Mirpur University of Science & Technology, MUST Mirpur AJ&K

Department of Software Engineering

Subject:	Introduction to DataScience	**Course Code:**	BSE-2304
Instructor:	Engr. Abdul Qadir Khan	**Date:**	28th Jan, 2024

Assignment 02 (Complex Engineering Problem)

Problem #	**CLO**	**Domain**	**BT-Level**
1	3,4	C	3,5

Assignment Title: Comprehensive Data Analysis, Visualization, and Decision Tree

Instructions:

Objective: This assignment aims to assess students' ability to apply data cleaning techniques, perform exploratory data analysis (EDA), and implement Python tools to process, transform, summarize, and visualize data effectively. Additionally, students will implement a decision tree model to demonstrate their understanding of supervised machine learning. The task integrates learning outcomes at cognitive levels C-3 (apply) and C-5 (implement).

1. **Dataset Selection:**

- Download a dataset of your choice from trusted sources such as Kaggle, UCI Machine Learning Repository etc
- Ensure the dataset has sufficient complexity, including missing values, categorical variables, and numerical variables.

2. **Tasks:**

- **Part 1: Data Cleaning and Preparation (C-3)**

1. Identify and handle missing values.
2. Identify and remove duplicate records.
3. Detect and address outliers using appropriate methods.
4. Normalize/scale numerical data for better analysis.
5. Encode categorical variables appropriately (e.g., one-hot encoding, label encoding).

- **Part 2: Exploratory Data Analysis and Visualization (C-3)**

1. Summarize the dataset by calculating descriptive statistics (mean, median, variance, etc.).
2. Identify correlations among numerical variables using correlation coefficients and heatmaps.
3. Explore and visualize key insights using at least three different types of visualizations (e.g., bar chart, scatter plot, histogram, boxplot).

- **Part 3: Python Implementation for Data Science Workflow and Decision Tree (C-5)**

1. **Collect:** Load the dataset into Python using pandas.
2. **Process:** Perform the data cleaning steps outlined in Part 1.
3. **Transform:** Apply necessary transformations to prepare data for analysis.
4. **Summarize:** Use Python libraries (e.g., pandas, numpy) to generate summary statistics.
5. **Visualize:** Utilize Python libraries (e.g., matplotlib, seaborn, plotly) to create meaningful visualizations that reflect the results of your EDA.
6. **Implement** a decision tree model:
7. **Split** the dataset into training and testing sets.
8. **Train** the model on the training data.
9. **Evaluate** the model's performance using metrics such as accuracy, precision, recall, or mean squared error (depending on whether the task is classification or regression).
10. **Visualize** the decision tree structure.

- **Report Submission:**

1. Submit a Python notebook (*.ipynb) that includes:
2. Code for all steps performed.
3. Inline comments explaining your code.
4. Visualization outputs embedded within the notebook.
5. Decision tree model implementation and evaluation results.
6. A PDF report summarizing:
7. Objectives and dataset description.
8. Key findings from your EDA and visualizations.
9. Reflection on the data cleaning and preparation process.
10. Insights from the decision tree implementation and evaluation.

Evaluation Notes:

- Use of innovative or advanced techniques for analysis, visualization, or decision tree implementation will receive extra credit.
- Ensure all submissions follow the prescribed format and include clear documentation.

Note: Plagiarism in any form will not be tolerated and may result in severe penalties.

Submission Deadline: 10 February, 2025

Rubrics:

Criteria	**Excellent (4)**	**Good (3)**	**Satisfactory (2)**	**Needs Improvement (1)**
Data Cleaning (C-3)	Thorough and accurate cleaning; all issues addressed.	Most issues addressed; minor errors.	Some issues addressed; significant errors.	Minimal effort; major issues remain.
Exploratory Data Analysis (C-3)	Comprehensive EDA with clear insights and diverse methods.	Adequate EDA with useful insights and visualization.	Basic EDA; limited insights or poor visualizations.	Minimal EDA with unclear or no insights.
Python Implementation and Decision Tree (C-5)	Efficient, well-commented, and error-free code; advanced usage of libraries; insightful model evaluation.	Functional code with minimal errors; good library usage; reasonable model evaluation.	Code runs with minor issues; basic library usage; limited model evaluation.	Non-functional or poorly written code; incomplete model evaluation.
Visualization (C-3, C-5)	Clear, insightful, and aesthetically pleasing visualizations.	Clear and useful visualizations; minor design issues.	Basic visualizations; limited insight.	Poor or missing visualizations.
Report Quality (C-5)	Well-structured, detailed, and error-free report.	Clear and mostly detailed report; few errors.	Basic report with some gaps or errors.	Poorly written report with major gaps.