KULLIYYAH OF INFORMATION & COMMUNICATION TECHNOLOGY

INFO 4313 DATA MINING

INDIVIDUAL ASSIGNMENT

SEMESTER 3, 2021/2022

LECTURER

ABDUL RAFIEZ BIN ABDUL RAZIFF

PREPARED BY

JOYADDAR MD JOBAYER 1731833

DATA MINING TOOLS AND TECHNIQUES:

INTRODUCTION:

Data mining, the process of sifting through massive databases to find hidden predictive information, is an exciting new development that has the potential to help businesses zero in on the data that will have the most impact on their operations and storage facilities for large amounts of data. Statistics Have Been Collected programme for mining, web mining, and gaining insight programmes like the RapidMiner Toolkit and the WEKA Toolkit. Throughout the book, the author includes examination of WEKA, RapidMiner and Net Tools spider tools KNIME and Orange. Several resources exist for data mining and web mining. Due of this, it is crucial that people understand the scientifically examining the efficacy of these instruments. The analysis demonstrates several advantages of these data mining instruments, together with the qualities and characteristics of current tools.

The data-mining process entails the following steps:

- Data Cleaning: The first phase of data cleaning is the removal of any invalid or blank records.
- Data Integration: Data must be gathered and integrated into a single structured structure before data mining can continue.
- Data Selection: Not all the information gathered has to be used. The ability to selectively remove non-essential data makes data selection a powerful tool.
- Data Transformation: Even once the cleaning process is complete, the data is not yet ready for data mining since it must be translated into a format that is understood by the algorithm.
- Data Mining: This process involves applying multiple algorithms to the data in search of previously unknown insights.
- Evaluation of Patterns: It is necessary to assess the significance of data mining results.

DATA MINING TOOLS:

The WEKA Tool:

The Waikato Environment for Knowledge Analysis (WEKA) is a data mining and machine learning platform created at the University of Hawaii at Manoa. University of Waikato, Department of Computing. It's a repository for several free data-mining tools. one of Weka's strengths is its compatibility with a wide variety of machine learning algorithms and its capability to receive data from a variety of sources, users may simply feed data into the programme, even if it isn't in a standard format that can be read by other data mining software. statistics, classifying, regressing, grouping, connecting the process of determining which rules to use and which features to use. It supports . arff (attribute relation file format) file format.

WEKA-API USE:

It has been observed that Weka's API capabilities allow its customers to accomplish more performance because there are so many open-source options Software development kits that can be downloaded from the internet. The additional
It's possible to accomplish more than a hundred distinct tasks using this software. data mining techniques, such as the Bayesian approach, rule-based procedures, and statistical analysis.

RapidMiner Software:

Using RapidMiner, we can perform machine learning and data
the steps of learning, such as importing new data and executing
data extraction, processing, and loading (also known as "ETL")
processing and display, modelling and analysis, assessment and
deployment. Java is the language used in the development of RapidMiner. language. It employs evaluators of qualities and learning techniques.

System Support for the RapidMiner Database:

Furthermore, RapidMiner supports the vast majority of database systems, allowing users to bring in data from a wide range of database systems for further examination and analysis. Similar to other data mining programmes, SQL queries provide the backbone of the database operations. Since SQL queries serve as the foundation for database support, it's possible that there are constraints on data import and database customization. Consequently, we'll need at least rudimentary programming skills to successfully convert database files by adding, removing, or modifying rows and columns of data.

RapidMiner-Visualization

RapidMiner offers extensive help with data and analytical visualisation. In-depth data analysis results can be crafted using the programme. Colourful and aesthetically pleasing visualisations of nodes and other information are possible.

ASSOCIATION METHOD:

One of the most used data mining methods is called "association," and it finds patterns by examining the connections between different parts of the same transaction. It's called a relation technique because it looks at the connections between things to find out which ones show up most often in a dataset. Association rules employ if-then statements to illustrate the likeliness of associations between data items or variables inside big data sets in different kinds of databases. The usage of association rules is widespread because of their usefulness in a variety of settings, such as the discovery of sales correlations in transactional data or medical datasets Because it aids in deducing shoppers' preferences, association is commonly employed in the retail industry.

CLUSTERING METHOD:

In data mining, clustering is a foundational early approach. Understanding the similarities and differences across datasets can be facilitated through the use of a process called "clustering," which involves the study of one or more attributes to locate data that are similar to each other. In a library, for instance, we can use the clustering technique to group books on the same topic together and label that section so that readers can quickly find the books they need without having to search the entire collection.

PREDICTION METHOD: Prediction covers a wide range of topics, from foreseeing component failures to unravelling fraud to projecting future financial gains for an organisation. Prediction utilises various data mining methods such as trend analysis, classification, pattern matching, and relational analysis. Estimate of future performance based on examination of historical data. In the case of credit card authorization, for instance, a combination of decision tree analysis of previous transactions and categorization historical pattern matches can be used to determine if a transaction is fraudulent.

Definition of Data Set:

Table 1 below summarises some of the key features of the data set that was used in this study. The dataset's full metadata is documented in the UCI repository Diabetes testing for Pima Indians was the driving force behind the collection of this data. A Pima Indian's diabetes status was predicted using demographic information (age, parity, medical history) and examination results (blood pressure, BMI, glucose tolerance test scores, etc.).

There are two possible outcomes, with the first meaning "diagnosed with diabetes." Around 500 Class 1 examples and Class 268 examples.

Table 1. Characteristics of data sets.

Data set	No. of example	Input attributes	Output classes	Total No. of attributes	Missing attributes status	Noisy attributes status
Pima	768	8	2	9	No	No

The characteristics are as follows:

First,

- the total number of pregnancies

- 4) Triceps skin fold thickness (mm) (mm)
- 5) 2-hour serum insulin (mu U/ml)
- 8) Age (years)
- Nine) Variable by class (0 or 1).

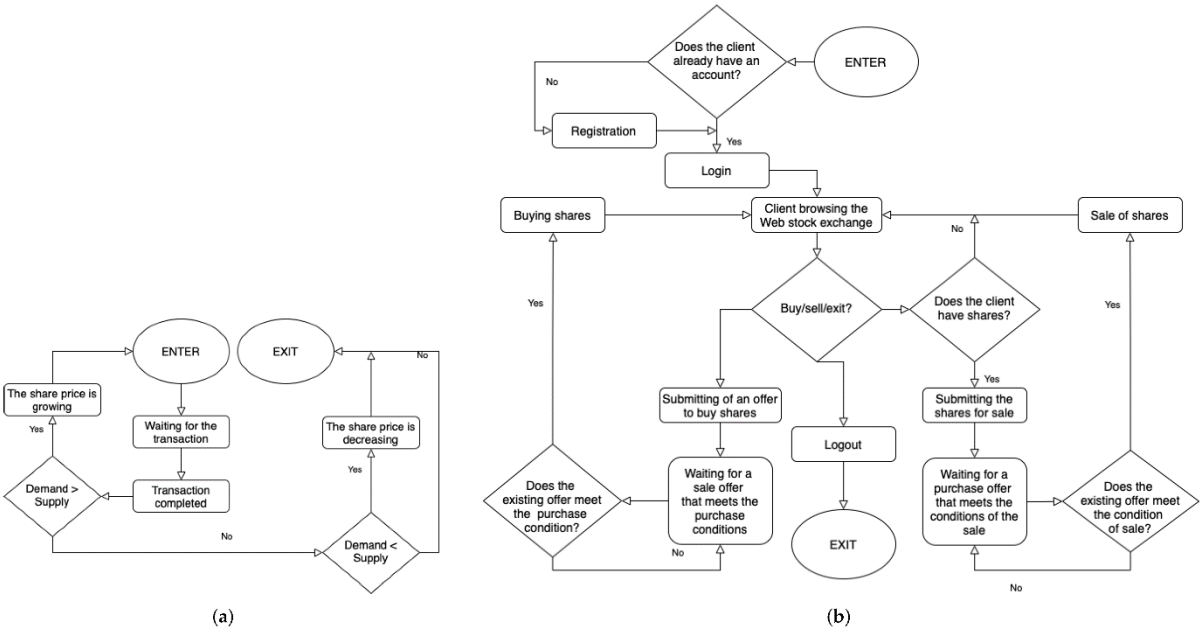
Web-Based Interactivity:

Based on IBM's Day Trader Benchmark, we developed a data mining model for behavioural analysis and a benchmark for the stock exchange's web-based application programming interface. To begin, we'll go through the study's transactional environment, which is an interactive one. A stock purchase or sell that has been finalised is considered a transaction. When a client makes a buy offer, we assume that the customer is specifying the highest possible price at which they are willing to buy; when a client makes a selling offer, we assume that the specified price is the lowest possible price at which the client is willing to sell.

(Figure 1a). In the case of a purchase offer, we presume the user has set their maximum acceptable purchase price, while in the case of a sale offer, we assume

the user has set their desired share price. Users and the company are the only parties with whom share purchases and sales are possible.

(Figure 1b) This schematic depicts the app's primary features. After registering and logging in, the customer can view the stock exchange's current offers and publish their own buy and sell offers for the shares they own. If the price of the selected share reaches the value provided by the user in the buy/sell offer, the sale and purchase are executed automatically.



Phishing Techniques and Security Measures:

The majority of phishing assaults today originate in online spaces like email and instant messaging apps.

- Indicators of Fraudulent Emails
- Phishing and other potentially malicious websites
- Most of these URLs lead to what appears to be a legitimate website, only they require you to submit information that the financial institution already has, including your social security number and mailing address. These phishing emails are typically spam that is sent to a huge number of people.
- The Identification of Phishing Web Addresses

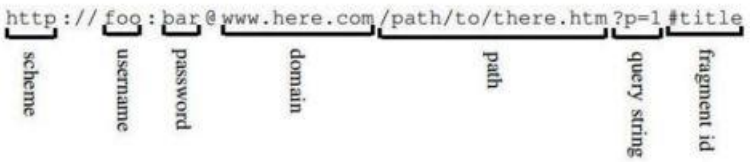


Fig. 2. The structure of a URL

CONCLUSION: Data mining is the process of discovering actionable insights in large datasets (databases, texts, the web, etc.). A wide range of free resources exist to help you decipher and extrapolate facts and information. In this study, we contrast several widely used data mining toolkits with web mining. When all factors are taken into account, a comprehensive examination of data mining and web mining software solutions highlights their value and importance. This analysis compares and contrasts the features, pros and cons, and utility of some popular data mining technologies. The evaluation considered each program's API support, database compatibility. Some Method used Clustering ,Association , Prediction as well as some important things cover like phishing and web based interactivity .

REFERENCES:

http://www.researchpublications.org/IJCSA/NCAICN-13/183.pdf

[1] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000. [2] Du Mouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999) Squashing flat files flatter. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM Press.

http://ojs.mediu.edu.my/index.php/IJDSR/article/view/1841/717

[1].

M. Rouse, "association rules (in data mining)," Techtarget, [Online].

Available:

https://searchbusinessanaly

tics.techtarget.com/definition/assoc
iation

https://www.scirp.org/pdf/JSEA_2013032915290889.pdf
[1] L. Carnimeo and A. Giaquinto, “An Intelligent System for Improving Detection of Diabetic Symptoms in Retinal Images,” IEEE International Conference on Information Technology in Biomedicine, Ioannina, 26-28 October 2006.
https://www.mdpi.com/2076-3417/12/12/6115/htm

1. Mughal, M.J. Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*. [**Google Scholar**] [**CrossRef**][**Green Version**]

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8093510
SpamAssassin, “Public corpus,” http://spamassassin.apache.org/ publiccorpus/, accessed January 2011. [1] F. Toolan and J. Carthy, “Phishing detection using classifier ensembles,” in eCrime Researchers Summit, 2009. eCRIME ’09., 20 2009. [2] T. Raffetseder, E. Kirda, and C. Kruegel “Building Anti-Phishing Browser Plug-Ins: An Experience Report” Third International Workshop on Software Engineering for Secure Systems (SESS'07)