Final Project

Alanda Cherestal

November 2022

Contents

1 Introduction

In this assignment, I implemented K-Nearest Neighbor and Naive Bayes Method to classify different type of rice using attributes from rice images. For example, if given the area of pixels within the border of a rice image, what type of rice is it? This project is written in python on Google Colab.

2 Data Processing

Every dataset needs to be split into train and test before machine learning can be implemented. Using Numpy.split(), I split the data set of 75000 samples into 80% for training and 20% for testing. I took both sets and separated the attributes from the class. That resulted in X and Y data for train and X and Y data for testing. The data set had over 100 attributes, so I decided to use the first five attributes. The classes for this data set are Basmati, Arborio, Jasmine, Ipsala, and Karacadag rice type.

2.1 K-Nearest Neighbor Method

I used the pixel area of each rice image to be the factor used to implement KNN. I imported KNeighborsClassifier from sklearn to implement the method. I pass both the areas and the classes to get the predicted labels that are used for accuracy.

2.2 Naive Bayes Method

Naive Bayes Method was tricky. I am still not sure If I used the method correctly. I first wanted to make a dictionary with ranges, since the attributes are integers, but that was not optimal. I imported GaussianNB from sklearn. I pass the attributes and the classes to get the predicted labels that are used for accuracy.

3 Evaluation

The predicted results are compared with the correct classes to give the accuracy of each method used. KNN resulted in a 75% accuracy and Naive Bayes resulted in a 93% accuracy.

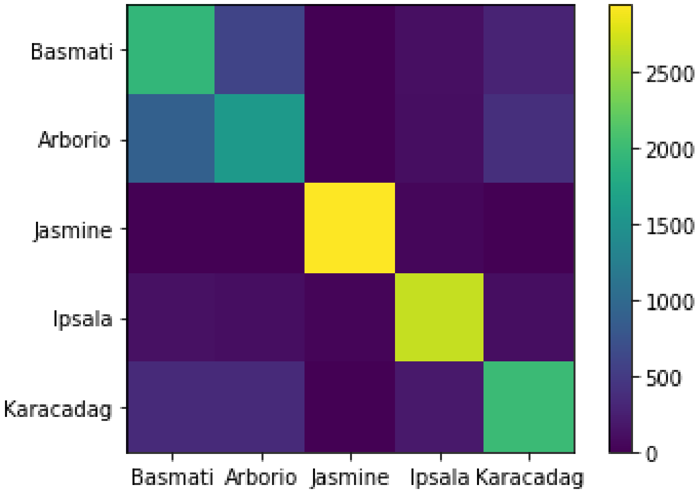

Figure 1: this Figure shows the confusion matrix from KNN method.

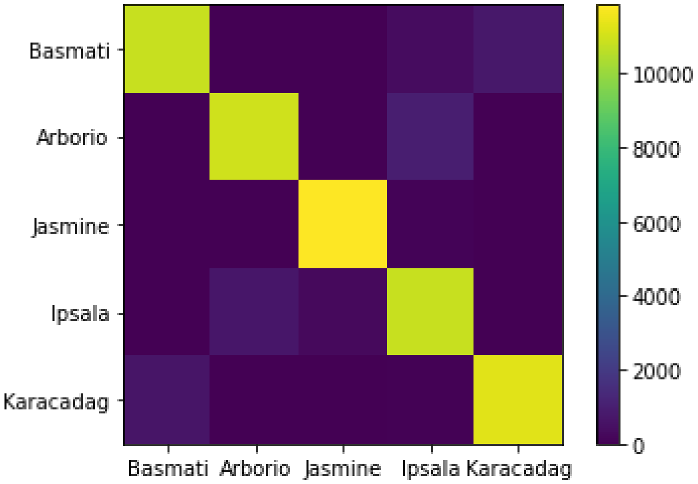

Figure 2: this Figure shows the confusion matrix from Naive Bayes method.

4 Discussion

In Figure 2, it shows the generated confusion matrix from KNN method. Basmati and Arborio rice types came up to be the most falsely predicted. It shows that both Basmati and Arborio are similar in size. I used K=5 which resulted in an accuracy of 0.75, but when I increased and decreased K, the accuracy went down. Maybe through more experimenting, I would be able to find the perfect K to result the highest accuracy. Figure 2 shows the generated confusion matrix from Naive Bayes method. It resulted in 0.93 accuracy for both the train and the test data.

5 What I've learned

I learned from the project that dataset searching is hard. I had a dataset in mind before starting the project but It was very hard to find the specific attributes I was thinking about. I wanted a data set that came with images, real attributes and categorical attribute. That was a fail.