Observations

Decision Tree	70.15%	Grid Search CV
K-Nearest Neighbors	83.72093023 %	Random Search CV
Logistic Regression	62.41 %	Grid Search CV
Support Vector Machine	67.83 %	Grid Search CV
Naive Bayes	55.81 %	-
Random Forest	76.62 %	Random Search CV
Multi Layer Perceptron	76.74 %	Grid Search CV
XGBoost	83.72093023 %	Grid Search CV

Naive Bayes

The Naive Bayes Classifier is often used in text classification applications and experiments because of its simplicity and effectiveness. Due to lack of reliable and confident scores, it does not model well.

TECHNIQUES TO IMPROVE BAYES:

> https://link.springer.com/chapter/10.1007/978-3-540-30586-6_76#:~:text=Naive Bayes is often used,lack of reliable confidence scores.

Even the scikit-learn documentation page says that : “On the flip side, although naive Bayes is known as a decent classifier, **it is known to be a bad estimator**, so the probability outputs from predict_proba are not to be taken too seriously”

But our problem does not classify text. Our problem statement is a multi way classification of glass types which are numeric.

Cites:

S. Hassan, M. Rafi and M. S. Shaikh, "Comparing SVM and naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment," 2011 IEEE 14th

International Multitopic Conference, Karachi, Pakistan, 2011, pp. 31-34, doi: 10.1109/INMIC.2011.6151495.

- Feature Independence : There is a possibility that the naive bayes algorithm assumed that the features are independent. However in the glass classification dataset , this assumption is clearly violated.
- We have **Continuous Features** in our dataset. This might have lead to the false assumptions by the model and maybe, gaussian Naive Bayes or other non-parametric classifiers can help in this case
- **Complex decision boundaries -** We can see that the features are having very close values from one type to another.
- Sensitive to Outliers.

Support Vector Machines

Good :

- Domain Complexity : Considering a very close call in classification, 67% is a very good score.
- We have a lot of rows on type 1 and type 2 glasses as compared to type 3 and type 5 glasses , which is clearly an imbalanced class distribution.
- We have about 9 features, which is pretty a decent number and SVMs can work well with large features.

Bad :

- The dataset of 215 rows are maybe not sufficient for the correct identification of the respective glass types.
- Underfitting might have been an issue as there are just 215 rows.
- The SVM might have assumed that the data is linearly separable or can be transformed into a higher dimensional space for separation.

K-Nearest Neighbors:

- KNN is non-parametric algorithm, so it can adapt to complex complex decision boundaries.
- The 81.39 % accuracy is also due to its lazy learning property. If does not explicitly build a model during the training phase. It memorizes the training data and makes decisions based on local patterns.
- Even if the data points are close to each other, we it can easily group them up and effectively capture the patterns.

However, it has got limitations.

- KNN needs to compute distances and find neighbors for each and every iteration, which can be very computationally expensive
- Its very sensitive to noise and outliers as KNNs are very influential by the outliers. In this dataset, the outliers are not very explicitly seen.
- The imbalanced data comes to play. Similar to SVM, this suffers from this problem. Hence, that loss of 18% might have been the cost of imbalanced data.

DOC TO BE UPDATED - KNN and XGBoost