Texas Christian University
CoSc 30103 – 55 Spring 2025

Lab 2 Assignment Report
Sakina Ghafoor

~~~~~~

I decided to change my dataset to a smaller one since my dataset from Lab 1 was too large and the model took too long to classify. In the new dataset there is enough historic data and it is clean.

~~~~~~

File: 1 ranker with gain ratio attribute eval
In this file the results for attribute ranking are listed. The top attribute is Extracurricular Activities which means this is a significant predictor for the selected prediction of Placement Status. Projects, Placement Training, and Internships are also highly relevant for Placement Status. Academic scores (HSC, SSC, CGPA) have lower importance, indicating that practical skills and activities may be stronger indicators of placement. StudentID has a score of 0, confirming it is not useful for prediction.

~~~~~~

File: 2 naive bayes

```
Correctly Classified Instances        7952               79.52   %
Incorrectly Classified Instances      2048               20.48   %
Kappa statistic                          0.5841
Total Number of Instances            10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4612 1191 |    a = NotPlaced
  857 3340 |    b = Placed
```

The model here is for Naive Bayes with the training set. This model has moderate accuracy and the kappa value shows that the results are fair.

~~~~~~

File: 3 naive bayes with cross validat 10 fold

```
Correctly Classified Instances        7950               79.5    %
Incorrectly Classified Instances      2050               20.5    %
Kappa statistic                          0.5838
Total Number of Instances            10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4610 1193 |    a = NotPlaced
  857 3340 |    b = Placed
```

This model is also Naive Bayes but with cross validation. There is not much difference with the kappa value between the 2 models.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 4 j48
Correctly Classified Instances        8806               88.06   %
Incorrectly Classified Instances      1194               11.94   %
Kappa statistic                          0.7538
Total Number of Instances            10000
=== Confusion Matrix ===
    a    b   <-- classified as
 5270  533 |    a = NotPlaced
  661 3536 |    b = Placed
```

The J48 model does a little better than the previous Naive Bayes models. The kappa model also suggests that this model has a high level of agreement between predictions and ground truth.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 5 j48 10 fold
Correctly Classified Instances        7785               77.85   %
Incorrectly Classified Instances      2215               22.15   %
Kappa statistic                          0.5427
Total Number of Instances            10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4781 1022 |    a = NotPlaced
 1193 3004 |    b = Placed
```

This is a J48 done with cross validation. There is a decrease in the kappa value which may show that there are some inaccuracies with the J48 model or that it may be optimistic.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 6 random forest
Correctly Classified Instances       10000              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Total Number of Instances            10000
=== Confusion Matrix ===
    a    b   <-- classified as
 5803    0 |    a = NotPlaced
    0 4197 |    b = Placed
```

This model is the Random Forest which shows a perfect classification, suggesting overfitting. Likely not generalizable with these suspicious results.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 7 random forest 10 fold
Correctly Classified Instances        7915               79.15   %
Incorrectly Classified Instances      2085               20.85   %
Kappa statistic                    0.5687
Total Number of Instances           10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4875  928 |    a = NotPlaced
 1157 3040 |    b = Placed
```

This model is the Random Forest with cross validation. Accuracy is comparable to Naïve Bayes with cross-validation when looking at the kappa value.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 8 one r
Correctly Classified Instances        7567               75.67   %
Incorrectly Classified Instances      2433               24.33   %
Kappa statistic                    0.5043
Total Number of Instances           10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4472 1331 |    a = NotPlaced
 1102 3095 |    b = Placed
```

This model is the OneR which shows that it might be the weakest model in terms of accuracy and kappa.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 9 one r 10 fold
Correctly Classified Instances        7547               75.47   %
Incorrectly Classified Instances      2453               24.53   %
Kappa statistic                    0.4997
Total Number of Instances           10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4478 1325 |    a = NotPlaced
 1128 3069 |    b = Placed
```

Even with cross validation the OneR model doesn’t show significant results or a good kappa value.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 10 random forest 20 fold
Correctly Classified Instances       7907               79.07   %
Incorrectly Classified Instances     2093               20.93   %
Kappa statistic                   0.567
Total Number of Instances          10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4872  931 |    a = NotPlaced
 1162 3035 |    b = Placed
```

Since the Random Forest model was the most successful, I decided to manipulate the number of folds to see if the kappa value could get better. Here there was no significance with 20 folds.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
File: 11 random forest 5 fold
Correctly Classified Instances       7936               79.36   %
Incorrectly Classified Instances     2064               20.64   %
Kappa statistic                   0.573
Total Number of Instances          10000
=== Confusion Matrix ===
    a    b   <-- classified as
 4886  917 |    a = NotPlaced
 1147 3050 |    b = Placed
```

Here I decreased the folds to 5 for the Random Forest model. There was not any major change.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

Summary

Model	Best Accuracy	Kappa
J48 (No CV)	88.06%	0.7538
Random Forest (No CV)	100%	1.000
Naïve Bayes (10-Fold CV)	79.50%	0.5838
OneR (10-Fold CV)	75.47%	0.4997

Conclusion

I think that the J48 model with no Cross-Validation is the best practical model due to high accuracy (88.06%) and good kappa (0.7538). Naïve Bayes with 10 fold cross validation is the most stable model, consistently reaching around 79% accuracy. Random Forest with no cross validation is overfitted and unrealistic for deployment because of the perfect results. OneR is the weakest classifier, unsuitable for this dataset.