Machine Learning Supplementary Details

No intermediary results – On one occasion we accidentally label the data wrong in matlab for train-test split and get a very large confusion matrix. The error was we accidentally standarised the y variable in python code.

**Glossary**

**Random Forest :** A supervised machine learning algorithm, which combines the use of multiple decision trees to make an average solution.

**KNN :** A supervised machine learning algorithm, which is more simple and calculates the difference

**n (number of neighbours) :** The K parameter symbolises the number of K neighbours which are close to the “c” to place it within a class, this is for the KNN.

**Precision :** The accuracy of positive predictions, Precision = True Positives / True Positives + False Positives

**Recall :** Recall measures the ability of a model to identify all relevant instances. It is the ratio of correctly predicted positive observations to all observations in an actual class. Recall = True Positives/ True Positives + False Negatives

**Accuracy :** This generally measures the accuracy of all predicted observations to the total observations, however it can be misleading when dealing with imbalanced classes.

Accuracy = True Positives + True Negatives / Total Observations

**F1-Score :** This is the harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when the class distribution is imbalanced. A high F1-score means the model has low false positives and low false negatives.

F1 – Score = 2 x Precision x Recall/ Precision + Recall

**Cohen Kappa Coefficient** : This is a statistic that measures the inter-rater agreement for qualitative categorical items. It is a more robust measure than simple percent agreement calculation. K = P0 – Pc/ 1 – Pc where Po is the relative observed agreement among rates and Pe is the hypothetical probability of chance agreement.

**Implementation Details**

**Other Model Architectures:**

Naïve Bayes: We did not choose, this model architecture, but this was most considered. We go with KNN and RF because they are best suited to complex multiclassification data. Furthermore, Naïve Bayes is better for textual data.

Decision Tree: We assume that decision trees would be better for binary classification and not for multiclassification, where random forest is a better version of these ensemble models.

**Parameterisations:**

KNN Distance Metric Hyperparameter: There are multiple ways the KNN can calculate the distance between a new point “c” and the number of points within the sample space. We choose Euclidean distance and not perform hyperparameter tuning on this as it is beyond the scope of the project.

**Issues of Errors:**

A blue and yellow squares with black and yellow squares

Description automatically generated