A logo with black text

Description automatically generated

CS770 Machine Learning

Final Project

12/03/2023

Submitted by,

Anusha Alishetty

B863T652

1. **Abstract**:

This study endeavours to predict the onset of Parkinson's disease by employing diverse machine learning models and assessing the influence of Principal Component Analysis (PCA) on model accuracy and performance. The dataset 'pd\_speech\_features.csv,' specifically curated for Parkinson's disease analysis, encompasses speech-related features crucial for prediction.

* **Purpose:** The primary objective involves leveraging machine learning models to discern early indications of Parkinson's disease onset and to comprehend how PCA affects the predictive prowess of these models.
* **Methodology:** Data Exploration and Preprocessing: The dataset undergoes thorough exploration, encompassing statistical insights, missing value analysis, and distribution examination. Segregation into independent features (X) and the target variable (y) is followed by standardization to enhance model performance.
* **PCA Implementation:** Calculating the covariance matrix from standardized data, determining eigenvalues and eigenvectors, and selecting principal components using sorted eigenvalues. The projection matrix is formulated to transform the dataset into a reduced feature space.
* **Model Building and Evaluation:** Splitting the dataset into training and test sets, training various classifiers (Logistic Regression, SVM, Decision Tree, and Random Forest) on PCA-transformed data, and evaluating models based on accuracy, confusion matrices, and classification reports.
* **Impact Analysis of PCA:** Comparative analysis of model performances before and after PCA, examining the advantages and disadvantages of employing PCA for predictive modelling.

1. **Introduction:**

Parkinson's disease, a neurodegenerative disorder, presents significant challenges in early detection and diagnosis. Utilizing machine learning techniques for predicting the onset of Parkinson's disease has garnered considerable attention due to its potential in facilitating early intervention and better patient care. The availability of the 'pd\_speech\_features.csv' dataset, focusing on speech-related features, serves as a pivotal resource in this pursuit.

* **Problem Statement:** The primary challenge lies in devising an efficient predictive model that can discern subtle indicators of Parkinson's disease onset from speech-related features. Addressing this challenge involves leveraging machine learning models capable of accurately identifying early signs of the disease based on the dataset's attributes.
* **Significance:** Early detection of Parkinson's disease plays a pivotal role in enhancing patient outcomes by allowing timely medical intervention and personalized care plans. The significance of this study lies in its potential to develop robust predictive models for early detection, thereby aiding clinicians in proactive disease management.
* **Objectives:** Develop and evaluate machine learning models capable of predicting the onset of Parkinson's disease based on speech-related features.
* Assess the impact of Principal Component Analysis (PCA) on model accuracy and performance. Analyse the advantages and disadvantages of employing PCA in enhancing the predictive capabilities of the models. Provide recommendations on the practical utility of PCA in similar predictive modelling scenarios for Parkinson's disease or related disorders.

1. **Methods:**

**1. Data Collection and Preprocessing:**

**Dataset Acquisition:** The 'pd\_speech\_features.csv' dataset was obtained, containing speech-related features pertinent to Parkinson's disease.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

**Exploratory Data Analysis (EDA**): Exploratory data analysis was conducted to understand the dataset's structure, including statistical properties, missing values, and feature distributions.

A screenshot of a computer code

Description automatically generated

**Data Segmentation:** Segregation into independent features (X) and the target variable (y) was performed, enabling separate handling of predictors and the variable to predict.

A screenshot of a computer

Description automatically generated

A screenshot of a number grid

Description automatically generated

**2. Data Standardization:**

Standardization: The data underwent standardization to ensure uniform scales across features, thereby mitigating the influence of varying scales on model training.

A screen shot of a computer code

Description automatically generated

**3. Principal Component Analysis (PCA):**

**Covariance Matrix Calculation**: The covariance matrix was computed from the standardized dataset to understand variable relationships.

A white background with numbers

Description automatically generated

A white sheet with numbers

Description automatically generated with medium confidence

**Eigenvalues and Eigenvectors Determination:** Eigenvalues and eigenvectors were derived from the covariance matrix.

A white background with black numbers

Description automatically generated



**Principal Components Selection:** Eigenvalues were sorted in descending order, aiding in the selection of principal components that encompass the most variance.

A screenshot of a computer

Description automatically generated

**Individual explained variance:**

A graph with a black background

Description automatically generated

**4. Model Building and Evaluation:**

Dataset Split: The dataset was split into training and test sets, ensuring separate data for model training and evaluation.

Classifier Training: Various classification models (Logistic Regression, SVM, Decision Tree, Random Forest) were trained using PCA-transformed data.

Model Evaluation: Models were evaluated based on accuracy, confusion matrices, and classification reports, providing insights into their predictive performance.

**5. Impact Analysis of PCA:**

Comparison of Model Performances: Comparative analysis was conducted to assess the performance of models before and after PCA application.

Advantages and Disadvantages of PCA: The influence of PCA on model accuracy and efficiency was scrutinized, discussing the strengths and limitations observed in using PCA for model enhancement.

**IV. Results & Discussion:**

**1. Performance of Machine Learning Models:**

The machine learning models were evaluated based on their predictive accuracy, confusion matrices, and classification reports. The results obtained from the analysis are summarized as follows:

**Logistic Regression:**

A screenshot of a computer

Description automatically generated

**Random Forest Classifier:**

A screenshot of a computer

Description automatically generated

**Support Vector Machine (SVM):**

A screenshot of a computer

Description automatically generated

**Decision Tree Classifier:**

A screenshot of a computer

Description automatically generated

**Confusion Matrix:**

1. **Logistic Regression:**

A red and blue squares with numbers

Description automatically generated

1. **Random Forest Classifier:**

A red and blue squares with numbers

Description automatically generated

1. **Support Vector Machine (SVM):**

A red and blue squares with white text

Description automatically generated

1. **Decision Tree Classifier:**

A red and blue squares

Description automatically generated

**2. Impact Analysis of PCA on Model Performance:**

The application of Principal Component Analysis (PCA) influenced the performance of machine learning models in predicting Parkinson's disease onset. The results of the models before and after PCA transformation demonstrated notable differences in their predictive capabilities.

**Logistic Regression (Using PCA):**

A screenshot of a computer

Description automatically generated

**Random Forest Classifier (Using PCA):**

A screenshot of a computer screen

Description automatically generated

**Support Vector Machine (Using PCA):**

A screenshot of a computer screen

Description automatically generated

**Decision Tree Classifier (Using PCA):**

A screenshot of a computer screen

Description automatically generated

**3.** **Confusion Matrix Using PCA:**

A red and blue squares with white text

Description automatically generated

A red and blue squares with white text

Description automatically generated

A red and blue squares with numbers

Description automatically generated

A red and blue squares with white text

Description automatically generated

**Discussion and Analysis:**

**Model Comparison**: Comparing model performances pre- and post-PCA, identifying models that benefitted from PCA and those less affected.

**Insights from Confusion Matrices:** Analysing misclassifications, identifying patterns or areas needing further investigation.

**PCA Impact:** Discussing how PCA influenced model accuracy, potential reasons for changes observed.

**Practical Implications**: Offering insights into practical applications and considerations for employing PCA in similar predictive modelling scenarios.

**V. Conclusion**

The analysis conducted to predict Parkinson's disease onset using machine learning models and PCA yielded valuable insights into the predictive capabilities and impact of dimensionality reduction techniques.

**1. Summary of Findings:**

The study evaluated various machine learning models for predicting Parkinson's disease onset, showcasing varying accuracies and performances.

Application of Principal Component Analysis (PCA) significantly influenced the predictive capacities of the models, showcasing both advantages and limitations.

PCA demonstrated efficacy in reducing feature dimensions but exhibited differing effects on model accuracies across classifiers.

**2. Implications:**

Early Detection Potential: The study underscores the potential for machine learning models in early detection of Parkinson's disease, laying the foundation for proactive intervention and personalized care.

PCA's Role: PCA's utility in reducing dimensionality while impacting model accuracy emphasizes its significance in preprocessing high-dimensional datasets.

**3. Future Research**:

Hyperparameter Tuning: Further exploration through hyperparameter tuning of models may enhance predictive performance.

Feature Engineering: Investigating additional speech-related features or engineering new features could potentially augment model predictive capabilities.

Comparative Studies: Conducting comparative studies with other dimensionality reduction techniques beyond PCA might provide deeper insights into their comparative efficacy.

**VI. Recommendations:**

Clinical Deployment: The findings could potentially aid clinicians in leveraging machine learning for early identification of Parkinson's disease, promoting timely interventions.

PCA Considerations: Practical considerations surrounding PCA implementation in similar predictive modelling scenarios, outlining its utility and caveats, would benefit practitioners.

**VII. References:**

* [**https://www.tutorialspoint.com/parkinson-disease-prediction-using-machine-learning-in-python**](https://www.tutorialspoint.com/parkinson-disease-prediction-using-machine-learning-in-python)**.**
* Sriram, T.V., Rao, M.V., Narayana, G.S., Kaladhar, D.S.V.G.K. and Vital, T.P.R., 2013. Intelligent Parkinson disease prediction using machine learning algorithms. *Int. J. Eng. Innov. Technol*, *3*, pp.212-215.
* Mandal, I. and Sairam, N., 2014. New machine-learning algorithms for prediction of Parkinson's disease. *International Journal of Systems Science*, *45*(3), pp.647-666.