2023 :: Advanced Topics in Data Science / Data Mining II :: Practical Assignment

Analysis of the Chicago Crimes over the last 6 years

Final project submission deadline: May 26th @ 23:30 (on MoodleUP)
Project proposal presentation: May 29th, during lectures.

Important:
- Be aware of the University of Porto Policy on Plagiarism.
- Properly <u>cite any external source</u> that you used.
- You are **not allowed to** use or reuse any piece of code **created by the other groups**.
- However, you can use any programming language for the development of your work, as well as code that is freely available (open source) on the Internet (provided you understand it, and properly refer to it).

General description:
You are provided with a subset of a well-known dataset - the Chicago Crimes - for the purpose of conducting specific analyses and making predictions. Your dataset covers the years 2017 to 2023. It is important to note that you must work with this given dataset and not obtain it from any other source.

The objectives of this project are twofold:

i) First, you are expected to conduct a comprehensive exploration of the data to gain insights into the nature of the crimes, their most common locations, and the periods during which they typically occur. Your findings should be presented using visualizations, association rules, recommendations, and link analysis.

ii) In the second part of the project, you are required to use four machine learning models capable of predicting the occurrence of crimes (or specific characteristics associated with them). These models should then be evaluated against actual data to assess their performance. In detail:

Part I

- Perform an exploratory data analysis on the dataset.
- Produce relevant visualizations to understand the crimes, or some subset of them.
- Identify association rules related to the crimes.
- Identify crime situations that typically co-occur.
- Create a graph for crime-related entities and perform link analysis.

Part II

- Identify relevant features to use in your machine learning model. Note that most features will need to be computed from the available information in the dataset.
- Use *four* machine learning algorithms to train the classifier
- Perform an evaluation of the model/classifiers used.

Remaining activities:

- Report your process in an article-type document (IEEE template, **max. 4 pages**)
- Do a presentation of your work.

Deliverables:
1. All the material needed to perform the analysis and the created plots/figures (R code)
   Filename: `Group##_code.zip`
2. The project report in the form of an article in PDF
   Filename: `Group##_article.pdf`
3. An electronic presentation (preferably in PPT/PPTX)
   Filename: `Group##_presentation.(pptx)`

1. **[30 points] The dataset exploration and creation of a knowledge base**
   a. Perform exploratory data analysis.
   b. Do all necessary data cleaning.
   c. Use visualization to better and more rapidly understand the data.
   d. Derive the association rules, recommendations and graphs based on your gained knowledge of the data.
   e. Analyze the results.

2. **[20 points] The classification System**
   a. Recall that for creating your ML model, the most interesting features to be included will derive from combinations of other fields and from the previous analysis you have made.
   b. You are expected to create a model with more than 8 features. Also, for each feature, there must be a particular motivation for including it in the model (which you should briefly describe in your report).
   c. You should also experiment your model with four machine learning algorithms.
   d. Evaluate the selected algorithms according to well-known metrics.

   Be advised: it is important that your group can add something (e.g., a different model, different features) that differentiates itself from any other group in the class. You will score for that creativity.

3. **[15 points] Report as an article**
   a. The project report should be written in the standard IEEE conference format template (**max. 4 pages**) and submitted in PDF.
   b. The report should include the title, authors, an abstract and a conclusion. The inclusion in the written report of at least 4 figures is mandatory in order to be more objective in the EDA.
   c. Explain the transformations you made to the dataset in order to obtain your final model, discuss the features you use to tackle the problem and your novelties.
   d. Report on the evaluation of your model.

   Note: reports that do not strictly follow the template will not be considered for assessment.

4. **[15 points] Final presentation**
   a. Prepare a 15-minute (maximum) presentation to explain what you did and the results you achieved using an electronic presentation as a support, mainly for graphics, illustrations, and tables. Note that every group member must participate in the presentation.
   b. You can schedule your presentation for one of the slots on an activity in MoodleUP (to be available). Please, choose a single group participant to register the whole group choice, as the number of time slots is limited.

   Note: You will also get points based on the **visual quality** of the electronic presentation.