1. Project Description

This project focuses on developing a machine learning model specifically designed to classify the gender of names from Myanmar . Myanmar names present unique challenges for gender classification due to their distinctive structure and diverse influences from various ethnic groups and languages. Unlike Western names, Myanmar names do not typically include surnames and often do not adhere to a consistent pattern that easily reveals gender .

2. Objective

The objective of this project is to build a robust machine learning model capable of classifying Myanmar names by gender with high accuracy. This model will leverage a dataset of Myanmar names labelled by gender,

3. Dataset

The dataset from (https://burmesenames.wordpress.com/) consists of 5323 Myanmar names, of which 2303 count 43.26% are female names and 3020 count 56.74% are male.

`

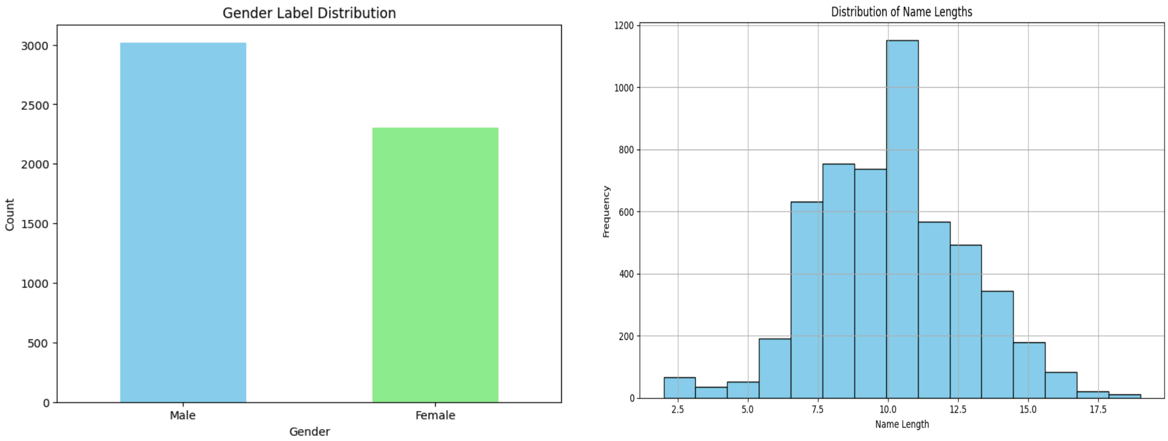


Figure 1:(a) Labels distribution and (b) Name length

4. Data Preprocessing

a. Data Cleaning:We have removed duplicate records from the dataset.

b. Text Normalisation :Used Lambda function to change name for lowercase .

Name Segmentation Breaking names into meaningful components (first name , middle name , last name) provides more granular information but when we run model accuracy is reduced and unable to be used in the final model.

5. Model Selection

We evaluate four machine learning models for the task of gender classification based on Myanmar names. The models considered are:

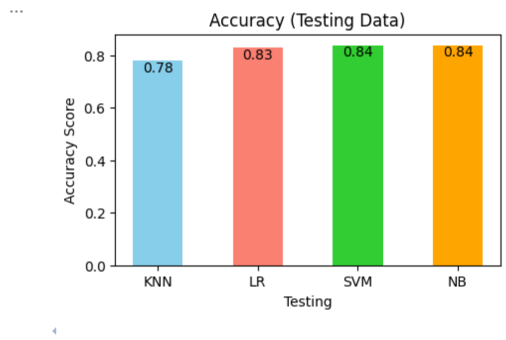
- K-Nearest Neighbors (KNN)
- Logistic Regression
- Support Vector Classifier (SVC)
- Naive Bayes

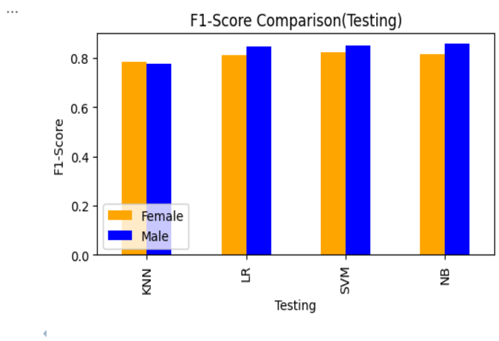
6. Model Evaluation Metrics

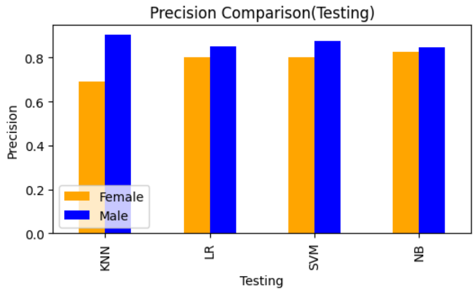
Define the metrics used to evaluate model performance:

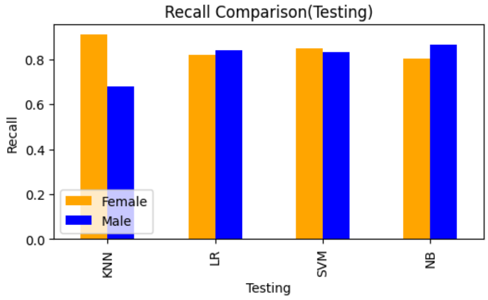
a. Accuracy: Overall correctness of predictions.
b. Precision and Recall: Trade-offs between correctly predicting each gender class.
c. F1 Score: Harmonic mean of precision and recall.

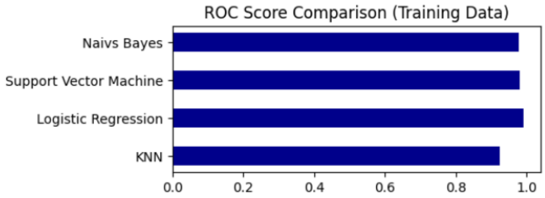
7. Model Performance

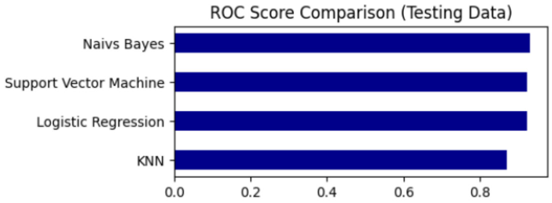












8. Discussion

a. Potential Improvements

i. Data related improvement : Collecting more data , especially from Myanmar different ethnic groups to cover their unique naming and naming style of different generations can be added into the dataset .

ii. Model Specific Improvement : Use Ensemble method ,combining multiple models in order to improve the accuracy.

iii. Model Interpretation and Fairness : For Logistic regression we would like to carry out Feature Importance Analysis to understand which parts of the name contribute most to the prediction.

9. Conclusion

In this report, we evaluated the performance of four machine learning algorithms—K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Classifier (SVC), and Naive Bayes—for the task of gender classification based on Myanmar names. Our analysis found that Logistic Regression , SVC and Naive Bayes results are approximately in the same range and KNN is inferior on Accuracy and F1 score .