Jefferson Veliz

Data mining for Business Analytics

Professor Avinash Jairam

July 15th, 2023

Lab #3

The motivation behind this project was to explore the potential of machine learning algorithms in accurately predicting an individual's gender based on specific facial features. Gender prediction based on physical characteristics has been an area of interest in various fields, including biometrics. Understanding how certain facial attributes relate to gender can have implications in areas such as facial recognition technology.

The Gender Classification dataset from Kaggle is used to predict an individual's gender based on specific facial features. In this project, I utilized two powerful machine learning algorithms, K-Nearest Neighbors (KNN) and Naive Bayes, to create predictive models. This report summarizes our data exploration, model training, cross-validation results, and performance evaluation on the test set.

The Gender dataset contains 5001 entries and 8 columns, including features such as long hair, forehead width and height, nose width and length, lips thinness, and the distance between the nose and lip. I ensured data completeness, as there were no missing values in any of the features. I split the dataset into training and test sets using a 70-30 split. Feature scaling was applied to standardize the data for both models using the StandardScaler from scikit-learn. I then trained the KNN model with five nearest neighbors and the Naive Bayes model with the Gaussian distribution. To assess model accuracy, I performed 5-fold and 10-fold cross-validation techniques. The results were as follows: KNN achieved an average accuracy of approximately 96.34% using 5-fold cross-validation and 96.54% using 10-fold stratified cross-validation. Naive Bayes obtained an average accuracy of approximately 97.14% using both 5-fold and 10-fold stratified cross-validation.

The KNN and Naive Bayes models were evaluated on the test set. The classification reports showed high precision, recall, and F1-scores for both genders, with overall accuracies of approximately 97%.

The results demonstrate the effectiveness of the KNN and Naive Bayes algorithms in accurately predicting gender based on facial features. The models' high accuracy suggests that certain facial attributes can be indicative of gender. However, it is essential to interpret the results with caution, considering potential biases and ethical implications related to gender prediction.

Overall, this lab showcases the potential of machine learning techniques in gender prediction and provides a foundation for exploring facial biometrics.