# 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.

Contents

[CCT College Dublin 1](#_Toc147527179)

[INTRODUCTION – An overview 3](#_Toc147527180)

[*DEEP LEARNING* 3](#_Toc147527181)

[*BIG DATA* 4](#_Toc147527182)

[Defining the question 5](#_Toc147527183)

[*Critical evaluation of the key findings* 8](#_Toc147527184)

[*Further Reading* 8](#_Toc147527185)

[*Conclusions* 9](#_Toc147527186)

[*Bibliography* 9](#_Toc147527187)

[*References* 9](#_Toc147527188)

Integrated CA1 Sem 2 MSc in Data Analytics: Deep Learning and Big data

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***Jose Luis Alvarado Carranza***

***Faculty of ICT***

***CCT College Dublin***

***County Dublin***

***2020092@student.cct.ie***

***ABSTRACT –*** Nowadays big data and deep learning are intertwined so it becomes of the utmost importance to learn its use as a data analyst and so this report shows two different examples of their use with real case studies proposed by the authors of this paper. It details the procedures followed and discussion of the results obtained.

***KEYWORDS: Hadoop, Keras, TensorFlow, Recurrent Neural Networks, Datanodes, Namenodes.***

# INTRODUCTION – An overview

Deep learning and big data constitute a big part of our daily lives. We use them intrinsically in the many apps we use. Human beings as well as companies around the world depend largely on their applications and they both benefit our lives while they influence our decision making. They establish a well-thought-out process so much that they have become the bedrock in which we set our future.

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.

## *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 1: 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).

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.

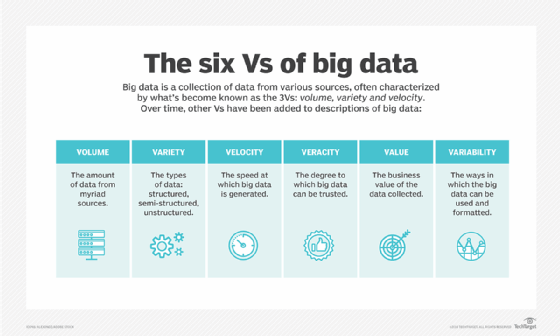
and neural networks make up the backbone of deep learning algorithms. It’s the number of node layers, or depth, of neural networks that distinguishes a single neural network from a deep learning algorithm, which must have more than three. (IBM, 2023).

## *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.

Now that we have established some definitions, we are going to start focusing on our analysis. Therefore, we want to define what exactly we want to gain from this continuous assessment.

There is a clear corelation between Big Data and Deep Learning as we have established previously. What we are going to do is provide two different scenarios in which big data and deep learning can help find a solution to a given problem. Therefore, we are going to use the techniques practiced in class to find an appropriate solution to a problem we establish ourselves.

There used to be a time when only companies had data but 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 we use big data tools to process the data from food brands in the US.

To put our understanding of deep learning to the test and obtain more understanding on the topic at hand, we are undertaking the study of minimum wages in Europe.

In case of big data, we have found a database which contains many different types of foods and their ingredients. This database was found in the FDA website. This data is over 800 MBs and contains all the products names, ingredients, amounts of ingredients in different units, brand names, etc., of all the products that have been approved by the FDA and therefore are regulated for human consumption.

In this scenario, we have been tasked to make a list of all the ingredients of the products. To do so, we are going to use what we have learned in class. We have downloaded the 800MBs file and get the information from there.

The file itself contains more than ten thousand lines of ingredients and to count each ingredient would take a lot of time and resources.

All this information has been gathered for a long period of time which makes its processing complicated. If we were to tackle this problem with traditional tools the system would crash.

A screenshot of a computer

Description automatically generated

Figure 3: System crashing

Our strategy is to utilize Hadoop data framework and sort the number of times a word is repeated and come up with the desired list of ingredients.

Hadoop is a framework used to store and process large datasets across computer clusters. It can be scaled from single computer system up to thousands of commodity systems that offer local storage and compute power. It can ingest big amounts of data easily by using shell. Hadoop is open source which in turn means that there is no cost for the software. Since the dataset we are analysing is static, Hadoop would be our option.

There are many other useful tools like Apache Spark. Spark is fast because it has in-memory processing. Additionally, it delivers near real-time analytics. However, due to the nature of the task at hand, it is more convenient to use Map Reduce to tackle the problem. This is, we just need a list of the ingredients.

Now, for our case study we had to make certain alterations to the dataset to be able to come up with a satisfactory result.

As seen in the figure 3 above, the data set used could not be opened by traditional means which made it difficult to gather the data. So, we decided to create a text file which includes a list of one thousand lines of ingredients. Then, we used Hadoop and downloaded the results from our cluster.

A screenshot of a computer program

Description automatically generated

Figure 4: Running Hadoop.

The figure 4 above shows that:

* Hadoop is running and to verify that we used the code JPS on our shell.
* The files are stored in a folder on the desktop called 2020092.
* In the same folder Mapper and Reduce are found.

Once the file has been uploaded to Hadoop, the file gets distributed among its namenodes and we use the Mapper and Reducer to get the desired results.

A screen shot of a computer

Description automatically generated

Figure 5: Obtaining results.

To verify the results, we access the Hadoop localhost and find the next figure.

A screenshot of a computer

Description automatically generated

Figure 6: Hadoop Directory.

In this figure 6 we can see that the file has been properly uploaded and sorted as per the requested task. The file has later been downloaded and included as part of the report and shared using Github with the name “ouput2020092”.

For our case study we also need to address the topic of Deep Learning. We have decided to approach it with a different example. This is because the data we found on the FDA branded foods does not allow us to make the best of recurrent neural networks as they are mainly suited for problems that involve sequential data. So, we found a different data set and we have tasked ourselves with predicting minimum wages in Europe.

This topic is relevant to us because firstly, we live in the European region. Secondly, it is widely accepted that minimum wages are an effective way to help workers at the lower end of the income distribution without damaging their job opportunities. The main concern that economists have is that there may be job loses if the minimum wage is set at too high a level. Firms will only employ workers if they think that the value of what they produce is greater than what the worker costs. If the minimum wage makes labour too expensive, there will be fewer jobs. (Observatory, 2022). So, being able to make a prediction could be beneficial to both saving jobs and protecting workers.

This data set includes data from many European countries as well as other countries not part of the European block like Argentina, Australia, Brazil, and many others which we will disregard for the time being. These data date back to the year 2000 and it is divided in semesters, and it ends in January of 2023. Thus, it is split in 48 semesters. It shows a continuous growth with the passing of time as shown in the figure.

A graph of the number of countries/regions

Description automatically generated

Figure 7: Statuary Minimum wages in Europe

In figure 7 we can appreciate:

* The countries ordered to highest to lowest in terms of their minimum wage.
* Luxembourg being the country with the highest minimum wage per hour in Europe at €13.80/hour while Bulgaria being the lowest at €2.41/hour.
* A steady growth of the minimum wage with the passing of time.

We are going to try to predict what would be the minimum wage in a year. We begin by loading the data set to a new Jupyter Notebook we have called 2020092AdvDA, as per the assessment requirements. We are going to use the notes given in class and follow them to the best of our knowledge and analyse the results.

It is worth mentioning that the data set was also too big to be handled by our tools and so we modified it to be able to work with it. As shown in the next figure.

A screen shot of a computer

Description automatically generated

Figure 8: Manipulation of the data set

The modification made was to choose the data we want to focus on. So, two sub datasets were made. Therefore, delimiting our research and setting us up to get the best results possible. Additionally, the processing times were diminished significantly as the data set was not as big as the original one.

This is reflected on the files uploaded. The file currestat.csv, shows the current minimum wage data available for Europe in between the year 2000 and 2023. This data set was used to train our recurrent neural network. The next file mweu.csv represents the test data.

In this part of the assessment, we focus on Keras API module and TensorFlow.

Keras is a high-level API that sits on top of three lower-level neural network APIs and acted as a wrapper to these lower-level libraries referred to as Keras backend engines (Keras, 2023).

While TensorFlow is an **infrastructure layer for differentiable programming**, dealing with tensors, variables, and gradients, Keras is a **user interface for deep learning**, dealing with layers, models, optimizers, loss functions, metrics, and more.

Keras serves as the high-level API for TensorFlow: Keras is what makes TensorFlow simple and productive.

TensorFlow is an end-to-end open-source platform for machine learning. TensorFlow is a rich system for managing all aspects of a machine learning system; however, this class focuses on using a particular TensorFlow API to develop and train machine learning models. (Flow, 2023).

The Layer class is the fundamental abstraction in Keras. A Layer encapsulates a state (weights) and some computation (defined in the call method). (Keras, 2023)

Since the Keras Module of TensorFlow only accepts arrays, we transform the data we want to use into an array. Then we will need to specify the number of timesteps. As we have 24 years divided into semesters, out timesteps will be of 48. Then we proceed to verify that our data is in an array that can be taken by our recurrent neural network.

We use the sequential class in TensorFlow in order to initialize the code of our recurrent neural Network. We also specify the units, return sequences, and shape of our input. Additionally, we need to include a rate that indicates how many neurons should be dropped in a specific layer of or neural network. To do so, we add a “dropout rate”. We defined our dropout as “0.2”.

We finish our recurrent neural network by setting an output layer. So, we use the Dense class. We set it to 1 as we only want the output of the next semester.

We specify important parameters for our recurrent neural network which are the optimizer (adam) and loss (mean squared error).

We proceed to train our recurrent network. We specify the variables using our data from the dataset, set epochs and the size of the batch.

In our study case, the training model results in an unexpected value. The loss is represented by a NaN (Not a number) which prevented us from advancing our study. However, using other techniques, resources, and the same data we were able to predict. Said results were not added to this report as is not in the scope of the course. We think that the obtained result is due to the imploding vanishing gradient problem.

Being that as it may, we were able to get results even it they were not the ones desired.

## *Critical evaluation of the key findings*

Throughout our assessment the importance of Big Data and Deep learning has been evident. These concepts are the building blocks to our future and learning how to use them represents the key to open new opportunities. So, with the continuous use of these technologies there is no doubt in m mind that humanity can strive for bigger and better things. That being said, all the aforementioned does not come without caveats.

While the tools and resources used in this continuous assessment are very powerful, they are not for everyone as they need a steep learning curve. We experienced this in trying to use Hadoop. It requires the user to learn a whole new set of commands to operate as it is used through Ubuntu (Linux) and it does not have a graphic design. Furthermore, in our study we were able to sort and count words, which is the equivalent to coding a “hello world” in other data frameworks. If something more complex needed to be done, there is a need to learn how to code through the use of java. That on its own is a whole different set of skills and understanding of other concepts not dealt with in the scope of this course.

Another downside to it is that since it does not have a GUI it is so easy to make mistakes when entering the commands.

In terms of our results on Big Data, since the main objective of this course is to be exposed to these technologies, I consider our findings a success because we did what we set ourselves to do. The tutorials were of major help and hopefully all these new concepts and tools become inscribed in our minds.

Now, in regard to our Deep learning exercise, We were not able to circumvent the imploding vanishing gradient. So, our results were not the ones we expected and if this were a different scenario, other tools could have been put to use to find a solution. Like we did but chose not to add in this report as is not part of the objectives of the course. Additionally, we understand that doing so defeats the purpose of this exercise and the course. Ideally there would be more examples and more in-depth explanations of how to properly manipulate the tools used in this study case.

## *Further Reading*

In our big data study case, we covered how to use Hadoop to sort out and count ingredients approved by the FDA and created a list. If there we had been able to use a different data set, a data set that included the amounts consumed by US citizens perhaps we could have been able to use that data and tried to predict how much of those ingredients would be consumed by a future date.

In our Deep Learning study case, if we had been able to predict at least one semester, it would have been interesting to find out the sustainability of increasing the minimum wage every year.

## *Conclusions*

* Hadoop is a very capable tool with a steep learning curve. This was evidence throughout our study because it depends on commodity hardware. However, this completely overshadowed by its non-existent cost.
* In our study of Deep learning, it became difficult to predict due to the nature of our recurrent neural network. This occurred because it needed an additional dataset that could be used as a testing data model in order to predict. If there were the cases that there is no more data available, more complex techniques have to be used.
* TensorFlow consistently takes longer times to train a model. This was evident to us in our study when we set our epochs to 100.
* Even though TensorFlow seems user-friendly, it adds a layer of complexity to the training models. Every code needs some platform or API (Keras) for its execution which increases dependency.

# *Bibliography*

Dataversity, 2022. *https://www.dataversity.net.* [Online]   
Available at: https://www.dataversity.net/brief-history-deep-learning/  
[Accessed 30 09 2023].

Flow, T., 2023. *https://www.tensorflow.org/.* [Online]   
Available at: https://www.tensorflow.org/  
[Accessed 06 10 2023].

Hassaballah, M. a. A. A. e., 2020. *Deep learning in Computer VIsion: principles and applications..* Boca Raton, FL: Taylor & Francis Group.

IBM, 2023. *https://www.ibm.com.* [Online]   
Available at: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/  
[Accessed 30 09 2023].

Keras, 2023. *https://keras.io/.* [Online]   
Available at: https://keras.io/why\_keras/  
[Accessed 06 10 2023].

Linkedin, 2023. *https://www.linkedin.com/advice.* [Online]   
Available at: https://www.linkedin.com/advice/0/what-can-you-learn-from-using-open-source-machine  
[Accessed 30 09 2023].

Microsoft, 2023. *https://azure.microsoft.com/.* [Online]   
Available at: https://azure.microsoft.com/en-us/solutions/ai/data-scientist-resources/  
[Accessed 30 09 2023].

Observatory, E., 2022. *https://www.economicsobservatory.com/.* [Online]   
Available at: https://www.economicsobservatory.com/minimum-wages-and-living-wages-what-happens-in-times-of-inflation  
[Accessed 06 10 2023].

Oracle, 2022. *The Evolution of Big Data and the Future of the Data Platform.* [Online]   
Available at: https://www.oracle.com/ie/a/ocom/docs/big-data/big-data-evolution.pdf  
[Accessed 05 10 2023].

Target, T., 2020. *https://www.techtarget.com/.* [Online]   
Available at: https://www.techtarget.com/searchdatamanagement/definition/big-data  
[Accessed 05 10 2023].

# *References*

Dataversity, 2022. *https://www.dataversity.net.* [Online]   
Available at: https://www.dataversity.net/brief-history-deep-learning/  
[Accessed 30 09 2023].

Flow, T., 2023. *https://www.tensorflow.org/.* [Online]   
Available at: https://www.tensorflow.org/  
[Accessed 06 10 2023].

Hassaballah, M. a. A. A. e., 2020. *Deep learning in Computer VIsion: principles and applications..* Boca Raton, FL: Taylor & Francis Group.

IBM, 2023. *https://www.ibm.com.* [Online]   
Available at: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/  
[Accessed 30 09 2023].

Keras, 2023. *https://keras.io/.* [Online]   
Available at: https://keras.io/why\_keras/  
[Accessed 06 10 2023].

Linkedin, 2023. *https://www.linkedin.com/advice.* [Online]   
Available at: https://www.linkedin.com/advice/0/what-can-you-learn-from-using-open-source-machine  
[Accessed 30 09 2023].

Microsoft, 2023. *https://azure.microsoft.com/.* [Online]   
Available at: https://azure.microsoft.com/en-us/solutions/ai/data-scientist-resources/  
[Accessed 30 09 2023].

Observatory, E., 2022. *https://www.economicsobservatory.com/.* [Online]   
Available at: https://www.economicsobservatory.com/minimum-wages-and-living-wages-what-happens-in-times-of-inflation  
[Accessed 06 10 2023].

Oracle, 2022. *The Evolution of Big Data and the Future of the Data Platform.* [Online]   
Available at: https://www.oracle.com/ie/a/ocom/docs/big-data/big-data-evolution.pdf  
[Accessed 05 10 2023].

Target, T., 2020. *https://www.techtarget.com/.* [Online]   
Available at: https://www.techtarget.com/searchdatamanagement/definition/big-data  
[Accessed 05 10 2023].