Cheng-Shao(Eric), Lin

Final Project Proposal

Project title: Dog Breed Rating Plot

Introduction

Having had dogs as pets for my whole life, I wanted my final project to be related to the subject of dogs. Recently, my friend was looking for puppies but was concerned about taking care of them based on multiple factors, specifically how well a puppy would adjust to living in a city apartment. For this project, I decided to find a dataset that rates dogs on multiple factors, specifically focusing on rating dogs based on how well they would adapt in an apartment, and by using linear regression, classification, and recommendation system, coming up with the best solution to the problem. In an academic article relating to dog behavior, a similar method was used to conclude whether a dog would be aggressive based on its height, bodyweight, and skull shape. Similarly, I am going to conclude whether a dog is fit for a college student based on adaptability in an apartment, tolerance to being alone, general friendliness, energy level, amount of shedding, etc.

Proposed Work

First, I will read the file and then find the missing value. For numerical features, I will input the mean value, and for the categorical features, I will first check whether there are any and, if so, then I will fill in with the mode, and then convert the categorical features into numerical features. Subsequently, I am going to give the user preferences, to calculate the score based on the condition that user provide and calculate the scores with linear regression. To classify the breeds as recommended or not recommended, I will set a threshold and if the score is bigger or equal to the threshold, I will change the value of recommendation. For the classification, I will separate the data into two classes, one for user preferences, one for recommendation, and split it into a training set and a testing set, while also making the predictions for classification. Then, I will add evaluate the classifier: accuracy, recall and F-1 score to compare the performance of the breed model. For the recommendation system, I will only show the breeds' name, their score, and the website, and sort the filtered dataset by the top 10 score column in descending order, to make sure the highest score appears first. Lastly, show the bar plot for the recommended breeds and the scores.

Timeline

Week 1: Researching the data and understanding its different components.

Week 2: Begin coding and finding the missing values, fill in with mean value and find whether there are categorical features. Write progress report on findings and coding progress. Mention any issues and troubleshoot.

Week 3: Continue working on code and refining it. Troubleshoot any issues.

Week 4: Finalize code, and summarize findings.

References

https://www.kaggle.com/datasets/mexwell/dog-breeds-dogtime-dataset/code

pone.0080529 1..7 (plos.org)