# CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission.

|  |  |
| --- | --- |
| Module Title: | Advanced Data Analytics  Big Data Storage and Processing |
| Assessment Title: | Integrated CA1 Sem 2 MSc in Data Analytics |
| Lecturer Name: | David McQuaid  Muhammad Iqbal |
| Student Full Name: | Jose Luis Alvarado Carranza |
| Student Number: | 2020092 |
| Assessment Due Date: | 06/09/2023 |
| Date of Submission: | 06/09/2023 |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Technical demonstration:**

Demonstrate a practical example of your chosen area of interest that should include Big Data and Advanced Data Analytics. This will be a coded example and will require you to upload your code in a Jupyter notebook.

Demonstrate your research along with your practical deployment using a video recording (5 – 7 minutes) or Screen cast. All code must in a Jupyter notebook and be uploaded as a zip file on Moodle. [0 - 20] Your paper should include:

* Overview of the chosen topic, including objective statement and Research Question. Presentation of state of the art, including research methodologies and key of the papers you reviewed. [0 - 20]
* Literature review based on the chosen papers for the topic of your choice including proper citations and Harvard style referencing. [0 - 15]
* Critical evaluation of the key findings, specifically their implications and limitations, and highlighting any contradicting viewpoints and research gaps. [0 - 30]
* Conclusions you have drawn based on your research. [0 - 15]

ABSTRACT

KEYWORDS

INTRODUCTION – An overview

DEEP LEARNING USING BIG DATA

Deep learning and big data. We are going to find a definition to these concepts, a bit of history and how the research made in this assessment relates one topic to the other. Additionally, we will try to gather enough concepts to be able to apply what has been learned in class, following the tutorials as much as needed.

Definition

DEEP LEARNING

Deep learning is defined by many different literature pieces as the inference of model parameters for decision making in a process mimicking the understanding process in the human brain. In other words, we can say that deep learning is a way of data inference in machine learning. These tools are among the main tools of artificial intelligence. (Hassaballah, 2020). However, this definition does not appear out of the thin air. Deep learning is a subfield of machine learning.

A diagram of a machine learning

Description automatically generated

Figure : Visual representation of Deep Learning

Deep Learning is a technique that learns features and tasks directly from data. It happens by running inputs through artificial neural networks which in turn have hidden layers where the data are processed allowing the machine to go deep in its learning. Making connections and weighing inputs for the best results.

The main difference between one another is that Machine Learning uses structured, labelled data to make predictions. This means that specific features are defined from the input data for the model and organized into tables (IBM, 2023). This is dependent on the type of data used, structured, semi structured, and unstructured. It generally means that there needs to be a pre-processing of data to organize it into a structured format. Whereas Deep Learning eliminates some of data pre-processing that is typically involved with machine learning. The algorithms can ingest and process unstructured data, like text and images, and it automates feature extraction, removing some of the dependency on human experts. (IBM, 2023).

To accomplish this task, Deep learning uses a combination of data inputs, weight, and biases. These elements work together to accurately recognize, classify, and describe objects within the data.

It is important to point out that the algorithms used in Deep Learning have existed for a long time. They have recently gained more importance because data have become much more pervasive and so, to effectively be put to use, massive amounts of data were required for its implementation.

The history of deep learning can be traced back to 1943, when Walter Pitts and Warren McCulloch created a computer model based on the neural networks of the human brain. (Dataversity, 2022)

They used a combination of algorithms and mathematics they called “threshold logic” to mimic the thought process. Since that time, Deep Learning has evolved steadily, with only two significant breaks in its development. (Dataversity, 2022)

Additionally, we now have hardware and architecture that are capable of handling the vast amounts of data and computational power these algorithms require. In short, the necessary technology and data that was needed to utilize these algorithms were not available a few decades ago. What is more, building and deploying these algorithms have become streamlined with open source. (Microsoft, 2023)

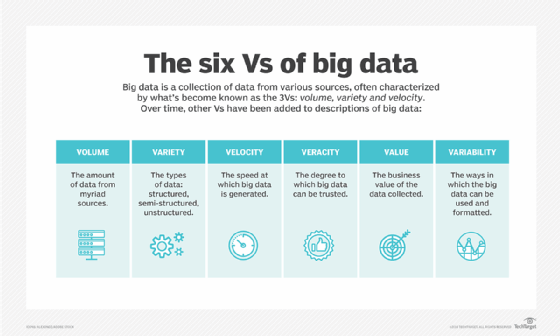
For instance, you can access the latest and most advanced features, algorithms, and models without spending time and money on developing them yourself. You can also learn from the code, documentation, tutorials, and examples of how to implement, test, and optimize various machine learning tasks and challenges. Additionally, you can get feedback, support, and collaboration from other users and developers who can help you solve problems or improve your code. Finally, you can contribute to the improvement and innovation of the open-source machine learning ecosystem by sharing your code, data, or insights with others. (Linkedin, 2023).

BIG DATA

Big data is a concept describing data sets that exceed the size that can be managed by traditional tools. It’s defined by six Vs: variety, volume, veracity, value, variability, and velocity to the growing variety of data sources that arrives in increasing volumes and with more velocity (the high rate at which data is received and acted on). (Oracle, 2022).

Although big data doesn't equate to any specific volume of data, big data deployments often involve terabytes, petabytes and even exabytes of data created and collected over time.

Why is big data important Companies use big data in their systems to improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profits. Businesses that use it effectively hold a potential competitive advantage over those that do not because they are able to make faster and more informed business decisions.

 Figure 2: Vs of big data.

For example, big data provides valuable insights into customers that companies can use to refine their marketing, advertising, and promotions in order to increase customer engagement and conversion rates. Both historical and real-time data can be analyzed to assess the evolving preferences of consumers or corporate buyers, enabling businesses to become more responsive to customer wants and needs. (Target, 2020)

Big data is also used by medical researchers to identify disease signs and risk factors and by doctors to help diagnose illnesses and medical conditions in patients. In addition, a combination of data from electronic health records, social media sites, the web and other sources gives healthcare organizations and government agencies up-to-date information on infectious disease threats or outbreaks.

To get valid and relevant results from big data analytics applications, data scientists and other data analysts must have a detailed understanding of the available data and a sense of what they are looking for in it. That makes data preparation, which includes profiling, cleansing, validation, and transformation of data sets, a crucial first step in the analytics process.

Once the data has been gathered and prepared for analysis, various data science and advanced analytics disciplines can be applied to run different applications, using tools that provide big data analytics features and capabilities. Those disciplines include machine learning and its deep learning offshoot, predictive modelling, data mining, statistical analysis, streaming analytics, text mining and more.

This is where we stop and think about what we want to get from this research paper.

I want to get a passing score in this project as a whole.

I want to be able to use deep learning regardless of what the data I find.

I want to be able to use at most 2 different examples of using big data.

Deep neural networks consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization. This progression of computations through the network is called forward propagation. The input and output layers of a deep neural network are called *visible*layers. The input layer is where the deep learning model ingests the data for processing, and the output layer is where the final prediction or classification is made.

Another process called backpropagationuses algorithms, like gradient descent, to calculate errors in predictions and then adjusts the weights and biases of the function by moving backwards through the layers in an effort to train the model. Together, forward propagation and backpropagation allow a neural network to make predictions and correct for any errors accordingly. Over time, the algorithm becomes gradually more accurate.

The above describes the simplest type of deep neural network in the simplest terms. However, deep learning algorithms are incredibly complex, and there are different types of neural networks to address specific problems or datasets. For example,

* [*Convolutional neural networks (CNNs),*](https://www.ibm.com/topics/convolutional-neural-networks)used primarily in computer vision and image classification applications, can detect features and patterns within an image, enabling tasks, like object detection or recognition. In 2015, a CNN bested a human in an object recognition challenge for the first time.
* [*Recurrent neural network (RNNs)*](https://www.ibm.com/topics/recurrent-neural-networks)are typically used in natural language and speech recognition applications as it leverages sequential or times series data.

(and neural networks make up the backbone of deep learning algorithms. It’s the number of node layers, or depth, of neural networks that distinguishes a single neural network from a deep learning algorithm, which must have more than three. (IBM, 2023) )

There used to be a time when only companies had data.

Nowadays, we have all become producers of data. This is evident to us in our daily lives. Every time we buy a product, watch content on any media platform or simply when we drive around our hometown, we generate data. Moreover, we not just generate data, we consume it too, and so companies rely heavily on the data we generate.

Most of the people want to have products and services that fit their needs. So, companies needed to be able to predict which products are likely to be bought by which customer and thus, maximizing sales and profits.

This keeps true regardless of the market. For example, any store being online or physical which sells all kinds of goods to its customers stores details of each transaction. Details such as date, amount, product sold, total money spent, etc.

This information is the one being used by companies and this information becomes huge when talking about a multinational.

In order to gain meaningful insights from data, we will perform a rigorous step-by-step process.

Defining the question

The first step for us will be to define the objective of the analysis, which is a ‘problem statement’. Essentially, we are asking a question with regards to a problem we are trying to solve.

For our case, we will try to answer the question of how do we use deep learning to process the data from food brands in the US.

A screenshot of a computer program

Description automatically generated

A screen shot of a computer

Description automatically generated