Pattern Recognition & Machine Learning

Lab-2 :- *"Decision Trees"*

Objectives

According to the Document provided, our goals were as follows :-

1. To construct Decision Tree Classifiers using Entropy and Gini as criterions. In support of this, use link as a Dataset.

2. To construct a Decision Tree Regressor, using dataset as a file for Input Dataset.

Procedures

For accomplishing these objectives, certain components would be required as follows :-

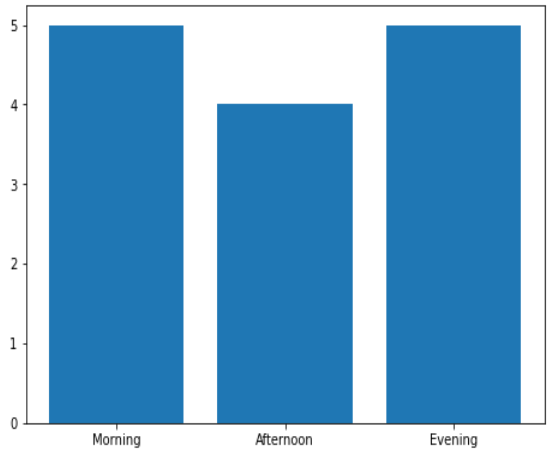
1. Preprocessing the Dataset
2. Shuffling and Splitting into Train and Test portions
3. Cross-Validating and Predicting through Models
4. Generating certain structural and parametric details about Models
5. Plotting these Trees and their corresponding Decision Surfaces.

Inputting the Datasets

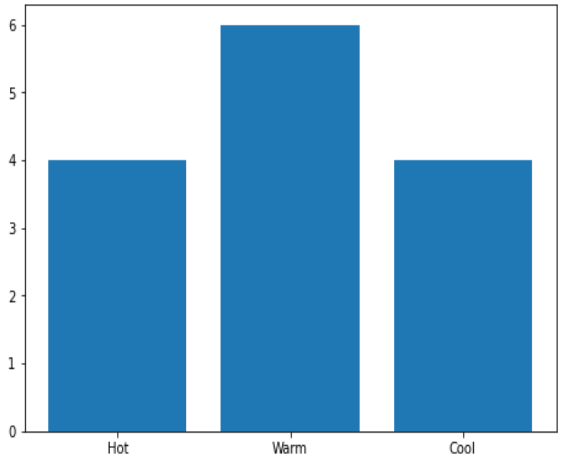
Through the given links, Datasets were obtained and their distribution details can be brought out as follows :-

Dataset from Question-I

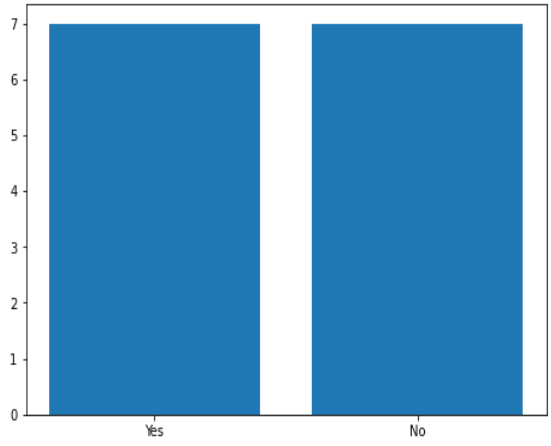
Time-wise Distribution



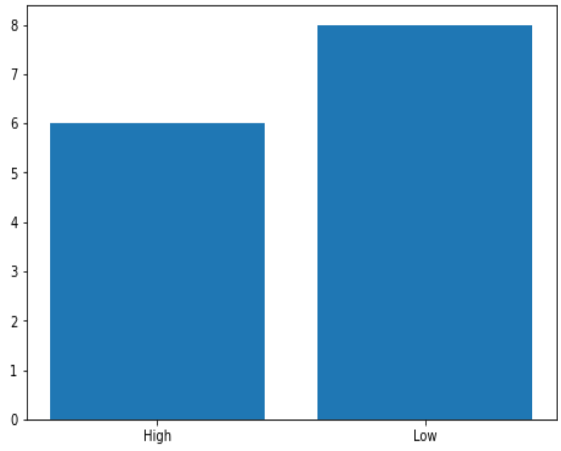
Temperature-wise Distribution



Friend-Wise Distribution



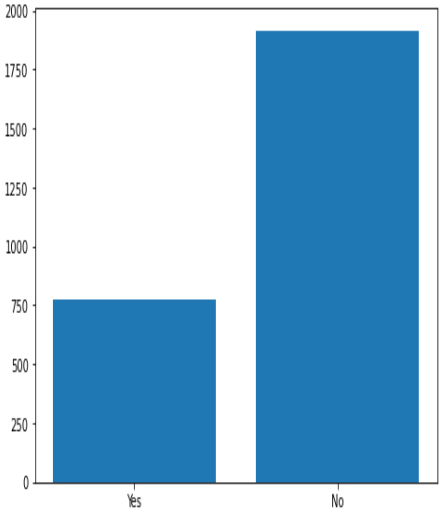
Wind-Wise Distribution



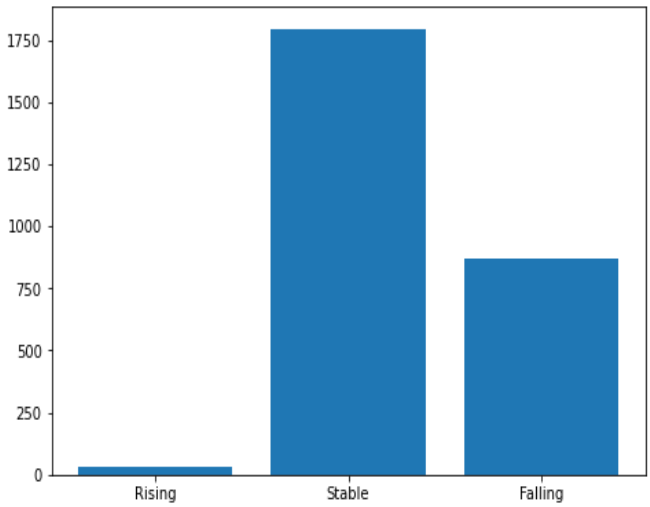
Seeing these Feature-Wise Distributions, there might be chances that Decision Tree Classifiers can tend towards some biasness, irrespective of whatsoever Criterion would be picked.

Dataset from Question-II

"45.5 Objective" based Distribution



Recent Trend-Wise Distribution



Again, seeing these Distributions, Decision Tree Regressor can also tend towards some biasness.

Libraries / Dependencies

1. Sklearn
2. Numpy
3. Pandas
4. Graphviz
5. Matplotlib

Extracting Features through pre-processing

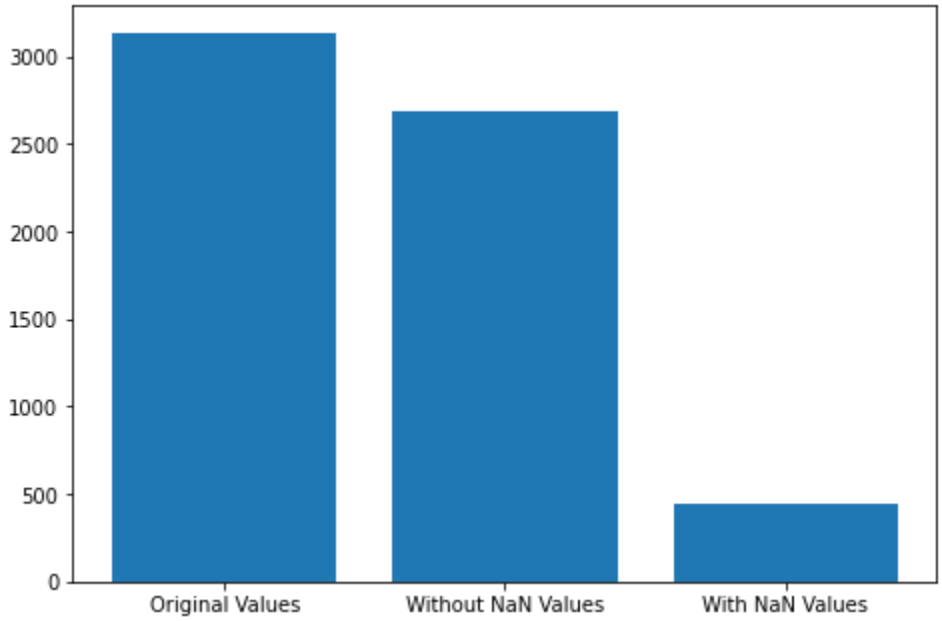
Both Datasets had some flaws and missing values. So, in order to resolve them according to the working of models, we incorporated some pre-processing measures as follows :-

1. Encoding Categorical Features :-

   For Categorical Labels, Ordinal Encoder was used, which assigns the float value of a particular label, according to its relative alphabetic order.
   Ex:- Afternoon as 0, Evening as 1 and Morning as 2.

2. Eliminating NaN Values :-

   In Dataset-II , * and ** were invalid values, in respect to features. So, they were converted to NaN value and then removed. As a proof, we can see the Dataset Distribution as follows :-



3. Converting Strings to Float Values :-

   In Dataset-II, Strings were used for big numeric figures. So, those values were converted to float64 format.

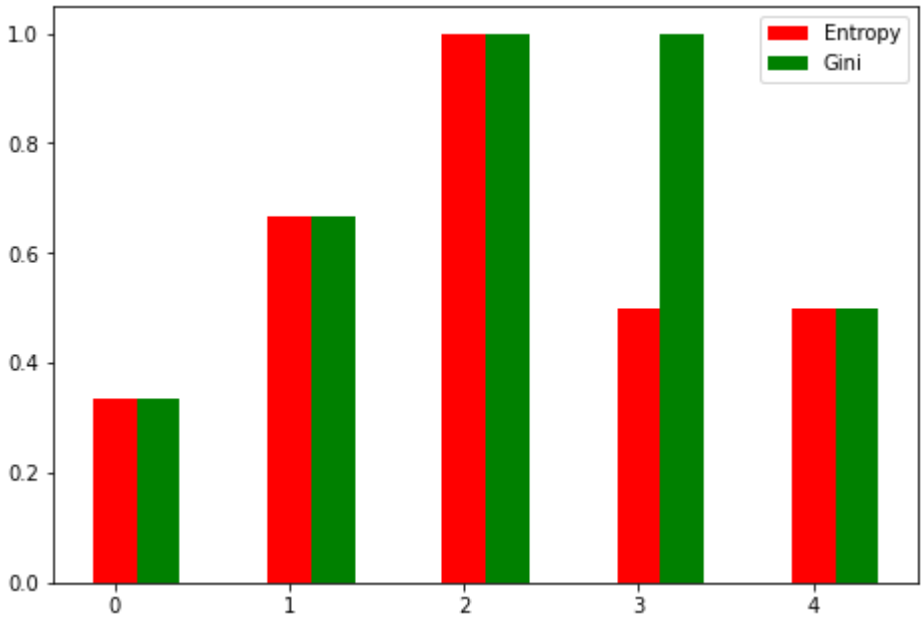
Splitting into Train and Test Sub-portions

For Decision Tree Classifiers as well as Decision Tree Regressor, the Ratio for Training and Test Dataset Portions, was kept to 9:1 . Although for smaller Datasets, it can lead to Overfitting, but the ratio behaves fine for Large Datasets.

Cross-Validation Details

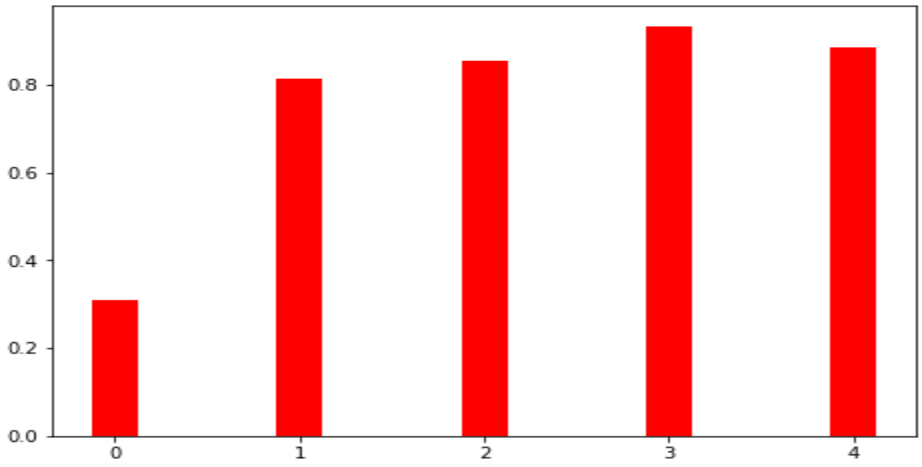
Cross Validation Scores were calculated for Decision Tree Classifiers and Decision Tree Regressor. The details with number of folds=5, were as follows :-

Decision Tree Classifiers



Criterion	Array of scores	Mean Score (Approx)	Standard Deviation
Entropy	0.33333333 0.66666667 1 0.5 0.5	0.6	0.2260
Gini	0.33333333 0.66666667 1 1 0.5	0.7	0.2666

Decision Tree Regressor



Criterion	Array of Scores	Mean Score (Approx)	Standard Deviation
Mean Squared Error	0.31068099 0.81292212 0.85281763 0.93396697 0.88365991	0.7588	0.2275

Details about Models

Some Characteristics about Decision Tree Classifiers and Decision Tree Regressor were given, as follows :-

Decision Tree Classifiers

Criterion	Depth of Tree	No. of Leaf Nodes	Gini Impurities
Entropy	4	7	Feature: 0, Score: 0.30355 Feature: 1, Score: 0.18150 Feature: 2, Score: 0.32757 Feature: 3, Score: 0.18739
Gini	4	7	Feature: 0, Score: 0.38393 Feature: 1, Score: 0.20000 Feature: 2, Score: 0.18750 Feature: 3, Score: 0.22857

Decision Tree Regressor

Criterion	Depth of Tree	No. of Leaf Nodes	Gini Impurities
Mean Squared Error	22	1518	Feature: 0, Score: 0.00475 Feature: 1, Score: 0.00003 Feature: 2, Score: 0.00189 Feature: 3, Score: 0.00052 Feature: 4, Score: 0.00318 Feature: 5, Score: 0.08402 Feature: 6, Score: 0.13662 Feature: 7, Score: 0.71412 Feature: 8, Score: 0.05487

Accuracy metrics of models

After constructing Decision Tree Classifiers and Decision Tree Regressor, their accuracies were brought out, upon their corresponding Test Dataset Portions. Those details were as follows :-

Criterion	Test Accuracy
Classifier :- Entropy	50%
Classifier :- Gini	50%
Regressor :- Mean Squared Error	mse = 0.9691