**Wine score prediction based on reviews**

|  |  |
| --- | --- |
| Name: | Xiaoyang Wang |
|  |
| Email: | gnayoaix@bu.edu |

# Description

1.

My dataset is basically about reviews of wines, which given by sommeliers. The main attributes is the description of wines.Besides this, there are country, designation, points, price, province, region, taster’s name.

Here are the explanations of each attributes:

Country: The country that the wine is from

Designation: The vineyard within the winery where the grapes that made the wine are from

Points: The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for

Price: The cost for a bottle of the wine

Province: The province or state that the wine is from

Region: The wine growing area in a province or state (ie Napa); Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley)

taster’s name: sommeliers

there are about 130k reviews.

2.

I want to create a predictive model to predict the possible point of wines based on the description from the taster. So basically , the distribution of points is from 80 to 100, and I build up a rule that if the point is greater than 88, it is a good wine, otherwise, it is a bad wine. So after I trained the model by descriptions, the model can tell me whether it is a good wine via the new description. In other words, in this scenario, some sentiments(words) can especially represent good wines, some word for bed wines.

So my model is a classification model

3.

I have some consideration of my rules: it is hard to set the boundary. We can say that a wine with 99 points is a good one, but what about 89 points? For instance, the true is the taste of a wine is bad, but it has some unique smell so the taster gave it 89 points, that is a mistake. However, for the most of the points, it is precise, so I think this model can be significance.

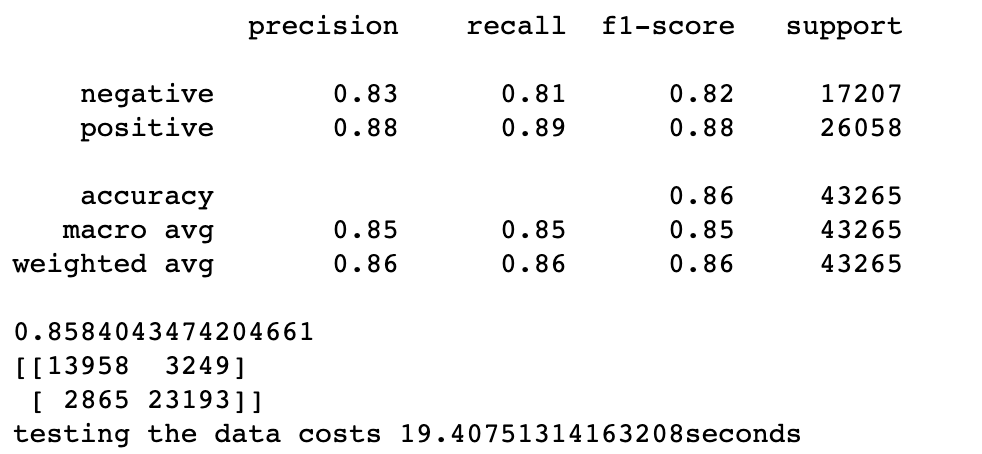
4.

I want to use F1-score, precision, recall, confusion metrics and weighted average accuracy as the evaluation of my project. I split the whole dataset into training and testing sets, the ratio of them is 2:1. I expected my model can have more than 80% accuracy.

# Result

I have tried three different machine learning models: logistic regression ,naïve bayes and SVM , here are the results:

Logistic regression (500 iterations):



Naïve bayes:

