Assignment Brief

Please read this coursework brief very carefully in order to do the tasks. You will need to submit a report of **4000 words**. This is an individual assignment and carries **100%** of the overall module marks. The assignment consists of the following components:

- Supervised and Unsupervised Learning (60%)
- Text Mining (40%)

Overview

In the following coursework, you will apply supervised and unsupervised learning and text mining techniques on different datasets which reflect some examples of real-world problems. In addition, you will have to reflect on what you have learnt and apply the knowledge you have gained to accomplish several tasks. You will need to produce a report of 4000 words documenting these tasks.

Task I: Supervised and Unsupervised Learning (60%)

1. Download the following dataset to use for the tasks below:

Classification	Heart failure dataset Heart failure dataset information
Regression	Fertility dataset Fertility dataset information
Clustering	Absenteeism at work dataset Absenteeism dataset information
Association Rules	Grocery dataset Grocery dataset information

2. You will need to accomplish the following:

- Perform **classification** using 3 algorithms, compare and discuss the results of the chosen algorithms.
- Perform **regression** using 2 algorithms, compare and discuss the results of the chosen algorithms.
- Perform **clustering** using 2 algorithms, compare and discuss the results of the chosen algorithms.
- Perform **association rule mining** using 1 algorithm and discuss the results (rules generated).

3. For the tasks above, it is also required to do the following:

- Use simple descriptive analytics to analyse the data (e.g. all attributes distribution, outliers).
- Use data exploratory techniques (e.g. three visualisations) to explore the dataset and analyse the results.
- Analyse the results with regards to the dataset properties.

Task 2: Text Mining (40%)

For this task, download Sentiment Labelled Sentences Dataset - Information .

Then select and download one of the data available (Amazon, IMDb, or Yelp).

- Amazon
- IMDb
- Yelp

You will need to accomplish the following:

1. Preprocess the textual data to transform it into a structured format.

2: Explain the changes to the data for each preprocessing method applied.

3: Use descriptive analytics for feature description of the dataset.

4: Perform clustering using 1 algorithm and discuss the results.

5: Perform classification using 2 algorithms, compare and discuss the results.

Submission

For this coursework, you need to submit the following:

- **A report** (*.pdf or *.word formats) documenting Task I and Task II; the report should include up to **4000 words** overall, excluding tables and figures. The report should only include relevant information about the results of the techniques or algorithms.

  Note: there is no hard requirement of words per section, the only hard requirement is the total of 4000 words for both Task I and Task II.

- **A zip file** with the KNIME processes used (or equivalent from other software). Upload the file to your Google Drive and include the link at the end of the report. To ensure your tutor can access the file, see **these instructions** for how to make a file viewable by anyone with the link.