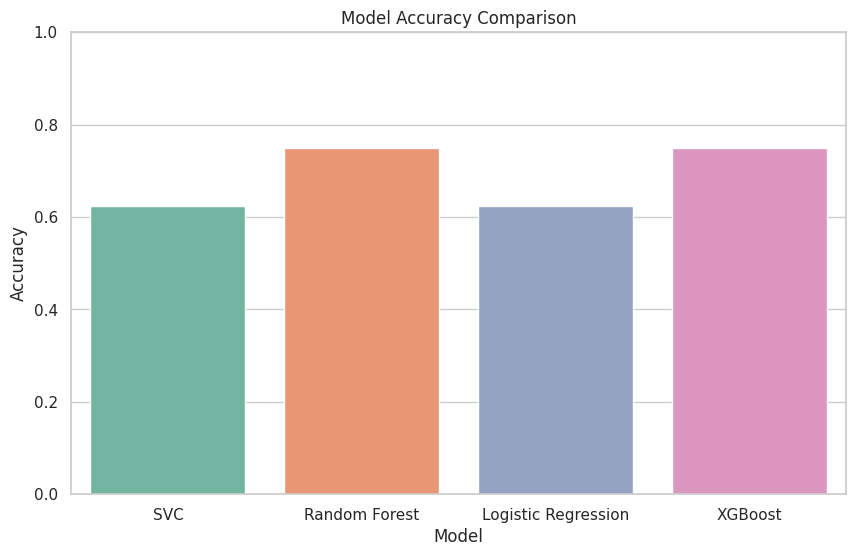
**Figure 5.4 showcases the model evaluation through a confusion matrix.**

** **

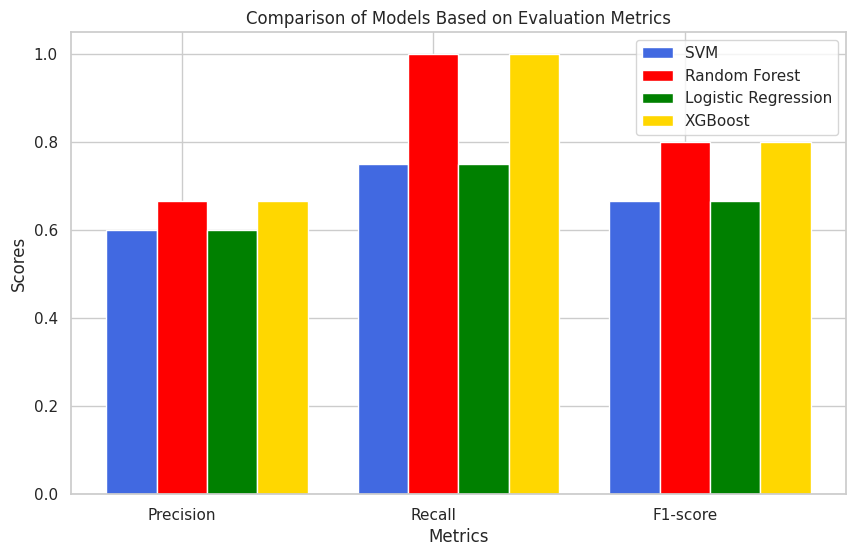
**Fig 5.4.1: Support Vector Machine Fig 5.4.2: Logistic Regression**

** **

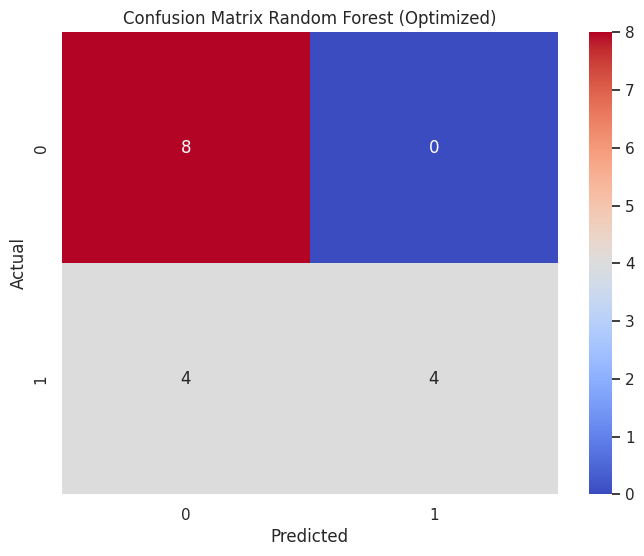
**Fig 5.4.3: XGBoost Classifier Fig 5.4.4: Random Forest**

****

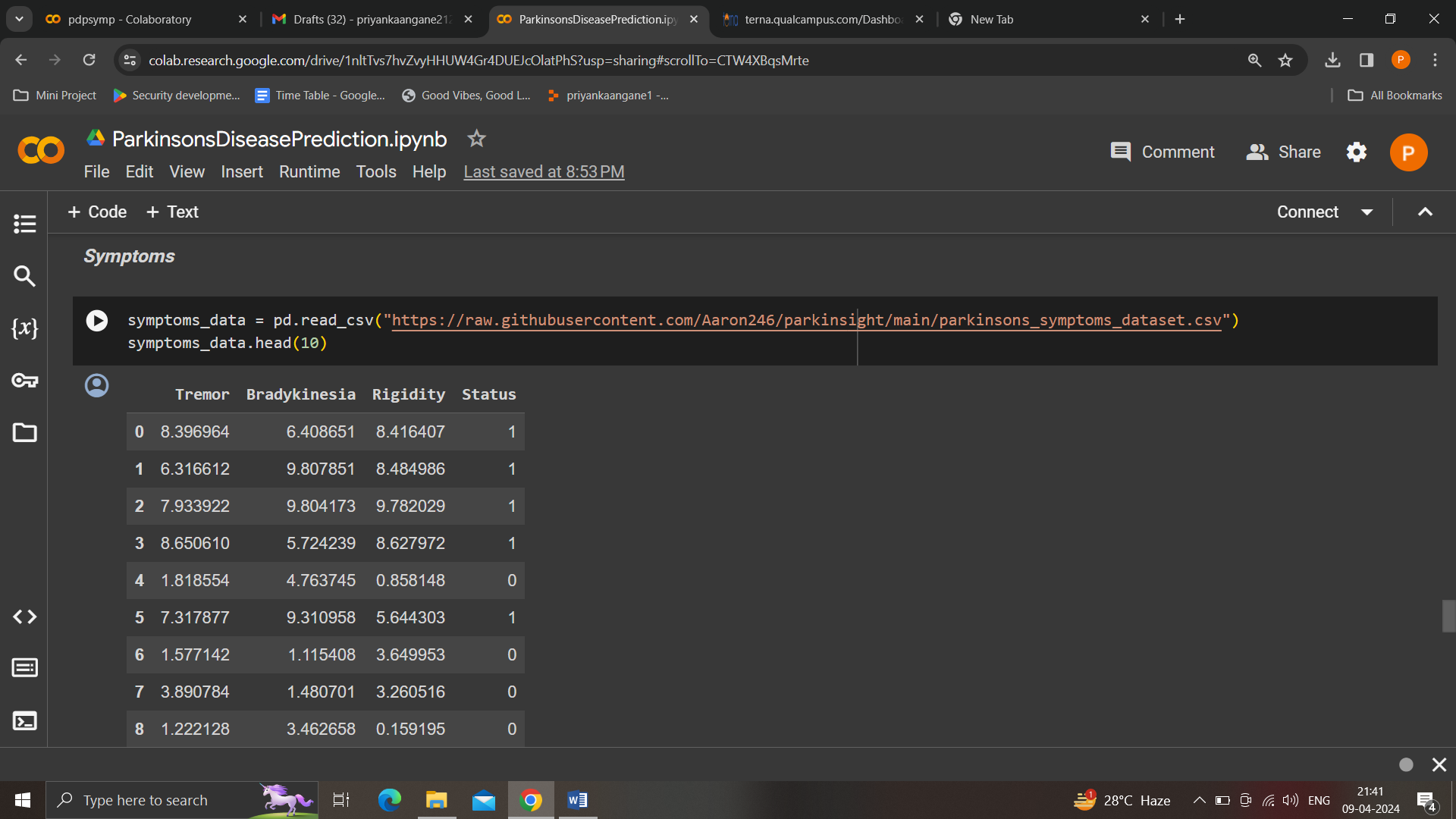
**Figure 5.5 compares the accuracy of different models including Random Forest, Support Vector Machine (SVM), XGBoost, and Logistic Regression.**

****

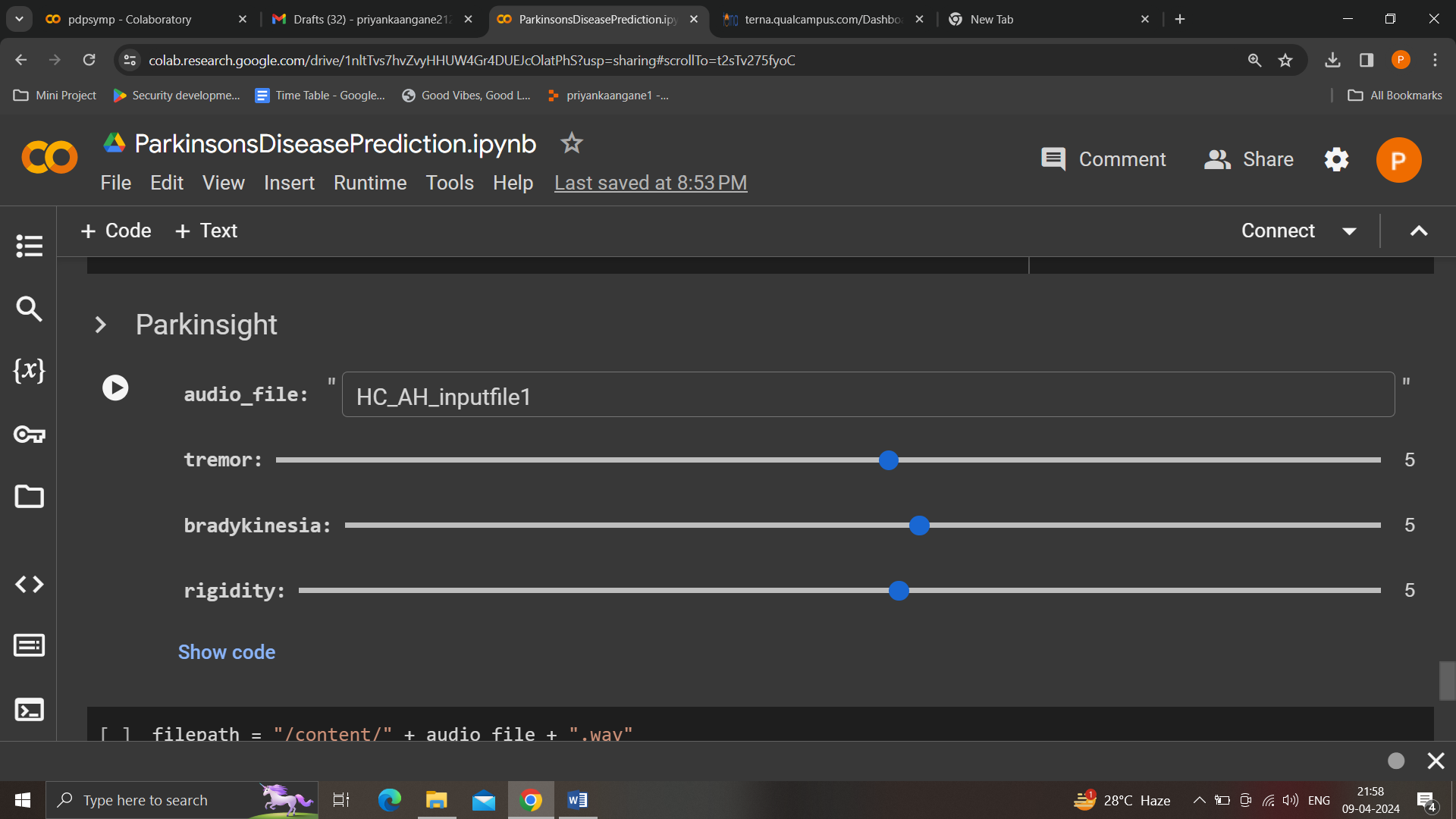
**Figure 5.5.1 illustrates a comparison of models based on various evaluation metrics such as precision, recall, F1-score, and accuracy.**



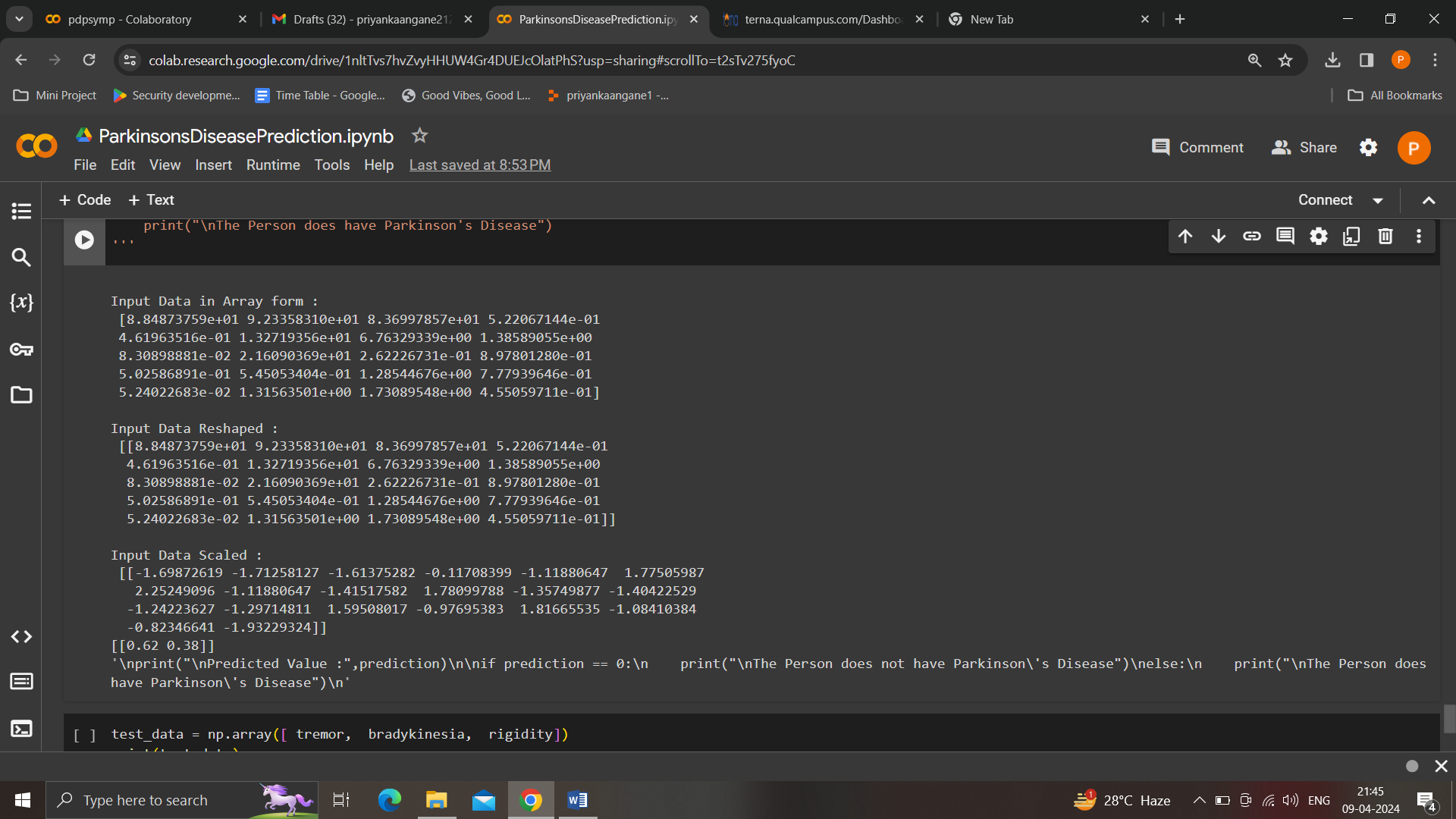
**Figure 5.6 represents the Random Forest model after hyper parameter tuning for improved performance.**



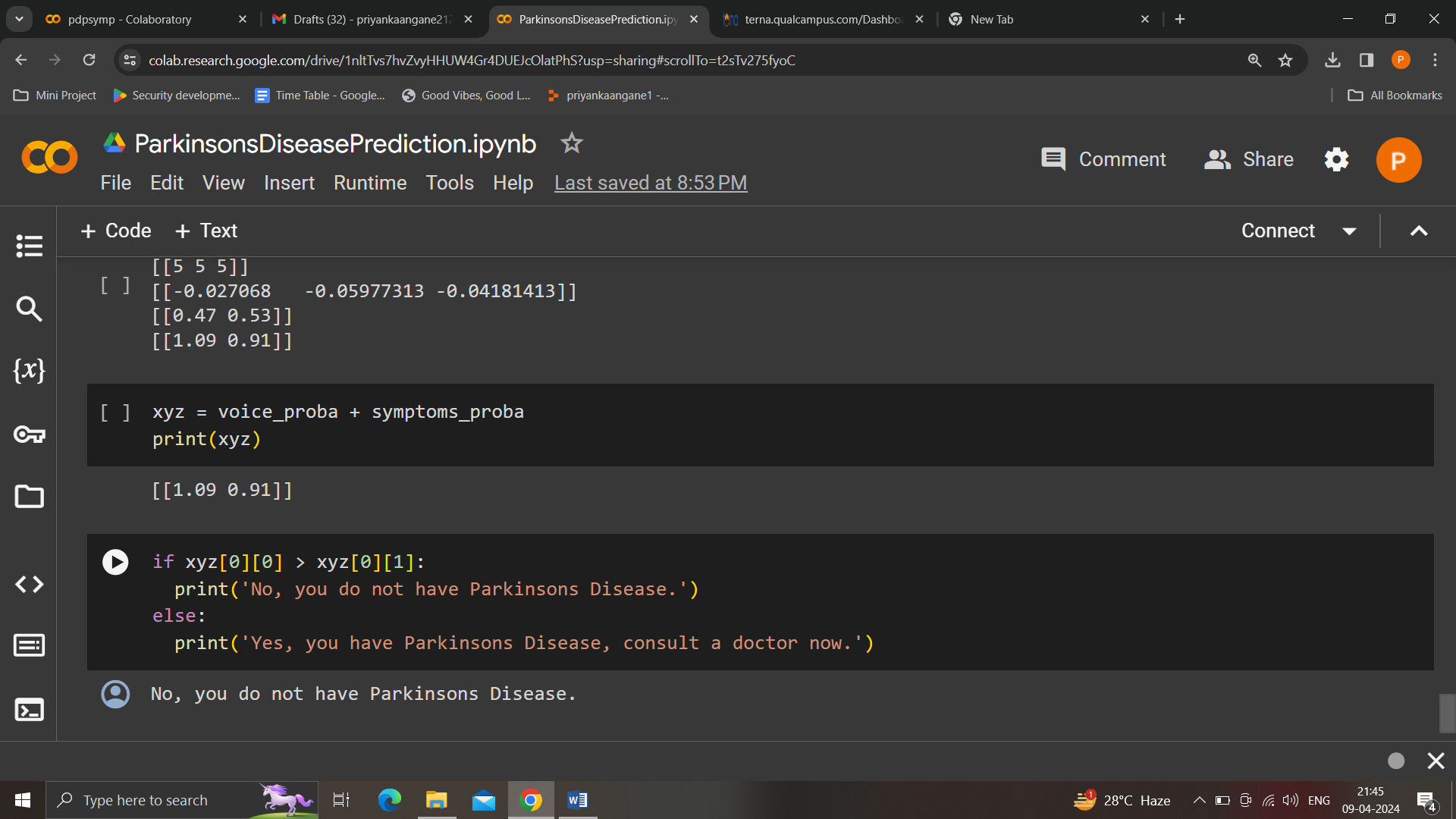
**Figure 5.7 presents the Symptoms Dataset, likely showcasing its structure, attributes, or summary statistics.**



**Figure 5.8 depicts the imported audio file used for voice-based Parkinson's disease prediction, a crucial element in the analysis process.**



**Figure 5.9 showcases the preprocessed input features derived from the audio file, essential for voice-based Parkinson's disease prediction.**



**Figure 5.10 displays the integrated prediction results from voice analysis and symptom data.**

## Chapter 6

**Conclusion and Future Scope**

##### Conclusion

In our project, titled "ParkinSight: Parkinson’s Disease Prediction Based on Voice & Symptoms Profiling," we embarked on a rigorous evaluation process to predict Parkinson's disease utilizing both voice and symptoms data. Initially, we meticulously evaluated various machine learning models using the voice dataset features to discern their predictive capabilities. Following this assessment, the Random Forest model emerged as the most promising candidate due to its superior accuracy in predicting Parkinson's disease.

Subsequently, we integrated the symptoms dataset into our predictive framework to further enhance the accuracy of our predictions. By combining the predictive power of the Random Forest model with the insights derived from symptoms data, we aimed to create a more comprehensive and robust predictive model for Parkinson's disease.

Through this integration, we were able to harness the complementary strengths of both voice-based features and symptoms data, resulting in a more accurate and reliable prediction of Parkinson's disease. Our approach allowed for a holistic assessment of the disease, considering multiple facets of the patient's condition to yield more informed and precise diagnostic outcomes.

##### Future Scope

##### 

##### The "ParkinSight" project offers promising avenues for future exploration. Potential directions include refining the predictive model by incorporating additional data sources and advanced machine learning techniques. Additionally, validating and deploying the system in real-world clinical settings and exploring mobile or wearable applications for continuous monitoring are important areas for future research. Overall, "ParkinSight" sets the stage for ongoing efforts to enhance early diagnosis and management of Parkinson's disease through innovative machine learning approaches and multidimensional data integration.

## References

1. Diogo Braga, Ana M. Madureira, Luis Coelho, Reuel Ajith, “Automatic detection of Parkinson’s disease based on acoustic analysis of speech”, Engineering Applications of Artificial Intelligence, Volume 77, 2019, Pages 148-158, ISSN 0952-1976, Elsevier.
2. Iqra Nissar1,, Danish Raza Rizvi1, Sarfaraz Masood1 and Aqib Nazir Mir, “Voice-Based Detection of Parkinson’s Disease through Ensemble Machine Learning Approach: A Performance Study”, Department of Computer Engineering, Jamia Millia Islamia, New Delhi-110025
3. Azian Abdullah,Nurul Norazman, Wan Ahmad, Saidatul Awang, Foong Jian, “Detection of Parkinson’s Disease (PD) Based On Speech Recordings using Machine Learning Techniques”, IEEE, 2020 Internation Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)
4. Ismail Cantürk, Fethullah Karabiber, “A Machine Learning System for the Diagnosis of Parkinson’s Disease from Speech Signals and Its Application to Multiple Speech Signal Type”, Springer, Arab J Sci Eng, DOI 10.1007/s13369-016-2206-3,King Fahd University of Petroleum & Minerals 2016
5. L. Brabenec, J. Mekyska, Z. Galaz, Irena Rektorova, “Speech disorders in Parkinson’s disease: early diagnostics and effects of medication and brain stimulation”, Springer, J Neural Transm, DOI 10.1007/s00702-017-1676-0, Springer-Verlag Wien 2017
6. Daniel Alonso, Guillermo Morales, Agustin Arribas, Carlos Carrascosa, Andres Rodellar, Pedro Vilda, “MonParLoc: A Speech-Based System for Parkinson’s Disease Analysis and Monitoring”, IEEE, Digital Object Identifier 10.1109/ACCESS.2020.3031646, date of current version October 27, 2020
7. Alice Rueda and Sridhar Krishnan, “Feature Analysis of Dysphonia Speech for Monitoring Parkinson's Disease”, IEEE, Department of Electrical & Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, 978-1-5090-2809-2/17/$31.00 ©2017 IEEE
8. Sura Mahmood Abdullah, Thekra Abbas, Munzir Hubiba Bashir, Ishfaq Ahmad Khaja, Musheer Ahmad, Naglaa f. Soliman, Walid El-Shafai, “Deep Transfer Learning Based Parkinson’s Disease Detection Using Optimized Feature Selection”, Department of Computer Sciences, University of Technology, Baghdad 10066, Iraq
9. Khaled M. Alalayah, Ebrahim Mogammed Senan, “Automatic and Early Detection of Parkinson’s Disease by Analyzing Acoustic Signals Using Classification Algorithms Based on Recursive Feature Elimination Method”, Department of Computer Science, Faculty of Science and Arts, Najran University, Sharurah 68341, Saudi Arabia
10. Basil K Varghese, Geraldine Bessie Amali D, Uma Devi K S, “Prediction of Parkinson’s Disease using Machine Learning Techniques on Speech dataset”, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Ind
11. Sandhiya S, Dr.Ashok.S, Mr.G.Vishnu ,Vardhan Rao, Dr.Prabhu V, “Parkinson's Disease Prediction Using Machine Learning Algorithm”, PG Scholar, Department of Electronics and Communication Engineering, Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering Chennai, India.
12. Jefferson S. Almeida, Pedro P. Reboucas Filho, Tiago Carneiro, Wei, Robertas Damasevicius, Rytis Maskeliunas, Hugo C. de Albuquerque, “Detecting Parkinson’s Disease with Sustained Phonation and Speech Signals using Machine Learning Techniques”, Pattern Recognition Letters (2019), doi:https://doi.org/10.1016/j.patrec.2019.04.005, Elsevier.
13. S. Almeida et al. [Detecting Parkinson’s disease with sustained phonation and speech signals using machine learning techniques](https://www.sciencedirect.com/science/article/pii/S0167865519301163) Pattern Recognition. Letter.(2019)
14. B. Harel et al. [Variability in fundamental frequency during speech in prodromal and incipient Parkinson’s disease: a longitudinal case study](https://www.sciencedirect.com/science/article/pii/S0278262604001393) Brain Cognition. (2004)
15. P. Sharma et al [Diagnosis of Parkinson’s disease using modified grey wolf optimization](https://www.sciencedirect.com/science/article/pii/S1389041718308726) Cogn. Syst. Res (2019)
16. O. Yaman et al. [Automated Parkinson’s disease recognition based on statistical pooling method using acoustic features](https://www.sciencedirect.com/science/article/pii/S0306987719311028) Med. Hypotheses (2020)
17. L. Ali et al.A multi-model framework for evaluating type of speech samples having complementary information about Parkinson’s disease International Conference on Electrical, Communication, and Computer Engineering(2019)
18. L. Ali et al. Automated detection of Parkinson’s disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network IEEE J. Transl. Eng. Health Med. (2019)

### [Optimizing Support Vector Machine Performance for Parkinson's disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction](https://doi.org/10.20473/jisebi.10.1.38-50) 2024, Journal of Information Systems Engineering and Business Intelligence