CENG 414 - Data Mining
Spring 2023
Homework 2

Kekevi, Kemal Anıl
2380608

May 9, 2023

2.1 In part 2.1 I upload the data as a pandas dataframe with corresponding column names. Than, I print the first 3 rows. Than, using the dataframe's shape printed the feature and observation numbers. Than, for all columns I printed the number of missing values. Than, I calculated the missing value percentage of all columns and dropped the ones that have more than 20 percent missing value. Than, using the dropna() function I create credit wo na dataframe and finally, I wrote it to csv file.

2.2 In the Multi-Layer Perceptron, I remove the index attribute because it does not give any contribute to my model. By default parameters and 5-fold cross validation:Learning rate (L): 0.3,Momentum (M): 0.2, Maximum number of epochs (N): 500, Validation set size (V): 0, Seed for random number generation (S): 0, Number of epochs before termination (E): 20, Number of hidden layers (H): 'a'

- Only 1 hidden layer created. In hidden layer there are 11 nodes created. Overall 13 nodes created 2 of them are output nodes.

- When I checked the default settings of the multi-layer proceptron classifier from Weka I see that normalize attributes option is true. So Weka normalized the attributes before classification. Effect of Normalizing attributes is: helping to improve the performance of the MLP classifier by ensuring that all input features are on a similar scale, which can prevent certain features from dominating the learning process.

- Since number of epochs before termination is 20 written in the run informations of classifier output and advance options of classifier says that validation threshold is 20, the halting strategy for the Weka MLP classifier is the "Maximum number of epochs" strategy.

- Accuracy: The proportion of cases that were correctly classified. The accuracy, according to the summary, is 65.85
- Precision: The ratio of actual positive results (TP) to occurrences that were projected to be positive (TP+FP). The precision for the "bad" class is 0.469, whereas the precision for the "good" class is 0.712. The average precision is 0.632.
- Recall: The ratio of instances that are truly positive (TP+FN) to the number of true positives (TP). The recall for the "bad" class is 0.313, while the recall for the "good" class is 0.827. The average recall is 0.659.
- The harmonic mean of recall and precision is the F1-measure. It creates a single value by combining the two measurements. The F1-measure for the "bad" class is 0.376, whereas the F1-measure for the "good" class is 0.765. The average F1-measure is 0.637.

The model appears to perform moderately overall, with high precision and recall for the "good" class but relatively low precision and recall for the "bad" class. The model's performance, which is indicated by the average F1-measure of 0.637, is not great but it is also not terrible either. The performance of the model may need to be further examined to determine whether it is adequate for the given application.

2.3 I remove the index attribute because it does not give any contribute to my model.

- The pruned decision tree has 16 nodes and 9 leaves. The root node splits the instances based on the duration attribute, with instances having "duration ¡= 42" going to the left and instances with "duration ¿ 42" going to the right. When we continue from the root node to right it gives "Bad Risk" leaf. When we continue from the root node to left it splits the data according to credit amount. If credit amount is greater than "7814" and duration is greater then "36" it gives "good Risk" label. If credit amount is greater than "7814" and duration is less then or

equal to "36" it gives "bad Risk" label. If credit amount is less then or equal to 7814 it does comparison based on Housing. If housing is "good or free" it gives "good Risk" label. If housing is "rent" then it compares instances Job. If Job value is greater then 2 it gives "good Risk" label. Else it compares duration. If duration is greater then "15" and age is greater than "45" it gives "good Risk" label. If duration is greater then "15" and age is less then or equal to "45" it gives "bad Risk" label. If duration is less then or equal to "15" it gives "good Risk" label.

- According to the Summary section, the tree successfully classified 532 out of 817 instances, a 65.12 percent accuracy rate. The Kappa value, which assesses the level of agreement between observed and expected classifications, is poor at 0.1149. The model's predictions are, on average, off by around 41.27 percent and 50.44 percent, as shown by the mean absolute error and root mean squared error of 0.4127 and 0.5044, respectively. The model's performance is poor, as shown by the high relative absolute error and root relative squared error.
- The model's performance on each class is displayed individually in the Detailed Accuracy By Class section. The model has a low precision (0.446), recall (0.261), and F-measure (0.329) for the "bad" class due to its low true positive rate (0.261) and high false positive rate (0.158) and low true positive rate (0.261). The model has a relatively high precision (0.700), recall (0.842), and F-measure (0.764) for the "good" class due to its high true positive rate (0.842) and low false positive rate (0.739).

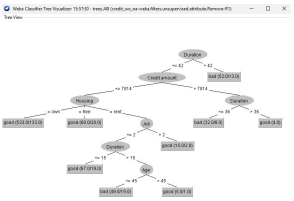

Figure 1: Rree visualization

2.4 I remove the index attribute because it does not give any contribute to my model.
I can calculate the following evaluation metrics from the output:

- Accuracy: The proportion of cases that were correctly classified. The accuracy, according to the summary, is 69.033
- Precision: The ratio of actual positive results (TP) to occurrences that were projected to be positive (TP+FP). The precision for the "bad" class is 0.556, whereas the precision for the "good" class is 0.716. The average precision is 0.664.
- Recall: The ratio of instances that are truly positive (TP+FN) to the number of true positives (TP). The recall for the "bad" class is 0.276, while the recall for the "good" class is 0.893. The average recall is 0.690.
- The harmonic mean of recall and precision is the F1-measure. It creates a single value by combining the two measurements. The F1-measure for the "bad" class is 0.369, whereas the F1-measure for the "good" class is 0.795. The average F1-measure is 0.655.

Where TP is true positive and Fp is false positive.
The model appears to perform moderately overall, with high precision and recall for the "good" class but relatively low precision and recall for the "bad" class. With an average F1-measure of 0.655, the model's performance is neither exceptional nor poor. The performance of the model may need to be further examined to determine whether it is adequate for the given application.