Comparison of Decision Tree and Random Forest Classifications

Raghda Al taei

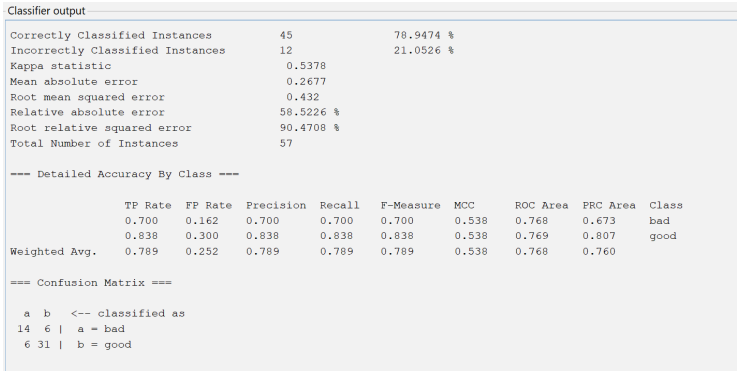
2024

1 Introduction

This report presents the results of classification using Decision Trees and Random Forest methods. We will compare the performance of a Decision Tree with pruning and without pruning, demonstrate how to classify new data, and analyze the performance of the Random Forest model.

2 Decision Tree Results

The Decision Tree was evaluated with and without pruning. Below are the summarized results.



Classifier output

```
Correctly Classified Instances          45               78.9474 %
Incorrectly Classified Instances        12               21.0526 %
Kappa statistic                          0.5378
Mean absolute error                      0.2677
Root mean squared error                  0.432
Relative absolute error                 58.5226 %
Root relative squared error             90.4708 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.700    0.162    0.700      0.700    0.700      0.538    0.768     0.673     bad
                 0.838    0.300    0.838      0.838    0.838      0.538    0.769     0.807     good
Weighted Avg.    0.789    0.252    0.789      0.789    0.789      0.538    0.768     0.760

=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |  a = bad
  6 31 |  b = good
```

Figure 1: Decision Tree Results

2.1 Results without Pruning

- Correctly Classified Instances: 42 (73.68%)
- Incorrectly Classified Instances: 15 (26.32%)

- Kappa Statistic: 0.4415
- Mean Absolute Error: 0.3192
- Root Mean Squared Error: 0.4669
- Total Number of Instances: 57

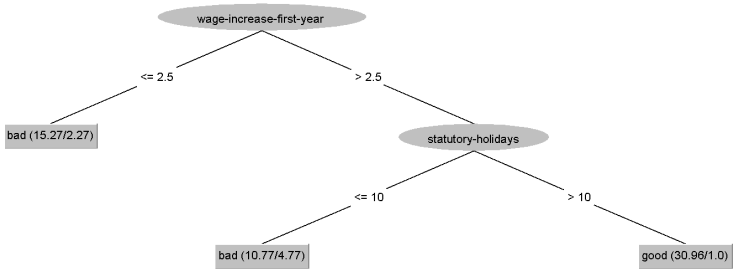


Figure 2: Decision Tree Structure pruning=TRUE

2.2 Results with Pruning

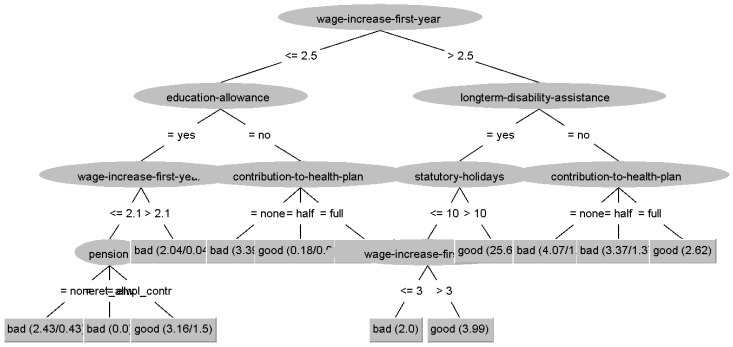
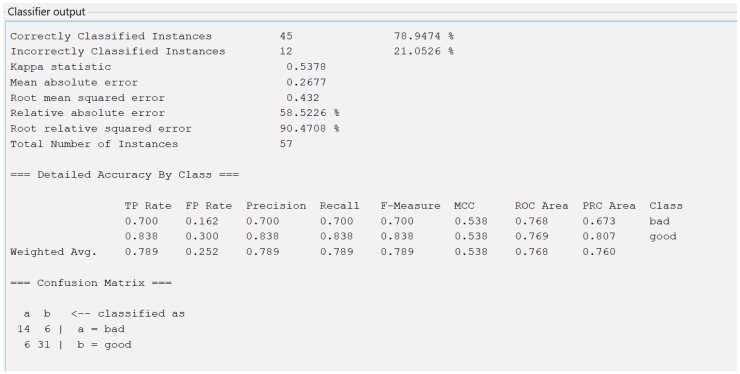


Figure 3: Decision Tree Structure pruning=False

- Correctly Classified Instances: 45 (78.95%)

- Incorrectly Classified Instances: 12 (21.05%)
- Kappa Statistic: 0.5378
- Mean Absolute Error: 0.2677
- Root Mean Squared Error: 0.4320
- Total Number of Instances: 57



Classifier output

```
Correctly Classified Instances          45               78.9474 %
Incorrectly Classified Instances        12               21.0526 %
Kappa statistic                          0.5378
Mean absolute error                      0.2677
Root mean squared error                  0.432
Relative absolute error                 58.5226 %
Root relative squared error             90.4708 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.700    0.162    0.700      0.700    0.700      0.538    0.768     0.673     bad
                 0.838    0.300    0.838      0.838    0.838      0.538    0.769     0.807     good
Weighted Avg.    0.789    0.252    0.789      0.789    0.789      0.538    0.768     0.760

=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |  a = bad
  6 31 |  b = good
```

Figure 4: Decision Tree Results pruning=TRUE

2.3 Comparison of Decision Tree Results

The results show an improvement in the performance of the Decision Tree after pruning. Specifically, the accuracy increased from 73.68% to 78.95%, and the Kappa statistic improved from 0.4415 to 0.5378. This indicates that pruning helps to reduce overfitting, enhancing the model's generalization capabilities.

3 Classifying New Data

To classify the provided data using the Decision Tree, we will follow the structure of the decision tree built from the training data. The feature values of the new data are:

Feature	Value
duration	1
shift-differential	20
wage-increase-first-year	3
education-allowance	yes
wage-increase-second-year	6
statutory-holidays	12
wage-increase-third-year	4
vacation	generous
cost-of-living-adjustment	tcf
longterm-disability-assistance	yes
working-hours	35
contribution-to-dental-plan	full
pension	ret_allw
bereavement-assistance	no
standby-pay	11
contribution-to-health-plan	half

By traversing the decision tree with these feature values, we can determine the classification (either "good" or "bad") based on the path taken through the nodes.

4 Random Forest Results

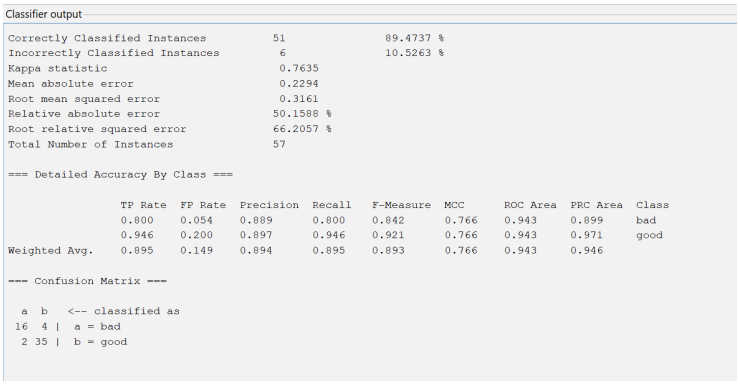


Figure 5: Random Forest Results

The Random Forest model was evaluated with the following results:

- Correctly Classified Instances: 51 (89.47%)
- Incorrectly Classified Instances: 6 (10.53%)

- Kappa Statistic: 0.7635
- Mean Absolute Error: 0.2294
- Root Mean Squared Error: 0.3161
- Total Number of Instances: 57

4.1 Detailed Accuracy By Class

Class	TP Rate
bad	0.800
good	0.946

5 Comparison between Random Forest and Decision Tree

When comparing the Random Forest model with both Decision Tree configurations (with and without pruning), the following observations can be made:

- **Correctly Classified Instances**:
  - Random Forest: 89.47%
  - Decision Tree (Pruned): 78.95%
  - Decision Tree (Unpruned): 73.68%
- **Kappa Statistic**:
  - Random Forest: 0.7635
  - Decision Tree (Pruned): 0.5378
  - Decision Tree (Unpruned): 0.4415
- **Mean Absolute Error**:
  - Random Forest: 0.2294
  - Decision Tree (Pruned): 0.2677
  - Decision Tree (Unpruned): 0.3192

5.1 Reasons for Random Forest's Superior Performance

1. **Ensemble Learning**: The Random Forest utilizes an ensemble of decision trees, leading to improved accuracy and robustness. 2. **Reduction of Overfitting**: By aggregating predictions from multiple trees, Random Forest reduces the risk of overfitting, especially compared to a single Decision Tree. 3. **Feature Randomization**: The inherent randomness in feature selection during the construction of each tree helps to enhance diversity, further improving generalization.

6 Conclusion

The Random Forest model outperformed the Decision Tree models in all evaluated metrics, demonstrating the advantages of ensemble learning techniques. Pruning the Decision Tree improved its performance but did not match the accuracy achieved by the Random Forest. Thus, for classification tasks similar to this, employing Random Forest is recommended for better predictive performance.