Quality Prediction

1.Data Preparation (What steps would you take to prepare your data and why?)

- We first import all the libraries and their sub -Libraries like NumPy and Pandas.
- Then we import dataset and examine the same to make sure to see all the columns of the output window or is there any missing value or any categorical feature to convert to numerical value
- We find our target variable which is **"is safe"** and is in a categorical feature state so we convert it into numerical feature to put it in common range of values.
- Then we divided dataset into label and feature sets by dropping target variable "**is safe**" from **X-axis** which was our target variable and putting it in **Y-axis** so to explicitly define which variable were independent and which is our target variable.
- Next we standardise our independent variables using standard scaler so to bring all the columns to a common range.
- Then balancing and tunning the parameters while choosing the metric accordingly.

2. Model Hyperparameter Tuning (Which hyperparameters would you tune and why? How would you tune them?

RFC-

- We first tune RFC hyperparameter n_ estimators and implementing cross-validation using grid Search .
- We will use n_ estimators as it contains multiple random forest trees which gives us the decision using fivefold cross validation .
- First, created multiple training set using bootstrap sampling starting with 10 decision trees and taking it till 300 using grid search, where we found 150 decision trees gives us the optimal score of 0.69.
- After that tried top 6 parameters variables, the result was 0.64 on 200 decision trees which is a lower score.

SVC-

- Tuned the SVC under Kernel parameter like linear, poly, rbf , sigmoid and classification _C- [.001,.01,.1,1,10,100] and implementing cross-validation using grid search.
- SVC is used to separate non-linear and linear data through its different mathematical function
- In this we used **recall** to minimise false negative.
- The best score was 0.73 on classification _ kernel – "Sigmoid" and classification _C – "0.001".

Quality Prediction

3. Choice of Evaluation Metric (Which metric would be suitable for model evaluation and why)?

- As per the results the suitable model would be **SVC**. The reason is the score of this model is 0.73 compared to RFC model which is 0.64 .
- We used **recall** metric to minimise false negative as we do not want unsafe water to be pretended as safe water for drinking which will be the worst-case scenario.

4. Overfitting avoidance mechanism (Which mechanism (feature Selection/ regularization) would you use and why)?

In RFC we used feature selection so that overfitting can be avoided by splitting data into test set and then by tuning hyperparameter while doing cross validation which minimise the overfitting to improve the score .The insights we get, helps to select the top parameters of the model which majorly contributes in the results.

5. Results analysis

a). Which of the two models (random forest or support vector classifier) would you recommend for deployment in the real-world?

Random Forest includes managing big datasets with lots of features and SVC requires data with a limited number of features and this algorithm handles both linear and non-linear data as well as multi-class classification problems. Therefore, we would like to go with SVC model for deployment in the real world.

b). Is any model underfitting? If yes, what could be the possible reasons?

Yes, both the modals are underfitting as the data provided is not enough as to predict the quality safe as both scores were less than 0.90 and when we took top 6 the score decreased.