Gender Classification Using Machine Learning Models

This presentation explores the use of predictive modeling in R for gender classification, comparing the performance of Naive Bayes, Random Forest, and SVM models.

CY **by Christelle Younan**

Project Overview

Objective

Compare predictive models for gender classification.

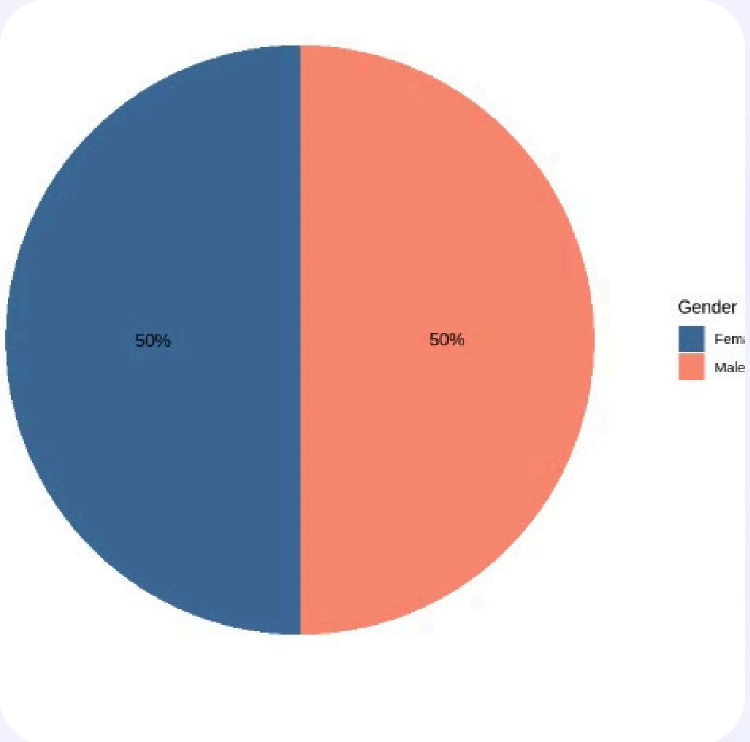
Dataset

Gender Classification Dataset with facial descriptive features.

Methods

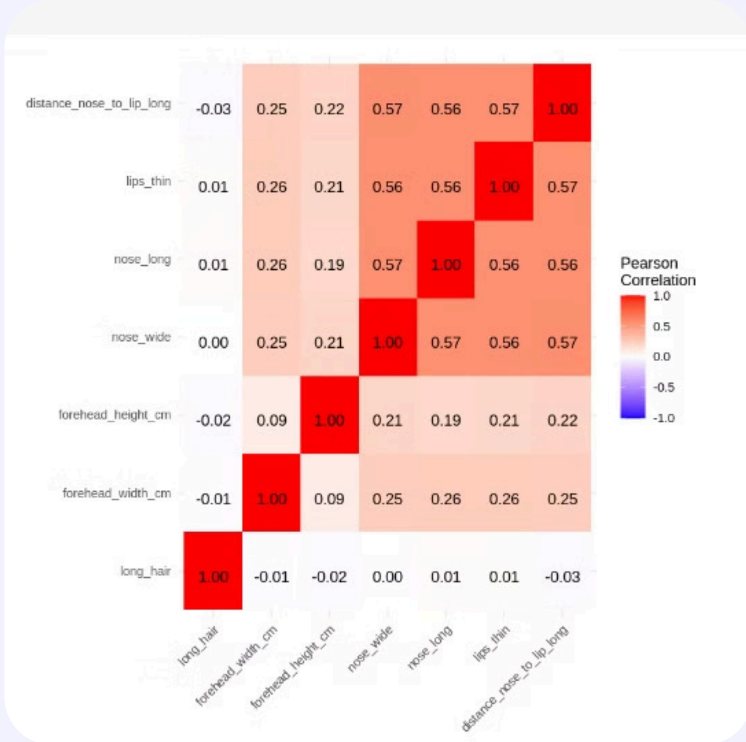
EDA, preprocessing, modeling, performance comparison, and hypothesis testing.

Exploratory Data Analysis



Gender Balance

Dataset is balanced (50% Male, 50% Female).



Correlation Matrix

Relationship between features.

Modeling Approaches

1

Naive Bayes

A probabilistic model based on Bayes' theorem.

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
   Female      Male
0.4483958 0.5516042

Conditional probabilities:
        long_hair
Y              [,1]      [,2]
  Female 0.8155172 0.3880447
  Male   0.8346181 0.3716551

        forehead_width_cm
Y             [,1]      [,2]
  Female 12.82724 0.8704819
  Male   13.55032 1.1868752

        forehead_height_cm
Y             [,1]      [,2]
  Female 5.801034 0.4268984
  Male   6.090119 0.5990471

        nose_wide
Y              [,1]      [,2]
  Female 0.1715517 0.3771530
  Male   0.8304135 0.3754007

        nose_long
Y              [,1]      [,2]
  Female 0.2068966 0.4052554
  Male   0.8542397 0.3529895

        lips_thin
Y              [,1]      [,2]
  Female 0.1956897 0.3969018
  Male   0.8156973 0.3878668

        distance_nose_to_lip_long
Y              [,1]      [,2]
  Female 0.1775862 0.3823289
  Male   0.8430273 0.3639025
```

2

Random Forest

An ensemble learning method that combines multiple decision trees.

```
Call:
 randomForest(formula = gender ~ ., data = train_df, ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 2

        OOB estimate of  error rate: 3.94%
Confusion matrix:
       Female Male class.error
Female   1123   37  0.03189655
Male       65 1362  0.04555011
```

Support Vector Machines

A supervised learning model that finds the optimal hyperplane to separate data points.

```
Call:
svm(formula = gender ~ ., data = train_df)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  323
```

Comparing Model Performance

94.73%

Naive Bayes

Accuracy: 94.73%.

95.66%

Random Forest

Accuracy: 95.66%.

95.05%

SVM

Accuracy: 95.05%.

NOTE

Confusion matrices highlight the strengths and weaknesses of each model in identifying both genders.

Confusion matrix for Naïve Bayes model :

	y_pred_test Female	y_pred_test Male
Female	281	9
Male	25	331

Confusion matrix for Random forest model :

	y_pred_test Female	y_pred_test Male
Female	282	8
Male	20	336

Confusion matrix for SVM model :

	y_pred_test Female	y_pred_test Male
Female	282	8
Male	24	332

Hypothesis Testing

Hypotheses

Null Hypothesis H0 : Models perform the same.

Alternate Hypothesis H1: Models do not perform the same.

Confidence Level

95% Confidence Level.

Z-critical=1.96.

Formula Used:

- Variance formula: $\sigma^2 = \frac{\text{Acc}\cdot(1-\text{Acc})}{n}$
- $Z\text{-score} = \frac{\text{Acc}_1 - \text{Acc}_2}{\sqrt{\sigma_1^2+\sigma_2^2}}$

With a threshold alfa = 5%, we have a z alpha/2 = 1.96 according to the z-table shown below:

Confidence Level	Alpha	Alpha/2	z alpha/2
90%	10%	5.0%	1.645
95%	5%	2.5%	1.96
98%	2%	1.0%	2.326
99%	1%	0.5%	2.576

Result

1

Z-value

0.658.

2

Decision

$|Z| < 1.96$

Fail to reject H0.

3

Conclusion

No significant difference between models.

Key Takeaways

Data Analysis

EDA is crucial for understanding data characteristics.

Model Selection

Choose models based on data and objective.

Performance Evaluation

Compare models using appropriate metrics.