AI-ML Internship

A Project Report submitted to the

GLOBAL NEXT CONSULTING INDIA PVT LTD

(Six – Week Internship Program)

By

D. Md Khizer Farhaan

Under the Supervision of

Dr. Anuradha Gupta
(Project Director)

Submitted To :

Global Next Consulting India Pvt. Ltd.

Duration of Internship :

23-March-2026 to 15-May-2026

May 2026

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report titled "AI-ML Internship (GNCIPL)" is the result of my own work carried out under the guidance of Ms. Anuradha Gupta during the period from March 2026 to May 2026.

I further declare that this report represents an authentic record of my work and does not contain any fabricated information. I have followed all principles of academic honesty and integrity in the preparation of this report.

D. Md Khizer Farhaan

CERTIFICATE

This is to certify that the project report entitled “Artificial Intelligence and Machine Learning Internship” has been successfully carried out by D. Md Khizer Farhaan.

The work was completed under the guidance of Ms. Anuradha Gupta during the period from March 2026 to May 2026. During this internship, I worked on various projects related to Artificial Intelligence and Machine Learning, demonstrating good analytical skills and practical understanding of the subject.

It is further certified that this work is an original record of my own efforts and has not been submitted, either in part or in full, to any other university or institution for the award of any degree, diploma, or certificate.

Ms. Anuradha Gupta

Program Director

GNCIPL

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude to all those who contributed to the successful completion of this project report.

I would like to express my sincere gratitude to my supervisor, Ms. Anuradha Gupta, for their invaluable guidance, encouragement, and constructive suggestions throughout the course of this work. Their expertise and constant support played a crucial role in the successful completion of this report.

I am also thankful to the staff of Global Next Consulting India Private Limited for providing the necessary resources, facilities and their assistance, without which this study would not have been possible.

Finally, I would also like to acknowledge my peers and teachers whose support and discussions have been helpful in the completion of this report.

D. Md Khizer Farhaan

ABSTRACT

This report presents the work completed during my six-week internship in the domain of Artificial Intelligence and Machine Learning. The internship focused on applying machine learning and data analysis techniques using Python and various AI/ML libraries to solve real-world problems.

The projects completed during the internship include Olympic Medal Count Analysis, Heart Disease Risk Analysis, Customer Segmentation using K-Means Clustering, Vehicle Feature Clustering, Parkinson's Disease Detection, and Financial Threat Monitoring Dashboard Using CTGAN. These projects involved data preprocessing, exploratory data analysis (EDA), clustering, classification, synthetic data generation, and visualization using real-world datasets from healthcare, sports, automobile, customer analytics, and fraud detection domains.

Various tools and technologies such as Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, CTGAN, and Streamlit were used throughout the internship. Different visualization techniques including histograms, scatter plots, heatmaps, and clustering visualizations were implemented to analyze patterns and generate meaningful insights from the datasets.

Overall, the internship provided practical exposure to machine learning workflows including data preprocessing, feature analysis, model development, visualization, and prediction systems, significantly improving my technical and analytical skills in Artificial Intelligence and Machine Learning.

INDEX

Chapter 1- Introduction

1.1 Company's Profile

Global Next Consulting India Private Limited (GNCIPL), headquartered in Greater Noida, Uttar Pradesh, is a cybersecurity-focused consulting firm dedicated to helping organizations protect their digital assets, data, and reputation. As threats evolve in today's digital world, GNCIPL offers proactive, customized solutions rather than reactive fixes. The company serves clients in diverse sectors including finance, healthcare, manufacturing, and technology, providing services like threat detection, risk assessment, incident response, compliance consulting, and 24/7 monitoring. GNCIPL's core values are integrity, innovation, customer-centricity, excellence, and collaboration - ensuring that technical solutions align with clients' specific needs and long-term goals.

Contact Details

Location- B5,402 P4 PHi2, CGEWHO TOWER, GREATER NOIDA 201310
Contact Numbers- 0120-4001768, +91-9315504902. +91-7666141260
Mail- hr@gncipl.com

1.2 Objectives of Internship

The main objectives of the internship were:

- To gain practical experience in Artificial Intelligence and Machine Learning
- To build predictive models using real-world datasets
- To understand classification, regression, and clustering algorithms
- To perform data preprocessing and feature engineering
- To evaluate model performance using appropriate metrics
- To deploy machine learning models using Streamlit
- To develop problem-solving skills using data-driven approaches

Chapter 2 - Projects

2.1 Olympic Medal Count Analysis by Country (Week 1)

2.1.1 Introduction

The Olympic Games are one of the largest international sporting events where athletes from different countries compete across multiple sports disciplines. Olympic data provides valuable insights into country-wise performance, sports dominance, medal trends, and athletic achievements over time.

This project focuses on performing Exploratory Data Analysis (EDA) on historical Olympic medal data using Python. The analysis helps understand medal distribution patterns, top-performing countries, sports with maximum medals, and medal trends across different Olympic years.

The project also evaluates performance efficiency using medals per capita by combining Olympic medal datasets with country population data. Various visualization techniques were used to identify trends and generate meaningful insights from the dataset.

2.1.2 Objectives

- To analyze Olympic medal distribution across countries
- To identify top-performing countries in Olympic history
- To study medal trends across different Olympic years
- To analyze sports contributing the highest number of medals
- To visualize medal distribution using graphs and charts
- To compare Gold, Silver, and Bronze medal trends
- To evaluate country performance using medals per capita
- To generate meaningful insights using Exploratory Data Analysis

2.1.3 Tools & Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Google Colab

2.1.4 Dataset Description

The project uses three datasets:

- Athlete Events Dataset
- NOC Regions Dataset
- Population Dataset

The datasets contain athlete records, country-wise Olympic performance information, medal details, and population statistics. These datasets were merged and processed to perform Olympic medal analysis and medals-per-capita evaluation.

2.1.5 Methodology

a) Data Collection
The datasets were collected from publicly available Olympic historical records and population datasets.

b) Data Preprocessing
The following preprocessing steps were performed:
- Loaded datasets using Pandas
- Selected relevant columns
- Removed rows with missing medal values
- Merged Olympic data with country region dataset
- Removed unnecessary columns
- Handled missing values
- Cleaned inconsistent country names
- Merged population dataset for performance analysis

c) Exploratory Data Analysis (EDA)
EDA was performed to understand trends and relationships within the data using:
- Bar charts
- Pie charts
- Line charts
- Horizontal bar charts
- Stacked bar charts

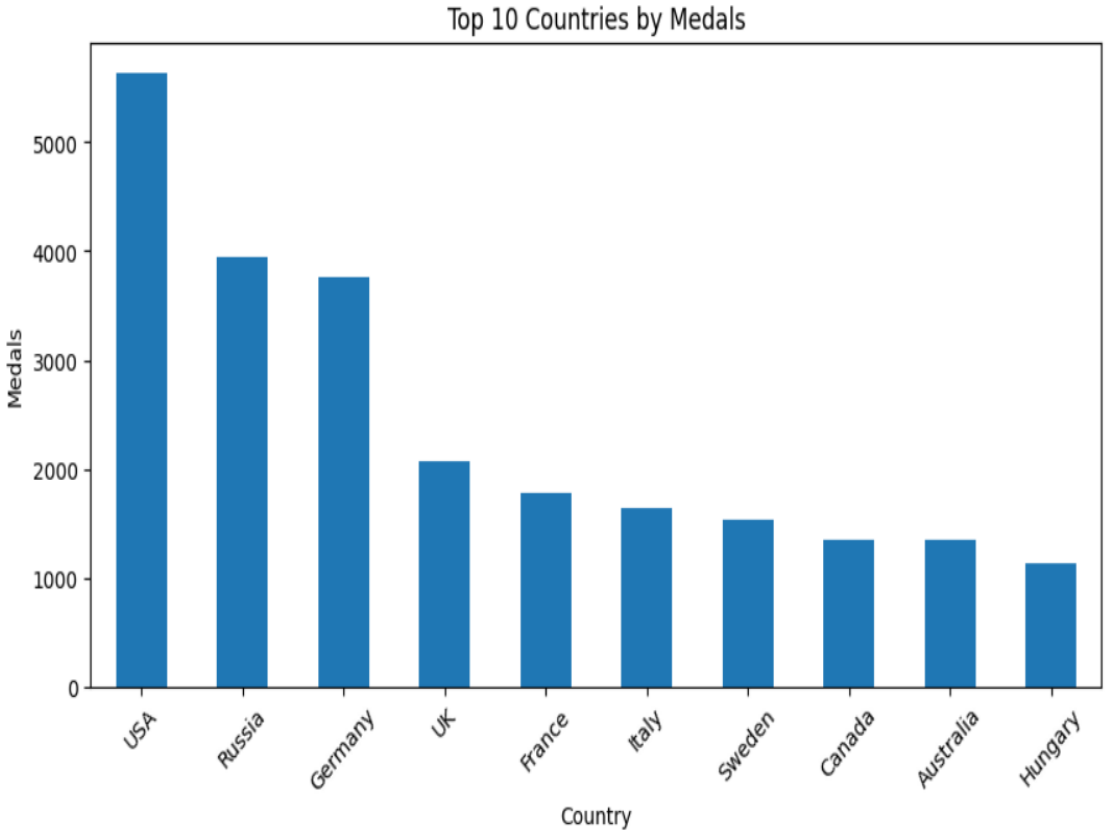
d) Data Visualization

Visualization techniques were used to analyze:

- Country-wise medal counts
- Medal distribution
- Olympic year trends
- Top sports analysis
- Medal type trends
- Medals per capita evaluation

2.1.6 Results & Insights

i) Top Countries by Medal Count



The **Figure 1: Top 10 Countries by Total Olympic Medal Count**

The analysis showed that countries such as the United States, Soviet Union/Russia, Germany, and the United Kingdom dominated Olympic history in terms of total medals won.

ii) Medal Distribution Analysis

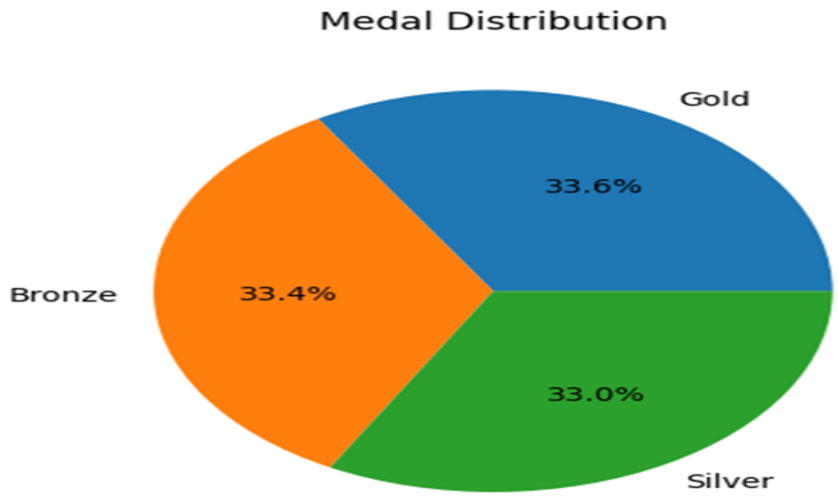


Figure 2: Distribution of Gold, Silver, and Bronze Medals

The pie chart visualization showed the proportion of Gold, Silver, and Bronze medals. The distribution remained relatively balanced across all medal categories.

iii) Medal Trends Across Olympic Years

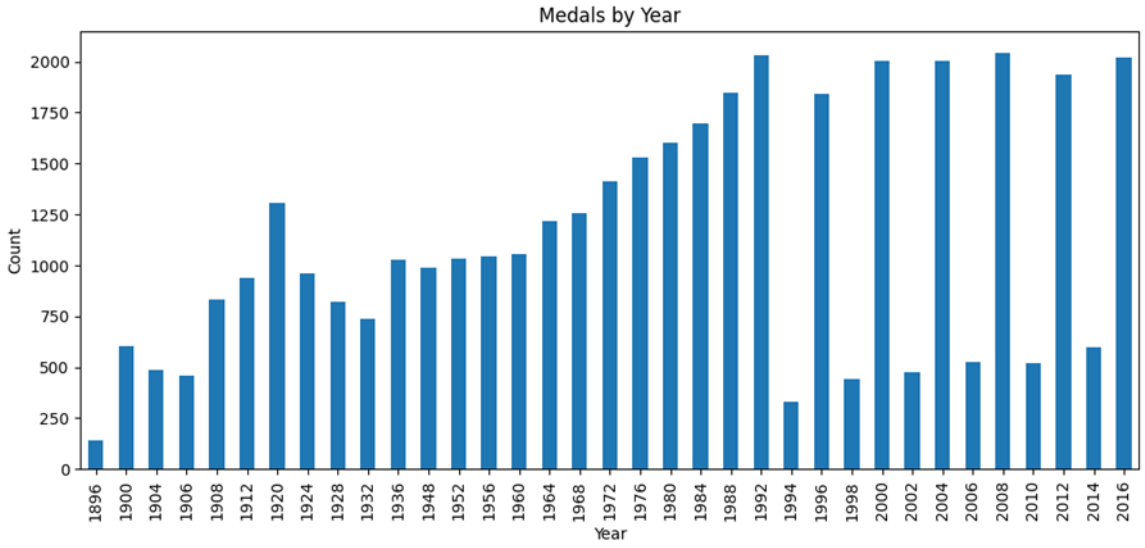


Figure 3: Olympic Medal Trends Across Different Years

The analysis of medals by year revealed fluctuations in medal counts across different Olympic years due to increasing participation and expansion of Olympic events.

iv) Country Performance Trends

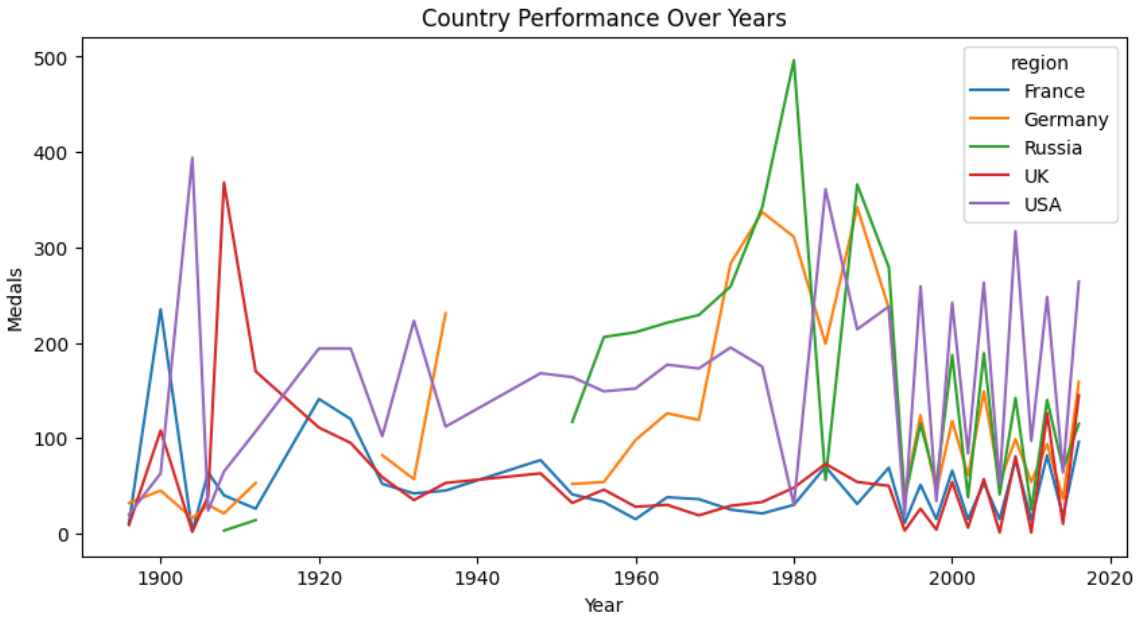


Figure 4: Country-wise Olympic Performance Trends Over Time

Line chart analysis showed how medal-winning performance changed over time for top-performing countries.

v) Top Sports by Medal Count

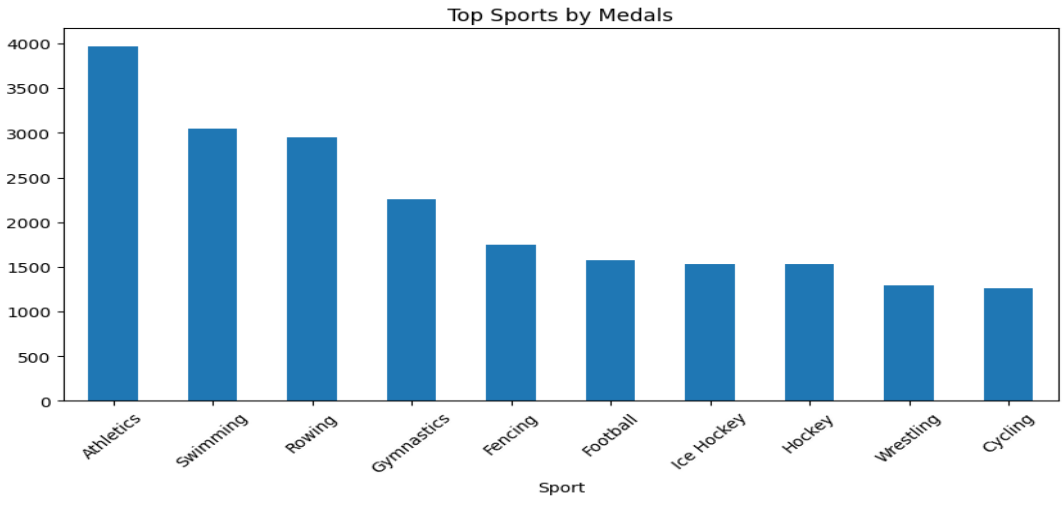


Figure 5: Top Sports Contributing the Highest Olympic Medals

Sports such as Athletics, Swimming, Gymnastics, Wrestling, and Rowing contributed the highest number of Olympic medals.

vi) Medal Type Trends Over Time

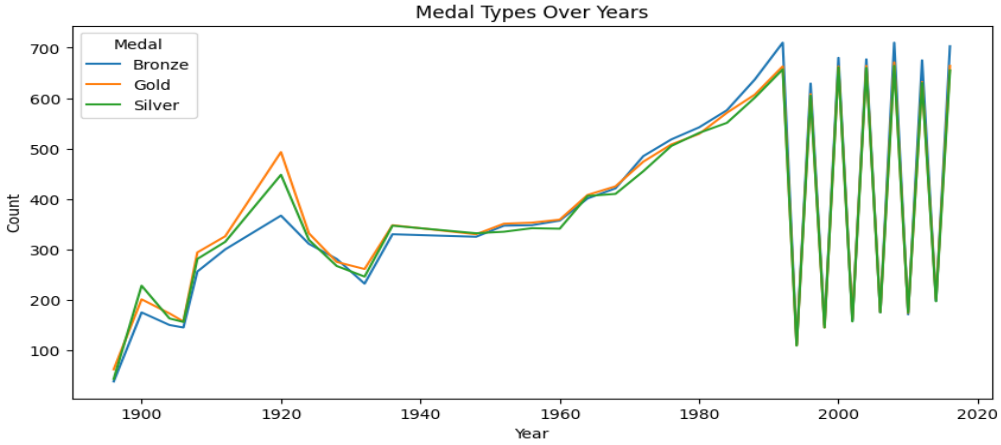


Figure 6: Gold, Silver, and Bronze Medal Trends Over Years

Gold, Silver, and Bronze medal counts increased gradually over time due to the growth of Olympic competitions.

2.1.7 Conclusion

This project successfully performed Exploratory Data Analysis on historical Olympic medal data using Python-based data analysis techniques.

The analysis revealed valuable insights regarding country-wise Olympic dominance, medal distribution patterns, sports contributing maximum medals, and Olympic trends across years.

The study also highlighted the importance of medals-per-capita analysis in evaluating country performance efficiency beyond total medal counts.

Overall, the project demonstrated how data visualization and EDA techniques can be used to uncover meaningful patterns from large historical sports datasets and support sports performance analysis.

2.2 Heart Disease Risk Analysis (Week 2)

2.2.1 Introduction

Heart disease is one of the leading causes of death worldwide and identifying major risk factors at an early stage is extremely important for effective healthcare management.

This project focuses on performing Exploratory Data Analysis (EDA) on heart disease datasets using Python to analyze relationships between medical attributes such as age, cholesterol, blood pressure, chest pain type, and heart rate.

2.2.2 Objectives

- To analyze medical data related to heart disease
- To identify important health-related risk factors
- To study the relationship between age and heart disease
- To analyze cholesterol and blood pressure patterns
- To perform Exploratory Data Analysis using Python
- To generate meaningful healthcare insights from the dataset

2.2.3 Tools & Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Google Colab

2.2.4 Dataset Description

The dataset contains medical information of patients including:

- Age
- Sex
- Chest Pain Type
- Blood Pressure
- Cholesterol Level
- Maximum Heart Rate
- Heart Disease Target Variable

The dataset contains both numerical and categorical variables used for healthcare analysis and visualization.

2.2.5 Methodology

a) Data Collection
The dataset was collected from publicly available healthcare datasets.

b) Data Preprocessing
• Loaded dataset using Pandas
• Checked missing values
• Removed duplicates
• Identified categorical and numerical features
• Prepared dataset for visualization

c) Exploratory Data Analysis (EDA)
EDA was performed using:
• Histograms
• Boxplots
• Count plots
• Scatter plots
• Correlation heatmaps

d) Data Visualization
Visualization techniques were used to analyze:
• Heart disease distribution
• Age distribution
• Cholesterol levels
• Blood pressure patterns
• Correlation between medical variables

2.2.6 Results & Insights

i) Heart Disease Distribution

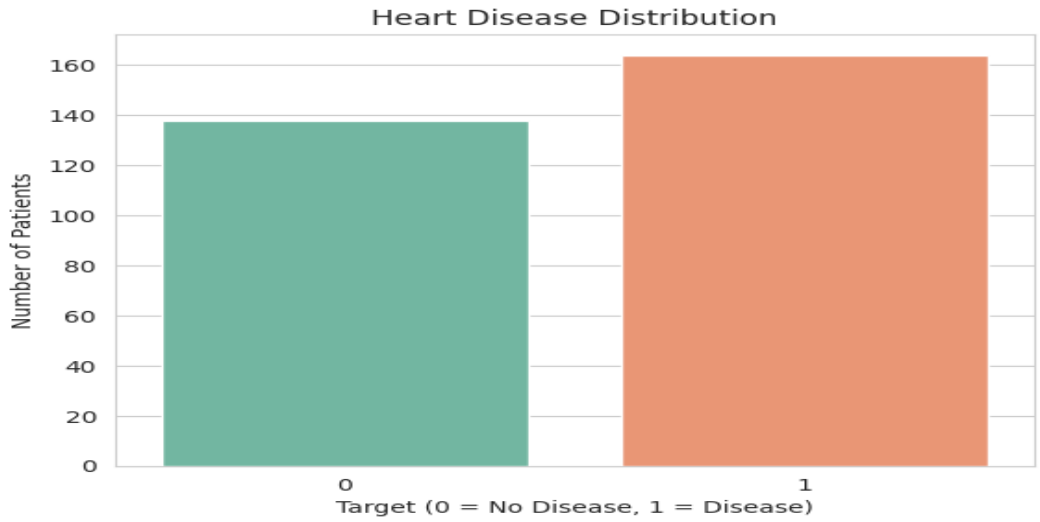


Figure 1: Distribution of Heart Disease Cases

• The dataset contains both heart disease and non-heart disease cases, showing balanced patient distribution.
• The graph highlights the importance of analyzing health-related risk factors associated with heart disease.

ii) Age Distribution Analysis

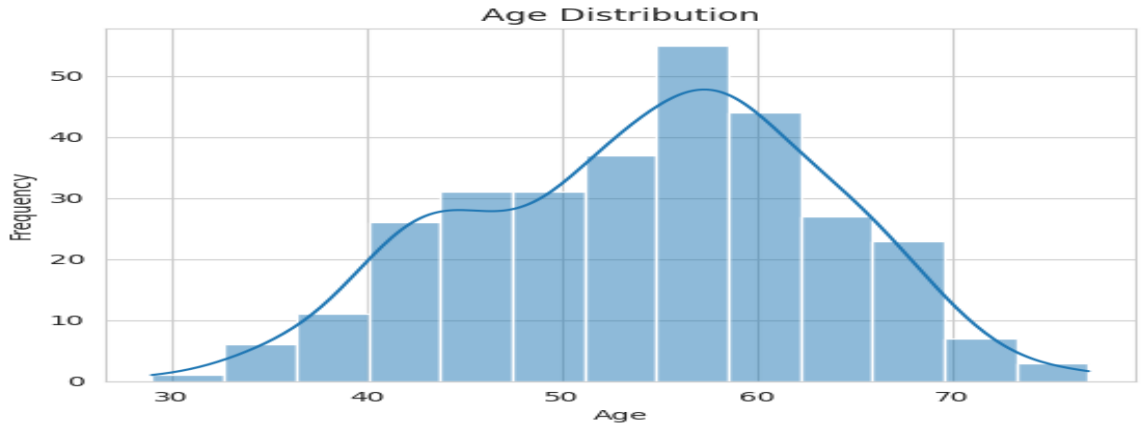


Figure 2: Distribution of Patient Age

• Most patients belong to the middle-aged and older age groups.
• The analysis indicates that heart disease risk increases with age.
stomers fall within the age group of 30–50 years, showing a balanced age distribution.

iii) Age vs Heart Disease Analysis

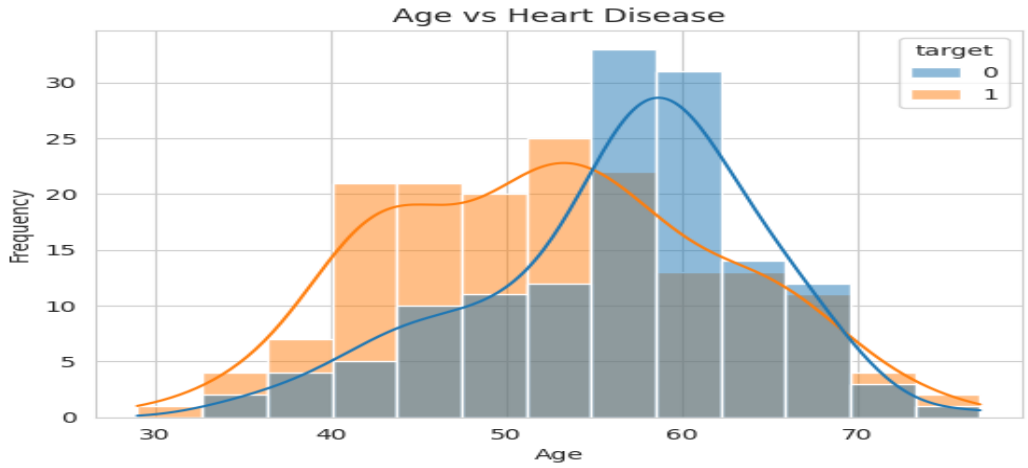


Figure 3: Relationship Between Age and Heart Disease

• Older patients recorded a higher number of heart disease cases compared to younger individuals.
• The graph shows that increasing age is an important cardiovascular risk factor.

iv) Boxplot Analysis of Age

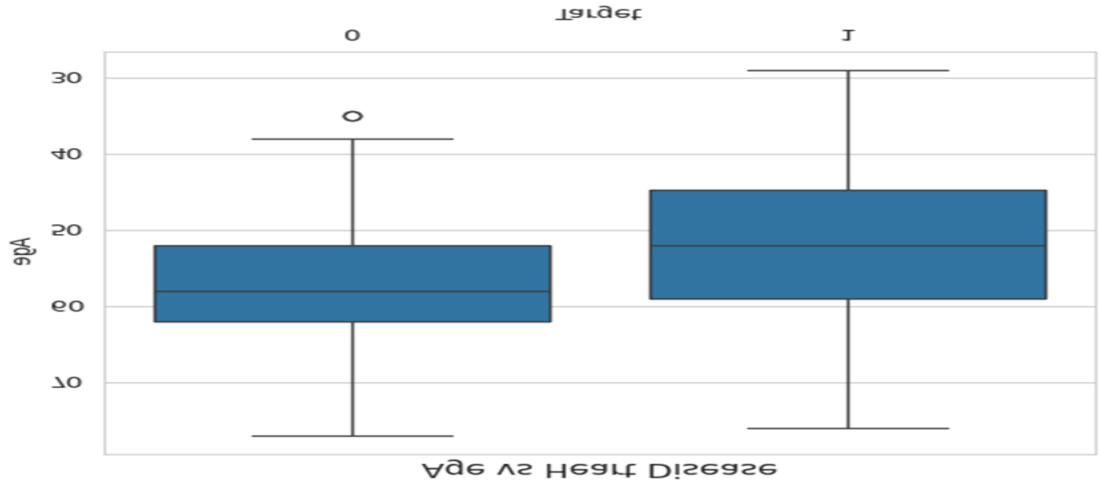


Figure 4: Age Distribution Based on Heart Disease

• Patients with heart disease showed a higher median age compared to healthy individuals.
• Some outliers indicate that heart disease may also occur in younger patients.

v) Cholesterol Distribution Analysis

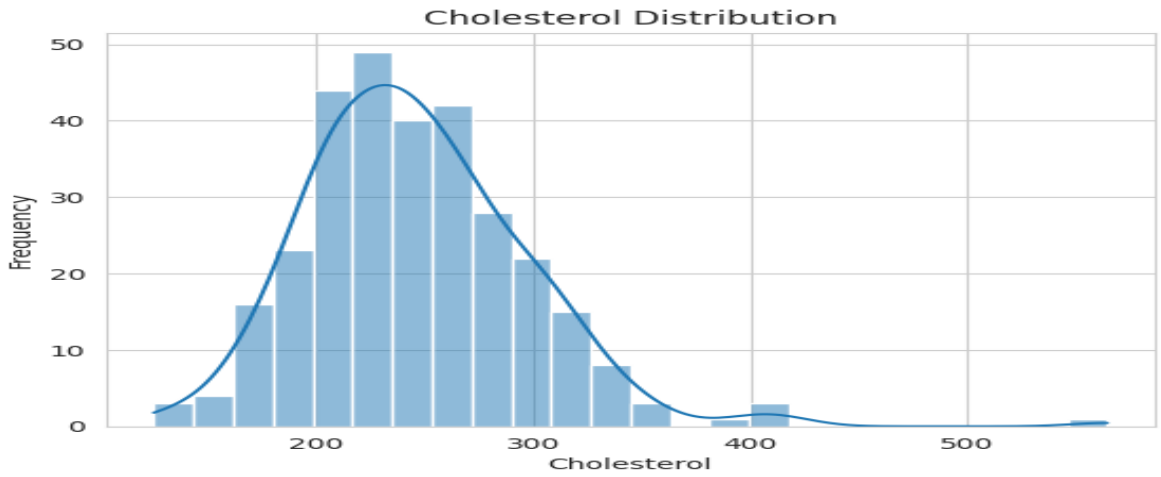


Figure 5: Distribution of Cholesterol Levels

- Several patients recorded cholesterol levels above the normal range.
- High cholesterol levels are strongly associated with cardiovascular diseases.

vi) Cholesterol vs Heart Disease

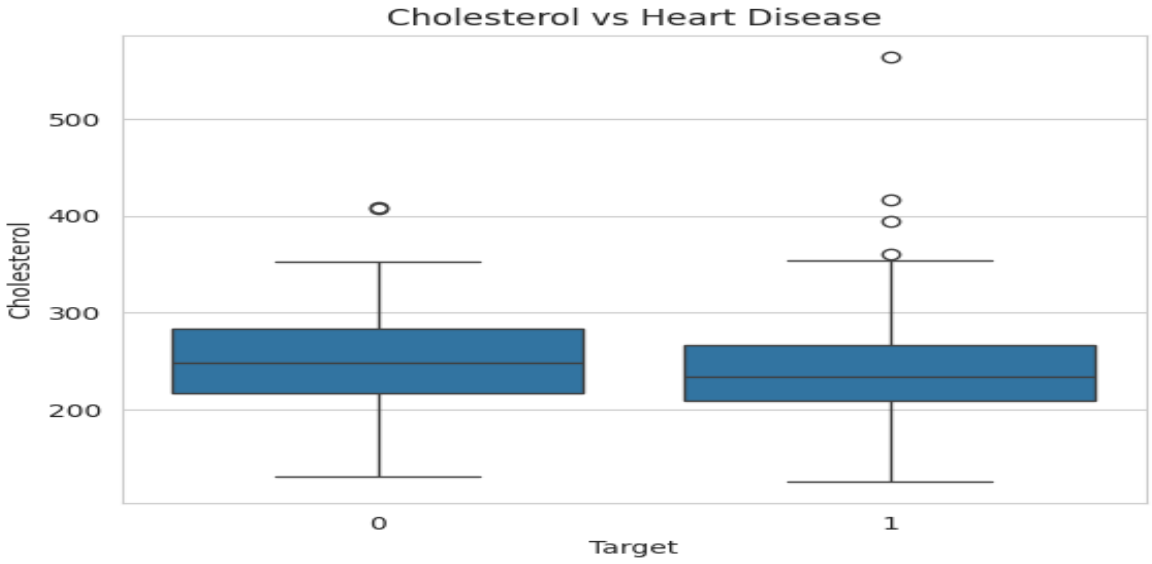


Figure 6: Cholesterol Levels Based on Heart Disease

- Patients diagnosed with heart disease generally showed higher cholesterol levels.
- The graph indicates variability in cholesterol levels among patients.

vii) Age vs Maximum Heart Rate Visualization

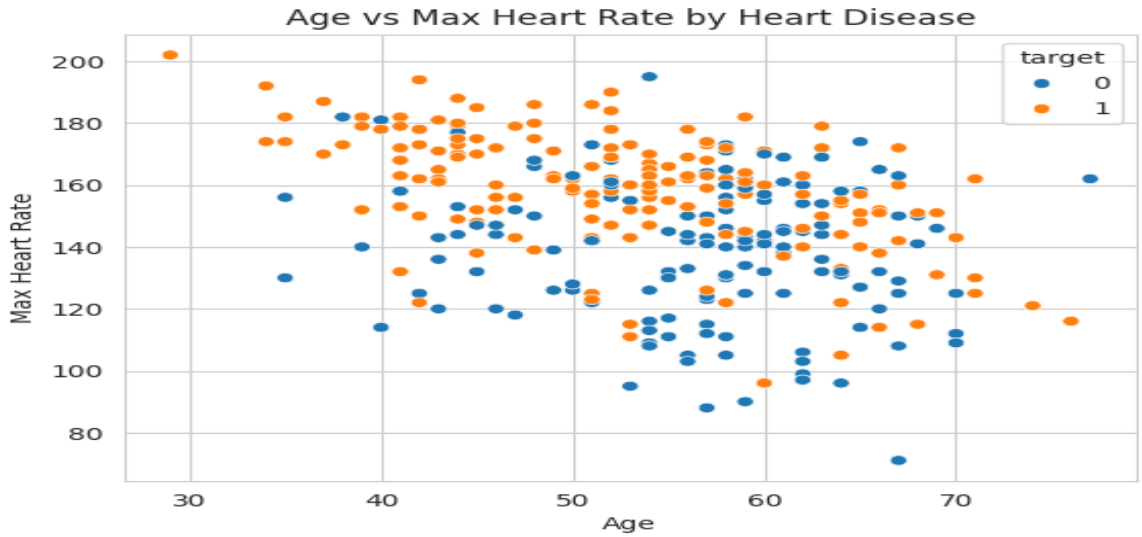


Figure 8: Age vs Maximum Heart Rate by Heart Disease

• Maximum heart rate generally decreases as patient age increases.
• Patients with heart disease showed lower maximum heart rates compared to healthier individual.

2.2.7 Conclusion

This project successfully analyzed heart disease risk factors using Python-based Exploratory Data Analysis techniques.

The study revealed that factors such as age, cholesterol levels, blood pressure, chest pain type, and heart rate are strongly associated with heart disease occurrence.

Overall, the project provided practical experience in healthcare data analysis, preprocessing, and visualization using Python.

2.3 Customer Segmentation for a Retail Store using K-Means (Week-3)

2.3.1 Introduction

Customer segmentation is one of the most important techniques used in retail businesses to understand customer behavior and improve marketing strategies. By grouping customers based on their purchasing patterns and income levels, businesses can target customers more effectively and improve customer satisfaction.

This project focuses on analyzing mall customer data using K-Means Clustering. The analysis was performed using Python to group customers into different segments based on annual income and spending score. Exploratory Data Analysis (EDA) and clustering techniques were used to identify customer groups and understand spending behavior patterns.

2.3.2 Objectives

- To analyze customer purchasing behavior
- To perform customer segmentation using K-Means Clustering
- To identify customer groups based on income and spending score
- To visualize customer patterns using graphs and charts
- To determine the optimal number of clusters using the Elbow Method
- To generate meaningful business insights from customer data

2.3.3 Tools & Technologies Used

- Python
- Pandas
- NumPy

• Matplotlib
• Seaborn
• Scikit-learn
• Google Colab

2.3.4 Dataset Description

The project uses the Mall Customers dataset containing customer demographic and spending information.

The dataset includes:
• Customer ID
• Gender
• Age
• Annual Income (k$)
• Spending Score (1-100)

The dataset contains both numerical and categorical variables used for customer analysis and clustering.

2.3.5 Methodology

a) Data Collection

The dataset was collected from publicly available retail customer datasets.

b) Data Preprocessing

• Loaded dataset using Pandas
• Checked missing values
• Removed unnecessary columns
• Selected important features for clustering

• Prepared data for visualization and model building

c) Exploratory Data Analysis (EDA)

EDA was performed using:

• Histograms

• Scatter plots

• Count plots

• Cluster visualizations

d) Clustering Technique

K-Means Clustering was used to divide customers into different groups based on spending behavior and annual income.

e) Data Visualization

Visualization techniques were used to analyze:

• Age distribution

• Gender distribution

• Income distribution

• Spending score distribution

• Customer clusters

• Relationship between income and spending score

2.3.6 Results & Insights

i) Customer Distribution Analysis

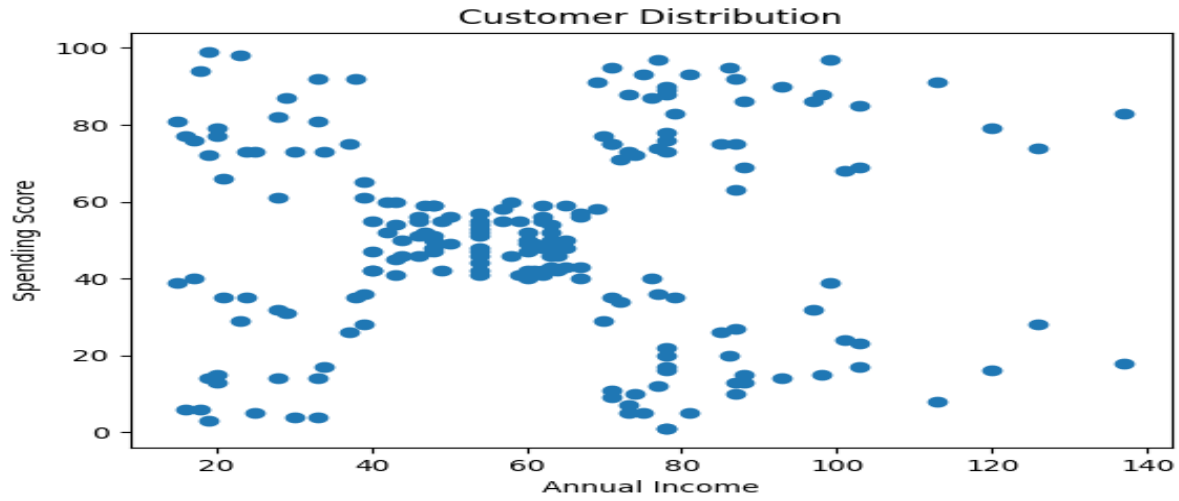


Figure 1: Customer Distribution based on Income and Spending Score

• The scatter plot shows the relationship between annual income and spending score of customers.

• The graph helps identify different customer behavior patterns before clustering.

ii) Elbow Method Analysis

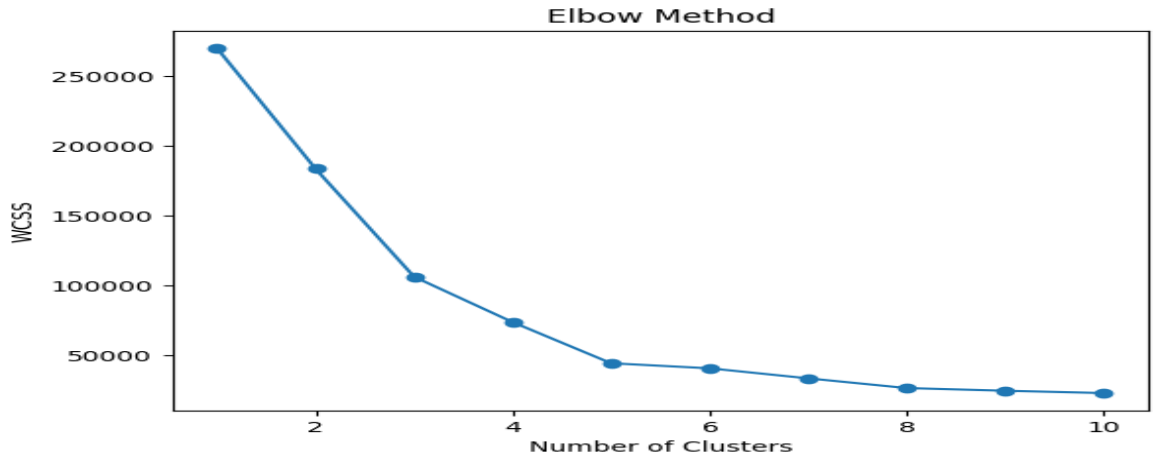


Figure 2: Elbow Method for Optimal Clusters

• The elbow graph was used to determine the optimal number of clusters for K-Means clustering.

• The graph indicated that 5 clusters provide the best segmentation result.

iii) Final Customer Clusters

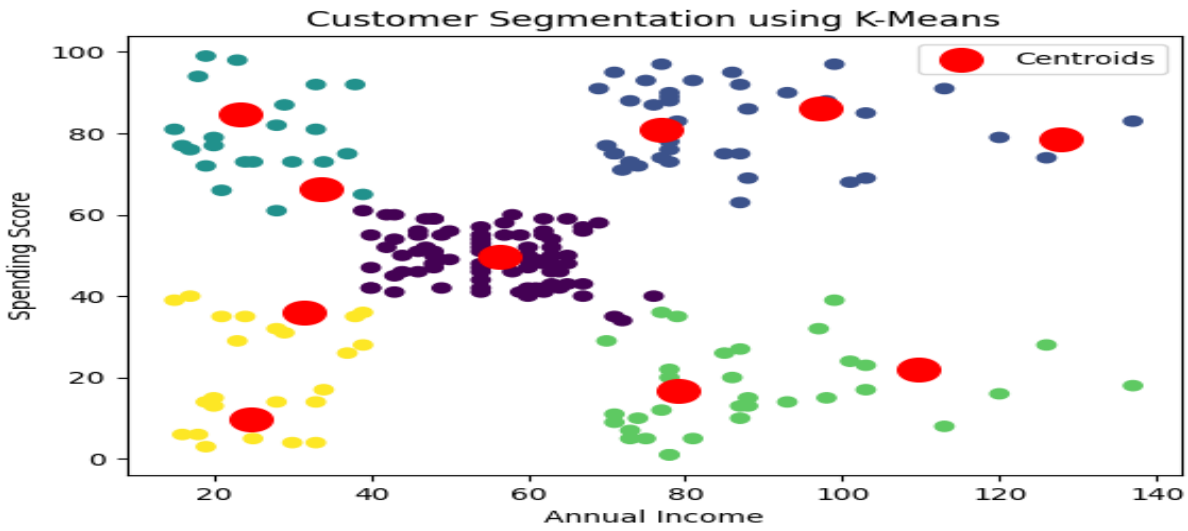


Figure 3: Customer Segments using K-Means Clustering

• Customers were divided into different groups based on annual income and spending behavior.

• The graph clearly shows high-spending, low-spending, and average customer segments.

iv) Age Distribution Analysis

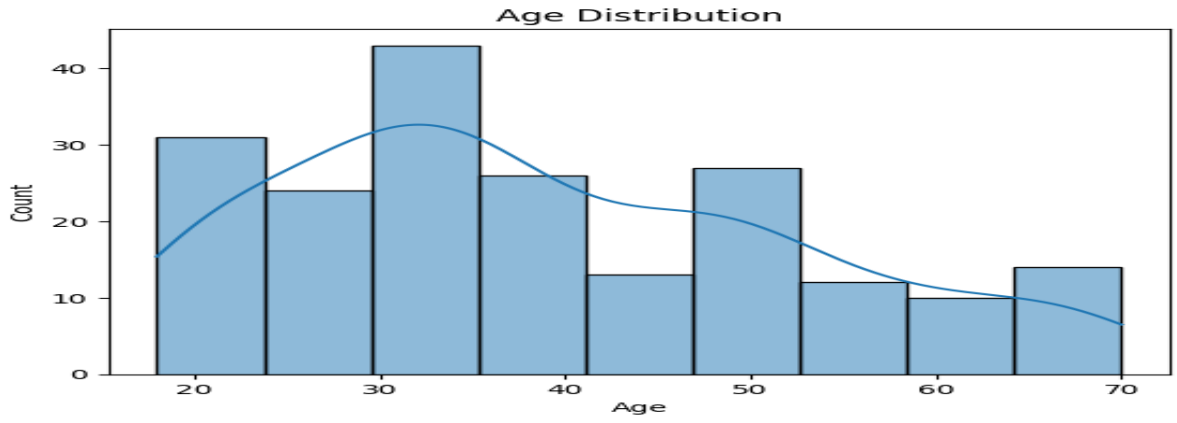


Figure 4: Distribution of Customer Age

• Most customers belong to middle-aged groups.

• The histogram helps understand customer age distribution patterns.

v) Gender Distribution Analysis

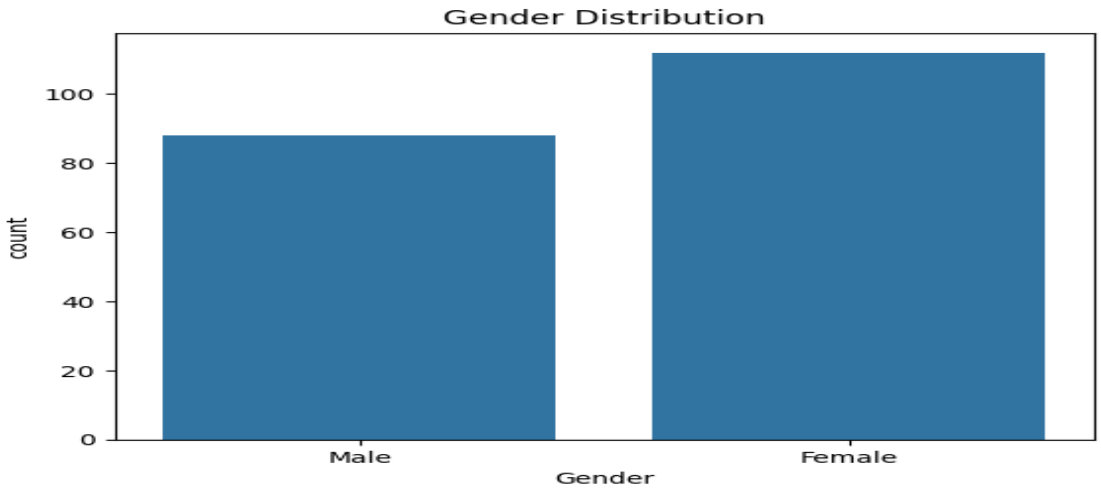


Figure 5: Gender Distribution of Customers

- The dataset contains both male and female customers.
- The count plot helps compare customer distribution based on gender.

vi) Income Distribution Analysis

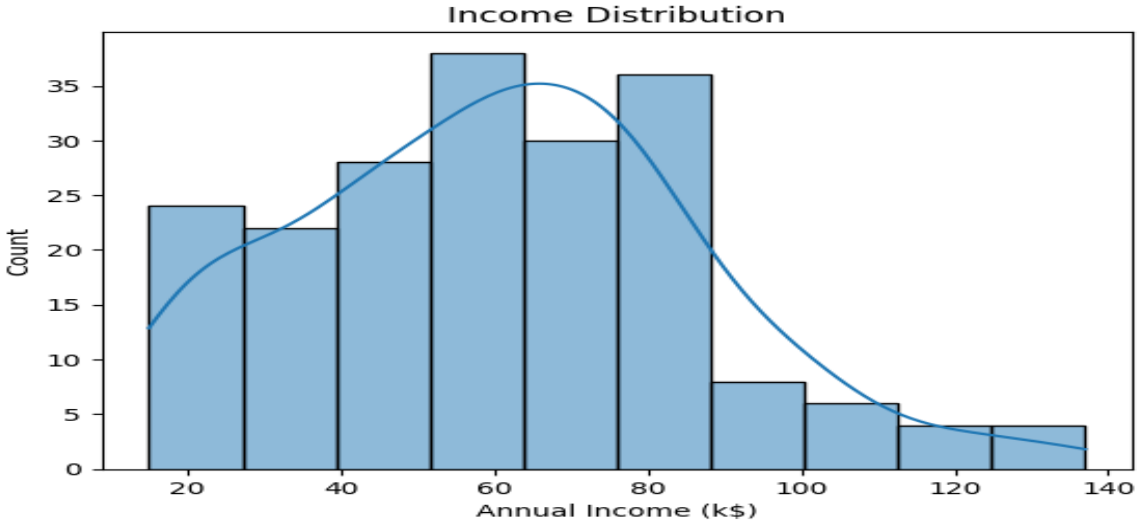


Figure 6: Distribution of Annual Income

- Customer income values are distributed across different income ranges.
- The graph helps analyze customer purchasing capacity.

vii) Spending Score Distribution

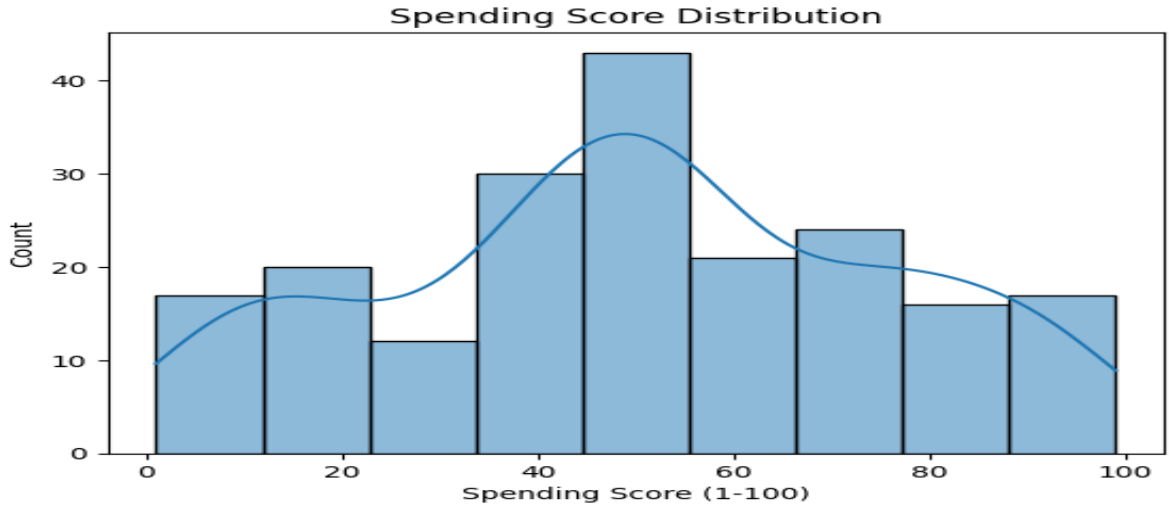


Figure 7: Distribution of Spending Score

- Spending scores vary significantly among customers.
- The graph helps identify customers with high and low spending behavior.

viii) Age vs Spending Score Analysis

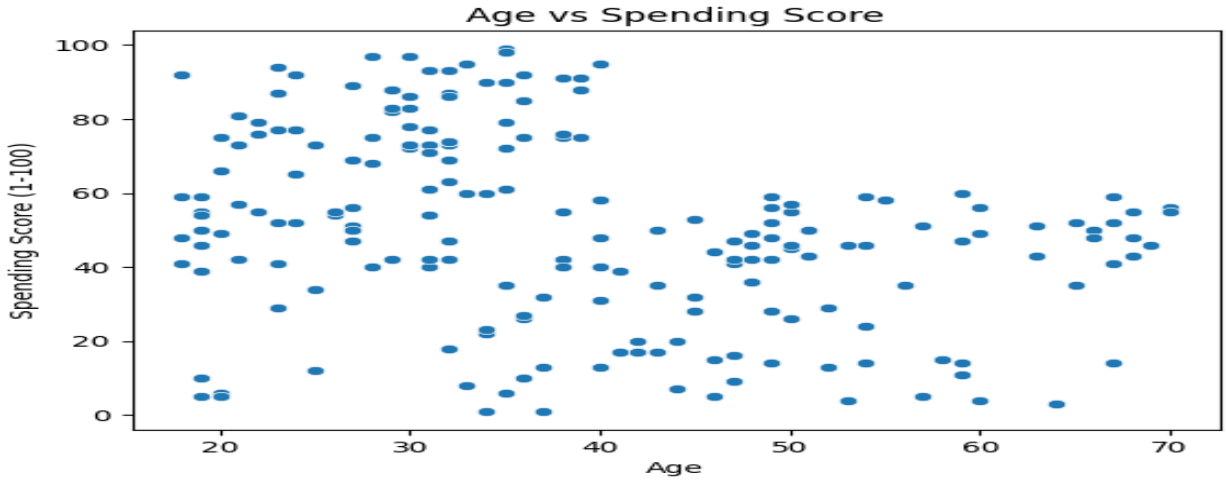


Figure 8: Age vs Spending Score

- Younger customers generally recorded higher spending scores.
- The scatter plot helps analyze the relationship between age and spending behavior.

2.3.7 Conclusion

This project successfully analyzed customer purchasing behavior using Exploratory Data Analysis and K-Means Clustering techniques.

The study identified different customer groups based on annual income and spending score. The clustering results help businesses understand customer behavior and improve marketing strategies.

Overall, the project provided practical experience in customer segmentation, clustering algorithms, data preprocessing, and visualization using Python.

2.4 Vehicle Feature Segmentation using Clustering (Week 4)

2.4.1 Introduction

Vehicle manufacturers and automotive companies often analyze vehicle characteristics to understand different categories of vehicles based on performance and fuel efficiency. Segmenting vehicles into meaningful groups helps in market analysis, product development, and customer targeting.

This project focuses on segmenting vehicles using unsupervised machine learning techniques based on features such as MPG (miles per gallon), horsepower, weight, acceleration, and cylinders. The project involves exploratory data analysis (EDA), data preprocessing, feature scaling, clustering using K-Means, and dimensionality reduction using Principal Component Analysis (PCA). The goal was to identify meaningful vehicle groups such as fuel-efficient vehicles and high-performance vehicles.

2.4.2 Objectives

• To analyze vehicle performance and efficiency characteristics
• To preprocess and clean automotive datasets
• To identify relationships between vehicle features
• To perform exploratory data analysis using visualization techniques
• To apply K-Means clustering for vehicle segmentation
• To determine the optimal number of clusters using Elbow Method and Silhouette Score
• To visualize clusters using PCA

2.4.3 Tools & Technologies Used

• Python
• Pandas
• NumPy
• Matplotlib
• Seaborn
• Scikit-learn

2.4.4 Dataset Description

The dataset contains vehicle-related attributes including fuel efficiency, engine performance, and physical characteristics.

Key features include:

• MPG (Miles Per Gallon)
• Horsepower
• Weight
• Acceleration
• Cylinders
• Displacement
• Origin

The dataset includes numerical variables related to vehicle performance and efficiency across different automobile types.

2.4.5 Methodology

a) Data Preprocessing

The following preprocessing steps were performed:

• Loaded the dataset using Pandas
• Checked missing values and duplicate records
• Handled missing values using median imputation
• Removed irrelevant columns
• Prepared the dataset for clustering

b) Exploratory Data Analysis (EDA)

EDA was performed using:

• Histograms
• Scatter plots
• Boxplots
• Correlation heatmaps

The analysis focused on understanding relationships among:

• Horsepower and MPG
• Weight and MPG
• Cylinders and MPG
• Horsepower and Weight

c) Feature Scaling

Standardization was applied using StandardScaler to normalize the dataset and ensure equal contribution of all features during clustering.

d) K-Means Clustering

K-Means clustering was applied to group vehicles into different segments based on their characteristics.

e) Optimal Cluster Selection

The following methods were used:

• Elbow Method
• Silhouette Score

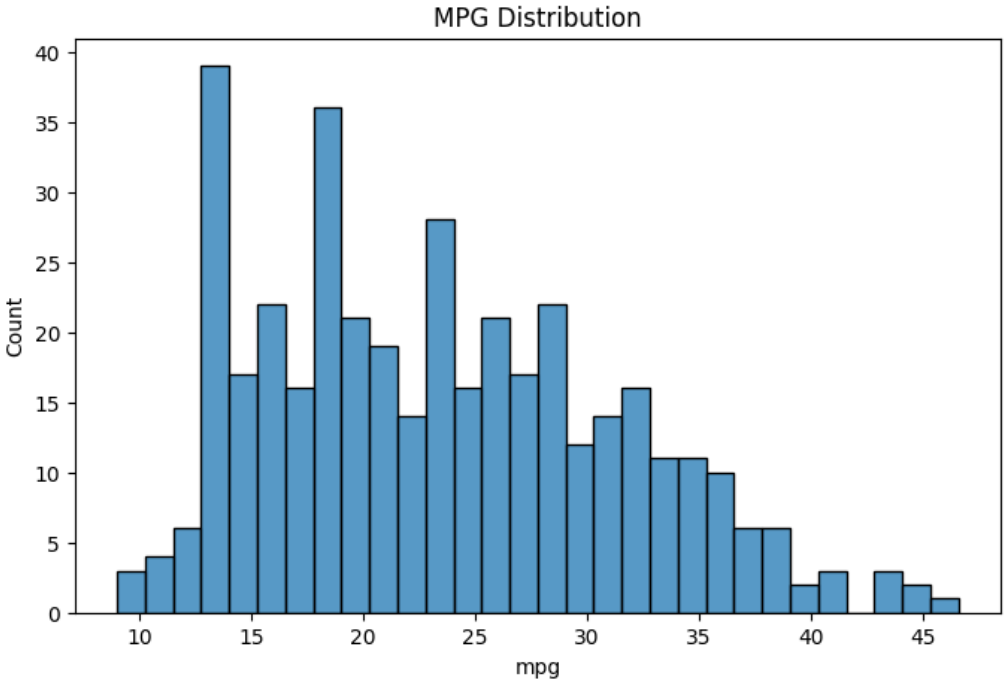
These techniques helped identify the most suitable number of clusters.

f) Principal Component Analysis (PCA)

PCA was applied to reduce dimensionality and visualize clusters in two-dimensional space.

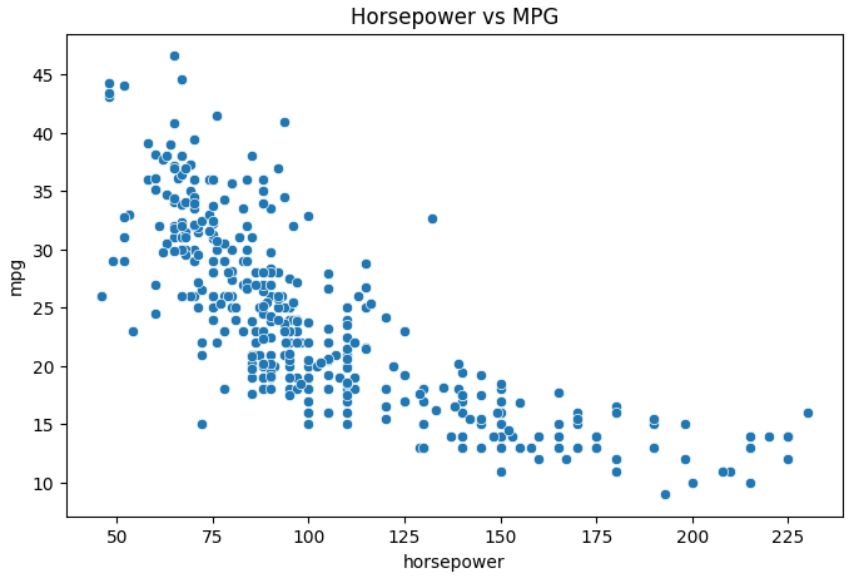
2.4.6 Results & Insights

i) MPG Distribution



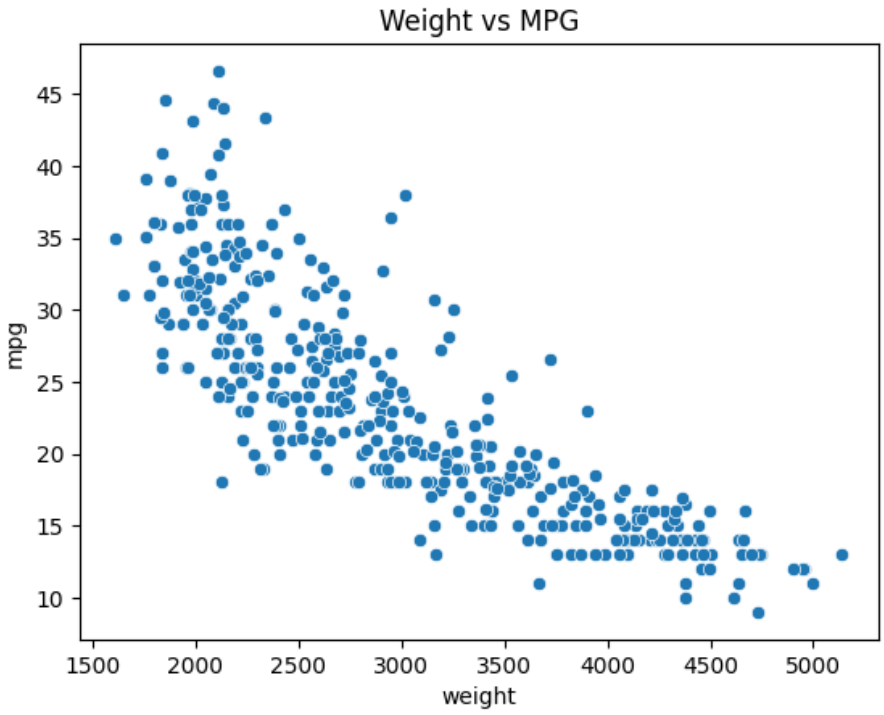
Most vehicles showed MPG values between 15–30, indicating moderate fuel efficiency across the dataset.

ii) Horsepower vs MPG



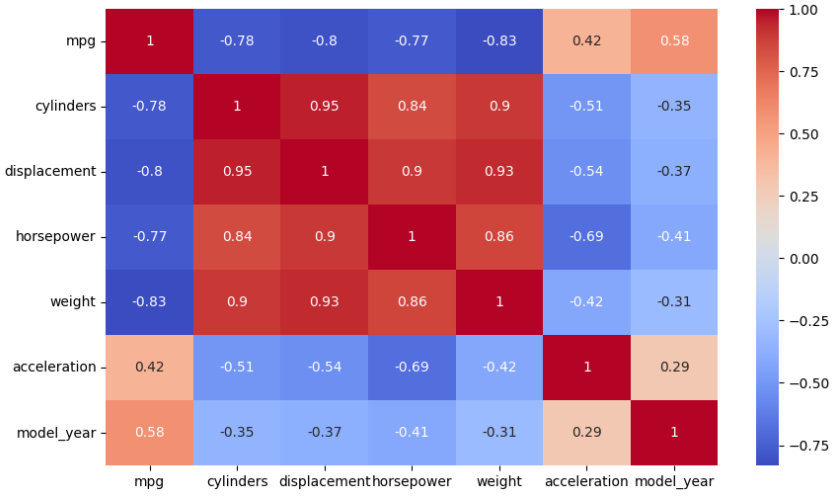
A strong negative relationship was observed between horsepower and MPG, meaning vehicles with higher horsepower generally had lower fuel efficiency.

iii) Weight vs MPG



Heavier vehicles showed lower MPG values, confirming that vehicle weight significantly affects fuel consumption.

iv) Correlation Analysis

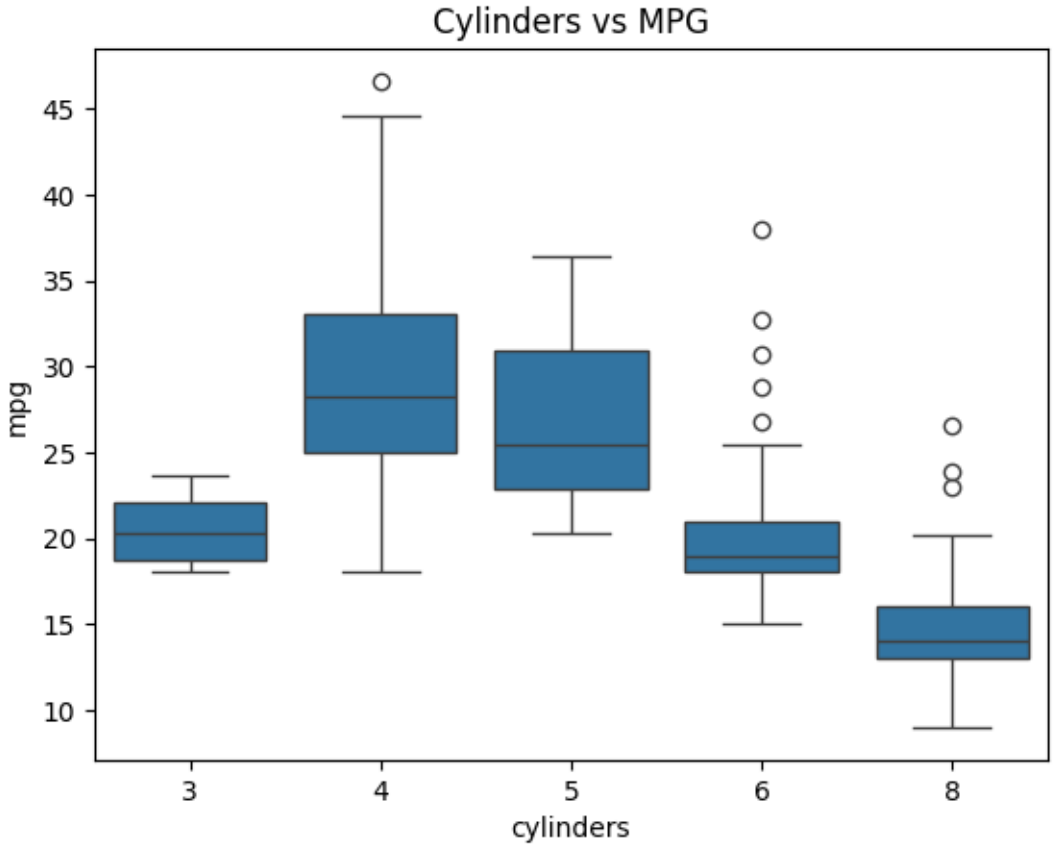


The heatmap revealed:

• MPG negatively correlated with weight and horsepower
• Weight positively correlated with horsepower

These relationships supported meaningful clustering of vehicles based on performance and efficiency.

v) Cylinders vs MPG



Vehicles with higher cylinder counts generally had lower fuel efficiency due to larger engine sizes.

vi) K-Means Clustering

K-Means clustering successfully grouped vehicles into distinct categories based on performance characteristics.

vii) PCA Visualization

PCA visualization showed reasonably clear separation between clusters, indicating effective segmentation of vehicles.

viii) Cluster Insights

The clustering results identified major vehicle groups:

• Cluster 0 → High-performance vehicles with higher horsepower and weight but lower MPG
• Cluster 1 → Fuel-efficient vehicles with lower horsepower and higher MPG

The clustering process successfully differentiated vehicles based on performance and efficiency characteristics.

2.4.7 Conclusion

This project successfully applied unsupervised machine learning techniques to segment vehicles based on their performance and fuel efficiency features.

The study revealed strong relationships between horsepower, weight, and fuel efficiency. K-Means clustering effectively grouped vehicles into meaningful categories, while PCA visualization helped interpret cluster structures.

Overall, the project provided practical experience in clustering algorithms, feature scaling, dimensionality reduction, and exploratory data analysis using real-world automotive datasets.

2.5 Parkinson’s Disease Detection (Python, ANN & Streamlit) (Week 5)

2.5.1 Introduction

Parkinson’s disease is a progressive neurological disorder that affects movement, speech, and motor control. Early detection is important for timely treatment and better disease management. Traditional diagnosis methods can be expensive and time-consuming, making machine learning-based prediction systems highly valuable.

This project focuses on detecting Parkinson’s disease using biomedical vocal measurements such as frequency, jitter, shimmer, and HNR (Harmonics-to-Noise Ratio). The project includes data preprocessing, exploratory data analysis (EDA), feature scaling, Artificial Neural Network (ANN) model building, evaluation, and deployment using Streamlit. A user-friendly web application was also developed to provide real-time disease prediction based on voice parameters.

2.5.2 Objectives

• To analyze biomedical voice data related to Parkinson’s disease
• To preprocess and clean healthcare datasets
• To identify important vocal features related to Parkinson’s disease
• To build a classification model using Artificial Neural Networks (ANN)
• To evaluate model performance using classification metrics
• To visualize model accuracy and loss trends
• To deploy the prediction model using Streamlit

2.5.3 Tools & Technologies Used

• Python
• Pandas
• NumPy
• Matplotlib
• Seaborn
• Scikit-learn

• TensorFlow / Keras
• Streamlit

2.5.4 Dataset Description

The dataset contains biomedical voice measurements of individuals classified as either healthy or Parkinson-affected.

Key features include:

• MDVP:Fo(Hz) → Fundamental Frequency
• MDVP:Jitter(%) → Frequency Variation
• MDVP:Shimmer → Amplitude Variation
• HNR → Harmonics-to-Noise Ratio
• Status → Target Variable (0 = Healthy, 1 = Parkinson)

The dataset contains both numerical features and classification labels used for disease prediction.

2.5.5 Methodology

a) Data Preprocessing

The following preprocessing steps were performed:

• Loaded dataset using Pandas
• Removed unnecessary columns
• Checked missing values and duplicates
• Separated features and target variable
• Applied feature scaling using StandardScaler

b) Exploratory Data Analysis (EDA)

EDA was performed using:

• Count plots
• Histograms
• Boxplots
• Scatter plots
• Correlation heatmaps

The analysis focused on identifying differences between healthy individuals and Parkinson patients based on vocal characteristics.

c) Feature Selection

Important features selected for model training included:

- Frequency (Fo)
- Jitter
- Shimmer
- HNR

d) Model Building (ANN)

An Artificial Neural Network (ANN) was developed using TensorFlow/Keras with:

- Input Layer
- Hidden Layers with ReLU activation
- Dropout layer for regularization
- Sigmoid output layer for binary classification

e) Model Training & Evaluation

The model was trained using scaled data and evaluated using:

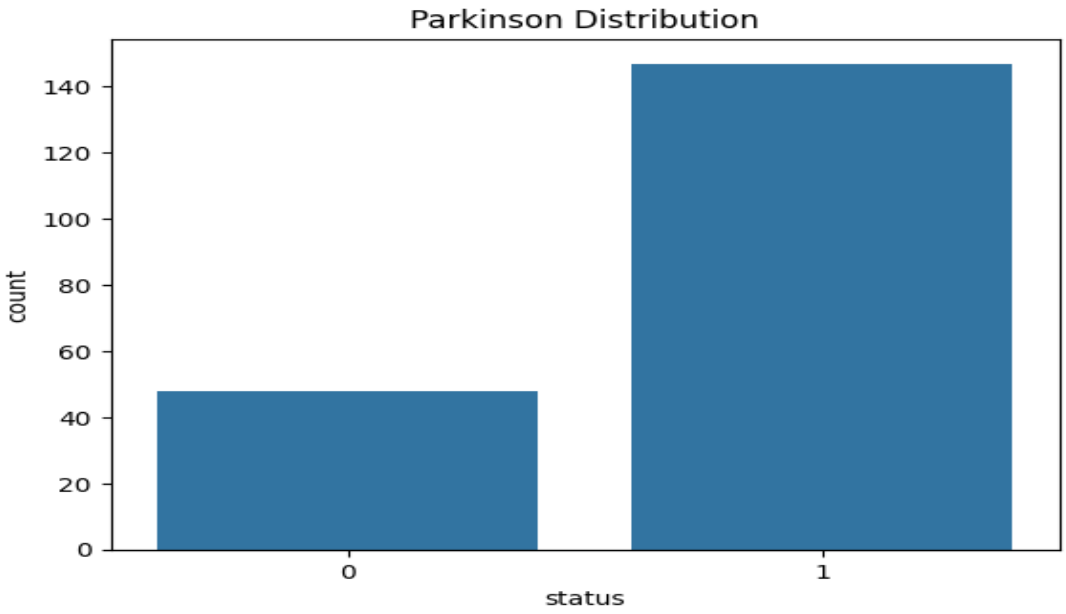
- Accuracy Score
- Confusion Matrix
- Classification Report

f) Streamlit Deployment

A Streamlit-based web application was created where users can input voice parameters and receive real-time Parkinson disease predictions.

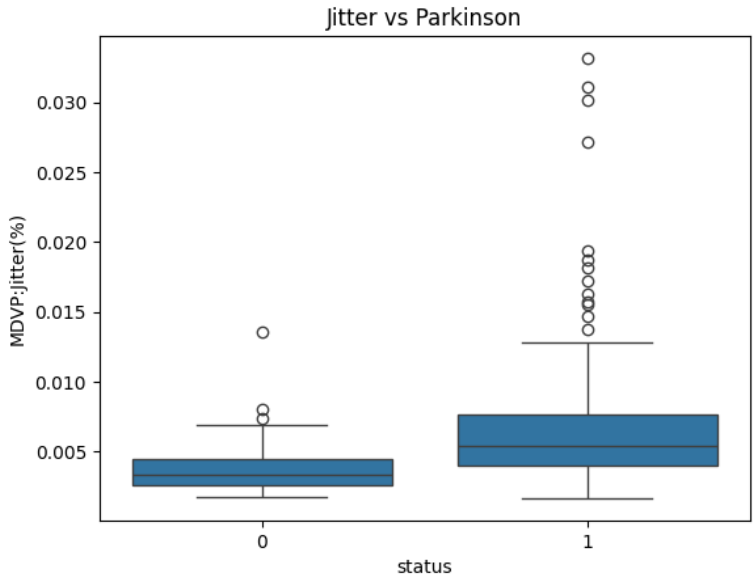
2.5.6 Results & Insights

i) Target Distribution



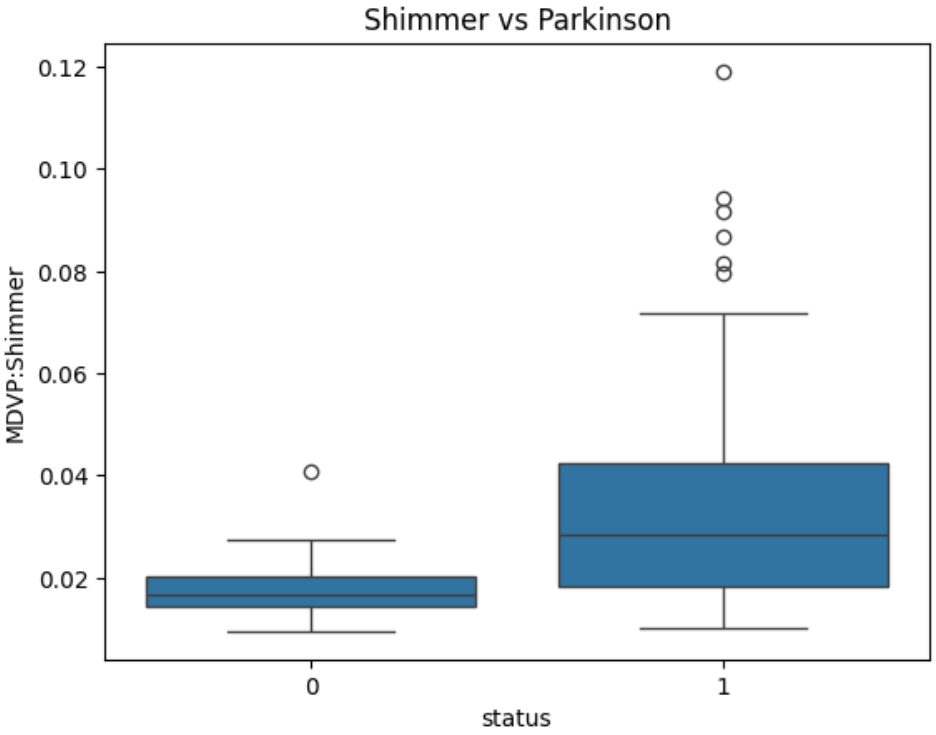
The dataset showed slight imbalance, with more Parkinson cases than healthy individuals.

ii) Jitter vs Parkinson



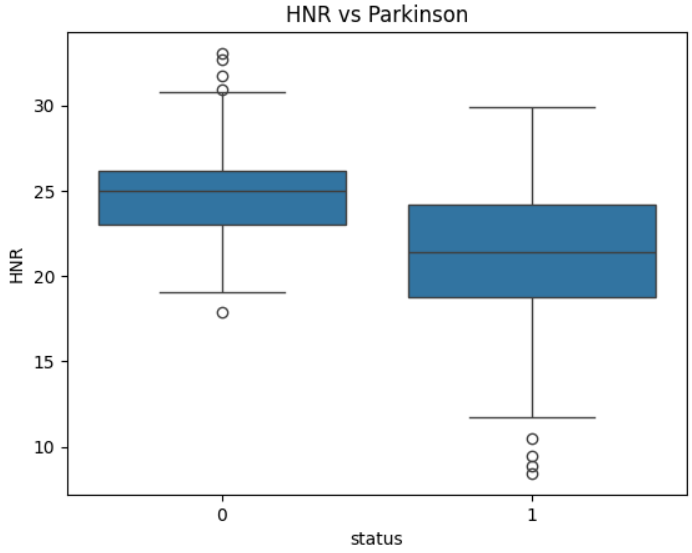
Parkinson patients generally showed higher jitter values, indicating instability in voice frequency.

iii) Shimmer vs Parkinson



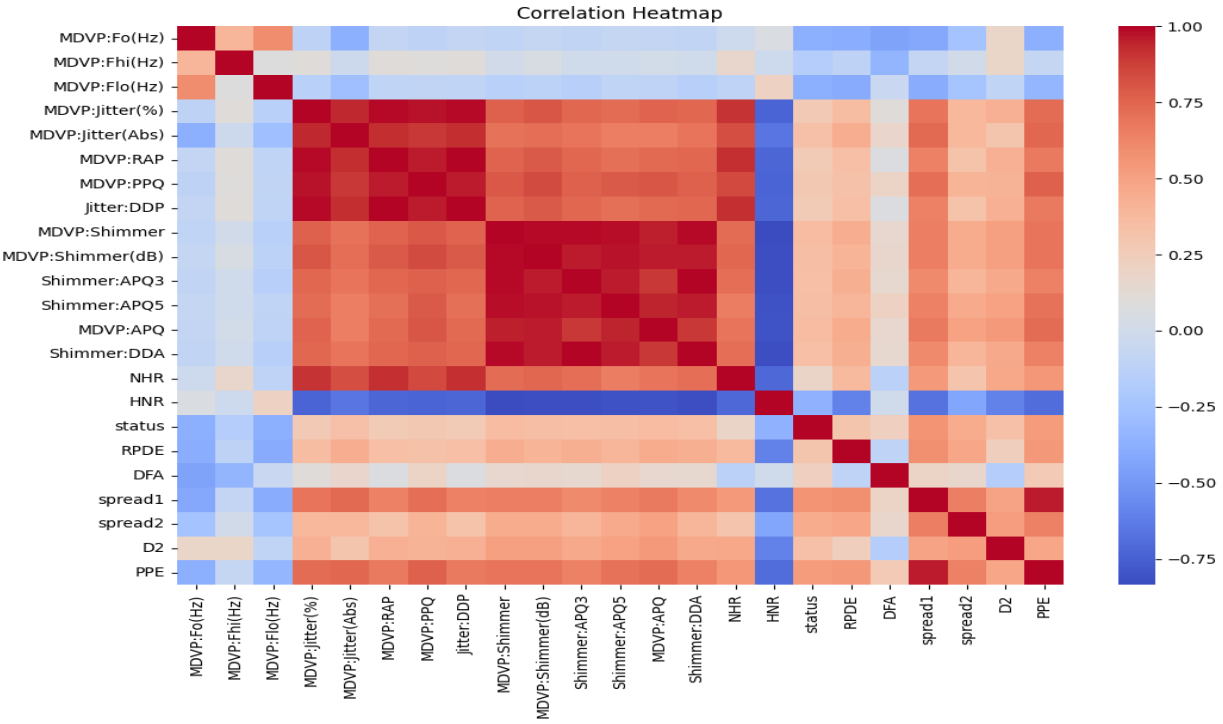
Higher shimmer values were observed among Parkinson patients, reflecting irregular voice amplitude patterns.

iv) HNR vs Parkinson



Healthy individuals showed higher HNR values, while Parkinson patients had lower HNR, indicating reduced voice clarity.

v) Correlation Analysis

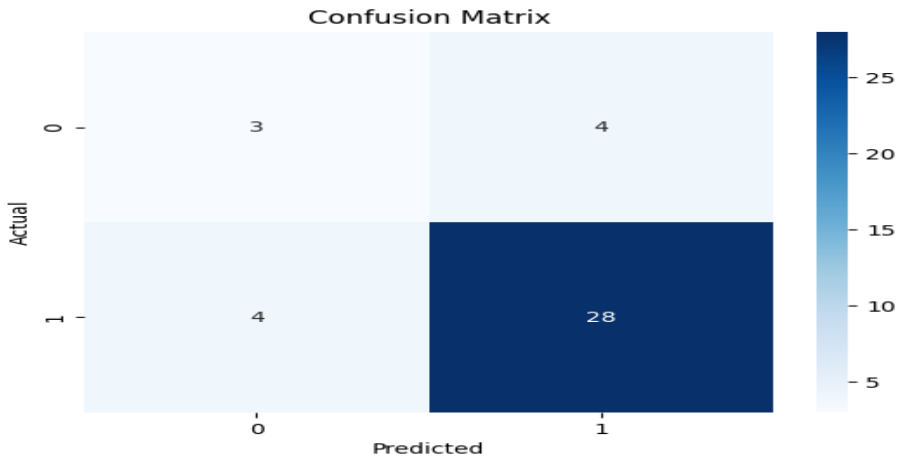


The heatmap revealed strong correlations among jitter and shimmer-related features, indicating interconnected voice abnormalities.

vi) Model Performance

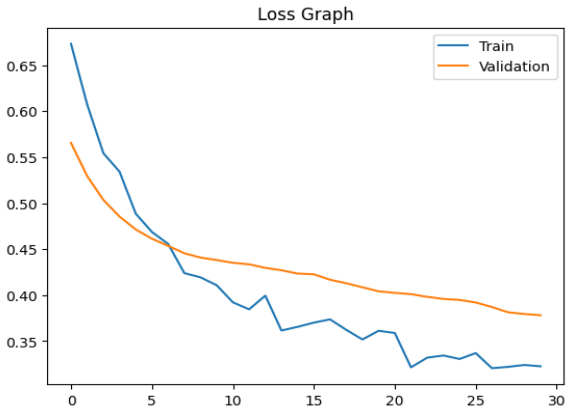
The ANN model achieved stable performance with good classification accuracy and minimal overfitting.

vii) Confusion Matrix



The confusion matrix showed that the model successfully classified most healthy and Parkinson cases with only a few misclassifications.

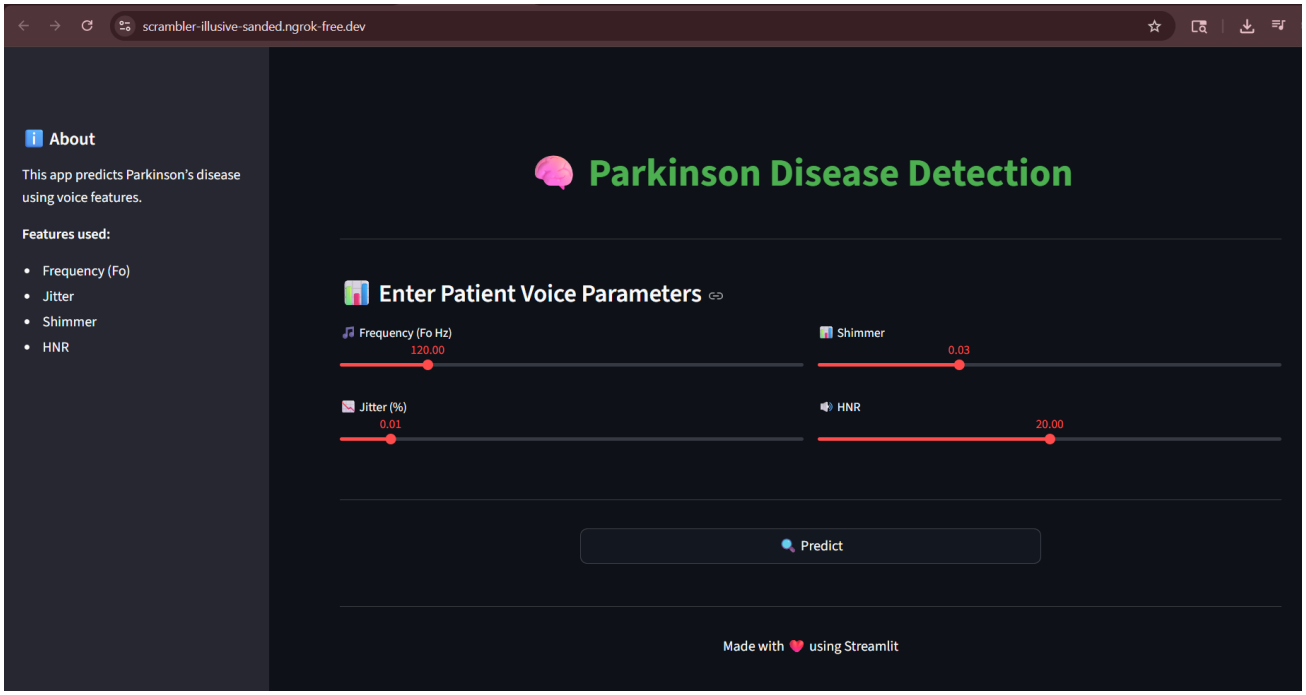
viii) Training & Validation Trends



Accuracy and loss graphs indicated stable model learning and smooth convergence during training.

ix) Streamlit Application

Figure : Streamlit-based Parkinson Disease Detection System



The above figure shows the Streamlit-based web application developed for Parkinson's Disease Detection using Machine Learning.

The application provides an interactive and user-friendly interface where users can enter important voice-related parameters such as Frequency (Fo), Jitter, Shimmer, and HNR using slider inputs.

After entering the values, the machine learning model processes the input data and predicts whether the patient is likely to have Parkinson's disease or not.

Features of the application include:

• Real-time prediction system

• Interactive slider-based input interface

• Simple and responsive UI design

• Machine learning model integration using ANN

• User-friendly healthcare prediction system

The deployment of the model using Streamlit makes the project more practical and demonstrates the complete end-to-end AI/ML workflow from model training to real-world application deployment.

2.5.7 Conclusion

This project successfully demonstrated the use of machine learning and deep learning techniques for Parkinson's disease detection using vocal features.

The analysis showed that jitter, shimmer, and HNR are strong indicators of Parkinson-related voice abnormalities. The Artificial Neural Network model achieved reliable classification performance and effectively distinguished healthy individuals from Parkinson patients.

The integration of the model into a Streamlit web application enhanced usability by allowing real-time disease prediction through a simple interface.

Overall, this project provided practical experience in healthcare AI applications, ANN model development, classification techniques, model evaluation, and deployment using Streamlit.

2.6 Financial Threat Monitoring Dashboard Using CTGAN (Week 6)

2.6.1 Introduction

Fraud detection is one of the most important applications of Artificial Intelligence and Machine Learning in the financial and banking sector. Detecting suspicious transactions helps organizations reduce financial losses and improve transaction security.

This project focuses on developing an AI-based Financial Threat Monitoring Dashboard using CTGAN (Conditional Tabular Generative Adversarial Network) and Machine Learning techniques. CTGAN was used to generate synthetic fraud transaction samples to improve fraud detection performance, especially in highly imbalanced datasets.

2.6.2 Objectives

- To analyze financial transaction datasets
- To identify suspicious transaction patterns
- To generate synthetic fraud data using CTGAN
- To handle class imbalance in fraud datasets
- To train Machine Learning models for fraud detection
- To visualize transaction risk analysis using graphs and dashboards

2.6.3 Tools & Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn
- CTGAN
- Streamlit
- Google Colab

2.6.4 Dataset Description

The dataset contains anonymized financial transaction records and fraud labels including:

• Transaction Amount
• Transaction Time
• PCA-transformed Features (V1–V28)
• Fraud Label

Target Variable:
• 0 → Legitimate Transaction
• 1 → Fraudulent Transaction

The dataset contains both fraudulent and non-fraudulent transaction records used for fraud analysis and classification.

2.6.5 Methodology

a) Data Collection

The financial transaction dataset was collected from publicly available fraud datasets.

b) Data Preprocessing

• Loaded dataset using Pandas
• Checked missing values
• Scaled numerical features
• Handled class imbalance
• Prepared dataset for model training

c) Exploratory Data Analysis (EDA)

EDA was performed using:
• Histograms
• Count plots
• Risk analysis charts
• Transaction distribution graphs

d) Synthetic Data Generation

CTGAN was used to generate synthetic fraud samples for balancing the dataset.

e) Machine Learning Model

Random Forest Classifier was used for fraud detection and transaction risk prediction.

f) Dashboard Deployment

A Streamlit dashboard was developed for real-time transaction analysis and fraud monitoring.

2.6.6 Results & Insights

i) Fraud Distribution Analysis

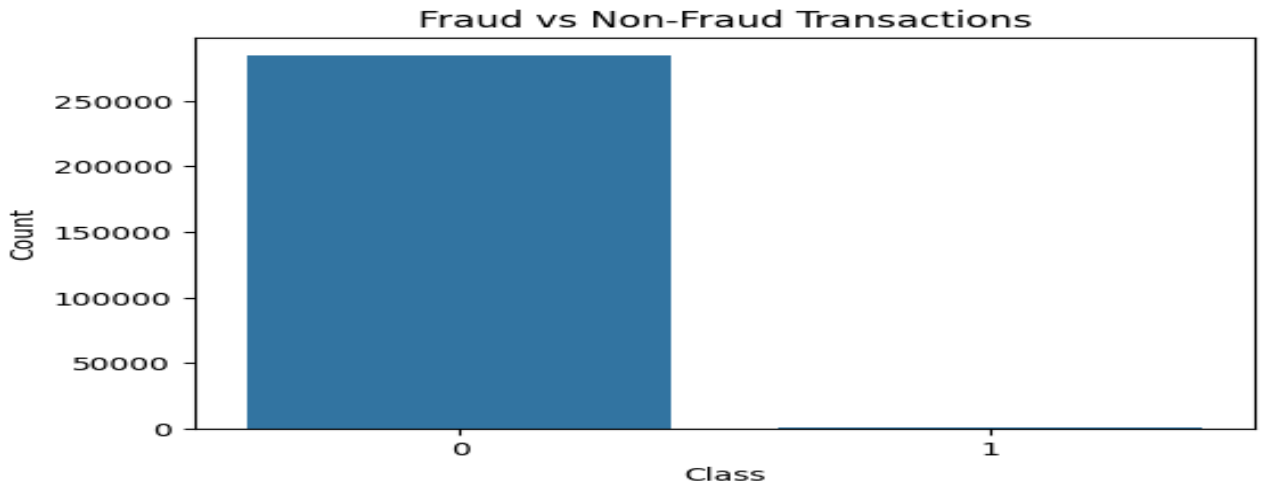


Figure 1: Fraud vs Non-Fraud Transactions

• The graph shows the distribution of fraudulent and non-fraudulent transactions.
• The dataset initially contained more normal transactions than fraud cases.

ii) Transaction Amount Analysis

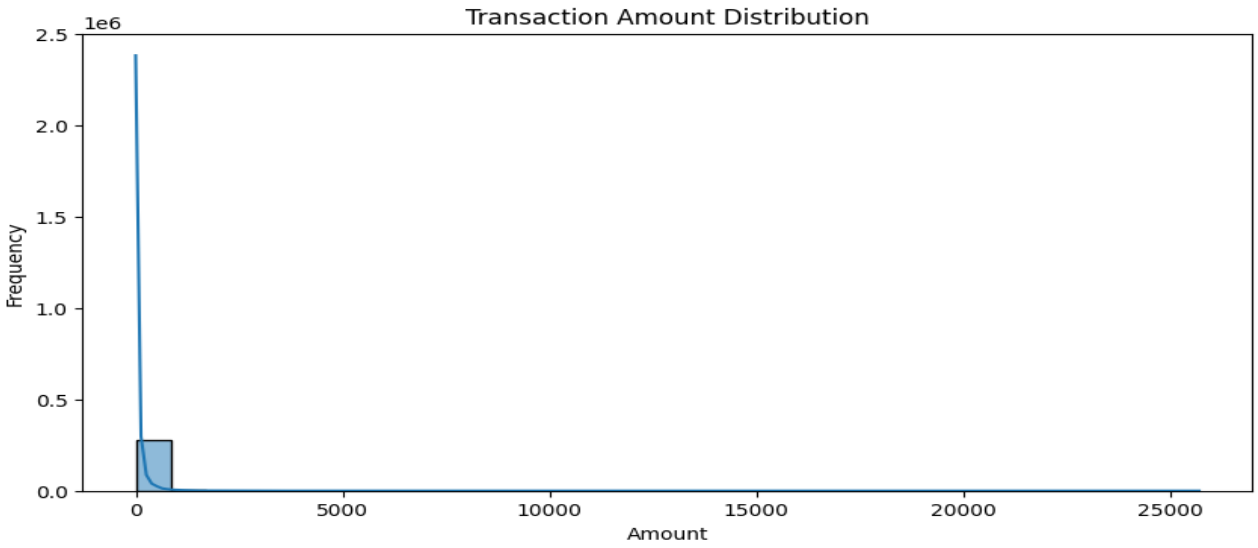


Figure 2: Transaction Amount Distribution

• Transaction amounts varied significantly across different records.
• The graph helps identify unusual transaction patterns related to fraud.

iii) CTGAN Synthetic Data Generation

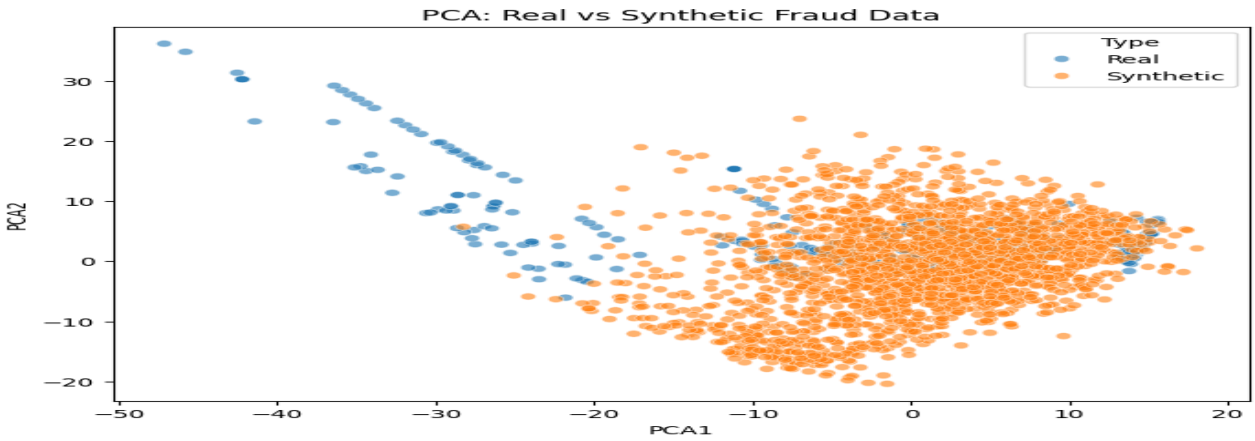


Figure 4: PCA Real vs Synthetic Fraud Data

• The graph compares real and synthetic fraud data generated using CTGAN.
• Synthetic data closely matched original transaction patterns.
ransactions existed across different transaction amounts, indicating that transaction amount alone is not a strong fraud indicator.

iv) Correlation Analysis

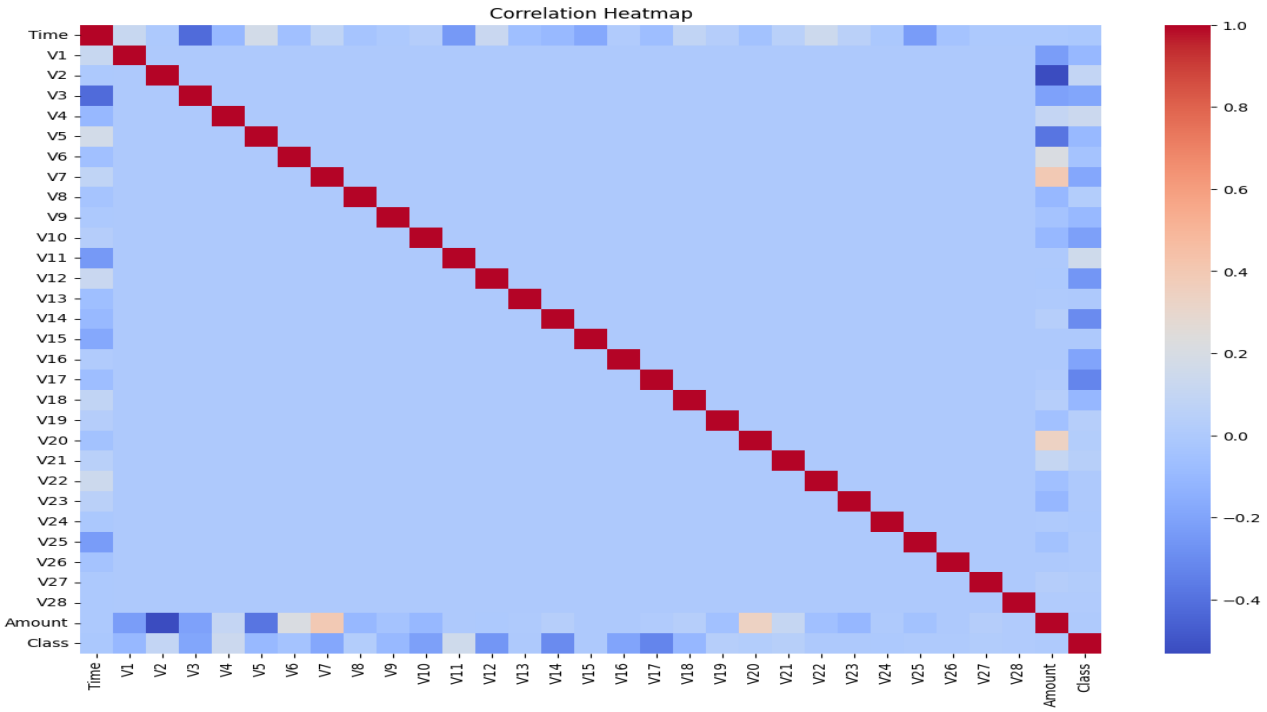


Figure 3: Correlation Heatmap

• The heatmap shows relationships among transaction-related numerical variables.
• Several variables showed meaningful correlation with fraudulent activity.

v) Fraud Comparison Analysis

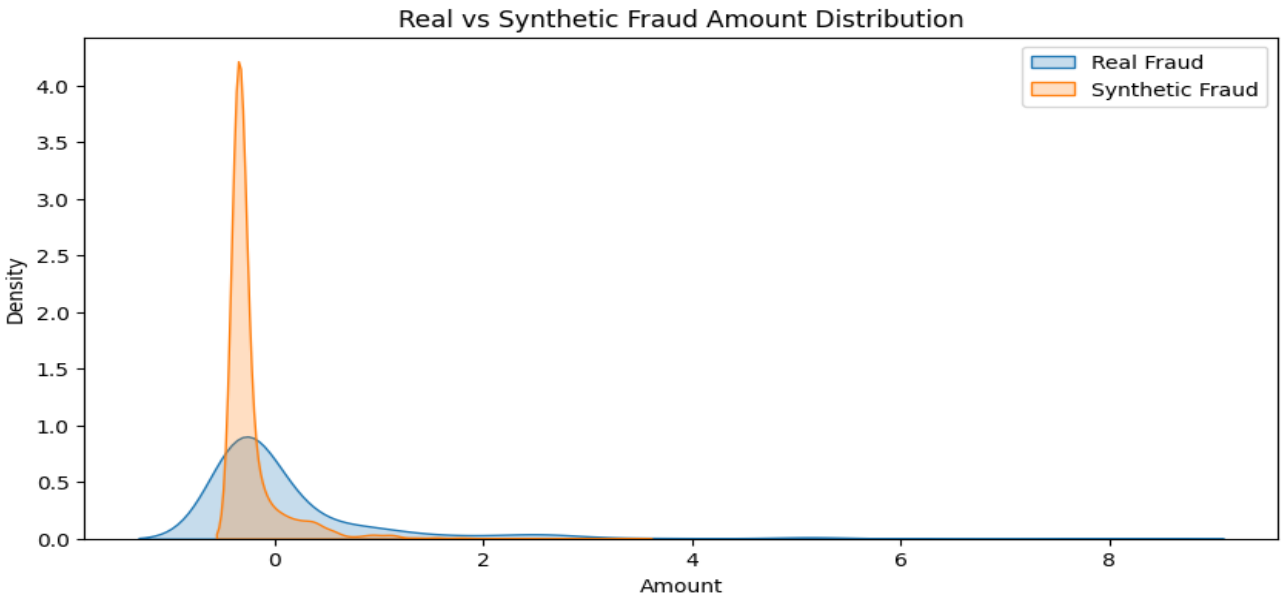


Figure 5: Real vs Synthetic Fraud Amount Distribution

- The graph compares original transaction data with synthetic fraud data.
- The synthetic data distribution closely matched the original dataset.

vi) Model Performance Analysis

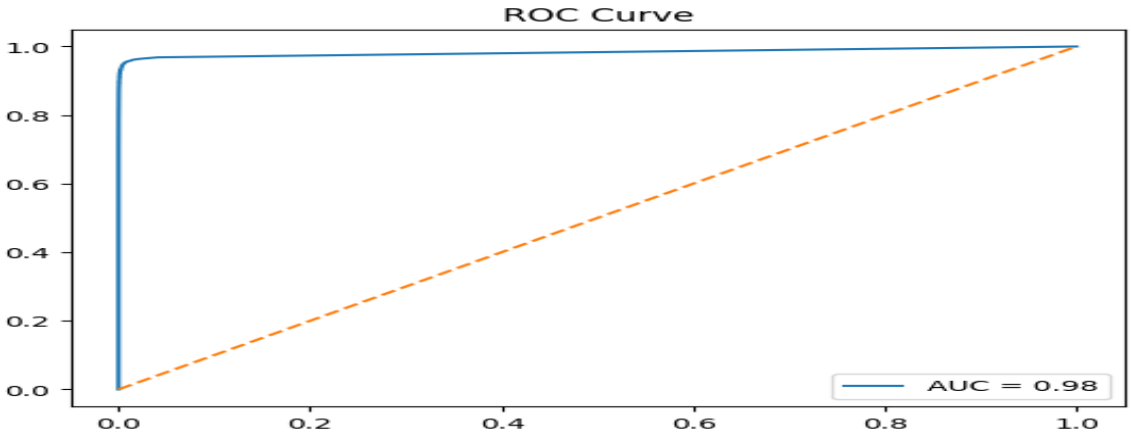


Figure 6: ROC Curve Comparison

- The ROC curve evaluates the fraud detection model performance.
- The model achieved high classification capability for fraud detection.

vii) Prediction Result Analysis

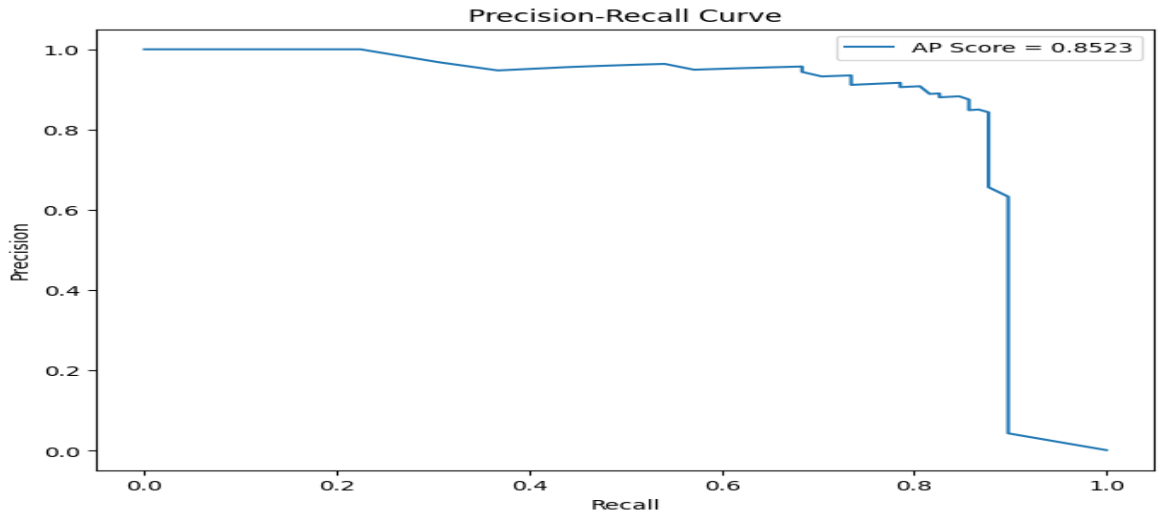


Figure 7: Precision-Recall Curve

- The precision-recall curve helps analyze prediction effectiveness.
- The graph shows strong fraud prediction performance.

viii) Final Fraud Detection Output

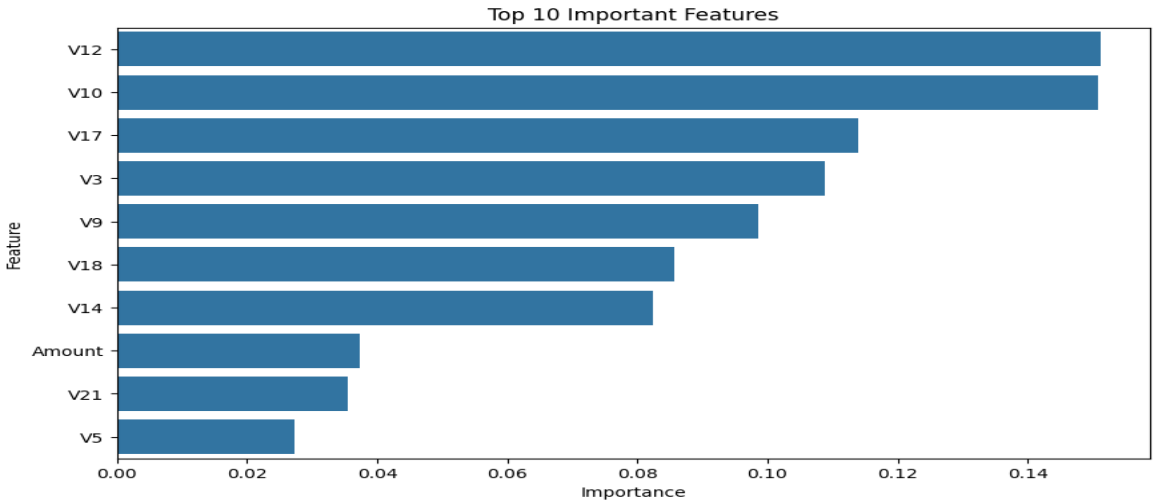


Figure 8: Top 10 Important Features

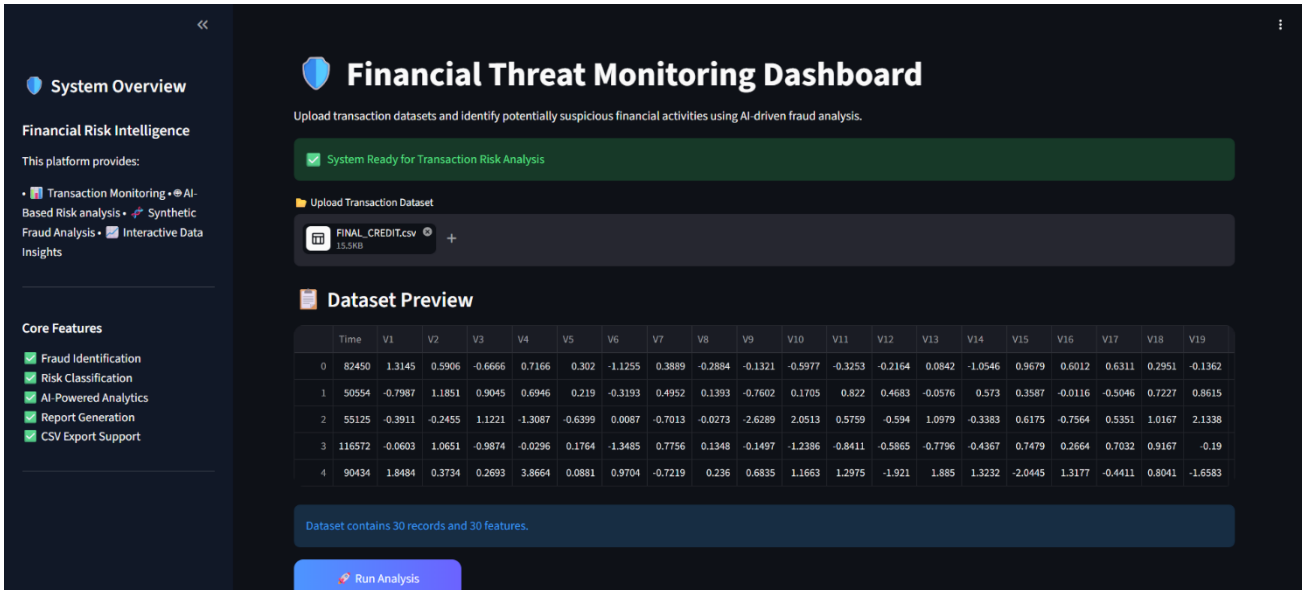
- The feature importance graph identifies major fraud detection factors.
- The final system successfully highlighted important fraud-related features.

ix) Streamlit Application

The deployed Streamlit application provided:

• Real-time fraud prediction
• Fraud probability visualization
• Interactive user interface for transaction analysis

Figure : Financial Threat Monitoring Streamlit Application:



The figure represents the final deployment stage of the **Financial Threat Monitoring Dashboard** through a Streamlit web application. The dashboard provides a modern, interactive, and user-friendly interface for analyzing uploaded financial transaction datasets and identifying potentially suspicious transactions in real time.

The deployed application integrates **Machine Learning** and **CTGAN-generated synthetic fraud data** to improve fraud detection performance and effectively handle the class imbalance problem commonly present in financial datasets. The system utilizes a trained **Random Forest Classifier** to classify transactions into legitimate and fraudulent categories based on multiple transaction-related features.

The dashboard includes several functionalities such as:

- Transaction dataset upload
- Dataset preview and transaction inspection
- AI-based fraud risk analysis
- Fraud probability prediction
- Risk classification (Low, Medium, High)
- Interactive visualizations and analytics
- Downloadable fraud analysis reports

The dashboard interface was designed using a dark-themed modern UI to improve user experience and provide professional financial monitoring visuals. Interactive charts and transaction analysis components were integrated to simplify fraud monitoring and enhance analytical interpretation.

The Streamlit deployment successfully converts the Machine Learning pipeline into a practical real-world web application that improves accessibility, scalability, usability, and demonstration capability for financial fraud detection systems.

Finally, the project was successfully deployed using **Streamlit Cloud**, allowing users to access the Financial Threat Monitoring Dashboard through a web browser for real-time fraud analysis and transaction risk monitoring.

Live Link : https://financial-threat-monitoring-dashboard-using-ctgan.streamlit.app/

2.6.7 Conclusion

This project successfully developed an AI-Based Financial Threat Monitoring Dashboard using CTGAN and Machine Learning techniques for fraud detection in financial transaction datasets. The system effectively handled class imbalance using synthetic fraud generation and improved fraud prediction performance.

The Streamlit dashboard enabled real-time transaction analysis, fraud risk monitoring, interactive visualizations, and downloadable reports. Overall, the project provided practical experience in fraud detection, synthetic data generation, machine learning, and dashboard deployment using Python and Streamlit. 🚀

CHAPTER 3: CONCLUSION

5.1 Overall Learning Outcomes

The six-week internship in the domain of Artificial Intelligence and Machine Learning provided practical exposure to the complete workflow of machine learning and data analysis projects, including data preprocessing, exploratory data analysis (EDA), clustering, classification, visualization, synthetic data generation, and deployment techniques.

Through six different projects, I gained hands-on experience in Python programming, machine learning, data visualization, clustering algorithms, fraud detection systems, and healthcare analytics. The internship involved working with real-world datasets from multiple domains such as sports analytics, healthcare, customer segmentation, automobile analysis, Parkinson's disease prediction, and financial fraud detection.

Each project helped improve my understanding of supervised and unsupervised learning techniques including classification, clustering, correlation analysis, and predictive modeling. I also learned how to handle real-world challenges such as missing values, imbalanced datasets, feature selection, data visualization, and model evaluation.

The major projects on Parkinson's Disease Detection and Financial Threat Monitoring Dashboard Using CTGAN provided practical exposure to machine learning, Artificial Neural Networks (ANN), Generative AI concepts, and real-time financial fraud monitoring systems. The fraud detection project also included deployment using Streamlit Cloud and interactive dashboard integration for transaction analysis, fraud prediction, and risk monitoring.

Throughout the internship, I improved my technical knowledge in Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, CTGAN, and machine learning workflows. The internship also enhanced my analytical thinking, problem-solving ability, report preparation, presentation skills, and understanding of real-world Artificial Intelligence and Machine Learning applications. 🚀

SUMMARY

The internship provided practical exposure to Artificial Intelligence and Machine Learning concepts through the implementation of multiple real-world projects using Python and various data science and machine learning libraries. During the internship, hands-on experience was gained in data preprocessing, exploratory data analysis (EDA), clustering, classification, machine learning model development, visualization, synthetic data generation, and deployment techniques.

Each week focused on solving real-world problems using structured AI/ML workflows including data collection, preprocessing, analysis, visualization, model development, evaluation, and deployment.

Across the six projects, the work covered multiple AI/ML domains:
• **Week 1 (Olympic Medal Count Analysis by Country)** – Performed exploratory data analysis on Olympic medal datasets using Python to analyze country-wise medal trends, sports performance, medal distribution, and historical Olympic patterns through data visualization and statistical analysis.

• **Week 2 (Heart Disease Risk Analysis)** – Analyzed healthcare datasets using Python to identify important heart disease risk factors such as age, cholesterol, blood pressure, chest pain type, and heart rate using data preprocessing, correlation analysis, and visualization techniques.

• **Week 3 (Customer Segmentation using K-Means Clustering)** – Applied unsupervised machine learning techniques to segment retail customers based on annual income and spending behavior using K-Means clustering, Elbow Method analysis, and customer cluster visualization.

• **Week 4 (Vehicle Feature Clustering using Machine Learning)** – Performed clustering and feature analysis on automobile datasets to group vehicles based on performance-related attributes such as horsepower, engine size, weight, and fuel efficiency using clustering techniques and visualization methods.

• **Week 5 (Parkinson's Disease Detection using Machine Learning)** – Developed a healthcare prediction system for Parkinson's disease detection using machine learning and Artificial Neural Network (ANN) techniques based on biomedical voice measurement data. The project included preprocessing, feature scaling, model training, evaluation, and prediction analysis.

• **Week 6 (Financial Threat Monitoring Dashboard Using CTGAN)** – Developed an AI-based financial fraud monitoring system using Random Forest Classification and CTGAN-generated synthetic fraud data. The project focused on handling class imbalance problems in financial transaction datasets and included transaction risk analysis, fraud prediction, interactive dashboard visualization, and deployment using Streamlit Cloud for real-time fraud monitoring and analysis.

Throughout the internship, strong technical and analytical skills were developed in Python programming, machine learning, clustering, data visualization, feature engineering, model evaluation, healthcare analytics, fraud detection, and AI-based system development.

The internship significantly improved problem-solving ability, analytical thinking, report preparation, and practical implementation skills while providing real-world exposure to end-to-end Artificial Intelligence and Machine Learning workflows. Overall, the internship successfully strengthened practical knowledge in Artificial Intelligence and Machine Learning and improved readiness for future projects and opportunities in the AI/ML domain.

REFERENCES

1. Kaggle Datasets – Olympic Medal Count Analysis, Heart Disease Risk Analysis, Customer Segmentation, Vehicle Feature Clustering, Parkinson's Disease Detection, and Financial Fraud Detection datasets.

2. UCI Machine Learning Repository – Heart Disease Dataset and Parkinson's Disease biomedical voice measurement dataset.

3. Scikit-learn Documentation – Machine learning algorithms, clustering techniques, Random Forest Classification, preprocessing methods, and model evaluation techniques.

4. CTGAN Documentation – Synthetic fraud data generation and handling imbalanced financial datasets using Generative AI techniques.

5. Streamlit Documentation – Development and deployment of interactive AI-based financial fraud monitoring dashboards.

6. Python Libraries Documentation :
   • Pandas Documentation
   • NumPy Documentation
   • Matplotlib Documentation
   • Seaborn Documentation

7. Research Articles & Online Resources –
   • Machine Learning Applications in Healthcare
   • AI-based Financial Fraud Detection Systems
   • Customer Segmentation using Clustering
   • Data Visualization and Exploratory Data Analysis
   • Generative AI for Synthetic Data Generation