WIE3007 2023/2024 S1

Tutorial 5

Q1: Briefly describe each step of the SEMMA methodology. Why is each step crucial in the data mining process?

- Sample
  - Brief description: Representative subset of a data from the entire dataset
  - Importance: Sampling helps reduce the volume of data, making it manageable and efficient to analyse, without losing the essential characteristics of the original data
- Explore
  - Investigate the sampled data to identify patterns, relationships, anomalies, and trends.
  - Importance: Exploration provides a deeper understanding of the data, allowing to uncover underlying characteristics and structures.
- Modify
  - Process and transform the data to prepare it for modelling
  - Importance: Modification ensures that the data is clean, properly formatted and relevant so that it perform better accuracy and effectiveness of the subsequent modelling step.
- Model
  - Develop a predictive or descriptive model to identify patterns, relationships or to make forecasts.
  - Importance: Modelling allows the extraction of knowledge from data, allowing for predictions, classifications, and insights generation.
- Assess
  - Evaluate the performance and validity of the developed models.
  - Assessment ensures that the models are accurate, reliable and generalizable to unseen data, confirming their utility and robustness.

Q2: You have a dataset of 1 million online transactions from an e-commerce website. Using the Sample step of SEMMA, outline a strategy to select a representative subset of this data for analysis. What sampling techniques would you consider, and why?

In the Sample step of the SEMMA framework, the goal is to select a representative subset of the dataset for analysis. Several sampling techniques can be considered, each with its own advantages and use cases. The choice of sampling technique should be based on the specific objectives, resources, and characteristics of the dataset. Here are some potential sampling techniques I would consider, along with their justifications:

Random Sampling:

Justification: Random sampling is a straightforward and unbiased method that provides each transaction an equal chance of being included in the sample. This is suitable when there are no specific criteria or factors to consider in the sampling process.

Stratified Sampling

Justification: Stratified sampling divides the dataset into subgroups based on relevant characteristics, such as product categories, customer segments, or transaction types. This approach ensures that each subgroup is adequately represented in the sample, which can be crucial for maintaining proportionality within the data.