CS699 A2 – Fall 2023

Homework Assignment 6

Due: 10/26

The goal of this assignment is to give students an opportunity to study a scalable machine learning system for tree boosting called *XGBoost*. XGBoost is a short for e**X**treme **G**radient **Boost**ing package. XGBoost is an efficient and scalable implementation of gradient boosting framework by J. Freedman et al.

This assignment has two parts: Part 1 is writing a short research paper and Part 2 is building and evaluating classification models using XGBoost using R. This is an individual work. You should not use all or any part of another students homework.

Part 1 (10 points)

You are required to write a short research paper about the XGBoost. Your research paper should be about 10 pages long (excluding an appendix). It should include, at minimum, the followings:

- An overview of XGBoost, including a brief description of how XGBoost works
- Important characteristics/features of XGBoost
- Advantages of XGBoost
- Disadvantages of XGBoost
- List of parameters of XGBoost and the meaning/description of each parameter

At the end of Part 1, you must include all material you used to write your paper as a list of references.

If you want, you may use a LLM or Generative AI tool, such as ChatGPT. In this case, you must include the followings as an appendix of your paper:

- The name and a brief description of the tool you used
- Why you used the tool (e.g., to save time, to generate ideas, etc.)
- The **entire exchange** between you and the tool, including questions/phrases you submitted and the responses you received.

Part 2 (10 points)

You are required to build and test a classification model using the *drug_consumption_cannabis.csv* dataset. You must study how to use XBoost in R yourself. You must write a R code implementing the following requirements:

- Read the *drug_consumption_cannabis.csv* dataset into *df*.
- Split *df* into a training set *tr* and a test set *ts* with the ratio of 75% - 25%.
- Use the training set *tr* to build the best model. When choosing the best model, you must perform parameter tuning.

- Test the model on the test set *ts* and calculate and include the following performance measures in your submission:

	TPR	FPR	Precision	Recall	F-measure	MCC	Kappa
Class 0							
Class 1							
Weighted average							

Deliverables

- Part 1: Your research paper (with references and an appendix if needed).
- Part 2:
    - A list of parameters you used for tuning
    - A pdf document including the above performance measures
    - A R source code file
- Combine all files into a single archive file and name it FirstName_LastName_hw6.EXT, where EXT is a appropriate archive file extension, such as *zip* or *rar*.