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

2203A52183 **THOKALA SATHWIKA**

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: 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: Fake News Identifier Application

The **Fake News Identifier** application utilizes a dataset comprising news articles and social media posts, labeled as either *real* or *fake*. Each entry includes the headline, article body, and sometimes the source and publication date. The dataset spans various topics including politics, health, science, and world news. This project focuses on the **classification of news content** using machine learning and natural language processing techniques to detect misinformation.

By analyzing the linguistic patterns and textual content, the aim is to develop a robust model that can automatically classify a given article as *fake* or *real*. This project is particularly valuable for combating the spread of misinformation, improving media literacy, and enhancing content credibility assessments. Additionally, it offers insights into the common characteristics of fake news and supports efforts in content moderation and digital journalism integrity.

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: Fake News Identifier Application

Data Collection and Preprocessing: The dataset comprises a collection of news articles labeled as either fake or true. Each entry typically includes the title and body text of the article. The data was first cleaned to remove punctuation, special characters, stopwords, and unnecessary whitespace. Text was then tokenized and converted into sequences of integers using a tokenizer. To ensure uniformity, all sequences were padded to the same length. These preprocessing steps help standardize the input for the LSTM model and improve learning performance.

Model Structure: The classification model uses a Long Short-Term Memory (LSTM) neural network, which is suitable for sequential text data. The architecture begins with an embedding layer that transforms each word into a dense vector representation. This is followed by two LSTM layers that capture contextual dependencies in the text. A dense layer with a sigmoid activation function is used at the end for binary classification, outputting values between 0 and 1 to indicate whether the news is real or fake.

Model Training: The model was compiled with the Adam optimizer and binary cross-entropy as the loss function, ideal for binary classification tasks. Training was conducted over 5 epochs with a batch size of 64, and 20% of the data was reserved for validation. During training, the model's accuracy and loss were monitored to ensure convergence and detect any signs of overfitting.

Evaluation Metrics: The trained model was evaluated using multiple metrics:

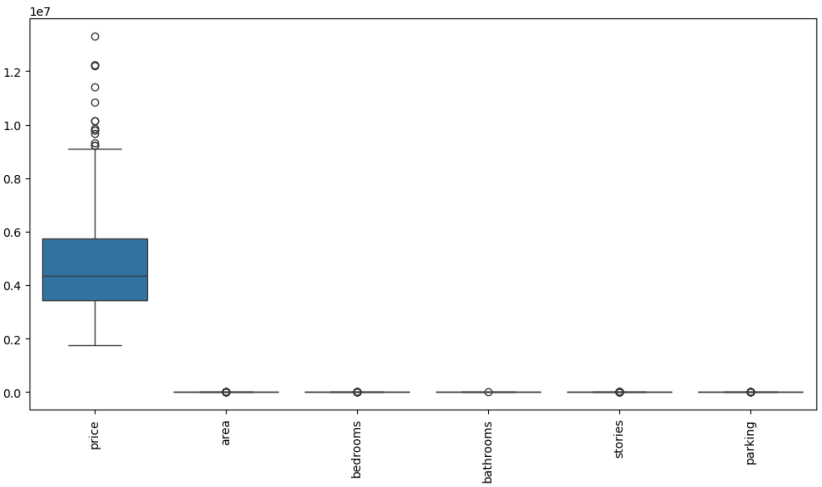
- Accuracy: Proportion of correct predictions.
- Precision: Proportion of true positive predictions among all positive predictions.
- Recall: Proportion of true positives among all actual positives.
- F1 Score: Harmonic mean of precision and recall.
- Confusion Matrix: Visualization of true positives, true negatives, false positives, and false negatives.
- ROC Curve: Plots true positive rate against false positive rate to evaluate threshold performance.

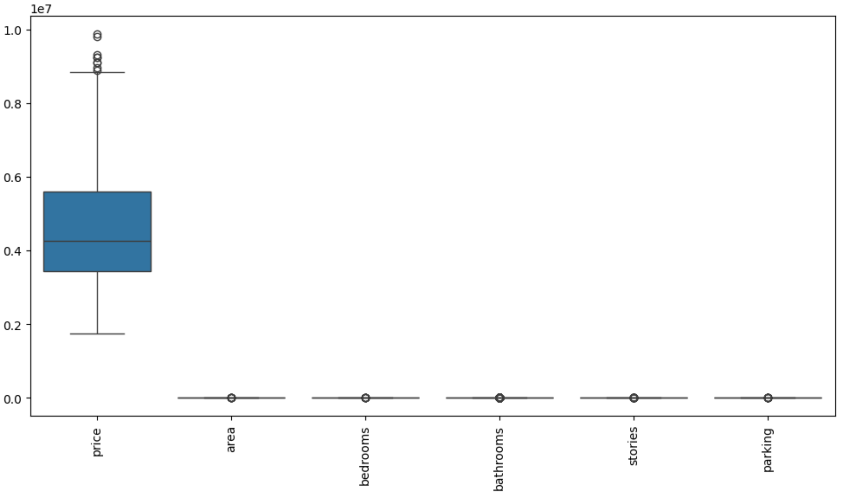
Visualizations

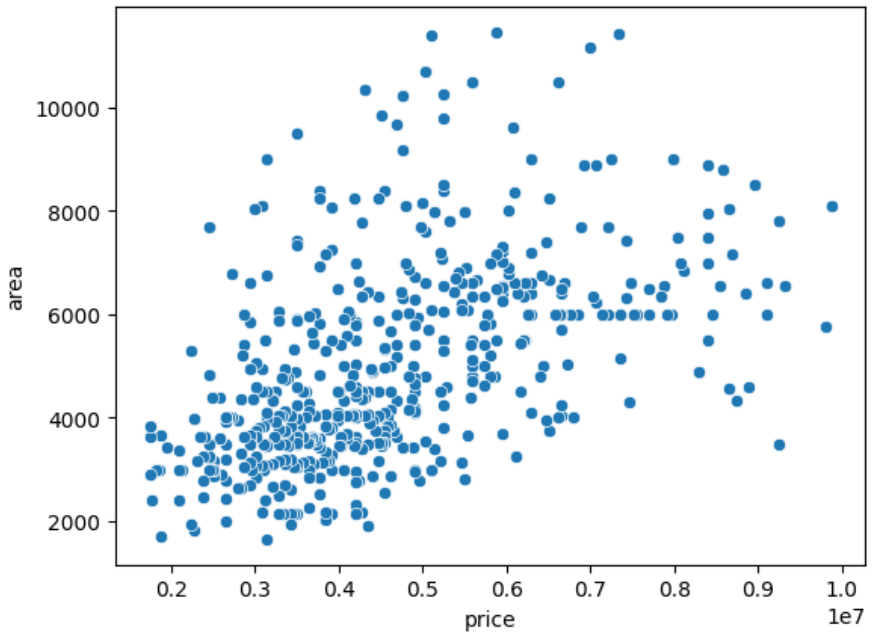
- Training and Validation Accuracy/Loss Curves: To observe how well the model learns over time.
- Confusion Matrix: To visually assess model predictions versus actual labels.
- ROC Curve and AUC Score: To evaluate classification performance across thresholds.
- Bar Charts of Precision, Recall, and F1 Score: To interpret model effectiveness in more detail.
- Prediction Samples: Random test samples were selected and classified by the model to visually inspect its performance, showing predicted vs. actual labels.

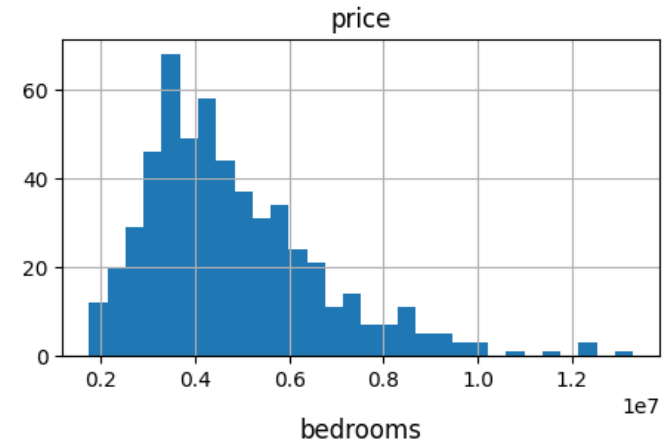
RESULTS

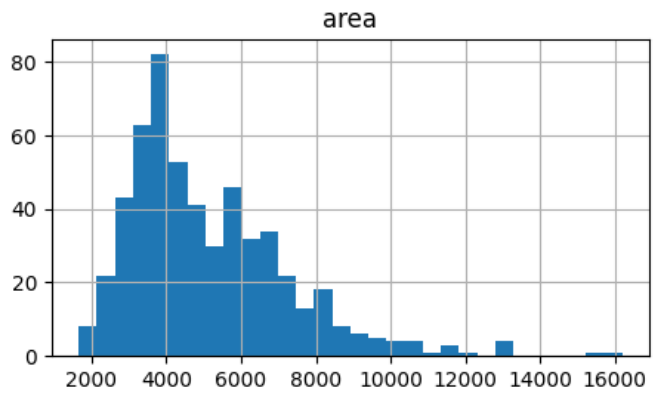
PROJECT-1

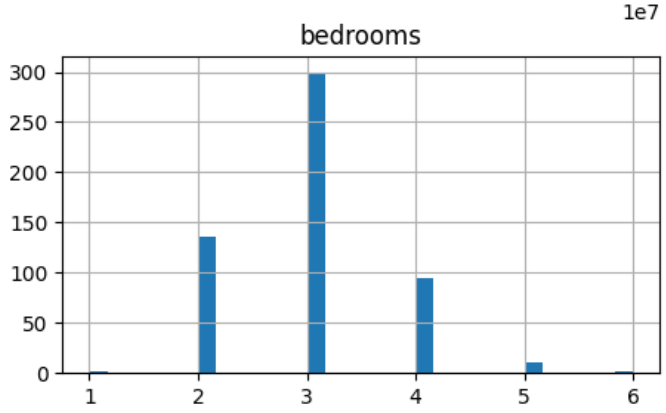


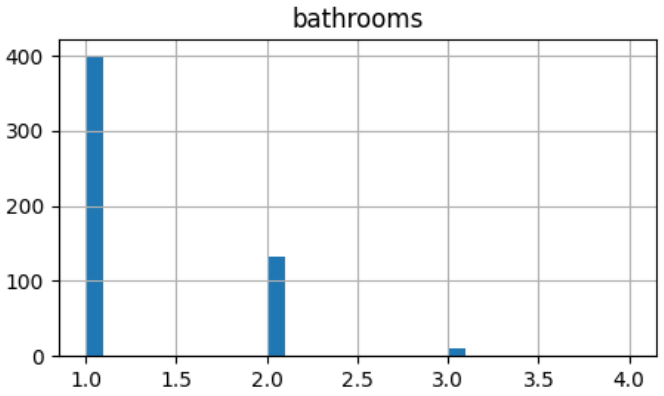


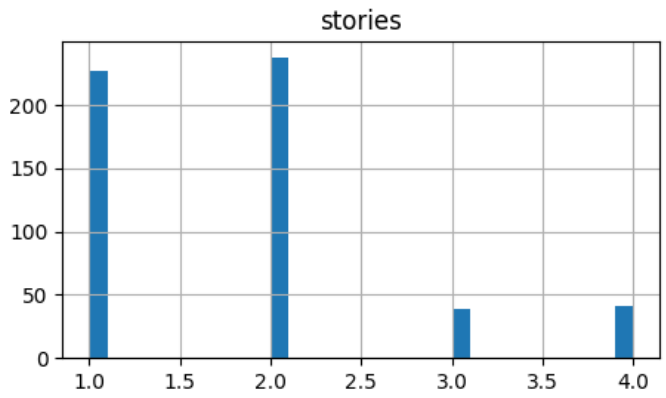


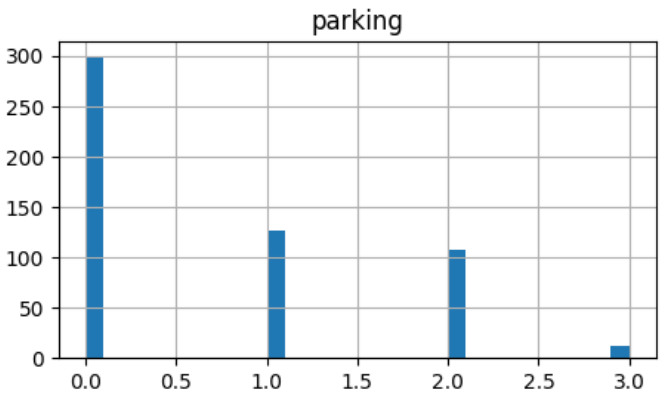












Skewness:
price 0.801228
area 0.844266
bedrooms 0.298626
bathrooms 1.187898
stories 1.089925
parking 0.900962
dtype: float64
Kurtosis:
price 0.224186
area 0.464279
bedrooms 0.050054
bathrooms -0.591177
stories 0.668355
parking -0.440962
dtype: float64

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:

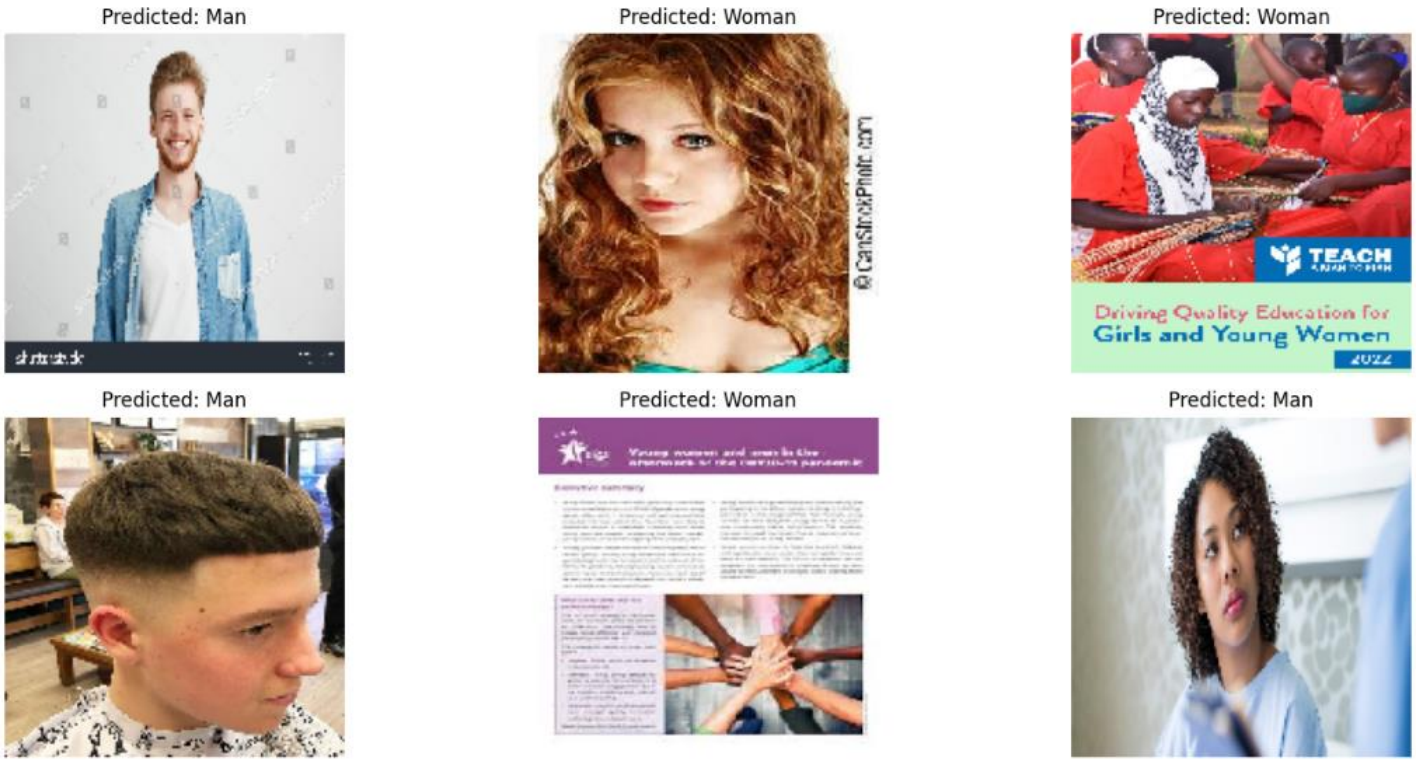
Metric	Linear Regression	Decision Tree	Random Forest
Skewness	0.853818	0.853818	0.853818
Kurtosis	0.0624557	0.0624557	0.0624557
RMSE	1.27707×10^6	1.77961×10^6	1.43442×10^6
R^2 Score	0.554137	0.134199	0.437502

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.

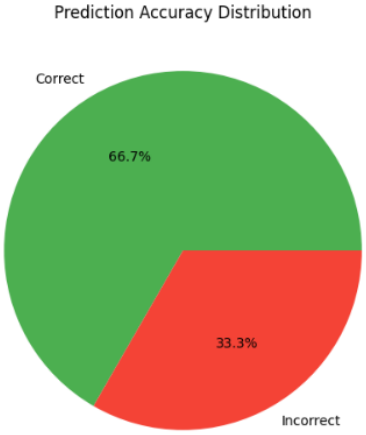
In terms of model performance:

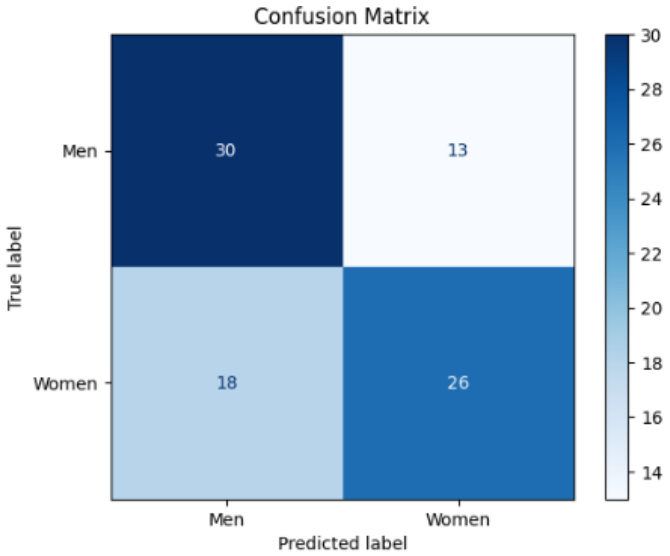
- **Linear Regression** performed best overall with the **lowest RMSE (1.27M)** and **highest R^2 score (0.55)**, indicating it explained about 55% of the variance in house prices.
- **Random Forest** came next with a slightly higher RMSE and lower R^2 (0.44).
- **Decision Tree** performed the worst,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

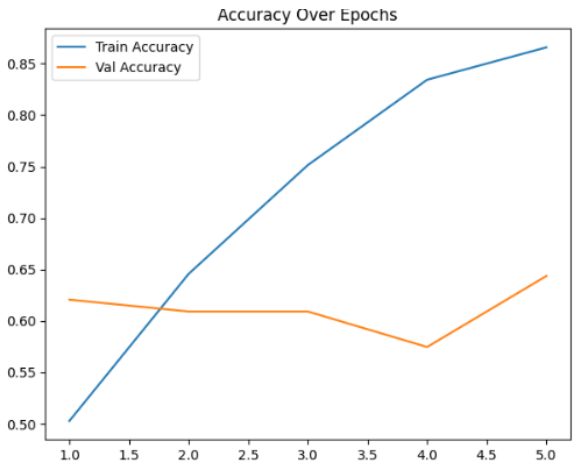
PROJECT-2

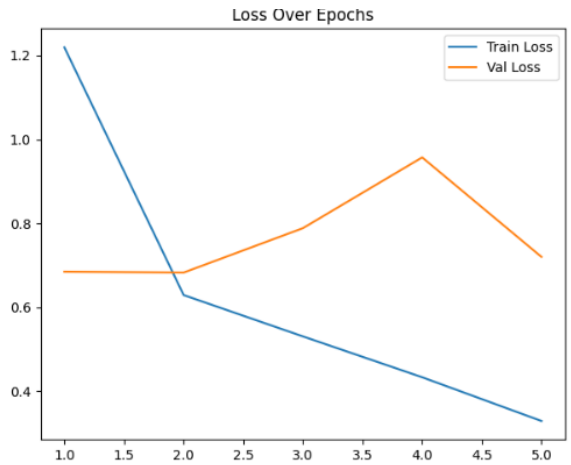


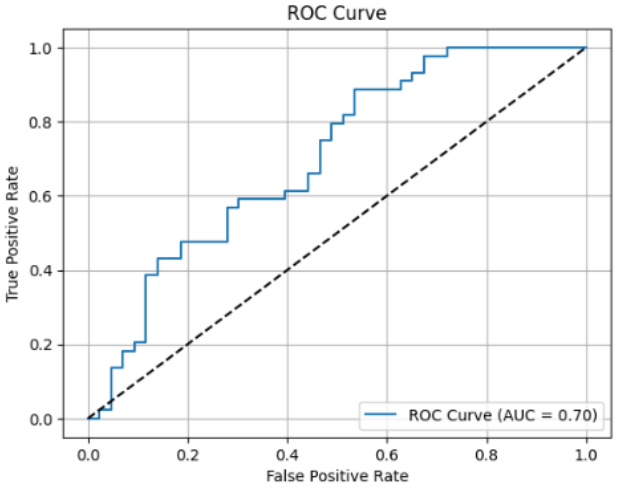
The figure visualizes the gender predictions from man vs woman model on nine sample images. The label above each image indicates the model's classification ("Predicted: Man" or "Predicted: Woman"). Based on this visual inspection, the model appears to perform well on this specific subset, correctly identifying 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.

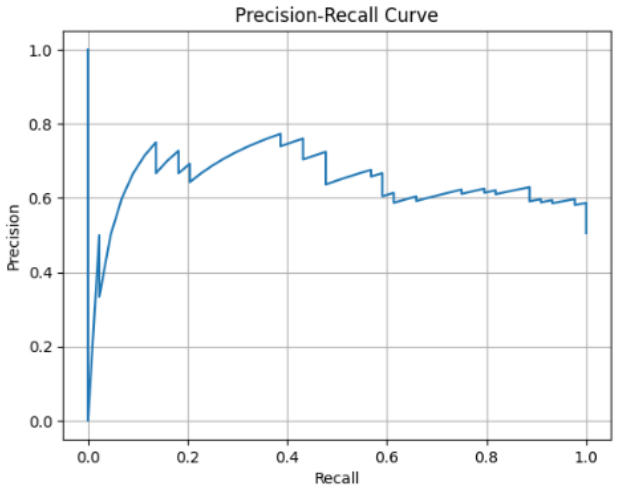



Confusion Matrix
Men
Women
True label
30
13
18
26
Men
Women
Predicted label
30
28
26
24
22
20
18
16
14


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





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

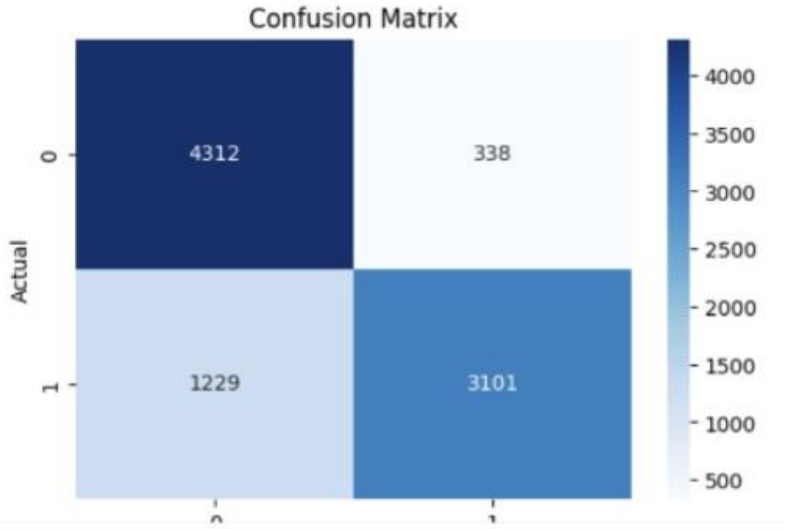
PROJECT-3

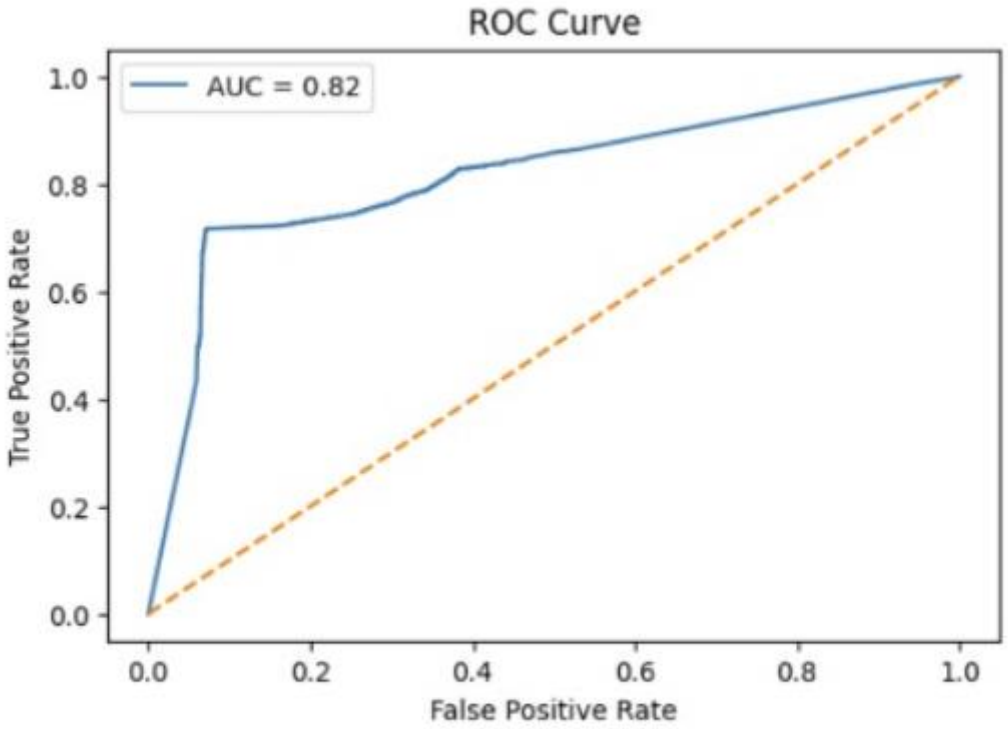
Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 500, 64)	320,000
lstm (LSTM)	(None, 500, 64)	33,024
dropout (Dropout)	(None, 500, 64)	0
lstm_1 (LSTM)	(None, 32)	12,416
dropout_1 (Dropout)	(None, 32)	0
dense (Dense)	(None, 1)	33

```
Epoch 1/5
20/20 ━━━━━━━━━━━━━━━━━━━━ 34s 2s/step - accuracy: 0.5193 - loss: 0.6935 - val_accuracy: 0.5178 - val_loss:
0.6924
Epoch 2/5
20/20 ━━━━━━━━━━━━━━━━━━━━ 29s 2s/step - accuracy: 0.5193 - loss: 0.6883 - val_accuracy: 0.4986 - val_loss:
0.6927
Epoch 3/5
20/20 ━━━━━━━━━━━━━━━━━━━━ 29s 2s/step - accuracy: 0.5353 - loss: 0.6681 - val_accuracy: 0.7993 - val_loss:
0.4499
Epoch 4/5
20/20 ━━━━━━━━━━━━━━━━━━━━ 29s 2s/step - accuracy: 0.8043 - loss: 0.4519 - val_accuracy: 0.8401 - val_loss:
0.4972
Epoch 5/5
20/20 ━━━━━━━━━━━━━━━━━━━━ 29s 1s/step - accuracy: 0.8491 - loss: 0.4524 - val_accuracy: 0.8255 - val_loss:
0.4565
```

```
Classification Report:
```

	precision	recall	f1-score	support
0	0.78	0.93	0.85	4650
1	0.90	0.72	0.80	4330
accuracy			0.83	8980
macro avg	0.84	0.82	0.82	8980
weighted avg	0.84	0.83	0.82	8980






```
Precision: 0.9017
Recall: 0.7162
F1 Score: 0.7983
```