Module: CMP-7023B - Data Mining
Assignment: Advanced Data Analysis Exercise

Set by: Kazhan Misri <k.misri@uea.ac.uk>
Date set: Week 6
Value: 65%

Date due: Wednesday 07/05/2025 3 pm [week 11]
Returned by: Assessment Period Week 3
Submission: Blackboard (A Turnitin point will be provided in Blackboard)
Checked by: Prof. Beatriz de la Iglesia

Learning outcomes

- Competence in using KDD software tools in medium to large databases.
- Competence in applying relevant techniques at each stage of the KDD process.
- Ability to evaluate the suitability of software tools in the context of different data analysis tasks.
- Competence in combining data manipulation and analysis approaches to improve the quality of input data.
- Understanding and identification of problems in input data such as outliers, missing data, unreliable data, differences in granularity, and others, and identifying an adequate strategy to deal with the problem data.
- Presentation of knowledge induced in a format suitable for the target audience and for the particular application.

Specification

Overview

Aim:

- To obtain an overall view of the complex process of Knowledge Discovery in Databases and understand the need for a methodical approach to KDD.
- To explore tools and algorithms available at each stage of the KDD process.
- To gain experience in using KDD software tools in a medium-sized database.
- To learn to combine data manipulation and analysis approaches to improve the quality of input data.
- To produce a suitable report describing the methods applied and the discussion of the findings.

Coursework Description

See attached at page 4 and 5.

Relationship to formative assessment

Lecture slides, lab exercises and links to resources provide the baseline to design experiments with data mining techniques applied to real data.

Deliverables and Submission Instructions

Document Format:
Your report should be typed. We recommend using LaTeX for efficient reference and cross-reference management (LaTex Template available on Blackboard). However, if you choose not to use LaTeX, please use Microsoft word ensure the use of a 12-point standard font like Arial or Comic Sans with a standard page layout, including margins.

- The report must collate all responses to the provided questions into a coherent and structured document.
- It should follow the structure outlined in the marking scheme and must **not exceed 12 pages** (excluding references and appendices).
- Include a clear and concise summary page outlining key findings for the insurance company's executive team (this page counts towards the 12-page limit).
- Append the cleaned code/notebook, produced to accomplish your tasks, in an appendix to the report.

Submission Instructions:
- Submit the report with the attached Jupyter notebook as a **single PDF file** to the Turnitin submission point on Blackboard **{002} {CMP-7023B-24-SEM2-B}**.
- Ensure you have successfully submitted the correct report to the Turnitin point and **received proof of submission.**

Resources

You can use the weekly lab documentations, lecture notes, library resources and other sources to accomplish your tasks. **Do not forget to cite** any external and online resources used. **Students are expected to work independently**, and any plagiarism or collusion will be penalised.

Marking scheme

Assessment criteria marks distributed as follows:	**Marks Approx.**
1. **Data exploration, dictionary, visualisation, and summary**	10%
2. **Data Cleaning, feature engineering and pre-processing**	15%

3. **Supervised model training, tuning and evaluation** including result interpretation and comparative analysis	30%
4. **Unsupervised learning using clustering algorithms**, including result interpretation, visualisation and comparative analysis	20%
5. **Overall presentation** including proper referencing, cross-referencing, interpretation, comparative analysis of models, well-documented Jupyter Notebook code, and a concise summary of key findings in the report.	25%
Total	**100%**

Plagiarism, collusion, and contract cheating

The University takes academic integrity very seriously. You must not commit plagiarism, collusion, or contract cheating in your submitted work. Our Policy on Plagiarism, Collusion, and Contract Cheating explains:

- what is meant by the terms 'plagiarism', 'collusion', and 'contract cheating'
- how to avoid plagiarism, collusion, and contract cheating
- using a proof reader
- what will happen if we suspect that you have breached the policy.

It is essential that you read this policy and you undertake (or refresh your memory of) our school's training on this. You can find the policy and related guidance here: https://my.uea.ac.uk/departments/learning-and-teaching/students/academic-cycle/regulations-and-discipline/plagiarism-awareness

The policy allows us to make some rules specific to this assessment. Note that:

> In this assessment, working with others is *not* permitted. All aspects of your submission, including but not limited to: research, design, development and writing, must be your own work according to your own understanding of topics. Please pay careful attention to the definitions of contract cheating, plagiarism and collusion in the policy and ask your module organiser if you are unsure about anything.

CMP-7023B - Data Mining
Second Assessed Exercise
Data Mining the Insurance Dataset

Date due: Wednesday 07/05/2025 3 pm [week 11]
Value: **65%**

Exercise Description

Understanding customer profiles is crucial in the insurance industry for tailoring services and predicting customer behaviour. In this coursework, you will analyse an insurance dataset to classify and segment customers based on various attributes.

The dataset '**Insurance_Data.csv**' contains 5,521 observations and 83 variables, including the target variable **Customer_Type**. The dataset includes information such as household size, education level, income, social class, and insurance contributions across different policy types (e.g. car, life, disability, and property insurance).

Your primary task is to classify customers into relevant subtypes and explore natural groupings using clustering techniques. You will apply supervised learning for classification and unsupervised learning for segmentation. Finally, you will interpret and report your findings.

To accomplish your task, you need to perform the following operations:

1. **Data Exploration, Visualisations, and Summary:**
   - Download the dataset, prepare a summary of features, including data type (numerical/categorical), and assess the amount of missing data and outliers in individual features. Conduct initial exploration with visualisation and statistical analysis of the features.
   - Provide a clear introduction to the dataset in your report, including relevant insights from visual analysis.

2. **Data Cleansing , Feature Engineering, and Pre-processing:**
   - Undertake any **cleansing or pre-processing** you think is necessary on the dataset**.**
   - **In your report, explain clearly what you have done and why you have done it.** Some cleaning could be to remove any feature/column if 60% values are missing, constant, or to remove duplicate and highly correlated information.

3. **Supervised Model Training, Tuning, and Evaluation:**
   - Split the data into a training set and a validation set once cleansing is done.
   - Use suitable toolkits and libraries (e.g. k-NN, Decision Tree, SVM, or more sophisticated models like Ensembles, ANN or deep learning) to train models on the training set to predict **Customer_Type**. Make sure to perform hyperparameter tuning to optimise model performance.
   - Experiment with **balancing the data, feature selection**, and other **adjustments/tuning** to enhance model quality.
   - Evaluate the performance of the models on the test/validation set.

- Use appropriate tools to clearly illustrate and identify the features that were deemed most crucial or had the most significant impact on decision-making within the best-performing model.
- **Present comparative analysis and interpretation of results:**
  - In your final report, describe and justify the decisions made during data processing.
  - Present and discuss the results, including model validation/evaluation techniques.
  - Discuss the effectiveness of the best model using metrics such as confusion matrices, ROC curve, precision, and recall performance in the context of a multi-class problem.
  - Ensure you include comparative analyses for your trained classifiers and if you have multiple pre-processing techniques for each classifier.

4. Unsupervised Learning Using Clustering Algorithms:

- Exclude the Customer_Type (target) field during clustering to ensure unsupervised learning.
- Apply unsupervised clustering algorithms (e.g. k-Means, hierarchical clustering, …) to the pre-processed dataset based on your judgment of the task.
- Use Scatter plots or t-SNE plots on the clusters to visually represent the data.
- Analyse the formed clusters, observing distinct patterns or groups related to different categories of Customer_Type.
- Discuss the effectiveness of the clustering models.
- Present comparative analysis and interpretation of results.
- Compare the clustering and classification outcomes and discuss your observations, i.e. do the clustering results align with the classification outcomes?