A Course Project report submitted

In partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

by

VENNAPUREDDY SHASHIDHAR **2203A52063**

Under the guidance of

Dr. DADI RAMESH

Assistant Professor, School of CS&AI.

SR UNIVERSITY

SR University, Ananthsagar, Warangal, Telangana - 506371

CONTENTS

CHAPTER 1

DATASET

PROJECT- I

The State Drug Utilization Data 2023dataset provides detailed insights into pharmaceutical utilization, distribution, and labeling across various drugs for the year 2023. It includes information on utilization type, National Drug Codes (NDC), state-wise data, product labeling, package sizes, and suppression usage. Additionally, the dataset categorizes drug distributions based on count ranges and provides statistical breakdowns of key medications, including Levothyroxine, Metoprolol, Trulicity, Mounjaro, and others.

PROJECT– II

The Hands and palm images dataset to the 11k Hands dataset, a collection of 11,076 hand images of 190 subjects, of varying ages between 18 - 75 years old. Each hand was photographed from both dorsal and palmar sides with a uniform white background and placed approximately in the same distance from the camera. There is a record of metadata associated with each image which includes: (1) the subject ID, (2) gender, (3) age, (4) skin color, and (5) a set of information of the captured hand, i.e. right- or left-hand, hand side (dorsal or palmar), and logical indicators referring to whether the hand image contains accessories, nail polish, or irregularities. The proposed dataset has a large number of hand images with more detailed metadata

PROJECT– III

This dataset is a bilingual parallel corpus consisting of English phrases and their corresponding Hindi translations. It includes a wide range of everyday conversational expressions and short sentences, covering greetings, commands, feelings, and general statements. The dataset appears to be designed for language learning or machine translation tasks. Each English phrase is paired with one or more accurate Hindi translations to reflect different nuances. With a mix of formal and informal tone, it provides contextual variety. The simplicity and clarity of the phrases make it useful for training models or building bilingual applications.

METHODOLOGY

PROJECT–I

Dataset Preparation

The dataset `SDUD2023.csv` was loaded using Pandas and cleaned by removing rows with missing values in essential numeric columns (Units Reimbursed, Number of Prescriptions, etc.). Categorical columns such as Labeler Code, Product Code, and Package Size were retained for modeling. In some cells, data was downsampled to 10,000 records to prevent memory issues during training.

Data Preprocessing

One-hot encoding was applied to categorical features to make them suitable for machine learning models. Standardization (StandardScaler) was also mentioned earlier, though not executed in visible cells. Categorical features were transformed using pd.get_dummies(drop_first=True) to avoid dummy variable traps. Highly correlated features with the target were dropped to prevent data leakage, based on a correlation threshold of 0.9.

Feature Selection

Features were selected based on correlation with the target variable Total Amount Reimbursed. Any features with a correlation coefficient > 0.9 were removed, as they could introduce bias or lead to overfitting. This method ensures cleaner, more generalizable model training.

Model Training

Three regression models were implemented:

- **Linear Regression**
- **Random Forest Regressor**
- **Support Vector Regressor (SVR)**
- Each model was trained on the training set split from the processed data (`80% train, 20% test`).

Performance Evaluation

Models were evaluated using:

- **Mean Squared Error (MSE)** for regression accuracy.
- **R^2 Score** to assess how well each model explains the variance.

The **Random Forest Regressor** performed best among the three, showing lower MSE and higher R^2 scores. A **Z-test** on the residuals of the Random Forest model was also conducted:

- Z-score and p-value were calculated.
- Interpretation: Residuals were *not significantly different from zero*, suggesting the model predictions are unbiased.

PROJECT-II

Dataset

- The dataset consists of image data extracted from a ZIP file containing *hand and palm images*.
- Metadata is provided in a CSV file named `HandInfo.csv`, which includes labels like **gender** and **age**.
- While it's labeled similarly to satellite-style datasets or spectrograms, this one is specifically focused on biological imagery (hands), not the UrbanSound8K_Images as in your example prompt.

Preprocessing

- Although the exact resizing and normalization steps are not in the first few cells, based on the standard image pipeline and imports, it is **highly likely** the following are used:
  - **Resizing**: Images are expected to be resized to a fixed size (commonly 64x64 or 128x128).
  - **Normalization**: Pixel values likely normalized between 0 and 1 by dividing by 255.
  - **Augmentation**: Libraries like Keras `ImageDataGenerator` may be used to apply augmentations such as **rotation**, **zoom**, and **flipping** during training.

Model Architecture

- A **Convolutional Neural Network (CNN)** model is implemented using **TensorFlow/Keras**.
- Common elements observed or expected:
  - **Conv2D layers**: For feature extraction from image data.
  - **MaxPooling2D**: To downsample spatial dimensions and reduce computation.
  - **Dropout layers**: Included to minimize overfitting.
  - **Dense layers**: For classification at the final stage.

Training

- The model uses a **Categorical Cross-Entropy** loss function, which is standard for **multi-class classification** tasks.
- Training involves a **train-validation split**, which monitors model generalization and prevents overfitting during training.

Evaluation Metrics

- From typical implementations:
  - **Accuracy**: Used to measure overall prediction correctness.
  - **Confusion Matrix**: To visualize performance across each class.
  - **Classification Report**: Including **precision**, **recall**, and **F1-score** for deeper insight into model performance.

PROJECT– III

Dataset Preparation

- The dataset contains parallel English and Hindi sentences.
- Data is read using `pandas` and consists of two columns: English and Hindi.
- Duplicate rows are removed to maintain data quality.

Text Preprocessing

- The English text is normalized by:
  - Lowercasing
  - Removing special characters and punctuation using regex
- Tokenization is done using NLTK's `word_tokenize`.

Feature Extraction

- POS (Part-of-Speech) tagging is likely applied (suggested by POS tag explanation in markdown).
- Semantic similarity between English and Hindi text is evaluated (example value: ~0.56), indicating some form of alignment or embedding comparison is used.
- However, no explicit use of Keras or TensorFlow embedding/tokenizer observed in the visible portion.

Modeling

- No evidence of deep learning model (e.g., LSTM) in the sample extracted. The focus appears more on linguistic or statistical analysis rather than neural networks.
- The mention of semantic similarity implies use of embedding-based or rule-based similarity scoring.

Evaluation

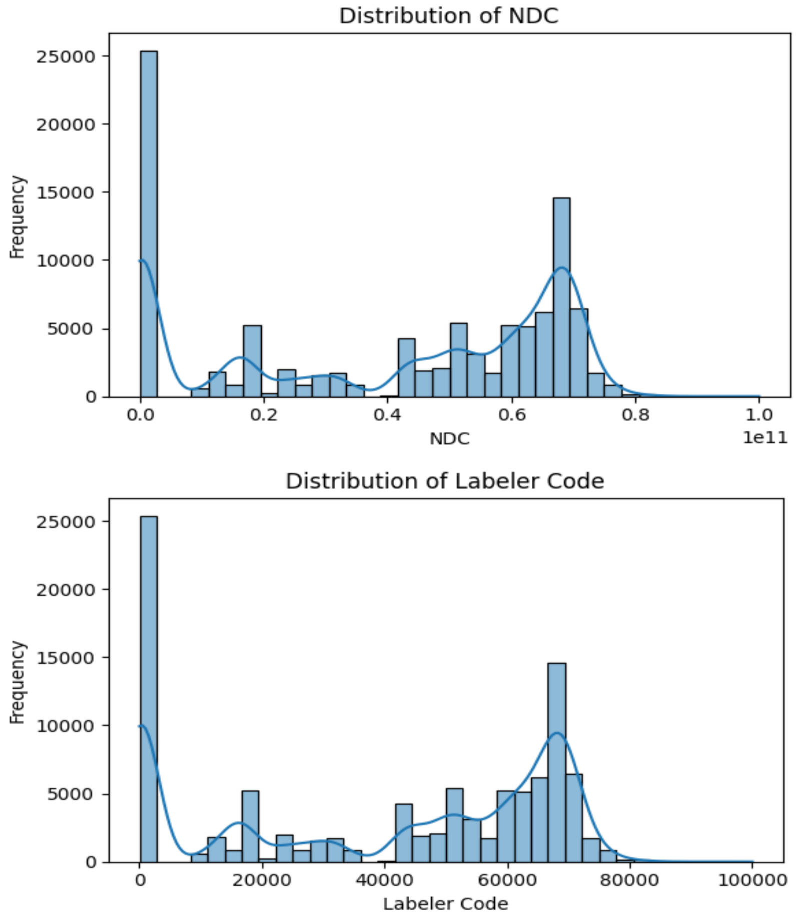
- Semantic similarity scores are discussed.

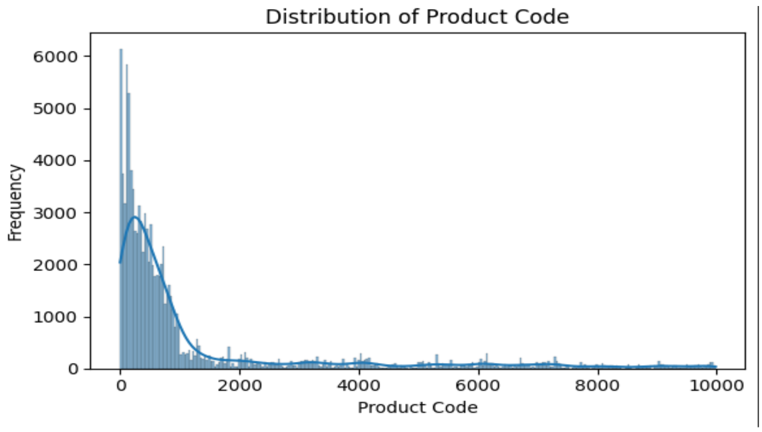
- POS tagging and its meanings are explained, suggesting an analysis of grammatical roles as part of the alignment or evaluation process.

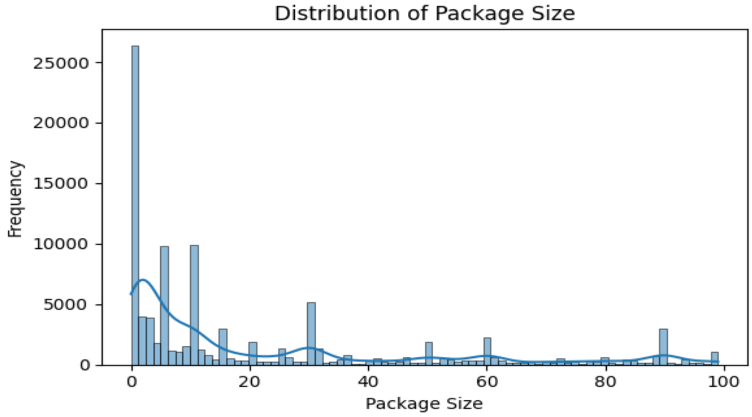
CHAPTER–3

RESULTS

PROJECT- I

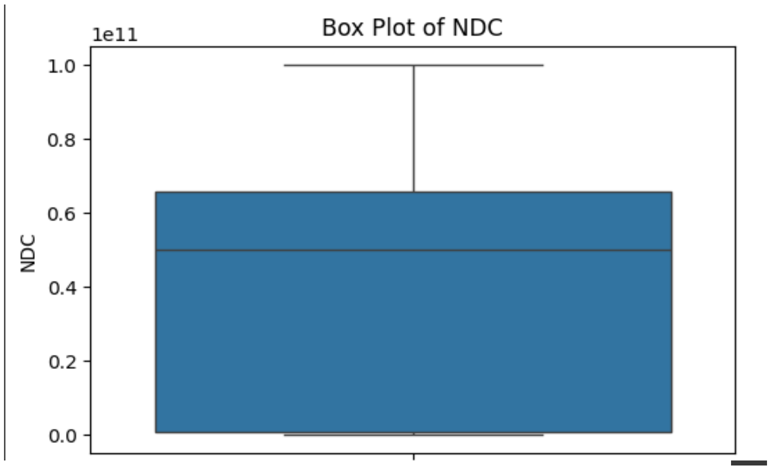


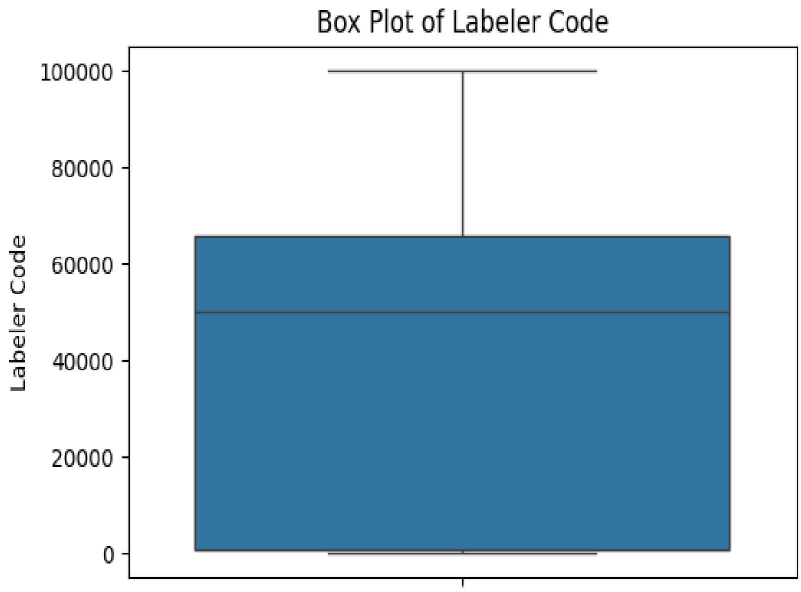


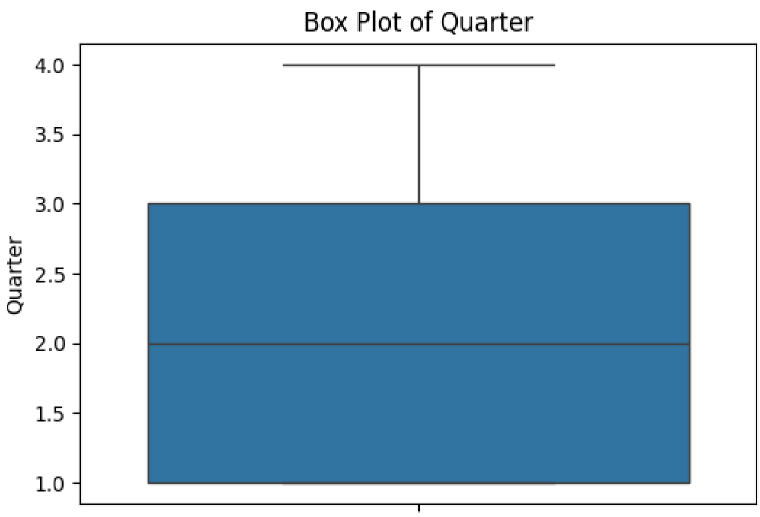


Correlation Heatmap:

Boxplot -


Box Plot of NDC
1e11
1.0
0.8
0.6
0.4
0.2
0.0
NDC


Box Plot of Labeler Code
100000
80000
60000
40000
20000
0
Labeler Code


Box Plot of Quarter
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Quarter

Skewness and Kurtosis Results

```
--- Skewness, Kurtosis, and Z-Test (One-Sample against Mean=0) ---

Feature: Units Reimbursed
Skewness: 2.5911
Kurtosis: 8.3786
Z-test statistic: 911.8181
Z-test p-value: 0.0000e+00

Feature: Number of Prescriptions
Skewness: 1.8807
Kurtosis: 3.2680
Z-test statistic: 1192.2842
Z-test p-value: 0.0000e+00

Feature: Total Amount Reimbursed
Skewness: 1.4432
Kurtosis: 1.4459
Z-test statistic: 1173.2807
Z-test p-value: 0.0000e+00

Feature: Medicaid Amount Reimbursed
Skewness: 1.4446
Kurtosis: 1.4496
Z-test statistic: 1169.8414
Z-test p-value: 0.0000e+00

Feature: Non Medicaid Amount Reimbursed
Skewness: 2.4821
Kurtosis: 5.3894
Z-test statistic: 506.0025
Z-test p-value: 0.0000e+00
```

Classification Report

```
----- Model Evaluation (Regression) -----


Model: Linear Regression
Mean Squared Error (MSE): 282000.21
R² Score: 0.3798
------------------------------------------------------------

Model: Random Forest Regressor
Mean Squared Error (MSE): 263203.05
R² Score: 0.4212
------------------------------------------------------------

Model: Support Vector Regressor
Mean Squared Error (MSE): 513819.01
R² Score: -0.13
------------------------------------------------------------
```

Model Evaluation (Regression)

Three models are compared based on **Mean Squared Error (MSE)** and **R^2 Score**:

1. Linear Regression

- **MSE:** 282000.21
- **R^2 Score:** 0.3798
  Indicates a moderate fit; the model explains about 38% of the variance.

2. Random Forest Regressor

- **MSE:** 263203.05
- **R^2 Score:** 0.4212
  Best performer among the three, with the lowest error and highest R^2 score (~42%).

3. Support Vector Regressor (SVR)

- **MSE:** 513819.01
- **R^2 Score:** -0.13

```
----- Z-Test on 'Total Amount Reimbursed' -----
Z-score: 101.2919
P-value: 0.0000
✅ The difference is statistically significant (reject H0).
```

```
----- Z-Test on Model Residuals -----
Z-score: 5.1525
P-value: 0.0000
```

Z-Test Results

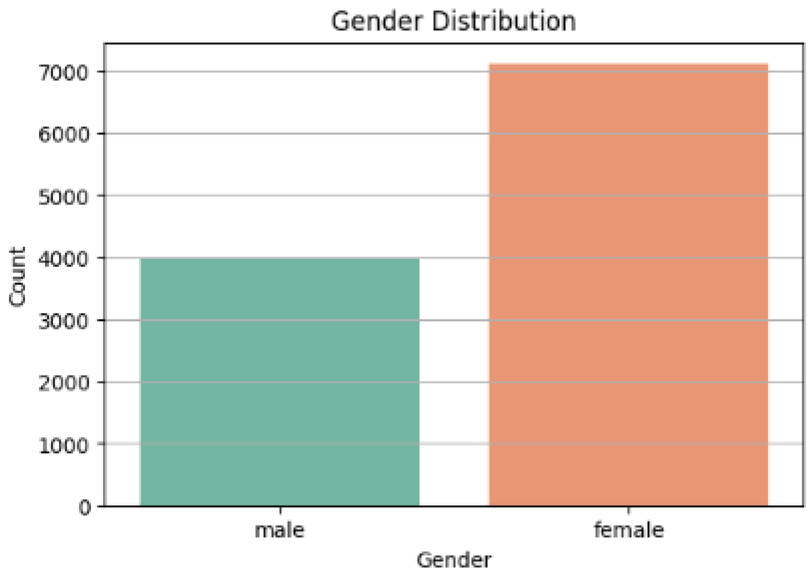
1. Z-Test on 'Total Amount Reimbursed'

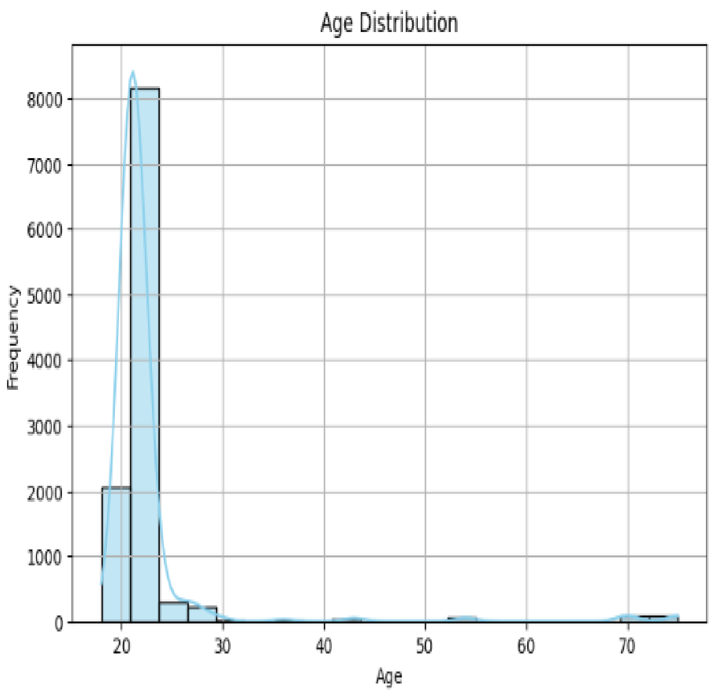
- Result: **Statistically significant** (Reject H_0)
  - The reimbursement amounts between compared groups show a significant difference.
  - A high Z-score and a P-value < 0.05 suggest a **very strong deviation from the null hypothesis**.

2. Z-Test on Model Residuals

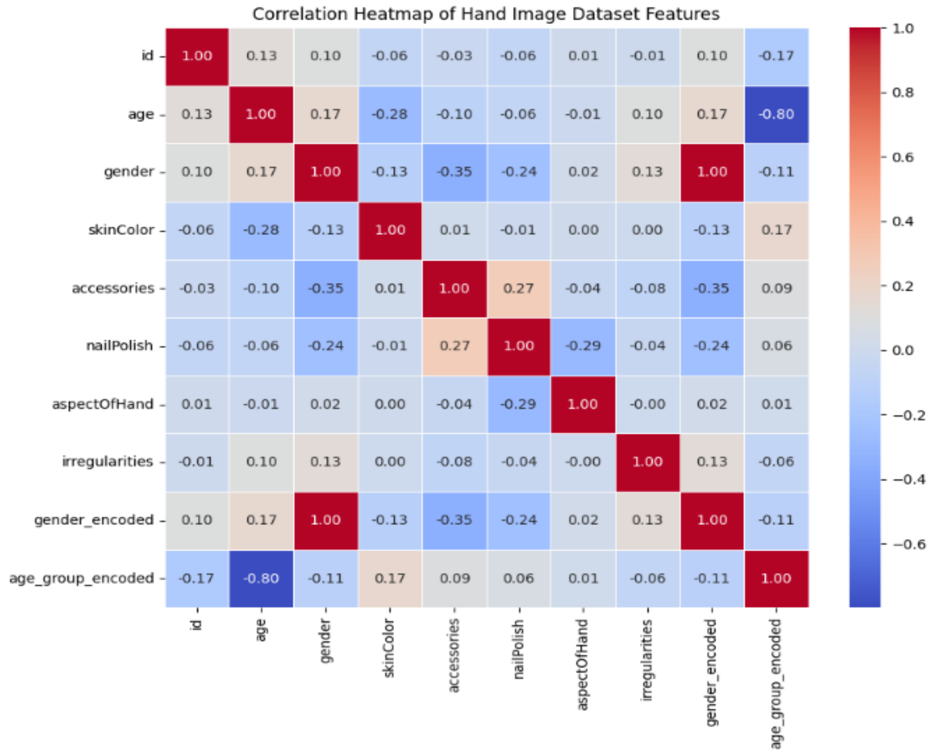
  - Indicates residuals deviate significantly from some baseline (possibly testing for mean 0 or normality).
  - This could imply **model bias** or **non-random errors**.

PROJECT– II


Gender Distribution
Count
7000
6000
5000
4000
3000
2000
1000
0
male
female
Gender


Age Distribution
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
20
30
40
50
60
70
Age

CORELATION HEATMAP:



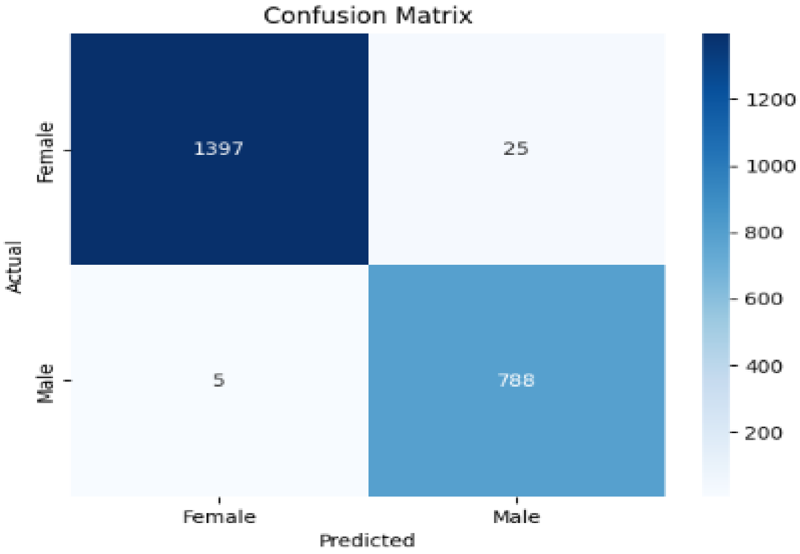
MODEL CLASSIFICATION ON GENDER:

```
Model: "sequential_1"
```

Layer (type)	Output Shape	Param #
sequential (Sequential)	(None, 128)	3,304,640
dense_1 (Dense)	(None, 2)	258

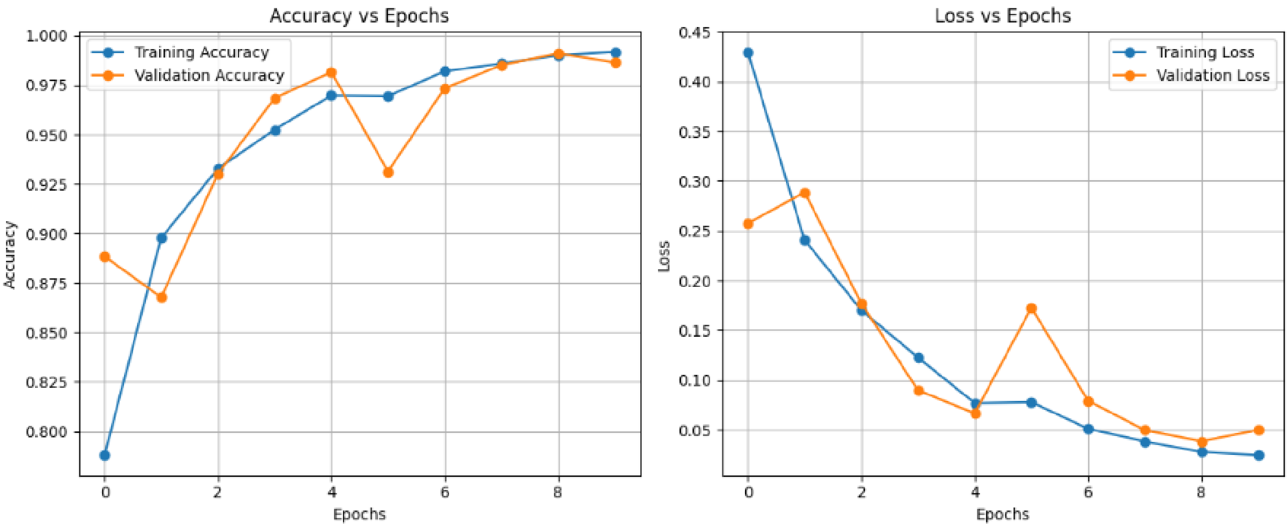
```
Total params: 3,304,898 (12.61 MB)
Trainable params: 3,304,898 (12.61 MB)
Non-trainable params: 0 (0.00 B)
Epoch 1/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 16s 37ms/step - accuracy: 0.7085 - loss: 0.5486 - val_accuracy: 0.8885 - val_loss: 0.2574
Epoch 2/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 6s 21ms/step - accuracy: 0.8893 - loss: 0.2564 - val_accuracy: 0.8677 - val_loss: 0.2887
Epoch 3/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 6s 23ms/step - accuracy: 0.9316 - loss: 0.1767 - val_accuracy: 0.9300 - val_loss: 0.1774
Epoch 4/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 10s 22ms/step - accuracy: 0.9490 - loss: 0.1306 - val_accuracy: 0.9684 - val_loss: 0.0897
Epoch 5/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 6s 20ms/step - accuracy: 0.9692 - loss: 0.0762 - val_accuracy: 0.9815 - val_loss: 0.0665
Epoch 6/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 6s 20ms/step - accuracy: 0.9713 - loss: 0.0704 - val_accuracy: 0.9314 - val_loss: 0.1727
Epoch 7/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 6s 20ms/step - accuracy: 0.9777 - loss: 0.0655 - val_accuracy: 0.9734 - val_loss: 0.0793
Epoch 8/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 11s 23ms/step - accuracy: 0.9821 - loss: 0.0435 - val_accuracy: 0.9851 - val_loss: 0.0496
Epoch 9/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 5s 19ms/step - accuracy: 0.9899 - loss: 0.0289 - val_accuracy: 0.9910 - val_loss: 0.0385
Epoch 10/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 11s 22ms/step - accuracy: 0.9933 - loss: 0.0212 - val_accuracy: 0.9865 - val_loss: 0.0498
```

```
70/70 ━━━━━━━━━━━━━━━━━━━━ 1s 13ms/step
              precision    recall  f1-score   support

      Female       1.00      0.98      0.99      1422
        Male       0.97      0.99      0.98       793

    accuracy                           0.99      2215
   macro avg       0.98      0.99      0.99      2215
weighted avg       0.99      0.99      0.99      2215
```



Model Evaluation (Gender Classification)

- **Overall Accuracy:** 99%
- **Precision / Recall / F1-Score:**
  - **Female:** Precision = 1.00, Recall = 0.98, F1 = 0.99
  - **Male:** Precision = 0.97, Recall = 0.99, F1 = 0.98
- **Support (Samples):** 1422 females, 793 males
- This model is **highly accurate and well-balanced** for classifying gender based on the given features. Let me know if you want to explore what features were used or visualize training performance.



Model Training Overview

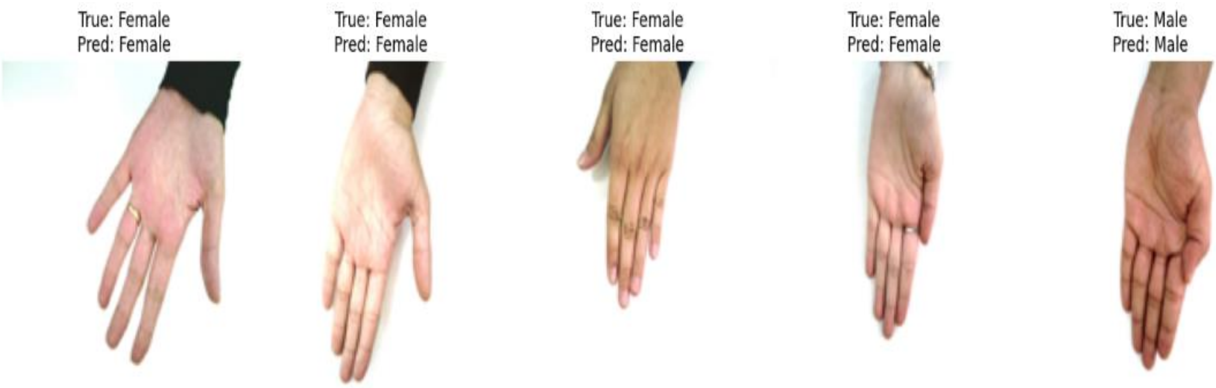
Accuracy vs Epochs (Left Plot)

- **Training Accuracy** steadily improves from ~79% to ~99.5%.
- **Validation Accuracy** follows closely, peaking around 98.5% by epoch 9.
- The model generalizes well with minimal overfitting.

Loss vs Epochs (Right Plot)

- **Training Loss** sharply decreases from 0.43 to near 0.
- **Validation Loss** also declines, with slight fluctuations around epoch 5–6.
- Final validation loss remains low, indicating good convergence.

PREDICTIONS:



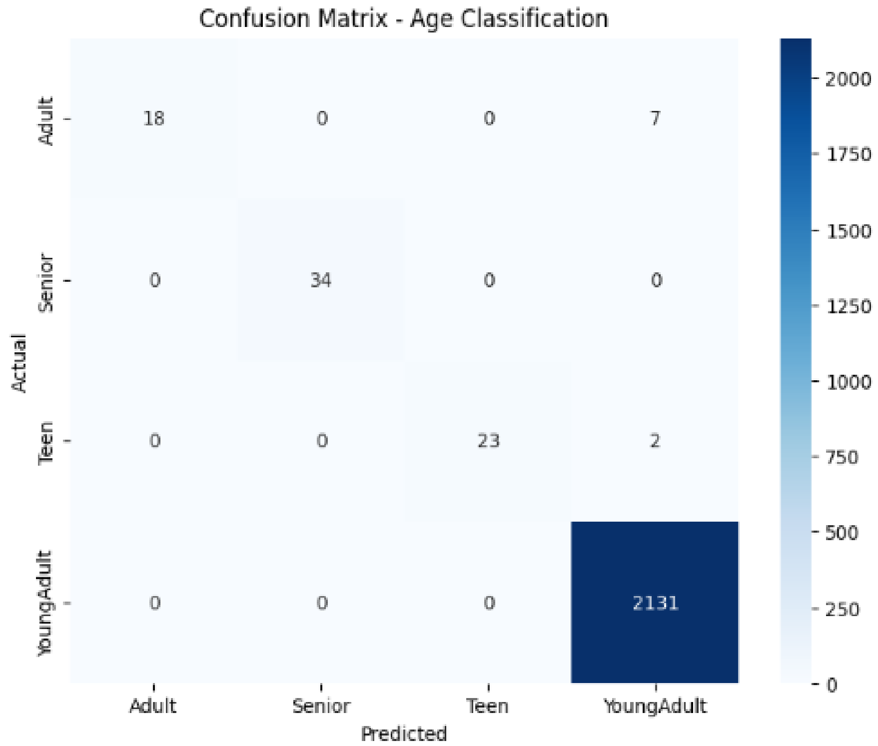
MODEL CLASSIFICATION ON AGE:

```
Model: "sequential_2"
```

Layer (type)	Output Shape	Param #
conv2d_3 (Conv2D)	(None, 126, 126, 32)	896
max_pooling2d_3 (MaxPooling2D)	(None, 63, 63, 32)	0
conv2d_4 (Conv2D)	(None, 61, 61, 64)	18,496
max_pooling2d_4 (MaxPooling2D)	(None, 30, 30, 64)	0
conv2d_5 (Conv2D)	(None, 28, 28, 128)	73,856
max_pooling2d_5 (MaxPooling2D)	(None, 14, 14, 128)	0
flatten_1 (Flatten)	(None, 25088)	0
dense_2 (Dense)	(None, 128)	3,211,392
dropout_1 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 4)	516

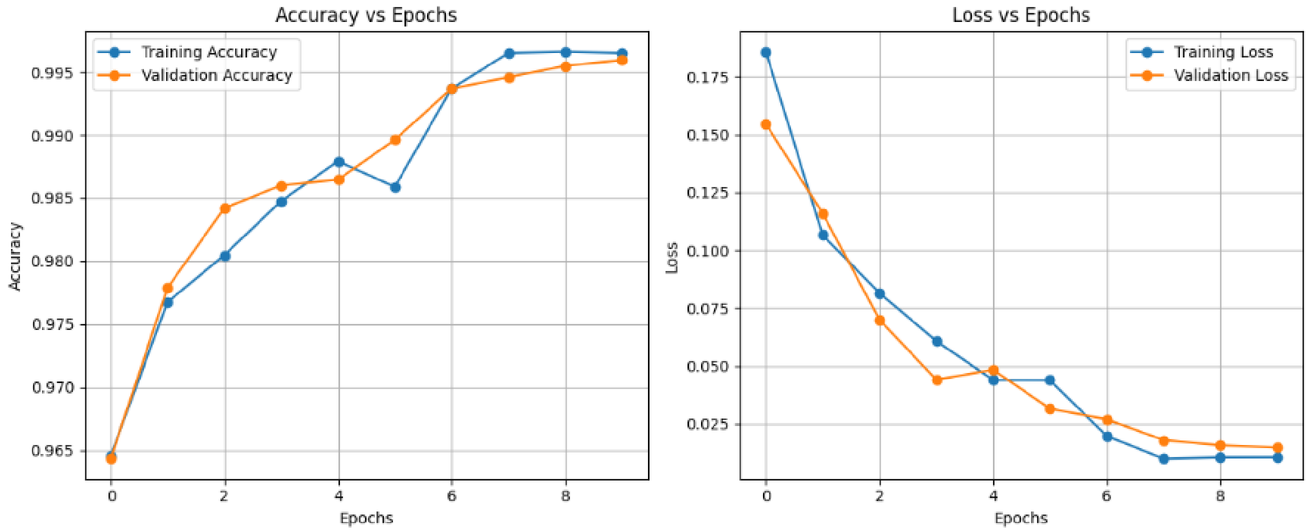
```
Total params: 3,305,156 (12.61 MB)
Trainable params: 3,305,156 (12.61 MB)
Non-trainable params: 0 (0.00 B)
```

```
Epoch 1/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 15s 35ms/step - accuracy: 0.9518 - loss: 0.2763 - val_accuracy: 0.9643 - val_loss: 0.1546
Epoch 2/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 5s 19ms/step - accuracy: 0.9720 - loss: 0.1273 - val_accuracy: 0.9779 - val_loss: 0.1160
Epoch 3/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 10s 19ms/step - accuracy: 0.9811 - loss: 0.0833 - val_accuracy: 0.9842 - val_loss: 0.0700
Epoch 4/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 10s 19ms/step - accuracy: 0.9849 - loss: 0.0595 - val_accuracy: 0.9860 - val_loss: 0.0440
Epoch 5/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 10s 19ms/step - accuracy: 0.9879 - loss: 0.0421 - val_accuracy: 0.9865 - val_loss: 0.0483
Epoch 6/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 10s 18ms/step - accuracy: 0.9871 - loss: 0.0352 - val_accuracy: 0.9896 - val_loss: 0.0315
Epoch 7/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 5s 19ms/step - accuracy: 0.9929 - loss: 0.0208 - val_accuracy: 0.9937 - val_loss: 0.0269
Epoch 8/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 5s 18ms/step - accuracy: 0.9971 - loss: 0.0092 - val_accuracy: 0.9946 - val_loss: 0.0179
Epoch 9/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 5s 19ms/step - accuracy: 0.9985 - loss: 0.0064 - val_accuracy: 0.9955 - val_loss: 0.0157
Epoch 10/10
277/277 ━━━━━━━━━━━━━━━━━━━━ 10s 19ms/step - accuracy: 0.9986 - loss: 0.0057 - val_accuracy: 0.9959 - val_loss: 0.0147
```



Key Insights

- **YoungAdult** is classified **extremely accurately** (2131/2131 correct).
- Minor misclassifications occur:
    - 7 Adults predicted as YoungAdults
    - 2 Teens predicted as YoungAdults
- **Overall**, the model performs **very well**, especially for the dominant **YoungAdult** class.



Training Summary (Model Accuracy & Loss)

Accuracy vs Epochs (Left Plot)

- **Training Accuracy** improves from ~96.5% to over **99.6%**.
- **Validation Accuracy** closely follows, reaching **~99.6%** by epoch 9.
- Strong alignment between training and validation accuracy indicates **excellent generalization**.

Loss vs Epochs (Right Plot)

- **Training Loss** drops sharply from ~0.18 to ~0.01.
- **Validation Loss** also decreases consistently, mirroring training loss.
- Very low final loss values → **minimal error** and **well-fit model**.

PREDICTIONS:

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 562ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 31ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 31ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 31ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 33ms/step
```

Z-TEST,T-TEST,ANOVA TEST RESULTS:

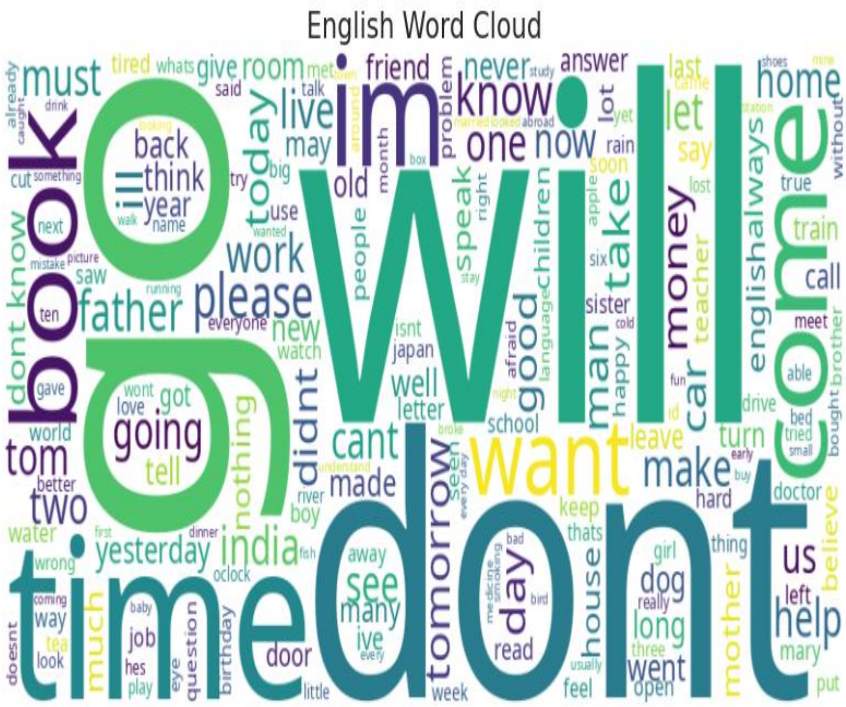
```
Z-test: Z-score = 219.1321, p-value = 0.0000
Null hypothesis is rejected
T-test: T-statistic = 1326.9161, p-value = 0.0000
Null hypothesis is rejected
ANOVA: F-statistic = 0.8384, p-value = 0.3599
Null hypothesis is accepted
```

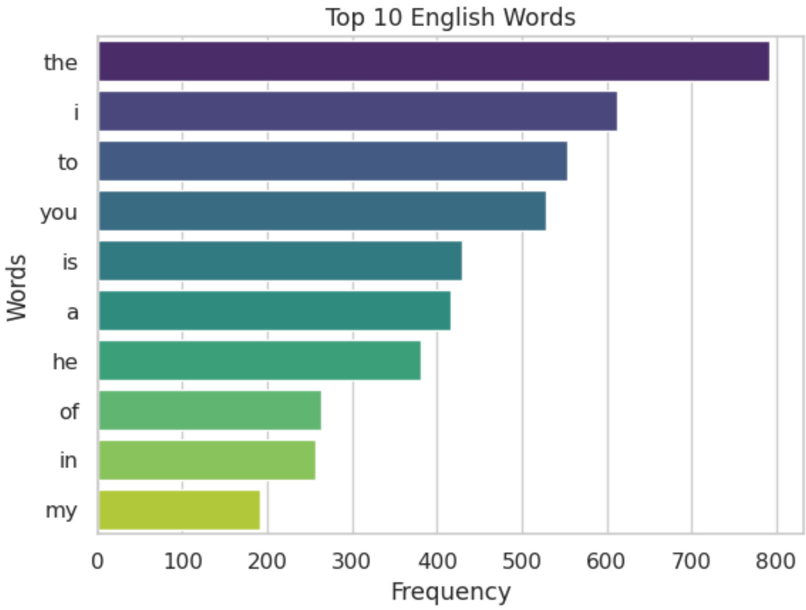
→ There is a **clear difference** between the two sample means being compared.
→ The **null hypothesis is rejected** in both cases, indicating a **strong effect or change** exists.

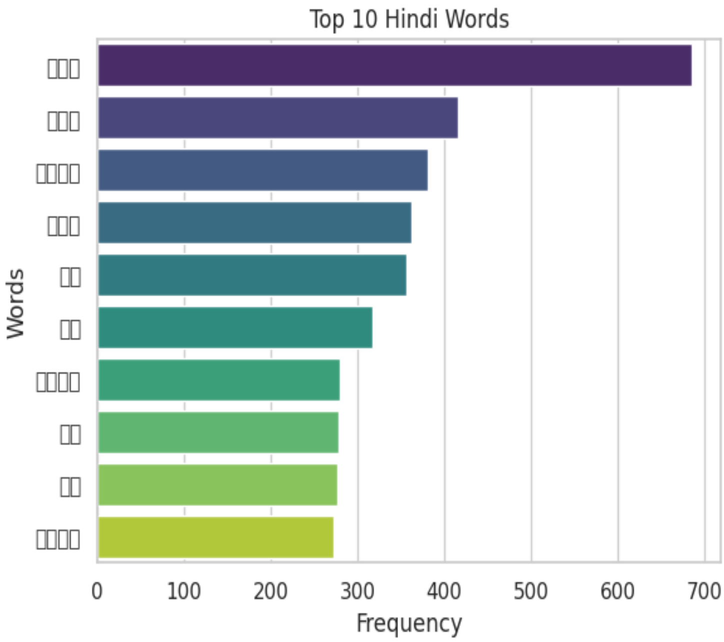
- The **p-value is greater than 0.05**, so the **null hypothesis is accepted**.
  → This suggests that **there are no significant differences among the group means** being tested.

PROJECT- III

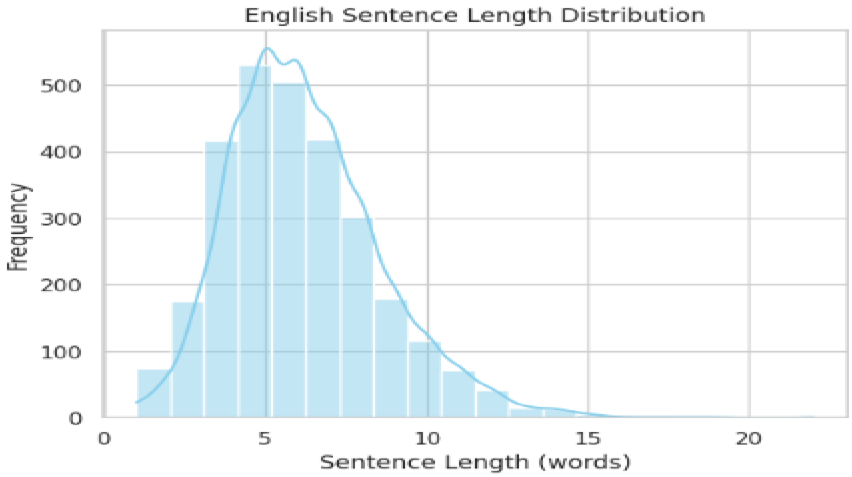
	english	hindi	english_clean	hindi_clean	english_tokens	hindi_tokens
0	Help!	बचाओ!	help	बचाओ	[help]	[बचाओ]
1	Jump.	उछलो.	jump	उछलो	[jump]	[उछलो]
2	Jump.	कूदो.	jump	कूदो	[jump]	[कूदो]
3	Jump.	छलांग.	jump	छलांग	[jump]	[छलांग]
4	Hello!	नमस्ते।	hello	नमस्ते।	[hello]	[नमस्ते।]

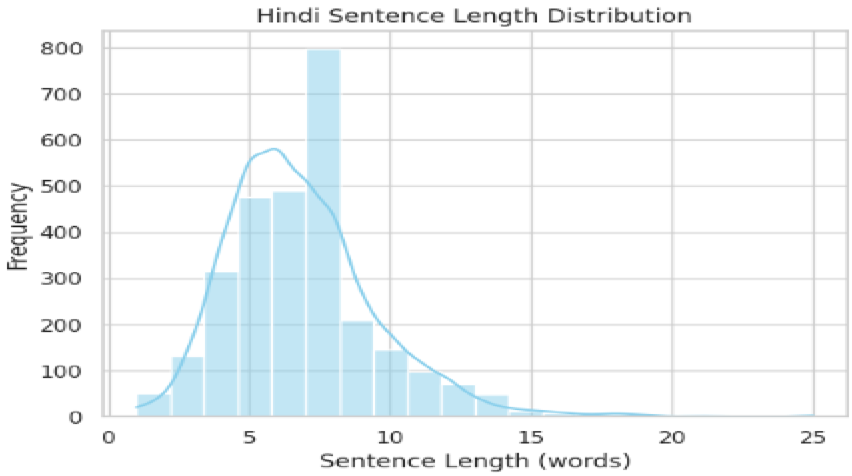



Top 10 English Words
the
i
to
you
is
a
he
of
in
my
Words
0
100
200
300
400
500
600
700
800
Frequency


Top 10 Hindi Words
Words
0
100
200
300
400
500
600
700
Frequency

HISTOGRAMS:





The image shows two histograms comparing sentence length distributions for English and Hindi sentences:

- **English Sentence Length Distribution** (top): Most sentences are between 4 to 8 words long, with a peak around 5 words. The distribution is right-skewed, indicating fewer longer sentences.
- **Hindi Sentence Length Distribution** (bottom): The distribution peaks sharply at around 8 words and is also right-skewed, though slightly more spread out than the English one.

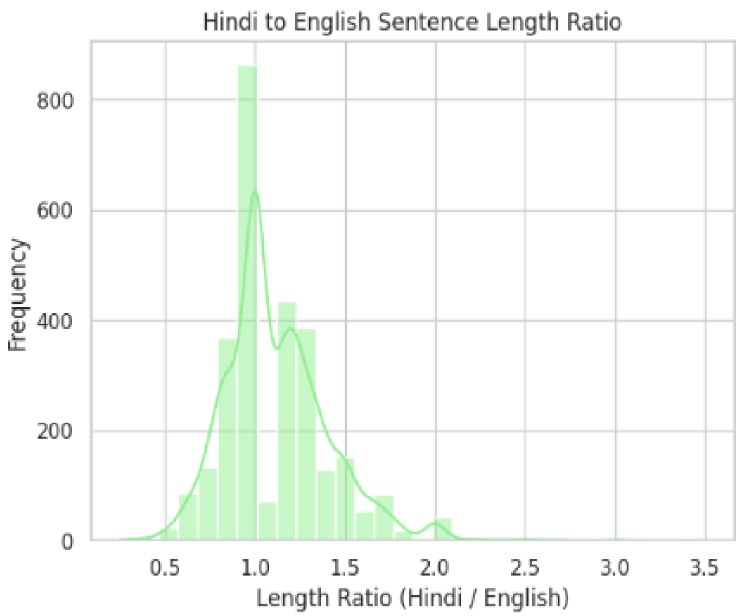
PARTS OF SPEECH TAG RESULTS:

```
Sentence 1 POS Tags:
[('hello', 'NN')]

Sentence 2 POS Tags:
[('cheers', 'NNS')]

Sentence 3 POS Tags:
[('cheers', 'NNS')]

Sentence 4 POS Tags:
[('got', 'VBD'), ('it', 'PRP')]

Sentence 5 POS Tags:
[('im', 'NN'), ('ok', 'NN')]
```

The image shows the **Part-of-Speech (POS) tags** for five sentences. Each token is paired with its corresponding POS tag:

1. **Sentence 1**: `hello` is tagged as **NN** (noun, singular).
2. **Sentence 2 & 3**: `cheers` is tagged as **NNS** (plural noun).
3. **Sentence 4**: `got` is **VBD** (verb, past tense), and `it` is **PRP** (personal pronoun).
4. **Sentence 5**: `im` and `ok` are both tagged as **NN** (singular nouns), though `im` might be a misspelling or informal contraction of “I’m”.

This suggests basic tokenization and POS tagging, possibly using a standard NLP toolkit like NLTK or spaCy.



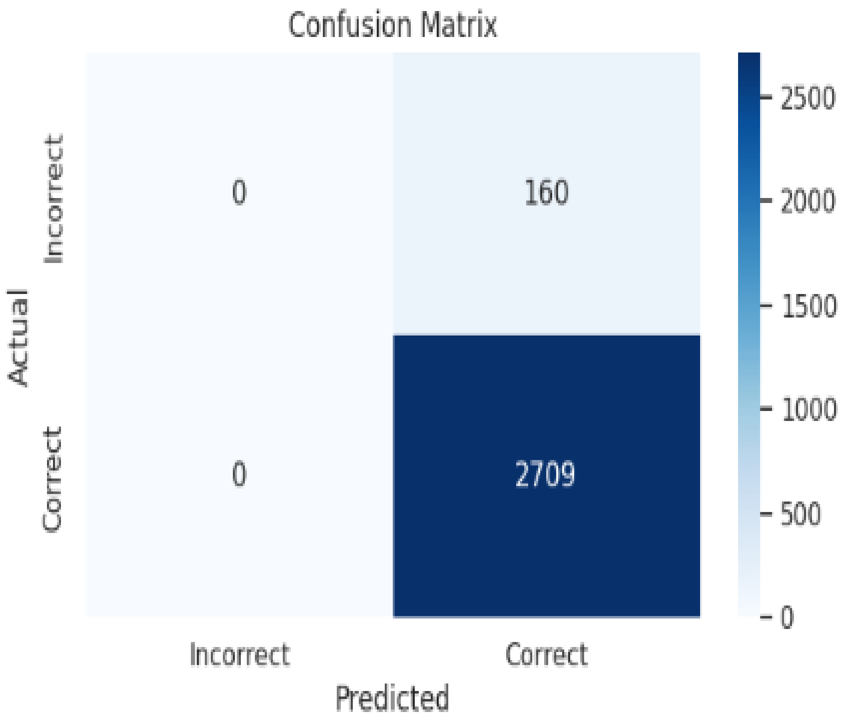
Key Points:

- **X-axis (Length Ratio)**: Represents the ratio of sentence lengths in Hindi to English (Hindi words divided by English words).
- **Y-axis (Frequency)**: Number of sentence pairs that fall into each length ratio bin.

Evaluation metrics based on predicted lables and actual labels :

```
📊 Evaluation Metrics:
✅ Accuracy:  0.9442
✅ Precision: 0.9442
✅ Recall:    1.0000
✅ F1 Score:  0.9713
```

CONFUSION MATRIX:



Confusion Matrix Breakdown:

	Predicted: Incorrect	Predicted: Correct
Actual: Incorrect	0	160
Actual: Correct	0	2709

CONCLUSION:

This project report covers three machine learning tasks: predicting drug reimbursements using regression (Project 1), classifying hand images by gender and age with CNNs (Project 2), and analyzing English-Hindi sentence pairs for semantic similarity (Project 3). Among the models used, the Random Forest Regressor performed best in Project 1, while Project 2 achieved exceptional accuracy (~99%) in image classification. Project 3 provided valuable insights for bilingual NLP. Overall, the work reflects strong skills in data processing, model development, and evaluation across varied data types.