INTRODUCTION TO DATA SCIENCE

BITI 2513

REPORT OF CLUSTERING HEART DISEASE PATIENT DATA

LECTURER:

PROFESSOR MADYA DR SHARIFAH SAKINAH BINTI SYED AHMAD

GROUP MEMBER:

No	Name	Matric No.
1	VISHWAREETA VANOO	B031810196
2	PREVINA MUNUGANAN	B031810286
3	JEYSHALINI TEVOSHA	B031810246
4	ZAITI AKTA BINTI ZAHARUDDIN	B031810365

INDEX

INTRODUCTION

Heart disease has been around as early as the 1900s and is the number one killer of men and women. There are several factors can increase the risk of developing heart disease. Some of these factors are beyond your control like family history, but deciding to lead a healthier lifestyle will prevent this problem most of the time. Getting a regular exercise, reducing smoking and eating healthier foods are ways of reducing heart diseases.

In this project, our problem domain is Healthcare. Doctors research previous cases to figure out how to better handle their patients. A patient with a similar history of health or symptoms to a prior patient will benefit from being treated the same way. This project examines whether we could group patients together to target treatments using some machine learning techniques.

OBJECTIVE

1. To cluster heart disease patients.
2. To learn the feature importance
3. To visualise the dataset
4. To explore machine learning techniques
5. To produce a clustering model with a better performance

GOAL

The main goal of this project is to ensure healthy lives and promote well-being for all at all ages. ‘Good health and well-being’ is the 3rd goal of Sustainable Development Goals. As a developing computer scientist, while advancing computer science as a discipline, we can and should play a key role in helping to address societal and environmental challenges in pursuit of a sustainable future. The goal of this project is to be a part of such a responsibility.

Besides that, our goal is also to contribute the knowledge of Artificial Intelligence in Healthcare. Although Artificial Intelligence has reached higher places, medical field still does not give a warming welcome for it in our country. It is mainly because of the concern on the medical errors it might cause. The goal of our project is to produce a better model that would reduce the errors and give confidence to the healthcare.

LITERATURE REVIEW

A lot of research has been carried out for medical data analysis to discover the hidden pattern and extract useful information from large clinical datasets by applying different supervised and unsupervised algorithms. The number of heart patients are increasing worldwide. Various heart disease datasets are provided and different techniques are being addressed in many existing literature regarding the heart disease patients. The gap in the existing literature was considered as one of the motivations for this study.

Firstly, a research had proposed the performance of clustering algorithm using heart disease dataset. They evaluated the performance and prediction accuracy of some clustering algorithms. The performance of clusters will be calculated using the mode of classes to clusters evaluation. They proposed Make Density Based Cluster with the prediction accuracy of 85.8086%, as the most versatile algorithm for heart disease diagnosis. Secondly, a research was proposed to predict heart disease using K-Means algorithm. The existing methods used to mine the hidden information in medical data for heart disease are decision tree, naive bayes and neural network. In this research, the K-Means clustering examines the prediction of high speed and accuracy result compared to those existing techniques. It is found that various clusters are commonly used for prediction of heart disease in different studies. A detailed analysis of clustering algorithms thus gives an insight into clustering techniques. This analysis is of great importance to medical practitioners who wish to predict heart failure at an appropriate step in their progress.

Our main focus is to process the data to get the useful information and explored the hidden pattern. In this paper, we use the dataset provided by data.gov repository. Preprocessing steps and performance of different supervised and unsupervised learning classifiers are described in the following section.

DATA SOURCE

The dataset used of this project is “austin-public-health-diabetes-self-management-education-participant-demographics-2015-2017-1” from data.gov Repository. Th dataset initially contained 25 variables and the unwanted variables which such as Class, Class language, Year, Insurance type, Medical home type, race, ethnicity, education and 4 types of zip codes which are not related to our country and also our problem domain.

Link : https://catalog.data.gov/dataset/austin-public-health-diabetes-self-management-education-participant-demographics-2015-2017

TOOL

The tools that were used in this project are Python programming language and Microsoft Excel.

METHOD

i) DATA MANAGEMENT

The goal of this data management is to describes the data type, to visualise the data and to improvise the structure to make it more usable for the research project.

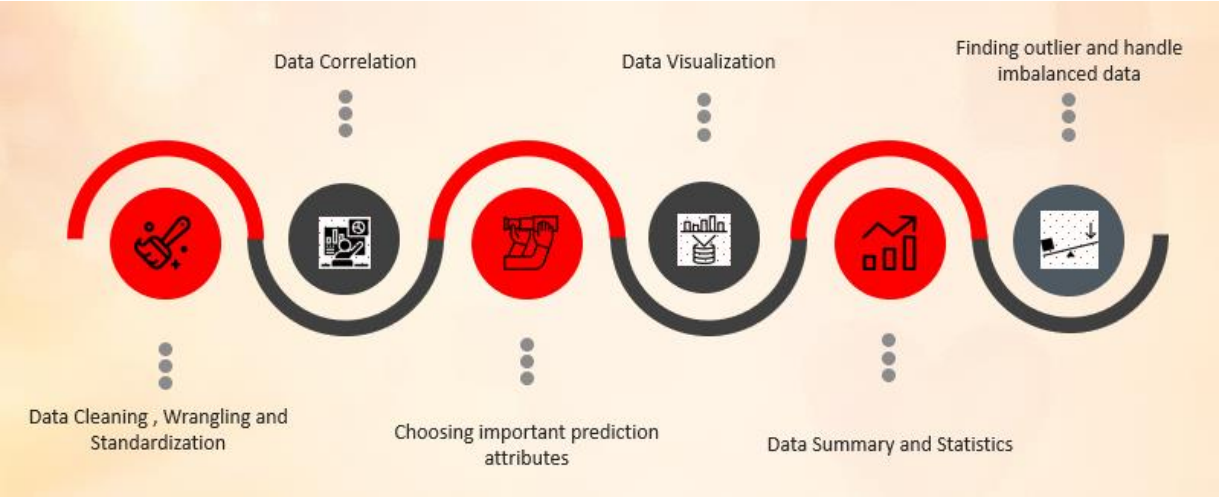
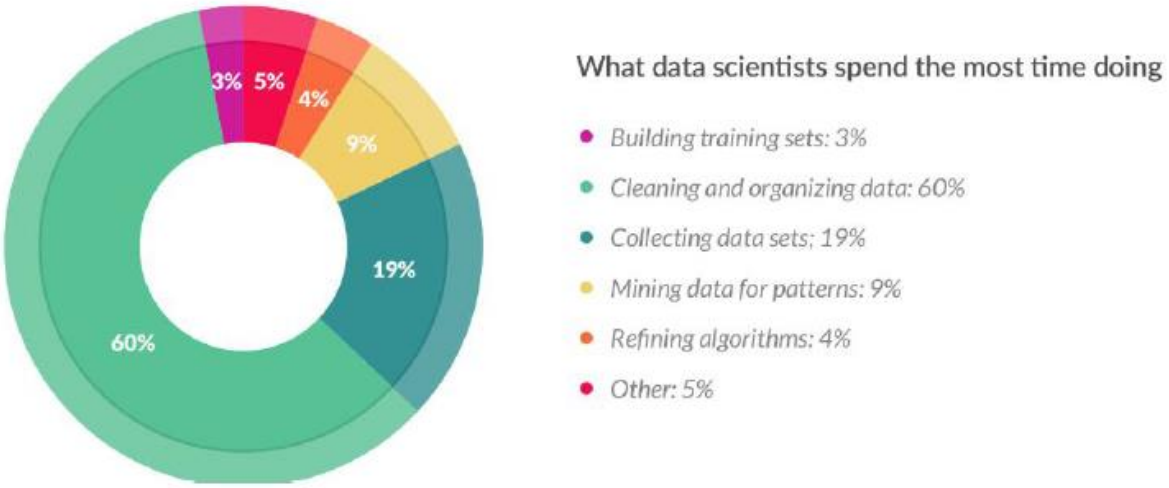


Figure 1: Data management flow

a) DATA CLEANING, WRANGLING AND STANDARDIZATION

Data collection in any organization is mostly done by non-experts. This leads to data being entered in various formats which leaves us with messy data. The pre-processing of dataset before using it to execute any process or evaluation is very important. Data cleaning plays a very important part although it can seem tedious.



From the pie chart above, cleaning and organization of data is very important and has the highest weightage among all other steps.

From the output of the first look into the dataset, we can see that it is a rather messy dataset. Thus, cleaning and preparation of the dataset were carried out. Next, we get rid of unwanted columns. The columns that are selected to be removed are redundant and unnecessary for our project. From 25 columns, it became 9 columns after the removal of unwanted columns. In the dataset, many rows had NaN values. Since our dataset is big enough, we removed rows with NaN values without risking the main outcome of our project. Initially we had 1688 rows in the dataset. After removing rows, we had 1369 rows. The dataset also was found to have many unstandardized inputs so we did standardisation too.

Apart from that, In the case of a dataset, not every column (variable) will necessarily have an effect on the output variable. If we add these irrelevant variables in the model, it will just make the model worst. Correlation is often the first step towards understanding these relationships, and then building better business and statistical models. Correlation or correlation coefficient numerically captures the association between two variables. High correlation variables are more linearly dependent, and thus have nearly the same effect on the dependent variable. And if there is a high correlation between two variables, we can remove one of the two features. Apart from that we could also remove variables that have no correlations with any of the variables or less correlation than that are intended.

b) DATA CORRELATION AND CHOOSING IMPORTANT ATTRIBUTES

Cramer's V and Chi square test

Cramér's V is a number between 0 and 1 that indicates how strongly two categorical variables are associated. A measure that does indicate the strength of the association is Cramér's V, defined as

The purpose of the Chi-square test is to check how likely an observed distribution is due to chance. It is often called a "goodness of fit" statistic, as it calculates how well the observed data distribution matches with the predicted distribution if the variables are independent. The chi-square independence test is a procedure for testing if two categorical variables are related in some population.

The result of the test is a test statistic that has a chi-squared distribution and can be interpreted to reject or fail to reject the assumption or null hypothesis that the observed and expected frequencies are the same. The variables are considered independent if the observed and expected frequencies are similar, that the levels of the variables do not interact, are not dependent. We can interpret the test statistic in the context of the chi-squared distribution with the requisite number of degrees of freedom and p- value.

The p-value and a chosen significance level (alpha), the test can be interpreted as follows:

- If p-value <= alpha: significant result, reject null hypothesis (H0), dependent.
- If p-value > alpha: not significant result, fail to reject null hypothesis (H0), independent.

The Cramer's V is applied to check whether any two variables are highly correlated so that they might have the chance of elimination. There are no high correlations between the potential predictive variables. There is no need for eliminating any variables as there are no higher correlations between the potential predictive variables. Next, we applied chi square test to see which variables are important and not important in predictive task. From the chi-square test we found that 'Tobacco use', 'Sugar-Sweetened Beverage Consumption' and 'Exercise' have higher p-value which makes them less important.

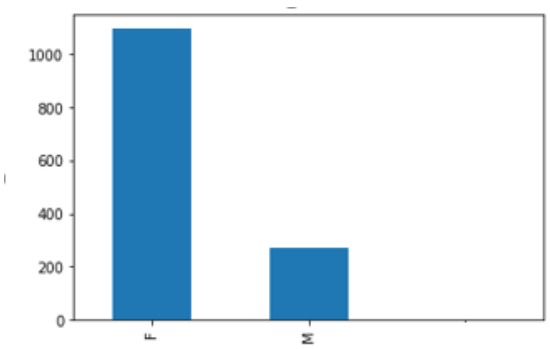
Based on our domain research we decided to keep 'Tobacco use' and 'Exercise'. According to (https://www.cdc.gov/heartdisease/index.htm) Centers for Disease Control and Prevention, they have created a quality improvement tool that related to tobacco usage named 'The Tobacco Cessation Change Package (TCCP)' which shows the important of that variable. Apart from that, the same organization promote physical activity as to reduce heart disease. Considering such knowledge, we do not want to eliminate 'Exercise' variable as well.

Why Cramer's V and Chi-square test?

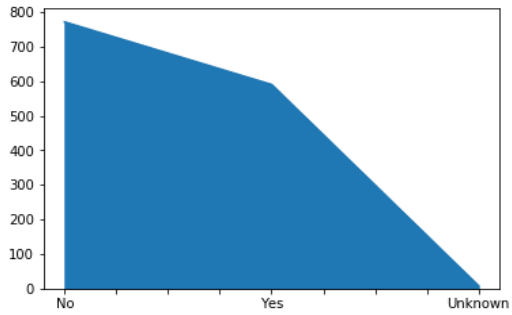
The advantages of Chi-square include its robustness in terms of data distribution, its ease of measurement, the detailed information that can be obtained from the study because it is a statistical method, its use in studies for which parametric assumptions cannot be fulfilled and its versatility in data handling.

Apart from that, there are lots of feature selection methods but it is important to choose the best one for a categorical variable. Since the variables in our dataset are mostly categorical, we found that Cramer's V and Chi square test are the widely used and best performed tests for feature selection with categorical variables.

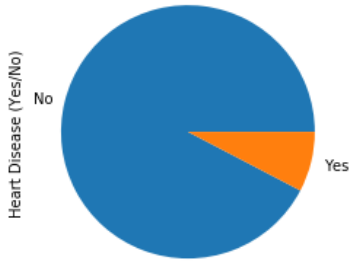
c) DATA VISUALISATION AND STATISTICAL SUMMARY



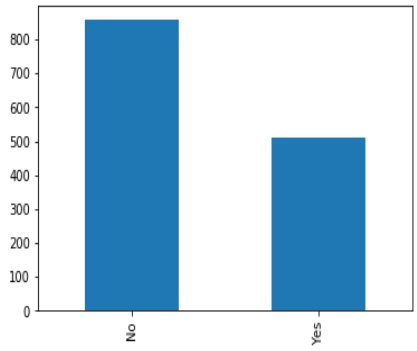
This figure shows that majority female takes the health test rather than male



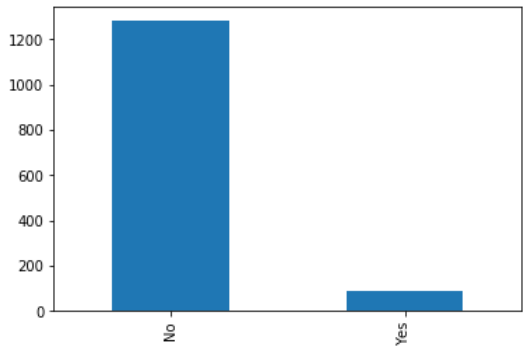
This figure shows that majority of the people had no diabetes status rather than the people that has diabetes status and less than 50 people doesn't take diabetes check-up.



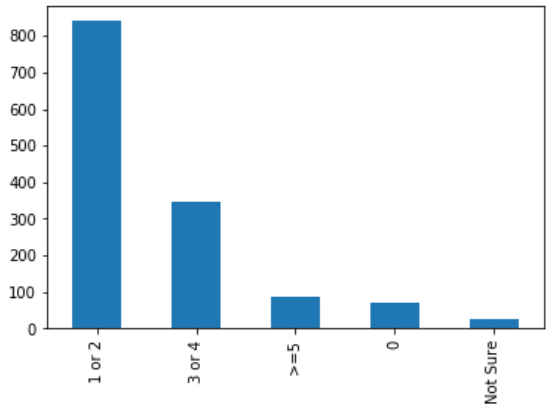
This figure shows the least amount of people that have heart disease rather than people that doesn't have heart disease.



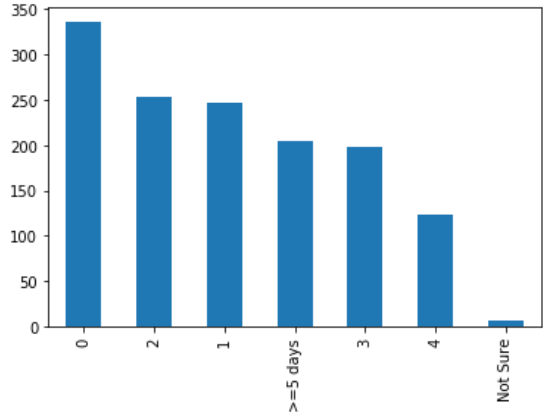
This figure shows more than 500 people have a high blood pressure problem and more than 800 people doesn't have high blood pressure.



This figure shows that the number of people that use tobacco is 10 times lesser than the people that didn't use tobacco.



This figure shows that there are still have least of people that doesn't sure about their fruit and vegetables intake in daily life.

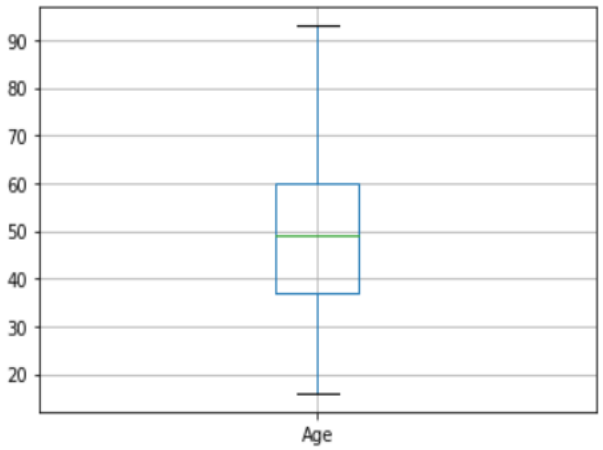


This figure shows that majority of people out there don't exercise in their daily life.

```
Mean Age: 49.18699780861943
Sum of Age: 67337
Max Age: 93
Min Age: 16
Count of Age: 1369
Median Age: 49.0
Std of Age: 15.16358481704277
Var of Age: 229.93430450365003
```

The statistical summary for age variable which is the only continuous variable.

d) HANDLING OUTLIER



From the figure above, we can know these statistical things:

- The bottom horizontal line of box plot is minimum value.
- First black horizontal line of rectangle shape of the box plot is First quartile of 25%
- Second black horizontal line of rectangle shape of the box plot is Second quartile or 50% or median.
- Third black horizontal line of rectangle shape of the box plot is third quartile or 75%
- Top black horizontal line of rectangle shape of the box plot is maximum value.
- There is no outlier detected, thus there is no outlier to be handled.

ii) DATA MODELLING AND VALIDATION

a) CLUSTERING TECHNIQUES

KMEANS CLUSTERING

Kmeans clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the dataset structures. The goal of kmeans is to group data points into distinct non-overlapping subgroups.

K-Means starts by randomly defining *k* centroids. From there, it works in iterative (repetitive) steps to perform two tasks:

1. Assign each data point to the closest corresponding centroid, using the standard Euclidean distance. In layman's terms: the straight-line distance between the data point and the centroid.

2. For each centroid, calculate the mean of the values of all the points belonging to it. The mean value becomes the new value of the centroid.

Once step 2 is complete, all of the centroids have new values that correspond to the means of all of their corresponding points. These new points are put through steps one and two producing yet another set of centroid values. This process is repeated over and over until there is no change in the centroid values, meaning that they have been accurately grouped. Or, the process can be stopped when a previously determined maximum number of steps has been met.

MEAN SHIFT CLUSTERING

Mean shift clustering is a simple and flexible clustering technique that has several nice advantages over other approaches. Mean shift clustering aims to discover “blobs” in a smooth density of samples. It is a centroid-based algorithm, which works by updating candidates for centroids to be the mean of the points within a given region. These candidates are then filtered in a post-processing stage to eliminate near-duplicates to form the final set of centroids.

The mean shift algorithm is a nonparametric clustering technique which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters.

Given n data points x_i , i = 1, ..., n on a d-dimensional space Rd , the multivariate kernel density estimate obtained with kernel K(x) and window radius h is

$$f(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right).$$

For radially symmetric kernels, it suffices to define the profile of the kernel k(x) satisfying

$$K(\mathbf{x}) = c_{k,d} k(\|\mathbf{x}\|^2)$$

where $c_{k,d}$ is a normalization constant which assures K(x) integrates to 1. The modes of the density function are located at the zeros of the gradient function ∇f(x) = 0.

The gradient of the density estimator is

$$\begin{aligned} \nabla f(\mathbf{x}) &= \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}) g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \\ &= \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)\right] \left[\frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}\right] \end{aligned}$$

where g(s) = −k 0 (s). The first term is proportional to the density estimate at x computed with kernel G(x) = $c_{g,d}$g(||x||^2) and the second term

$$\mathbf{m}_h(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}$$

is the mean shift. The mean shift vector always points toward the direction of the maximum increase in the density. The mean shift procedure, obtained by successive

• computation of the mean shift vector $m_h(x^t)$,

• translation of the window $x^{t+1} = x^t + m^h(x^t)$

is guaranteed to converge to a point where the gradient of density function is zero.

RANDOM FOREST CLASSIFICATION

Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object. It is an ensemble tree-based learning algorithm. Ensemble algorithms are those which combines more than one algorithm of same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object. Random Forest Prediction for a classification problem:

f(x) = majority vote of all predicted classes over B trees

These above results are aggregated, through model votes or averaging, into a single ensemble model that ends up outperforming any individual decision tree's output.

b) VALIDATION PROCESS

CROSS VALIDATION

Cross-validation is a statistical method used to estimate the skill of machine learning models. It is also a resampling procedure used to evaluate machine learning models on a limited data sample.

It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

The general procedure is as follows:

1. Shuffle the dataset randomly.

2. Split the dataset into k groups

3. For each unique group:

    1) Take the group as a hold out or test data set
    2) Take the remaining groups as a training data set
    3) Fit a model on the training set and evaluate it on the test set
    4) Retain the evaluation score and discard the model

4. Summarize the skill of the model using the sample of model evaluation scores

c) EVALUATION

ACCURACY

Accuracy refers to how close a sample statistic is to a population parameter . Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. The accuracy formula with using confusion matrix data:

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

PRECISION

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. The precision formula:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Immediately, you can see that Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive.

RECALL

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

F1-SCORE

$$F1 = 2\ x\ \frac{Precision * Recall}{Precision + Recall}$$

F1 Score is needed when you want to seek a balance between Precision and Recall. F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives.

IMPLEMENTATION

First of all, the non-numerical values were changed into numerical values. The dependent variable and the independent variables were separated accordingly for training purpose. Cross validation with 8-folds was used for validation process. All the parameters for used techniques (Kmeans clustering, Mean shift clustering and Random forest classification) were changed few times in order to get the best of the process.

RESULT

KMEANS CLUSTERING

8-Fold accuracy

```
========================================
8-Fold Accuracy :  50.40884672922616
========================================
```

Total accuracy

```
========================================
Accuracy(Total) =  57.85244704163623
========================================
```

Confusion Matrix

```
[[0.51862673 0.40467495]
 [0.01680058 0.05989774]]
============================
```

Precision, recall, f1score and support

```
============================================================
              precision    recall  f1-score   support

           0       0.97      0.56      0.71      1264
           1       0.13      0.78      0.22       105

    accuracy                           0.58      1369
   macro avg       0.55      0.67      0.47      1369
weighted avg       0.90      0.58      0.67      1369

============================================================
```

Execution Time (s)

```
==============================
Program Executed in 3.1415468
==============================
```

MEAN-SHIFT CLUSTERING

8-Fold accuracy

```
========================================
8-Fold Accuracy :  87.06905344757241
========================================
```

Total accuracy

```
====================================
Accuracy(Total) =  87.0708546384222
====================================
```

Confusion matrix

```
[[0.86486486 0.05843682]
 [0.07085464 0.00584368]]
==========================
```

Precision, recall, f1score, support

```
                    precision    recall  f1-score   support

                 0       0.92      0.94      0.93      1264
                 1       0.09      0.08      0.08       105

          accuracy                           0.87      1369
         macro avg       0.51      0.51      0.51      1369
      weighted avg       0.86      0.87      0.87      1369

============================================================
```

Execution time(s)

```
====================================
Program Executed in 85.61345349999999
====================================
```

RANDOM FOREST CLASSIFICATION

Confusion Matrix

```
=== Confusion Matrix ===
[[401  14]
 [ 32   5]]
```

Precision, Recall, f1-score, and support

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.93      0.97      0.95       415
           1       0.26      0.14      0.18        37

    accuracy                           0.90       452
   macro avg       0.59      0.55      0.56       452
weighted avg       0.87      0.90      0.88       452
```

AUC Scores

```
=== All AUC Scores ===
[0.78910488 0.70618306 0.77361246 0.69498539 0.8218111  0.68232717
 0.72176241 0.72930867]


=== Mean AUC Score ===
Mean AUC Score - Random Forest:  0.739886893170121
```

ANALYSIS

Comparison of the accuracy, precision, recall, f1score between kmeans clustering, mean shift clustering and random forest classification

	Kmeans Clustering	Mean Shift Clustering	Random Forest Classification
Accuracy	57.85	87.07	**90.00**
Precision (Macro avg)	0.55	0.51	**0.59**
Precision (Weighted avg)	**0.90**	0.86	0.87
Recall (Macro avg)	**0.67**	0.51	0.55
Recall (Weighted avg)	0.58	0.87	**0.90**
F1-Score (Macro avg)	0.47	0.51	**0.56**
F1-Score (Macro avg)	0.67	0.87	**0.88**

From the comparison above we found that Random forest classification is a better performing model for the dataset.

Why the accuracy of Mean shift clustering is higher than k means clustering?

The main reason is because k-means clustering works better only with data that is balanced. It does not work well with clusters of different size and different density. This is why some of the accuracy measures like weighted average for precision and macro average for rtecall for the biggest cluster is better than the others.

Moreover, as the number of dimensions increases, a distance-based similarity measure converges to a constant value between any given examples. This makes it perform poorly with our dataset which contains more dimensions and it is less accurate in updating the centroid.

Furthermore, mean shift clustering also models non-convex shaped data and doesn't assume any prior shape like spherical, elliptical and so on which shows it does not affect by the balance of the dataset.

Apart from that, mean shift clustering works well with medium and small size dataset where it manages to cover all the datapoints faster with its window in hierarchical method.

Why Random forest classification is a better model?

The difference is that random forests build multiple decision trees on random data subsets and then average results. This often allows predictions on unseen data to be much more accurate than with the decision trees.

Moreover, random forest classification is aided by the predefined labels which makes them a better model when encountering new data. Whereas, in k-means clustering and mean shift clustering there is no predefined label so the prediction will be less aided.

FUTURE IMPLEMENTATION

For our project, the dataset was found to be rather imbalanced. An imbalanced dataset means instances of one of the two classes is higher than the other. For future implementations, techniques such as oversampling or undersampling can be used to handle the imbalanced dataset. Handling imbalanced data is very prominent as imbalanced datasets can affect the accuracy of the model and fails most algorithms in finding a proper solution. Besides handling imbalanced data, data mining techniques can be implemented to produce more accurate predictions like neural networks.

CONCLUSION

Heart disease is one of the biggest problems faced by the healthcare industry. Prediction models which are accurate can help in detecting heart disease at an earlier stage thus, enabling prevention. As the saying goes, 'Prevention is better than cure', it is solely better to take preventive measures than healing measures when it may be too late for some cases.

REFFERENCE

1. https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15
2. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
3. https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/
4. https://scikit learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html#:~:text=Mean%20shift%20clustering%20using%20a,points%20within%20a%20given%20region.&text=If%20not%20set%2C%20the%20seeds%20are%20calculated%20by%20clustering.
5. https://towardsdatascience.com/3-steps-to-a-clean-dataset-with-pandas-2b80ef0c81ae
6. https://github.com/mramshaw/Data-Cleaning
7. https://www.kdnuggets.com/2019/11/data-cleaning-preprocessing-beginners.html
8. https://medium.com/@rrfd/cleaning-and-prepping-data-with-python-for-data-sciencebest-practices-and-helpful-packages-af1edfbe2a3
9. https://blogs.oracle.com/datascience/introduction-to-correlation
10. https://statisticsbyjim.com/basics/correlations/
11. https://cmdlinetips.com/2019/08/how-to-compute-pearson-and-spearman-correlationin-python/
12. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6172294/
13. https://www.cdc.gov/heartdisease/index.htm
14. https://www.geeksforgeeks.org/box-plot-visualization-with-pandas-and-seaborn/
15. https://www.researchgate.net/publication/335827126_Heart_Disease_Prediction_with_Data_Mining_Clustering_Algorithms

Github link : https://github.com/HeraGoodwood/ClusteringHeartDisease