|  |  |
| --- | --- |
| **Name: Shubhra Jain** | **Roll No.: 22/AI/55** |
| **Reg. No: PCE22CA054** | **Branch: CS(AI)** |

**ASSIGNMENT**

**POORNIMA COLLEGE OF ENGINEERING, JAIPUR**

**III B.TECH. (VI Sem.) SEC- D**

**Code: 6CAI6-02**

**Subject Name–Machine Learning**

**(BRANCH: ADVANCE COMPUTING (AI))**

**Max. Time: 2 hrs. Max. Marks: 20 Marks**

**INSTRUCTIONS: UPLOAD THE SOLUTION ON YOUR GITHUB REPOSITORY and MENTIONED THE URL OF THE REPOSITORY ON TCSION**

**ASSIGNMENT QUESTION 1: CO3**

**Fault Prediction Using Supervised Machine Learning**

**Problem Context**

**You are an engineer working for a power distribution company responsible for maintaining and ensuring the reliability of the electrical grid. Your task is to develop a system for detecting and classifying electrical faults in the grid. Electrical faults can lead to disruptions, damage equipment, and pose safety hazards. The company is interested in a predictive maintenance system that can identify and classify different types of electrical faults to facilitate timely intervention.**

**Fault Prediction Dataset:** [**https://www.kaggle.com/code/pythonafroz/fault-prediction-usingdecision-tree-algorithm**](https://www.kaggle.com/code/pythonafroz/fault-prediction-using-decision-tree-algorithm)

|  |  |  |
| --- | --- | --- |
| S/r No. | Question | Marks |
| Q1. | Name any 4 libraries required for the implementation of the problem statement using python | 2 |
| Ans1. | 1. **Pandas:** For Data Manipulation(loading,cleaning and preprocessing the dataset) 2. **NumPy:** For numerical computations and to work with arrays. 3. **Matplotlib/Seaborn:** For visualizing data distributions, correlations, and model performance. 4. **Scikit-learn(sklearn):** For applying ML algors (Eg. Decision Tree) |  |
| Q2. | Go through the above Kaggle link and answer the following: About this dataset file:  [https://www.kaggle.com/code/pythonafroz/fault-prediction-using-decisiontree-algorithm](https://www.kaggle.com/code/pythonafroz/fault-prediction-using-decision-tree-algorithm)   1. Total no. of columns in the dataset: 10 columns 2. Write and count input columns: 6 columns(la, lb, lc, Va, Vb, Vc) 3. Write and count the output column: 4 columns[G C B A] | 3 |
| Q3. | What is the purpose this library used for the given problem statement:   |  | | --- | | from sklearn.preprocessing import LabelEncoder | | 1 |
| Ans3. | Convert target class (fault type) or any categorical feature into numeric values before training the model. |  |
| Q4. | What is the purpose this library used for the given problem statement:   |  | | --- | | from sklearn.model\_selection import train\_test\_split | | 1 |
| Ans4. | Used to split the dataset into training and testing sets. |  |
| Q4. | List all the algorithms through which you can able to find Electrical Faults Detection and Classification | 3 |
| Ans4. | 1. **Decision Tree:** Simple, interpretable model ideal for classification tasks like fault type prediction. 2. **Random Forest:** Ensemble of decision trees; improves accuracy and reduces overfitting. 3. **Naïve Bayes:**. Useful for quick and probabilistic fault classification. 4. **Logistic Regression:** Good for binary fault detection problems (fault vs. no fault). 5. **Support Vector Machine:** Effective for high-dimensional spaces & fault pattern recognition. |  |
| Q5. | How to read the Classification\_Report generated by several models in the given problem statement:  [https://www.kaggle.com/code/pythonafroz/fault-prediction-using-decisiontree-algorithm](https://www.kaggle.com/code/pythonafroz/fault-prediction-using-decision-tree-algorithm) | 5 |
| Ans5. | The **classification\_report** in scikit-learn provides a detailed summary of a model’s classification performance. It includes **precision**, **recall**, **f1-score**, and **support** for each class.   1. **Precision: It** measures the accuracy of positive predictions. It tells us **how many of the samples predicted as a particular fault class were actually correct. High precision** means **low false positives.**      1. **Recall: Recall (True Positive Rate**) tells us **how well the model identifies all relevant instances** of a particular class. High recall means the model catches most of the faults. 2. **F1-Score:** It is the **harmonic mean** of Precision and Recall. It provides a **single metric** that balances the trade-off between the two. F1 Score is ideal for **imbalanced datasets.** 3. **Support: It** refers to the **number of actual occurrences** of each class (fault type) in the dataset. If a class has **very low support**, a model may perform poorly on it. |  |
| Q6. | From the mentioned link: [https://www.kaggle.com/code/pythonafroz/faultprediction-using-decision-tree-algorithm](https://www.kaggle.com/code/pythonafroz/fault-prediction-using-decision-tree-algorithm)  Do one sight analysis and figure out which algorithms work well on the given dataset. And on what basis are Model comparisons done over there? | 5 |
| Ans6. | The algorithms that work the best for this dataset are:   1. **Decision Tree Classifier:**   - Accuracy: 100%  -It is interpretable and simple.  - It works well when data is clean and not noisy.   1. **Random Forest Classifier:**   - Accuracy: 99-100%  - It handles overfitting better than a single Decision Tree.   1. **XGBoost:**   - Accuracy: 99-100%  - It is great with tabular data.   1. **Support Vector Machine(SVM):**   - Accuracy: 98-99%  - Effective in high-dimensional spaces.  - It is good when margin between classes exists.  These metrics are used for the analysis :   * **Accuracy** → Total correct predictions / Total samples * **Precision** → TP / (TP + FP): Focuses on exactness * **Recall** → TP / (TP + FN): Focuses on completeness * **F1-Score** → Harmonic mean of precision and recall |  |

**ASSIGNMENT QUESTION 2: CO4**

**Customer Segmentation using Unsupervised Problem Context:**

|  |  |
| --- | --- |
| **Name: Shubhra Jain** | **Roll No.: 22/AI/55** |
| **Reg. No: PCE22CA054** | **Branch: CS(AI)** |

**You are a data scientist working for a retail company that wants to improve its marketing strategies by better understanding customer behaviour. One approach is to segment customers into distinct groups based on their purchasing habits. This will allow the company to tailor marketing campaigns to specific groups, ultimately increasing sales and customer satisfaction.**

**Task:**

**Design and explain the customer segmentation model using any one of the unsupervised algorithms.**

1. Data Collection: Obtain a dataset containing customer purchase history, including details such as purchase frequency, amount spent, types of products purchased, etc. You may use publicly available datasets or simulate data for this assignment Include the first 10 rows of the dataset that you are going to consider. 5 marks

1. Data Preprocessing: Clean the dataset and perform necessary preprocessing steps such as normalization, handling missing values, and feature engineering. 5 marks
2. Unsupervised Learning (Clustering): Apply an unsupervised learning algorithm (e.g., Kmeans clustering, hierarchical clustering) to segment customers into distinct groups based on their purchasing behaviour. 5 marks
3. Evaluate the clustering results using appropriate evaluation metrics. 5 marks





