Decision Tree Assignment

support@intellipaat.com

+91-7022374614

US: 1-800-216-8930 (Toll-Free)

Problem Statement:

Mike is working as a Machine Learning engineer, help her build a Decision Tree model and solve the following tasks:

Dataset used: Chronic Kidney Disease Data Set

Data Set Information:

Column Names	Description
Age	Age of the patient
Bp	Blood pressure of the patient
Sg	Specific gravity
Al	Albumin level
Su	Sugar level
Rbc	Red blood cells
Pc	Pus cell
Pcc	Pus cell clump
Ba	Bacteria
Bgr	Blood glucose random
Bu	Blood urea
Sc	Serum creatinine
Sod	Sodium level
Pot	Potassium level
Hemo	Hemoglobin level
Pcv	Packed cell volume
Wc	White blood cell count
Rc	Red blood cell count
Htn	Hyper tension
Dm	Diabetes Mellitus
Cad	Coronary artery disease

Appet	Appetite
Pe	Peda anema
Ane	Anemia
Classification	Classification of disease

1. Which of the following is the drawback of Decision Tree?

a. able to generate understandable rules
b. able to handle both continuous and categorical variable
c. less appropriate for estimation task
d. perform classification without requiring much computation

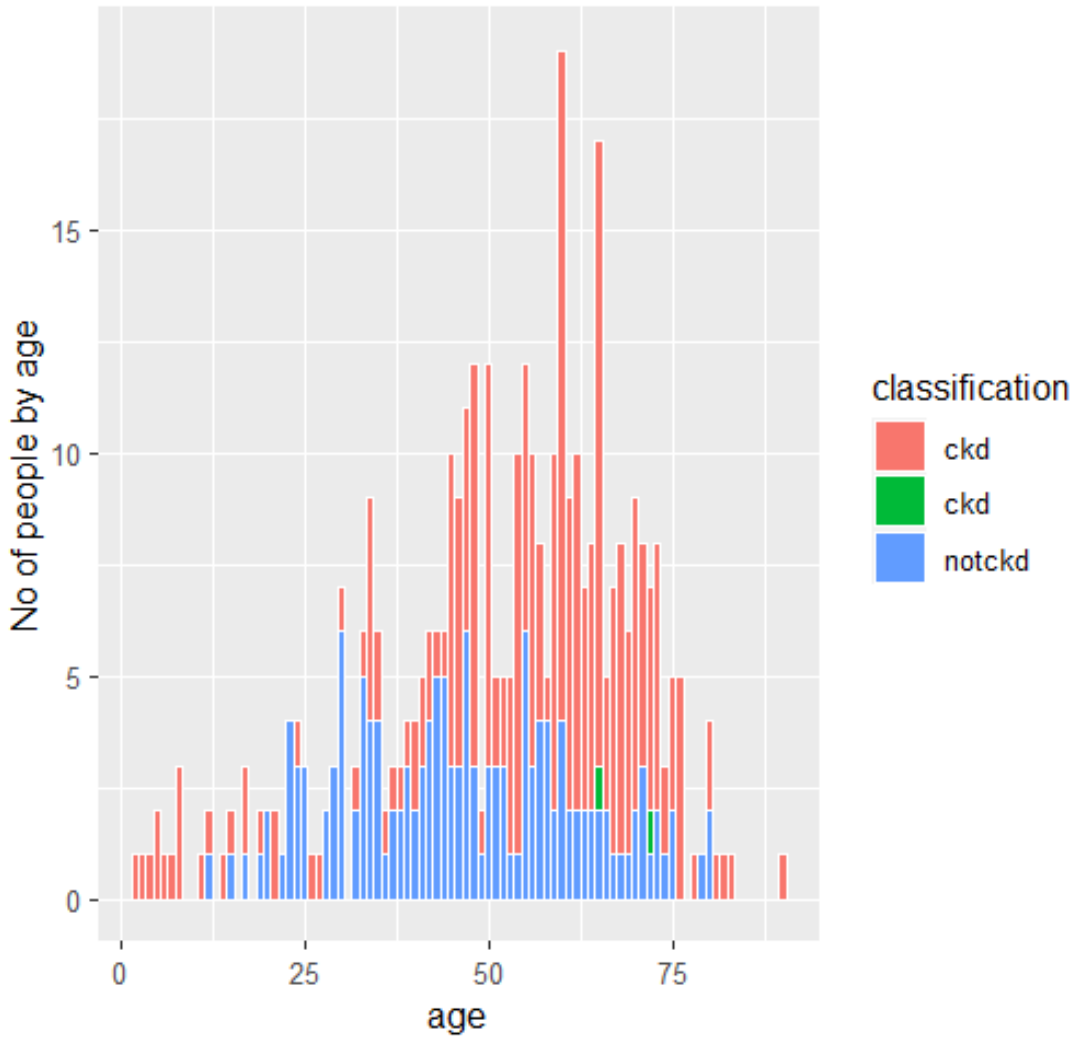
2. The measure of uncertainty of a random variable that characterizes the impurity of an arbitrary collection of examples is:

a. Information Gain
b. Gini Index
c. Entropy
d. None of these

3. Which of the following statements is not true about the Gini index?

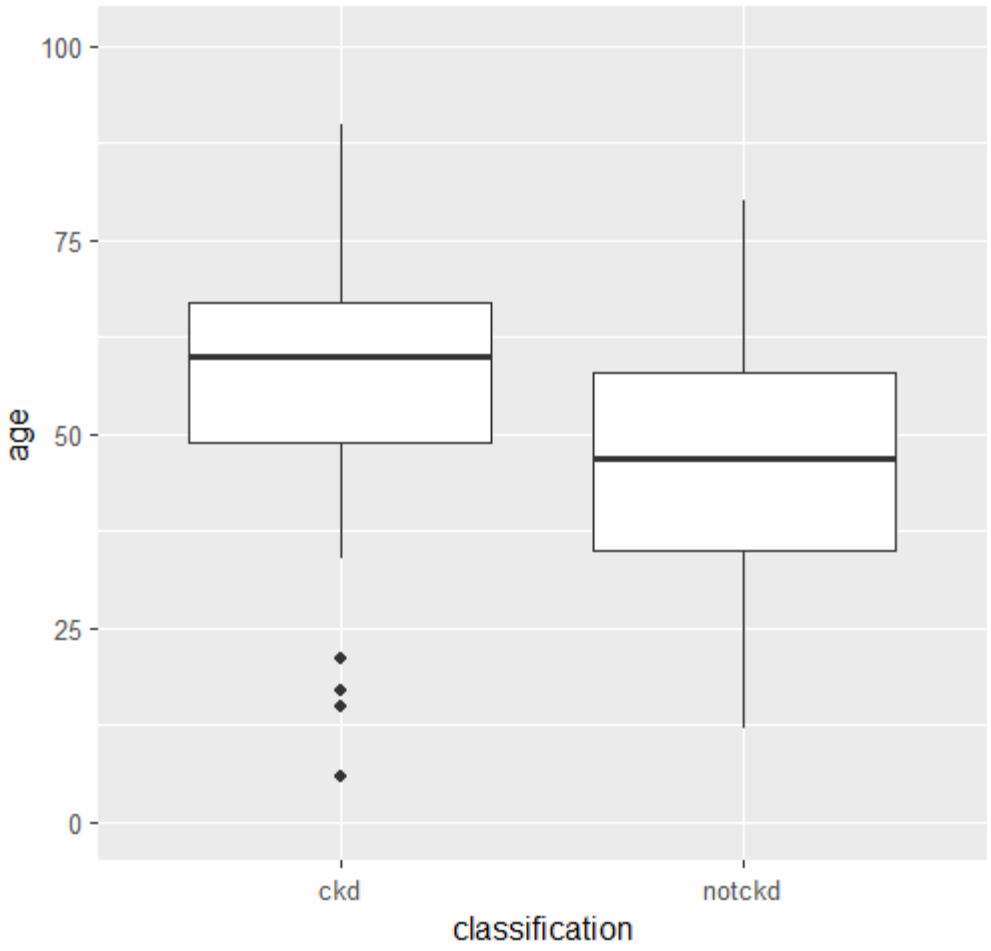
a. Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified
b. Zero gini index implies perfect classification.
c. It varies between 0 and (1-1/n) where n is the number of categories in a dependent variable.
d. None of the above

4. What inferences can be drawn from the below plot?



a. Histogram of Age distribution
b. Histogram of Age distribution grouped by Classification
c. Histogram of Age and Number of people by age
d. Histogram of Classification grouped by Age

5. From the below boxplot, identify the IQR of the Age column:

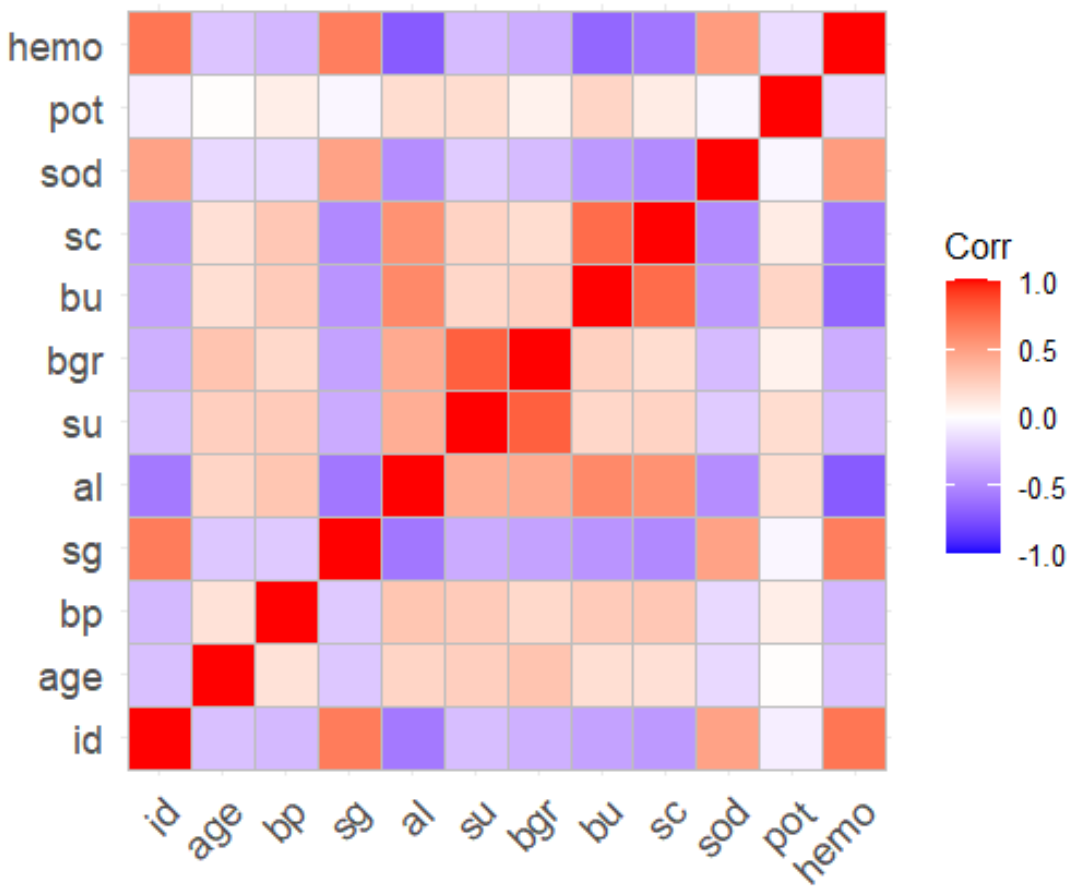


a. A.25
b. 27
c. 21
d. 20

6. Get the count of all the patients whose age id greater than 50 years and has normal red blood cells.

a. 77
b. 78
c. 73
d. 70

Given below the correlation plot of the kidney disease data, analyze the same and the following questions.



7. Which column has the best positive correlation with ‘su’ column?

   a. Al
   b. Bp
   c. Bgr
   d. hemo

8. The column ‘hemo’ has highest negative correlation with:

   a. age
   b. su
   c. bgr
   d. pot

9. _____ is the least correlated column with ‘al’ column:

   a. Sod

b. Hemo
c. Age
d. pot

10. What does the correlation value 0 or near to zero between two columns signify?

a. The two columns are independent of each other
b. The two columns have very less dependency on each other
c. Both a and b
d. None

Build a Decision tree model on the given dataset considering Classification as the target variable and Age as an independent variable, split the data based on the target variable in 80:20 ratio. Predict the values on top of the test set & store the result. Build a confusion matrix for actual and predicted values using the inbuilt function confusionMatrix() from caret package.

Answer the questions that follow considering the decision tree model built:

11. What is the accuracy of the model?

a. 0.6 to 0.7
b. 0.5 to 0.55
c. 0.75 to 0.8
d. 0.85 to 0.9

14. What is the specificity of the model?

a. 0.6 to 0.7
b. 0.75 to 0.8
c. 0.85 to 0.9
d. 0.7 to 0.75

15. What is the Misclassification rate of the above model?

a. 0.5 to 0.6
b. 0.6 to 0.7
c. 0.75 to 0.8
d. 0.7 to 0.75