Water Potability Prediction with Machine Learning

Darshika Mishra[1], Constance Ferragu[2]

Supporting Information

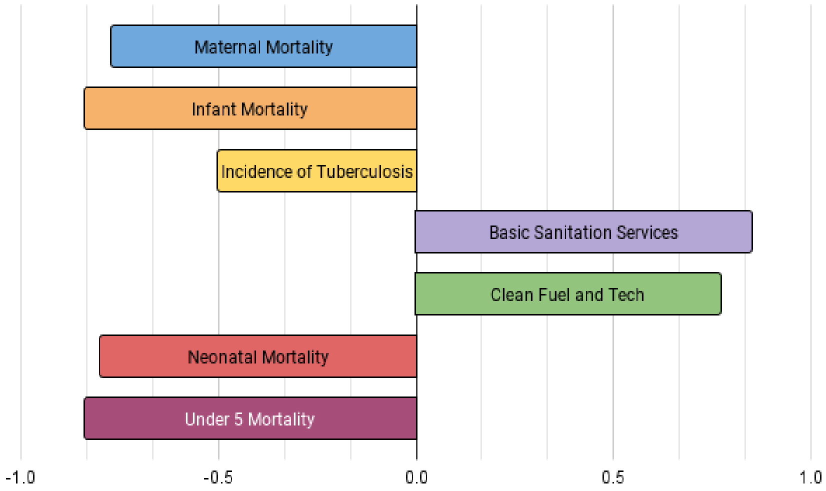

Figure S1. Visual representation of pairwise correlations within dataset. Offers a clearer picture of the correlations depicted in (**Fig 1).**

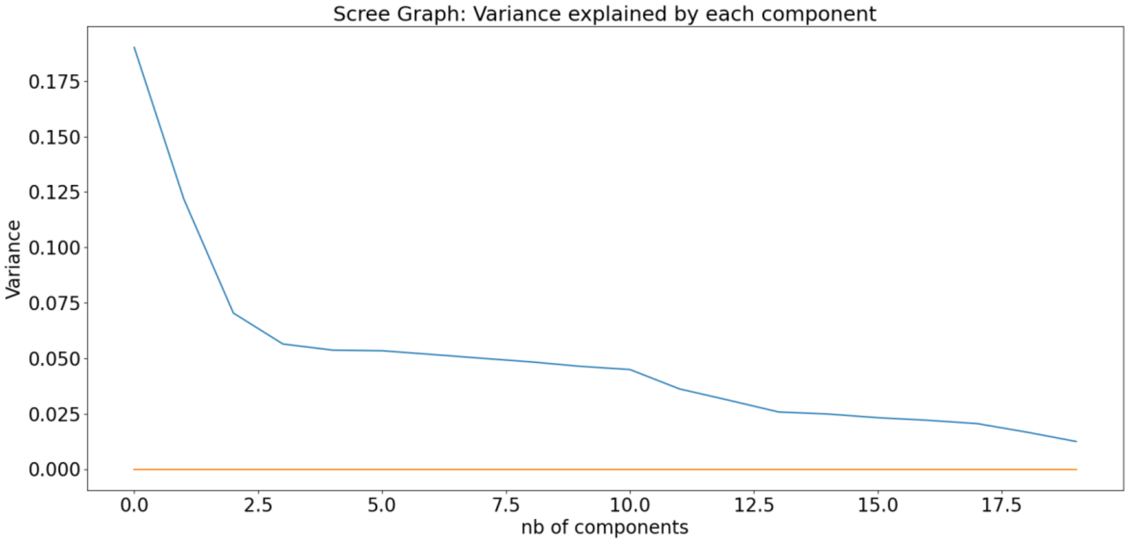

Figure S2. Variance graph that accounts for the variability that the different components of the dataset account for. Indicates that the dimensionality of the dataset can be reduced using PCA Analysis.

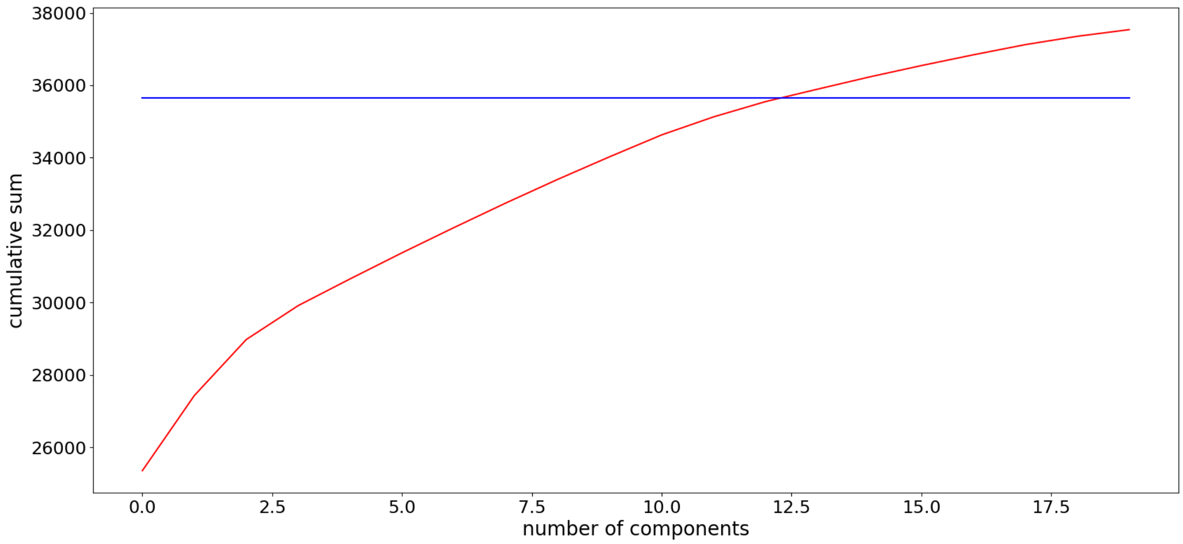

Figure S3. Cumulative graph plotting cumulative sum of the eigenvalues. Around 12 components are necessary to capture 95% of the variance (x-value where red and blue line intersect).

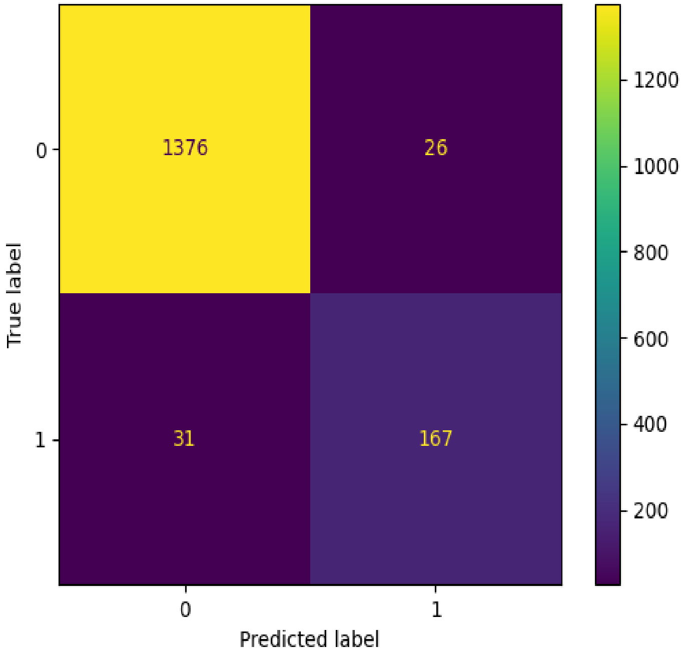

Figure S4. Decision Tree Classifier Confusion Matrix. Depicts the number of type I errors, or false positives (top right box) and type II errors, or false negatives (bottom left box).

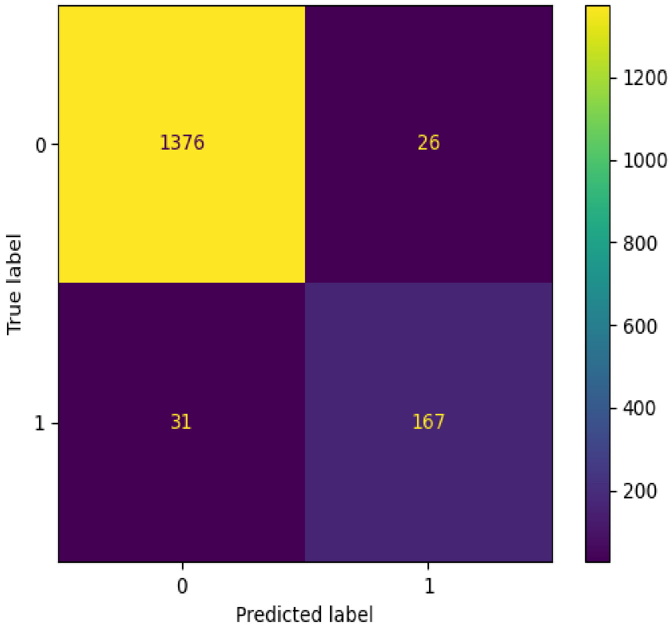

Figure S5. XGB Confusion Matrix. Depicts the number of type I errors, or false positives (top right box) and type II errors, or false negatives (bottom left box).

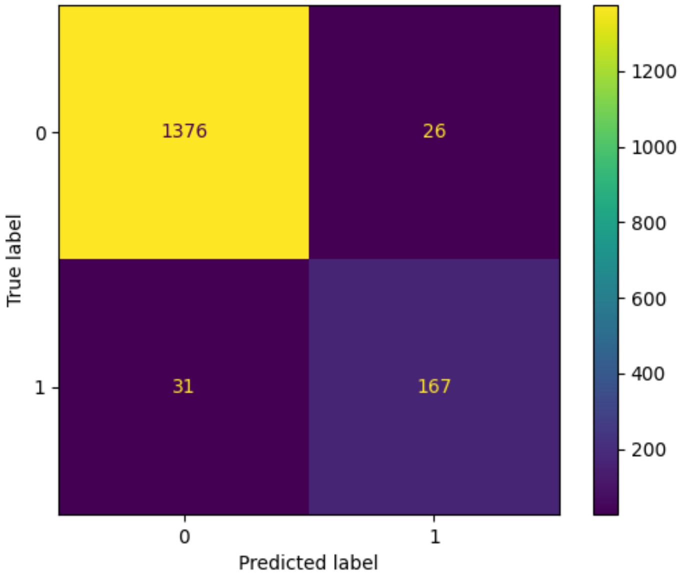

Figure S6. Random Forest Confusion Matrix. Depicts the number of type I errors, or false positives (top right box) and type II errors, or false negatives (bottom left box).

Table S1. Summary of Model Accuracies. The accuracies and weighted accuracies each model produced.

Model	Normal Accuracies	Weighted Accuracies
XGB Classifier with Optimal Parameters	96.38%	89.5%
Decision Tree Classifier with Optimal Parameters	96.44%	87%
Random Forest Classifier with Optimal Parameters	95.88%	82%

Table S2. Optimal Parameters. These are the optimal parameters determined from performing the grid search and used to achieve the best possible accuracies.

Model	Optimal Parameters
XGB Classifier	'n_estimators': 750, 'learning_rate': 0.1
Decision Tree Classifier	'criterion': 'entropy', 'max_depth': 91, 'min_samples_leaf': 10
Random Forest Classifier	'min_samples_leaf': 1, 'n_estimators': 350