Water Potability Classification

Tan Xiao Xuan

Oral: https://imperial.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=19e62f57-013a-4a9d-a100-ad480146ea81

1. Introduction

1.1 Objective

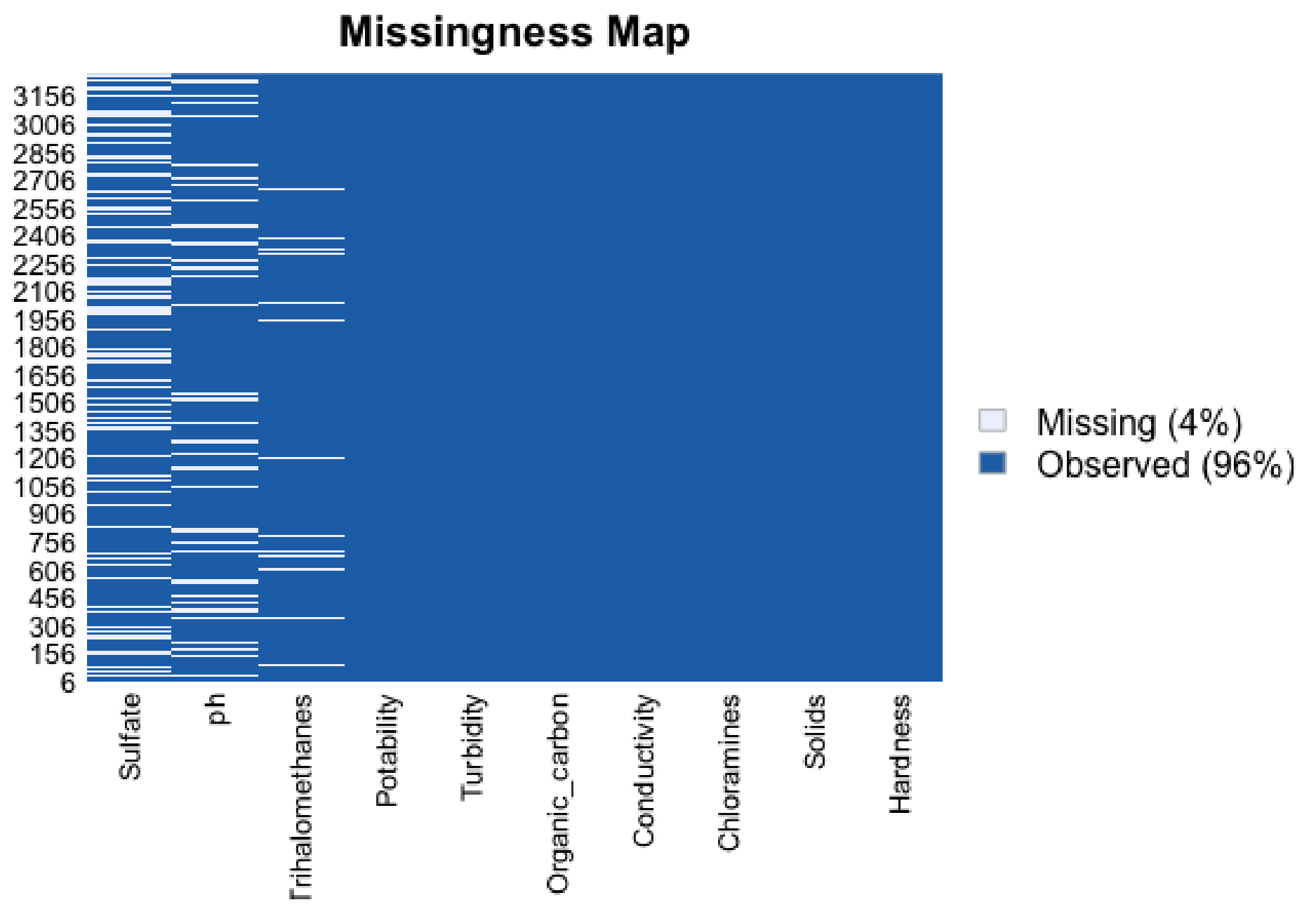
The purpose of this project is to conduct an experimental study on different classification methods using the water potability data set and recommend the preferred classifier for the data set.

1.2 Data source [2]

The data set is obtained from kaggle, consisting of information about 3276 different water sources. It contains 9 features, namely pH value, hardness, total solids dissolved, chloramines, sulfate, conductivity, total organic carbon, trihalomethanes and turbidity. The water bodies are distinguished into 2 classes, where 0 implies Not Potable and 1 implies Potable.

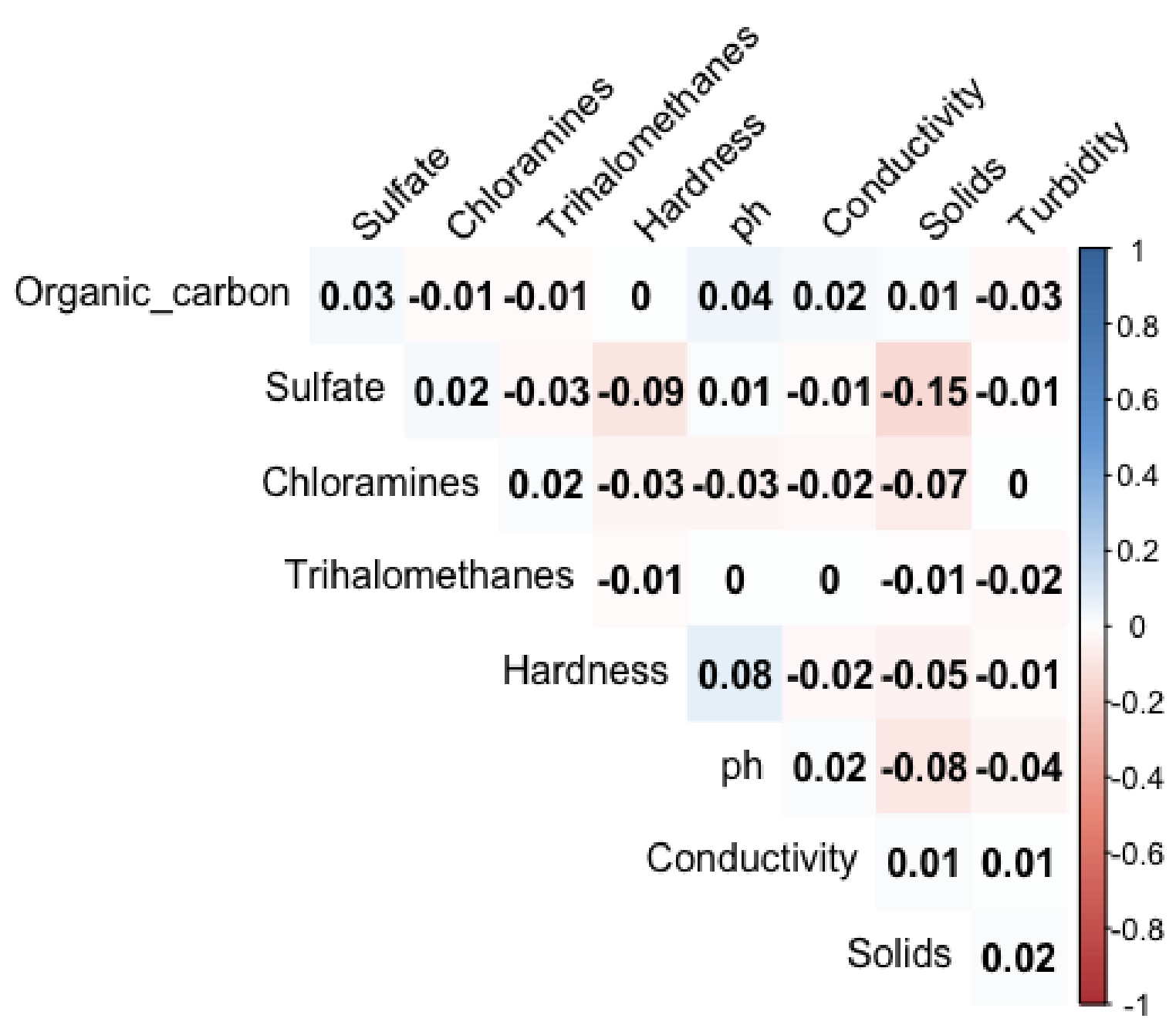
2. Data preparation

2.1 Handling missing data [4]



The data set contains missing values for the features Sulfate, pH and Trihalomethanes. Imputation is an approach to replace missing values with estimations such as mean and mode or values obtained by an estimated distribution of the feature. Here the missing values are replaced by the means of the respective features.

2.2 Removing redundant features [5]



The correlation matrix above shows that the the correlation between the features in the data set are not highly correlated, implying the features are not redundant.

2.3 Normalising the data [1]

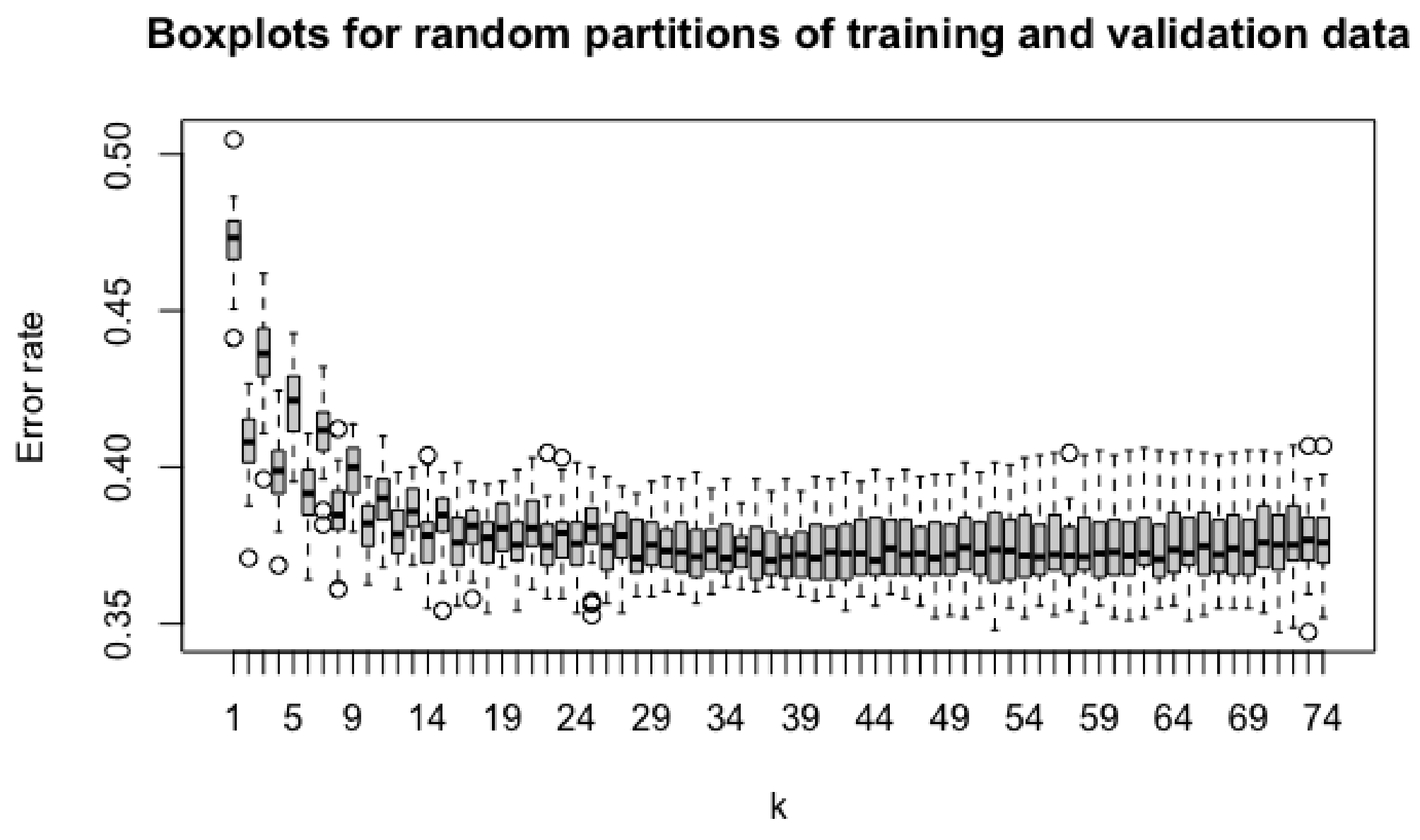
A summary of the data set shows that the values range from zero to hundreds. To improve the performance of the classifiers, the data are normalised using the formula $x' = (x - X_{min})/(X_{max} - X_{min})$.

2.4 Splitting the data

The data set is split into training and test set with respect to ratio 80:20. The training set contains 2620 data. The test set contains 656 data.

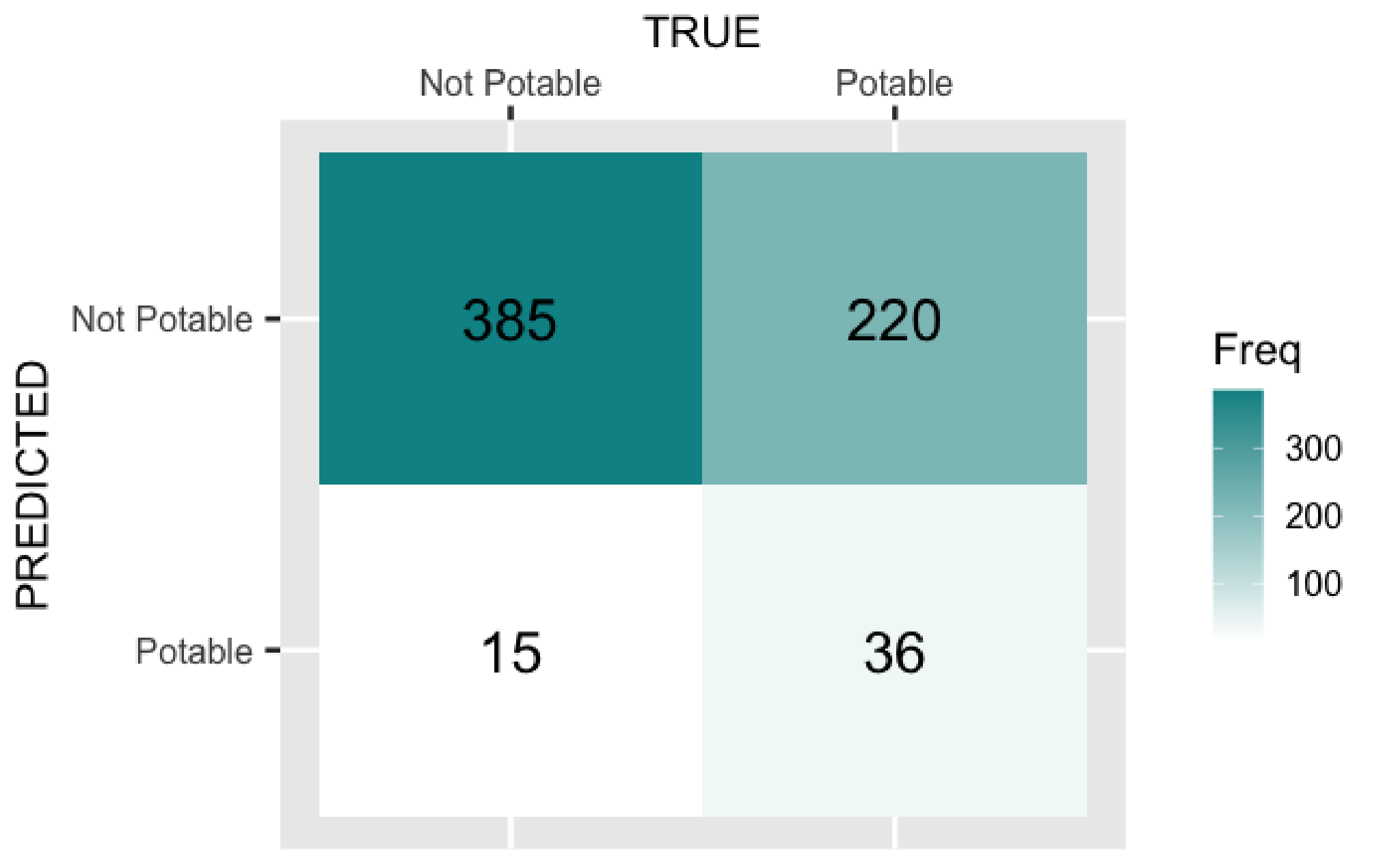
3. K Nearest Neighbours

3.1 Choosing best k value



The training set is randomly split into training and validation set with respect to ratio 50:50 for 30 times and the error rates are calculated. The diagram above shows that k=37 has the lowest median error rate (0.3702) and thus is chosen.

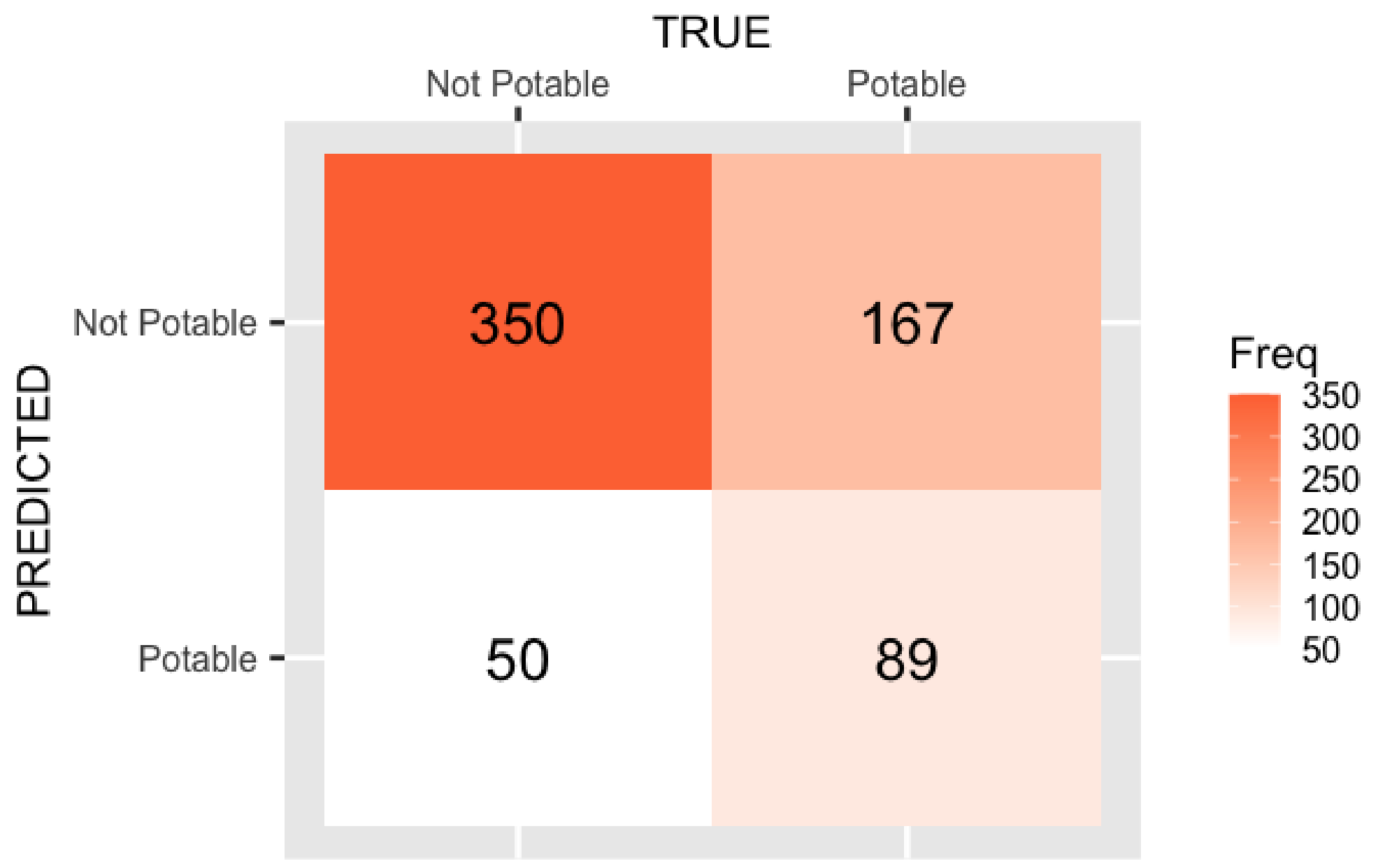
3.2 Confusion matrix for test set



Setting k=37, the error rate for the training set is 0.3416. From the confusion matrix above, the error rate for the test set is 0.3582.

4. Random Forest

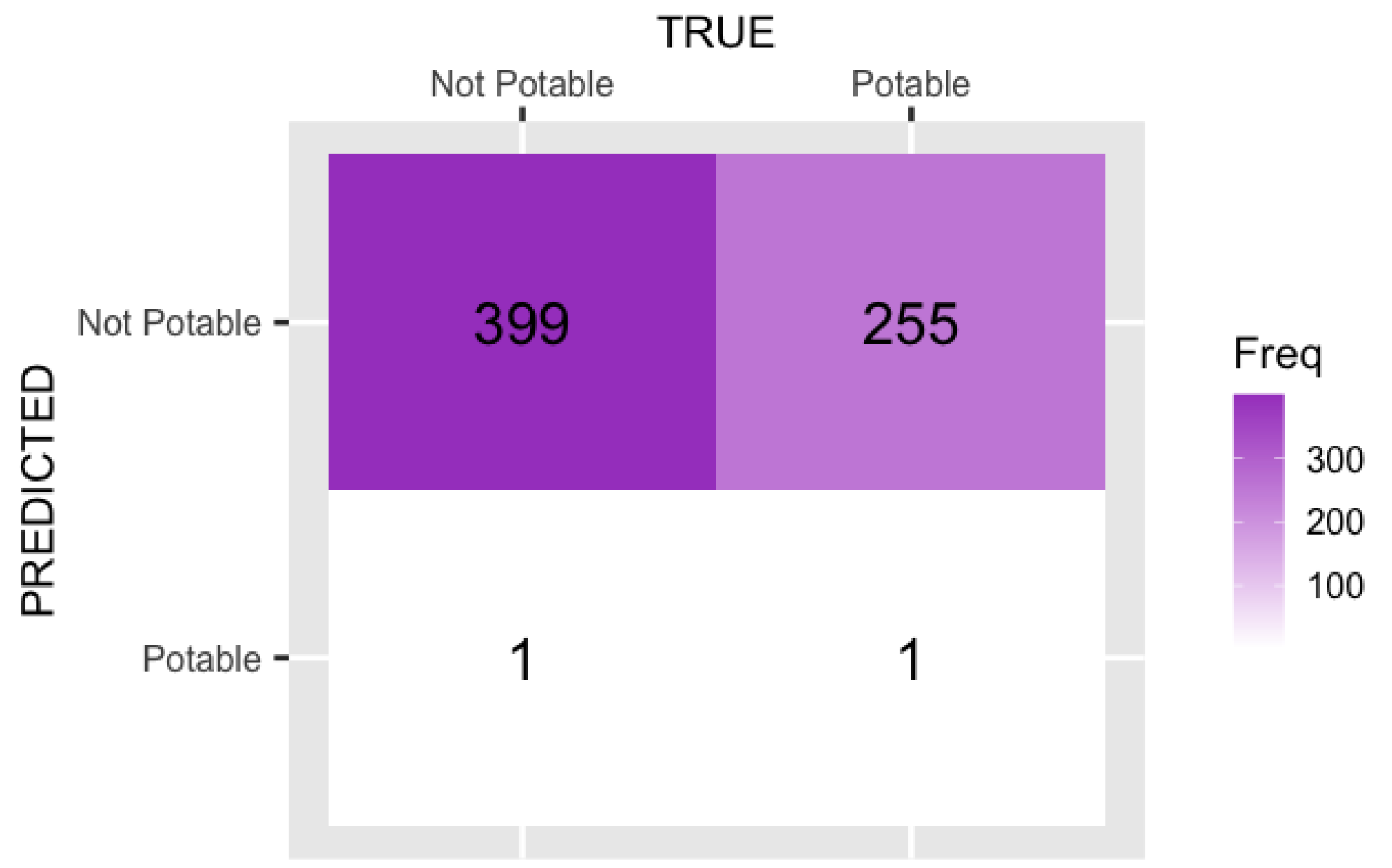
Confusion matrix for test set



The error rate for the training set is 0.3260. From the confusion matrix above, the error rate for the test set is 0.3308.

5. Logistic Regression

Confusion matrix for test set

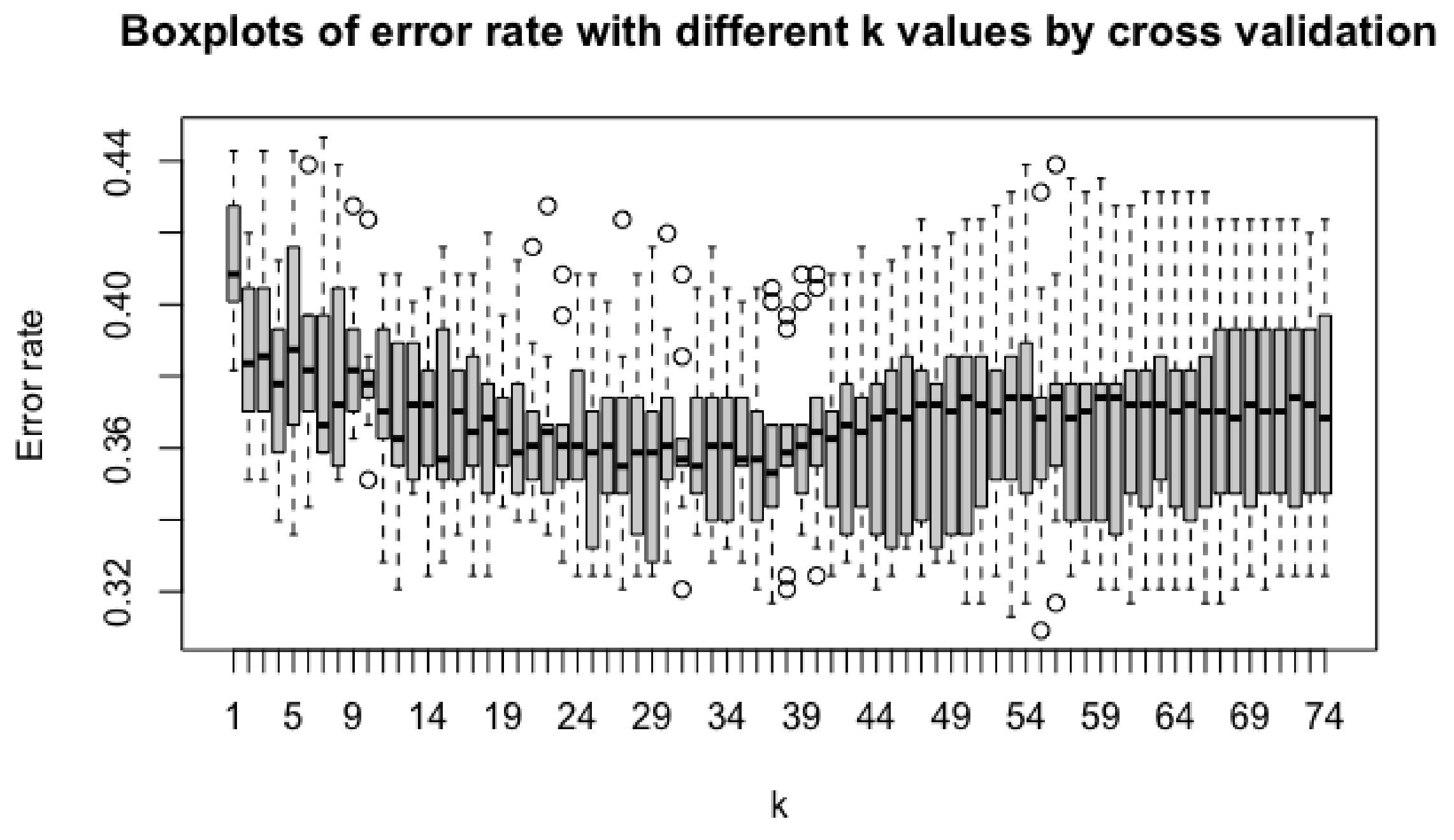


The error rate for the training set is 0.3897. From the confusion matrix above, the error rate for the test set is 0.3902.

6. Cross Validation

10-Fold cross validation is applied to each of the classifiers.

6.1 K Nearest Neighbours



From the diagram above, the k value which has the lowest median error rate is k=37 again. The error rate on the training data by cross validation is 0.3531 while the standard error is 0.0087.

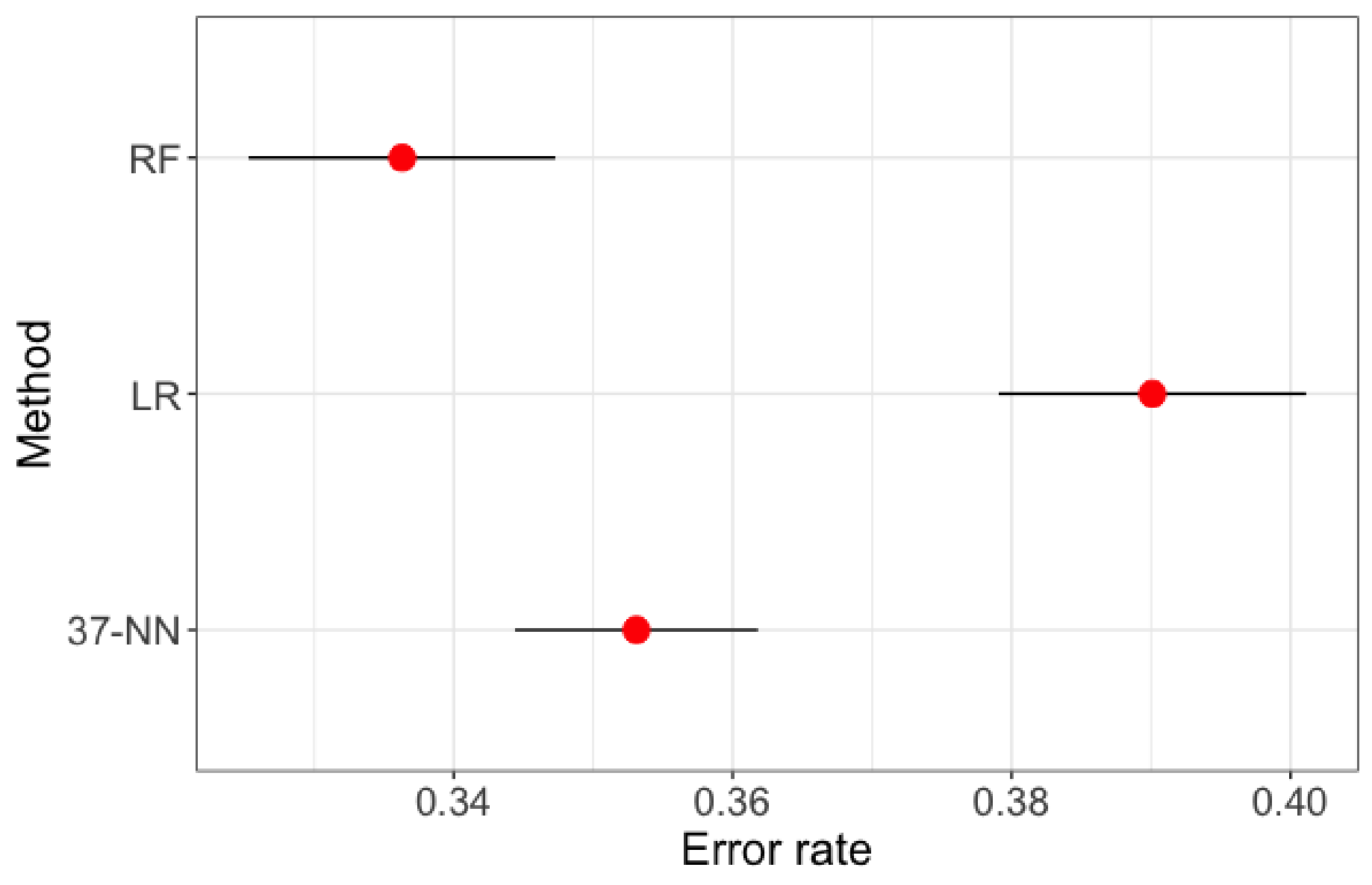
6.2 Random Forest

The error rate on the training data by cross validation is 0.3363 while the standard error is 0.011.

6.3 Logistic Regression

The error rate on the training data by cross validation is 0.3901 while the standard error is 0.011.

6.4 Summary



7. Conclusion

7.1 Summary

	Error rate		
Method	Train	CV on train (SE)	Test
37-NN	0.3416	0.3531 (0.0087)	0.3582
RF	0.3260	0.3363 (0.011)	0.3308
LR	0.3897	0.3901 (0.011)	0.3902

7.2 Hypothesis testing [3]

Applying the McNemar's test, the null hypothesis is that the population error rate of 37-NN is the same as the population error rate of random forest. Using $n_{37-NN} = 62$ and $n_{RF} = 44$, the test statistic is $\frac{|n_{37-NN} - n_{RF}| - 1}{\sqrt{n_{37-NN} + n_{RF}}} = 1.651$. The p-value=0.0987>0.05=α, where α is the significance level. There is insufficient evidence to reject the null hypothesis.

7.3 Recommendation

The confusion matrix for logistic regression shows that it is not efficient in classifying potable water sources. It also has the highest error rate for training and test data among all the classifiers. Thus, it is eliminated.

The McNemar's test shows that there is no statistically significant difference between 37-NN and random forest. However, the error rate for random forest is lower than 37-NN for training set, cross validation on training set as well as test set. Furthermore, the random forest classifier can be further improved by tuning the hyperparameters. Hence, I would recommend random forest classifier for the water potability data set.

8. References

[1] K. Doherty, R. Adams, and N. Davey. "Non-Euclidean norms and data normalisation". In: *ESANN'2004 proceedings - European Symposium on Artificial Neural Networks Bruges (Belgium)*. 2004, pp. 181–186.

[2] Aditya Kadiwal. *Water Quality Version 3*. Available from: https://www.kaggle.com/adityakadiwal/water-potability [Accessed 15/06/2021]. 2021.

[3] Brian D. Ripley. "Statistical Decision Theory". In: *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996, pp. 17–90. DOI: 10.1017/CBO9780511812651.003.

[4] Maytal Saar-Tsechansky and Foster Provost. "Handling missing values when applying classification models". In: *Journal of Machine Learning Research 8*. 2007, pp. 1625–1657.

[5] Lei Yu and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution". In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, pp. 856–863.