


| | | |
|---|-------------|---------------------------|
|  | Assessment: | Project Milestone 3 |
| | Subject: | Business Intelligence 381 |
| | Total: | 50 |

Project Milestone 3

Introduction

This project will investigate various health and demographic datasets to identify meaningful patterns and trends. The details of the project requirements are described in the **Project Outline** document. The project outline underscores the use of CRISP-DM as a structured approach to guide you through the data science project, emphasizing its importance in the context of real-world data analysis and modelling.

Outline

The project is divided into six (6) milestones, where at each stage some deliverables are to be produced in terms of a series of reports that describe the project plan, and the work carried out during the iterative process of data preparation, modelling, evaluation and eventual deployment.

The CRISP-DM methodology breaks down a data mining project life cycle into six phases with each phase consisting of many secondary tasks. The focus of this milestone is on the data modelling phase of CRISP-DM.

Data Modelling

Data modelling is the fourth phase of the CRISP-DM methodology and follows the data preparation phase. The goal of this stage is to select a suitable model that solves the business problem and meets the objectives defined in the Business Understanding phase, design procedures for testing and evaluating the model's quality, build the model using the dataset prepared in the Data Preparation phase and assess the model from both the business problem domain and technical perspective.


The modelling phase, much like the rest of the data science project phases, is iterative and you are likely to experiment with different algorithms and techniques to find the most suitable model. Your objective at this stage is to develop the data mining models that solve the given business problem and aligns with the pre-defined business criteria. You are encouraged to gradually explore, refine and extend your models to uncover new insights, improve accuracy, explore various options and better align with real-world organisational needs. Describe all the tasks or iterations leading to the final model(s) and their performance in relation to all other models experimented with. A well-developed version of this project can be a valuable addition to your portfolio, GitHub repository, or professional profile.

Tasks

This milestone consists of the following tasks that should be submitted as an assignment on given due date.

Select Modelling Technique

- Select the modelling technique (e.g., logistic regression, decision tree with C4.5, random forest, etc.).
- Justify the choice of algorithm(s) based on the data characteristics.

| | | |
|---|-------------|---------------------------|
|  | Assessment: | Project Milestone 3 |
| | Subject: | Business Intelligence 381 |
| | Total: | 50 |

- Document the actual modelling techniques that are used.
- Describe the assumptions. Some techniques are based on specific assumptions such as the quality, format, and distribution of data. Make sure that these assumptions hold and go back to the Data Preparation Phase if necessary.

Generate Test Design

- Describe the procedures for training, testing, and evaluating the models. For example, supervised data mining tasks such as classification generally to use error rates as quality measures for data mining models.
- Specify how the dataset should be separated into training data, test data, and validation sets. The model is built on the training set and its quality estimated on the test set.

Build Model

- Develop the model(s) using the R programming language.
- Run the selected technique using the prepared dataset to create one or more models.
- Document the model parameters. With any modelling tool, there are often several parameters that can be adjusted. Describe the parameters and their chosen values, along with the rationale for the choice.
- Describe the model's behaviour and interpretation in terms of accuracy, robustness, possible shortcomings, etc.


Assess the Model

- Summarise the results of generated model(s)
- Assess the model performance metrics in terms of graphs, confusion matrices and other statistical measures.
- Interpret the model based on domain knowledge, predefined success criteria, and test design.

IMPORTANT: Any plagiarised work will receive a zero (0) for the project and disciplinary action will be taken.

Deliverables:

- Modelling Report (in PDF format).
- Complete code (R and R Markdown files) used for data preparation, analysis, modelling and evaluation.
- Any other code or files if other tools were used during this phase of the project

| | | |
|---|-------------|---------------------------|
|  | Assessment: | Project Milestone 3 |
| | Subject: | Business Intelligence 381 |
| | Total: | 50 |

Grading Criteria:

| Criteria | Weight |
|---------------------|-----------|
| Introduction | 5 |
| Modelling Technique | 10 |
| Test Design | 10 |
| Model Description | 10 |
| Model Assessment | 10 |
| Conclusion | 5 |
| TOTAL | 50 |

Additional Information

- This is a group assignment. Please continue with your project groups. Remember that a group must have at least 2 members but may not exceed four people.
- All work must be original. Copying another group's work or using any work available on other repositories will not be tolerated.
- Includes names of all group members on the Cover Page.
- Submit your project electronically on Moodle (BC Connect) before the due date.
- All writing must be correctly cited and referenced.
- **Plagiarism is a serious offence.** Belgium Campus uses software that can scan for plagiarism and a student caught doing this will get 0 for this assignment.
- No mark will be awarded if the assignment is not uploaded via BC Connect.
- Late assignments will not be accepted; missing the deadline is an automatic 0.