EUGENIO, GIAN ORLAND E.

CDT 1101 -ITBD

|  |  |
| --- | --- |
| **Activity Number: 1** | **What is Data Mining?** |
| **Instruction:**  **Answer the following questions briefly (1-2 sentences):**   * 1. **What is data mining?** * It is discovering hidden patterns from already available data. In other words, Extraction of interesting non trivial implicit previously unknown and potentially useful patterns or knowledge from huge amount of data.   1. **What are the different phases to process data mining?** * **Data cleaning** - It can be applied to remove noise and correct inconsistencies in the data. * **Data Integration** - Merges data from multiple sources into a coherent data store, such as a data warehouse. * **Data selection** - Where data relevant to the analysis task are retrieved from the database. * **Data transformation** - Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. * **Data mining** - An essential process where intelligent methods are applied in order to extract data patterns.   1. **What are the different techniques of data mining?** * **Tracking patterns** - is usually a recognition of some aberration in your data happening at regular intervals, or an ebb and flow of a certain variable over time. * **Classification -** is a more complex data mining technique that forces you to collect various attributes together into discernable categories, which you can then use to draw further conclusions, or serve some function. * **Association -** is related to tracking patterns, but is more specific to dependently linked variables. * **Outlier detection** - simply recognizing the overarching pattern can’t give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. * **Clustering -** is very similar to classification, but involves grouping chunks of data together based on their similarities. * **Regression -** used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. * **Prediction -** is one of the most valuable data mining techniques, since it’s used to project the types of data you’ll see in the future.   1. **Where can we use data mining?** * Data mining is used in diverse applications such as banking, marketing, healthcare, telecom industries, and many other areas. Data mining techniques help companies to gain knowledgeable information, increase their profitability by making adjustments in processes and operations.   1. **List and describe five free/open-source data mining applications** * **Xplenty** - provides a platform that has functionalities to integrate, process, and prepare data for analytics. * **Rapid Miner -** is one of the best predictive analysis-system developed by the company with the same name as the Rapid Miner. * **Orange -** is a perfect software suite for machine learning & data mining. It best aids the data visualization and is a component-based software. * **Weka** - is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for data analysis and predictive modeling. It contains algorithms and visualization tools that support machine learning. * **Apache Mahout** is a project developed by Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering.   1. **List and describe five commercial data mining applications** * **Marketing** - is used to explore increasingly large databases and to improve market segmentation. By analyzing the relationships between parameters such as customer age, gender, tastes, etc., it is possible to guess their behavior in order to direct personalized loyalty campaigns. * **Retail** - Supermarkets, for example, use joint purchasing patterns to identify product associations and decide how to place them in the aisles and on the shelves. Data mining also detects which offers are most valued by customers or increase sales at the checkout queue. * **Banking** - Banks use data mining to better understand market risks. It is commonly applied to credit ratings and to intelligent anti-fraud systems to analyze transactions, card transactions, purchasing patterns and customer financial data. * **Medicine** - Data mining enables more accurate diagnostics. Having all of the patient's information, such as medical records, physical examinations, and treatment patterns, allows more effective treatments to be prescribed. * **Television and Radio** - These systems collect and analyze, on the fly, anonymous information from channel views, broadcasts and programming. Data mining allows networks to make personalized recommendations to radio listeners and TV viewers, as well as get to know their interests and activities in real time and better understand their behavior. | |