AI/ML Internship

A Project Report submitted to the

GLOBAL NEXT CONSULTING INDIA PVT LTD

(Six – Week Internship Program)

By

AMEYA JOSE CHAZHOOR

Under the Supervision of

Dr. Anuradha Gupta

(Project Director)

Submitted To :

Global Next Consulting India Pvt. Ltd.

Duration of Internship :

27-March-2026 to 15-may-2026

December 2025

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report, **"AI/ML Internship (GNCIPL)"**, submitted as per the requirements for the Data Analyst/ Business Analyst/ Data Science role, This is the result of original work carried out by me under the guidance of **Ms. Anuradha Gupta** during the time period from March 2026 to May 2026.

I further declare that this report represents authentic record of my own work and does not contain any falsely fabricated ideas, data, facts or sources. I also declare that I have adhered to all principles of academic honesty and integrity and that this report has not been submitted, either in part or in full, to any other institute, university, or organization for the award of any degree, diploma, or certification.

AMEYA JOSE CHAZHOOR

CERTIFICATE

This is to certify that the project report entitled **"AI/ML Internship Report"** has been carried out by **Ameya Jose Chazhoor ,** a Fresher in job search and improve skill in Data analyst & Data Science role with the Past Experience in Medical Coding Analyst around one and half year. This work was carried out under the guidance of **Ms. Anuradha Gupta** from October 2025 to December 2025. It is further certified that this work has not been submitted to any other university or institution for the award of any other 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.

Ameya Jose Chazhoor

ABSTRACT

Fraud detection in financial transactions is a critical challenge due to the highly imbalanced nature of real-world datasets, where fraudulent activities represent only a small fraction of total transactions. This imbalance significantly affects the performance of traditional machine learning models, leading to poor detection of fraud cases.

This project focuses on enhancing fraud detection using Generative Artificial Intelligence, specifically the Conditional Tabular Generative Adversarial Network (CTGAN). The objective is to generate realistic synthetic fraud data to address the issue of class imbalance and improve the effectiveness of predictive models.

The project begins with exploratory data analysis (EDA) to understand the dataset characteristics, including class distribution, feature relationships, and presence of outliers. The analysis reveals that fraud instances are extremely rare, which justifies the need for data augmentation.

CTGAN is then trained on the minority class (fraud transactions) to learn its underlying distribution and generate synthetic samples. These generated samples are combined with the original dataset to create a more balanced dataset.

A machine learning model, specifically a Random Forest classifier, is trained on both the original and augmented datasets. The performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. The results demonstrate a significant improvement in fraud detection, particularly in recall and F1-score, after applying synthetic data augmentation.

This project highlights the effectiveness of Generative AI techniques in handling imbalanced datasets and improving model performance. It demonstrates how synthetic data generation can be used as a powerful tool in real-world applications such as fraud detection.

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

During my six-week internship at GNCIPL as a Data Analyst Intern, the main objectives were:

- To gain hands-on experience in data analytics tools and techniques, especially using Python (Google Colab, Jupyter Notebook), R, ETL Process and Microsoft Excel.

- To work on real-world datasets and deliver meaningful insights, visualizations, and dashboard reports.

- To learn data preprocessing, cleaning, transformation, and applying formulas and classification logic.

- To enhance analytical thinking, effective communication, and presentation skills through weekly minor projects and a major end project.

Chapter 2 - Projects

2.1 Basic Data Analysis using Excel (Week 1)

2.1.1 Introduction

Data analysis is a fundamental step in understanding patterns, trends, and insights from raw data. In business environments, data-driven decision-making plays a crucial role in improving performance and efficiency.

This project focuses on performing basic data analysis using Microsoft Excel. The objective is to explore and analyse a structured dataset to identify meaningful insights using Excel tools such as formulas, pivot tables, charts, and dashboards.

The dataset used for this project contains structured information related to business operations, including categories such as sales, customer details, product performance, and regional distribution.

Excel was used to clean, organize, and analyze the dataset, enabling the creation of visual dashboards and key performance indicators (KPIs) to summarize the data effectively.

2.1.2 Objectives

Primary

Objectives

• To understand the fundamentals of data analysis using Excel
• To clean and prepare raw data for analysis
• To identify patterns and trends in the dataset
• To create meaningful visualizations using charts and dashboards
• To generate insights that support decision-making

Specific Analytical Goals

• Analyse distribution of key variables in the dataset
• Create pivot tables to summarize large datasets
• Identify top-performing categories and segments
• Compare performance across different regions or groups
• Develop an interactive dashboard using Excel tools

2.1.3 Methodology

a) Dataset Preparation

• Loaded the dataset into Excel
• Reviewed column structure and data types
• Cleaned the dataset by:

- Removing duplicates
- Handling missing values
- Correcting inconsistent entries
  • Standardized formats (dates, numbers, text)
  • Organized data into structured tables

b) Data Analysis Techniques

• Used Excel formulas such as:

- SUM, AVERAGE, COUNT
- IF conditions for classification

  • Created pivot tables to:
- Summarize data
- Group by categories
- Calculate totals and averages

  • Applied filtering and sorting for deeper analysis

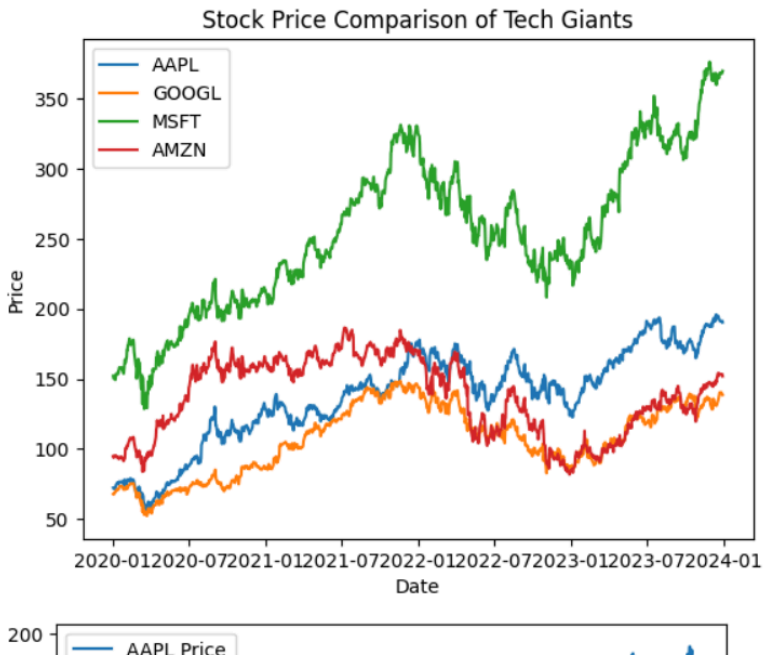
c) Visualization and Dashboard

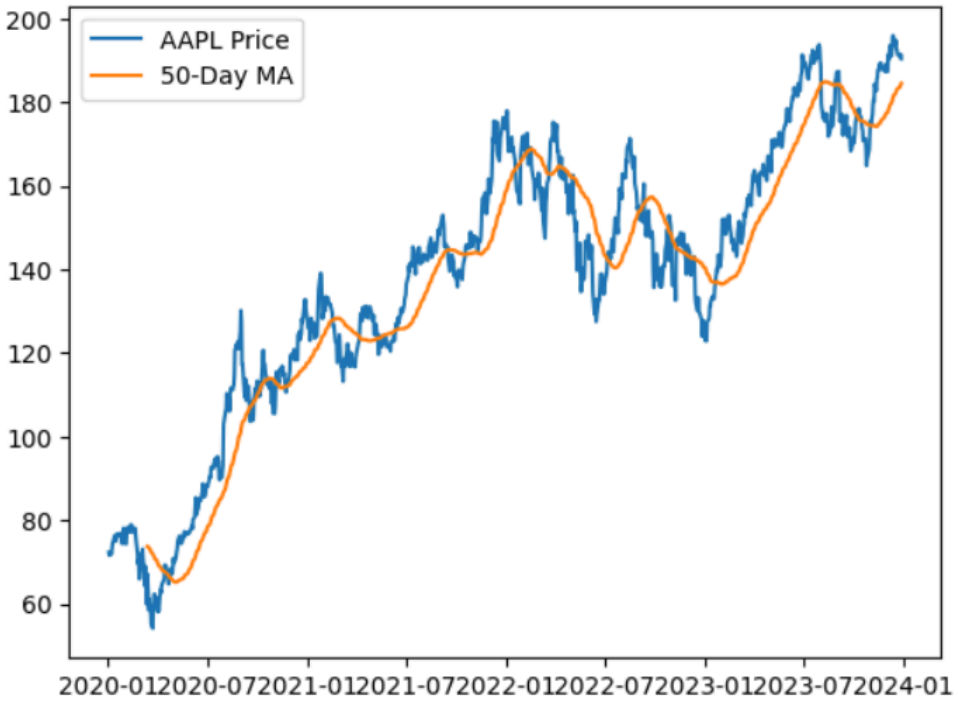
• Created charts such as:

- Bar charts
- Pie charts
- Line graphs

  • Designed a dashboard with:
- KPI indicators (Total, Average, Counts)
- Category-wise analysis
- Region-wise comparison

  • Used slicers for interactivity





2.1.4 Results and Insights

a) Overall Performance Insights

• Identified key trends in the dataset
• Highlighted high-performing categories
• Observed variations across different groups

b) Category Analysis

• Some categories showed significantly higher performance
• Certain segments contributed more to overall results

c) Regional Analysis

• Performance varied across regions
• Certain regions showed higher activity levels

d) Data Trends

• Patterns were identified over time or across variables
• Visualization helped in understanding relationships

2.1.5 Recommendations

• Focus on high-performing categories to maximize output
• Improve performance in low-performing segments
• Use dashboards for regular monitoring
• Maintain clean and structured data for better analysis
• Automate reporting using Excel features

2.1.6 Conclusion

This project provided a strong foundation in data analysis using Excel. It demonstrated how raw data can be transformed into meaningful insights using simple tools and techniques.

The use of pivot tables, charts, and dashboards enabled effective visualization and interpretation of data. This project highlights the importance of Excel as a powerful tool for data analysis in real-world business scenarios

2.2 Renewable Energy (Wind) Forecasting Analysis (Week 2)

2.2.1 Introduction

Renewable energy has become a crucial alternative to conventional power sources, with wind energy playing a significant role in sustainable electricity generation. This project, *Wind Energy Forecast Analysis*, focuses on understanding and predicting wind-based power generation patterns using meteorological parameters such as wind speed, temperature, humidity, and pressure.

The dataset consists of hourly wind energy data collected from multiple Indian locations, including Davanagere, Kanniyakumari, Jaisalmer, and Kutch. It includes parameters such as wind speed at 10m and 100m heights, temperature, humidity, air pressure, and energy output (kWh).

The project aims to analyse wind energy generation patterns from January to March 2024. Using MySQL for data processing and Excel for visualization, the objective is to uncover trends, relationships, and insights to support efficient energy forecasting and decision-making.

2.2.2 Objectives

• To collect and integrate wind energy datasets from multiple locations using MySQL
• To clean and preprocess data containing wind speed, temperature, humidity, pressure, and energy output
• To extract time-based features such as date, month, and hour for temporal analysis
• To analyse relationships between environmental factors and energy generation

• To forecast future wind energy production using Excel
• To compare energy generation across different locations
• To design an interactive dashboard for visualization and insights

2.2.3 Methodology

a) SQL Phase – Data Integration and Cleaning

• Created a database and imported district-level datasets into MySQL
• Combined datasets using **UNION ALL** to form a unified dataset
• Performed data cleaning:

- Removed duplicate records
- Handled missing values using mean imputation
- Standardized column names and data types
  • Extracted date-based features (year, month, day)
  • Created derived fields such as humidity categories
  • Generated a cleaned dataset for analysis

b) SQL-Based Analysis

• Calculated key metrics such as:

- Maximum and minimum energy output
- Average energy generation across states
  • Analysed wind speed distribution at different heights
  • Studied relationship between wind speed and energy output
  • Evaluated impact of temperature and humidity on energy generation
  • Identified hourly, daily, and monthly energy patterns
  • Compared performance across locations

c) Excel Phase – Dashboard and Visualization

An interactive Excel dashboard was created including:

• KPI Cards:

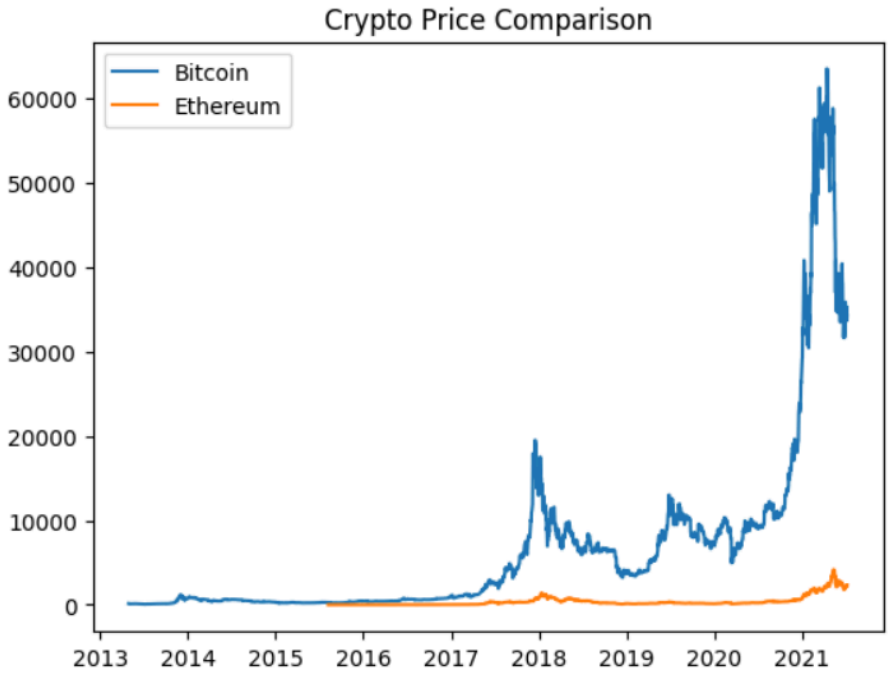
- Total energy generated
- Average wind speed
- Peak production hour

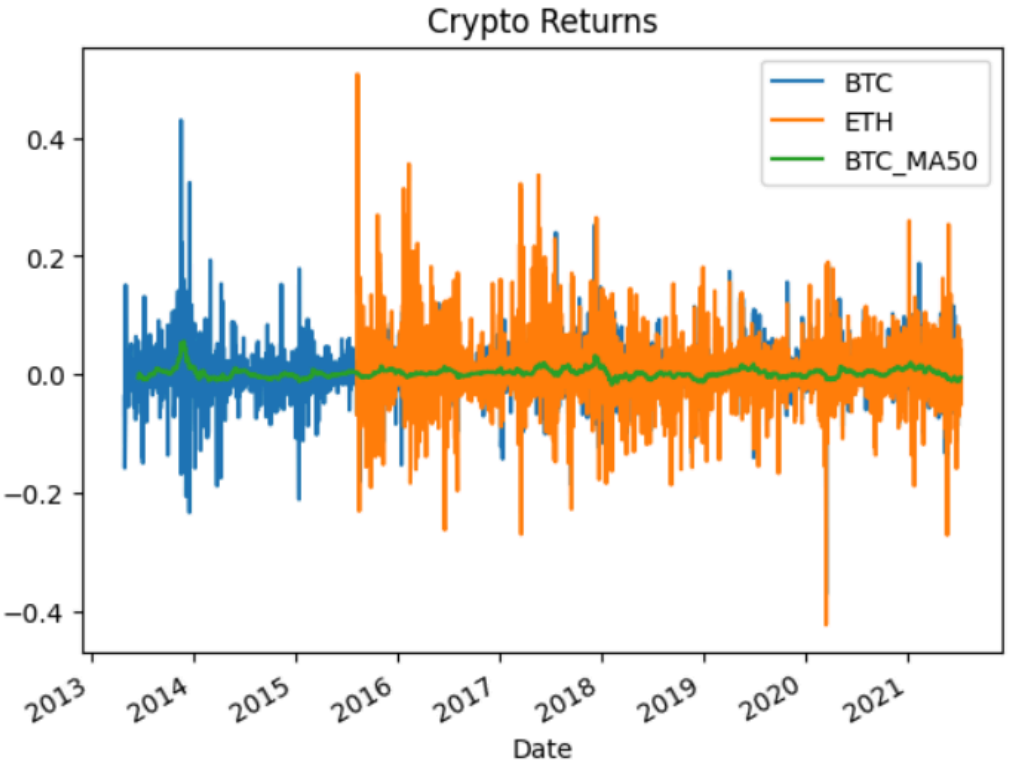
• Charts:

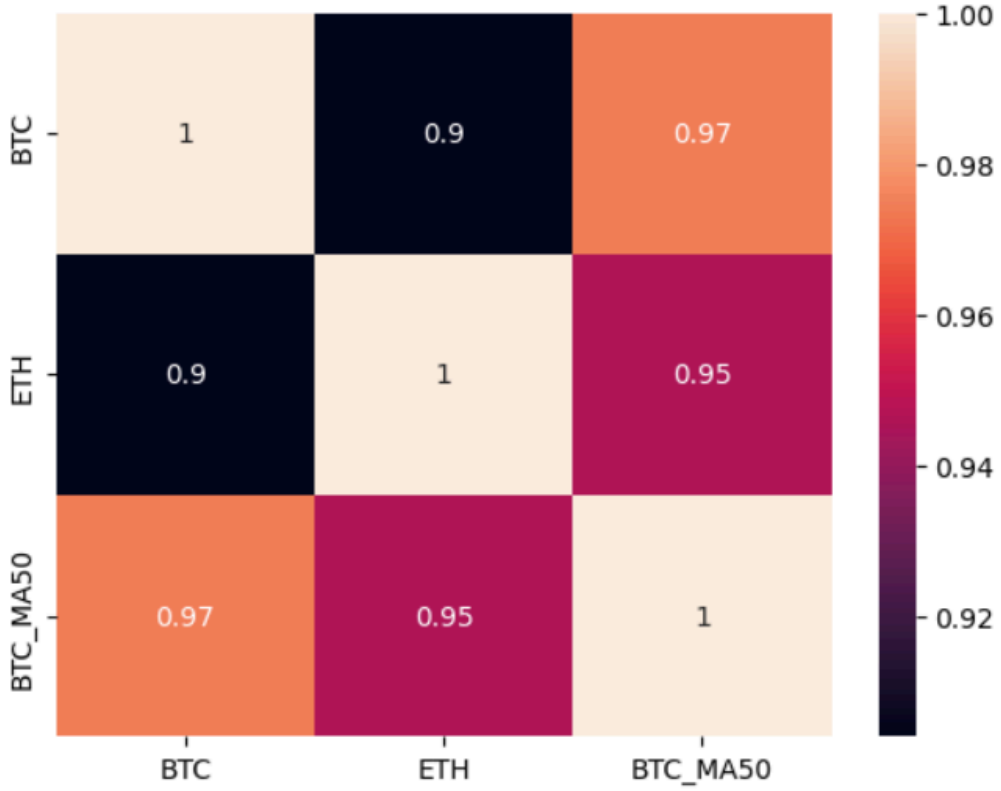
- Hourly wind speed trend (Line Chart)
- Temperature & humidity vs energy (Bar Chart)
- Monthly energy comparison (Column Chart)
- Daily energy trends (Line Chart)
- Forecasting future energy generation

• Additional Features:

- Slicers for filtering (Location, Date, Time, Humidity)
- Comparative charts for different regions







2.2.4 Results and Insights

a) Overall Performance

• Total energy generated: ~0.32M kWh

• Average wind speed (100m): ~11.79 m/s

• Peak energy production observed during evening hours

b) Wind Speed Analysis

• Strong positive relationship between wind speed and energy output
• Higher wind speeds significantly increase power generation

c) Temperature and Humidity Impact

• Moderate temperature (24°C – 28°C) supports optimal energy generation
• Medium humidity levels produce the highest energy output
• Extreme humidity reduces turbine efficiency

d) Time-Based Trends

• Wind speeds remain stable during night and peak in evening
• Daily energy output shows natural fluctuations
• February recorded the highest energy generation

e) Location-Based Insights

• Davanagere recorded the highest wind speed and energy output
• Coastal and semi-arid regions showed consistent performance
• Locations like Kutch and Kanniyakumari demonstrated stable generation

f) Forecasting Insights

• Forecast results indicate steady future energy production
• Suggests stable wind conditions for upcoming months

2.2.5 Conclusion

The Wind Energy Forecast Analysis demonstrates that environmental factors such as wind speed, temperature, and humidity significantly influence energy production. Among these, wind speed is the most critical factor, showing a strong positive relationship with power generation.

The project highlights the importance of data integration, preprocessing, and visualization in understanding renewable energy trends. The use of SQL and Excel enabled efficient analysis and forecasting of wind energy.

The results indicate that regions like Davanagere and Kutch have strong potential for sustainable energy production. This study supports the role of data analytics in improving renewable energy planning and operational efficiency.

2.3 Customer Segmentation and Spending Prediction using Machine Learning (Week 3)

2.3.1 Introduction

Customer behavior analysis is an essential aspect of modern business strategies, as it helps organizations understand purchasing patterns and improve decision-making. This project focuses on analyzing customer data and predicting spending behavior using supervised machine learning techniques.

The dataset used in this project is the *Mall Customers Dataset*, which includes features such as age, gender, annual income, and spending score. The spending score represents customer purchasing behavior and engagement level.

The objective of this project is to build a predictive model that estimates customer spending based on demographic and financial attributes. By understanding these patterns, businesses can improve marketing strategies, target the right customers, and increase profitability.

2.3.2 Objectives

Primary

Objectives

• To analyze customer data and understand behavioral patterns
• To build a predictive model for customer spending
• To apply supervised learning techniques for regression analysis
• To evaluate model performance using appropriate metrics

Specific Analytical Goals

• Perform data cleaning and preprocessing
• Conduct exploratory data analysis (EDA) to identify patterns
• Select relevant features for model training
• Train a regression model to predict spending score
• Evaluate model performance using error metrics
• Interpret results to derive business insights

2.3.3 Methodology

a) Dataset Preparation

• Loaded the *Mall Customers dataset* into Python
• Identified key features:

- Age
- Annual Income
- Spending Score (Target variable)
  • Checked for missing values and inconsistencies
  • Cleaned and prepared dataset for analysis

b) Exploratory Data Analysis (EDA)

• Performed univariate analysis using histograms to understand distribution of:

- Age

- Annual Income
- Spending Score

• Conducted bivariate analysis:

- Scatter plot of income vs spending
- Observed customer behavior patterns

• Generated correlation heatmap:

- Analysed relationships between variables
- Identified weak correlation between income and spending

c) Feature Selection

• Selected input features:

- Age
- Annual Income

• Selected target variable:

- Spending Score

d) Model Building

• Applied **Linear Regression**, a supervised learning algorithm
• Split dataset into:

- Training data (80%)
- Testing data (20%)
  • Trained model to learn relationship between input features and target

e) Model Evaluation

• Predicted spending score using test data

• Evaluated model using:

- Mean Squared Error (MSE)
  • Compared predicted values with actual values

2.3.4 Results and Insights

a) Data Distribution Insights

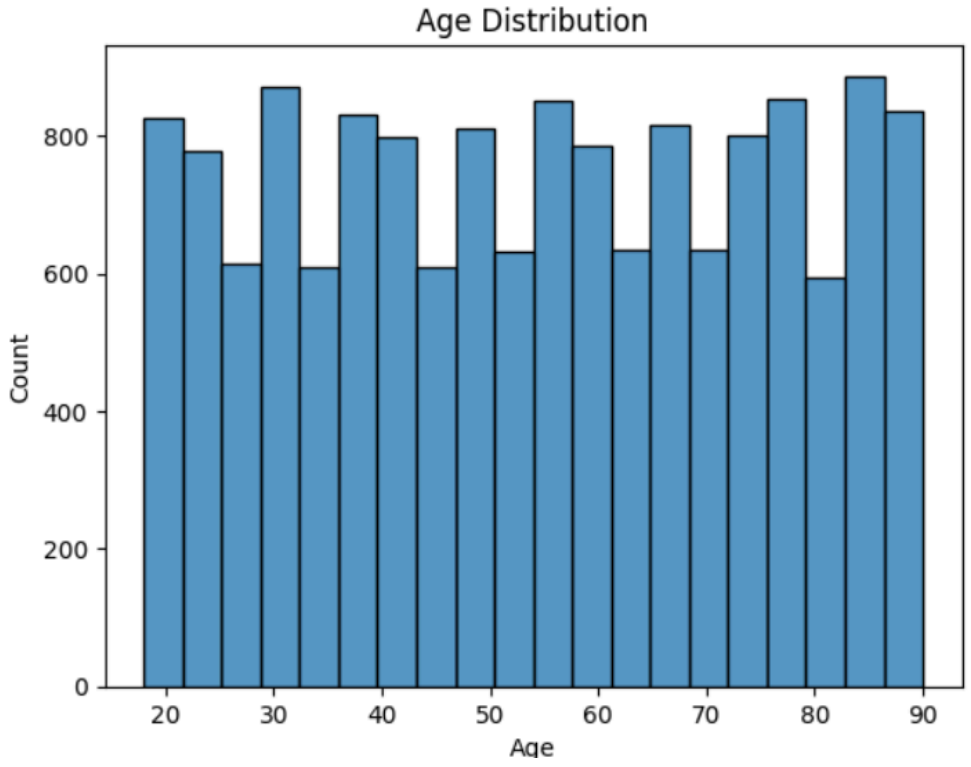
• Age and income are spread across multiple ranges

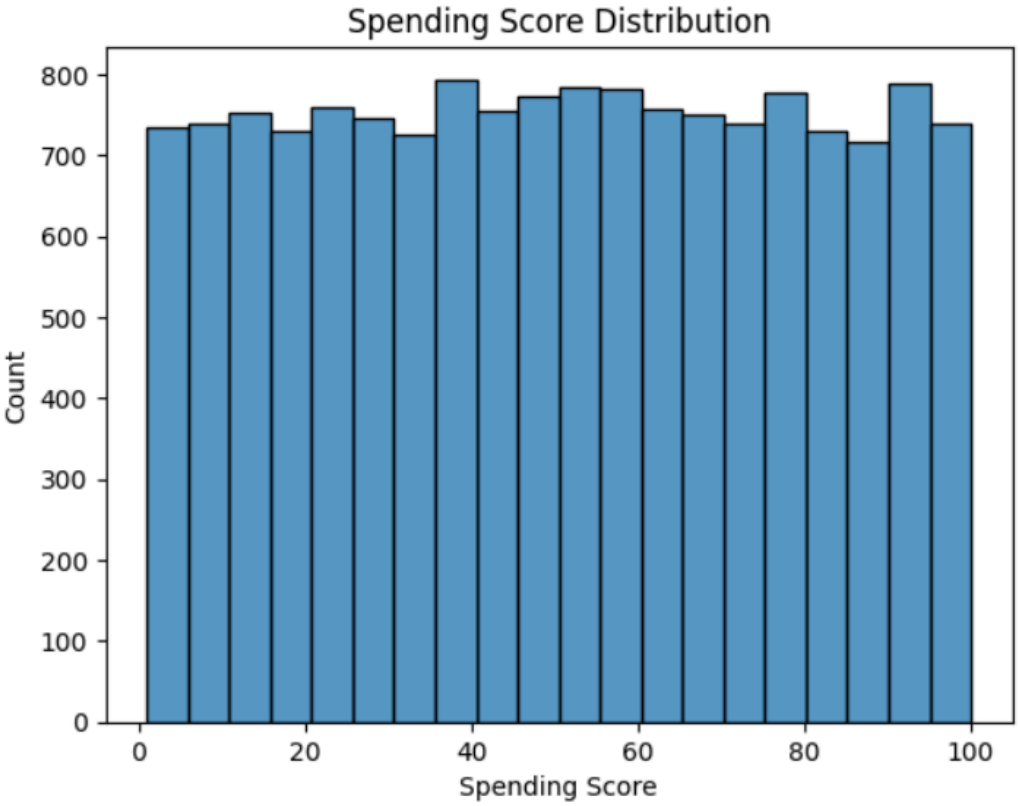
• Spending score shows varied customer behavior

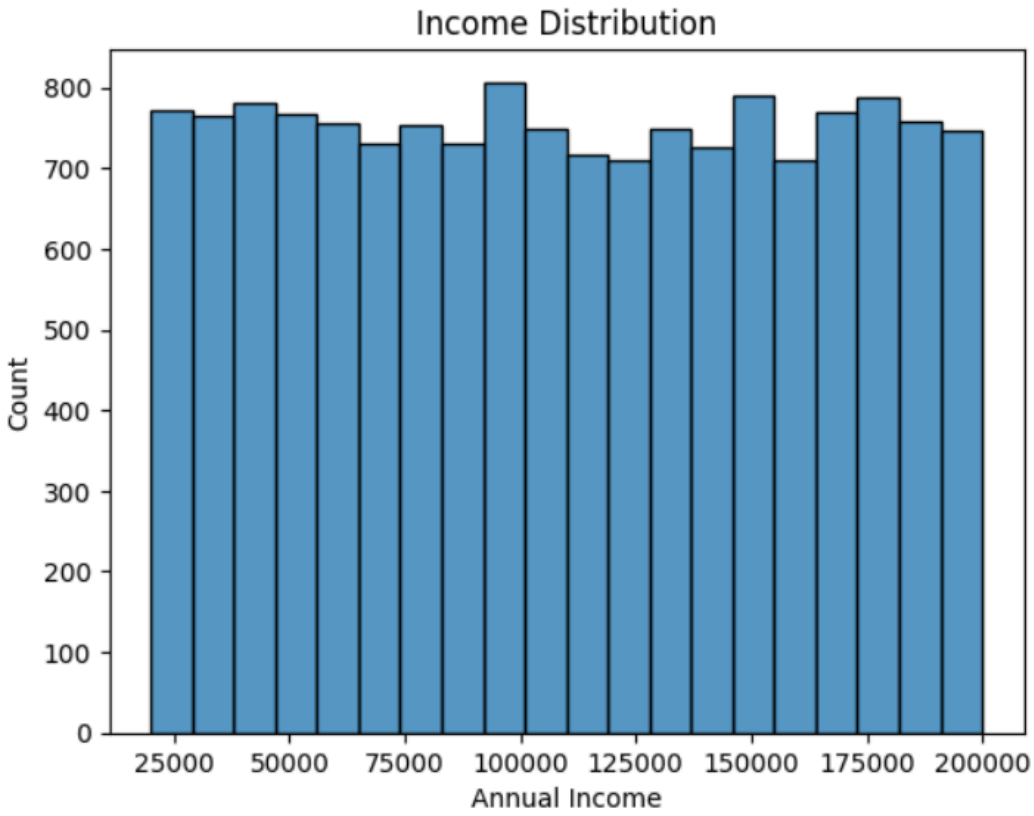
b) Customer Behavior Analysis

• Customers with similar income show different spending patterns

• Indicates that spending is influenced by multiple factors





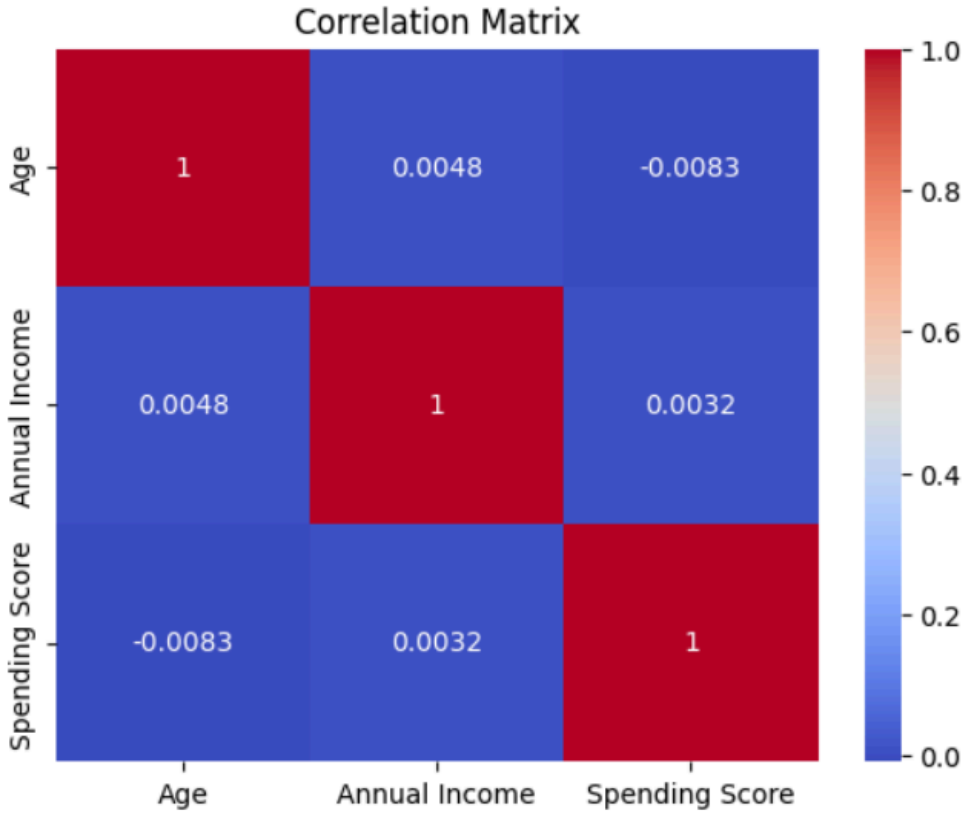


c) Model Performance

• Model successfully predicted spending score

• Error was within acceptable range

• Demonstrates that regression can estimate customer behavior



d) Key Insights

• Income alone does not determine spending behavior

• Customer segmentation is necessary for better targeting

• Machine learning helps in predicting customer actions

2.3.5 Recommendations

• Use customer segmentation for targeted marketing
• Personalize offers based on predicted spending behavior
• Include more features such as lifestyle and purchase history
• Apply advanced models like Random Forest for better accuracy

2.3.6 Conclusion

This project demonstrates the application of supervised learning for predicting customer spending behavior. By using linear regression, it was possible to estimate spending scores based on demographic and financial features.

The results highlight the importance of data-driven decision-making in business. The project also shows that machine learning can help organizations understand customer behavior and improve marketing strategies.

2.4 Cryptocurrency Market Segmentation using Unsupervised Learning (Week 4)

2.4.1 Introduction

The cryptocurrency market has grown rapidly over the years, with thousands of digital assets exhibiting diverse behaviors in terms of price, market capitalization, trading volume, and volatility. Understanding these patterns is essential for investors, analysts, and financial institutions to make informed decisions.

This project focuses on segmenting cryptocurrencies into meaningful groups using unsupervised machine learning techniques. Unlike supervised learning, unsupervised learning does not rely on labeled data but instead identifies hidden patterns and structures within the dataset.

The dataset used includes key financial attributes such as market capitalization, trading volume, price changes, and other performance indicators. By applying clustering techniques such as K-Means, the goal is to group cryptocurrencies based on similar characteristics and uncover market behavior patterns.

The insights derived from this segmentation help in identifying high-value assets, volatile cryptocurrencies, and stable investment options.

2.4.2 Objectives

Primary

Objectives

- To analyse cryptocurrency market data and identify patterns
- To apply unsupervised learning techniques for clustering
- To segment cryptocurrencies based on similar characteristics
- To visualize clusters and interpret their meaning

Specific Analytical Goals

- Perform data cleaning and preprocessing
- Standardize features for clustering
- Apply K-Means clustering algorithm
- Determine optimal number of clusters using Elbow Method
- Evaluate cluster quality using silhouette score
- Visualize clusters using PCA and t-SNE
- Interpret clusters to derive financial insights

2.4.3 Methodology

a) Dataset Preparation

• Loaded cryptocurrency dataset into Python
• Selected key features:

- Market capitalization
- Trading volume
- Price
- Percentage change

• Cleaned dataset:

- Removed duplicates
- Handled missing values
- Ensured consistent data types

• Prepared dataset for analysis

b) Data Preprocessing

• Applied **StandardScaler** to normalize features
• Ensured all variables are on the same scale
• Improved clustering performance

c) Exploratory Data Analysis (EDA)

• Analysed feature distributions
• Used heatmaps to understand correlations
• Identified outliers using boxplots
• Applied PCA for dimensionality reduction

d) Clustering using K-Means

• Applied **K-Means algorithm**

• Used Elbow Method to determine optimal number of clusters

• Assigned cluster labels to each cryptocurrency

e) Cluster Evaluation

• Used **Silhouette Score** to evaluate cluster quality

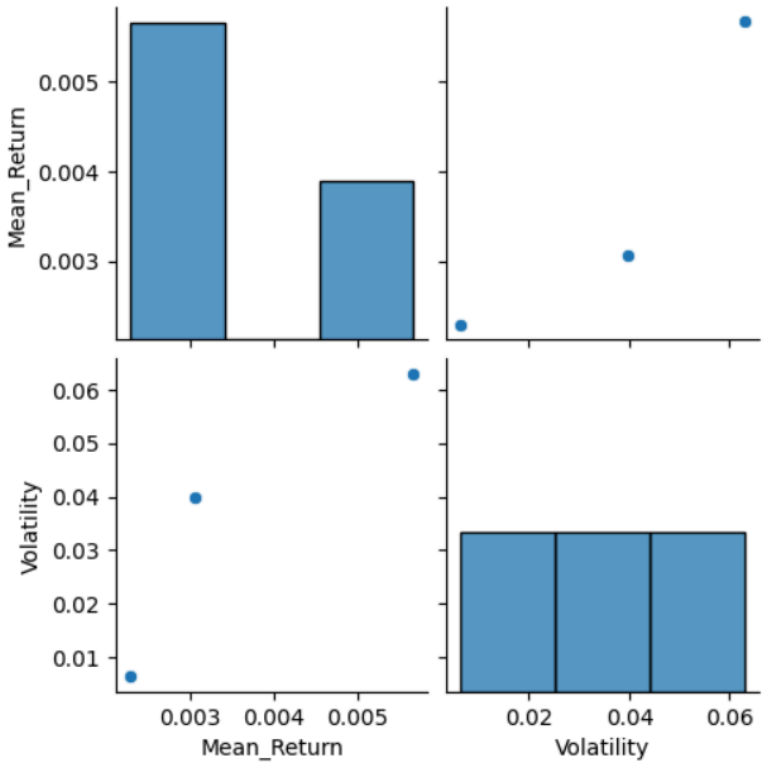
• Analysed cluster separation and cohesion

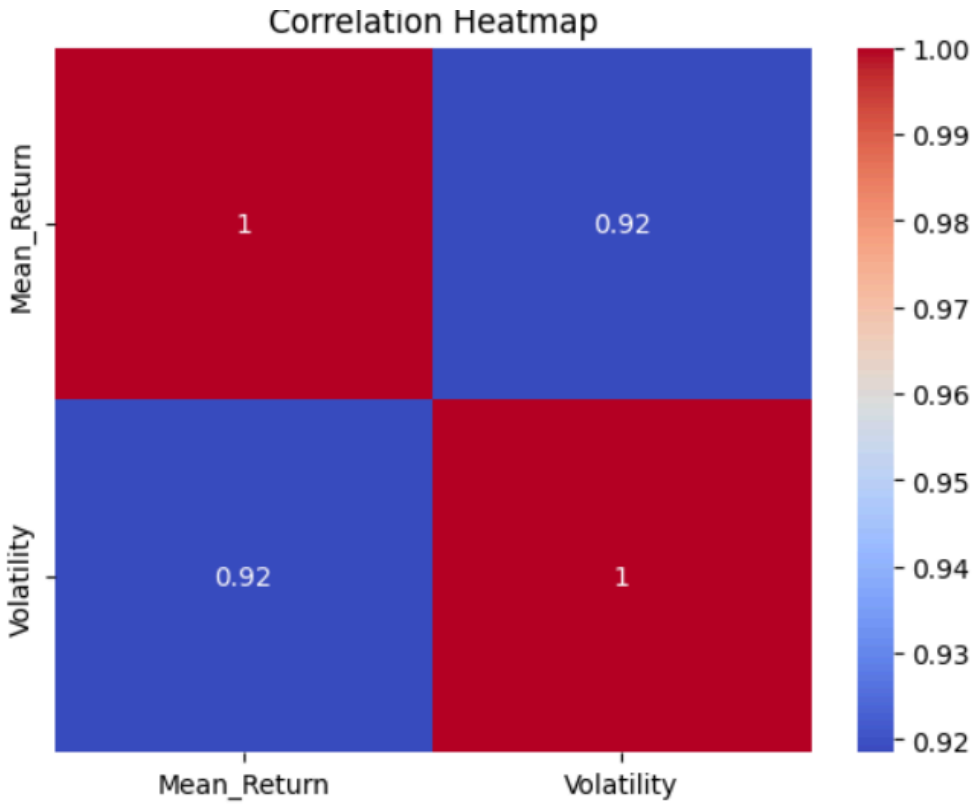
f) Visualization Techniques

• PCA plots to visualize clusters in 2D space

• t-SNE visualization for better cluster representation

• Scatter plots for cluster distribution



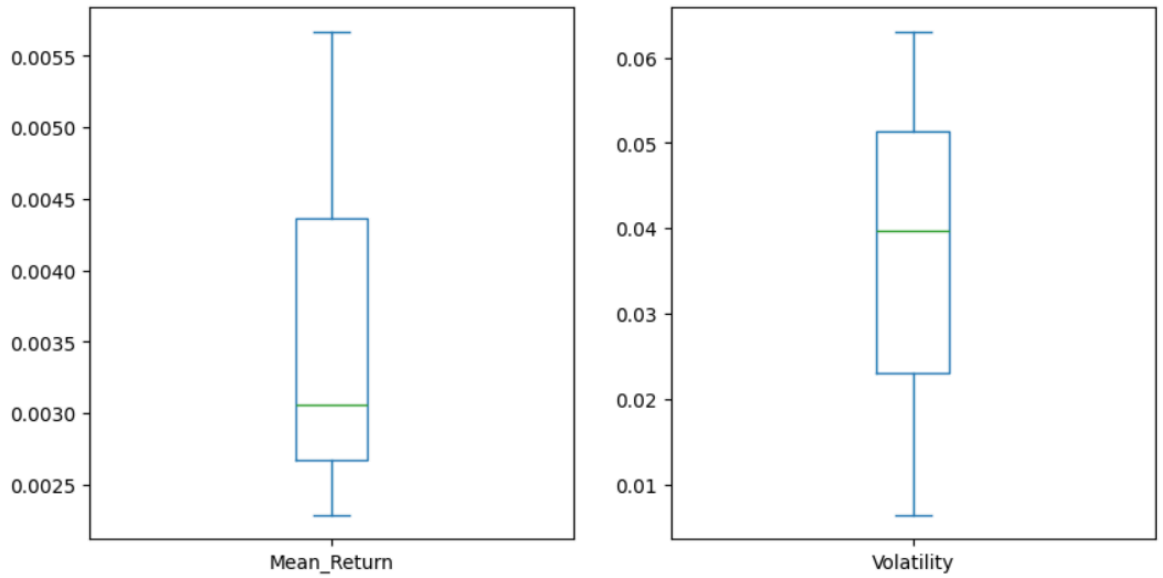


2.4.4 Results and Insights

a) Cluster Formation

• Cryptocurrencies were grouped into distinct clusters

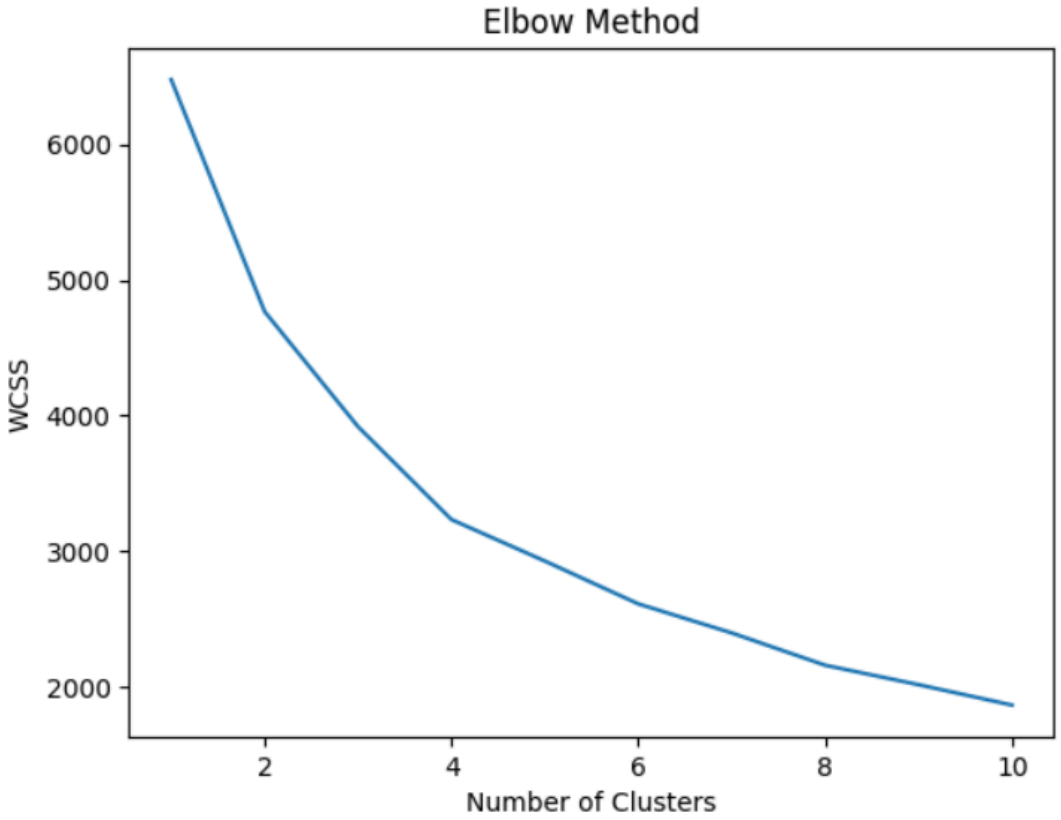
• Each cluster represents a different market segment

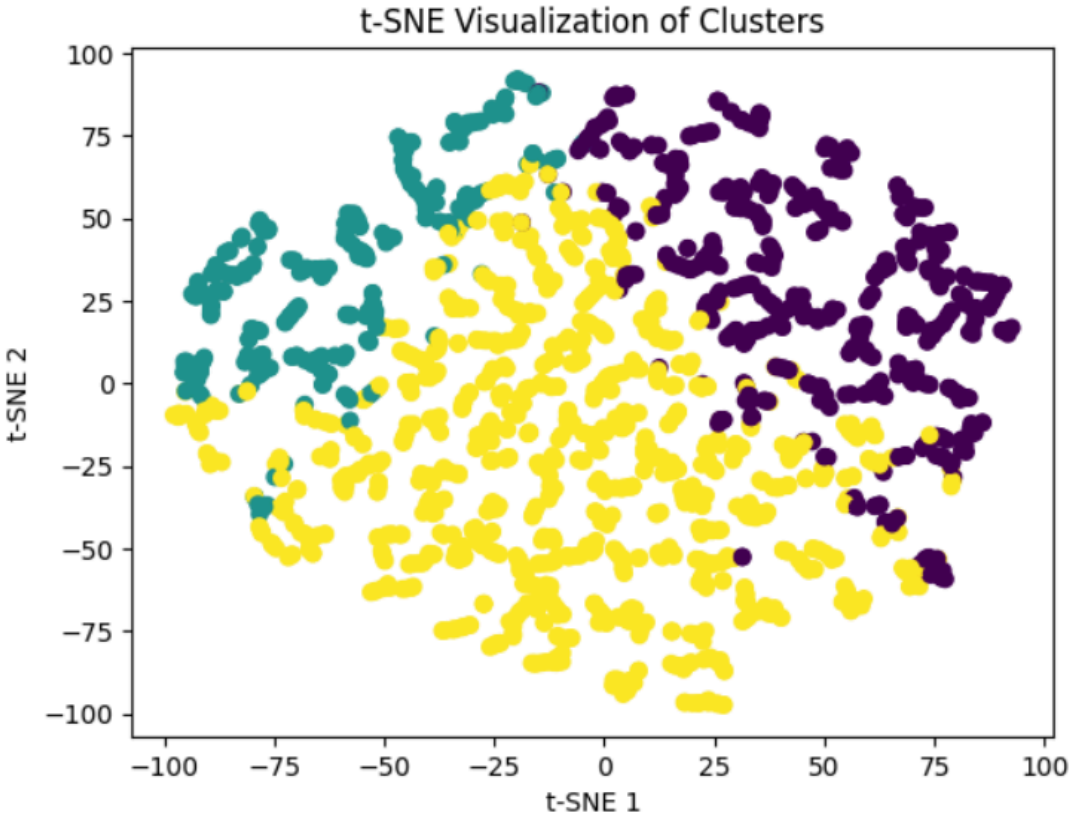


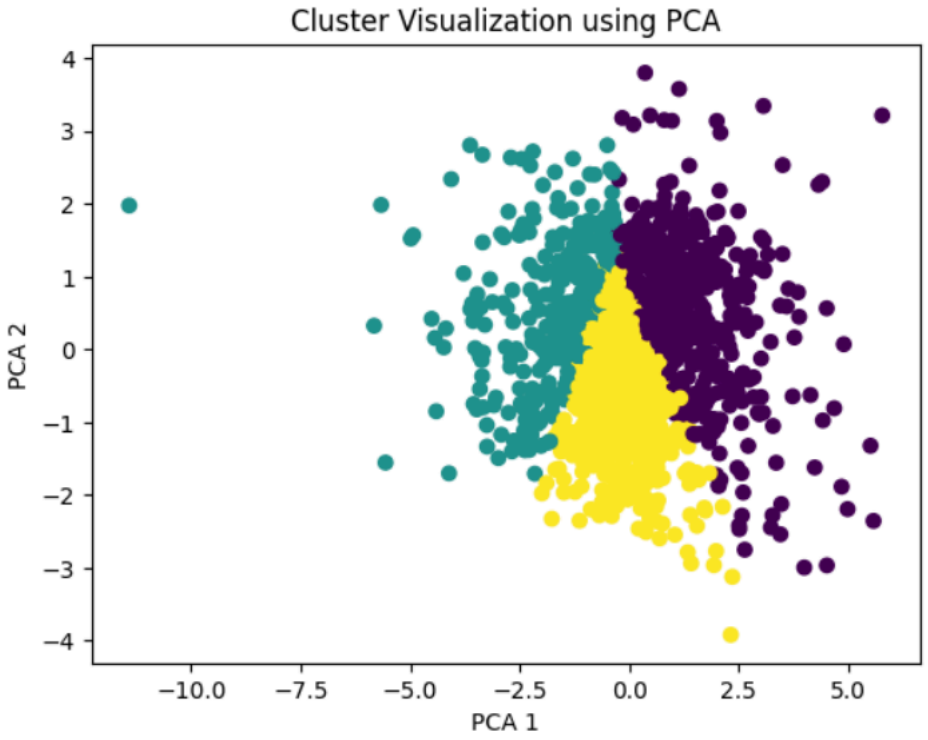
b) High-Value Assets

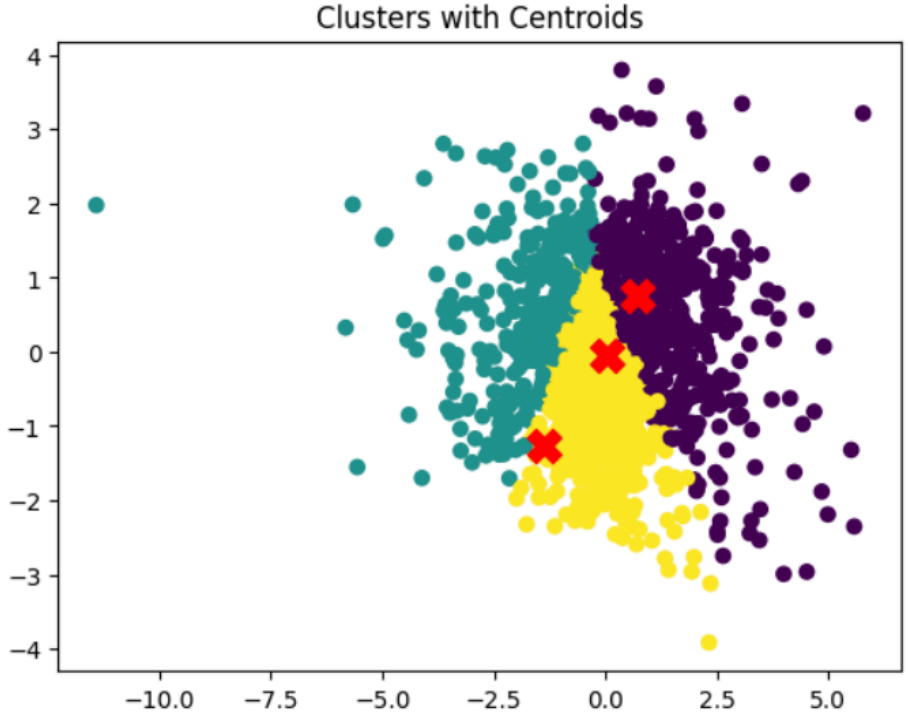
• Some clusters contain cryptocurrencies with:

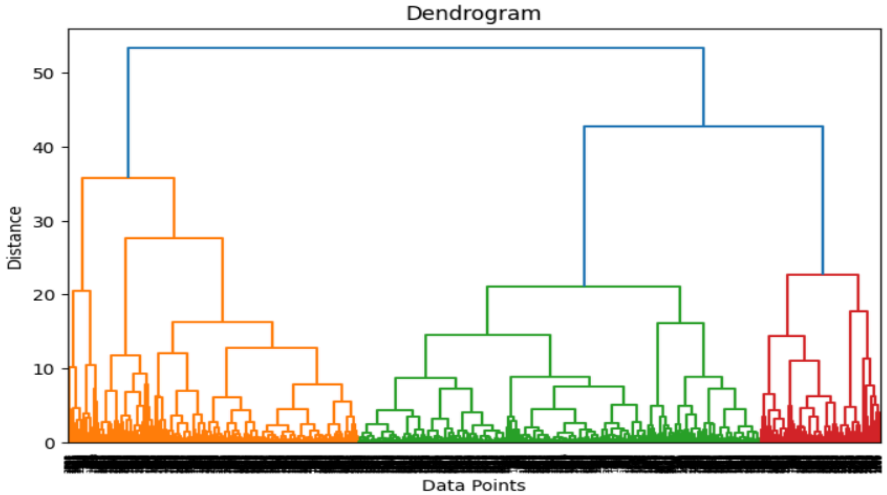
- High market capitalization
- High trading volume
  • Represent stable and popular assets

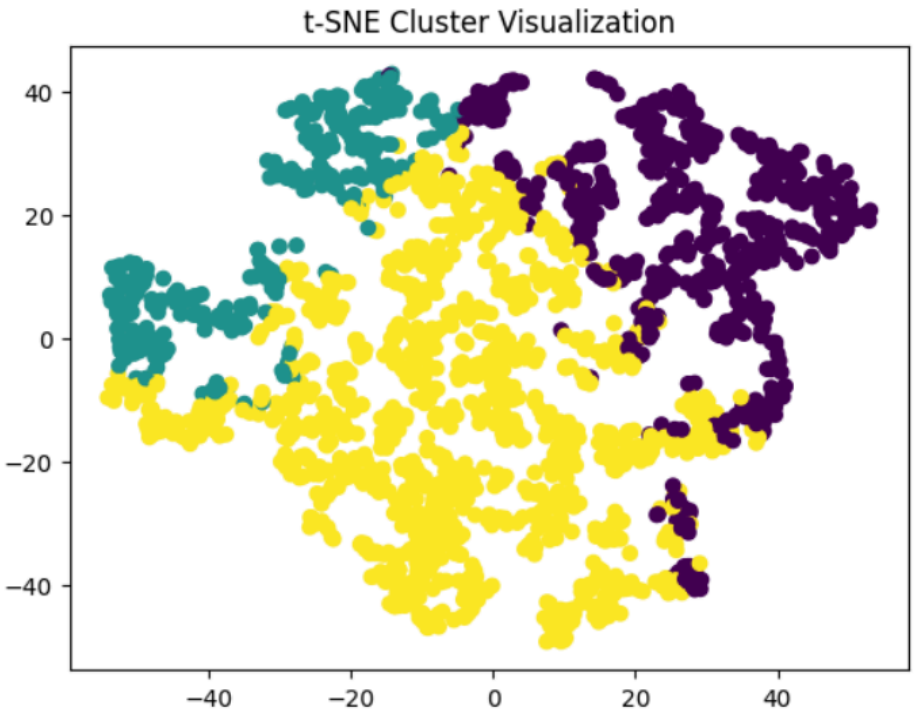





Cluster Visualization using PCA
PCA 2
PCA 1
4
3
2
1
0
−1
−2
−3
−4
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0


Clusters with Centroids
4
3
2
1
0
−1
−2
−3
−4
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0


Dendrogram
Distance
Data Points
50
40
30
20
10
0



c) Volatile Assets

• Certain clusters show:

- High price fluctuations
- Low stability
  • Indicate high-risk investment options

d) Low-Activity Assets

• Some clusters include:

- Low market capitalization
- Low trading volume
  • Represent less active or emerging cryptocurrencies

e) Visualization Insights

• PCA plots show overlapping but distinguishable clusters
• t-SNE provides clearer separation of clusters
• Confirms effectiveness of clustering approach

2.4.5 Conclusion

This project demonstrates the application of unsupervised learning techniques in analyzing cryptocurrency market data. By using K-Means clustering, cryptocurrencies were successfully grouped into meaningful segments based on their financial characteristics.

The analysis highlights the presence of different categories of assets, such as high-value, volatile, and low-activity cryptocurrencies. These insights are valuable for investors and analysts in understanding market behavior and making informed decisions.
The project also emphasizes the importance of data preprocessing, feature scaling, and visualization in achieving effective clustering results. Overall, it showcases how machine learning can be applied to extract actionable insights from complex financial data.

2.4.6 Key Findings

• Cryptocurrency market can be segmented into distinct groups
• High-value assets show stable behavior
• Volatile assets exhibit high risk and fluctuations
• Low-cap assets represent emerging or inactive markets
• Feature scaling is essential for clustering accuracy
• Visualization techniques help interpret clustering results

2.4.7 Business Actions

• Use clustering to identify investment strategies
• Focus on stable assets for long-term investment
• Monitor volatile assets for short-term opportunities
• Track emerging assets for potential growth
• Apply clustering in portfolio management systems

2.5 Neural Network Model for Classification using Deep Learning (Week 5)

2.5.1 Introduction

Deep Learning has become a powerful approach in modern data science, enabling machines to learn complex patterns from large datasets. Artificial Neural Networks (ANNs) are one of the fundamental architectures in deep learning, inspired by the structure and functioning of the human brain.

This project focuses on building a classification model using a neural network to solve a real-world problem such as fraud detection, customer classification, or pattern recognition. The objective is to design, train, and evaluate an ANN model using deep learning techniques.

The dataset used contains multiple input features representing real-world attributes, along with a target variable indicating class labels. Using libraries such as TensorFlow and Keras, the neural network is trained to learn patterns and make accurate predictions.

The project demonstrates how deep learning can outperform traditional machine learning methods in handling complex, non-linear relationships in data.

2.5.2 Objectives

Primary

Objectives

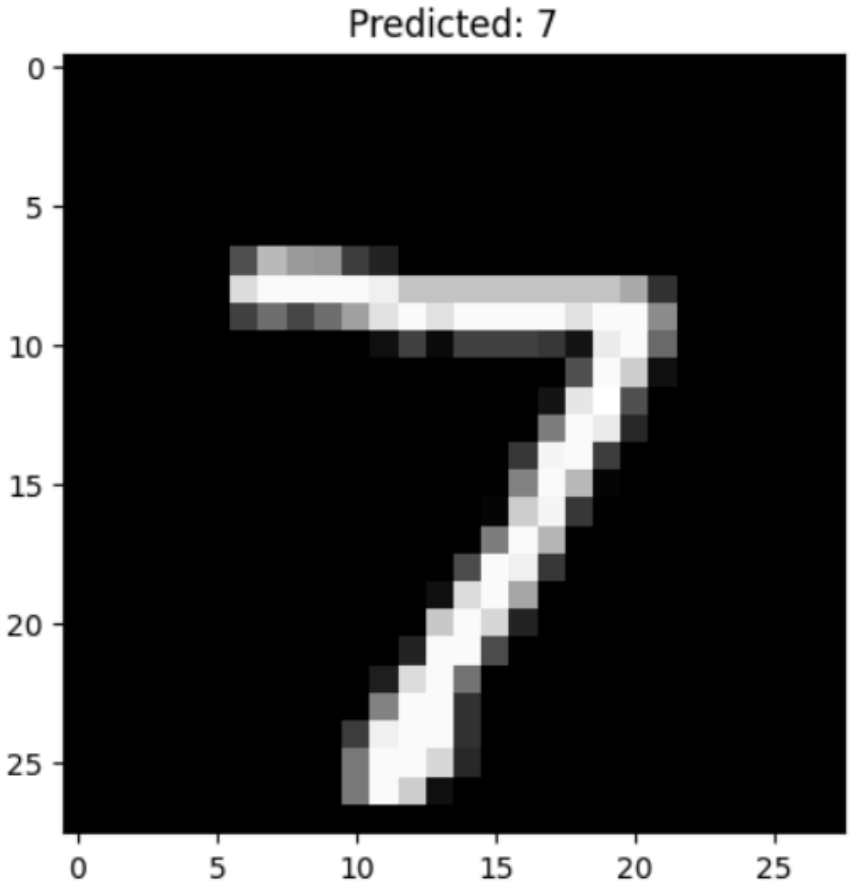
- To understand the fundamentals of neural networks and deep learning
- To build a classification model using ANN
- To train the model using real-world data
- To evaluate model performance using classification metrics

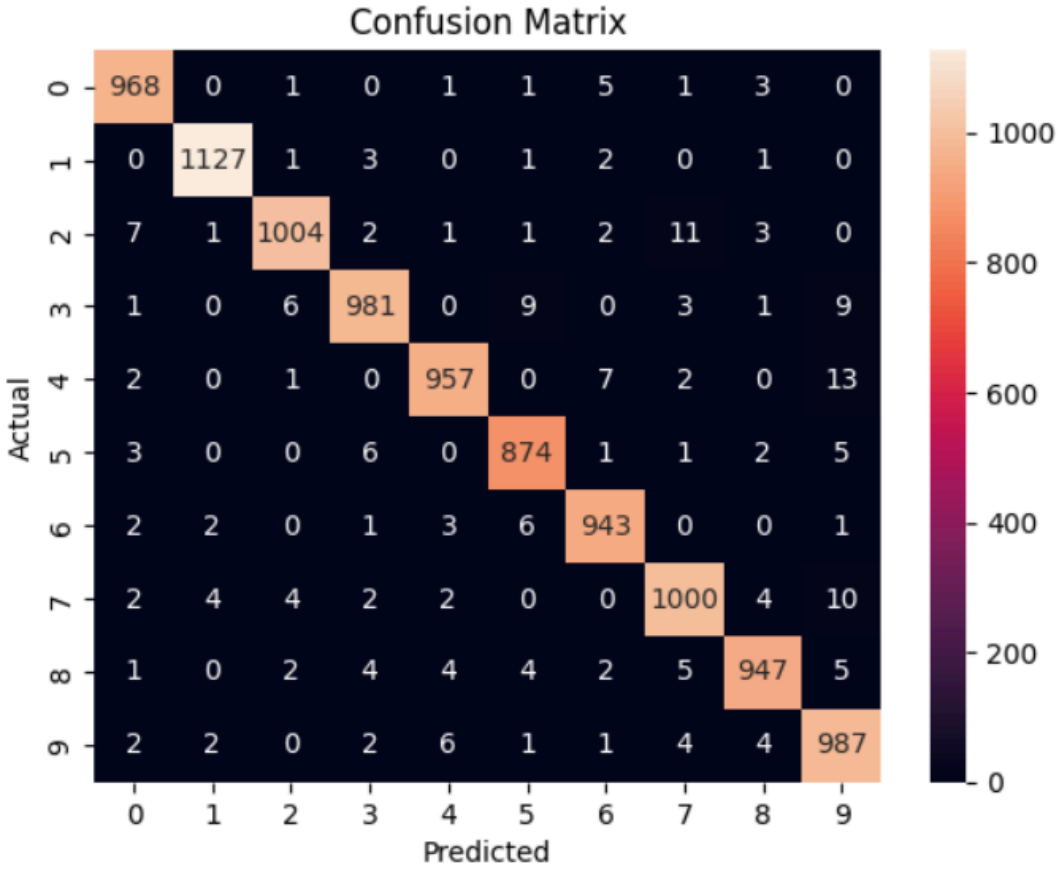
Specific Analytical Goals

- Perform data preprocessing and feature scaling
- Design neural network architecture

• Apply activation functions and optimization techniques

• Train the model using epochs and batch processing





• Evaluate performance using accuracy, precision, recall, and F1-score
• Analyze model performance using confusion matrix

2.5.3 Methodology

a) Dataset Preparation

• Loaded dataset into Python environment
• Identified input features and target variable
• Checked for missing values and inconsistencies
• Cleaned and structured dataset for training

b) Data Preprocessing

• Encoded categorical variables if present
• Applied feature scaling using StandardScaler
• Split dataset into training and testing sets

c) Model Architecture (ANN Design)

• Designed a neural network using Keras:

- Input layer
- Hidden Layer 1 (Dense + ReLU activation)
- Dropout layer (to reduce overfitting)
- Hidden Layer 2 (Dense + ReLU)
- Output layer (Sigmoid for binary classification / Softmax for multi-class)

d) Model Training

• Compiled model using:

- Optimizer: Adam
- Loss function: Binary Crossentropy
  • Trained model using:
- Number of epochs (e.g., 20)
- Batch size (e.g., 32)

e) Model Evaluation

• Evaluated model performance using:

- Accuracy
- Precision
- Recall
- F1-score
  • Generated confusion matrix
  • Compared predicted vs actual values

2.5.4 Results and Insights

a) Model Performance

• Neural network achieved high classification accuracy
• Model successfully learned patterns in the dataset

b) Prediction Insights

• Model was able to distinguish between different classes effectively
• Reduced classification errors compared to simpler models

c) Confusion Matrix Analysis

• High number of correct predictions
• Lower false positives and false negatives
• Improved reliability of predictions

d) Learning Behavior

• Model accuracy improved over epochs
• Loss decreased steadily during training
• Indicates effective learning process

2.5.5 Conclusion

This project demonstrates the application of deep learning using Artificial Neural Networks for classification tasks. The model successfully learned complex relationships within the data and provided accurate predictions.

The results highlight the effectiveness of neural networks in solving real-world problems, especially where traditional machine learning methods may struggle. Proper preprocessing, architecture design, and training techniques are essential for achieving optimal performance.

Overall, this project showcases the importance of deep learning in modern data analytics and its potential for future applications.

2.5.6 Key Findings

• Neural networks can model complex relationships effectively
• Feature scaling significantly improves performance
• Dropout helps prevent overfitting
• Model performance improves with proper tuning
• Deep learning provides better accuracy than traditional methods

2.5.7 Business Applications

- Fraud detection systems
- Customer classification and segmentation
- Recommendation systems
- Medical diagnosis prediction
- Risk assessment models

2.6 Fraud Detection using Generative AI (CTGAN) (Week 6)

2.6.1 Introduction

Fraud detection is a critical challenge in financial systems, where fraudulent transactions account for only a very small portion of the total data. This extreme class imbalance makes it difficult for traditional machine learning models to detect fraud accurately, as they tend to bias toward normal transactions.

This project focuses on enhancing fraud detection using Generative Artificial Intelligence, specifically the Conditional Tabular Generative Adversarial Network (CTGAN). CTGAN is designed to generate realistic synthetic tabular data by learning the underlying distribution of the original dataset.

The dataset used in this project is the Credit Card Fraud Detection dataset, which contains anonymized transaction features (V1 to V28), along with transaction time, amount, and a target variable indicating fraud status.

The main objective is to generate synthetic fraud data using CTGAN, augment the dataset, and improve the performance of machine learning models in detecting fraudulent transactions.

2.6.2 Objectives

Primary

Objectives

- To address class imbalance in fraud detection datasets
- To generate realistic synthetic fraud data using CTGAN
- To improve model performance using augmented data
- To enhance fraud detection accuracy and recall

Specific Analytical Goals

• Analyse fraud vs non-fraud distribution
• Train CTGAN model on fraud transactions
• Generate synthetic fraud samples
• Combine synthetic and original data
• Train machine learning model on augmented dataset
• Compare performance before and after augmentation

2.6.3 Methodology

a) Dataset Preparation

• Loaded the *creditcard.csv* dataset into Python
• Identified features:

- Time
- Amount
- V1 to V28 (PCA transformed features)
  • Target variable:
- Class (0 = Normal, 1 = Fraud)

• Checked for missing values and ensured data consistency
• Observed severe class imbalance (fraud < 1%)

b) Exploratory Data Analysis (EDA)

• Analysed class distribution using count plots
• Identified imbalance between fraud and non-fraud cases
• Used boxplots to analyse skewness in transaction amount
• Generated correlation heatmap for feature relationships
• Applied PCA to visualize data distribution

c) Generative AI using CTGAN

• Selected fraud transactions as training data
• Trained CTGAN model to learn fraud patterns
• Generated synthetic fraud samples (e.g., 5000 records)
• Verified similarity between original and synthetic data distributions

d) Data Augmentation

• Combined original dataset with synthetic fraud data
• Shuffled dataset to remove bias
• Created a balanced dataset for training

e) Model Training and Evaluation

• Applied **Random Forest Classifier**
• Split dataset into training and testing sets
• Trained model on:

- Original dataset (before augmentation)
- Augmented dataset (after synthetic data)

• Evaluated model using:

- Accuracy
- Precision
- Recall
- F1-score

• Generated confusion matrix for performance comparison

2.6.4 Results and Insights

a) Class Imbalance Insight

• Fraud transactions constituted less than 1% of the dataset

• Initial model performance was biased toward normal transactions

b) Model Performance (Before Augmentation)

• High accuracy but poor fraud detection

• Low recall for fraud class

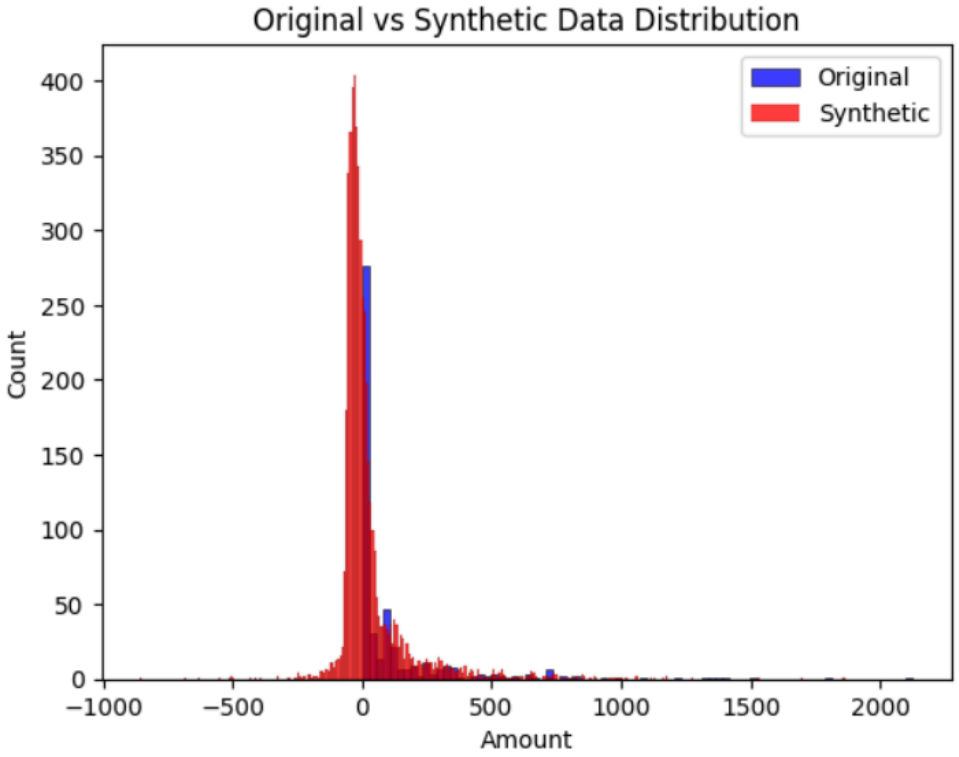
• Many fraud cases were misclassified

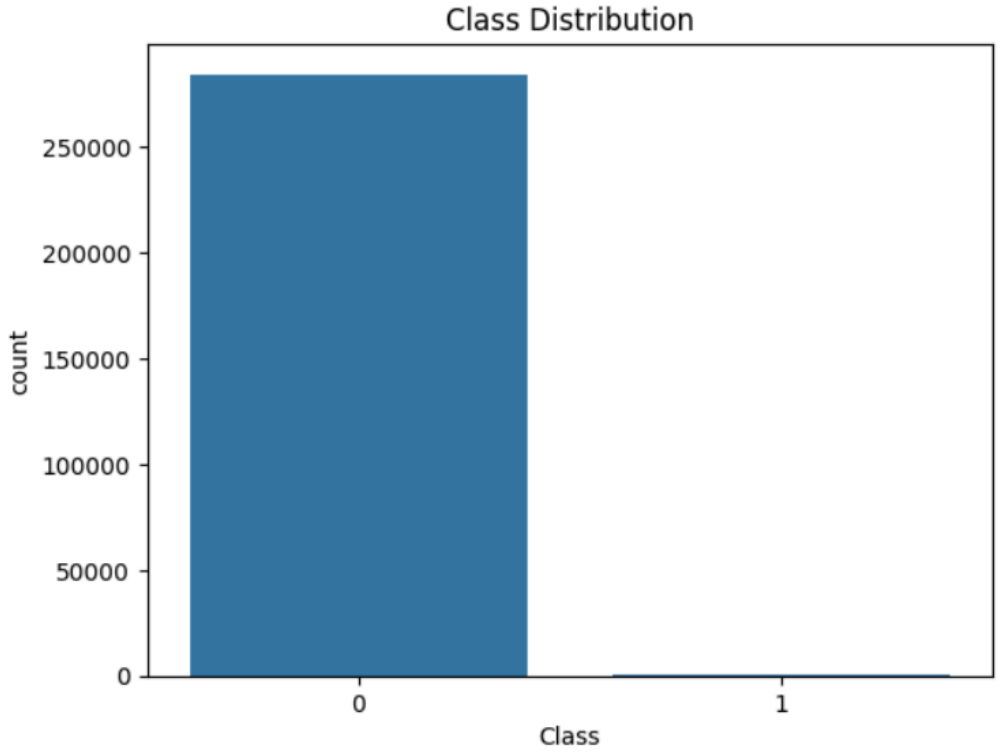
c) Model Performance (After Augmentation)

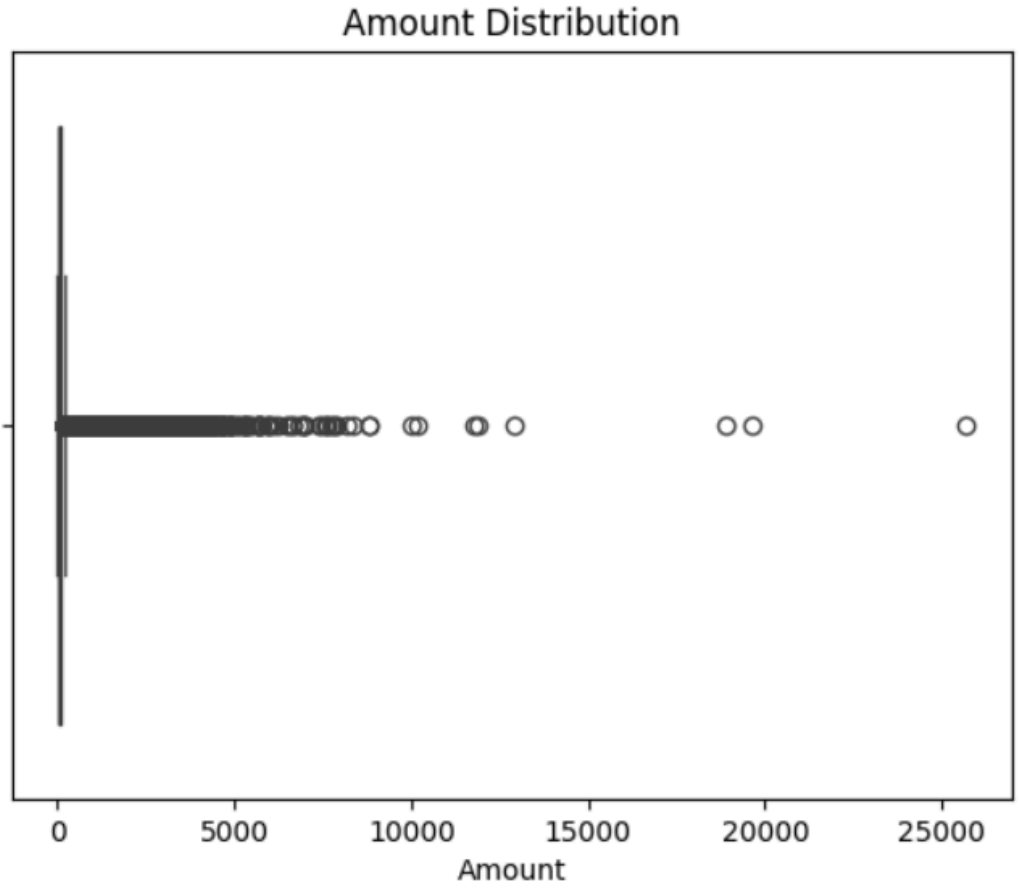
• Significant improvement in fraud detection

• Increased recall and F1-score

• Better identification of fraud cases







d) Confusion Matrix Analysis

• Increase in correctly detected fraud transactions

• Reduction in false negatives (missed fraud cases)

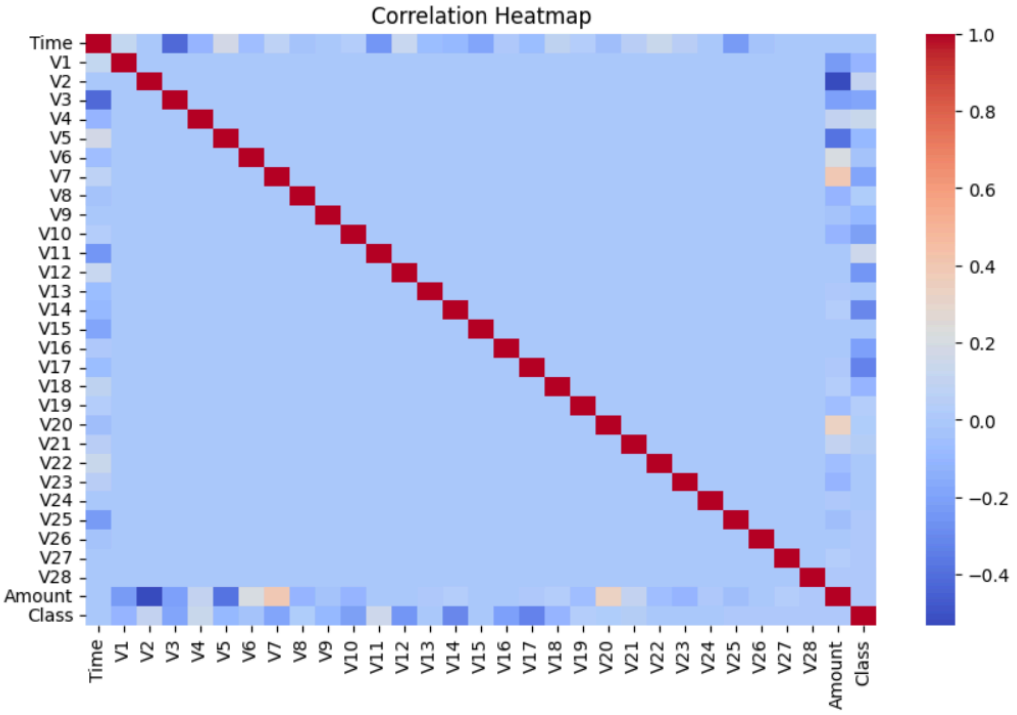
• Improved reliability of the model

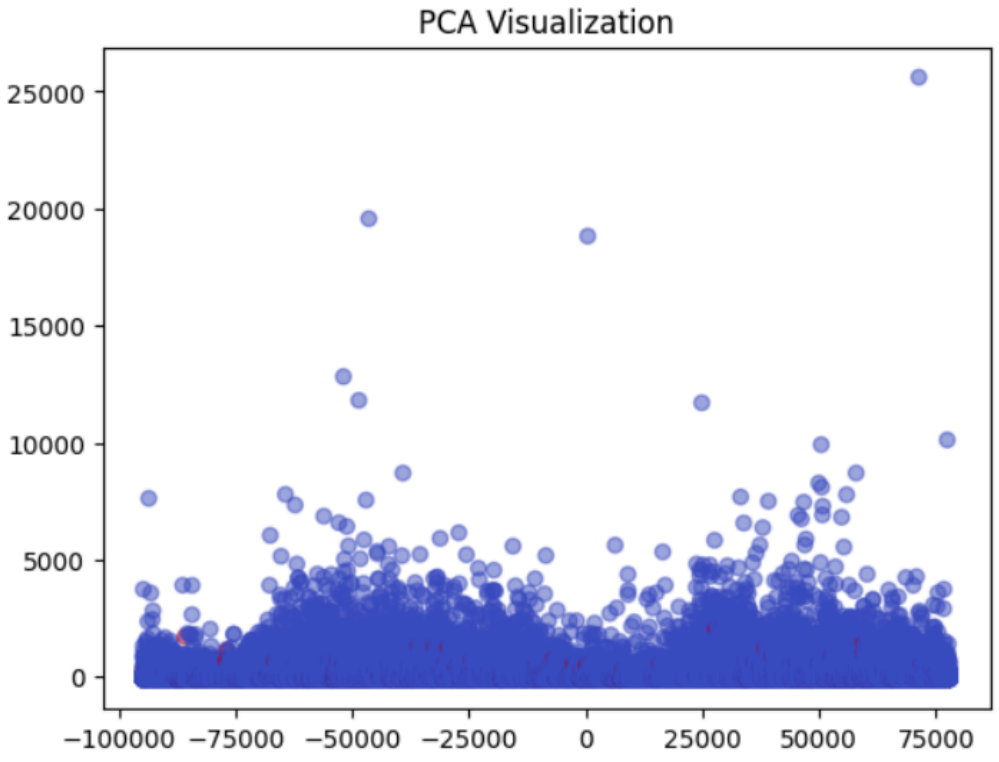
e) Synthetic Data Insights

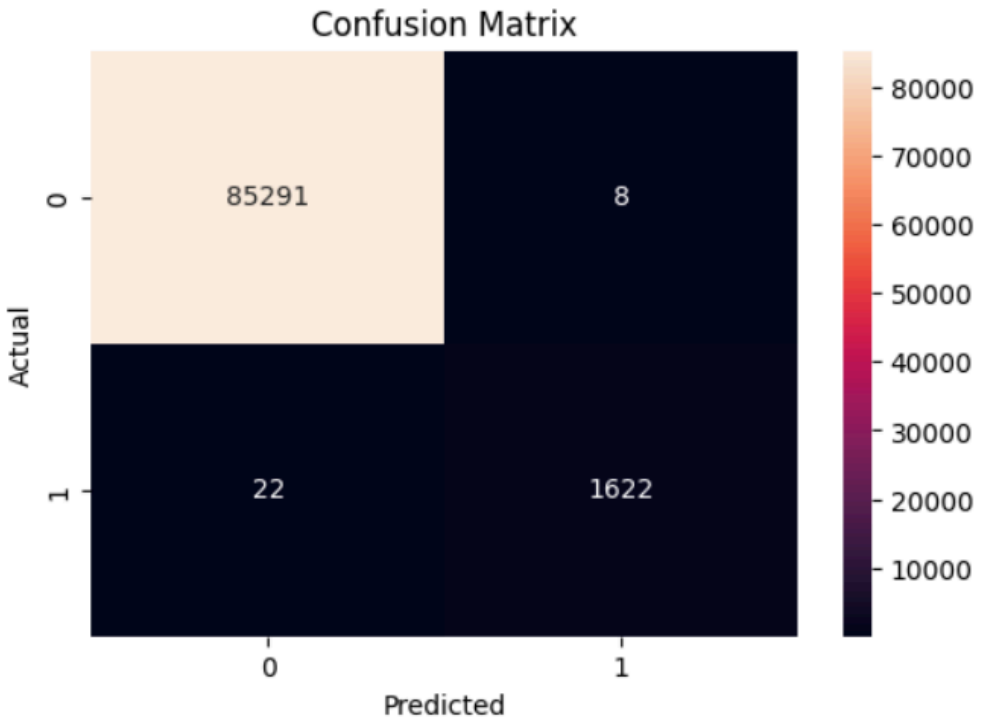
• Generated data closely resembled real fraud patterns

• Helped model generalize better

• Demonstrated effectiveness of Generative AI in tabular data







2.6.5 Conclusion

This project successfully demonstrates the use of Generative AI (CTGAN) to address the problem of class imbalance in fraud detection. By generating synthetic fraud data, the model's ability to detect fraudulent transactions improved significantly.

The results highlight that data augmentation using CTGAN enhances recall and overall model performance. This approach can be effectively applied to other real-world problems involving imbalanced datasets.

The project emphasizes the importance of combining machine learning with generative models to build more robust and reliable predictive systems.

2.6.6 Key Findings

- Fraud datasets are highly imbalanced
- Synthetic data improves model performance
- Recall is a critical metric in fraud detection
- CTGAN effectively generates realistic tabular data
- Data augmentation enhances model reliability

2.6.7 Business Applications

• Real-time fraud detection systems
• Banking and financial security
• Risk management systems
• Transaction monitoring platforms
• Insurance fraud detection

CHAPTER 5: CONCLUSION

5.1 Overall Learning Outcomes

The internship at Global Next Consulting India Pvt. Ltd. provided comprehensive exposure to the end-to-end workflow of data science and machine learning, including data collection, preprocessing, analysis, modeling, and visualization.

Through six structured projects, I gained hands-on experience with tools and technologies such as Python, SQL, Excel, Power BI, Tableau, and machine learning libraries. These projects covered key domains of data analytics, including data analysis, supervised learning, unsupervised learning, deep learning, and generative AI.Each weekly project contributed to building a strong foundation in different areas:

- Week 1 focused on basic data analysis using Excel
- Week 2 introduced advanced data handling using SQL and Excel
- Week 3 covered supervised learning for prediction tasks
- Week 4 involved unsupervised learning for clustering and segmentation
- Week 5 explored deep learning using neural networks
- Week 6 integrated Generative AI using CTGAN for synthetic data generation

The major learning outcome was the ability to apply theoretical concepts to real-world problems. The final project on fraud detection using Generative AI combined machine learning and data augmentation techniques to solve the issue of class imbalance, demonstrating a practical industry-level solution.The internship also enhanced essential professional skills such as analytical thinking, problem-solving, logical reasoning, and effective communication.

5.2 Applications of Work

The knowledge and skills acquired during this internship can be applied across multiple real-world domains:

- **Business Analytics:** Developing dashboards and models to analyze trends and improve decision- making

- **Finance & Fraud Detection:** Identifying fraudulent transactions using machine learning and Generative AI techniques

• **Customer Analytics:** Predicting customer behavior and segmenting users for targeted marketing

• **Market Analysis:** Using clustering techniques to understand patterns in financial or cryptocurrency markets

• **Artificial Intelligence Systems:** Building predictive and intelligent systems using deep learning models

• **Data Science & Research:** Applying data-driven approaches to solve complex real-world problems

Internship Certificate

SUMMARY

The internship provided in-depth and practical exposure to various data analysis, machine learning, and visualization techniques, enabling hands-on experience with multiple tools such as Python, Advanced Excel, SQL, R, Tableau, Power BI, and deep learning frameworks. Each week focused on solving real-world problems through structured workflows — from data collection and preprocessing to modeling, visualization, and interpretation.

Across the six projects, the work covered a wide spectrum of analytical and machine learning domains:

- **Week 1 (Basic Data Analysis using Excel)** – Performed data cleaning, preprocessing, and analysis using Excel. Utilized pivot tables, charts, and statistical functions to extract insights and create interactive dashboards for data visualization.

- **Week 2 (Data Analysis using Advanced Excel & MySQL)** – Processed and analyzed structured datasets using SQL and Excel by performing data integration, cleaning, and transformation. Generated insights through queries, aggregations, and dashboard-based reporting.

- **Week 3 (Customer Segmentation and Spending Prediction using Supervised Learning)** – Applied machine learning techniques using Python to analyze customer data and predict spending behavior. Performed data preprocessing, exploratory data analysis, feature selection, and implemented regression models for prediction.

- **Week 4 (Cryptocurrency Market Segmentation using Unsupervised Learning)** – Used clustering techniques such as K-Means to segment cryptocurrencies based on market features. Applied preprocessing, scaling, PCA, and visualization techniques to identify patterns and group similar assets.

- **Week 5 (Neural Network Model using Deep Learning)** – Developed a classification model using Artificial Neural Networks (ANN). Applied feature scaling, designed model architecture using Keras, trained the model using optimization techniques, and evaluated performance using classification metrics.

- **Week 6 (Fraud Detection using Generative AI – CTGAN)** – Implemented Generative AI techniques to address class imbalance in fraud detection. Trained CTGAN to generate synthetic fraud data, combined it with original data, and improved model performance using machine learning algorithms.

Through these projects, both technical and analytical competencies were significantly enhanced, including data preprocessing, statistical reasoning, machine learning, deep learning, and business intelligence. The internship strengthened the ability to transform raw data into meaningful insights and develop predictive models for real-world applications.

Overall, this experience successfully bridged the gap between theoretical knowledge and practical implementation, building confidence and readiness for professional roles in data science, machine learning, and analytics.

REFERENCES

1. Kaggle Datasets – Credit Card Fraud Detection Dataset, Mall Customers Dataset, Cryptocurrency Market Dataset
2. Kaggle / Public Data Sources – Sample datasets used for Excel-based data analysis and SQL projects Microsoft Documentation – Excel Analytics, Advanced Excel Functions, Data Cleaning, and Dashboard Creation
3. MySQL Documentation – Database Management, SQL Queries, Joins, Aggregation Functions, and Data Integration Techniques

4. Python.org – Documentation for Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn
5. TensorFlow & Keras Documentation – Neural Network Architecture, Model Training, and Deep Learning Techniques
6. CTGAN Documentation (SDV / CTGAN Library) – Synthetic Data Generation using Generative Adversarial Networks
7. Scikit-learn Documentation – Supervised Learning, Unsupervised Learning, Clustering Algorithms, and Model Evaluation Metrics
8. Research Articles & Online Resources –
    - "Applications of Machine Learning in Customer Behavior Prediction"
    - "Unsupervised Learning Techniques in Market Segmentation"
    - "Generative AI for Tabular Data Synthesis"
9. Tableau & Power BI Documentation – Data Visualization, Dashboard Creation, and Business Intelligence Techniques
10. W3Schools / HackerRank – SQL Queries, Joins, CTEs, UNION, and Aggregation Functions
11. Online Tutorials & Learning Platforms – Concepts related to Machine Learning, Deep Learning, and Data Science