

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 Batches	23CSBTB44		
Assignment Number: 02					
Q. No.	Question				Expected Time to complete
1	Assignment: Feature Importance Analysis using SHAP				
<div>Objective</div> <div>To select a publicly available dataset from any domain, apply SHAP (SHapley Additive exPlanations) to identify important features, build a predictive model, and interpret the results in detail.</div> <div>Dataset Selection Guidelines</div> <div>Students choose datasets from the domain:</div> <div>❖ Retail &amp; E-commerce – e.g., purchase prediction, customer segmentation.</div> <div>Requirements for dataset selection:</div> <div><div>At least 500 rows of data.</div><div>Minimum 5 independent variables (features).</div><div>A clear target variable for classification or regression.</div><div>Dataset must be publicly accessible (Kaggle, UCI Repository, government portals, etc.).</div></div> <div>Tasks</div> <div><div>Data Collection &amp; Preprocessing</div><div>Download the chosen dataset in .csv format/ or any.</div><div>Load it into Python using Pandas.</div><div>Handle missing values, duplicates, and outliers.</div><div>Encode categorical variables if needed.</div><div>Normalize or standardize data when required.</div></div> <div>Model Building</div> <div><div>Split the dataset into training (80%) and testing (20%) sets.</div><div>Choose a suitable model (e.g., Random Forest, Logistic Regression, XGBoost).</div><div>Train and evaluate the model using relevant metrics:</div><div>Classification: Accuracy, Precision, Recall, F1-score, ROC.</div><div>Regression: RMSE,MSE, MAPE,MPE, MAE, R<sup>2</sup> score.</div></div> <div>SHAP Implementation</div> <div><div>Install and import SHAP (pip install shap).</div><div>Select an appropriate SHAP explainer (TreeExplainer, KernelExplainer, etc.).</div><div>Compute SHAP values for the test set.</div></div> <div>Generate and include:</div>					

- Summary plot – overall feature importance.
- Force plot – individual prediction explanation.
- Waterfall plot – step-by-step feature contribution.

### **Result Interpretation**

- Identify and explain the top 5 most influential features.
- Compare SHAP feature importance with the model's built-in feature importance (if available).
- Discuss whether the results are meaningful in the chosen domain.

### **Report Preparation**

- Title Page – Assignment title, student name, roll number, date.
- Introduction – Problem statement and dataset overview.
- Dataset Description – Source, size, features, target variable.
- Preprocessing Steps – Cleaning and transformation details.
- Model & Performance – Algorithm choice, parameters, evaluation metrics.
- SHAP Analysis – Plots and explanations.

**Conclusion** – Key insights, limitations, and possible improvements.

### **Submission Requirements**

- 🚩 Python code file (.ipynb or .py).
- 🚩 Dataset file (.csv).
- 🚩 Report (.pdf) including SHAP plots and explanations.