CIS 6930- INTRODUCTION TO DATA MINING

INDIVIDUAL PROJECT-I REPORT

In this project, we are asked to apply different classification methods on the given datasets:

* **Iris Flower Dataset:** This dataset contains 50 samples from each of three species of Iris flower and each sample has four features: length and width of the sepals and petals.
* **Life Expectancy Dataset:** This dataset contains 194 sample containing the name of the countries, continents and the life expectancy in each of them.

The following are the classification methods used in this project:

* Decision Tree:
  + RIPPER
  + C4.5
  + Oblique tree
* Naïve Bayes
* k-Nearest Neighbor(kNN)

The datasets are always split into training set (80%) and test set (20%).

**Decision Tree RIPPER method for IRIS and Life Expectancy dataset:**

Here we use the JRIP function to implement the RIPPER algorithm with the IRIS and Life Expectancy dataset.

The JRIP function gives the various rules for the training and test data. The following is the confusion matrix for the RIPPER method in IRIS and Life Expectancy:

prediction setosa versicolor virginica

setosa 10 0 0

versicolor 0 7 0

virginica 0 3 10

prediction Africa Asia Europe North America South America

Africa 0 0 0 0 0

Asia 40 16 3 6 2

Europe 0 2 5 0 0

North America 0 0 0 0 0

South America 0 0 0 0 0

**Decision Tree C4.5 method for IRIS and Life Expectancy dataset:**

Here we use the J48 function to implement the C4.5 algorithm with the IRIS and Life Expectancy dataset.

The J48 function gives the various size and leaves of the tree for the training and test data. The following is the confusion matrix for the C4.5 method in IRIS and Life Expectancy:

testPred setosa versicolor virginica

setosa 10 0 0

versicolor 0 7 1

virginica 0 3 9

testPred1 Africa Asia Europe North America South America

Africa 5 0 0 1 0

Asia 3 9 2 1 0

Europe 0 2 5 1 1

North America 2 1 1 1 1

South America 0 0 0 0 0

**Decision Tree Oblique method for IRIS and Life Expectancy dataset:**

Here we use the oblique.tree function to implement the oblique tree algorithm with the IRIS and Life Expectancy dataset.

The oblique.tree function gives the arranges the objects of the tree for the training and test data.

**Naïve Bayes for IRIS and Life Expectancy dataset:**

Here we use the naiveBayes function to implement the Naïve Bayes algorithm with the IRIS and Life Expectancy dataset.

The naiveBayes function gives the A-Priori and the conditional probabilities of the tree for the training and test data. The following is the confusion matrix for the Naïve Bayes method in IRIS and Life Expectancy:

testPred2 setosa versicolor virginica

setosa 10 0 0

versicolor 0 10 0

virginica 0 0 10

testPred3 Africa Asia Europe North America South America

Africa 5 0 0 1 0

Asia 2 7 1 2 0

Europe 0 2 5 0 0

North America 3 3 2 1 2

South America 0 0 0 0 0

**k-Nearest Neighbor for IRIS and Life Expectancy dataset:**

Here we use the IBk function to implement the kNN algorithm with the IRIS and Life Expectancy dataset.

The IBk function gives the nearest neighbors of the tree for the training and test data used in the classification. The following is the confusion matrix for the kNN method in IRIS and Life Expectancy:

prediction setosa versicolor virginica

setosa 10 0 0

versicolor 0 8 0

virginica 0 2 10

prediction Africa Asia Europe North America South America

Africa 5 0 1 1 0

Asia 4 8 2 2 0

Europe 1 2 5 0 2

North America 0 1 0 1 0

South America 0 1 0 0 0

**IRIS DATASET FINAL TABLE**

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| RIPPER | 0.9 |
| C4.5 | 0.87 |
| Oblique | 0.72 |
| Naïve Bayes | 1 |
| kNN | 0.93 |

**LIFE EXPECTANCY FINAL TABLE**

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| RIPPER | 0.21 |
| C4.5 | 0.42 |
| Oblique | 0.39 |
| Naïve Bayes | 0.5 |
| kNN | 0.53 |

**CONCLUSION:**

In conclusion we can see that the Naïve Bayes and kNN algorithms have the most accuracy and are the most suitable for the classification of both the datasets. The C4.5 algorithm give consistent output for all types of datasets but is not up to par with Naïve Bayes and kNN.