188.429 Business Intelligence (VU 4,0) – WS 2025

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 we recommend to use the Jupyter Notebook template provided to conduct the **Data Mining and Machine Learning tasks** below and (semi-)automatically documenting the provenance in the graph database provided.

- Note that documentation (and grading) of your experiments is predominantly performed based on the provenance information and descriptions thereof that you provide as integral part of your notebook in PROV-O and automatically logged in the provenance knowledge graph for this course. (Note that the focus in grading will not be on the correctness of the usage of the ontologies. Similarly, if you fail to find a specific concept in an ontology to model information it is ok to use a free-form field (*rdfs:comment*) instead. You should aim for as much machine-actionability as reasonably possible, but it should not turn into a exercise on ontology exegesis.)
- In addition to the provenance information collected you will have to submit a short **report** based on the structure in this assignment paper which summarizes the key aspects of each processing step, and which can be (semi-)automatically created from your provenance logs in the graph database (if you want to) by exporting the respective information into e.g. a LaTeX document, augmented with any additional observations and discussions, specifically interpretations of results if they are not logged in the knowledge graph.

(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 provide this information in the jupyter notebook, creating the root of your provenance documentation. 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. When documenting your activites from within the notebook, be aware that there are two types of roles defined: code-writer und code-executor:

the former ist usually set once to the person who is responsible for the code in the respective cell (and hence usually will not change once the code has been written), the latter should always be set to the person actually executing the notebook cells at that very moment. Hence, make sure that whenever you start executing cells that you set your own student id as the code-executor so that the activities are accredited to you and the provenance is correctly captured.

Use PROV-O to automatically document all experiments / runs of individual cells in your notebook in a **Knowledge Graph (KG)** via the infrastructure provided, relying on suitable ontologies wherever possible. Specifically, we recommend using the ontologies listed below. If you find that some information you want to represent cannot be properly represented by these, feel free to use other ontologies that you are aware of, or use controlled vocabularies or free-form text to represent this information. Recommended ontologies include:

- Provenance
  - PROV-O:
    - doc: https://www.w3.org/TR/prov-o/
    - serialization: https://www.w3.org/ns/prov-o
- Data:
  - schema.org:
    - doc: https://schema.org/Dataset
    - serialization: https://schema.org/version/latest/schemaorg-current-https.ttl
  - Crossaint
    - doc: https://docs.mlcommons.org/croissant/docs/croissant-spec.html
    - serialization: https://github.com/mlcommons/croissant/blob/main/docs/croissant.ttl
  - Units of Measurement:
    - SI Digital Framework (preferred!)
      - doc: https://github.com/TheBIPM/SI_Digital_Framework/blob/main/SI_Reference_Point/docs/README.md
      - doc: https://si-digital-framework.org/
      - doc: https://si-digital-framework.org/SI
      - serialization: https://github.com/TheBIPM/SI_Digital_Framework/blob/main/SI_Reference_Point/TTL/si.ttl
    - Quantities and Units (if not covered by SI Digital Framework)
      - doc: https://www.omg.org/spec/Commons
      - serialization: https://www.omg.org/spec/Commons/QuantitiesAndUnits.ttl
- ML Experiments:
  - MLSO:
    - doc: https://github.com/dtai-kg/MLSO
    - doc: https://dtai-kg.github.io/MLSO/#http://w3id.org/
    - serialization: https://dtai-kg.github.io/MLSO/ontology.ttl
- Terminology:
  - ISO22989: Artificial intelligence concepts and terminology (use as controlled vocabulary)

Assignment 3: Data Analytics

See end of this assignment sheet for formatting and submission information. The information should be automatically logged from within your notebook. It must include all runs of your notebook cells, i.e. it also includes the failed iterations, exploration of parameter spaces, attempts at re-running, with design decisions and justifications for (re-)runs being provided as additional fields in the provenance documentation using *rdfs:comment*. Note that you should aim to document your analysis as far as possible in a machine-interpretable manner, with natural language text being limited largely to actual descriptions and interpretations of, e.g. the business goals, ethical constraints – while success criteria may already be specified in structured form. Specifically, justifications for any decision taken will usually be documented using *rdfs:comment* fields. It should cover 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:

(1) Business Understanding (Responsible: A + B jointly, part of interim submission) (Note: these fields will mostly be logged "manually" into the knowledge graph by providing the according documentation / justifications, rationale etc. as textual input in *rdfs:comment* into the provenance tree)

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)
(Items a) to d) will consist of automatically logged provenance information based on the execution of the code, complemented by the interpretation of the results logged as *rdfs:comment*; items e) to g) are manually logged into the knowledge graph in dedicated cells solely as *rdfs:comment*)

a. **Attribute types**, units of measurement, and the semantics of attributes,
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 contains 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)
(Item a) will consist of provenance information based on the execution of the code, complemented by the interpretation of the results logged as *rdfs:comment*; items b) to d) are manually logged into the knowledge graph in dedicated cells solely as *rdfs:comment*)

a. Perform **necessary pre-processing actions** based on the results of the Data Understanding phase and document these at a level of detail that ensures reproducibility of changes to the data.
b. Describe other **pre-processing steps** considered but not applied due to which reason. (e.g. data cleansing, transformations, binning, scaling, outlier removal, attribute removal, transcoding, …).

c. 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)
d. 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)

(4) Modeling (Responsible: A)

(Item a) is manually logged into the knowledge graph in dedicated cells solely as *rdfs:comment*; Item b) to d) will consist of provenance information based on the execution of the code, complemented by the interpretation of the results logged as *rdfs:comment*.

a. Identify suitable **data mining algorithms** and select one of these as the most suitable for your experiments, providing a **justification** for the selection as part of the provenance information.
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** for its selection and the tuning performed (e.g. interval step-width, autotuning, impact on compute effort required).
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, and ensuring that this splitting is reproducible.
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.
f. **Select the most suitable model** based on the performance on the validation data and document the decision.
g. **Re-train** the model with identical hyper-parameters using the **full train and validation data** as the final model.

(5) Evaluation (Responsible: B)

(Items a), b) and e) will consist of provenance information based on the execution of the code, complemented by the interpretation of the results logged as *rdfs:comment*; Item b) and d) are manually logged into the knowledge graph in dedicated cells solely as *rdfs:comment*.

a. Apply the final model on the test data, documenting and reflecting on the performance.
b. 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. expected **base-line performance** of a trivial acceptor / rejecter or random classifier
c. **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…)

d. **Compare the performance obtained with the success criteria defined in the Business Understanding phase**.

e. **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)**

Most items are manually logged into the knowledge graph in dedicated cells solely as *rdfs:comment*. For a) you should link to the respective business Objectives and Success Criteria identified in the Business Understanding Phase, similarly for b).

a. Compare the performance obtained with the **Business Success Criteria** specified and identify in how far the **Business Objectives** are met (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 document potential ethical aspects as well as impact assessment / risks identified in deployment, linking to statements documented as part of answering questions 1.f) and 2.e)

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 / the provenance ontology.

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

Usually, you will not need to log this into the provenance graph. Add this only in the final report. You will be able to create a large part of this report automatically via the documentation provided in the provenance graph. Once you have generated the according LaTeX output (c.f. template notebook provided) you can save the file and manually edit it using any eidotr or upload it into Overleaf (*https://overleaf.com*), using your TUWIEN account via the single-sign on (SSO) option to continue working on it collaboratively. Alternatively, you can, of course, also simply export plain text output and use any other word processing software of your choice to prepare the final report.

a. Add comments, explanations beyond the comments added into and retrieved from the provenance graph.

b. Briefly summarize your overall findings and lessons learned

c. (**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 2025

Assignment 3: Data Analytics

Submission guidelines:

- **Upload ONE [zip/tgz/rar] file** to TUWEL that **contains (1) your report as a PDF file, 2) your notebook, and (3) any auxiliary files needed for reproducing your experiments.** You **must follow this naming convention**:
    - BI2025_gr<groupno>_<Matnr.1>_<Matnr.2>.zip
    - Example: A submission of group 5 with 2 students (ids: 00059999, 00039999) looks like this: `BI2025_gr05_00059999_00039999.[zip/tgz/rar]`
    - Example: A submission of a single student (with group no. 99) (id: 00987654) looks like this: `BI2025_gr99_00987654.[zip/tgz/rar]`
    - Apply the same naming convention to the report (but obviously with pdf extension)
- The notebook you provided should be uploaded (ideally: created and worked on) using a code repository (such as TU Gitlab, *https://gitlab.tuwien.ac.at*), with an appropriate license (e.g. GLP for code, c.f. *https://choosealicense.com/* or CC-BY for artifacts, images, etc.), and referred to from the structured description as part of the provenance information via a persistent identified (PID, e.g. DOI, URI) using the Prov-O ontology.
- **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/latex/templates/acm-conference-proceedings-primary-article-template/wbvnghjbzwpc*), 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**!
- **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.
- Extensive tables and auxiliary figures may be included in an **appendix** – but make sure the key information (i.e. core parts of tables that have an impact on the interpretation) are provided in the core part of the report.
- 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 **14. 12. 2025**
- **Submit the entire report** (including the previously submitted Section 1-3) by **18. 1. 2026**

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.

Assignment 3: Data Analytics

FAQ:

- **“Do we have to write a report as a text document?”**
  - You should try to automate the report generation as much as possible, relying on the information documented as part of the provenance documentation. However, comparative discussions, elaboration and summarization of insights generated across experiment runs etc. need not be documented as part of the provenance (though you may decide to do so if you want). These descriptive texts you can also add to the text document created from the provenance graph.
- **“Do we need extensive knowledge of ontologies?”**
  - No. The focus of this exercise is to make documentation machine-readable as far as reasonably possible. If you are unable to find a fitting ontology field, use free-form fields instead.
- **“Do we need some specific setup for this to work?”**
  - No. The template notebook provides everything necessary. You do not need to run a local instance of e.g. GraphDB; all logging will happen on the TU Starvers Server.
- **“My machine learning algorithm did not perform well – will I get point reductions?”**
  - No. As long as you can argue for the decisions you made regarding the choice of algorithms, parameter-tuning and data preparation, we will not deduct points for not meeting a certain benchmark.
- **“What is code_executor and code_writer about?”**
  - These roles are used to document **who wrote** a piece of code and **who executed** it in the notebook. In many groups this will often be the same person, but in principle they can differ (e.g., one student implements a cell, another one runs the experiments). The purpose is to capture realistic provenance information about responsibilities in the workflow, not to grade you on how perfectly you distinguish these roles. If you clearly indicate, for the key cells, who is responsible for writing and executing the code, that is sufficient.
- **“How do we handle licensing for our code and artifacts?”**
  - For this assignment you are not expected to do a deep legal analysis, but you are expected to make a conscious, documented choice. In practice, this means that you pick an appropriate open license for your work (e.g. one for your code, one for figures/text), add it to your code repository in a standard way (typically via a LICENSE file and/or a short note in the README), and then reference that choice in your provenance graph. Concretely, you can treat your repository or artifact as an entity in the knowledge graph and attach the license information (and, ideally, the repository URL) to it so that it is clear under which terms others may reuse your work.