PROJECT REPORT:

ASSOCIATION MINNING AND CLUSTERING

PROJECT REPORT BY:

TUSHAR JINDAL

ROLL NUMBER: 18CSU218

NCU
THE NORTHCAP UNIVERSITY
NAAC ACCREDITED
(Formerly ITM University, Gurugram)

INDEX

NCU
THE NORTHCAP UNIVERSITY
NAAC ACCREDITED
(Formerly ITM University, Gurugram)

APRIORI ALGORITHM

1. OBJECTIVE

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction. Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database

Important Terms:

1. Minimum support: Support is an indication of how frequently the itemset appears in the dataset.

$$\text{Support (A)} = \frac{\text{Number of transaction in which A appears}}{\text{Total number of transactions}}$$

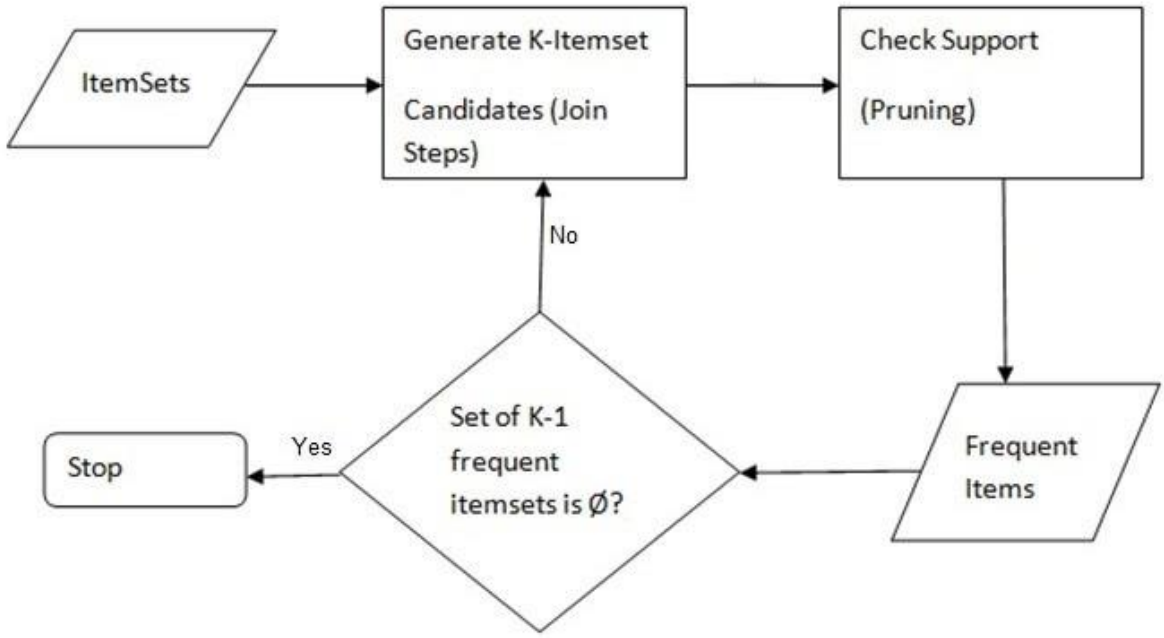
2. Confidence: confidence constraint is applied to these frequent itemsets in order to form rules.

$$\text{Confidence } (A \longrightarrow B) = \frac{\text{Support(AUB)}}{\text{Support(A)}}$$

3. Lift: This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

$$\text{lift}(A \Rightarrow B) = \frac{\text{confidence}(A \Rightarrow B)}{P(B)} = \frac{P(A \cup B)}{P(A)P(B)}$$

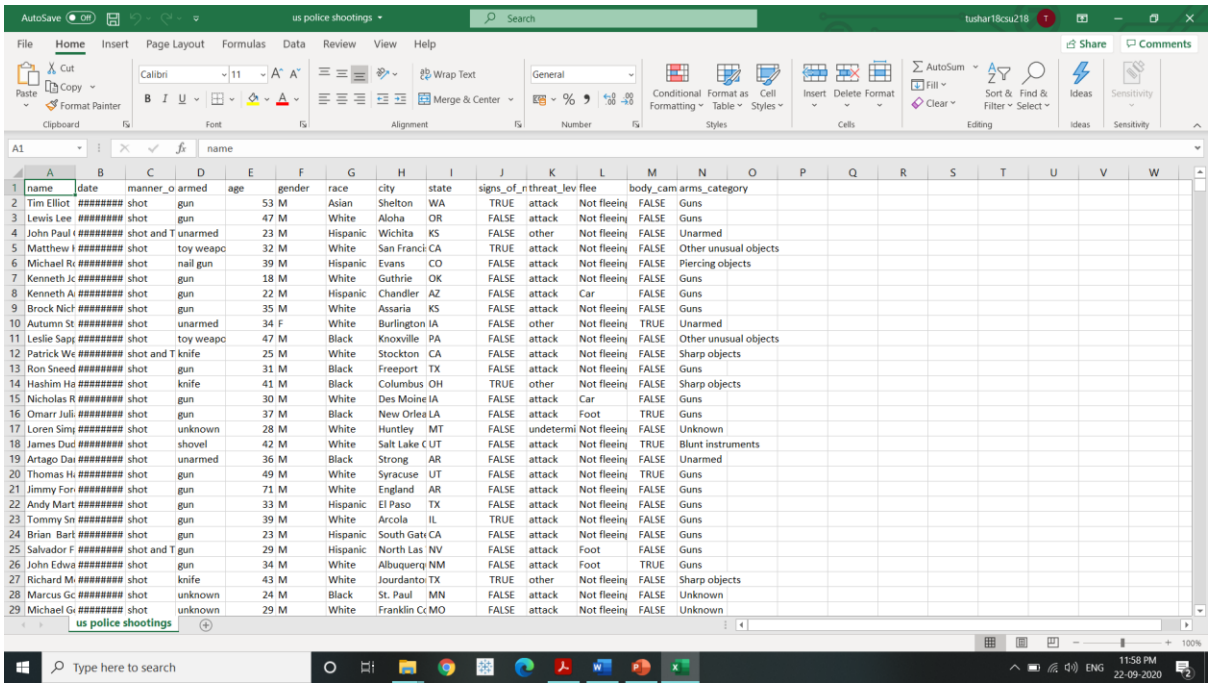
Flow Diagram for Apriori algorithm:



NCU
THE NORTHCAP UNIVERSITY
NAAC ACCREDITED
(Formerly ITM University, Gurugram)

2. DATA SET DESCRIPTION

The data set I have chosen for the project is the **US police shooting dataset.** The reason for choosing this dataset was to address the recent growing concerns about police shooting people of a certain community without required permissions. This gave rise to a lot of social concerns about the growing racism shown by the US police. I have tried to find patterns in the shooting incidents and tried to verify if this is true or not.



name	date	manner_o	armed	age	gender	race	city	state	signs_of_r	threat_lev	flee	body_cam	arms_category
Tim Elliot	########	shot	gun	53	M	Asian	Shelton	WA	TRUE	attack	Not fleeing	FALSE	Guns
Lewis Lee	########	shot	gun	47	M	White	Aloha	OR	FALSE	attack	Not fleeing	FALSE	Guns
John Paul	########	shot and T	unarmed	23	M	Hispanic	Wichita	KS	FALSE	other	Not fleeing	FALSE	Unarmed
Matthew	########	shot	toy weapo	32	M	White	San Franci	CA	TRUE	attack	Not fleeing	FALSE	Other unusual objects
Michael R	########	shot	nail gun	39	M	Hispanic	Evans	CO	FALSE	attack	Not fleeing	FALSE	Piercing objects
Kenneth Jc	########	shot	gun	18	M	White	Guthrie	OK	FALSE	attack	Not fleeing	FALSE	Guns
Kenneth A	########	shot	gun	22	M	Hispanic	Chandler	AZ	FALSE	attack	Car	FALSE	Guns
Brock Nic	########	shot	gun	35	M	White	Assaria	KS	FALSE	attack	Not fleeing	FALSE	Guns
Autumn St	########	shot	unarmed	34	F	White	Burlington	IA	FALSE	other	Not fleeing	TRUE	Unarmed
Leslie Sap	########	shot	toy weapo	47	M	Black	Knoxville	PA	FALSE	attack	Not fleeing	FALSE	Other unusual objects
Patrick We	########	shot and T	knife	25	M	White	Stockton	CA	FALSE	attack	Not fleeing	FALSE	Sharp objects
Ron Sneed	########	shot	gun	31	M	Black	Freeport	TX	FALSE	attack	Not fleeing	FALSE	Guns
Hashim Ha	########	shot	knife	41	M	Black	Columbus	OH	TRUE	other	Not fleeing	FALSE	Sharp objects
Nicholas R	########	shot	gun	30	M	White	Des Moine	IA	FALSE	attack	Car	FALSE	Guns
Omarr Juli	########	shot	gun	37	M	Black	New Orlea	LA	FALSE	attack	Foot	TRUE	Guns
Loren Sim	########	shot	unknown	28	M	White	Huntley	MT	FALSE	undetermi	Not fleeing	FALSE	Unknown
James Dud	########	shot	shovel	42	M	White	Salt Lake C	UT	FALSE	attack	Not fleeing	TRUE	Blunt instruments
Artago Dai	########	shot	unarmed	36	M	Black	Strong	AR	FALSE	attack	Not fleeing	FALSE	Unarmed
Thomas Ha	########	shot	gun	49	M	White	Syracuse	UT	FALSE	attack	Not fleeing	TRUE	Guns
Jimmy For	########	shot	gun	71	M	White	England	AR	FALSE	attack	Not fleeing	FALSE	Guns
Andy Mart	########	shot	gun	33	M	Hispanic	El Paso	TX	FALSE	attack	Not fleeing	FALSE	Guns
Tommy Sn	########	shot	gun	39	M	White	Arcola	IL	TRUE	attack	Not fleeing	FALSE	Guns
Brian Bart	########	shot	gun	23	M	Hispanic	South Gate	CA	FALSE	attack	Not fleeing	FALSE	Guns
Salvador F	########	shot and T	gun	29	M	Hispanic	North Las	NV	FALSE	attack	Foot	FALSE	Guns
John Edwa	########	shot	gun	34	M	White	Albuquerq	NM	FALSE	attack	Foot	TRUE	Guns
Richard M	########	shot	knife	43	M	White	Jourdanto	TX	TRUE	other	Not fleeing	FALSE	Sharp objects
Marcus Gc	########	shot	unknown	24	M	Black	St. Paul	MN	FALSE	attack	Not fleeing	FALSE	Unknown
Michael G	########	shot	unknown	29	M	White	Franklin C	MO	FALSE	attack	Not fleeing	FALSE	Unknown

My Dataset contains 4895 rows and 14 columns.

Data pre-processing and reasons:

1. **Dropping the unnecessary columns**: I have Dropped the columns 'name', 'date', 'body_camera', 'signs_of_mental_illness', 'city' and 'state' as values of name and date were not frequent and columns like 'body_camera', 'signs_of_mental_illness', 'city' and 'state' were not required according to the objective of my project.

2. **Checking of missing values:** I checked if my dataset contained any missing values by which I found that there were no missing values in my dataset.

3. RULE MINING PROCESS

Parameter setting:

1. Minimum length: we want at least 3 items to be associated. No point in having a single or two items in the result.
2. Minimum lift: Minimum of 3 (less than that is too low)
3. Minimum support: 0.05 (Randomly taken)
4. Minimum confidence: At least 60%

Choice of algorithm:

I have applied apriori to find the frequent patterns in the shooting incidents and also applied K-means Clustering to try to verify if both are showing same results.

4. RESULTING RULES

1. General description: The reason for choosing this dataset was to address the recent growing concerns about police shooting people of a certain community without required permissions. This gave rise to a lot of social concerns about the growing racism shown by the US police. I have tried to find patterns in the shooting incidents and tried to verify if this is true or not.

2. Number of Rules: 33

3. Selection Of rules for client: Top 20 rules having highest lift.

Out[13]:

	Association	Lift
1	[Unarmed, unarmed]	14.0661
4	[Unarmed, unarmed, M]	14.0661
11	[Unarmed, unarmed, shot]	14.0661
18	[Unarmed, shot, M, unarmed]	14.0661
2	[Unknown, unknown]	11.7105
5	[Unknown, M, unknown]	11.7105
19	[shot, Unknown, M, unknown]	11.7105
12	[Unknown, shot, unknown]	11.7105
28	[shot, White, M, Sharp objects, knife]	6.17925
23	[Sharp objects, knife, shot, White]	6.157
14	[Sharp objects, knife, M, White]	6.11893
7	[Sharp objects, knife, White]	6.10869
30	[shot, White, Not fleeing, Sharp objects, knife]	6.08026
8	[Sharp objects, knife, attack]	6.06055
15	[Sharp objects, knife, M, attack]	6.05245
25	[White, M, Not fleeing, Sharp objects, knife]	6.04122
20	[Sharp objects, knife, Not fleeing, White]	6.02853
22	[Sharp objects, knife, shot, Not fleeing]	5.99842
10	[Sharp objects, knife, shot]	5.99791
27	[shot, M, Not fleeing, Sharp objects, knife]	5.9954

5. RESULTS

By the study, police does not seem to target any particular community therefore showing no sign of racism. Many people shot were unarmed or very few of the suspects had any kind of sharp objects like a knife.

K-MEANS CLUSTERING

K-means is a simple unsupervised machine learning algorithm that groups data into a specified number (k) of clusters.

Objective:

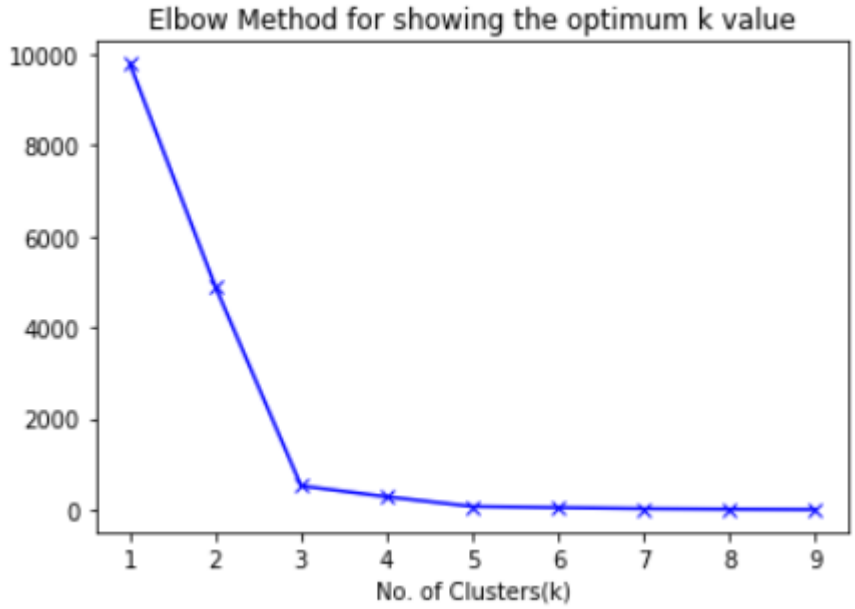
I have made clusters using attributes “manner of death” and “race” to verify the results of apriori algorithm.

Data pre-processing:

1. Dropping the columns: I have made clusters using attributes “manner of death” and “race” because I wanted to check if any particular race is being targeted showing the sign of racism and compare it with the apriori algorithm.
2. Checking for missing values: My dataset does not contain any missing value.
3. Encoding: I have encoded attributes “manner of death” and “race” in order to convert them from categorical to numerical values so that clustering can be applied to them.
4. Scaling data: It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

K-MEANS CLUSTERING PROCESS:

1. Elbow method: Implements the elbow method for determining the optimal number of clusters. Choose a number of clusters that covers most of the variance.



Since we are getting elbow at k=3 therefore number of clusters for this project is 3.

2. Let X = {x_1,x_2,x_3,……..,x_n} be the set of data points and V = {v_1,v_2,…….,v_c} be the set of centers.

   1) Randomly select *'c'* cluster centers. (Elbow method)

   2) Calculate the distance between each data point and cluster centers.

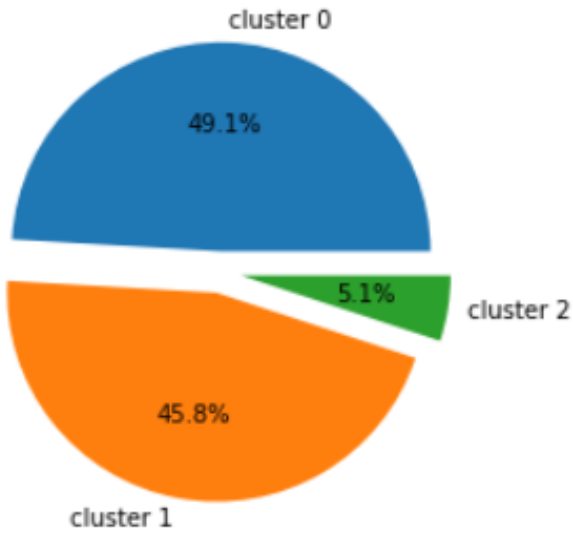
   3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

   4) Recalculate the new cluster center using new centers

   5) Repeat the same steps till the centers of clusters do not change.

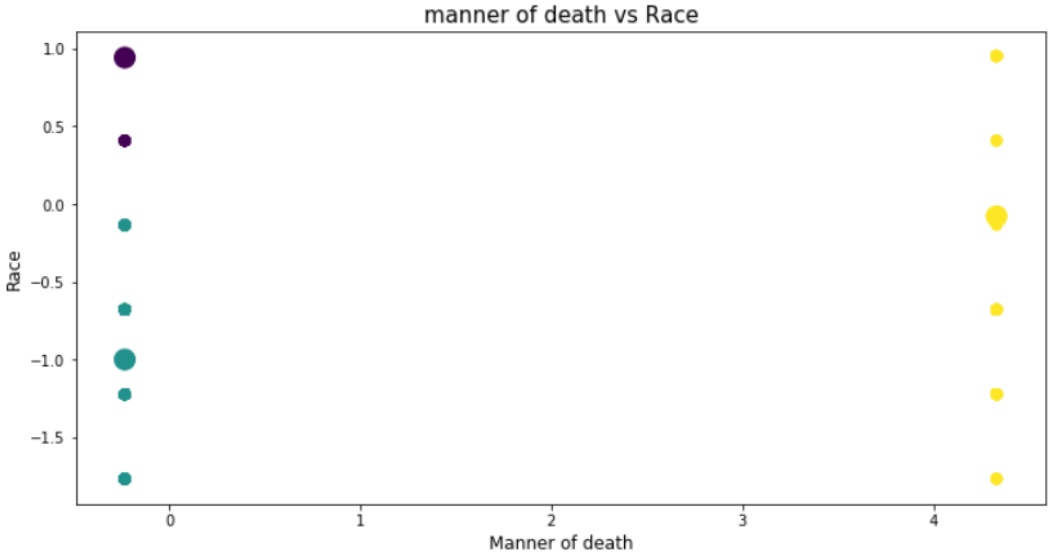
K-MEANS CLUSTERING RESULTS

1. After performing clustering on the data, the data was divided into 3 clusters. Here is the pie chart visualising clusters telling what percentage of the data has been assigned to that cluster.



1. cluster 0 consists of 49.1% of data
2. cluster 1 consists of 45.8% of data
3. cluster 2 consists of 5.1% of data

2. Plotting Scatter plot to show the three clusters with their respective centres.



All three clusters were of mixed "race" which shows that there was no selected targeting of people of particular community.

CONCLUSION

Both Apriori and k-means clustering algorithm verified that that was no racism shown by US police. They are not targeting any particular community.