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

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

Scheme of Evaluation with solutions
Internal Assessment Test 1 – MAY 2021

Sub:	Big data Analytics					Sub Code:	17CS82	Branch:	ISE
Date:	22/05/2021	Duration:	90 min's	Max Marks:	50	Sem / Sec:	VIII A,B		OBE

Answer any FIVE FULL Questions

		MARKS	CO	RBT
1 (a)	**What are the key elements of a data warehouse? Explain its Architecture ?** **Scheme:** Four key elements with explanation 4*2 = 8M **Solution:** 1. Subject oriented: To be effective, a DW should be designed around a subject domain, i.e. to help solve a certain category of problems. 2. Integrated: The DW should include data from many functions that can shed light on a particular subject area. Thus the organization can benefit from a comprehensive view of the subject area. 3. Time-variant (time series): The data in DW should grow at daily or other chosen intervals. That allows latest comparisons over time. 4. Nonvolatile: DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, DW is consistently available for analysis, across the organization and over time. 5. Summarized: DW contains rolled-up data at the right level for queries and analysis. The process of rolling up the data helps create consistent granularity for effective comparisons. It also helps reduces the number of variables or dimensions of the data to make them more meaningful for the decision makers. 6. Not normalized: DW often uses a star schema, which is a rectangular central table, surrounded by some look-up tables. The single table view significantly enhances speed of queries. 7. Metadata: Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in the DW should be sufficiently well-defined. 8. Near Real-time and/or right-time (active): DWs should be updated in near real-time in many high transaction volume industries, such as airlines. The cost of implementing and updating DW in real time could be discouraging though. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.	[08]	CO3	L1
(b)	**Describe two Business Intelligence tools and its applications.** **Scheme:** Two tools and application 2M **Solution:** A spreadsheet tool, such as Microsoft Excel, can act as an easy but effective BI tool by itself. Data can be downloaded and stored in the spreadsheet, then analyzed to produce insights, then presented in the form of graphs and tables. This system offers limited automation using macros and other features. The analytical features include basic statistical and financial functions. Pivot tables help do sophisticated what-if analysis. Add-on modules can be installed to enable moderately sophisticated	[02]	CO3	L1

	statistical analysis. b)A dashboarding system, such as IBM Cognos or Tableau, can offer a sophisticated set of tools for gathering, analyzing, and presenting data. At the user end, modular dashboards can be designed and redesigned easily with a graphical user interface. The back-end data analytical capabilities include many statistical functions. The dashboards are linked to data warehouses at the back end to ensure that the tables and graphs and other elements of the dashboard are updated in real time			
2 (a)	**What is CRISP-DM data mining cycle? Compare and contrast supervised and unsupervised learning techniques?** **Scheme :** CRISP CYCLE 3M , Comparison 2M **Solution :** 1.Business Understanding: The first and most important step in data mining is asking the right business questions. A related important step is to be creative and open in proposing imaginative hypotheses for the solution. 2. Data Understanding: A related important step is to understand the data available for mining. One needs to be imaginative in scouring for many elements of data through many sources in helping address the hypotheses to solve a problem. Without relevant data, the hypotheses cannot be tested. 3. Data Preparation: The data should be relevant, clean and of high quality. It's important to assemble a team that has a mix of technical and business skills, who understand the domain and the data. It helps to improve predictive accuracy. 4. Modeling: This is the actual task of running many algorithms using the available data to discover if the hypotheses are supported. Patience is required in continuously engaging with the data until the data yields some good insights. 5. Model Evaluation: One should not accept what the data says at first. It is better to triangulate the analysis by applying multiple data mining techniques, and conducting many what-if scenarios, to build confidence in the solution 6. Dissemination and rollout: It is important that the data mining solution is presented to the key stakeholders and is deployed in the organization. The model should be eventually embedded in the organization's business processes. Supervised and Unsupervised Leaning Data may be mined to help make more efficient decisions in the future. Or it may be used to explore the data to find interesting associative patterns. The right technique depends upon the kind of problem being solved.	[05]	CO4	L2
(b)	**Describe three business applications where cluster analysis will be useful. Develop a pseudo code for K-Means algorithm.** **Scheme**: Three applications 3M, K-means -2M **Solution:** **Applications of Cluster Analysis** Cluster analysis is used in almost every field where there is a large variety of transactions. It helps provide characterization, definition, and labels for populations. It can help identify natural groupings of customers, products, patients, and so on. It can also help identify outliers in a specific domain and thus decrease the size and complexity of problems. A prominent business application of cluster analysis is in market research. Customers are segmented into clusters based on their characteristics—want and needs, geography, price sensitivity, and so on. Here are some examples of clustering:	[05]	CO4	L2

	1.*Market Segmentation*: Categorizing customers according to their similarities, for instance by their common wants and needs, and propensity to pay, can help with targeted marketing. 2. *Product portfolio*: People of similar sizes can be grouped together to make small, medium and large sizes for clothing items. 3. *Text Mining*: Clustering can help organize a given collection of text documents according to their content similarities into clusters of related topics. Here is the pseudo code for implementing a K-means algorithm. Algorithm K-Means (K number of clusters, D list of data points) 1. Choose K number of random data points as initial centroids (cluster-centers) 2. Repeat till cluster-centers stabilize a. { *Allocate each point in D to the nearest of K centroids;* b. *Compute centroid for the cluster using all points in*			
3 (a)	**What are the data visualization techniques? When would you use tables or graphs?** **Scheme:** Five techniques 5 M , Reason when to choose tables or graph 1M **Solution:** 1. Line graph. This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables. 2. Scatter plot: This is another very basic and useful graphic form. It helps several the relationship between two variables. In the above case let, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points. 3. Bar graph: A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate. 4. Stacked Bar graphs: These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar. 5. Histograms: These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.6. Pie charts: These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value. 7. Box charts: These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction. 8. Bubble Graph: This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is	[06]	CO3	L2

	a bubble (or a circle) ... the size of the circle and the color fill in the circle could represent two additional dimensions. 9. Dials: These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range. These ranges could be colored red, yellow and green to give an instant view of the data. 10. Geographical Data maps are particularly useful maps to denote statistics.			
(b)	**List Advantages and Disadvantages of Regression Models** **Scheme:** Write any two advantages and disadvantages 2*2 =4M **Solution:** **Regression Models are very popular because they offer many advantages.** 1. Regression models are easy to understand as they are built upon basic statistical principles such as correlation and least square error. 2. Regression models provide simple algebraic equations that are easy to understand and use. 3. The strength (or the goodness of fit) of the regression model is measured in terms of the correlation coefficients, and other related statistical parameters that are well understood. 4. Regression models can match and beat the predictive power of other modeling techniques. 5. Regression models can include all the variables that one wants to include in the model. 6. Regression modeling tools are pervasive. They are found in statistical packages as well as data mining packages. MS Excel spreadsheets can provide simple regression modeling capabilities. **Regression models can however prove inadequate under many circumstances.** 1. Regression models can not cover for poor data quality issues. If the data is not prepared well to remove missing values or is not well-behaved in terms of a normal distribution, the validity of the model suffers. Regression models suffer from collinearity problems (meaning strong linear correlations among some independent variables). If the independent variables have strong correlations among themselves, then they will eat into each other's predictive power and the regression coefficients will lose their ruggedness. Regression models will not automatically choose between highly collinear variables, although some packages attempt to do that. 3. Regression models can be unwieldy and unreliable if a large number of variables are included in the model. All variables entered into the model will be reflected in the regression equation, irrespective of their contribution to the predictive power of the model. There is no concept of automatic pruning of the regression model. 4. Regression models do not automatically take care of non-linearity. The user needs to imagine the kind of additional terms that might be needed to be added to the regression model to improve its fit. Regression models work only with numeric data and not with categorical variables. There are ways to deal with categorical variables though by creating multiple new variables with a yes/no value.	[04]	CO3	L2
4 (a)	**Differentiate between C4.5, CHART, CHAID decision tree algorithms.** **Scheme:** **6 comparisons 6M**	[06]	CO4	L2

Characteristic(→) / Algorithm(↓)	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection
ID3	Information Gain	Handles only Categorical value	Do not handle missing values.	No pruning is done	Susceptible on outliers
CART	Towing Criteria	Handles both Categorical and Numeric value	Handle missing values.	Cost-Complexity pruning is used	Can handle Outliers
C4.5	Gain Ratio	Handles both Categorical and Numeric value	Handle missing values.	Error Based pruning is used	Susceptible on outliers

Table 3: basic characteristic of decision tree algorithms

- **Chi-Squared Automatic Interaction Detection(CHAID)** It is one of the oldest tree classification methods originally proposed by Kass in 1980
- The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations
- The next step is to cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable
- The next step is to choose the split the predictor variable with the smallest adjusted *p*-value, i.e., the predictor variable that will yield the most significant split
- Continue this process until no further splits can be performed

(b) **Explain ID3 Algorithm.** [04] CO4 L2

Scheme:

steps and algorithm 4M

Solution:

ID3 (Iterative Dichotomiser 3): Basic Idea

• Invented by J.Ross Quinlan in 1975.

• Used to generate a decision tree from a given data set by employing a top-down, greedy search, to test each attribute at every node of the tree.

• The resulting tree is used to classify future samples

ALGORITHM

• Calculate the entropy of every attribute using the data set

• Split the set into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)

• Make a decision tree node containing that attribute

• Recurse on subsets using remaining attributes

Entropy

• In order to define information gain precisely, we need to discuss entropy first

. • A formula to calculate the homogeneity of a sample.

• A completely homogeneous sample has entropy of 0 (leaf node).

• An equally divided sample has entropy of 1.

• The formula for entropy is: **Entropy(S) = -p(I) log2 p(I)**

• where p(I) is the proportion of S belonging to class I.

∑ is over total outcomes. Log2 is log base 2.

Information Gain (IG)

• The information gain is based on the decrease in entropy after a dataset is split on

	an attribute. • The formula for calculating information gain is: Gain(S, A) = Entropy(S) - ((|Sv | / |S|) * Entropy(Sv)			
5	**Create a decision tree for the following data set. The objective is to predict the class category. (loan approved or not)**	[10]	CO5	L3

Age	Job	House	Credit	Loan Approved
Young	False	No	Fair	No
Young	False	No	Good	No
Young	True	No	Good	Yes
Young	True	Yes	Fair	Yes
Young	False	No	Fair	No
Middle	False	No	Fair	No
Middle	False	No	Good	No
Middle	True	Yes	Good	Yes
Middle	False	Yes	Excellent	Yes
Middle	False	Yes	Excellent	Yes
Old	False	Yes	Excellent	Yes
Old	False	Yes	Good	Yes
Old	True	No	Good	Yes
Old	True	No	Excellent	Yes
Old	False	No	Fair	No

Age	Job	House	Credit	Loan Approved
Young	False	No	Good	?

Scheme:

calculating Errors three cycles 3*3=9M

Tree construction and predicting solution 1M

Solution:

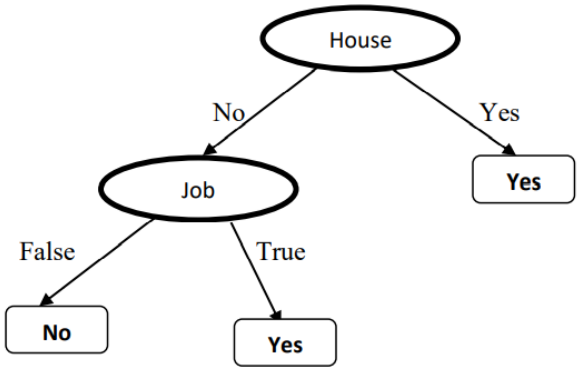
Solution :

Attributes	Rules	Error	Total Error
Age	Young →No	2/5	5/15
	Middle→Yes	2/5	
	Old→Yes	1/5	
Job	False→No	4/10	4/15
	True→Yes	0/5	
House	No→No	3/9	3/15
	Yes→Yes	0/6	
Credit	Fair→	1/5	3/15
	Good→	2/6	
	Excellent→	0/4	

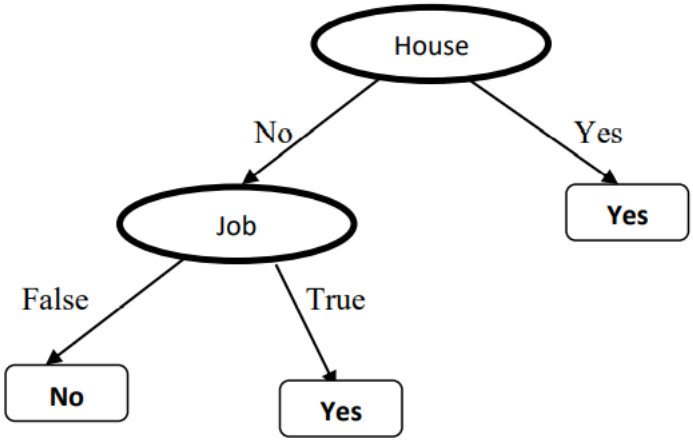
- To select the root node, we find the attribute which is having least nu
  there is a tie between two attributes, House and Credit.
- Select attribute House as a root node as it has two branches as com
  attribute

Attributes	Rules	Error	Total Error
Age	Young →No	1/4	2/9
	Middle→No	0/2	
	Old→Yes	1/3	
Job	False→No	0/6	0/9
	True→Yes	0/3	
Credit	Fair→No	0/4	2/9
	Good→Yes	2/4	
	Excellent→Yes	0/1	

- Job attribute is having least error than others



- Job attribute is having least error than others



- For the following test data **Answer is No**

Age	Job	House	Credit	LoanApproved
Young	False	No	Good	*No*

---

6 **Consider the following dataset.** [10] CO5 L3

Student	Test_mark	Grade
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

a) What linear regression equation best predicts statistics performance, based on math aptitude scores?
b) If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
c) How well does the regression equation fit the data?
Scheme:
a) 3M
b) 3M
c) 4M
Solution:

In the table below, the x_i column shows scores on the aptitude test. Similarly, the
y_i column shows statistics grades. The last two columns show deviations scores - thedifference between the student's score and the average score on each test. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	(x_i-x)	(y_i-y)
1	95	85	17	8
2	85	95	7	18
3	80	70	2	-7
4	70	65	-8	-12
5	60	70	-18	-7
Sum	390	385		
Mean	78	77		

And for each student, we also need to compute the squares of the deviation scores(the last two columns in the table below).

Student	x_i	y_i	$(x_i-x)^2$	$(y_i-y)^2$
1	95	85	289	64
2	85	95	49	324
3	80	70	4	49
4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630

And finally, for each student, we need to compute the product of the deviation scores.

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below.Notice that all of our inputs for the regression analysis come from the above three tables.

First, we solve for the regression coefficient (b_1):

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_1 = 470/730$$

$$b_1 = 0.644$$

Once we know the value of the regression coefficient (b_1), we can solve for theregression slope (b_0):

$b_0 = y - b_1 * x$

$b_0 = 77 - (0.644)(78)$

$b_0 = 26.768$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

b) In our example, the independent variable is the student's score on the aptitude test.The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade ($\hat{y}$) would be:

$\hat{y} = b_0 + b_1x$

$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80$

$\hat{y} = 26.768 + 51.52 = 78.288$

c) Whenever you use a regression equation, you should ask how well the equation fits the data. One way to assess fit is to check the coefficient of determination, whichcan be computed from the following formula.

$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value forobservation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

Computations for the sample problem of this lesson are shown below. We begin by computing the standard deviation of x (σ_x):

$\sigma_x = sqrt [\Sigma (x_i - x)^2 / N]$

$\sigma_x = sqrt(730/5) = sqrt(146) = 12.083$

Next, we find the standard deviation of y, (σ_y):

$\sigma_y = sqrt [\Sigma (y_i - y)^2 / N]$

$\sigma_y = sqrt(630/5) = sqrt(126) =$

11.225 And finally, we compute the coefficient of determination (R^2):

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2$$

$$R^2 = [(1/5) * 470 / (12.083 * 11.225)]^2$$

$$R^2 = (94 / 135.632)^2 = (0.693)^2 = 0.48$$

A coefficient of determination equal to 0.48 indicates that about 48% of the variation in statistics grades (the dependent variable) can be explained by the relationship to math aptitude scores (the independent variable). This would be considered a good fitto the data, in the sense that it would substantially improve an educator's ability to predict student performance in statistics class.