1.

The dataset is 1984 United States Congressional Voting Records Database.

2.

The dataset has 17 attributes, 435 instances.

3.

3.1

Accuracy: Democrat: 0.891

Accuracy: Republication: 0.917

F-Measure: Democrat: 0.917

F-Measure: Republication:0.877

AUC: Democrat: 0.973

AUC: Republication: 0.973

3.2

The overall accuracy is 90.1149 %

The overall F-Measure is 0.902

The overall AUC is 0.973

3.3

Confusion matrix

|  |  |  |
| --- | --- | --- |
| a | b |  |
| 238 | 29 | democrat |
| 14 | 154 | republican |

3.4

Accuracy: Democrat: 0.891(in 3.1) to 0.827

Accuracy: Republication: 0.917(in 3.1) to 0.943

The performance has become lower in Naïve bayes. The accuracy of Democrat has decreased; The accuracy of Republication has increased.

3.5

10-fold cross validation:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Train | Train | Train | Train | Train | Train | Train | Train | Train | Test |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Train | Train | Train | Train | Train | Train | Train | Train | Test | Train |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Train | Train | Train | Train | Train | Train | Train | Test | Train | Train |

…

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Train | Train | Train | Train | Train | Train | Train | Train | Test |

The data set will be divided into 10 equal parts. The 10-fold cross validation would take 9 parts of data as train data and the rest one part as test data. Each part would take turn to become test data set. Finally, the calculate the average of the 10 times of performance results.

Split data:

|  |  |
| --- | --- |
| Train data | Test data |

Split-data would only be divided into 2 parts including train data and test data. There are only one time calculation.

4.

4.1

The overall accuracy is 98.8506 %

The overall F-Measure is 0.988

The overall AUC is 0.992

4.2

Compared the three metrics: accuracy, F-measure, AUC showed in 3.2 and 4.1.

We can find the Logistic Regression has better performance.

5.

5.1

Size of the tree : 11

5.2

A close up of a map

Description automatically generated

5.3

The accuracy is 94.2529 %

The F-Measure is 0.943

The overall AUC is 0.979

6.

A screenshot of a cell phone

Description automatically generated

1.After using the t-test, we can find the Logistic Regress and J 48 have better performance than Naïve bayes and the the risk is 5%

2. Naïve bayes is lowest and Logistic Regress is similar to J48