Data mining project

Bots in Twitter

A.Y. 2022/2023

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 contains information about tweets. This dataset is composed of 2 csv files: users.csv, tweets.csv.
In users.csv there are the following variables:

1. User Id: a unique identifier of the user
2. Statues Count: the count of the tweets made by the user at the moment of data crawling
3. Lang: the user's language selected
4. Created at: the timestamp in which the profile was created
5. Label: a binary variable that indicates if a user is a *bot* or a *genuine* user

In tweets.csv each row contains information about a single tweet. In this case the variables are:

1. ID: a unique identifier for the tweet
2. User Id: a unique identifier for the user who wrote the tweet
3. Retweet count: number of retweets for the tweet in analysis
4. Reply count: number of reply for the tweet in analysis
5. Favorite count: number of favorites (likes) received by the tweet
6. Num hashtags: number of hashtags used in the tweet
7. Num urls: number of urls in the tweet
8. Num mentions: number of mentions in the tweet
9. Created at: when the tweet was created
10. Text: the text of the tweet

Task1: Data Understanding and Preparation (30 points)

Task 1.1: Data Understanding

Explore the 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 user and his/her behavior derived from the information collected from the tweets.
Examples of indicators to be computed are:

- How many tweets were published by the user?
- How many tweets are published by the user in a given period of time?
- Total number of tweets
- Total number of likes and comments
- Ratio between the number of tweets and the number of likes
- Entropy of the user
- Average length of the tweets per user
- Average number of special characters in the tweets per user

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

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 profile 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.

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

Based on the user's profile 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:
    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**:**
    1. Study of the clustering parameters

2. Characterization and interpretation of the obtained clusters

- Analysis by hierarchical clustering
    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/

Delivery of the first draft of the report with Task 1.1, Task 1.2 and Task 2: 5 November 2022

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 user the label which is a binary variable that indicates if a user is a *bot* or a *genuine* user.

. The students need to:

1) define a user profile that enables the classification. Please, reason on the suitability of the user profile, defined for the clustering analysis. In case this profile is 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.