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.

Task 4: Address one of the two tasks (32 POINTS)

Task 4.1: Time Series Analysis

Consider the incidents dataset and only incidents that happened in the years [2014, 2015, 2016, 2017]. Extract a time series for each city, computing for each week of the 4 years a score. The score can be an index created in one of the previous tasks or a new one and it can be different for each subtask. Therefore, each value of the time series (one for each city) corresponds to the score value for a certain week of 2014, 2015, 2016, 2017. You can filter the cities, excluding the ones with a low number of weeks with incidents. For example, you can consider only cities with a number of weeks with incidents greater than 15% of the total number of the weeks of the 4 years.

Task 4.1.1: Clustering and motif/anomalies extraction

The goal of this task is grouping similar cities through the use of the created time series, based on the defined score. Analyze the results of the clustering and extract motifs and anomalies in the time series for a deep understanding and exploration..

Task 4.1.2: Shapelet extraction

Exploiting the created time series, extract the shapelet according to the class of the binary variable *isKilled*.

Note: For this subtask there must be no relationship between the score used for the time series and the *n_killed* variable, as *isKilled* is derived from this.

Task 4.2: Explanation Analysis

In Task 3 you trained several machine learning models. The major drawback of some models is that, even if they have good predictive performance, they are uninterpretable, e.g. the internal reasonings of the model are difficult (or impossible) to understand. To overcome this limitation, there are two approaches: either explain with post-hoc methods the machine learning models already trained, such as SHAP and LIME; or train interpretable by design machine learning models, such as EBM. In this task we ask you to:

Explain locally some non interpretable model (trained in task 3) using LIME and SHAP;
Train an interpretable by design model (either EBM or TabNet)

You can use the following libraries:
https://github.com/interpretml/interpret for EBM
https://github.com/slundberg/shap for SHAP
https://github.com/marcotcr/lime for LIME
https://github.com/titu1994/tf-TabNet for TabNet

After the application of post-hoc methods and interpretable models, provide some explanation examples (plots of different kinds) and apply the evaluation metrics presented during the lectures to find out the best explanation approach among the ones proposed.

Rules for final delivery and Exam

Project Delivery. The final deadline of the project is **8th January 2024 at 23:59**. This deadline is **STRICT**. No extension is possible because then the winter session of exams starts. Each group must deliver by email to anna.monreale@unipi.it and lorenzo.mannocci@phd.unipi.it a zipped folder named **DM_GroupID.zip** and containing 4 folders and 1 pdf file:

1. a folder named **DM_GroupID_TASK1**, containing source code of data understanding
2. a folder named **DM_GroupID_TASK2**, containing source code of data clustering
3. a folder named **DM_GroupID_TASK3**, containing source code of classification
4. a folder named **DM_GroupID_TASK4**, containing source code of time series analysis/explanation analysis
5. a pdf file with maximum 25 pages including figures discussing the results of the 4 tasks. The name of this file must be: **DM_Report_GroupID.pdf.** The file must contain the list of authors (i.e., members of the group).

Remember that the final submission can contain updated versions of the work already delivered in the previous deadlines.

ATTENTION: students that did not deliver the project by 8th January need to ask the new project description to teachers that must be delivered in 20 days.

Exam

There are two possible options for the exam:

1. project presentation + questions on the whole program
2. project presentation + paper presentations (in the dates already fixed) + question on the topics not covered by the project

I prefer to have group presentations of the project. If this is impossible otherwise we can find a solution together.

How to book for the exam colloquium?

In https://esami.unipi.it/ you can find the dates for the exam: one for January and one for February. Each student must do the registration on one of the 2 dates. These are not the dates of the colloquium but we will use the list of registered students for organizing the exam dates. We will share with you a calendar for the oral exam.