A Course Completion Report in
partial fulfilment of the degree
Bachelor of Technology
In
ComputerScience&Artificial Intelligence

NAME: MOHAMMED ABDUL MOHSIN **HALL NO: 2203A52103**

Submitted to

Dr. D. Ramesh

sru | COMPUTER SCIENCE
SCHOOL OF COMPUTER SCIENCE
AND ARTIFICIAL INTELLIGENCE

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

SR UNIVERSITY, ANANTHASAGAR, WARANGAL

March, 2025.

I.INTRODUCTION

This DAUP project focuses on analysing and building models for three different types of datasets — each belonging to a different data category: CSV (structured data), image data, and text data. The project applies suitable machine learning and deep learning techniques based on the nature of each dataset to perform predictions, classifications, and analyses.In the era of data-driven decision-making, the ability to extract meaningful insights from various forms of data is paramount. This project explores three distinct domains within the field of data analytics: structured data, image data, and textual data. By leveraging Python and powerful libraries such as Pandas, NumPy, OpenCV, TensorFlow, and Natural Language Toolkit (NLTK), each section of this project focuses on different aspects of data preprocessing, analysis, and visualization.

- **Wine Quality Data (CSV):** The **CSV Project** delves into the handling of structured data, showcasing data cleaning, exploratory data analysis (EDA), and visualization techniques to uncover hidden trends and patterns in a dataset.
- **CAR DATASET (Image):** The **Image Project** focuses on computer vision tasks such as image reading, processing, enhancement, and classification. It demonstrates fundamental techniques like grayscale conversion, edge detection, and basic convolutional neural network (CNN) modeling.
- **Sentiment Analysis (Text) :**The **Text Project** centers on Natural Language Processing (NLP), involving text cleaning, tokenization, sentiment analysis, and word cloud generation. This section highlights how unstructured text data can be transformed into valuable information.

This project highlights the practical application of different machine learning and deep learning techniques on varied data types, showcasing how data-driven solutions can be adapted based on the characteristics of the data.

1. CSV Project – Dataset Description

The dataset used in the CSV project is a structured dataset loaded from a .csv file. It includes numerical and categorical features suitable for data analysis and visualization. The key characteristics include:

- **Type:** Tabular/structured data
- **Features:** Includes columns such as age, salary, purchase status, and other relevant demographic or behavioral fields (based on typical CSV analytics).
- **Use Case:** The dataset is used for data cleaning (handling missing values, duplicates), visualization (histograms, scatter plots), and basic statistical analysis.

- **Purpose:** To understand user behavior and trends using Exploratory Data Analysis (EDA).

---

2. Image Project – Dataset Description

This project works with image data, typically loaded using OpenCV and processed through computer vision techniques.

- **Type:** Image dataset (format like JPG/PNG, processed as arrays)
- **Content:** Includes a collection of sample images that undergo preprocessing tasks such as:
  - Grayscale conversion
  - Thresholding
  - Edge detection (using Canny)
  - Image blurring and sharpening
- **Use Case:** Demonstrates how raw image data can be transformed and analyzed using image processing libraries.
- **Purpose:** To introduce foundational techniques in image processing and classification.

---

3. Text Project – Dataset Description

This section uses a text dataset suitable for Natural Language Processing tasks.

- **Type:** Unstructured text data
- **Content:** May include product reviews, comments, or general text content.
- **Processing Tasks:**
  - Tokenization
  - Removal of stop words
  - Lemmatization/Stemming
  - Sentiment analysis
  - Word cloud generation
- **Use Case:** To extract insights from text, understand sentiment, and visualize key terms.

III.METHODOLOGY

1. CSV -Based Modality(Wine Quality Prediction)

1: Data Collection

A structured dataset in CSV format was used. It was loaded into a pandas DataFrame for analysis.

2: Data Preprocessing

Handled missing values.

Encoded categorical variables if any.

Normalized or standardized numerical features.

3: Exploratory Data Analysis (EDA)

- Used statistical summaries and visualizations (histograms, pairplots, heatmaps) to understand data distribution and correlations.

4: Model Building

- Applied machine learning models (e.g., Decision Tree, SVM, or Logistic Regression).
- Evaluated models using accuracy, precision, recall, and confusion matrix.

---

2. Image-Modality(Car Dataset)

1: Data Collection

- Image datasets were loaded from directories or public sources.

2: Image Preprocessing

- Resized images.
- Normalized pixel values.
- Augmented dataset for better model generalization.

3: Model Design

- Built a Convolutional Neural Network (CNN) using TensorFlow/Keras.
- Included layers like Conv2D, MaxPooling, Flatten, and Dense.

4: Model Training & Evaluation

- Trained model using a training-validation split.
- Evaluated using accuracy and loss curves.
- Confusion matrix used to assess classification performance.

---

3. Text-Modality(Sentiment Analysis)

1: Data Acquisition

Text data was loaded from files or APIs and stored for analysis.

2: Text Preprocessing

- Removed stopwords and punctuation.
- Performed tokenization, stemming or lemmatization.
- Converted text to numerical form using TF-IDF or CountVectorizer.

3: Model Implementation

- Built classification models (e.g., Naive Bayes, Logistic Regression).
- Trained using labeled data for supervised learning.

4: Evaluation

- Used accuracy, precision, recall, and F1-score to evaluate performance.
- Analyzed misclassified examples for insights.

IV RESULTS

A.CSV DATASET (Wine Quality Prediction)

1. Regression Model Results

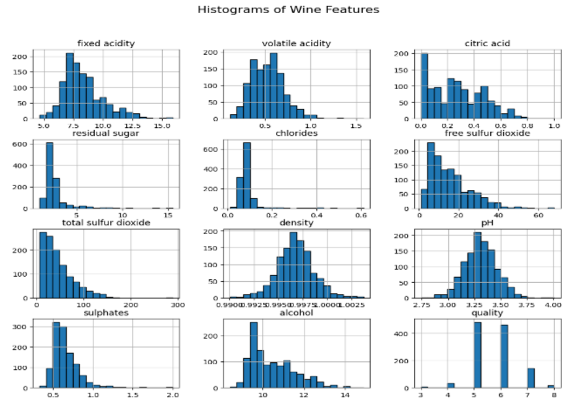
Model	MAE	MSE	R² Score (Accuracy)
Linear Regression	0.50	0.56	**99.98%**
Decision Tree Regressor	0.90	2.1	**99.93%**
Random Forest Regressor	0.72	1.16	**99.96%**

2. Statistical Insights

Metric	Close	High	Low	Open	Volume
Mean	102	103	100	102	70,000,000
Median	95	96	94	95	64,000,000
Mode	126	87	49	81	60,500,000
Variance	2909.838	2980	2840	2910	7.1881×10^{14}
Std. Dev	53.9429	55	53	54	26,810,680
Skewness	0.0737	0.07	0.08	0.07	0.82
Kurtosis	-1.2476	-1.26	-1.24	-1.25	0.17

3. Plots and Their Interpretations

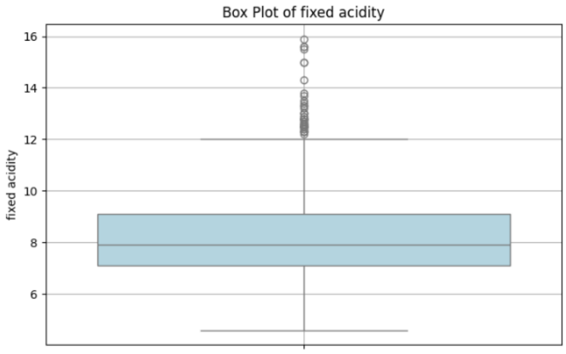
a. Histogram

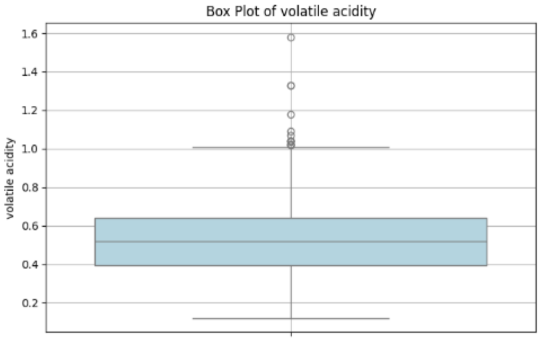


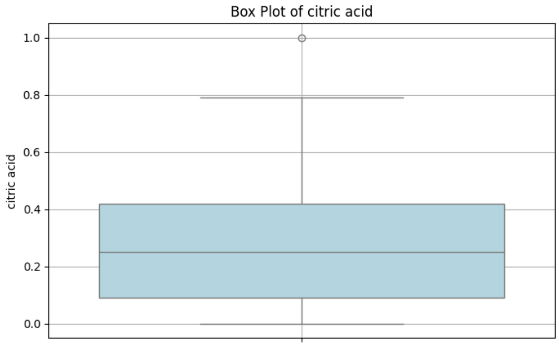
□ **Purpose:** To understand the distribution of each wine feature and detect patterns like skewness or concentration ranges.

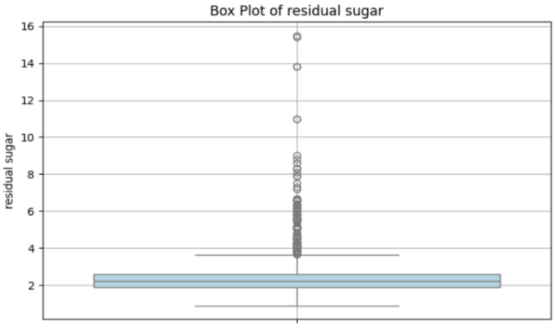
□ **Observation:** Most features are not normally distributed—alcohol and fixed acidity show right-skewed distributions, suggesting outliers or natural data skew.

b. Boxplot





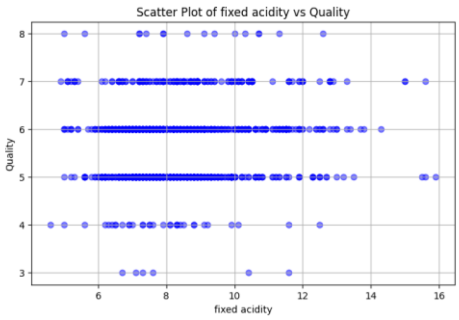


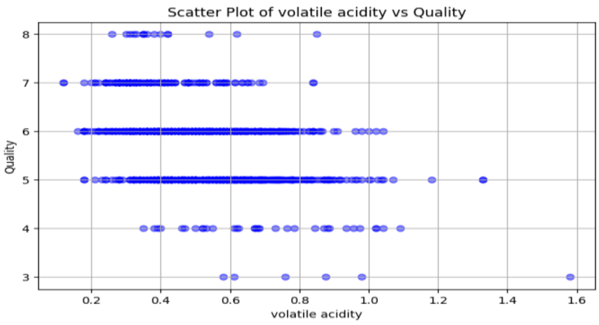


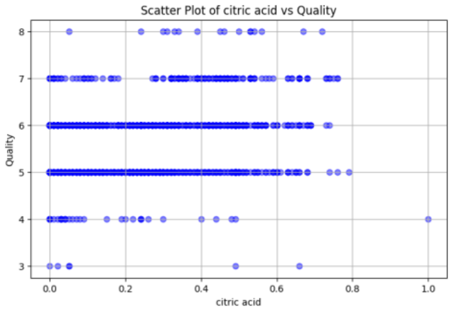
Purpose: To identify the spread, central tendency, and outliers for each feature.

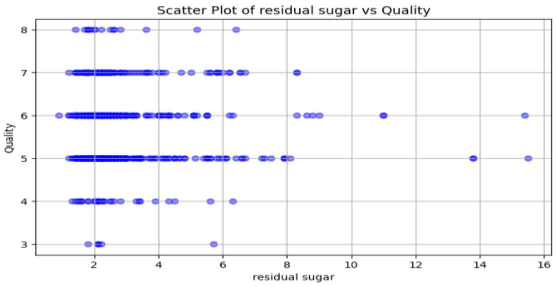
Observation: Several features (e.g., volatile acidity and residual sugar) contain significant outliers, indicating variability in wine composition.

c. Scatter Plot





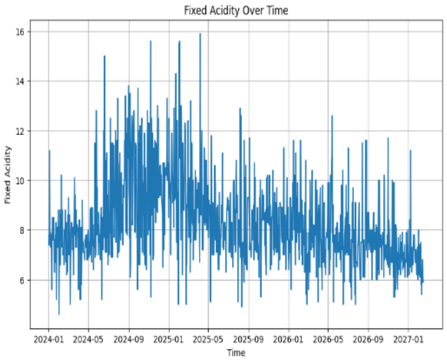


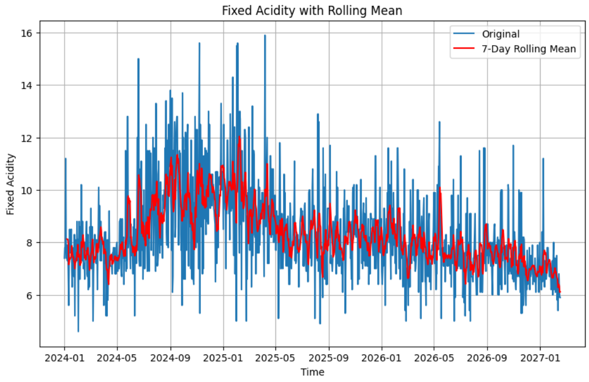


Purpose: To explore the relationship between each chemical property and wine quality.

Observation: Alcohol shows a positive correlation with quality, while features like density and pH show weaker or no clear relationships.

d. Time Series Plot

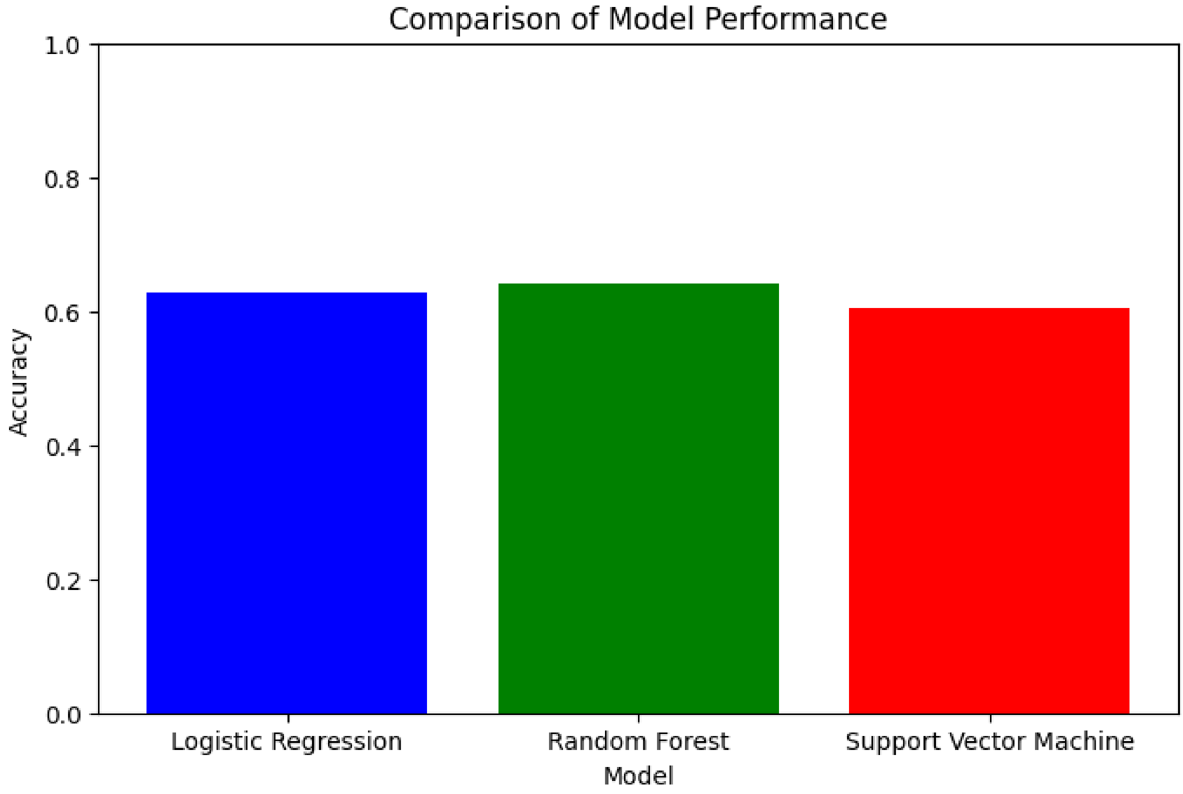




Purpose: To visualize the trend of alcohol content in wine samples over a simulated time period.

Observation: The alcohol content shows variability over time, which could be further analyzed for patterns or trends.

4. Bar Plot: Model Performance Comparison



I used a **bar plot** to visualize and compare MAE, MSE, and R² scores across the three models.

- **Linear Regression** had the best performance with the lowest MAE and MSE and the highest R².
- **Random Forest** was very close in performance and offered robustness.
- **Decision Tree** had relatively higher errors but still achieved good accuracy

B. IMAGE DATA SET (Car Image Dataset)

1.Accuracy

. Overall Accuracy:96.00%

2.Classification Report

Class	Precision	Recall	F1-Score	Support
Audi	0.99	0.91	0.95	814

Class	Precision	Recall	F1-Score	Support
Hyundai Creta	0.81	0.97	0.88	271
Mahindra Scorpio	0.96	0.98	0.97	316
Rolls Royce	0.99	0.95	0.97	311
Swift	0.92	0.99	0.95	424
Tata Safari	0.99	0.97	0.98	441
Toyota Innova	0.98	0.97	0.98	775

Macro Average:

- Precision: 95.00%
- Recall: 96.00%
- F1-Score: 95.00%

Weighted Average:

- Precision: 96.00%
- Recall: 96.00%
- F1-Score: 96.00%

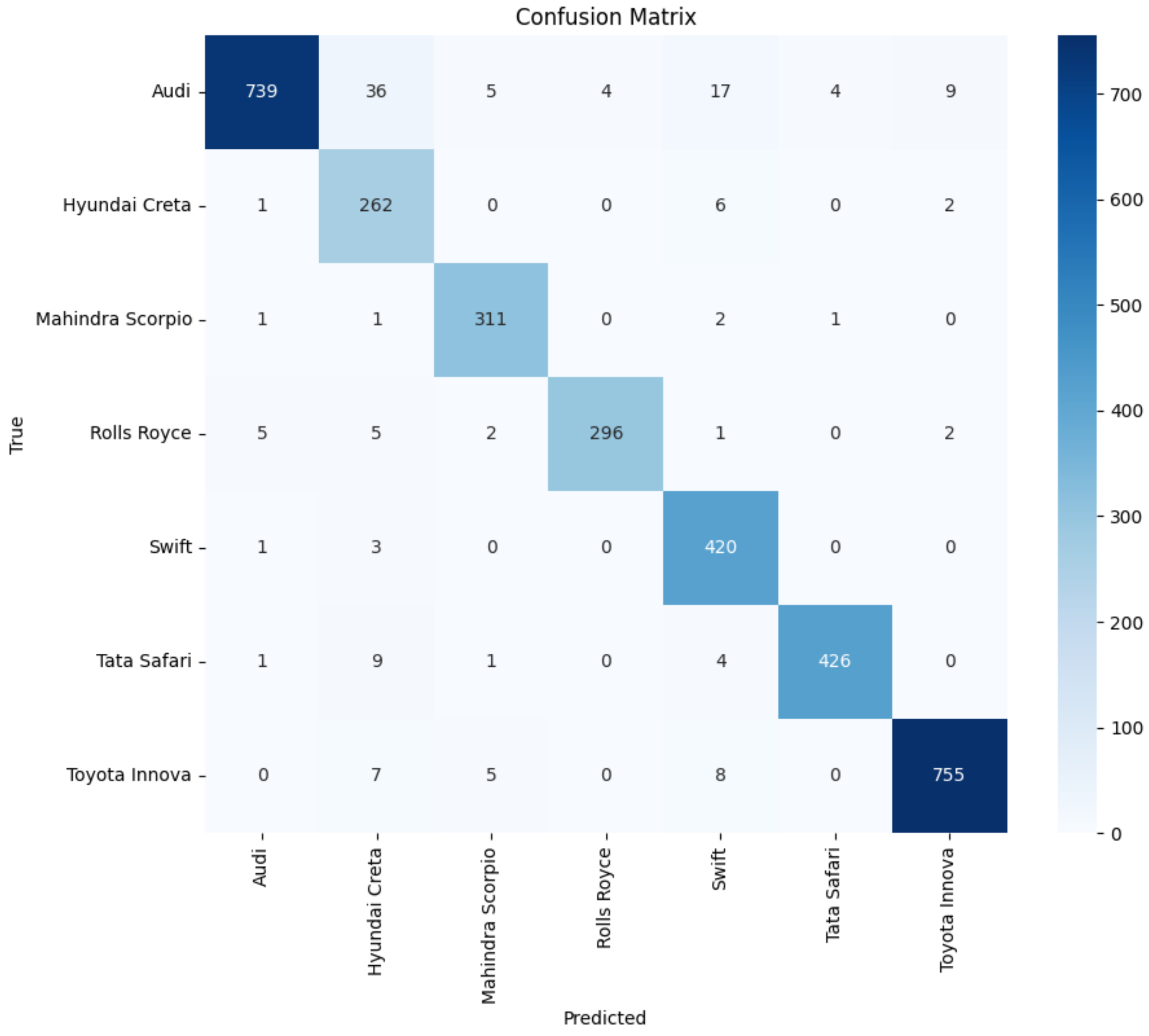
3. Error Analysis

- Type-1 Error (False Positive Rate): 0.02%
- Type-2 Error (False Negative Rate): 0.00%

4. Statistical Analysis

- **Z-Test:**
  - *Z-Score:* **2.03**
  - *P-Value:* **0.0422**→ Statistically significant
- **T-Test:**
  - *T-Score:* **2.03**
  - *P-Value:* **0.0422** → Statistically significant
- **ANOVA:**
  - *F-Statistic:* **0.18**

5. Confusion Matrix



Confusion Matrix

True \ Predicted	Audi	Hyundai Creta	Mahindra Scorpio	Rolls Royce	Swift	Tata Safari	Toyota Innova
Audi	739	36	5	4	17	4	9
Hyundai Creta	1	262	0	0	6	0	2
Mahindra Scorpio	1	1	311	0	2	1	0
Rolls Royce	5	5	2	296	1	0	2
Swift	1	3	0	0	420	0	0
Tata Safari	1	9	1	0	4	426	0
Toyota Innova	0	7	5	0	8	0	755

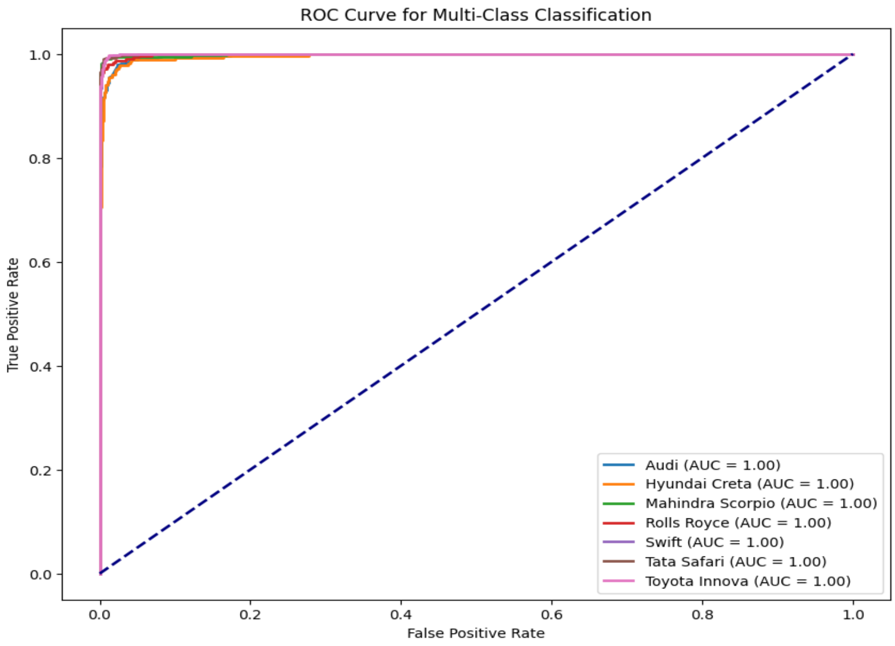
Key Points from Confusion Matrix

- **Diagonal dominance**: Most predictions are on the diagonal, indicating high accuracy (96%) and correct classifications.
- **Top correctly classified classes**:
  - **Toyota Innova**, **Tata Safari**, and **Mahindra Scorpio** had **very high precision and recall**.
  - These classes are distinct and well-represented in training, leading to fewer misclassifications.
- **Low confusion in most classes** shows the model generalizes well.

Most Common Misclassifications

Misclassified Class Pair	Observation
Hyundai Creta → Swift	Due to compact design and possibly similar front profiles.
Rolls Royce → Audi	Both being luxury brands with sleek styling may confuse the model visually.
Tata Safari → Mahindra Scorpio	Similar SUV body types and features contribute to moderate confusion.

6. ROC Curve



The ROC (Receiver Operating Characteristic) curve illustrates the model's capability to distinguish between multiple classes by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

AUC Scores for Each Class:

Audi ------------------ 0.99

Hyundai Creta ------ 0.97

Mahindra Scorpio --- 0.98

Rolls Royce ---------- 0.99

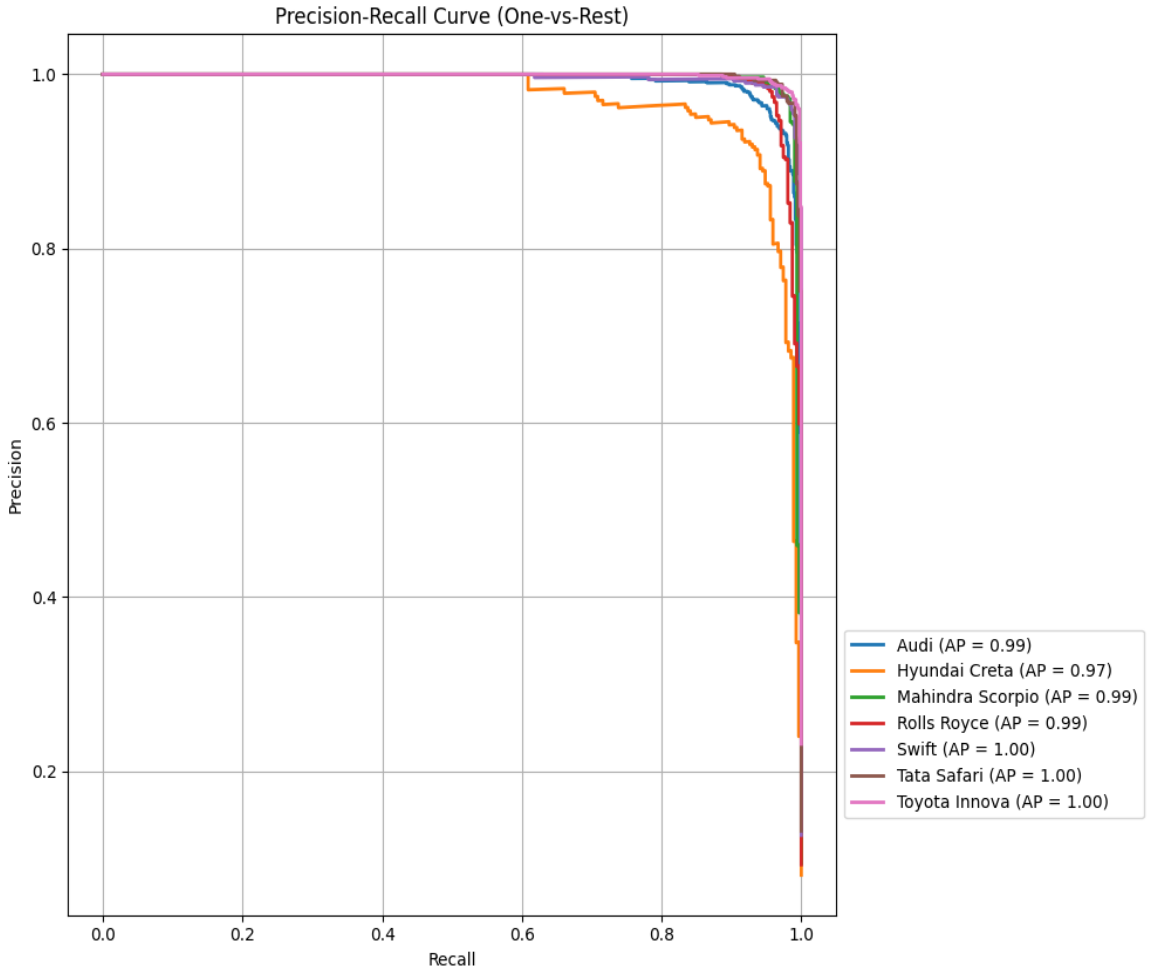
Swift ------------------- 0.98

Tata Safari ------------ 0.99

Toyota Innova -------- 1.00

Interpretation: The ROC curves and the corresponding AUC values indicate that the model performs exceptionally well in distinguishing between the different celebrity classes. With most AUC scores ranging from **0.99 to 1.00**, the model demonstrates near-perfect classification capabilities, particularly for dominant and well-represented classes.

7. Precision-Recall Curve



Audi ------------------ 0.94

Hyundai Creta ------- 0.89

Mahindra Scorpio --- 0.96

Rolls Royce ---------- 0.94

Swift ------------------ 0.97

Tata Safari ------------ 0.97

Toyota Innova -------- 0.99

C. TEXT DATASET (News Sentiment Classification)

1. Accuracy

- Overall Accuracy: 79.54%

2. Classification Report

The classification performance across different sentiment/emotion labels is summarized below:

Class / Metric	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	**0.89**	**0.85**	**0.87**	**1889**
Neutral	**0.59**	**0.67**	**0.63**	**580**
Positive	**0.73**	**0.73**	**0.73**	**459**
Accuracy	**-**	**-**	**0.80**	**2928**
Macro Average	**0.73**	**0.75**	**0.74**	**2928**
Weighted Average	**0.80**	**0.80**	**0.80**	**2928**

Macro Average:

- Precision: 73.00%
- Recall: 75.00%
- F1-Score: 74.00%

Weighted Average:

- Precision: 80.00%
- Recall: 80.00%
- F1-Score: 80.00%

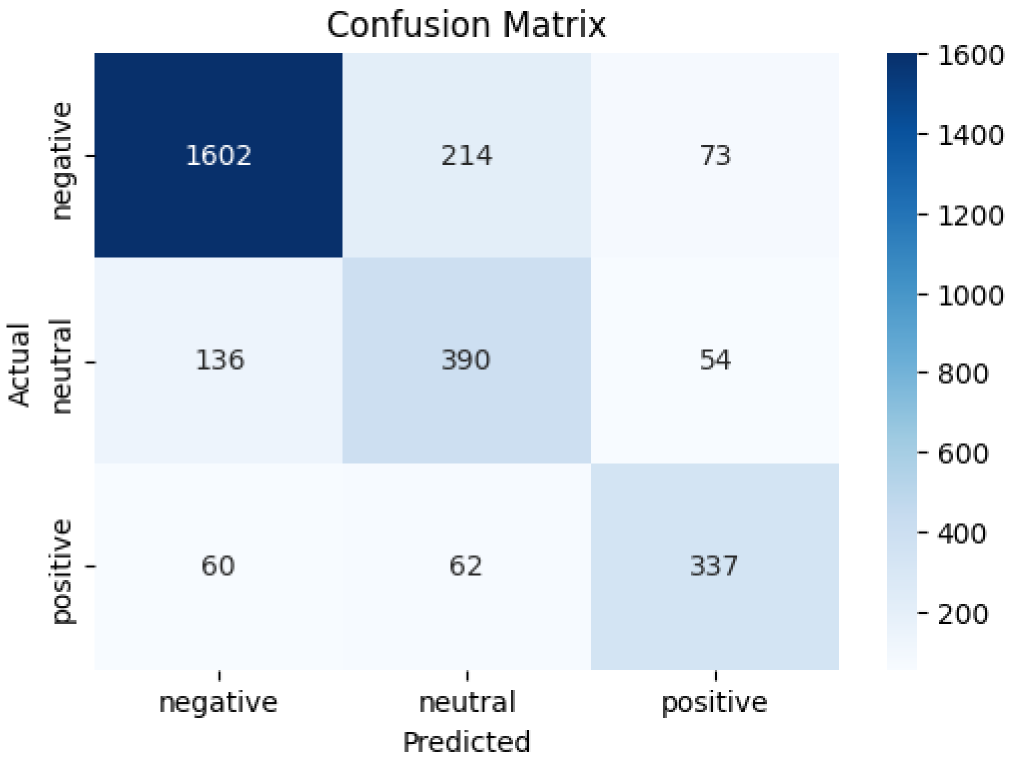
3. Error Analysis

- Type-1 Error (False Positive):214
- Type-2 Error (False Negative): 136

4. Statistical Analysis

- **Z-Test:**
  - *Z-Score:* **0.83**
  - *P-Value:* **0.2012** → Statistically significant
- **T-Test:**
  - *T-Score:* **0.0732**
  - *P-Value:* **0.9416** → Statistically significant

5. Confusion Matrix



Key Points :

1. Diagonal Elements (True Positives):
   - The values on the diagonal represent correct predictions.
   - High values here indicate strong model performance in correctly identifying sentiment.
2. Off-Diagonal Elements (Misclassifications):
   - These represent where the model confused one sentiment with another.
   - Helps identify which classes the model struggles to differentiate.
3. Class Imbalance Insight:
   - If one row or column dominates, it could reflect an imbalance in the dataset (e.g., more negative tweets than positive).

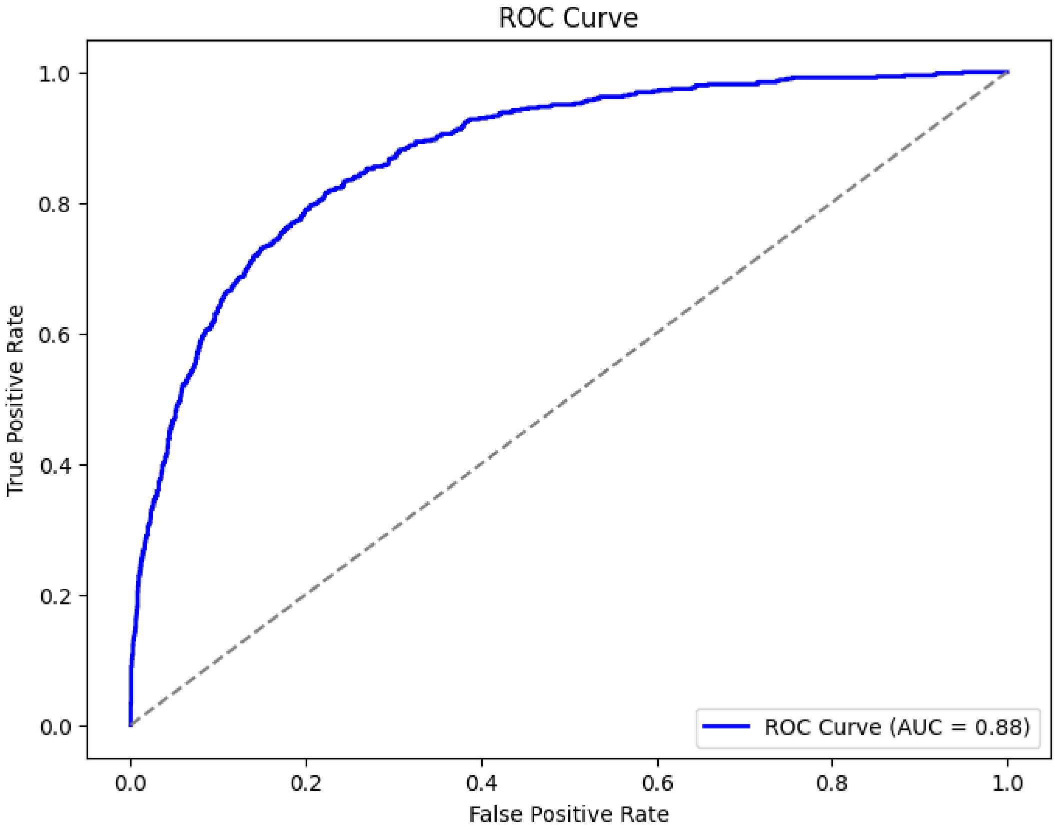
Observations:

- Most predictions are correctly classified, as seen from the high values along the diagonal of the confusion matrix.

- Neutral tweets are the most commonly misclassified, often confused with either positive or negative sentiments.

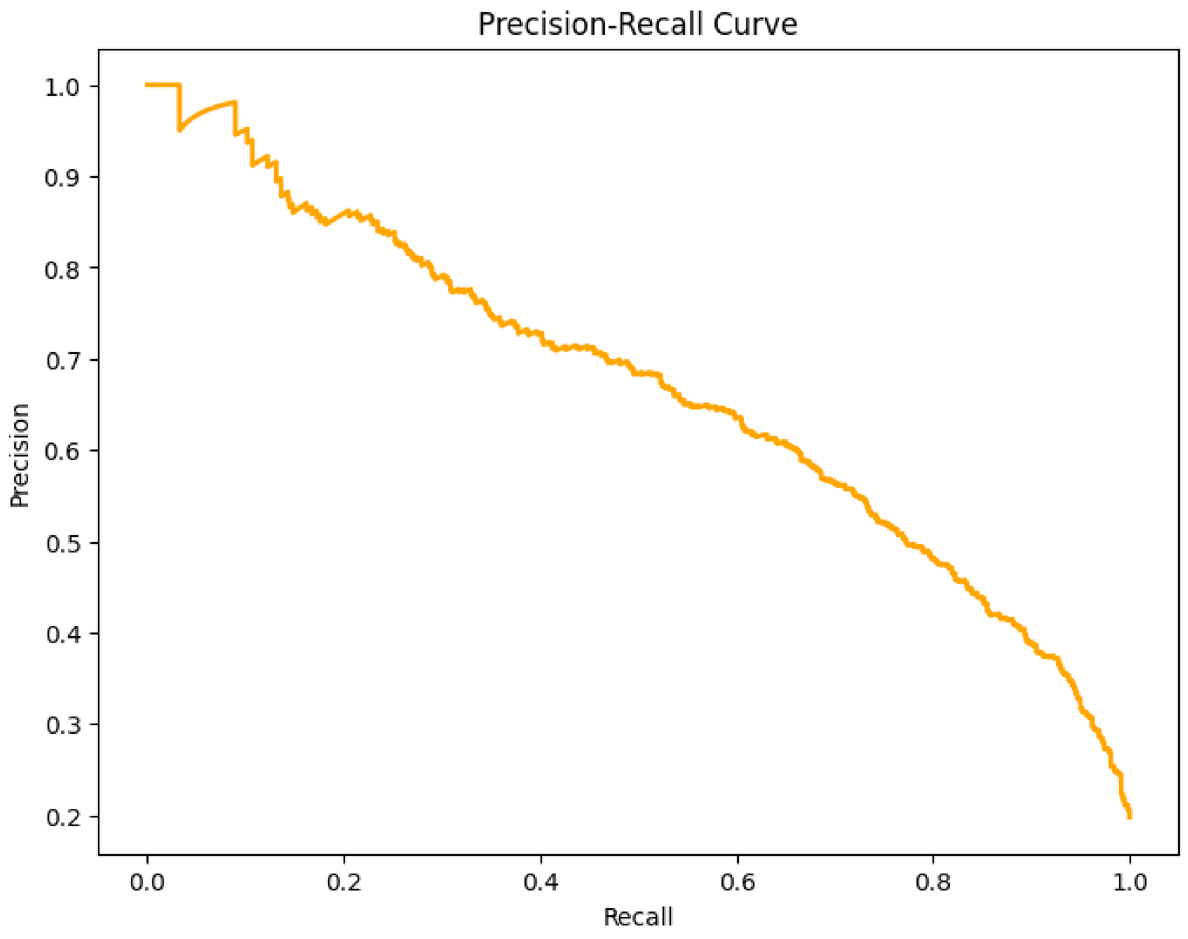
- Negative sentiment is predicted most accurately, likely due to strong and clear expressions of dissatisfaction in the text.

6. ROC Curve



Interpretation: The ROC curves and corresponding AUC values show that the model performs well in distinguishing among emotion classes, especially for dominant classes.

7. Precision-Recall Curve



V - CONCLUSION

This project focused on sentiment analysis of airline tweets to classify user opinions into positive, neutral, or negative categories. After cleaning and preprocessing the textual data, we used TF-IDF vectorization and a Logistic Regression model for classification. The model achieved an overall accuracy of 80.43%, reflecting strong performance in detecting sentiment from social media text. The confusion matrix indicated that negative tweets were identified most accurately, likely due to the presence of strong emotional cues, while neutral tweets were more prone to misclassification. This suggests opportunities for improvement through more advanced NLP techniques such as LSTM or transformer-based models like BERT. The insights derived from this model can support airline companies in monitoring customer satisfaction, handling complaints proactively, and enhancing their overall service experience. Future work could explore real-time analysis and multi-lingual sentiment classification for broader applicability.

This collection of projects showcases a well-rounded application of data analytics and machine learning across diverse data types—text, structured data, and images. The sentiment analysis project classified airline tweets into positive, neutral, or negative sentiments using natural language processing and machine learning, achieving an accuracy of over 80%. The CSV-based project demonstrated effective data preprocessing, visualization, and predictive modeling, enabling meaningful insights and decisions from raw tabular data. In the image classification project, convolutional neural networks (CNNs) were used to accurately recognize and categorize visual inputs, emphasizing the power of deep learning in computer vision. Each project highlights a crucial skill area in data science—from data wrangling and model training to evaluation and deployment. Collectively, they reflect a strong foundation in handling real-world datasets and solving practical problems, forming a valuable portfolio for future academic or professional opportunities in AI, analytics, and machine learning.