Big Data Analysis
Assignment 2
Workflow and Discussion

Earass Ahmad 19I-1254
Asfand Yar Ali 19I-1212

19 Apr 2021

1 Introduction

This assignment is about using the Pokec social network data to generate meaningful insights from it and to implement multiple frequent pattern mining algorithms. We create a thorough data pipeline to implement these tasks. The core components of the pipeline are discussed in the later sections.

2 Data Pipeline

Following are the core modules in the data pipeline:

2.1 Dataset

The dataset used in this work is a cleaned dataset obtained as the output of the first assignment. This dataset has approximately 1.1 million records. Some of the columns are excluded from FPM algorithms which are discussed in the upcoming sections.

2.2 Pre-Processing

Some pre-processing is done on the dataset to make it more meaningful. We add some relevant text to some columns to make the values self explanatory. For example, added "Age: " before the age values so we know that the value being shown, independent of the column name, is Age. This also helps the algorithms to distinguish between values of Age and, let's say, BMI. Apart from this change, we have selected only the following columns for the FPM implementation:

1. AgeGroup
2. BMIGroup
3. Gender
4. MaritalStatus
5. SmokingStatus
6. DrinkingStatus
7. ProfileStatus
8. BodyType
9. EyeSight
10. EyeColor

Figure 1 shows how the values look for each column in the processed dataset.

	AgeGroup	BMIGroup	GenderMF	MaritalStatus	SmokingStatus	DrinkingStatus	ProfileStatus	BodyType	EyeSight	EyeColor
0	Age: 25 - 34	BMI: 26 - 30	Male	Marital Status: Other	Smoking Status: Other	Drinking Status: Other	Public	Body Type: Other	Eye Sight: Other	Eye Color: Other
1	Age: 15 - 24	BMI: 18 - 25	Female	Marital Status: Single	Smoking Status: Non Smoker	Drinking Status: Occasional Drinker	Public	Body Type: Average	Eye Sight: Average	Eye Color: Green
2	Age: 15 - 24	BMI: 18 - 25	Male	Marital Status: Single	Smoking Status: Other	Drinking Status: Other	Public	Body Type: Other	Eye Sight: Other	Eye Color: Brown
3	Age: 25 - 34	BMI: 18 - 25	Male	Marital Status: Single	Smoking Status: Other	Drinking Status: Other	Not Public	Body Type: Other	Eye Sight: Other	Eye Color: Green
5	Age: 25 - 34	BMI: 18 - 25	Female	Marital Status: Other	Smoking Status: Non Smoker	Drinking Status: Occasional Drinker	Public	Body Type: Average	Eye Sight: Other	Eye Color: Green
6	Age: 25 - 34	BMI: 18 - 25	Male	Marital Status: Single	Smoking Status: Non Smoker	Drinking Status: Other	Public	Body Type: Sported	Eye Sight: Other	Eye Color: Brown
7	Age: 15 - 24	BMI: 18 - 25	Male	Marital Status: Other	Smoking Status: Non Smoker	Drinking Status: Occasional Drinker	Public	Body Type: Other	Eye Sight: Other	Eye Color: Blue
8	Age: 15 - 24	BMI: Missing	Male	Marital Status: Single	Smoking Status: Other	Drinking Status: Other	Public	Body Type: Other	Eye Sight: Good	Eye Color: Other

Figure 1: Preview of the Processed dataset ready to be passed to the FPM algorithms

2.3 Frequent Pattern Mining

Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. We have implemented 3 different frequent pattern mining algorithms and compared their performance and efficiency. The algorithms are applied on multiple iterations changing the minimum support support threshold and number of records. The minimum support thresholds applied are: 0.004, 0.006, 0.008. Following number of records are changed on each iteration: 50000, 100000, 200000, 500000, 700000.

These three algorithms are discussed in detail in the other algorithm's review report, however, we would provide an overview of how they are applied on our dataset.

2.3.1 Apriori Association Rules

The Apriori algorithm is applied on the dataset and it generates some meaning full results. We have used the apyori package available on PyPi for this.

2.3.2 PrefixSpan

The other algorithm we have used is PrefixSpan, which is from a sequential pattern algorithm. This is from the Extended Data Types section. This algorithm is applied on the same sets of minimum support and number of records. This algorithm is obtained from the Prefixspan package available on PyPi.

2.3.3 Colossal Pattern Mining

The third technique we have applied is the Colossal Patterns Mining which is from the Extended Patterns section. This algorithm is applied on the same sets of minimum support and number of records. This algorithm is obtained from the following Github repo: https://github.com/GENU05/mining-colossal-patterns-in-high-dimensional-databases

2.3.4 Algorithms Comparison

The performance of these three algorithms is then compared based on the minimum support threshold and number of records.

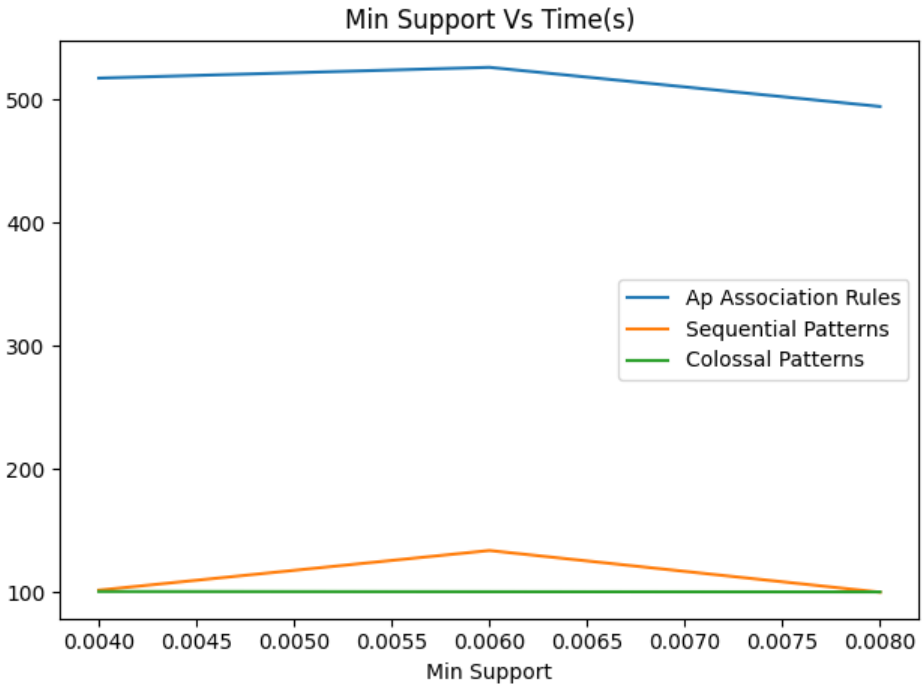

Figure 2: Min Support Vs Time(s)

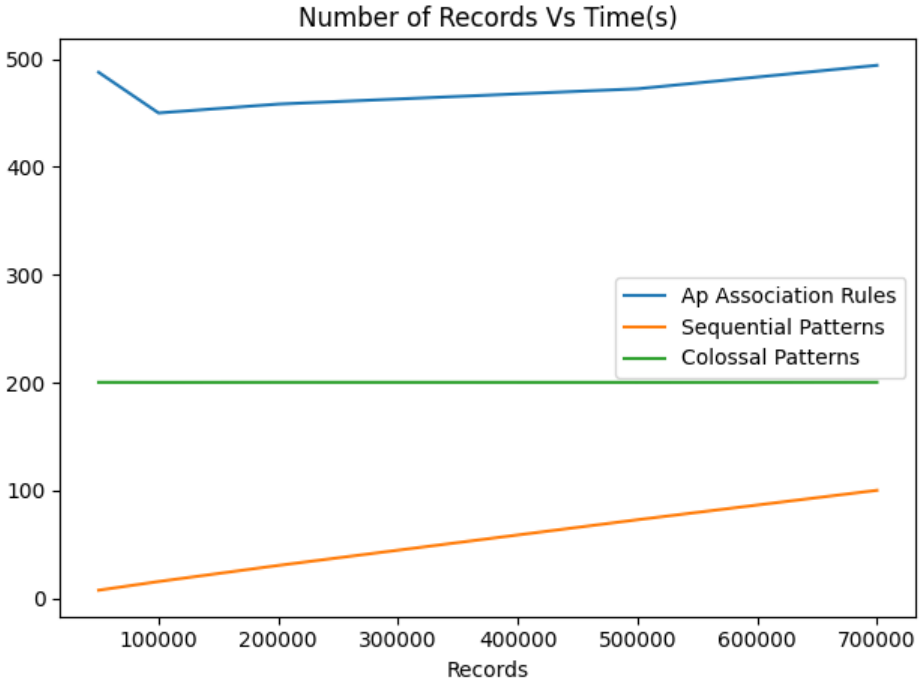

Figure 3: Number of Records Vs Time(s)

It is evident from the above plots that Apriori algorithm takes most of the time as compared to the other algorithms.

3 GUI

A web app GUI is created in Python's Flask framework to demonstrate the visualizations from assignments 1 and 2. Also, we can download the reports of the assignments from the front-end.

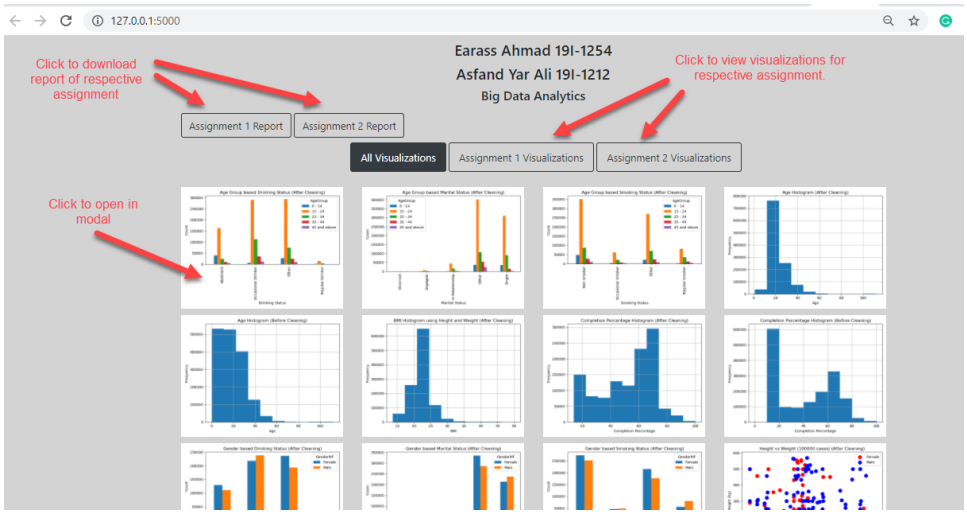

Figure 4: Home page layout

Figure 5: Enlarged view of visualization

4 Tools

In this work, we have primarily used Python programming language. Following packages of Python are used:

1. Pandas: Used for data reading, wrangling and manipulation
2. Numpy: Used for performing data operations of large arrays
3. Matplotlib: Used for generating visualizations
4. Logging: To log the process flow
5. Flask: Used as the web framework for creating the web app

Apart from the above, HTML, CSS, and Javascript are used in the GUI front-end.