KDD 2022
MAPi
UM

jgama@fep.up.pt

Schedule

- 18/10 (JGama) Introduction to Machine Learning and Data Mining.
- Classification problems. Basic algorithms
- 1º Assessment: Classification Problems.
-
- 25/10 (JGama) Classification Algorithms: Multiple Models
- (Rita P. Ribeiro) Evaluation, ROC curves, Imbalanced Domain Learning
-
- 8/11 (J. Gama) Advanced Topics in Classification:
- Novelty Detection, Structured Output Prediction
-
- 16/11 (Alipio J.) (Wednesday)
- Text mining, Natural Language Processing
- 2º Assessment: Text Mining.
-
- 22/11 9h30 (P. Azevedo)
- Clustering.
- Frequent pattern mining, Sequence mining.
-
- 30/11 (Wednesday)
- (Alipio J.) Web Mining, Recommendation Systems

- 6/12
- (J.Gama) Semi-supervised Learning, Auto-ML
- (JGama) Data stream analysis, Social network analysis
-
-
- 13/12
- (Rita P. Ribeiro) Predictive maintenance
- (J. Gama) Students Presentations: How did I solve the Kaggle competition?
-
- 10/1/2023 (J. Gama) Ehsan Aminian, Nuno Paiva, Thiago Andrade:
- Presentations of PhD works in progress.
- Nuno Paiva, Thiago Andrade, Ehsan 11h

Evaluation

- Report on Kaggle competition
- Text mining

Kaggle Competition

- **Predict when a product goes for sale**
  - The published data contained information on a range of products in different establishments. Occasionally these products are sold on sale. Based on a set of characteristics, it is intended to classify products with this status.
  - The variable to predict is y (y=1 means with discount), all the other variables can be used to predict y.
    - 'i' is the record number (should be ignored in the analysis)
    - 'd' variables are discrete variables.
    - 'x' variables are continuous variables.
- Link:
  - https://www.kaggle.com/t/07da21c6c83c8ca1d3a3dd605552921c
  - (copy and paste in your browser)

  Students must organize themselves in groups of 2 elements.

1) The report should be delivered until 30 **November 2022** via Moodle. Authors must upload the report has a **pdf** document.
2) You can use any tool or combination of tools for the work (R, Python, Excel, Weka, KNIME, …). The report must have **at most** 12 pages.
3) The report consists of:
   - Provide basic descriptive statistics for some of the variables in the training data set.
   - Study the relevance of attributes to discriminate classes. Identify irrelevant attributes.
   - Using the **kaggle** platform find the most promising method to predict the test set.
     - This involve pre-processing methods, learning algorithms, parameters of the algorithm, ensemble models, post processing, etc.
   - Using ROC analysis which classifier would you choose?

The report will be evaluated considering the following parameters:

Structure of the report. The report must contain (at least ...):

▪ Cover page with identification of students

Brief description of the techniques used

▪ Description of the algorithms used

Experiences and analysis of results

Conclusions

- Evaluation will consider:
  - Critical analysis of the results.
  - Argumentation and justification of the choices made.