rule-me-maybe

COMS 6111 Project 3

Group

Helena He (hh3090) Kristine Pham (klp2157)

Files

- main.py
- daily_inmates.csv
- example-run.txt
- README.md

To Run the Program

NOTE: The Google Cloud VM we used for this project is the same as the one we used for Project 2.

To install all packages needed to run the program, use this command to install numpy and pandas:

```
pip install numpy pandas
```

To run the program, use this command:

```
python3 main.py daily_inmates.csv <min_sup> <min_conf>
```

Dataset Description

(a) We used NYC Open Data's "Daily Inmates In Custody" dataset.

(b) The function `update_csv()`, commented out at the bottom of `main.py`, takes in the original csv file and expands the race and Y/N items. For example, we expanded the 'BRADH' column items, which tells you if an inmate is under mental observation, to "N_MENTAL_OBSV" or "Y_MENTAL_OBSV". In addition, we converted the numerical variable, Age, to a categorical variable using an age range instead.

(c) This dataset is compelling because it can provide insight about not only criminal activity within the city, but also the state of the inmates. The dataset shares with you their age, if they are affiliated with a gang, and if they are under mental observation. We can find patterns to see how all these attributes relate to other attributes, like their top charge, their infractions, and the reason they are being held in custody.

Project Design

The apriori algorithm is used to generate all frequent itemsets. There are two functions used for this, `apriori` and `apriori_gen`. These are implemented using the pseudocode given in Section 2.1 of the Agrawal and Srikant paper in VLDB 1994.

The following happens when the `apriori` function is called:

1. Scan the transactions once to build frequent 1-itemsets that meet min_sup.
2. Repeat for k = 2, 3, … until no new frequent itemsets appear:
    - Join step: take every pair of frequent (k–1)-itemsets that share their first k–2 items and merge them into a sorted k-tuple.
    - Prune step: for each k-tuple candidate, generate all its (k–1)-sized subsets. Delete the candidate if any subset isn't already in the previous level's frequent set Lk-1.
    - Calculate the supports of each valid generated candidate. Discard any candidates that don't meet the min_sup threshold. This filtered set becomes Lk

The functions `build_k_itemset_rules` and `build_high_conf_rules` are used for rule generation. For each frequent k-itemset, we try moving each element to the RHS and compute:

$$\text{confidence} = \frac{\text{support}(LHS)}{\text{support}(LHS \cup RHS)}$$

Finally, we output only the association rules that meet the min_conf threshold.

Compelling Sample Run

Run this command to see a compelling sample run:

```
python3 main.py daily_inmates.csv 0.01 0.7
```

These results are compelling because we can gain insight on the state of inmates currently being processed in our criminal justice system.

For example, one compelling rule we see is: [130.35] => [MED] (Conf: 80.2083%, Supp: 1.0505%)

The number 130.35 refers to the New York Penal Law § 130.35, which describes rape in the first degree. This rule indicates that inmates with a rape charge against them are placed in medium custody. Intuitively, this reveals that correctional systems often reserve maximum level custody for the highest risk inmates and minimum for low risk ones. Medium custody is designed for individuals who pose a serious risk but who aren't considered the very highest flight or violence risk.

Another interesting rule we see is: [120.05,N_GANG,N_INFRACTION] => [Y_MENTAL_OBSV] (Conf: 70.1389%, Supp: 2.7558%)

This rule indicates that if an inmate has a charge of assault in the 2nd degree, no gang affiliation, and no previous infractions, then they are most likely under mental health observation. This is interesting because an otherwise "clean" record (no gang ties, no infractions) combined with a serious assault suggests a one-off episode driven by distress, paranoia, or psychosis rather than criminality.

Finally, we have a rule that may indicate troubling racial bias in our criminal justice system: [125.25,18-25,MAX] => [BLACK] (Conf: 70.0787%, Supp: 1.2142%)

This rule says that when an inmate is 18–25 years old, held in maximum security, and charged with second degree murder, then they are most likely black. Taken together, this rule shows that age, custody level, and crime severity (factors that should be race neutral) have become a proxy for race in our data. This disparity reflects structural biases at every stage of our criminal justice system, such as who is stopped and charged, who gets held without bail, and who is classified as "maximum risk", rather than any real connection between race and criminal behavior.