USN | | | | | | | | | | |

CMRIT
CMR INSTITUTE OF TECHNOLOGY, BENGALURU.
ACCREDITED WITH A++ GRADE BY NAAC

Internal Assessment Test 2 – February 2024

Sub:	Artificial Intelligence and Machine Learning - Set 3					Sub Code:	21CS54	Branch:	ISE
Date:	01-02-2024	Duration:	90 Minutes	Max Marks:	50	Sem / Sec:	**5 A,B,C**		**OBE**

Answer any FIVE FULL Questions

<table>
<tr><th></th><th></th><th>MARKS</th><th>CO</th><th>RBT</th></tr>
<tr><td>1a</td><td>Explain in detail about the Data Visualization with different forms of Graph representation.</td><td>8</td><td>CO2</td><td>L2</td></tr>
<tr><td>1b</td><td>Define Covariance with the formula and find the covariance of data.
X= {1,2,3,4,5}and Y= {1,4,9,16,25}.</td><td>2</td><td>CO2</td><td>L1</td></tr>
<tr><td>2a</td><td>Apply Find-S algorithm for the below dataset and specify the limitation of the same.
Training Dataset:
<table>
<tr><th>CGPA</th><th>Interactiveness</th><th>Practical Knowledge</th><th>Communication Skills</th><th>Logical Thinking</th><th>Interest</th><th>Job Offer</th></tr>
<tr><td>>=9</td><td>Yes</td><td>Excellent</td><td>Good</td><td>Fast</td><td>Yes</td><td>Yes</td></tr>
<tr><td>>=9</td><td>Yes</td><td>Good</td><td>Good</td><td>Fast</td><td>Yes</td><td>Yes</td></tr>
<tr><td>>=8</td><td>No</td><td>Good</td><td>Good</td><td>Fast</td><td>No</td><td>No</td></tr>
<tr><td>>=9</td><td>Yes</td><td>Good</td><td>Good</td><td>Slow</td><td>No</td><td>Yes</td></tr>
</table></td><td>8</td><td>CO2</td><td>L3</td></tr>
<tr><td>2b</td><td>Differentiate Classical and Adaptive Machine Learning Systems</td><td>2</td><td>CO2</td><td>L2</td></tr>
<tr><td>3a</td><td>Write the algorithm for the Nearest Centroid Classifier. Consider the sample data shown in the table with two features x and y. The target classes are ‘A’ or ‘B’. Predict the class using the same.
Sample Data:
<table>
<tr><th>X</th><th>Y</th><th>Class</th></tr>
<tr><td>3</td><td>1</td><td>A</td></tr>
<tr><td>5</td><td>2</td><td>A</td></tr>
<tr><td>4</td><td>3</td><td>A</td></tr>
<tr><td>7</td><td>6</td><td>B</td></tr>
<tr><td>6</td><td>7</td><td>B</td></tr>
<tr><td>8</td><td>5</td><td>B</td></tr>
</table></td><td>8</td><td>CO2</td><td>L3</td></tr>
<tr><td>3b</td><td>Define Heuristic Space Search</td><td>2</td><td>CO2</td><td>L1</td></tr>
<tr><td>4a</td><td>Define Weighted K-NN. Consider the student performance training dataset of 8 Instances, Classify whether a student will pass or fail using Weighted K-Nearest Neighbour Algorithm for the given test Instance (7.6, 60, 8)
Training Dataset T:
<table>
<tr><th>S.No</th><th>CGPA</th><th>Assessment</th><th>Project Submitted</th><th>Result</th></tr>
<tr><td>1</td><td>9.2</td><td>85</td><td>8</td><td>Pass</td></tr>
<tr><td>2</td><td>8</td><td>80</td><td>7</td><td>Pass</td></tr>
<tr><td>3</td><td>8.5</td><td>81</td><td>8</td><td>Pass</td></tr>
<tr><td>4</td><td>6</td><td>45</td><td>5</td><td>Fail</td></tr>
<tr><td>5</td><td>6.5</td><td>50</td><td>4</td><td>Fail</td></tr>
<tr><td>6</td><td>8.2</td><td>72</td><td>7</td><td>Pass</td></tr>
<tr><td>7</td><td>5.8</td><td>38</td><td>5</td><td>Fail</td></tr>
<tr><td>8</td><td>8.9</td><td>91</td><td>9</td><td>Pass</td></tr>
</table></td><td>8</td><td>CO2</td><td>L3</td></tr>
<tr><td>4b</td><td>Write a short note non-parametric algorithm</td><td>2</td><td>CO2</td><td>L1</td></tr>
</table>

<table>
<tr><td>5a</td><td>Consider the following training set for predicting the sales of the Items.
Training Item Table:
<table><tr><th>Items</th><th>Actual Sales(In Thousands)</th></tr><tr><td>I_1</td><td>80</td></tr><tr><td>I_2</td><td>90</td></tr><tr><td>I_3</td><td>100</td></tr><tr><td>I_4</td><td>110</td></tr><tr><td>I_5</td><td>120</td></tr></table>The test items actual and prediction is given,
<table><tr><th>Test Items</th><th>Actual Value(y_i)</th><th>Predicted Value($\bar{y}_i$)</th></tr><tr><td>I_6</td><td>80</td><td>75</td></tr><tr><td>I_7</td><td>75</td><td>85</td></tr></table>Find MAE, MSE, RMSE, RelMSE, and CV for the above table with the formula.</td><td>8</td><td>CO2</td><td>L3</td></tr>
<tr><td>5b</td><td>Define Lazy Learner Algorithm</td><td>2</td><td>CO2</td><td>L1</td></tr>
<tr><td>6a</td><td>Construct a Decision Tree using C4.5 to assess a student performance during his course of study and predict whether a student will get a job offer or not in his final year of the course. The training dataset T consists of 10 data instances with attributes such as ‘CGPA’, ‘Interactiveness’,’Practical Knowledge’, and ‘Communication Skills’. The target class attribute is the ‘Job Offer’.
Training Dataset:
<table><tr><th>S.No</th><th>CGPA</th><th>Interactiveness</th><th>Practical Knowledge</th><th>Communication Skills</th><th>Job Offer</th></tr><tr><td>1</td><td>>=9</td><td>Yes</td><td>Very good</td><td>Good</td><td>Yes</td></tr><tr><td>2</td><td>>=8</td><td>No</td><td>Good</td><td>Moderate</td><td>Yes</td></tr><tr><td>3</td><td>>=9</td><td>No</td><td>Average</td><td>Poor</td><td>No</td></tr><tr><td>4</td><td><8</td><td>No</td><td>Average</td><td>Good</td><td>No</td></tr><tr><td>5</td><td>>=8</td><td>Yes</td><td>Good</td><td>Moderate</td><td>Yes</td></tr><tr><td>6</td><td>>=9</td><td>Yes</td><td>Good</td><td>Moderate</td><td>Yes</td></tr><tr><td>7</td><td><8</td><td>Yes</td><td>Good</td><td>Poor</td><td>No</td></tr><tr><td>8</td><td>>=9</td><td>No</td><td>Very good</td><td>Good</td><td>Yes</td></tr><tr><td>9</td><td>>=8</td><td>Yes</td><td>Good</td><td>Good</td><td>Yes</td></tr><tr><td>10</td><td>>=8</td><td>Yes</td><td>Average</td><td>Good</td><td>Yes</td></tr></table></td><td>10</td><td>CO2</td><td>L3</td></tr>
</table>

CI **CCI** **HOD**

USN | | | | | | | | | | |

Internal Assessment Test 2 – February 2023

Sub:	Artificial Intelligence and Machine Learning Set 2					Sub Code:	21CS54	Branch:	ISE
Date:	1-2-2024	Duration:	90 Minutes	Max Marks:	50	Sem / Sec:	**5 A,B,C**		**OBE**

Answer any FIVE FULL Questions			MARKS	CO
1a	Data Visualization with different forms of Graph representation. Bar chart Pie Chart Histogram Dot Plots	 2M 2M 2M 2M	8	CO2
1b	Covariance: Formula Solution	 1M 1M	2	CO2
2a	Solution Limitation	6M 2M	8	CO2
2b	Classical and Adaptive Machine Learning Systems	2M	2	CO2
3a	Nearest Centroid Classifier Algorithm Solution	 2M 6M	8	CO2
3b	Heuristic Space Search	2M	2	CO2
4a	Weighted K-NN Definition Solution	 2M 6M	8	CO2
4b	Non-parametric algorithm	2M	2	CO2
5a	MAE MSE RMSE RelMSE CV for the above table with the formula.	1M 2M 2M 2M 1M	8	CO2
5b	Lazy Learner Algorithm	2M	2	CO2
6a	C4.5 Algorithm Iteration: 1 Iteration: 2 Decision Tree	 4M 4M 2M	10	CO3

1a. Data Visualization with different forms of Graph representation.

Bar chart

Pie Chart

Histogram

Dot Plots.

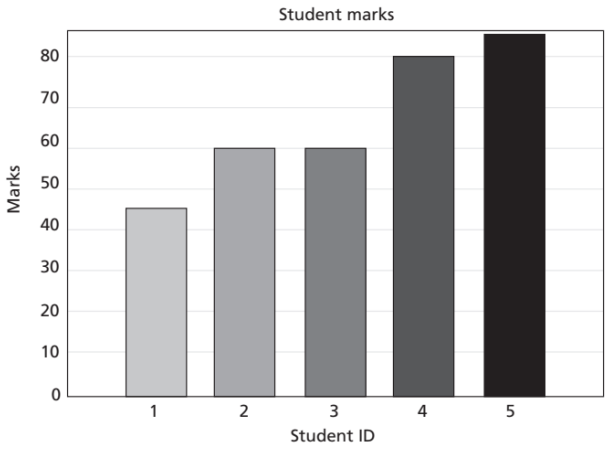
- To understand data, graph visualization is must. Data visualization helps to understand data.

- It helps to present information and data to customers. Some of the graphs that are used in univariate data analysis are **bar charts, histograms, frequency polygons and pie charts.**
- The advantages of the graphs are presentation of data, summarization of data, description of data, exploration of data, and to make comparisons of data.

Bar Chart

- A Bar chart (or Bar graph) is used to display the frequency distribution for variables.
- Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups.

The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown below in Figure



Pie Chart These are equally helpful in illustrating the univariate data. The percentage frequency distribution of students' marks {22, 22, 40, 40, 70, 70, 70, 85, 90, 90} is below in Figure 2.4.

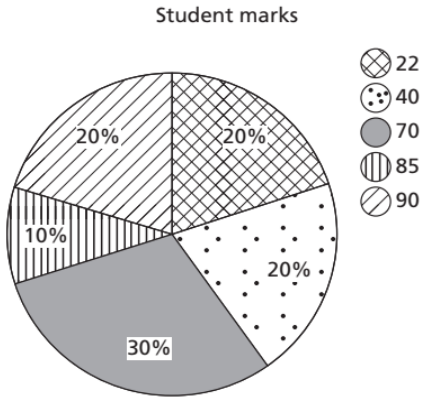


Figure 2.4: Pie Chart

It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, $2/10 \times 100 = 20\%$ space in a pie of 100% is allotted for marks 22 in Figure 2.4.

Histogram It plays an important role in data mining for showing frequency distributions. The histogram for students' marks {45, 60, 60, 80, 85} in the group range of 0–25, 26–50, 51–75, 76–100 is given below in Figure 2.5. One can visually inspect from Figure 2.5 that the number of students in the range 76–100 is 2.

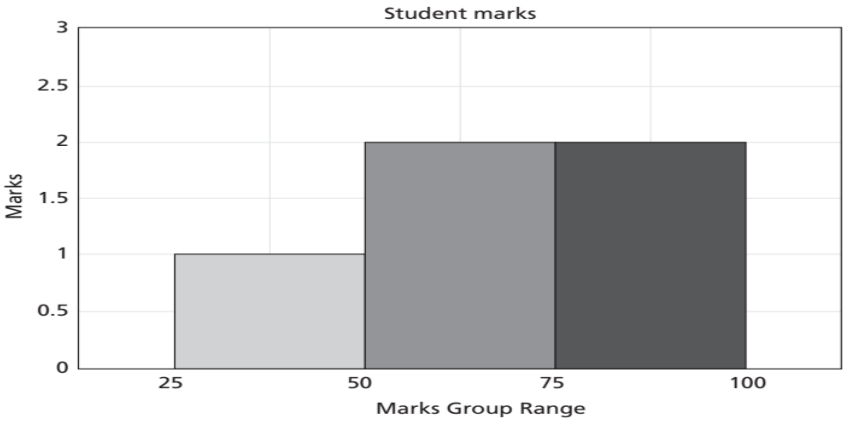
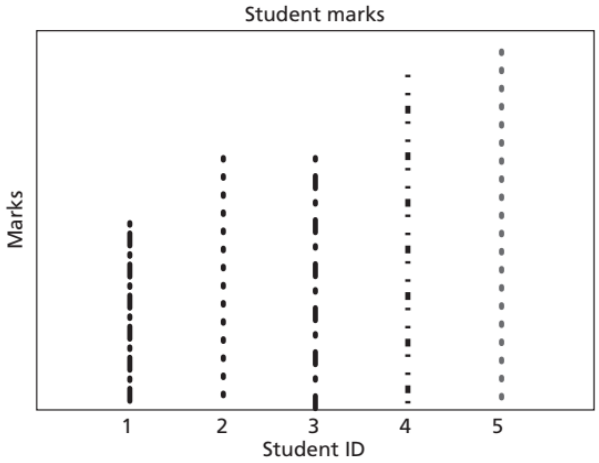


Figure 2.5: Sample Histogram of English Marks

Histogram conveys useful information like nature of data and its mode. Mode indicates the peak of dataset. In other words, histograms can be used as charts to show frequency, skewness present in the data, and shape.

Dot Plots These are similar to bar charts. They are less clustered as compared to bar charts, as they illustrate the bars only with single points. The dot plot of English marks for five students with ID as {1, 2, 3, 4, 5} and marks {45, 60, 60, 80, 85} is given in Figure 2.6. The advantage is that by visual inspection one can find out who got more marks.



1b. Covariance with the formula and find the covariance of data.
X= {1,2,3,4,5}and Y= {1,4,9,16,25}.

$$\text{cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(X))(y_i - E(Y))$$

Solution: Mean(X) = $E(X) = \frac{15}{5} = 3$, Mean(Y) = $E(Y) = \frac{55}{5} = 11$. The covariance is computed using Eq. (2.17) as:

$$\frac{(1-3)(1-11)+(2-3)(4-11)+(3-30)(9-11)+(4-3)(16-11)+(5-3)(25-11)}{5} = 12$$

The covariance between X and Y is 12. It can be normalized to a value between −1 and +1. This is done by dividing it by the correlation of variables. This is called Pearson correlation coefficient. Sometimes, $N - 1$ is also can be used instead of N. In that case, the covariance is 60/4 = 15.

2a. Find-S algorithm:

Training Dataset:

CGPA	Interactiveness	Practical Knowledge	Communication Skills	Logical Thinking	Interest	Job Offer
>=9	Yes	Excellent	Good	Fast	Yes	Yes
>=9	Yes	Good	Good	Fast	Yes	Yes
>=8	No	Good	Good	Fast	No	No
>=9	Yes	Good	Good	Slow	No	Yes

Solution:

Step 1: Initialize 'h' to the most specific hypothesis. There are 6 attributes, so for each attribute, we initially fill 'φ' in the initial hypothesis 'h'.

h = <φ φ φ φ φ φ>

Step 2: Generalize the initial hypothesis for the first positive instance. $I1$ is a positive instance, so generalize the most specific hypothesis 'h' to include this positive instance. Hence,

$I1$: ≥9 Yes Excellent Good Fast Yes **Positive instance**

h = <≥9 Yes Excellent Good Fast Yes>

Step 3: Scan the next instance $I2$, since $I2$ is a positive instance. Generalize 'h' to include positive instance $I2$. For each of the non-matching attribute value in 'h' put a '?' to include this positive instance. The third attribute value is mismatching in 'h' with $I2$, so put a '?'.

$I2$: ≥9 Yes Good Good Fast Yes **Positive instance**

h = <≥9 Yes ? Good Fast Yes>

Now,scan $I3$. Since it is a negative instance, ignore it. Hence, the hypothesis remains the same without any change after scanning $I3$.

I3: ≥8 No Good Good Fast No **Negative instance**

h = <≥9 Yes ? Good Fast Yes>

Now scan $I4$. Since it is a positive instance, check for mismatch in the hypothesis 'h' with $I4$. The 5th and 6th attribute value are mismatching, so add '?' to those attributes in 'h'.

I4: ≥9 Yes Good Good Slow No **Positive instance**

h = <≥9 Yes ? Good ? ?>

Now, the final hypothesis generated with Find-S algorithm is:

h = <≥9 Yes ? Good ? ?>

It includes all positive instances and obviously ignores any negative instance.

2b. Differentiate Classical and Adaptive Machine Learning Systems

Classical Learning Systems:

- A classical machine learning system has components such as Input, Process and Output.
- The input values are taken from the environment directly.

- These values are processed and a hypothesis is generated as output model.
- This model is then used for making predictions. The predicted values are consumed by the environment.

Adaptive Machine Learning Systems:

- The systems interact with the input for getting labelled data as direct inputs are not available.
- This process is called reinforcement learning.
- In reinforcement learning, a learning agent interacts with the environment and in return gets feedback.
- Based on the feedback, the learning agent generates input samples for learning, which are used for generating the learning model.
- Such learning agents are not static and change their behaviour according to the external signal received from the environment.
- The feedback is known as reward and learning here is the ability of the learning agent adapting to the environment based on the reward. These are the characteristics of an adaptive system.
-

3a. Write the algorithm for the Nearest Centroid Classifier. Consider the sample data shown in the table with two features x and y. The target classes are 'A' or 'B'. Predict the class using the same.

- Sample Data:

X	Y	Class
3	1	A
5	2	A
4	3	A
7	6	B
6	7	B
8	5	B

Solution:

Step 1: Compute the mean/centroid of each class. In this example there are two classes called 'A' and 'B'.

Centroid of class 'A' = (3 + 5 + 4, 1 + 2 + 3)/3 = (12, 6)/3 = (4, 2)

Centroid of class 'B' = (7 + 6 + 8, 6 + 7 + 5)/3 = (21, 18)/3 = (7, 6)

Now given a test instance (6, 5), we can predict the class.

Step 2: Calculate the Euclidean distance between test instance (6, 5) and each of the centroid.

$$\text{Euc\_Dist}[(6, 5) ; (4, 2)] = \sqrt{(6-4)^2 + (5-2)^2} = \sqrt{13} = 3.6$$

$$\text{Euc\_Dist}[(6, 5) ; (7, 6)] = \sqrt{(6-7)^2 + (5-6)^2} = \sqrt{2} = 1.414$$

The test instance has smaller distance to class B. Hence, the class of this test instance is predicted as 'B'.

3b. Define Heuristic Space Search

Heuristic search is a search strategy that finds an optimized hypothesis/solution to a problem by iteratively improving the hypothesis/solution based on a given heuristic function or a cost measure.
Heuristic search methods will generate a possible hypothesis that can be a solution in the hypothesis space or a path from the initial state

4a. Define Weighted K-NN. Consider the student performance training dataset of 8 Instances, Classify whether a student will pass or fail using Weighted K-Nearest Neighbour Algorithm for the given test Instance (7.6, 60, 8)

S.No	CGPA	Assessment	Project Submitted	Result
1	9.2	85	8	Pass
2	8	80	7	Pass
3	8.5	81	8	Pass
4	6	45	5	Fail
5	6.5	50	4	Fail
6	8.2	72	7	Pass
7	5.8	38	5	Fail
8	8.9	91	9	Pass

Solution:

Step 1: Given a test instance (7.6, 60, 8) and a set of classes {Pass, Fail}, use the training dataset to classify the test instance using Euclidean distance and weighting function.

Assign $k = 3$. The distance calculation is shown in Table 4.5.

Table 4.5: Euclidean Distance

S.No.	CGPA	Assessment	Project Submitted	Result	Euclidean Distance
1.	9.2	85	8	Pass	$\sqrt{(9.2-7.6)^2+(85-60)^2+(8-8)^2}$ $= 25.05115$
2.	8	80	7	Pass	$\sqrt{(8-7.6)^2+(80-60)^2+(7-8)^2}$ $= 20.02898$
3.	8.5	81	8	Pass	$\sqrt{(8.5-7.6)^2+(81-60)^2+(8-8)^2}$ $= 21.01928$

S.No.	CGPA	Assessment	Project Submitted	Result	Euclidean Distance
4.	6	45	5	Fail	$\sqrt{(6-7.6)^2+(45-60)^2+(5-8)^2}$ $= 15.38051$
5.	6.5	50	4	Fail	$\sqrt{(6.5-7.6)^2+(50-60)^2+(4-8)^2}$ $= 10.82636$
6.	8.2	72	7	Pass	$\sqrt{(8.2-7.6)^2+(72-60)^2+(7-8)^2}$ $= 12.05653$
7.	5.8	38	5	Fail	$\sqrt{(5.8-7.6)^2+(38-60)^2+(5-8)^2}$ $= 22.27644$
8.	8.9	91	9	Pass	$\sqrt{(8.9-7.6)^2+(91-60)^2+(9-8)^2}$ $= 31.04336$

Step 2: Sort the distances in the ascending order and select the first 3 nearest training data instances to the test instance. The selected nearest neighbors are shown in Table 4.6.

Table 4.6: Nearest Neighbors

Instance	Euclidean Distance	Class
4	15.38051	Fail
5	10.82636	Fail
6	12.05653	Pass

Step 3: Predict the class of the test instance by weighted voting technique from the 3 selected nearest instances.

- Compute the inverse of each distance of the 3 selected nearest instances as shown in Table 4.7.

Table 4.7: Inverse Distance

Instance	Euclidean Distance	Inverse Distance	Class
4	15.38051	0.06502	Fail
5	10.82636	0.092370	Fail
6	12.05653	0.08294	Pass

- Find the sum of the inverses.

  Sum = 0.06502 + 0.092370 + 0.08294 = 0.24033

- Compute the weight by dividing each inverse distance by the sum as shown in Table 4.8.

Table 4.8: Weight Calculation

Instance	Euclidean Distance	Inverse Distance	Weight = Inverse distance/Sum	Class
4	15.38051	0.06502	0.270545	Fail
5	10.82636	0.092370	0.384347	Fail
6	12.05653	0.08294	0.345109	Pass

- Add the weights of the same class.

  Fail = 0.270545 + 0.384347 = 0.654892

  Pass = 0.345109
- Predict the class by choosing the class with the maximum vote.

The class is predicted as '**Fail**'.

4b. Write a short note non-parametric algorithm

• A natural approach to similarity-based classification is k-Nearest-Neighbors (k-NN), which is a non-parametric method used for both classification and regression problems.

It is a simple and powerful non-parametric algorithm that predicts the category of the test instance according to the **'k' training samples which are closer to the test instance and classifies it to that category which has the largest probability**

5a. Consider the following training set for predicting the sales of the Items.

Training Item Table:

Items	Actual Sales(In Thousands)
I_1	80
I_2	90
I_3	100
I_4	110
I_5	120

The test items actual and prediction is given,

Test Items	Actual Value(y_i)	Predicted Value($\bar{y}_i$)
I_6	80	75
I_7	75	85

Find MAE, MSE, RMSE, RelMSE, and CV for the above table with the formula.

Mean Absolute Error (MAE) using Eq. (5.12) is given as:

$$MAE = \frac{1}{2} \times |80 - 75| + |75 - 85| = \frac{15}{2} = 7.5$$

Mean Squared Error (MSE) using Eq. (5.13) is given as:

$$MSE = \frac{1}{2} \times |80 - 75|^2 + |75 - 85|^2 = \frac{125}{2} = 62.5$$

S.No	CGPA	Interactiveness	Practical Knowledge	Communication Skills	Job Offer
1	>=9	Yes	Very good	Good	Yes
2	>=8	No	Good	Moderate	Yes
3	>=9	No	Average	Poor	No
4	<8	No	Average	Good	No
5	>=8	Yes	Good	Moderate	Yes
6	>=9	Yes	Good	Moderate	Yes
7	<8	Yes	Good	Poor	No
8	>=9	No	Very good	Good	Yes
9	>=8	Yes	Good	Good	Yes
10	>=8	Yes	Average	Good	Yes

Root Mean Square error using Eq. (5.14) is given as:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{62.5} = 7.91$$

For finding RelMSE and CV, the training table should be used to find the average of y. The average of y is $\frac{80 + 90 + 100 + 110 + 120}{5} = \frac{500}{5} = 100.$

RelMSE using Eq. (5.15) can be computed as:

$$\text{RelMSE} = \frac{(80 - 75)^2 + (75 - 85)^2}{(80 - 100)^2 + (75 - 100)^2} = \frac{125}{1025} = 0.1219$$

CV can be computed using Eq. (5.16) as $\frac{\sqrt{62.5}}{100} = 0.08.$

5b. Define Lazy Learner Algorithm

• Instance-based methods also referred to as lazy learning methods since it does not generalize any model from the training dataset but just keeps the training dataset as a knowledge base until a new instance is given.

6a. Construct a Decision Tree using C4.5 to assess a student performance during his course of study and predict whether a student will get a job offer or not in his final year of the course. The training dataset T consists of 10 data instances with attributes such as 'CGPA', 'Interactiveness','Practical Knowledge', and 'Communication Skills'. The target class attribute is the 'Job Offer'.

Training Dataset:

Iteration 1:

Step 1: Calculate the Class_Entropy for the target class 'Job Offer'.

$$\text{Entropy\_Info(Target Attribute = Job Offer)} = \text{Entropy\_Info}(7, 3) = -\left[\frac{7}{10}\log_2\frac{7}{10} + \frac{3}{10}\log_2\frac{3}{10}\right]$$
$$= (-0.3599 + -0.5208)$$
$$= 0.8807$$

Step 2: Calculate the Entropy_Info, Gain(Info_Gain), Split_Info, Gain_Ratio for each of the attribute in the training dataset.

CGPA:

$$\text{Entropy Info}(T, \text{CGPA}) = \frac{4}{10}\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right] + \frac{4}{10}\left[-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right] + \frac{2}{10}\left[-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right]$$
$$= \frac{4}{10}(0.3111 + 0.4997) + 0 + 0$$
$$= 0.3243$$
$$\text{Gain(CGPA)} = 0.8807 - 0.3243$$
$$= 0.5564$$
$$\text{Split\_Info}(T, \text{CGPA}) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{4}{10}\log_2\frac{4}{10} - \frac{2}{10}\log_2\frac{2}{10}$$
$$= 0.5285 + 0.5285 + 0.4641$$
$$= 1.5211$$
$$\text{Gain Ratio(CGPA)} = (\text{Gain(CGPA)})/(\text{Split\_Info}(T, \text{CGPA}))$$
$$= \frac{0.5564}{1.5211} = 0.3658$$

Interactiveness:

$$\text{Entropy Info}(T, \text{Interactiveness}) = \frac{6}{10}\left[-\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6}\right] + \frac{4}{10}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$$
$$= \frac{6}{10}(0.2191 + 0.4306) + \frac{4}{10}(0.4997 + 0.4997)$$
$$= 0.3898 + 0.3998 = 0.7896$$
$$\text{Gain(Interactiveness)} = 0.8807 - 0.7896 = 0.0911$$
$$\text{Gain(Interactiveness)} = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.9704$$
$$\text{Gain\_Ratio(Interactiveness)} = \frac{\text{Gain(Interactiveness)}}{\text{Split\_Info}(T, \text{Interactiveness})}$$
$$= \frac{0.0911}{0.9704}$$
$$= 0.0939$$

Practical Knowledge:

$$\text{Entropy\_Info}(T, \text{Practical Knowledge}) = \frac{2}{10}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{3}{10}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right] + \frac{5}{10}\left[-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right]$$

$$= \frac{2}{10}(0) + \frac{3}{10}(0.5280 + 0.3897) + \frac{5}{10}(0.2574 + 0.4641)$$

$$= 0 + 0.2753 + 0.3608 = 0.6361$$

$$\text{Gain(Practical Knowledge)} = 0.8807 - 0.6361$$

$$= 0.2448$$

$$\text{Split\_Info}(T, \text{Practical Knowledge}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10}$$

$$= 1.4853$$

$$\text{Gain\_Ratio(Practical Knowledge)} = \frac{\text{Gain(Practical Knowledge)}}{\text{Split\_Info}(T, \text{ Practical Knowledge})}$$

$$= \frac{0.2448}{1.4853}$$

$$= 0.1648$$

Communication Skills:

$$\text{Entropy\_Info}(T, \text{Communication Skills}) = \frac{5}{10}\left[-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right] + \frac{3}{10}\left[-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right] + \frac{2}{10}\left[-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right]$$

$$= \frac{5}{10}(0.5280 + 0.3897) + \frac{3}{10}(0) + \frac{2}{10}(0)$$

$$= 0.3609$$

$$\text{Gain(Communication Skills)} = 0.8813 - 0.36096$$

$$= 0.5202$$

$$\text{Split\_Info}(T, \text{ Communication Skills}) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10}$$

$$= 1.4853$$

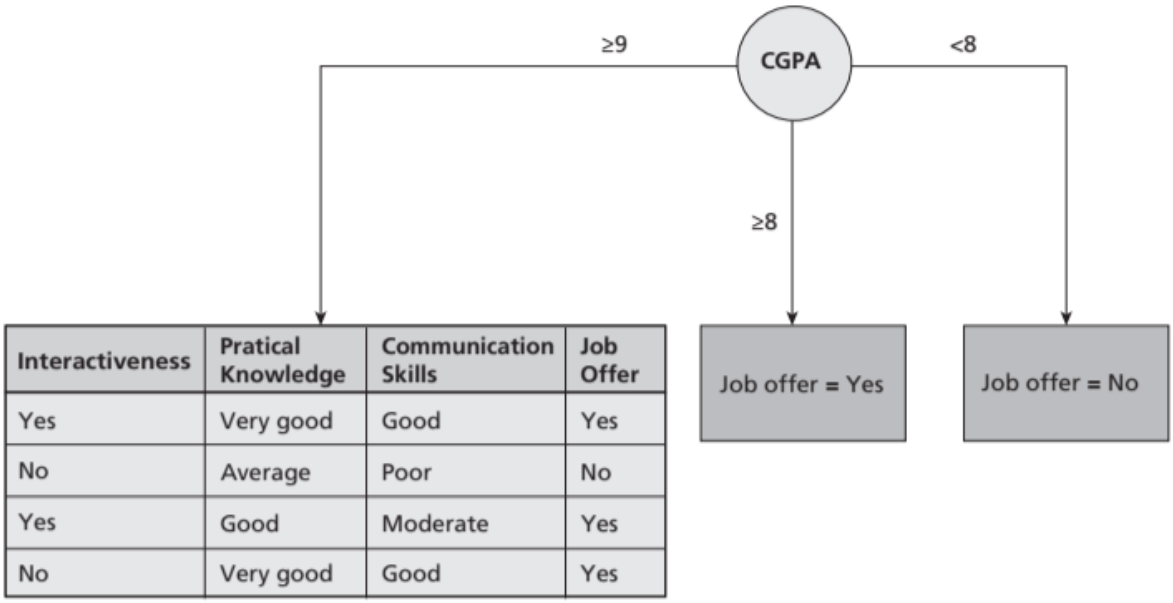
$$\text{Gain\_Ratio(Communication Skills)} = \frac{\text{Gain(Communication Skills)}}{\text{Split\_Info}(T, \text{ Communication Skills})}$$

$$= \frac{0.5202}{1.4853} = 0.3502$$

Attribute	Gain_Ratio
CGPA	**0.3658**
INTERACTIVENESS	0.0939
PRACTICAL KNOWLEDGE	0.1648
COMMUNICATION SKILLS	0.3502

Step 3: Choose the attribute for which Gain_Ratio is maximum as the best split attribute.

From Table 6.10, we can see that CGPA has highest gain ratio and it is selected as the best split attribute. We can construct the decision tree placing CGPA as the root node shown in Figure 6.5. The training dataset is split into subsets with 4 data instances.



Interactiveness	Pratical Knowledge	Communication Skills	Job Offer
Yes	Very good	Good	Yes
No	Average	Poor	No
Yes	Good	Moderate	Yes
No	Very good	Good	Yes

Figure 6.5: Decision Tree after Iteration 1

Iteration 2:

Total Samples: 4

Repeat the same process for this resultant dataset with 4 data instances.

Job Offer has 3 instances as Yes and 1 instance as No.

$$\text{Entropy\_Info(Target Class = Job Offer)} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$
$$= 0.3112 + 0.5$$
$$= 0.8112$$

Interactiveness:

$$\text{Entropy\_Info}(T, \text{Interactiveness}) = \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{2}{4}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right]$$
$$= 0 + 0.4997$$
$$\text{Gain(Interactiveness)} = 0.8108 - 0.4997 = 0.3111$$

$$\text{Split\_Info}(T, \text{Interactiveness}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 0.5 + 0.5 = 1$$

$$\text{Gain\_Ratio(Interactiveness)} = \frac{\text{Gain(Interactiveness)}}{\text{Split\_Info}(T, \text{Interactiveness})}$$
$$= \frac{0.3112}{1} = 0.3112$$

Practical Knowledge:

$$\text{Entropy\_Info}(T, \text{Practical Knowledge}) = \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{1}{4}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right] + \frac{1}{4}\left[-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$= 0$$

$$\text{Gain(Practical Knowledge)} = 0.8108$$

$$\text{Split\_Info}(T, \text{Practical Knowledge}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} = 1.5$$

$$\text{Gain\_Ratio(Practical Knowledge)} = \frac{\text{Gain(Practical Knowledge)}}{\text{Split\_Info}(T,\ \text{Practical Knowledge})} = \frac{0.8108}{1.5} = 0.5408$$

Communication Skills:

$$\text{Entropy\_Info}(T, \text{Communication Skills}) = \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right] + \frac{1}{4}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right] + \frac{1}{4}\left[-\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1}\right]$$

$$= 0$$

$$\text{Gain(Communication Skills)} = 0.8108$$

$$\text{Split\_Info}(T, \text{Communication Skills}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} = 1.5$$

$$\text{Gain\_Ratio(Communication Skills)} = \frac{\text{Gain(Practical Knowledge)}}{\text{Split\_Info}(T,\ \text{Practical Knowledge})} = \frac{0.8108}{1.5} = 0.5408$$

Attributes	Gain_Ratio
Interactiveness	0.3112
Practical Knowledge	0.5408
Communication Skills	0.5408