**Water Potability Modelling Report**

**Introduction:** The data contains 3276 rows of data and 10 columns. Each row represents a different water body with the target variable being ‘Potability’ which is potable if 1 and non-potabile if 0. This data is important as drinking water is a human right and access of it should be a priority for any government of humanitarian organization.

**Data exploration:** I first checked for nonvalues and found some in the ‘ph’, ‘Sulfates’ and 'Trihalomethanes' variables. These are replaced with the average of the variable values which share the same output. Replacing nonvalues rather than removing them ensures all information is used making a stronger model. I then plotted a distribution plot and a boxplot against the ‘Potability’ variable for each variable. The distribution plots show that each variable has a different range of values, so standardization is needed. Most variables appear to not be normal which suggests that linear regression or Gaussian Naive models may not be effective as they assume normality for each variable. The boxplots suggest the distribution of potable and non-potable appear to be similar for each variable. This suggests a linear model may not be effective. This is reconfirmed with the pair plot. This suggests that either a decision tree, Kneighbors or non-linear support-vector machine will be our strongest model. After standardizing the data, I fitted it with a random forest classifier. This was to use the feature importance function to rank the variables by information gained. All variables appear to be useful especially ‘ph’ and ‘Sulfates’. Due to its speed, we use a Gaussian naïve model and one by one by importance we add a variable and evaluate the model. This is to decide which variables to use in our model. For this data adding all the variables does not significantly lower the accuracy of the model and so I will keep all the data.

**Modelling:** The data is split into a training and validation set. Using the training data, we evaluate basic algorithms; logistic regression, linear discriminant analysis, k neighbors, decision tree, gaussian naïve and support-vector machine. Cross validation is used to prevent overfitting as different data is used to train and test and it also tests each model multiple times to take an average of the results which identifies outliers. A decision tree model is our strongest model as predicted in the date exploration section. I can improve our decision tree model by ensemble and in particular boosting or bagging. After checking some ensemble models with the same method as before we have 2 strong models, random forest a bagging model and gradient boosting a boosting model which I will optimize and select the stronger model. This is done with grid search and checking combinations of variables. For random forest number of estimators is not searched to save computational time. After optimization both models are compared by cross validation. The strongest model ends up being random forest and as such it will be used for the validation.

**Validation:** The random forest model with optimized variables is used to create a model using the training data. This model is now tested with the validation data. The accuracy of the model is about 80.5% which appears to be strong. Using a confusion matrix and a classification report we can investigate where the model’s limitations are. From the confusion matrix we can see that 106 errors are false non potables and 22 are false potables. This is seen in the classification report with the high recall score for non potability and a high precision score for potability. The model’s limitation is that if often predicts potable water as non-potable.

**Conclusion:** In this report we have tidied a data source and explored its properties for data modelling. After checking a range of models, the best were selected and optimized. From this I found the strongest model and validated it. The model selected; random forest is a combination of many decision trees which was our strongest basic model suggesting why it did so well. The accuracy of around 80.5% is the strongest model I could find on Kaggle suggesting this is a very strong model. The limitation of the model being over predicting non-potability could be viewed as a strength as using this model in a practical sense would reduce the chance of a dangerous water source being used. In fact, an improvement of this model could be to remove all false potabile results. This would reduce the chance of anyone using a deadly water source at the cost of the overall accuracy of the model and as such some potable water sources not being used.