Ensemble of SVM's

Ensemble learning combines multiple models to improve prediction. Ensembles are preferred over a single model because they can balance out overall scores, but they are more time-consuming since results must be combined.

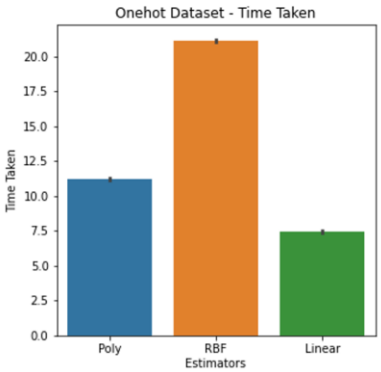
In the next tests, ensemble of SVMs with Polynomial, RBF, and Linear kernels will be used along with the bagging method. Bagging can help generalize the model by reducing the variance of individual models and improving prediction results.

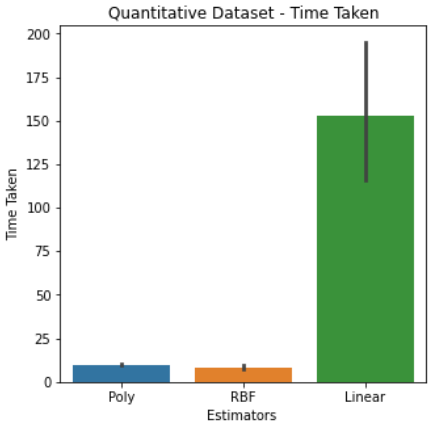
C Parameter Tuning

The focus of these tests is on the C parameter, which balances the trade-off between misclassification and margin width. A smaller C value leads to a soft margin which may result in underfitting, while a larger C value leads to a harder margin which may result in overfitting. Therefore, these tests are aimed at finding the optimal C value for the ensemble.

Although a range of C values from 1 to 8 were tested all models achieved a perfect F1 score and accuracy of 100% so the time taken was measured.

In the onehot dataset, the RBF kernel took the longest to complete, likely due to the increased number of features it had to process. In contrast, in the quantitative dataset, the linear kernel took substantially longer to complete than the polynomial or RBF kernel. The time range for the linear kernel was between 115 to 195 seconds, while the difference between RBF and polynomial kernel was marginal. Overall, the RBF and polynomial kernels performed better on the quantitative dataset.

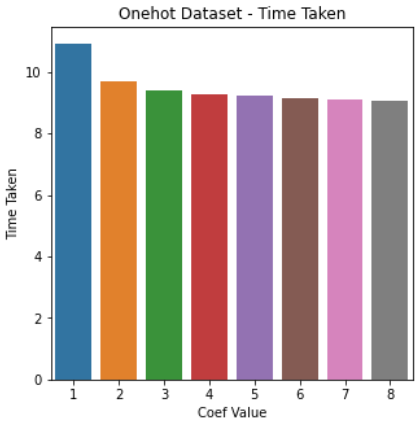


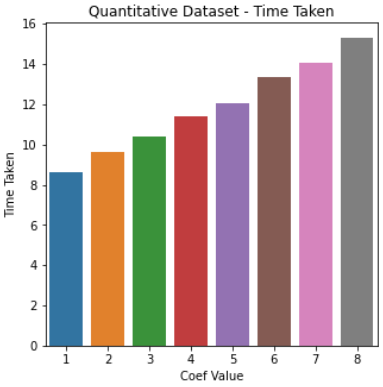


Polynomial Coefficient

The coef0 parameter adjusts the independent term of the kernel, but a high value can increase model complexity and reduce generalizability.

In the tests, a C value of 1 and coef0 values ranging from 1 to 8 are used. The onehot dataset achieved perfect accuracy and F1 scores of 100%. The coef0 value of 1 resulted in slightly longer completion times than other values, with a difference of 1.2 seconds. In contrast, the quantitative dataset did not initially achieve 100% accuracy or F1 scores until the coef0 value was set to 3, which was also the fastest among the tested values. Therefore, the optimal value for coef0 on the quantitative dataset is 3, as it achieved a perfect score and was the fastest value.

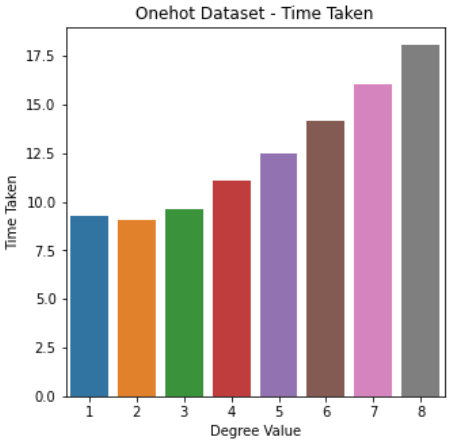


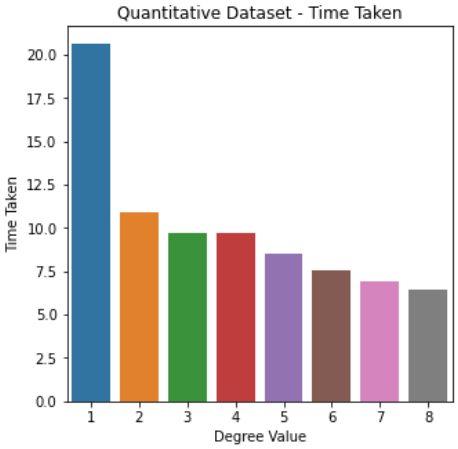


Polynomial Degree

The degree parameter is used to control the decision boundary. A high degree value creates a more flexible decision boundary, however too high can cause the model to overfit.

The results of onehot show that a degree value of 2 is optimal. The values after taking longer to compute and the degree value of 1 doesn't reach 100% accuracy or f1 score. However, in the quantitative dataset, an accuracy and f1 score of 100% is only reached by value 4.



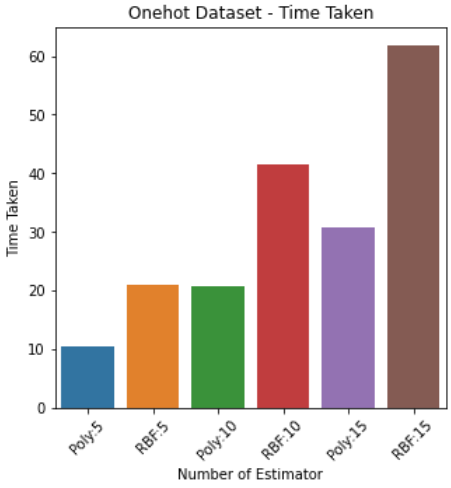


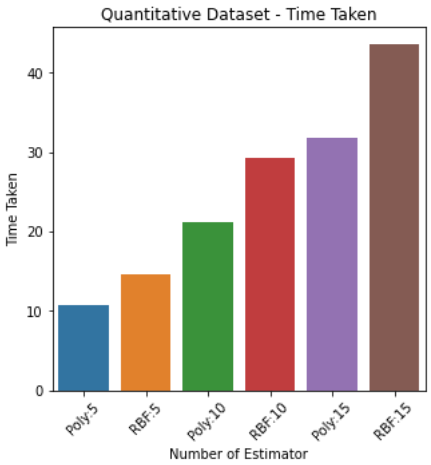
Number of Estimations

The number of estimators refers to the number of base estimators used to create the ensemble model. The outputs of these base estimators are then combined to create the final output.

After hyper-tuning the kernel hyperparameters, testing is carried out on the bagging n_estimators hyperparameter. The range of the number of estimators is 5, 10, and 15 as anything larger will be computationally expensive. The graph shows only the RBF and polynomial kernels since they have produced the best results so far. Both RBF and polynomial kernels achieved 100% accuracy and f1 scores on the test data, so the results are measured by time taken.

The polynomial kernel performed faster than the RBF kernel for both onehot and quantitative datasets, with an n_estimator value of 5 producing the best results. These findings show that using a smaller number of base estimators can lead to faster result times, while still achieving 100% accuracy and f2 scores.





Final Test

Finally, the remaining 20% of the quantitative and onehot dataset are used to compare the results of RBF and polynomial kernels.

These results show that overall, polynomial performs faster, with both datasets. However, the quantitative dataset results show a marginal difference of 0.45 seconds between the two kernels. But overall, the Polynomial quantitative dataset is the best performance.