Big Data Assignment Report

Kausik N

COE17B010

Steps to RUN:

Pip Install Modules –

pickle, numpy, pandas, matplotlib seaborn, plotly, tqdm, functools, mlxtend, sklearn

Run - ***main.py***

Pre-processing

File: Preprocessing.py

Methods:

1. Reading Dataset – ReadCSVFile(path)
2. Frequency Distribution – FreqDist(Data)
3. Missing Count – MissingCount(Data)
4. Cleaning – MissingClean(Data)

Remove Technique

If missing data in field (Symbol or Scientific Name) – Remove Row as this is unique data

Ignore Technique

If missing data in Synonym field – no need to do anything as it is a optional field for data

Subset Technique

If missing data in Common Name or Family – replace smartly

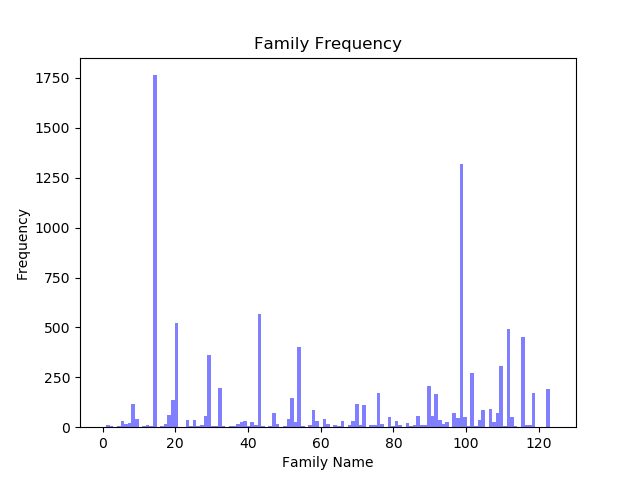
Family of subspecies will be same as subspecies

So, if data missing, using Scientific Name, identify its species and use its Family to fill in missing family name (Subset Technique)

Redundant Technique – RedundantClean(Data)

If data rows are repeated, remove those redundant rows and keep only unique rows

1. Visualising – Histogram(Data) for Family Frequency



Part A

File: Algorithms.py

Algorithms:

Encoding – TransactionalEncoding(Data)

FIM

1. Apriori
2. FPGrowth

CFI

1. Charm
2. Apriori-Close

MFI

1. Pincer Search
2. Mafia

LFI

1. Apriori Based LFI
2. FPGrowth Based LFI

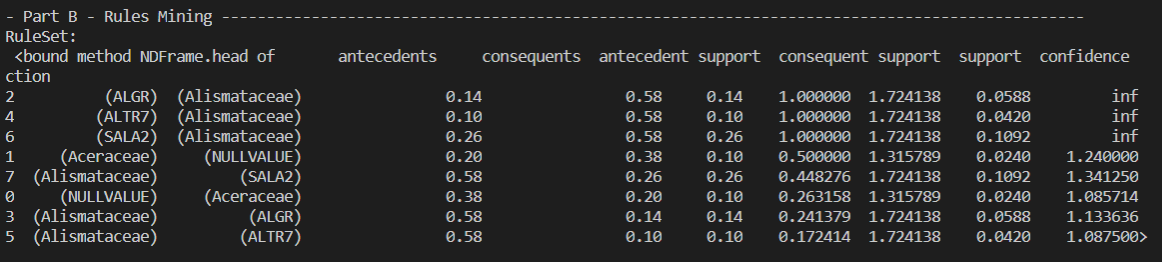
Part B

File: Algorithms.py

Function: RuleMining(FrequentItemsets)

Done for FPGrowth generated frequent itemsets

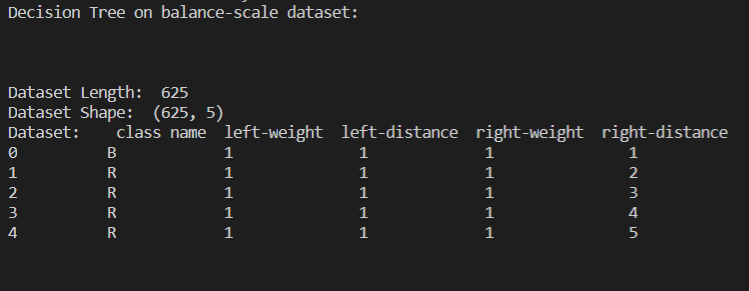
Confidence = 1.0

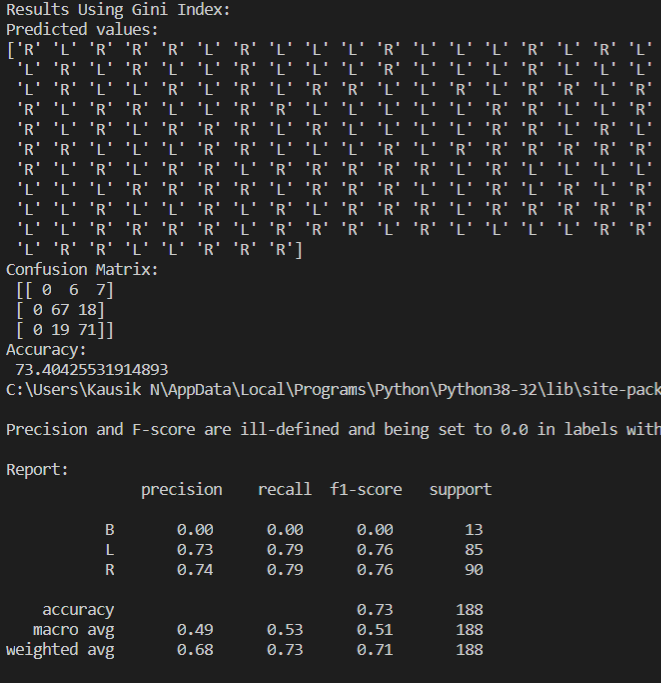


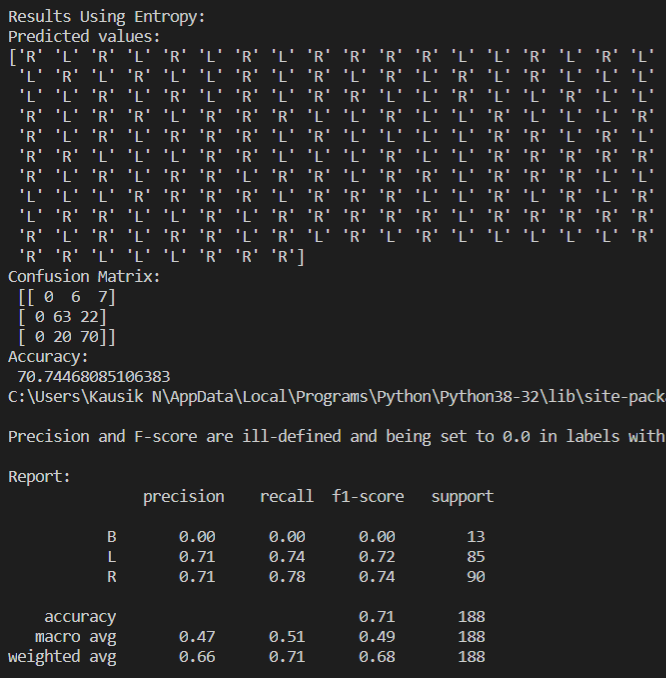
Part C

File: DecisionTree.py, BayesClassifier.py

Decision Tree done for balance-scale.csv dataset







Bayes Classifier done for Iris Dataset

