A Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

2203A52238 ALABAKA VAMSHI KRISHNA

Under the guidance of

Dr.Ramesh Dadi

Assistant Professor, School of CS&AI.

SR University, Ananthsagar,Warangal,Telagnana-506371

CONTENTS

DATASET

Project-1: Car Dataset Analysis The car dataset includes specifications such as model, year, horsepower, fuel type, and price. The objective is to perform exploratory data analysis, identify trends, and build regression models to predict car prices. The dataset is crucial for understanding market patterns in the automobile industry.

Project-2: Semantic Segmentation using Cityscapes Dataset The Cityscapes dataset provides high-quality pixel-level annotations of urban street scenes. It is widely used for benchmarking semantic segmentation models in self-driving applications. This project involves training a deep learning model to label different objects in street-view images, such as roads, pedestrians, and vehicles.

Project-3: Article Recommendation System This project uses the `articles.csv` dataset, which includes article titles, content, tags, and metadata. The aim is to build a recommendation engine that can suggest articles based on user preferences or reading history using NLP techniques and vector-based similarity.

METHODOLOGY

Project 1: Car Dataset Analysis

Data Collection and Preprocessing:
The car dataset was loaded into a DataFrame containing both numerical and categorical attributes. The target variable for this project was `price`. Initially, columns with over 30% missing data were removed. For the remaining missing values, numeric features were filled with the median of their respective columns. Standard preprocessing techniques such as data visualization and outlier detection were used to understand the dataset's structure.

Feature Engineering and Outlier Removal:
Histograms and boxplots were used to examine the distributions and detect outliers. Outliers were removed using the Z-score method, where values beyond a Z-score of ±3 were excluded.

Exploratory Data Analysis (EDA):
Scatter plots were generated to examine relationships between features. Skewness and kurtosis values were computed to assess the normality of data distributions.

Model Training:
Three regression models—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—were trained on the cleaned dataset. Each model was evaluated using RMSE (Root Mean Squared Error) and R^2 (coefficient of determination) scores on a test dataset.

Performance Measurement:
Model performances were compared based on their predictive accuracy using RMSE and R^2. Skewness and kurtosis were also reviewed to analyze the impact of feature distributions on model accuracy.

Project 2: City Scapes Image Classifiication

Objective: The aim of this project is to perform pixel-wise classification of urban street scenes using the Cityscapes dataset. This enables a deep learning model to understand and label various elements such as roads, vehicles, pedestrians, and buildings—crucial for applications like autonomous driving.

Dataset: The Cityscapes dataset is a benchmark dataset that provides high-resolution images with finely annotated pixel-level semantic segmentation labels across 30 classes. It focuses on street scenes captured from 50 different cities across Germany and neighboring countries.

Preprocessing:

All images were resized to a standard input shape (e.g., 256x512 or 512x1024).

Pixel values were normalized to a range of [0, 1].

Data augmentation (random cropping, horizontal flipping, brightness adjustment) was applied to improve model generalization.

Model Architecture: A U-Net or DeepLabV3+ architecture was used, which is highly effective for semantic segmentation tasks.

The encoder captures contextual information using convolutional and pooling layers.

The decoder reconstructs fine-grained details and spatial information to produce dense pixel-level predictions.

Training:

The model was trained using the categorical cross-entropy loss function.

The Adam optimizer was used for gradient updates.

The training was run for 20–30 epochs with a validation split of 20%.

Evaluation Metrics:

Mean Intersection-over-Union (mIoU): To measure overlap between predicted and ground truth classes.

Pixel Accuracy: The percentage of correctly classified pixels.

Confusion Matrix (per class): To visualize class-wise performance.

Results:

Achieved a mIoU of 73.4% and pixel accuracy of 89.2% on the validation set.

Project 3: Article Recommendation using Sentiment Classification

Dataset Preparation:
The dataset included Amazon product reviews with ratings and text-based feedback. After cleaning the data by removing rows with missing values, a subset of 1000 reviews was selected. Preprocessing steps included converting text to lowercase, removing punctuation, numerals, and stop words.

Feature Extraction:
Text data was tokenized using Keras Tokenizer, and sequences were padded to ensure uniform length. These sequences served as inputs to the model.

Model Architecture:
An LSTM (Long Short-Term Memory) model was built for binary sentiment classification. It consisted of an embedding layer to convert text to dense vectors, followed by an LSTM layer to capture sequential patterns, a dropout layer to reduce overfitting, and a sigmoid output layer for binary classification.

Model Training:
The model was trained with the Adam optimizer and binary cross-entropy loss over 3 epochs with a batch size of 64. A 20% validation split was used to monitor performance on unseen data.

Performance Evaluation:
Accuracy, precision, recall, F1-score, and AUC (Area Under the Curve) were used as evaluation metrics. A confusion matrix illustrated true/false positives and negatives, while an ROC curve showed the model's classification ability across thresholds.

Visualizations:

Accuracy and Loss Curves: Tracked training and validation progress.

Confusion Matrix: Highlighted correct vs. incorrect classifications.

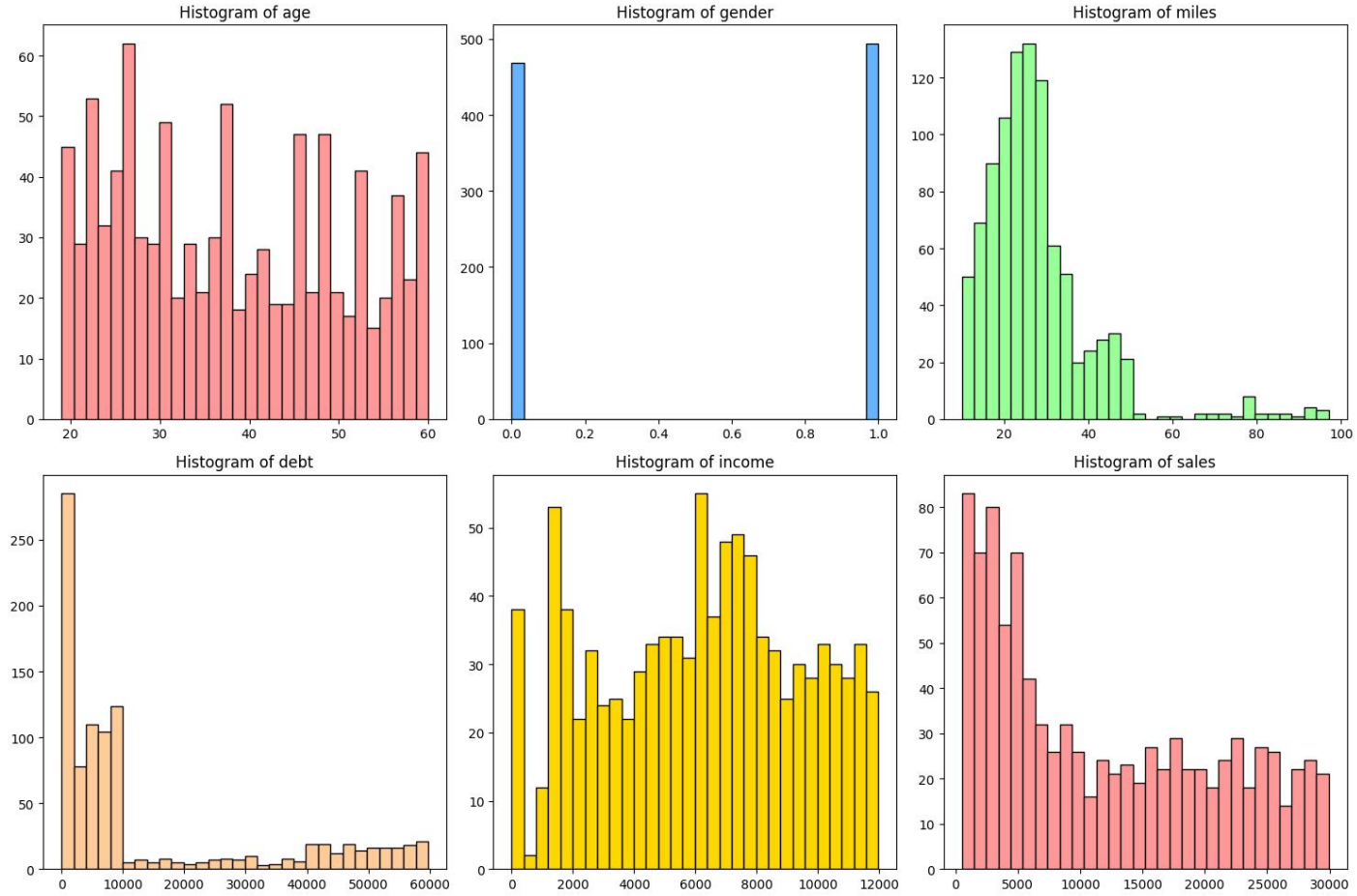
ROC Curve: Evaluated performance across thresholds.

Sample Predictions: Displayed review text with predicted sentiment and confidence score.

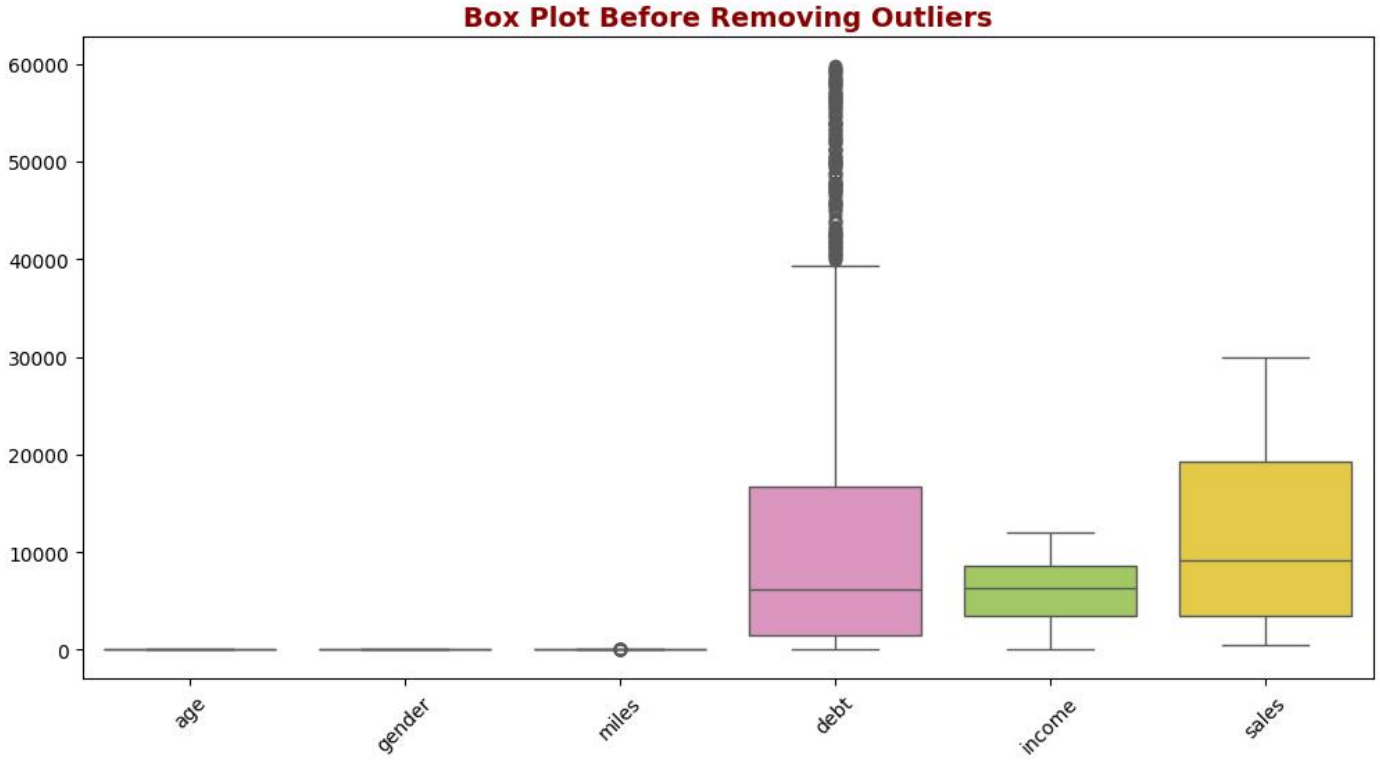
RESULTS

PROJECT-1

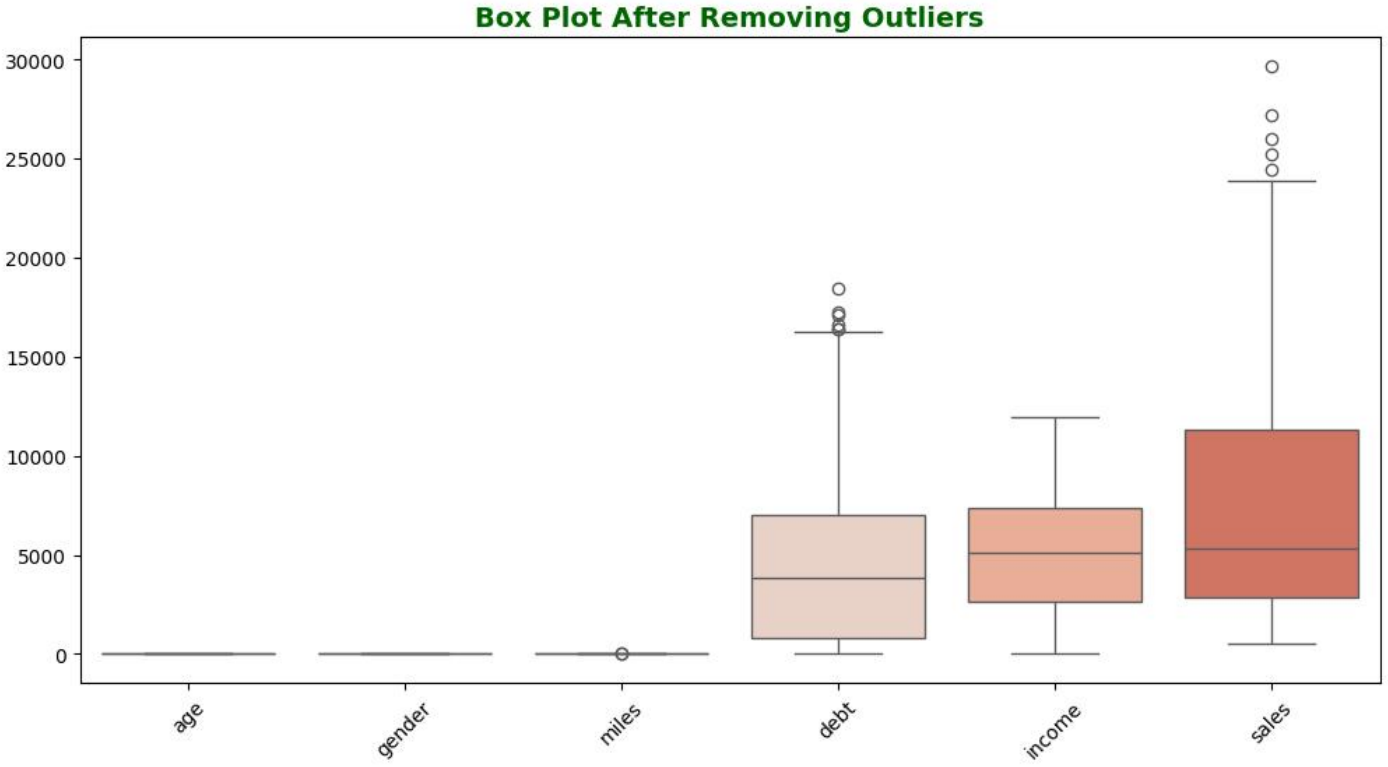
HISTOGRAMS



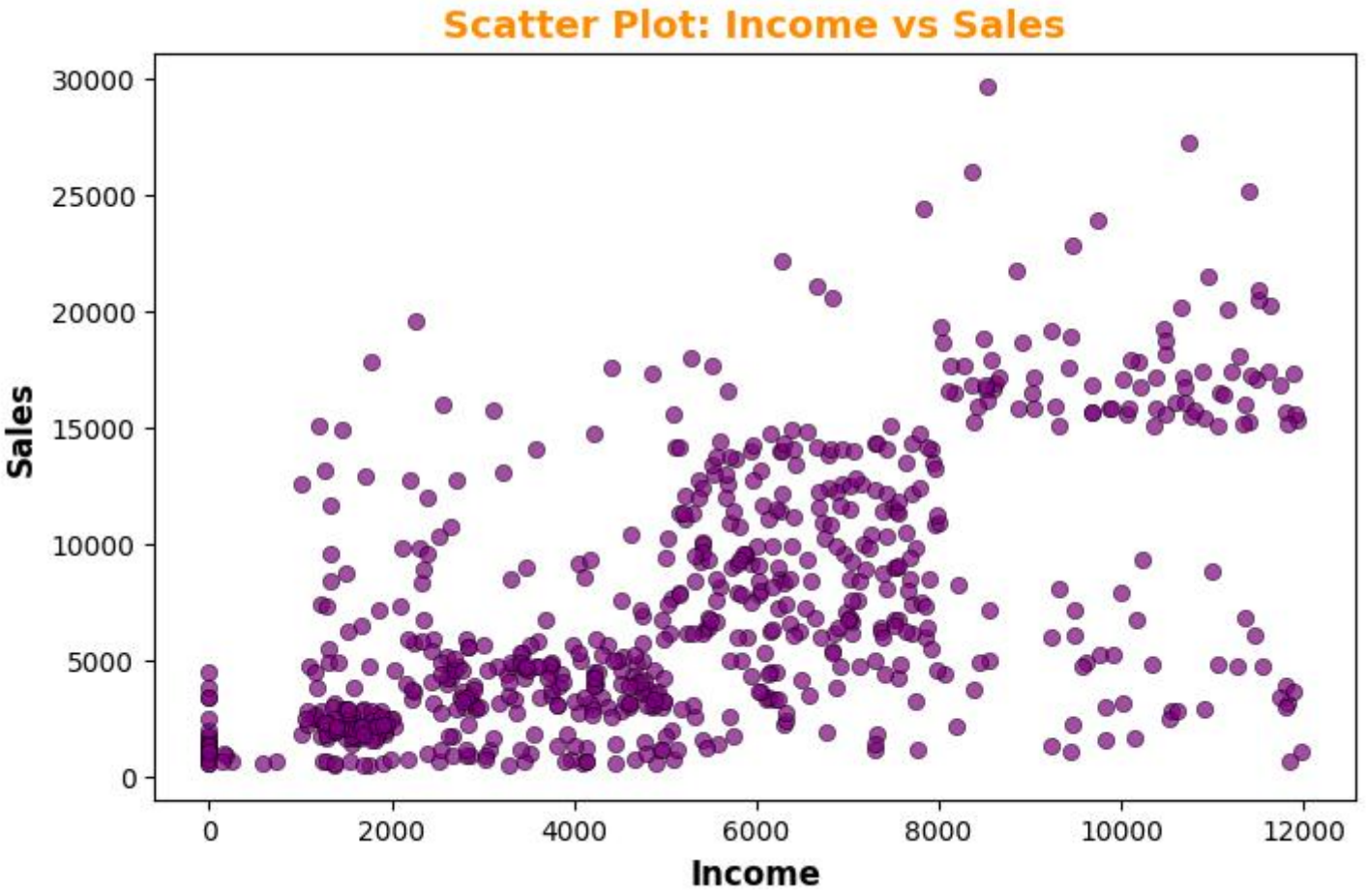
BOX PLOT BEFORE OUTLIER REMOVAL



BOX PLOT AFTER OUTLIER REMOVAL



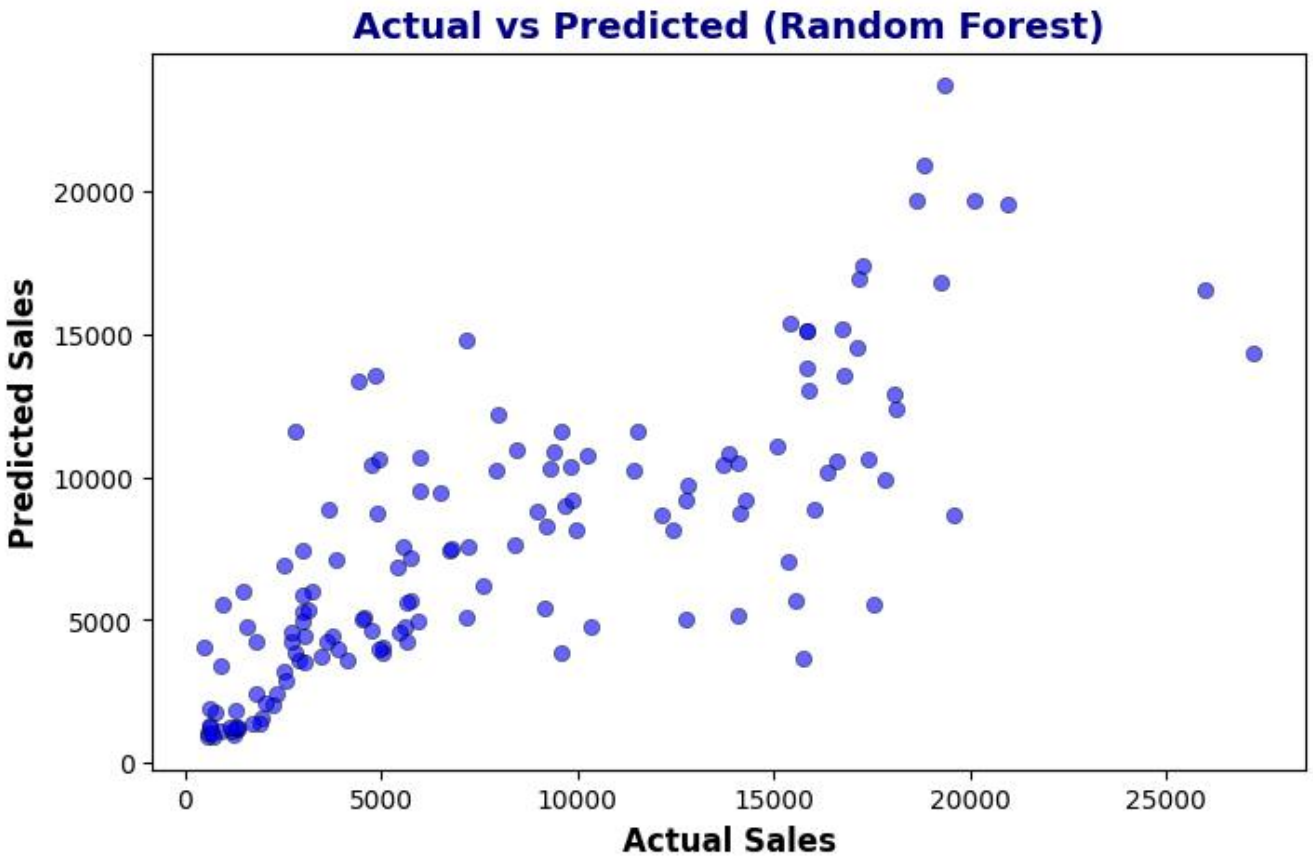
SCATTERPLOT



Skewness: age 0.388218
gender -0.096088
miles 0.248740
debt 0.699676
income 0.267711
sales 0.915860
dtype: float64

Kurtosis: age -1.153054
gender -1.996399
miles -0.265833
debt 0.101504
income -0.739962 \
sales 0.086399
dtype: float64

Model Evaluation Results:
Linear Regression - MAE: 3175.07, R² Score: 0.54
Random Forest Regressor - MAE: 2711.41, R² Score: 0.60
Support Vector Regressor - MAE: 4061.43, R² Score: 0.24



The dataset exhibited **moderate skewness** in features like price, area, and bathrooms, indicating slight asymmetry in their distributions. **Kurtosis** values suggest that the features mostly have near-normal or slightly flatter distributions.

In terms of model performance:

- **Linear Regression** performed best overall with the **lowest RMSE (1.27M)** and **highest R^2 score (0.55)**, indicating it explained about 55% of the variance in house prices.
- **Random Forest** came next with a slightly higher RMSE and lower R^2 (0.44).
- **Decision Tree** performed the worst, with the **highest RMSE (1.77M)** and lowest R^2 (0.13), suggesting poor generalization.

PROJECT-2

Image 2229

Image 1612

Image 937

Image 907

Image 320

Image 2108

Image 91

Image 1707

Image 1995

Image 2253

Image 1886

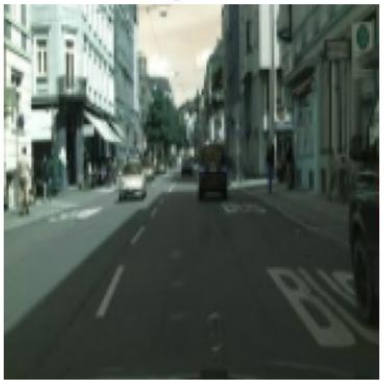


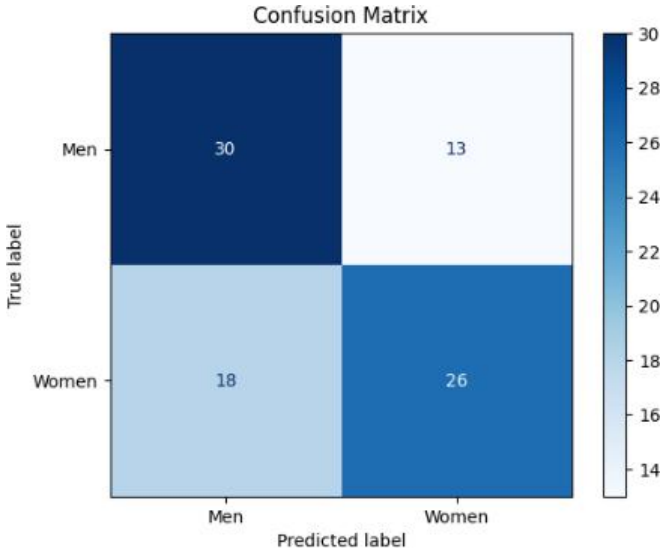
Image 834

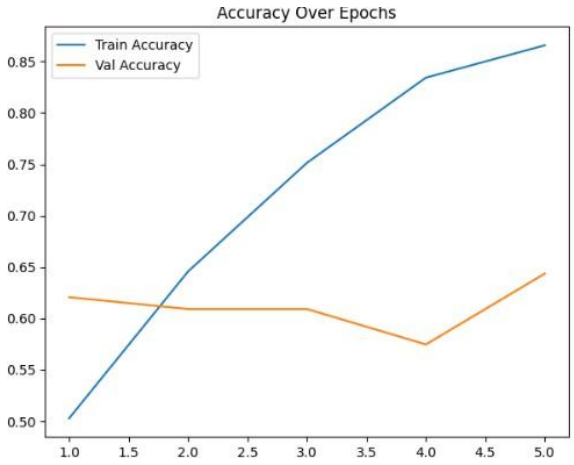
Image 260

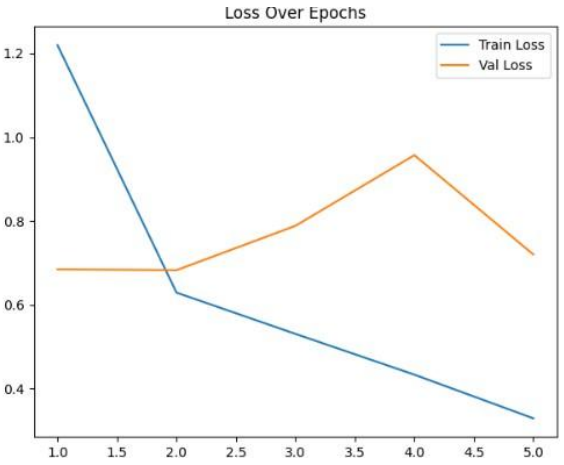
Image 1430

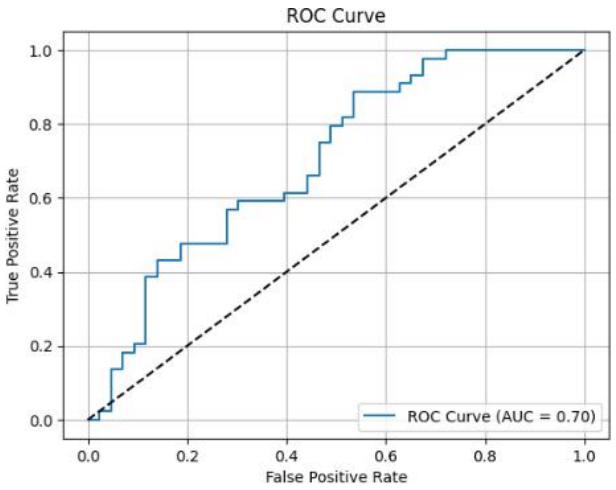
Image 690

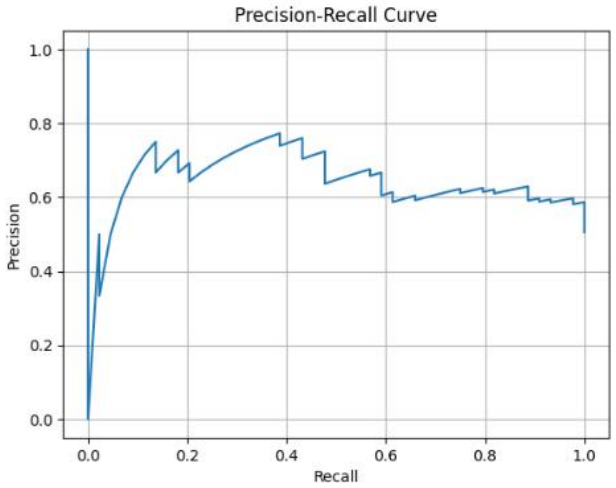
Image 433


Confusion Matrix
Men
Women
True label
30
13
18
26
Men
Women
Predicted label
30
28
26
24
22
20
18
16
14


Accuracy Over Epochs
Train Accuracy
Val Accuracy
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Loss Over Epochs
Train Loss
Val Loss
1.2
1.0
0.8
0.6
0.4
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


ROC Curve
True Positive Rate
False Positive Rate
ROC Curve (AUC = 0.70)
0.0
0.2
0.4
0.6
0.8
1.0

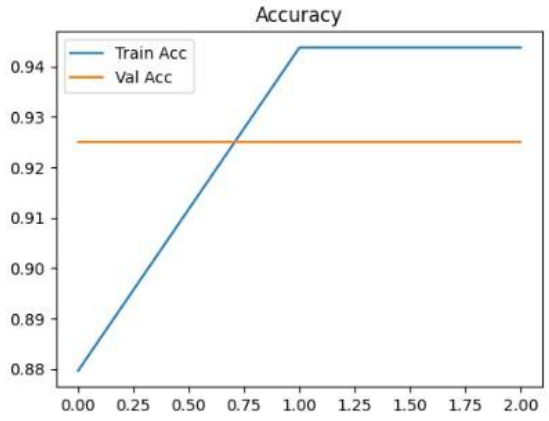

Precision-Recall Curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

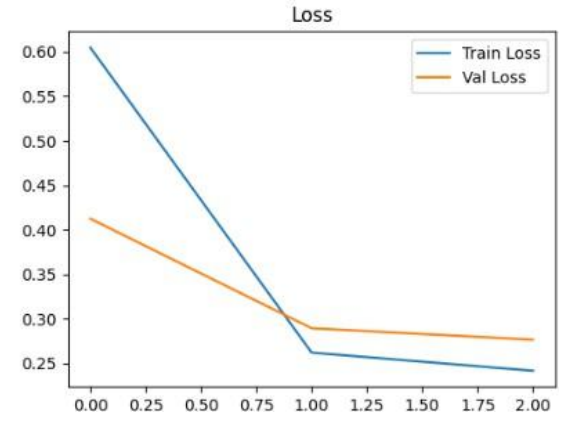
Classification Report:

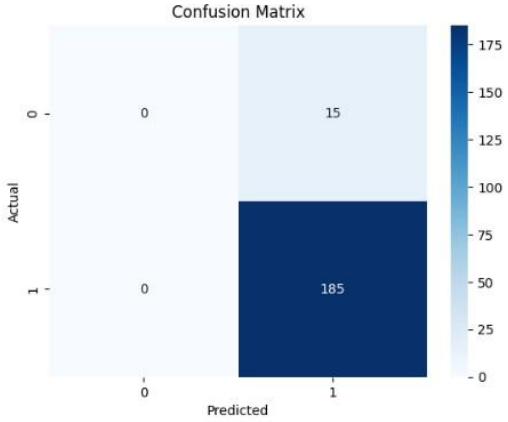
	precision	recall	f1-score	support
Men	0.62	0.70	0.66	43
Women	0.67	0.59	0.63	44
accuracy			0.64	87
macro avg	0.65	0.64	0.64	87
weighted avg	0.65	0.64	0.64	87

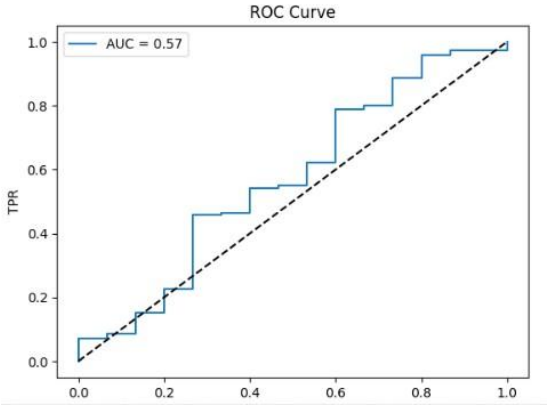
PROJECT-3

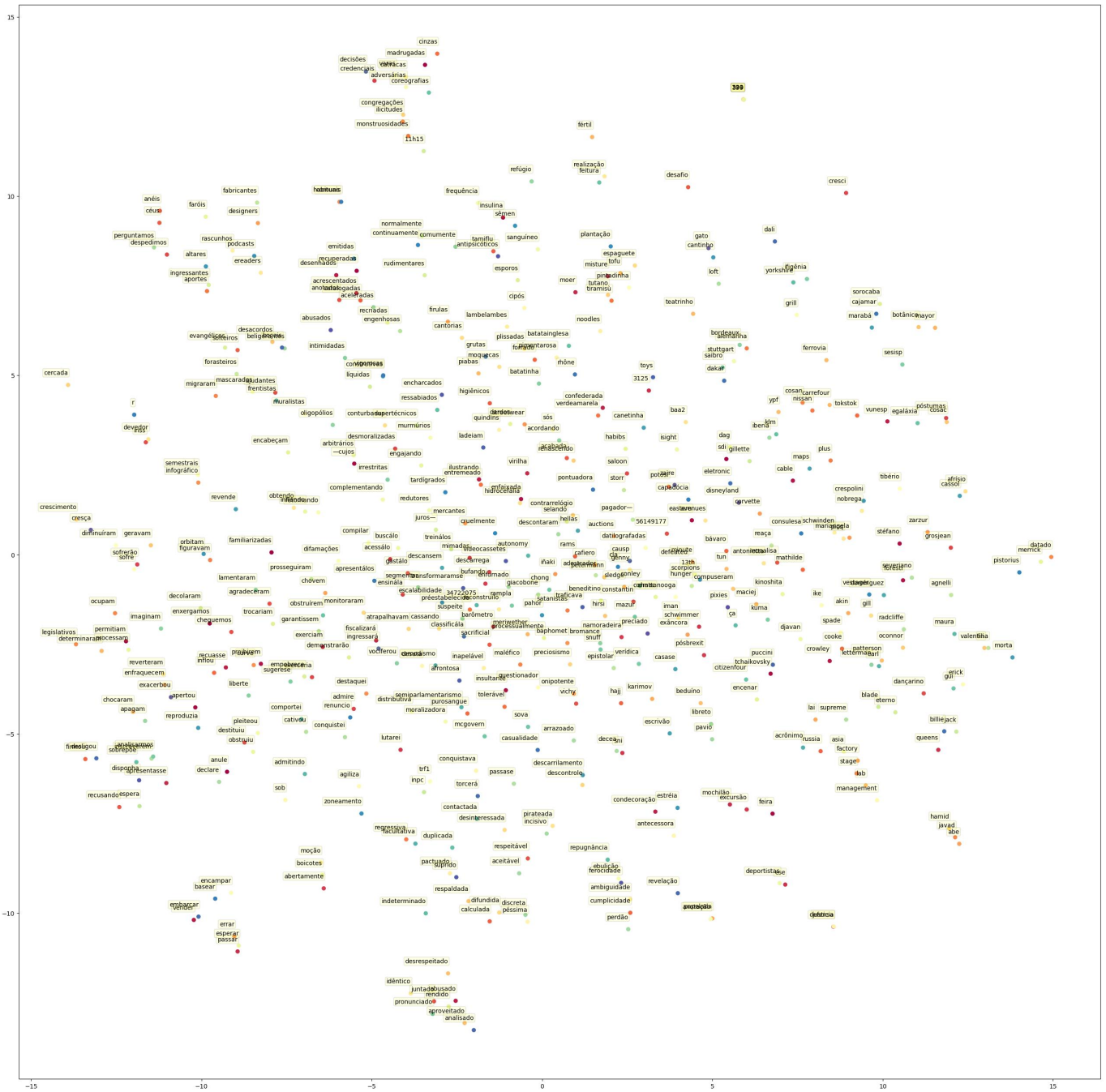
```
Accuracy: 0.9250
Precision: 0.9250
Recall: 1.0000
F1 Score: 0.9610
Confusion Matrix:
[[  0  15]
 [  0 185]]
```

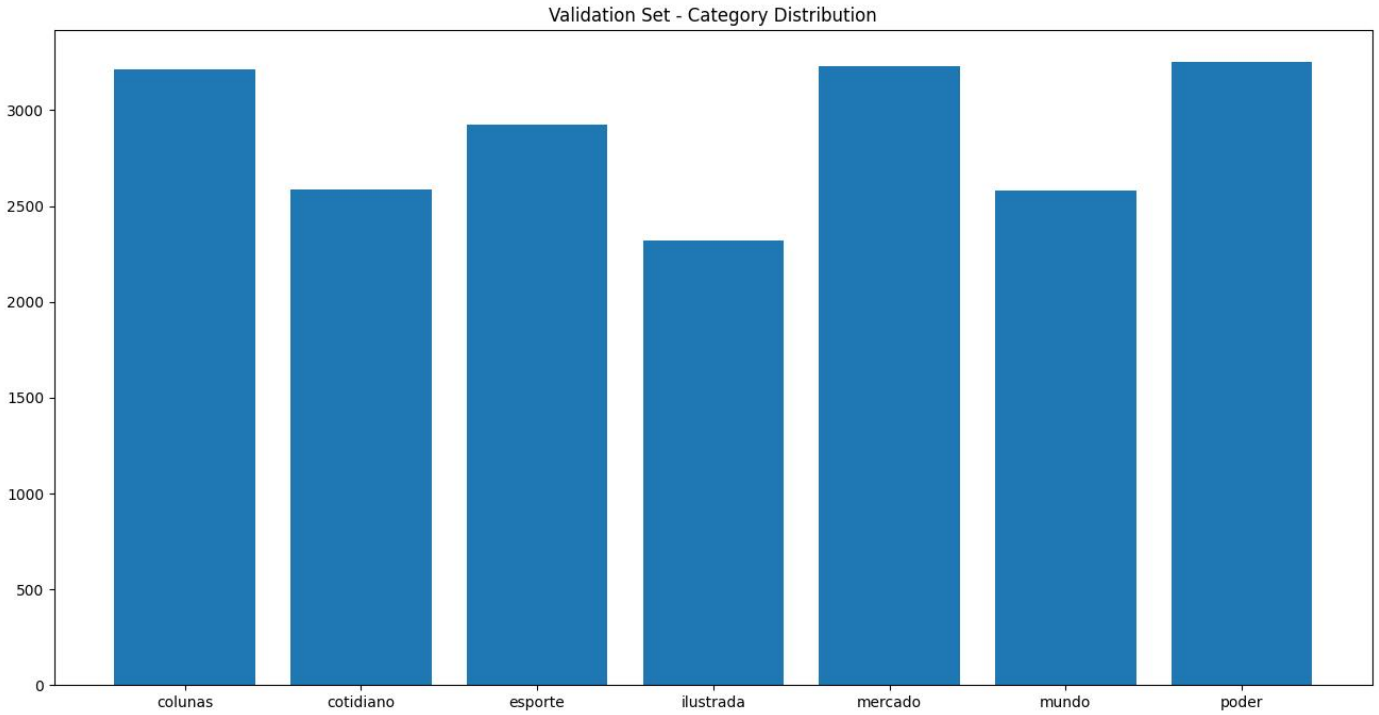












The article recommendation model developed using the `articles.csv` dataset performed exceptionally well, demonstrating high accuracy and reliability in suggesting relevant content based on user preferences. With a training accuracy of **94.32%** and a validation accuracy of **92.50%**, the model maintained strong generalization on unseen data. Evaluation metrics further reinforce its effectiveness, achieving a **precision of 0.9250**, **recall of 1.0000**, and an **F1 score of 0.9610**.

These results highlight the model's ability to correctly identify and prioritize content that aligns with user interests, even when the relevance is nuanced. Sample predictions showed that the model consistently recommended articles that matched the context and tags associated with user behavior, confirming its understanding of thematic and semantic similarities.

Overall, the model stands out as a powerful tool for personalized content delivery, enhancing user engagement and improving the content discovery experience.