

| 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 | 23CSBTB32                                  |  |                           |
| Assignment Number: 02  |  |                       |  |  |                           |
| Q. No.   | Question   |                       |  |  | Expected Time to complete |
| 1  | Assignment: Feature Importance Analysis using SHAP |                       |  |  |                           |
| <p><b>Objective</b></p> <p>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.</p> <p><b>Dataset Selection Guidelines</b></p> <p><b>Students choose datasets from the domain:</b></p> <p>❖ <b>Finance &amp; Business – e.g., credit risk, fraud detection, marketing campaigns.</b></p> <p><b>Requirements for dataset selection:</b></p> <ul style="list-style-type: none"><li>• At least 500 rows of data.</li><li>• Minimum 5 independent variables (features).</li><li>• A clear target variable for classification or regression.</li><li>• Dataset must be publicly accessible (Kaggle, UCI Repository, government portals, etc.).</li></ul> <p><b>Tasks</b></p> <ul style="list-style-type: none"><li>• Data Collection &amp; Preprocessing</li><li>• Download the chosen dataset in .csv format/ or any.</li><li>• Load it into Python using Pandas.</li><li>• Handle missing values, duplicates, and outliers.</li><li>• Encode categorical variables if needed.</li><li>• Normalize or standardize data when required.</li></ul> <p><b>Model Building</b></p> <ul style="list-style-type: none"><li>• Split the dataset into training (80%) and testing (20%) sets.</li><li>• Choose a suitable model (e.g., Random Forest, Logistic Regression, XGBoost).</li><li>• Train and evaluate the model using relevant metrics:</li><li>• Classification: Accuracy, Precision, Recall, F1-score, ROC.</li><li>• Regression: RMSE,MSE, MAPE,MPE, MAE, R<sup>2</sup> score.</li></ul> <p><b>SHAP Implementation</b></p> <ul style="list-style-type: none"><li>• Install and import SHAP (pip install shap).</li><li>• Select an appropriate SHAP explainer (TreeExplainer, KernelExplainer, etc.).</li><li>• Compute SHAP values for the test set.</li></ul> |  |                       |  |  |                           |

**Generate and include:**

- 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.