MESIIN481222 Advanced Machine Learning
Lab Assignment 1 – Decision Tree and Random Forest in Healthcare

Graded assignment (Continuous assessment)

You are required to submit your solutions of this assignment on DeVinci Online, with respect to the deadline set by your instructor. Submit your own work. Cheating will not be tolerated and will be penalized.

Requirement: Lab activity 1 – Loan Use Case

Before you solve this lab assignment, make sure you finished the lab activity 1 which requires the implementation of decision tree and random forest models to solve a classification problem helping investors take the decision regarding investing or not in people.

Kyphosis Use Case

In this lab assignment, you will repeat what we have done in the previous lab activity when you we implemented a decision tree and a random forest models to predicts whether a borrower will pay or not its loan in order to help the investors to take decisions in investments.

In this lab activity, you will apply decision tree and random forest models on a medical classification use case.

You are going to train a model that can predict whether a child will have Kyphosis after its spinal surgery or not.

The Kyphosis dataset has 3 features: Age of the child (in months), Number (of vertebrae) and the Start (top vertebra) operated on.

You are going to train two models using the same dataset: decision tree classifier and random forest classifier.

Start by opening the file DT_Kyphosis.ipynb. This file contains a fully working example of the Kyphosis problem using sklearn. It contains all the code along with all the explanation needed. The first part of this lab is to familiarize yourself with the sklearn library functions.

Scikit learn actually has some built-in visualization capabilities for decision trees. The part of this code is given. Explore it and try to understand it. Add internal comments for your explanation.

Follow the instructions given in the python notebook to implement your machine learning techniques. The steps that you have to follow are the following:

- Import the needed python libraries
- Load your dataset

- Explore your data by using the info(), head(), and describe() methods
- Conduct an exploratory analysis with some data visualization
- Split your dataset into train and test sets.
- Create a decision tree model and train it on the train dataset.
- Explore the visualization of the decision tree and add your comments of explanation and interpretation.
- Make predictions with this model and evaluate its performance using the classification report and the confusion matrix.
- Create a random forest model and train it on the train dataset.
- Make predictions with this model and evaluate its performance using the classification report and the confusion matrix.

NB: Do not forget to add internal comments for interpretation and explanation in each step. This is mandatory and will be graded.