University of Pisa

Department of Computer Science

Data Mining Report

Gun Incidents in the USA

Andrea Franceschi (*mat1*)
Vincenzo Gargano (*mat2*)
Luca G. Pinta (*579458*)

A.Y. 2023-2024

Contents

Chapter 1

Data Understanding

List main numerical measure about the dataset (*e.g. number of rows, number of attributes, domain of attributes, etc*).

1.0.1 Data semantics

Describe briefly the semantics of each attribute, specify the exact measure (e.g a certain ID can have numerical values, char value that if start with C identify a specific category, etc)

1.0.2 Data quality

List incosistency with meaning described before, statistical useless data if present, repetitions, if some text needed standardization (lower/uppercase), if dropped missing value record of applied manipulations. Search for outliers, present box plots of outliers and distriubtion histograms: specify which method is used for outliers removal and why, describing the size of the DS after manipulation and the percentage of decrease respect to original one.

1.0.3 Data distribution, statistics & correlation

Perform a statistical analysis to assess the distribution of each attribute and find hidden information in the dataset. Compute the Pearson correlation matrix, shows if useful result are obtained, otherwise analyze attribute or pair of attribute, showing their distribution. Compare attribute distribution and range of values for specific range of time.

Chapter 2

Data Preparation

Join datasets, refining specific events if possible, create new attributes by join them

2.0.1 Data semantics

2.0.2 Data quality

2.0.3 Data distribution, statistics & correlation

Chapter 3

Clustering

In this chapter we discuss **3 different clustering algorithms** (*K-means, DBScan and Hierarchical*), fine-tuning their hyperparameters and evaluating results from both the statistical and semantical point of view.

Chapter 4

Predictive Analysis

Chapter 5

Pattern Mining

Chapter 6

Conclusion