INFO6105 – FINAL PROJECT

Wine dataset - https://www.kaggle.com/datasets/zynicide/wine-reviews

a) **Problem Statement:** The objective is to build a machine learning model that can accurately classify the type or variety of wine based on various attributes and features associated with each wine. The target variable for the classification is the 'variety' column, which contains the specific wine varieties, such as 'Chardonnay', 'Pinot Noir', 'Cabernet Sauvignon', and others.

Significance:

- Classifying wine varieties can be beneficial for organizing and categorizing a large collection of wines in databases or wine cellars.
- Can be used in wine recommendation systems to suggest wines to consumers based on their preferences.
- Can be used to educate consumers about the different types of wines available and help them make more informed choices when purchasing wines, matching wines with their preferences, and exploring new varieties.
- Understanding popularity of wine varieties can be crucial for market analysis, allowing wine producers and sellers to identify trends and make informed decisions about their product offerings and marketing strategies.

b) **Rationale of the target variable:** The target variable of a dataset is the feature of a dataset about which you want to gain a deeper understanding. The 'variety' column is a fundamental attribute of wine dataset. The type of grape or blend used to produce a wine has a significant impact on its flavor, aroma, and characteristics. It is a key feature that wine enthusiasts, sommeliers, and wine consumers often consider when choosing or reviewing wines. By predicting the wine variety, the model can provide valuable insights into which features are more indicative of a specific wine type. This can be useful for various applications, such as automating wine categorization, assisting in wine recommendation systems, or identifying mislabeled wines.

c) Supervised ML involves creating models that can learn from labelled data to make predictions. This classification problem maps features (description selected by RFC) to the target variable (Wine Variety) in the dataset. The various ML algorithms we used are as follows:

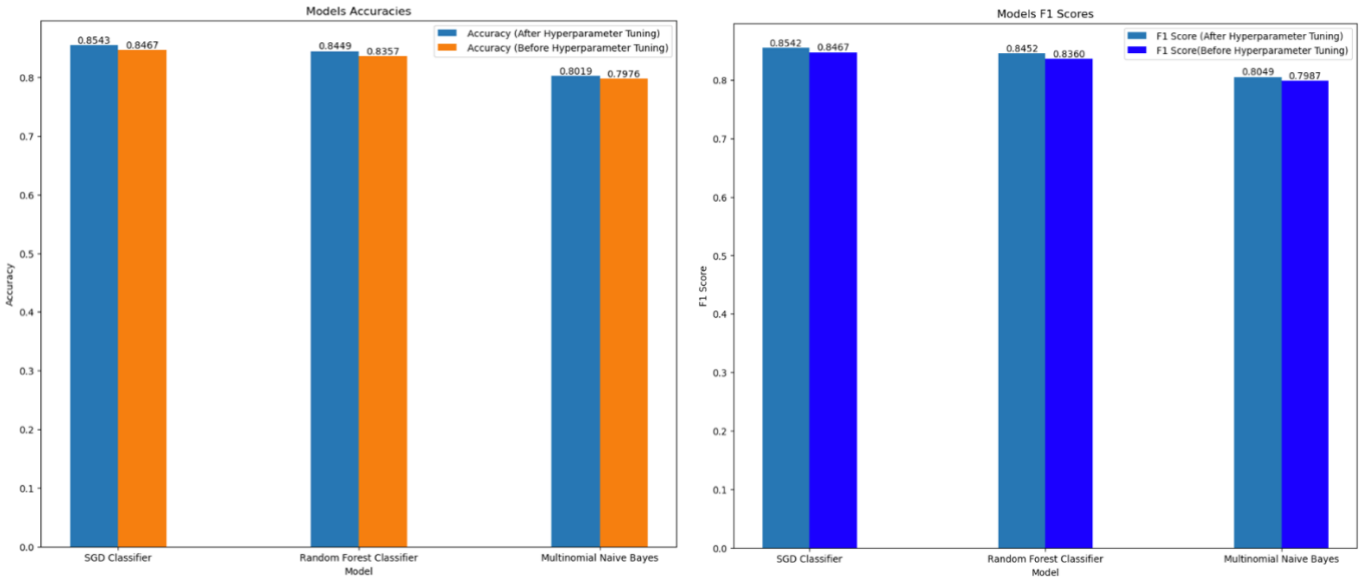
Stochastic Gradient Descent (SGD) Classifier: The SGD Classifier is optimized for large-scale machine learning tasks and is particularly effective for text classification and natural language processing applications due to its capability to handle sparse data and high dimensionality.

Random Forest Classifier (RFC): The Random Forest Classifier, an ensemble learning method, is known for its robustness and high performance in dealing with high-dimensional spaces and large training sets.

Multinomial Naive Bayes (NB): The Multinomial Naive Bayes classifier is often used in text classification tasks where the data is represented as word vector counts. It is particularly suitable for tasks with discrete features.

Among these, the Stochastic Gradient Descent (SGD) Classifier showed the best performance compared to other models.

d) **Conclusion:** The accuracy and F1 scores of all the models (Stochastic Gradient Descent, Random Forest Classifier, Multinomial Naive Bayes) with default hyperparameters and after hyperparameter tuning are shown in the images below. After hyperparameter tuning, the models perform significantly better in terms of accuracy, F1 score, etc., across all the models. Particularly, the Stochastic Gradient Descent model achieves the highest accuracy of 85.43% and F1 score of 85.42% and outperforms the other models.



e) **Scope for future work:**

- Try different or advanced machine learning models such as deep learning models (CNNs, RNNs, Transformers) or ensemble methods (e.g., Gradient Boosting, XGBoost).
- Handling descriptions in multiple languages
- As data grows, we can explore continual learning approaches to update the model periodically with new data
- The trained model can be deployed as a web service or API to enable real-time predictions or integration with other applications.

f) **Additional Efforts:** In addition to the core tasks of data preprocessing, model training, and performance evaluation, we conducted **feature analysis** to identify the most important features for the target wine variety. This step enhances the project's comprehensiveness and usefulness by deepening our understanding of influential features, potentially improving model accuracy. We fitted our dataset to three different ML models: **SGDC, RFC, and NBC**, and performed a comparative performance evaluation. Among them, SGDC demonstrated superior performance. Our team's effort in recognizing the most influential features, displaying a proactive approach, and ensuring continual progression justifies allocating extra points due to exceeding standard project requirements.