University of Monastir
Higher School of Computer Science and Mathematics
Software Engineering – I1- 2023-2024

Machine Learning Algorithms

- The assignment is due : 15 April 2024.
- Write your name at the top right-hand of each page submitted (hard copy).
- Prepare a 12 minutes presentation of your machine learning project.
- To submit your code, please send it as an attachment via email to my address. Package your code as a ZIP file with the name "fistName_LastName". You have to submit a clear code with comments. I will run it on https://www.python.org/downloads/. Try your code before submitting.
- You Will wok on one of the following data sets. You must work in groups of 2 or 3 students.
    A. Stroke Prediction Dataset
    https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download
    B. Online Retail Customer Churn Dataset
    https://www.kaggle.com/datasets/hassaneskikri/online-retail-customer-churn-dataset
    C. Water Quality
    https://www.kaggle.com/datasets/adityakadiwal/water-potability
    D. Employee dataset
    https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset
    E. Banking Dataset Classification
    https://www.kaggle.com/datasets/rashmiranu/banking-dataset-classification
    F. Cardio Vascular Disease Detection
    https://www.kaggle.com/datasets/bhadaneeraj/cardio-vascular-disease-detection
    G. Predicting Credit Card Approvals
    https://www.kaggle.com/datasets/devzohaib/predicting-credit-card-approvals
    H. Water Quality
    https://www.kaggle.com/datasets/adityakadiwal/water-potability
    I. Water quality
    https://www.kaggle.com/datasets/mssmartypants/water-quality

1. Data Engineering

a. Download the data and report how you prepare the data for machine learning:
b. Load and represent the data using an appropriate data structure.
c. Give a description of the data set like size, features, target variables, predictive variables, feature types, etc.

d. Apply any preprocessing steps that might be required to clean or filter the data before analysis.
e. Analyze, characterize, and summarize the cleaned dataset, using tables and plots where appropriate. Clearly explain and interpret any analysis results which are produced.
f. Summarize any insights which you gained from your analysis of the data.
g. Suggest ideas for further analysis which could be performed on the data like data transformation and data reduction. Conduct the suggested analysis and clearly explain your results.

2. Model Engineering

a. Explain how you split the data into training and test sets
b. Run the decision tree algorithm on the training data without pruning (mention the parameter setting).
c. Give the graphical and textual representation of the learned decision tree.
d. Which features are most relevant for the classification task. Explain how the overall importance of a feature is a decision tree can be computed.
e. Show how data are separated for 3 to 5 leaf nodes.
f. What is your learned decision tree's accuracy over the training set?
g. what is your learned decision tree's accuracy over the test set?
h. Run the decision tree algorithm on the training data with pre-pruning (mention the parameter setting). Explain which thresholds you use and how you set them to obtain optimum results.
i. What are the sizes of your original tree and your pruned tree?
j. What are the accuracies of your unpruned and pruned trees over the training set? Over the test set?
k. Implement a post-pruning procedure for your learned decision tree (without pruning). You may choose reduced error pruning, rule post-pruning, or any other method you choose so long as you describe it precisely (Note: if you use rule post-pruning, make up a reasonable definition of "size"). Explain if you need a validation set for post-pruning. If yes, explain how you split your data into training/validation/test sets. (Note: some methods make post-pruning on training set).
l. What are the sizes of your original tree and your pruned tree?
m. What are the accuracies of your unpruned and pruned trees over the training set? Over the test set? Over the validation set.
n. Which of these trees (pruned or unpruned) would you recommend using to classify future data (justify your answer in terms of your actual observed accuracies).
o. Which of these trees (pre-pruned or post-pruned) would you recommend using to classify future data (justify your answer in terms of your actual observed accuracies).
p. Explain if it is interesting to combine pre-pruning and post-pruning. If yes, show how you combine them than evaluate you model.

Good Luck