**On-line Appendix**

This appendix contained tables for the following cases:

1. There are total 4 tables provided in the appendix corresponding to the results of different prediction models.
2. First table is corresponding to the results of the GANSYN based resampled models. Second and third tables are corresponding to the results of the SMOTE and ADSYN approaches based prediction models, respectively. Last table is corresponding to the results of baseline prediction models on the original imbalanced datasets.
3. Each table shows results of all performance evaluation measures for all the used 7 datasets.
4. Fifth table provides the details of source code and complexity metrics available in the used 7 datasets.
5. The sixth and seventh tables show the parameter values of different used machine learning techniques and GAN model.

Table 1: Results of the presented GANSYN approach based prediction models for the all five datasets in terms of considered performance measures

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Techniques** | **Precision** | **Recall (PD)** | **F1\_score** | **AUC** | **PF** | **MCC** | **G-mean** | **G-measure** |
| **DATASET-1** | | | | | | | | |
| KNN | 0.9500 | 0.9500 | 0.95 | 0.9375 | 0.1250 | 0.8977 | 0.9117 | 0.9110 |
| Logistic Regression | 0.7600 | 0.7300 | 0.7300 | 0.7450 | 0.1667 | 0.4836 | 0.7800 | 0.7783 |
| Naive Bayes | 0.8600 | 0.8600 | 0.8600 | 0.8520 | 0.2083 | 0.7171 | 0.8251 | 0.8244 |
| Decision Tree | 0.9000 | 0.8800 | 0.8800 | 0.8540 | 0.2917 | 0.7683 | 0.7895 | 0.7849 |
| Random Forest | 0.9400 | 0.9300 | 0.9300 | 0.9160 | 0.1667 | 0.8648 | 0.8803 | 0.8790 |
| SVM | 0.7600 | 0.7600 | 0.7600 | 0.7410 | 0.3750 | 0.5003 | 0.6892 | 0.6859 |
| **DATASET-2** | | | | | | | | |
| KNN | 1.0000 | 1.0000 | 1.0000 | 0.9950 | 0.0087 | 0.9908 | 0.9956 | 0.9956 |
| Logistic Regression | 0.9000 | 0.8900 | 0.8900 | 0.8900 | 0.0522 | 0.7901 | 0.9185 | 0.9180 |
| Naive Bayes | 0.9300 | 0.9200 | 0.9200 | 0.9250 | 0.1391 | 0.8522 | 0.8899 | 0.8895 |
| Decision Tree | 0.9900 | 0.9900 | 0.9900 | 0.9860 | 0.0261 | 0.9727 | 0.9819 | 0.9819 |
| Random Forest | 1.0000 | 1.0000 | 1.0000 | 0.9951 | 0.0087 | 0.9908 | 0.9956 | 0.9956 |
| SVM | 0.9300 | 0.9100 | 0.9100 | 0.9170 | 0.1652 | 0.8389 | 0.8716 | 0.8708 |
| **DATASET-3** | | | | | | | | |
| KNN | 0.9900 | 0.9900 | 0.9900 | 0.9880 | 0.0108 | 0.9760 | 0.9896 | 0.9896 |
| Logistic Regression | 0.9800 | 0.9800 | 0.9800 | 0.9820 | 0.0054 | 0.9656 | 0.9873 | 0.9872 |
| Naive Bayes | 0.9200 | 0.9200 | 0.9200 | 0.9220 | 0.0838 | 0.8453 | 0.9181 | 0.9181 |
| Decision Tree | 0.9900 | 0.9900 | 0.9900 | 0.9930 | 0.0027 | 0.9867 | 0.9936 | 0.9936 |
| Random Forest | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.0001 | 0.999 | 0.9986 | 0.9986 |
| SVM | 0.8800 | 0.8600 | 0.8600 | 0.8580 | 0.0162 | 0.7395 | 0.9198 | 0.9177 |
| **DATASET-4** | | | | | | | | |
| KNN | 0.9900 | 0.9900 | 0.9900 | 0.9848 | 0.0303 | 0.9720 | 0.9798 | 0.9797 |
| Logistic Regression | 0.8800 | 0.8700 | 0.8700 | 0.8770 | 0.0606 | 0.7552 | 0.9040 | 0.9034 |
| Naive Bayes | 0.9500 | 0.9400 | 0.9400 | 0.9410 | 0.0909 | 0.8877 | 0.9244 | 0.9243 |
| Decision Tree | 0.9300 | 0.9300 | 0.9300 | 0.9300 | 0.0606 | 0.8591 | 0.9347 | 0.9347 |
| Random Forest | 0.9600 | 0.9600 | 0.9600 | 0.9540 | 0.0606 | 0.9153 | 0.9496 | 0.9496 |
| SVM | 0.9000 | 0.9000 | 0.9000 | 0.8990 | 0.1212 | 0.8018 | 0.8893 | 0.8893 |
| **DATASET-5** | | | | | | | | |
| KNN | 0.9600 | 0.9600 | 0.9600 | 0.9460 | 0.1071 | 0.9105 | 0.9258 | 0.9252 |
| Logistic Regression | 0.9000 | 0.9000 | 0.9000 | 0.8900 | 0.1071 | 0.7867 | 0.8964 | 0.8964 |
| Naive Bayes | 0.9600 | 0.9600 | 0.9500 | 0.9460 | 0.1071 | 0.9105 | 0.9258 | 0.9252 |
| Decision Tree | 0.9700 | 0.9700 | 0.9700 | 0.9640 | 0.0714 | 0.9398 | 0.9491 | 0.9488 |
| Random Forest | 0.9700 | 0.9700 | 0.9700 | 0.9640 | 0.0714 | 0.9398 | 0.9491 | 0.9488 |
| SVM | 0.9700 | 0.9600 | 0.9500 | 0.9464 | 0.1071 | 0.9105 | 0.9258 | 0.9252 |
| **DATASET-6** | | | | | | | | |
| KNN | 0.99 | 0.99 | 0.99 | 0.997 | 0.0053 | 0.9943 | 0.9924 | 0.9924 |
| Logistic Regression | 0.89 | 0.88 | 0.88 | 0.883 | 0.1980 | 0.7761 | 0.8401 | 0.8392 |
| Naive Bayes | 0.96 | 0.96 | 0.96 | 0.963 | 0.0683 | 0.9279 | 0.9457 | 0.9456 |
| Decision Tree | 0.99 | 0.99 | 0.99 | 0.993 | 0.0079 | 0.9878 | 0.9911 | 0.9911 |
| Random Forest | 0.99 | 0.99 | 0.99 | 0.989 | 0.0206 | 0.9792 | 0.9847 | 0.9847 |
| SVM | 0.95 | 0.95 | 0.95 | 0.949 | 0.0955 | 0.9026 | 0.9270 | 0.9267 |
| **DATASET-7** | | | | | | | | |
| KNN | 0.99 | 0.99 | 0.99 | 0.994 | 0.0104 | 0.9896 | 0.9898 | 0.9898 |
| Logistic Regression | 0.94 | 0.94 | 0.94 | 0.935 | 0.0303 | 0.8733 | 0.9547 | 0.9546 |
| Naive Bayes | 0.77 | 0.59 | 0.5 | 0.587 | 0.2386 | 0.3107 | 0.3226 | 0.2716 |
| Decision Tree | 0.99 | 0.99 | 0.99 | 0.996 | 0.0063 | 0.9937 | 0.9919 | 0.9919 |
| Random Forest | 0.99 | 0.99 | 0.99 | 0.995 | 0.0084 | 0.9917 | 0.9908 | 0.9908 |
| SVM | 0.73 | 0.73 | 0.73 | 0.731 | 0.2296 | 0.4636 | 0.7499 | 0.7496 |
| **AVERAGE** | | | | | | | | |
| KNN | 0.9775 | 0.9775 | 0.9867 | 0.9725 | 0.0528 | 0.9536 | 0.9621 | 0.9618 |
| Logistic Regression | 0.8763 | 0.8650 | 0.8650 | 0.8684 | 0.0984 | 0.7393 | 0.8826 | 0.8819 |
| Naive Bayes | 0.9013 | 0.8763 | 0.8638 | 0.8735 | 0.2162 | 0.7711 | 0.8221 | 0.8154 |
| Decision Tree | 0.9575 | 0.9525 | 0.9525 | 0.9463 | 0.0948 | 0.9095 | 0.9277 | 0.9265 |
| Random Forest | 0.9737 | 0.9712 | 0.9712 | 0.9661 | 0.0632 | 0.9397 | 0.9536 | 0.9533 |
| SVM | 0.8600 | 0.8538 | 0.8525 | 0.8478 | 0.1856 | 0.7072 | 0.8327 | 0.8314 |

Table 2: Results of the SMOTE oversampling approach based prediction models for the all five datasets in terms of considered performance measures

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Techniques** | **Precision** | **Recall (PD)** | **F1\_score** | **AUC** | **PF** | **MCC** | **G-mean** | **G-measure** |
| **DATASET-1** | | | | | | | | |
| KNN | 0.7700 | 0.7500 | 0.7500 | 0.7500 | 0.3750 | 0.5164 | 0.6847 | 0.6818 |
| Logistic Regression | 0.8100 | 0.8100 | 0.8100 | 0.8125 | 0.2083 | 0.6255 | 0.8008 | 0.8007 |
| Naive Bayes | 0.7000 | 0.6900 | 0.7300 | 0.6875 | 0.2083 | 0.3834 | 0.7391 | 0.7373 |
| Decision Tree | 0.8400 | 0.8300 | 0.8300 | 0.8330 | 0.2500 | 0.6761 | 0.7890 | 0.7880 |
| Random Forest | 0.8600 | 0.8300 | 0.8300 | 0.8330 | 0.2917 | 0.6885 | 0.7668 | 0.7644 |
| SVM | 0.7400 | 0.7300 | 0.7300 | 0.7200 | 0.3750 | 0.4686 | 0.6755 | 0.6734 |
| **DATASET-2** | | | | | | | | |
| KNN | 0.8500 | 0.8300 | 0.8300 | 0.8340 | 0.2522 | 0.6799 | 0.7878 | 0.7868 |
| Logistic Regression | 0.9400 | 0.9400 | 0.9400 | 0.9390 | 0.0756 | 0.8788 | 0.9322 | 0.9321 |
| Naive Bayes | 0.9000 | 0.9000 | 0.9000 | 0.9040 | 0.1130 | 0.8092 | 0.8935 | 0.8934 |
| Decision Tree | 0.9800 | 0.9800 | 0.9800 | 0.9780 | 0.0259 | 0.9654 | 0.9771 | 0.9771 |
| Random Forest | 0.9800 | 0.9800 | 0.9800 | 0.9800 | 0.0348 | 0.9568 | 0.9726 | 0.9726 |
| SVM | 0.8900 | 0.8500 | 0.8500 | 0.8300 | 0.2957 | 0.7373 | 0.7738 | 0.7703 |
| **DATASET-3** | | | | | | | | |
| KNN | 0.8600 | 0.8500 | 0.8500 | 0.8510 | 0.2135 | 0.7087 | 0.8176 | 0.8170 |
| Logistic Regression | 0.8000 | 0.8000 | 0.8000 | 0.8000 | 0.2189 | 0.6004 | 0.7905 | 0.7904 |
| Naive Bayes | 0.7600 | 0.7200 | 0.7100 | 0.7220 | 0.0973 | 0.4779 | 0.8062 | 0.8011 |
| Decision Tree | 0.9300 | 0.9300 | 0.9300 | 0.9300 | 0.1000 | 0.8667 | 0.9149 | 0.9148 |
| Random Forest | 0.9200 | 0.9100 | 0.9100 | 0.9100 | 0.1405 | 0.8260 | 0.8844 | 0.8840 |
| SVM | 0.7500 | 0.7500 | 0.7500 | 0.7470 | 0.2865 | 0.4957 | 0.7315 | 0.7313 |
| **DATASET-4** | | | | | | | | |
| KNN | 0.9100 | 0.9100 | 0.9100 | 0.9090 | 0.0909 | 0.8182 | 0.9095 | 0.9095 |
| Logistic Regression | 0.9200 | 0.9200 | 0.9200 | 0.9200 | 0.0909 | 0.8489 | 0.9145 | 0.9145 |
| Naive Bayes | 0.8900 | 0.8900 | 0.8900 | 0.8930 | 0.0909 | 0.7882 | 0.8995 | 0.8994 |
| Decision Tree | 0.9400 | 0.9400 | 0.9400 | 0.9390 | 0.0909 | 0.8804 | 0.9244 | 0.9243 |
| Random Forest | 0.9200 | 0.9200 | 0.9200 | 0.9240 | 0.0909 | 0.8489 | 0.9145 | 0.9145 |
| SVM | 0.9100 | 0.9100 | 0.9100 | 0.9090 | 0.1212 | 0.8197 | 0.8943 | 0.8941 |
| **DATASET-5** | | | | | | | | |
| KNN | 0.9500 | 0.9500 | 0.9500 | 0.9460 | 0.1071 | 0.8980 | 0.9210 | 0.9205 |
| Logistic Regression | 0.9100 | 0.8900 | 0.8900 | 0.8920 | 0.2143 | 0.8044 | 0.8362 | 0.8346 |
| Naive Bayes | 0.9500 | 0.9500 | 0.9500 | 0.9460 | 0.0714 | 0.8934 | 0.9392 | 0.9392 |
| Decision Tree | 0.9500 | 0.9500 | 0.9500 | 0.9460 | 0.0714 | 0.8934 | 0.9392 | 0.9392 |
| Random Forest | 0.9700 | 0.9600 | 0.9600 | 0.9490 | 0.0714 | 0.9309 | 0.9442 | 0.9440 |
| SVM | 0.9300 | 0.9300 | 0.9300 | 0.9280 | 0.1071 | 0.8593 | 0.9112 | 0.9111 |
| **DATASET-6** | | | | | | | | |
| KNN | 0.95 | 0.94 | 0.94 | 0.944 | 0.0889 | 0.8916 | 0.9254 | 0.9253 |
| Logistic Regression | 0.91 | 0.89 | 0.89 | 0.893 | 0.2124 | 0.8060 | 0.8372 | 0.8357 |
| Naive Bayes | 0.87 | 0.87 | 0.87 | 0.868 | 0.0749 | 0.7423 | 0.8971 | 0.8967 |
| Decision Tree | 0.99 | 0.99 | 0.99 | 0.994 | 0.0092 | 0.9886 | 0.9904 | 0.9904 |
| Random Forest | 0.99 | 0.99 | 0.99 | 0.989 | 0.0193 | 0.9791 | 0.9854 | 0.9853 |
| SVM | 0.91 | 0.91 | 0.91 | 0.909 | 0.1205 | 0.8206 | 0.8946 | 0.8945 |
| **DATASET-7** | | | | | | | | |
| KNN | 0.92 | 0.91 | 0.91 | 0.914 | 0.1451 | 0.8347 | 0.8820 | 0.8816 |
| Logistic Regression | 0.89 | 0.86 | 0.86 | 0.86 | 0.2787 | 0.7511 | 0.7876 | 0.7846 |
| Naive Bayes | 0.85 | 0.79 | 0.78 | 0.79 | 0.4186 | 0.6402 | 0.6777 | 0.6698 |
| Decision Tree | 0.99 | 0.99 | 0.99 | 0.995 | 0.0084 | 0.9906 | 0.9908 | 0.9908 |
| Random Forest | 0.99 | 0.99 | 0.99 | 0.991 | 0.0146 | 0.9834 | 0.9877 | 0.9877 |
| SVM | 0.85 | 0.84 | 0.84 | 0.843 | 0.1910 | 0.6885 | 0.8243 | 0.8242 |
| **Average** | | | | | | | | |
| KNN | 0.8871 | 0.8771 | 0.8771 | 0.8783 | 0.1818 | 0.7639 | 0.8469 | 0.8461 |
| Logistic Regression | 0.8829 | 0.8729 | 0.8729 | 0.8738 | 0.1856 | 0.7593 | 0.8427 | 0.8418 |
| Naive Bayes | 0.8457 | 0.8300 | 0.8329 | 0.8301 | 0.1535 | 0.6764 | 0.8360 | 0.8338 |
| Decision Tree | 0.9457 | 0.9443 | 0.9443 | 0.9450 | 0.0794 | 0.8945 | 0.9323 | 0.9321 |
| Random Forest | 0.9471 | 0.9400 | 0.9400 | 0.9378 | 0.0947 | 0.8877 | 0.9222 | 0.9218 |
| SVM | 0.8543 | 0.8457 | 0.8457 | 0.8409 | 0.2139 | 0.6985 | 0.8150 | 0.8141 |

Table 3: Results of the ADASYN oversampling approach based prediction models for the all five datasets in terms of considered performance measures

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Techniques** | **Precision** | **Recall (PD)** | **F1\_score** | **AUC** | **PF** | **MCC** | **G-mean** | **G-measure** |
| **DATASET-1** | | | | | | | | |
| KNN | 0.8600 | 0.8500 | 0.8500 | 0.8540 | 0.2083 | 0.7139 | 0.8203 | 0.8198 |
| Logistic Regression | 0.8300 | 0.8100 | 0.8100 | 0.8120 | 0.2917 | 0.6390 | 0.7575 | 0.7558 |
| Naive Bayes | 0.7600 | 0.7500 | 0.7500 | 0.7500 | 0.1667 | 0.5071 | 0.7906 | 0.7895 |
| Decision Tree | 0.8800 | 0.8800 | 0.8700 | 0.8740 | 0.1667 | 0.7526 | 0.8563 | 0.8560 |
| Random Forest | 0.8700 | 0.8500 | 0.8500 | 0.8540 | 0.2500 | 0.7242 | 0.7984 | 0.7969 |
| SVM | 0.8700 | 0.8500 | 0.8500 | 0.8540 | 0.2500 | 0.7242 | 0.7984 | 0.7969 |
| **DATASET-2** | | | | | | | | |
| KNN | 0.8700 | 0.8500 | 0.8500 | 0.8510 | 0.2435 | 0.7166 | 0.8019 | 0.8005 |
| Logistic Regression | 0.9700 | 0.9700 | 0.9700 | 0.9730 | 0.0348 | 0.9477 | 0.9676 | 0.9676 |
| Naive Bayes | 0.9200 | 0.9200 | 0.9200 | 0.9210 | 0.1043 | 0.8440 | 0.9077 | 0.9077 |
| Decision Tree | 0.9700 | 0.9700 | 0.9700 | 0.9730 | 0.0348 | 0.9477 | 0.9676 | 0.9676 |
| Random Forest | 0.9800 | 0.9800 | 0.9800 | 0.9820 | 0.0174 | 0.9651 | 0.9813 | 0.9813 |
| SVM | 0.8800 | 0.8500 | 0.8500 | 0.8520 | 0.2783 | 0.7288 | 0.7832 | 0.7806 |
| **DATASET-3** | | | | | | | | |
| KNN | 0.8600 | 0.8500 | 0.8500 | 0.8490 | 0.2216 | 0.7056 | 0.8134 | 0.8126 |
| Logistic Regression | 0.8000 | 0.7900 | 0.7900 | 0.7920 | 0.2568 | 0.5886 | 0.7663 | 0.7659 |
| Naive Bayes | 0.7300 | 0.6900 | 0.6800 | 0.6940 | 0.1000 | 0.4265 | 0.7880 | 0.7811 |
| Decision Tree | 0.9200 | 0.9200 | 0.9200 | 0.9160 | 0.1162 | 0.8350 | 0.9017 | 0.9015 |
| Random Forest | 0.9100 | 0.9000 | 0.9000 | 0.9040 | 0.1649 | 0.8167 | 0.8670 | 0.8664 |
| SVM | 0.7400 | 0.7400 | 0.7400 | 0.7410 | 0.2811 | 0.4842 | 0.7294 | 0.7293 |
| **DATASET-4** | | | | | | | | |
| KNN | 0.8900 | 0.8900 | 0.8900 | 0.8930 | 0.1212 | 0.7882 | 0.8844 | 0.8844 |
| Logistic Regression | 0.9600 | 0.9500 | 0.9500 | 0.9540 | 0.0909 | 0.9129 | 0.9293 | 0.9291 |
| Naive Bayes | 0.8300 | 0.8300 | 0.8300 | 0.8330 | 0.1515 | 0.6670 | 0.8392 | 0.8391 |
| Decision Tree | 0.9200 | 0.9200 | 0.9200 | 0.9240 | 0.0909 | 0.8489 | 0.9145 | 0.9145 |
| Random Forest | 0.9200 | 0.9100 | 0.9100 | 0.9090 | 0.1515 | 0.8243 | 0.8787 | 0.8782 |
| SVM | 0.9600 | 0.9500 | 0.9500 | 0.9540 | 0.0909 | 0.9129 | 0.9293 | 0.9291 |
| **DATASET-5** | | | | | | | | |
| KNN | 0.9100 | 0.9100 | 0.9100 | 0.9100 | 0.0714 | 0.8220 | 0.9192 | 0.9192 |
| Logistic Regression | 0.8800 | 0.8400 | 0.8400 | 0.8390 | 0.3214 | 0.7166 | 0.7550 | 0.7507 |
| Naive Bayes | 0.8900 | 0.8900 | 0.8900 | 0.8920 | 0.1071 | 0.7857 | 0.8914 | 0.8914 |
| Decision Tree | 0.9500 | 0.9500 | 0.9500 | 0.9460 | 0.0357 | 0.8934 | 0.9571 | 0.9571 |
| Random Forest | 0.9700 | 0.9600 | 0.9600 | 0.9640 | 0.0714 | 0.9309 | 0.9442 | 0.9440 |
| SVM | 0.9700 | 0.9600 | 0.9600 | 0.9640 | 0.0714 | 0.9309 | 0.9442 | 0.9440 |
| **DATASET-6** | | | | | | | | |
| KNN | 0.92 | 0.91 | 0.91 | 0.91 | 0.1375 | 0.8284 | 0.8859 | 0.8856 |
| Logistic Regression | 0.91 | 0.9 | 0.9 | 0.896 | 0.2076 | 0.8100 | 0.8445 | 0.8428 |
| Naive Bayes | 0.86 | 0.86 | 0.86 | 0.858 | 0.0806 | 0.7223 | 0.8892 | 0.8887 |
| Decision Tree | 0.99 | 0.99 | 0.99 | 0.996 | 0.0074 | 0.9921 | 0.9913 | 0.9913 |
| Random Forest | 0.99 | 0.99 | 0.99 | 0.985 | 0.0267 | 0.9723 | 0.9816 | 0.9816 |
| SVM | 0.9 | 0.9 | 0.9 | 0.89 | 0.1713 | 0.7869 | 0.8636 | 0.8629 |
| **DATASET-7** | | | | | | | | |
| KNN | 0.92 | 0.91 | 0.91 | 0.9 | 0.1534 | 0.8247 | 0.8777 | 0.8771 |
| Logistic Regression | 0.89 | 0.85 | 0.85 | 0.852 | 0.2944 | 0.7382 | 0.7745 | 0.7711 |
| Naive Bayes | 0.86 | 0.8 | 0.79 | 0.79 | 0.4050 | 0.6506 | 0.6899 | 0.6824 |
| Decision Tree | 0.99 | 0.99 | 0.99 | 0.994 | 0.0094 | 0.9885 | 0.9903 | 0.9903 |
| Random Forest | 0.99 | 0.99 | 0.99 | 0.99 | 0.0188 | 0.9803 | 0.9856 | 0.9856 |
| SVM | 0.84 | 0.84 | 0.84 | 0.836 | 0.1879 | 0.6748 | 0.8259 | 0.8258 |
| **Average** | | | | | | | | |
| KNN | 0.8900 | 0.8814 | 0.8814 | 0.8810 | 0.1653 | 0.7713 | 0.8575 | 0.8570 |
| Logistic Regression | 0.8914 | 0.8729 | 0.8729 | 0.8740 | 0.2139 | 0.7647 | 0.8278 | 0.8261 |
| Naive Bayes | 0.8357 | 0.8200 | 0.8171 | 0.8197 | 0.1593 | 0.6576 | 0.8280 | 0.8257 |
| Decision Tree | 0.9457 | 0.9457 | 0.9443 | 0.9461 | 0.0659 | 0.8940 | 0.9398 | 0.9398 |
| Random Forest | 0.9471 | 0.9400 | 0.9400 | 0.9411 | 0.1001 | 0.8877 | 0.9195 | 0.9191 |
| SVM | 0.8800 | 0.8700 | 0.8700 | 0.8701 | 0.1901 | 0.7490 | 0.8391 | 0.8384 |

Table 4: Results of the baseline prediction models on the original imbalanced datasets for the all five datasets in terms of considered performance measures

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Techniques** | **Precision** | **Recall (PD)** | **F1\_score** | **AUC** | **PF** | **MCC** | **G-mean** | **G-measure** |
| **DATASET-1** | | | | | | | | |
| KNN | 0.73 | 0.75 | 0.74 | 0.488 | 0.1287 | -0.0259 | 0.8084 | 0.8061 |
| Logistic Regression | 0.77 | 0.71 | 0.73 | 0.57 | 0.2277 | 0.1186 | 0.7405 | 0.7398 |
| Naive Bayes | 0.74 | 0.71 | 0.72 | 0.506 | 0.1980 | 0.0114 | 0.7546 | 0.7532 |
| Decision Tree | 0.75 | 0.7 | 0.72 | 0.54 | 0.2277 | 0.0751 | 0.7353 | 0.7344 |
| Random Forest | 0.42 | 0.5 | 0.46 | 0.5 | 0.0098 | -0.0394 | 0.7036 | 0.6645 |
| SVM | 0.7 | 0.8 | 0.75 | 0.47 | 0.0495 | -0.0904 | 0.8720 | 0.8688 |
| **DATASET-2** | | | | | | | | |
| KNN | 0.49 | 0.5 | 0.5 | 0.497 | 0.0043 | -0.0068 | 0.7056 | 0.6657 |
| Logistic Regression | 0.49 | 0.48 | 0.49 | 0.49 | 0.0452 | -0.0224 | 0.6770 | 0.6388 |
| Naive Bayes | 0.49 | 0.47 | 0.48 | 0.466 | 0.0667 | -0.0276 | 0.6623 | 0.6252 |
| Decision Tree | 0.49 | 0.5 | 0.5 | 0.497 | 0.0043 | -0.0068 | 0.7056 | 0.6657 |
| Random Forest | 0.49 | 0.5 | 0.5 | 0.5 | 0.0021 | -0.0048 | 0.7063 | 0.6662 |
| SVM | 0.49 | 0.5 | 0.5 | 0.497 | 0.0043 | -0.0068 | 0.7056 | 0.6657 |
| **DATASET-3** | | | | | | | | |
| KNN | 0.55 | 0.52 | 0.52 | 0.515 | 0.0195 | 0.0556 | 0.7141 | 0.6796 |
| Logistic Regression | 0.62 | 0.62 | 0.62 | 0.619 | 0.0550 | 0.2410 | 0.7654 | 0.7488 |
| Naive Bayes | 0.61 | 0.7 | 0.64 | 0.698 | 0.1080 | 0.3012 | 0.7902 | 0.7844 |
| Decision Tree | 0.55 | 0.51 | 0.51 | 0.512 | 0.0168 | 0.0489 | 0.7081 | 0.6716 |
| Random Forest | 0.55 | 0.51 | 0.49 | 0.5 | 0.0060 | 0.0342 | 0.7120 | 0.6741 |
| SVM | 0.56 | 0.51 | 0.51 | 0.514 | 0.0134 | 0.0605 | 0.7093 | 0.6724 |
| **DATASET-4** | | | | | | | | |
| KNN | 0.46 | 0.5 | 0.48 | 0.5 | 0.0071 | -0.0229 | 0.7046 | 0.6651 |
| Logistic Regression | 0.6 | 0.65 | 0.62 | 0.652 | 0.0709 | 0.2523 | 0.7771 | 0.7649 |
| Naive Bayes | 0.56 | 0.59 | 0.57 | 0.589 | 0.0935 | 0.1513 | 0.7313 | 0.7148 |
| Decision Tree | 0.55 | 0.53 | 0.53 | 0.527 | 0.0360 | 0.0731 | 0.7148 | 0.6840 |
| Random Forest | 0.46 | 0.5 | 0.48 | 0.5 | 0.0071 | -0.0229 | 0.7046 | 0.6651 |
| SVM | 0.97 | 0.56 | 0.6 | 0.56 | 0.0000 | 0.3442 | 0.7483 | 0.7179 |
| **DATASET-5** | | | | | | | | |
| KNN | 0.66 | 0.62 | 0.63 | 0.616 | 0.0619 | 0.2704 | 0.7626 | 0.7466 |
| Logistic Regression | 0.57 | 0.58 | 0.57 | 0.58 | 0.1327 | 0.1508 | 0.7092 | 0.6951 |
| Naive Bayes | 0.76 | 0.76 | 0.76 | 0.763 | 0.0619 | 0.5263 | 0.8443 | 0.8397 |
| Decision Tree | 0.71 | 0.72 | 0.72 | 0.724 | 0.0796 | 0.4390 | 0.8140 | 0.8079 |
| Random Forest | 0.43 | 0.5 | 0.47 | 0.5 | 0.0088 | -0.0339 | 0.7040 | 0.6647 |
| SVM | 0.66 | 0.62 | 0.63 | 0.616 | 0.0619 | 0.2704 | 0.7626 | 0.7466 |
| **DATASET-6** | | | | | | | | |
| KNN | 0.5 | 0.5 | 0.5 | 0.499 | 0.0005 | -0.0016 | 0.7069 | 0.6666 |
| Logistic Regression | 0.5 | 0.5 | 0.5 | 0.499 | 0.0016 | -0.0027 | 0.7066 | 0.6663 |
| Naive Bayes | 0.5 | 0.43 | 0.39 | 0.431 | 0.3596 | -0.0195 | 0.5248 | 0.5145 |
| Decision Tree | 0.5 | 0.5 | 0.5 | 0.5 | 0.0003 | 0.1611 | 0.7070 | 0.6666 |
| Random Forest | 0.5 | 0.5 | 0.5 | 0.499 | 0.0005 | 0.1309 | 0.7069 | 0.6666 |
| SVM | 0.5 | 0.5 | 0.5 | 0.5 | 0.0003 | 0.1611 | 0.7070 | 0.6666 |
| **DATASET-7** | | | | | | | | |
| KNN | 0.5 | 0.5 | 0.5 | 0.5 | 0.0001 | 0.1172 | 0.7071 | 0.6666 |
| Logistic Regression | 0.53 | 0.55 | 0.54 | 0.55 | 0.0060 | 0.0835 | 0.7394 | 0.7082 |
| Naive Bayes | 0.51 | 0.72 | 0.5 | 0.71 | 0.0785 | 0.0990 | 0.8145 | 0.8084 |
| Decision Tree | 0.5 | 0.5 | 0.5 | 0.5 | 0.0038 | 0.1663 | 0.7058 | 0.6658 |
| Random Forest | 0.5 | 0.5 | 0.5 | 0.5 | 0.0001 | 0.1172 | 0.7071 | 0.6666 |
| SVM | 0.5 | 0.5 | 0.5 | 0.499 | 0.0002 | -0.0009 | 0.7070 | 0.6666 |
| **Average** | | | | | | | | |
| KNN | 0.56 | 0.553 | 0.552 | 0.5245 | 0.0304 | 0.07104 | 0.72929 | 0.69891 |
| Logistic Regression | 0.577 | 0.579 | 0.576 | 0.5658 | 0.07284 | 0.12102 | 0.72964 | 0.70721 |
| Naive Bayes | 0.604 | 0.627 | 0.585 | 0.6055 | 0.14957 | 0.18501 | 0.72813 | 0.71788 |
| Decision Tree | 0.581 | 0.569 | 0.571 | 0.5536 | 0.04652 | 0.16057 | 0.73197 | 0.70421 |
| Random Forest | 0.483 | 0.502 | 0.486 | 0.4998 | 0.00497 | 0.03125 | 0.70674 | 0.66732 |
| SVM | 0.61 | 0.562 | 0.563 | 0.5286 | 0.02052 | 0.12301 | 0.73907 | 0.70902 |

Table 5: Description of the software complexity, source code, and aging-related metrics

|  |  |  |
| --- | --- | --- |
| Type | Metrics | Description |
| Program size | AltAvgLineBlank, AltAvgLineCode, AltAvgLineComment, AltCountLineBlank, AltCountLineCode, AltCountLineComment, AvgLine, AvgLineBlank, AvgLineCode, AvgLineComment, CountDeclClass, CountDeclFunction, CountLine, CountLineBlank, CountLineCode, CountLineCodeDecl, CountLineCodeExe, CountLineComment, CountLineInactive, CountLinePreprocessor, CountSemicolon, CountStmt, CountStmtDecl, CountStmtEmpty, CountStmtExe, RatioCommentToCode | Metrics related to the amount of lines of code, declarations, statements, and files |
| McCabe’s cyclomatic complexity | AvgCyclomatic, AvgCyclomaticModified, AvgCyclomaticStrict, AvgEssential, MaxCyclomatic, MaxCyclomaticModified, MaxCyclomaticStrict, SumCyclomatic, SumCyclomaticModified, SumCyclomaticStrict, SumEssential | Metrics related to the control flow graph of functions and methods |
| Halstead metrics | Program Volume, Program Length, Program Vocabulary, Program Difficulty, Effort, N1, N2, n1, n2 | Metrics based on operands and operators in the program |
| Aging-Related Metrics (ARMs) | AllocOps, DeallocOps, DerefSet, DerefUse, UniqueDerefSet, UniqueDerefUse | Metrics related to memory usage |

Table 6: Parameter values set for different machine learning techniques

|  |  |
| --- | --- |
| Technique | Parameter values |
| K-Nearest Neighbour (KNN) | n\_neighbors = range(1, 21, 2), weights = ['uniform', 'distance'], metric = ['euclidean', 'manhattan', 'minkowski'] |
| Naïve Bayes (NB) | With default param eter in scikitlearn |
| Logistic Regression (LR) | solvers = ['liblinear'],  penalty = ['l2'], c\_values = [100, 10, 1.0, 0.1, 0.01] |
| Decision Tree (DT) | mdepth=np.arange(1,40), grid={"criterion":["gini","entropy"],"max\_depth":mdepth} |
| Random Forest (RF) | 'bootstrap': [True],  'max\_depth': [2,4,5],  'max\_features': ["log2","sqrt"],  'min\_samples\_leaf': [10,30], 'min\_samples\_split': [10,12], 'n\_estimators': [100, 350, 500] |
| Support Vector Machine (SVM) | kernel = ['poly', 'rbf', 'sigmoid'], C parameters= [50, 10, 1.0, 0.1, 0.01] |

Table7: Parameter values set for Generative Adversarial Network (GAN)

|  |  |
| --- | --- |
| Parameter | Values |
| max\_epoch | 10 |
| steps\_per\_epoch | 10 |
| save\_checkpoints | True |
| restore\_session | True |
| batch\_size | 1 |
| z\_dim | 50 |
| noise= | 0.2 |
| l2norm | 0.00001 |
| learning\_rate | 0.001 |
| num\_gen\_rnn | 100 |
| num\_gen\_features | 100 |
| num\_dis\_layers | 2 |
| num\_dis\_hidde | 100 |
| optimizer | 'AdamOptimizer' |