IT2011 - Artificial Intelligence and Machine Learning

Department of Information Technology, Faculty of Computing

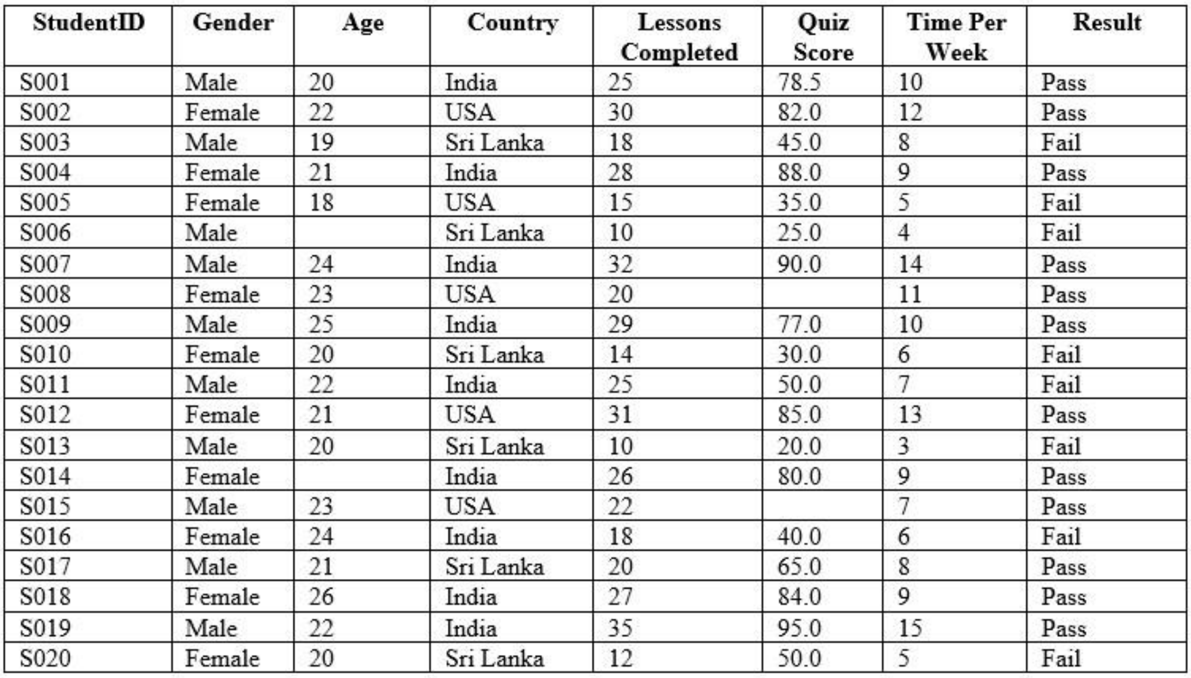
Year 2 semester 1 (2025)

Tutorial 04

Question 01: Uncovering Learning Pattern

BrightFuture Academy, an innovative online learning platform, offers digital courses to thousands of students across the globe. But despite their engaging content and experienced instructors, the academy noticed a trend where some students were dropping out or failing the final exams without any clear reasons. To solve this challenge, the school's director wants to develop a machine learning model which can predict whether a student is at risk of failing, based on their learning behavior and background. However, the data retrieved from the system is incomplete and inconsistent, and therefore, you are required to apply the data preprocessing technique.

You are provided with the following sample data set:



StudentID	Gender	Age	Country	Lessons Completed	Quiz Score	Time Per Week	Result
S001	Male	20	India	25	78.5	10	Pass
S002	Female	22	USA	30	82.0	12	Pass
S003	Male	19	Sri Lanka	18	45.0	8	Fail
S004	Female	21	India	28	88.0	9	Pass
S005	Female	18	USA	15	35.0	5	Fail
S006	Male		Sri Lanka	10	25.0	4	Fail
S007	Male	24	India	32	90.0	14	Pass
S008	Female	23	USA	20		11	Pass
S009	Male	25	India	29	77.0	10	Pass
S010	Female	20	Sri Lanka	14	30.0	6	Fail
S011	Male	22	India	25	50.0	7	Fail
S012	Female	21	USA	31	85.0	13	Pass
S013	Male	20	Sri Lanka	10	20.0	3	Fail
S014	Female		India	26	80.0	9	Pass
S015	Male	23	USA	22		7	Pass
S016	Female	24	India	18	40.0	6	Fail
S017	Male	21	Sri Lanka	20	65.0	8	Pass
S018	Female	26	India	27	84.0	9	Pass
S019	Male	22	India	35	95.0	15	Pass
S020	Female	20	Sri Lanka	12	50.0	5	Fail

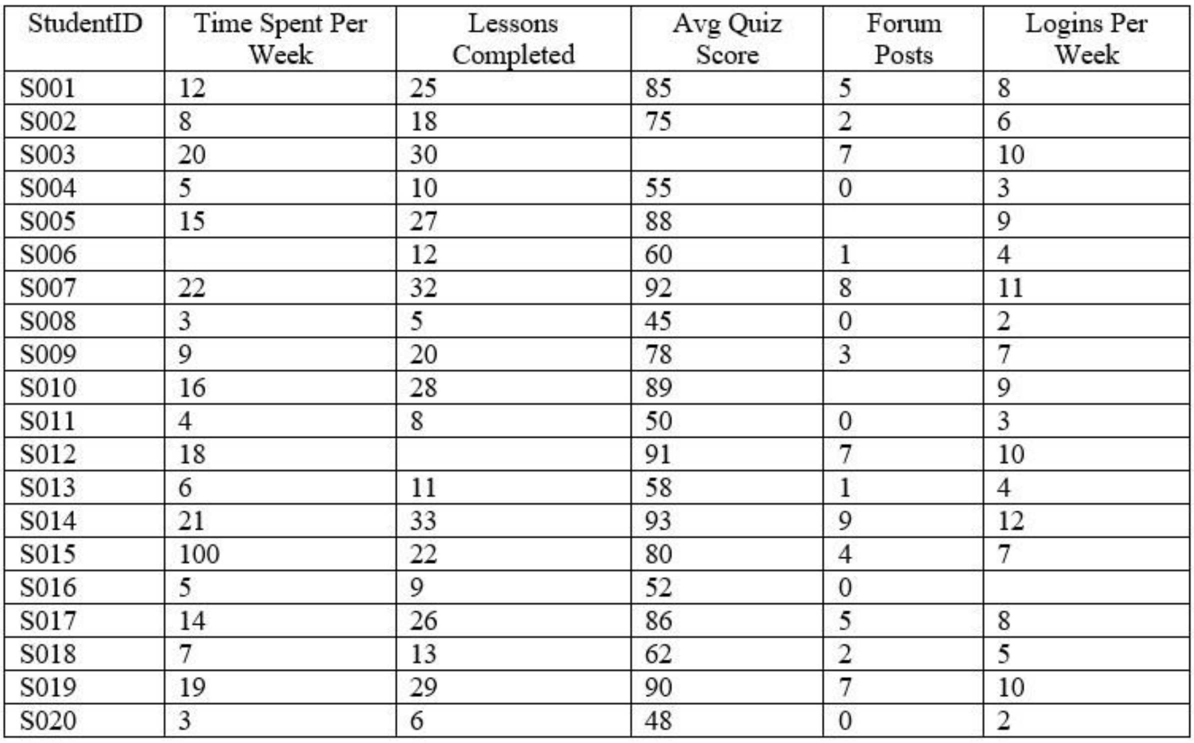
Based on the data given above,

1. Identify the type of data provided, structured or unstructured.
2. Check whether the dataset has any missing values. If yes, explain how those missing values can be handled.
3. Are there any categorical values in the data set?
   (a) If yes, mention the categorical data
   (b) Explain what encoding techniques are appropriate for the categorical data mentioned above.
4. Apply the data scaling technique to the appropriate features from the data set. Give reasons for your selection.
5. Create a new feature called EngagementLevel using the formula:
   EngagementLevel = LessonsCompleted / (TimeSpentPerWeek + 0.0001)
6. Out of all the features, which ones are likely to be the most predictive of Result? Why
7. Can we apply dimensionamilty reduction to the given data set? Justify
8. List the features of the dataset after preprocessing and feature engineering is applied.
9. Provide the preprocessed data set (Final Output of the data set)

Question 02: Understand Student Behaviour

The institute further wants to understand how students engage with their learning platform. Unfortunately, the institute does not have any labeled outcomes for this, and only raw behavioral data is available. The director wants to analyze this data and group students into meaningful categories or clusters based on how they interact with the system to tailor personalized learning experiences for each type of student. However, the dataset is inconsistent and messy, which requires a preprocessing task.

You are Given a Dataset with these Features as below:



StudentID	Time Spent Per Week	Lessons Completed	Avg Quiz Score	Forum Posts	Logins Per Week
S001	12	25	85	5	8
S002	8	18	75	2	6
S003	20	30		7	10
S004	5	10	55	0	3
S005	15	27	88		9
S006		12	60	1	4
S007	22	32	92	8	11
S008	3	5	45	0	2
S009	9	20	78	3	7
S010	16	28	89		9
S011	4	8	50	0	3
S012	18		91	7	10
S013	6	11	58	1	4
S014	21	33	93	9	12
S015	100	22	80	4	7
S016	5	9	52	0	
S017	14	26	86	5	8
S018	7	13	62	2	5
S019	19	29	90	7	10
S020	3	6	48	0	2

There is no Final Result and only the raw student interaction data is available.

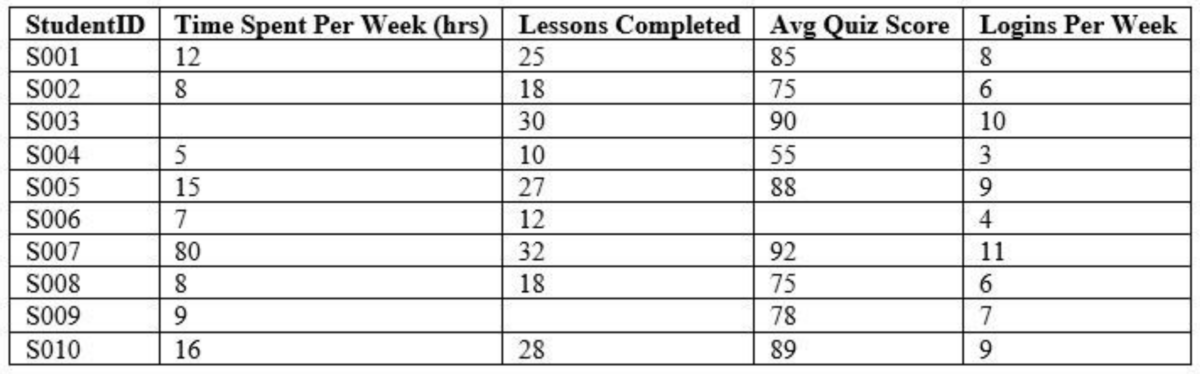
Using the provided dataset, answer the following questions:

1. Using the provided dataset, answer the following questions:
   (a) If missing data is found, describe the methods that can be used to address these gaps.
   (b) Apply the preferred method for the given dataset.
2. Identify if there are any categorical variables present in the dataset.
   (a) If categorical variables exist, specify them.
   (b) Discuss the appropriate encoding techniques for the identified categorical variables.
3. Apply a suitable data scaling method to the relevant features in the dataset and explain your choice of scaling technique.
4. Create a new feature that could provide more insights into student behavior: Learning Efficiency = Time Spent Per Week / Avg Quiz Score
5. Out of all the features, which ones are likely to be the most predictive?

6. List the features of the dataset after preprocessing and feature engineering is applied.
7. Is it suitable to apply dimensionality reduction to the data given. Give reasons.
8. Provide the preprocessed data set (Final Output of the data set)

Activity

You are given a study activity log of 10 students from a short online course. Unfortunately, the data is messy. Your task is to clean and prepare the data before it can be analyzed.



StudentID	Time Spent Per Week (hrs)	Lessons Completed	Avg Quiz Score	Logins Per Week
S001	12	25	85	8
S002	8	18	75	6
S003		30	90	10
S004	5	10	55	3
S005	15	27	88	9
S006	7	12		4
S007	80	32	92	11
S008	8	18	75	6
S009	9		78	7
S010	16	28	89	9

1. Apply data cleaning and preprocessing techniques to the data.
2. Provide the output data set (After preprocessing)