PROJECT REPORT

FML Case Studies

SUBMITTED TO MIT SCHOOL OF COMPUTING

BACHELOR OF TECHNOLOGY
(Information Technology)

Information Technology- (CORE)

BY

Prajwal Kadbane	MITU22BTEC0050
Kedar Kulkarni	MITU22BTIT0041
Anisha Jain	MITU22BTIT0017
Jaydeep Ghule	MITU22BTAE0021

Under The Guidance of
Prof. Palash Sontakke

ABSTRACT

CASE STUDY 1:

This project involves the application of Principal Component Analysis (PCA) to the MNIST handwritten digits dataset to reduce its dimensions for visualization and classification. PCA is a statistical technique used to transform high-dimensional data into a lower-dimensional space while retaining the maximum variance present in the data. The primary goals are to enhance computational efficiency, simplify data visualization, and assess the impact of dimensionality reduction on classification accuracy.

The MNIST dataset, consisting of 70,000 images of handwritten digits (0–9), is widely used for training and testing in the field of machine learning. Each image is a 28x28 pixel grid, flattened into a 784-dimensional vector. This high dimensionality poses challenges for computation and visualization. By reducing the dataset to 2D or 3D using PCA, we aim to better understand the structure of the data and maintain classification performance.

CASE STUDY 2:

The early detection of heart disease is crucial for timely medical intervention and effective treatment. This project aims to predict the likelihood of heart disease in patients based on their medical history and diagnostic test results using a machine learning approach. Specifically, a **Decision Tree classifier** is employed to classify patients into two categories: those with heart disease (label 1) and those without (label 0).

The dataset used for this project is the **Heart Disease UCI Dataset** from Kaggle, which contains various features related to heart health, such as age, cholesterol levels, blood pressure, and ECG results. The primary goal of the project is to preprocess this data, handle any missing or unbalanced values, and build an effective predictive model.

THEORY

CASE STUDY 1: Dimensionality Reduction of Handwritten Digits Using PCA

The field of machine learning often deals with high-dimensional data, where each data point can be represented by a large number of features. In such scenarios, it is common to face challenges related to the "curse of dimensionality," which includes increased computational cost, data sparsity, and overfitting. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), are vital tools that help mitigate these issues by projecting high-dimensional data onto a lower-dimensional space while retaining most of the original information.

This project focuses on applying PCA to the well-known MNIST handwritten digits dataset, which consists of 70,000 grayscale images of digits (0–9). Each image is represented as a 28x28 pixel grid, resulting in 784 features per image when flattened. Although the dataset is manageable for basic machine learning tasks, visualizing and analyzing such high-dimensional data can be difficult, and models trained on such data may become computationally expensive or inefficient.

Motivation and Goals: The main motivation for using PCA in this project is to simplify the dataset by reducing its dimensionality while preserving as much variance as possible. This simplification allows for:

1. **Enhanced Data Visualization**: Projecting the dataset to a 2D or 3D space for easy visual analysis, which can provide insights into the distribution and clustering of the data.
2. **Computational Efficiency**: Reducing the number of features can speed up training and inference for machine learning models.
3. **Preservation of Key Information**: While PCA reduces the dimensionality, it aims to maintain the most important structures and relationships within the data.

Objectives of the Project:

- **Dimensionality Reduction**: Apply PCA to reduce the original 784-dimensional dataset to 2 or 3 dimensions.
- **Visualization**: Create visual representations of the reduced-dimension data to identify patterns and relationships between different digit classes.
- **Classification Performance Analysis**: Evaluate the impact of dimensionality reduction on the accuracy of a simple classification algorithm, such as K-Nearest Neighbors (KNN), and compare it to the performance on the original high-dimensional dataset.

Background on Principal Component Analysis (PCA):

PCA is a linear dimensionality reduction technique that transforms a dataset into a new coordinate system such that the greatest variance lies along the first coordinate (principal

component), the second greatest variance along the second coordinate, and so on. The process involves:

1. **Mean Centering the Data**: Shifting the data such that its mean is zero.
2. **Computing Covariance Matrix**: Calculating the covariance matrix to understand how features vary together.
3. **Eigenvalue Decomposition**: Finding the eigenvalues and eigenvectors of the covariance matrix.
4. **Selecting Principal Components**: Choosing the top components that capture the highest variance.
5. **Projecting Data**: Reconstructing the dataset in the lower-dimensional space using the selected principal components.

Importance of PCA in Machine Learning: PCA plays a significant role in pre-processing data, especially when dealing with high-dimensional datasets. By reducing the dimensionality, PCA helps prevent overfitting in models, reduces noise, and makes the analysis more interpretable. In visualization tasks, reducing complex datasets to two or three dimensions allows researchers and practitioners to visually inspect data distributions, detect outliers, and gain insights into potential clustering or classification boundaries.

MNIST Dataset Overview: The MNIST dataset has been a benchmark in the field of machine learning for training various algorithms. Each image in the dataset represents a digit from 0 to 9, making it a multi-class classification problem. The dataset is relatively balanced, with each digit represented approximately equally. However, visualizing or comprehending the 784-dimensional feature space directly is impractical, making PCA an ideal tool for this project.

Project Relevance: By applying PCA to the MNIST dataset, this project bridges the gap between complex high-dimensional data and interpretable lower-dimensional representations. It demonstrates the practical applications of dimensionality reduction for visualization and analysis and highlights the balance between maintaining data integrity and achieving computational benefits.

CASE STUDY 2: Predicting Heart Disease Using Decision Tree Classifier

Heart disease remains one of the most prevalent and life-threatening health conditions worldwide, posing significant challenges to healthcare systems. According to global health statistics, millions of individuals suffer from cardiovascular diseases annually, making early diagnosis and treatment paramount. Predictive models based on machine learning have the potential to revolutionize preventive healthcare by identifying patients at risk before severe complications arise. This project focuses on building a predictive model that uses medical data to forecast the likelihood of heart disease in patients.

Project Overview

The objective of this project is to design and implement a **Decision Tree classifier** capable of predicting whether a patient is at risk of heart disease based on historical medical data and diagnostic test results. The model is trained to classify patients into one of two categories: those diagnosed with heart disease (1) and those not diagnosed with heart disease (0). This binary classification approach is essential for prioritizing cases that require further medical attention and monitoring.

Dataset Description

The dataset employed for this project is the **Heart Disease UCI Dataset**, available on Kaggle. It is a comprehensive collection of patient data that includes medical attributes commonly associated with heart health, such as:

- **Age**: The age of the patient.
- **Sex**: The gender of the patient (e.g., male, female).
- **Chest Pain Type (CP)**: Describes the type of chest pain experienced (e.g., typical angina, atypical angina).
- **Resting Blood Pressure (trestbps)**: The patient's blood pressure at rest.
- **Cholesterol Levels (chol)**: Serum cholesterol in mg/dl.
- **Fasting Blood Sugar (fbs)**: Indicates if blood sugar is greater than 120 mg/dl.
- **Resting Electrocardiographic Results (restecg)**: Results from the patient's ECG.
- **Maximum Heart Rate Achieved (thalach)**: The maximum heart rate recorded during testing.
- **Exercise-Induced Angina (exang)**: Indicates whether the patient experiences angina due to exercise.
- **ST Depression (oldpeak)**: A measure related to exercise stress testing.
- **Slope of Peak Exercise ST Segment (slope)**: Refers to the slope of the peak exercise segment.
- **Number of Major Vessels (ca)**: Number of major blood vessels colored by fluoroscopy.

- **Thalassemia (thal)**: A blood disorder status.

These features are analyzed to understand their correlation with the risk of heart disease, which provides insight into which factors most significantly influence the likelihood of developing cardiovascular conditions.

Importance of Predictive Modeling in Healthcare

The use of machine learning models, such as Decision Trees, in medical diagnostics is gaining traction due to their interpretability and effectiveness. Decision Trees are particularly valuable because they can provide clear decision-making paths that are easily understood by healthcare professionals, which helps build trust in machine-assisted decisions. Unlike black-box models, Decision Trees illustrate how different features contribute to the final prediction, facilitating better clinical interpretations and discussions with patients.

Methodology Outline

1. **Data Collection**: The dataset is downloaded and examined for structure and quality.
2. **Data Preprocessing**: This step involves cleaning the data by addressing any missing values, encoding categorical features, and scaling numerical data to ensure a well-prepared dataset.
3. **Exploratory Data Analysis (EDA)**: A thorough analysis is conducted to visualize data distributions, correlations, and highlight any significant patterns or trends that impact heart disease predictions.
4. **Model Building**: The project employs a Decision Tree classifier, initially trained with default parameters. The model is then optimized using hyperparameter tuning techniques such as **GridSearchCV** to enhance its predictive performance.
5. **Model Evaluation**: Performance metrics, including **Accuracy**, **Precision**, **Recall**, **F1-score**, and **ROC-AUC**, are utilized to evaluate the effectiveness of the model.
6. **Decision Tree Visualization**: Tools like graphviz or matplotlib are used to present a visual representation of the tree structure, making it easier to understand how decisions are made.

Expected Outcomes

The anticipated result of this project is a well-tuned Decision Tree model that can accurately classify patients at risk of heart disease. The project seeks to identify critical features that are most influential in predicting heart disease, which can provide medical professionals with data-driven insights to support early diagnosis and intervention. Ultimately, this project underscores the potential of leveraging machine learning in clinical settings to improve patient outcomes and optimize healthcare resources.

CONCLUSION

CASE STUDY 1: Dimensionality Reduction of Handwritten Digits Using PCA

The completion of this project provides valuable insights into the use and effectiveness of Principal Component Analysis (PCA) as a dimensionality reduction technique applied to a well-known dataset, the MNIST handwritten digits dataset. The project demonstrates how PCA can transform a high-dimensional dataset into a lower-dimensional space, facilitating data visualization, enhancing computational efficiency, and preserving important structures within the data.

Key Findings and Observations:

1. **Dimensionality Reduction and Variance Retention**:
    - By applying PCA to the 784-dimensional feature space of the MNIST dataset, we successfully reduced it to 2D and 3D representations. Despite the significant reduction in dimensionality, PCA managed to capture a substantial proportion of the variance. For example, reducing the dataset to the first two principal components often retained around 70–80% of the original variance, ensuring that the most relevant data characteristics were preserved.
    - The explained variance ratio analysis helped determine the number of principal components needed to strike a balance between dimensionality reduction and information retention. This aspect is crucial when using PCA in real-world applications, as it shows how much of the data's variability is maintained in the reduced representation.
2. **Visualization Benefits**:
    - The 2D and 3D scatter plots created using the first few principal components provided a clear visual representation of the dataset. Distinct clusters were observed for different digit classes, which indicated that even in a reduced space, PCA was effective in preserving the separability of different classes.
    - Visualization with PCA helped highlight overlaps between certain digit classes (e.g., 3 and 8 or 4 and 9), which are inherently similar and can be harder to distinguish. These insights are valuable for understanding the challenges of classification in high-dimensional data and how dimensionality reduction affects class boundaries.
3. **Impact on Classification Performance**:
    - The evaluation of a basic K-Nearest Neighbors (KNN) classifier on both the original high-dimensional data and the PCA-reduced data provided an important benchmark. The classification accuracy on the original 784-dimensional data

served as the baseline, while the accuracy on the 2D/3D PCA-transformed data illustrated the trade-off between reduced dimensionality and model performance.

- While the accuracy of the KNN classifier on the PCA-reduced data was lower than that on the full-dimensional data, it remained competitive. This result highlighted that PCA can significantly reduce the feature space while maintaining a reasonable level of classification accuracy, which is beneficial when computational resources or processing time are limited.

4. **Computational Efficiency**:
    - Reducing the dimensionality of the data led to a substantial decrease in computational complexity. The training and testing times for the KNN classifier were noticeably shorter when working with PCA-transformed data compared to the original data. This finding underlines PCA's practical advantage for large-scale or real-time applications where processing speed is critical.

Challenges and Limitations:

- One of the main challenges observed was the trade-off between reducing dimensionality and preserving classification performance. Although PCA effectively captured most of the variance in the first few components, some loss of information was inevitable, impacting model accuracy.
- The application of PCA, which is inherently a linear transformation, may not capture non-linear relationships within the data as effectively as more sophisticated non-linear dimensionality reduction techniques like t-SNE or UMAP.

Potential Future Work:

- **Enhanced Techniques**: Exploring non-linear dimensionality reduction methods such as t-SNE or UMAP could provide a more detailed representation of complex datasets, potentially capturing more nuanced relationships than PCA.
- **Hybrid Approaches**: Combining PCA with other feature selection methods could help retain additional useful information that PCA alone might miss.
- **Deep Learning**: Implementing autoencoders, a type of neural network designed for data compression, could provide an alternative approach to dimensionality reduction with potentially better retention of complex data structures.

Concluding Remarks: This project successfully demonstrated the application of PCA for dimensionality reduction on the MNIST handwritten digits dataset. The reduction to lower dimensions enabled insightful visualization and maintained reasonable classification performance, proving the effectiveness of PCA for data simplification tasks. The balance between preserving data variance and achieving computational benefits underscores PCA's utility in both exploratory data analysis and machine learning pipelines. By understanding how PCA affects data representation and model performance, researchers and practitioners

can make more informed decisions when pre-processing high-dimensional datasets for visualization and modeling purposes.

CASE STUDY 2: Predicting Heart Disease Using Decision Tree Classifier

The "Predicting Heart Disease Using Decision Tree Classifier" project demonstrates the powerful application of machine learning in medical diagnostics. By leveraging the **Heart Disease UCI Dataset**, we developed an effective Decision Tree model that classifies patients based on their risk of heart disease, using a variety of clinical and demographic features. The project underscores the potential of predictive analytics in aiding early diagnosis and improving treatment outcomes for patients.

Summary of Key Findings

1. **Model Effectiveness**: The Decision Tree classifier, with proper preprocessing and hyperparameter tuning, achieved a balance of predictive accuracy and interpretability. Metrics such as **accuracy, precision, recall, F1-score, and ROC-AUC** provided a comprehensive view of the model's performance. The model's ability to correctly identify patients at risk (sensitivity) and those without risk (specificity) was demonstrated to be effective, providing reliable early detection support.

2. **Feature Importance**: The project highlighted which features from the dataset played the most critical roles in predicting heart disease. Attributes such as age, cholesterol levels, chest pain type, and maximum heart rate emerged as significant predictors. Understanding the influence of these factors allows healthcare practitioners to prioritize further diagnostic tests or preventive measures for high-risk individuals.

3. **Model Interpretability**: One of the strengths of using a Decision Tree classifier is its interpretability. Unlike complex black-box models, Decision Trees provide a visual and logical flow of decisions that can be easily understood by clinicians. This transparency helps build trust in machine learning models, fostering wider acceptance and integration into clinical workflows.

4. **Data Preprocessing and EDA Insights**: The data preprocessing phase, which included handling missing values, encoding categorical variables, and normalizing numerical features, proved crucial for model performance. **Exploratory Data Analysis (EDA)** revealed important trends and correlations, such as the association between chest pain type and heart disease, which deepened the understanding of the dataset and informed the model-building process.

Challenges and Considerations

While the Decision Tree classifier showed strong performance, the project encountered challenges that are important to acknowledge:

- **Overfitting Risk**: Decision Trees can be prone to overfitting, especially when grown without constraints. To mitigate this, the model was pruned, and hyperparameters were fine-tuned to achieve generalizability across unseen data.
- **Class Imbalance**: Addressing imbalanced classes using techniques such as **SMOTE** or undersampling was necessary to ensure that minority classes were represented accurately, avoiding biased model predictions.

Future Work

To further enhance the project, several extensions and improvements can be considered:

- **Ensemble Learning**: Implementing ensemble methods such as **Random Forests** or **Gradient Boosted Trees** can improve model robustness and predictive performance.
- **Feature Engineering**: Creating new features from existing data (e.g., interaction terms between variables) could uncover deeper relationships and improve model accuracy.
- **Cross-Validation**: Performing cross-validation can provide more robust evaluations and prevent data leakage from single train-test splits.
- **Integration with Clinical Systems**: Partnering with healthcare providers to integrate such predictive tools into electronic health records (EHR) systems could help automate risk assessment processes in real-time.

Final Thoughts

This project successfully demonstrated that a well-constructed Decision Tree classifier can serve as a valuable tool for early heart disease detection. By facilitating proactive healthcare measures and reducing the burden of late-stage diagnosis, predictive models can empower healthcare professionals to deliver timely interventions, potentially saving lives and optimizing treatment outcomes. However, real-world implementation of such models requires collaboration between data scientists and medical experts, ensuring that machine learning insights are validated and ethically applied within clinical contexts.

In conclusion, machine learning models like the Decision Tree offer significant promise in transforming healthcare diagnostics. With continuous improvements in model sophistication and data quality, these tools can play an integral role in enhancing patient care, aiding early diagnosis, and fostering preventive medicine.