Name: Kunal Sinha Roll no: 2K17/CO/164 Batch: CO-A3

Experiment 2

Aim: List out various open-source data mining tools and techniques and explain them.

Theory:

Data Mining

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.
Some of the most popular open-source tools used for Data Mining are:

1. **Weka:** Also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for data analysis and predictive modelling. It contains algorithms and visualization tools that support machine learning. Weka has a GUI that facilitates easy access to all its features. It is written in the JAVA programming language. Weka supports major data mining tasks including data mining, processing, visualization, regression etc. It works on the assumption that data is available in the form of a flat-file. Weka can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query.

2. **Rapid Miner:** RapidMiner is one of the best predictive analysis systems developed by the company with the same name as Rapid Miner. It is written in the JAVA programming language. It provides an integrated environment for deep learning, text mining, machine learning & predictive analysis. The tool can be used for a vast range of applications including business applications, commercial applications, training, education, research, application development, machine learning.
   Rapid Miner offers the server as both on-premise & in public/private cloud infrastructures. It has a client/server model as its base. Rapid Miner comes with template-based frameworks that enable speedy delivery with a reduced number of errors (which are quite commonly expected in the manual code writing process). Rapid Miner constitutes of three modules, namely
   - RapidMiner Studio- This module is for workflow design, prototyping, validation etc.
   - RapidMiner Server- To operate predictive data models created in the studio
   - RapidMiner Radoop- Executes processes directly in the Hadoop cluster to simplify predictive analysis.

3. **Orange:** Orange is a perfect software suite for machine learning & data mining. It best aids data visualization and is a component-based software. It has been written in Python computing language. As it is a component-based software, the components of orange are

called ‘widgets’. These widgets range from data visualization & pre-processing to an evaluation of algorithms and predictive modelling. Widgets offer major functionalities like

- Showing data table and allowing to select features
- Reading the data and training predictors and to compare learning algorithms
- Visualizing data elements etc.

Additionally, Orange brings a more interactive and fun vibe to the dull analytic tools. It is quite interesting to operate. Data coming to Orange gets quickly formatted to the desired pattern and it can be easily moved where needed by simply moving/flipping the widgets. Users are quite fascinated by Orange. Orange allows users to make smarter decisions in a short time by quickly comparing & analyzing the data.

4. **KNIME:** KNIME is the best integration platform for data analytics and reporting developed by KNIME.com AG. It operates on the concept of the modular data pipeline. KNIME constitutes various machine learning and data mining components embedded together. KNIME has been used widely for pharmaceutical research. In addition, it performs excellently for customer data analysis, financial data analysis, and business intelligence. KNIME has some brilliant features like quick deployment and scaling efficiency. Users get familiar with KNIME in quite a lesser time and it has made predictive analysis accessible to even naive users. KNIME utilizes the assembly of nodes to pre-process the data for analytics and visualization.

5. **Apache Mahout:** Apache Mahout is a project developed by the Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering. Mahout is written in JAVA and includes JAVA libraries to perform mathematical operations like linear algebra and statistics. Mahout is growing continuously as the algorithms implemented inside Apache Mahout are continuously growing. The algorithms of Mahout have implemented a level above Hadoop through mapping/reducing templates. To key up, Mahout has the following major features
   - Extensible programming environment
   - Pre-made algorithms and math experimentation environment
   - GPU computes for performance improvement.

6. **DataMelt:** DataMelt, also known as DMelt is a computation and visualization environment that provides an interactive framework to do data analysis and visualization. It is designed mainly for engineers, scientists & students. DMelt is written in JAVA and is a multi-platform utility. It can run on any operating system which is compatible with JVM. It contains Scientific & mathematical libraries. Scientific libraries: To draw 2D/3D plots. Mathematical libraries: To generate random numbers, curve fitting, algorithms etc.DataMelt can be used for the analysis of large data volumes, data mining, and stat analysis. It is widely used in the analysis of financial markets, natural sciences & engineering.

There are several major data mining techniques that have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree.

1. **Association:** Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using the association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for the customer and increase sales.

2. **Classification:** Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. The classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. In classification, we develop software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

3. **Regression:** Regression, used primarily as a form of planning and modelling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

4. **Clustering:** Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label them with a meaningful name. If readers want to grab books on that topic, they would only have to go to that shelf instead of looking for the entire library.

5. **Prediction** The prediction, as its name implied, is a data mining technique that discovers the relationship between independent variables and the relationship between dependent and

independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

6. **Sequential Patterns**: Sequential patterns analysis is one of the data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period. In sales, with historical transaction data, businesses can identify a set of items that customers buy together at different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

7. **Decision trees:** A decision tree is one of the most commonly used data mining techniques because its model is easy to understand for users. In this technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

Findings and Learnings :

We learned about data mining and its respective significance. We also learned about the open-source tools which can be used to assist us in the practice of data warehousing and data mining. Different techniques are available for data mining depending on the type of data and the amount of data available. Finally we learned about the features of the tools used for data mining. They also vary a lot in the way they approach data mining and the various algorithms they use.