[Project Code : RFDT]

Short Term Rain Forecasting using Decision Tree based Learning Model

Project Duration : 22-Jan-2023 ~~ 11-Feb-2023
Submission Information : (via) CSE-Moodle

Objective:
Short-term rainfall forecasting plays an important role in hydro-logic modeling and water resource management problems such as flood warning and real time control of urban drainage systems. Climatic changes for a location are dependent on variable factors like temperature, precipitation, atmospheric pressure, humidity, wind speed and combination of other such factors.

Your task is to build a decision tree learning model to predict whether it will rain on the following day using the information from the present day. In particular, you shall be doing the following tasks:

Your Tasks:

1. *Decision Tree Model without Pruning:*
    a. Build a decision tree by taking maximum depth as input and by randomly splitting the train set as 80/20 split i.e., 80% for training and 20% for validation. Provide the validation set accuracy by averaging over 10 random 80/20 splits. Consider that particular tree which provides the best test-set accuracy as for further experiments.
    b. Implement the standard **ID3 Decision tree algorithm** as discussed in class, using information gain to choose which attribute to split at each point. Do NOT use scikit-learn for this part.
    c. Find out the best possible depth limit to be used for your dataset. Provide a depth V/S test accuracy plot explaining the same.
    d. Test out the implementation of Decision Tree Classifier from scikit-learn package, using information gain only for the best dataset split.

2. *Revised Decision Tree Model with Pruning:*
    a. To prune the tree obtained in part (1), you have to use Reduced Error Pruning. Do NOT use scikit-learn for this part.
    b. Test out the implementation of Decision Tree Classifier from scikit-learn package, using information gain and reduced error pruning.
    c. Finally, use the *entire* training set to learn a decision tree with and without pruning and use the test set to estimate their accuracy.

3. *Classification Report:*
    a. Create a classification report for both the trees in tabular form. (with and without pruning).
    b. You need to calculate accuracy, precision , recall , f1-score and support on the test-set of both the models for both the classes(Yes and No).

4. *Visualization:* Print the final decision tree obtained from question 1 & 2 following the hierarchical levels of data attributes as nodes of the tree. You may use sklearn.tree.plot_tree for this task.

5. *Report:* Write a short report (2-3 pages) detailing your work, results, observations, and justifications. Include an overview of methods, results, and interpretation.

Note: *The program can be written in C / C++ / Java / Python programming language from scratch. No machine learning /data science /statistics package / library should be used for model creation.*

Dataset:

Data Filename: **rain_predict_train.csv** contains training data
rain_predict_test.csv contains test data
(Please note that both sets may contain missing values. To handle these missing values, you should use appropriate techniques and clearly explain your methods in the report)

Data Description: This data was extracted Open Data Portal - https://data.gov.in/. It has approximately 120,000 data points and contains 12 continuous variables and 2 categorical variables. The target variable is binary, with the two classes being**:** Yes and No.

Here's detailed description of all the **attributes**:

Attribute ID	Attribute	Data type	Values type	Description
1	mint	float64	Continuous	Min Temp
2	maxt	float64	Continuous	Max Temp
3	rainfall	float64	Continuous	Rainfall Amount
4	windd3	object	Categorical	Wind Direction
5	winds9	int64	Continuous	Wind Speed at 9am
6	winds3	int64	Continuous	Wind Speed at 3pm
7	hum9	int64	Continuous	Humidity at 9am
8	hum3	int64	Continuous	Humidity at 3pm
9	pres9	float64	Continuous	Pressure at 9am
10	pres3	float64	Continuous	Pressure at 3pm
11	temp9	float64	Continuous	Temperature at 9am
12	temp3	float64	Continuous	Temperature at 3pm
13	rain	float64	Categorical	Rain Today
14	riskmm	float64	Continuous	RiskMM
15	raint	object	Categorical	Rain Tomorrow

Submission Details: (to be submitted in CSE-Moodle, **by one representative of the group**)

1. ZIPPED folder containing code (with comments) and the dataset files
2. Report (in pdf format)

Submission Guidelines:

1. You may use one of the following languages: C / C++ / Java / Python. No machine learning /data science /statistics package / library should be used for model creation.
2. Your Programs should run on a Linux Environment.
3. Your program should be standalone and should **not** use any special purpose library. **Numpy or Pandas may be used.** And, you can use libraries for other purposes, such as formatting and visualization of data.
4. You should submit the program file and README file and **not** the output/input file.
5. You should name your file as <GroupNo_ProjectCode.extension>.
   (e.g., *Group99_RFDT.zip* for code-distribution and *Group99_RFDT.pdf* for report)
6. The submitted program file *should* have the following header comments:
   # Group Number
   # Roll Numbers : Names of members (listed line wise)
   # Project Number

Project Title

7. Submit through CSE-MOODLE only.
   Link to our Course page: https://moodlecse.iitkgp.ac.in/moodle/course/view.php?id=508

You should not use any code available on the Web. Submissions found to be plagiarized or having used ML libraries (except for parts where specifically allowed) will be awarded zero marks.

For any questions about the assignment, contact the following TA:
Abhinav Bohra (Email: abhinavbohra09@gmail.com)