**Parameter Optimization**

After completing the bagging and normalization stages of the preprocessing pipeline we selected the highest performing ensemble model for each of the three imputed data sets. We then set out to optimize the parameters for the algorithms contained within each of these highest-performing ensemble classifiers. In order to find the optimal parameters for each different ensemble and data pairing, we split the training data and performed cross validation while iterating through different values for adjustable parameters of each individual algorithm contained within the ensemble.

For example, after bagging and normalization, Ensemble B yielded the best performance for our Mode-imputed data. Since Ensemble B consisted of three algorithms — KNN, Random Forest and Naïve Bayes — we iterated through different values of number of neighbors for KNN and number of estimators for Random Forest (excluding Naïve Bayes from this process since it does not provide any adjustable parameters) using the normalized Mode-imputed training data. We then concluded that the optimal number of neighbor for KNN was 30 and that the optimal number of estimators for Random Forest was 50.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ensemble | Data | KNN | Random Forest | Logistic Regression | SVM | Naïve Bayes |
| B | Mode | n\_neighbors = 30 | n\_estimators = 50 | N/A | N/A | N/A |
| B | KNN | n\_neighbors = 50 | n\_estimators = 90 | N/A | N/A | N/A |
| C | RF | n\_neighbors = 15 | n\_estimators = 90 | C = 10 | C = 10,  gamma = 0.1,  kernel = 'rbf' | N/A |

**Feature Selection**

For our feature selection stage, we took the Filter Method approach, which ranks features by how closely correlated they are to the target variable (absolute value of Pearson Correlation Coefficient). We ran each of our optimized models with three different set amounts of features: the first 20 most correlated features, the first 40 most correlated features, and the first 60 correlated features. These results were then compared to the performance results generated by using all features.

Reducing the number of features only improved the performance for the KNN Data, Ensemble B pairing, with the optimal feature number being 60.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data | Method | Features | accuracy | precision | recall | f1-score |
| Mode | Filter | 20 | 81.59% | 58.09% | 61.71% | 59.84% |
| Mode | Filter | 40 | 82.48% | 63.13% | 62.87% | 63.00% |
| Mode | Filter | 60 | 82.38% | 63.21% | 62.59% | 62.90% |
| Mode | Regular | All | 82.51% | 62.38% | 63.15% | 62.76% |
| KNN | Filter | 20 | 85.19% | 66.54% | 69.48% | 67.98% |
| KNN | Filter | 40 | 84.98% | 67.86% | 68.34% | 68.10% |
| KNN | Filter | 60 | 84.91% | 68.54% | 67.89% | 68.21% |
| KNN | Regular | All | 85.98% | 58.16% | 76.85% | 66.21% |
| Random Forest | Filter | 20 | 85.34% | 61.78% | 72.17% | 66.57% |
| Random Forest | Filter | 40 | 83.42% | 14.54% | 20.92% | 17.16% |
| Random Forest | Filter | 60 | 85.41% | 63.83% | 71.39% | 67.40% |
| **Random Forest** | **Regular** | **All** | **85.73%** | **63.65%** | **72.55%** | **67.81%** |