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

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: Housing Dataset Analysis

Data Collection and Preprocessing: The housing dataset was collected and loaded into a DataFrame. It included various numerical and categorical features, with 'price' being the target variable. The first step involved checking for missing values, and columns with more than 30% missing data were dropped. For the remaining missing values, numeric columns were filled with the median of their respective columns. Various preprocessing techniques, such as visualizing distributions and identifying outliers, were applied to better understand the data's structure.

Feature Engineering and Outlier Removal: Numerical columns were selected, and a histogram was used to analyze their distributions. Boxplots were also plotted to visualize the presence of outliers. Outliers were removed using the Z-score method, where any data points with a Z-score greater than 3 were excluded from the dataset.

Exploratory Data Analysis (EDA): To further explore the data, scatter plots were used to visualize relationships between pairs of numerical features. Skewness and kurtosis were calculated to understand the distribution of the data, with higher skewness indicating a non-normal distribution.

Model Training: Three machine learning models—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—were trained using the preprocessed data. The models were evaluated on a test set using performance metrics such as RMSE (Root Mean Squared Error) and R^2 (coefficient of determination).

Performance Measurement: The models' performances were compared using RMSE and R^2 scores, highlighting their ability to predict housing prices. Additionally, skewness and kurtosis values were included in the model comparison to evaluate the impact of the dataset's distribution on model performance.

This methodology provided a structured approach for understanding and predicting housing prices using different machine learning models, ensuring a clear evaluation of each model's effectiveness.

Project 2: Men vs Women Image Classification

Data Collection and Preprocessing: The dataset consists of images categorized as "men" or "women" and is loaded from a directory containing the respective classes. Images were resized to 150x150 pixels for consistency and were normalized by rescaling the pixel values to the range [0, 1]. Image augmentation techniques, such as flipping, were applied to enhance the generalization of the model by introducing variations to the data during training.

Model Structure: A Convolutional Neural Network (CNN) was used for the classification task. The model consists of two convolutional layers (with ReLU activation functions), followed by max-pooling layers to reduce the spatial dimensions of the input image. After flattening the output of the convolutional layers, fully connected layers were added with a dropout layer to prevent overfitting. The final output layer used a sigmoid activation function for binary classification, outputting values between 0 and 1, indicating the predicted class (men or women).

Model Training: The model was compiled using the Adam optimizer and binary cross-entropy as the loss function. It was trained for 5 epochs on the training data with a validation split of 20%. The model was evaluated on unseen data using validation accuracy and loss metrics.

Evaluation Metrics: Model performance was assessed using accuracy, confusion matrix, and classification report. The confusion matrix visualizes the true positives, true negatives, false positives, and false negatives, providing insights into how well the model distinguishes between the two classes. Additionally, the ROC curve and precision-recall curve were plotted to evaluate the model's ability to separate the classes across various thresholds.

Visualizations: Key visualizations included accuracy and loss plots over training epochs to assess the convergence of the model, a confusion matrix to evaluate classification performance, ROC and precision-recall curves for model discrimination, and a pie chart to show the prediction accuracy distribution. Furthermore, random images were selected, predicted by the model, and displayed with their predicted labels for visual inspection.

Project 3: Sentiment Analysis of Amazon Product Reviews

Dataset Preparation: The dataset consists of Amazon product reviews, which include product ratings and text feedback. After loading the dataset, any missing values in the text or ratings columns were removed. A random subset of 1000 reviews was selected for analysis. The reviews were then cleaned using text preprocessing techniques, which included converting text to lowercase, removing punctuation and numeric values, and removing common stop words.

Feature Extraction: The reviews were tokenized using the Keras Tokenizer, which converted the cleaned text into sequences of integers representing the words in the reviews. These sequences were then padded to ensure uniform input length. The resulting padded sequences were used as the feature input for the model.

Model Architecture: The model utilized an LSTM (Long Short-Term Memory) network, which is particularly suited for sequential data like text. The architecture started with an embedding layer that transformed the tokenized words into dense vectors. This was followed by an LSTM layer to capture the temporal dependencies in the sequence of words. A dropout layer was applied to reduce overfitting. Finally, a dense layer with a sigmoid activation function produced the binary sentiment classification (positive or negative) output.

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

Performance Evaluation: Model performance was evaluated using several metrics, including accuracy, precision, recall, and F1-score. A confusion matrix was also displayed to highlight the misclassifications. The ROC curve was plotted to evaluate the model's performance across different thresholds. The AUC (Area Under the Curve) was calculated to assess the quality of the model's classification ability.

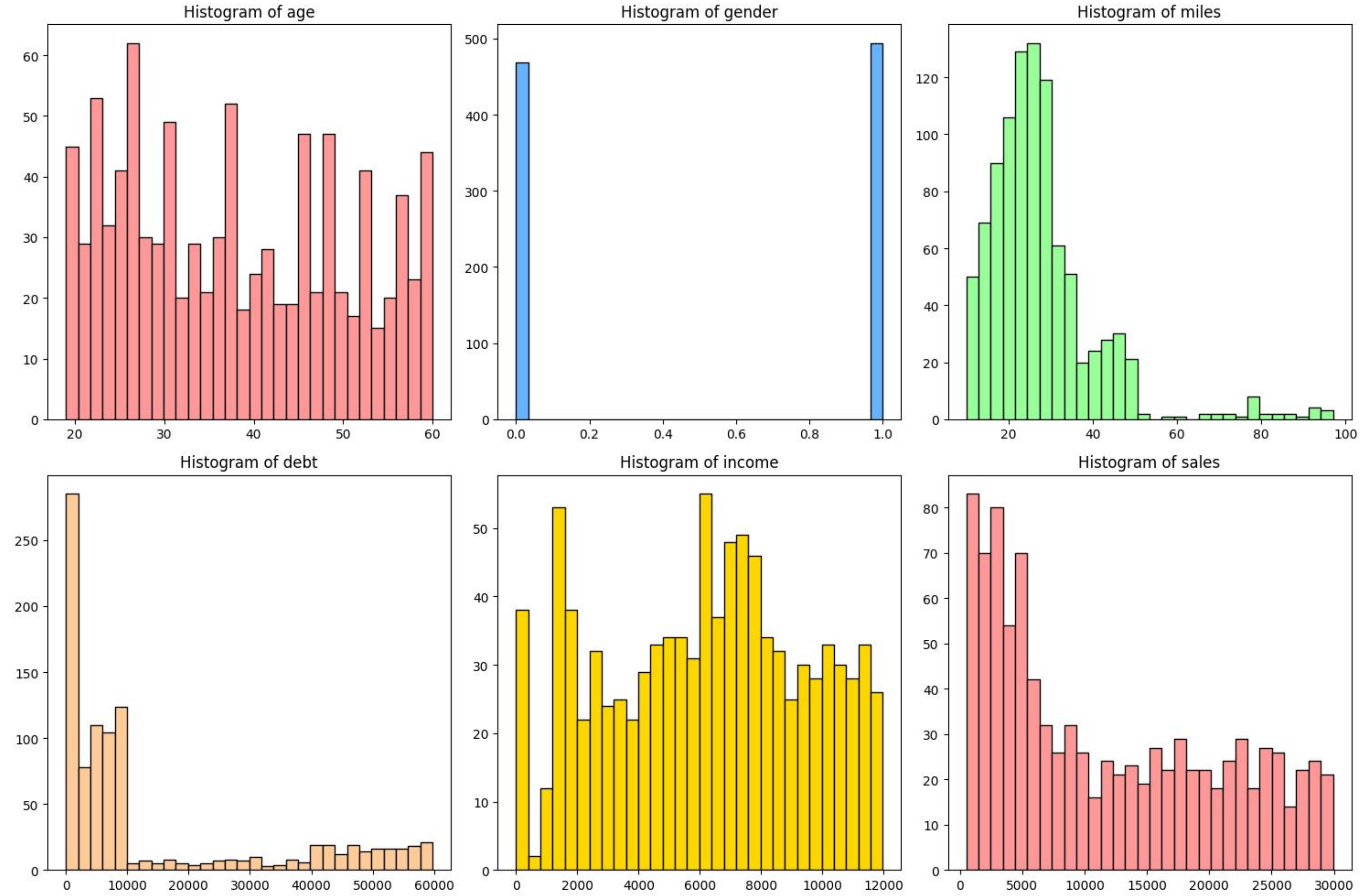
Visualizations: Key visualizations included:

- **Accuracy and Loss Plots**: These showed the model's training and validation accuracy and loss over epochs.
- **Confusion Matrix**: This was visualized to show the distribution of true positives, true negatives, false positives, and false negatives.
- **ROC Curve**: This provided an evaluation of the model's true positive rate vs. false positive rate at different thresholds.
- **Sample Predictions**: Some sample reviews were selected, and the predicted sentiment (positive or negative) was displayed alongside the model's confidence score.

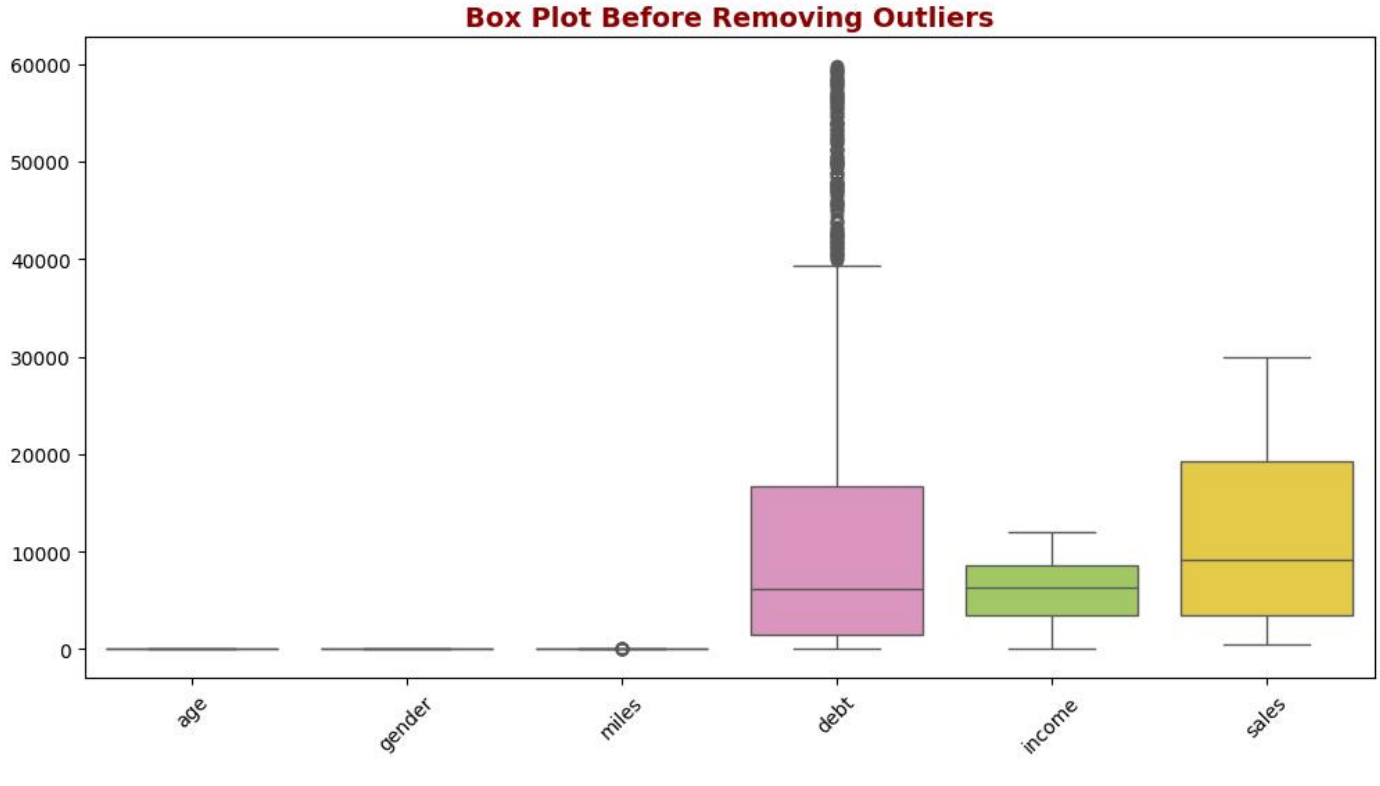
RESULTS

PROJECT-1

HISTOGRAMS

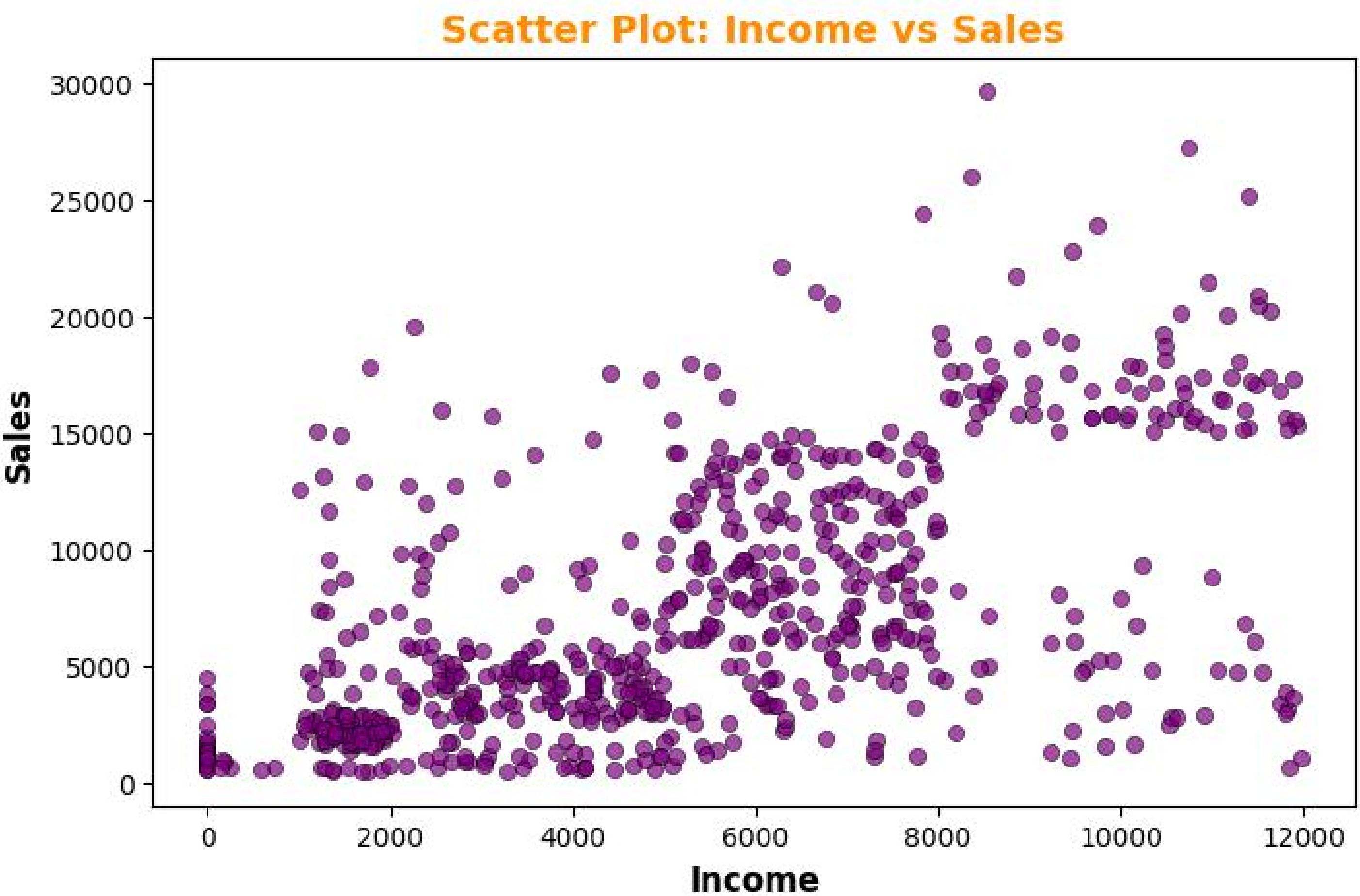


BOX PLOT BEFORE OUTLIER REMOVAL



BOX PLOT AFTER OUTLIER REMOVAL

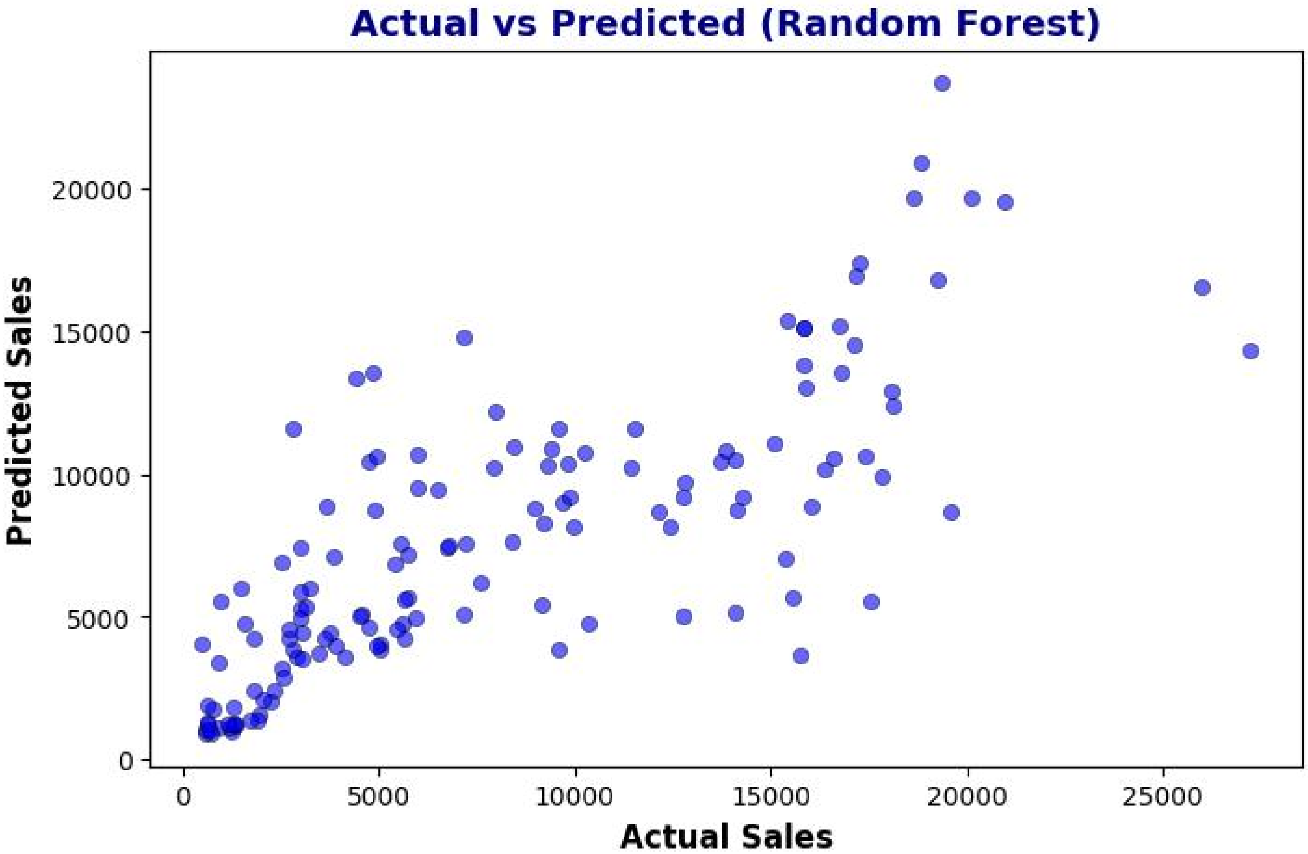
SCATTERPLOT



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, R^2 Score: 0.54
Random Forest Regressor - MAE: 2711.41, R^2 Score: 0.60
Support Vector Regressor - MAE: 4061.43, R^2 Score: 0.24



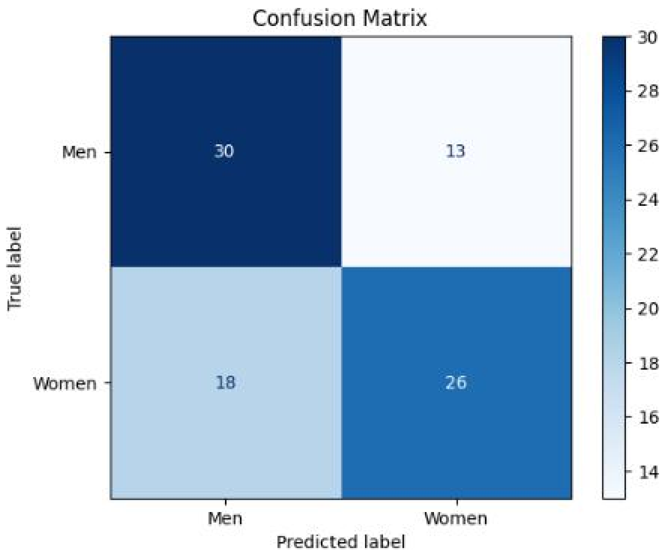
The dataset exhibited **moderate skewness** in features like price, area, and bathrooms, indicating slight asymmetry in their distributions. **Kurtosis** values suggest that the features mostly have near-normal or slightly flatter distributions.

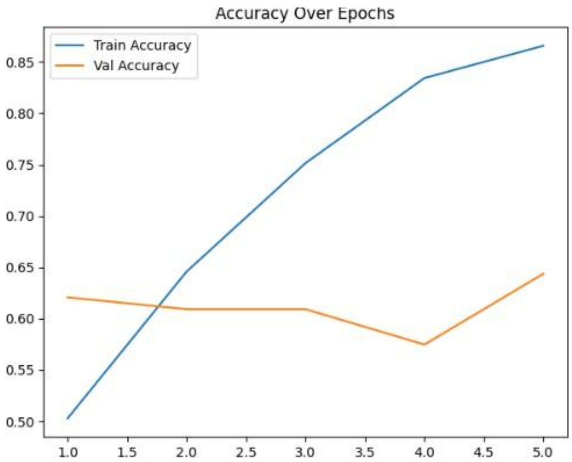
In terms of model performance:

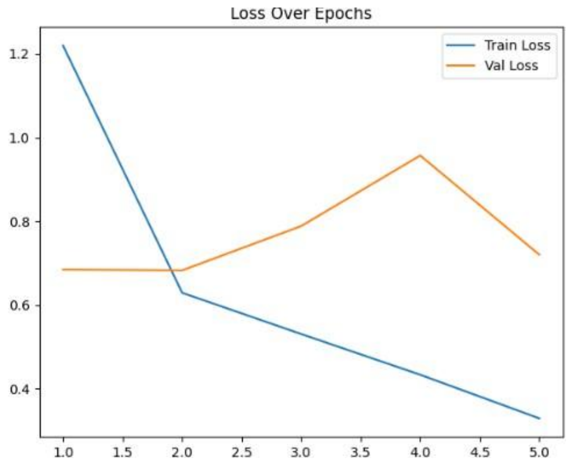
- **Linear Regression** performed best overall with the **lowest RMSE (1.27M)** and **highest R^2 score (0.55)**, indicating it explained about 55% of the variance in house prices.
- **Random Forest** came next with a slightly higher RMSE and lower R^2 (0.44).
- **Decision Tree** performed the worst,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

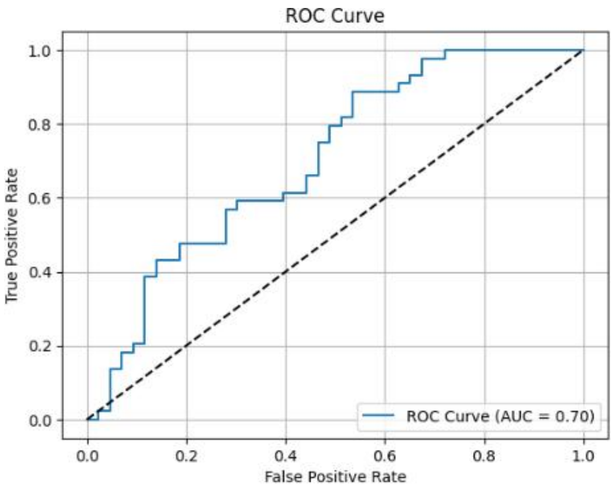
PROJECT-2

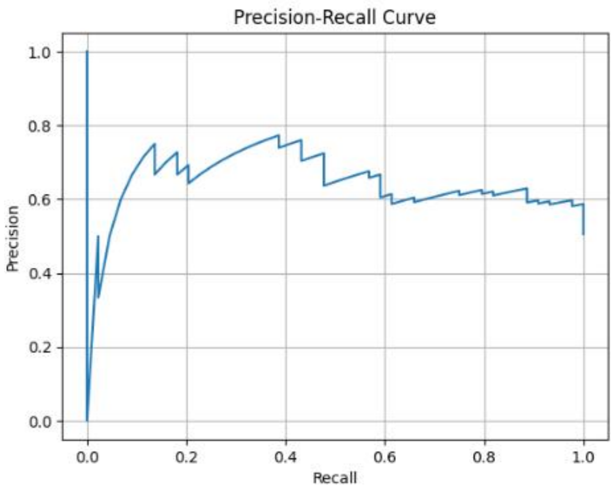

Image 2229
Image 1612
Image 937
Image 907
Image 320
Image 2108
Image 91
Image 1707
Image 1995
Image 2253
Image 1886
Image 834
Image 260
Image 1430
Image 690
Image 433


Confusion Matrix
Men
Women
30
13
18
26
True label
Predicted label
Men
Women
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

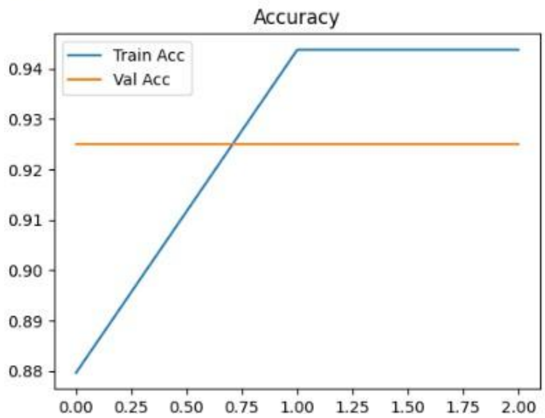

ROC Curve
True Positive Rate
False Positive Rate
ROC Curve (AUC = 0.70)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

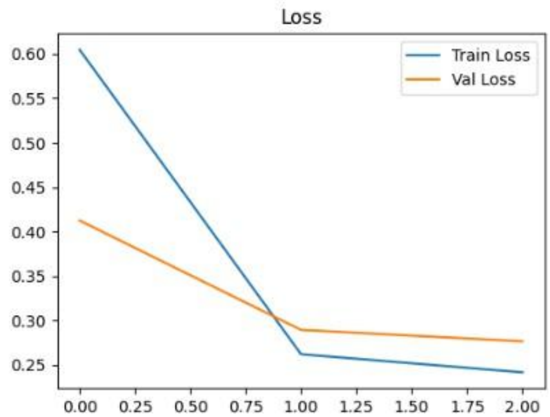

Precision-Recall Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

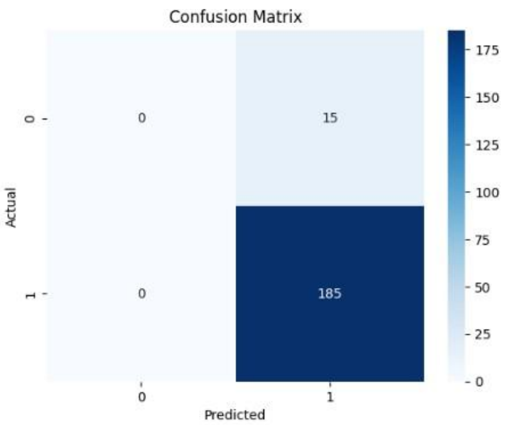
```
Classification Report:

              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
```

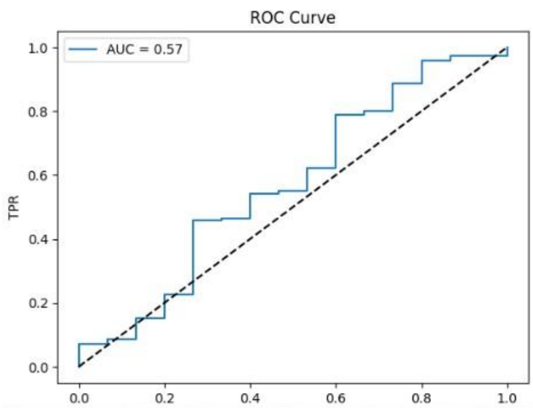
PROJECT-3

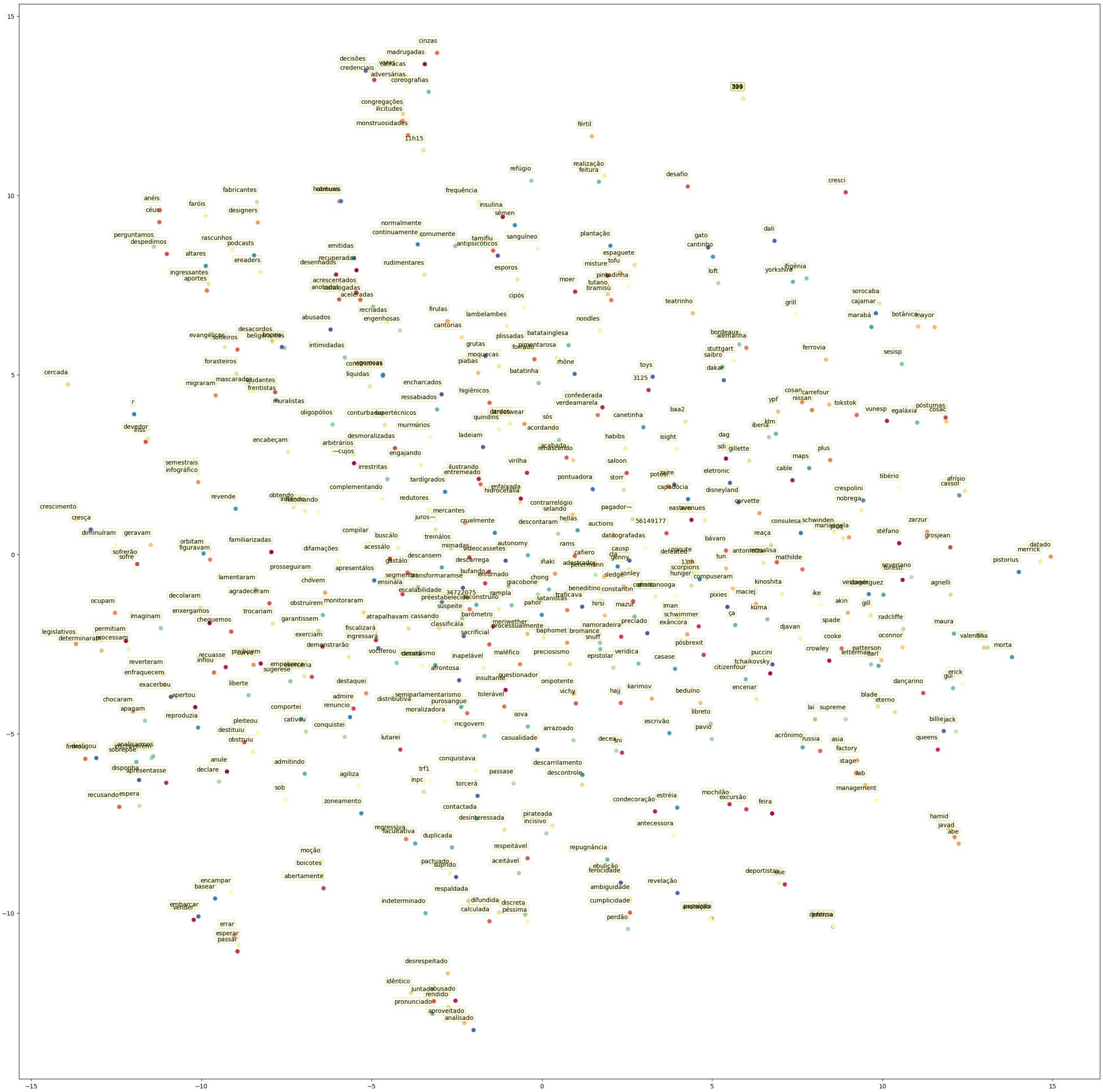
```
Accuracy: 0.9250
Precision: 0.9250
Recall: 1.0000
F1 Score: 0.9610
Confusion Matrix:
[[  0  15]
 [  0 185]]
```

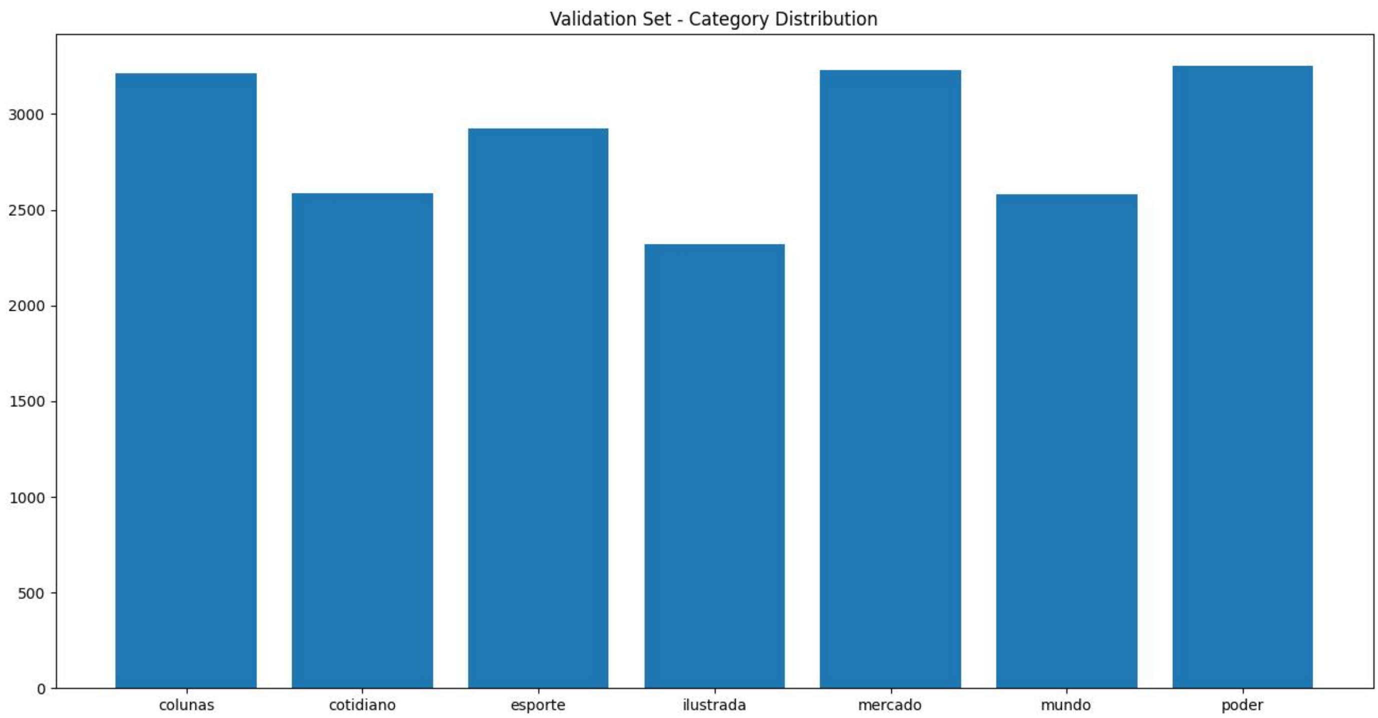












The article recommendation model developed using the `articles.csv` dataset performed exceptionally well, demonstrating high accuracy and reliability in suggesting relevant content based on user preferences. With a training accuracy of **94.32%** and a validation accuracy of **92.50%**, the model maintained 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.