**ASSIGNMENT 4**

Note: the calculation for performance metrics can be seen in the excel that is attached with this document.

Create a classification model for letter recognition using decision trees as a classification method with a holdout partitioning technique for splitting the data into training versus testing.

* Changing the values for the depth, number of cases per parent and number of cases per leaf produces different tree configurations with different accuracies for training and testing. Choose at least five different configurations and report the accuracy for training and testing for each one of them. Which configuration will you choose as the best model? Explain your answer.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| #of parent | #of Child | Max tree Depth | # of nodes | # of terminal nodes | Depth | Training Accuracy | Testing Accuracy |
| 18 | 09 | 10 | 295 | 148 | 10 | 72% | 69.3% |
| 50 | 25 | 12 | 187 | 94 | 10 | 66.8% | 65.1% |
| 140 | 70 | 30 | 123 | 62 | 10 | 62.1% | 60.8% |
| 170 | 85 | 20 | 105 | 53 | 10 | 61% | 60.5% |
| 200 | 100 | 15 | 83 | 42 | 10 | 56.8% | 57.2% |

The model with least number of Parent and child i.e., NP =18 and Nc =09 is considered to be the best model, since it provides the highest accuracy for the data set.

* For the best tree configuration, report the misclassification matrix and interpret it. In your opinion, is accuracy a good way to interpret the performance of the model? If not, suggest other measures.



In this matrix, of the 789 actual A’s, the system predicted that 5 were D, 2 were G, 52 were K and so on. All correct guesses are located in the diagonal of the table, so it’s easy to visually inspect the table for errors, as they will be represented by values outside the diagonal.

Accuracy is not only the option to determine the performance of the classification model. There are different performance metrics to calculate the performance of the model they are: Sensitivity, Specificity and Precision

We get Sensitivity 72.07%, Specificity 98.26% and Precision 77.15% for the training set and Sensitivity 70.5%, Specificity 98.23% and Precision 74.42%

Hence these performance metrics are taken into consideration to decide the performance of the model apart from accuracy.

* What are the most important three attributes for recognizing the letters?

The three important features that can be derived from the decision tree model are x2ybr, y\_bar, xybar.

On the same data from Problem 1, apply a K-nearest neighbor classifier to classify the data. Report the following:

1. If you are doing any data transformation, explain the transformation and why it is needed.

The transformation I have chosen is Z-score. It allows us to calculate the probability of a score occurring within our normal distribution and enables us to compare two scores that are from different normal distributions. This can be achieved by converting scores in a normal distribution to z-scores in what becomes a standard normal distribution. When the distribution is normally distributed, we can find out the probability of a score occurring by standardizing the scores, known as Z-scores. The standard normal distribution converts the group of data in our frequency distribution such that the mean is 0 and standard deviation is 1.

The Z-score transformation ranges from [-3, 3], any cases beyond this range is considered to be outliers, hence the number of cases are reduced accordingly. Also the number of features selected in this case are limited to be three, considering the important three features that provides best accuracy for the model.

1. Report the misclassification matrix and the appropriate performance metrics for different values of K (K=1, 3, 5, and 7).

Misclassification matrix for untransformed data

|  |  |
| --- | --- |
| K Values | Accuracy % |
| K=1 | 83.52% |
| K=3 | 83.34% |
| K=5 | 83.15% |
| K=7 | 83% |

For K=1, is considered to be the best model since it has the highest accuracy.

K=1



K=3



K=5



K=7



Misclassification matrix after Z-score transformation:

The accuracy of the transformed data increases drastically, since the Z-score transformation is applied and the outliers are excluded and also the number of features are decreased while model building.

|  |  |
| --- | --- |
| K Values | Accuracy % |
| K=1 | 95.22% |
| K=3 | 94.92% |
| K=5 | 94.32% |
| K=7 | 94.21% |

We get Sensitivity as 95.65%, Specificity as 99% and Precision as 95.69%. Hence from this it can be noted that with the aid of Performance metrics one can choose the best model.

K=1



K=3



K=5



K=7



1. Interpret the results and also compare them with the ones obtained by using the decision trees.

The accuracy of the best model obtained in the decision tree is 72% whereas in the KNN the accuracy of the best model is K=1 with 95.22%. Since in KNN the data is transformed and also the number of features are reduced. Hence KNN is considered best compared to decision tree model.