Lab Assignment 1

Due date: 15 April 2022

Question: Read the COVID-19 mobility and new cases datasets. The goal of this assignment is to take new cases as target variable and predict them from mobility data. Take “India” as the case study for this assignment.

You are given two dataset files i.e., “**changes-visitors-covid.csv**” and “**covid-data.csv**”.

changes-visitors-covid.csv: This file contains statistics related to mobility in different countries during covid-19.

Entity : Name of countries
Code : Country’s code
Day : Date
retail_and_recreation : mobility for retail and recreation
grocery_and_pharmacy : mobility for grocery and pharmacy
residential : mobility in residential areas
transit_stations : mobility in transit stations
parks : mobility in parks
workplaces : mobility in workplaces

covid-data.csv: This file contains statistics related to covid-19 new cases in different countries.

iso_code : Country’s code
continent : Country’s continent
location : Name of countries
date : Date
new_cases : New cases on a particular day.

*You can ignore rest features.
MSE: mean squared error
RMSE: root mean squared error

I. Design a variation of the ID3 algorithm named as “ID3-A” that uses the approach used by CART for regression analysis. Instead of Gini Impurity, use entropy for ID3-A. Compare the accuracy and performance of CART and ID3-A. Apply this approach considering the entire dataset.
II. Perform K-Means clustering on data examples. Using subsets in different clusters, design a “K-means Forest”. Predict results by taking the average (or weighted average by comparing test examples with cluster centroid) of results predicted by trees in the forest.
    A. Use ID3-A to create trees in the forest.
    B. Use CART to create trees in the forest.

C. Compare the accuracy and performance of K-means Forests created using CART and ID3-A for parts A. and B.

Example: Let K=3, Then part A. 3 ID3-A trees: Forest 1. Compute result 1. Then part B. 3 CART trees: Forest 2. Compute Result 2. Compare Result 1 and Result 2.

Note:

Language: Python, Java, C++, MATLAB - - any of your choice
Use of any in-built libraries is discouraged.
Appropriate data pre-processing, data validation approaches and post-processing approaches such as pruning should be used at requisite places.
Make sure the model is not under-fit and over-fit.
Report your code, results and observations.
The assignment can be done in groups of less than or equal to 2. The group will remain for further assignments or till any further notice is given.