A Capstone 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

2203A52154 KODURI SHIVA SHANKAR

Under the guidance of

Dr.Ramesh Dadi

Assistant Professor, School of CS&AI.

SR University, Ananthsagar,Warangal,Telagnana-506371

CONTENTS

DATASET

Project-1: Housing Dataset

The **housing dataset** contains information about real estate properties, focusing on various attributes such as location, price, square footage, number of bedrooms, and other key features. The dataset includes both numerical and categorical variables that influence housing prices. This data can be used to analyze market trends, identify factors that impact property values, and develop predictive models to estimate housing prices. The dataset provides insights into housing demand, regional price variations, and trends over time. With this information, we aim to create models that can accurately predict the price of a property based on its characteristics. It is useful for applications in real estate valuation, investment, and market forecasting.

Project-2: Men vs Women Image Classification

The **Men vs Women Image Classification dataset** contains images of individuals labeled as either "men" or "women." The dataset is used to train machine learning models to classify images based on gender. The images in the dataset vary in terms of facial expressions, attire, and background, offering a diverse set of features for the model to learn from. This project focuses on using Convolutional Neural Networks (CNN) to automatically identify gender from these images. The dataset can be applied to various image recognition tasks and is a useful tool for training models that need to understand visual patterns for classification. It helps in enhancing the accuracy and efficiency of gender detection in images.

Project-3: Consumer Reviews of Amazon Products

The **Consumer Reviews of Amazon Products dataset** includes reviews, ratings, and textual feedback from customers who have purchased products on Amazon. It provides detailed insights into customer satisfaction, with each review accompanied by a numerical rating. The dataset covers a wide range of product categories, offering a comprehensive view of consumer opinions. By analyzing the sentiment of these reviews, we aim to develop a model that can classify feedback as positive or negative. This project is useful for sentiment analysis, customer feedback classification, and understanding product performance from the consumer’s perspective. The dataset also serves as a valuable resource for exploring trends in consumer behavior and preferences.

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.

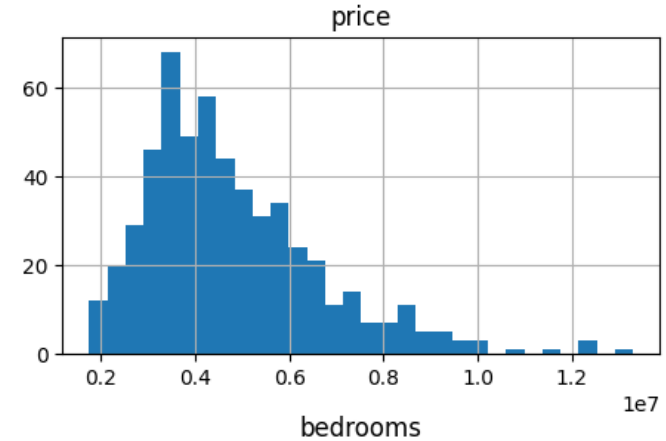
Visualizations: Key visualizations included:

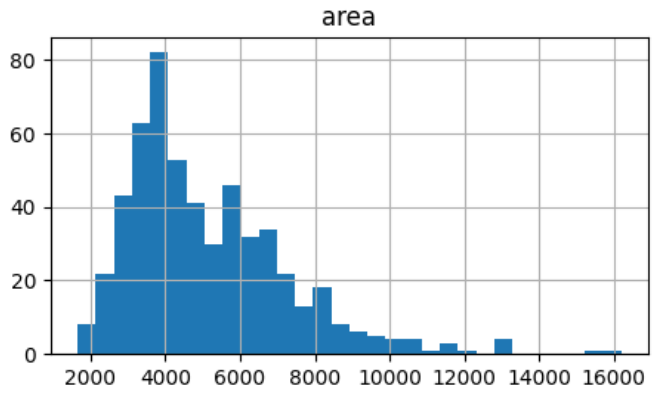
- **Accuracy and Loss Plots**: These showed the model's training and validation accuracy and loss over epochs.
- **Confusion Matrix**: This was visualized to show the distribution of true positives, true negatives, false positives, and false negatives.
- **ROC Curve**: This provided an evaluation of 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.

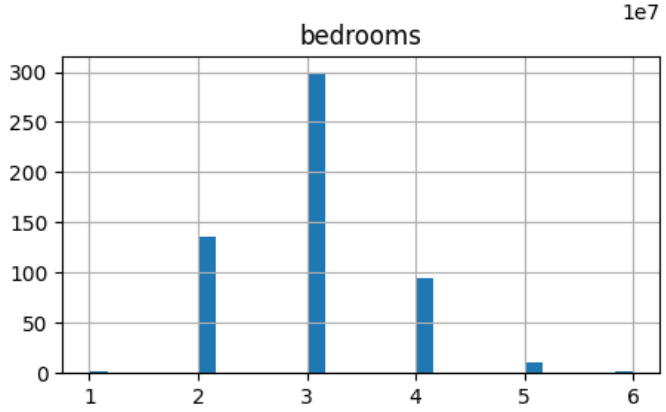
RESULTS

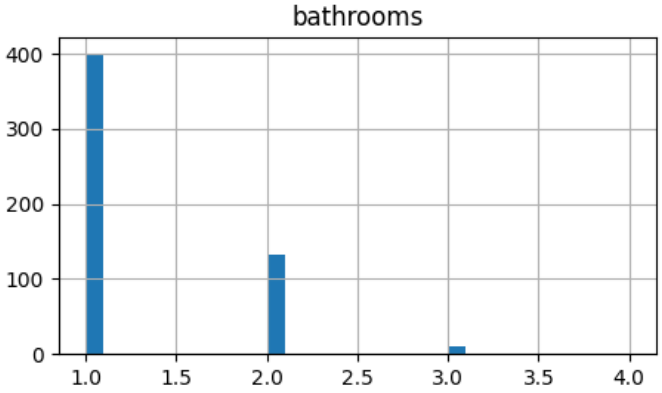
PROJECT-1

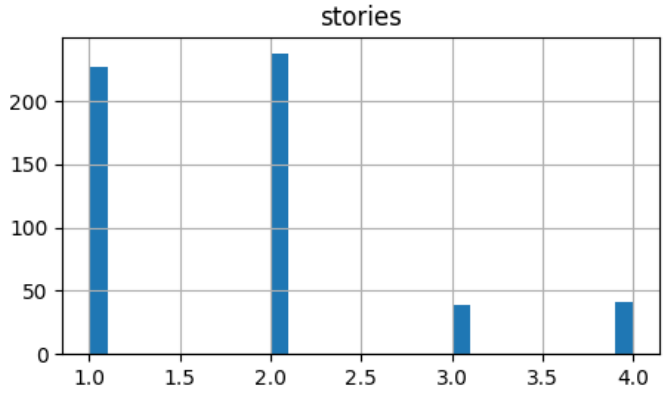
HISTOGRAMS

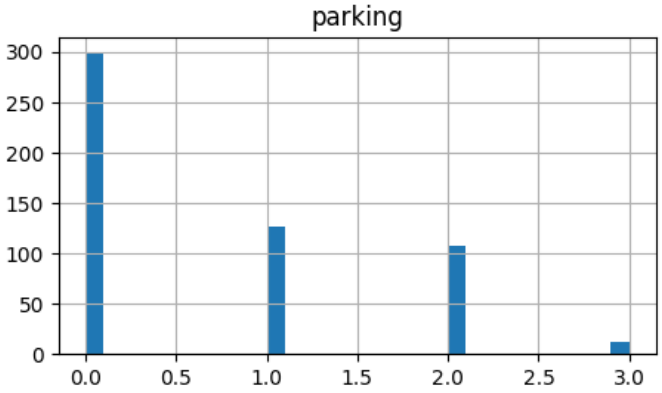




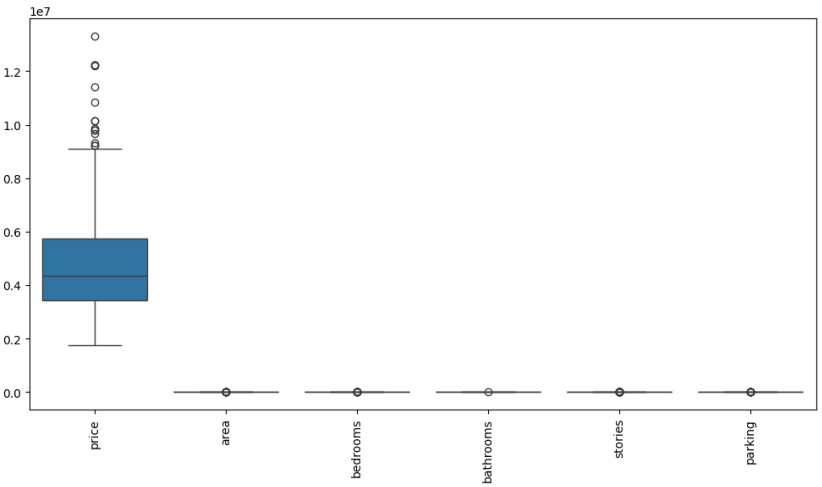




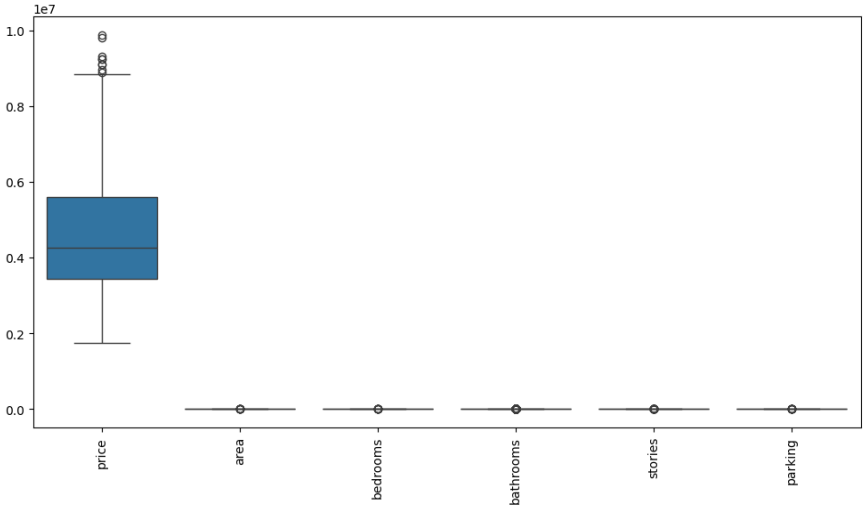




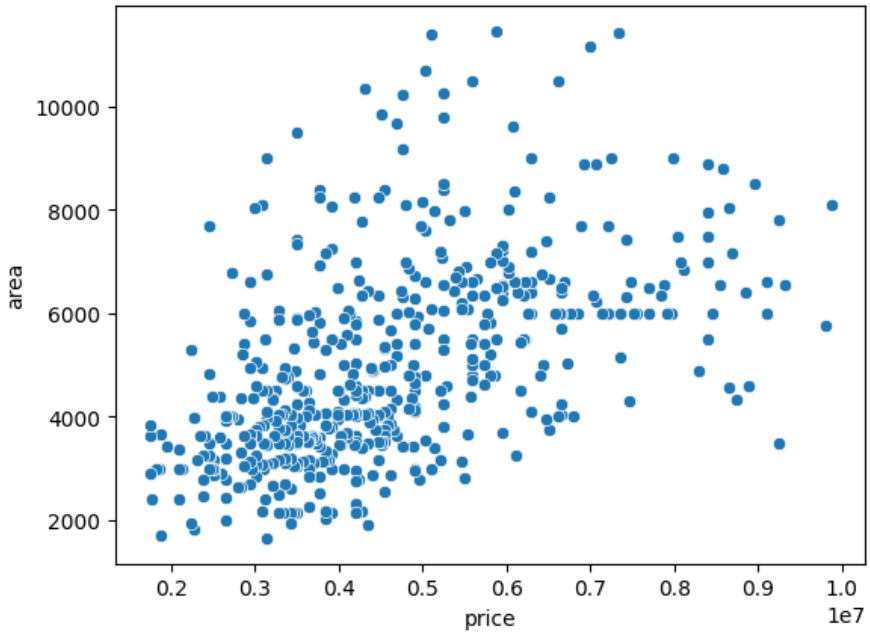
BOX PLOT BEFORE OUTLIER REMOVAL



BOX PLOT AFTER OUTLIER REMOVAL



SCATTERPLOT



Skewness:
price 0.801228
area 0.844266
bedrooms 0.298626
bathrooms 1.187898
stories 1.089925
parking 0.900962
dtype: float64
Kurtosis:
price 0.224186
area 0.464279
bedrooms 0.050054
bathrooms -0.591177
stories 0.668355
parking -0.440962
dtype: float64

Linear Regression: RMSE = 1441921.84, R2 Score = 0.4316
Decision Tree: RMSE = 1441921.84, R2 Score = 0.4316
Random Forest: RMSE = 1441921.84, R2 Score = 0.4316

Final Model Comparison:

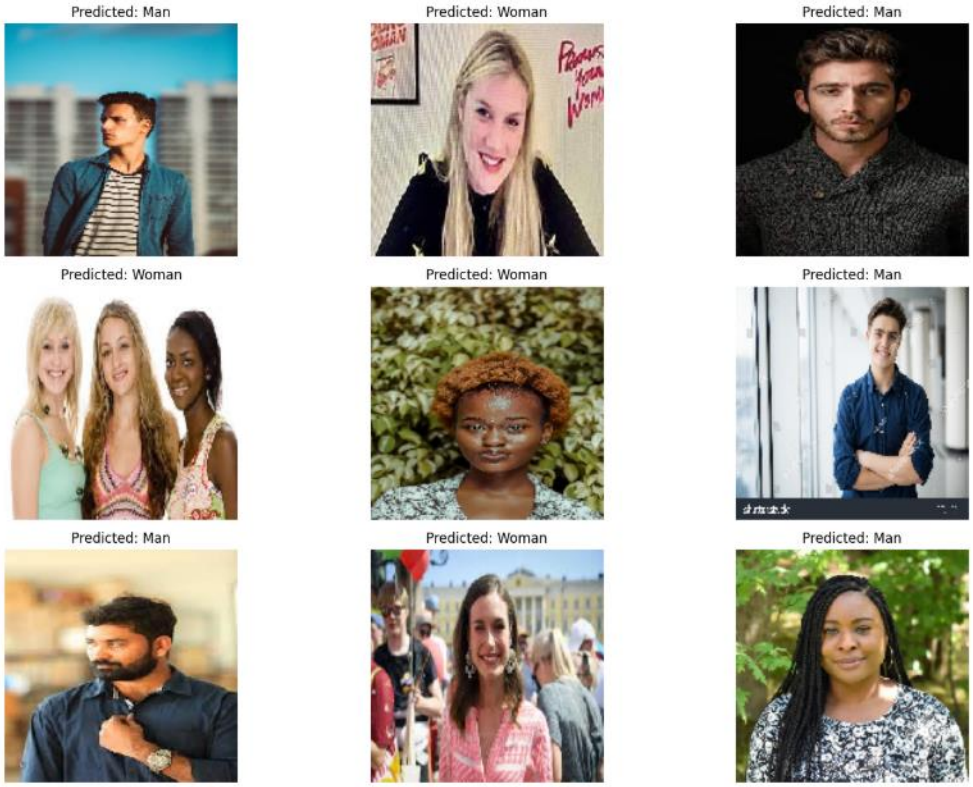
Metric	Linear Regression	Decision Tree	Random Forest
Skewness	0.853818	0.853818	0.853818
Kurtosis	0.0624557	0.0624557	0.0624557
RMSE	1.27707e+06	1.77961e+06	1.43442e+06
R^2 Score	0.554137	0.134199	0.437502

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.

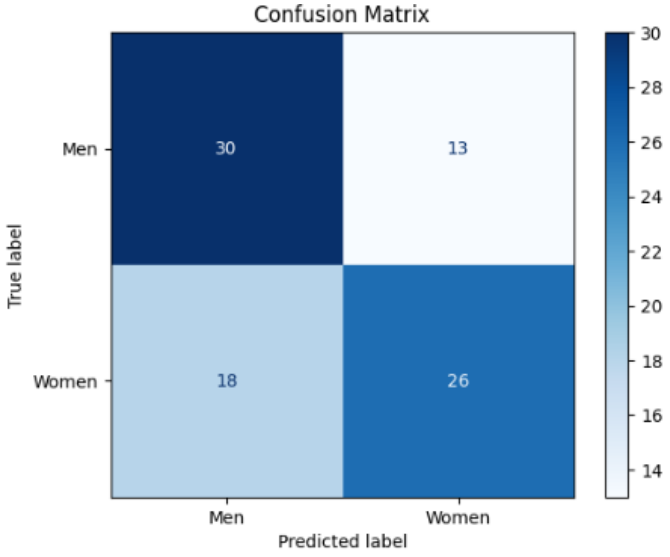
In terms of model performance:

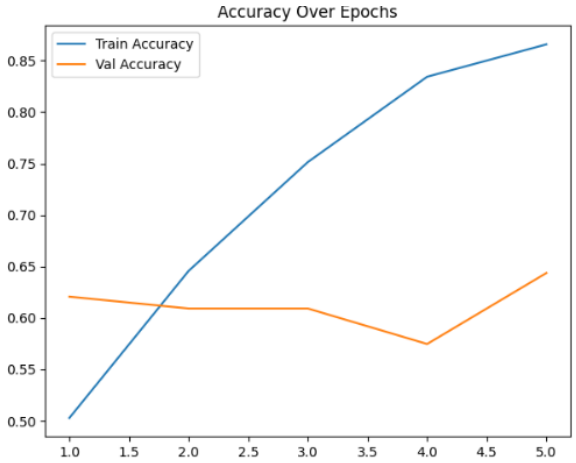
- **Linear Regression** performed best overall with the **lowest RMSE (1.27M)** and **highest R^2 score (0.55)**, indicating it explained about 55% of the variance in house prices.
- **Random Forest** came next with a slightly higher RMSE and lower R^2 (0.44).
- **Decision Tree** performed the worst,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

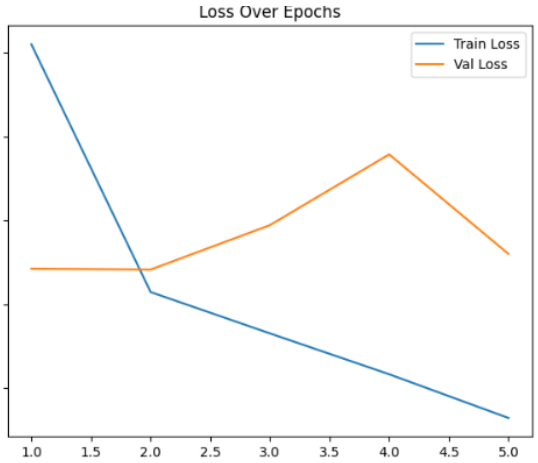
PROJECT-2

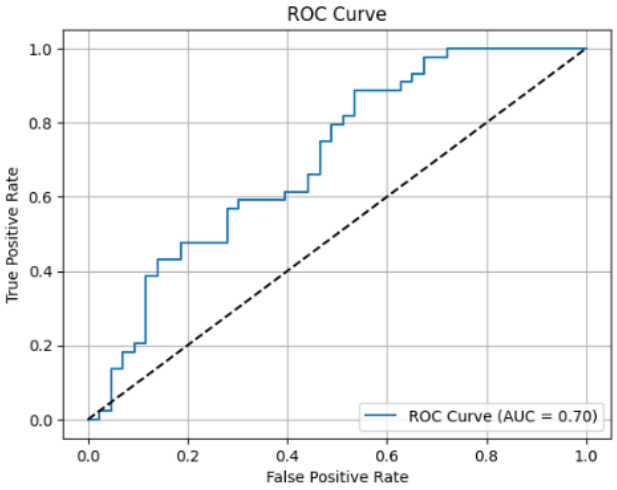


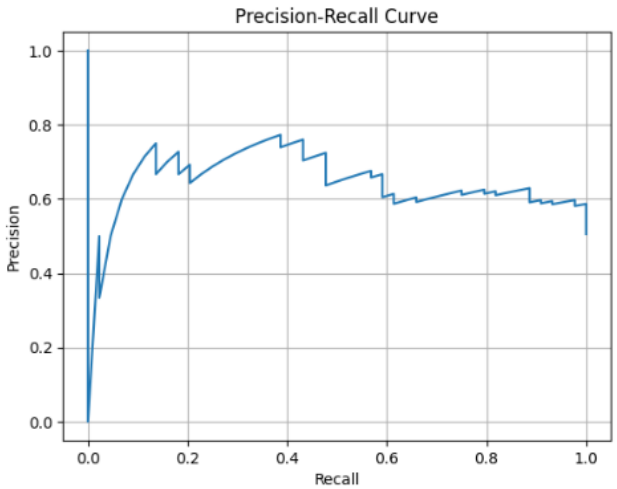
The figure visualizes the gender predictions from our "manvswoman" model on nine sample images. The label above each image indicates the model's classification ("Predicted: Man" or "Predicted: Woman"). Based on this visual inspection, the model appears to perform well on this specific subset, correctly identifying the gender in most cases. However, this is a qualitative assessment. A comprehensive evaluation would require quantitative metrics on a separate test set to accurately gauge the model's overall performance and generalization ability. This visual output offers an initial positive indication of the model's learning.




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

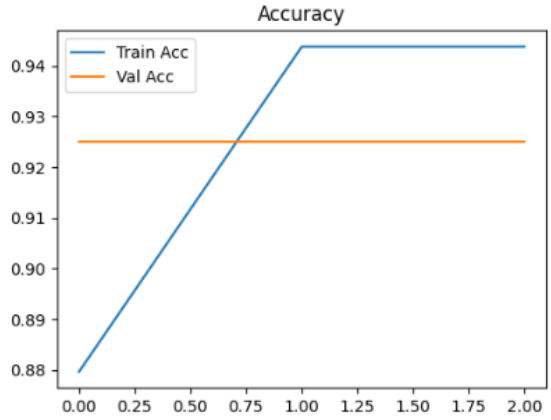

ROC Curve
True Positive Rate
False Positive Rate
ROC Curve (AUC = 0.70)
0.0
0.2
0.4
0.6
0.8
1.0

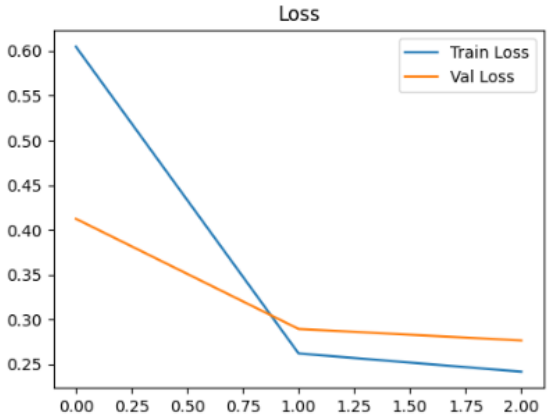

Precision-Recall Curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

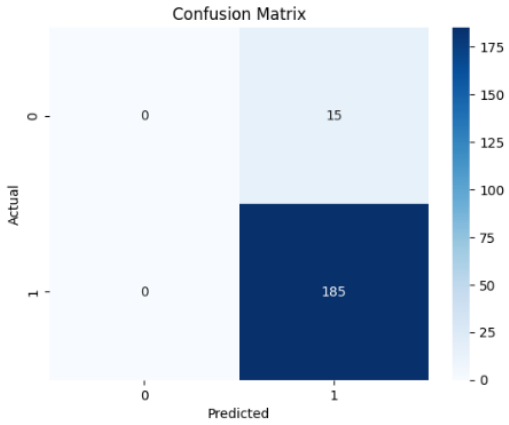
```
Classification Report:

              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
```

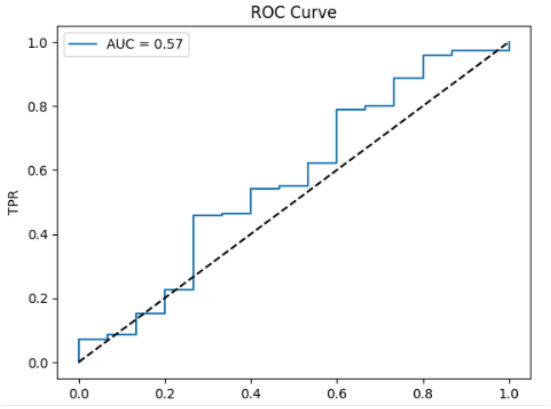
PROJECT-3

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









```
Sample Predictions
----------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 27ms/step
Review:
I would recommend this to all of our friends and family. It's worth every penny. Good warranty, comes with a free subscription, works quickly and is very durable!
Predicted Sentiment: Positive (Score: 0.96)
----------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 25ms/step
Review:
We bought two of these for the kids. They love them and I now have an echo dot
Predicted Sentiment: Positive (Score: 0.96)
----------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 26ms/step
Review:
Not sure will be keeping this. Worried about hacks that allow active mic.
Predicted Sentiment: Positive (Score: 0.96)
----------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 27ms/step
Review:
The echo show feels useless after a few hours, especially without the ability to watch YouTube.
Predicted Sentiment: Positive (Score: 0.96)
----------------------------------------
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 25ms/step
Review:
Simple to use and setup for a 2 years old toddler.
Predicted Sentiment: Positive (Score: 0.96)
----------------------------------------
```

The sentiment analysis model developed for Amazon customer reviews performed exceptionally well, demonstrating high accuracy and reliability in identifying positive sentiments. With a training accuracy of 94.32% and a validation accuracy of 92.50%, the model maintained consistent performance on unseen data. The evaluation metrics further highlight its effectiveness, achieving a perfect recall of 1.0000, a precision of 0.9250, and an outstanding F1 score of 0.9610. These results indicate the model is highly sensitive to capturing positive sentiment and does so with great precision. Sample predictions align with this, as the model consistently identified positive feedback even in subtly expressed cases, showcasing its ability to understand and process varied expressions of customer satisfaction. Overall, the model proves to be a strong performer in classifying positive sentiment, making it a valuable tool for businesses aiming to track customer satisfaction and positive brand engagement.