SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE		DEPARTMENT OF COMPUTER SCIENCE ENGINEERING	
Program Name: B. Tech		**Assignment Type:** Lab	**Academic Year:** 2025-26
Course Coordinator Name	Dr.Vairachilai Shenbagavel		
Instructor(s) Name	Srinivas Komakula		
Course Code	23CA201SE402	**Course Title**	Explainable & AI (P)
Year/Sem	III/V	**Regulation**	R24
Date and Day of Assignment	28-07-2025	**Time(s)**	09:00AM -05:00PM
Duration	2 Hours	**Applicable to Batch**	**23CSBTB41**
Assignment Number: 01			

Q. No.	Question	Expected Time to complete
1	Skill Boost – Coding Bootcamp	

Context:

Skill Boost sees that attending free webinars leads to higher course sign-ups.

Webinars Attended (x)	Sign-ups (y)
0	5
1	15
2	25
0	8
1	18

Objective:

Analyze the effect of attending free webinars on the number of sign-ups for Skill Boost's coding boot camp by performing Linear Regression and interpreting SHAP values.

Requirements:

1. **Perform Linear Regression Analysis**
   - Use the given dataset where:
     - ▪ Independent Variable (x): Webinars Attended
     - ▪ Dependent Variable (y): Sign-ups
2. **Calculate the Baseline Value**
   - Compute the mean of all sign-up values (y values).
3. **Calculate SHAP Values**
   - For each record, calculate the difference between the predicted value and the baseline.
   - This difference is the SHAP value, attributed to the number of webinars attended.
4. **Compute Final Prediction**
   - Use the linear regression model to calculate predicted sign-ups for each webinar count.
   - Confirm that:
     Final Prediction=Baseline+SHAP Value\text{Final Prediction} = \text{Baseline} + \text{SHAP Value}Final Prediction=Baseline+SHAP Value
5. **Interpret the Results**
   - Explain how the number of webinars attended influenced each predicted sign-up count.
   - Compare the predicted value to the actual value for each row.
   - Identify under prediction or over prediction, and provide reasoning.

Deliverables:

- Linear regression implementation with coefficients
- Baseline (mean of y)
- Table of SHAP values and predictions
- Explanation of how each input influenced the prediction
- Comparison of predicted vs actual values, with over/under prediction notes
- Summary analysis covering:
  - Accuracy of the model
  - Trend analysis
  - SHAP interpretation insights

Q. No.	Question	Expected Time to complete
2	FinTrack – Investment Signup Prediction using Multiple Linear Regression and SHAP Analysis	

Objective:

Evaluate how the number of webinars and blog posts published influence new user signups using Multiple Linear Regression and explain the outcomes using SHAP value interpretation.

Given Dataset:

Webinars (x_1)	Blogs (x_2)	Signups (y)
3	5	60
2	3	45
1	4	40
4	2	55
2	1	35

Tasks:

1. **Perform Multiple Linear Regression Analysis**
   - Use Webinars and Blogs as independent variables
   - Use Signups as the dependent variable
2. **Calculate the Baseline Value**
   - Compute the mean of all signup values
3. **Calculate SHAP Values**
   - Calculate SHAP Value
   - Distribute SHAP contributions between Webinars and Blogs based on model coefficients
4. **Compute Final Prediction for Each Record**
   - Use the regression equation
   - Verify: Prediction = Baseline + SHAP (Webinars) + SHAP (Blogs)
5. **Interpret the Results**
   - For each record, explain how webinars and blogs affected the prediction
   - Compare predicted vs actual signup values
   - Indicate if the model overpredicted or underpredicted and suggest possible causes

Q. No.	Question	Expected Time to complete
3	**Regression with Diabetes Dataset**	

Objective:

Understand how patient features influence disease progression using Multiple Linear Regression and SHAP value analysis.

Tasks

1. Perform Multiple Linear Regression Analysis

- Use all available features from the Diabetes dataset as independent variables.

- Fit a Multiple Linear Regression model to predict disease progression.

2. Calculate the Baseline Value

- Compute the **mean** of the target variable (disease progression scores) from the training data.
- This will serve as the **baseline prediction**.

3. Calculate SHAP Values

- Apply SHAP to compute **feature contributions** to each prediction.
- Use model coefficients to proportionally attribute the difference from the baseline to each feature.

4. Compute Final Prediction for Each Record

- For every test record, verify that:

Prediction = Baseline + SHAP(Feature$_1$) + SHAP(Feature$_2$) + ... + SHAP(Feature$_n$)

5. Interpret the Results

- For each patient record:
  - Explain how each feature contributed to the predicted disease progression.
  - Compare the **predicted value** vs the **actual observed value**.
  - Comment on whether the model **overpredicted or underpredicted** and **why**, based on SHAP values.

Q. No.	Question	Expected Time to complete
4	**Regression with Student Performance Dataset**	

Objective:

Investigate how student background and behavior influence final exam scores using Multiple Linear Regression and SHAP value analysis.

Tasks

1. Perform Multiple Linear Regression Analysis

- Use all relevant student attributes (e.g., study time, parental education, absences, etc.) as independent variables.
- Fit a regression model to predict the **final exam score**.

2. Calculate the Baseline Value

- Compute the **mean of the final exam scores** from the training set.
- This serves as the **baseline prediction** (expected value).

3. Calculate SHAP Values

- Use SHAP to compute the contribution of each student attribute to the final exam score prediction.
- Distribute the prediction deviation from the baseline among the features.

4. Compute Final Prediction for Each Record

- For each student record, confirm:

Predicted Score = Baseline + SHAP(Feature$_1$) + SHAP(Feature$_2$) + ... + SHAP(Feature$_n$)

5. Interpret the Results

- For every prediction:
  - Explain how different features (e.g., study time, failures, health) impacted the exam score.
  - Compare predicted score to actual score.
  - Comment on overprediction or underprediction and possible reasons behind it.