RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Faculty of Engineering and Technology			
Ramaiah University of Applied Sciences			
Department	Computer Science and Engineering	Programme	B. Tech
Semester/Batch	7th Sem /2017		
Course Code	CSC402A	Course Title	Data Mining
Course Leader	Prof. N D Gangadhar, Prof. Mohan Kumar, Prof. Santoshi Kumari		

Assignment			
Register No		Name of Student	

Sections		Marking Scheme	Marks		
			Max Marks	First Examiner Marks	Moderator Marks
Part-A					
	A 1	Data Cleaning: Redundant and Inconsistent Data	05		
	A 2	Data Cleaning: Missing Values and Outliers	05		
	A 3	Data Normalization	05		
	A 4	Data Transformation	05		
	A 5	Interpretation of Results	05		
		Part-A Max Marks	**25**		
Part -B					
	B 1.	Supervised Learning	10		
	B 2.	Un-supervised Learning	10		
	B 3	Comparative Analysis	05		
		Part-B Max Marks	**25**		
		Total Assignment Marks	**50**		

Course Marks Tabulation				
Component- CET B Assignment	**First Examiner**	**Remarks**	**Second Examiner**	**Remarks**
A				
B				
Marks (Max 50)				
Marks (out of 25)				

Signature of First Examiner

Signature of Second Examiner

RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Please note:

1. Documental evidence for all the components/parts of the assessment such as the reports, photographs, laboratory exam / tool tests are required to be attached to the assignment report in a proper order.
2. The First Examiner is required to mark the comments in RED ink and the Second Examiner's comments should be in GREEN ink.
3. The marks for all the questions of the assignment have to be written only in the **Component – CET B: Assignment** table.
4. If the variation between the marks awarded by the first examiner and the second examiner lies within +/- 3 marks, then the marks allotted by the first examiner is considered to be final. If the variation is more than +/- 3 marks then both the examiners should resolve the issue in consultation with the Chairman BoE.

Assignment

Instructions to students:

1. The assignment consists of **5** questions: Part A–**5** Question, Part B--1 Question.
2. Maximum marks is **50**.
3. The assignment has to be neatly word processed as per the prescribed format.
4. The maximum number of pages should be restricted to **20**.
5. Restrict your report for Part-A and Part-B to 20 pages each only.
6. The printed assignment must be submitted to the course leader.
7. **Submission Dates:** Part-A: 28/11/2021; Part-B: 02/01/2021
8. **Submission after the due date is not permitted.**
9. **IMPORTANT**: It is essential that all the sources used in preparation of the assignment must be suitably referenced in the text.
10. Marks will be awarded only to the sections and subsections clearly indicated as per the problem statement/exercise/question

RAMAIAH UNIVERSITY OF APPLIED SCIENCES

Preamble:

The course is intended to teach the principles, methods and techniques of data mining and its applications. Data mining algorithms, tuning them for a given application and actionable interpretations are emphasized. Students are trained to analyses, visualize and interpret the data and associated implicit insights.

Part-A

Part–A **(5 + 5 + 5 + 5 + 5 = 25 Marks)**

Consider dataset chosen in consultation with the Course Leader. Perform the following data mining operations:

1. Choose and apply suitable data cleaning methods to clean the data by eliminating redundant values and delete the records with inconsistent values
2. Implement methods to fix missing values and to identify the outliers. Eliminate the outliers from the dataset
3. Design and implement the following data transformation methods and plot the distribution of the data:
    i. Min-Max Normalization
    ii. Z-score Standardization
    iii. Decimal Scaling
4. Study and discuss how normality of data can be achieved using any two of the following transformation techniques:
    i. Square root transformation
    ii. Natural Log transformation
    iii. Inverse square root transformation
5. Preform exploratory data analysis with appropriate visualization and interpret the results obtained in the above tasks

Part-B

Part-B **(20 + 5 = 25 Marks)**

Consider the cleaned and preprocessed data from part A, and perform the following methods.

6. Design and implement any two models each from the following types to classify and categorize the data:
    i. Supervised learning
    ii. Un-supervised learning

    Input to the above mentioned model should be the normalized data mentioned in Questions 3 and/or 4 of Part-A. Every model should tested on the normalized datasets obtained there
7. Analyse the outcome various models obtained in Question 6 and discuss which data model and data transformation combination best suits the context. Justify.