ourdiu.com | Shared by SUPAN ROY

Daffodil International University
Faculty of Science & Information Technology
Department of Software Engineering
Midterm Examination, Summer-2026
Course Code: DS421; Course Title: Machine Learning Driven Data Analysis I
Sections & Teachers: A, B, C [FA]

Time: 1:30 Hours | Marks: 25

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1.	A football analytics company wants to build a machine learning model to estimate Cristiano Ronaldo's match rating from the number of shots on target he makes during a match. The model follows a simple linear regression equation, $\hat{y} = mx$, where m is the model parameter. The model is trained using Stochastic Gradient Descent (SGD) with a learning rate of $\alpha = 0.05$, and the initial value of the weight is $m = 1.2$. The training data collected from different matches of Ronaldo is given below:	[Marks-5 +5=10]	CO-1 *Level-5*

Shots on Target (x)	Ronaldo's Match Rating (y)
2	5.8
3	6.9
5	8.1
6	9.0

a) Using this dataset, **estimate** the predicted match rating $\hat{y}$ and determine the squared error for each training sample sequentially. Then update the weight (m) using SGD after each sample and determine the final updated value of (m) after completing one full epoch over the dataset.

b) Explain how SGD helps train this Ronaldo rating prediction model. Discuss the importance of the learning rate and compare SGD with Batch Gradient Descent in terms of training speed, update behavior, and performance. Finally, mention one advantage and one limitation of SGD for football analytics applications.

2.	A football analytics team is studying the performance of FIFA World Cup winning teams and wants to reduce the number of performance variables	[Marks-10]	CO-2 *Level-3*

ourdiu.com | Shared by SUPAN ROY

before building a machine learning model for team performance analysis. Because some features are strongly correlated, the analyst uses Principal Component Analysis (PCA) to transform the data into a lower-dimensional space while preserving the maximum possible information. The analyst selects the following two features for three World Cup-winning teams:

- x_1: Average number of goals scored per match
- x_2: Average number of shots on target per match

The dataset is given below:

Team	Goals per Match (x_1)	Shots on Target per Match (x_2)
T1	2	3
T2	4	7
T3	7	12

Apply Principal Component Analysis (PCA) to the given FIFA World Cup winner dataset and determine the transformed dataset in reduced-dimensional form. Show all necessary mathematical steps and obtain the final principal component representation of the World Cup teams' performance data. Also briefly explain how PCA reduces dimensionality while preserving the most significant variance in football performance analysis.

3. A sports analytics company develops an Artificial Neural Network (ANN) model to predict whether a football player should be selected for a FIFA World Cup starting lineup based on performance statistics. After using the system for some time, analysts notice that players from less globally recognized football teams are less likely to be recommended by the model than players from top-ranked football nations, even when both groups show similar match performance data.

[Marks-5]

CO-3 *Level-4*

Analyze the fairness, ethical, and societal issues that may arise from this ANN model. **Identify** possible sources of bias in the system and suggest two measures that can be implemented to improve fairness and ensure more reliable model outcomes.