**Walchand College of Engineering, Sangli**

## **Machine Learning Lab (6CS372)**

**TY BTech | AY 2023-2024 | Even Sem**

**Assignment 6**

**Decision tree and Random Forest**

**Task 1: Introduction to Random Forest**

* Explore and study Decision Tree classifier algorithm and Random Forest algorithm. Understand how they work and their advantages over other machine learning algorithms.
* Understand the key parameters of Random Forest such as number of trees, depth of trees, and splitting criteria.
* Understand key parameters of Decision Tree Classifier.
* Understand how Random Forest utilizes ensemble learning.

**Task 2: Dataset Selection**

* Select a dataset suitable for classification or regression tasks. You can use built-in datasets from scikit-learn or external datasets from sources like Kaggle or UCI Machine Learning Repository. Prefer built-in dataset. Later on you can simply change dataset and repeat experiment.
* Load the dataset into your Python environment.

**Task 3: Data Pre-processing**

* Perform necessary data pre-processing steps such as handling missing values, encoding categorical variables, and scaling numerical features.

**Task 4: Model Implementation**

* Split the dataset into training and testing sets.
* Implement Decision Tree classifier.
* Implement a Random Forest classifier (or regressor) using the scikit-learn library.
* Train the model on the training dataset.

**Task 5: Model Evaluation**

* Evaluate the performance of the trained model using appropriate evaluation metrics such as accuracy, precision, recall, F1-score (for classification), or mean squared error, R-squared (for regression).
* Compare the performance of Random Forest with decision tree algorithm if applicable.

**Task 6: Hyperparameter Tuning**

* Perform hyperparameter tuning to optimize the performance of the Random Forest model.
* Experiment with different values of key hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf.
* Evaluate the tuned model and compare its performance with the default model.

**Task 7: Visualization**

* Visualize feature importance to understand which features are most influential in the classification/regression process.

**Task 8: Reporting analysis**

Report your justifications, the findings of each task, including model performance metrics, insights gained from visualization, and any observations made during the experimentation process in a markdown cell at the end.