Universidade do Minho

Aprendizagem Automática em Sistemas Empresariais (Predictive Business Analytics)

Mestrado (Integrado) em Engenharia e Gestão de Sistemas de Informação, Mestrado em Sistemas de Informação

Semester 1 2023/2024

PROJECT 1

Problem statement

Using the Crisp-DM methodology it is intended to generate / induce one or several models / patterns for predicting, in the most efficient way, future values for a variable (target).

One of the resulting models should be integrated into the analytical component of a business intelligence system to be developed in the context of project 2 (deployment phase).

Data Set / Competition

Five data sets are considered for competition. The assignment of competitions to the working groups is defined as follows (Kaggle platform):

	Problem	URL	DM Goal	#col	#row
TP1	Used Car Price Predictions	https://www.kaggle.com/t/95601e086d1d4ef7bf0b28adbc2b7e1f	Regression	12	4000
TP2	Credit Card Fraud Detection (2023)	https://www.kaggle.com/t/df8effe4d67a4a07b4ceab8002c33fc5	Classification Binary	31	550K
TP3	Mobile Price Classification	https://www.kaggle.com/t/b0941383718748ec85cce4e437c0b9d1	Classification Multi-class	21	2000
TP4	Water Quality and Potability	https://www.kaggle.com/t/baaf2af079b44eaba6e25a76f11a4e70	Classification Binary	10	~3200
TP5	Car resale data (2023)	https://www.kaggle.com/t/d3649ac7f5a6454a91b8c77f2423758a	Regression	15	~17K

Deadline for submission of final reports

Week 16/10/2023 - P1 - Business Understand
Week 20/11/2023 – P1 - Data Understand + Data Preparation
Week 11/12/2023 - P1 - Modeling + Evaluation

Working Groups

4 elements (max.).

Assessment Methodology

To determine the grading, the following components and respective weights will be considered:

Report	20%;	
Application of the methodology	20%;	
Results/model performance	30%	(This parameter will be evaluated by applying a ranking among the working groups in each competition using F1-Score / RMSE);
Continuous assessment	25%;	
Peer assessment	5%	

The elements of the working group will be consulted for a differentiation in the grades.
The CRISP-DM methodology should be followed.

Tools

For datamining:
R (data mining package), Python, WEKA, RapidMiner, Other

For data processing
MS-Excel, DBMS (MS-SQL Server, MS-Access, MySQL, other), Other

References

1. Santos, M.F., Azevedo, C. Data Mining - Descoberta de Conhecimento em Bases de Dados, FCA, Portugal, 2005.
2. Azevedo, Ana; Santos, Manuel (2008). KDD, SEMMA and CRISP-DM: a parallel overview; Proceedings of DM 2008 – IADIS European Conference on Data Mining 2008, pp 182-185.
3. Hastie, T., Tibshirani, Friedman J., The Elements of Statistical Learning – Data Mining, Inference and Prediction, Springer.
4. Han, J., Kamber, M., Data Mining: Concepts and Techniques, Morgan Kaufmann, New York, USA.
5. Berthold, M., Hand, D., Intelligent Data Analysis – An Introduction, Springer.
6. Crisp-DM, http://www.the-modeling-agency.com/crisp-dm.pdf
7. ML Mastery https://machinelearningmastery.com
8. DM Competitions and Tools www.kaggle.com
9. DM tool Documentation.
10. Azevedo, Ana; Santos, Manuel (2009). An architecture for an effective usage of data mining in business intelligence systems. In Proceedings of the 13th IBIMA Conference on Knowledge Management and Innovation in Advancing Economies, pp. 1319 – 1325.
11. Azevedo, Ana; Santos, Manuel (2009). BUSINESS INTELLIGENCE - State of the Art, Trends, and Open Issues. In Proceedings of KMIS 2009 – 1st International Conference on Knowledge Management and Information Sharing, pp. 296-300.
12. Azevedo, A., & Santos, M. F. (2013). A Perspective on Data Mining Integration with Business Intelligence. In Data Mining: Concepts, Methodologies, Tools, and Applications (pp. 1873-1892). Hershey, PA: Information Science Reference. doi:10.4018/978-1-4666-2455-9.ch097
13. Business Intelligence and Analytics: Systems for Decision Support, Global Edition (10e), By Efraim Turban, Ramesh Sharda, Dursun Delen, Pearson Higher Ed USA, ISBN: 9781292009209
14. Business Intelligence, Analytics, and Data Science A Managerial Perspective, Ramesh Sharda; Dursun Delen; Efraim Turban, Pearson, ISBN: 9780134633282, 0134633288, Edition: 4th, 2017
15. Ana Azevedo, Manuel Filipe Santos (2015-2020). Integration of Data Mining in Business Intelligence Systems, IGI.
16. Machine Learning Algorithms: Popular algorithms for data science and machine learning, 2nd Edition, Giuseppe Bonaccorso, Pack Publishing, 2018.

PROJECT 2

Problem statement

It is intended to complete the deployment phase by building a Business Intelligence (BI) system based on data and data mining models resulting from project 1, allowing to achieve the goals defined in the phase of business understanding. The BI system should include:

- A database / Data Warehouse / Data Marts to store the data;
- An ETL component to populate the database (it can be in offline mode);
- An analytical component based on Data Mining models for predicting and a set of dashboards for visualizing management information and performance indicators (KPI).

Important dates

Groups will be asked to make a short demonstration between 8 January and 15 January 2024. The report can be delivered by January 15, 2024.

Working Groups

4 elements (max.).

Assessement Methodology

To determine the evaluation, the following components and respective weights will be considered:

Component	Weight
BI System	**70**
System architecture	10
Multidimensional model	10
Analytical component (data mining models)	20
Indicators	10
Dashboards	10
Quality and overall efficiency	10
Demonstration	**10**
Results presentation	5
Individual performance	5
Report	**15**
Peer assessment	**5**

The elements of the working group will be consulted for a differentiation in the grades.

Tools

Groups are free to select from (not limited to):

- Suites - MS SQL Server, Oracle BI
- Data Base Systems – MS SQL, Oracle, MySQL, MS Excel, MS Access, Oracle
- Visualization / Dashboarding – Tableau, PowerBI

References

17. BI tool documentation;
18. Scientific and technological documentation related to the application area.