DMG ASSIGNMENT : 03

Pradeep Kumar (MT20036)
Akanksha Pandey (MT20048)

1. Methodology:

Approach and Reasons:

Association rule mining means finding interesting relationships between dataitems.

The training data had 2 types of data items, nominal (Aspect, Elevation, etc) and numeric (Horizontal_Distance_To_Hydrology, Vertical_Distance_To_Hydrology and Label). We removed all the numeric data items except Label and converted Label to nominal data.

Rules:

1. Attributes : 8

   Elevation
   Aspect
   Slope
   Soil_Type
   Hillshade_9am
   Hillshade_noon
   Horizontal_Distance_To_Fire_Points
   Label

   Minimum Support : 0.01

   Minimum Metric (Confidence) : 0.6

Number of Rules : 528

F1 Score : 0.47541

2. Attributes : 8

    Elevation
    Aspect
    Slope
    Soil_Type
    Hillshade_9am
    Hillshade_noon
    Horizontal_Distance_To_Fire_Points
    Label

    Minimum Support : 0.01

    Minimum Metric (Confidence) : 0.7

    Number of Rules : 5000

    F1 Score : 0.44139

3. Attributes : 8

    Elevation
    Aspect
    Slope
    Soil_Type
    Hillshade_9am
    Hillshade_noon
    Horizontal_Distance_To_Fire_Points
    Label

    Minimum Support : 0.01

Minimum Metric (Confidence) : 0.7

Number of Rules : 7000

F1 Score : 0.47138

After finding the rules we removed non-decimal numbers and symbols from the rule file generated. We then removed numeric data items from our test data also.

Pick a row from test data and check for the longest rule that is matching to it and apply the label of that rule to it. If no rule is matching then we assumed it belongs to class one.

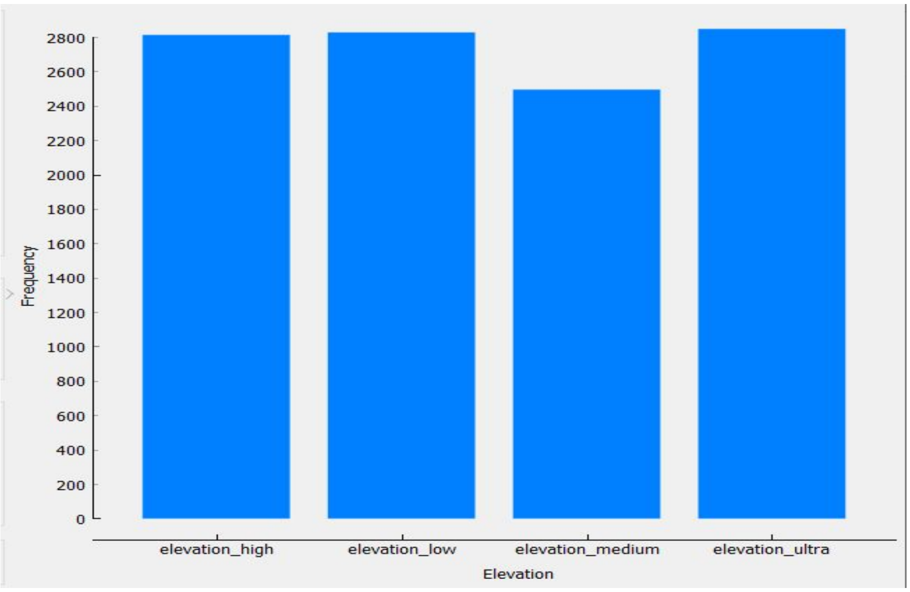
Reason: Rules have a specific pattern and if the data have the same components as rules then there is a high chance they will belong to the same class.

2. Visualization of Skewness:

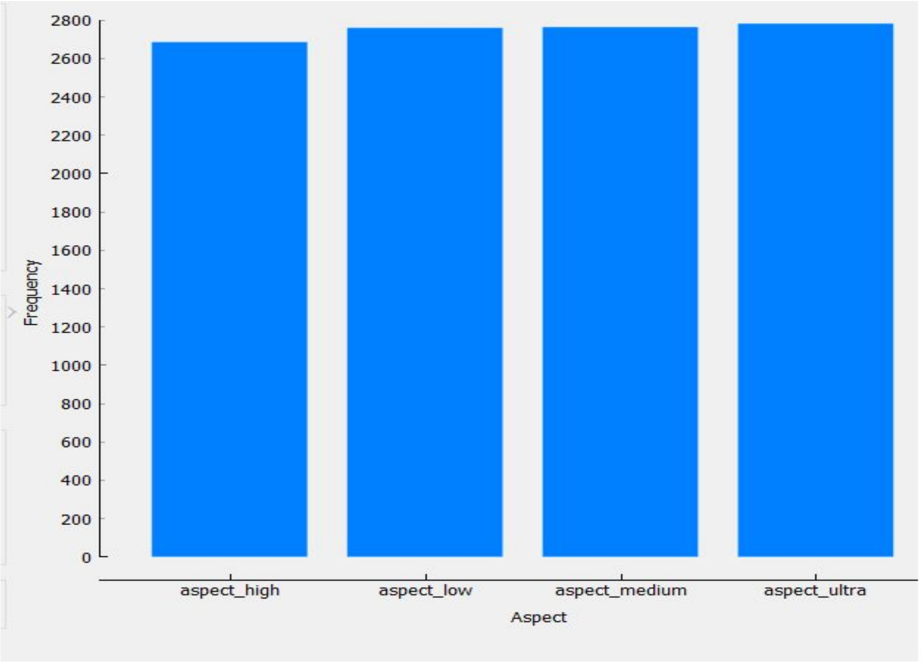
We have not done any preprocessing except deleting id, Horizontal_Distance_To_Hydrology and Vertical_Distance_To_Hydrology and converting Label to nominal.

Following are the visualization of given data :

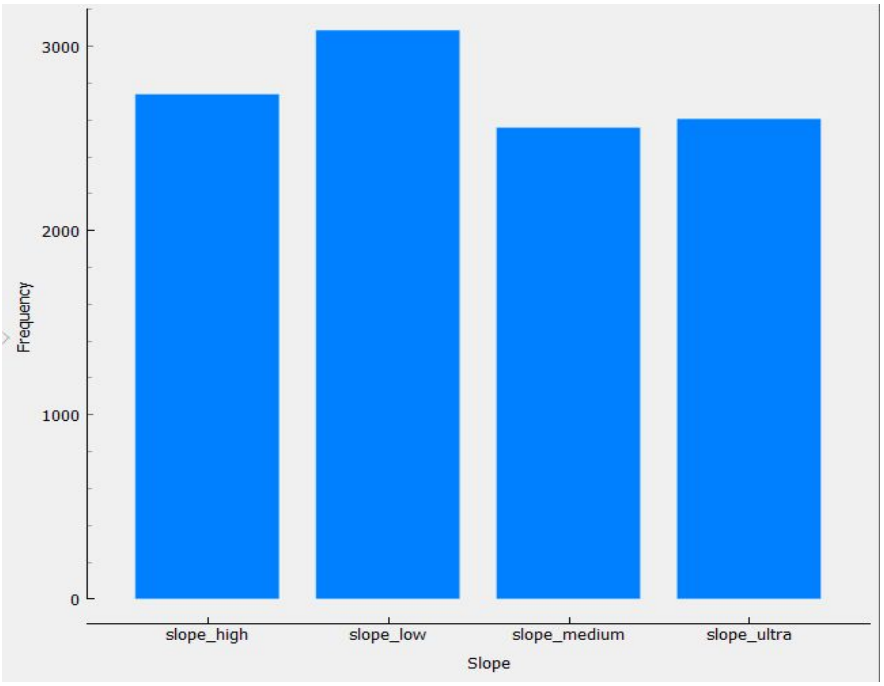
Elevation:



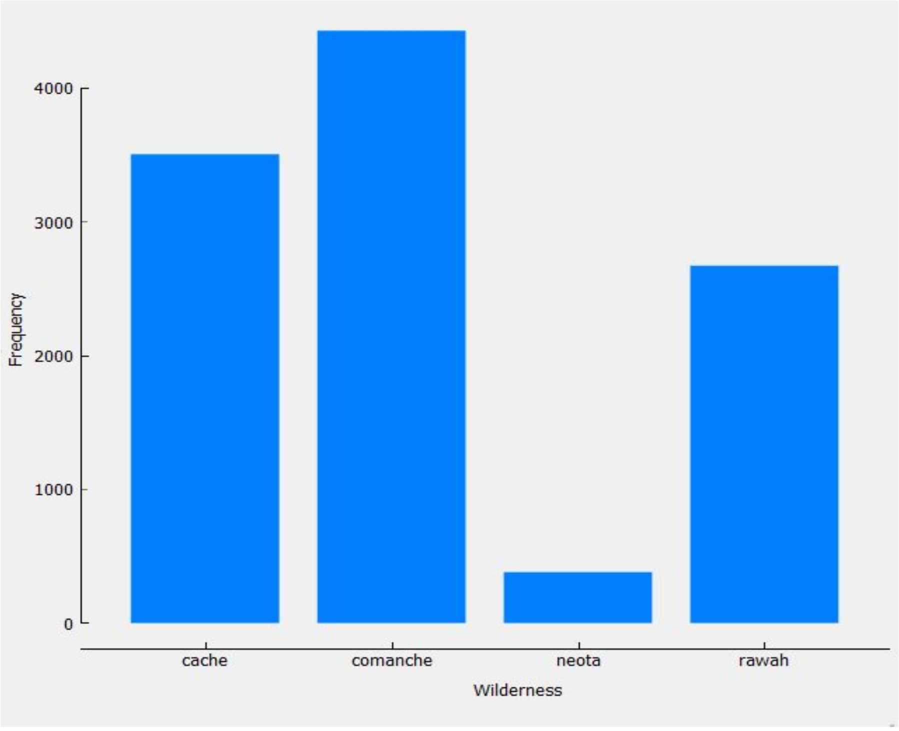
Aspect:



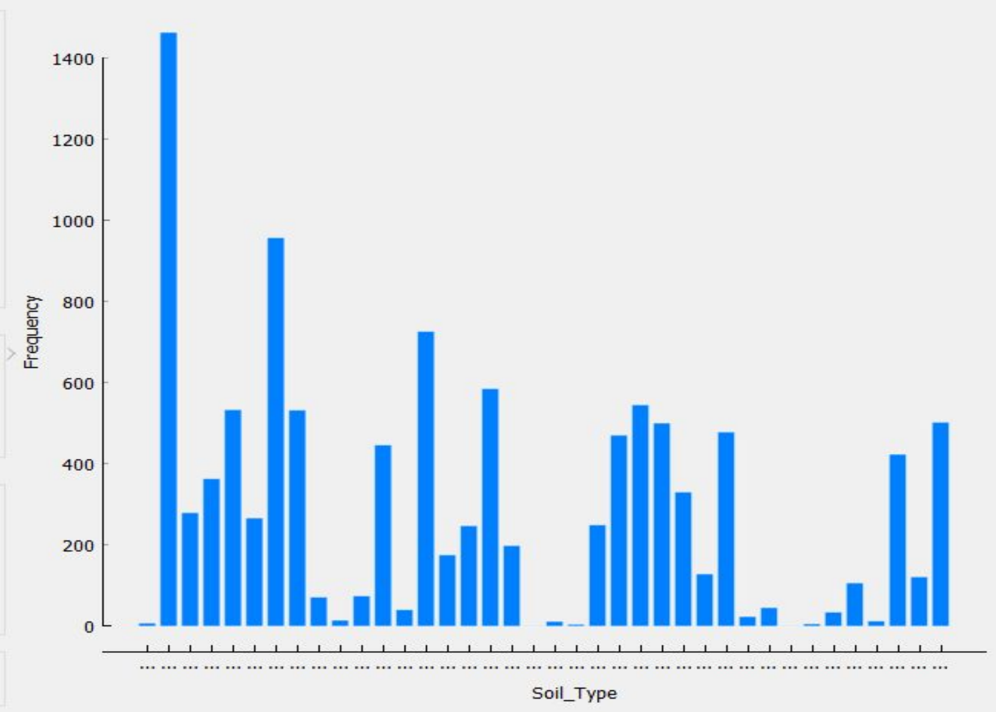
Slope:



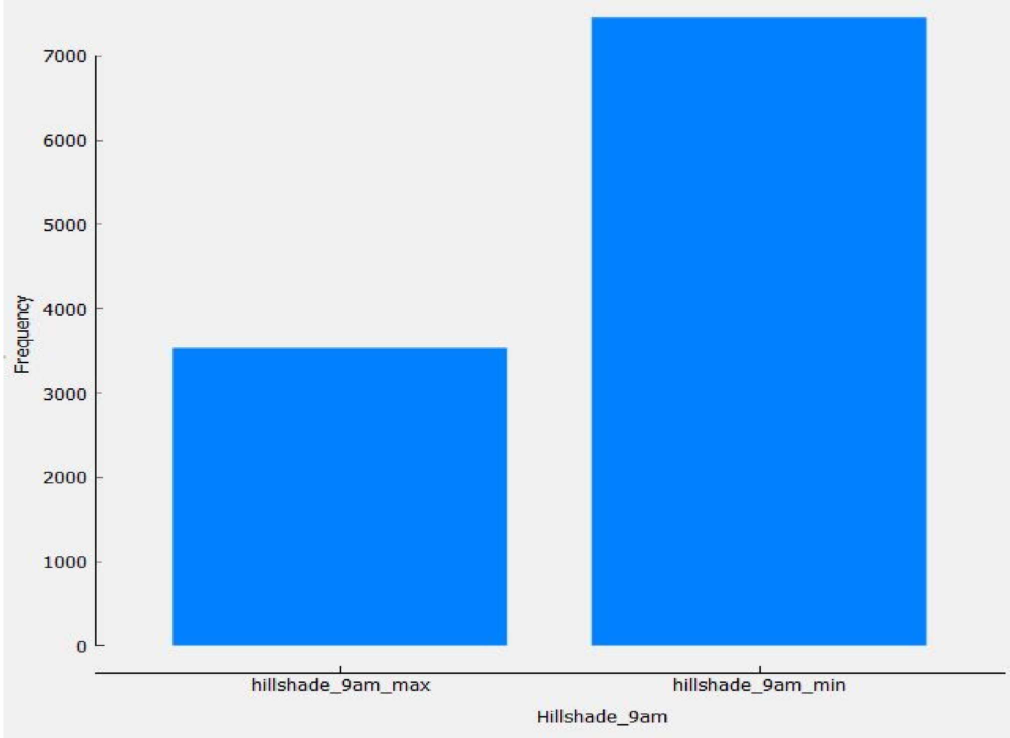
Wilderness:



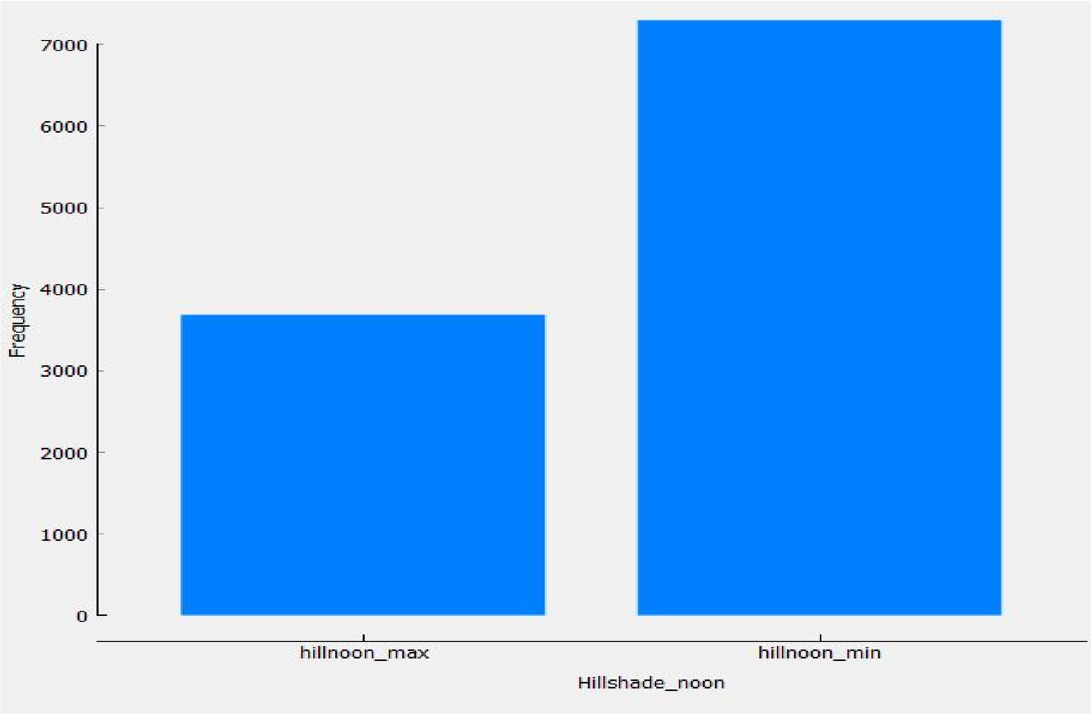
Soil_Type:



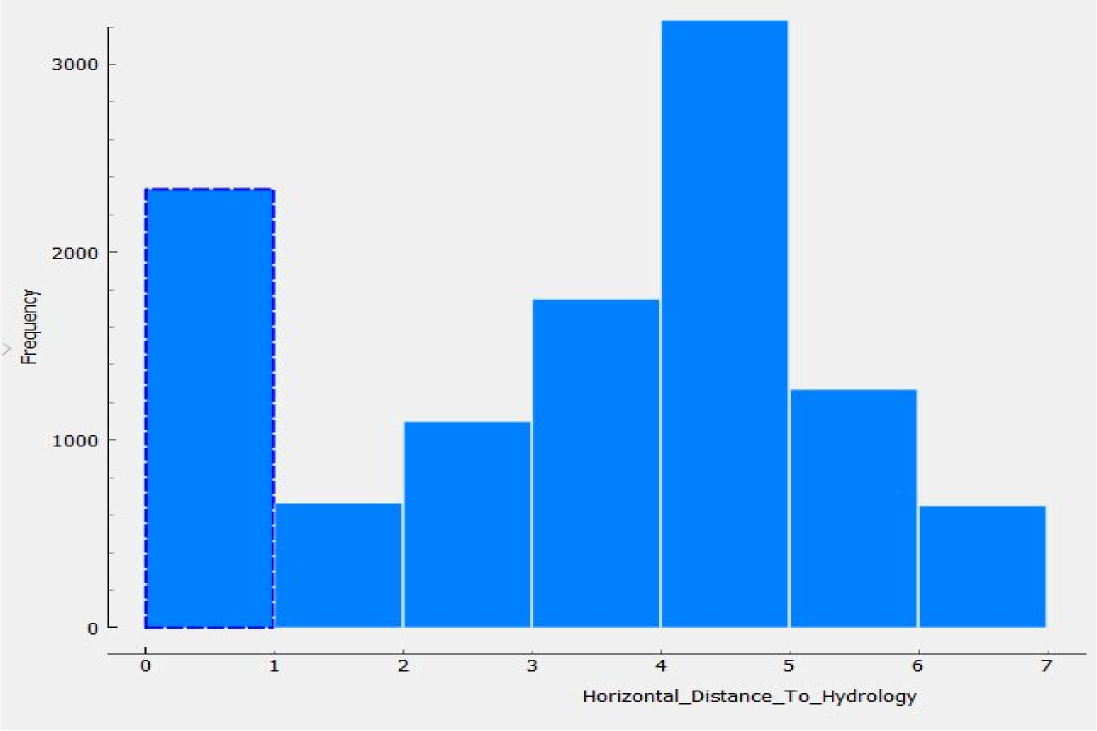
Hillshade_9am:



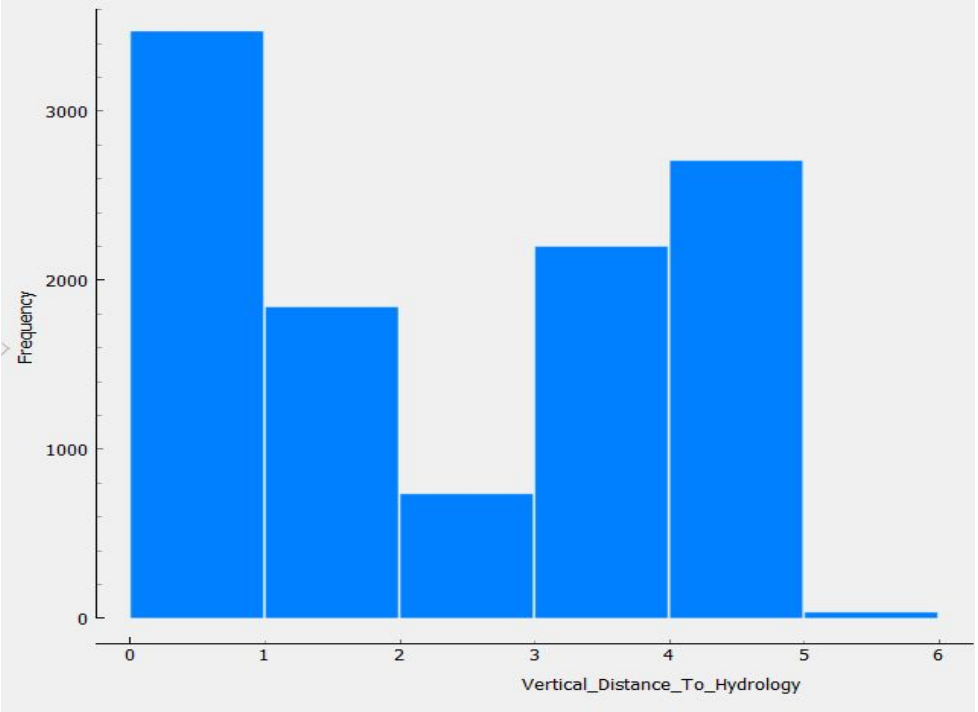
Hillshade_noon:



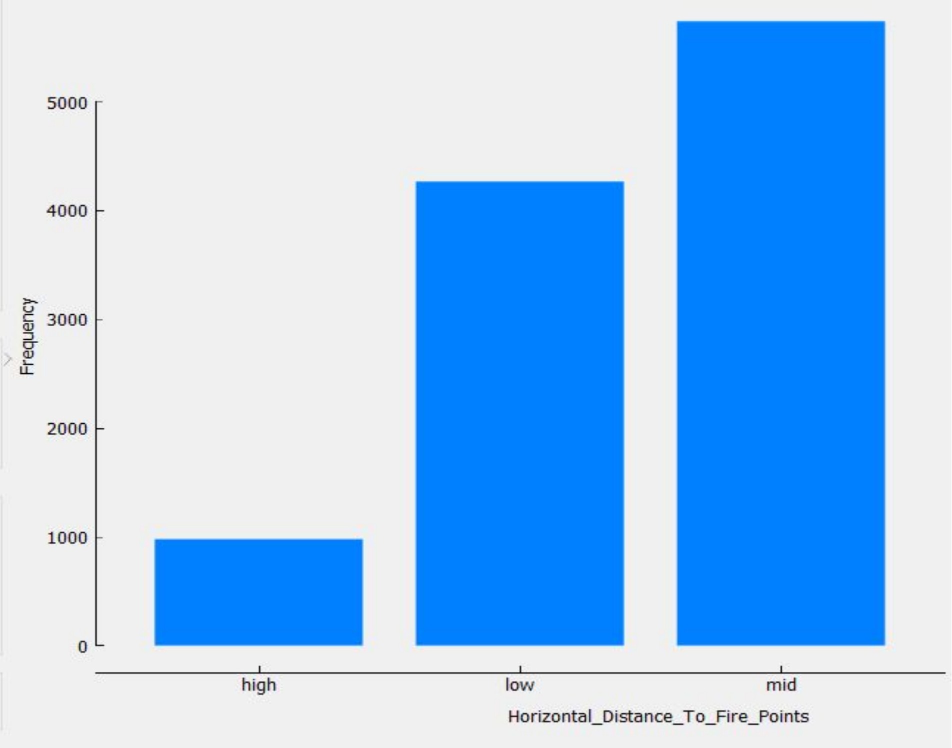
Horizontal_Distance_To_Hydrology:



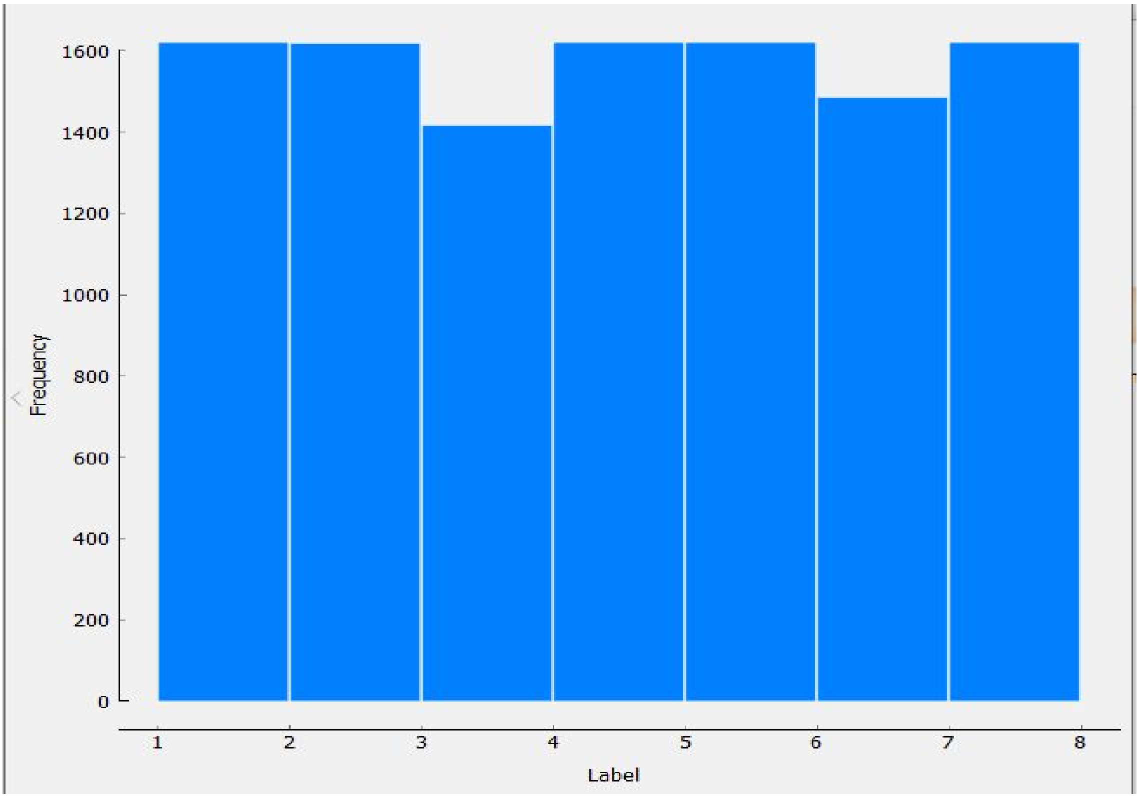
Vertical_Distance_To_Hydrology:



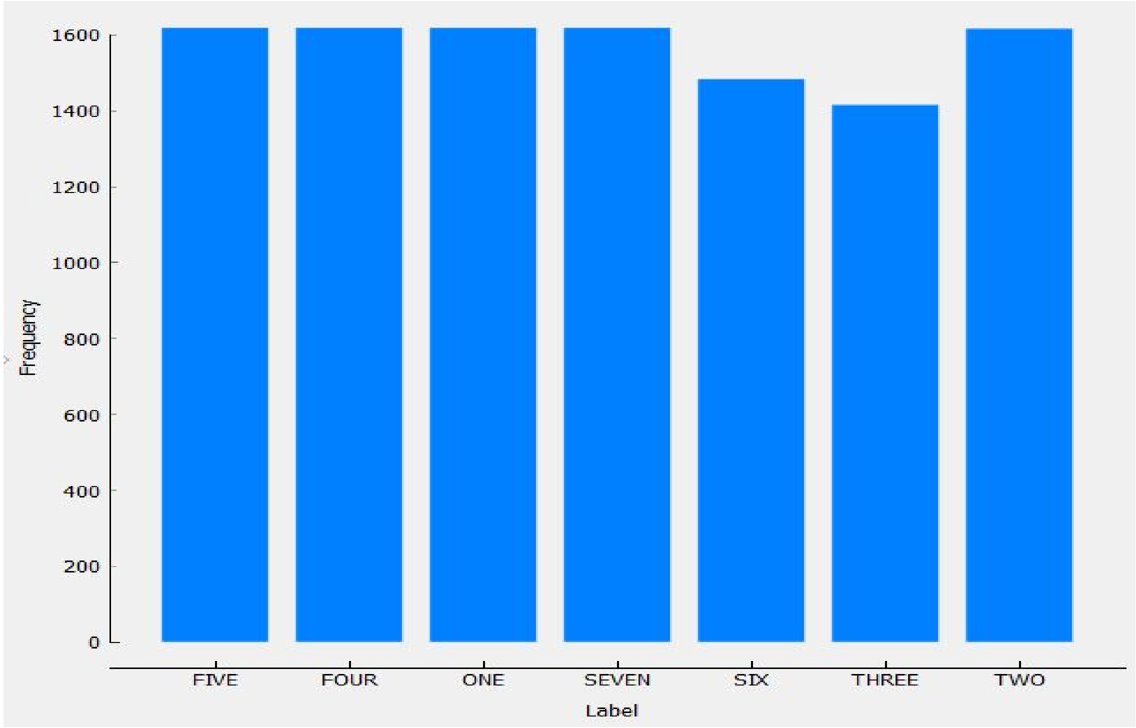
Horizontal_Distance_To_Fire_Points:



Label (Before):



Label (After):



3. Data Analysis Steps:

1. See the type of data given for the training.
2. Drop id
3. Drop numeric data Horizontal_Distance_To_Hydrology and Vertical_Distance_To_Hydrology
4. Convert label to nominal data
5. Set minimum support and minimum confidence value.
6. Make sure all the classes are present in the generated rules.

4. Comparison of F1 Score:

For our best 3 models F1 scores are 0.47541, 0.47138, 0.44139 and the F1 score of the Random forest-based baseline given on the kaggle Leaderboard is 0.47451 which is less than our first model but greater than other two models.

5. Learning:

1. Use of kaggle
2. Use of WEKA
3. Use of Orange
4. Use of SPMF
5. How to do classification using Association Rule mining from scratch.

6. Contribution:

Pradeep Kumar: Processing, Code, Report
Akanksha Pandey: Code, Report