COMP47750

Machine Learning with Python

# Assignment 2

# **Feature Selection**

## Objective

The objective for this assignment is to compare the performance of two machine learning models that incorporate implicit feature selection, that is decision trees and linear (logistic) regression.

The decision tree implementation in scikit-learn includes a number of mechanisms to control the complexity of the tree. These mechanisms effectively control the number of features used in the model.

For Logistic Regression L1 regularisation can be used to eliminate less useful features from the model.

The dataset for analysis in this assignment is available in the file **Student\_Perf.csv** on Brightspace. The dataset is derived from the original UCI version here <[link](https://archive.ics.uci.edu/dataset/320/student+performance)>; the descriptions of the features there will give you some information on the data.

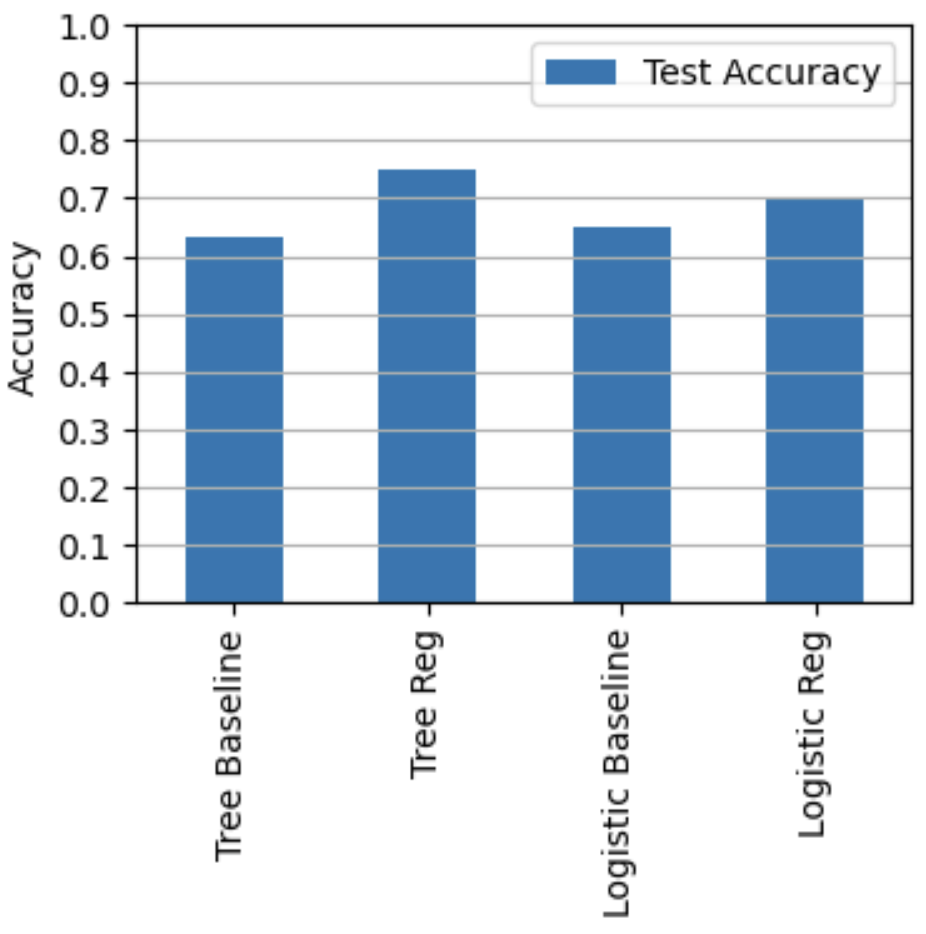
## Requirements

**Task 1: Accuracy**

* First assess the baseline generalisation accuracy of decision trees and logistic regression without any regularisation. Testing can be done in a repeated hold-out or cross-validation framework.
* For the decision tree implementation, select one of the parameters used to control model complexity and perform a grid search on that parameter to find a good value. Assess the accuracy of this strategy.
* Apply a similar model tuning strategy to logistic regression with L1 regularisation.
* Report these estimates of generalisation accuracy in a chart such as is shown in Figure 1.
* Comment on the results.

**Task 2: Feature Selection**

* Select sample regularised models (from Task 1) for decision trees and logistic regression and report what features are selected and not selected.
* Comment on these results. Are the selections stable (consistent across different runs)? Is there agreement between the logistic regression and decision tree models?



**Fig. 1:** Bar chart for results (these numbers are random).

**Submission:** This is an individual (not group) project. Submission is through the Brightspace page. Your submission should comprise two files, your notebook in ipymb and in html format. If your notebook takes a long time to run, submit without clearing the outputs. You should use markdown cells in the notebook to report your findings and conclusions.