**Supplementary material 4 – Iterations of ensemble models for machine learning**

Table S2: Performance metrics for ensemble machine learning models on test data (n = 936)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Voting Ensemble Model ^ | Period of Analysis:  October 2019 to December 2019 | | | | | | |
| Training data  (n = 3742) | Testing data  (n = 936) | | | | | |
| Accuracy | Accuracy | Label | No. of records | Precision | Recall | F1-score |
| Model 1 – RFC + SVC + GBC | 0.99 | 0.98 | Substandard | 203 | 0.97 | 0.95 | 0.96 |
| Non-substandard | 733 | 0.99 | 0.99 | 0.99 |
| Model 2 – RFC + SVC + LRC | 0.99 | 0.98 | Substandard | 203 | 0.97 | 0.95 | 0.96 |
| Non-substandard | 733 | 0.99 | 0.99 | 0.99 |
| Model 3 – SVC + LRC + GBC | 0.99 | 0.98 | Substandard | 203 | 0.97 | 0.95 | 0.96 |
| Non-substandard | 733 | 0.99 | 0.99 | 0.99 |
| Model 4 – RFC + LRC + GBC | 0.99 | 0.98 | Substandard | 203 | 0.97 | 0.95 | 0.96 |
| Non-substandard | 733 | 0.99 | 0.99 | 0.99 |

**\*** Values correct to 2 decimal places

^ GBC: Gradient Boosting Classifier; LRC: Logistic Regression Classifier; RFC: Random Forest Classifier; SVC: Support Vector Classifier

Table S3: Performance metrics for ensemble machine learning models on validation set 1 data (n = 4,920)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Voting Ensemble Model ^ | Period of Analysis:  January 2020 – March 2020 | | | | | |
| Validation Set 1  (n = 4920) | | | | | |
| Accuracy | Label | No. of records | Precision | Recall | F1-score |
| Model 1 – RFC + SVC + GBC | 0.97 | Substandard | 733 | 0.94 | 0.84 | 0.89 |
| Non-substandard | 4187 | 0.97 | 0.99 | 0.98 |
| Model 2 – RFC + SVC + LRC | 0.96 | Substandard | 733 | 0.95 | 0.80 | 0.87 |
| Non-substandard | 4187 | 0.97 | 0.99 | 0.98 |
| Model 3 – SVC + LRC + GBC | 0.97 | Substandard | 733 | 0.95 | 0.84 | 0.89 |
| Non-substandard | 4187 | 0.97 | 0.99 | 0.98 |
| Model 4 – RFC + LRC + GBC | 0.96 | Substandard | 733 | 0.94 | 0.80 | 0.86 |
| Non-substandard | 4187 | 0.97 | 0.99 | 0.98 |

**\*** Values correct to 2 decimal places

^ GBC: Gradient Boosting Classifier; LRC: Logistic Regression Classifier; RFC: Random Forest Classifier; SVC: Support Vector Classifier

Table S4: Performance metrics for ensemble machine learning models on validation set 2 data (n = 2,458)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Voting Ensemble Model ^ | Period of Analysis:  April 2020 – May 2020 | | | | | |
| Validation Set 2  (n = 2458) | | | | | |
| Accuracy | Label | No. of records | Precision | Recall | F1-score |
| Model 1 – RFC + SVC + GBC | 0.98 | Substandard | 427 | 0.96 | 0.93 | 0.94 |
| Non-substandard | 2031 | 0.98 | 0.99 | 0.99 |
| Model 2 – RFC + SVC + LRC | 0.97 | Substandard | 427 | 0.96 | 0.88 | 0.92 |
| Non-substandard | 2031 | 0.98 | 0.99 | 0.98 |
| Model 3 – SVC + LRC + GBC | 0.98 | Substandard | 427 | 0.97 | 0.91 | 0.94 |
| Non-substandard | 2031 | 0.98 | 0.99 | 0.99 |
| Model 4 – RFC + LRC + GBC | 0.98 | Substandard | 427 | 0.96 | 0.90 | 0.93 |
| Non-substandard | 2031 | 0.98 | 0.99 | 0.99 |

**\*** Values correct to 2 decimal places

^ GBC: Gradient Boosting Classifier; LRC: Logistic Regression Classifier; RFC: Random Forest Classifier; SVC: Support Vector Classifier