Universidade do Minho
Departamento de Informática
Mestrado [integrado] em Engenharia Informática

Dados e Aprendizagem Automática
1°/4° Ano, 1° Semestre
Ano letivo 2024/2025

Practical Exercise no. 3

Theme

Data Exploration and Preparation

Exercise

An in-depth exploration of the data makes it possible to draw conclusions, often hidden, that may be important for understanding the domain and the problem at hand. In this practical statement, it is expected that a set of techniques will be applied to explore and process datasets.

Tasks

I - Sentiment Analysis Dataset

Download the dataset available at https://bit.ly/3ufnsgF which contains data from a set of users of a web platform and their sentiment towards it.

Carry out the following tasks:

T1. Load the dataset. Apply methods for data exploration, i.e. analyzing the data in relation to their own:

a. Central tendency;
b. Statistical dispersion;
c. Correlation between features.

T2. Create plots to visualize the data;

T3. Apply different data treatments such as:

a. Treating missing values;
b. Removing duplicate records;
c. Create an 'age_binned' attribute as a result of creating 3 bins of equal frequency on the 'age' feature;
d. For each record, extract the year, month and day of the week from the 'birthday' feature;
e. Remove users from the platform who have, at the same time, an activity on the platform ('WebActivity') of less than 1 hour and who are over 70 years old;
f. Use 'Sentiment Analysis' to determine the mode of the remaining attributes;
g. Using 'Sentiment Analysis' and 'MaritalStatus', obtain the average of the other attributes.

T4. Critically analyze the previous tasks. What conclusions can the company draw??

II - EPL Players Stats Dataset

Download the dataset available at https://bit.ly/3525yDr. This dataset contains data on the performance of various football players in the 2017/2018 edition of the Premier League.

Complete the following task:

T5. Load the dataset. Explore the data, look for relevant information and display it.

T6. Critically analyze the dataset: for example, which team is the most undisciplined? What are the top 10 assists for goals? What are the top 5 nationalities in this edition of the Premier League?