Predicting Dog Stages in WeRateDogs Data

There are 84% Dog Stages was null, so I decide to predict the dog stages from another value that was not null. I use simple decision tree. The step I do are:

1. Make sure the type of our table
2. Define X and Y as predictor and label
3. Encode categorical predictor
4. Split train, and test data
5. Make prediction
6. Show Metrics from prediction

I use some library in sklearn to make predictions, such as sklearn metrics, sklearn model selection, and sklearn preprocessing. The result also can be checked in the notebook section prediction.

1. Make sure the type of our table

   In this step I make sure that all data type is right, and define what variable I need.

2. Define X and Y as predictor and label

   Because the unbalanced data, I try to upsampling so the model will learn much data. After that, I define the predictor (X) and the label (Y). Of course, the label is dog_stages, and the predictor is all the variables except 'dog_stage','timestamp','tweet_id','expanded_urls', and 'jpg_url' because dog_stage is our label, another variable is unique, and for timestamp I think I don’t need it for now.

3. Encode categorical predictor

   Because some value is object or categorical so we must encode that into numeric nominal variable so the model can learn by that.

4. Split train, and test data

   I just split the dataset into train and test, I’m not use data validation and cross validation technique because the data have small size.

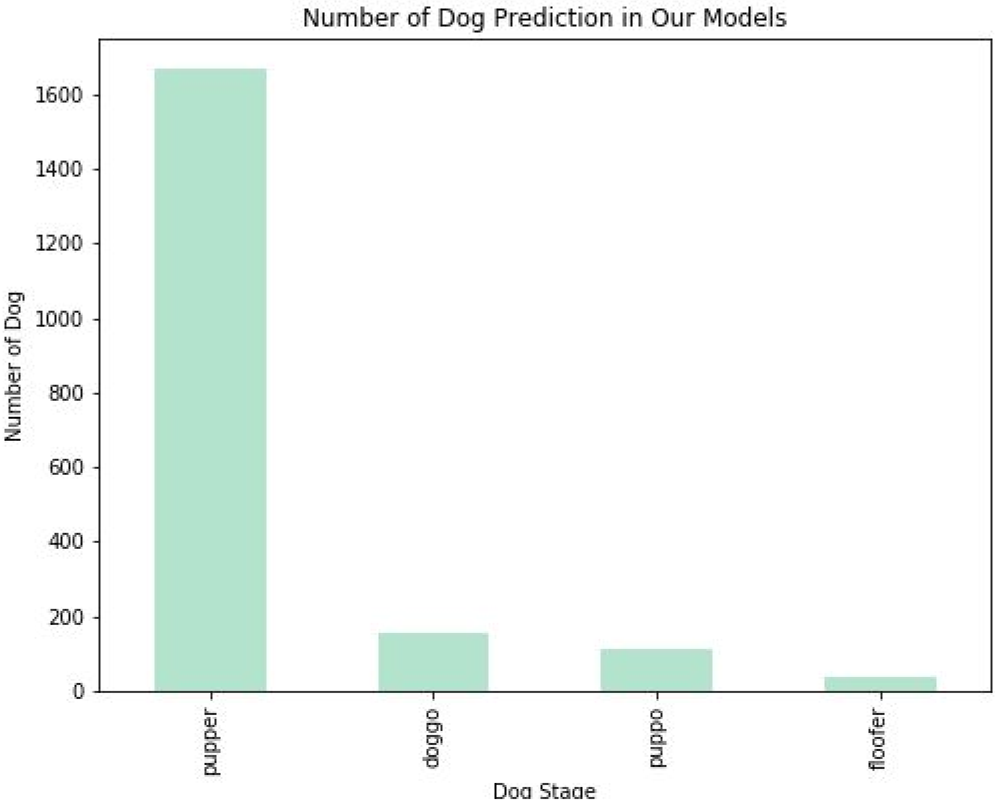
5. Make prediction

   I make predictions with sklearn decision tree gini.

6. Show Metrics from prediction

The metrics are good enough so I decide to make predictions in data with missing values in dog_stage.

The Result



Just like the value before upper sampling, the popular dog_stage is pupper.