1. **Dataset**

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

1. **Preprocessing**

* Remove rows with missing value(0)
* Scale every features to have mean 0 and standard variation 1.
* Transform target label “B” and “M” to numeric value.

1. **Parameter tuning log**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment #** | **Classifier** | **Cross-Validation fold** | **Parameters** | **Average Accuracy of 10 folds** |
| 1 | Decision Tree | 10 | Max\_depath = 4 | 0.928 |
| **2** | **Decision Tree** | **10** | **Max\_depath = 8** | **0.930** |
| 3 | Decision Tree | 10 | Max\_depath = 12 | 0.923 |
| 4 | Decision Tree | 10 | Max\_depath = 16 | 0.924 |
| 5 | Decision Tree | 10 | All leaf nodes are pure | 0.924 |
| 6 | Neural Network | 10 | N\_hiddden\_layers = 3 | 0.968 |
| 7 | Neural Network | 10 | N\_hiddden\_layers = 6 | 0.973 |
| 8 | Neural Network | 10 | N\_hiddden\_layers = 10 | 0.971 |
| 9 | Neural Network | 10 | N\_hiddden\_layers = 15 | 0.973 |
| **10** | **Neural Network** | **10** | **N\_hiddden\_layers = 20** | **0.977** |
| 11 | Neural Network | 10 | N\_hiddden\_layers = 50 | 0.977 |
| 12 | Neural Network | 10 | N\_hiddden\_layers = 100 | 0.977 |
| 13 | SVM | 10 | Penalty term C = 0.05 | 0.932 |
| 14 | SVM | 10 | Penalty term C = 0.2 | 0.952 |
| 15 | SVM | 10 | Penalty term C = 1 | 0.977 |
| **16** | **SVM** | **10** | **Penalty term C = 5** | **0.979** |
| 17 | SVM | 10 | Penalty term C = 20 | 0.971 |
| 18 | SVM | 10 | Penalty term C = 200 | 0.957 |
| 19 | Logstic Reg | 10 | Penalty term C = 0.001 | 0.942 |
| 20 | Logstic Reg | 10 | Penalty term C = 0.01 | 0.968 |
| 21 | Logstic Reg | 10 | Penalty term C = 0.05 | 0.977 |
| **22** | **Logstic Reg** | **10** | **Penalty term C = 0.2** | **0.980** |
| 23 | Logstic Reg | 10 | Penalty term C = 1 | 0.977 |
| 24 | Logstic Reg | 10 | Penalty term C = 5 | 0.971 |
| 25 | Logstic Reg | 10 | Penalty term C = 20 | 0.962 |
| 26 | Logstic Reg | 10 | Penalty term C = 200 | 0.960 |
| 27 | KNN | 10 | Neighbors = 2 | 0.953 |
| 28 | KNN | 10 | Neighbors = 3 | 0.966 |
| **29** | **KNN** | **10** | **Neighbors = 4** | **0.966** |
| 30 | KNN | 10 | Neighbors = 5 | 0.966 |
| 31 | KNN | 10 | Neighbors = 6 | 0.962 |
| 32 | Bagging | 10 | N\_estimators = 4 | 0.935 |
| 33 | Bagging | 10 | N\_estimators = 6 | 0.953 |
| 34 | Bagging | 10 | N\_estimators = 8 | 0.941 |
| **35** | **Bagging** | **10** | **N\_estimators = 10** | **0.962** |
| 36 | Bagging | 10 | N\_estimators = 15 | 0.946 |
| 37 | Random Forest | 10 | N\_estimators = 4 | 0.941 |
| 38 | Random Forest | 10 | N\_estimators = 6 | 0.952 |
| 39 | Random Forest | 10 | N\_estimators = 8 | 0.946 |
| **40** | **Random Forest** | **10** | **N\_estimators = 10** | **0.961** |
| 41 | Random Forest | 10 | N\_estimators = 15 | 0.950 |
| 42 | Adaboost | 10 | N\_estimators = 10 | 0.941 |
| 43 | Adaboost | 10 | N\_estimators = 20 | 0.955 |
| 44 | Adaboost | 10 | N\_estimators = 30 | 0.968 |
| 45 | Adaboost | 10 | N\_estimators = 40 | 0.961 |
| **46** | **Adaboost** | **10** | **N\_estimators = 50** | **0.973** |
| 47 | Adaboost | 10 | N\_estimators = 60 | 0.969 |

1. **Result**

Total 556 instance and 30 features in the dataset after preprocessing, use 10-fold cross validation to fit and evaluate model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | Parameter | Accuracy | Precision | Recall | F1 |
| Decision Tree | Max\_depth = 8 | 0.916 | 0.877 | 0.903 | 0.886 |
| Neural Network | Hidden\_layers = 20 | 0.975 | 0.960 | 0.970 | 0.963 |
| SVM | Penalty term C = 5 | 0.979 | 0.984 | 0.959 | 0.971 |
| Logistic Reg | Penalty term C = 0.2 | 0.980 | 0.986 | 0.961 | 0.973 |
| KNN | Neighbors = 4 | 0.966 | 0.995 | 0.907 | 0.949 |
| Bagging | N\_estimators = 10 | 0.953 | 0.952 | 0.920 | 0.934 |
| Random Forest | N\_estimators = 10 | 0.944 | 0.944 | 0.914 | 0.927 |
| AdaBoost | N\_estimators = 50 | 0.973 | 0.971 | 0.948 | 0.959 |