**Assignment 4**

**Total Marks: 50**

**Objective**:

In this assignment, you are given two datasets: ***fraud\_detection.xls*** and ***smart\_city\_air\_quality.xls***. You can choose which dataset to work with, but it is strongly recommended that students from the IoT class work with the *smart\_city\_air\_quality.xls* dataset.

Your task is to analyze the dataset using machine learning and data analytics techniques, depending on the dataset you choose.

**Task 1: Fraud Detection**

For the *fraud\_detection.xls* dataset, your goal is to build a model that can predict whether a transaction is fraudulent based on various features in the dataset. You will need to perform the following steps:

1. **Data Preprocessing (5 Marks)**: Load and explore the dataset, handle missing values and outliers, and split it into training (80%) and testing (20%) sets.
2. **Modeling (35 Marks)**: Implement either GDA or Naive Bayes, train the model, and evaluate performance using metrics like accuracy, precision, recall, and F1 score. Perform cross-validation for generalization.
3. **Interpretation (10 Marks)**: Analyze important features, discuss potential issues like overfitting, and suggest improvements.

**Deliverables**:

* + Jupyter Notebook code that implements the solution (using libraries such as sklearn or similar).

**Task 2: Smart City Air Quality Analysis**

For the *smart\_city\_air\_quality.xls* dataset, your goal is to explore the dataset and generate meaningful insights (through classification tasks) related to air quality in a smart city using machine learning and data analysis techniques.

1. **Data Preprocessing (10 Marks)**: Load, clean, and scale the dataset, handling missing values and visualizing trends in air quality indicators (temperature, humidity, etc.).
2. **Exploratory Data Analysis (EDA) (5 Marks)**: Identify correlations and patterns related to time, season, or location; visualize key findings with plots.
3. **Predictive Modelling (25 Marks)**: Use either GDA or Naive Bayes to find gain interesting insights.
4. **Model Evaluation (10 Marks)**: Assess the model’s performance and identify the most important features affecting predictions.

**Deliverables**:

* + Jupyter notebook that implements your data preprocessing, analysis, and predictive model (using libraries like pandas, sklearn, matplotlib, etc.).

**Due Date**: 15th November 2024