ADVANCED DATABASE SYSTEMS
PROJECT 3

a. Team Members

Rohan Kulkarni (rk2845)
Aishwarya Rajesh (ar3567)

b. Project Files

The uploaded project directory contains the following files:

apriori_algorithm.py : Contains the Python source code.
README.pdf: Includes the required README information.
INTEGRATED-DATASET.csv: Contains the integrated dataset used from NYC Open Data.
example-run.txt: Contains the results of an interesting sample run.
run.sh: Script to run the Python code.

c. Project Description

(a) **Dataset used**: The NYPD *Stop, Question and Frisk Data* from year 2014. https://data.cityofnewyork.us/Public-Safety/The-Stop-Question-and-Frisk-Data/ftxv-d5ix

(b) **Dataset pre-processing**:

Attribute selection :
The original dataset is trimmed of its attributes to obtain only those records of the instances when people who were stopped and frisked were also searched for possession of knives (represented by the *knifcuti* attribute) and contraband (represented by the *contrabn* attribute). The pre-processing of this dataset includes the careful selection of 7 interesting attributes from the dataset.The attributes selected are :

Borough (*city* attribute) : Borough where the frisking took place.
(Bronx,Brooklyn,Manhattan,Queens,Staten Island)
Build (*build* attribute) : The build of the person who was frisked. *(H,M,T,U,Z)*
Age (*age* attribute) : The age of the person who was frisked. *(<15-30>,<30-50>,<50 above>)*
Race (*race* attribute) : The race of the person who was frisked. *(A,B,I,P,Q,U,W,Z)*
Sex (*sex* attribute) : The sex of the person who was frisked. *(Male, Female)*
Knife found (*knifcuti* attribute) : Was the person carrying a knife. *(1,0)*

Contraband found (*contrabn* attribute) : Was the person found with contraband *(1,0)*

Attribute Transformation :

Age : All the age values were grouped into buckets of *<15-30>,<30-50>,<50 above>* years.

Sex : All 'M' and 'F' values were converted to 'Male' and 'Female' respectively.

Knife found : Tuples with missing values were eliminated and 'Y' and 'N' were converted to '1' and '0' respectively. In case of a '1' value in a record from the dataset, **'Knife'** was added as an item, to the corresponding transaction in the transaction database.

Contraband found : Tuples with missing values were eliminated and 'Y' and 'N' were converted to '1' and '0' respectively. In case of a '1' value in a record from the dataset, **'Contraband'** was added as an item, to the corresponding transaction in the transaction database.

(c) **Interestingness of the chosen dataset**:

There has been varied criticism on the *Stop, Question and Frisk* procedure followed by the NYPD in an attempt to reduce street crime, wherein many innocents have been claimed to be detained by the department.
(http://www.nydailynews.com/new-york/nypd-stop-and-frisk-detains-millions-results-article-1.1307179).

We chose this dataset to understand the other side of the issue, to get a perspective about the kind of people detained (based on their age group, sex and race) who were found to be guilty of possessing knives or contraband, that are commonly found with people guilty of illegal possession. The chosen dataset alone provided us with an interesting perspective on this, based on the most recent update to the *Frisk* database maintained by the NYPD available for use; we have thus not integrated other datasets since simple modifications to the chosen dataset gave us interesting results to gain a good insight for this purpose.

d. Running the program

Type **'sh run.sh INTEGRATED-DATASET.csv <min_supp> <min_conf>'** in the project directory to run the program, where:

i. *<min_supp>* is the minimum support for a considered itemset in the transaction database and is a value between 0 and 1. Eg. *0.1*
ii. *<min_conf>* is the minimum confidence desired and is a value between 0 and 1. Eg. *0.6*

e. Internal design

The apriori algorithm to obtain frequent itemsets described in [1] is run on the data derived in the INTEGRATED-DATASET.csv file. No variations have been made to this algorithm in the implementation, since the original algorithm gives us interesting results. The program is designed using functions to define various stages of the algorithm as described below:

i. *generateDatabase*: Generates the 'transactions' from the given dataset file.
ii. *generateInitialLargeItemsets*: Generates the unique 'items' present in the dataset along with their support count and returns the 1-large itemset.
iii. *genCandidateItemsets*: Generates the candidate itemsets for a next pass of the algorithm.
iv. *pruneItemsets*: Filters and prunes a generated candidate set by removing the itemsets that have at least one subset missing from the large itemset generated in the previous pass.
v. *calculateSupport*: Calculates the supports of the itemsets derived in iv. and returns the large itemset for the current pass, to be used to generate the candidate set for the next pass of the algorithm.

The algorithm stops when the large itemset obtained in a current pass has no members. The large itemsets recorded during every pass of the algorithm constitute the frequent itemsets, by the apriori rule.

The frequent itemsets obtained from the algorithm are scanned and high-confidence association rules are mined using the *mineStrongRules* function. These rules contain exactly one item on the right hand side and at least one item on the left hand side as required. The confidence of a rule is calculated as described in class as the ratio of the support of all items constituting the rule to the support of the set of items on the left hand side of the rule.

The *createOutputFile* function logs the results, i.e. the frequent itemsets and the high-confidence association rules in the *output.txt* file in the format as required.

f. Interesting run

Command-line specification of run: sh run.sh INTEGRATED-DATASET.csv 0.3 0.7

Interestingness of results:
It is observed that most of the people who were stopped, frisked and found with contraband or knives were *Males*. For instance, around 97% of the medium-built (*M*) people found with contraband were males, around 91% of the people in the age group of *<15-30>* years found with contraband were males and around 97% of the people found with knives were males. Interestingly enough, it is also observed that around 70% of the

people who were stopped, frisked, searched and found with contraband were *Males* in the age group of <*15-30*> years, around 71% were medium-built (*M*) and around 75% were *Black (B) Males*.

These results thus present a general inference about the description and the kind of people who have previously been found to be guilty of possession of knives and contraband upon being frisked and searched, thereby supporting the claim that people belonging to a particular age group,gender,race and having a certain body type are more or less likely to be stopped, questioned and frisked to check for illegal possession of goods by the NYPD.

REFERENCES

[1] Agrawal, Rakesh; Ramakrishnan, Srikant; "*Fast Algorithms for Mining Association Rules*", VLDB 1994.

[2] Gravano, Luis; *Lecture Notes*, COMS E6111: Advanced Database Systems; Columbia University, New York; Fall 2015.