

# PROJECT #2

## OVERVIEW:

In this project, we have been provided with a wine reviews dataset with two columns: “review\_text” and “wine\_variant” and the goal is to create a wine recommendation system using test classification.

Data:

**Target variable** – ‘wine\_variant’

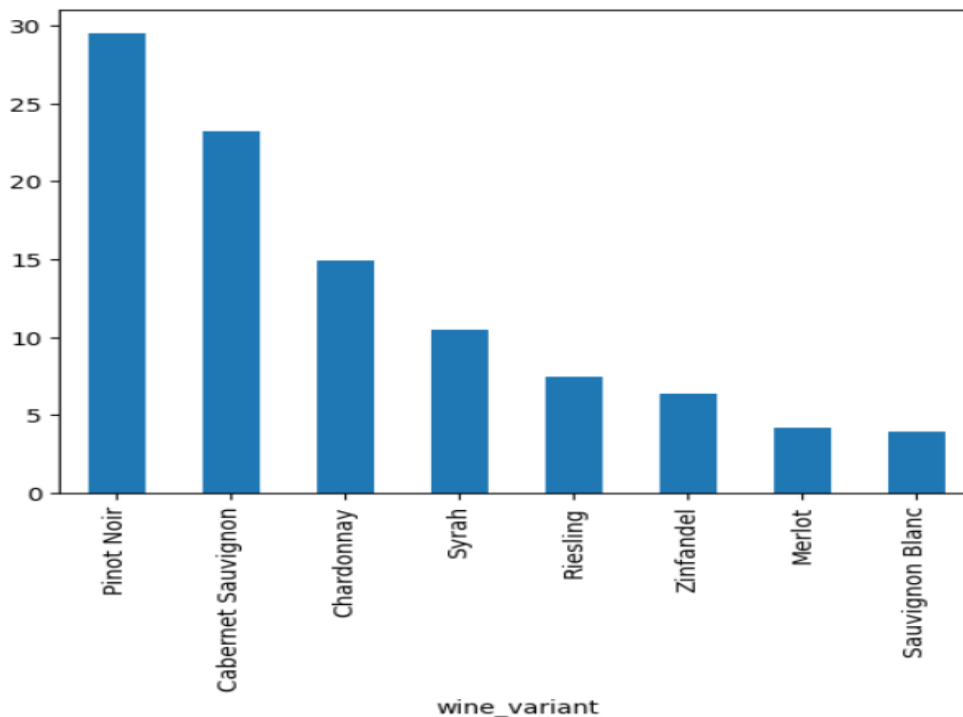
**Categories** – 8

**Types** - 'Pinot Noir', 'Sauvignon Blanc', 'Cabernet Sauvignon', 'Chardonnay', 'Syrah', 'Riesling', 'Merlot', 'Zinfandel'

**Train data** – 10000 observations were split into **test set** of sample size 25% (**2500**). Stratified sampling used for appropriate representation of above-mentioned classes. An additional validation data with 5000 observations has been used.

**Distribution** – In percentage

Pinot Noir	29.48
Cabernet Sauvignon	23.17
Chardonnay	14.91
Syrah	10.49
Riesling	7.43
Zinfandel	6.37
Merlot	4.23
Sauvignon Blanc	3.92



## Machine Learning (Models and Results):

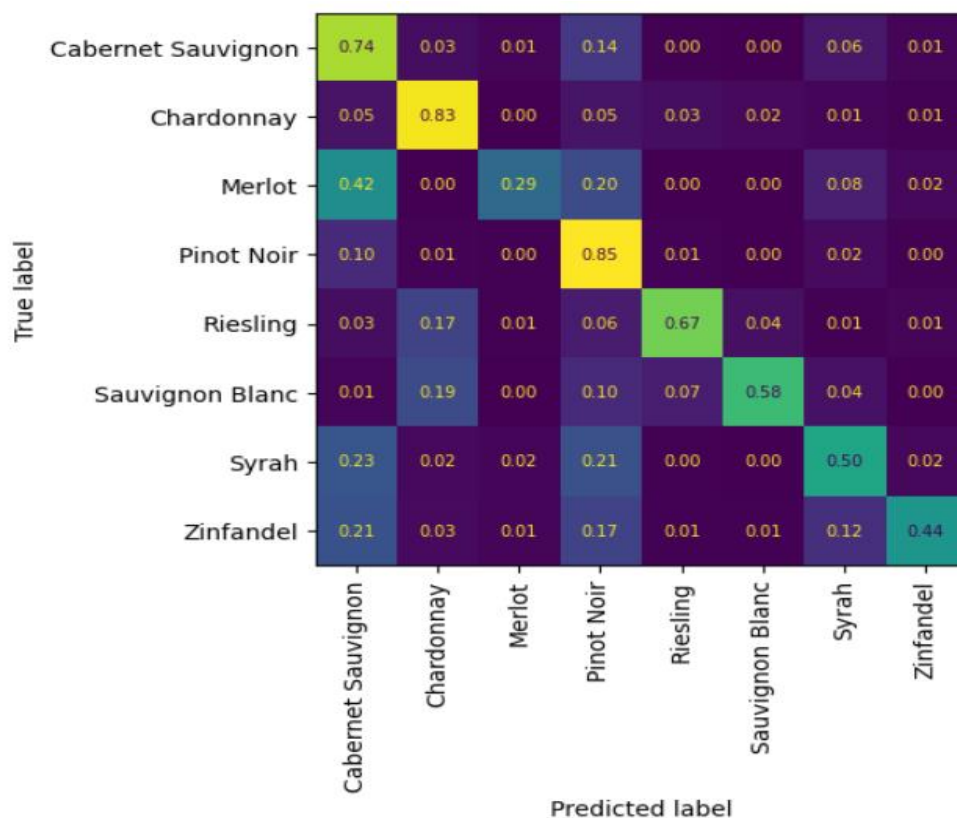
1. **Linear SVM** was used on the given data with **TF-IDF vectorization** which resulted in a macro average f1 score of **0.65**

	precision	recall	f1-score	support
Cabernet Sauvignon	0.64	0.74	0.69	579
Chardonnay	0.78	0.83	0.81	372
Merlot	0.62	0.29	0.40	106
Pinot Noir	0.73	0.85	0.79	737
Riesling	0.81	0.67	0.73	185
Sauvignon Blanc	0.78	0.58	0.67	98
Syrah	0.60	0.50	0.55	263
Zinfandel	0.76	0.44	0.56	160
accuracy			0.71	2500
macro avg	0.72	0.61	0.65	2500
weighted avg	0.71	0.71	0.70	2500

With hyperparameter tuning for **50 iterations**, the following model was selected as the best model. However, results remained unchanged.

Fitting 5 folds for each of 50 candidates, totalling 250 fits

```
{'linearsvc_C': 1, 'linearsvc_loss': 'hinge', 'linearsvc_penalty': 'l2', 'tfidfvectorizer_lowercase': True, 'tfidfvectorizer_min_df': 1, 'tfidfvectorizer_stop_words': None}
```



2. For **non-linear Support Vector Machine**, I experimented with the two following algorithms:

### 2.1 Polynomial kernel

The macro average f1 score worsened to **0.19** an accuracy of only **0.44**

	precision	recall	f1-score	support
Cabernet Sauvignon	0.60	0.79	0.68	579
Chardonnay	0.76	0.78	0.77	372
Merlot	0.95	0.18	0.30	106
Pinot Noir	0.64	0.88	0.74	737
Riesling	0.90	0.51	0.66	185
Sauvignon Blanc	0.93	0.43	0.59	98
Syrah	0.73	0.33	0.46	263
Zinfandel	0.90	0.27	0.41	160
accuracy			0.67	2500
macro avg	0.80	0.52	0.58	2500
weighted avg	0.72	0.67	0.65	2500

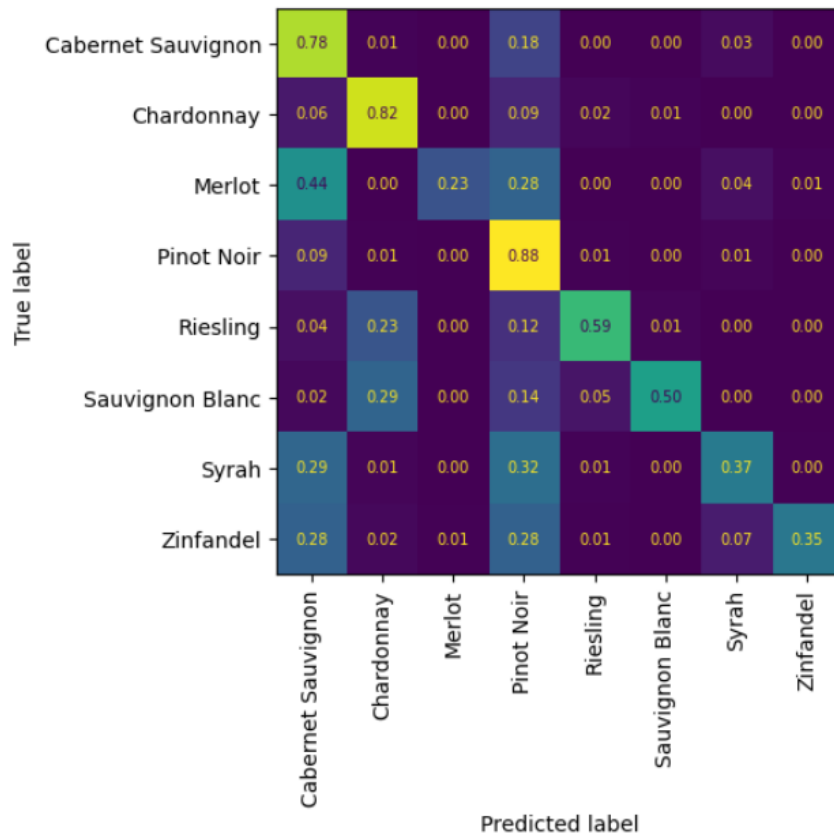
### 2.2 RBF kernel

With RBF kernel and tfidf vectorization we obtained a macro average f1 score of **0.58** and an improvement of accuracy score to **0.67**

	precision	recall	f1-score	support
Cabernet Sauvignon	0.60	0.79	0.68	579
Chardonnay	0.76	0.78	0.77	372
Merlot	0.95	0.18	0.30	106
Pinot Noir	0.64	0.88	0.74	737
Riesling	0.90	0.51	0.66	185
Sauvignon Blanc	0.93	0.43	0.59	98
Syrah	0.73	0.33	0.46	263
Zinfandel	0.90	0.27	0.41	160
accuracy			0.67	2500
macro avg	0.80	0.52	0.58	2500
weighted avg	0.72	0.67	0.65	2500

3. **Stochastic Gradient Descent Classifier** with tfidf vectorization did not provide any improvement is results as seen below.

	precision	recall	f1-score	support
Cabernet Sauvignon	0.93	0.05	0.09	579
Chardonnay	0.82	0.08	0.14	372
Merlot	0.00	0.00	0.00	106
Pinot Noir	0.30	1.00	0.46	737
Riesling	0.00	0.00	0.00	185
Sauvignon Blanc	0.00	0.00	0.00	98
Syrah	0.00	0.00	0.00	263
Zinfandel	0.00	0.00	0.00	160
accuracy			0.32	2500
macro avg	0.26	0.14	0.09	2500
weighted avg	0.43	0.32	0.18	2500



**Observations:**

From the model evaluation metrics and confusion metrics we can clearly see that 'Cabernet Sauvignon', 'Chardonnay' and 'Pinot Noir' have better chances of classification. So we can conclude that the discrepancy in results is primarily due to class imbalance.

**Tacking Class Imbalance:**

Using domain knowledge and looking at representation of classes in our data I have recategorized the 'wine\_variants' as follows:

```
wine_variant
Cabernet Sauvignon/Zinfandel    29.55
Pinot Noir                      29.48
Chardonnay                     14.91
Merlot/Syrah                   14.72
Riesling/Sauvignon Blanc       11.35
Name: count, dtype: float64
```

We can find the categories along with their distribution in percentages in the above screenshot. I have tried fit the above-mentioned algorithms to the new categorization.

## Machine Learning (Models and Results):

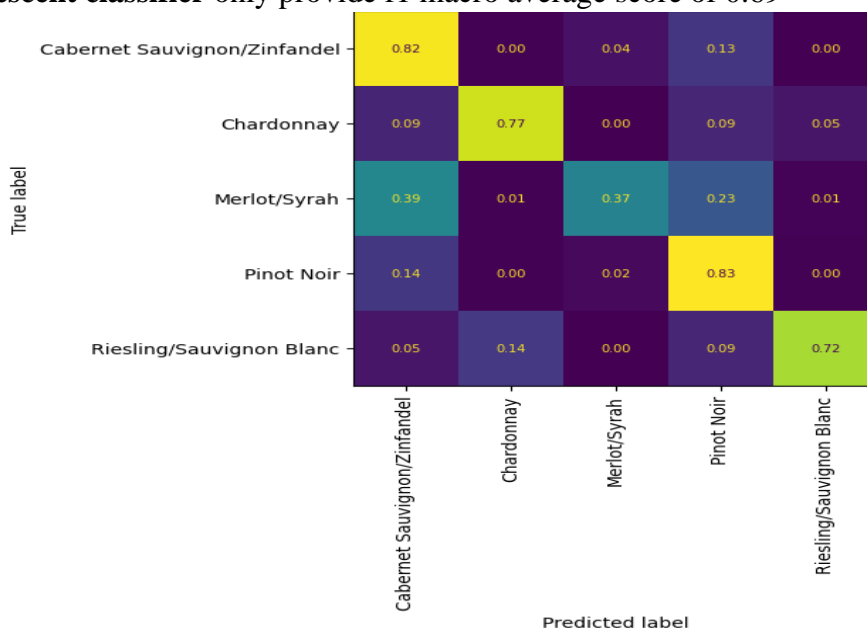
1. **Multinomial Naïve Bayes** model shows a significant improvement in results to **0.73** as macro average f1 score.

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.70	0.76	0.73	739
Chardonnay	0.83	0.81	0.82	372
Merlot/Syrah	0.62	0.47	0.54	369
Pinot Noir	0.74	0.81	0.77	737
Riesling/Sauvignon Blanc	0.84	0.75	0.79	283
accuracy			0.74	2500
macro avg	0.75	0.72	0.73	2500
weighted avg	0.74	0.74	0.73	2500

2. **Linear SVM** was used on the given data with **TF-IDF vectorization** also resulted in a macro average f1 score of **0.73**

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.71	0.81	0.76	739
Chardonnay	0.80	0.81	0.80	372
Merlot/Syrah	0.75	0.40	0.52	369
Pinot Noir	0.75	0.84	0.79	737
Riesling/Sauvignon Blanc	0.82	0.73	0.77	283
accuracy			0.75	2500
macro avg	0.76	0.72	0.73	2500
weighted avg	0.75	0.75	0.74	2500

3. **Non-linear Support Vector Machine with RBF kernel and Stochastic gradient descent classifier** only provide f1 macro average score of 0.69



In order to improve results, I have tried working with Latent Semantic Analysis and Contextual embeddings. The results have been summarized below:

### 1. LSA + Linear SVC

---

```
{'C': 66.32394340124732, 'loss': 'squared_hinge', 'penalty': 'l2'}
```

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.70	0.77	0.73	739
Chardonnay	0.79	0.81	0.80	372
Merlot/Syrah	0.62	0.38	0.47	369
Pinot Noir	0.73	0.81	0.77	737
Riesling/Sauvignon Blanc	0.79	0.71	0.75	283
accuracy			0.73	2500
macro avg	0.73	0.70	0.70	2500
weighted avg	0.72	0.73	0.72	2500

### 2. Sentence Transformer (all-mpnet-base-v2) + LinearSVC

---

```
{'C': 1.5675906888544584, 'loss': 'squared_hinge', 'penalty': 'l2'}
```

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.68	0.73	0.70	739
Chardonnay	0.75	0.73	0.74	372
Merlot/Syrah	0.64	0.38	0.48	369
Pinot Noir	0.70	0.80	0.75	737
Riesling/Sauvignon Blanc	0.75	0.73	0.74	283
accuracy			0.70	2500
macro avg	0.71	0.68	0.68	2500
weighted avg	0.70	0.70	0.69	2500

No significant improvements in results were observed.

### 3. Sentence Transformer (all-mpnet-base-v2) + Non-linear SVC(polynomial kernel)

```
{'C': 7.51714305198504}
```

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.64	0.70	0.67	739
Chardonnay	0.73	0.76	0.74	372
Merlot/Syrah	0.54	0.44	0.48	369
Pinot Noir	0.72	0.74	0.73	737
Riesling/Sauvignon Blanc	0.78	0.69	0.74	283
accuracy			0.68	2500
macro avg	0.68	0.67	0.67	2500
weighted avg	0.68	0.68	0.68	2500

#### 4. Sentence Transformer (all-mpnet-base-v2) + Non-linear SVC(rbf kernel)

```
{'C': 1.5125648028067211, 'gamma': 0.09999999999999999}
```

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.60	0.80	0.68	739
Chardonnay	0.79	0.71	0.75	372
Merlot/Syrah	0.95	0.23	0.37	369
Pinot Noir	0.66	0.78	0.71	737
Riesling/Sauvignon Blanc	0.80	0.60	0.69	283
accuracy			0.67	2500
macro avg	0.76	0.62	0.64	2500
weighted avg	0.72	0.67	0.66	2500

#### 5. Sentence Transformer (all-mpnet-base-v2) + Random Forest Classifier

```
{'n_estimators': 30, 'min_samples_split': 5, 'max_depth': 9, 'criterion': 'entropy'}
```

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.48	0.75	0.58	739
Chardonnay	0.73	0.52	0.61	372
Merlot/Syrah	0.85	0.15	0.25	369
Pinot Noir	0.57	0.66	0.61	737
Riesling/Sauvignon Blanc	0.72	0.36	0.48	283
accuracy			0.56	2500
macro avg	0.67	0.49	0.51	2500
weighted avg	0.62	0.56	0.53	2500

### Neural Networks:

#### 1. CNN Classifier

```
params = {'MAX_LENGTH': 600,  
          'BATCH_SIZE': 512,  
          'EMBED_DIM': 256,  
          'EPOCHS': 50,  
          'MAX_GRADIENT': 2,  
          'LR': 0.01,  
          'ALPHA': 1e-3,  
          'N_FILTERS': 500,  
          'FILTER_SIZES': [1, 2],  
          'DROPOUT': 0.3  
}
```

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.71	0.71	0.71	739
Chardonnay	0.85	0.73	0.79	372
Merlot/Syrah	0.57	0.49	0.53	369
Pinot Noir	0.72	0.81	0.76	737
Riesling/Sauvignon Blanc	0.75	0.80	0.77	283
accuracy			0.72	2500
macro avg	0.72	0.71	0.71	2500
weighted avg	0.72	0.72	0.72	2500

## 2. LSTM Classifier

```
params = {'MAX_LENGTH': 600,
          'BATCH_SIZE': 256,
          'EMBED_DIM': 512,
          'EPOCHS': 130,
          'MAX_GRADIENT': 2,
          'LR': 0.1,
          'ALPHA': 1e-3,
          'HIDDEN_DIM': 64,
          'DROPOUT': 0.2
        }
```

	precision	recall	f1-score	support
Cabernet Sauvignon/Zinfandel	0.00	0.00	0.00	739
Chardonnay	0.00	0.00	0.00	372
Merlot/Syrah	0.00	0.00	0.00	369
Pinot Noir	0.29	1.00	0.46	737
Riesling/Sauvignon Blanc	0.00	0.00	0.00	283
accuracy			0.29	2500
macro avg	0.06	0.20	0.09	2500
weighted avg	0.09	0.29	0.13	2500

### Observations:

From the above results we can conclude that neither contextual embedding nor latent semantic analysis is causing improvement in results. As observed earlier, the re-categorizing does improve our results considerably but we also note that the category “**Merlot/Syrah**” is consistently performing poorly across all algorithms. Since it has a fair representation in the data, we can infer that cause is not class imbalance. This points us towards a possible chance of mis-grouping and initiates a need for further regrouping.

### REGROUPING THE TARGET VARIABLES:

We have used the following domain knowledge to categorize the target variables into appropriate groups as required for a wine recommendation system.

#### Group 1: Light-bodied, Crisp Whites

1. Sauvignon Blanc
2. Riesling



Both Sauvignon Blanc and Riesling are known for their bright acidity, refreshing qualities, and often fruity or floral aromas. They are typically light-bodied wines that pair well with seafood, salads, and lighter dishes.

#### Group 2: **Full-bodied Whites**

##### 1. **Chardonnay**

Chardonnay stands out as a full-bodied white wine with a wide range of flavors, from crisp and unoaked styles to rich and buttery expressions. It pairs well with a variety of foods, including poultry, seafood, and creamy pasta dishes.

#### Group 3: **Medium to Full-bodied Reds**

1. **Pinot Noir**
2. **Merlot**
3. **Syrah**

These red wines span the spectrum from medium to full-bodied, offering varying levels of tannins and fruitiness. Pinot Noir tends to be lighter-bodied with red fruit flavors, Merlot is medium-bodied with softer tannins, and Syrah is bold with spicy characteristics. They pair well with red meats, pasta dishes, and hearty stews.

#### Group 4: **Bold Reds**

1. **Cabernet Sauvignon**
2. **Zinfandel**

Cabernet Sauvignon and Zinfandel are both bold, full-bodied red wines with rich flavors of dark fruits and firm tannins. They pair well with robust dishes like grilled meats, steak, and aged cheeses.

These smaller groupings highlight the distinct characteristics of each wine and offer insight into their taste profiles and food pairings.

### **UPDATED RESULTS:**

#### 1. **TFIDF Vectorization + Linear SVC**

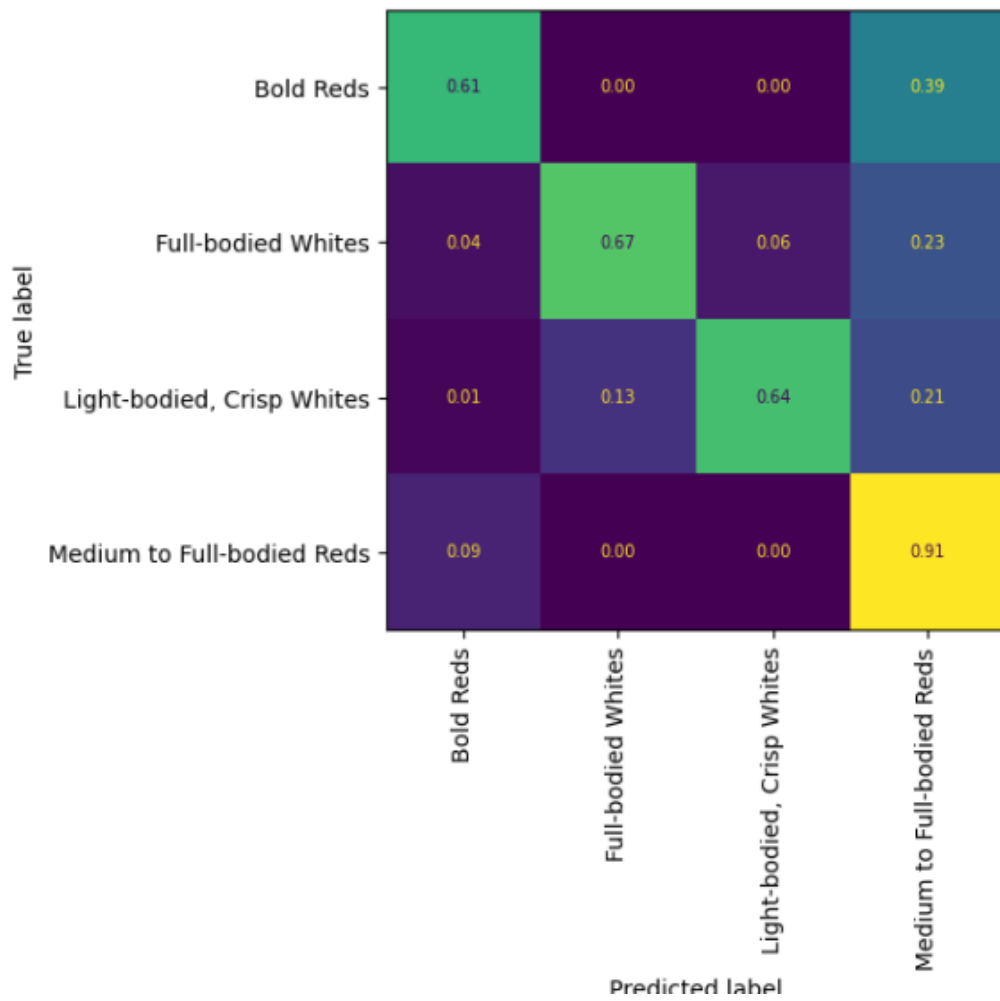
	precision	recall	f1-score	support
Bold Reds	0.73	0.70	0.71	739
Full-bodied Whites	0.86	0.79	0.82	372
Light-bodied, Crisp Whites	0.88	0.74	0.80	283
Medium to Full-bodied Reds	0.77	0.83	0.80	1106
accuracy			0.78	2500
macro avg	0.81	0.77	0.78	2500
weighted avg	0.78	0.78	0.78	2500

#### 2. **TFIDF Vectorization + Linear SVC with hyperparameter tuning**

Fitting 5 folds for each of 100 candidates, totalling 500 fits

```
{'linearsvc_C': 0.2321388625555897, 'linearsvc_loss': 'squared_hinge', 'linearsvc_max_iter': 2002, 'linearsvc_penalty': 'l2', 'tfidfvectorizer_lowercase': True, 'tfidfvectorizer_min_df': 3, 'tfidfvectorizer_stop_words': 'english'}
```

	precision	recall	f1-score	support
Bold Reds	0.77	0.69	0.72	739
Full-bodied Whites	0.86	0.76	0.81	372
Light-bodied, Crisp Whites	0.88	0.72	0.79	283
Medium to Full-bodied Reds	0.76	0.88	0.81	1106
accuracy			0.79	2500
macro avg	0.82	0.76	0.78	2500
weighted avg	0.79	0.79	0.78	2500



### 3. Contextual Embedding + Linear SVC with hyperparameter tuning

---

```
{'C': 12.865472168183121, 'loss': 'hinge', 'penalty': 'l2'}
```

	precision	recall	f1-score	support
Bold Reds	0.72	0.65	0.68	739
Full-bodied Whites	0.77	0.71	0.74	372
Light-bodied, Crisp Whites	0.75	0.71	0.73	283
Medium to Full-bodied Reds	0.74	0.82	0.78	1106
accuracy			0.74	2500
macro avg	0.75	0.72	0.73	2500
weighted avg	0.74	0.74	0.74	2500

### 4. Contextual Embedding + RBF kernel SVM

---

```
{'C': 3.8701165859335274, 'gamma': 10.0}
```

	precision	recall	f1-score	support
Bold Reds	0.85	0.27	0.41	739
Full-bodied Whites	0.90	0.26	0.40	372
Light-bodied, Crisp Whites	0.77	0.13	0.22	283
Medium to Full-bodied Reds	0.51	0.97	0.67	1106
accuracy			0.56	2500
macro avg	0.75	0.41	0.42	2500
weighted avg	0.69	0.56	0.50	2500

### 5. Convolutional Neural Network

Parameters –

```
params = {'MAX_LENGTH': 600,  
          'BATCH_SIZE': 512,  
          'EMBED_DIM': 512,  
          'EPOCHS': 50,  
          'MAX_GRADIENT': 2,  
          'LR': 0.01,  
          'ALPHA': 1e-3,  
          'N_FILTERS': 500,  
          'FILTER_SIZES': [1, 2],  
          'DROPOUT': 0.3  
        }
```

Results on test set –

	precision	recall	f1-score	support
Bold Reds	0.69	0.73	0.71	739
Full-bodied Whites	0.83	0.74	0.79	372
Light-bodied, Crisp Whites	0.86	0.69	0.77	283
Medium to Full-bodied Reds	0.77	0.81	0.79	1106
accuracy			0.76	2500
macro avg	0.79	0.74	0.76	2500
weighted avg	0.77	0.76	0.76	2500

Results on validation set –

	precision	recall	f1-score	support
Bold Reds	0.69	0.77	0.73	1477
Full-bodied Whites	0.80	0.78	0.79	745
Light-bodied, Crisp Whites	0.84	0.70	0.77	567
Medium to Full-bodied Reds	0.80	0.78	0.79	2211
accuracy			0.77	5000
macro avg	0.78	0.76	0.77	5000
weighted avg	0.77	0.77	0.77	5000

## 6. DistilBERT Model

Training –

 [3750/3750 29:42, Epoch 10/10]

Epoch	Training Loss	Validation Loss	Accuracy	F1 Score
1	0.592100	0.572056	0.739200	0.741621
2	0.468400	0.513672	0.777600	0.777547
3	0.293900	0.561780	0.784400	0.778364
4	0.123900	0.691120	0.789600	0.788829
5	0.147600	0.857492	0.785200	0.789751
6	0.084600	1.045856	0.781200	0.782769
7	0.026600	1.267746	0.785600	0.791238
8	0.057500	1.310684	0.781200	0.784031
9	0.001100	1.370054	0.782800	0.786848
10	0.005200	1.388676	0.779200	0.785061

Results on test set –

	precision	recall	f1-score	support
Bold Reds	0.73	0.72	0.73	739
Full-bodied Whites	0.87	0.78	0.83	372
Light-bodied, Crisp Whites	0.78	0.84	0.81	283
Medium to Full-bodied Reds	0.79	0.82	0.81	1106
accuracy			0.79	2500
macro avg	0.79	0.79	0.79	2500
weighted avg	0.79	0.79	0.79	2500

Results on Validation set –

	precision	recall	f1-score	support
Bold Reds	0.72	0.76	0.74	1477
Full-bodied Whites	0.82	0.79	0.80	745
Light-bodied, Crisp Whites	0.79	0.81	0.80	567
Medium to Full-bodied Reds	0.81	0.79	0.80	2211
accuracy			0.78	5000
macro avg	0.78	0.79	0.79	5000
weighted avg	0.78	0.78	0.78	5000


## 7. RoBERTa

Training –

 [3752/3752 35:47, Epoch 8/8]


Epoch	Training Loss	Validation Loss	Accuracy	F1 Score
1	0.673100	0.533543	0.754000	0.743172
2	0.520500	0.561946	0.761800	0.764767
3	0.418500	0.596411	0.745000	0.753272
4	0.321000	0.565207	0.795800	0.792943
5	0.259800	0.712381	0.790600	0.790530
6	0.271200	0.896873	0.785800	0.789710
7	0.079700	1.076672	0.790600	0.791532
8	0.046100	1.160723	0.792600	0.795253

Results on test set –

100%  2500/2500 [00:34<00:00, 72.06it/s]

	precision	recall	f1-score	support
Bold Reds	0.75	0.71	0.73	739
Full-bodied Whites	0.80	0.79	0.80	372
Light-bodied, Crisp Whites	0.81	0.77	0.79	283
Medium to Full-bodied Reds	0.79	0.83	0.81	1106
accuracy			0.78	2500
macro avg	0.79	0.78	0.78	2500
weighted avg	0.78	0.78	0.78	2500

Results on validation set –

100%  5000/5000 [01:09<00:00, 71.88it/s]

	precision	recall	f1-score	support
Bold Reds	0.75	0.75	0.75	1477
Full-bodied Whites	0.81	0.81	0.81	745
Light-bodied, Crisp Whites	0.80	0.82	0.81	567
Medium to Full-bodied Reds	0.81	0.81	0.81	2211
accuracy			0.79	5000
macro avg	0.79	0.80	0.80	5000
weighted avg	0.79	0.79	0.79	5000

## **CONCLUSION:**

From the above results we have the four best classifier along list in the order of descending macro average f1 score on validation set:

1. RoBERTa (0.80)
2. DistilBERT (0.79)
3. TFIDF Vectorization + Linear SVC (with hyperparameter tuning) (0.78)
4. CNN (0.77)

We can conclude two things from the above analysis:

1. Given the size of the training set, the transfer learning algorithms(RoBERTa and DistilBERT) are likely to provide much better results as seen in the table above.
2. Given the class imbalance in the dataset, the best way to group the categories is on the basis of domain knowledge as stated above. Grouping on the basis of taste and flavour is more appropriate when building a wine recommendation system rather looking at the distribution of target variables. This has led to a significant improvement in results improving classification accuracy from low 70s to almost 80%.
3. Although our model has shown a significant improvement in results from the baseline SVC model, the macro f1 score does not go above 80% even after working with

multiple models. This is a clear indication that we need more training data to improve our classification report.

### **ERROR ANALYSIS:**

We have used the RoBERTa model for performing error analysis using SHAP. We have taken a sample of 30 mis-predicted observations from the provided test set of sample size 500 for this analysis.

	wine_group	error_count
0	Medium to Full-bodied Reds	22
1	Bold Reds	21
2	Light-bodied, Crisp Whites	9
3	Full-bodied Whites	8

The above table shows that the most mis-predicted category was “Medium to Full-bodied Reds” i.e Pinot Noir, merlot and Syrah with appearing either as the actual label predicted as something else or the wrongly predicted label with the actual label as something else. This was followed by “Bold Reds” i.e Cabernet Sauvignon and Zinfandel.

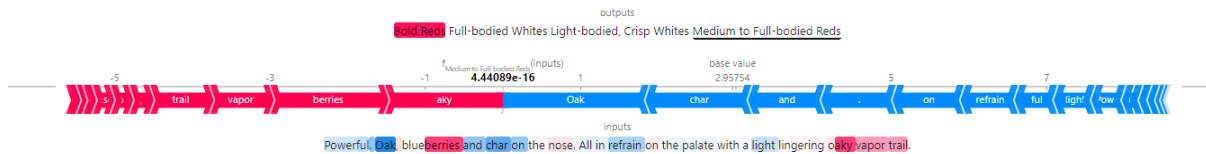
This leads us to the following table where we have tried to track the distribution of mis-predicted combinations.

	combination	count
0	Medium to Full-bodied Reds+Bold Reds	11
1	Bold Reds+Medium to Full-bodied Reds	9
2	Light-bodied, Crisp Whites+Full-bodied Whites	4
3	Full-bodied Whites+Light-bodied, Crisp Whites	3
4	Medium to Full-bodied Reds+Light-bodied, Crisp Whites	1
5	Bold Reds+Light-bodied, Crisp Whites	1
6	Medium to Full-bodied Reds+Full-bodied Whites	1

As per our guess from the first table, our classifier is confusing the two categories “Medium to Full-bodied Reds” and “Bold Reds” as evident from 11 and 9 mis predicted observations out of the sample of 30 (63.33%).

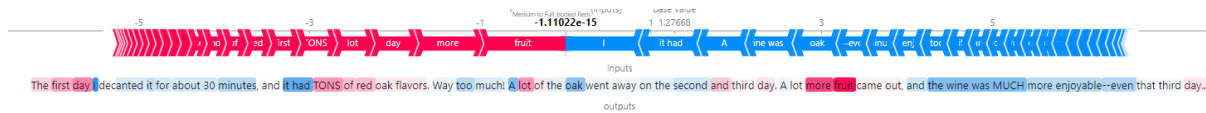
We have then used SHAP to identify the points on confusion in the text we are analysing and the results are as follows. We will look into a few samples for our report, for a model detailed analysis please refer to the code.

Example 1: “Medium to Full-bodied Reds” classified as “Bold Reds”



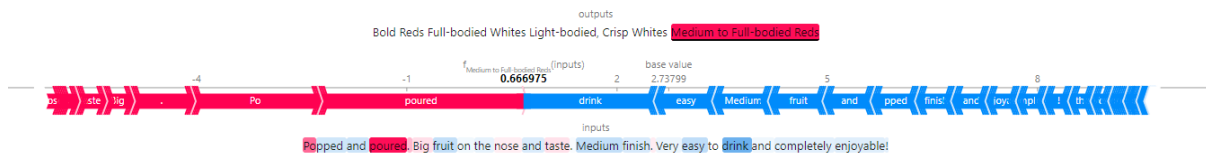
While words like “light” and “oak” incline the results towards “Medium to Full-bodied Reds”, the final outcome seems to be influenced by the use of “powerful”, “refrain” and “berries”.

### Example 2: “Medium to Full-bodied Reds” classified as “Bold Reds”



In this example we see that the use of words like, “TONS” and “more fruit” has pushed the classifier to predict “Bold Red”

### Example 3: “Bold Reds” classified as “Medium to Full-bodied Reds”



In the given scenario, the word “medium” clearly influences the result

### Example 4 : “Light-bodied, Crisp Whites” classified as “Full-bodied Whites”



The use of the word “champagne” which is a “Full-bodied white” has stirred the prediction to be as such.

From the above analysis we see errors that are primarily domain knowledge related. However, in the reviews we also have text that are redundant and do not contribute to the classification with respect to taste of quality of wine as seen below. Hence, a recommendation from this would be to carefully curate samples that are used to train the wine-recommendation model in order to obtain more accurate results.

