AI-ML Internship

A Project Report submitted to the

GLOBAL NEXT CONSULTING INDIA PVT LTD

(Six – Week Internship Program)

By

Pranav Thakur

Under the Supervision of

Dr. Anuradha Gupta
(Project Director)

Submitted To :

Global Next Consulting India Pvt. Ltd.

Duration of Internship :

23-March-2026 to 9-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.

Pranav Thakur

CERTIFICATE

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

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.

Pranav Thakur

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 building intelligent systems using real-world datasets across multiple domains such as healthcare, fraud detection, music analytics, and customer segmentation.

The projects involved data preprocessing, exploratory data analysis, feature engineering, and the development of machine learning models for classification, regression, and clustering tasks. Key implementations include Healthcare Cost Prediction, Insurance Fraud Detection, Spotify Popularity Prediction, Vehicle Segmentation using clustering, Parkinson Disease Detection, and Credit Card Fraud Detection System.

Additionally, model deployment was performed using Streamlit to create user-friendly applications for real-time predictions. The internship enhanced my understanding of end-to-end machine learning pipelines and strengthened my skills in Python, Scikit-learn, and model evaluation techniques.

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 Healthcare Cost Analysis by Country (Week 1)

2.1.1 Introduction

Healthcare expenditure plays a significant role in improving the quality of life and increasing life expectancy across countries. Understanding how healthcare investment affects population health can help governments and policymakers make better healthcare decisions.

This project focuses on analyzing the relationship between healthcare expenditure per capita and life expectancy across multiple countries using Python-based data analysis techniques. The analysis was performed using datasets collected from reliable global sources such as the World Health Organization (WHO) and OECD. The project includes data preprocessing, exploratory data analysis (EDA), visualization, and interpretation of healthcare trends across countries.

2.1.2 Objectives

- To analyze healthcare expenditure patterns across countries
- To study the relationship between healthcare spending and life expectancy
- To preprocess and merge healthcare datasets for analysis
- To perform exploratory data analysis using Python
- To visualize healthcare trends using graphs and charts
- To identify countries with high healthcare expenditure and life expectancy
- To generate meaningful insights from global healthcare data

2.1.3 Tools & Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Google Colab / Jupyter Notebook

2.1.4 Dataset Description

The project uses two datasets:

- Healthcare Expenditure per Capita Dataset
- Life Expectancy Dataset

The datasets contain country-wise and year-wise information related to healthcare spending and average life expectancy. Since the healthcare expenditure dataset extended up to 2023 while the life expectancy dataset was available only till 2021, the analysis was restricted to common years up to 2021 for consistency.

2.1.5 Methodology

a) Data Collection
The datasets were collected from reliable global healthcare sources such as WHO and OECD.

b) Data Preprocessing
The following preprocessing steps were performed:
- Loaded datasets using Pandas
- Checked dataset structure and missing values
- Filtered healthcare expenditure data till 2021
- Merged both datasets using Country and Year columns
- Renamed columns for better readability
- Removed inconsistencies and handled missing values

c) Exploratory Data Analysis (EDA)
EDA was performed to understand trends and relationships within the data using:
- Line charts
- Scatter plots
- Histograms
- Heatmaps
- Boxplots
- Country-wise comparison charts

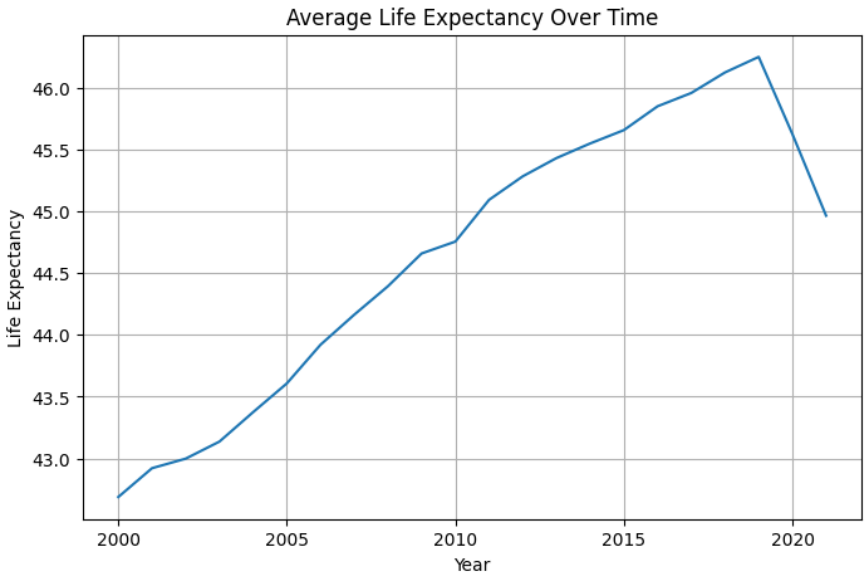
d) Data Visualization

Visualization techniques were used to analyze:

- Healthcare expenditure trends over time
- Life expectancy trends
- Correlation between expenditure and life expectancy
- Top countries by healthcare expenditure
- Distribution of healthcare expenditure across countries

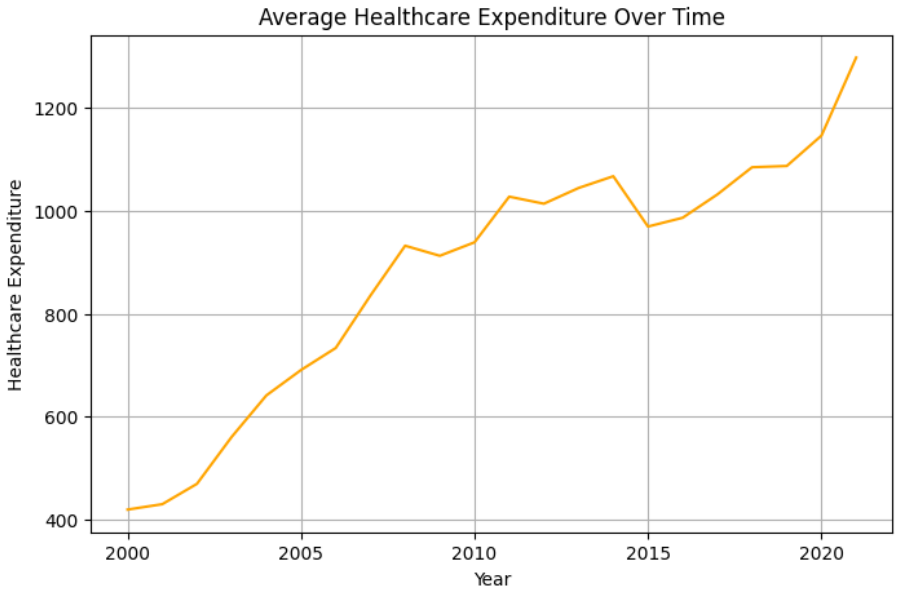
2.1.6 Results & Insights

i) Life Expectancy Trend



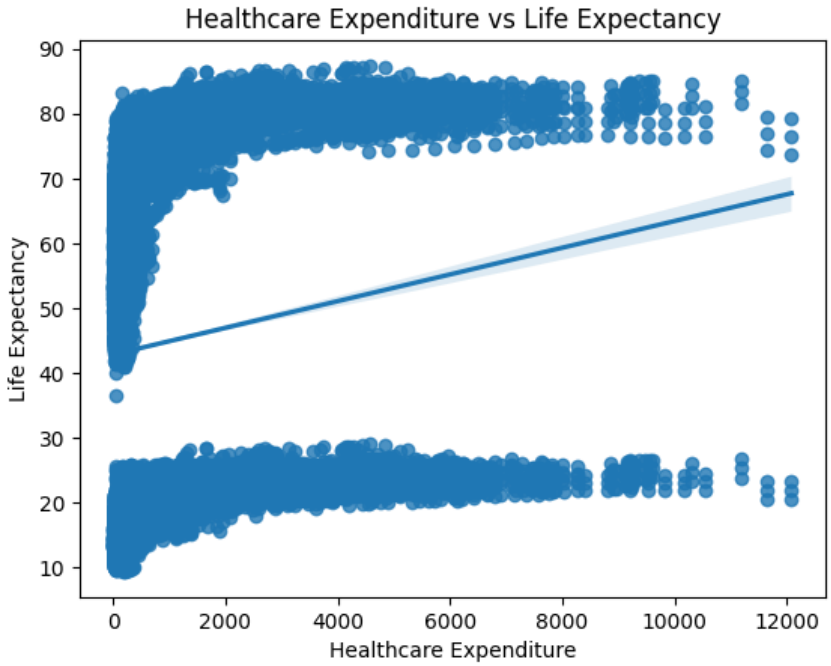
The analysis showed that average life expectancy has steadily increased over time across countries, indicating improvements in healthcare systems and living conditions.

ii) Healthcare Expenditure Trend



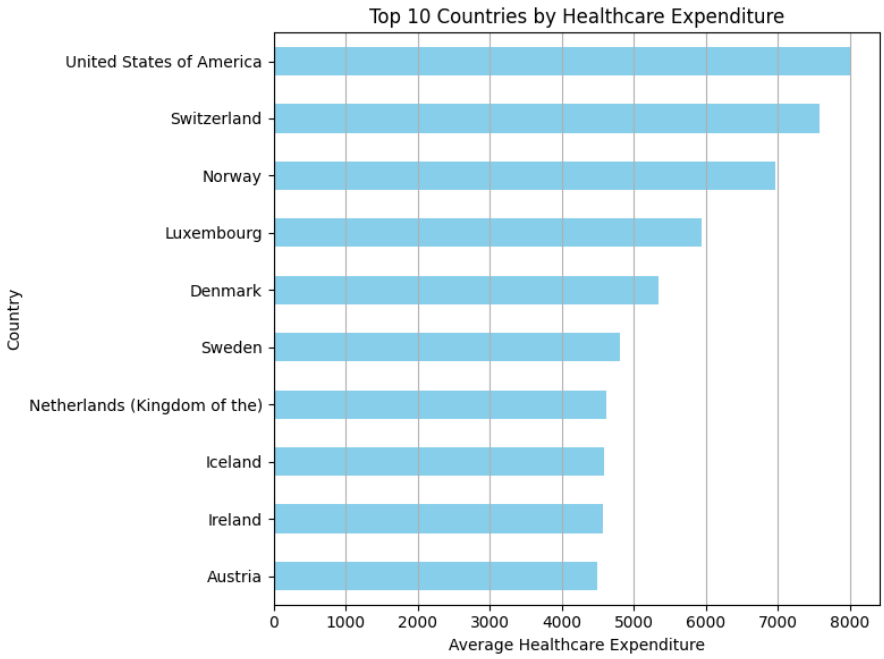
Healthcare expenditure per capita also showed a continuous upward trend, reflecting increased investment in healthcare infrastructure and services.

iii) Relationship Between Expenditure & Life Expectancy



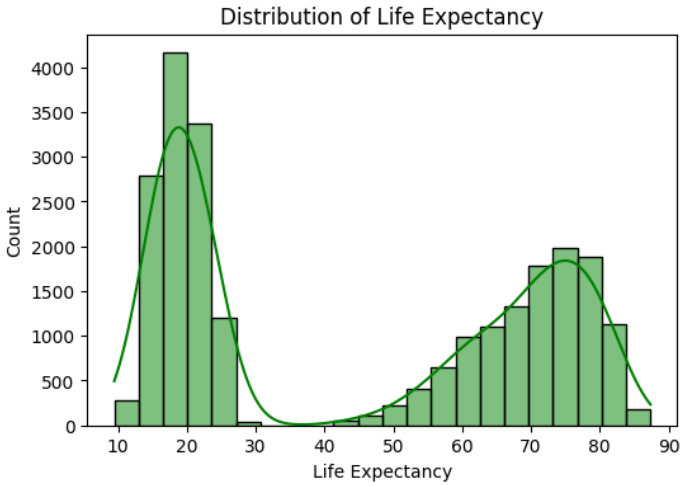
Scatter plot analysis revealed a positive correlation between healthcare expenditure and life expectancy. Countries spending more on healthcare generally recorded higher life expectancy values.

iv) Top Healthcare Spending Countries



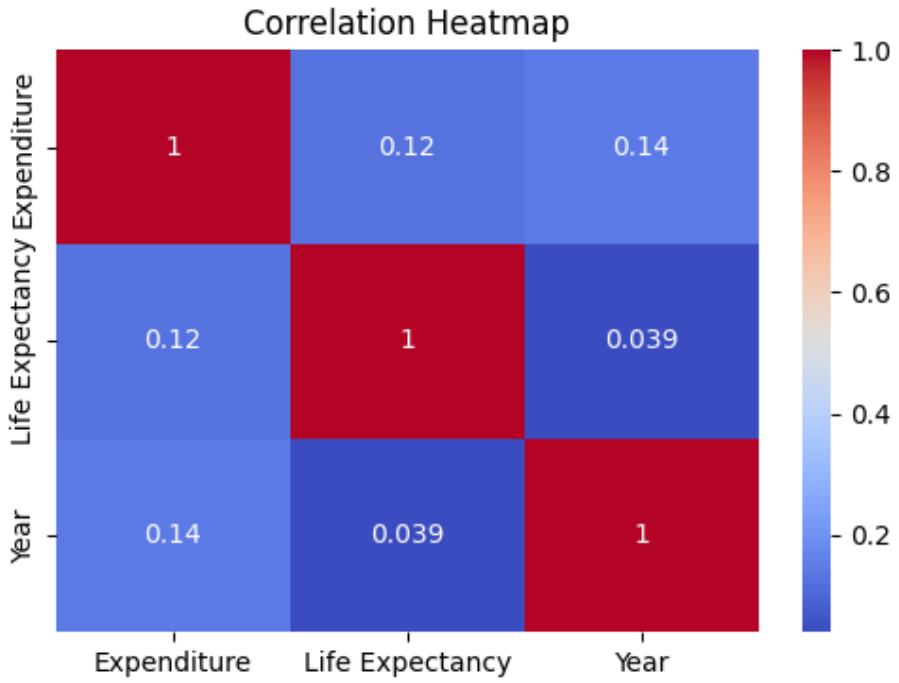
Developed countries dominated healthcare expenditure rankings, highlighting stronger economic capacity and healthcare infrastructure.

v) Distribution Analysis



Healthcare expenditure distribution was highly right-skewed, indicating that only a few countries spend significantly higher amounts on healthcare compared to others.

vi) Correlation Analysis



The correlation heatmap showed a strong positive relationship between healthcare expenditure and life expectancy, supporting the importance of healthcare investment.

2.1.7 Conclusion

This project successfully analyzed global healthcare expenditure and life expectancy trends using Python-based exploratory data analysis techniques.

The study revealed that countries with higher healthcare expenditure generally achieve better life expectancy outcomes. However, healthcare spending alone does not determine population health, as factors such as lifestyle, environment, and healthcare efficiency also influence outcomes.

Overall, the project demonstrates how data analytics and visualization techniques can be used to generate meaningful insights from global healthcare data and support better healthcare decision-making.

2.2 Insurance Claims & Fraud Analysis (Week 2)

2.2.1 Introduction

Insurance fraud is one of the major challenges faced by insurance companies, leading to significant financial losses every year. Detecting fraudulent claims manually is difficult due to the large volume and complexity of insurance data.

This project focuses on analyzing insurance claim data to identify hidden patterns, trends, and anomalies that may indicate fraudulent activities. The project involves data cleaning, preprocessing, exploratory data analysis (EDA), visualization, and dimensionality reduction using Principal Component Analysis (PCA). Various factors such as claim amount, customer age, incident type, property damage, and police reports were analyzed to understand their relationship with fraud occurrence.

2.2.2 Objectives

- To analyze insurance claims data for fraud-related patterns
- To preprocess and clean the dataset for analysis
- To identify relationships between claim features and fraud occurrence
- To perform exploratory data analysis using visualization techniques
- To detect outliers and unusual claim patterns
- To apply Principal Component Analysis (PCA) for dimensionality reduction
- To generate meaningful insights for insurance fraud detection

2.2.3 Tools & Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Plotly
- Scikit-learn
- Missingno

2.2.4 Dataset Description

The dataset contains insurance claim records with customer information, policy details, claim amounts, incident reports, and fraud labels.

Key attributes include:
• Customer Age
• Policy Premium
• Claim Amount
• Incident Type
• Property Damage
• Police Report Availability
• Fraud Reported Status
The dataset consists of approximately 1000 records and includes both numerical and categorical variables.

2.2.5 Methodology

a) Data Preprocessing
The following preprocessing steps were performed:
• Loaded the dataset using Pandas
• Removed duplicate records
• Handled missing values using median and mode imputation
• Replaced invalid symbols with null values
• Performed one-hot encoding for categorical variables
• Applied Min-Max Scaling for normalization

b) Exploratory Data Analysis (EDA)
EDA was performed to understand fraud patterns and claim behavior using:
• Count plots
• Histograms
• Scatter plots
• Heatmaps
• Boxplots

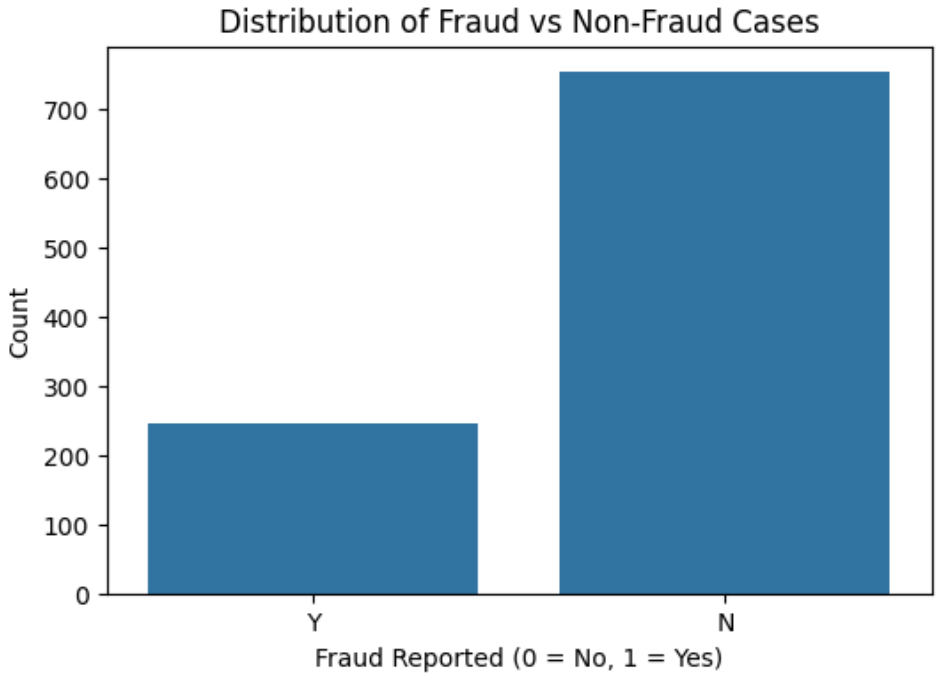
c) Outlier Detection
Boxplots were used to identify extreme claim values and unusual customer behavior.

d) Correlation Analysis
A correlation heatmap was created to identify relationships among numerical variables.

e) Principal Component Analysis (PCA)
PCA was applied to reduce dimensionality and visualize fraud vs non-fraud cases in two-dimensional space.

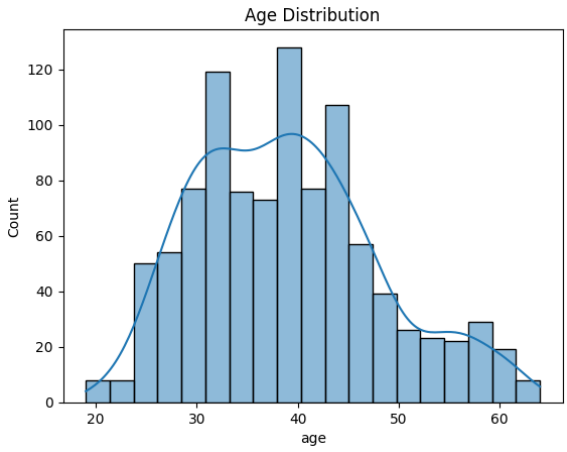
2.2.6 Results & Insights

i) Fraud vs Non-Fraud Distribution



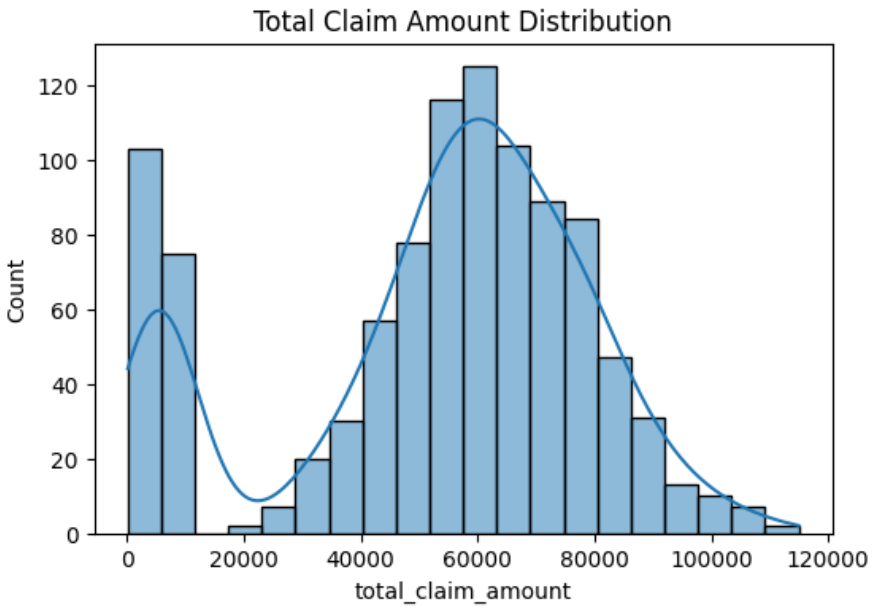
The analysis showed that non-fraudulent claims are significantly higher than fraudulent claims, indicating class imbalance in the dataset.

ii) Customer Age Distribution



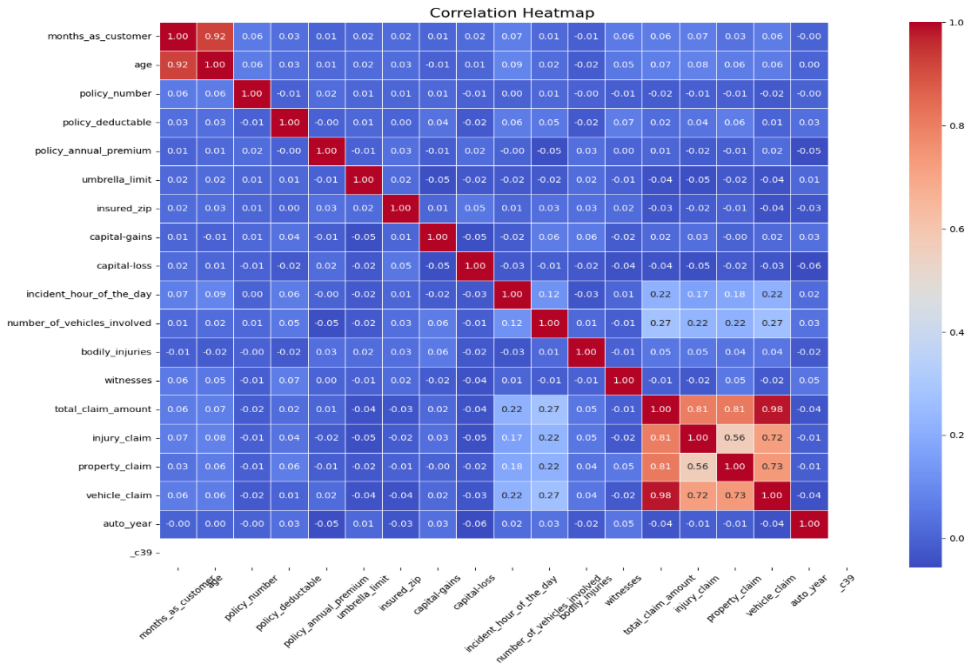
Most customers fall within the age group of 30–50 years, showing a balanced age distribution.

iii) Claim Amount Distribution



Most claim amounts were concentrated between moderate ranges, while a few extremely high claims appeared as outliers.

iv) Correlation Analysis

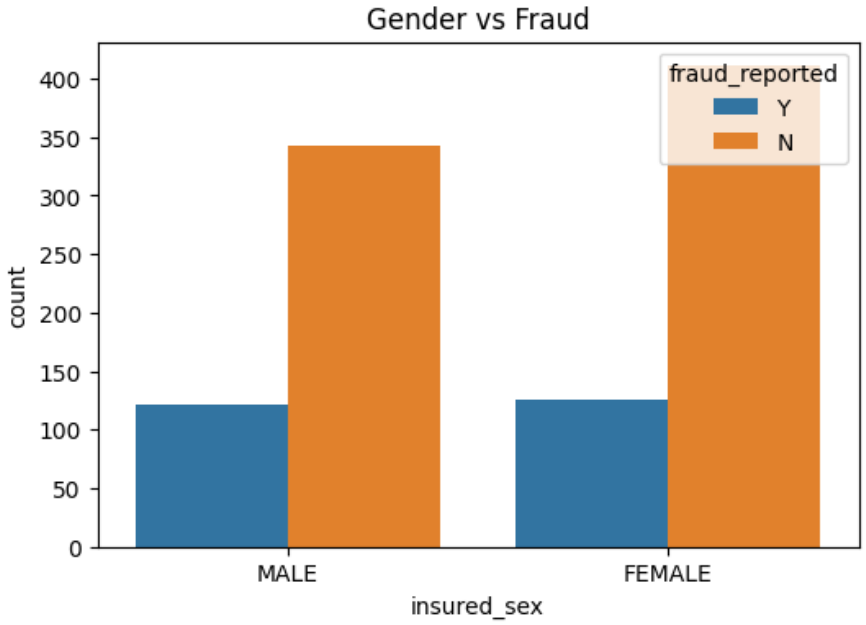


Strong correlations were observed among claim-related features such as:

- Total Claim Amount
- Injury Claim
- Vehicle Claim
- Property Claim

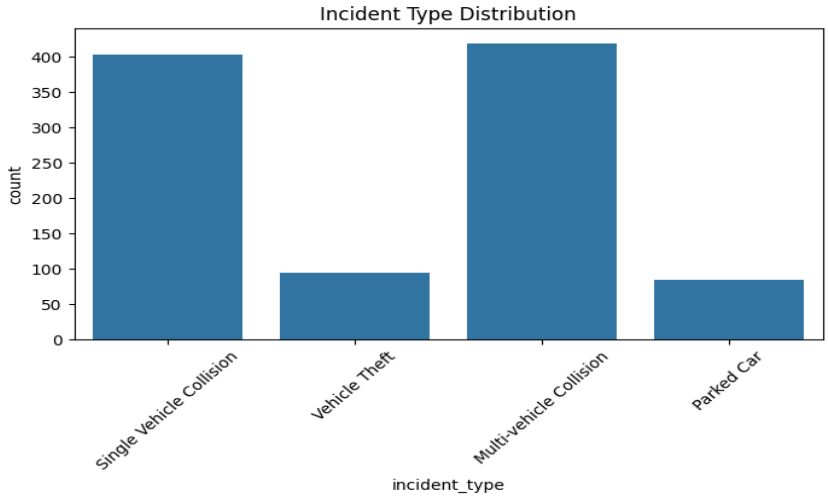
This indicates that total claim amount depends heavily on these components.

v) Gender vs Fraud Analysis



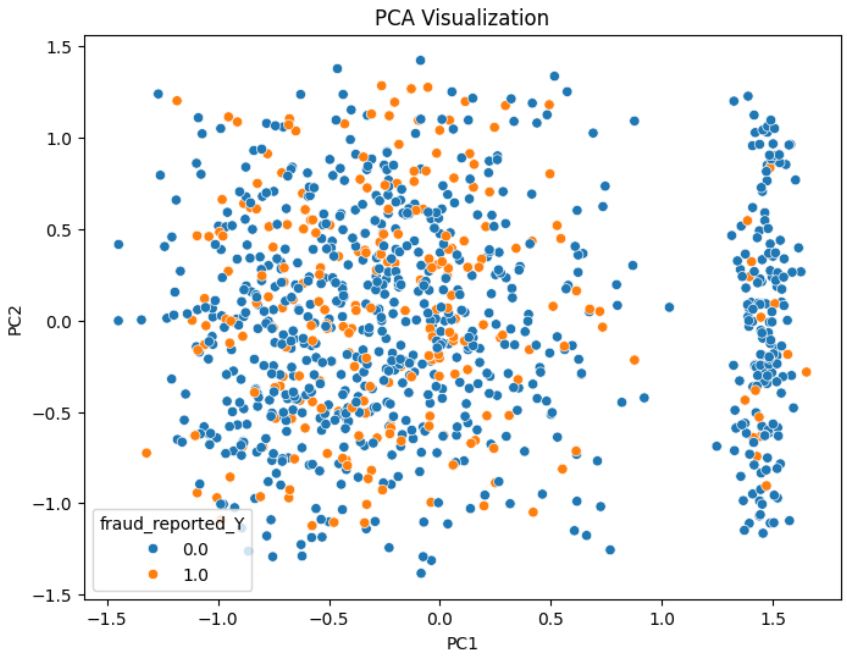
The analysis showed no significant difference in fraud occurrence based on gender.

vi) Incident Type Analysis



Vehicle collision incidents were the most common compared to theft and parked car incidents.

vii) PCA Visualization



The PCA results showed significant overlap between fraud and non-fraud cases, indicating that fraud detection is a complex problem requiring advanced machine learning techniques.

2.2.7 Conclusion

This project successfully analyzed insurance claims data to identify fraud-related patterns and customer claim behavior.

The study revealed that fraudulent claims form a smaller portion of the dataset, while most claims are genuine. Features such as claim amount, incident type, and claim-related variables showed useful patterns for fraud analysis, whereas demographic factors like gender had minimal influence.

The PCA analysis demonstrated that fraud detection cannot rely on simple feature separation alone and may require more advanced machine learning approaches for accurate prediction.
Overall, this project provided valuable insights into insurance fraud analysis using Python-based exploratory data analysis and preprocessing techniques.

2.3 Spotify Popularity Prediction Analysis (Week-3)

2.3.1 Introduction

Music streaming platforms such as Spotify generate massive amounts of data related to songs, artists, and listener preferences. Predicting song popularity is a challenging task because popularity depends on multiple factors including audio characteristics, listener behavior, trends, and audience preferences.

This project focuses on analyzing Spotify music data to understand the relationship between various audio features and song popularity. The project includes data preprocessing, exploratory data analysis (EDA), dimensionality reduction using Principal Component Analysis (PCA), and machine learning model building for popularity prediction. Regression models such as Linear Regression and Random Forest Regressor were used to predict song popularity based on audio features like danceability, energy, loudness, tempo, and acousticness.

2.3.2 Objectives

• To analyze Spotify song data and identify popularity patterns
• To preprocess and clean music-related datasets
• To study relationships between audio features and popularity
• To perform exploratory data analysis using visualization techniques
• To apply PCA for dimensionality reduction and pattern analysis
• To build machine learning models for popularity prediction
• To evaluate model performance using regression metrics

2.3.3 Tools & Technologies Used

• Python
• Pandas

• NumPy
• Matplotlib
• Seaborn
• Scikit-learn
• Missingno

2.3.4 Dataset Description

The dataset contains Spotify track information including audio features and popularity scores.

Key attributes include:
• Track Name
• Artist Name
• Popularity Score
• Danceability
• Energy
• Loudness
• Tempo
• Acousticness
• Instrumentalness
• Valence
• Track Genre
The dataset includes both numerical and categorical features related to music characteristics and listener engagement.

2.3.5 Methodology

a) Data Preprocessing

The following preprocessing steps were performed:

• Loaded dataset using Pandas

• Removed duplicate tracks

• Dropped irrelevant columns

• Handled missing values in artist and track information

• Selected relevant features for machine learning

b) Exploratory Data Analysis (EDA)

EDA was conducted using multiple visualization techniques:

• Histograms

• Scatter plots

• Boxplots

• Correlation heatmaps

• Genre distribution charts

• Artist comparison charts

c) Principal Component Analysis (PCA)

PCA was applied after feature scaling to reduce dimensionality and analyze hidden patterns within the dataset.

d) Machine Learning Models

Two regression models were implemented:

• Linear Regression

• Random Forest Regressor

e) Model Evaluation

The models were evaluated using:

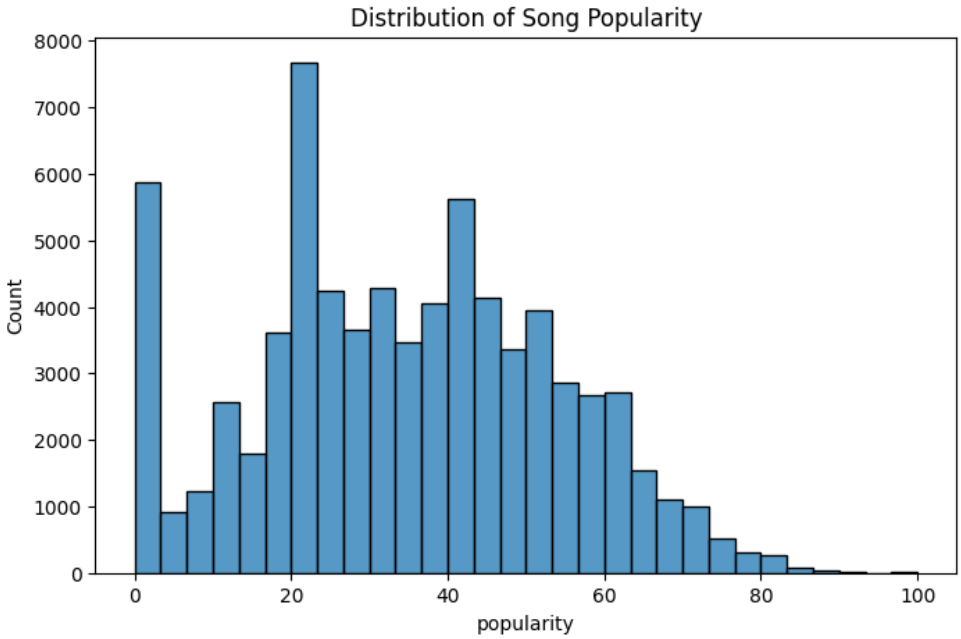
• R^2 Score

• Mean Squared Error (MSE)

• Mean Absolute Error (MAE)

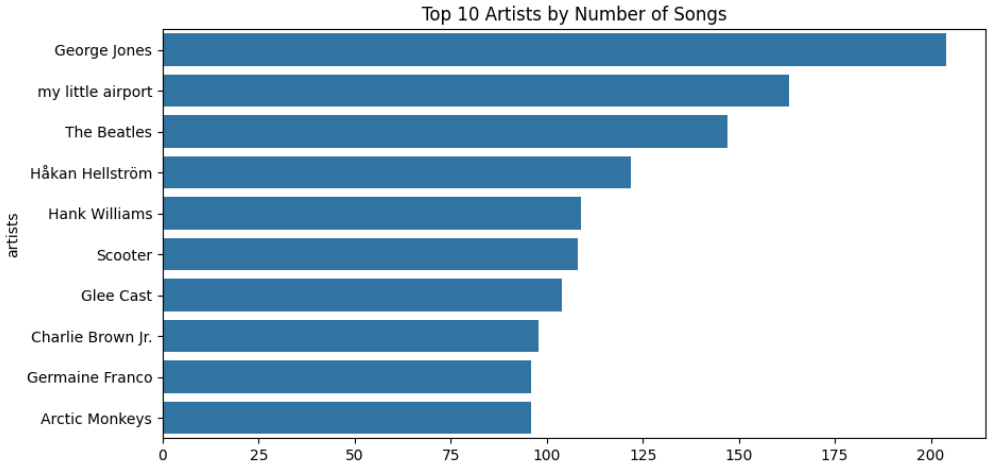
2.3.6 Results & Insights

i) Popularity Distribution



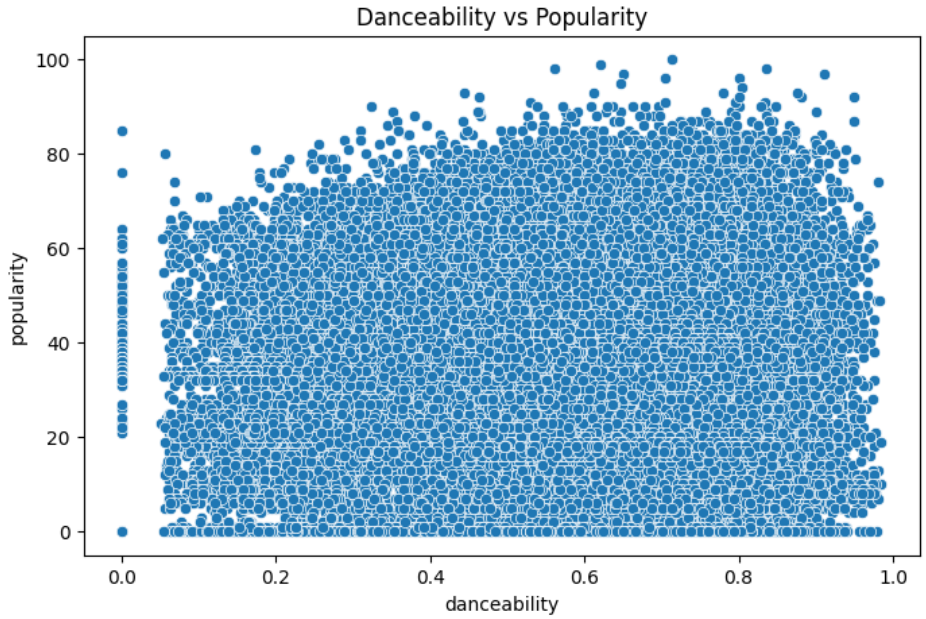
The popularity distribution showed that most songs fall within low to medium popularity ranges, while only a few songs achieve extremely high popularity.

ii) Top Artists Analysis



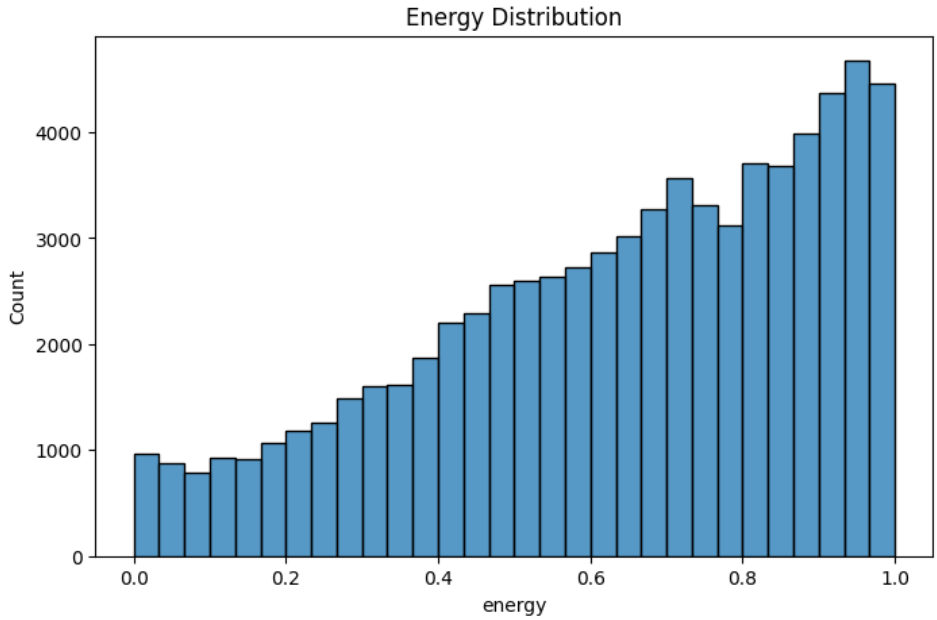
Some artists contributed a significantly larger number of tracks, indicating dataset imbalance toward certain artists.

iii) Danceability vs Popularity



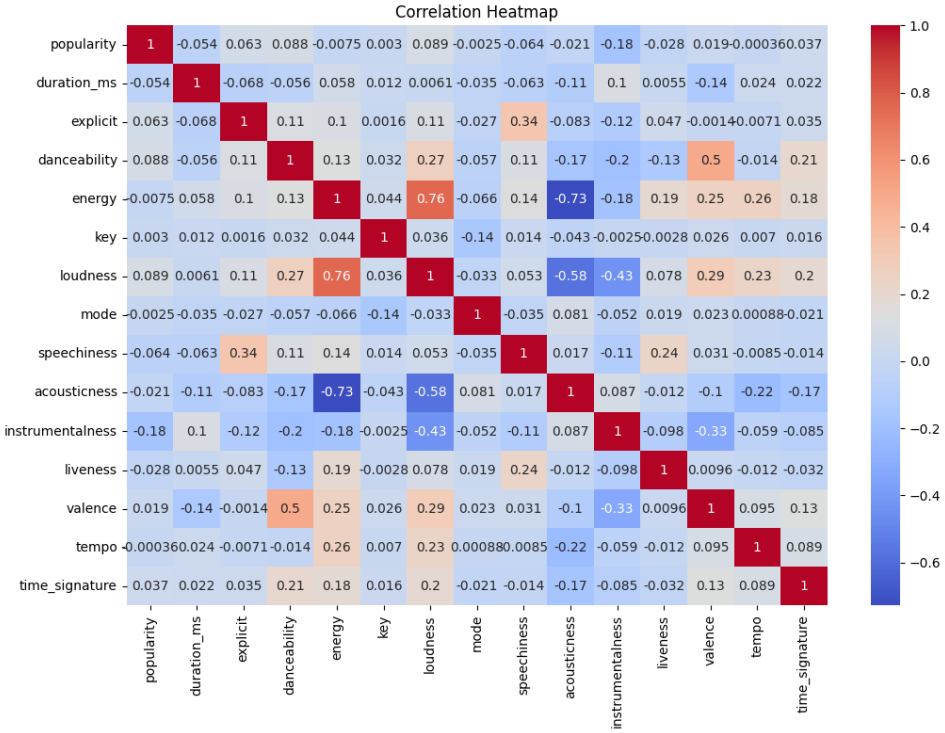
Scatter plot analysis revealed that higher danceability alone does not guarantee higher song popularity.

iv) Energy Distribution



Most songs in the dataset had relatively high energy levels, suggesting energetic tracks are common on Spotify.

v) Correlation Analysis

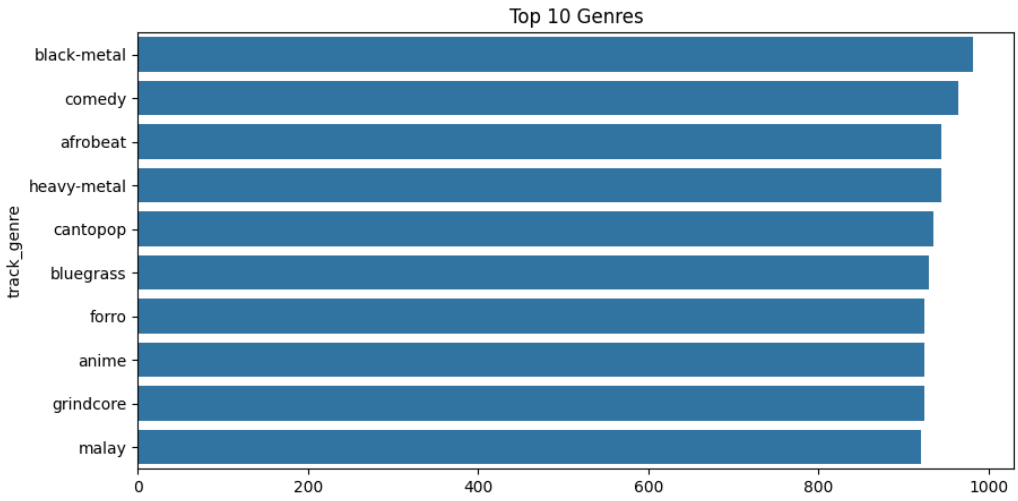


The heatmap showed weak correlations between popularity and most audio features. However, strong correlation existed between:

• Energy and Loudness

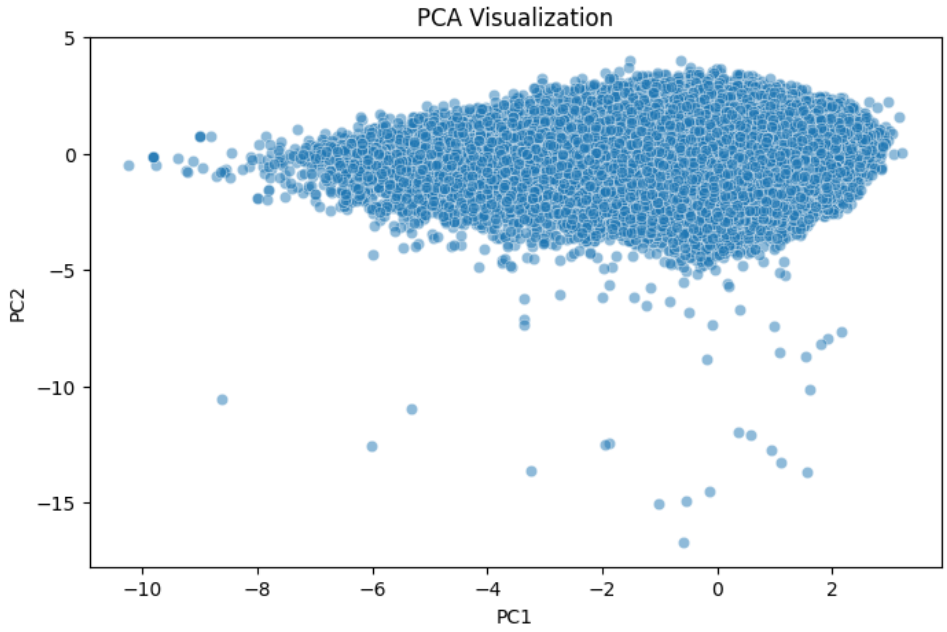
This indicates feature dependency among some audio characteristics.

vi) Genre Analysis



The dataset included multiple music genres, showing diversity in track categories and listener preferences.

vii) PCA Results



PCA visualization showed no clear separation or clustering in the data, indicating that song popularity is influenced by complex factors beyond measurable audio features.

viii) Machine Learning Performance

Both Linear Regression and Random Forest models showed limited predictive performance, with predicted popularity values not closely matching actual values.

2.3.7 Conclusion

This project successfully analyzed Spotify music data and explored the relationship between audio features and song popularity using Python and machine learning techniques.

The study revealed that popularity prediction is a complex problem because most audio features show weak correlation with popularity. Although machine learning models were implemented, the prediction accuracy remained limited due to the influence of external factors such as audience trends, marketing, and listener behavior.

Overall, the project demonstrated the challenges of music popularity prediction and provided practical experience in exploratory data analysis, PCA, feature engineering, and regression modeling using real-world Spotify datasets.

2.4 Vehicle 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

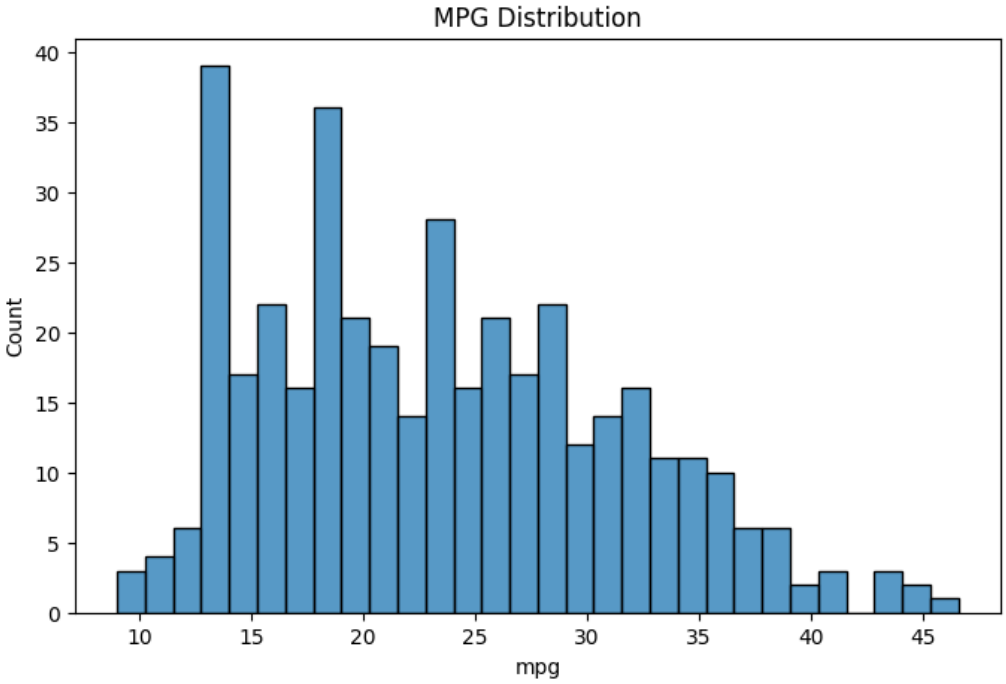
These techniques helped identify the most suitable number of clusters.

f) Principal Component Analysis (PCA)

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

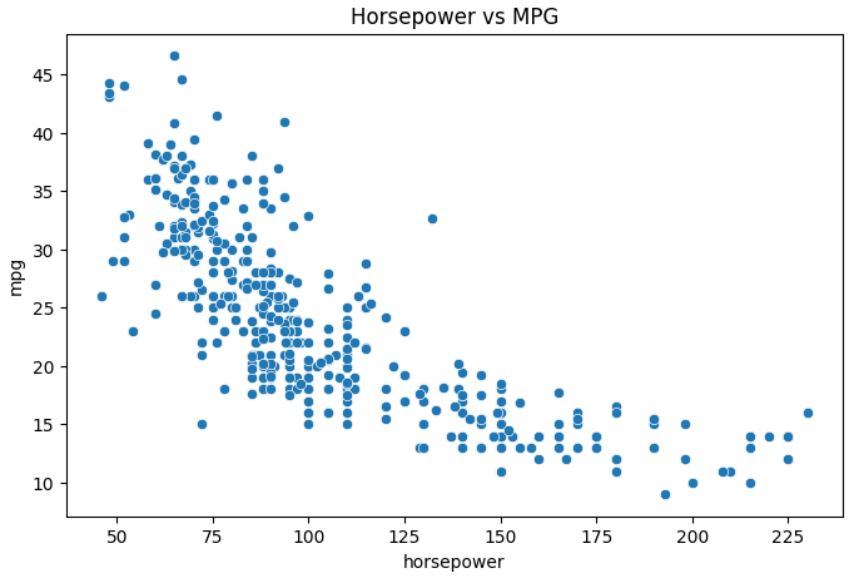
2.4.6 Results & Insights

i) MPG Distribution



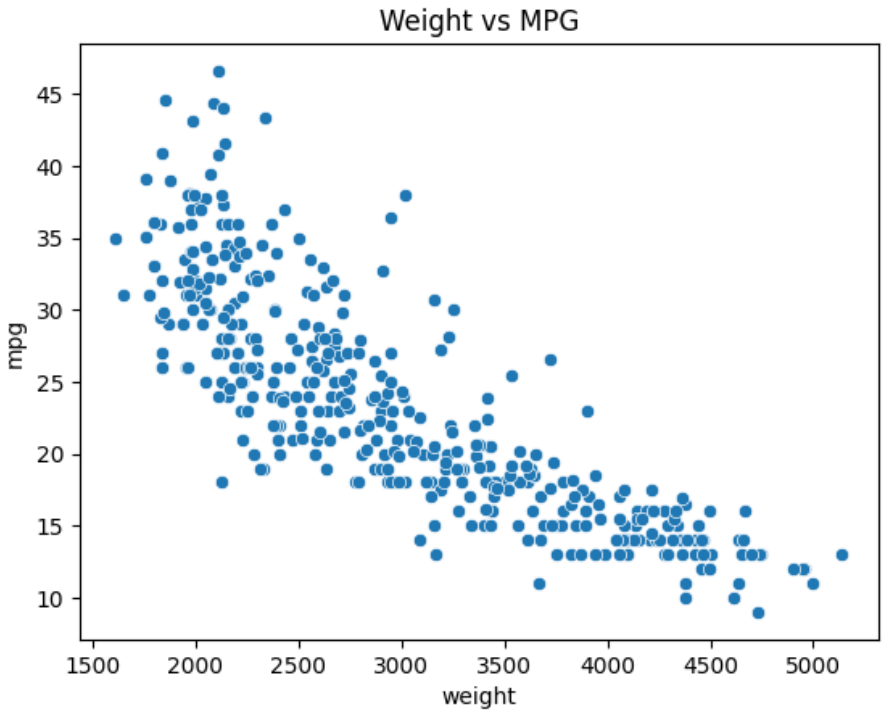
Most vehicles showed MPG values between 15–30, indicating moderate fuel efficiency across the dataset.

ii) Horsepower vs MPG



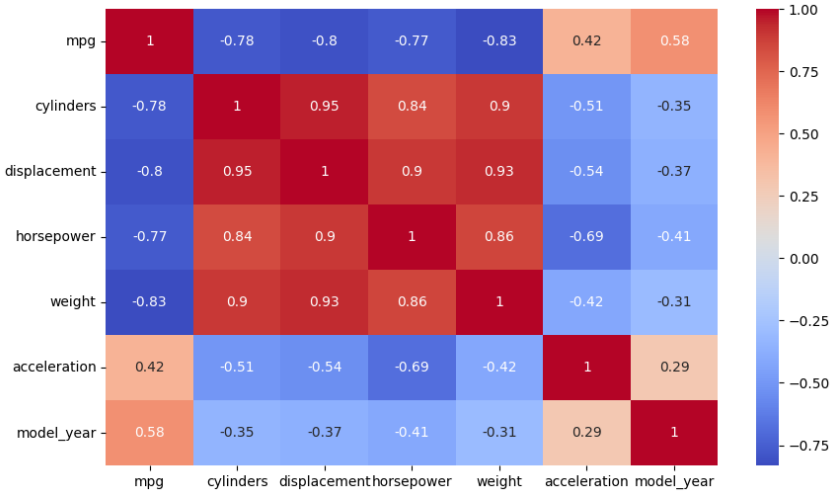
A strong negative relationship was observed between horsepower and MPG, meaning vehicles with higher horsepower generally had lower fuel efficiency.

iii) Weight vs MPG



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

iv) Correlation Analysis

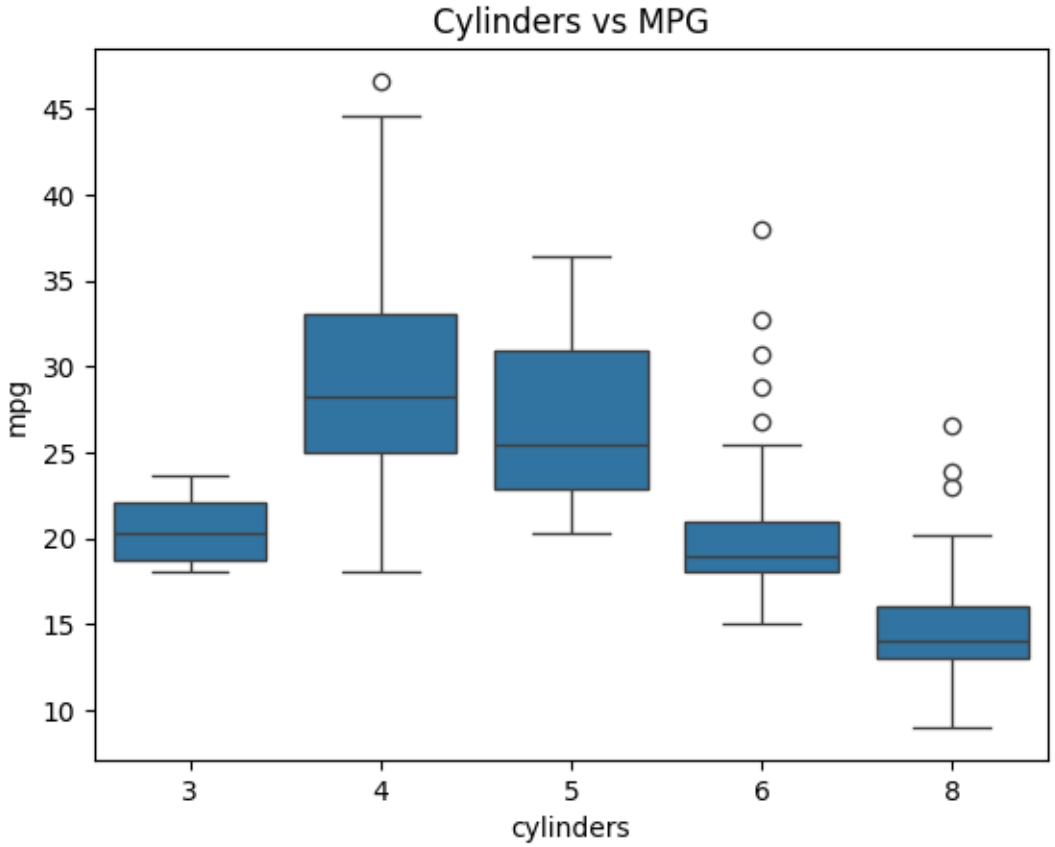


The heatmap revealed:

• 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 Using Machine Learning (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:

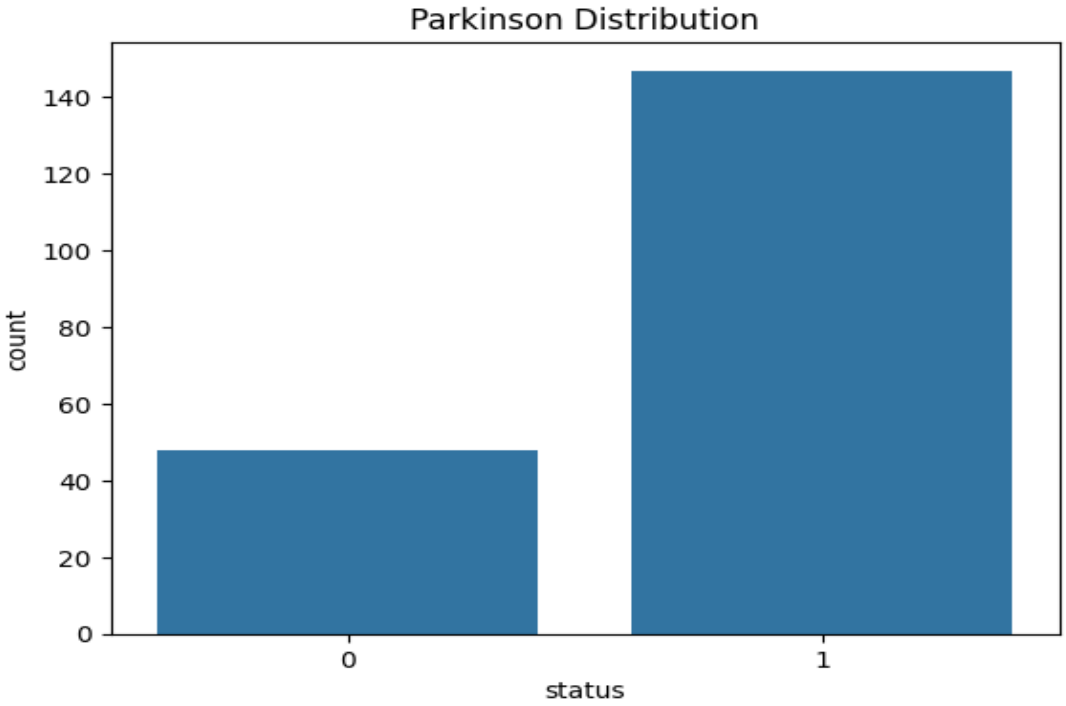
• Accuracy Score
• Confusion Matrix
• Classification Report

f) Streamlit Deployment

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

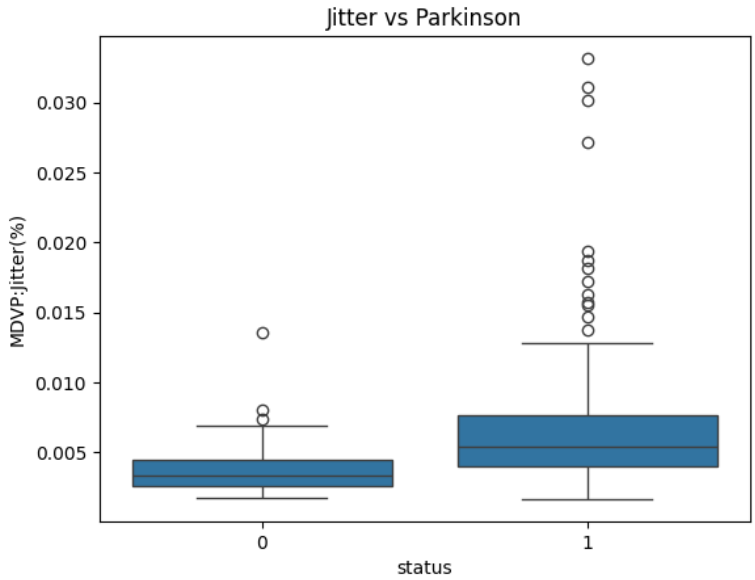
2.5.6 Results & Insights

i) Target Distribution



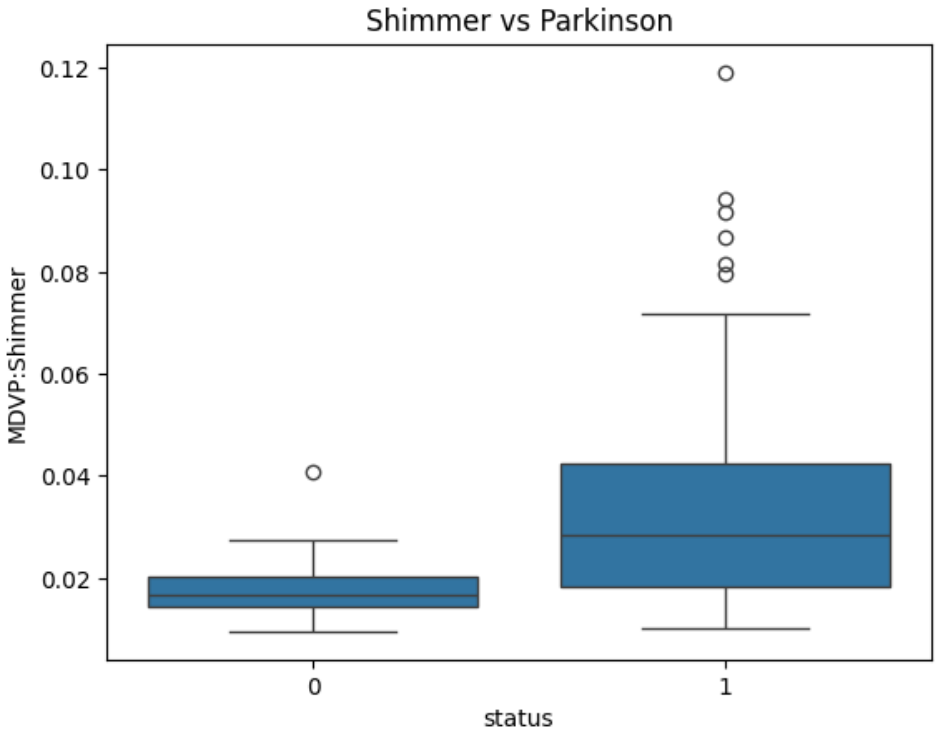
The dataset showed slight imbalance, with more Parkinson cases than healthy individuals.

ii) Jitter vs Parkinson



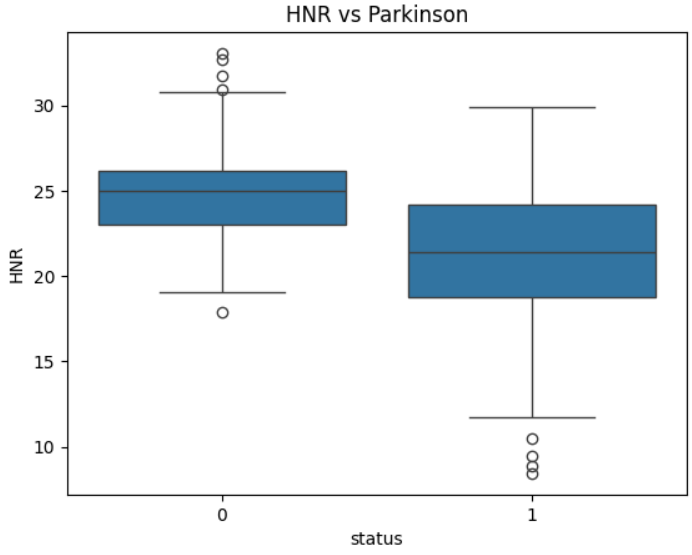
Parkinson patients generally showed higher jitter values, indicating instability in voice frequency.

iii) Shimmer vs Parkinson



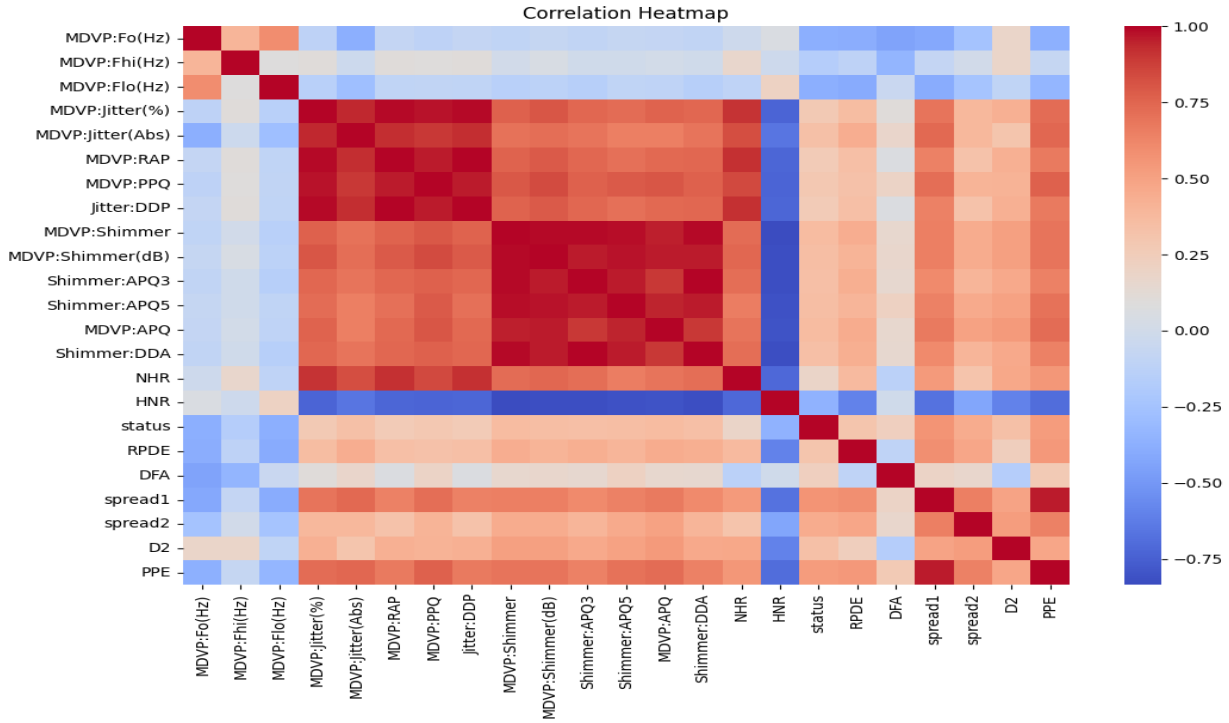
Higher shimmer values were observed among Parkinson patients, reflecting irregular voice amplitude patterns.

iv) HNR vs Parkinson



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

v) Correlation Analysis

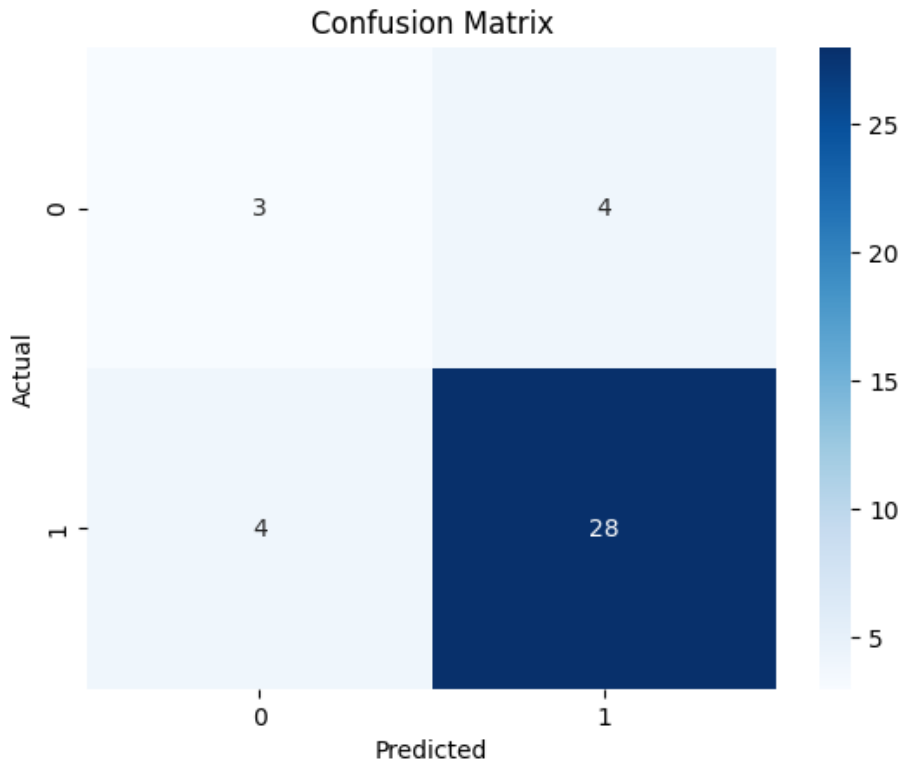


The heatmap revealed strong correlations among jitter and shimmer-related features, indicating interconnected voice abnormalities.

vi) Model Performance

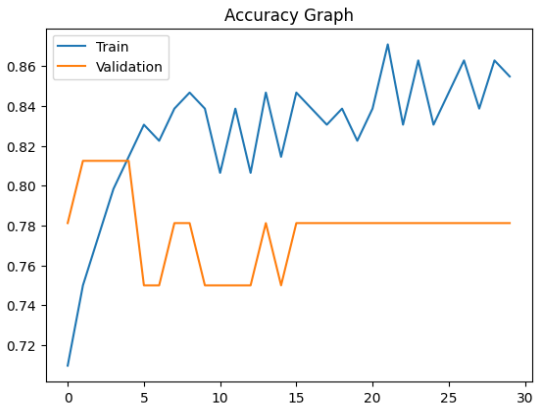
The ANN model achieved stable performance with good classification accuracy and minimal overfitting.

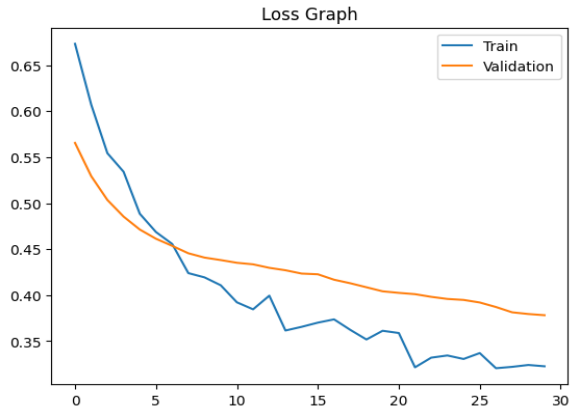
vii) Confusion Matrix



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

viii) Training & Validation Trends



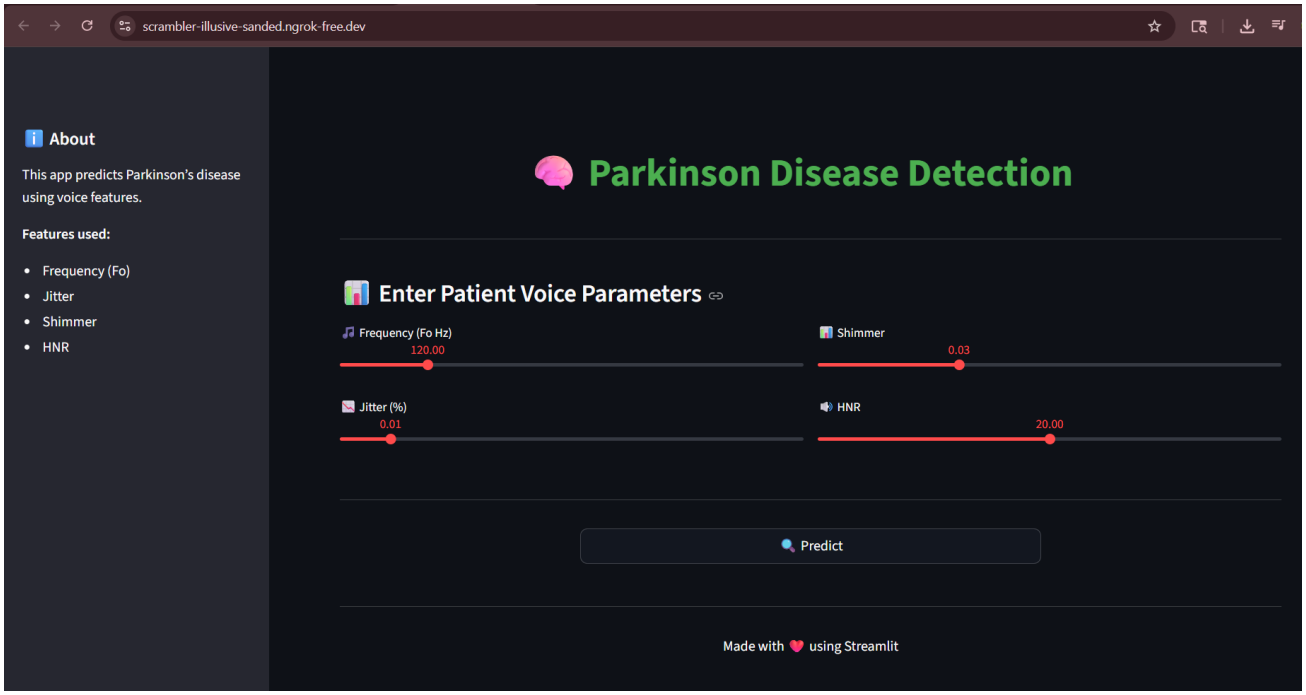


Accuracy and loss graphs indicated stable model learning and smooth convergence during training.

ix) Streamlit Application

The deployed Streamlit application enabled real-time predictions using user-provided voice parameters, making the system interactive and user-friendly.

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 Credit Card Fraud Detection System using CTGAN & Machine Learning (Python, Generative AI & Streamlit) (Week 6)

2.6.1 Introduction

With the rapid growth of online transactions and digital payment systems, fraudulent financial transactions have become a major challenge for banks and financial institutions. Traditional rule-based fraud detection systems are often slow, inflexible, and unable to identify complex fraud patterns effectively.

This project focuses on building an intelligent fraud detection system using machine learning and Generative AI techniques. A Random Forest Classifier was developed to classify transactions as either fraudulent or legitimate. Since the dataset was highly imbalanced, CTGAN (Conditional Tabular Generative Adversarial Network) was used to generate synthetic fraud samples and improve model performance.

The project also includes exploratory data analysis (EDA), preprocessing, model evaluation, and deployment using Streamlit for real-time fraud prediction.

2.6.2 Objectives

- To analyze credit card transaction data for fraud detection
- To preprocess and clean transaction datasets
- To identify fraud-related transaction patterns
- To handle class imbalance using synthetic data generation (CTGAN)
- To build a machine learning classification model for fraud detection
- To evaluate model performance using classification metrics and ROC curve
- To deploy the fraud detection system using Streamlit

2.6.3 Tools & Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn
- CTGAN (Generative AI)

• Streamlit
• Joblib

2.6.4 Dataset Description

The dataset contains financial transaction records labeled as either legitimate or fraudulent.
Key features include:
• Transaction Amount
• Transaction Time
• V1 to V28 anonymized features
• Class (0 = Legitimate, 1 = Fraudulent)
The dataset is highly imbalanced, with fraudulent transactions representing only a very small percentage of total records.

2.6.5 Methodology

a) Data Preprocessing

The following preprocessing steps were performed:
• Loaded dataset using Pandas
• Removed duplicate records
• Checked missing values
• Applied feature scaling on Amount and Time columns using StandardScaler
• Reformatted the dataset for model training

b) Exploratory Data Analysis (EDA)

EDA was conducted using:
• Count plots
• Histograms
• Boxplots
• KDE plots
• Correlation heatmaps
The analysis focused on understanding fraud patterns and transaction behavior.

c) Class Imbalance Handling

Since fraud cases were significantly fewer than legitimate transactions, CTGAN was used to generate synthetic fraud data and balance the dataset.

d) Machine Learning Model

A Random Forest Classifier was trained on:

• Original dataset
• Augmented dataset containing synthetic fraud samples

e) Model Evaluation

The model was evaluated using:
• Accuracy Score
• Precision
• Recall
• F1-score
• Confusion Matrix
• ROC Curve & AUC Score

f) Streamlit Deployment

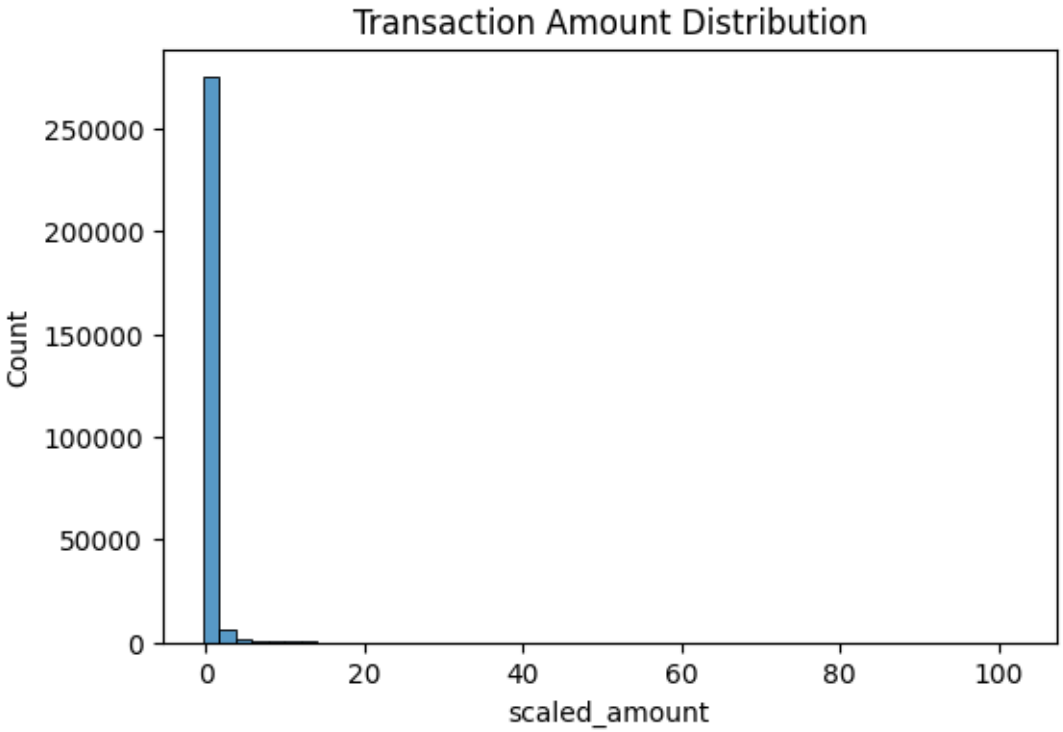
A Streamlit-based web application was developed to allow users to:
• Enter transaction details
• Receive real-time fraud predictions
• View fraud probability scores

2.6.6 Results & Insights

i) Class Distribution

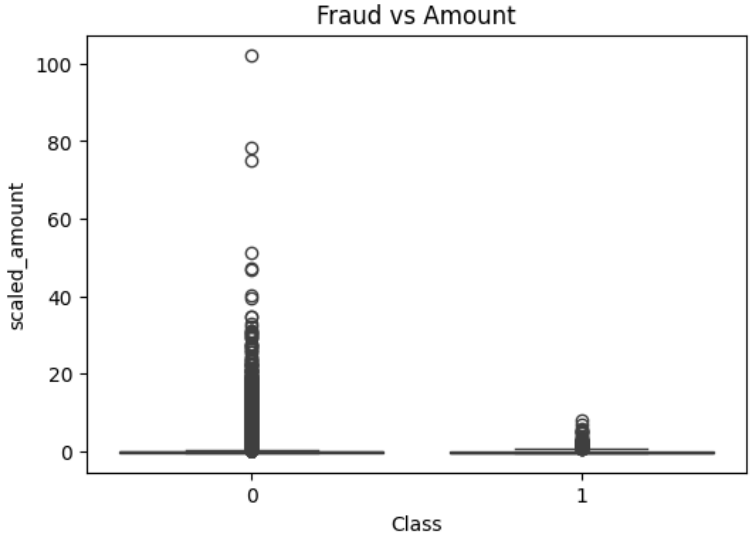
The dataset was heavily imbalanced, with legitimate transactions greatly outnumbering fraudulent transactions.

ii) Transaction Amount Distribution



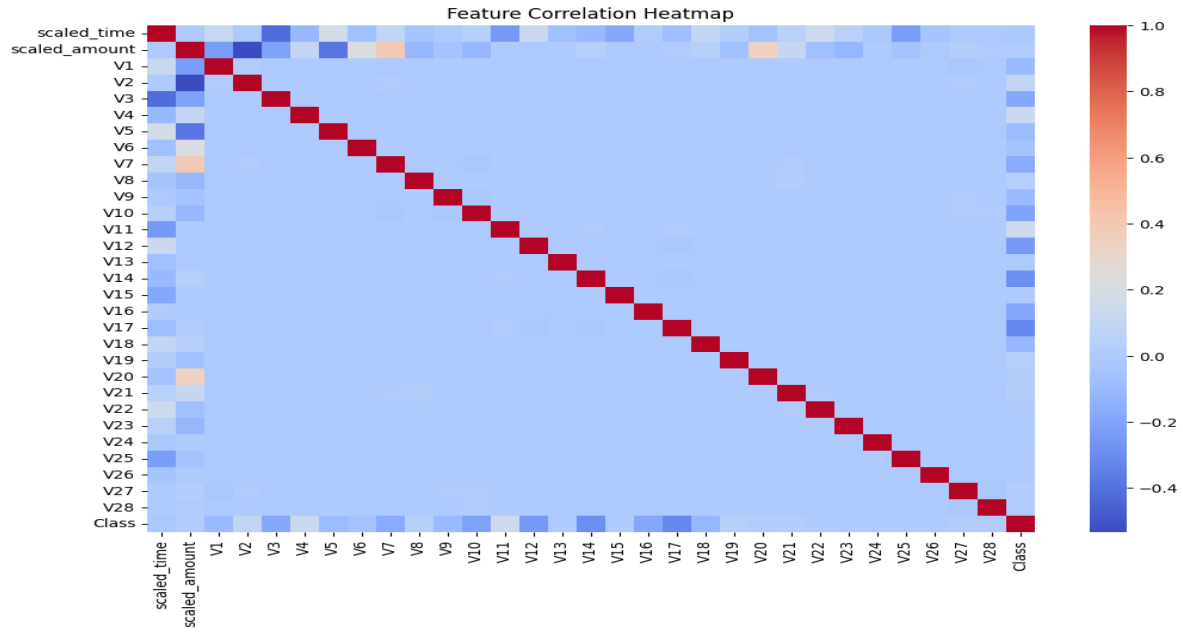
Most transactions were low-value transactions, while high-value transactions appeared less frequently.

iii) Fraud vs Amount Analysis



Fraudulent transactions existed across different transaction amounts, indicating that transaction amount alone is not a strong fraud indicator.

iv) Correlation Analysis

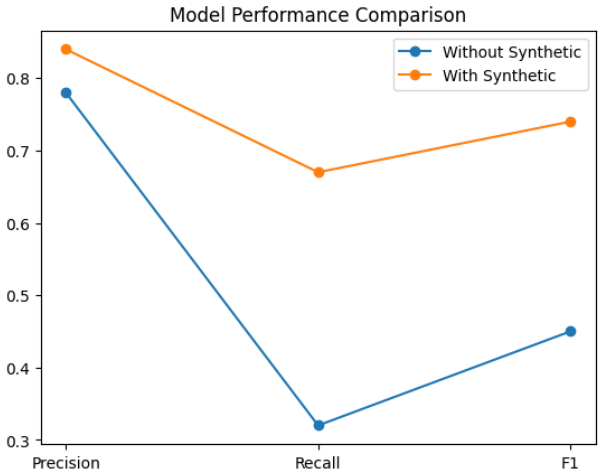


The correlation heatmap showed low correlation among most features, which is beneficial for machine learning model performance.

v) CTGAN Synthetic Data Generation

CTGAN successfully generated synthetic fraud samples that improved class balance within the dataset.

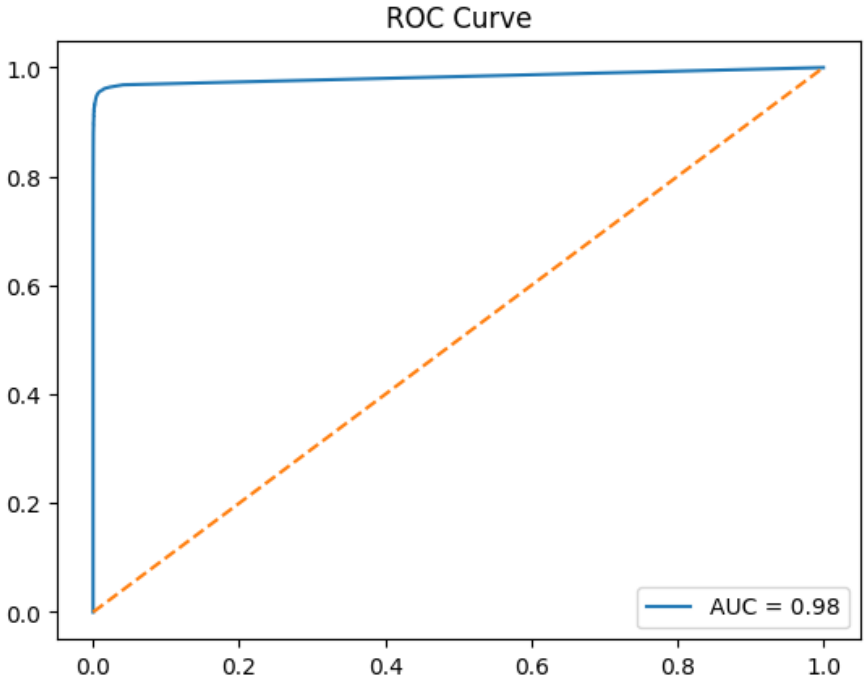
vii) Model Performance Improvement



The Random Forest model trained with synthetic data performed significantly better compared to the model trained only on real data.
Key improvements were observed in:

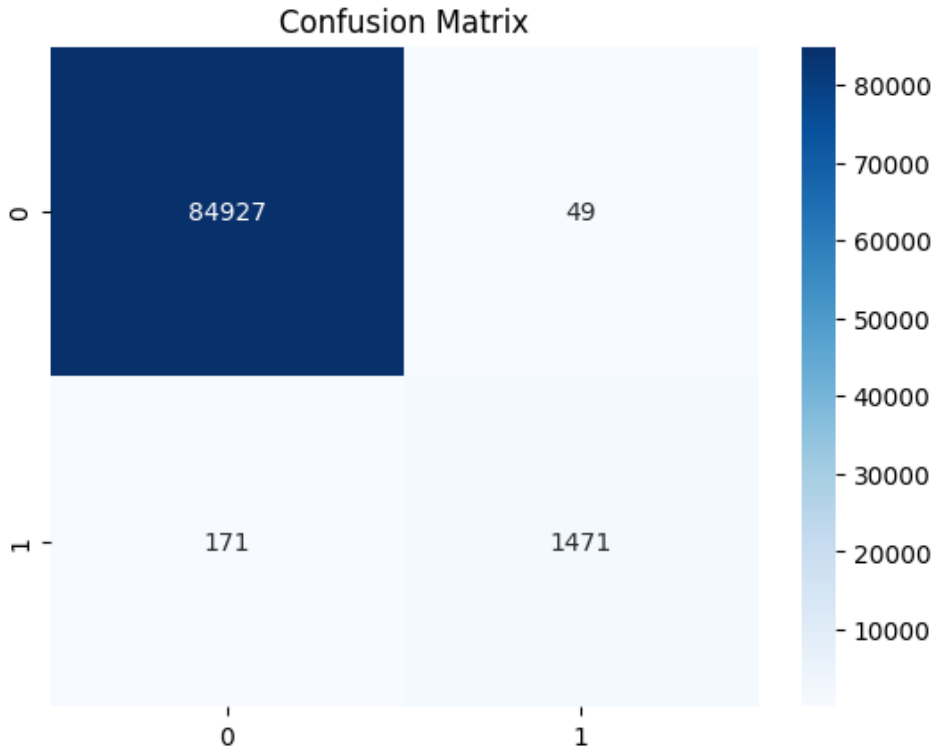
- Recall
- F1-score
- Fraud detection capability

viii) ROC Curve Performance



The ROC curve showed excellent classification performance with a very high AUC score, indicating strong fraud detection capability.

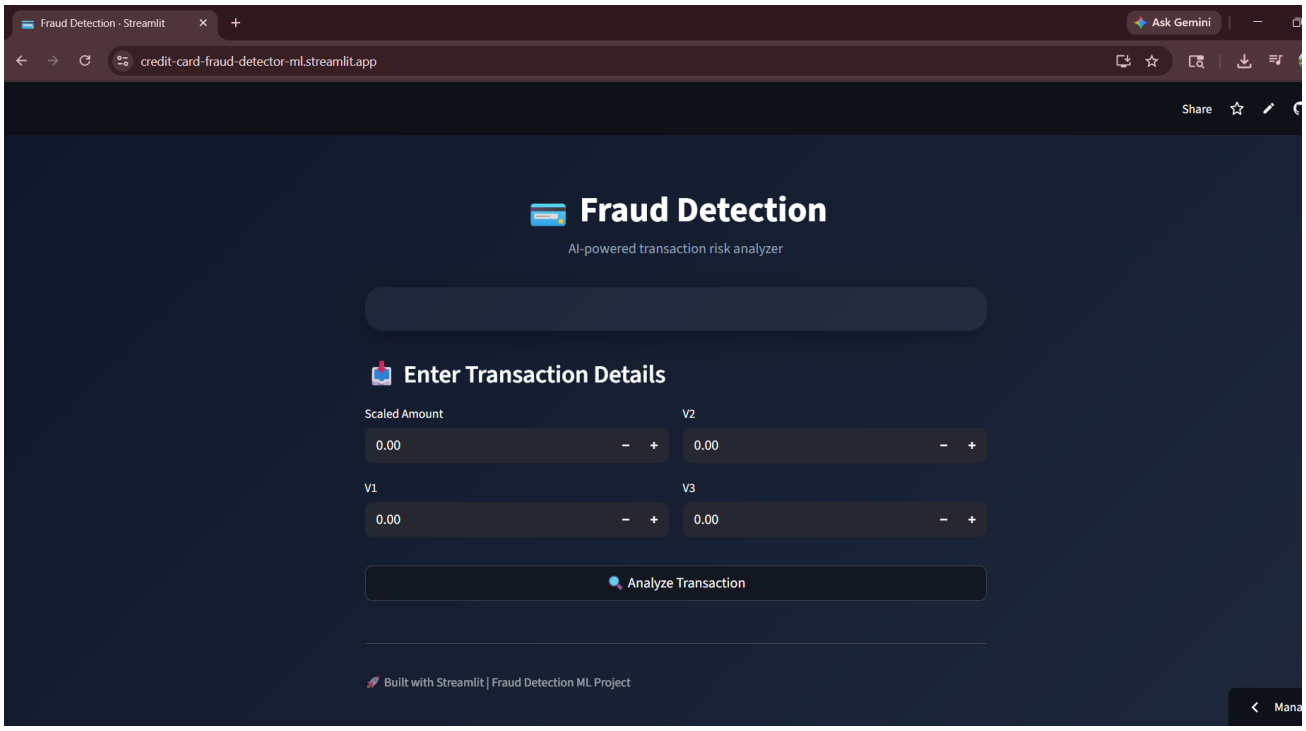
ix) Confusion Matrix



The confusion matrix revealed:
• High true negative rate
• Very low false positive rate
• Improved fraud detection accuracy

ix) Streamlit Application
The deployed Streamlit application provided:
• Real-time fraud prediction
• Fraud probability visualization
• Interactive user interface for transaction analysis

Figure : Fraud Detection Streamlit Application:



The figure represents the final deployment stage of the fraud detection project through a Streamlit web application. The application was designed to provide a responsive and interactive environment for performing fraud prediction tasks in real time.

Users can input transaction attributes through the graphical interface, after which the pre-trained Random Forest model processes the data and predicts whether the transaction is legitimate or fraudulent. The deployment also incorporates preprocessing techniques and CTGAN-generated synthetic samples to improve model robustness and handling of class imbalance.

The Streamlit deployment enhances accessibility, scalability, and usability of the developed system by converting the machine learning pipeline into a practical web-based solution suitable for demonstration and real-world analytical applications.

And at the last project has been deployed at Streamlit Cloud so that everyone can able to access the app .

Live Link : https://credit-card-fraud-detector-ml.streamlit.app/

2.6.7 Conclusion

This project successfully developed a machine learning-based fraud detection system capable of identifying suspicious financial transactions in real time.

The integration of CTGAN for synthetic data generation significantly improved fraud detection performance by handling dataset imbalance more effectively. The Random Forest model demonstrated strong classification capability with high accuracy and improved recall.

The Streamlit deployment made the system interactive and practical for real-world usage by enabling real-time transaction analysis and fraud probability prediction.

Overall, this project provided practical experience in fraud analytics, Generative AI, machine learning classification, imbalance handling, model evaluation, and deployment using Streamlit.

CHAPTER 3: CONCLUSION

5.1 Overall Learning Outcomes

The internship at Global Next Consulting India Pvt. Ltd. provided practical exposure to the complete workflow of Artificial Intelligence and Machine Learning projects, including data preprocessing, exploratory data analysis, model building, evaluation, and deployment.

Through six structured projects, I gained hands-on experience in Python, Machine Learning, Deep Learning, data visualization, and Streamlit application development. The internship involved working on real-world datasets across multiple domains such as healthcare analytics, fraud detection, music popularity prediction, automotive segmentation, and financial transaction analysis.

Each project helped strengthen my understanding of supervised and unsupervised learning techniques, including regression, classification, clustering, and Artificial Neural Networks (ANN). I also learned how to handle real-world challenges such as missing values, class imbalance, feature scaling, dimensionality reduction, and model evaluation.

The major projects on Parkinson's Disease Detection and Credit Card Fraud Detection integrated machine learning, deep learning, Generative AI concepts, and deployment techniques into complete end-to-end AI solutions.

The internship also enhanced my problem-solving ability, analytical thinking, project implementation skills, and understanding of real-world AI/ML applications. Additionally, it improved my professional skills such as communication, report preparation, presentation, and independent learning.

SUMMARY

The internship provided in-depth practical exposure to Artificial Intelligence and Machine Learning concepts through the implementation of multiple real-world projects. During the internship, hands-on experience was gained in data preprocessing, exploratory data analysis (EDA), machine learning, deep learning, clustering, classification, regression, and deployment techniques using tools and technologies such as Python, Scikit-learn, TensorFlow/Keras, Streamlit, and Generative AI techniques.

Each week focused on solving a real-world problem using structured AI/ML workflows — from data collection and preprocessing to model development, evaluation, visualization, and deployment.

Across the six projects, the work covered multiple AI/ML domains:

• **Week 1 (Healthcare Cost Analysis)** – Analyzed healthcare expenditure and life expectancy data using Python by performing data preprocessing, exploratory data analysis, correlation analysis, and visualization to identify healthcare trends and relationships between medical expenditure and life expectancy.

• **Week 2 (Insurance Claims & Fraud Analysis)** – Performed insurance fraud analysis using Python by cleaning and preprocessing claim data, analyzing fraud patterns, applying PCA for dimensionality reduction, and generating insights related to fraudulent insurance claims.

• **Week 3 (Spotify Popularity Prediction Analysis**) – Built machine learning models to predict song popularity using Spotify audio features such as danceability, energy, loudness, and tempo. Applied regression techniques, PCA, feature analysis, and model evaluation to understand factors affecting music popularity.

• **Week 4 (Vehicle Segmentation using Clustering)** – Applied unsupervised machine learning techniques using K-Means clustering to segment vehicles based on performance and fuel efficiency features. Used PCA visualization and clustering evaluation methods such as Elbow Method and Silhouette Score.

• **Week 5 (Parkinson's Disease Detection System)** – Developed a healthcare AI application using Artificial Neural Networks (ANN) to detect Parkinson's disease based on biomedical voice measurements. The project included feature scaling,

model training, evaluation, and deployment through a Streamlit web application for real-time disease prediction.

• **Week 6 (Credit Card Fraud Detection System using CTGAN & Machine Learning)** – Built an intelligent fraud detection system using Random Forest Classification and Generative AI techniques. CTGAN was used to generate synthetic fraud samples for handling class imbalance, and the final model was deployed through a Streamlit web application for real-time fraud prediction and transaction analysis.

Through these projects, strong technical and analytical skills were developed in machine learning, deep learning, data visualization, feature engineering, model evaluation, clustering, classification, and AI model deployment.

The internship significantly improved problem-solving ability, analytical thinking, and practical implementation skills while providing real-world exposure to end-to-end AI/ML workflows.

Overall, the internship successfully bridged the gap between theoretical knowledge and industry-level AI/ML applications, enhancing confidence and readiness for future opportunities in Artificial Intelligence and Machine Learning.

REFERENCES

1. Kaggle Datasets – Healthcare Cost Analysis, Insurance Claims Fraud Analysis, Spotify Popularity Prediction Analysis, Parkinson's Disease Detection, and Credit Card Fraud Detection datasets.

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

3. TensorFlow / Keras Documentation – Artificial Neural Network (ANN) model development and deep learning implementation.

4. Streamlit Documentation – Deployment of machine learning models through interactive web applications.

5. CTGAN Documentation – Synthetic data generation and handling class imbalance using Generative AI techniques.

6. Research Articles & Online Resources –
   - Machine Learning Applications in Fraud Detection
   - Deep Learning Applications in Healthcare
   - AI-based Disease Prediction Systems
   - Clustering Techniques in Machine Learning
   - Generative AI for Synthetic Data Generation