Academic Year: 2023-24
Semester: VI
Class / Branch: TE-IT
Subject: BI Lab
Name of Instructor: Prof. Apeksha Mohite

Name of Student: Manjiri Gole
Student ID: 21104006
Date of Performance: 30/03/2024
Date of Submission: 30/03/2024

Experiment No. 13

Aim: Business Intelligence Mini Project.

1. **Problem Definition:**

   The problem addressed by this dataset is the classification of congressmen into two categories based on their voting records: democrat or republican. This classification task aims to predict the party affiliation of a congressman given their votes on various legislative issues.

2. **Data mining task to be performed:**
   a. **ZeroR Classification:** ZeroR is a baseline classification algorithm used in data mining, serving as a simple reference point for comparison with other more complex models. It makes predictions solely based on the majority class in the training dataset, ignoring all input features. This means that regardless of the input data, ZeroR will always predict the most frequent class label in the training set. While ZeroR is extremely simplistic and lacks predictive power compared to more sophisticated algorithms, it can provide a benchmark for evaluating the performance of other classification models.
   b. **Random Tree Classification:** Random Tree classification is a decision tree-based algorithm that constructs a collection of decision trees during the training phase. Each tree is grown randomly, selecting a subset of features at each node split to reduce correlation between trees and improve generalization. During classification, predictions are made by aggregating the votes or predictions from all the individual trees. Random Tree classification offers simplicity and efficiency while reducing the risk of overfitting compared to traditional decision trees.

c. **Random Forest Classification:** Random Forest is an ensemble learning method based on the Random Tree algorithm. It builds multiple decision trees during training and combines their predictions through averaging or voting to improve accuracy and robustness. Random Forest introduces additional randomness by bootstrapping the training data and randomly selecting subsets of features for each tree. This helps in reducing overfitting and makes Random Forest one of the most popular and powerful classification algorithms, capable of handling large datasets with high dimensionality.

d. **Naive Bayes:** Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem and the assumption of feature independence. Despite its simplicity, Naive Bayes often performs well in practice, especially for text classification and spam filtering tasks. It calculates the probability of each class given a set of input features and selects the class with the highest probability as the prediction. Although Naive Bayes assumes independence between features, it can still be effective even when this assumption is violated, making it a versatile and efficient choice for many classification problems.

3. **Dataset identified: Votes Dataset**

4. **Source of dataset: WEKA**

5. **Details of the dataset:**

   a. **Source Information:** The data was sourced from the Congressional Quarterly Almanac for the specified session and congress. The donor of the dataset is Jeff Schlimmer, and the dataset was provided on April 27, 1987.

   b. **Past Usage:** The dataset has been used in academic research, particularly in Jeffrey Schlimmer's doctoral dissertation at the University of California, Irvine, focusing on concept acquisition through representational adjustment.

   c. **Relevant Information:** The dataset includes votes on 16 issues, with each issue being represented as a binary attribute (voted for or against). Additionally,

there's an attribute indicating party affiliation. The voting options are simplified to 'y' (voted for), 'n' (voted against), and '?' (unknown or abstained).

d. **Number of Instances:** There are a total of 435 instances (representing individual Congressmen), with 267 being Democrats and 168 being Republicans.
e. **Missing Attribute Values:** Missing attribute values are denoted by "?".
f. **Class Distribution:** Approximately 45.2% of the instances are Democrats, while 54.8% are Republicans.
g. **Attribute Information:** Each attribute represents a vote on a specific issue, with the last attribute indicating the party affiliation.
h. **Class Predictiveness and Predictability:** The dataset provides conditional probabilities for each attribute given a particular class, as well as conditional probabilities for each class given a particular attribute.

6. Algorithms to accomplish the task:

This section should include the brief description of the algorithm being implemented, the results of the applied algorithm and the data visualization (Screenshots of both to be included).

a. Zero Classification

- Description: ZeroR is a simple baseline classifier that predicts the majority class for all instances. In this case, it predicted that all instances belong to the "democrat" class.
- Performance: Achieved an accuracy of around 61.38%, which is not very high.
- Insights: Since it always predicts the majority class, it doesn't provide meaningful insights into the data or help in accurately predicting the class labels.

```
=== Summary ===

Correctly Classified Instances         267               61.3793 %
Incorrectly Classified Instances       168               38.6207 %
Kappa statistic                          0
Mean absolute error                      0.4742
Root mean squared error                  0.4869
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.614      1.000    0.761      ?        0.491     0.609     democrat
                 0.000    0.000    ?          0.000    ?          ?        0.491     0.382     republican
Weighted Avg.    0.614    0.614    ?          0.614    ?          ?        0.491     0.521

=== Confusion Matrix ===

   a   b   <-- classified as
 267   0 |   a = democrat
 168   0 |   b = republican
```

b. RandomForest Classifier:

- Description: RandomForest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.
- Performance: Achieved an accuracy of around 93.56%, which is significantly higher than the ZeroR classifier.
- Insights: RandomForest likely captures more complex patterns in the data compared to ZeroR, leading to better classification performance. It's generally a robust classifier suitable for various types of datasets.

```
=== Summary ===

Correctly Classified Instances         418               96.092  %
Incorrectly Classified Instances        17                3.908  %
Kappa statistic                          0.9175
Mean absolute error                      0.0714
Root mean squared error                  0.1742
Relative absolute error                 15.0587 %
Root relative squared error             35.7776 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.970    0.054    0.966      0.970    0.968      0.917    0.993     0.996     democrat
                 0.946    0.030    0.952      0.946    0.949      0.917    0.993     0.988     republican
Weighted Avg.    0.961    0.044    0.961      0.961    0.961      0.917    0.993     0.993

=== Confusion Matrix ===

   a   b   <-- classified as
 259   8 |   a = democrat
   9 159 |   b = republican
```

c. Naive Bayes Classifier:

- Description: Naive Bayes is a probabilistic classifier based on Bayes' theorem with the "naive" assumption of independence between features.
- Performance: Achieved an accuracy of around 90.11%, which is also better than ZeroR but slightly lower than RandomForest.
- Insights: Naive Bayes is efficient and often performs well on text classification tasks. It assumes that features are conditionally independent given the class, which might not always hold true in real-world datasets.

```
=== Summary ===

Correctly Classified Instances         392               90.1149 %
Incorrectly Classified Instances        43                9.8851 %
Kappa statistic                          0.7949
Mean absolute error                      0.0995
Root mean squared error                  0.2977
Relative absolute error                 20.9815 %
Root relative squared error             61.1406 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.891    0.083    0.944      0.891    0.917      0.797    0.973     0.984     democrat
                 0.917    0.109    0.842      0.917    0.877      0.797    0.973     0.957     republican
Weighted Avg.    0.901    0.093    0.905      0.901    0.902      0.797    0.973     0.973

=== Confusion Matrix ===

   a   b   <-- classified as
 238  29 |   a = democrat
  14 154 |   b = republican
```

d. RandomTree Classifier:

- Description: RandomTree classifier constructs decision trees using random attribute selection.
- Insight: It provides insights into attribute importance and decision-making processes.
- Performance: Achieves 93.56% accuracy with substantial agreement (kappa = 0.8636).

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         407               93.5632 %
Incorrectly Classified Instances        28                6.4368 %
Kappa statistic                          0.8636
Mean absolute error                      0.0699
Root mean squared error                  0.2379
Relative absolute error                 14.7341 %
Root relative squared error             48.8605 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.955    0.095    0.941      0.955    0.948      0.864    0.966     0.971     democrat
                 0.905    0.045    0.927      0.905    0.916      0.864    0.967     0.937     republican
Weighted Avg.    0.936    0.076    0.936      0.936    0.935      0.864    0.966     0.958

=== Confusion Matrix ===

   a   b   <-- classified as
 255  12 |   a = democrat
  16 152 |   b = republican
```

7. **Conclusion:** In summary, RandomForest is the best classifier among the three for this dataset, considering its high accuracy as 96.092% and ability to capture complex patterns. Naive Bayes also performed well and is simpler and more interpretable, but RandomForest achieved better accuracy in this case. ZeroR, being a baseline classifier, provided the lowest accuracy and minimal insights into the dataset.