# Task One – Narrative

(Plan)

Chosen Scenario

NHS Covid-19 Test Data & Vaccine Data

Approaches used in big data analysis

* Modelling
  + Advantages
    - Allows for easy understanding and use of the data we have
  + Disadvantages
    - Overly Approximate relationships can lead to inaccurate or vague conclusions
* Visualization
  + Advantages
    - Easier to spot trends for the average person as well as data scientists
  + Disadvantages
    - Inaccurate sets of data can badly affect the visualisation of the data
* Machine Learning
  + Advantages
    - More data often means better learning and more accurate predictions or diagnostics <https://ieeexplore.ieee.org/abstract/document/7906512>
  + Disadvantages
    - Traditionally, ML algorithms were designed with “the assumption that entire dataset can fit in memory”. <https://ieeexplore.ieee.org/abstract/document/7906512>
* Data Mining
  + Advantages
    - Efficient way to extract trends from data
    - Good at predicting future trends like machine learning and making prediction models
  + Disadvantages
    - Demanding. Requires large amounts of hardware.

Sources:

<https://www.sciencedirect.com/science/article/pii/S0268401219302890>

(Actual Research Review)

## Introduction

In this research review, I will consider four approaches to big data analysis. Each method will be explained, advantages and disadvantages of each will be considered as well as fundamental limitations all in the context of a problem scenario. The problem scenario chosen is the processing of vaccination and Covid-19 data during the ongoing pandemic done by the NHS in the UK.

## Data Modelling

In big data, Data modelling is stated to be “methods of fast and cost-effective mathematical analysis with approximate relationships between variables” (Galetsi, Katsaliaki, & Kumar, 2020). In our problem scenarios’ case, this can be finding links in our data set between the severity of covid cases, and the vaccinations. The data, once modelled, could indicate that people that have had both doses of a vaccine and a booster dose have a reduced chance of hospitalisation following catching covid, for example.

Data Modelling has the advantage of allowing for easy understanding of data and the relationship between it and other data entities. It can help companies understand their data better to establish how it should be used and for what purpose (Mullins, 2021). However, if the relationships that are made are overly approximate or broad, this can lead to conclusions that are inaccurate or vague which could impact the results taken from that data and the understanding gained from analysing it.

This method has the strength is that it can be used to easily gain a better understanding of the data that the health sector has in our case. This can be useful to put the necessary measures in place for Covis-19 for example. However, the links made within the variables in the data have to be accurate. Otherwise, findings from the data can influence potentially bad decisions being made which is one of the limitations.

## Data Visualization

Another approach is Data visualisation. This can be simply defined as “Creating tables, images, diagrams, and other intuitive ways to understand data” (Galetsi, Katsaliaki, & Kumar, 2020). In context, this can be interpreted to be graphs to show the current daily positive cases for covid, for example or it also could be a diagram to illustrate the vaccine uptake in the population including first, second, and booster doses.

An advantage of this method is that it makes it much easier for the average person to understand. Other methods tend to be much less accessible to the public as the trends are less obvious often appearing in numbers rather than being illustrated as seen here. On the flipside, trends can be skewed due to inaccurate data. This is true for most big data techniques but especially so here as it is much more obvious in graphs and charts and can appear more skewed than in other approaches. Furthermore, some can only look at the important data without taking into consideration the bigger picture. This can lead to biased results in cases.

Overall, this method’s strengths lie in the ease of illustrating information and sharing it with almost anyone which is important when thinking about Covid-19 data. However, its limitations are that it is sensitive to inaccurate data which must be removed before the data can be processed. It can also lead to biased results as discussed formerly.

## Machine Learning

Thirdly, we have the machine learning approach. It is defined as “AI aimed to design algorithms that allow computers to evolve behaviours based on empirical data” (Galetsi, Katsaliaki, & Kumar, 2020). These algorithms “learn from data and provide data driven insights, decisions, and predictions” (L'Heureux, Grolinger, Elyamany, & Capretz, 2017). Such Algorithms could be used to predict how the pandemic will advance based on the trends of data that are present in the dataset.

An advantage of ML is the “presumption that algorithms can learn better with more data” (L'Heureux, Grolinger, Elyamany, & Capretz, 2017). This is perfect for big data as they have access to large datasets and therefore can make more accurate predictions and insights. This comes with drawbacks, however. Traditional ML algorithms “were designed for smaller datasets” (L'Heureux, Grolinger, Elyamany, & Capretz, 2017) assuming that “the entire dataset can fit in memory”. In big data, this is not practical nor true as the datasets are simply too sizable.

This method’s strengths lie in the fact that the more data it is supplied with. The better the predictions are. This is important when predicting data for any application including our own scenario. However, while the memory can cause a limitation as discussed previously, some of these have been alleviated with “paradigms” and “frameworks” such as MapReduce and Hadoop (L'Heureux, Grolinger, Elyamany, & Capretz, 2017)

## Data Mining

Data mining can be defined as “A set of techniques that extract information about data” (Galetsi, Katsaliaki, & Kumar, 2020). The most popular techniques used in data mining are “decision trees, artificial neural networks and support vector machines”. In our chosen scenario, this could be used to make a prediction models for the survivability of Covid-19 based on certain conditions on may be suffering from and vaccinations status.

Data Mining has the advantage that it is a very efficient way of extracting trends from data in the sample. This means that it can be scaled up by large amount without having the processing time go up significantly which is very important in big data. As well as that, like machine learning, it is a very good way of making predictions based on the trends extracted from the data and models based upon this such as prediction models. However, Data mining is very hardware demanding and requires a large amount of it. This means that it is not very accessible.

This method’s strengths rely on the fact that it is very efficient and accurate and that it is a good way of making predictions which can be relied upon to make decisions. This is very important in our scenario. However this method isn’t very accessible which can be a limitation. This could, on the other hand, be solved by outsourcing.

## Summative Table

|  |  |  |
| --- | --- | --- |
| Big Data Approaches | Advantages | Disadvantages |
| Data Modelling | Easy and accessible understanding of data | Conclusions can be vague or inaccurate if the trends are overly approximate |
| Data Visualization | Easy to spot trends for the average person | Can be badly affected by inaccurate data fields |
| Machine Learning | More data means more accurate predictions made | Traditionally, ML algorithms were designed with “the assumption that entire dataset can fit in memory”. |
| Data Mining | Efficiency at extracting data trends  Good at making predictions from data | Demanding. Requires large amounts of hardware |

# Task Two – Analysis

## Section I: Data Summary, Understanding and Visualisation

### Task 1; Missing Values:

In this dataset there was found to be no missing values from any fields in the dataset as shown in Fig 1 below. To achieve this, each column had a statement as shown in Fig 1 to count the null values in each column and display the results.

As no missing data was found, this means that the dataset doesn’t have to be altered before we can start processing it. In the event that there was missing data, there are a few ways to remedy this. The first way would be to insert data into the missing fields to complete them. However, this means that the extra new data in the dataset has to be accurate which is difficult in this case as, unless we take new readings which would mean a new dataset, we cannot do this accurately enough. The second approach would be to just simply delete the rows which have missing data on them. This allows us to have a complete dataset minus the empty rows. While this means that we will have less data to work with, a complete dataset is required to extract trends and make accurate predictions.

### Task 2; Summary Statistics:

In this section,

### Task 3; Correlation Matrix:

## Section 2: Classification

# Bibliography

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big Data analytics in health sector: Theorical Framework, Techniques and Prospects . *International Journal of Information Management* , 206-216.

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017, April 20 ). *Machine Learning with Big Data: Challenges and Approaches.* Retrieved from IEEE Explore: https://ieeexplore.ieee.org/abstract/document/7906512

Mullins, C. S. (2021, May 3). *The Importance of Data Modeling in a Big Data World.* Retrieved from Database, Trends and Applications: https://www.dbta.com/Editorial/Think-About-It/The-Importance-of-Data-Modeling-in-a-Big-Data-World-145915.aspx