ECS 308 – Data Science and Machine Learning

Analysis Report

Report: [redacted]

Name: Subham Das

Roll No: [redacted]

Date: 09/05/2021

INTRODUCTION

Parkinson's disease is a progressive disorder of the central nervous system affecting movement and inducing tremors and stiffness. It has 5 stages to it and affects more than 1 million individuals every year in India. This is chronic and has no cure yet. It is a neurodegenerative disorder affecting dopamine-producing neurons in the brain.

We have used UCI ML Parkinsons dataset for this. The dataset has 24 columns and 195 records. UCI Parkinson's dataset is composed of 22 features extracted from voice measurements of 32 subjects (24 PD cases and 8 healthy controls). Each subject contributed 6 or 7 records, generating a total of 195 records.

The various features included are:

1. MDVP:Fo(Hz)	2. Jitter:DDP	3. NHR
4. MDVP:Fhi(Hz)	5. MDVP:Shimmer	6. HNR
7. MDVP:Flo(Hz)	8. MDVP:Shimmer(dB)	9. RPDE
10. MDVP:Jitter(%)	11. Shimmer:APQ3	12. DFA
13. MDVP:Jitter(Abs)	14. Shimmer:APQ5	15. spread1
16. MDVP:RAP	17. MDVP:APQ	18. spread2
19. MDVP:PPQ	20. Shimmer:DDA	21. D2
22. PPE		

We have analysed the data using the following algorithms:

1. Logistic Regression 2. Support Vector Machine (SVM) 3. Decision Tree

4. Linear Regression 5. Artificial Neural Netwok (ANN) 6. K-Means Clustering

7. Gaussian Mixture Model 8. K Nearest Neighbours (KNN)

Each of the algorithms has a different accuracy and prediction value and is used to determine whether a patient has Parkinson's disease or not.

We will analyse each of these algorithms separately and then look at a comparative study.

NOTE:

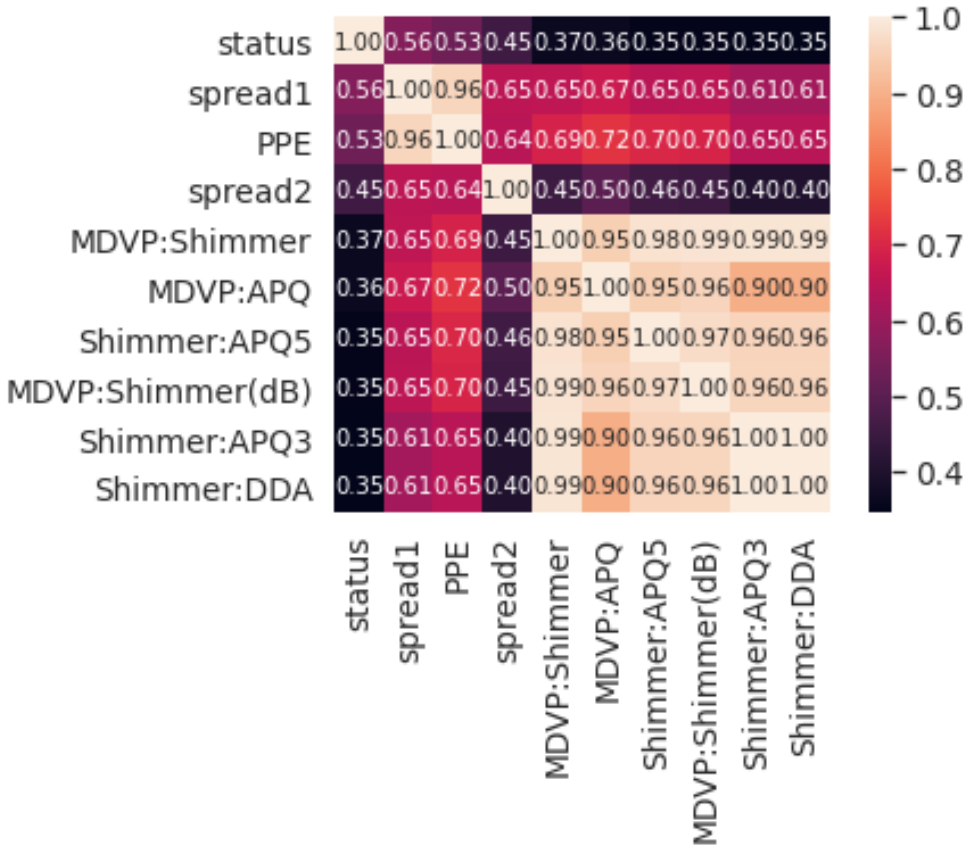
1. There are no Null values.

---

2. Encoding the Categorical values into numerical values is not required in this dataset because all values we have floating type only. We have name column as a categorical values but we are not going to use that column in model prediction.

---

3. There are no outliers in our dataset



Above is the correlation values in descending order. We are going to drop MDVP:RAP column to MDVP:Fhi(Hz) because it has less correlation with other columns.

If we decrease the column count then accuracy will increase gradually because we are not keeping the irrelevant features.

Feature Split: We performed a feature split, i.e., splitting the dataset into input and output attributes and dropping irrelevant column values from our dataset so that we can get better accuracy.

CASE STUDY 1: Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression).

Under different cases the algorithm performed as follows:

1. Prediction we got without applying feature scaling –
   Logistic Regression Classification Algorithm: 0.859583

2. Prediction we got by applying feature scaling –
   Logistic Regression Classification Algorithm: 0.859583

3. Prediction we got after tuning –
   Logistic Regression Classification Algorithm: 0.853333

Finally after fitting we obtained the following result:

→ Training accuracy of **86.5%** and the test set accuracy is **84.6%** for Logistic Regression

CASE STUDY 2: Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. SVMs are one of the most robust prediction methods, being based on statistical learning frameworks. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

It constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection.

Under different cases the algorithm performed as follows:

1. Prediction we got without applying feature scaling –
   Support Vector Machine classification Algorithm: 0.821667

2. Prediction we got by applying feature scaling –
   Support Vector Machine classification Algorithm: 0.821667

3. Prediction we got after tuning –
   Support Vector Machine Classification Algorithm: 0.82

Finally after fitting we obtained the following result:

→ Training accuracy of **89.7%** and the test set accuracy is **84.6%** for SVM

CASE STUDY 3: Decision Tree

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

There are two main types of Decision Trees:

1. Classification trees (Yes/No types)
2. Regression trees (Continuous data types)

In our study we have used **Classification Tress**.

Under different cases the algorithm performed as follows:

1. Prediction we got without applying feature scaling –
   Decision Tree Classification Algorithm: 0.840000

2. Prediction we got by applying feature scaling -
   Decision Tree Classification Algorithm: 0.865833

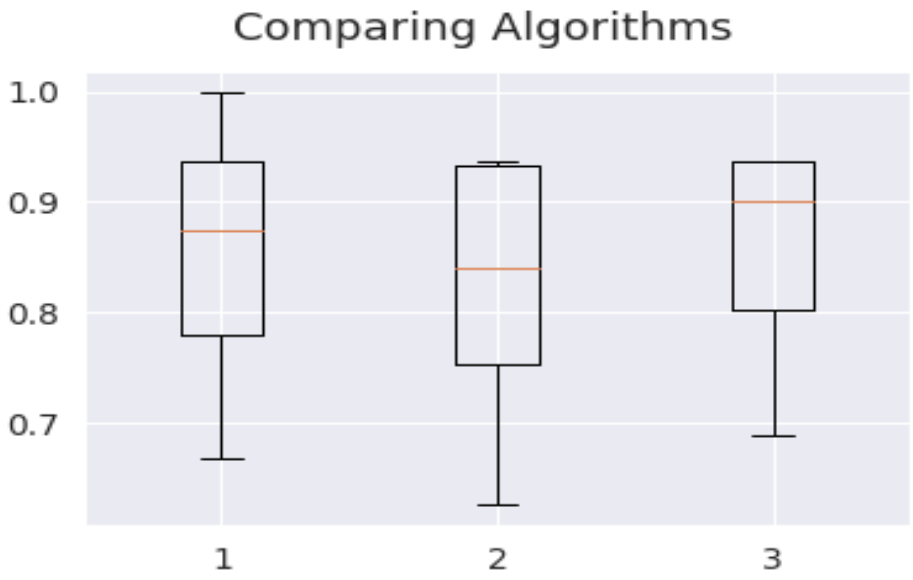
3. Prediction we got after tuning –

   Decision Tree Classification Algorithm: 0.859583

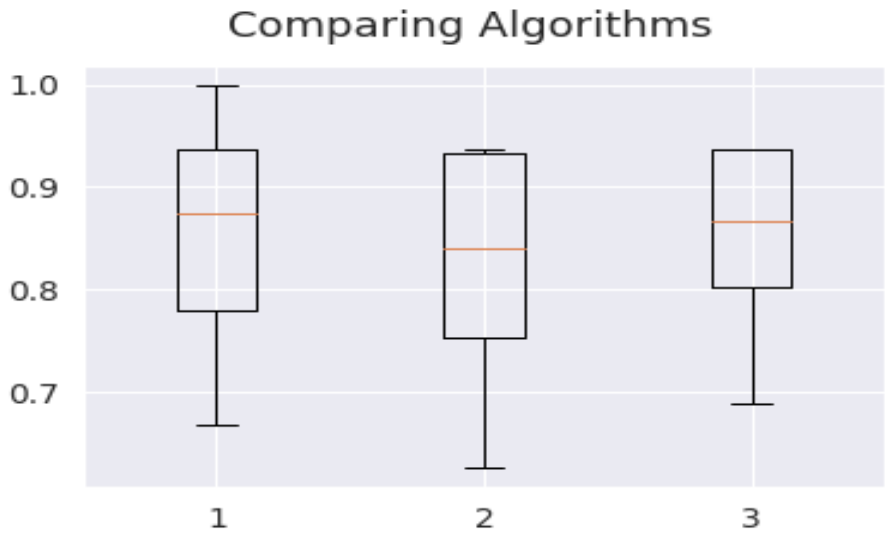
Finally after fitting we obtained the following result:

→ Training accuracy of **100%** and the test set accuracy is **89.7%** for Decision Tree

(i) Comparing the three algorithms before feature scaling:



(ii) Comparing the three algorithms after feature scaling:



CASE STUDY 4: Artificial Neural Networks (ANN)

Neural networks learn (or are trained) by processing examples, each of which contains a known "input" and "result," forming probability-weighted associations between the two, which are stored within the data structure of the net itself. The training of a neural network from a given example is usually conducted by determining the difference between the processed output of the network (often a prediction) and a target output.

After a sufficient number of these adjustments the training can be terminated based upon certain criteria.

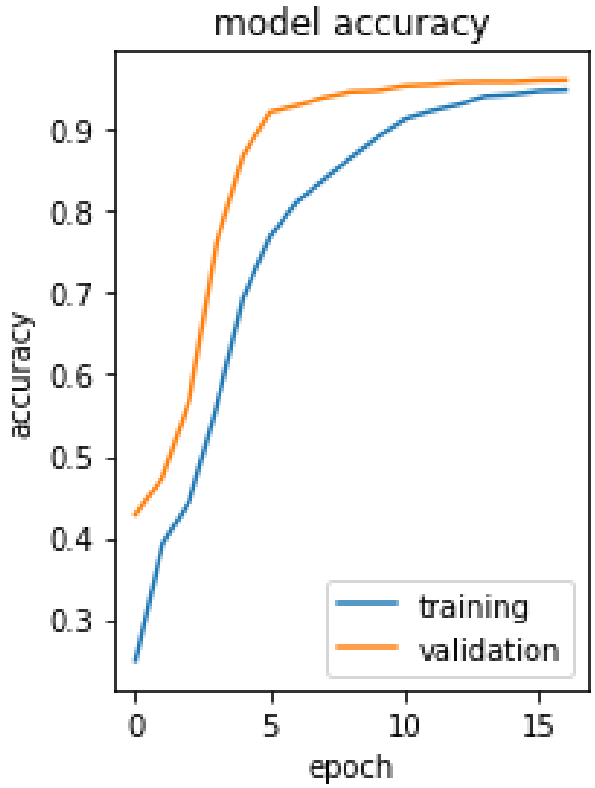
The **batch size** is a hyperparameter that defines the number of samples to work through before updating the internal model parameters.

The number of **epochs** is a hyperparameter that defines the number times that the learning algorithm will work through the entire training dataset. Over here he have set the epoch to **100**.

Dense layer is the regular deeply connected neural network layer.

Finally after fitting we obtained the following result:

→ Test set accuracy of **93.37%** using ANN



CASE STUDY 5: Linear Regression

It is one of the best statistical models that studies the relationship between a dependent variable (Y) with a given set of independent variables (X). The relationship can be established with the help of fitting a best line.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Finally after fitting we obtained the following result:

→ Test set accuracy of **25%** using Linear Regression

The metrics obtained from analyzing the algorithm is as follows:

Mean squared error: 0.15
Coefficient of determination: 0.25

If the coefficient of determination is 1 it is perfect prediction.

CASE STUDY 6&7: Unsupervised Learning – K Means Clustering & Gaussian Mixture Model

Unsupervised Learning is a machine learning technique in which the users do not need to supervise the model. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data.

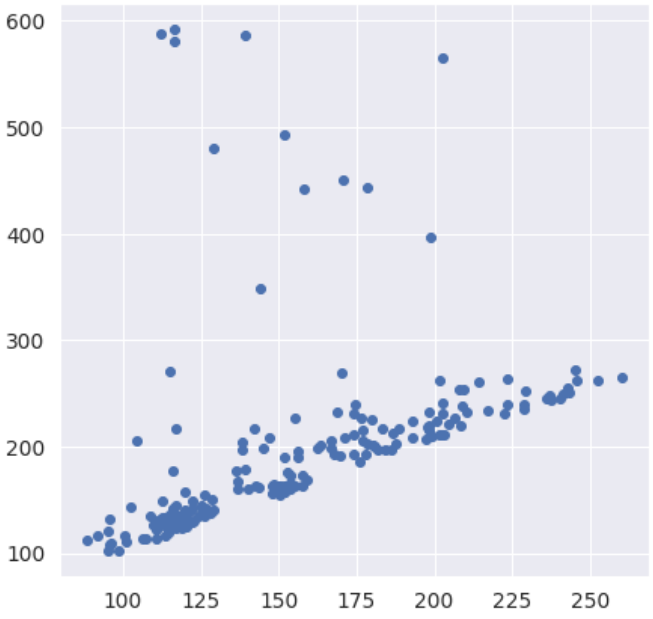
***k*-means clustering** is a method of vector quantization, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult.

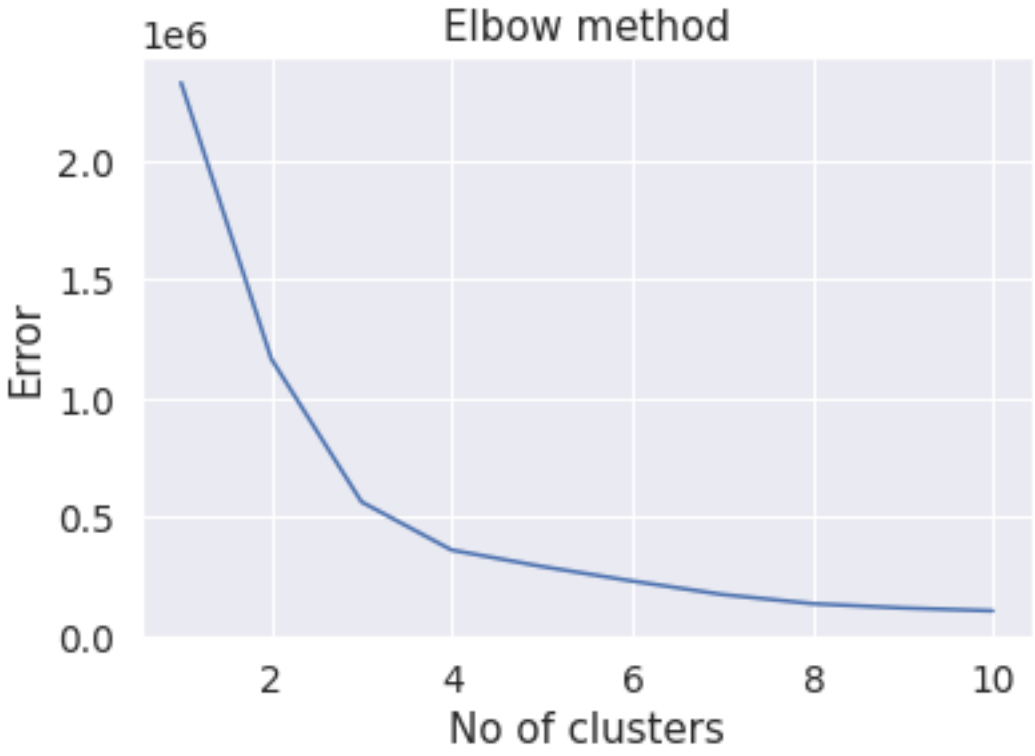
Gaussian mixture models are a probabilistic model for representing normally distributed subpopulations within an overall population. Mixture models in general don't require knowing which subpopulation a data point belongs to, allowing the model to learn the

subpopulations automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning.

The input data is as follows:

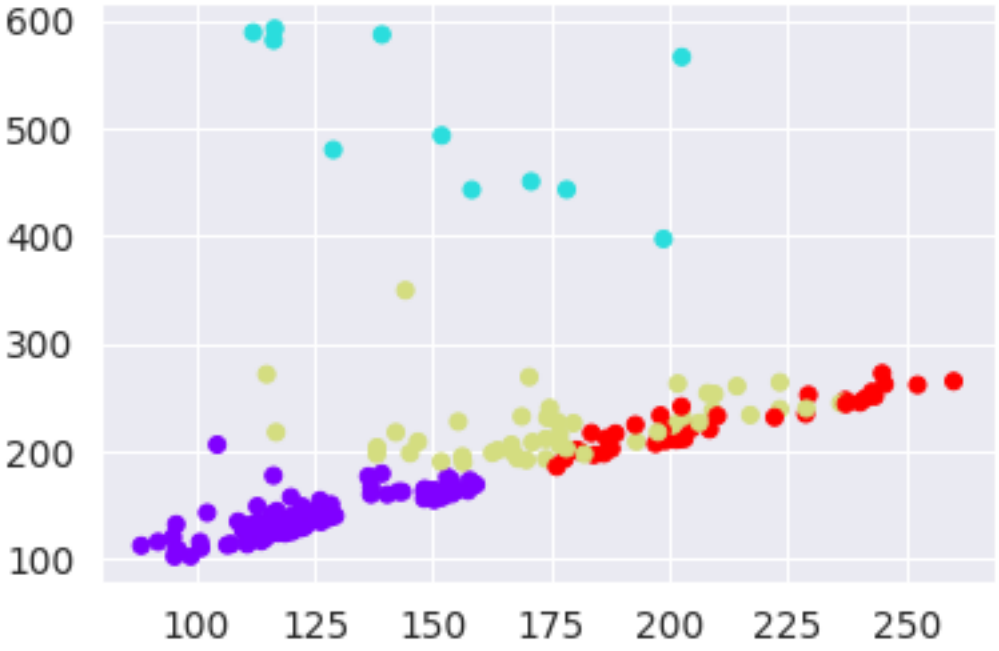


In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

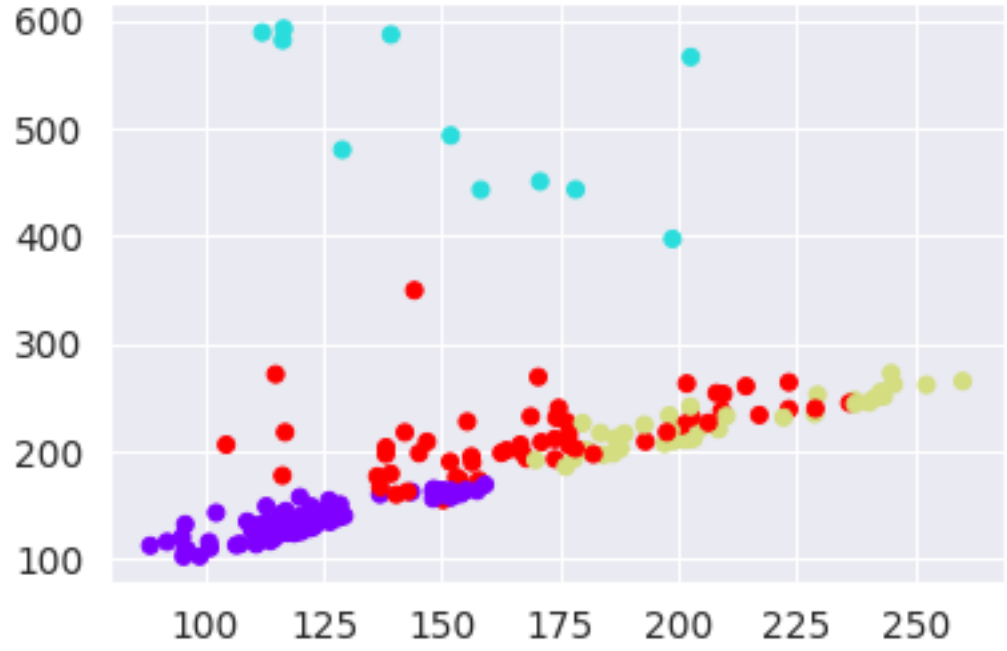


Based on the plot we determine that we have to use a cluster value of K = 4.

(i) Plot from K Means Clustering



(ii) Plot from Gaussian Mixture Model



From the graphs we observe that Gaussian Mixture Model gives better clusters and better categorizes the input data. Due to overlap the K-Means Clustering cannot provide ideal clusters.

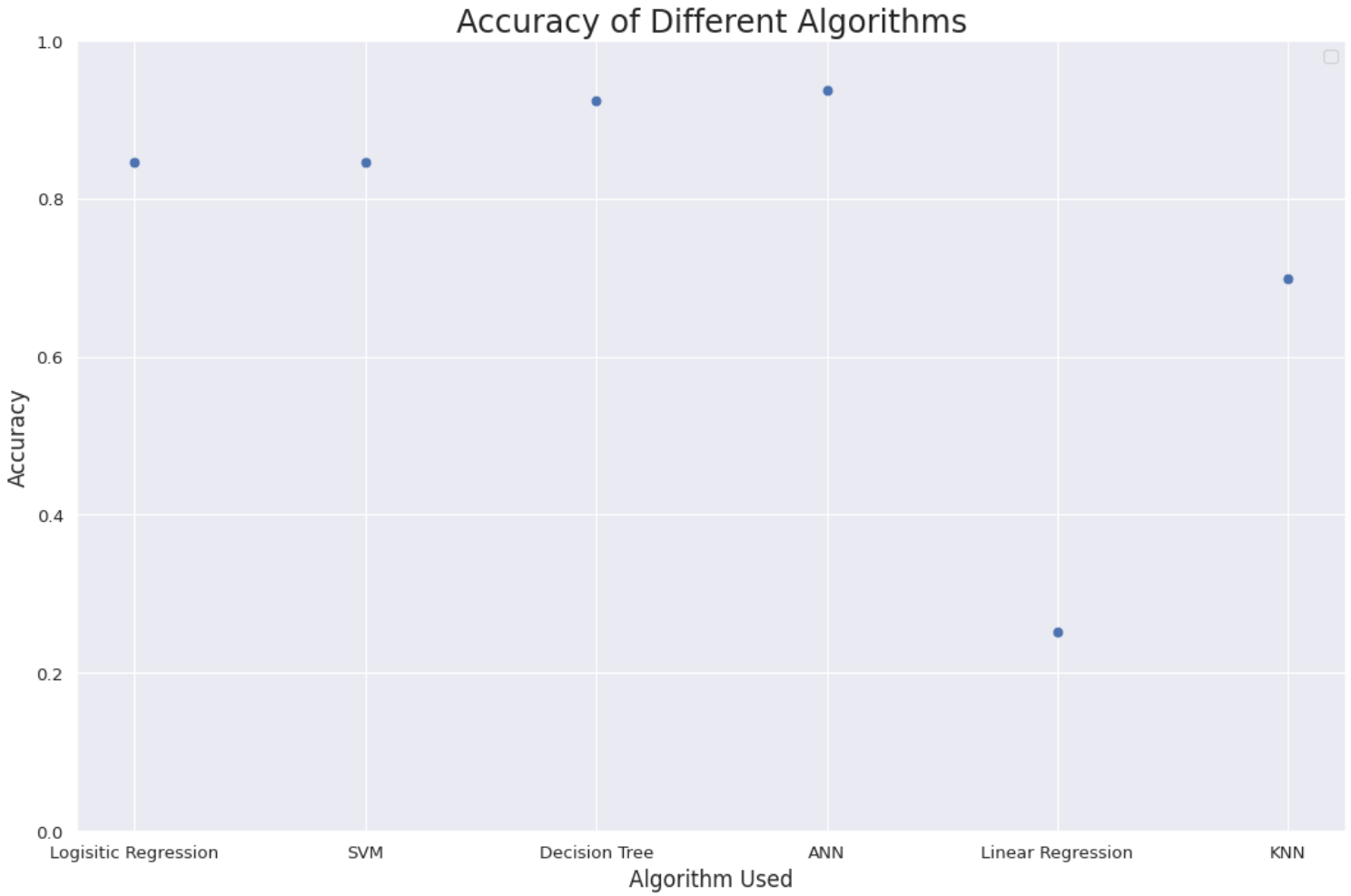
CASE STUDY 8: K Nearest Neighbours (KNN)

The *k*-nearest neighbours algorithm (*k*-NN) is a non-parametric classification method. *k*-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales
then normalizing the training data can improve its accuracy dramatically

Finally after fitting we obtained the following result:

→ Training accuracy of **87.65%** and the test set accuracy is **69.93%** for KNN

COMPARING THE MODELS



Of the different algorithms used Decision Tree Classifier and Artificial Neural Network Algorithm provided the best prediction values of 89.7% and 93.37% respectively.

Linear Regression Algorithm performed the poorest (25%) as it tries to find a linear relation between input and output data, however, that is not true for our dataset. It has been subject to underfitting.

SVM and Logistic Regression also provide a fairly decent prediction value (84.6%), however, the reduced values maybe due to the overlapping nature of dataset which the two could not take into account properly and underperformed.
