*The application of unsupervised clustering for classification of brain tumour, breast cancer, and housing data with the use of K-means, Mean-shift, and DBSCAN with implementation of hyperparameter tuning*

Contents

[***Abstract*** 3](#_Toc101733617)

[***Introduction*** 4](#_Toc101733618)

[***What is machine learning?*** 5](#_Toc101733619)

[***Literature Review*** 6](#_Toc101733620)

[**Supervised clustering** 6](#_Toc101733621)

[**Unsupervised clustering** 8](#_Toc101733622)

[**Literature Conclusion** 11](#_Toc101733623)

[***The artefact*** 12](#_Toc101733624)

[**Introduction** 12](#_Toc101733625)

[***The data*** 15](#_Toc101733626)

[***Testing the artefact*** 18](#_Toc101733627)

[***K-means*** 18](#_Toc101733628)

[***Mean-shift*** 24](#_Toc101733629)

[***DBSCAN*** 28](#_Toc101733630)

[***Results*** 31](#_Toc101733631)

[***Conclusion*** 33](#_Toc101733632)

[***Future work*** 33](#_Toc101733633)

[***Project management*** 35](#_Toc101733634)

[***Bibliography*** 37](#_Toc101733635)

# ***Abstract***

*This project researches the effectiveness of unsupervised learning to classify and understand brain tumour data. During research the application of the K-means, Mean-shift, and DBSCAN unsupervised learning algorithms were used on brain tumour, breast cancer data and housing data to find which was the most effective. It was found that the K-means algorithm was the most accurate with a 66.8% accuracy while the Mean-shift and DBSCAN algorithms were unable to create a reliable and accurate model. As the data used was not suitable and did not allow for a classification to be made.*

# ***Introduction***

The term ‘AI’ or ‘artificial intelligence’ is an umbrella term used to define different approaches for a computer to demonstrate a level of intelligence. AI is a simulation of the computer’s intelligence compared to a human’s ability to understand a problem, create a solution, and execute tasks. However, one main method of AI is machine learning, which can be defined as “computational methods using experience to improve performance or to make accurate predictions” (Mohri, Rostamizadeh and Talwalker, 2018). The computer’s given tasks can range from large scale problems such as the use of autonomous vehicles that use image processing to a simple problem such as facial recognition. The difference between the two is that the autonomous vehicles need to be able to process many sources of information and interpret data while facial recognition only must perform one task. Due to this fact AI has branched into many other subsections as it can have diverse purposes that affect the way data is interpreted.

Machine learning has been used to create many methods to sort data, including algorithms such as text-processing, which has been used in the detection of spam emails and classifying of a webpage. An algorithm can detect features such as, titles, subheading, text, images, graphs, maps, diagrams, and equations (Charles and Yasotha, 2016) that are added to the document or webpage. This allows for large amount of data to be automatically categorised and sorted by an algorithm that can learn and determine the type of document or webpage that it is viewing. The computer uses this experience to teach itself and improve its methods based upon the dataset provided or via interaction with the environment. The dataset gathered is used as a ‘training set’ for the computer to create more accurate predictions (Géron, 2019). Allowing the computer to train itself using a dataset allows for further improvement, this means that using an active environment such as Amazon’s Alexa/Google Assistant can greatly improve the user’s experience as the algorithm is able to understand the user’s preferences and learn accordingly.

Data mining is another tool used by machine learning to process and sort information that otherwise performed by a human is a complicated and long task. A machine learning classification algorithm takes large datasets and categorise them into a set of results. This method of machine learning is used in the stock market to make future predictions based on current trends. This is done with a large dataset of the past trends of the company. This is then analysed to make a prediction for the future trends “since the data is in large volume and it can’t be processed instantly so machine learning algorithms are used for analyses” (Vazirani, Sharma and Sharma, 2020) and using these predictions allows for future investments to be made. Although the algorithm predicts future trends this does not mean that they will be 100% correct, due to the different types of machine learning techniques and algorithms used.

With the many methods of machine learning this implies that there are many different algorithms that are all able to do different things, therefore, there is no one definitive algorithm meaning that the task that must be completed needs to be paired with the correct algorithm. Consequently, careful planning is needed before starting a project as to ensure that the correct method is picked.

## ***What is machine learning?***

Machine learning is a subsect of Artificial Intelligence. It can be explained as a computer’s ability to learn behaviour and problem solve. Machine learning’s early beginnings included playing checkers; this was achieved by creating an algorithm where “the computer plays by looking ahead a few moves and by evaluating the resulting board positions” (Samuel 1959). The computer was able to look at some of the possibilities of the opponent and calculate which move was the best to make for highest odds of winning. By saving all the board positions encountered during play, the computer could predict the outcome because it could process successful board positions and translate it into moves. Sorting the data while playing is a form of environmental learning as the computer stored the information needed to learn during play. Modern machine learning programs still use this method, but the computer is also trained using training datasets.

Machine learning consists of “designing efficient and accurate predictions algorithms” (Mohri, Rostamizadeh and Talwalker, 2018). This is done with the use of data because the more data the computer has, the more accurate the prediction will be. This data is stored in the form of a dataset, which is used as training data for the model to learn and predict new, unknown data presented to the algorithm. By storing data, this allows for the computer to look at the new data from the current environment it is in and compare that to the training data. By doing this, the algorithm made can compare and predict the best possible outcome. This can be applied to many forms of data including text based or image-based data.

Clustering algorithms are a form of machine learning that accepts data presented to produce a partitioning of the data. The data given is then used to plot a graph to visually represent the data. “This can be data used to group new articles by topics, or by author, or by language” (Finley and Joachims, 2005). Clustering allows for many datasets to be used to confirm or discover new patterns in data. There are two subsets of clustering - supervised and unsupervised clustering.

Machine learning has a large influence in modern times because of its efficiency to learn from data presented and interpret it to problem solve. This can be seen with the use of natural language processing (NLP). According to Quarteroni, NLP has been used to create technologies such as the Amazon Echo Devices, Google Home, and Apple’s HomePod (Quarteroni, 2018). By using NPL it allows for the API to pick up on language used by the user and interpret them into commands. This can be as simple as requesting the local weather forecast, setting an alarm, and creating a shopping list. These devices use NPL to help the API to understand the user’s request, so if the user stumbled on a word, it is still able to process what it heard as a command and execute what is wanted from the user.

# ***Literature Review***

Machine learning is used to make predictions of data and two of these methods are supervised and unsupervised learning. These techniques are used to predict data but are done with different methods, therefore their uses are different. As the uses are different the outcome of the data is processed and classified with different methods. Supervised learning is used to classify data where the outcome is known, and this is useful as it allows for a trend to be found within the data. This method of machine learning can be used to sort and classify a large dataset quickly. Unsupervised learning is used for data where the outcome is not known, which is useful because it allows for classification and prediction of data where the outcome is not known. The algorithm uses the data, of which the outcome is already known, to train until the accuracy is acceptable and can be implemented into real world use. This is achieved through multiple methods, one of which is hyperparameter tuning. This means that new data with the same labels can be used on the algorithm and a prediction can be made. This method of classification is important to research and improve as the more accurate the prediction, the more suitable unsupervised learning is for real-world implementation.

## **Supervised clustering**

Supervised clustering is defined using labelled datasets, so the labels and the values associated to them are already known. This means that the data can be manipulated until the desired outcome is met. As the input data is fed to the model, it can adjust the weights until the model has been fitted appropriately. This allows for the use of supervised learning in such applications as sorting spam emails into the correct inbox. There are many benefits of supervised clustering such as:

* Creating background knowledge for a dataset.
* Dataset compression and editing.
* Learning subclasses and to use these subclasses to enhance classification algorithms.
* Evaluating distance functions in distance function learning.

(Erick, Zedaat and Zhao, 2005).

There are advantages and disadvantages for using supervised learning which affects how suitable it is for data analysis. Using supervised learning allows for data to be classified into specific labels so it can be easily processed. This allows the user to modify information to manipulate a predetermined outcome. This model is effective because it allows the algorithm to recognise patterns and confirm the outcome. The user knows the input data therefore, the model can be adapted to get the desired effect from the algorithm. Since the data is known before it has been processed, the results wanted from the algorithm can be pursued, meaning that the answers generated are more accurate to unsupervised clustering because the desired outcome is known already.

Although a high level of accuracy is crucial, predetermined labels do not allow for real time learning, which limits the effectiveness of the algorithm outside of its environment. This means that the algorithm cannot adapt to unknown datasets. A supervised-learning method is required for the algorithm to group and display the data. However, this data needs to be constantly checked because new data, such as a new label, can complicate the effectiveness of the algorithm, meaning that accuracy is negatively impacted. If a new label is added, then the algorithm must be manually adapted to incorporate the new change as it cannot do it independently.

This classifying method leads to a high accuracy, but this does not allow the algorithm to be used and trained on a different dataset as it would have been tuned and made for the data that it is using.

Supervised learning is used to process and find patterns within a large dataset. This is due to the methodology that supervised learning uses. By using data that is already labelled the algorithm used can easily detect a common trend as the outcome is already known and the algorithm is able to use that label to train and predict the module. This makes it a very powerful tool when looking at data such as the stock market due to large volumes of data that can be produced. Using supervised learning in the stock market allows for “predicting the future stock and price.” (Yelen and Theng, 2020) using data that otherwise would be too difficult for a human to process and plot accurately. Supervised learning allows for this task to be completed quickly and accurately. This proved to be successful as each algorithm used “has an increased in predictive proformas resulting in favorable outcomes” (ibid) for the task of understanding and predicting using supervised learning.

Supervised learning is better when used on completed datasets because introducing new data can restrict the algorithm from completing its task to high accuracy. Supervised learning is a useful tool to understand large and complicated data as it can categorise the data and show a pattern, but supervised learning does not allow for future predictions. This means that if an unlabelled dataset was used then the algorithm would not be able to predict a common trend accurately.

## **Unsupervised clustering**

Unsupervised clustering is a type of algorithm that can learn from unlabelled data and create a model to find groups within the data given. This means that when data is given to the algorithm, it can detect a pattern within it and plot a graph. The algorithm can detect data points near each other and group them to highlight the patterns found. This is useful to help classify data that does not have any labels associated to it, therefore the gathering of data can be grouped to better understand what it shows. If done correctly, datasets that are too large for a human to go through can be sorted and grouped with unsupervised learning and given labels. For example, “the main advantage of clustering is that interesting patterns and structures can be found directly from very large data sets with little or none of the background knowledge” (Jain, Rajavat and Bharitya, 2012). This means that large data can be sorted efficiently and accurately. Even though it takes longer for an unsupervised algorithm to be completely accurate, compared to supervised clustering, it can lead to more efficient data handling.

Unsupervised clustering also has its advantages and disadvantages. Unsupervised learning is flexible because it allows for real time learning, so the dataset can be unlabelled. This means that once processed and organised by the algorithm, the data can be categorised. The use of unsupervised learning allows for the data to be understood clearly by different models. This is because the unlabelled data can be used on multiple algorithms without the restrictions that supervised learning has. Although the flexibility of unsupervised learning is a positive, it allows for less accurate results as the data is unlabelled, and the desired outcome has not been previously found.

Unsupervised learning is used to predict unlabelled data and find a common trend. New data that has not been labelled can be used allowing for new data on a topic to be collected and processed to predict a future outcome. This method of learning has been used for identifying patterns in cancer data. Unsupervised learning is chosen for areas like biomedical science because “medical applications and models are non-linear and therefore there is no defined training set” (Hamoudi, Bettayeb, Alssaafin, Hachim, Nassir, Nassif, 2019). The adaptive nature of biomedical science such as a virus, and diseases constantly changing making it difficult to create and define one training module for future works. The methodology used for this project is Hierarchical Clustering because this method uses geometry allowing for the algorithm to define the cluster boundaries using simple matrix algebra. This was determined as the best approach as the algorithm allows for tuning to get the desired clusters. By filtering the data using the Ward-Euclidean method the results show an 80% true positive from the testing as by cleaning and filtering the data. This research shows the potential of this type of learning as with the data given and after cleaning the results were positive leading to a high accuracy.

Unsupervised learning is done through many types of algorithms to understand data. These algorithms are designed to read, group and plot data that is unlabelled. Some of these methods are:

* K-means
* DBSCAN
* OPTICS
* BIRCH
* Affinity propagation
* Mean-shift

K-means is one of the oldest algorithms used for clustering as it is one of the most basic forms for unsupervised learning, therefore making it one of the most popular forms of clustering. It is an algorithm that is used for “finding cluster structure in a data set that is characterized by the greatest similarity within the same cluster” (Sinaga and Yang, 2020). This allows for the algorithm to find multiple clusters within a large amount of data. The K-means algorithm has been used for research such as data mining, where large amounts of data are given to the algorithm, and it can cluster and extract this data into a readable format. The application of data mining using K-means has been researched before, in the paper it is stated that K-means “is wildly used in data mining as an unsupervised clustering algorithm” (Shen and Duan, 2020). Throughout the paper the K-means algorithm is researched as an effective method for unsupervised clustering. It is concluded that K-means is a fast and efficient way of classifying data.

DBSCAN (Density-Based Clustering Algorithm) is an algorithm that works by calculating the distance between points and plotting a graph accordingly. This means that the clusters are calculated by the distinctive distances between the data points. Therefore, the data generated from this algorithm can be any shape and could merge two clusters together so data that is more compact and can be clustered with more accuracy rather than K-means. The DBSCAN algorithm “relies on density-based notion of clusters” (Ester, Kiregel, Sander and Xu, 1996), meaning that it relies on the dataset holding two points of high density and low density. The data required to create a DBSCAN model is required to have a minimum number of points. Anomalies are the point of data that do not show normal behaviour and can be defined as “the problem of finding patterns in data that do not conform to expected normal behaviour” (Çelik, Dadaşer-Çelik and Dokuz, 2011). Using DBSCAN, it can define anomalous points in data that do not fit to any of the other clusters found. DBSCAN has been used to classify a large dataset because it is able to discover all clusters’ points in the data and is able to detect the noise that was added to the dataset. It is concluded that “DBSCAN is significantly more effective in discovering clusters or arbitrary shape” (Ester, Kiregel, Sander and Xu, 1996), proving the effectiveness of the DBSCAN algorithm in discovering clusters using a large dataset.

The algorithm OPTICS (Ordering Points To Identify the Clustering Structure) is an algorithm like DBSCAN with the difference being that “OPTICS works in principle like such an extended DBSCAN algorithm for an infinite number of distance parameters” (Ankerst, Breunig, Kriegel and Sander, 1999). The OPTICS algorithm is used to cluster large datasets with the main attraction being there is not one global point that the algorithm relies on therefore, as stated by Ankerst, Breunig, Kiregel and Sander, OPTICS allows for a broad range of parameters settings and thus is a versatile base for both automatic and interactive cluster analysis. This algorithm is used for large datasets of high density as Ankerst, Breunig, Kiregel and Sander’s research shows that the OPTICS algorithm for extraction of hierarchical is highly effective and highly accurate.

The BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) method is a clustering algorithm thatuses distanced-based hierarchies to create a clustering model. To build a model the algorithm creates a CFT (Clustering Feature Tree). This is where the data is compressed into a set of CF nodes. These nodes have several subclusters which hold the information needed to create the model. The BIRCH algorithm is appropriate for “large-scale data, and the algorithm sets the same threshold for all clusters” (Sun, Duan, Lui and Li, 2019) but it is not an effective algorithm if the difference between the datapoints is too great. This means that the dataset used needs to be accessed before the implementation of this algorithm; if this algorithm is used with high dimensional data, it is not efficient. This problem has been tackled with the implication of M-BIRCH (Multi-threshold), which is an adaptation of the original algorithm because it can “set a threshold for each CF” (ibid). This version of the algorithm was seen to be very accurate and an effective way to cluster data.

The affinity propagation is an algorithm that is used for data mining and statistics. This algorithm takes “input measures of similarity between pairs of data points” as stated by (Frey and Dueck, 2007). All the data points are considered as exemplars and by exchanging data between points, the algorithm can find the ideal exemplars, and this is where the clusters are formed. The affinity propagation algorithm can be used in many ways, for data mining and computer vision. This method used for the classification of imaging can be an effective tool due to its implication and how it can process data, but this is also one of its drawbacks. Due to its complexity, it can be hard to implicate properly and efficiently give the desired outcome. This means it is most efficient with the use of smaller datasets as the large the dataset the more complex the problem becomes.

The mean-shift algorithm is a centroid-based algorithm. This algorithm works by updating multiple centroid candidates, which are filtered in a post-processing stage where near-duplicates centroids are eliminated when the final centroids are calculated and formed. The algorithm automatically sets the clusters when all the centroids are processed. Mean-shift is a powerful algorithm as “there are no requirements regarding the prior information of the number of clusters” (Nguyen, Khosravi, Hettiarachchi, Creighton, and Nahavandi, 2014) meaning that this algorithm can be used on new data without having prior knowledge of the field. Isolating what is specifically needed from an image can be difficult, however this is ideal for Mean-shift because the algorithm does not require any training dataset to classify new data. This can lead to high accuracy of new data and be an efficient tool when dealing with unknown data.

## **Literature Conclusion**

In conclusion, machine learning is an efficient and effective tool that has a large influence throughout the field of artificial intelligence. The methods researched have shown two types of computer learning, supervised and unsupervised. Supervised learning allows for more accurate models, but the data is more restricted as it must be labelled and can be used as an analytic tool rather than a tool to predict an outcome based on data. Unsupervised learning allows for more flexible use of data but for a more inaccurate model and results. It also allows for the use of new data that has not yet been processed and an outcome found. Therefore, it is an ideal tool to use when looking at data that may not be the same, such as a virus or cancer cells as both have the potential to change, and a new sample may share a common trend but not have the same outcome. This use of unsupervised learning can lead to false positives but tuning the algorithm to detect and exclude any false positives found it can lead to a much higher accuracy and more efficient type of learning than supervised learning. The algorithms that have been researched have clear strengths and weaknesses, for example, the K-means algorithm is one of the most basic forms of clustering but has the most general use that can be applied to multiple datasets; this means it is quick and efficient in classifying data. Another algorithm that was researched was the DBSCAN algorithm, which is used for large datasets with data points that are closely grouped together. This is where the K-means algorithm would struggle but DBSCAN is efficient and is able to organise data where other algorithms may struggle. Even though each algorithm has its weaknesses they all have their use cases and can lead to a highly accurate model.

# ***The artefact***

## **Introduction**

This project will explore the implementation of different algorithms to a dataset and the outcome of each. To develop the artefact, it was necessary to thoroughly research different methods that could be applied to ensure the algorithms achieved their aims. This research helped understand how different algorithms classify data and how this could be implemented into the research. Unsupervised learning has many techniques for predictions, and it is important to identify which algorithms can make the best prediction based on the primary brain tumour dataset. By using multiple algorithms this meant that multiple possibilities could be explored and that the best algorithm could be chosen and tuned to predict the best outcome. Using multiple algorithms demonstrates that each algorithm has their own uses.

During the research for the project many languages were found that can use machine learning. The three languages that were up for debate were C++, Java, and Python. Each language has their particular strengths and weaknesses.

The language C++ is a high-level language which can be used for machine learning. The language is fast but at the downside of being complex. Complexity is not necessarily a bad property to have with a language as it allows for a large range of tools to use, but this means that debugging and testing can also become very complex, which can slow down a project as you can spend more time debugging rather than testing to improve the algorithms.

Java is an object-orientated language that is fast but complex, similar to C++. Coding in Java can be confusing and complicated so does not allow for easy implementation and testing; this means that a project can be slowed down as it can take much longer to test efficiently. Even though Java can run fast and execute programmes quickly the time saved with this can easily be taken up with debugging thus making the time saved negligible.

Python is an ideal language for machine learning because of its simplicity and large access to many libraries that use machine learning, such as TensorFlow, Pandas, Keras, Scikit-learn. The language is fast and quick to debug meaning that testing can be a smooth process allowing for quick changes and implementation with multiple libraries which are a key element of this project as it allows for simple interpretation and quick easy testing. Therefore, it was decided to use Python for the project as this is the most efficient language.

During the research machine learning was a proposed method for the classification of data as it is efficient due to many machine learning methods that can be used. One method is used to classify data is clustering. Clustering is decided by parameters and each method uses different parameters to decide the clusters.

These methods can be used in supervised and unsupervised learning. These types of learning are used for different reasons, for instance, unsupervised learning can be used when a large dataset that does not seem to have a common trend and an algorithm is needed to classify to highlight a pattern within the data. Supervised learning may be used on data that has been surveyed to understand the purchasing power of certain age groups. Supervised learning has also been used to filter spam emails. This is completed by categorising a sample of spam emails, then a characteristic is assigned to each feature, which is then fed to an algorithm that can disguise those characteristics and filter the new emails that are sent to the inbox. The new data is compared to the already known data and if a match is found it is filtered.

Unsupervised learning is used on a dataset where the outcome is not yet defined, allowing for the algorithm to interpret the data and classify it from patterns and trends that it finds. The reason unsupervised learning is favoured is because it allows for an algorithm that can create a classified dataset using a dataset without labels as they are not needed to complete an analysis. This allows for the data to be interpreted differently by each algorithm, due to each algorithm having a different set of parameters and methodology to determine the outcome.

Unsupervised learning was chosen for this project because of the real-world implications. This is due to how unsupervised learning makes its predictions. The data is then checked with the known outcome to test its accuracy. Using supervised learning leads to higher accuracies as the outcome is already known but this compromises the predictions of data with an unknown outcome, resulting in a model that has a higher accuracy but without the ability to predict the outcome of future data.

The use of unsupervised learning in the medical field is an important tool as it can determine a trend from the data that is provided. It has been used in the past to find and determine breast cancer genomes (Hamoudi, Bettayeb, Alssaafin, Hachim, Nassir, Nassif, 2019). Using unsupervised learning is a useful tool as it can predict a common trend within data of a changing nature such as a virus or cancer. Finding the correct algorithm to use is important as each have their own uses and methodology on how to classify the data. This means that exploring multiple methods and different algorithms is important to further the development of machine learning and its real-world use. Unsupervised learning creates models with less accuracy then supervised learning. Unsupervised learning is important as the more the technology is the developed the more it can be used to create models to predict a pattern of a new illness that is not yet completely known to the medical community. Using unsupervised learning in medicine would be beneficial as it could be used to create a module of an unknown illness, this could be used to find a pattern within a new dataset of symptoms to diagnose and classify a new illness. This use of unsupervised learning could grant professionals the ability to diagnose an illness before traditional medicine could using the predictions of unsupervised algorithm.

Unsupervised clustering also has its advantages and disadvantages. Unsupervised learning is flexible because it allows for real time learning, so this allows the dataset to be unlabelled. This means that once processed and organised by the algorithm, the data is categorised. The use of unsupervised learning allows for the data to be understood clearly by different models. This is because unlabelled data can be used with multiple algorithms without the restrictions that supervised learning has of needing the outcome labelled for the prediction to work. Although the flexibility of unsupervised learning is a positive it allows for less accurate data as it is unlabelled, and the desired outcome has not been found.

## ***The data***

To ensure that each algorithm is tested two datasets were used. These datasets were:

* A brain tumour dataset due to the large number of features for the algorithms to use and the datapoints are varied but follow a common trend.
* Breast cancer data because the data points are close, and the features share similar points.
* Housing data as the dataset is larger and a well-known training dataset.

It is important to use multiple varying datasets because this highlights the strength and weaknesses of each algorithm as this is key to understanding the use cases for each algorithm.

During the research, it was found that large datasets are the best for unsupervised learning as the algorithm has more information to learn and plot its data. Therefore, the datasets decided needed to fit the requirements for unsupervised learning. It was decided that multiple datasets should be used as this would be suitable as the data is varying. These datasets were chosen as they all contain a large amount of data; this means that the algorithm will have access to more data to classify the outcome and become more accurate. Unsupervised learning needs many points to allow for accurate classifications, for instance, if the datasets are small and do not contain many features the algorithms do not have enough information to make accurate predictions.

Each dataset is different but have similar trait; both are high density datasets and have many features, which is important as the more features the algorithm has access to, the more accurate the predictions will be. The primary dataset used for the project was the brain tumour dataset because it will be most effective with all the algorithms due to its large number of features and data entries. The data also follows a clear trend, implying it can be easily tested with a small sample size of the dataset to test the accuracy of the algorithm early. The secondary dataset is a breast cancer dataset similar to the primary dataset but with less features and fewer data entries. This is positive as the data points are close together so it will show the different accuracies of the algorithms. The third dataset chosen is a housing dataset because it is a well-known and popular dataset used to train algorithms. This is because the dataset is dense and large with datapoints having a large difference between them which makes it easier for an algorithm to detect and classify.

The data collected needed to be processed before it could be used to ensure that the data was valid for the algorithms to analyse. This meant using a library called PairPlots, which allows for the data to be visualised into graphs. It displays the data that the dataset contains. Using PairPlots meant that the data could analysed to see if there are any clear patterns within the data that the algorithms should easily pick up on. After running the data through PairPlots it showed that there was a clear pattern that the data followed therefore this meant that the data is suitable for cluster analysis. As seen in ***figure 1***.

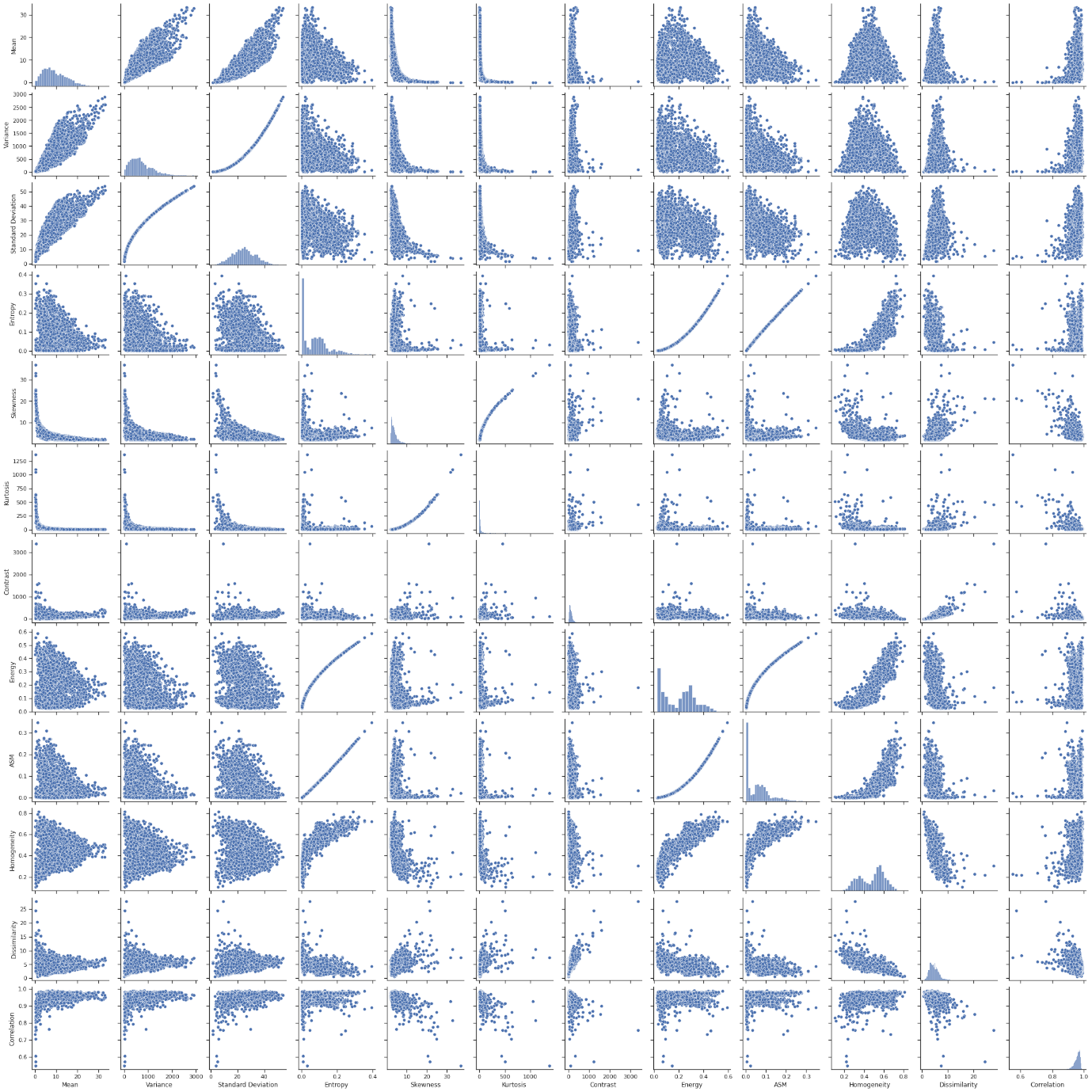


Figure 1: PairPlots of Brain tumour data

Visualising the data is a useful tool as it meant that picking a couple of features to use for early testing of the artefact was easier. This is important as it is possible to test the accuracy early on and ensure that the algorithms picked are suitable because by only using two features the algorithm can easily sort and label correctly. Doing this allowed for exploration of multiple algorithms quickly rather than picking one algorithm and testing it against the full dataset, then further in the project realising that it is not the appropriate algorithm for this project. By doing smaller scale testing it was found that the K-means, DBSCAN, and Mean-shift algorithm were the most appropriate algorithms for this project. These were chosen because the algorithms all test and process the data in different ways but use the same base principle of checking the data points’ distance from each other to create the model. Using different algorithms is important as it ensures there is a varied interpretation of the data despite using similar methods.

The data used is important as it determines the accuracy of the algorithms. Therefore, it is important to ensure that the current features of each dataset are chosen because if the all the data features are given to the algorithm at once it can affect the accuracy as some of the features may not have a commonality to the rest of the data. Choosing the current features to use is important and by visualising the data it allowed for the correct features to be picked. By doing choosing the correct features this allows for better results and more accurate testing. To pick the correct features the data had to be viewed again in PairPlots to find the features that seem to share a common trend meaning they are suitable to use.

# 

# ***Testing the artefact***

The purpose of testing is to improve the algorithms accuracy to create a better model. Testing is an important aspect of machine learning. To ensure that the correct method of testing was chosen more than one method needed to be researched. While researching hyperparameter tuning and neural networks were two methods that were found. Both methods are good for testing and lead to better results. Neural networks are a complex method that uses multiple layers of data to create the prediction. This leads to a high level of accuracy as the algorithm has access to multiple layers of data to learn form to create its prediction but using this method can be ineffective as the multiple layers of data can pose an issue. If the algorithm is under performing it can be hard to find the cause and fix it. Hyperparameter tuning takes the existing algorithm and used the built-in parameters to improve the accuracy. This is done by changing the values of one or more of the parameters and testing the new values against the dataset. Hyperparameter tuning can also be completed much quicker than deep learning as changing the value of each parameter can be done quickly and tested.

One of the methods used to improve algorithm accuracy is deep learning. According to (Deng and Dong, 2013) the definition of deep learning is “a class of machine learning techniques that exploit many layers of non-linear information processing for supervised or unsupervised learning feature extraction and transformation, and for pattern analysis and classification.” Using this type of learning can lead to high test results but can be hard to implement. For this reason, hyperparameter tuning was the chosen method for improvement for the algorithms.

All data shown is the hyperparameter tuning of the primary brain tumour dataset.

## ***K-means***

The K-means algorithm is considered one of the basic methods of unsupervised learning as it can sort small datasets accurately but when dealing with large datasets the algorithm struggles to sort and clusters the data, which is a limitation. One major limitation of the K-means algorithm is that is the algorithm is not as efficient when that data being used has anomalies, as the algorithm may place data that is a clear outliner with the rest of the data into one of the clusters. With the algorithm not being efficient at picking up anomaly data it may classify the data inaccurately making the accuracy reported back wrong and making the results wrong. Another problem with this is that data that should be classified one way can end up being misplaced within the clusters made due to its close relativity to the other data again making the clusters generated inaccurate. On the other hand, K-means is a successful algorithm to use when hyperparameter tuning, as shown by (Korovkinas, Danėnas and Garšva, 2019). In this paper it shows that the K-means algorithm is tuned by implementing SVM. This showed an improvement to the total accuracy.

During testing K-means algorithm it was important to understand the workings through the documentation of the algorithms. Further research into the documentation allowed for the proper tuning to be conducted. Without any tuning done to the K-means algorithm accuracy was averaging 33.8% as shown in ***figure 2***. This accuracy is considered low.

Figure 2: K-means testing with tuning.

As seen with ***figure 2*** there is no improvement over the iterations meaning this is the highest accuracy that can be achieved without any hyperparameter tuning. The highest accuracy achieved with this method was 33.81%.

While tuning the random\_state parameter there is an improvement. The random\_state parameter “determines the number of generations for centroid initialization” (scikit learn, K-means). The results reach a peak and then drops back down to the original accuracy. From this the appropriate value is chosen. Once this value has been chosen it is set aside to allow for further tuning.

Figure 3: K-means random\_state

The random\_state parameter is set at NONE by default, this parameter is unable to go below 0. Running the algorithm with the random\_state decreasing shows poor accuracy but by increasing this value the accuracy increases but with negative peaks, therefore this parameter is sensitive and can greatly affect the accuracy. As seen in ***figure 3***. The highest accuracy achieved with this method of hyperparameter tuning was 66.1% accuracy.

The next parameter chosen to test was the n\_init parameter. The n\_init parameter effects the number of times the K-means algorithm will be run with different centroid seeds. Tuning this parameter shows a positive increase in total accuracy ***figure(x)*** but there are also peaks in the accuracy meaning that the parameter can have a positive and negative effect on the data.

Figure 4: K-means n\_init tuning.

As seen with ***figure 4*** there is improvement to the total accuracy but with dips of negative accuracy. In-between iteration 3 and 6 there is a period of stabled increase accuracy but at iteration 7 there is a negative peak of accuracy. The highest accuracy achieved with this hyperparameter tuning was 66.13%.

The last parameter chosen to be implemented is the max\_iter, this parameter determines the amount the algorithm is allowed to iterate for a single run. This parameter allows for the algorithm to process the data a set number of times, by default this is set to 300 this was increased and decreased to see the change in accuracy. By tuning this parameter, the algorithm can process the data more frequently and create a better model. This default value was used to increment up and increment down. The best results were found when incrementing down.

Figure 4: K-means testing with tuning.

The max\_iter parameter is the most stable parameter out of the 3 chosen as seen in ***figure 4***, this parameter has a consistently close accuracy with a margin of 0.1% error, this means that this is the most consistent and accurate standalone parameter. The highest accuracy achieved was 66.18%.

The next step of testing the K-means algorithm is to implement these 3 parameters into the same algorithm and tuning all of them at once to achieve the highest value. By doing this it allows for more control of the algorithm due to being to edit more than one parameter at once. This method can cause issues of conflict and make the model unstable but also can achieve the highest accuracy once stabilised.

This methodology is repeated with all the parameters implemented into the algorithm. The optimal value from the previous testing is chosen and the algorithm is run. Therefore, further testing is needed to be completed, using the methodology for the previous testing the algorithm is ran again.

Figure 5: K-means testing with tuning.

As seen in ***figure 5*** there is conflict between these values meaning that further testing needs to be completed. By applying the previous methodology there is a great improvement and the algorithm becomes stable as seen in ***figure 5***. This is the highest accuracy that can be performed for the brain tumour dataset. The highest accuracy achieved is 66.18% which is a 33% improvement compared to the standard K-means algorithm with no implementation of hyperparameter tuning.

To create the highest accuracy model possible, it was important to use the previous methodology and implement all the parameters into the algorithm. This was done using the previous values to tune the algorithm. With the improvement that is seen with the previous testing the parameters can be implemented into the algorithm. The chosen values for each parameter to create the highest accuracy:

* random\_state tuned to 15
* max\_inter tuned to 275
* n\_init tuned to 20

Figure 6: K-means testing with tuning.

These are the values that perform the best to give the highest accuracy using the brain tumour dataset. This methodology was repeated for the breast cancer and housing data.

## ***Mean-shift***

Mean-shift is an algorithm where centroids are created by checking the distance in-between points and if the points are too close then it will not use those points to create a cluster. This means that data points that are close together should be suitable for this algorithm but if the data is too close then it will be unable to distinguish clusters. This makes the Mean-shift algorithm successful at understanding large datasets that may seem to not have a common trend as the algorithm is able to create centroids based on the distance between the data points. As seen in (Kim, Park, Lee and Choi, 2018) Mean-shift is implemented with SPCC distancing to create a module which results in a higher accuracy model. This algorithm is used to classify data that otherwise may seem too complex for a basic algorithm like k-means to be used.

To see the progress of tuning Mean-shift the algorithm needed to be run without any parameters edited and to know which parameters to tune to increase its accuracy. Further research into the parameters of Mean-shift allowed for proper tuning, which was done by looking at the documentation (sci-kit learn, Mean-shift). The base accuracy of the Mean-shift algorithm was at 0.0% ***figure 7***. This was believed to be due to the algorithm not processing the data correctly to create the clusters desired.

Figure 7: K-means untuned.

As seen in ***figure 8*** there is no improvement to the accuracy of this algorithm. This is due to the algorithm detecting 22 clusters where there should be 2. This means that the accuracy will always be low until a parameter is found that can change the cluster amount.

While reading the documentation for Mean-shift it was decided that the bandwidth parameter will be changed using estimate bandwidth, using this parameter prevents Mean-shift creating its own value. This prevents a more random classification as having control of this parameter allows for more accurate clusters.

Figure 8: Mean-shift bandwidth

Tuning the bandwidth parameter sees an improvement because the accuracy is 0.26%. This accuracy is very low, but there is an improvement signifying that it is possible to tune Mean-shift according to the data.

The next parameter chosen to be tuned is max\_iter. This parameter determines how many iterations for each loop Mean-shift can use. By editing this parameter Mean-shift has a longer period to classify the data. This is set to a default of 300; by iterating up and down it allows for the most appropriate value to be found.

Figure 9: Mean-shift max\_iter

As seen in ***figure 9*** there is no improvement to the accuracy but the max\_iter parameter stabilises the number of clusters found. This is due to not being able to set the clusters to a certain amount but by using this parameter it allows for a consistent cluster count of 2. The highest accuracy found was 0.26%.

The next parameter chosen for tuning was the min\_bin\_freq. This parameter determines what the algorithm allows to be used because the parameter will only accept the data that matches the seeds, in turn limiting what data is used.

Figure 10: Mean-shift min\_freq

As seen in ***figure 10*** there is no improvement, this is shared with the accuracy of 0.26%. This means that this parameter does not improve the accuracy for this data.

With these hyperparameters researched, all the parameters are entered into the algorithm to achieve the best accuracy. Using these parameters show that it can be used for improvement with a starting accuracy of 0% to 0.26%. This improvement is minimal and only slightly improves the model. The highest accuracy achieved for the breast cancer dataset was 0.18% which was also a minimal improvement from 0.0% from the base algorithm. This is due to the data not being suitable for the algorithm used. Mean-shift has been used on image data as used by (Kim, Park, Lee and Choi, 2018). Therefore, it would not be possible to increase the accuracy further as there was a high chance of a false positive, implying that the algorithm was not suitable for the data used and a high accuracy was unachievable.

## ***DBSCAN***

The DBSCAN algorithm is used to classify a large dataset where the points are close. This algorithm requires the data points to be close and dense; if the data points are far apart the algorithm would struggle to create a centre point. The DBSCAN algorithm is used for large datasets that do not seem to contain a common trend. This means that the algorithm has its limitations as it is not a suitable algorithm to use with a small dataset. When used with a small dataset the algorithm cannot create a model as the dataset is too small, also if the dataset is large but it is not dense with the data points being far apart, the algorithm will not be able to classify this data. DBSCAN struggles with data that has a high number of anomalies. This is due to that data not being near the centroid, so the algorithm is not able to classify the anomaly. DBSCAN has been tuned with a fuzzy extension (Bechini, Criscione, Ducange, Marcelloni, and Renda, 2020) which had a higher accuracy than without. This shows that hyperparameter tuning is possible with the algorithm.

When testing the DBSCAN algorithm it was important to check the documentation, (Scikit-learn, DBSCAN) allowing for the correct parameters to be tuned to ensure that the highest accuracy model could be found. To begin testing the base accuracy needed to be found which was zero. This was done for all the datasets, and this was the common result. The brain tumour dataset was the first to be tested and its accuracy was 0.0%.

Figure 11: DBSCAN untuned.

As seen in ***figure 11*** there is no improvement with the base algorithm, due to DBSCAN only being able to create 1 centroid. This is accuracy is very low due to the algorithm detecting 1 cluster where 2 should be found.

The next step of tuning was to tune the min\_samples. The min\_samples parameter controls the total weight required to create a neighbourhood. By setting this value to a higher value rather than the default of 5 then the algorithm should be able to create more than 1 cluster for the dataset.

Figure 12: DBSCAN min\_smaples

As seen in ***figure 12*** this shows that by tuning the min\_samples parameter the algorithm can successfully create 2 clusters constantly, but the accuracy does not change. This means that the algorithm creates 2 clusters but does not correctly label the data.

The next stop to test DBSCAN was to tune the eps parameter. This parameter edits the distance needed to create a neighbourhood. By tuning this parameter, it is possible to increase and decrease the distance needed to create a valid cluster, seeming that the data being used has minimal changes in distance it was important to tune this down from the default value of 0.5.

Figure 13: DBSCAN eps.

In ***figure 13*** it shows that the eps value has a large control on the amount of clusters made. Iterating from the starting value of 0.1 through to 1 this shows that the most stable value is 0.6 as this is where the algorithm can produce 2 clusters with an accuracy of 0.26%.

The next step is to implement all the parameters into the algorithm is to create the best accuracy. This is achieved by using the previous testing methods and implementing the parameters into the algorithm. This model will have the highest accuracy for DBSCAN.

Figure 13: DBSCAN tuned

In ***figure 13*** it is evident that the algorithm is able to produce a stable model of 2 clusters with a low accuracy of 0.26%. This is due to the data not being suitable for this algorithm. DBSCAN is used on data sets that are “non-uniform density with broad values and gradually spare forwards” as stated by (Bechini, Criscione, Ducange, Marcelloni, and Renda, 2020). The data used is highly dense, but the data points are narrow and close together meaning that the algorithm would not perform well on this data. This is the reason why with tuning no improvement to the accuracy was made.

## ***Results***

The results show that for the data used the algorithm that achieved the highest accuracy was the K-means algorithm. The K-means algorithm paired with the brain tumour dataset had a base accuracy of 33.8% but once tuned this accuracy doubled to the highest accuracy of 66.18%. The breast cancer dataset had a base accuracy of 14.72% but when the algorithm was tuned the accuracy reached 85.23%. The housing dataset had a base accuracy of 15.5% but once turned this accuracy reached 41.6%. This is due to the amount of parameters that can be tuned as the algorithm has a lot of flexibility that the user can use. The breast cancer data was the most successful data used with K-means because the datapoints are close but have a significant change meaning that the algorithm can create accurate clusters.

The Mean-shift algorithm was able to create reliable clusters once tuned but did not achieve a high accuracy due to the datasets not being suitable for the algorithm. The Mean-shift algorithm was harder to implement and tune, as shown by the results. The highest accuracy found for the brain tumour dataset using the Mean-shift algorithm was 0.26%. The highest accuracy for the breast cancer dataset using the Mean-shift algorithm was 0.26%. The highest accuracy for the housing dataset was 0.5%. The accuracy did not improve during tuning, but the algorithm did stabilise and was consistently able to create the correct number of clusters for each dataset.

The DBSCAN algorithm was the hardest algorithm to implement and tune. This is because of the limited parameters that could be tuned to fit the data that was being used. The highest accuracy achieved for the brain tumour dataset was 0.26%. The highest accuracy achieved for the breast cancer dataset was 0.26%. The DBSCAN algorithm was unable to produce an accuracy for the housing dataset. The data being used was not suitable for the algorithm, even though that data was dense the data points were too close together meaning that DBSCAN was unable to distinguish between the datapoints to create the create number of clusters consistently.

In ***figure 14*** it shows the highest accuracy achieved by each algorithm with all the datasets. This table highlights the importance of the data used as the K-means algorithm performed the best out of the 3 algorithms tested due to the data.

Figure 14: all accuracy’s

# ***Conclusion***

Unsupervised learning is powerful but complex and hard to implement into an algorithm that can classify the data successfully. The use of unsupervised learning is plentiful as the methodology of unsupervised learning can be used of all types of data as the data being used does not need to be labelled. This allows for an algorithm to be tuned to previous data than if new data is formatted the same it can have a high accuracy. Supervised learning does not allow for this, but it tends to have a higher overall accuracy as the data being used outcome is known. This means that for applications such a medical data unsupervised learning is a better tool as new data of a new virus can be classified without the outcome. Unsupervised learning complexity is what allows for the high accuracy of unknown data due to the tuning that can be done.

The setback of unsupervised learning clustering is the data as each algorithm needs the data to be formatted in a certain way. The Mean-shift algorithm is complex but does allow for multiple parameters to be tuned similarity to the K-means algorithm but due to the limitations on what data can be used it did not achieve a high accuracy even with the ability to edit many parameters. The Mean-shift algorithm became stable and created the correct numbers of clusters but was not able to classify the data correctly. The DBSCAN algorithm did not allow for many parameters to be tuned meaning that the flexibility of this algorithm is low and is depended on the data available. Therefore, the accuracy achieved was low, but the algorithm was stabilised at 2 clusters. Due to the way that DBSCAN works, by requiring the clusters to be dense but far apart, it will classify it all as 1 cluster. The lack of distinct data points required to achieve a high accuracy were unavailable, therefore the DBSCAN algorithm was unable to perform as intended. Even though the K-means algorithm is considered one of the most basic algorithms that can be used for unsupervised learning, the results show that it is one of the most consistent. The K-means algorithm was used with all three datasets and made a significant improvement. With all the datasets, being able to be run successfully and achieve a higher accuracy once tuned means that unsupervised learning combined with hyperparameter tuning is a reliable and stable method of classifying data. The K-means algorithm was able to be tuned and fit well to each dataset event though they all varied in size and data. This was due to the flexibility of the algorithm’s parameters because being able to use multiple parameters this meant that the algorithm was able to achieve a high accuracy with the minimum improvement of 32.38% accuracy increase and with the maximum of 70.51% improvement. Therefore, the most efficient algorithm to use on data that has not been formatted to fit other unsupervised learning algorithms is the K-means algorithm.

## ***Future work***

To improve the accuracy of all three algorithms pre-processing could be used - one example of this is data normalisation. The use of data normalisation is to edit the dataset to create a more suitable data for the algorithm to use. This is done by using the cleaned dataset and having a separate algorithm to plot the dataset into more suitable data points. With this method the Mean-shift and DBSCAN algorithm could have been able to create a classification. Data classification has successfully been implemented with medical data to improve the accuracy (Singh and Singh, 2021). With data normalisation there could be further improvement for the K-means algorithm as well due to the data being more suitable for classification to occur.

While the K-means algorithm is the most accurate at 66.18% this accuracy is still considered low. A higher accuracy could be achieved by implementing a neural network. Neural networks “consist of many samples, connected processors called neurons” (Schmidhuber, 2015). These neurons all use different methods to classify the data meaning that the data can be processed through multiple methods leading to a higher accuracy. Using a neural network allows for data to be complex where the data does not need to be in a certain format to achieve a high accuracy. A neural network is complex but by using multiple neurons this allows for the data to be processed multiple time before creating the model. This would be equivalent to using the data through each of the algorithm proposed and more to find the most suitable algorithm for the data.

# ***Project management***

Project management is important as it allows for each task to be allocated its own time slot to ensure that the project is completed on time. One of the methods used to manage this project was the use of a Gantt chart, which allows for the process of a project to be visualised so it is easy to follow and ensure that the current task is being completed. With a Gantt chart the project was able to be managed and kept on track. As seen in ***Figure 15,*** it was crucial to multitask and keep to each milestone to ensure the project flowed effectively and that development did not stagnate.

*Application

Description automatically generated with low confidence*

Figure 15: Gannt Chart

The Gantt chart shows the process that ensured the project would be achieved. The first step was to find research for the project, which allowed for the project to be outlined and to determine which method of machine learning to be used. Next was to write the literature review because it allowed for unsupervised learning to be chosen and explored which algorithm should be used. After choosing the algorithms it was time to start creating the artefact and have a prototype. When the first prototype was created testing could be completed, while testing notes were made to keep track of the project that could be added to the writing. The next step was to write about the artefact and testing.

To keep track of the artefact and any milestones that were achieved GIT was used. Using GIT meant that any changes that were made could be kept track of and that if any references to old versions could be accessed, while also maintaining version control.

Using GIT allowed for testing to be completed easily as once 1 of the algorithms were being tested then the work could be committed, and notes could be recorded and used for reference. The application GitKraken was used as this allowed for easy visualisation of the artefact and steps achieved. As seen in ***Figure 16*** these were the steps taken regarding development of the artefact including each algorithm and hyperparameter tuning.

Graphical user interface

Description automatically generated

Figure 16: GitKraken

# ***Bibliography***

Ankerst M Breunig M M Kiregel HP Sander J 1999 OPTICS Ordering Points To Identify the Clustering Structure 1999 SIGMOD 1999 Proceeding ACM SIGMOD International Conference on Management of Data 1999 Philadelphia Pennsylvania USA DOI101145304182304187

Bechini A Criscione M Ducange P Marcelloni F and Renda A 2020 FDBSCANAPT A Fuzzy Densitybased Clustering Algorithm with Automatic Parameter Tuning 2020 IEEE International Conference on Fuzzy Systems FUZZIEEE

Çelik M DadaşerÇelik F and Dokuz A Ş 2011 Anomaly dectection in temperature data using DBSCAN algorithm 2011 International Symposium on Innovations in Intelligent Systems and Applications Instanbul Turkey

Charles EYA Yasotha R 2016 Yasotha R Automated Text Document Catorization 2015 IEEE Seventh International Conference on Intelligent Computing and Information Cairo Egypt Section IV DOI 101109IntelCIS20157397271

Deng L and Dong Y 2013 “Deep learning Methods and Applications” 734 pp 199200

Erick F C Zedaat N Zhao Z 2005 Supervised clustering – algorithms and benefits 2004 16th IEEE Internation Conference on Tools with Artificial Intelligence Boca Raton FL USA

Ester M Kiregel HP Sander J Xu X A 1996 A DensityBased Algorithm for Discovering Clusters in Large Spatial Databases with Noise University Of Munich

Finley T Joachims T 2005 Supervised Clustering With Support Vector Machines Department of Computer Science Cornell University Ithaca

Frey Brenden J Dueck D 2007 Clustering by Passing Messages Between Data Points Science Express Vol 315 Issue 5814 pp 972976 DOI 101126science1136800

Hamoudi R Bettayeb M Alssaafin A Hachim M Nassir Q Nassif A B 2019 Identifying Patterns of Breast Cancer Genetic Signatures using Unsupervised Machine Learning 2019 EEE International Conference on Imaging Systems and Techniques IST DOI 101109IST4802120199010510

Jain A Rajavat A Bharitya R 2012 Design Analysis and Implementation of Modified Kmean Algorithm for Large Dataset Increase Scalability and Efficiency 2012 2012 Fourth International Conference on Computational Intelligence and Communication pp 627631 Mathura India

Kim N Park S Lee J and Choi J 2018 Load Profile Extraction by MeanShift Clustering with Sample Pearson Correlation Coefficient Distance Energies 119 p2397

Korovkinas K Danėnas P and Garšva G 2019 SVM and kMeans Hybrid Method for Textual Data Sentiment Analysis Baltic Journal of Modern Computing 71

Mohri M Rostamizadeh A Talwalker A 2018 Foundations Of Machine Learning 2nd EditionMassachusetts Massachusetts Institute of Technology

Nguyen T Khosravi A Hettiarachchi I Creighton D Nahavandi S 2014 Classfication of neural action potentials using mean shift clustering 2014 2014 IEEE international Conference on Systems Man and Cybernetics SMC San Diego CA USA

Quarteroni S 2018 Natural Language Processing for Industrial Applications Spektrum 41 p105

Samuel A L 1959 Some Studies In Machine Learning Using The Game Of Checkers Volume 3 issue 3 httpsieeexploreieeeorgabstractdocument5392560

Schmidhuber J 2015 Deep learning in neural networks An overview Neural Networks 61 pp85117

Shen H and Duan Z 2020 Application Research of Clustering Algorithm Based on KMeans in data mining 2020 International Conference on Computer Information and Big Data Applications CIBDA Guiyanf China

Singh, N. and Singh, P., 2021. Exploring the effect of normalization on medical data classification. 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)

Sinaga K P Yang M 2020 Unsupervised KMeans Clustering Algorithm 8 pp 8071680727 DOI 101109ACCESS20202988796

Sun Q Duan Y Liu F Li H 2019 Application of Improved MultiThreshold Birch Clustering in Reservoir Prediction 2019 2019 6th International Conference on Systems and Informatics ICSAI Shanghai China

Vazirani S Sharma A and Sharma P 2020 Analysis of various machine learning algorithm and hybrid model for stock market prediction using python 2020 International Conference on Smart Technologies in Computing Electrical and Electronics pp 203207 Bengaluru India

Yelne A Thend D 2021 Stock prediction and analysis of Using Supervised Machine Learning Algorithms 2021 Internation Conference on Computaatuional Intelligence and Computing Applications ICCICA pp 16 Nagpur India

Scikit-learn K-means parameters:

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

Scikit-learn Mean-shift parameters:

[https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html#](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html)

Scikit-learn DBSCAN parameters:

[https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html)

Data sources:

The brain tumour dataset was found at:

<https://www.kaggle.com/jakeshbohaju/brain-tumor>

The breast cancer dataset was found at:

<https://www.kaggle.com/code/buddhiniw/breast-cancer-prediction>