**CIS 6930: Introduction to Data Mining**

**Individual Project 1 – Report**

The project implements the following classification algorithms for the Iris Flower dataset and the Life Expectancy dataset.

1. C4.5 (Decision Tree)
2. RIPPER (Decision Tree)
3. Oblique (Decision Tree)
4. Naive Bayes
5. k-Nearest Neighbor (kNN)

Following are the confusion matrices of each classification method for the Iris Flower dataset. The target class here is the species class.

(Horizontal Set=Actual Class, Vertical Set=Predicted Class)

The plots and other details can be obtained by running the R scripts.

**C4.5 (Decision Tree)**

**Setosa versicolor virginica**

**Predictiontest**

**setosa** 10 0 0

**versicolor**  0 10 0

**virginica**  0 0 10

**RIPPER (Decision Tree)**

**setosa versicolor virginica**

**predictiontestrip**

**setosa** 10 0 0

**versicolor** 0 9 0

**virginica** 0 1 10

**Oblique (Decision Tree)**

**setosa versicolor virginica**

**prediction**

**setosa**  10 0 0

**versicolor** 0 8 0

**virginica**  0 2 10

**Naive Bayes**

**setosa versicolor virginica**

**prediction**

**setosa**  10 0 0

**versicolor** 0 9 0

**virginica**  0 1 10

**k-Nearest Neighbor (kNN)**

**setosa versicolor virginica**

**prediction**

**setosa**  10 0 0

**versicolor**  0 10 0

**virginica** 0 0 10

Following are the details of each classification method for the Life Expectancy dataset. The target class here is the continent class.

(Horizontal Set=Actual Class, Vertical Set=Predicted Class)

The plots and other details can be obtained by running the R scripts.

**C4.5 (Decision Tree)**

**Africa Asia Europe North America South America**

**predictiontestle**

**Africa**  7 0 0 0 0

**Asia**  2 8 1 1 1

**Europe**  1 3 7 1 1

**North America** 0 0 0 2 0

**South America** 0 1 0 0 0

**RIPPER (Decision Tree)**

**Africa Asia Europe North America South America**

**predictiontestlerip**

**Africa** 7 0 0 0 0

**Asia** 3 8 2 3 1

**Europe**  0 4 6 1 1

**North America** 0 0 0 0 0

**South America** 0 0 0 0 0

**Oblique (Decision Tree)**

**Africa Asia Europe North America South America**

**Prediction**

**Africa** 6 2 0 0 0

**Asia**  2 4 1 0 1

**Europe** 0 6 6 2 1

**North America** 2 0 1 2 0

**South America** 0 0 0 0 0

**Naive Bayes**

**Africa Asia Europe North America South America**

**Prediction**

**Africa** 7 1 0 0 0

**Asia**  2 5 1 0 1

**Europe** 0 6 6 2 1

**North America** 1 0 1 2 0

**South America** 0 0 0 0 0

**k-Nearest Neighbor (kNN)**

**Africa Asia Europe North America South America**

**prediction**

**Africa**  10 0 0 0 0

**Asia**  0 8 3 1 1

**Europe**  0 4 5 2 1

**North America** 0 0 0 1 0

**South America** 0 0 0 0 0

The following table indicates the accuracy of the above classifiers on the Iris Flower dataset

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| C4.5 (Decision Tree) | 1 |
| RIPPER (Decision Tree) | 0.966 |
| Oblique (Decision Tree) | 0.933 |
| Naïve Bayes | 0.966 |
| kNN | 1 |

The following table indicates the accuracy of the above classifiers on the LifeExpectancy dataset

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| C4.5 (Decision Tree) | 0.666 |
| RIPPER (Decision Tree) | 0.583 |
| Oblique (Decision Tree) | 0.5 |
| Naïve Bayes | 0.555 |
| kNN | 0.666 |

**Conclusion**

Thus from the accuracy measures for the classification algorithms for the Iris Flower dataset and Life Expectancy dataset, it can be observed that C4.5 and kNN classifiers are more accurate than the other classifiers. Also, the oblique decision tree classifier is the least accurate for both the datasets.