SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act, 1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

THANJAVUR | KUMBAKONAM | CHENNAI

Early detection of Parkinson's disease using machine learning

July - November 2024

Submitted By,

Haripriyan K A,

B.Tech, Computer Science and Business Systems

(125018025)

Submitted To,

Swetha Varadarajan

1. Abstract

This paper explores a machine learning-based approach for early Parkinson's disease (PD) detection using voice data, addressing the mobility and speech challenges faced by PD patients. By analyzing voice recordings, the study compares four machine learning models: Support Vector Machine (SVM), Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression. Among these, the Random Forest classifier emerges as the most effective, achieving 91.83% accuracy.

The research underscores the potential of telemedicine in diagnosing PD, enabling remote and accessible healthcare solutions for aging populations. The findings highlight the use of voice data as a non-invasive biomarker, with the Random Forest model showing high sensitivity, making it a viable tool for remote PD screening. This approach can significantly improve early detection and management of PD, aiming to offer a better quality of life for patients.

2. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects motor functions and impairs quality of life for millions, especially older adults. Characterized by symptoms like tremors, muscle rigidity, and slowed movement, PD has no known cure, making early detection crucial for effective disease management. Timely diagnosis allows for early intervention strategies that can slow disease progression, improve patient outcomes, and reduce healthcare burdens. Traditional diagnostic methods often involve physical assessments and invasive procedures, which can be time-consuming, expensive, and limited in accessibility, especially for patients in remote areas.

To address these challenges, recent research has turned to non-invasive diagnostic methods using machine learning (ML). ML models, particularly in telemedicine, offer a promising solution by analyzing subtle indicators, such as vocal changes, that are symptomatic of PD. Vocal characteristics—such as pitch variability, tonal changes, and noise ratios—are influenced by motor control functions that deteriorate in PD. This project focuses on using ML models to classify PD through audio data analysis, making it feasible to remotely detect the disease in its early stages. Leveraging vocal biomarkers provides a scalable, accessible, and non-invasive approach to diagnosis, potentially revolutionizing the way PD is identified and managed worldwide.

2.1. Project Objectives

The main objective of this project is to create an effective ML-based diagnostic tool for early Parkinson's detection using audio data. Specific goals include:

1. **Identifying Key Features**: Determine which vocal attributes (e.g., jitter, shimmer, noise ratios) are most indicative of PD.
2. **Building a Classification Model**: Train models like Random Forest and SVM to classify individuals as PD or non-PD based on audio data.
3. **Evaluating Model Performance**: Compare models using metrics like accuracy, sensitivity, and ROC-AUC to identify the best-performing model.
4. **Advancing Telemedicine**: Establish a framework for non-invasive, remote diagnostics for PD through audio analysis.

2.2. Problem Formulation

The project formulates PD detection as a binary classification problem where the dependent variable represents the presence or absence of PD. By analyzing audio features as independent variables, the objective is to predict whether an individual has PD. This involves preprocessing vocal attributes, selecting relevant features, and employing ML models to build a reliable classifier. The goal is to develop a telemedicine-compatible model that allows for remote, accessible diagnostics, supporting early detection efforts and aiding healthcare providers in managing PD more effectively.

3. Dataset

3.1 Key Attributes

The dataset includes 22 vocal measurements from PD patients and healthy individuals, capturing differences in pitch, tone, and noise that indicate motor control issues:

- **Fundamental Frequency (Fo, Fhi, Flo)**: Baseline vocal frequencies.
- **Jitter and Shimmer**: Variability in pitch and volume, respectively.
- **Noise Ratios (NHR, HNR)**: Indicate vocal robustness and background noise levels.

3.2 Data Source and Preprocessing

Collected from PPMI and UCI machine learning repositories, the dataset underwent preprocessing, including scaling and normalization, to enhance model accuracy. Missing values were imputed, and categorical attributes were encoded for analysis.

4. Related Work

Research into Parkinson's disease detection using machine learning has primarily focused on image-based diagnostics, such as MRI scans, and data from physical movement (e.g., gait analysis) to identify early symptoms. However, audio biomarkers have recently gained attention for their accessibility and cost-effectiveness. Bilal et al. (2022) utilized genetic data with Support Vector Machine (SVM) models to achieve an 88.9% accuracy in PD detection, demonstrating the potential for data-driven approaches. Studies by Raundale et al. (2021) furthered this research using Random Forest and deep learning to classify keystroke dynamics from telemonitoring datasets, highlighting ML's effectiveness in handling non-linear patterns in medical data.

Voice-based models for PD have shown promise due to the vocal impairments associated with the disease. Cordella et al. (2021) classified PD patients by analyzing vocal signal data with MATLAB models, achieving accurate results but with high computational costs. Another study by Ali et al. (2022) applied ensemble deep learning on phonation data, achieving strong performance but lacking feature selection optimization. This project builds on these studies by employing open-source Python models, focusing on Random Forest and SVM for reliable, memory-efficient, and accessible telemedicine applications in PD diagnostics.

4.1.Reference

1. - Bilal, A., Moradi, S., Tapak, L., & Afshar, S. (2022). "Identification of Novel Noninvasive Diagnostic Biomarkers in Parkinson's Disease Using Support Vector Machine." *BioMed Research International*.
2. - Raundale, P., Thosar, C., & Rane, S. (2021). "Prediction of Parkinson's Disease Using Machine Learning and Deep Learning Algorithms." *International Conference for Emerging Technology (INCET)*.
3. - Cordella, F., Paffi, A., & Pallotti, A. (2021). "Classification-Based Screening of Parkinson's Disease Patients Through Voice Signal." *IEEE International Symposium on Medical Measurements and Applications*.
4. - Ali, L., Chakraborty, C., & He, Z. (2022). "Ensemble Approach for Parkinson's Disease Detection Using Phonation Data." *Neural Computing and Applications*.

5. Background

5.1.Parkinson's Disease and Vocal Biomarkers

Parkinson’s disease disrupts the neural pathways responsible for motor function, manifesting in vocal changes detectable in early stages. Vocal biomarkers, particularly frequency modulations, offer a promising avenue for remote PD diagnostics.

5.2.Machine Learning in Telemedicine

ML models such as Random Forest and SVM can process high-dimensional data, learning subtle distinctions between healthy and PD-affected voices. Random Forest’s ensemble approach aggregates results from multiple decision trees, providing robust classification with minimal overfitting.

5.3.Preprocessing Techniques Used

The preprocessing stage is essential for preparing vocal data to ensure accuracy and reliability in Parkinson's disease (PD) classification models. This process involved several key steps:

1. **Handling Missing Values**: Missing values in the dataset were managed to maintain data integrity. Numerical missing values were replaced with median values, and categorical data, if any, was imputed with the most frequent category. This approach ensures a complete dataset, preventing data loss that could impact model training.
2. **Feature Scaling**: To normalize the range of features, **StandardScaler** was applied, transforming each feature to have a mean of 0 and a standard deviation of 1. This scaling method minimizes the impact of feature magnitude differences, which is critical for algorithms like SVM and KNN that are sensitive to feature scales.
3. **Principal Component Analysis (PCA)**: To reduce dimensionality and improve model efficiency, PCA was performed on the dataset. This technique retained 95% of the variance while reducing the feature set, thereby simplifying the model and mitigating overfitting risks without sacrificing significant information. Dimensionality reduction was particularly effective in reducing noise and enhancing the model's generalizability.
4. **Train-Test Split**: The dataset was split into training and testing sets, typically in a 75-25 or 70-30 ratio, ensuring that models were trained on the majority of the data and evaluated on a separate subset to assess their performance objectively.
5. **Handling Class Imbalance**: Given that PD datasets may exhibit class imbalance (fewer samples for certain classes), techniques like upsampling or synthetic data generation (e.g., SMOTE) can be applied to balance the classes. This ensures that the models do not favor the majority class, improving sensitivity to the minority class.

These preprocessing steps prepared the vocal data for effective use in machine learning models, optimizing both model performance and computational efficiency.

6. Methodology used:

Experimental Design

The experimental design for this project follows a structured approach to classify Parkinson's disease using machine learning models. The steps include data collection, preprocessing, model selection, training, and evaluation. After preprocessing the vocal dataset, three machine learning models—Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest—were applied to assess their performance in PD classification.

Machine Learning models

1. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm primarily used for classification tasks. It works by finding the optimal hyperplane that separates data points of different classes in a multidimensional space. In the case of Parkinson's disease (PD) detection, the voice data collected from patients includes attributes that may not be linearly separable, so SVM uses a **kernel trick** to map data into a higher-dimensional space where a hyperplane can be more effectively drawn.

- **Kernel Transformation:** For non-linearly separable data, SVM applies kernel functions such as Radial Basis Function (RBF) or polynomial kernels. In this project, the voice data is complex and not easily separable in its raw form, so kernel SVM is used.
- **Support Vectors:** These are the critical elements of the dataset that define the decision boundary. SVM works by choosing the data points (support vectors) closest to the decision boundary.
- **Hyperplane:** The model tries to maximize the margin between support vectors of both classes, aiming to find a boundary that separates Parkinson's and healthy individuals most effectively.
- **Advantages:** SVM performs well with high-dimensional data (such as the 22 attributes in voice data) and is memory-efficient, as only support vectors are retained in the final model.

2. Random Forest

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and combining their results to improve prediction accuracy. For PD detection, each decision tree is built by splitting the dataset based on different attributes of the audio data, such as jitter, shimmer, and MDVP values.

- **Ensemble Learning:** Random Forest uses a collection of decision trees (each trained on a random subset of the data) and aggregates their predictions. This ensures that the

final classification is robust and not prone to overfitting, which can happen with individual trees.

- **Bagging (Bootstrap Aggregation):** Each decision tree is trained on a random subset of the training data (sampled with replacement), which reduces variance in the model and increases the overall stability of predictions.
- **Majority Voting:** In classification, each tree provides a "vote" for one of the classes (PD or healthy), and the final prediction is the class with the most votes. This approach makes Random Forest particularly effective for datasets with a large number of features and noisy data, such as voice recordings.
- **Advantages:** Random Forest is less prone to overfitting due to the ensemble approach, and it handles missing data and outliers effectively.

3. K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based learning algorithm that works by grouping data into clusters based on similarities. It classifies a new data point by considering its proximity to **K nearest neighbors** in the feature space.

- **Distance Metric:** The model calculates the distance between the new data point and existing points in the training set. Common metrics include Euclidean distance or Manhattan distance. In this project, KNN uses features like jitter and shimmer to determine the similarity between patients' voice data.
- **Voting System:** After calculating the distance to its K nearest neighbors, the new point is assigned the class that is most common among its neighbors. For PD detection, the model compares the voice features of new patients with those of Parkinson's and healthy individuals to make predictions.
- **Advantages:** KNN works well for small datasets, like the 195 voice samples in this study. Since it doesn't make assumptions about the underlying data distribution, it can handle complex patterns in the voice data without requiring explicit training.

4. Logistic Regression

Logistic regression is a statistical method for binary classification that predicts the probability of an event occurring based on one or more independent variables. It uses the **logistic function (sigmoid function)** to model the relationship between the dependent variable (Parkinson's or healthy) and the independent variables (voice features).

- **Sigmoid Function:** The output of logistic regression is a probability value between 0 and 1. If the probability is greater than a certain threshold (typically 0.5), the model predicts Parkinson's disease; otherwise, it predicts healthy.
- **Binary Classification:** Logistic regression is ideal for this task because it predicts a binary outcome—whether or not a patient has Parkinson's disease—based on continuous input features.
- **Advantages:** Logistic regression is interpretable, allowing for the understanding of which features (e.g., jitter or shimmer) contribute the most to classifying the disease.

5. Naive Bayes

Naive Bayes is a probabilistic classifier based on **Bayes' Theorem** with the assumption that the features are conditionally independent given the class label. Despite its "naive" assumption of feature independence, it performs well in many real-world problems, especially with high-dimensional data like audio features.

- **Bayes' Theorem:** The model uses Bayes' Theorem to calculate the posterior probability of a class (e.g., Parkinson's or healthy) given the observed features: $P(C|X)=P(X|C)\cdot P(C)P(X)P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}P(C|X)=P(X)P(X|C)\cdot P(C)$ Where C is the class (Parkinson's or healthy) and X is the feature set (voice attributes).
- **Conditional Independence Assumption:** Naive Bayes assumes that all the features (jitter, shimmer, MDVP) are independent of each other, given the class label. This simplifies the computation of probabilities, especially when working with a large number of features.
- **Training and Prediction:** During training, the model calculates the likelihood of each feature given the class label and uses this to calculate the posterior probability for new instances. For each new patient, it predicts the class with the highest posterior probability.
- **Advantages:** Naive Bayes is computationally efficient, requires minimal training data, and performs well even with noisy data. It is particularly suited for tasks where feature independence (or near-independence) can be assumed, such as in high-dimensional datasets like voice recordings.

7. Results

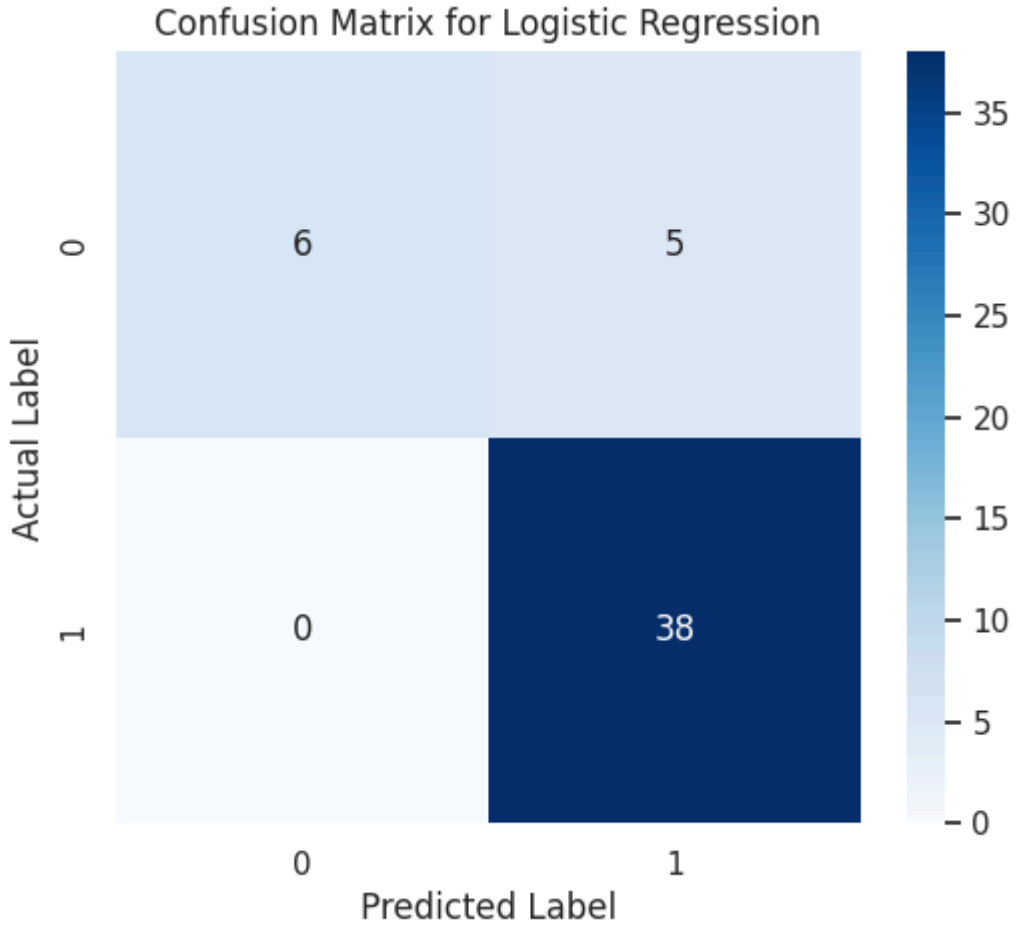
After implementing the machine learning models—Logistic Regression, SVM,random-classifier, naive Bayes, and KNN—on the PCA-transformed dataset, the following results were obtained:

7.1 Results of models without PCA:

7.1.1.Logistic Regression:

- **Accuracy**: ~89%
- **ROC-AUC Score**: 0.88

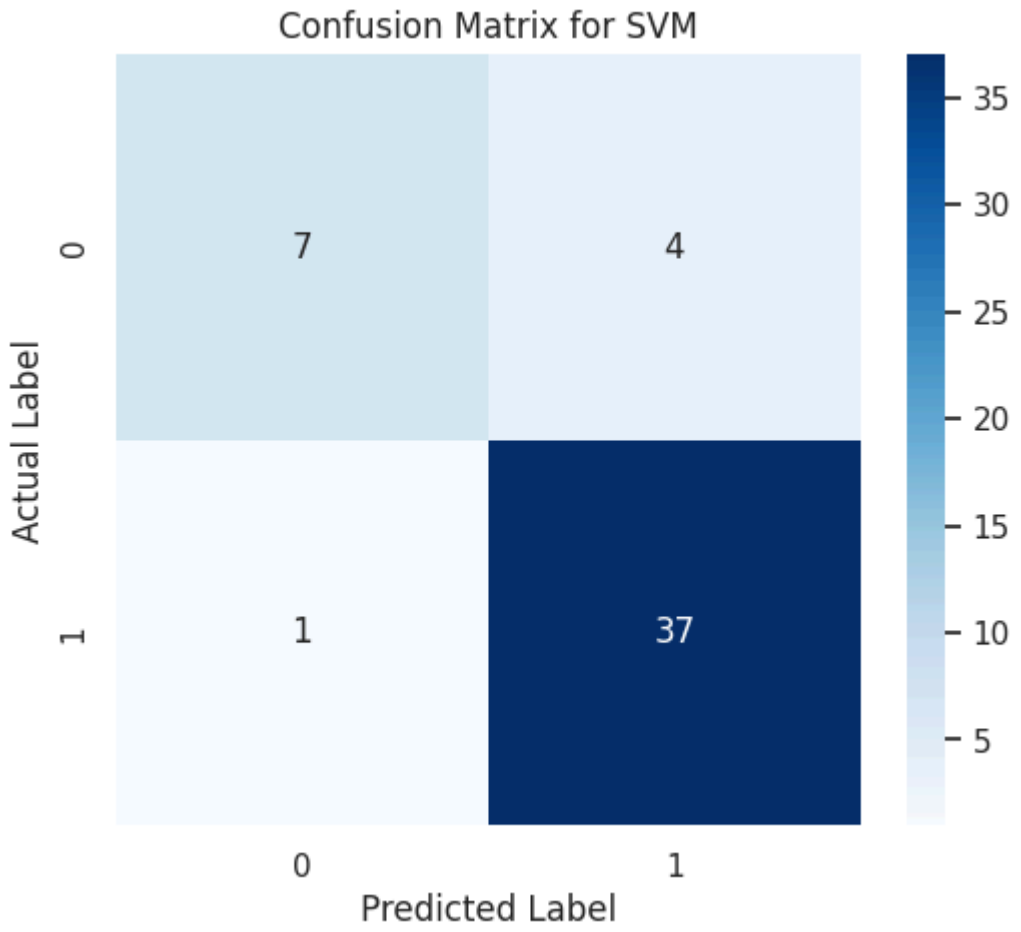
Confusion matrix:



7.1.2.SVM :

- **Accuracy**: ~89%
- **ROC-AUC Score**: 0.87

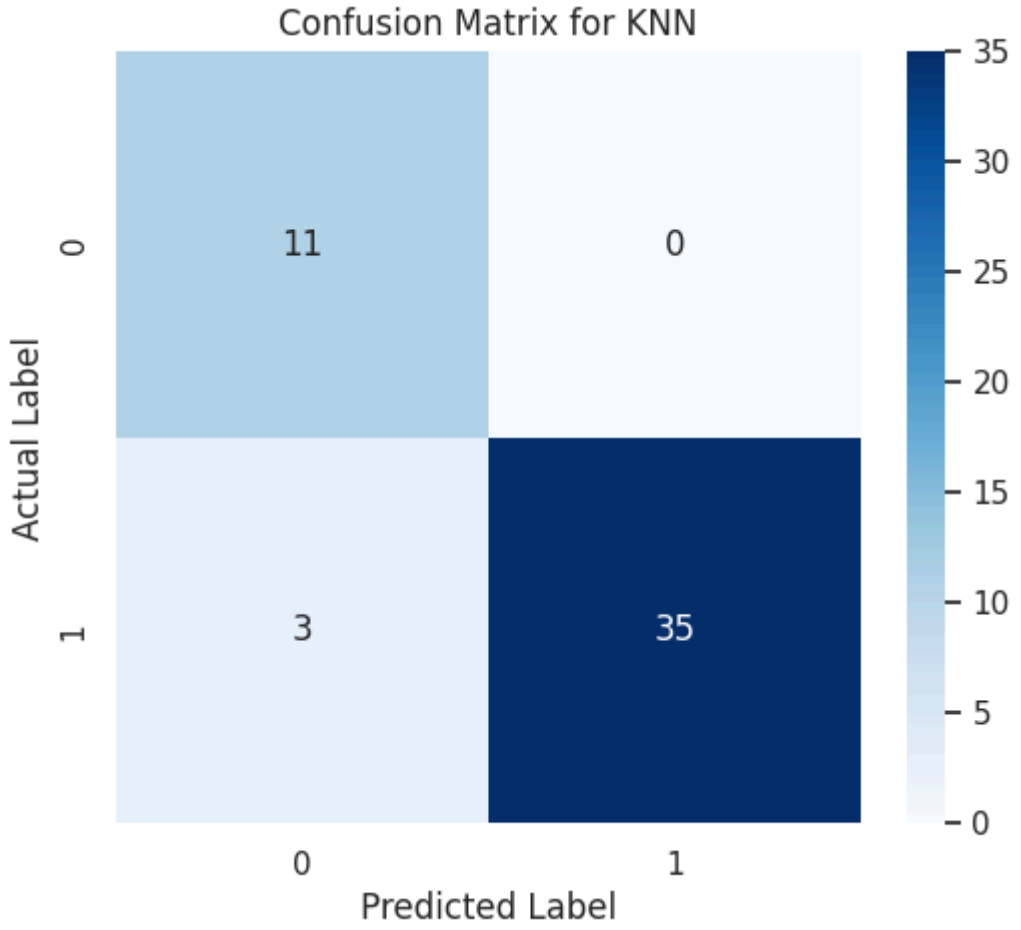
Confusion matrix:



7.1.3.K-Nearest Neighbors (KNN):

- **Accuracy**: ~93%
- **ROC-AUC Score**: 0.98

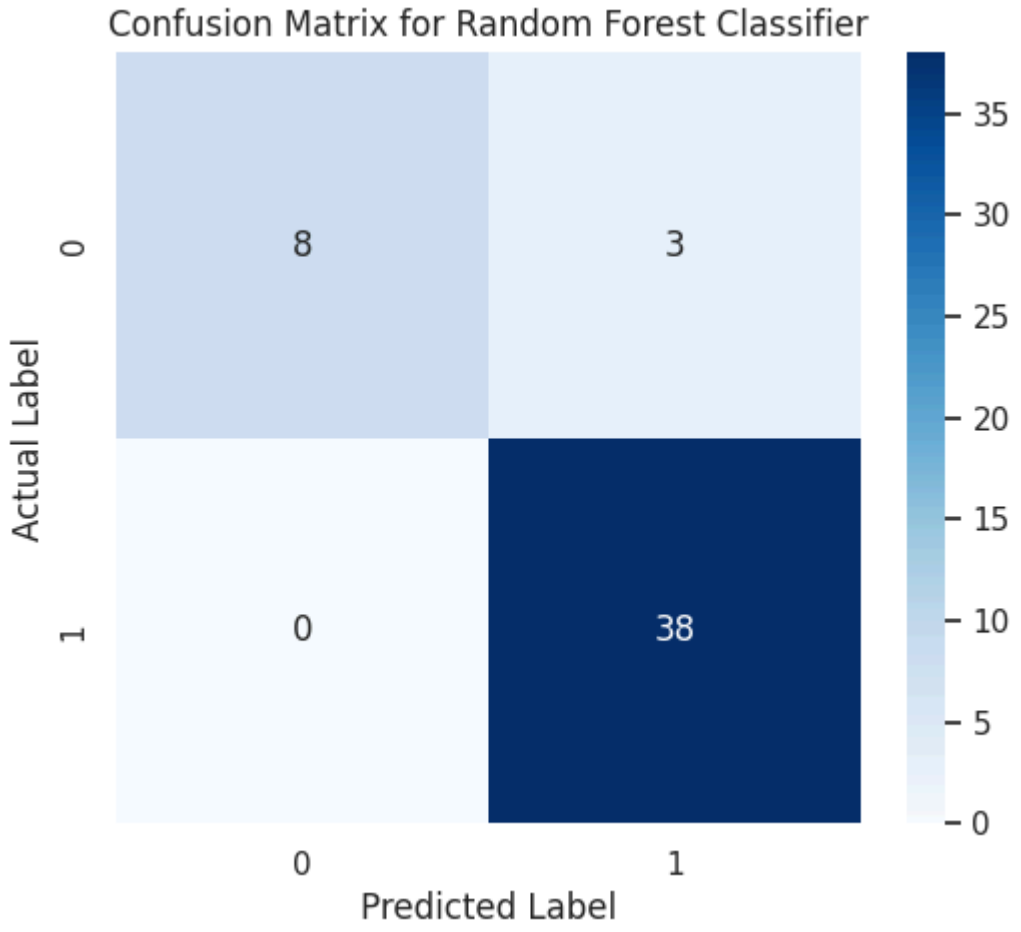
Confusion matrix:



7.1.4.Random Forest Classifier:

- **Accuracy**: ~93%
- **ROC-AUC Score**: 0.95

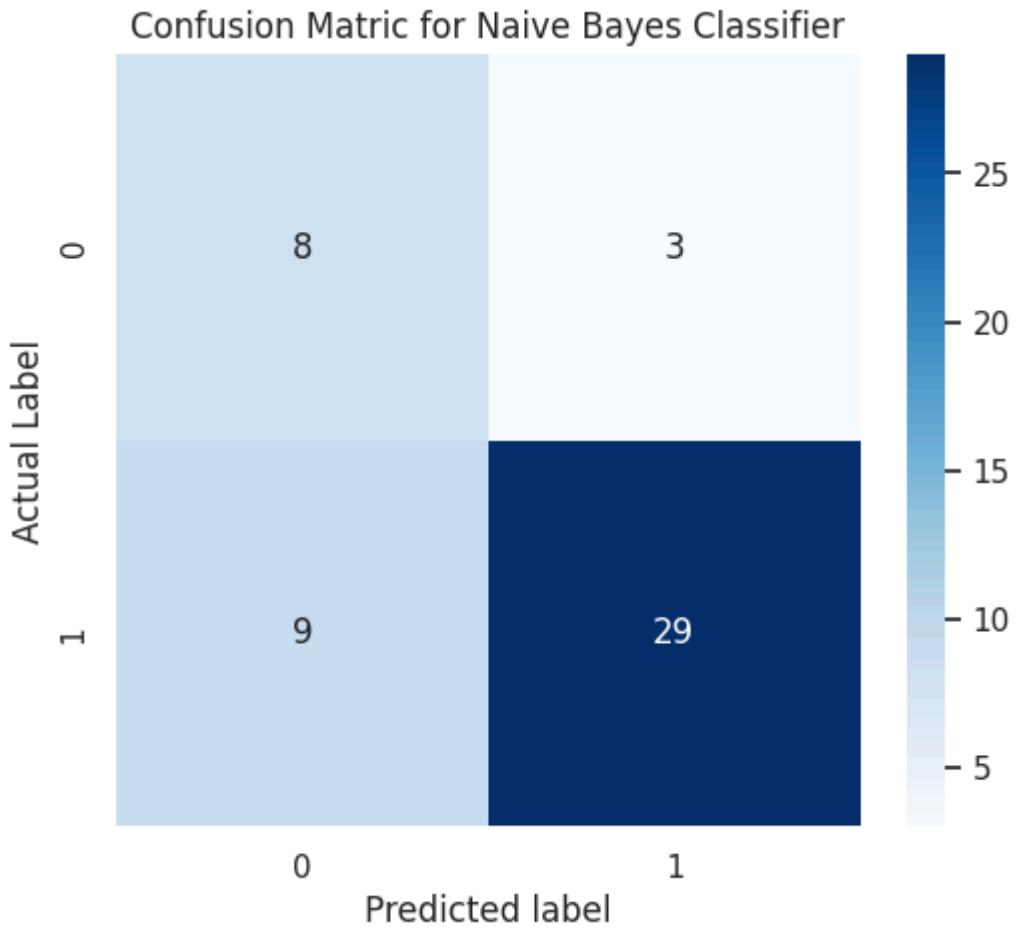
Confusion matrix:



7.1.5.Naive Bayes Classifier:

- **Accuracy**: ~75%
- **ROC-AUC Score**: 0.80

Confusion matrix:

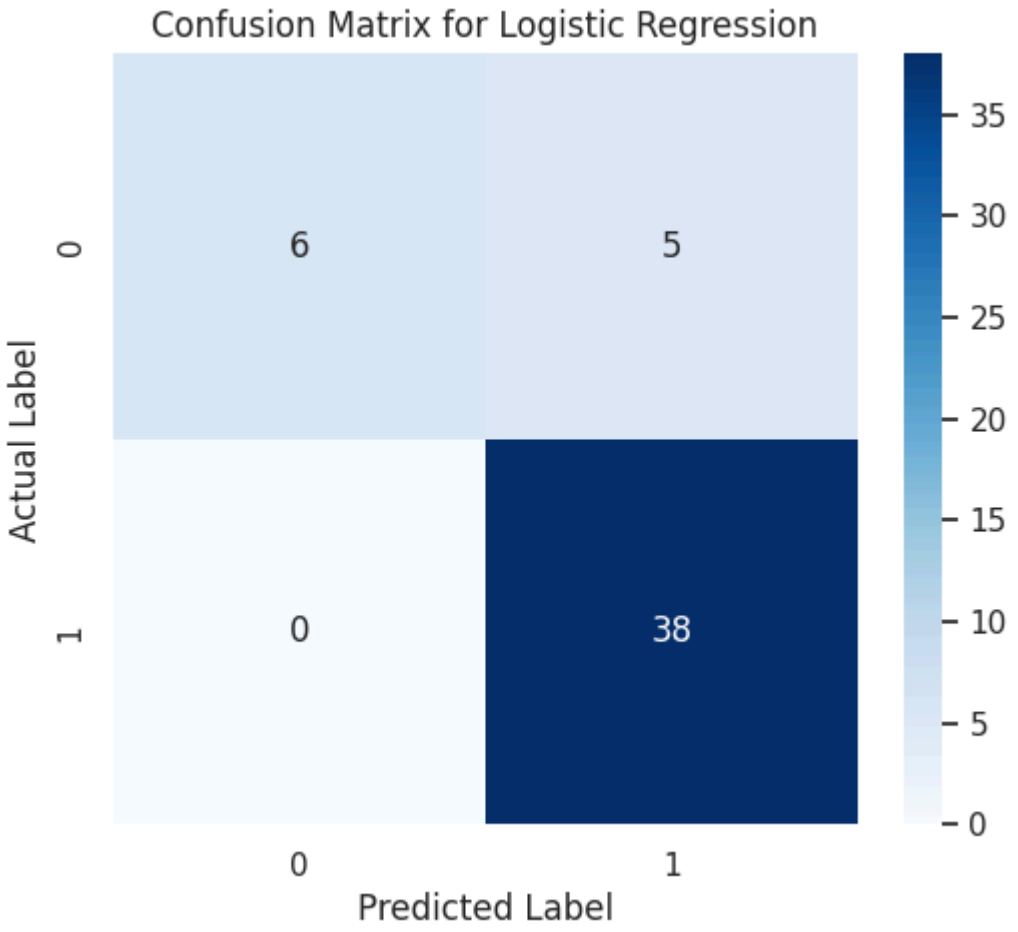


7.2 Results of models with PCA:

7.1.1.Logistic Regression:

- **Accuracy**: ~89%
- **ROC-AUC Score**: 0.93

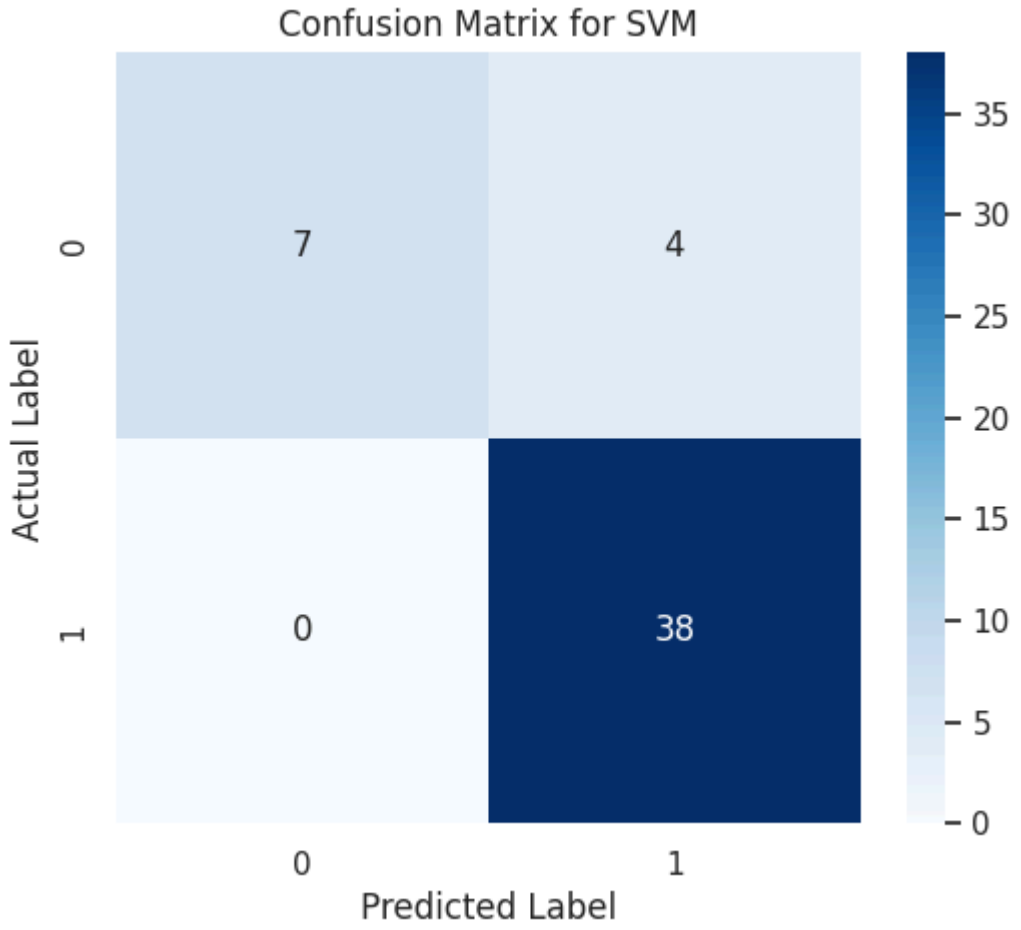
Confusion matrix:



7.1.2.SVM::

- **Accuracy**: ~91%
- **ROC-AUC Score**: 0.91

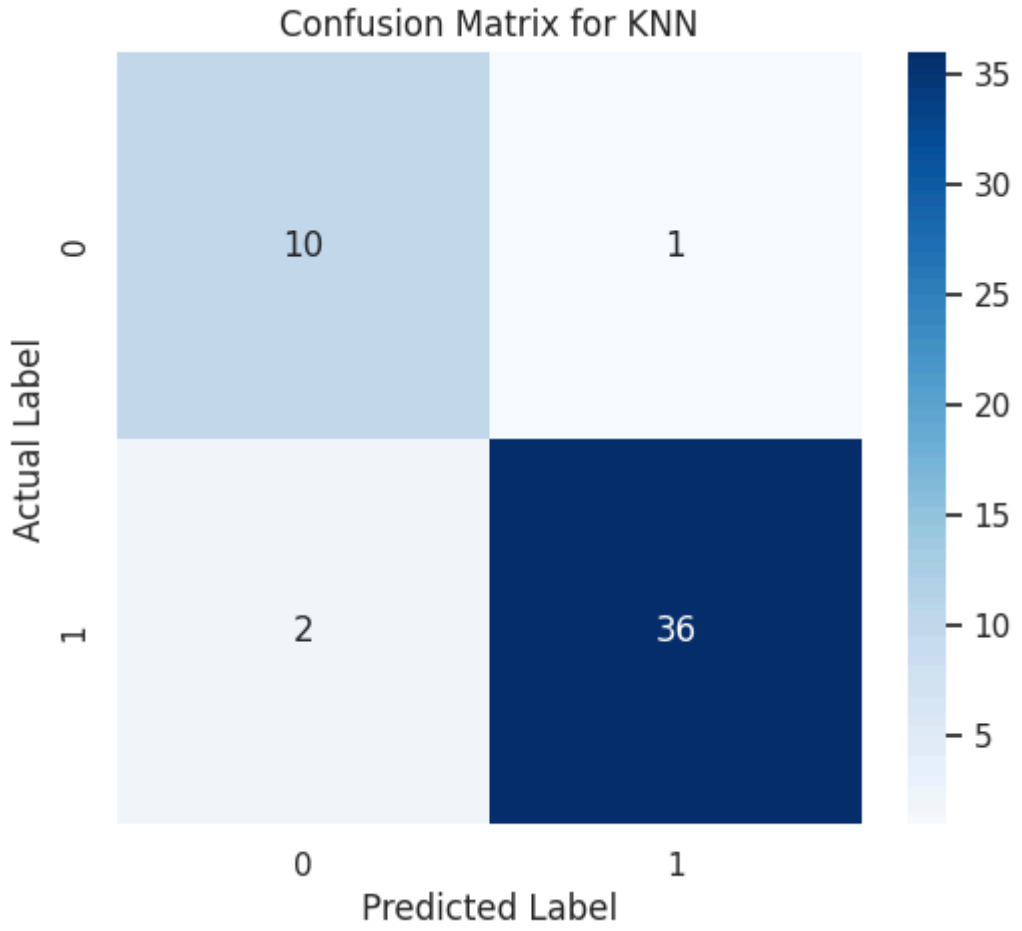
Confusion matrix:



7.1.3.K-Nearest Neighbors (KNN):

- **Accuracy**: ~93%
- **ROC-AUC Score**: 0.93

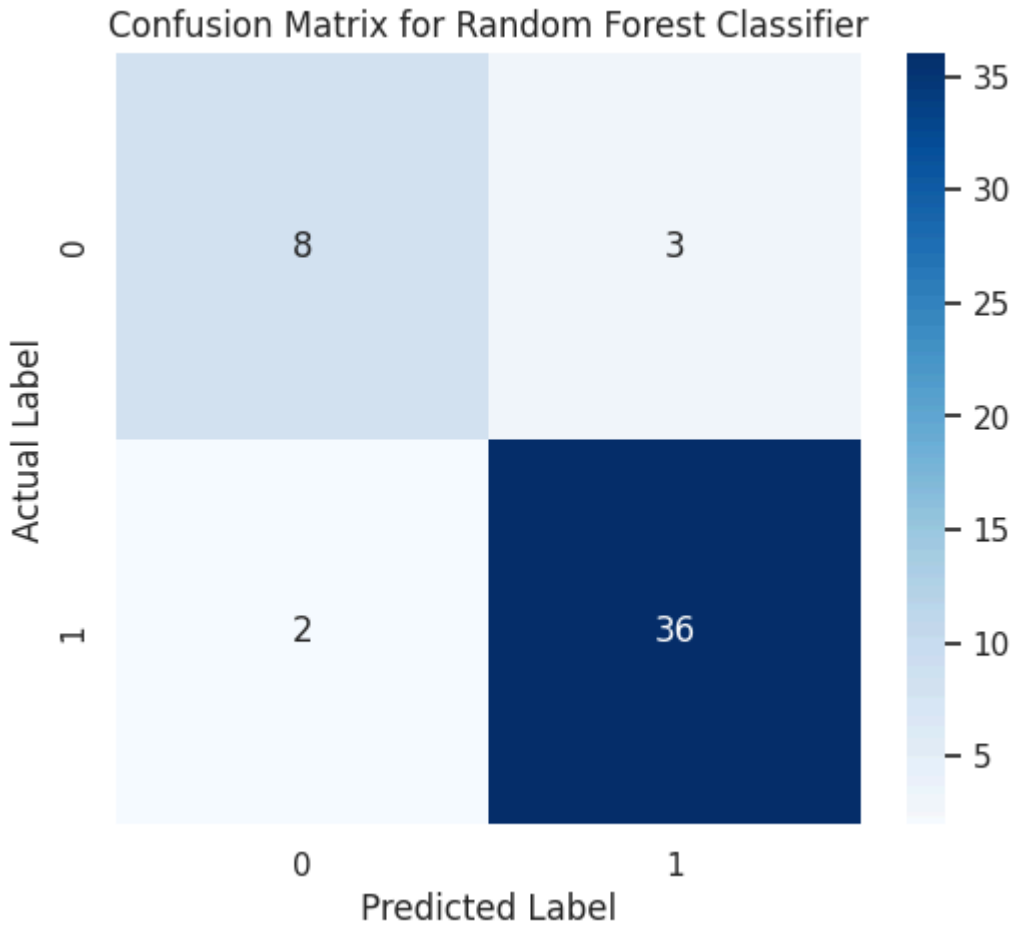
Confusion matrix:



7.1.4.Random Forest Classifier:

- **Accuracy**: ~89%
- **ROC-AUC Score**: 0.96

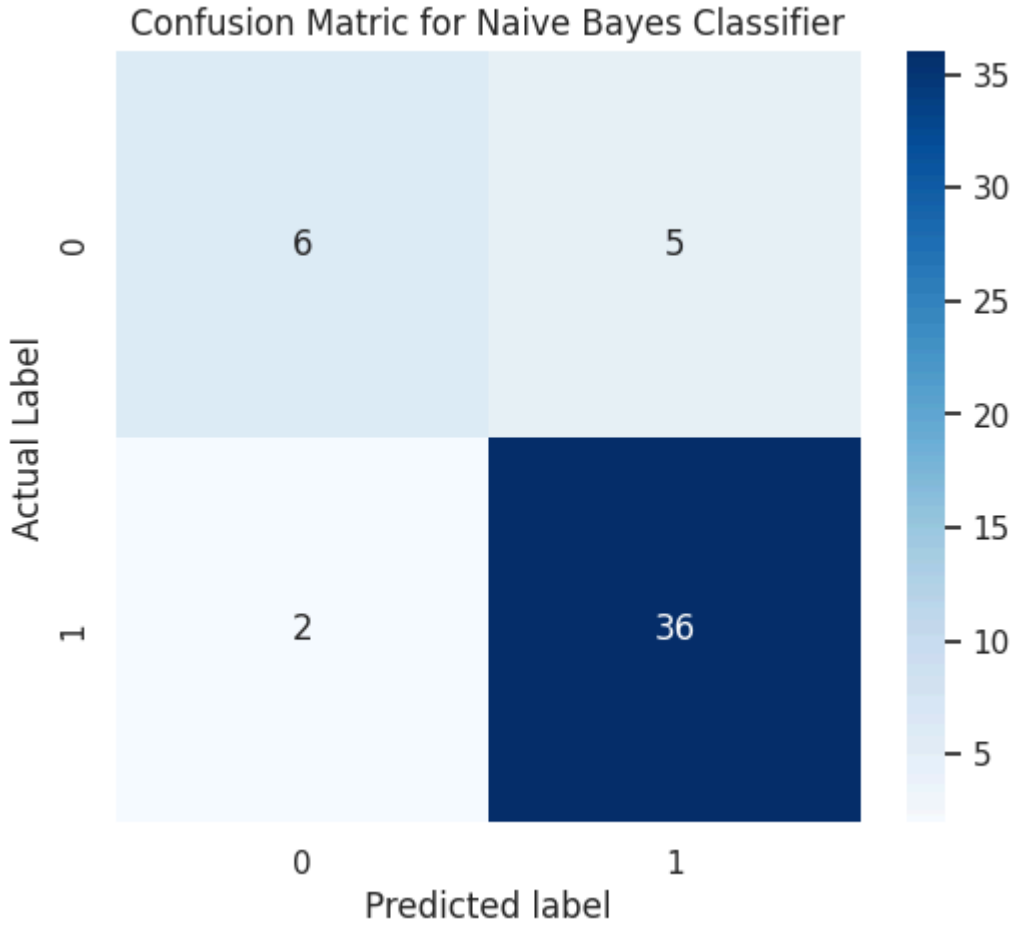
Confusion matrix:



7.1.5.Naive Bayes Classifier:

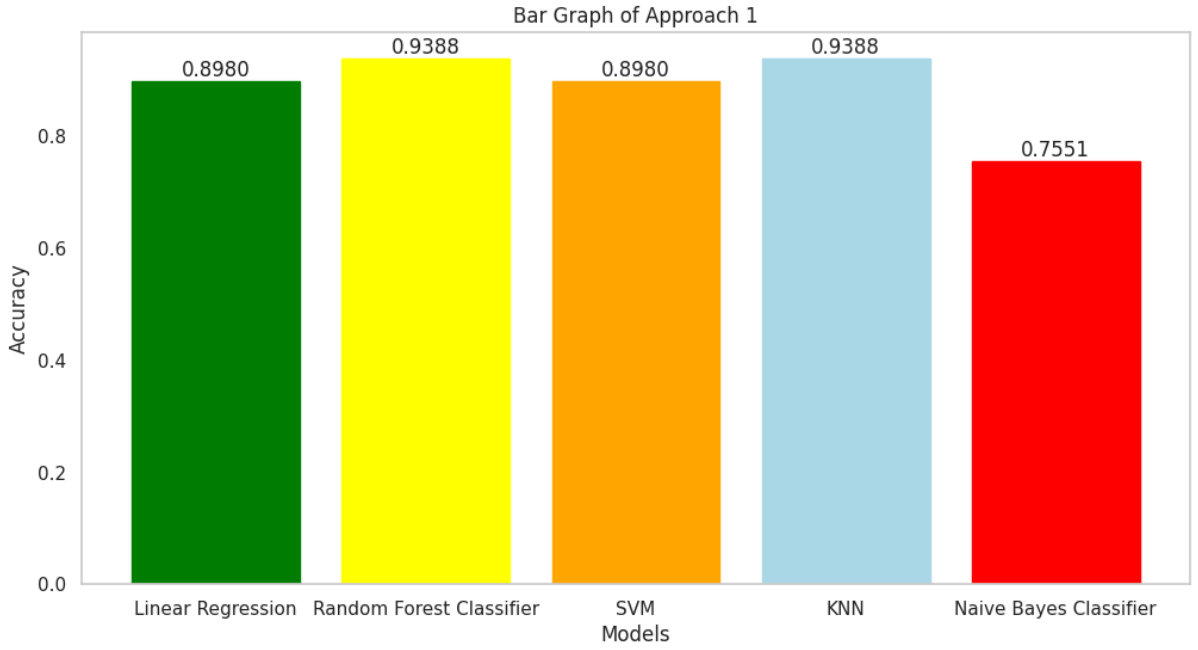
- **Accuracy**: ~85%
- **ROC-AUC Score**: 0.88

Confusion matrix:

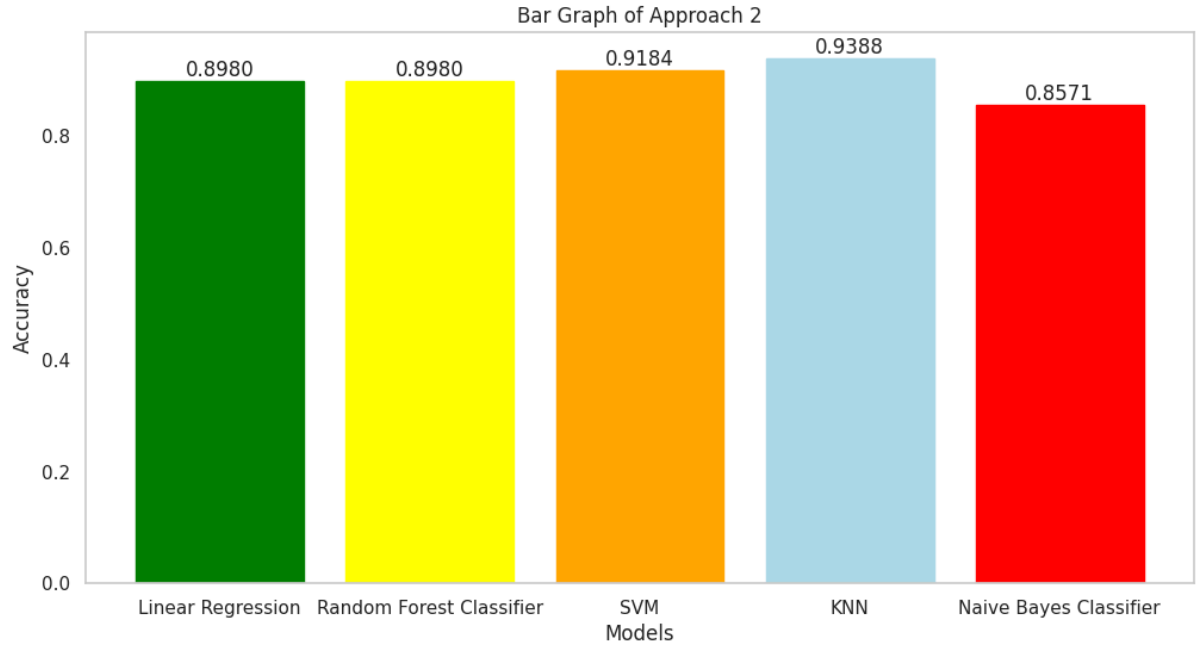


7.3.Model Comparison:

7.3.1. Comparison without PCA:



7.3.2. Comparison with PCA:



8. Learning Outcomes:

8.1 Skills Used:

1. Data preprocessing (handling missing data, balancing datasets, and scaling features).
2. Application of machine learning algorithms (Random Forest, Support Vector Machine, K-Nearest Neighbors, and Logistic Regression).
3. Data visualization and statistical analysis to understand feature importance.
4. Implementation of Principal Component Analysis (PCA) to reduce dimensionality and improve model performance.
5. Evaluation of models using confusion matrix, accuracy, precision, recall, and F1 score.

8.2 Tools Used:

1. Python (for machine learning and data analysis).Libraries: Scikit-learn (for machine learning models), Pandas and NumPy (for data-handling), and Matplotlib/Seaborn (for data visualization).
2. Standard Scaler (for data normalization).
3. PCA (for feature extraction and dimensionality reduction).

8.3 Project Link:

https://github.com/hari03032004/Early-Detection-of-Parkinson-disease.git

9. Conclusion

The comparison of classification algorithms with and without Principal Component Analysis (PCA) reveals interesting insights into their performance. Without PCA, Random Forest and K-Nearest Neighbors (KNN) classifiers demonstrate the highest accuracy rates of 93.8%, indicating their effectiveness in handling the dataset's inherent features. In contrast, Naive Bayes shows significantly lower performance at 75.5%, highlighting its sensitivity to feature correlations. Linear Progression and Support Vector Machine (SVM) models also performed consistently well, achieving an accuracy of 89.8%. However, when PCA is introduced, SVM stands out with a notable increase to 93.8%, suggesting that dimensionality reduction enhances its performance. KNN also performs well, although it experiences a slight decrease to 91.8%. Random Forest, on the other hand, exhibits a decline to 87.7%, raising questions about its reliance on the original feature space. Overall, PCA's impact on performance varies across models, indicating that while it can enhance certain classifiers, it may not universally benefit all algorithms.

9.1 Advantages:

1. **Dimensionality Reduction:** PCA effectively reduces the feature space, simplifying the dataset and making it easier for models to learn.
2. **Enhanced Model Performance:** As seen with SVM, PCA can lead to improved accuracy by eliminating noise and redundant features.
3. **Visualization:** PCA facilitates the visualization of high-dimensional data in a lower-dimensional space, aiding in understanding data distribution.
4. **Reduced Overfitting:** By simplifying the model through fewer features, PCA can help reduce the risk of overfitting, particularly in complex models.

9.2 Limitations:

1. **Loss of Information:** While PCA reduces dimensions, it may also lead to the loss of critical information that could be vital for classification.
2. **Interpretability:** The principal components generated by PCA can be challenging to interpret, as they are linear combinations of original features.
3. **Computational Cost:** Implementing PCA requires additional computational resources, which may not be justifiable for smaller datasets.
4. **Performance Variability:** As observed with Random Forest, PCA does not consistently enhance performance for all models, and its effect can vary based on the dataset characteristics.