Supply Chain Digitization (NPTEL)

Assignment 06

1. I am a retailer. Which approach should I use to anticipate and recommend products to my customers?

   a) Prescriptive analytics

   b) Common sense

   c) Predictive analytics

   d) Descriptive analytics

Ans: Predictive analytics

2. Which of the following techniques is commonly used in diagnostic analytics?

   a) Data visualization and reporting

   b) Root cause analysis and drill-down analysis

   c) Predictive modelling and machine learning

   d) Optimization algorithms and simulations

Ans: Root cause analysis and drill-down analysis

3. Why is veracity important in big data analytics?

   a) It ensures the data is stored efficiently.

   b) It guarantees high processing speed for real-time analytics.

   c) It enhances the reliability and validity of insights derived from the data.

   d) It increases the volume of data collected for better decision-making.

Ans: It enhances the reliability and validity of insights derived from the data.

4. What is a significant challenge posed by big data analytics?

   a) Slow data processing speed

   b) High data accuracy and consistency

   c) Managing and processing large volumes of diverse data

   d) Limited data sources for analysis

Ans: Managing and processing large volumes of diverse data

5. Examining historical patient data, forecasting potential health risks, and suggesting personalized treatment strategies for improved patient care includes which of the following analytics?

   a) Descriptive, Diagnostic, Prescriptive

b) Diagnostic, Prescriptive

c) Predictive, Prescriptive, Descriptive

d) Descriptive, Predictive, Prescriptive

Ans: Descriptive, Predictive, Prescriptive

6. If a decision tree splits a dataset into two groups based on a feature, and the probabilities of the two groups being positive (O) and negative (N) are known, which of the following represents the entropy of the split?

   a) $-O \log_2(O) - N \log_2(N)$

   b) $O \log_2(O) - N \log_2(N)$

   c) $1 - O^2 - Q^2$

   d) 2ON

Ans: $-O \log_2(O) - N \log_2(N)$

7. Which of the following statements is true about the values of entropy and Gini impurity index when the dataset is perfectly balanced (i.e., equal proportion of positives and negatives)?

   a) The value of entropy is always higher than the value of the Gini index.

   b) The value of entropy is always lower than the value of the Gini index.

   c) The value of entropy is greater than equal to the value of the Gini index for this dataset.

   d) Both have identical values for this dataset.

Ans: The value of entropy is always higher than the value of the Gini index.

8. Which of the following is/are used as a stopping criterion in classification tree?

   a) Levels of tree from root node.

   b) Minimum reduction in impurity

   c) Minimum reduction in MSE

   d) Minimum number of observations in each node

Ans: Levels of tree from root node., Minimum reduction in impurity, Minimum number of observations in each node

9. What will happen if the splitting of nodes is continued and there is no stopping criterion?

   a) Accuracy of the training data will increase.

   b) Accuracy of the testing data will increase.

   c) Accuracy of the training data will decrease.

   d) Overfitting of the model.

Ans: Accuracy of the training data will increase., Overfitting of the model.

10. What is the difference in entropy and gini index at Node 1 of the machine breakdown case classification tree discussed in class?

Ans: 0.46 – 0.48

Q11 – 15 are related to Machine breakdown case. You may need to use Maintenance.ipynb and Manintenance.csv provided to you. Consider 70% of the data as training data as mentioned in class.

11. If utilization is chosen as predictor variable and cut off value as 85% at Node 0, what will be the probability of machine failure when utilization <= 85%?

Ans: 0.58 – 0.59

12. Let utilization <= 85% be Node1 and utilization > 85% as Node 2. What is the gini index at Node 2?

Ans: 0.43 – 0.44

13. What is the overall reduction in impurity after one depth with utilization as predictor variable as mentioned in previous questions? Assume the gini index at Node 0 as 0.461596. (Please answer in order of 10^(3)).

Ans: 5.10 – 5.13

14. What will be the overall reduction in impurity after two depth if the first branching happened at Utilization <= 92.05 (Node 1) and second at Oil contamination <= 5.5 (Node 3)?

Ans: 0.08– 0.09

15. What is the accuracy score when max_depth = 3? Is it improved?

    (a) 0.743, yes

    (b) 0.743, no

    (c) 0.734, yes

    (d) max_depth does not play any role in accuracy score

Ans: 0.743, yes