Rica Mae Jin C. Calugtong
BSIT 3-5

IT Elective 2

MIDTERM ASSIGNMENT # 1

EXPLORING THE LANDSCAPE OF DATA MINING

1. Data mining and traditional statistical analysis both aim to extract valuable insights from data, but they take different approaches. Data mining explores large, often unstructured datasets and idetifies hidden patterns using algorithms such as clustering and decision trees, whereas traditional statistical analysis employs organized, smaller datasets to test hypothesis and make population conclusions. Data mining is usually predictive and exploratory, whereas traditional statistical analysis is inferential and theory-driven. Both require high-quality data, but data mining frequently deals with more complicated data type and focuses on pattern discovery rather than testing pre-defined hypothesis.

2. In retail, customer segmentation aims to target specific customer groups with personalized offers to improve sales. Data used includes transactional data (ex: purchases, spending) and demographic data (ex: age, income). Techniques like clustering are applied to identify distinct customer segments based on purchasing behavior. Benefits include more effective marketing campaigns, improved customer satisfaction, and increased sales and engagement.

3. Data preparation is crucial beoause raw data often contains inconsistencies, missing values, or noise that can distort model results. Common preprocessing techniques include data cleaning, normalization, and feature engineering. These techniques ensure that data is in a suitable format for analysis, improving the accuracy and effectiveness of the model.

4. Data mining raises significant ethical concerns regarding privacy and security, as it involves collecting and analyzing large amounts of personal data, which can potentially be misused if not handled responsibly, key considerations include obtaining informed consent, anonymizing data, implementing robust security measures, and maintaining transparency with users about data usage to reduce these issues.

5. Bias in data can result in distorted findings, Increasing disparities, particularly in employment and financing. This might happen if the data used to train models is not representative or has historical biases. To reduce bias, it is critical to maintain data diversity, use fair algorithms, and conduct regular fairness tests to verify that outcomes are equitable across demographic groups.

6. A decision tree is a machine learning algorithm that splits data into subsets based on feature values, forming a tree-like structure that leads to a prediction or classification. It is suited for classification and regression tasks, such as predicting customer loss or house prices. Strengths include its interpretability and ability to handle both numerical and categorical data. However, decision trees are prone to overfitting and instability, especially with deep trees, which can be reduced by pruning or limiting tree depth.