Data mining project

Gun Incidents in the USA

A.Y. 2023/2024

A **project** consists of data analysis based on data mining tools. The project has to be performed by a team of 3 students. It has to be performed by using Python. The guidelines require addressing specific tasks, and results must be reported in a unique paper. This paper's total length must be **25 pages** of text including figures. The students must deliver both: paper and well-commented Python notebooks.

Dataset description

The data are divided in 3 csv files. The main one, *incidents.csv*, contains information about gun incidents in the USA.
In the dataset there are the following variables:

1. date: date of incident occurrence
2. state: state where incident took place
3. city_or_county: city or county where incident took place
4. address: address where incident took place
5. latitude: latitude of the incident
6. longitude: longitude of the incident
7. congressional_district: congressional district where the incident took place
8. state_house_district: state house district
9. state_senate_district: state senate district where the incident took place
10. participant_age1: exact age of one (randomly chosen) participant in the incident
11. participant_age_group1: exact age group of one (randomly chosen) participant in the incident
12. participant_gender1: exact gender of one (randomly chosen) participant in the incident
13. min_age_participants: minimum age of the participants in the incident
14. avg_age_participants: average age of the participants in the incident
15. max_age_participants: maximum age of the participants in the incident
16. n_participants_child: number of child participants 0-11
17. n_participants_teen: number of teen participants 12-17
18. n_participants_adult: number of adult participants (18 +)
19. n_males: number of males participants
20. n_females: number of females participants
21. n_killed: number of people killed
22. n_injured: number of people injured
23. n_arrested: number of arrested participants
24. n_unharmed: number of unharmed participants
25. n_participants: number of participants in the incident

26. notes: additional notes about the incident
27. incident_characteristics1: incident characteristics
28. incident_characteristics2: incident characteristics (not all incidents have two available characteristics)

The second file, *povertyByStateYear.csv* contains information about the poverty percentage for each USA state and year, so it includes the following variables:

1. state
2. year
3. povertyPercentage: poverty percentage for the corresponding state and year

The third file, *year_state_district_house.csv* contains information about the winner of the congressional elections in the USA, for each year, state and congressional district. It includes the following variables:

1. year
2. state
3. congressional_district
4. party: winning party fort the corresponding congressional_district in the state, in the corresponding year
5. candidateVotes: number of votes obtained by the winning party in the corresponding election
6. totalVotes: number total votes for the corresponding election

Task1: Data Understanding and Preparation (30 points)

Task 1.1: Data Understanding

Explore the incidents dataset with the analytical tools studied and write a concise “data understanding” report assessing data quality, the distribution of the variables and the pairwise correlations.

Task 1.2: Data Preparation

Improve the quality of your data and prepare it by extracting new features interesting for describing the incidents. Therefore, you are going to describe the single incident and examples of indicators to be computed are:

- How many males are involved in the incident w.r.t. the total number of males involved in incidents for the same city and in the same period?
- How many injured and killed people have been involved w.r.t the total injured and killed people in the same congressional district in a given period of time?
- Ratio of the number of the killed people in the incident w.r.t. the number of participants in the incident
- Ratio of unharmed people in the incident w.r.t. the average of unharmed people involved in incidents for the same period

Note that these examples are not mandatory. You can derive indicators that you prefer and that you consider interesting for describing the incidents.

It is MANDATORY that each team defines some indicators. Each of them has to be correlated with a description and when it is necessary also its mathematical formulation. The extracted variables will be useful for the clustering analysis (i.e., the second project's task). Once the set of indicators is computed, the team has to explore the new features for a statistical analysis (distributions, outliers, visualizations, correlations).

Subtasks of DU:

- Data semantics for each feature that is not described above and the new one defined by the team
- Distribution of the variables and statistics
- Assessing data quality (missing values, outliers, duplicated records, errors)
- Variables transformations
- Pairwise correlations and eventual elimination of redundant variables.

Nice visualization and insights can be obtained, exploiting the latitude and longitude features (e.g. https://plotly.com/python/getting-started/).

Task 2: Clustering analysis (30 POINTS - 32 with optional subtask)

Based on the features extracted in the previous task, explore the dataset using various clustering techniques. Carefully describe your decisions for each algorithm and which are the advantages provided by the different approaches.

Subtasks

- Clustering Analysis by K-means on the entire dataset:
    1. Identification of the best value of k
    2. Characterization of the obtained clusters by using both analysis of the k centroids and comparison of the distribution of variables within the clusters and that in the whole dataset
    3. Evaluation of the clustering results
- Analysis by density-based clustering**.** In this task, choose **one state** in the dataset:
    1. Study of the clustering parameters
    2. Characterization and interpretation of the obtained clusters
- Analysis by hierarchical clustering. In this task, choose **one state** in the dataset:
    1. Compare different clustering results got by using different version of the algorithm
    2. Show and discuss different dendrograms using different algorithms
- Final evaluation of the best clustering approach and comparison of the clustering obtained
- **Optional (2 points):** Explore the opportunity to use alternative clustering techniques in the library: https://github.com/annoviko/pyclustering/

Note: The final report delivered within the end of December can also improve the already delivered tasks.

Task 3: Predictive Analysis (30 POINTS)

Consider the problem of predicting for each incident (considering the whole dataset for this task) the label which is a binary variable that indicates if in the incident there have been **at least a killed** person or not.

The students need to:

1) define new features that enable the classification. Please, reason on the suitability of the features defined for the clustering analysis. In case these features are not suitable for the above prediction problem you can also change the indicators.
2) perform the predictive analysis comparing the performance of different models discussing the results and discussing the possible preprocessing applied to the data for managing possible identified problems that can make the prediction hard. Note that the evaluation should be performed on both training and test sets.

Note: The final report delivered within 8 January can also improve the already delivered tasks.