Program Project

Program Identification

Program ID: T5

Program Title: Machine Learning Module

Program Information:

Project Title:	**Comparative Analysis of Machine Learning Models on the Iris Dataset**

Description

Project Overview: The aim of this project is to provide hands-on experience to trainees in both supervised and unsupervised machine learning tasks using the famous Iris dataset. The project involves two main tasks:

1.Unsupervised Learning: Clustering and outlier detection on the dataset.

2.Supervised Learning: Building baseline models, comparing multiple algorithms, tuning the best-performing model, and evaluating it against an ensemble of algorithms.

Dataset: **Iris Dataset**

- The Iris dataset is a well-known dataset in the machine learning community, containing 150 samples of iris flowers, each with four features (sepal length, sepal width, petal length, and petal width) and a target variable specifying the iris species (setosa, versicolor, or virginica).
- **Dataset Link: Iris Dataset: https://archive.ics.uci.edu/ml/datasets/iris**

Tasks and Mark Distribution:

1. Data Preprocessing (10 marks):

- Load the Iris dataset.
- Perform data exploration and visualization.
- Check for missing values and handle them if any.
- Split the dataset into features and target variables.

2. Unsupervised Learning: Clustering and Outlier Detection (20 marks):

- Apply K-means clustering algorithm to cluster the data.
- Visualize the clusters.
- Detect outliers using appropriate techniques such as isolation forest or DBSCAN.
- Evaluate the clustering results.

3. Supervised Learning: Baseline Model (10 marks):

- Choose an appropriate evaluation metric based on the problem (classification).
- Split the dataset into training and testing sets.
- Build a baseline model (e.g., logistic regression or decision tree) using default parameters.
- Evaluate the baseline model's performance.

4. Model Comparison (30 marks):

- Select 3-4 machine learning algorithms (e.g., SVM, Random Forest, Gradient Boosting) suitable for the problem.
- Implement each algorithm and evaluate its performance using cross-validation.
- Compare the performance of algorithms based on evaluation metrics.
- Select the best-performing algorithm.

5. Model Tuning and Ensemble (20 marks):

- Perform hyperparameter tuning on the best-performing algorithm using Grid Search or Random Search.
- Evaluate the tuned model's performance.
- Implement an ensemble of the top-performing algorithms and compare its performance with the tuned model.

6. Documentation and Presentation (10 marks):

- Provide a clear and concise report documenting the project process, including data preprocessing, model implementation, evaluation, and conclusions.
- Prepare a presentation summarizing the key findings and insights from the project.

Total Marks: 100

Note:

- Trainees are encouraged to seek guidance from instructors, conduct additional research, and experiment with different approaches to enhance their understanding and skills in machine learning. The project aims to provide a comprehensive learning experience in both supervised and unsupervised learning tasks using a real-world dataset.
- Maximum team count is three

Project Outcomes

By the end of this **project** trainee will deliver:

A. Python code in a Notebook
B. Presentation for the project showing their results