|  |  |
| --- | --- |
| **Dataset1:** |  |
| Algorithm1: | Random Forest |
| Algorithm2: | Logistic Regression |
| Algorithm3: | Support Vector Machine (Linear) |
| Algorithm4: | Support Vector Machine (RBF) |
|  | |
| **Datase2:** |  |
| Algorithm1: |  |
| Algorithm2: |  |
| Algorithm3: |  |
| Algorithm4: |  |

**Dataset and Algorithm description:**

**Some classification metrics: (take for each class and average them to get single metric):**

1. PPV = (TP)/(TP+FP) # Precision or Positive\_Precdictive\_Value (PPV)
2. Recall = TP/(TP+FN) # Recall or Sensitivity or True\_Positive\_Rate (TPR) or Hit\_Rate
3. F1\_S = (2\*PPV\*Recall)/(PPV+Recall) # F1 Score or Harmonic Mean
4. F1\_M = (PPV+Recall)/2 # F1 Measure
5. Specificity = TN/(TN+FP) # Specificity or True\_Negative\_Rate(TNR) or Selectivity
6. NPV = TN/(TN+FN) # Negative\_Predictive\_Value
7. FPR = FP/(FP+TN) # False\_Positive\_Rate
8. FNR = FN/(TP+FN) # False\_Negative\_Rate or Miss\_Rate
9. FDR = FP/(TP+FP) # False\_Discovery\_Rate
10. CSI = TP/(TP+FN+FP) # Critical\_Success\_Index or Threat\_Score(TS)
11. FM = sqrt(PPV\*Recall) # Fowlkes\_Mallows\_Index
12. BA = (Recall+Specificity)/2 # Balanced\_Accuracy
13. MCC = (TP\*TN-FP\*FN)/(sqrt((TP+FP)\

\*(TP+FN)\*(TN+FP)\*(TN+FN))) # Mathews\_Correlation\_Coefficient

1. BI = Recall+Specificity-1 or TPR-FPR # Bookmaker\_Informedness or Informedness
2. MK = PPV+NPV-1 # Markedness or delta

Optional

1. FOR = FN/(FN+TN) # False\_Omission\_Rate
2. PLR = Recall/FPR # Positive\_Likelihood\_Ratio
3. NLR = FNR/Specificity # Negative\_Likelihood\_Ratio
4. PT = sqrt(FPR)/(sqrt(Recall)\

+sqrt(FPR)) # Prevalence\_Threshold

1. DOR = PLR/NLR # Diagnostic\_Odds\_Ratio

* 21. Accuracy # Overall accuracy, not for each class
* 22. Cohen Kappa score # Overall Kappa score, not for each class

**Result analysis procedure:**

1. **First try all algorithm for each dataset.**
2. **Now for each dataset choose best 3 or 4 algorithms**

* One best from Normal ML Classification methods,
* One best from Deep learning Classification methods,
* One best from Ensemble learning Classification methods,
* And one best from unsupervised or semi supervised methods (optional)

1. **Now Do the following tables for each dataset:**

**For Dataset1:**

**Table 1 ( For Normal Split [ 80:20, 90:10, 85:15 ] )**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Split Ratio** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Random Forest | Train 90%, Test 10% | 70.13 | 57.14 | 46.15 | 51.06 |
| Random Forest | Train 80%, Test 20% | 77.92 | 72.09 | 58.49 | 64.58 |
| Random Forest | Train 75%, Test 25% | 73.44 | 59.70 | 62.5 | 61.07 |
| Random Forest | Train 85%, Test 15% | 75.00 | 58.82 | 57.14 | 57.97 |
| Logistic Regression | Train 90%, Test 10% | 76.62 | 70.00 | 53.84 | 60.89 |
| Logistic Regression | Train 80%, Test 20% | 79.22 | 75.61 | 58.49 | 65.96 |
| Logistic Regression | Train 75%, Test 25% | 74.48 | 62.71 | 57.81 | 60.16 |
| Logistic Regression | Train 85%, Test 15% | 78.44 | 69.23 | 51.42 | 59.02 |
| SVM(Linear) | Train 90%, Test 10% | 76.62 | 70.00 | 53.85 | 60.89 |
| SVM(Linear) | Train 80%, Test 20% | 78.57 | 73.81 | 58.49 | 65.26 |
| SVM(Linear) | Train 75%, Test 25% | 75.52 | 64.91 | 57.81 | 61.16 |
| SVM(Linear) | Train 85%, Test 15% | 76.72 | 64.28 | 51.43 | 57.14 |
| SVM(Poly) | Train 90%, Test 10% | 76.62 | 72.22 | 50.00 | 59.09 |
| SVM(Poly) | Train 80%, Test 20% | 74.67 | 70.59 | 45.28 | 55.17 |
| SVM(Poly) | Train 75%, Test 25% | 73.96 | 61.67 | 57.81 | 59.68 |
| SVM(Poly) | Train 85%, Test 15% | 73.27 | 60.00 | 34.28 | 43.64 |

**Observation**: Take the best split ratio for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1: Random Forest

Best Split=80:20

Reason: Random forest with 80:20 split works better than other split set.

Updated Algorithm2: Logistic Regression

Best Split=80:20

Reason: Logistic Regression with 80:20 works far better than other split set.

Updated Algorithm3: SVM(Linear)

Best Split= 80:20

Reason: SVM(Linear ) with 80:20 works better than other split set

Updated Algorithm4: SVM(ploy)

Best Split=90:10

Reason: Here we can see SVM(ploy) works better with 90:10 splits other than other other split set.

**# Result needed without any tuning operation perform # TABLE 2**

**Table 3 (For** **Model Optimization using Hyperparameter Tuning) (With out Feature sealection)**

**[CV = best CV techniques for each algo)]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Model Optimization** | **Accuracy** | **Precision** | **F1-Score** | **Recall** |
| Random Forest | GridSearchCV | 75.97 | 77.5 | 62.63 | 52.54 |
| Random Forest | RandomizedSearchCV | 74.02 | 74.36 | 59.18 | 49.15 |
| Random Forest | Hyperopt | 74.67 | 77.78 | 58.95 | 47.46 |
| Algorithm1 | Nature-Inspired |  |  |  |  |
| Logistic Regression | GridSearchCV | 74.67 | 75.00 | 60.61 | 50.85 |
| Logistic Regression | RandomizedSearchCV | 76.62 | 78.05 | 64.00 | 54.23 |
| Logistic Regression | Hyperopt | 77.27 | 80.00 | 64.65 | 54.24 |
| Logistic Regression | Nature-Inspired |  |  |  |  |
| SVM (Linear) | GridSearchCV | 76.62 | 78.05 | 64.00 | 54.24 |
| SVM (Linear) | RandomizedSearchCV | 76.62 | 79.49 | 63.26 | 52.54 |
| SVM (Linear) | Hyperopt | 77.92 | 79.07 | 66.67 | 57.63 |
| SVM (Linear) | Nature-Inspired |  |  |  |  |
| SVM (RBF) | GridSearchCV | 61.69 | - | 0 | - |
| SVM(RBF) | RandomizedSearchCV | 74.67 | 73.81 | 61.38 | 52.54 |
| SVM(RBF) | Hyperopt | 65.58 | 65.00 | 32.91 | 22.03 |

**Observation**: Take the best Model Optimization using Hyperparameter tuningtechnique for each algorithm basis on the classification metrics and do the following table with that.

**Updated Algorithm1: Random Forest**

Best Split= 80:20

Best CV= GridSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm2: Logistic Regression**

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm3: SVM (Linear)**

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm4: SVM(RBF)**

Best Split= 80 : 20

Best CV= RandomizedSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Table 4 (Apply feature selection Method)**

**[After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms]**

**(Split – 80:20)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Random Forest | 74.67 | 61.36 | 55.10 | 58.06 |
| Logistic Regression | 75.97 | 63.04 | 59.18 | 61.05 |
| SVM (Linear) | 76.62 | 64.44 | 59.18 | 61.70 |
| SVM(RBF) | 75.97 | 65.79 | 51.02 | 57.47 |

**Observation**: Take the best feature selection technique for each algorithm basis on the classification metrics and do the following tables with that.

**Updated Algorithm1: Random Forest**

Best Split= 80:20

Best CV= GridSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm2: Logistic Regression**

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm3: SVM (Linear)**

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm4: SVM(RBF)**

Best Split= 80 : 20

Best CV= RandomizedSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Table 5 ( Perform Model Optimization using Hyperparameter After Feature Selection)**

**[CV = best CV techniques for each algo)]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Model Optimization** | **Accuracy** | **Precision** | **F1-Score** | **Recall** |
| Random Forest | GridSearchCV | 75.97 | 77.5 | 62.63 | 52.54 |
| Random Forest | RandomizedSearchCV | 74.02 | 74.36 | 59.18 | 49.15 |
| Random Forest | Hyperopt | 74.67 | 77.78 | 58.95 | 47.46 |
| Algorithm1 | Nature-Inspired |  |  |  |  |
| Logistic Regression | GridSearchCV | 74.67 | 75.00 | 60.61 | 50.85 |
| Logistic Regression | RandomizedSearchCV | 76.62 | 78.05 | 64.00 | 54.23 |
| Logistic Regression | Hyperopt | 77.27 | 80.00 | 64.65 | 54.24 |
| Logistic Regression | Nature-Inspired |  |  |  |  |
| SVM (Linear) | GridSearchCV | 76.62 | 78.05 | 64.00 | 54.24 |
| SVM (Linear) | RandomizedSearchCV | 76.62 | 79.49 | 63.26 | 52.54 |
| SVM (Linear) | Hyperopt | 77.92 | 79.07 | 66.67 | 57.63 |
| SVM (Linear) | Nature-Inspired |  |  |  |  |
| SVM (RBF) | GridSearchCV | 61.69 | - | 0 | - |
| SVM(RBF) | RandomizedSearchCV | 74.67 | 73.81 | 61.38 | 52.54 |
| SVM(RBF) | Hyperopt | 65.58 | 65.00 | 32.91 | 22.03 |

**Observation**: Take the best Model Optimization using Hyperparameter tuningtechnique for each algorithm basis on the classification metrics and do the following table with that.

**Updated Algorithm1: Random Forest**

Best Split= 80:20

Best CV= GridSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm2: Logistic Regression**

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm3: SVM (Linear)**

Best Split= 80:20

Best CV= HyperOpt

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**Updated Algorithm4: SVM(RBF)**

Best Split= 80 : 20

Best CV= RandomizedSearchCV

Best Feature selection= After running 5 Feature Selection Method i.e., K-Best, Mutual Info Classif, Chi-square, Correlation matrix we have inferred by Selecting 5 Best Features from the Dataset and then run the previous algorithms

**# Conclusion Model with feature selection VS Model without Feature Selection**

**# model performance with hyperparameter tuning VS model performance without hyperparameter tuing**

**Table 6 (For Choosing best model)**

**Reason:**

|  |  |
| --- | --- |
| **Best algorithm Name** |  |
| **Model description** |  |
|  | |
| **Classification metric 1** |  |
| **Classification metric 2** |  |
| **Classification metric 3** |  |
| **Classification metric 4** |  |
| **Classification metric 5** |  |
| **Classification metric 6** |  |
| **Classification metric 7** |  |
| **Classification metric 8** |  |
| **Classification metric 9** |  |
| **Classification metric 10** |  |
| **Classification metric 11** |  |
| **Classification metric 12** |  |
| **Classification metric 13** |  |
| **Classification metric 14** |  |
| **Classification metric 15** |  |
| **Classification metric 16** |  |
| **Classification metric 17** |  |
| **Classification metric 18** |  |
| **Classification metric 19** |  |
| **Classification metric 20** |  |
| **Classification metric 21** |  |
| **Classification metric 22** |  |

**For Dataset2:**

**Table 1 (For Normal Split)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Split Ratio** | **Classification metric 1** | **Classification metric 2** | **Classification metric 3** |
| Algorithm1 | Train 90%, Test 10% |  |  |  |
| Algorithm1 | Train 80%, Test 20% |  |  |  |
| Algorithm1 | Train75%, Test 25% |  |  |  |
| Algorithm2 | Train 90%, Test 10% |  |  |  |
| Algorithm2 | Train 80%, Test 20% |  |  |  |
| Algorithm2 | Train75%, Test 25% |  |  |  |
| Algorithm3 | Train 90%, Test 10% |  |  |  |
| Algorithm3 | Train 80%, Test 20% |  |  |  |
| Algorithm3 | Train75%, Test 25% |  |  |  |
| Algorithm4 | Train 90%, Test 10% |  |  |  |
| Algorithm4 | Train 80%, Test 20% |  |  |  |
| Algorithm4 | Train75%, Test 25% |  |  |  |

**Observation**: Take the best split ratio for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1:

Best Split=

Reason:

Updated Algorithm2:

Best Split=

Reason:

Updated Algorithm3:

Best Split=

Reason:

Updated Algorithm4:

Best Split=

Reason:

**Table 2 (For Features Encoding) (optional)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Encoding** | **Classification metric 4** | **Classification metric 5** | **Classification metric 6** |
| Algorithm1 | Label |  |  |  |
| Algorithm1 | One-hot |  |  |  |
| Algorithm1 | Response |  |  |  |
| Algorithm2 | Label |  |  |  |
| Algorithm2 | One-hot |  |  |  |
| Algorithm2 | Response |  |  |  |
| Algorithm3 | Label |  |  |  |
| Algorithm3 | One-hot |  |  |  |
| Algorithm3 | Response |  |  |  |
| Algorithm4 | Label |  |  |  |
| Algorithm4 | One-hot |  |  |  |
| Algorithm4 | Response |  |  |  |

**Observation**: Take the best encoding technique for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1:

Best Split=

Best Encoding=

Reason:

Updated Algorithm2:

Best Split=

Best Encoding=

Reason:

Updated Algorithm3:

Best Split=

Best Encoding=

Reason:

Updated Algorithm4:

Best Split=

Best Encoding=

Reason:

**Table 3 (For Cross Validation)**

**[Kfold or stratified Kfold (K=10 or 5 or 4 based on best split 90-10 or 80-20 or 75-25 respectively for each algo.)]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Cross Validation** | **Classification metric 7** | **Classification metric 8** | **Classification metric 9** |
| Algorithm1 | Kfold |  |  |  |
| Algorithm1 | Stratified Kfold |  |  |  |
| Algorithm1 | Holdout |  |  |  |
| Algorithm2 | Kfold |  |  |  |
| Algorithm2 | Stratified Kfold |  |  |  |
| Algorithm2 | Holdout |  |  |  |
| Algorithm3 | Kfold |  |  |  |
| Algorithm3 | Stratified Kfold |  |  |  |
| Algorithm3 | Holdout |  |  |  |
| Algorithm4 | Kfold |  |  |  |
| Algorithm4 | Stratified Kfold |  |  |  |
| Algorithm4 | Holdout |  |  |  |

**Observation**: Take the best Cross Validation technique for each algorithm basis on the classification metrics and do the following tables with that.

Updated Algorithm1:

Best Split=

Best Encoding=

Best CV=

Reason:

Updated Algorithm2:

Best Split=

Best Encoding=

Best CV=

Reason:

Updated Algorithm3:

Best Split=

Best Encoding=

Best CV=

Reason:

Updated Algorithm4:

Best Split=

Best Encoding=

Best CV=

Reason:

**Table 4 (For feature selection) (optional)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Feature selection** | **Classification metric 10** | **Classification metric 11** | **Classification metric 12** |
| Algorithm1 | Without |  |  |  |
| Algorithm1 | Chi-Square |  |  |  |
| Algorithm1 | Mutual Info Classif |  |  |  |
| Algorithm1 | Correlation Coefficient |  |  |  |
| Algorithm2 | Without |  |  |  |
| Algorithm2 | Chi-Square |  |  |  |
| Algorithm2 | Mutual Info Classif |  |  |  |
| Algorithm2 | Correlation Coefficient |  |  |  |
| Algorithm3 | Without |  |  |  |
| Algorithm3 | Chi-Square |  |  |  |
| Algorithm3 | Mutual Info Classif |  |  |  |
| Algorithm3 | Correlation Coefficient |  |  |  |
| Algorithm4 | Without |  |  |  |
| Algorithm4 | Chi-Square |  |  |  |
| Algorithm4 | Mutual Info Classif |  |  |  |
| Algorithm4 | Correlation Coefficient |  |  |  |

**Observation**: Take the best feature selection technique for each algorithm basis on the classification metrics and do the following tables with that.

**Updated Algorithm1:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Reason:

**Updated Algorithm2:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Reason:

**Updated Algorithm3:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Reason:

**Updated Algorithm4:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Reason:

**Table 5 (For Model Optimization using Hyperparameter Tuning) (optional)**

**[CV = best CV techniques for each algo. And Nature-Inspired means any one recent NIOA Published in between 2021 to 23 like MGO, NOA, MFO\_SFR)]**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Model Optimization** | **Classification metric 13** | **Classification metric 14** | **Classification metric 15** |
| Algorithm1 | GridSearchCV |  |  |  |
| Algorithm1 | RandomizedSearchCV |  |  |  |
| Algorithm1 | Hyperopt |  |  |  |
| Algorithm1 | Nature-Inspired |  |  |  |
| Algorithm2 | GridSearchCV |  |  |  |
| Algorithm2 | RandomizedSearchCV |  |  |  |
| Algorithm2 | Hyperopt |  |  |  |
| Algorithm2 | Nature-Inspired |  |  |  |
| Algorithm3 | GridSearchCV |  |  |  |
| Algorithm3 | RandomizedSearchCV |  |  |  |
| Algorithm3 | Hyperopt |  |  |  |
| Algorithm3 | Nature-Inspired |  |  |  |
| Algorithm4 | GridSearchCV |  |  |  |
| Algorithm4 | RandomizedSearchCV |  |  |  |
| Algorithm4 | Hyperopt |  |  |  |
| Algorithm4 | Nature-Inspired |  |  |  |

**Observation**: Take the best Model Optimization using Hyperparameter tuningtechnique for each algorithm basis on the classification metrics and do the following table with that.

**Updated Algorithm1:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Best Model optimization=

Reason:

**Updated Algorithm2:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Best Model optimization=

Reason:

**Updated Algorithm3:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Best Model optimization=

Reason:

**Updated Algorithm4:**

Best Split=

Best Encoding=

Best CV=

Best Feature selection=

Best Model optimization=

Reason:

**Table 6 (For Choosing best model)**

**Reason:**

|  |  |
| --- | --- |
| **Best algorithm Name** |  |
| **Model description** |  |
|  | |
| **Classification metric 1** |  |
| **Classification metric 2** |  |
| **Classification metric 3** |  |
| **Classification metric 4** |  |
| **Classification metric 5** |  |
| **Classification metric 6** |  |
| **Classification metric 7** |  |
| **Classification metric 8** |  |
| **Classification metric 9** |  |
| **Classification metric 10** |  |
| **Classification metric 11** |  |
| **Classification metric 12** |  |
| **Classification metric 13** |  |
| **Classification metric 14** |  |
| **Classification metric 15** |  |
| **Classification metric 16** |  |
| **Classification metric 17** |  |
| **Classification metric 18** |  |
| **Classification metric 19** |  |
| **Classification metric 20** |  |
| **Classification metric 21** |  |
| **Classification metric 22** |  |