جامعة الملك سعود
King Saud University
1957

STROKE PREDICATION

SUPERVISED BY: DR. MASHAEL ALDAYEL

TABLE OF CONTENT

OUR PROBLEM:

Stroke is the second-leading cause of death and the most common global cause of disability. WHO estimates that 1 in 4 persons may experience a stroke during their lifetime; because strokes can occur at any time and to anyone, regardless of age, we have chosen to concentrate on this dataset. Given the sudden nature of strokes, we intend to investigate and analyze the data to provide predictions on what are some risk factors and shed light on the types of people who are likely to experience one, allowing for future changes in lives.

DATA:

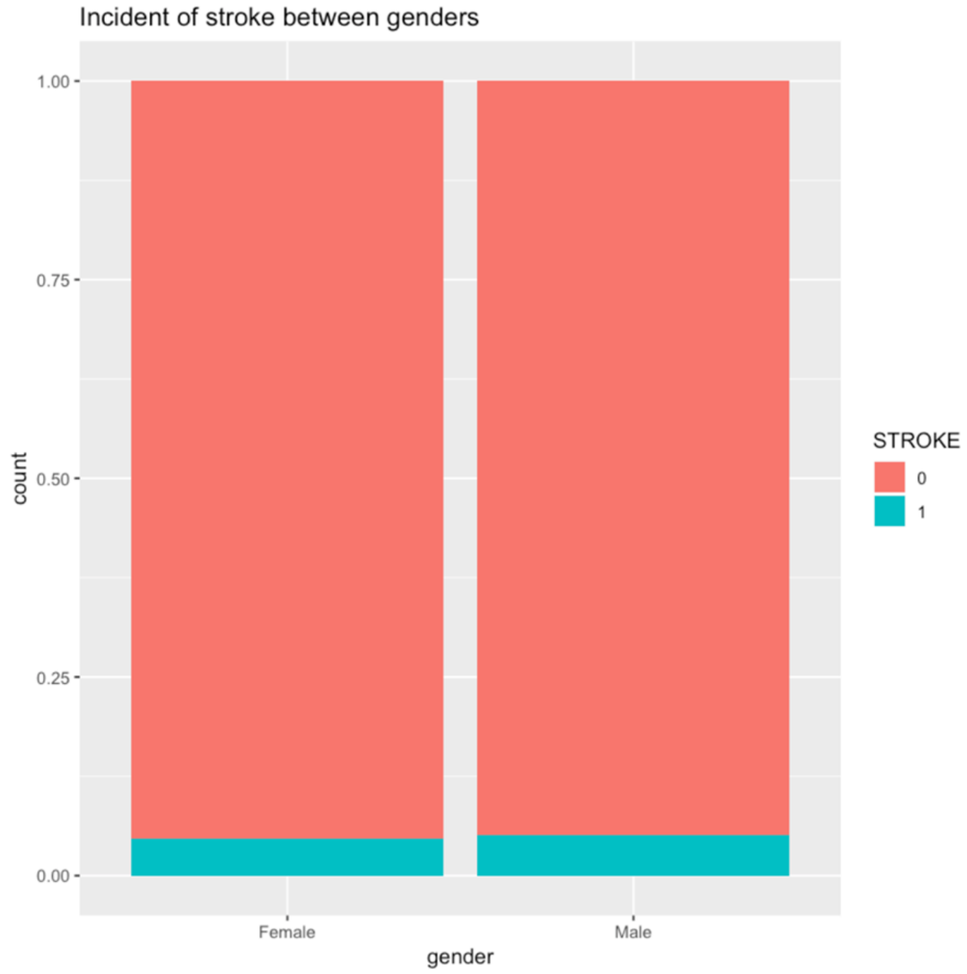
Among the 5110 objects in our dataset sample, 12 attributes are used to describe them. Our characteristics values are utilized to identify their types, such as the nominal for id, binary for gender, and numeric for age.

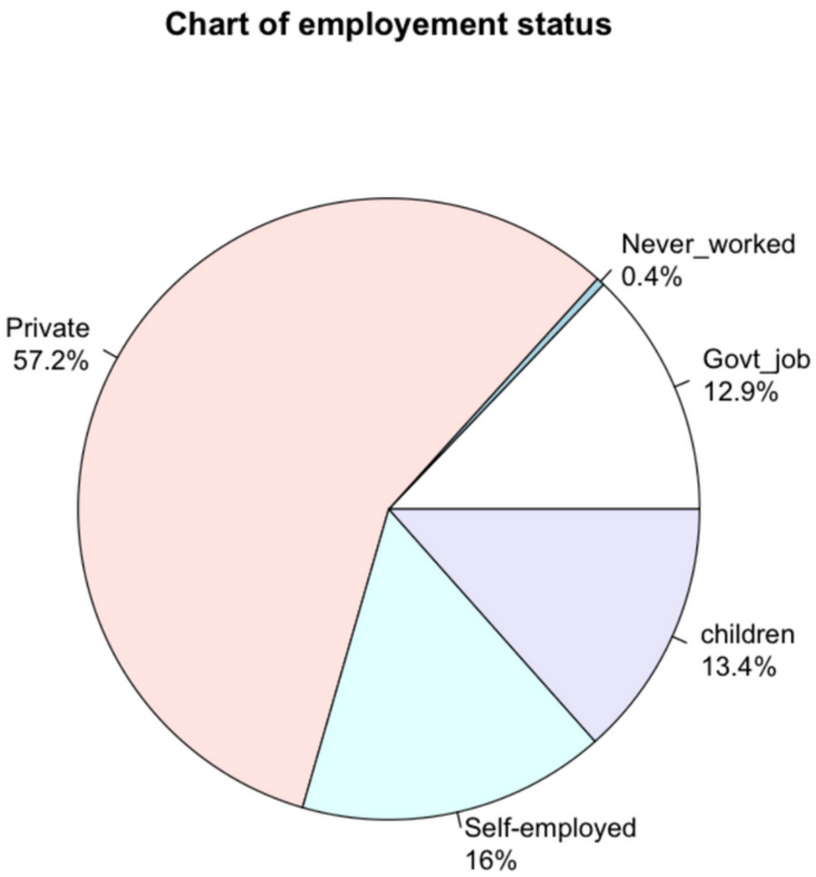
We had ever_married, work_type and residence_type.also and two attributes for hypertension and heart disease that took two values 1 and 0 to indicate whether they are suffered from it or not. The last attribute, "stroke", was described by two values 0 and 1 for the possibility of having a stroke or not as a result of analysis of the previous data, , which is what we aim to train our model to predict.

DATA DICTIONARY:

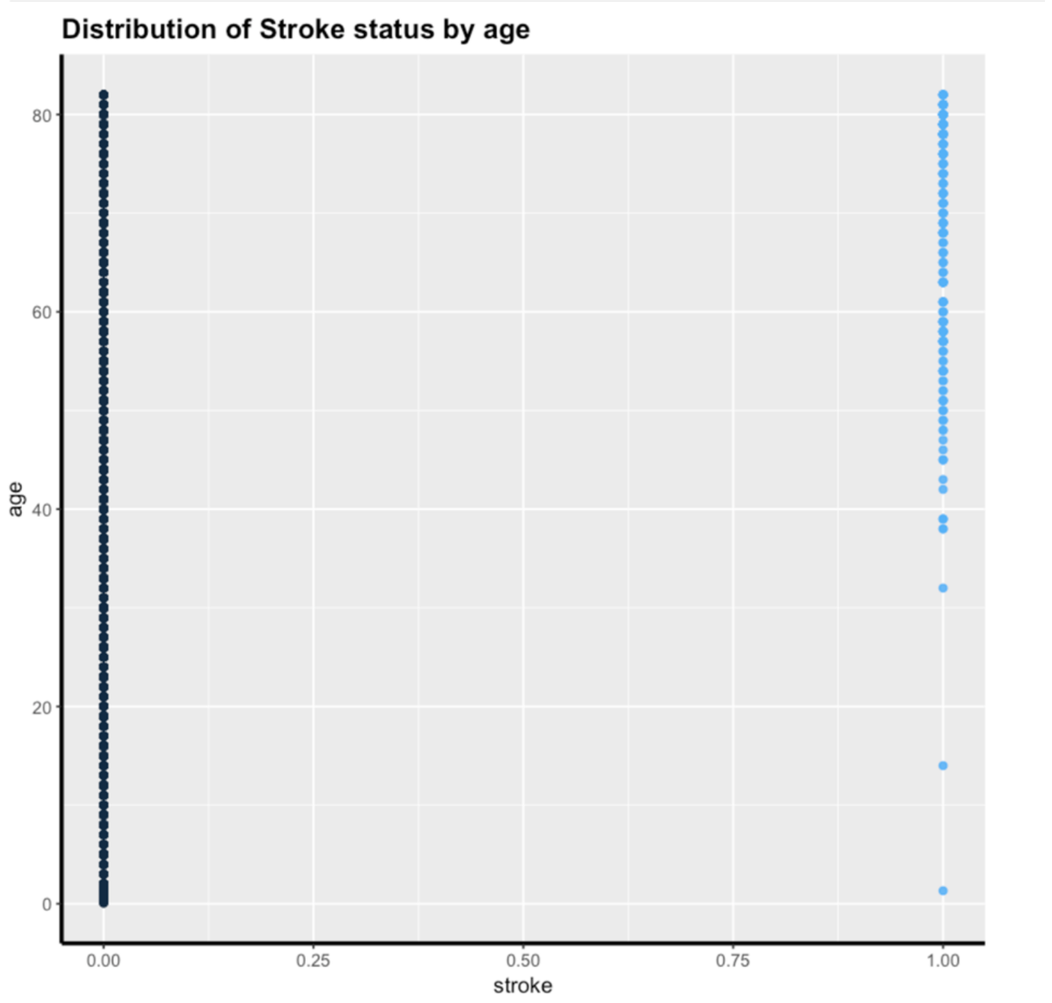
Attribute Name	Description	Data Type	Possible values
id	Unique id of the patient	Nominal	Range between 67-72940
gender	Gender of the patient	Binary	Female Male
age	Age of the patient	Numeric	Range between 0.08-82
hypertension	Hypertension binary feature, 1 means the patient has hypertension, 0 means they do not.	Binary	0,1
heart_disease	Heart disease binary feature, 1 means the patient has heart disease, 0 means they do not.	Binary	0,1
ever_married	Has the patient ever been married?	Binary	Yes No
work_type	Work type of the patient	Nominal	"Private" "Self-employed" "children" "Govt_job" "Never_worked"
residence_type	Residence type of the patient	Binary	"Urban" "Rural"
avg_glucose_level	Average glucose level in blood	Numeric	Range between 55.1-272
bmi	Body Mass Index	Numeric	Range between 10.3-97.6
smoking_status	Smoking status of the patient	Nominal	"never smoked" "Unknown" "formerly smoked" "smokes"
stroke	Stroke event, 1 means the patient had a stroke, 0 means not	Binary	0,1

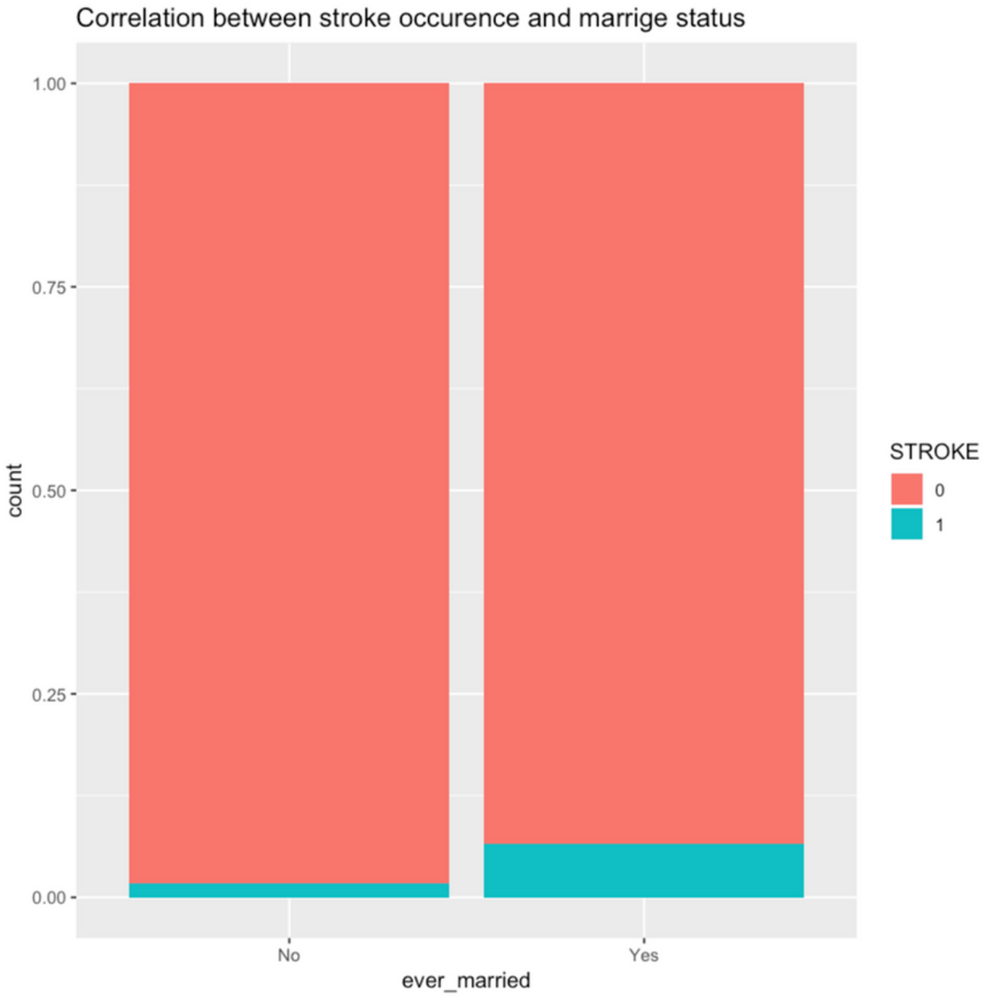
DATA REPRESENTATION





DATA REPRESENTATION





02.

DATA PREPROCESSING

DATA PREPROCESSING

- **Data Cleaning**
  - -Missing Values
  - -Noisy Data
- **Data Reduction**
  - -Attribute Subset Selection
- **Data Transformation**
  - -Encoding
  - -Normalization

DATA CLEANING

-Removed Null values in “bmi” by replacing them with the mean

```
# Checking missing values
sum(is.na(data))
```

```
## [1] 201
```

```
# Replacing null values with the mean
data$bmi[is.na(data$bmi)]<-mean(data$bmi, na.rm = TRUE)
```

-Detect outliers in “age”, “avg_glucose_level” and “bmi” and remove them

```
#detect Average glucose level outliers
OutAvg <- outlier(data$avg_glucose_level)
print(OutAvg)
```

```
## [1] 271.74
```

```
#Remove Average glucose level outlier
data <- data[data$avg_glucose_level != OutAvg, ]
```

DATA TRANSFORMATION

Encoding catagorical data to numeric representation [0,1,2,..]

```
## 6 56669   Male  81            0             0          Yes       Private
##   Residence_type avg_glucose_level      bmi  smoking_status stroke
## 1          Urban            228.69 36.60000 formerly smoked      1
## 2          Rural            202.21 28.89456    never smoked      1
## 3          Rural            105.92 32.50000    never smoked      1
## 4          Urban            171.23 34.40000          smokes      1
## 5          Rural            174.12 24.00000    never smoked      1
## 6          Urban            186.21 29.00000 formerly smoked      1
```

```
data$work_type = factor(data$work_type,levels = c("Govt_job","Private", "Self-employed","children","Never_worked"
), labels = c(5,4,3,2,1))
data$gender = factor(data$gender, levels = c("Male", "Female"), labels = c(1, 2))
data$ever_married= factor(data$ever_married, levels = c("No", "Yes"), labels = c(0, 1))
data$Residence_type= factor(data$Residence_type, levels = c("Urban", "Rural"), labels=c(1,2))
data$smoking_status= factor(data$smoking_status, levels = c("Unknown","never smoked", "formerly smoked","smokes")
, labels=c(1,2,3,4))
head(data)
```

```
##      id gender age hypertension heart_disease ever_married work_type
## 1  9046      1  67            0             1            1         4
## 2 51676      2  61            0             0            1         3
## 3 31112      1  80            0             1            1         4
## 4 60182      2  49            0             0            1         4
## 5  1665      2  79            1             0            1         3
## 6 56669      1  81            0             0            1         4
##   Residence_type avg_glucose_level      bmi smoking_status stroke
```

DATA TRANSFORMATION

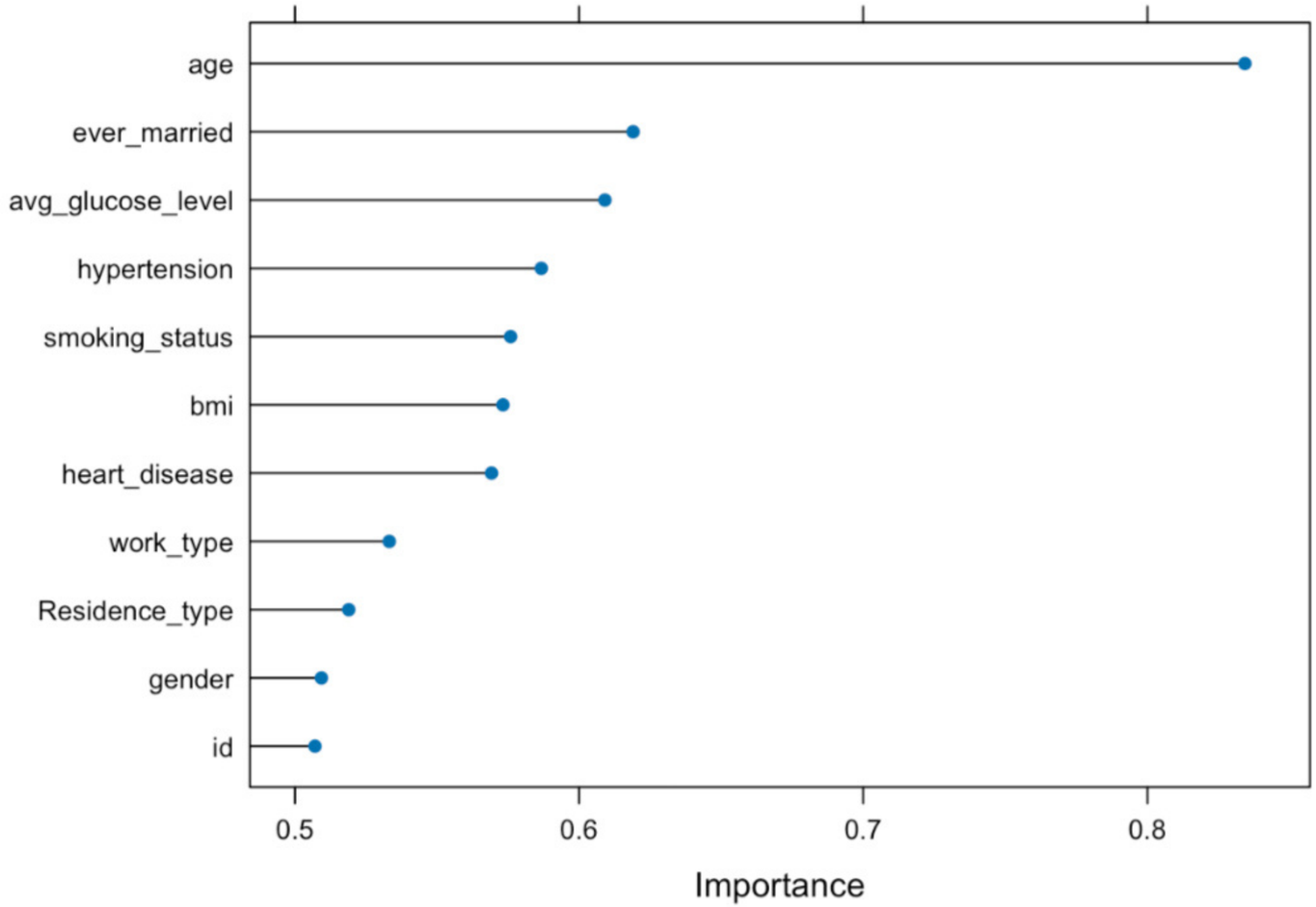
Normalize the numerical data (age, average glucose level, and BMI) using min-max scaling

```
#normalize data
normalize <- function(x){ return ((x - min(x))/ (max(x)- min(x)))}
data$avg_glucose_level= normalize(data$avg_glucose_level)
data$age= normalize(data$age)
data$bmi= normalize(data$bmi)
head(data)
```

```
##      id gender       age hypertension heart_disease ever_married work_type
## 1  9046      1 0.8167155            0             1            1         4
## 2 51676      2 0.7434018            0             0            1         3
## 3 31112      1 0.9755621            0             1            1         4
## 4 60182      2 0.5967742            0             0            1         4
## 5  1665      2 0.9633431            1             0            1         3
## 6 56669      1 0.9877810            0             0            1         4
##   Residence_type avg_glucose_level       bmi smoking_status stroke
## 1              1         0.8162622 0.3219094              3      1
## 2              2         0.6917325 0.2275956              2      1
## 3              2         0.2389014 0.2717258              2      1
## 4              1         0.5460403 0.2949816              4      1
## 5              2         0.5596313 0.1676867              2      1
## 6              1         0.6164880 0.2288862              3      1
```

DATA REDUCTION

After using random forest selection function to perform feature selection, we've decided to take the top 4 features

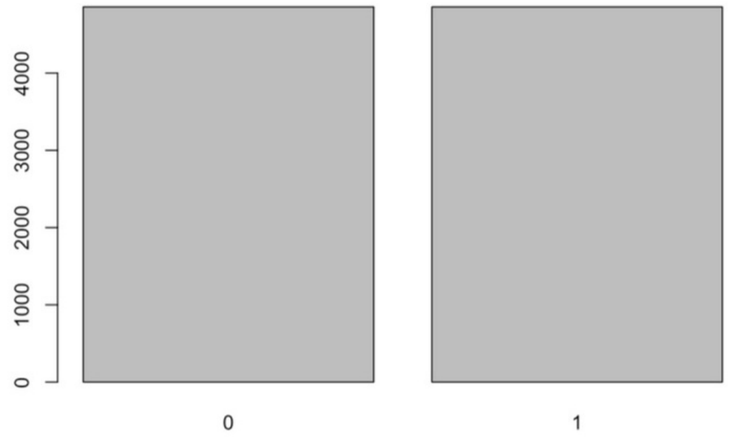


IMBALANCED DATASET

In the dataset, only 5% of all the individuals have experienced a stroke. Consequently, the model would achieve an accuracy of 95% by consistently predicting that individuals do not have a stroke.

Therefore, we oversampled our minority class to make ths stroke and non-stroke tuples 50/50

```
# upscaling the data
library('caret')
library('dplyr')
data<-upSample(data[,-5],data$stroke, yname="stroke")
plot(data$stroke)
```



03.

CLASSIFICATION

CLASSIFICATION

We took a balanced sample (1000 tuples) from our dataset. Our goal is to predict the class label (stroke) which has two values, yes or no.

We tried three different size of partitions and three attribute selection measures:

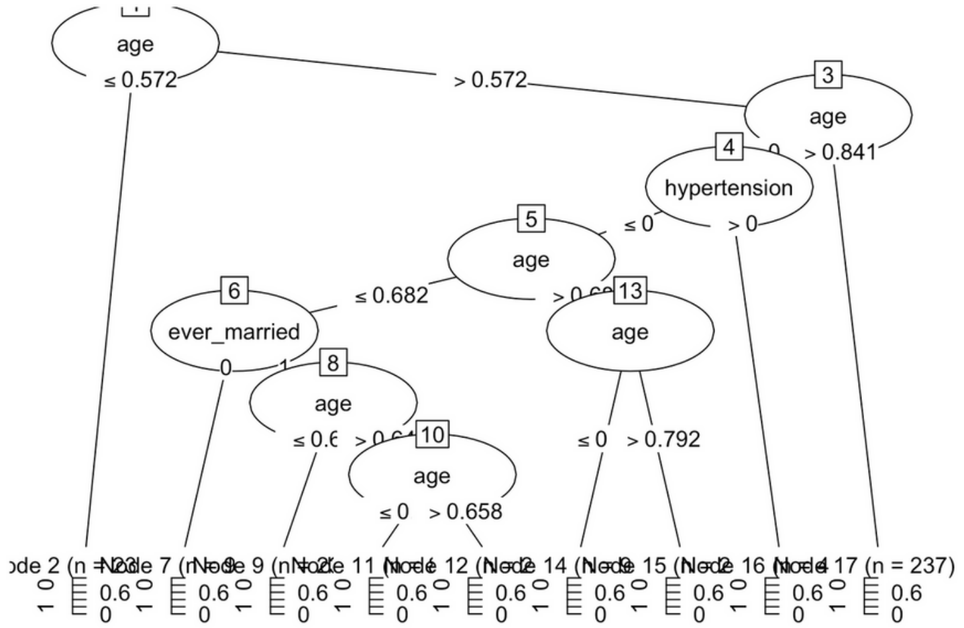
Gain Ratio
(C.50)

Gini Index
(rpart)

Information gain
(ctree)

GAIN RATIO (C.50)

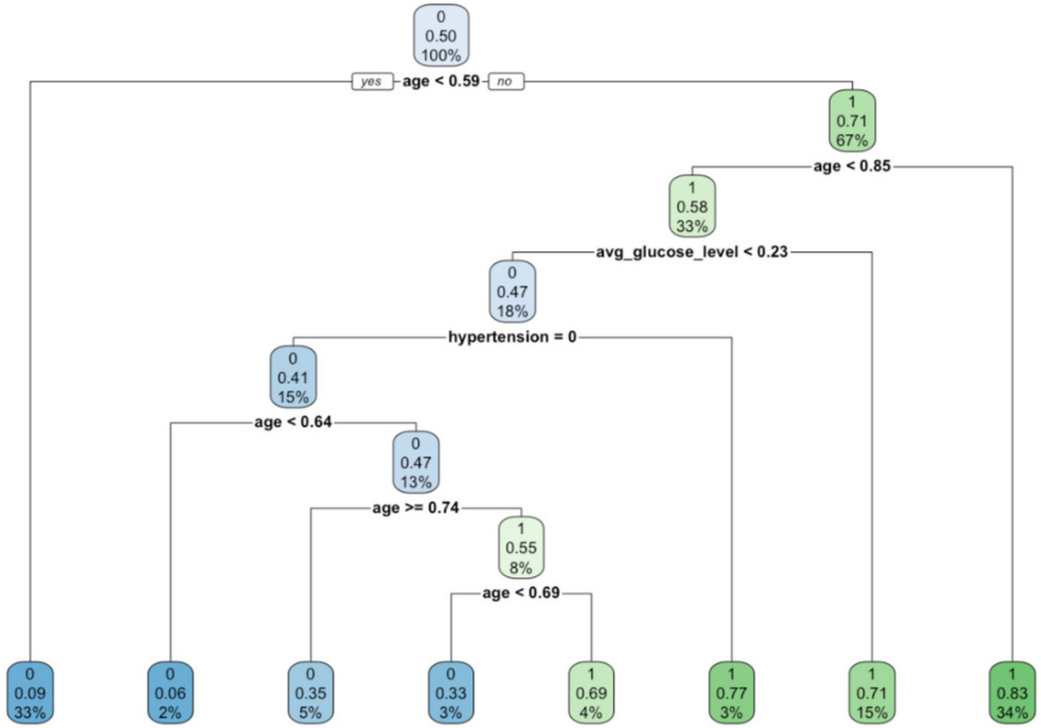
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 103  14
         1  47 136

               Accuracy : 0.7967
                 Sensitivity : 0.9067
                 Specificity : 0.6867
            Pos Pred Value : 0.7432
```



	70 % training set 30% testing set:	80 % training set 20% testing set:	85 % training set 15% testing set:
Accuracy	0.7967	0.795	0.78
Precision	0.7432	0.7323	0.7442
Sensitivity	0.9067	0.9300	0.8533
Specificity	0.6867	0.6600	0.7067

GINI INDEX (RPART)

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 70 14
         1 30 86

               Accuracy : 0.78
             Sensitivity : 0.8600
             Specificity : 0.7000
          Pos Pred Value : 0.7414
```



	70 % training set 30% testing set:	80 % training set 20% testing set:	85 % training set 15% testing set:
Accuracy	0.74	0.78	0.74
Precision	0.6875	0.7414	0.7045
Sensitivity	0.8800	0.8600	0.8267
Specificity	0.6000	0.7000	0.6533

INFORMATION GAIN (CTREE)

The “ctree” methods requires the continuous attributes to be discretized, and char values to be factors

```
cutPoints(data_class$age,data_class$stroke)
```

```
## [1] 0.3829423 0.5784457 0.8472630
```

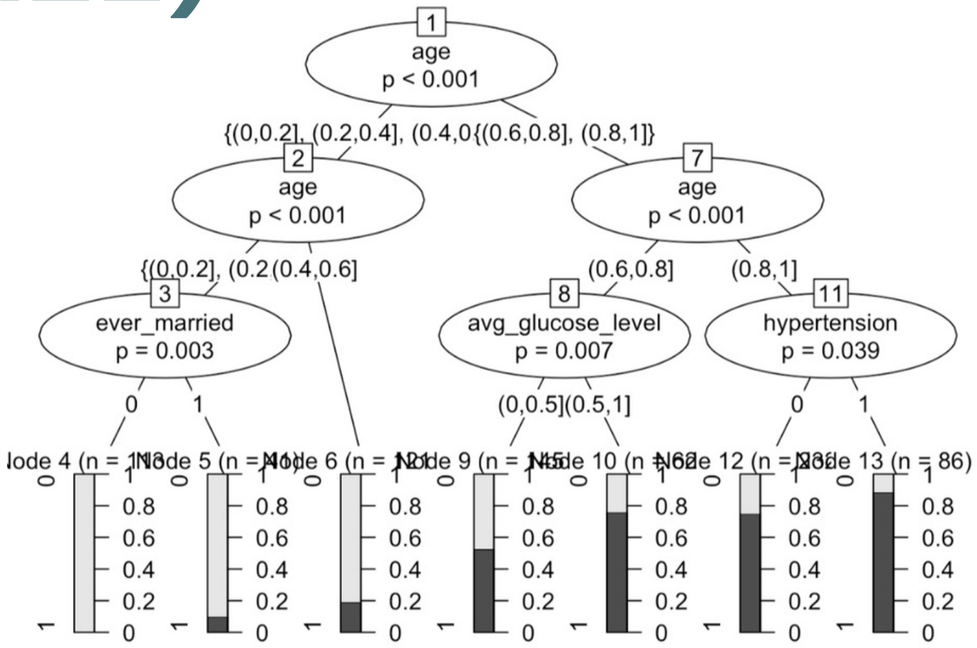
```
data_class$age= cut(data_class$age, breaks= seq(0,1, by=0.2),right=TRUE)

cutPoints(data_class$avg_glucose_level,data_class$stroke)
```

```
## [1] 0.5006114
```

```
data_class$avg_glucose_level    = cut(data_class$avg_glucose_level, breaks= seq(0,1, by=0.5),right=TRUE)

data_class$hypertension <- as.factor(data_class$hypertension )
data_class$ever_married   <- as.factor(data_class$ever_married  )
```

INFORMATION GAIN (CTREE)

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 60  8
         1 40 92

               Accuracy : 0.76
           Sensitivity : 0.9200
           Specificity : 0.6000
        Pos Pred Value : 0.6970
```



	70 % training set 30% testing set:	80 % training set 20% testing set:	85 % training set 15% testing set:
Accuracy	0.7433	0.76	0.7067
Precision	0.6780	0.6970	0.6505
Sensitivity	0.9267	0.9200	0.8933
Specificity	0.5600	0.6000	0.5200

04.

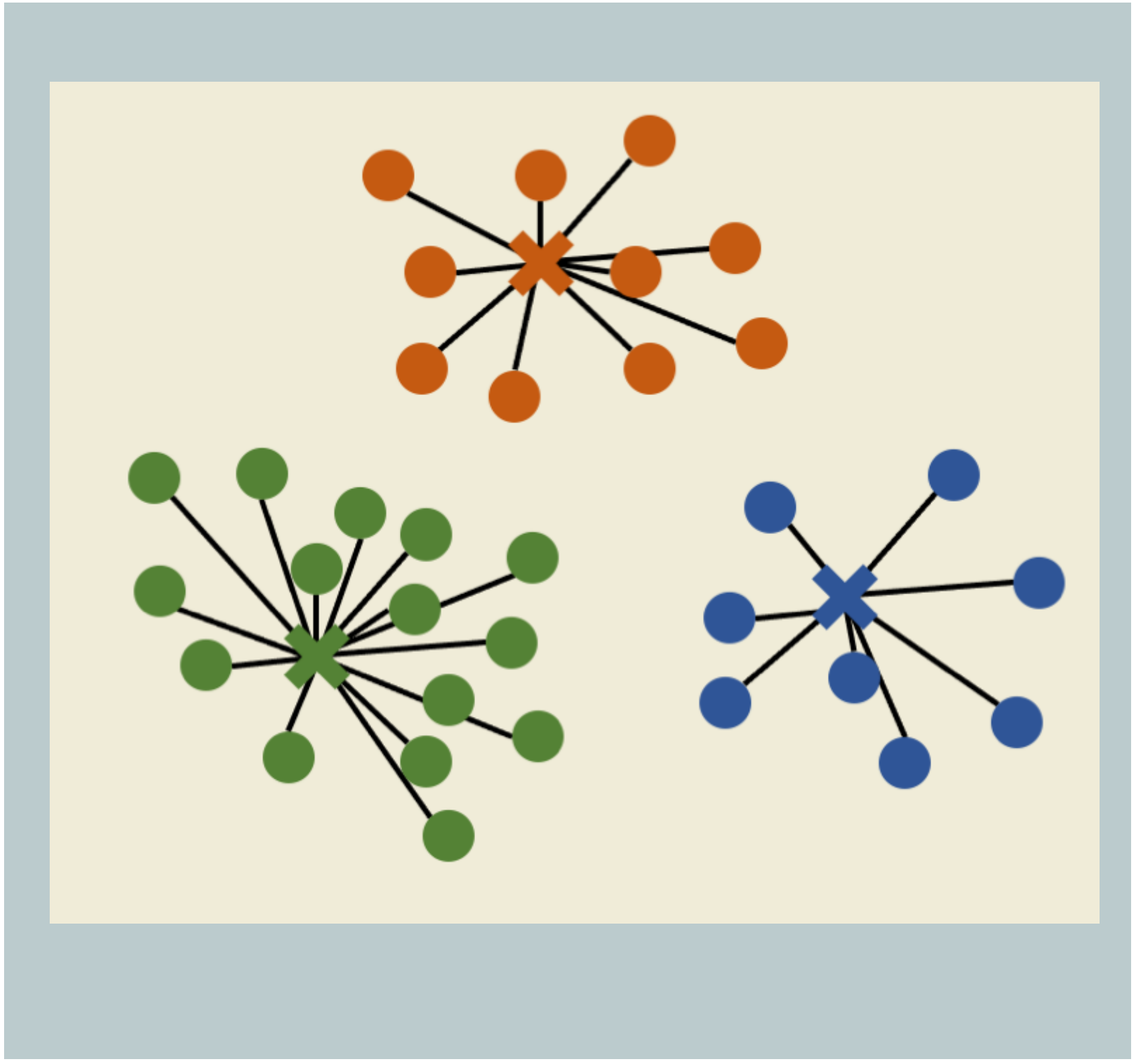
CLUSTERING

Clustering

It’s an **unsupervised** learning that split the data into groups that have a high intra-class similarity and low inter-class similarity after removing the class label “stroke”.

Clustering Methods Used:

Partitioning approach: **K-means**



K-means

- Choose a random centers
- Assign each object to the cluster with the nearest center.
- Iteratively, update the cluster centroids until reach the appropriate center for each cluster

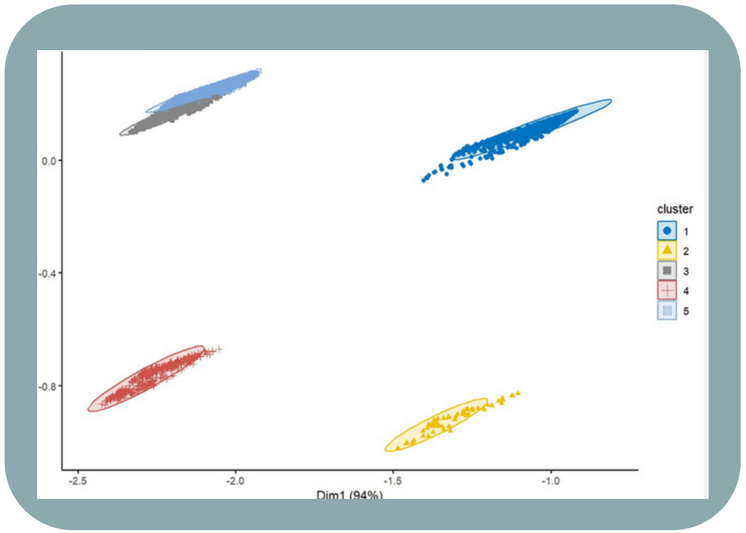
Why K-means?

It’s suitable for large datasets and simpler than the other algorithms.

RESULTS

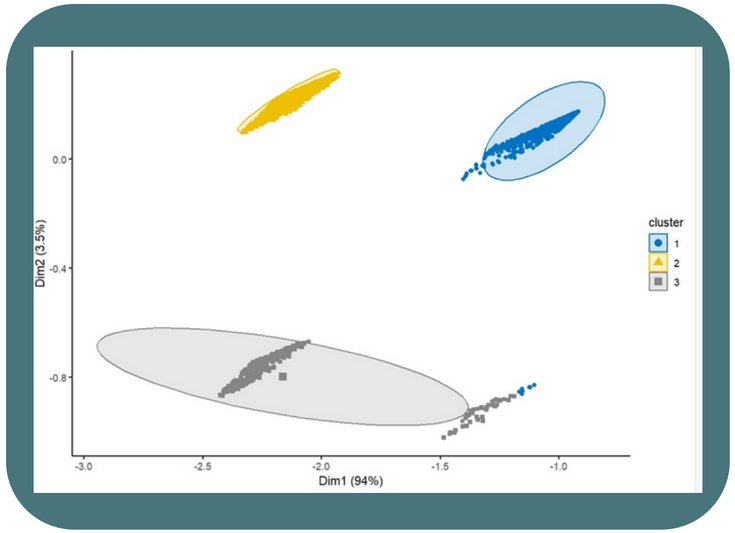
Cluatering results after selecting the top four attributes (age, hypertension, ever_married, avg_glucose_level) and Average silhouette width for each cluater.

K=5



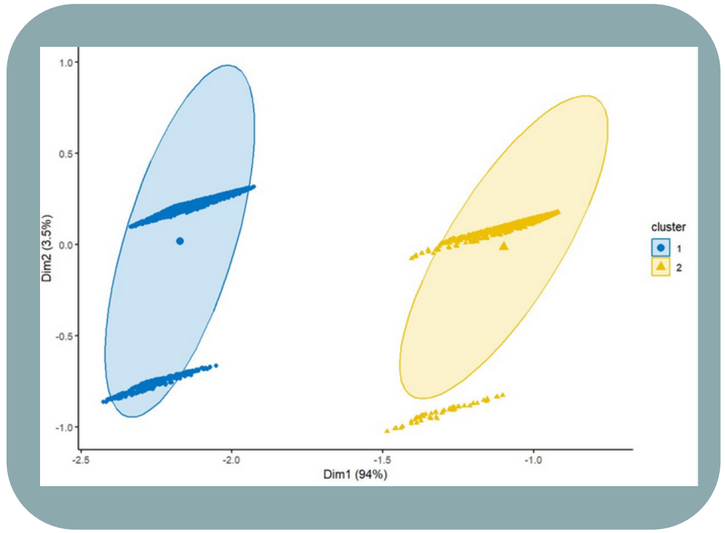
0.59

K=3



0.62

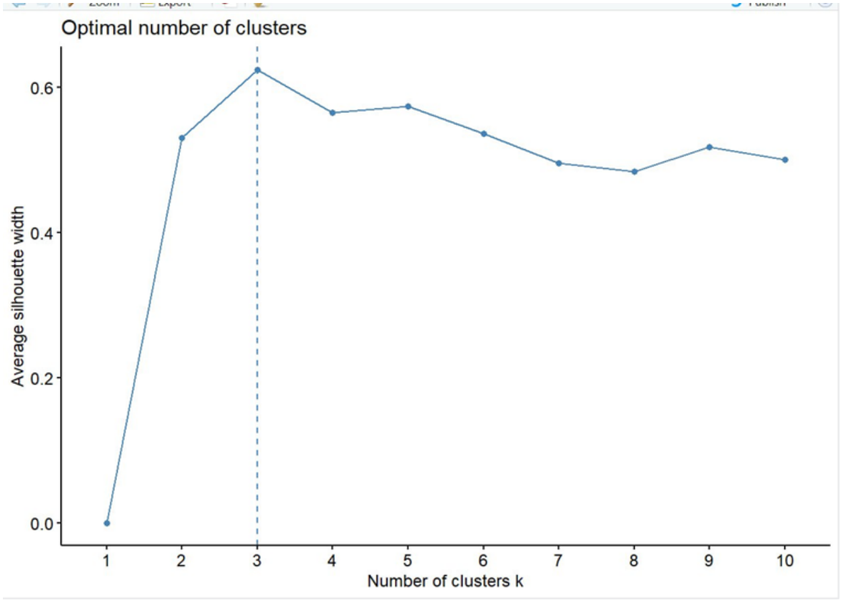
K=2



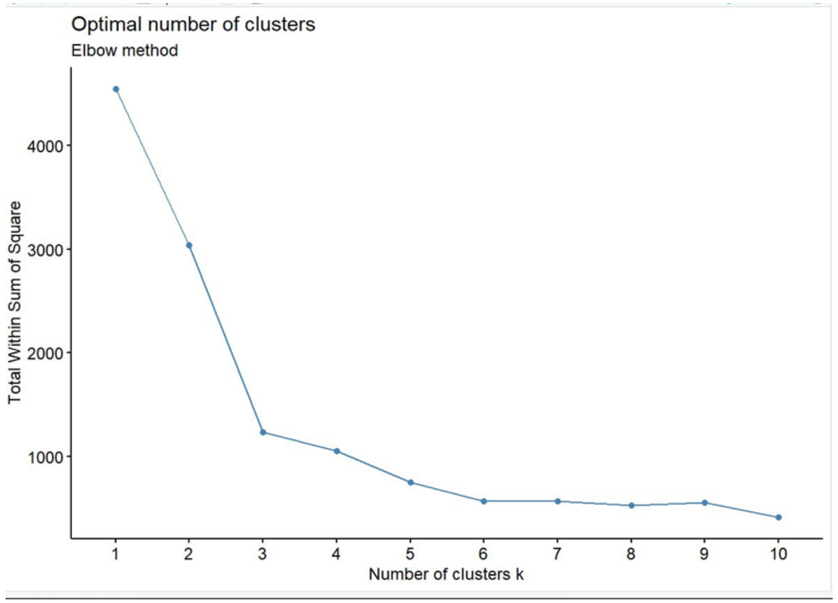
0.53

HOW TO KNOW OPTIMAL NUMBER OF CLUSTERS?

Average silhouette:



Elbow Method:



Average silhouette: **0.63**
Recall: **0.48174**
Precision: **0.56934**
Total within-cluster sum of square: **1232.939**

05.

FINDINGS

SUMMERY OF CLASSIFICATION OUTCOMES

	70 % training set 30% testing set:	80 % training set 20% testing set:	85 % training set 15% testing set:
Accuracy	0.7967	0.795	0.78
Precision	0.7432	0.7323	0.7442
Sensitivity	0.9067	0.9300	0.8533
Specificity	0.6867	0.6600	0.7067

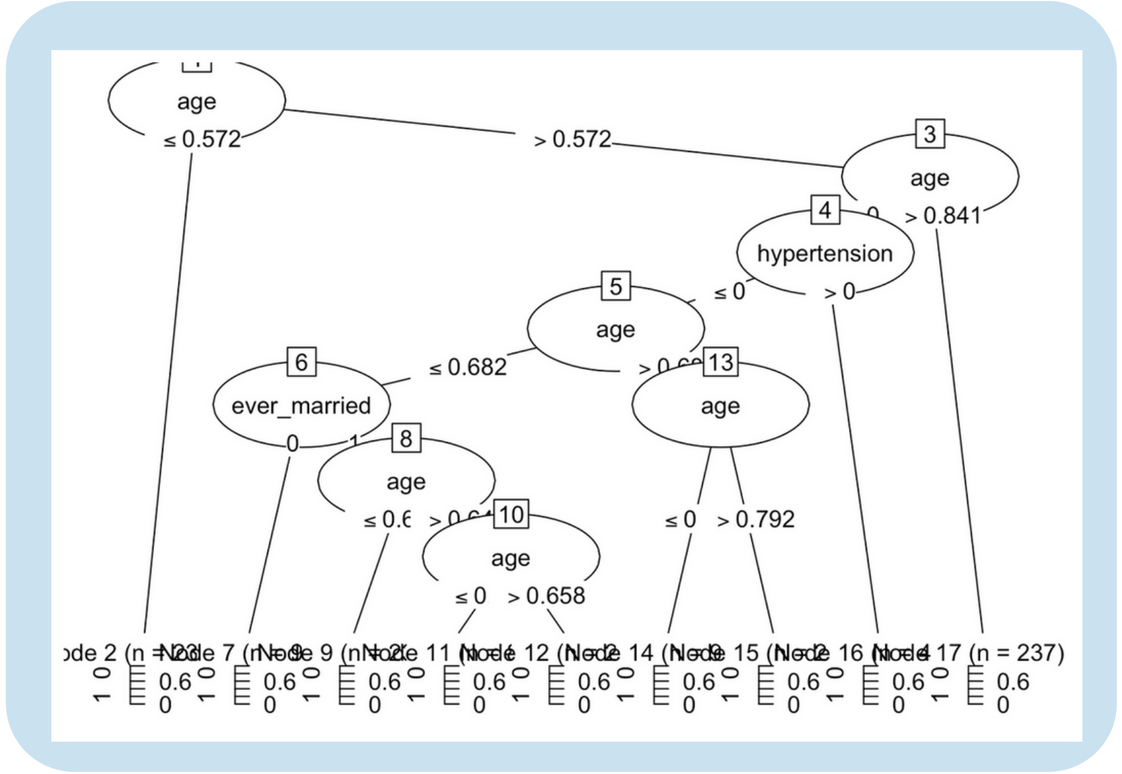
####GINI INDEX:

	70 % training set 30% testing set:	80 % training set 20% testing set:	85 % training set 15% testing set:
Accuracy	0.74	0.78	0.74
Precision	0.6875	0.7414	0.7045
Sensitivity	0.8800	0.8600	0.8267
Specificity	0.6000	0.7000	0.6533

####INFORMATION GAIN:

	70 % training set 30% testing set:	80 % training set 20% testing set:	85 % training set 15% testing set:
Accuracy	0.7433	0.76	0.7067
Precision	0.6780	0.6970	0.6505
Sensitivity	0.9267	0.9200	0.8933
Specificity	0.5600	0.6000	0.5200

SELECTED DECISION TREE



GAIN RATIO METHOD WITH (70% TRAINING SET AND 30% TESTING SET) YIELDED THE MOST ACCURATE AND RELIABLE RESULTS FOR OUR DATASET.

ANALYSIS OF DECISION TREE

- **Number of Leaf Nodes:** The decision tree consists of 9 leaf nodes, indicating that 9 rules or distinct conditions have been extracted from the tree.

- **Important Attributes**: the attributes of age, ever married, and hypertension

play a crucial role in making decisions and classifying instances within the tree structure.

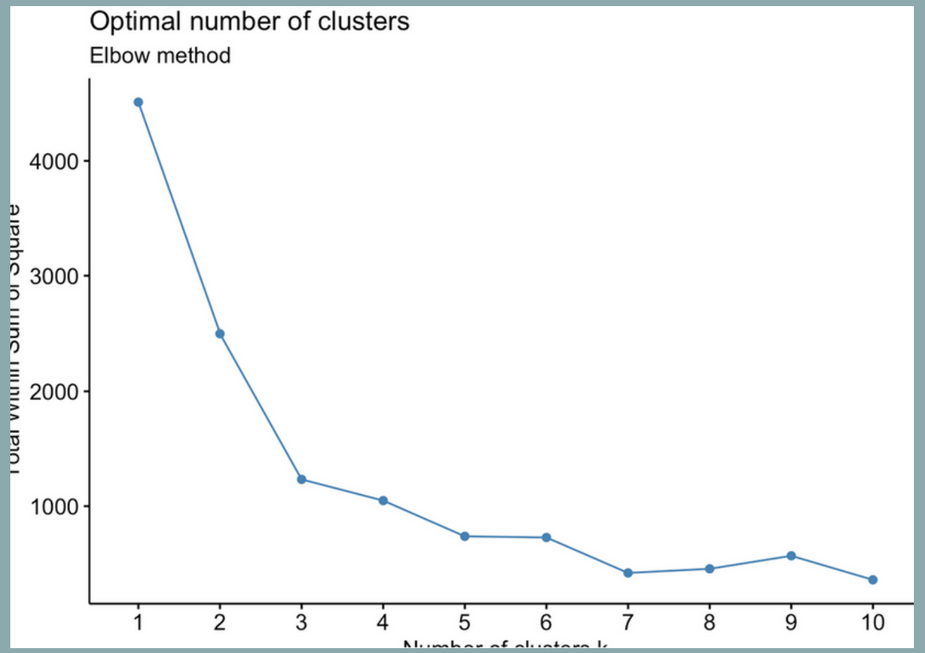
- **Age and Stroke**:Analysis indicates a correlation between older age and a higher likelihood of experiencing a stroke.

- **Hypertension and Stroke:**

The tree reveals that individuals with elevated levels of hypertension are more susceptible to experiencing a stroke.

CLUSTER OUTCOMES

	K-means		
Number of clusters	**K=2**	**K=3**	**K=5**
Average silhouette width for each cluster	0.53	0.63	0.59
total within-cluster sum of square	2500.985	1232.939	670.6099
BCubed precision	0.5404063	0.5693427	0.5866183
BCubed recall	0.669987	0.4817475	0.3322907



BEST NUMBER OF CLUSTER:

BINARY CLASS LABELS:

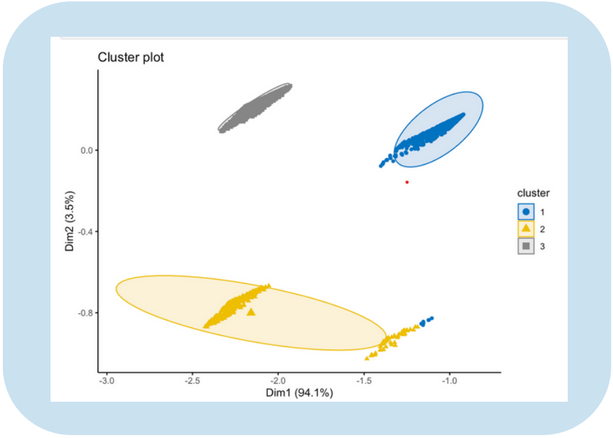
SUGGEST **TWO DISTINCT CLUSTERS** FOR STROKE PREDICTION;SINCE IT INVOLVES BINARY CLASS LABELS (0 FOR NO STROKE, 1 FOR STROKE)

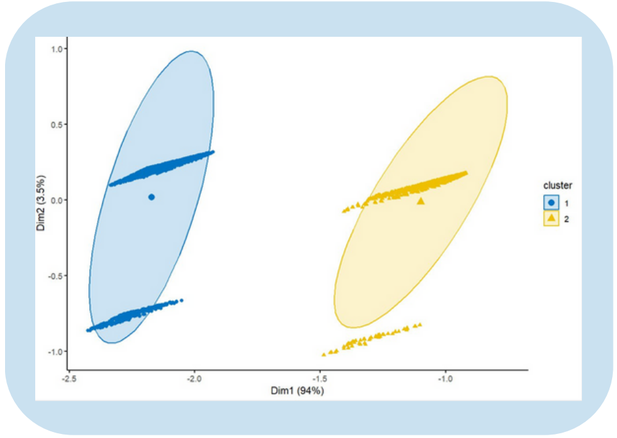
VS
EVALUATION METRICS :

LIKE SILHOUETTE ANALYSIS, BCUBED&RECALL PRECISION MAY FAVOR THE **3-MEAN DISTINCT CLUSTERS** DUE TO MANY REASON SUCH AS COMPLEX DATA OR VARIABLES NOT CONSIDERED

WHAT IS THE OPTIMAL DECISION?

DESPITE POTENTIAL EXTERNAL FACTORS, CHOOSING THE 2-MEAN CLUSTERING IS MORE OPTIMAL FOR OUR DATA SET.





CLASSIFICATION VS CLUSTERING

IN OUR DATA SET, CLASSIFICATION IS RECOMMENDED OVER CLUSTERING.

WHILE CLUSTERING REVEALS INSIGHTS INTO DATA PATTERNS AND GROUPINGS, IT CANNOT PREDICT INDIVIDUAL STROKE OCCURRENCES—**OUR PRIMARY GOAL.** CLASSIFICATION IS MORE SUITABLE FOR THIS SPECIFIC PREDICTION TASK OF DETERMINING WHETHER AN INDIVIDUAL WILL EXPERIENCE A STROKE OR NOT.

THANK YOU

- Bashair Alsadhan
- Maryam Altuwaijri
- Rama Alshebel
- Yasmen Alsuhaibani