Wine Variety Prediction Report

Submitted By – Abhinav Garg

A model to predict the variety of grape used in wine using the given dataset, with the help of natural language processing and other data analysis techniques.

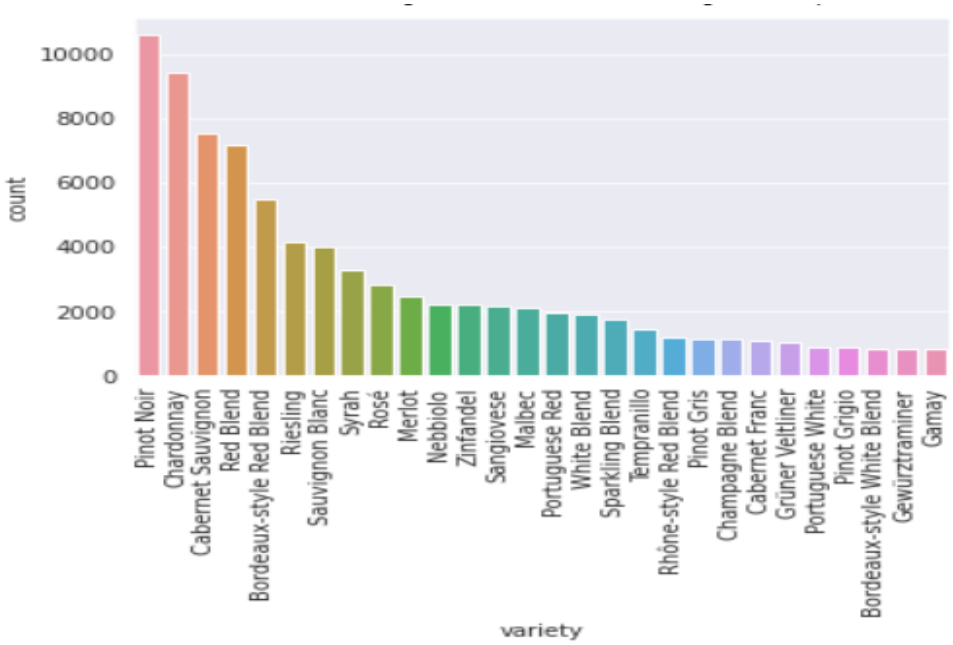
1. Dataset Description

The given dataset has 82,657 samples with the following features:

FEATURE	DESCRIPTION	NULL VALUE %AGE
USER_NAME	The user name of the reviewer	23.462018
COUNTRY	The country that the wine is from.	0.042344
REVIEW_TITLE	The title of the wine review, which often contains the vintage.	0.000000
REVIEW_DESCRIPTION	A verbose review of the wine.	0.000000
DESIGNATION	The vineyard within the winery where the grapes that made the wine are from.	28.608587
POINTS	The ratings given by the user (between 0 -100)	0.000000
PRICE	The cost for a bottle of the wine	6.737481
PROVINCE	The province or state that the wine is from.	0.042344
REGION_1	The wine-growing area in a province.	15.430030
REGION_2	More specific regions specified within a wine-growing area.	56.508221
WINERY	The winery that made the wine	0.000000
VARIETY	The type of grapes used to make wine.	0.000000

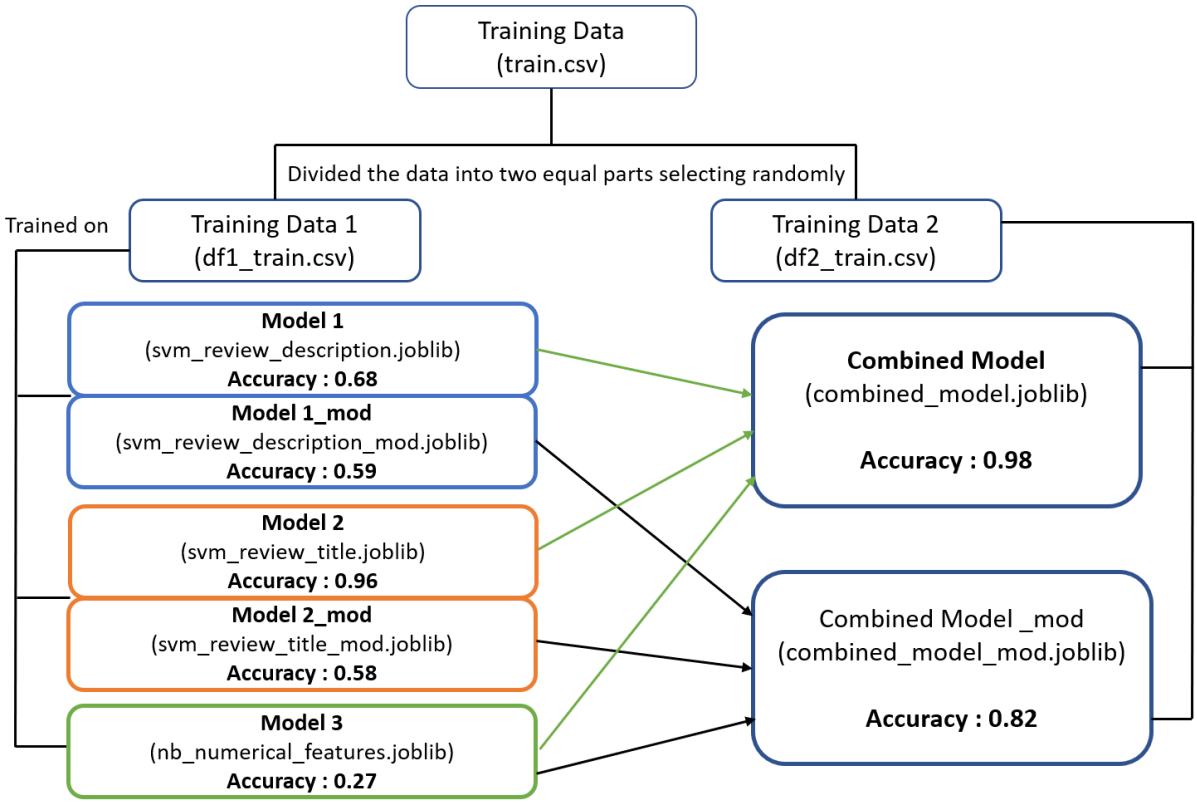
We can see that there are many missing values in the data provided so we either dropped of the columns which has large number of missing values and are calculated / filled in the notebook (*Data PreProcessing and Visualization.ipynb*)

The below graph was plotted to observe the “distribution of reviews with respect to variety” and we observed that the dataset is slightly unbalanced.



2. Workflow

The workflow is depicted through the below flow diagram:



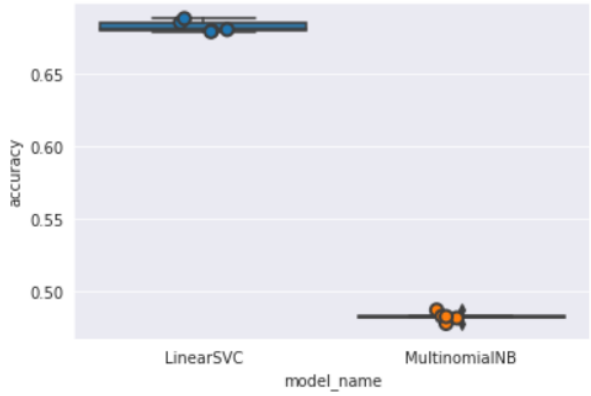
Step 1:
The Training dataset is divided randomly into two equal parts.
**Refer Notebook "Splitting the Training Data.ipynb"*

Step 2:
Different features of training dataset part 1 was used such as review description, review title, country, points.

Step 2.1: Model 1
Applied Natural Language Processing to extract out the features from the review description provided, then used these features to train two different machine learning algorithms (LinearSVC and Multinomial-NB) and got the following results. Hence selected LinearSVC.

```
model_name
LinearSVC        0.682931
MultinomialNB    0.482296
Name: accuracy, dtype: float64
```

```
Started Training LinearSVC
[LibLinear][LibLinear][LibLinear][LibLinear][LibLinear]Completed Training LinearSVC
Started Training MultinomialNB
Completed Training MultinomialNB
```

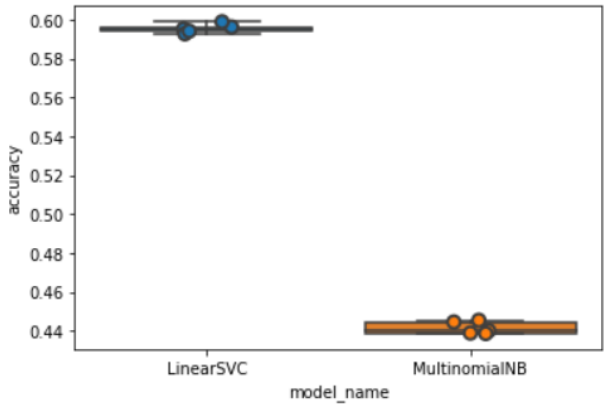


**Refer Notebook "Classification based on Review Description.ipynb"*

Step 2.2: Model 1_mod
While applying NLP it was observed that description contains the name of the variety we want to predict hence trained another model excluding out the names of the variety of the features and got the following results. Hence selected LinearSVC.

```
model_name
LinearSVC        0.595416
MultinomialNB    0.441419
Name: accuracy, dtype: float64
```

```
Started Training LinearSVC
[LibLinear][LibLinear][LibLinear][LibLinear][LibLinear]Completed Training LinearSVC
Started Training MultinomialNB
Completed Training MultinomialNB
```



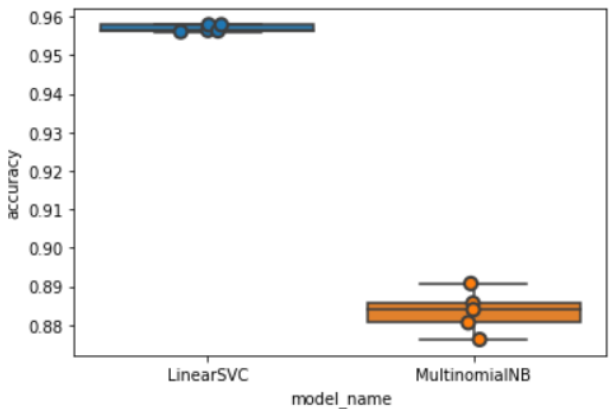
**Refer Notebook "(MODF)Classification based on Review Description.ipynb"*

Step 2.3: Model 2

Applied Natural Language Processing to extract out the features from the review title provided, then used these features to train two different machine learning algorithms (LinearSVC and Multinomial-NB) and got the following results. Hence selected LinearSVC.

```
model_name
LinearSVC        0.956921
MultinomialNB    0.883419
Name: accuracy, dtype: float64
```

```
Started Training LinearSVC
[LibLinear][LibLinear][LibLinear][LibLinear][LibLinear]Completed Training LinearSVC
Started Training MultinomialNB
Completed Training MultinomialNB
```



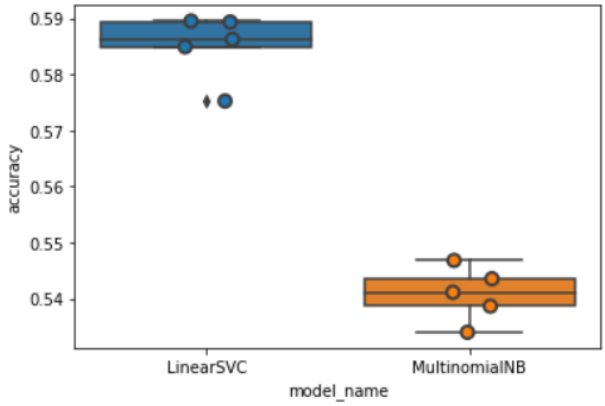
**Refer Notebook "Classification based on Review Title.ipynb"*

Step 2.4: Model 2_mod

While applying NLP it was observed that title contains the name of the variety we want to predict hence trained another model excluding out the names of the variety of the features and got the following results. Hence selected LinearSVC.

```
model_name
LinearSVC        0.585082
MultinomialNB    0.540841
Name: accuracy, dtype: float64
```

```
Started Training LinearSVC
[LibLinear][LibLinear][LibLinear][LibLinear][LibLinear]Completed Training LinearSVC
Started Training MultinomialNB
Completed Training MultinomialNB
```



**Refer Notebook "(MODF)Classification based on Review Title.ipynb"*

Step 2.5: Model 3

In this Model features such as country and points were used to train and it was trained on MultinomialNB and got the following results.

accuracy			0.27
macro avg	0.16	0.20	0.16
weighted avg	0.21	0.27	0.21

**Refer Notebook "Classification based on Numeric Features.ipynb"*

Step 3:

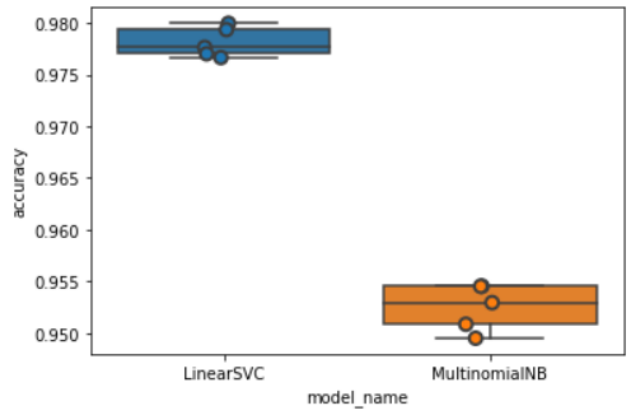
The above models were used to predict classes on Training data 2 and the above predictions were used as features to train out a combined model.

Step 3.1: Combined Model

Combining the model 2.1, 2.3, 2.5 and the results were used as features to train out this combined model and got the following results.

```
model_name
LinearSVC        0.978132
MultinomialNB    0.952480
Name: accuracy, dtype: float64
```

```
Started Training LinearSVC
[LibLinear][LibLinear][LibLinear][LibLinear][LibLinear]Completed Training LinearSVC
Started Training MultinomialNB
Completed Training MultinomialNB
```



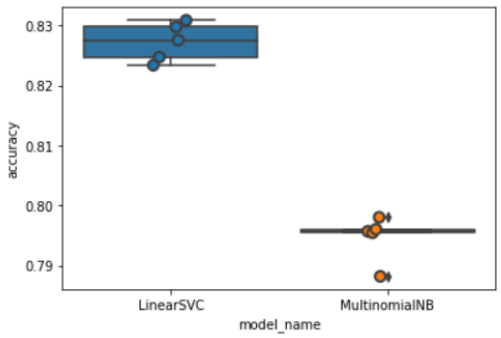
**Refer Notebook "Combined Model Classification.ipynb"*

Step 3.2: Combined Model_mod

Combining the model 2.2, 2.4, 2.5 and the results were used as features to train out this combined model and got the following results.

```
model_name
LinearSVC        0.827219
MultinomialNB    0.794688
Name: accuracy, dtype: float64
```

```
Started Training LinearSVC
[LibLinear][LibLinear][LibLinear][LibLinear][LibLinear]Completed Training LinearSVC
Started Training MultinomialNB
Completed Training MultinomialNB
```



**Refer Notebook "(MODF) Combined Model Classification.ipynb"*

3. Data Insights

Please refer this notebook for data insights. (*"Data Insights.ipynb"*)

4. Conclusion

Hence, we have studied the data and performed various data analysis on it. We trained different models and achieved an accuracy of 98% when we use the variety names present in given title and description and an accuracy of 82% without using the variety names present in given title and description.