Cheng-Shao(Eric), Lin

Prof. Hongchang Gao

Principles of Data Science

12 April 2024

**Final Project Progress Report II**

**Project title: Dog Breed Rating Plot**

I have made substantial progress on my code, while also encountering a few errors. I was able to refine some of the code I had created last week, and continue working on other aspects of the project in order to start getting the project closer to completion.

**Week 2 Process**

I have finished the user interface in which the user can type what attributes they want and then give the attribute they have chosen, a rating. First, the system will show all the attributes of the breed information. I have additionally assigned each row a code, so that user does not need to type the entire name description of the breed. In the user interface, I have made the code so that whatever the user types, it is in lowercase. Therefore, the user does not need to worry about whether it is uppercase or lowercase. Simultaneously, the system will detect if the code of the breed has existed already. When the user finishes their selection, they simply must type “done”. Once the system receives the command “done”, it will show what the user has selected. Additionally, I also created a system in which users can give a score between 1 and 5 of whatever attribute they have chosen, in order to show their preference, and to also detect if the number is valid.

Furthermore, I picked the category of the breed being good at living apartment as the target and used the tendency to bark or howl as its comparison. The crucial part of my process was loading the dataset containing the dog breed trait information from the CSV file. Then I checked to confirm if there were any missing values, and I found 7 missing. The values were ‘weight’ and ‘height’. Therefore, I removed the non-numeric characters from them and converted the cleaned values to numeric format. I used errors=’coerce’ to ensure that whether any values could not be converted to numeric, they would be replaced with NaN which acts as a missing value to avoid error. I did this in order to check if the dataset is balanced or not. After that, I visualized the distribution of the target variable.

**Problems Encountered**

While normalizing the features of the sample, I have encountered a problem in which the system shows that I cannot convert the breed’s name and URL to float. These are important factors that are needed to demonstrate to the user what breeds are best suited for them, and the website as well. The problem that I encounter is as follows:

**ValueError: could not convert string to float: 'Stabyhoun'**

This error indicates that there is a string value of the different breeds that cnot be converted to a float. This occurs often when I am trying to use a string column as a feature in any machine learning model that requires numerical data.

**Project completion plan**

For this week, setting user interface proved to be a bit difficult because we need to take into consideration what the user might be thinking. Checking the balance of data is relatively easy. All I had to do was simply follow the formula. Other than that, I am working on solving a few minor issues and refining the code to avoid encountering any future errors.

In the next few weeks, I will figure the problem out first, and get the system to stop giving me an error. Additionally, I will implement the K-nearest neighbors (KNN) method, and also use Linear Regression to predict continuous outcomes based on the input features and Logistic Regression. . Additionally, I will be calculating the accuracy, recall, and F-1 score in order to compare the performance of each different breed model.