Gender Identification from Facial Features

Xiang Liu, Nikolay Sizov, Xingyu Jiang

CMPSC 441 Spring 2024

Artificial Intelligence

Md Hossain

21 April 2024

Table of Contents

[Team Members 2](#_Toc1327566448)

[Introduction 2](#_Toc1709199510)

[Problem Description 2](#_Toc174698142)

[Objective 2](#_Toc1262261661)

[Datasets 2](#_Toc1443242538)

[Data Source and format 3](#_Toc844365241)

[Data example 4](#_Toc655561247)

[Preprocessing Data 5](#_Toc657791932)

[Features 6](#_Toc1068235763)

[Feature details 6](#_Toc2016762538)

[Feature Extraction Process 8](#_Toc491105615)

[Experimental Details 9](#_Toc1132146047)

[Diagram: Framework 9](#_Toc107460271)

[Data Division: Training and Testing 11](#_Toc774849130)

[Classifiers: Implementation and Tuning 11](#_Toc834880773)

[Evaluation metrics 12](#_Toc1679828855)

[Result and Analysis 12](#_Toc376348285)

[Results 13](#_Toc1484421)

[Analysis of the results 13](#_Toc1674447859)

[Conclusion 13](#_Toc1736083355)

[Discussing limitation 13](#_Toc1416187195)

[Future Direction 14](#_Toc1436788053)

[Appendix 14](#_Toc1114830746)

## **Team Members**

* Xiang Liu – initial data understanding and exploration, data preprocessing, code quality check, development of data classification reporting.
* Nikolay Sizov – responsible for classification algorithms, training and. accuracy developments.
* Xingyu Jiang – responsible for data extraction, data preprocessing, classification models’ hyperparameters tuning.

## **Introduction**

### Problem Description

In this project, we aim to develop a system capable of identifying a person's gender from their facial image. The process begins by using machine learning algorithms to create distinct templates that capture the typical facial features of males and females. These templates serve as a benchmark for comparison.

Once a new facial image is input into the system, it is analyzed and compared against these predefined templates. If the facial characteristics of the input image align more closely with the male template, the system classifies the individual as male. Conversely, if the image aligns more closely with the female template, the individual is classified as female.

### Objective

Utilizing the AR face database, which features facial images of 136 individuals. Each image is annotated with 22 distinct facial landmarks, captured under various expressions and lighting conditions. For our purpose we are to implement, and test multiple machine learning algorithms sourced from the Scikit-learn library. Our main goal is to evaluate these classifiers' effectiveness to determine which performs best in accurately predicting gender from facial analysis. Ultimately, we aim to develop a robust system that utilizes these insights to enhance gender classification based on facial features.

**Datasets**

### Data Source and format

The utilized AR face database is sourced from public domain, it consists of images from 76 males and 60 females. Each image includes manual annotations of 22 distinct facial points. These images are presented in frontal views and are varied by facial expressions and different lighting settings, enhancing the depth of our analysis. The annotations in these images are formatted as follows:

* Images of males: m-xx-yy.pts
* Images of females: w-xx-yy.pts

Here, xx stands for a unique number identifying an individual, ranging from "00" for the first individual to "76" for males and "60" for females, while yy indicates the type of expression or lighting condition:

1. Neutral expression
2. Smile
3. Anger
4. Left side lighting

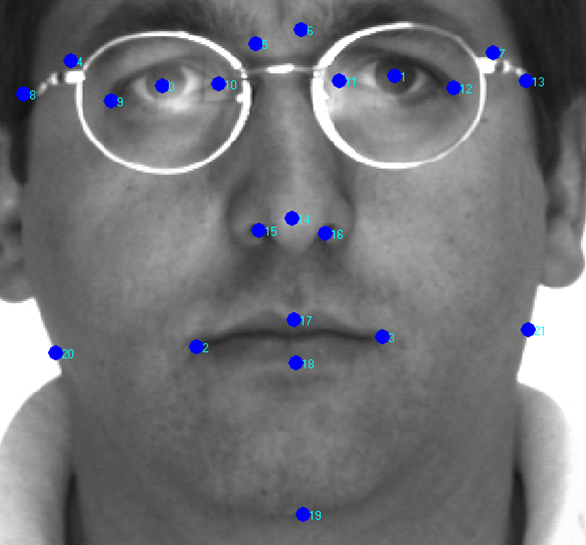
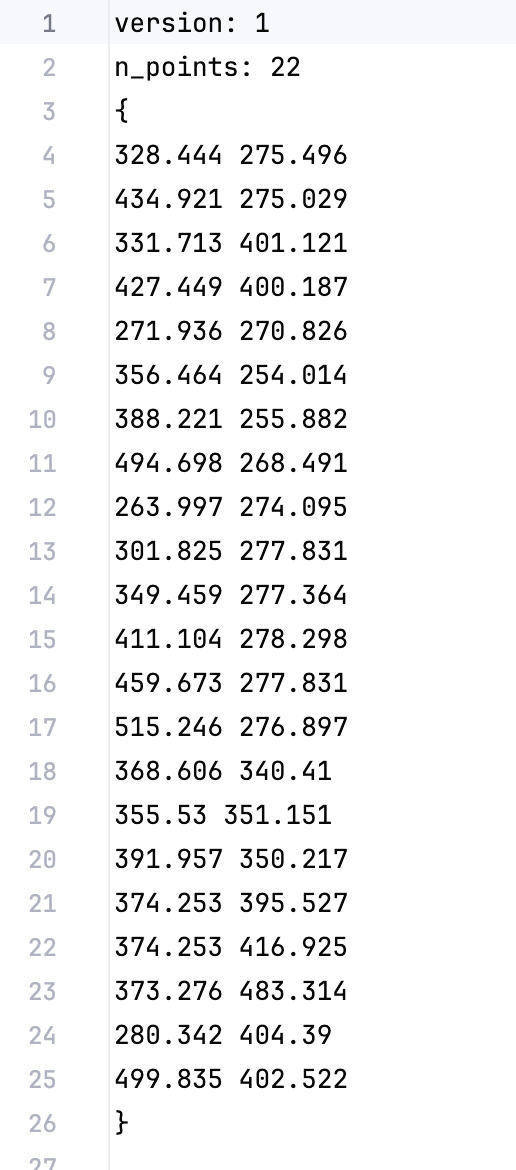


Figure 1: A sample face image marked with the 22 points.

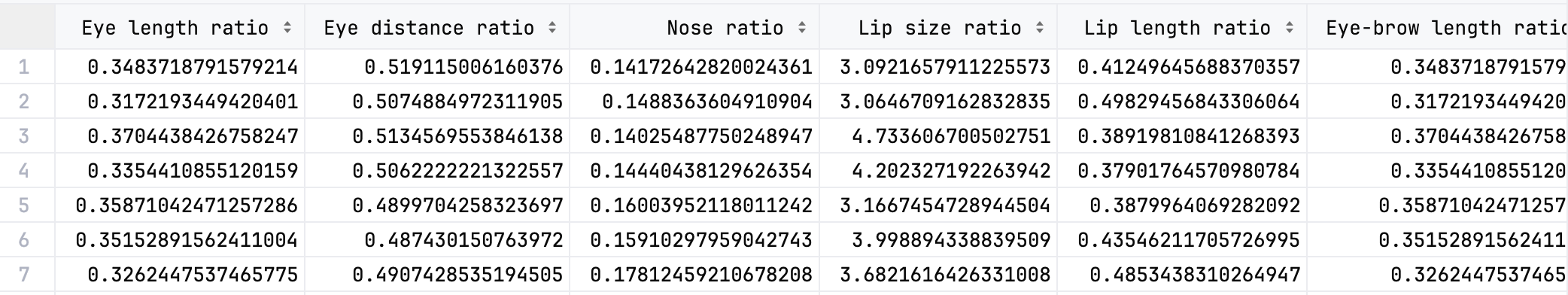
### Data example

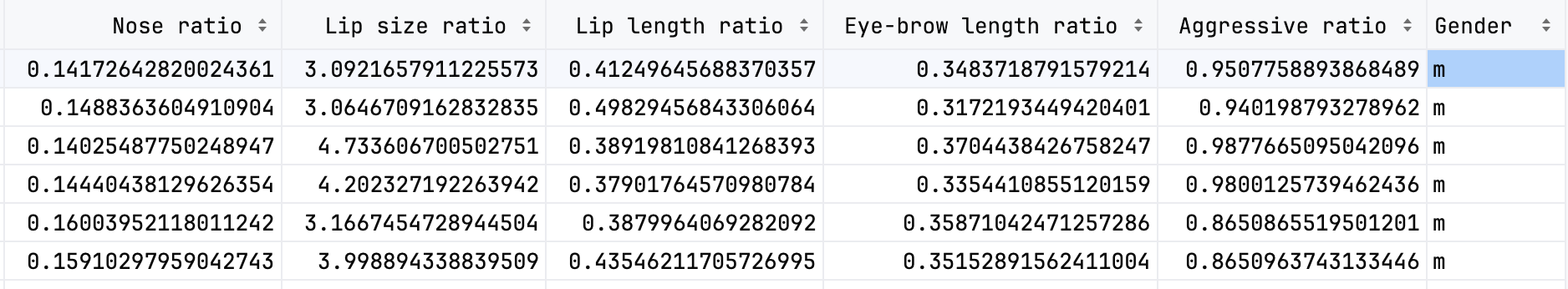
Each file captures the 22 facial landmarks through (X, Y) coordinates, detailing the precise position of each feature on the face. Below is an example of the data format:



Preprocessing Data (For example, min-max or z-score normalization)

Preprocessed data example:





**Features**

### Feature details

A set of key facial features are defined and extracted, based on precise measurements between specific anthropometric points annotated on each face image. These features are designed to capture distinctive aspects of facial geometry that are potentially indicative of gender differences. A detailed description of each feature is provided below:

1. Eye Length Ratio
   * Definition: The maximum eye length (either left or right eye) divided by the distance between points 8 and 13.
   * Purpose: This ratio helps gauge the relative size of the eyes in relation to the width of the face, which can vary between genders.
2. Eye Distance Ratio
   * Definition: The distance between the centers of the two eyes divided by the distance between points 8 and 13.
   * Purpose: This measures the spacing of the eyes, another facial feature that can differ significantly between males and females.
3. Nose Ratio
   * Definition: The distance between points 15 and 16 over the distance between points 20 and 21.
   * Purpose: This ratio examines the width of the nose relative to the width of the mouth area, helping to highlight facial proportion differences.
4. Lip Size Ratio
   * Definition: The horizontal distance between points 2 and 3 over the vertical distance between points 17 and 18.
   * Purpose: This captures the thickness and width of the lips, which are prominent gender-differentiating features.
5. Lip Length Ratio
   * Definition: The distance between points 2 and 3 over the distance between points 20 and 21.
   * Purpose: This ratio assesses the width of the lips relative to the lower face width, important for distinguishing facial structure variations.
6. Eye-brow Length Ratio
   * Definition: The longer distance (either between points 4 and 5 or points 6 and 7) over the distance between points 8 and 13.
   * Purpose: Eyebrow length relative to the upper face width can provide insights into typical gender-specific facial features.
7. Aggressive Ratio
   * Definition: The distance between points 10 and 19 over the distance between points 20 and 21.
   * Purpose: Although termed "aggressive," this ratio may correlate with perceived facial expression intensity, which can vary by gender and cultural perceptions.
8. Jaw-line Ratio
   * Definition: The distance between nose tip and chin of point 14 and 19, divided by the horizontal distance across the lower face between points 20 and 21.
   * Purpose: This ratio measures the prominence and width of the jawline relative to the lower face, which can vary distinctly between genders and is a key indicator of facial form.
9. Jaw Horizontal length ratio
   * Definition: The horizontal distance between points 2 and 3 over the distance from the nose tip (point 14) to the chin (point 19)
   * Purpose: The ratio between lip length and jaw length can vary distinctly between genders as well
10. Nose Tip to Upper Lip Distance
    * Definition: The distance from the nose tip (point 14) to the upper lip (point 17).
    * Purpose: This measurement highlights the vertical spacing between the nose and the lips, which can reflect gender-specific facial proportions and is crucial for understanding the overall facial harmony.
11. Lower Lip to Chin Distance
    * Definition: The distance from the lower lip (point 18) to the chin (point 19).
    * Purpose: This ratio evaluates the length of the lower facial region below the lips, an area that often exhibits significant differences between male and female facial structures in terms of length and proportion.
12. Nose tip to chin Distance
    * Definition: The distance from the nose tip (point 14) to the chin (point 19)
    * Purpose: This is the length of the nose tip to chin, an area that often has significant differences between male and female facial structures.
13. Lip ratio
    * Definition: The horizontal distance between points 2 and 3 over horizontal distance across the lower face between points 20 and 21.
    * This ratio measures the prominence and width of the jawline relative to the lower face, which can vary distinctly between genders and is a key indicator of facial form.

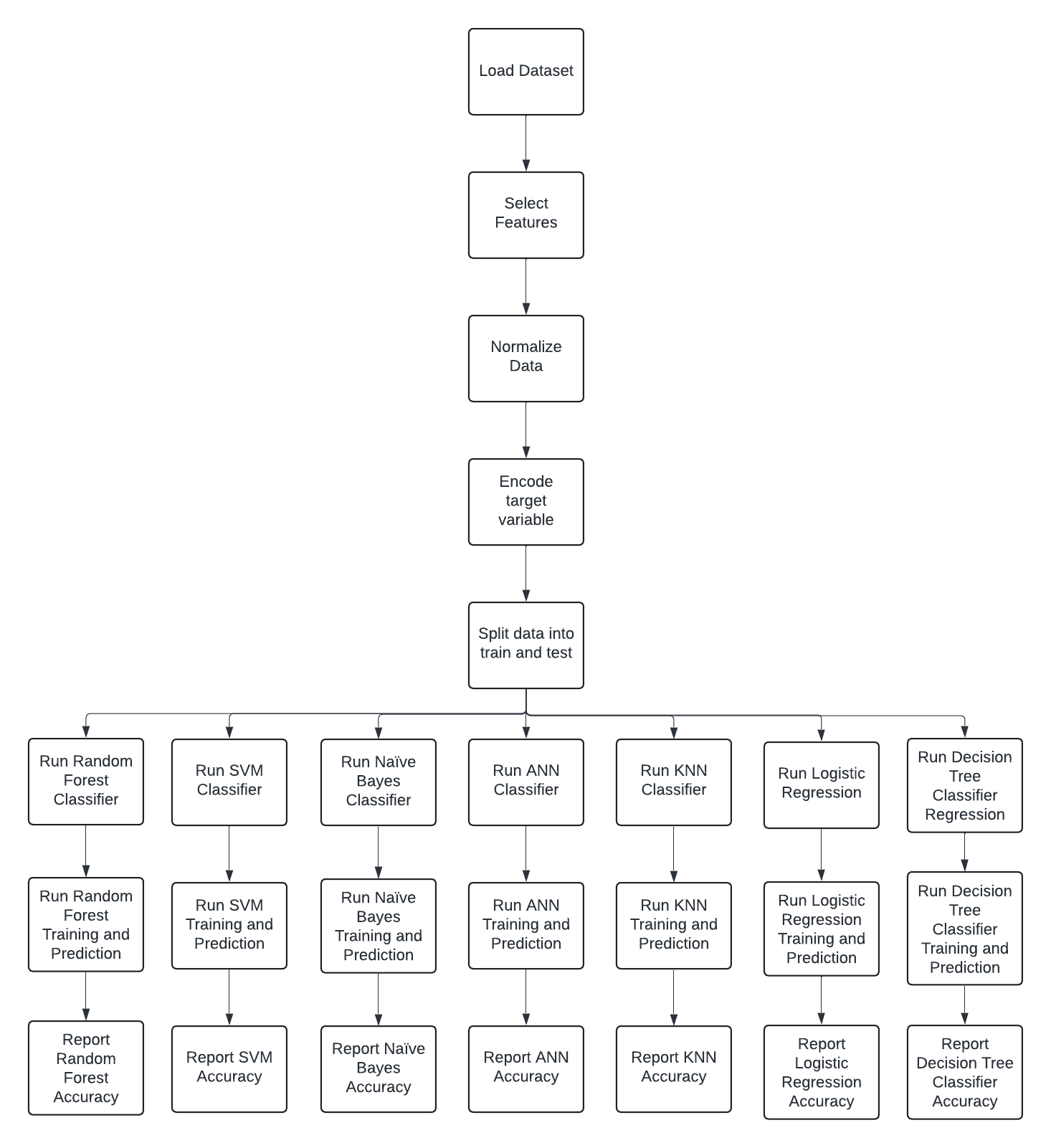
### Feature Extraction Process

For each of the above features, distances are computed using the Euclidean distance formula, which calculates the straight-line distance between two points in Euclidean space. This measurement is critical in maintaining the geometric integrity of the facial features across different images. The calculated ratios and distances are then compiled to form a comprehensive feature set that feeds into the machine learning models for gender classification.

**Experimental Details**

### Diagram: Framework

The diagram below outlines the program flow, from data loading to the application of multiple classification algorithms. This flowchart provides an overview of the steps involved in our experimental setup, including data preprocessing, model training, and the evaluation phase. Each stage is designed to ensure a systematic approach to handling and analyzing the data, maximizing the accuracy and efficiency of our classification results.



First, when we select seven features which are (Eye Length Ratio, Eye Distance Ratio, Nose Ratio, Lip Size Ratio, Lip Length Ratio, Eye-brow Length Ratio, and Aggressive Ratio). We to trained and tested with 4 different classifiers, we get have highest accuracy of 71.57% of logistic regression classifier:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision (Men) | Recall (Men) | F1-Score (Men) | Precision (Female) | Recall (Female) | F1-Score (Female) |
| Logistic Regression Classifier | 71.57% | 73% | 84% | 78% | 68% | 53% | 59% |

After we add one more feature “Jaw-line Ratio” when training the model, we get better results of accuracy of 74.47% from K-Nearest Neighbors (KNN) classifier. We continue to add one more feature “Nose Tip to Upper Lip Distance” it was able to increase by 6 % from KNN classifier again:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision (Men) | Recall (Men) | F1-Score (Men) | Precision (Female) | Recall (Female) | F1-Score (Female) |
| K-Nearest Neighbors (KNN) | 80.39% | 80% | 90% | 85% | 81% | 65% | 72% |

Extend to add in one more additional feature which is “Lower Lip to Chin Distance” we get a best accuracy of 88% from Artificial Neural Network (ANN) classifier. We have tried to use different techniques to improve the performance of all the models, but it did not help. The ANN classifier performed the best with the three additional features:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision (Men) | Recall (Men) | F1-Score (Men) | Precision (Female) | Recall (Female) | F1-Score (Female) |
| Artificial Neural Network (ANN) | 88.24% | 87% | 95% | 91% | 91% | 78% | 84% |

Despite 88% is already a great accuracy we continued test our experiment by add two more feature of “Jaw Horizontal Length Ratio” and “Lip Ratio” we were able to increase the accuracy by 5% which comes to 93.14%:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision (Men) | Recall (Men) | F1-Score (Men) | Precision (Female) | Recall (Female) | F1-Score (Female) |
| Artificial Neural Network (ANN) | 93.14% | 90% | 100% | 95% | 100% | 82% | 90% |

### Data Division: Training and Testing

We divided our data into training and testing, with proportions of 80 % and 20% from preprocessed data.

### Classifiers: Implementation and Tuning

* K-Nearest Neighbors (KNN):
  + Implemented using the KNeighborsClassifier from scikit-learn.
  + The number of neighbors is set to 5 (n\_neighbors=5).
* Artificial Neural Network (ANN):
  + Implemented using the MLPClassifier from scikit-learn.
  + The hidden layer sizes are set to (100,), indicating a single hidden layer with 100 neurons.
  + The maximum number of iterations is set to 1000 (max\_iter=1000).
* Naïve Bayes:
  + Implemented using the GaussianNB classifier from scikit-learn.
* Logistic Regression:
  + Implemented using the LogisticRegression classifier from scikit-learn.
  + The maximum number of iterations is set to 1000 (max\_iter=1000).
* Decision Tree Classifier:
  + Implemented using the DecisionTreeClassifier from scikit-learn.
* Support Vector Machine (SVM):
  + Implemented using the SVC classifier from scikit-learn.
  + The kernel is set to 'linear', indicating a linear SVM.
* Random Forest Classifier:
  + Implemented using the RandomForestClassifier from scikit-learn.
  + The number of trees in the forest is set to 100 (n\_estimators=100).

### Evaluation metrics

In our project, we evaluated the gender classification model using the following metrics:

1. Precision: Indicates the proportion of positive identifications that were correct. A higher precision relates to a lower false-positive rate.
2. Recall (Sensitivity): Measures the proportion of positives identified correctly. It signifies the model's ability to find all relevant cases within a dataset.
3. F1-Score: Combines precision and recall into a single metric by taking their harmonic mean. It gives a balance between the precision and recall measurements.
4. Support: The number of actual occurrences of the class in the specified dataset. For our evaluation, the support is the number of faces for each gender that the model attempted to classify.

**Result and Analysis**

### Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | Precision (M) | Recall (M) | F1-Score (M) | Precision (F) | Recall (F) | F1-Score (F) |
| K-Nearest Neighbors | 80.39% | 83% | 85% | 84% | 76% | 72% | 74% |
| Artificial Neural Network | 93.14% | 90% | 100% | 95% | 100% | 82% | 90% |
| Naïve Bayes | 78.43% | 83% | 81% | 82% | 71% | 75% | 73% |
| Logistic Regression | 90.20% | 86% | 100% | 93% | 100% | 75% | 86% |
| Decision Tree | 70.59% | 74% | 79% | 77% | 64% | 57% | 61% |
| Support Vector Machine | 91.18% | 87% | 100% | 93% | 100% | 78% | 87% |
| Random Forest | 81.37% | 82% | 89% | 85% | 80% | 70% | 75% |

### Analysis of the results

The classification performance of each machine learning model was evaluated based on accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of how well each classifier performed. In the context of this project, '0' represents males and '1' represents females.

The ANN model outperformed all other models with the highest accuracy of 88.24%, indicating a strong fit for this classification task. It showed high precision and recall for both classes, indicating its balanced classification capability

**Conclusion**

### Discussing limitation

**Biological Variation and Overlap**

Gender is not always clearly delineated by physical features alone. There is considerable overlap in the facial measurements between genders, which can lead to misclassification errors. Models that overly rely on these measurements may struggle to achieve high accuracy because of the biological and genetic diversity within and across gender groups.

**Cultural and Demographic Variability**

Facial features can vary significantly across different ethnicities and populations. If the training data is not representative of this diversity, the model may perform well only on the demographic like that of the training set, resulting in biased predictions when applied to other groups.

**Changes Over Time**

Features like facial measurements can change over time due to aging, health changes, or other factors. A model trained on data from a particular age group might not perform well when applied to another age group.

**Data Quality and Quantity**

High-quality, annotated data is crucial for training accurate models. Insufficient or poor-quality data can lead to models that are less accurate and generalize poorly to real-world conditions.

**Feature Selection and Engineering**

Selecting the right features and engineering them appropriately is critical. Misjudgments in these areas can lead to models that either capture irrelevant variability or miss important subtleties, affecting overall performance.

**Algorithm Choice and Tuning**

The choice of algorithm and its tuning is crucial. Some algorithms might not handle the subtlety and complexity of the task well, especially if the feature space is highly dimensional or if the relationship between features and gender is non-linear.

### Future Direction

Moving forward, we aim to refine our approach by delving into more complex facial measurements and exploring automated feature extraction techniques. We will also enhance our dataset variety to strengthen the model's reliability. We would also like to improve our model so that it can adapt to various groups and settings, while also employing sophisticated machine learning tactics to improve its accuracy.

## **Appendix**

Dataset of feature: <https://github.com/xiangliu1123/Gender-Identification/blob/main/all_csv/preProcess_df.csv>

Source Code: <https://github.com/xiangliu1123/Gender-Identification/tree/main/src>