UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECP2753-01

DATA MINING

ASSIGNMENT 1

PREPARED BY :

NO	NAME	MATRIC NO
1	NUR ALEEYA SYAKILA BINTI MUHAMAD SUBIAN	A19EC0127
2	AINA ALIAH BINTI RUSLAN	A19EC0012
3	JEGGEAN A/L RAJENDRAN	A19EC0059

LECTURER'S NAME	DR ROZILAWATI BINTI DOLLAH @ MD ZAIN
SUBMITTED ON	29/03/2021

Table of Contents

1.0 Knowledge Discovery in Databases (KDD)

1.1 Definition

The process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. [13]

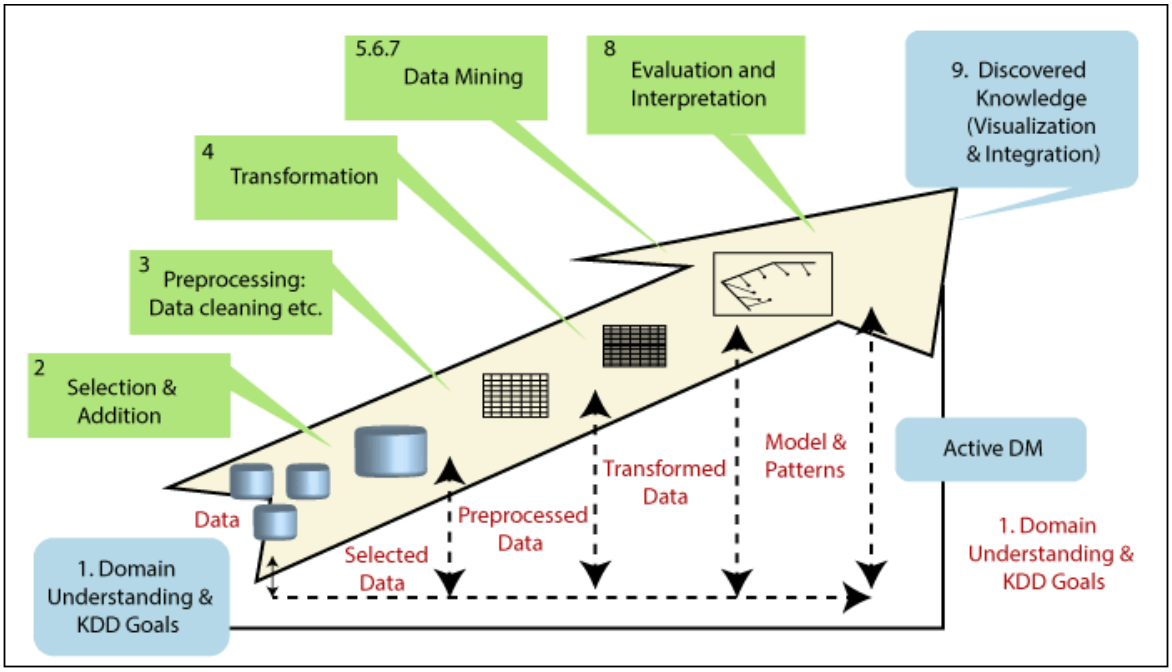

Figure 1 : KDD Process [15]

1.2 KDD Process

1. Building up an understanding of the application domain

This is the initial preliminary step. It develops the scene for understanding what should be done with the various decisions like transformation, algorithms, representation, etc. The individuals who are in charge of a KDD venture need to understand and characterize the objectives of the end-user and the environment in which the knowledge discovery process will occur (involves relevant prior knowledge). [3]

2. Choosing and creating a data set on which discovery will be performed

Once defined the objectives, the data that will be utilized for the knowledge discovery process should be determined. This incorporates discovering what data is accessible, obtaining important data, and afterward integrating all the data for knowledge discovery onto one set involves the qualities that will be considered for the process. This process is important because Data Mining learns and discovers from the accessible data. This is the evidence base for building the models. If some significant attributes are missing, at that point, then the entire study may be unsuccessful from this respect, the more attributes are considered. On the other hand, to organize, collect, and operate advanced data repositories is expensive, and there is an arrangement with the opportunity for best understanding the phenomena. This arrangement refers to an aspect where the interactive and iterative aspect of the KDD is taking place. [3]

3. Preprocessing and cleansing

In this step, data reliability is improved. It incorporates data clearing, for example, Handling the missing quantities and removal of noise or outliers. It might include complex statistical techniques or use a Data Mining algorithm in this context. For example, when one suspects that a specific attribute is lacking reliability or has many missing data, at this point, this attribute could turn into the objective of the Data Mining supervised algorithm. A prediction model for these attributes will be created, and after that, missing data can be predicted. The expansion to which one pays attention to this

level relies upon numerous factors. Regardless, studying the aspects is significant and regularly revealing by itself, to enterprise data frameworks. [3]

4. Data Transformation

In this stage, the creation of appropriate data for Data Mining is prepared and developed. Techniques here incorporate dimension reduction(for example, feature selection and extraction and record sampling), also attribute transformation(for example, discretization of numerical attributes and functional transformation). This step can be essential for the success of the entire KDD project, and it is typically very project-specific. For example, in medical assessments, the quotient of attributes may often be the most significant factor and not each one by itself. In business, we may need to think about impacts beyond our control as well as efforts and transient issues. For example, studying the impact of advertising accumulation. However, if we do not utilize the right transformation at the starting, then we may acquire an amazing effect that gives us the transformation required in the next iteration. Thus, the KDD process follows upon itself and prompts an understanding of the transformation required. [3]

5. Prediction and description

We are now prepared to decide on which kind of Data Mining to use, for example, classification, regression, clustering, etc. This mainly relies on the KDD objectives, and also on the previous steps. There are two significant objectives in Data Mining, the first one is a prediction, and the second one is the description. Prediction is usually referred to as supervised Data Mining, while descriptive Data Mining incorporates the unsupervised and visualization aspects of Data Mining. Most Data Mining techniques depend on inductive learning, where a model is built explicitly or implicitly by generalizing from an adequate number of preparing models. The fundamental assumption of the inductive approach is that the prepared model applies to future cases. The technique also takes into account the level of meta-learning for the specific set of accessible data. [3]

6. Selecting the Data Mining algorithm

Having the technique, we now decide on the strategies. This stage incorporates choosing a particular technique to be used for searching patterns that include multiple inducers. For example, considering precision versus understandability, the previous is better with neural networks, while the latter is better with decision trees. For each system of meta-learning, there are several possibilities of how it can be succeeded. Meta-learning focuses on clarifying what causes a Data Mining algorithm to be fruitful or not in a specific issue. [3]

7. Utilizing the Data Mining algorithm

At last, the implementation of the Data Mining algorithm is reached. In this stage, we may need to utilize the algorithm several times until a satisfying outcome is obtained. For example, by turning the algorithms control parameters, such as the minimum number of instances in a single leaf of a decision tree. [3]

8. Evaluation

In this step, we assess and interpret the mined patterns, rules, and reliability to the objective characterized in the first step. Here we consider the preprocessing steps as for their impact on the Data Mining algorithm results. For example, including a feature in step 4, and repeat from there. This step focuses on the comprehensibility and utility of the induced model. In this step, the identified knowledge is also recorded for further use. The last step is the use, and overall feedback and discovery results acquired by Data Mining. [3]

9. Using the discovered knowledge

Now, we are prepared to include the knowledge into another system for further activity. The knowledge becomes effective in the sense that we may make changes to the system and measure the impacts. The accomplishment of this step decides the effectiveness of the whole KDD process. There are numerous challenges in this step, such as losing the "laboratory conditions" under which we have worked. For example, the

knowledge was discovered from a certain static depiction, it is usually a set of data, but now the data becomes dynamic. Data structures may change certain quantities that become unavailable, and the data domain might be modified, such as an attribute that may have a value that was not expected previously. [3]

2.0 Data Mining

2.1 Definition

Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).

2.2 DM techniques/tasks

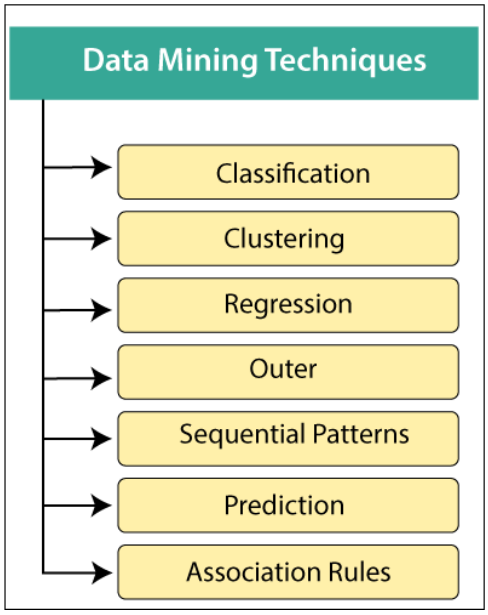

Figure 2 : Data Mining Techniques [14]

1. Predictions.
2. Sequential patterns.
3. Regression.
4. Clustering.
5. Classification.
6. Outlier detection.
7. Association.

2.3 DM algorithms

Table 1 : Data Mining Algorithms

Technique / Task	Algorithm
Classification	C4.5, Support Vector Machines (SVM), Adaboost, k Nearest Neighbour (kNN), Naive Bayes, Statistical Procedure Based Approach, Neural Network (NN),Decision Tree
Clustering	K-mean, Expectation-Maximization (EM), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Mean-Shift, hierarchical
Association	Apriori Algorithm
Outlier Detection	k Nearest Neighbour (kNN), Support Vector Clustering, Support Vector Machines (SVM)
Decision trees	Classification and regression (CART), ID3, C4.5, Chi-square Automatic Interaction Detection (CHAID)
Predictions	Neural Network (NN)
Sequential patterns	FARPAM, Bitmap, OpenMP, Apriori, SPAM, RobustSPAM
Regression.	Support Vector Machines (SVM)

2.4 DM applications/tools

Table 2 : Data Mining Applications

Technique / Task	Application Tool
Classification	Weka,ApacheMahout,Oracle,Mallet,XLMiner, ENVI
Clustering	Cluster 3.0, Java TreeView,PYCLUSTER
Outlier Detection	CMSR Data Mine
Association	FPM, Bart Goethals, FrIDA,KNIME, Magnum Opus
Predictions	Rapid Miner, Orange, Weka, Oracle, XLMiner
Sequential patterns	XAffinity(TM), SPMF,Miningco
Regression.	Weka, Oracle, XLMiner, IBM SPSS Modeler

3.0 Machine Learning

3.1 Definition

Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data without being explicitly programmed. It is a method of data analysis that automates analytical model building. Machine learning is a branch of artificial intelligence and this algorithm builds a model based on observations or data, such as examples, direct experience, or instruction, to identify patterns and make decisions with minimal human intervention or assistance and adjust actions accordingly. Currently, machine learning has been used in multiple fields and industries. For example, medical diagnosis, image processing, prediction, classification, learning association, regression etc. The intelligent systems built on machine learning algorithms also have the capability to learn from past experience or historical data.

3.2 Machine Learning Methods

Machine learning algorithms are often categorized as supervised or unsupervised. **Supervised learning** algorithms are trained using labeled examples, such as an input where the desired output is known. The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data. Supervised learning is commonly used in applications where historical data predicts likely future events. For example, you could use supervised machine learning techniques to help a service business that wants to predict the number of new users who will sign up for the service next month. **Unsupervised learning** is used against data that has no historical labels. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data. For example, you could use unsupervised learning techniques to help a retailer that wants to segment products with similar characteristics — without having to specify in advance which characteristics to use.

1. Regression

   Regression methods fall within the category of supervised machine learning. They help to predict or explain a particular numerical value based on a set of prior data, for example predicting the price of a property based on previous pricing data for similar properties.

2. Classification

   Another class of supervised ML, classification methods predict or explain a class value. For example, they can help predict whether or not an online customer will buy a product. The output can be yes or no: buyer or not buyer. But classification methods aren't limited to two classes. For example, a classification method could help to assess whether a given image contains a car or a truck. In this case, the output will be 3 different

values: 1) the image contains a car, 2) the image contains a truck, or 3) the image contains neither a car nor a truck.

3. Clustering

    Clustering methods is in the category of unsupervised machine learning because their goal is to group or cluster observations that have similar characteristics. Clustering methods don't use output information for training, but instead let the algorithm define the output. In clustering methods, we can only use visualizations to inspect the quality of the solution.

4. Dimensionality Reduction

    The dimensionality reduction removes the least important information (sometimes redundant columns) from a data set. In practice, the data is often seen as sets with hundreds or even thousands of columns (also called features), so reducing the total number is vital. For instance, images can include thousands of pixels, not all of which matter to the analysis. Or when testing microchips within the manufacturing process, it might have thousands of measurements and tests applied to every chip, many of which provide redundant information. In these cases, we need dimensionality reduction algorithms to make the data set manageable.

5. Ensemble Methods

    Ensemble methods use the idea of combining several predictive models (supervised ML) to get higher quality predictions than each of the models could provide on its own. For example, the Random Forest algorithms is an ensemble method that combines many Decision Trees trained with different samples of the data sets. As a result, the quality of the predictions of a Random Forest is higher than the quality of the predictions estimated with a single Decision Tree.

6. Neural Nets and Deep Learning

    The objective of neural networks is to capture non-linear patterns in data by adding layers of parameters to the model. Neural networks, a beautiful biologically-inspired programming paradigm which enables a computer to learn from

observational data and Deep learning, a powerful set of techniques for learning in neural networks. Neural networks and deep learning currently provide the best solutions to many problems in image recognition, speech recognition, and natural language processing.

7. Transfer Learning

Transfer Learning refers to re-using part of a previously trained neural net and adapting it to a new but similar task. Specifically, once we train a neural net using data for a task, we can transfer a fraction of the trained layers and combine them with a few new layers that you can train using the data of the new task. By adding a few layers, the new neural net can learn and adapt quickly to the new task.

8. Reinforcement Learning

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

9. Natural Language Processing

Natural Language Processing (NLP) is the part of AI that studies how machines interact with human language. NLP works behind the scenes to enhance tools we use every day, like chatbots, spell-checkers, or language translators. Combined with machine learning algorithms, NLP creates systems that learn to perform tasks on their own and get better through experience.

10. Word Embeddings

Word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

3.3 Machine Learning vs Data Mining

Table 3 : Differences between Machine Learning and Data Mining

Machine Learning	Data Mining
Introduce new algorithms from data as well as past experience.	Extracting knowledge from a large amount of data.
Introduce in near 1950, the first program was Samuel's checker-playing program	Introduce in 1930, initially referred as knowledge discovery databases
Existing data as well as algorithms.	Traditional databases with unstructured data.
Goes beyond what's happened in the past to predict future outcomes based on the pre-existing data.	Simply looking for patterns that already exist in the data
The machine is usually given some rules or variables to understand the data and learn.	The 'rules' or patterns are unknown at the start of the process.
The initial rules are in place, the process of extracting information and 'learning' and refining is automatic, and takes place without human intervention. In other words, the machine becomes more intelligent by itself.	More manual process that relies on human intervention and decision making.
Trained on a 'training' data set, which teaches the computer how to make sense of data, and then to make predictions about new data sets.	Used on an existing dataset (like a data warehouse) to find patterns.
Can be used in a vast area.	Applied in the limited area.

3.4 Why Data Mining is Important

There is a wide availability of huge amounts of data in this world and that data needs for turning into such useful information and knowledge. Back to the definition of Data Mining where it is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data and these tools help to predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

Based on the definition, Data Mining can be viewed as a result of the natural evolution of information technology especially in the database industry (Data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining)). The fast-growing and huge amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large databases become “data tombs" data archives that are seldom revisited. Consequently, important decisions are often made based not on the information-rich data stored in databases but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. In this case, Data mining tools which perform data analysis may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. The widening gap between data and information calls for a systematic development of data mining tools which will turn data tombs into “golden nuggets" of knowledge

3.5 What Kind of Data To Be Mined

The data from multiple sources are integrated into a common source known as **Data Warehouse**. There are few types of data that can be mined:

1.Files

- Flat files are defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.
- Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on a flat file, then there will be no relations between the tables.
- Flat files are represented by a data dictionary. Eg: CSV file.
- **Application**: Used in DataWarehousing to store data, Used in carrying data to and from server, etc.

2.Relational Databases

- A Relational database is defined as the collection of data organized in tables with rows and columns.
- Physical schema in Relational databases is a schema which defines the structure of tables.
- Logical schema in Relational databases is a schema which defines the relationship among tables.
- Standard API of relational databases is SQL.
- **Application**: Data Mining, ROLAP model, etc.

3.DataWarehouse

- A data warehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.
- There are three types of data warehouse: **Enterprise** data warehouse, **Data Mart** and **Virtual** Warehouse.
- Two approaches can be used to update data in DataWarehouse: **Query-driven** Approach and **Update-driven** Approach.
- **Application**: Business decision making, Data mining, etc.

4.Transactional Databases

- Transactional databases is a collection of data organized by time stamps, date, etc to represent transactions in databases.
- This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
- Highly flexible system where users can modify information without changing any sensitive information.
- Follows ACID property of DBMS.
- **Application**: Banking, Distributed systems, Object databases, etc.

5.Multimedia Databases

- Multimedia databases consist of audio, video, images and text media.
- They can be stored on Object-Oriented Databases.
- They are used to store complex information in a pre-specified format.
- **Application**: Digital libraries, video-on demand, news-on demand, musical database, etc.

6.Spatial Database

- Store geographical information.
- Stores data in the form of coordinates, topology, lines, polygons, etc.
- **Application**: Maps, Global positioning, etc.

7.Time-series Databases

- Time series databases contain stock exchange data and user logged activities.
- Handles array of numbers indexed by time, date, etc.
- It requires real-time analysis.
- **Application**: eXtremeDB, Graphite, InfluxDB, etc.

8.WWW

- WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
- It is the most heterogeneous repository as it collects data from multiple resources.
- It is dynamic in nature as Volume of data is continuously increasing and changing.
- **Application**: Online shopping, Job search, Research, studying, etc

4.0 Artificial Intelligence (AI)

4.1 Definition

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

The ideal characteristic of artificial intelligence is its ability to rationalize and take actions that have the best chance of achieving a specific goal. A subset of artificial intelligence is machine learning, which refers to the concept that computer programs can automatically learn from and adapt to new data without being assisted by humans. Deep learning techniques enable this automatic learning through the absorption of huge amounts of unstructured data such as text, images, or video. [17]

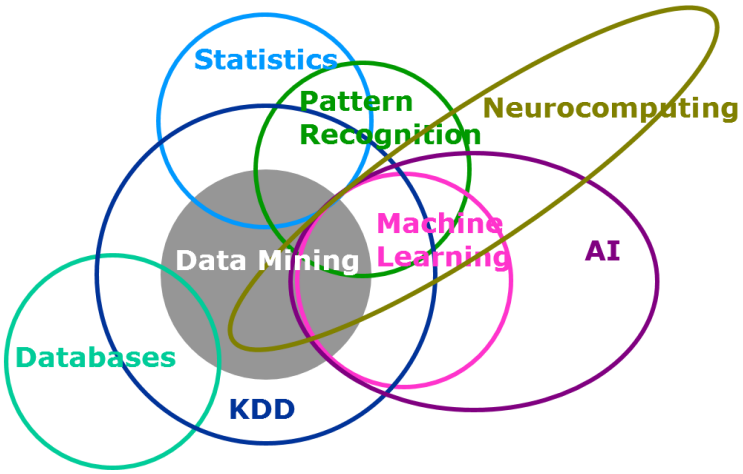

Figure 3 : Machine Learning, AI, Data Mining [16]

4.2 Machine Learning vs Artificial Intelligence

Table 4 : Differences between Machine Learning and Artificial Intelligence

Machine Learning	Artificial Intelligence
Machine learning is a subset of AI which allows a machine to automatically learn from past data without programming explicitly.	Artificial intelligence is a technology which enables a machine to simulate human behavior.
The goal of ML is to allow machines to learn from data so that they can give accurate output.	The goal of AI is to make a smart computer system like humans to solve complex problems.
In ML, we teach machines with data to perform a particular task and give an accurate result.	In AI, we make intelligent systems to perform any task like a human.
Deep learning is a main subset of machine learning.	Machine learning and deep learning are the two main subsets of AI.
Machine learning has a limited scope.	AI has a very wide range of scope.
Machine learning is working to create machines that can perform only those specific tasks for which they are trained.	AI is working to create an intelligent system which can perform various complex tasks.
Machine learning is mainly concerned about accuracy and patterns.	The AI system is concerned about maximizing the chances of success.
The main applications of machine learning are the Online **recommender system**, **Google search algorithms**, **Facebook auto friend tagging suggestions**, etc.	The main applications of AI are **Siri, customer support using catboats**, Expert System, Online game playing, intelligent humanoid robot, etc.
Machine learning can also be divided into mainly three types that are **Supervised learning**, **Unsupervised learning**, and **Reinforcement learning**.	On the basis of capabilities, AI can be divided into three types, which are, **Weak AI**, **General AI**, and **Strong AI**.
It includes learning and self-correction when introduced with new data.	It includes learning, reasoning, and self-correction.
Machine learning deals with Structured and semi-structured data.	AI completely deals with Structured, semi-structured, and unstructured data.

5.0 References

1. *KDD Process in Data Mining: What You Need To Know? | upGrad blog*. (2020, November 23). UpGrad Blog. https://www.upgrad.com/blog/kdd-process-data-mining/
2. DBD, R. (2021). *KDD Process/Overview*. Uregina.ca. http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
3. *KDD Process in Data Mining - Javatpoint*. (2011). Www.javatpoint.com. https://www.javatpoint.com/kdd-process-in-data-mining
4. *Data Mining Processes | Data Mining tutorial by Wideskills*. (2015). Wideskills.com. https://www.wideskills.com/data-mining-tutorial/data-mining-processes#:~:text=Data%20mining%20is%20the%20core,series%20analysis%20and%20so%20on.
5. *What Is The Difference Between Data Mining And Machine Learning?* (2020). Bernard Marr. https://bernardmarr.com/default.asp?contentID=1741
6. *Data Mining vs Machine Learning | Top 10 Best Differences To Learn*. (2018, March 9). EDUCBA. https://www.educba.com/data-mining-vs-machine-learning/
7. *Machine learning: What it is and why it matters. (n.d.).* Retrieved March 19, 2021, https://www.sas.com/en_my/insights/analytics/machine-learning.html#:~:text=Machine%20learning%20is%20a%20method,decisions%20with%20minimal%20human%20intervention.
8. Castañón, J. (2019, September 5). *10 Machine Learning Methods that Every Data Scientist Should Know*. Medium. https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9
9. What is Machine Learning? A definition - Expert System. (2021, February 11). Retrieved from https://www.expert.ai/blog/machine-learning-definition/
10. Brownlee, J. (2019, August 7). *What Are Word Embeddings for Text?* Machine Learning Mastery. https://machinelearningmastery.com/what-are-word-embeddings/

11. *Word embedding machine learning models*. Algorithmia Blog. (2020, July 29). https://algorithmia.com/blog/word-embedding-machine-learning-models
12. *Difference between Artificial intelligence and Machine learning - Javatpoint*. (2011). Www.javatpoint.com. https://www.javatpoint.com/difference-between-artificial-intelligence-and-machine-learning#:~:text=Artificial%20intelligence%20is%20a%20technology%20which%20enables%20a%20machine%20to,humans%20to%20solve%20complex%20problems
13. Shivali, Birla, J., & Gurpreet. (2018). Knowledge Discovery in Data-Mining. *International Journal of Engineering Research & Technology*, *3*(10). https://www.ijert.org/knowledge-discovery-in-data-mining#:~:text=WHAT%20DOES%20KNOWLEDGE%20DISCOVERY%20IN%20DATABASES%20(KDD)%20MEAN%3F
14. *Data Mining Techniques - Javatpoint*. (2011). Www.javatpoint.com. https://www.javatpoint.com/data-mining-techniques
15. *KDD Process in Data Mining - Javatpoint*. (2011). Www.javatpoint.com. https://www.javatpoint.com/kdd-process-in-data-mining
16. Mitchell-Guthrie, P. (2014, August 22). *Looking backwards, looking forwards: SAS, data mining, and machine learning*. The SAS Data Science Blog. https://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-looking-forwards-sas-data-mining-and-machine-learning/
17. *How Artificial Intelligence Works*. (2021). Investopedia. https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp
18. Alexander, D. (n.d.). *Data mining*. https://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/#:~:text=For%20businesses%2C%20data%20mining%20is,and%20accurately%20predict%20customer%20loyalty