SECP2753 – DATA MINING

SEMESTER 2 2021/2022

GROUP 2

ASSIGNMENT 1

Members:

1. **GROUP LEADER**: NAYLI NABIHAH JASNI (A20EC0105)
2. MIKHEL ADAM BIN MUHAMMAD EZRIN (A20EC0237)
3. MUHAMMAD DINIE HAZIM BIN AZALI (A20EC0084)
4. MADINA SURAYA BINTI ZHARIN (A20EC0203)

TABLE OF CONTENTS

1.0 KNOWLEDGE DISCOVERY IN DATABASE (KDD)

1.1 Definition

According to Frawley et al. (1992), knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley et al., 1992).

While in another article, Knowledge Discovery in Database (KDD) is an automatic, exploratory analysis and modelling of large data repositories. KDD is the organised process of identifying valid, novel, useful, and understandable patterns from large and complex data sets (Maimon & Rokach, 2005).

1.2 Process in KDD

1.2.1 Domain Understanding & KDD Goals

- Understand the purpose of decision making such as transformation, algorithms, representation, etc.
- Understand and characterise the objectives of the end-user and the environment of the KDD process to occur, according to relevant prior knowledge.

1.2.2 Selection & Addition

- Select and create a data set on which discovery will be performed.
- Discover the accessibility of data, obtaining important data and integrating the data for KDD, involving the qualities needed.

1.2.3 Preprocessing: Data cleaning etc.

- Improve data reliability.
- Creating a prediction model is one of the strategies to help find missing data.

1.2.4 Transformation

- Creation of appropriate data for data mining is prepared and developed.

- Using dimensionality reduction or transformation methods to reduce the number of variables under consideration or finding invariant representations for the data.

1.2.5 Data Mining: Prediction and Description

- Prediction: Point out on as supervised data mining
- Descriptive: Containing the visualisation and unsupervised aspects in data mining.

1.2.6 Data Mining: Selection

- This is where the particular technique is chosen to be used when searching patterns that include multiple inducers.

1.2.7 Data Mining: Utilisation

- This is where the data mining algorithms start to get involved in the process.
- For certain cases, the algorithm will be used more than once to get the desired outcome.

1.2.8 Evaluation and Interpretation

- This process will be the one where the assessments and interpretation of the mined patterns, rules and reliability of the objective characterised in the earliest process.

1.2.9 Discovered Knowledge

- Ready to include the knowledge into another system for further activity.
- An effective knowledge in the sense that we may make changes to the system and measure the impacts

1.3 Figure in KDD

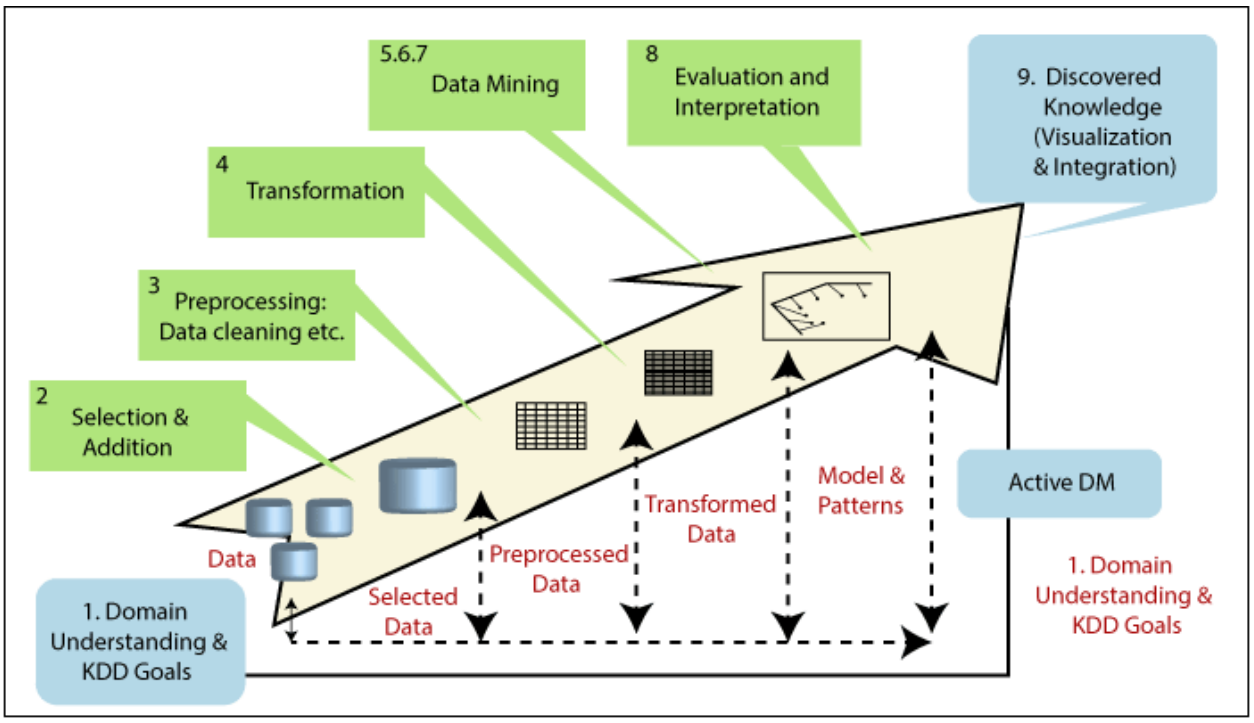

Figure 1: Knowledge Discovery in Database

1.4 Step of KDD

Below are the steps of KDD briefly explained as simple as possible:

1.4.1 Data Cleaning

Removing noisy, inconsistent, and irrelevant data from the collection.

1.4.2 Data Integration

Combining multiple data sources into a common source.

1.4.3 Data Selection

Data relevant to the analysis are retrieved from the database.

1.4.4 Data Transformation

Data is transformed into appropriate forms required for the mining procedure.

1.4.5 Data Mining

Intelligent techniques that are used to extract potentially useful patterns.

1.4.6 Pattern Evaluation

Identifying and evaluating data patterns representing knowledge based on a given measure.

1.4.7 Knowledge Representation

Representing or visualising the data mining results.

2.0 DATA MINING (DM)

2.1 Definition

According to Raval K.,(2012), a process involving the extraction of functional information and patterns from large data is well known as data mining. Data mining also interprets mathematical analysis to derive the patterns and trends that are discovered in the data.

2.2 Data Mining Steps

Figure 2: Data Mining Steps

The four steps/stages of data mining are as follows explained as simple as possible:

2.2.1 Data gathering

Data relevant to the analysis is identified and gathered. The data may be located from different sources.

2.2.2 Data preparation

Pre-process the data to prepare it to be mined. In certain cases the data may be transformed to make consistent data sets.

2.2.3 Mining the data

Once the data is prepared, an appropriate mining technique will be used to mine the data.

2.2.4 Data analysis and interpretation

Interpret or visualise the results by using analytical models to further make decisions based on the results

2.3 Data Mining Techniques/ Tasks

Below are the techniques involved in data mining:

2.3.1 Association rule mining

A process where "If-then" statements are used to identify relationships between data elements. Support and confidence criteria are used to assess the relationships. Support measures the frequency of related elements that appear in the data set while confidence reflects the frequency of accurate "if-then" statements.

2.3.2 Classification

Assigns the elements in data sets to different categories defined as part of the data mining process.

2.3.3 Clustering

Data elements that share particular characteristics are grouped together into clusters as part of data mining applications.

2.3.4 Regression

Another method to find relationships in datasets by calculating predicted data values based on a set of variables.

2.3.5 Sequence and path analysis

Data can also be mined to look for patterns in which a particular set of events or values leads to later ones.

2.3.6 Neural networks

A set of algorithms that simulate the activity of a human brain. Useful in complex recognition applications involving deep learning.

2.4 Data Mining Algorithm

DM Techniques	DM algorithms
Association rule mining	- AIS algorithm - SETM algorithm - Apriori algorithm
Classification	- Logistic Regression - Naive Bayes - K-Nearest Neighbours - Decision trees - Support Vector Machines
Clustering	- K-Means - Mean-Shift - Density-Based Spatial Clustering of Applications with Noise (DBSCAN) - Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM) - Agglomerative Hierarchical
Regression	- Simple Linear Regression - Lasso Regression - Logistic Regression - Support Vector Machines - Multivariate Regression Algorithm - Multiple Regression Algorithm
Sequence and path analysis	- AprioriAll Algorithm - AprioriSome Algorithm
Neural networks	- Hopfield network

	- Multilayer perceptron - Boltzmann machine - Kohonen network

2.5 Issue(s) in Data Mining

2.5.1 Mining methodology and user interaction issues

i. Mining different kind of knowledge in databases

Different users have different knowledge and use it in different ways. Thus, it is difficult to cover a vast range of data to meet client requirements.

ii. Interaction mining of knowledge at multiple levels of abstractions

An interactive data mining process are needed to focus more on searching patterns from different angles

iii. Query languages and ad hoc mining

Language of the data mining query language should be perfectly matched with the query language of the data warehouse.

iv. Handling missing or incomplete data

Many attribute values will be incorrect due to human error or instrument fail. Data cleaning helps this.

2.5.2 Performance issues

i. Efficiency and scalability of data mining algorithms

ii.Parallel, distributed, and incremental mining algorithms

Caused by a huge size of databases, wide distribution of data, and complexity of some data mining methods.

2.5.3 Diversity of database type issues

i. Handling of relational complex types of data

ii. Mining information from heterogeneous databases and global information systems

Data fetched from different sources on LAN and WAN

2.6 Data Mining Tools/Applications

Tools	Application
MonkeyLearn	MonkeyLearn is a machine learning platform that specialises in text mining. Available in a user-friendly interface, you can easily integrate MonkeyLearn with your existing tools to perform data mining in real-time.
RapidMiner	RapidMiner is a free open-source data science platform that features hundreds of algorithms for data preparation, machine learning, deep learning, text mining, and predictive analytics.
Oracle Data Mining	Oracle Data Mining is a component of Oracle Advanced Analytics that enables data analysts to build and implement predictive models. It contains several data mining algorithms for tasks like classification, regression, anomaly detection, prediction, and more.
iBM SPSS Modeler	IBM SPSS Modeler is a data mining solution, which allows data scientists to speed up and visualise the data mining process. Even users with little or no programming experience can use advanced algorithms to build predictive models in a drag-and-drop interface.
Weka	Weka is an open-source machine learning software with a vast collection of algorithms for data mining. It was developed by the University of Waikato, in New Zealand, and it's written in JavaScript.
KNIME	KNIME is a free, open-source platform for data mining and machine learning. Its intuitive interface allows you to create end-to-end data science workflows, from modelling to production. And different pre-built components enable fast modelling without entering a single line of code.
H2O	H2O is an open-source machine learning platform, which aims to make AI technology accessible to everyone. It supports the most common ML algorithms and offers Auto ML functions to help

	users build and deploy machine learning models in a fast and simple way, even if they are not experts.
Orange	Orange is a free, open-source data science toolbox for developing, testing, and visualising data mining workflows.
Apache Mahout	Apache Mahout is an open-source platform for creating scalable applications with machine learning. Its goal is to help data scientists or researchers implement their own algorithms.
SAS Enterprise Mining	SAS Enterprise Miner is an analytics and data management platform. Its goal is to simplify the data mining process to help analytics professionals turn large volumes of data into insights.

2.7 Importance of Data Mining

Points below mentioning about why data mining is important:

a. It is the procedure of capturing large sets of data in order to identify the insights and visions of that data. The demand of data in industry is rapidly growing which has also increased the demands for data analysts and data scientists.

b. We can analyse the data and convert it into meaningful information by using all the techniques. This will help the business to make accurate and better decisions in an organisation.

c. It helps to develop smart market decisions, run accurate campaigns, make predictions and more.

d. We can analyse customer behaviours and their insights with the help of Data Mining. This will lead to great success and data driven business.

2.8 Type of Data to be Mined

These are the types of data that can be mined:

2.8.1 Spatial Database

Store the geographical information, such as coordinates and lines.

2.8.2 Flat Files

These are known as the binary form or text

These files can be easily extracted using data mining algorithms

2.8.3 Relational Database

An organised collection of related data

Can be found in form of rows and columns

2.8.4 Transactional Database

An organised collection of related data that is determined by timestamps

2.8.5 Multimedia Database

Can be any multimedia, in example, videos, images, audio and text

2.8.6 Data Warehouse

Collection of data from one or more sources

2.8.7 World Wide Web (WWW)

Collection of documents and resources and each data

2.8.8 Time Series Database

Mainly can store the stock exchange data.

One of the examples is eXtremeDB

3.0 MACHINE LEARNING

3.1 Definition

Since a few decades ago, Machine Learning models have effectively illustrated the complex patterns of unobserved sets of data (Murdoch et, al., 2019) . Machine Learning is known as the one of the applications of Artificial Intelligence that targets the utilisation of data and algorithms to emulate the way of human learning as well as increasing the accuracy gradually.

3.2 Types of Machine Learning

Below are the three (3) main types of machine learning as well as its definition that is generally acknowledged by the mass.

3.2.1 Supervised learning

One of the most basic kinds of machine learning. This algorithm is trained using a small dataset (training dataset) that needs to be labelled accurately. The dataset should be fairly similar to the final dataset in terms of its characteristics. It will find cause and effect relationships based on the given parameters and in the end, the algorithm will have an idea on how the data works. If done properly, supervised learning can be extremely powerful when used in the right circumstances when used with the final dataset as it will also continue to learn even after being deployed.

3.2.2 Unsupervised learning

Similar to supervised learning, this type of learning has an advantage where it's able to work with unlabelled data eliminating the need for human input to create the dataset "machine readable". This allows larger datasets to be worked on. The relationship between data points are perceived in an abstract manner, resulting in the creation of hidden structures. This makes the algorithm versatile in a sense that it can adapt to the data dynamically and offers more post-deployment development as compared to supervised learning.

3.2.3 Reinforcement learning

This type of learning is very similar to how human beings learn in their lives, meaning that it has an algorithm that will learn and improve itself based on new situations using trial and error methods. Desired outputs will be reinforced and non-desired outputs will be discouraged. The algorithm is put to work in a work environment with an interpreter and a reward system and with each iteration, the interpreter will decide whether an outcome is favourable or not. In most cases, the reward system is directly related to the efficacy of the outcome. The solution is not an absolute value in typical reinforcement learning use-cases, such as finding the shortest route between two points on a map. Instead, it adopts an effectiveness score, which is expressed as a percentage value. The greater this percentage value, the greater the reward for the algorithm. As a result, the program is trained to give the best possible solution for the best possible reward.

3.3 Machine Learning Process

There are 5 keywords in the Machine Learning Process. Therefore, below are overview of Machine Learning Process:

3.3.1 Get Data

Understand the purpose of the project to recognise the input and output formats. Machine Learning model is used to retrieve meaningful insights and compatible results.

3.3.2 Clean, Prepare and Manipulate data

Raw data is constantly unorganised and missing its elements. Those unorganised data can contribute to the failure of Machine Learning. After choosing the suitable data sets, all of these data sets need to be converted to valid formats that have been chosen in the Machine Learning platform. Then, finally the data sets can be split into training and test data sets.

3.3.3 Train Model

This is where the data set will be connected to a suitable algorithm. These mathematical modelling based algorithms are used to study and create some predictions.

These algorithms are usually directed into one of these three categories:

- Binary (Classified into 2 categories)
- Classification (Classified into many categories)
- Regression (Predicted by numeric means)

3.3.4 Test Model

By using the test data, the model's accuracy can be checked to verify the trained model made before.

3.3.5 Improve

This process is included if the results from the test model are not good enough. Some of the things that can be considered while refining the model is by reviewing the model's results, reconsider the algorithm chosen as well as adjust the parameters of the chosen algorithm.

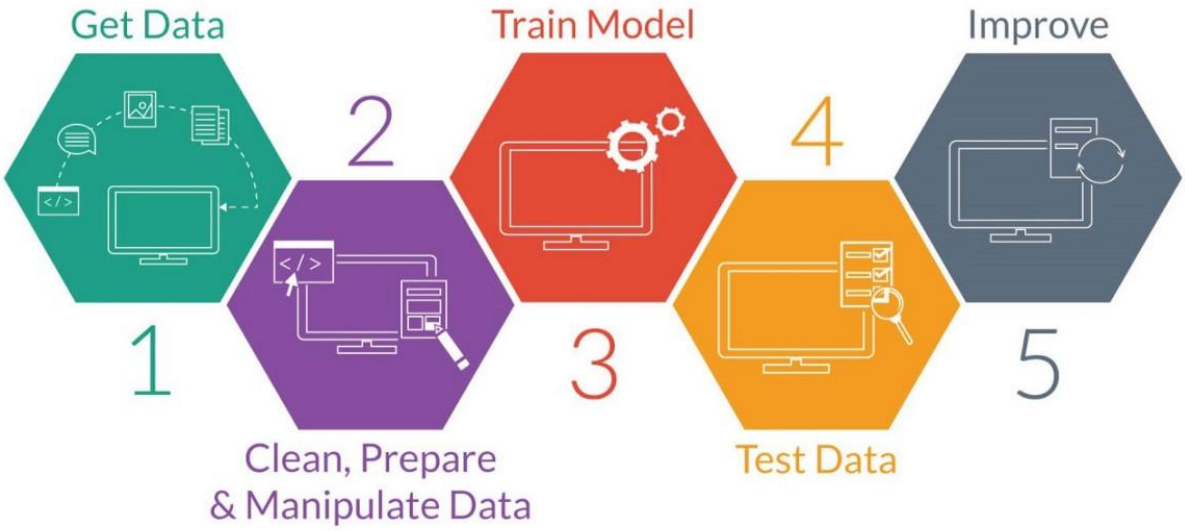

Figure 3: Machine Learning Process

3.4 Data Mining VS Machine Learning

Data Mining	Machine Learning
Has been around since 1930s	Appears in the 1950s
Created to extract rules from vast amounts of data	Teaches a computer how to understand and learn the parameters supplied to it
A research approach for determining a certain outcome based on the sum of the data collected	Teaches a machine to execute complex tasks and uses data and experience to improve its intelligence
Based on large data sets which are then utilised to produce forecasts for corporations and other organisations.	Algorithms are used instead of raw data
Dependent on human involvement and is ultimately designed for human use	It has the ability to train itself and is not reliant on human influence or activities
Data mining will never work unless it is used and interacted with by a real person	Human interaction with machine learning is largely limited to the early setup of algorithms
Data mining needs machine learning	Machine learning does not necessarily need data mining
Static and follows pre-determined parameters	As the correct circumstances arise, the algorithms are adjusted
Simply looking for patterns that already exists in the data	Goes beyond what's happened in the past to predict future outcomes based on the pre-existing data
The rules or patterns are unknown at the	Usually given some rules or variables to

start of the process	understand the data and learn

4.0 ARTIFICIAL INTELLIGENCE (AI)

4.1 Definition

Artificial Intelligence (AI) is a wider branch of computer science concerned with building smart machines capable of performing tasks that require human intelligence. AI leverages computers and machines to impersonate human mind capabilities by solving problems and making a decision. For instance, Siri, Alexa, self-driving cars, robots, etc (BuiltIn, 2022).

4.2 Applications of Artificial Intelligence

There are basically an infinite number of applications for AI but all of them share one goal, which is to replace the human interaction aspect of something and use a computer to mimic it. Below are some examples where AI is majorly used.

Field	Application
Autonomous Vehicles	- Safety and convenient features such as autonomous emergency braking to avoid hitting object in front of the car - Fully autonomous self driving
Gaming	- AI opponents a.k.a bots that mimic a human player to play against in games such as chess.
Business	- Chatbots that replace the conventional human service agent to speed things up

4.3 Career in Artificial Intelligence

Due to AI surging demands in industries, AI promises tons of job opportunities rather than any other career path. Most of the candidates require at least a bachelor's degree in the field listed below:

Field	Career
Computer Science	Big Data Engineer Data Scientist Machine Learning Engineer Research Scientist AI Data Analyst Business Intelligence Developer AI Engineer Robotics Scientist Product Manager
Mathematics	Big Data Engineer Data Scientist Machine Learning Engineer Research Scientist AI Data Analyst Business Intelligence Developer
Engineering	Business Intelligence Developer Robotics Scientist
Data Science	AI Engineer
Statistics	AI Engineer
Robotics	Robotics Scientist
Product Management	Product Manager

Business Administration	Product Manager
Management Sciences	Product Manager

4.4 Machine Learning VS Artificial Intelligence

Machine Learning	Artificial Intelligence
Allows machines to automatically learn from past data without explicitly being programmed. It is basically a subset of AI	Allows machines to simulate human behaviour.
The goal is to learn from data so that it can give an accurate output.	The goal is to make a smart computer which can do complex problem solving like human minds.
Teach machines with data to give an accurate result by performing a particular task.	Make an intelligence system to perform tasks like a human.
Deep learning is the main subset.	Machine learning and deep learning are the main subsets
Limited scope	Very wide range of scope
Perform only specific tasks which they are trained.	Create an intelligent system which performs various complex tasks.
Mainly concerned about accuracy and patterns.	Mainly concerned with maximising the chances of success.
Example of applications: ● Online recommender system ● Google search algorithms	Example of applications: ● Siri ● Customer support using catboats

• Facebook auto friend tagging suggestions	• Expert systems • Online game playing • Intelligent humanoid robot
Types of capabilities: • Supervised learning • Unsupervised learning • Reinforcement learning	Types of capabilities: • Weak AI • General AI • Strong AI
Includes learning and self-correction when introduced with new data.	Includes learning, reasoning, and self-correction.
Deal with: • Structured data • Unstructured data	Deal with: • Structured data • Semi-structured data • Unstructured data

Figure of KDD, DM, ML and AI

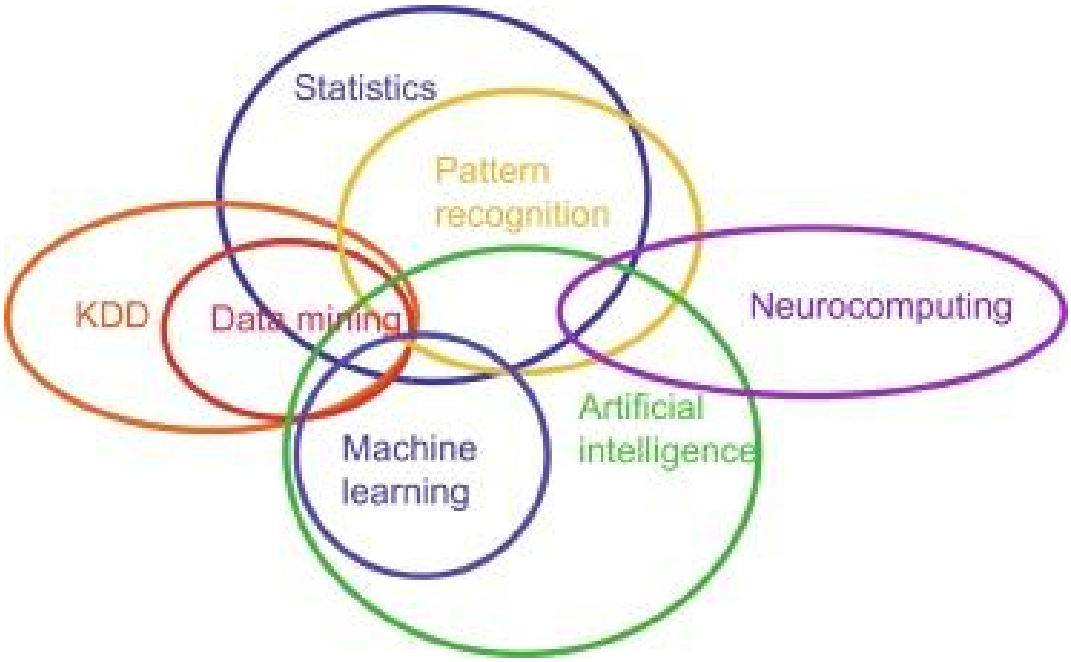

Figure 4: Relationship of KDD, DM, ML and AI

Based on Figure 4, it can clearly be seen that these unfolding technologies are indeed helping humans to solve real problems, either simple problems or complex problems. Simple problems usually require simple analysis which can be solved by using data analysis. Complex problems need complex analysis and this can be solved using machine learning. As shown above, Knowledge Discovery in Database, Data Mining, Machine Learning are the subunits of Artificial Intelligence. These are some of the ways that can be used to solve real problems.

REFERENCES

Arora, S. (2022, February 10). Data Mining Vs. Machine Learning: The Key Difference. simplilearn. https://www.simplilearn.com/data-mining-vs-machine-learning-article#:~:text=Data%20mining%20is%20designed%20to,total%20of%20the%20gathered%20data.

Artificial Intelligence. BuiltIn. (n.d.). Retrieved March 29, 2022, from https://builtin.com/artificial-intelligence

Career Opportunities in Artificial Intelligence: List of various job roles. upGrad blog.(2021, December 7). Retrieved March 29, 2022 from https://www.upgrad.com/blog/career-opportunities-in-artificial-intelligence/

DBD, U. of R. (n.d.). Overview of the KDD process. KDD Process/Overview. Retrieved March 22, 2022, from http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Difference between Artificial Intelligence and machine learning - javatpoint. www.javatpoint.com. (n.d.). Retrieved March 29, 2022, from https://www.javatpoint.com/difference-between-artificial-intelligence-and-machine-learning

Expert.AI Team. (2020, May 6). What is Machine Learning? A definition - Expert System. Expert.ai. Retrieved March 29,2022, from https://www.expert.ai/blog/machine-learning-definition/

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI magazine, 13(3), 57-57.

Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier Gezondheidszorg.

KDD process in data mining - javatpoint. www.javatpoint.com. (n.d.). Retrieved March 22, 2022, from https://www.javatpoint.com/kdd-process-in-data-mining

Machine Learning: A Quick Introduction and Five Core Steps. (2019, April 10). Centric Consulting. Retrieve March 29,2022, from https://centricconsulting.com/blog/machine-learning-a-quick-introduction-and-five-core-steps/

Maimon, O., & Rokach, L. (2005). Introduction to Knowledge Discovery in Databases. In Data Dining and Knowledge Discovery Handbook (pp. 1-17). Springer, Boston, MA.

Marr, B. (n.d.). What Is The Difference Between Data Mining And Machine Learning? Retrieved March 29, 2022, from https://www.scribbr.com/apa-examples/website/

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080.

M.Vineeth. (2020, June 22). MACHINE LEARNING : An Overview (Part-3). Analytics Vidhya. Retrieve March 29,2022, from https://medium.com/analytics-vidhya/machine-learning-an-overview-part-3-515b560b8fc9

Partner, B. (2022, February 28). Major Issues and Challenges in Data Mining. Bench Partner. Retrieved March 22, 2022, from https://benchpartner.com/major-issues-and-challenges-in-data-mining

Potentia Analytics (2019, December 19). What Is Machine Learning: Definition, Types, Applications and Examples. https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/#:%7E:text=These%20are%20three%20types%20of,unsupervised%20learning%2C%20and%20reinforcement%20learning

Raval, K. M. (2012). Data mining techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 2(10).

What is the relationship between machine learning, optimization theory, statistical analysis, data mining, neural networks, artificial intelligence, and pattern recognition? (n.d.). Retrieved March 29, 2022, from https://theaiupdates.blogspot.com/2020/02/what-is-relationship-between-machine.html