**Brief Experimental Design Summary.**

We endeavoured to find the best ensemble learning algorithm for the Arrhythmia, Caesarean and Website-Phishing datasets respectively, out of the Decision Tree, Random Forest, Bagging and AdaBoost methods. We began by splitting each dataset into their feature set and target set. For all datasets, the target was the final column ‘class’.

From this, we distinguished the training and test set at an 80:20 split. Splitting the data into 80% training and 20% testing means the model will assess more known data and hopefully come to an accurate solution, when tested against the unseen data in the test set.

We run the ensemble algorithms on the training set after, derive an accuracy score for each algorithm and then cross-validate. Cross-validation evaluates models on the unseen data in the test set, by a resampling procedure that summarises the skill of the model using the sample of model evaluation scores. Parameter k=10 was chosen as it is a common tactic that has been found to generally result in a model skill estimate with low bias and a modest variance.

Finally, we created a data frame comprised of the evaluation scores of each algorithm and their standard deviations to pass through autorank. Autorank (in particular, the Frequentist approach) determines if the data are normal, the populations are homogenous (equal variances) and the number of populations and selects the appropriate statistical from there to procure the most statistically significant ensemble learning algorithm on a dataset.

Arrhythmia Dataset Analysis

According to the autorank report, the best algorithm for the Arrhythmia Dataset is the Bagging algorithm. A Bagging algorithm functions by splitting the data into multiple training sets upon which a class of learning or optimising methods such as decision trees and neural networks are applied. After training these multiple models on different samples of the same data, the prediction is averaged into a single summary with the reasoning that the averaging of misclassification errors on different data splits gives a better estimate of the predictive ability of a learning method.

The Bagging algorithm outperformed the other algorithms on the Arrhythmia dataset due to the fact that it is a multi-variate dataset with many attributes that would benefit from Bagging’s reduction of variance and yet not overfit.

*Mean Test Accuracy and Standard Deviation Output and Head of Autorank Results*

A picture containing text, receipt

Description automatically generatedText

Description automatically generated

Caesarian Dataset Analysis

The best algorithm for the Caesarian Dataset according to the Autorank report is the AdaBoost algorithm. Boosting trains each tree on a modified version of the original dataset. AdaBoost does this by assigning higher weights to wrongly classified observations as well as the trained classifiers according to accuracy, while it iteratively trains. This process iterates until complete training data fits without any error.

The characteristics of the Caesarian dataset that enabled the AdaBoost algorithm to outperform the others is due to the relatively small size and simplicity of the dataset since Boosting methods would falter for particularly noisy data and logically perform better on simpler data.

*Mean Test Accuracy and Standard Deviation Output and Head of Autorank Results*

Text

Description automatically generatedText

Description automatically generated

Website-Phishing Dataset Analysis

For the Website-Phishing Dataset the best method from the Autorank report is the Random Forest algorithm. Random Forest is a Bagging algorithm that combines various decision trees to produce a more generalised model. Individual decision trees are generated using a random selection of attributes at each node to determine the split and during classification, each tree votes with the most popular class returned.

Random forests are more robust to errors and outliers as the generalisation error for a forest converges as long as the number of trees in the forest is large and considers many fewer attributes for each split which makes it efficient for large databases. For this reason, the characteristics of the Website-Phishing dataset that enabled the Random Forest algorithm to outperform the others is due to the vast number of instances and number of attributes present in the multivariate dataset.

*Mean Test Accuracy and Standard Deviation Output and Head of Autorank Results*

Text

Description automatically generatedText

Description automatically generated

Task 2: Gradient Boosting

Text

Description automatically generated with medium confidenceGradient Boosting trains many models sequentially and each new model gradually minimises the loss function of the whole system using the Gradient Descent method. I applied the exact same experimental design as was done for the four initial algorithms in their exploration of the given three datasets as GradientBoosting is similarly an ensemble learner(one that is even derived from AdaBoost) that would have likely benefitted from the same experimental design.

Task 3: Iris Dataset: Head of Auto Rank Results

**SUMMARY**

The best performing algorithm was the Bagging method, ranking 1st for the Arrhythmia dataset and 2nd in both the Caesarian and Website Phishing datasets. In consideration of these successes across different types of datasets, Bagging would likely be the ensemble algorithm I would try first in the future.