188.429 Business Intelligence (VU 4,0) – WS 2022/23

Assignment 3: Data Analytics

The goal of this assignment is to solve a data analytics problem following the CRISP-DM Process. This being a class assignment rather than a real-life setting, several simplifications will have to be made. Particularly, you will need to make certain assumptions and simplifications in the course of the project, both because a real problem owner and data expert is not available, and deployment of the solution is obviously out of scope.

For performing the experiments you can use any Machine Learning platform of your choice, e.g. WEKA, Scikit-Learn, Spark / MLlib, Matlab, R, … according to your preferences, to conduct the **Data Mining and Machine Learning tasks** below.

- Information on how to obtain and use **WEKA**, the open Source Machine Learning Platform from Waikato University, is available at https://wekatutorial.com/.
- For **Scikit-Learn** you may consult the tutorial available at https://scikit-learn.org/stable/tutorial/index.html
- For **Spark/MLlib** you can consult the manual available at https://spark.apache.org/docs/latest/ml-guide.html)
- Structure your **report** for this assignment based on the structure in this assignment paper**.** For the experiments, provide detailed documentation of all steps to **ensure reproducibility** of all results based on the information provided.

(1) Form groups of two persons and **select a data set** from the OpenML Machine Learning Repository (http://www.openml.org), Kaggle (https://www.kaggle.com/datasets), or a similar benchmark data repository with the **following requirements***:*
   - posing a classification or regression problem
   - minimum **1.000 instances**,
   - minimum **15 attributes**,
   - minimum 4 class labels if it is a classification task
   - not an "artificial" dataset, i.e., a dataset consisting of synthesized, sampled or interpolated values (e.g. the BNG* datasets on OpenML)
   - where the **features carry semantics that can be interpreted** by you (i.e. not a collection of image files where features still would need to be extracted by you)
   - with a certain **variety of feature semantics and** (preferably) also **feature types** (i.e. not a data set with just 5000 bag-of-words features or greyscale histogram features of images)
   - where you **understand the semantics of the data and the domain** so that you can make reasonable assumptions on its use, the goals to be met.
   -

(2) ***Register*** **the dataset** you picked in the TUWEL Wiki. Each dataset can be used by no more than two groups! (first come, first serve - do it early to get a data set that you also find interesting to work with.)

(3) **Determine, who is "Person-A" and "Person-B" in your group** and list this in the report. The work has to be done in groups of 2 persons. However, there is always one individual responsible (and graded) for each section of the final report to ensure proper "load-balancing" in terms of responsibilities and coordinating the joint input into each section.

Prepare a report (see end of this assignment sheet for formatting and submission information) which documents your analysis containing at least (additional material/sections may be provided if you think they are important in your specific setting) the following sections as a reduced subset of the CRISP-DM process:

Assignment 3: Data Analytics

(1) Business Understanding (Responsible: A + B jointly, part of interim submission)

a. Define and describe the data source and a **scenario** in which a business analytics task based on the data set you identified should be solved
b. Clearly define and describe the **Business Objectives**
c. Clearly define and describe the **Business Success Criteria**
d. Clearly define and describe the **Data Mining Goals**
e. Clearly define and describe the **Data Mining Success Criteria**
f. Are there any **AI risk aspects** that may require specific consideration?

(2) Data Understanding: Data Description Report presenting (Responsible: A, part of interim submission)

a. **Attribute types** and their semantics,
b. **Statistical properties** describing the dataset including correlations
c. **Data quality** aspects, e.g. missing values and their potential effects and reasons, uneven distributions in certain attribute types, plausibility of values, outliers, information available on data provenance and data cleansing applied before, etc.
d. **Visual exploration** of data properties and hypotheses
e. Evaluate and document whether the data set contain attributes that are potentially **ethically sensitive,** minority classes or underrepresented data groups, unbalanced distributions with respect to bias (to guide over- and under-sampling, micro- and macro evaluation criteria).
f. What potential **risks** and additional types **of bias** exist in the data? What questions would you need to have answered by an external expert in order to determine potential bias or data quality issues?
g. Which actions are likely required in data preparation based on this analysis?

(3) *Data Preparation report* (Responsible: B, part of interim submission)

a. Analyze and perform **necessary actions** based on analysis performed in the Data Understanding phase.
b. Analyze options and potential for **derived attributes** (note: if the potential is considered low, these obviously do not necessarily have to be applied for your analysis, but options should be documented)
c. Analyze options for additional **external data sources**, attributes that might be useful to better address the business objectives or data mining goals (Note: this description may be hypothetical, i.e. you are not necessarily required to actually obtain and integrate the external data for the analysis)
d. Describe other **pre-processing steps** considered, specifying which ones were applied or not applied due to which reason. (e.g. data cleansing, transformations, binning, scaling, outlier removal, attribute removal, transcoding, …) at a level of detail that ensures reproducibility of changes to the data.

(4) Modeling (Responsible: A)

a. Identify suitable **data mining algorithms** and select one of these as the most suitable for your experiments, providing a **justification** for the selection.
b. Identify the **hyper-parameters** available for tuning in your chosen model and select one that you deem most relevant for tuning, providing a **justification**.
c. Define and document a **train / validation / test set split**, considering where necessary appropriate stratification, any dependencies between data instances (e.g. time series data) and relative sizes of the respective subsets.
d. **Train the model** on the training set and compare the performance on the validation set to identify the best hyper-parameter setting, explicitly **documenting all parameter settings** tested (avoid stating simply to have used “default parameters”, focus on reproducibility of the results you report).
e. Report suitable **performance metrics** supported, where possible, by figures/graphs showing the tuning process of the hyper parameter.

Assignment 3: Data Analytics

(5) Evaluation (Responsible: B)

a. **Apply the final model** on the test data and document performance.
b. **Re-train** the model with identical hyper-parameters using the **full train and validation data** and again apply it on the test data, documenting and reflecting on the performance.
c. Identify and document
    i. **state-of-the-art performance, i.e. the performance obtained by others** using the same (albeit potentially slightly differently pre-processed) data set as reported in literature (preferably in peer-reviewed papers, in absence of these grey literature or solid internet publications are fine as well). If no baseline performance can be identified for your task, report on other analyses/tasks using the same dataset.
    ii. the expected **base-line performance** of a trivial acceptor / rejecter or random classifier
d. **Compare the performance achieved with the benchmark and baseline** performances according to different metrics (i.e. overall, but also on per-class level (confusion matrix), micro/macro precision/recall in the case of classification tasks, regression errors in certain parts of the data space, … (Note your goal is not necessarily to obtain a better result than what has been reported in the state of the art, this is not a grading criterion! On the other hand, if the performance of your classifiers is below a random baseline or trivial acceptor / rejecter you may want to investigate the reason…)
e. **Compare the performance obtained with the success criteria defined in the Business Understanding phase**.
f. **Identify a "protected attribute"** and evaluate whether the **model exhibits a bias towards that group**. The attribute can be one that may be considered sensitive or – in absence of any actually sensitive attributes – any attribute that identifies a subgroup of the data for which you may want to identify skewed performance of the model.

(6) **Deployment: (Responsible: A+B)**

a. Compare the performance obtained with respect to the needs for addressing the **Business Objectives** (in how far are the results obtained sufficient to take decisions needed for achieving the Business Objectives, which other analyses or aspects would still be missing) and provide recommendations for deployment (fully automatic, hybrid solutions, deploying only for a part of the data space, …) as well as recommendations for subsequent analysis.
b. Consider and briefly document potential ethical aspects as well as impact assessment / risks identified in deployment
c. Document aspects to be monitored during deployment, specifying triggers that should lead to intervention.
d. Briefly re-visit reproducibility aspects reflecting on aspects well documented and those that might pose a risk in terms of reproducibility based solely on the information provided in this report

(7) **Summarize your findings (Responsible: A+B)**

a. Briefly summarize your overall findings and lessons learned
b. (**optional**) Provide **feedback on this exercise** in general: which parts were useful / less useful; which other kind of experiment would have been interesting, … (this section is, obviously, optional and will not be considered for grading. You may also decide to provide that kind of feedback anonymously via the feedback mechanism in TISS – in any case we would appreciate learning about it to adjust the exercises for next year following a major re-structuring this year based on feedback obtained.)

188.429 Business Intelligence (VU 4,0) – WS 2022/23

Assignment 3: Data Analytics

Submission guidelines:

- **Upload ONE [zip/tgz/rar] file** to TUWEL that **contains (1) your report as a PDF file** (no Word files, no TEX sources) **and (2) any auxiliary files needed for reproducing your experiments** (i.e. any scripts, transformation tools, config files etc. that you produced and that represent information not sufficiently documented in the report)**.** You **must follow this naming convention**:
  - BI2023_gr<groupno>_<Matnr.1>_<Matnr.2>.zip
  - Example: A submission of group 5 with 2 students (ids: 00059999, 00039999) looks like this: BI2023_gr05_00059999_00039999.[zip/tgz/rar]
  - Example: A submission of a single student (with group no. 99) (id: 00987654) looks like this: BI2023_gr99_00987654.[zip/tgz/rar]
  - Apply the same naming convention to the report (but obviously with pdf extension)
- **Follow the ACM formatting guidelines, using the templates provided at** https://www.acm.org/publications/proceedings-template. (Conference Proceedings Style File, 2-column layout) LaTeX recommended (you may use the Overleaf Template provided at https://www.overleaf.com/gallery/tagged/acm-official#.WOuOk2e1taQ), but Word/OpenOffice template is obviously also ok.
- **Put your names, group number and your student IDs in the report!** (as author)
- Clearly identify who is **person A** and who is **person B**!
- **Report page limit:** Maximum **12 pages. Focus on the key aspects!**
- **Use graphs** to visualize findings. Do not just print graphs, also **describe** what they mean.
- **Use tables** to combine findings and other information for maximum overview whenever possible. Describe what you show and explain the data. Clarify, don't mystify.
- Consider issues of **reproducibility**: ensure you provide sufficient information allowing others to re-produce your experiments.
- **Enumerate and label ALL figures, equations and tables** and refer to them in the report --- describe, explain and integrate them with the text. It must be clear to the reader what information can be learned from them.
- **Submit Sections 1, 2 and 3** as the **interim submission** by Tue, **19. 12. 2023**
- **Submit the entire report** (including the previously submitted Section 1-3) by Tue, **23. 1. 2024**

General advice:

- Reserve plenty of **time for "playing" with the data** and start early.
- **Collaboration between groups** is welcome, **but** ensure your group uses a **unique data set.**
- **Collaboration inside the group**: Try to perform at least part of the tasks within the group together. Specifically, discuss the results amongst each other. Subdividing and **solving tasks alone will cost you more time and not meet the goals of the exercise.** Specifically, we discourage completely splitting the assignment into sub-parts distributed across group members. Collaborate, brainstorm and discuss what you find. In an eventual review meeting, **every group member has to demonstrate knowledge of each aspect of the work and the steps taken**.
- Make sure the **structure of the report** follows the **structure of the tasks** provided here.