# MARK5836 Assignment 1

## 1. Discuss what are the key similarities and differences between machine learning and data mining (2.5 Marks).

Data mining is the process of identifying interesting and useful patterns in large databases using automated computational methods (Arentze, 2009). Data mining falls under the field of Knowledge Discovery in Databases (KDD) that encompasses theories, methods, and techniques; attempting to find pattern and meaning in data and extract useful knowledge from them. According to Fayyad et. al. (1996), KDD is a process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The KDD process involves several steps such as selection, pre-process, transformation, data mining, interpretation, and evaluation. Among these steps, the most important one is data mining, exemplifying the application of machine learning algorithms in analysing data (Kavakiotis et. al., 2017).

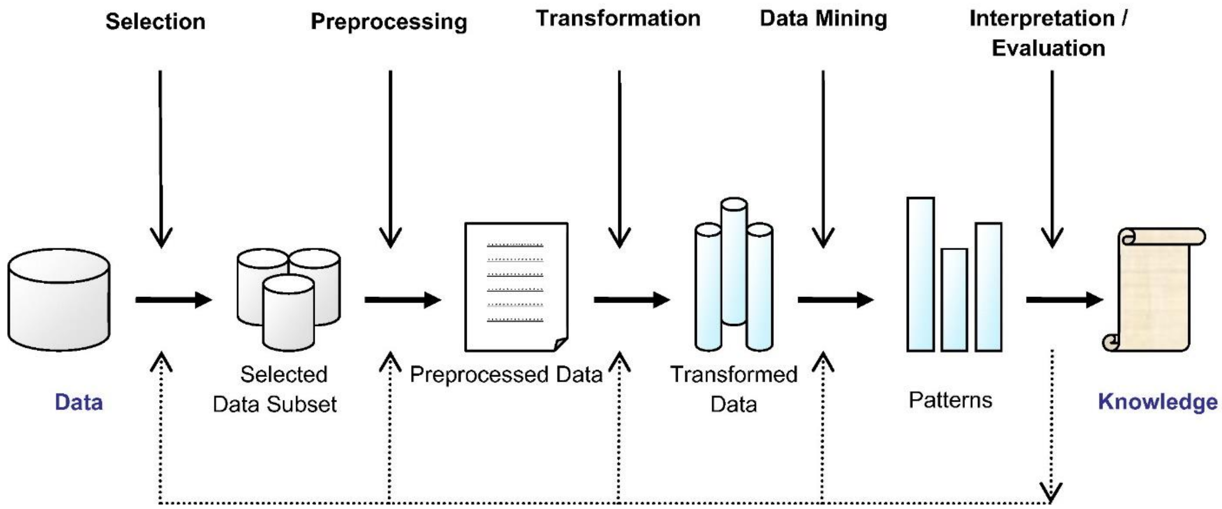

Fig. 1. The basic steps of the KDD process (Kavakiotis et. al., 2017).

On the contrary, Machine Learning (ML) is often considered same as “artificial intelligence”, since in a broader sense learning is one of the main characteristics of an entity called intelligent (Kavakiotis et.al., 2017). The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience (Wilson & Keil, 2001).

With the increasing availability of very large datasets and evolving power of machine learning methods, the field of data mining gained more popularity in the late 90’s due to the through automated data recording methods and (Arentze, 2009). There are some key differences between data mining and machine learning. Data mining focuses on the extraction of information, typically from large data sets, into some useful structure for data analysis by employing machine learning, statistics, and database systems; whereas machine learning focuses on the design and study of algorithms to build mathematical models based on data sets (Chandra, 2021).

Some standard tasks in data mining include clustering, pattern recognition, regression, summarization, dependency modelling and change, and deviation detection (Arenze 2009). Kavakiotis et.al. (2017) categorised machine learning tasks into three broad categories. These are: a) supervised learning, in which the system infers a function from labelled training data, b) unsupervised learning, in which the learning system tries to infer the structure of unlabelled data, and c) reinforcement learning, in which the system interacts with a dynamic environment.

Lastly, the application of the two are also difference. Zhou (2003) stated that, data mining has received contributions from a lot of disciplines such as databases, machine learning, statistics, information retrieval, data visualization, parallel and distributed computing, etc. It is reasonable to argue that the powerful data management techniques donated by the database community and the practical data analysis techniques donated by the machine learning community, data mining would be seeking a needle in the haystack (Zhou, 2003). Machine learning is generally more efficient than traditional mathematical and statistical models. ML driven tools and techniques have been widely and successfully applied to many different fields (such as health, education, biostatistics, manufacturing wireless sensor networks, and finance), since they remain incapable of understanding complex relations among features of data samples and predicting unknown feature values for a new sample (Dogan 2021).

**2. Discuss the similarities and differences between linear and logistic regression models with applications (2.5 Marks).**

Linear Regression Model and Logistic Regression Model are two of the most popular statistical techniques used in data science. Although, both of these models fall fundamentally under predictive analysis, however, they are different in nature and are used for different kind of problems.

The equation for linear regression is [y = a + bx]. A continuous value can take any value within a specified interval (range) of values. Whereas logistic regression equation is [f(x) = ex / ex + 1]. Here, the output is a probability ranging from 0 (not going to happen) to 1 (going to happen), or a categorization that says something is either part of the category or not part of the category.

Both linear and logistic regression have one dependent variable and one independent variable, however logistic regression has two types of output: a) *Classification* **-** decides between two available outcomes, b) *Probability* - determines the probability that something is true or false, with specific meaning. When the dependent variables are binary logistic regression is considered and when dependent variables are continuous then linear regression is used.

The key similarity between linear and logistic regression is both are machine learning algorithms that are part of supervised learning models. Since both are part of a supervised model so they make use of labelled data for making predictions. Having said that, the applications of these two models are quite different. Linear regression is used for predicting continuous values whereas logistic regression is widely used as a classification algorithm.

A logistic approach fits best when the task is to predict the likelihood of an event happening or a choice being made. For instance, if a business wants to know the likelihood of a visitor choosing an offer made on their website, logistic regression models can help them determine a probability of what type of visitors are likely to accept the offer or not. As a result, the business can make better decisions about promoting their offer or make decisions about the offer itself. As more data is provided, it could learn how to do this better over time (IBM Analytics, 2016). Manufacturer’s analytics team can use logistic regression analysis as part of a statistics software package to discover a probability between part failures in machines and the length of time those parts are held in inventory. With the information it receives from this analysis, the team can decide to adjust delivery schedules or installation times to eliminate future failures. Some other real-life examples of using logistic regression for classification algorithm include assessing credit risk and credit scoring, profiling the consumers of packaged goods, increasing profits in the banking industry, predicting customer churn and so on (IBM Analytics, 2016).

## References

Arentze, T. A. (2009). Spatial data mining, cluster and pattern recognition. *International Encyclopedia of Human Geography*, 325-331.

Dogan, A., & Birant, D. (2020). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 114060.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37-37.

IBM Analytics. (2016). A glimpse inside the mind of a data scientist, *IBM*, retrieved from < https://www.ibm.com/topics/logistic-regression>

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, *15*, 104-116.

Wilson, R. A., & Keil, F. C. (Eds.). (2001). *The MIT Encyclopedia of the Cognitive Sciences*. MIT press.

Zhou, Z. (2003). Three perspectives of data mining. *Artificial Intelligence, 143*(1),139-146.