BELGIUM CAMPUS iTversity It's the way we're wired	**Assessment:**	**Project Milestone 2**
	Subject:	**Business Intelligence 381**
	Total:	**50**

Project Milestone 2

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, the work carried out during the iterative process of data preparation, modelling, and evaluation.

While CRISP-DM 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 third phase.

Data Preparation

Start off with the initial data collection and proceed with activities that enable you to become familiar with the data, identifying data quality problems, discovering first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

Tasks

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

Select Data:

- Decide on the data to be used for analysis.
- Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.

Verify Data Quality:

- Perform significance and correlation tests to decide if fields should be included.
- Reconsider data selection criteria considering experiences of data quality and data exploration (i.e., you may wish include or exclude other sets of data)
- Reconsider data selection criteria considering your modelling requirements (i.e., model assessment may show that other datasets are needed)
- Based on data selection criteria, decide if one or more attributes are more important than others and weight the attributes accordingly. Decide, based on the context (i.e., application, tool, etc.), how to handle the weighting.
- Select relevant data subsets (e.g., significant attributes, and only data which meet certain conditions or using other advanced data reduction techniques such Principal Component Analysis)
- Document the rationale for inclusion or exclusion.

Clean Data:

Raise the data quality to the level required by the selected analysis techniques. This may involve the selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modelling.

- Clean and preprocess the data to address quality issues (e.g., impute missing values, remove duplicates, handle outliers).
- Reconsider how to deal with any observed type of noise.
- Correct, remove, or ignore noise.
- Decide how to deal with special values and their meaning. The area of special values can give rise to many strange results and should be carefully examined.
- Reconsider Data Selection Criteria considering experiences of data cleaning (i.e., you may wish to include or exclude other sets of data)

Remember that some fields may be irrelevant to the data mining goals and, therefore, noise in those fields has no significance. However, if noise is ignored for these reasons, it should be fully documented as the circumstances may change later.

Prepare data for modelling:

- Transform and encode categorical variables as needed.
- Discretize or scale numerical variables if necessary.
- Check available techniques for sampling data.
- Consider the use of sampling techniques (e.g., A quick solution may involve splitting test and training datasets or reducing the size of the test dataset, if the tool cannot handle the full dataset. It may also be useful to have weighted samples to give different importance to different attributes or different values of the same attribute.)
- Split the data into training and testing sets for in preparation for the modelling phase.

Describe the decisions and actions that were taken to address the data quality problems reported during the Verify Data Quality Task in **Project Milestone 1**. If the data are to be used in the data mining exercise, the report should address outstanding data quality issues and what possible effect this could have on the results. Remember that visualizations are also important to effectively communicate the data preparation tasks and to verify the accuracy and consistency of the data.

IMPORTANT NOTE: *This project encourages originality and recognizes that there may not be a single correct answer to every aspect of the problem. Even though the emphasis is on CRISP-DM, applying the same methodology to the same business problem does not necessarily yield the same outcomes. The key is to make informed and sound decisions that align with the expectations of a data science project. The success of this project will be judged not only by the final outcome but also by the rigour, creativity, and thoughtfulness applied to the various stages of the project. A well-documented and well-justified project demonstrates a thorough understanding of data science principles and practices, which is a key objective of this project.* ***Any plagiarised work will receive a zero (0) for the project and disciplinary action will be taken.***

Deliverables:

- Data Preparation Report (in PDF or document format).
- Code (R, and R Markdown) used for data analysis and modelling.
- Power BI project file if Power BI was used during this stage of the project.

Grading Criteria:

Criteria	Weight
Data Description	5
Data Selection (Inclusion/Exclusion Criteria)	10
Data Cleaning Process	15
Attribute/Feature Selection	10
Data Transformations and Aggregation	10
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 must 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.