FACULTY OF COMPUTING AND INFORMATICS

SEMESTER 1, SESSION 2023/2024

KK04703 Data Mining

GROUP PROJECT

Assignment 2

PREPARED FOR,

ASSOC. PROF. TS. DR. MOHD HANAFI BIN AHMAD HIJAZI

PREPARED BY GROUP 2:

No.	NAME	MATRIC NUMBER	EACH MEMBER CONTRIBUTION
1	EMERLYN TRISHA MERING AK JAMES	BI20160322	Partitional Clustering
2	JULIANA EKOT	BI20110242	Bayesian Classifier (Naïve Bayes Algorithm)
3	KHOO HUANG KWANG	BI20110253	Decision Tree

TABLE OF CONTENT

TITLE **PAGE**

LIST OF FIGURES

LIST OF TABLES

1.0 Introduction

The American Stroke Association (ASA) defines a stroke as a blockage in the arteries, which are the blood vessels that supply the brain with nutrition and oxygen. The disruption can be caused by a blood clot that obstructs the brain's blood supply (ischemic stroke), a blood vessel rupture (hemorrhagic stroke), or a transient clot (transient ischemic attack). Because stroke is associated with the blood arteries that supply the brain, it is also classified as a cerebrovascular illness.In Malaysia, stroke is the 3rd leading causes of death as it reached 15,642 (11.31%) of the total death in 2017 but many have survived from it (World Health Ranking et al. 2019). Unfortunately, those who have survived tend to have certain impairments in their senses, hearing, speech, vision, mobility, and IQ. Because some parts of the brain are deprived of oxygenated blood, rain cannot regulate specific bodily functions. (Nurul Fatin Rakib et al. 2020)

The difficulty in recognizing the symptoms of stroke has been identified as one big problem encountered when identifying an individual who should be undergoing treatment for his or her condition. The symptoms of stroke may be difficult to notice because they are subtle or manifest differently, complicating the interpretation and diagnosis process. At times strokes may occur with no well defined symptoms. The speed and accuracy of detection are critical for prompt medical attention, as early intervention significantly increases the likelihood that things will work out well in this respect. It is important to enhance the diagnostics tools and make the public more aware of improving diagnostic ability relating to faster detection that helps minimise stroke occurrence.

Prediction of stroke is critical due to the need for immediate treatment in order to prevent irreversible damage or death. Due to the technological development in medicine, machine learning techniques can now predict a possible onset of stroke. This challenge can be addressed by developing an effective stroke prediction model based on data mining and machine learning techniques. The stroke classification process makes use of the data mining technique Partitional Clustering, Decision Tree, and Bayesian Classifier.

Stroke prediction is relevant because it needs immediate treatment to prevent tissue damage, disability or even death. Technological advancements in the medical industry now allow machine learning techniques to forecast when a stroke might take place. This challenge can be addressed by developing an effective stroke prediction model based on data mining and machine learning techniques. The data mining technique used to classify types of strokes is the Naïve Bayes algorithm. For the purpose of this task, a stroke prediction dataset from Kaggle (Fedesoriano, 2021) is used. In this dataset, 12 columns are present with a total of 5110 rows. This algorithm may divide patients into two groups: As stroke and not Stroke, based on their characteristics and symptoms.

2.0 Data Preparation

The Kaggle Stroke Prediction Dataset was used in this project. The patient's physical characteristics and stroke status are included in this dataset. This dataset contains information on 5110 participants with 12 attributes. It analyses the relationship between several factors, including gender, age, type of disease, and smoking status, and the individual's risk of having a stroke. The dataset's attributes are displayed in table 2.1, and the first 32 rows of the dataset are displayed in Figure 2.1.

Table 2.1: Dataset attribute

Attribute	Description
id	A unique identifier for the individual
Gender	“Male”, “Female” or “Other”
age	Age of individual
hypertension	0 for the individual don’t have hypertension, while 1 is for the individual has hypertension

heart_disease	0 for the individual don't have heart disease, while 1 is for the individual has heart disease
ever_married	"No" or "Yes"
work_type	"Private", "Self-employed", "children", "Govt_job", or "Never_worked"
Residence_type	"Rural" or "Urban"
avg_glucose_level	Average glucose level in blood of individual
bmi	Body mass index
smoking_status	"formerly smoked", "never smoked", "smokes", or "Unknown"
stroke	1 is has stroke, 0 is not have stroke

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	gender	age	hypertensi	heart_dise	ever_marr	work_type	Residence_	avg_gluco:	bmi	smoking_s	stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly sı	1
3	51676	Female	61	0	0	Yes	Self-emplc	Rural	202.21	N/A	never smo	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smo	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-emplc	Rural	174.12	24	never smo	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly sı	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smo	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smo	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smo	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
16	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smo	1
17	58202	Female	50	1	0	Yes	Self-emplc	Rural	167.41	30.9	never smo	1
18	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
19	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
20	27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smo	1
21	25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
22	70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
23	13861	Female	52	1	0	Yes	Self-emplc	Urban	233.29	48.9	never smo	1
24	68794	Female	79	0	0	Yes	Self-emplc	Urban	228.7	26.6	never smo	1
25	64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
26	4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly sı	1
27	70822	Male	80	0	0	Yes	Self-emplc	Rural	104.12	23.5	never smo	1
28	38047	Female	65	0	0	Yes	Private	Rural	100.98	28.2	formerly sı	1
29	61843	Male	58	0	0	Yes	Private	Rural	189.84	N/A	Unknown	1
30	54827	Male	69	0	1	Yes	Self-emplc	Urban	195.23	28.3	smokes	1
31	69160	Male	59	0	0	Yes	Private	Rural	211.78	N/A	formerly sı	1
32	43717	Male	57	1	0	Yes	Private	Urban	212.08	44.2	smokes	1

Figure 2.1: First 32 rows of the dataset.

For this project, the dataset is obtained from Kaggle and stored as predictStroke.csv. The R studio will first read the dataset. After importing the dataset into the R studio, each attribute's data frame is displayed in Figure 2.2.

```
> stroke <- read.csv("predictStroke.csv")
> str(stroke)
'data.frame':	5110 obs. of  12 variables:
 $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
 $ gender           : chr  "Male" "Female" "Male" "Female" ...
 $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
 $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
 $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
 $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type        : chr  "Private" "Self-employed" "Private" "Private" ...
 $ Residence_type   : chr  "Urban" "Rural" "Rural" "Urban" ...
 $ avg_glucose_level: num  229 202 106 171 174 ...
 $ bmi              : chr  "36.6" "N/A" "32.5" "34.4" ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "never smoked" "smokes" ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

Figure 2.2: Import dataset to R studio

2.1 Data Cleaning

For the attribute "bmi," there are 201 null values. Since the dataset contains null values, data cleaning was used to ensure that each attribute is free of null values and NA values. To eliminate the rows that have null values, the easiest method is to delete them. Furthermore, gender needs to be classified as a binary variable. Thus, the data that has "Other" as the gender has likewise been deleted. Figure 2.3 illustrates the findings after testing the null values remaining after data cleansing.

```
> stroke[stroke== "N/A"]<- NA
> stroke[stroke == "Other"] <- NA
> colSums(is.na(stroke))
               id            gender               age      hypertension     heart_disease
                0                 1                 0                 0                 0
     ever_married         work_type    Residence_type avg_glucose_level               bmi
                0                 0                 0                 0               201
   smoking_status            stroke
                0                 0
> stroke <- na.omit(stroke)
> colSums(is.na(stroke))
               id            gender               age      hypertension     heart_disease
                0                 0                 0                 0                 0
     ever_married         work_type    Residence_type avg_glucose_level               bmi
                0                 0                 0                 0                 0
   smoking_status            stroke
                0                 0
```

Figure 2.3: Null values remaining after data cleansing

After data cleaning has done, the number of subject reduced from 5110 to 4908 observation with 12 variable, as shown in figure 2.4

```
> str(stroke)
'data.frame':	4908 obs. of  12 variables:
 $ id               : int  9046 31112 60182 1665 56669 53882 10434 60491 12109 12095 ...
 $ gender           : chr  "Male" "Male" "Female" "Female" ...
 $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension     : int  0 0 0 1 0 1 0 0 1 0 ...
 $ heart_disease    : int  1 1 0 0 0 1 0 0 0 1 ...
 $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ work_type        : chr  "Private" "Private" "Private" "Self-employed" ...
 $ Residence_type   : chr  "Urban" "Rural" "Urban" "Rural" ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi              : chr  "36.6" "32.5" "34.4" "24" ...
 $ smoking_status   : chr  "formerly smoked" "never smoked" "smokes" "never smoked" ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:202] 2 9 14 20 28 30 44 47 51 52 ...
  ..- attr(*, "names")= chr [1:202] "2" "9" "14" "20" ...
```

Figure 2.4: Dataframe after data cleaning

The class characteristics are then transformed into factors for the category variables. The class designations "0, 1" for heart disease, stroke, and hypertension have been changed to "No, Yes." We can see that there are more females in the dataset than men based on the structure and summary shown in Figure 2.5. For various age groups, the age attribute distribution column is normal. In addition, there is bias in the predictor class stroke data points.

```
> stroke$stroke<- factor(stroke$stroke, levels = c(0,1), labels = c("No", "Yes"))
> stroke$gender<- as.factor(stroke$gender)
> stroke$hypertension<- factor(stroke$hypertension, levels = c(0,1), labels = c("No", "Yes"))
> stroke$heart_disease<- factor(stroke$heart_disease, levels = c(0,1), labels = c("No", "Yes"))
> stroke$ever_married<- as.factor(stroke$ever_married)
> stroke$work_type<- as.factor(stroke$work_type)
> stroke$Residence_type<- as.factor(stroke$Residence_type)
> stroke$smoking_status<- as.factor(stroke$smoking_status)
> stroke$bmi<- as.numeric(stroke$bmi)
> summary(stroke)
       id           gender          age        hypertension heart_disease ever_married
 Min.   :   77   Female:2897   Min.   : 0.08   No :4457     No :4665      No :1704    
 1st Qu.:18603   Male  :2011   1st Qu.:25.00   Yes: 451     Yes: 243      Yes:3204    
 Median :37581                 Median :44.00                                          
 Mean   :37060                 Mean   :42.87                                          
 3rd Qu.:55182                 3rd Qu.:60.00                                          
 Max.   :72940                 Max.   :82.00                                          
         work_type    Residence_type avg_glucose_level      bmi                smoking_status stroke    
 children     : 671   Rural:2418     Min.   : 55.12    Min.   :10.30   formerly smoked: 836   No :4699  
 Govt_job     : 630   Urban:2490     1st Qu.: 77.07    1st Qu.:23.50   never smoked   :1852   Yes: 209  
 Never_worked :  22                  Median : 91.68    Median :28.10   smokes         : 737             
 Private      :2810                  Mean   :105.30    Mean   :28.89   Unknown        :1483             
 Self-employed: 775                  3rd Qu.:113.50    3rd Qu.:33.10                                    
                                     Max.   :271.74    Max.   :97.60                                    
```

Figure 2.5: Summary after converting 1, 0 to yes,no.

2.2 Data Exploration

The dataset will be divided into training and test sets upon the completion of the data cleaning procedure. 30% of the dataset will be utilised as the test set, while the remaining 70% will be used as the training set. The package "caTools" is loaded in order to do data partitioning. The training and test data may be created again using the same random number set.seed (123) method. Figure 2.6 displays the ratios of response variables in the test and training sets.

```
> install.packages("caTools")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the
appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/rayson/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/caTools_1.18.2.zip'
Content type 'application/zip' length 246159 bytes (240 KB)
downloaded 240 KB

package 'caTools' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\rayson\AppData\Local\Temp\RtmpAtMCfz\downloaded_packages
> library(caTools)
Warning message:
package 'caTools' was built under R version 4.2.3
> set.seed(123)
> split <- sample.split(stroke, SplitRatio = 0.7)
> train_stroke <- subset(stroke, split == "TRUE")
> test_stroke <- subset(stroke, split == "FALSE")
> prop.table(table(train_stroke$stroke))

        No        Yes 
0.95751834 0.04248166 
> prop.table(table(test_stroke$stroke))

        No        Yes 
0.95721271 0.04278729 
> 
```

Figure 2.6: Data Training and Data Test

3.0 Data Mining

We may define Data mining process as- discovering impressions in enormous amounts of unseen data that is available in the record (database). A range of approaches to retrieve information from record (database). The process of extracting useful information from a bigger collection of unprocessed data is known as data mining. It involves utilising various applications to analyse massive volumes of data for data patterns. (Shukla, R.K. et al. 2020).By using data mining techniques, we may better comprehend the data and identify opportunities to solve problems and support others in making critical decisions. This project aims to create a predictive model for stroke based on age, gender, hypertension, employment type, housing type, average glucose level, BMI, smoking status, and history of stroke. The data mining approach that will be used is Partitional Clustering, Decision Tree, and Bayesian classifier. Before going deeply on the data mining approach, let's look at some visualisations and extract data from the dataset. The ggplot2 package in R was used to create and plot all the graphs, as shown in Figure 3.1.

```
> install.packages("ggplot2")
WARNING: Rtools is required to build R packages but is not currently installed. Please dow
nload and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Warning in install.packages :
  package ‘ggplot2’ is in use and will not be installed
> library(ggplot2)
> gg<-ggplot(stroke, aes(x = gender, fill = stroke))+geom_bar(position = "fill")+stat_coun
t(geom = "text",aes(label= after_stat(count)),position = position_fill(vjust = 0.5), color
="black")
> gg
>
```

Figure 3.1: The Coding for Plotting Graph using ggplot2 Package

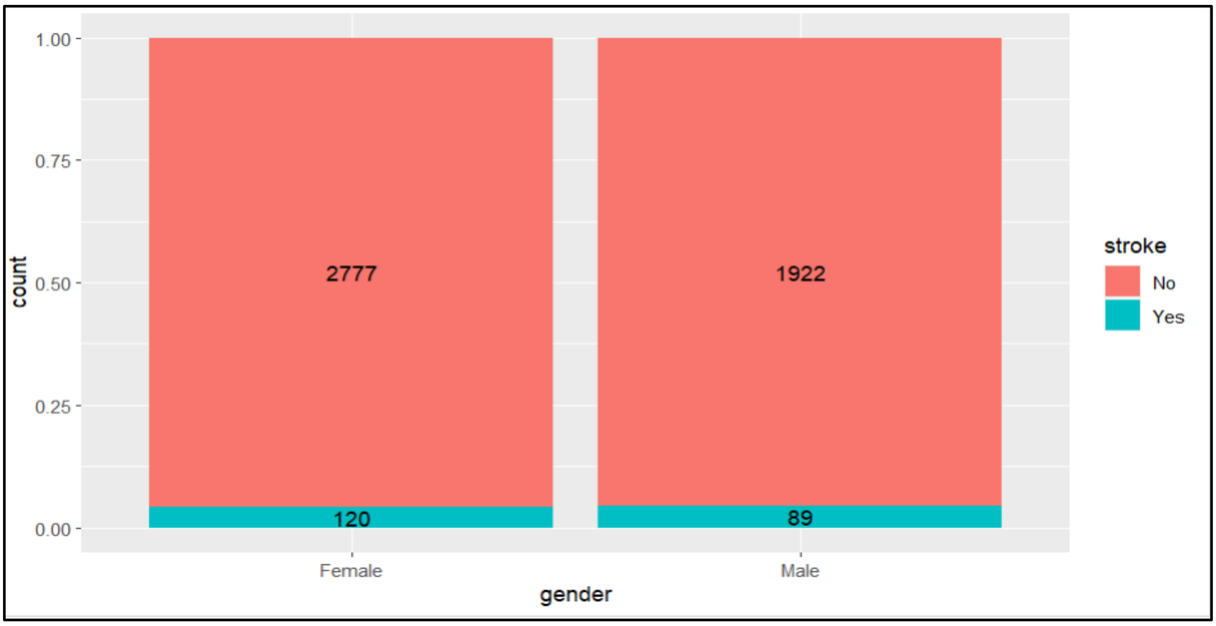

Figure 3.2: Stroke Diagnosis and Gender

The gender and stroke diagnosis of the patients are given in Figure 3.3. There is an imbalance in the number of male and female patients when there are more female patients than male patients. The observation indicates that a somewhat higher proportion of female stroke patients than male patients exists: 120 patients will have a stroke, compared to just 89 instances for males. At the same time, the data show that a higher percentage of women than men do not have strokes, with 2777 female patients not having a stroke.Male patients in 1922 do not get strokes. These results might be the result of a gender gap in the patient group.

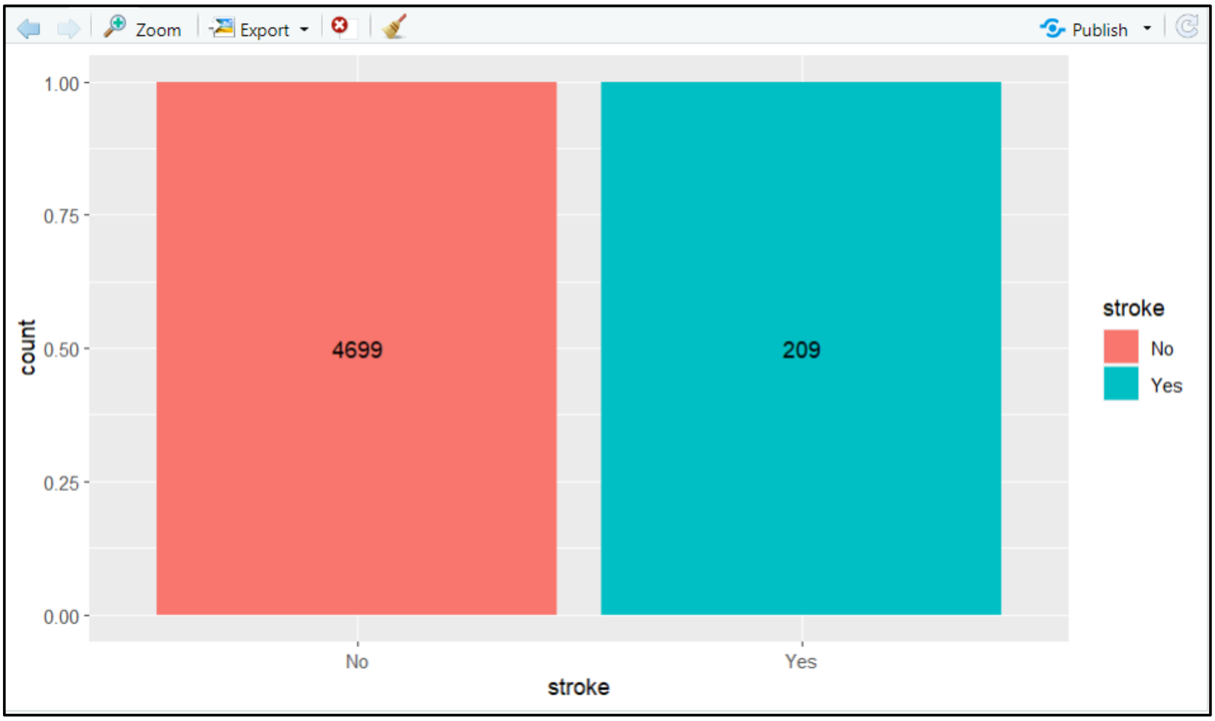

Figure 3.3: Stroke Vs Without Stroke

The number of patients with and without stroke is depicted in Figure 3.2. The dataset exhibits a clear class imbalance, with a much greater number of patients (4699) without a stroke than those (209) with a stroke.

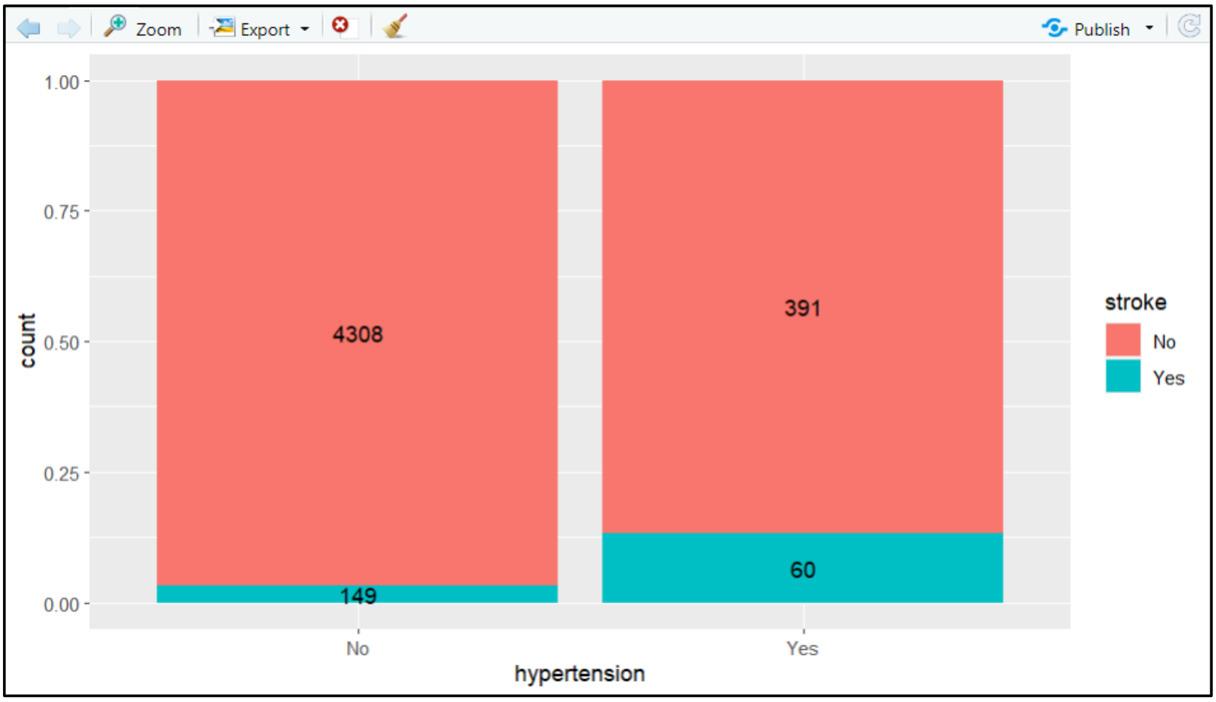

Figure 3.4: Hypertension Status and Stroke

The patient's hypertension status and stroke information are displayed in Figure 3.4. The proportion of patients without hypertension is greater than that of patients with hypertension (451 individuals without hypertension against 4457 patients with hypertension). The finding indicates that a greater proportion of individuals without hypertension—149 of those without hypertension and 60 of those with hypertension—are experiencing strokes than those with hypertension. Therefore, it may be said that most stroke patients do not have a history of hypertension.

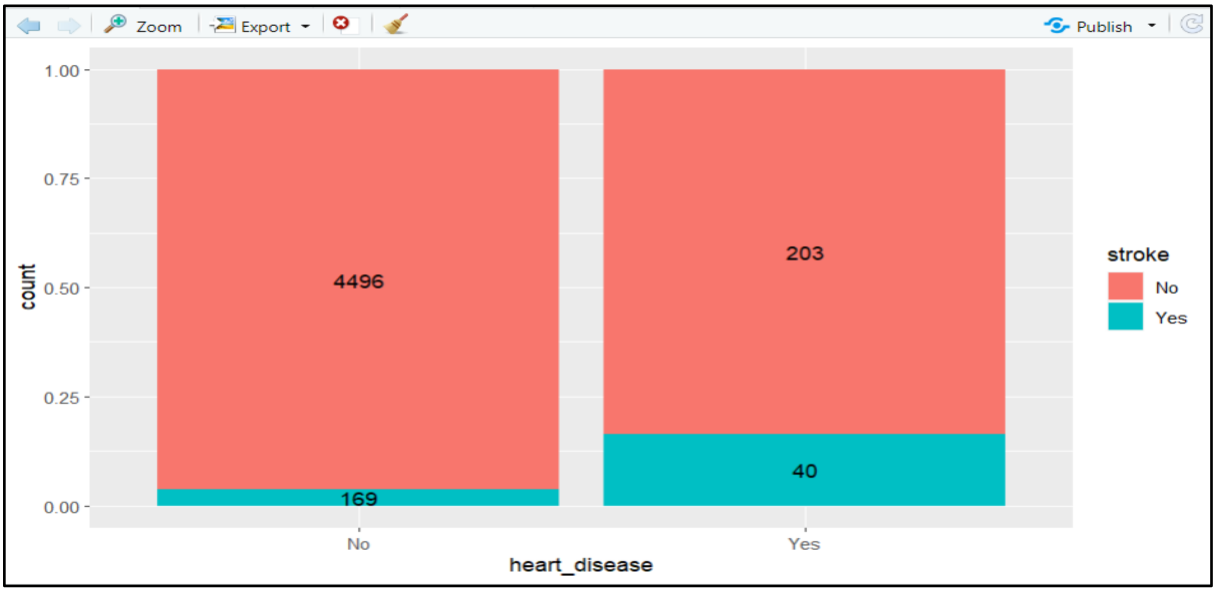

Figure 3.5: Heart Disease and Stroke

The data on heart disease and stroke is displayed in Figure 3.5. The number of patients who do not have heart disease is higher compared to the patients with heart disease which is 4665 for the patient with no heart disease and 243 for the patient with heart disease. Furthermore, there are fewer individuals with heart disease who experience strokes than there are patients without heart disease (40 patients with heart disease and 169 patients without heart disease). It may thus be concluded that heart disease is not a factor in the majority of stroke victims.

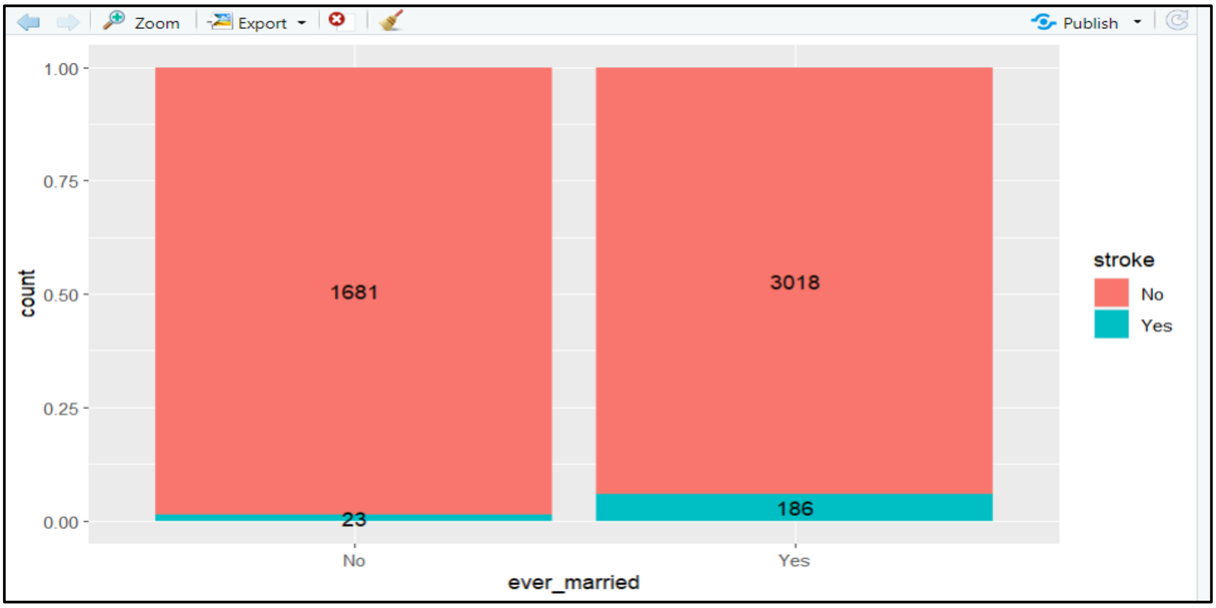

Figure 3.6: Ever Married and Stroke

The patient's marital status is displayed in Figure 3.6 in opposition to the stroke diagnosis. Compared to single patients, married individuals get strokes at a significantly greater rate. Just 23 people do not have a spouse when they suffer a stroke, compared to 186 married patients. Therefore, it may be concluded that patients who are married have a higher risk of stroke than patients who are single.

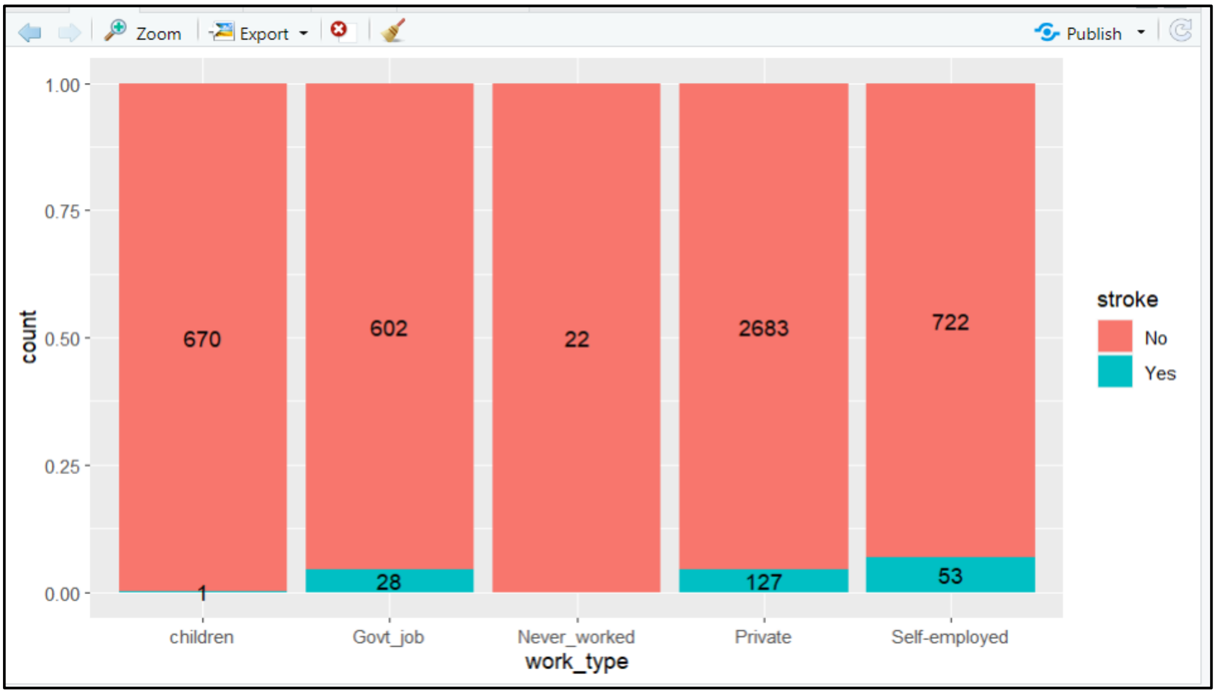

Figure 3.7: Work Type and Stroke

One of the things that might lead to a stroke in a patient is their sort of work. Figure 3.7 displays the statistics related to the type of job and stroke. The most of the patients in this dataset work in the private sector which is 127 patients, followed by self-employed which is 53 patients. For patients whose work type is govt_job have 28 get the stroke and for patients work type is children and never_worked which is 1 and 0 respectively.

Figure 3.8: Residence Type and Stroke

The types of residence and stroke data are displayed in Figure 3.8. In this dataset, there are two distinct dwelling types: rural and urban. According to the bar chart, there is just a small difference between the two types of residences when it comes to the risk of having a stroke or not—that is, 100 for patients residing in rural areas and 109 for patients residing in urban areas. Thus, it can be said that there is no conclusive link between the kind of habitation and having a stroke.

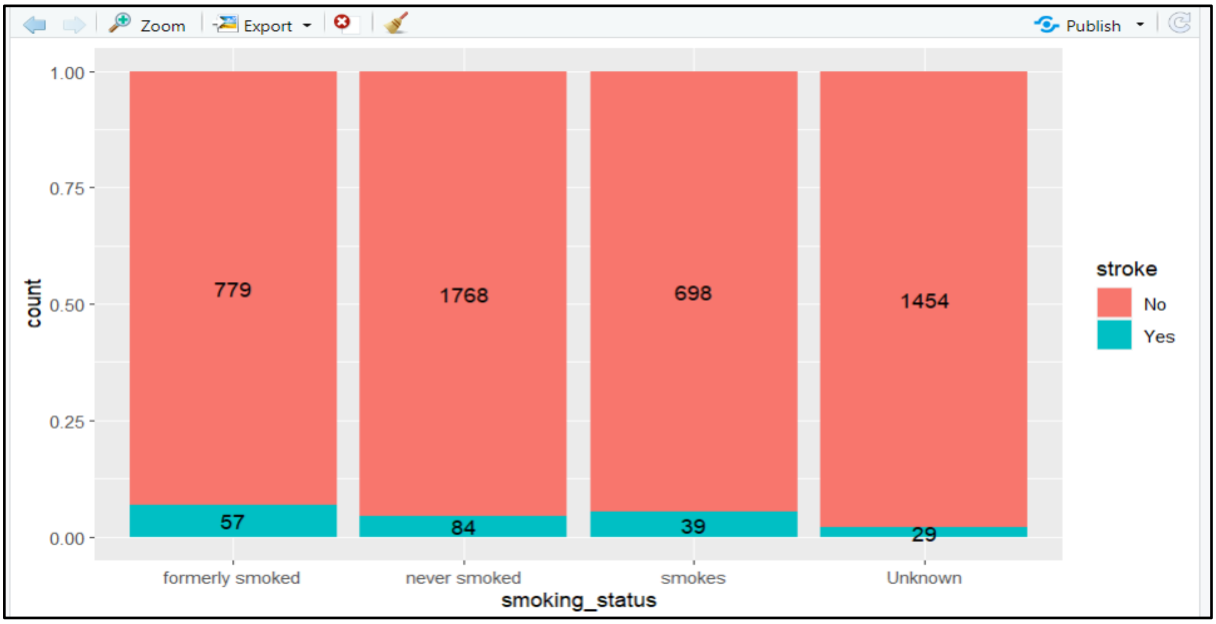

Figure 3.9: Smoking Status and Stroke

Every patient's smoking status is displayed in Figure 3.9. There are four options for indicating one's smoking status: "never smoke," "smokes," "formerly smoked," and "unknown for the smoking attribute." It was noticed that the number of patients who do not smoke without suffering a stroke is the largest compared to the patients who have other smoking statuses which is 1768 patients who have never smoked and without stroke. The number of patients who do not smoke, however, is the largest at 84, although it is smaller than the number of patients who have a smoking history (smokers and former smokers, at 39 and 57, respectively). This is because, the total number of patients for formerly smoked (836) and smokes (737) are lower compared to the never smoked (1852). The percentage of smokers, formerly smoked and never smoked, is 5.29%, 6.82% and 4.54%. Consequently, people who have smoked in the past are more likely to get strokes.

3.1 Decision Tree

After the Partitional Clustering is done, the Decision Tree will be used for data training. Decision trees are a popular machine learning technique for uncovering patterns from existing data. (Intan Rahmatillah, Eriana Astuty and Ivan Diryana Sudirman et al., 2023).

Firstly, the packages such as “C50” need to be installed and loaded.The C50 package in R offers an implementation of the C5.0. They operate by recursively dividing the dataset according to features value in order to make a decision.

After the process of data training, The decision tree contains one node and “No” is a majority class. The figures in parentheses stand for instances at that point corresponding to Figure 3.10.

```
> set.seed(9850)
> g<- runif(nrow(stroke))
> stroker<- stroke[order(g),]
> str(stroke)
'data.frame':	4908 obs. of  12 variables:
 $ id               : int  9046 31112 60182 1665 56669 53882 10434 60491 12109 12095 ...
 $ gender           : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 2 1 1 1 1 ...
 $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 1 ...
 $ heart_disease    : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 2 1 1 1 2 ...
 $ ever_married     : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1 2 2 2 ...
 $ work_type        : Factor w/ 5 levels "children","Govt_job",..: 4 4 4 5 4 4 4 4 4 2 ...
 $ Residence_type   : Factor w/ 2 levels "Rural","Urban": 2 1 2 1 2 1 2 2 1 1 ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi              : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
 $ smoking_status   : Factor w/ 4 levels "formerly smoked",..: 1 2 3 2 1 2 2 4 2 3 ...
 $ stroke           : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 - attr(*, "na.action")= 'omit' Named int [1:202] 2 9 14 20 28 30 44 47 51 52 ...
  ..- attr(*, "names")= chr [1:202] "2" "9" "14" "20" ...
> m1<-C5.0(stroker[1:4000,-12], stroker[1:4000,12])
> m1
```

```
Call:
C5.0.default(x = stroker[1:4000, -12], y
 = stroker[1:4000, 12])

Classification Tree
Number of samples: 4000
Number of predictors: 11

Tree size: 1

Non-standard options: attempt to group attributes

> summary(m1)

Call:
C5.0.default(x = stroker[1:4000, -12], y
 = stroker[1:4000, 12])


C5.0 [Release 2.07 GPL Edition]         Tue Jan 16 23:27:37 2024
-------------------------------

Class specified by attribute `outcome'

Read 4000 cases (12 attributes) from undefined.data

Decision tree:
 No (4000/170)


Evaluation on training data (4000 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	     1  170( 4.3%)   <<
```

```
	   (a)   (b)    <-classified as
	  ----  ----
	  3830          (a): class No
	   170          (b): class Yes


 Time: 0.0 secs
```

Figure 3.10: Decision Tree Training Model on Data

3.2 Bayesian Classifier

Globally regarded as a critical public health concern, stroke strongly contributes to morbidity and mortality rates. The desire for early detection motivates the investigation of advanced machine learning methods, with the Naive Bayes classifier emerging as a powerful tool for predicting stroke risk based on patient data. This study's major purpose is to construct a predictive model utilizing the Naive Bayes algorithm, exploiting a dataset rich in health-related variables. In order to reduce the overall incidence of strokes, the model has the ability to help healthcare providers identify patients at higher risk of stroke early on. This, in turn, might lead to specific therapies and lifestyle modifications.

The Naïve Bayes algorithm will be used to train the data after the data preparation process is over. Naïve Bayes is a simple classification method based on the Bayes Theorem and probabilities. Training sets of data make classification work well (Tempola et al., 2021).

Before anything else, tools like "e1071" and "caret" need to be loaded and installed. The "e1071" package's Naïve Bayes function enables the usage of both numeric and component variables across the Naive Bayes model. Besides that, the "caret" package is also full because it has features that make training models for classification problems faster.

Once the training data was finished, the model created the conditional probability for each trait on its own. It is possible to figure out the prior chances, which display the data distribution. A summary of the a priori probabilities findings can be seen in Figure 3.2.

```
> install.packages("e1071")
WARNING: Rtools is required to build R packages but is not currently ins
opriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/ASUS/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/e1071_1.7-1
Content type 'application/zip' length 664240 bytes (648 KB)
downloaded 648 KB

package ‘e1071’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ASUS\AppData\Local\Temp\Rtmp0KIZsP\downloaded_packages
> install.packages("caret")
WARNING: Rtools is required to build R packages but is not currently ins
opriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/ASUS/AppData/Local/R/win-library/4.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/caret_6.0-9
Content type 'application/zip' length 3579102 bytes (3.4 MB)
downloaded 3.4 MB

package ‘caret’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\ASUS\AppData\Local\Temp\Rtmp0KIZsP\downloaded_packages
> library(e1071)
```

```
> library(caret)
Loading required package: lattice
Warning messages:
1: package ‘caret’ was built under R version 4.2.3
2: package ‘lattice’ was built under R version 4.2.3
> set.seed(123)
> classifier_stroke <- naiveBayes(stroke ~ ., data = train_stroke)
> classifier_stroke

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        No        Yes
0.95751834 0.04248166

Conditional probabilities:
     id
Y         [,1]     [,2]
  No  37060.67 20808.43
  Yes 37383.86 22002.30

     gender
Y        Female      Male
  No  0.5933610 0.4066390
  Yes 0.5107914 0.4892086
```

```
     age
Y         [,1]     [,2]
  No  41.80158 22.34578
  Yes 67.45324 12.78089

     hypertension
Y             No        Yes
  No  0.91701245 0.08298755
  Yes 0.70503597 0.29496403

     heart_disease
Y             No        Yes
  No  0.95435685 0.04564315
  Yes 0.75539568 0.24460432

     ever_married
Y             No        Yes
  No  0.35812320 0.64187680
  Yes 0.09352518 0.90647482

     work_type
Y        children    Govt_job Never_worked     Private Self-employed
  No  0.143313118 0.120970316  0.004468560 0.574210022   0.157037983
  Yes 0.007194245 0.122302158  0.000000000 0.640287770   0.230215827

     Residence_type
Y         Rural     Urban
  No  0.4985637 0.5014363
  Yes 0.4964029 0.5035971
```

```
     avg_glucose_level
Y         [,1]     [,2]
  No  104.7862 43.69757
  Yes 138.2239 63.59848

     bmi
Y         [,1]     [,2]
  No  28.84156 7.982554
  Yes 30.29065 6.331494

     smoking_status
Y      formerly smoked never smoked    smokes   Unknown
  No         0.1678902    0.3705713 0.1512927 0.3102458
  Yes        0.2302158    0.4100719 0.2374101 0.1223022

> |
```

Figure 3.11: Bayesian Classifier Training Model on Data

3.3 Partitional Clustering

Partitional clustering is an essential method in unsupervised machine learning, aimed at organising a dataset into separate groups or clusters that do not overlap. This is done by identifying patterns or similarities among the data points. Sorting the enormous amount of data is crucial for effective analysis, reasoning, and decision-making. Due to its broad application, clustering has gained significant importance in various fields in recent years (Kutbay, 2018).

The k-means algorithm operates through a series of iterations to accomplish its objective. The process starts by selecting a set number of cluster centroids, which act as the central points for each cluster. Points are assigned to the cluster that has the closest centroid, typically determined using Euclidean distance calculations. The algorithm then proceeds to update the centroids by calculating the average of the data points within each cluster. This process of updating assignments continues iteratively until convergence, refining the cluster assignments and centroids.

Firstly, in order to use partitional clustering in R, it is necessary to install and load the relevant packages, such as "cluster." The "cluster" package offers a comprehensive range of partitional clustering algorithms, including the widely used k-means algorithm. These algorithms divide the dataset into separate clusters by comparing the similarity of data points.

The partitional clustering procedure can start as soon as the packages are loaded and installed. When using k-means clustering, the algorithm first chooses the initial cluster centres at random before allocating each data point to the closest cluster centre. In Figure 3.3, the results of partitional clustering illustrate the distinct partitioning of the dataset into clusters based on similarity, revealing inherent patterns and structures within the data.

```
> stroke <- read.csv("stroke.csv")
> stroke[stroke == "N/A" | stroke == "Other"] <- NA
> colSums(is.na(stroke))
               id            gender               age      hypertension
                0                 1                 0                 0
    heart_disease      ever_married         work_type    Residence_type
                0                 0                 0                 0
avg_glucose_level               bmi    smoking_status            stroke
                0               201                 0                 0
> stroke <- na.omit(stroke)
> colSums(is.na(stroke))
               id            gender               age      hypertension
                0                 0                 0                 0
    heart_disease      ever_married         work_type    Residence_type
                0                 0                 0                 0
avg_glucose_level               bmi    smoking_status            stroke
                0                 0                 0                 0
> 
```

```
> stroke$stroke <- factor(stroke$stroke, levels = c(0, 1), labels = c("No", "Yes"))
> stroke$gender <- as.factor(stroke$gender)
> stroke$hypertension <- factor(stroke$hypertension, levels = c(0, 1), labels = c("No", "Yes"))
> stroke$heart_disease <- factor(stroke$heart_disease, levels = c(0, 1), labels = c("No", "Yes"))
> stroke$ever_married <- as.factor(stroke$ever_married)
> stroke$work_type <- as.factor(stroke$work_type)
> stroke$Residence_type <- as.factor(stroke$Residence_type)
> stroke$smoking_status <- as.factor(stroke$smoking_status)
> stroke$bmi <- as.numeric(stroke$bmi)
> summary(stroke)
       id            gender          age        hypertension heart_disease
 Min.   :   77   Female:2897   Min.   : 0.08   No :4457     No :4665     
 1st Qu.:18603   Male  :2011   1st Qu.:25.00   Yes: 451     Yes: 243     
 Median :37581                 Median :44.00                             
 Mean   :37060                 Mean   :42.87                             
 3rd Qu.:55182                 3rd Qu.:60.00                             
 Max.   :72940                 Max.   :82.00                             
 ever_married         work_type    Residence_type avg_glucose_level
 No :1704     children     : 671   Rural:2418     Min.   : 55.12   
 Yes:3204     Govt_job     : 630   Urban:2490     1st Qu.: 77.07   
              Never_worked :  22                  Median : 91.68   
              Private      :2810                  Mean   :105.30   
              Self-employed: 775                  3rd Qu.:113.50   
                                                  Max.   :271.74   
      bmi                smoking_status stroke    
 Min.   :10.30   formerly smoked: 836   No :4699  
 1st Qu.:23.50   never smoked   :1852   Yes: 209  
 Median :28.10   smokes         : 737             
 Mean   :28.89   Unknown        :1483             
 3rd Qu.:33.10                                    
 Max.   :97.60                                    
```

```
> stroke
      id gender age hypertension heart_disease ever_married     work_type
1   9046   Male  67           No           Yes          Yes       Private
3  31112   Male  80           No           Yes          Yes       Private
4  60182 Female  49           No            No          Yes       Private
5   1665 Female  79          Yes            No          Yes Self-employed
6  56669   Male  81           No            No          Yes       Private
7  53882   Male  74          Yes           Yes          Yes       Private
8  10434 Female  69           No            No           No       Private
10 60491 Female  78           No            No          Yes       Private
11 12109 Female  81          Yes            No          Yes       Private
12 12095 Female  61           No           Yes          Yes      Govt_job
13 12175 Female  54           No            No          Yes       Private
15  5317 Female  79           No           Yes          Yes       Private
16 58202 Female  50          Yes            No          Yes Self-employed
17 56112   Male  64           No           Yes          Yes       Private
18 34120   Male  75          Yes            No          Yes       Private
19 27458 Female  60           No            No           No       Private
21 70630 Female  71           No            No          Yes      Govt_job
22 13861 Female  52          Yes            No          Yes Self-employed
23 68794 Female  79           No            No          Yes Self-employed
24 64778   Male  82           No           Yes          Yes       Private
25  4219   Male  71           No            No          Yes       Private
26 70822   Male  80           No            No          Yes Self-employed
27 38047 Female  65           No            No          Yes       Private
29 54827   Male  69           No           Yes          Yes Self-employed
31 43717   Male  57          Yes            No          Yes       Private
32 33879   Male  42           No            No          Yes       Private
33 39373 Female  82          Yes            No          Yes Self-employed
34 54401   Male  80           No           Yes          Yes Self-employed
35 14248   Male  48           No            No           No      Govt_job
36   712 Female  82          Yes           Yes           No       Private
37 47269   Male  74           No            No          Yes       Private
```

38	24977	Female	72	Yes	No	Yes	Private
39	47306	Male	58	No	No	No	Private
40	62602	Female	49	No	No	Yes	Private
41	4651	Male	78	No	No	Yes	Private
42	1261	Male	54	No	No	Yes	Private
43	61960	Male	82	No	Yes	Yes	Private
45	7937	Male	60	Yes	No	Yes	Govt_job
46	19824	Male	76	Yes	No	Yes	Private
48	47472	Female	58	No	No	Yes	Private
49	35626	Male	81	No	No	Yes	Self-employed
50	36338	Female	39	Yes	No	Yes	Private
53	59190	Female	79	No	Yes	Yes	Private
54	47167	Female	77	Yes	No	Yes	Self-employed
56	25831	Male	63	No	Yes	Yes	Private
57	38829	Female	82	No	No	Yes	Private
59	58631	Male	73	Yes	No	Yes	Self-employed
60	5111	Female	54	Yes	No	Yes	Govt_job
61	10710	Female	56	No	No	Yes	Private
62	55927	Female	80	Yes	No	Yes	Private
63	65842	Female	67	Yes	No	Yes	Self-employed
64	19557	Female	45	No	No	Yes	Private
66	17013	Male	78	Yes	No	No	Private
67	17004	Female	70	No	No	Yes	Private
68	72366	Male	76	No	No	Yes	Private
69	6118	Male	59	No	No	Yes	Private
70	7371	Female	80	Yes	No	Yes	Self-employed
72	2326	Female	67	Yes	No	Yes	Private
73	27169	Female	66	Yes	No	Yes	Govt_job
74	50784	Male	63	No	No	Yes	Private
75	19773	Female	52	No	No	Yes	Private
76	66159	Female	80	No	Yes	Yes	Self-employed
77	36236	Male	80	Yes	No	Yes	Private
78	71673	Female	79	No	No	Yes	Private

80	42117	Male	43	No	No	Yes	Self-employed
81	57419	Male	59	No	No	Yes	Private
83	26727	Female	79	No	No	No	Private
84	66638	Female	68	Yes	No	No	Self-employed
86	32399	Male	54	No	No	Yes	Private
87	3253	Male	61	No	Yes	Yes	Private
88	71796	Female	70	No	Yes	Yes	Private
89	14499	Male	47	No	No	Yes	Private
90	49130	Male	74	No	No	Yes	Private
91	28291	Female	79	No	Yes	Yes	Private
92	51169	Male	81	No	No	Yes	Private
93	66315	Female	57	No	No	No	Self-employed
94	37726	Female	80	Yes	No	Yes	Self-employed
95	54385	Male	45	No	No	Yes	Private
96	2458	Female	78	No	No	Yes	Private
97	35512	Female	70	No	No	Yes	Self-employed
98	56841	Male	58	No	Yes	Yes	Private
99	8154	Male	57	Yes	No	Yes	Govt_job
100	4639	Female	69	No	No	Yes	Govt_job

	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	Urban	228.69	36.6	formerly smoked	Yes
3	Rural	105.92	32.5	never smoked	Yes
4	Urban	171.23	34.4	smokes	Yes
5	Rural	174.12	24.0	never smoked	Yes
6	Urban	186.21	29.0	formerly smoked	Yes
7	Rural	70.09	27.4	never smoked	Yes
8	Urban	94.39	22.8	never smoked	Yes
10	Urban	58.57	24.2	Unknown	Yes
11	Rural	80.43	29.7	never smoked	Yes
12	Rural	120.46	36.8	smokes	Yes
13	Urban	104.51	27.3	smokes	Yes
15	Urban	214.09	28.2	never smoked	Yes
16	Rural	167.41	30.9	never smoked	Yes
17	Urban	191.61	37.5	smokes	Yes
18	Urban	221.29	25.8	smokes	Yes
19	Urban	89.22	37.8	never smoked	Yes
21	Rural	193.94	22.4	smokes	Yes
22	Urban	233.29	48.9	never smoked	Yes
23	Urban	228.70	26.6	never smoked	Yes
24	Rural	208.30	32.5	Unknown	Yes
25	Urban	102.87	27.2	formerly smoked	Yes
26	Rural	104.12	23.5	never smoked	Yes
27	Rural	100.98	28.2	formerly smoked	Yes
29	Urban	195.23	28.3	smokes	Yes
31	Urban	212.08	44.2	smokes	Yes
32	Rural	83.41	25.4	Unknown	Yes
33	Urban	196.92	22.2	never smoked	Yes
34	Urban	252.72	30.5	formerly smoked	Yes
35	Urban	84.20	29.7	never smoked	Yes
36	Rural	84.03	26.5	formerly smoked	Yes
37	Rural	219.72	33.7	formerly smoked	Yes
38	Rural	74.63	23.1	formerly smoked	Yes
39	Rural	92.62	32.0	Unknown	Yes
40	Urban	60.91	29.9	never smoked	Yes
41	Rural	78.03	23.9	formerly smoked	Yes
42	Urban	71.22	28.5	never smoked	Yes
43	Urban	144.90	26.4	smokes	Yes
45	Urban	213.03	20.2	smokes	Yes
46	Rural	243.58	33.6	never smoked	Yes
48	Urban	107.26	38.6	formerly smoked	Yes
49	Urban	99.33	33.7	never smoked	Yes
50	Rural	58.09	39.2	smokes	Yes
53	Rural	127.29	27.7	never smoked	Yes
54	Urban	124.13	31.4	never smoked	Yes
56	Rural	196.71	36.5	formerly smoked	Yes
57	Rural	59.32	33.2	never smoked	Yes
59	Urban	194.99	32.8	never smoked	Yes
60	Urban	180.93	27.7	never smoked	Yes
61	Urban	185.17	40.4	formerly smoked	Yes
62	Rural	74.90	22.2	never smoked	Yes
63	Rural	61.94	25.3	smokes	Yes
64	Rural	93.72	30.2	formerly smoked	Yes
66	Urban	113.01	24.0	never smoked	Yes
67	Urban	221.58	47.5	never smoked	Yes
68	Urban	104.47	20.3	Unknown	Yes
69	Urban	86.23	30.0	formerly smoked	Yes
70	Rural	72.67	28.9	never smoked	Yes
72	Rural	179.12	28.1	formerly smoked	Yes
73	Rural	116.55	31.1	formerly smoked	Yes
74	Rural	228.56	27.4	never smoked	Yes
75	Rural	96.59	26.4	never smoked	Yes
76	Rural	66.72	21.7	formerly smoked	Yes
77	Urban	240.09	27.0	never smoked	Yes
78	Urban	110.85	24.1	formerly smoked	Yes

80	42117	Male	43	No	No	Yes	Self-employed
81	57419	Male	59	No	No	Yes	Private
83	26727	Female	79	No	No	No	Private
84	66638	Female	68	Yes	No	No	Self-employed
86	32399	Male	54	No	No	Yes	Private
87	3253	Male	61	No	Yes	Yes	Private
88	71796	Female	70	No	Yes	Yes	Private
89	14499	Male	47	No	No	Yes	Private
90	49130	Male	74	No	No	Yes	Private
91	28291	Female	79	No	Yes	Yes	Private
92	51169	Male	81	No	No	Yes	Private
93	66315	Female	57	No	No	No	Self-employed
94	37726	Female	80	Yes	No	Yes	Self-employed
95	54385	Male	45	No	No	Yes	Private
96	2458	Female	78	No	No	Yes	Private
97	35512	Female	70	No	No	Yes	Self-employed
98	56841	Male	58	No	Yes	Yes	Private
99	8154	Male	57	Yes	No	Yes	Govt_job
100	4639	Female	69	No	No	Yes	Govt_job

Figure 3.12: Partitional Clustering Training Model on Data

4.0 Evaluation

4.1 Decision Tree

In any data mining program, evaluation is a crucial step aimed at verifying the findings obtained through the selection of specific methods. However, in this project Partitional Clustering, Decision Tree, Bayesian classifiers are used as data mining approaches where the predict () function is applied to the stroke test dataset using these algorithms.

In Partitional Clustering, the aim is to partition similar data points into clusters in order to understand patterns within a dataset. Contrastingly, the Decision Tree constructs a tree-like model based on observations including features upon which new cases are classified. In the case of the Bayesian classifier, this approach entails creation of a model through reviewing training data towards predicting unseen category labels.

The classification process presents the challenge of identifying which class an observation belongs to. All methods require a process of analysing training data to obtain either model or classifier, which is then applied for predicting the labels in new incoming data. The understanding of classification metrics, more specifically the confusion matrix is important in interpreting results. Figure 4.1 is a summary of prediction outcomes based on the Decision Tree approach.

```
> p1<-predict(m1, stroker[4001:4900,])
> p1
  [1] No No No No No No No No No No No No No No No No
 [17] No No No No No No No No No No No No No No No No
 [33] No No No No No No No No No No No No No No No No
 [49] No No No No No No No No No No No No No No No No
 [65] No No No No No No No No No No No No No No No No
 [81] No No No No No No No No No No No No No No No No
 [97] No No No No No No No No No No No No No No No No
[113] No No No No No No No No No No No No No No No No
[129] No No No No No No No No No No No No No No No No
[145] No No No No No No No No No No No No No No No No
[161] No No No No No No No No No No No No No No No No
[177] No No No No No No No No No No No No No No No No
[193] No No No No No No No No No No No No No No No No
[209] No No No No No No No No No No No No No No No No
[225] No No No No No No No No No No No No No No No No
[241] No No No No No No No No No No No No No No No No
[257] No No No No No No No No No No No No No No No No
[273] No No No No No No No No No No No No No No No No
[289] No No No No No No No No No No No No No No No No
[305] No No No No No No No No No No No No No No No No
[321] No No No No No No No No No No No No No No No No
[337] No No No No No No No No No No No No No No No No
[353] No No No No No No No No No No No No No No No No
[369] No No No No No No No No No No No No No No No No
[385] No No No No No No No No No No No No No No No No
[401] No No No No No No No No No No No No No No No No
[417] No No No No No No No No No No No No No No No No
[433] No No No No No No No No No No No No No No No No
[449] No No No No No No No No No No No No No No No No
[465] No No No No No No No No No No No No No No No No
[481] No No No No No No No No No No No No No No No No
[497] No No No No No No No No No No No No No No No No
[513] No No No No No No No No No No No No No No No No
[529] No No No No No No No No No No No No No No No No
[545] No No No No No No No No No No No No No No No No
[561] No No No No No No No No No No No No No No No No
[577] No No No No No No No No No No No No No No No No
[593] No No No No No No No No No No No No No No No No
[609] No No No No No No No No No No No No No No No No
```

```
[625] No No No No No No No No No No No No No No No No
[641] No No No No No No No No No No No No No No No No
[657] No No No No No No No No No No No No No No No No
[673] No No No No No No No No No No No No No No No No
[689] No No No No No No No No No No No No No No No No
[705] No No No No No No No No No No No No No No No No
[721] No No No No No No No No No No No No No No No No
[737] No No No No No No No No No No No No No No No No
[753] No No No No No No No No No No No No No No No No
[769] No No No No No No No No No No No No No No No No
[785] No No No No No No No No No No No No No No No No
[801] No No No No No No No No No No No No No No No No
[817] No No No No No No No No No No No No No No No No
[833] No No No No No No No No No No No No No No No No
[849] No No No No No No No No No No No No No No No No
[865] No No No No No No No No No No No No No No No No
[881] No No No No No No No No No No No No No No No No
[897] No No No No
Levels: No Yes
> table(stroker[4001:4900,12],Predicted= p1)
     Predicted
      No Yes
  No  861   0
  Yes  39   0
> plot(m1)
```

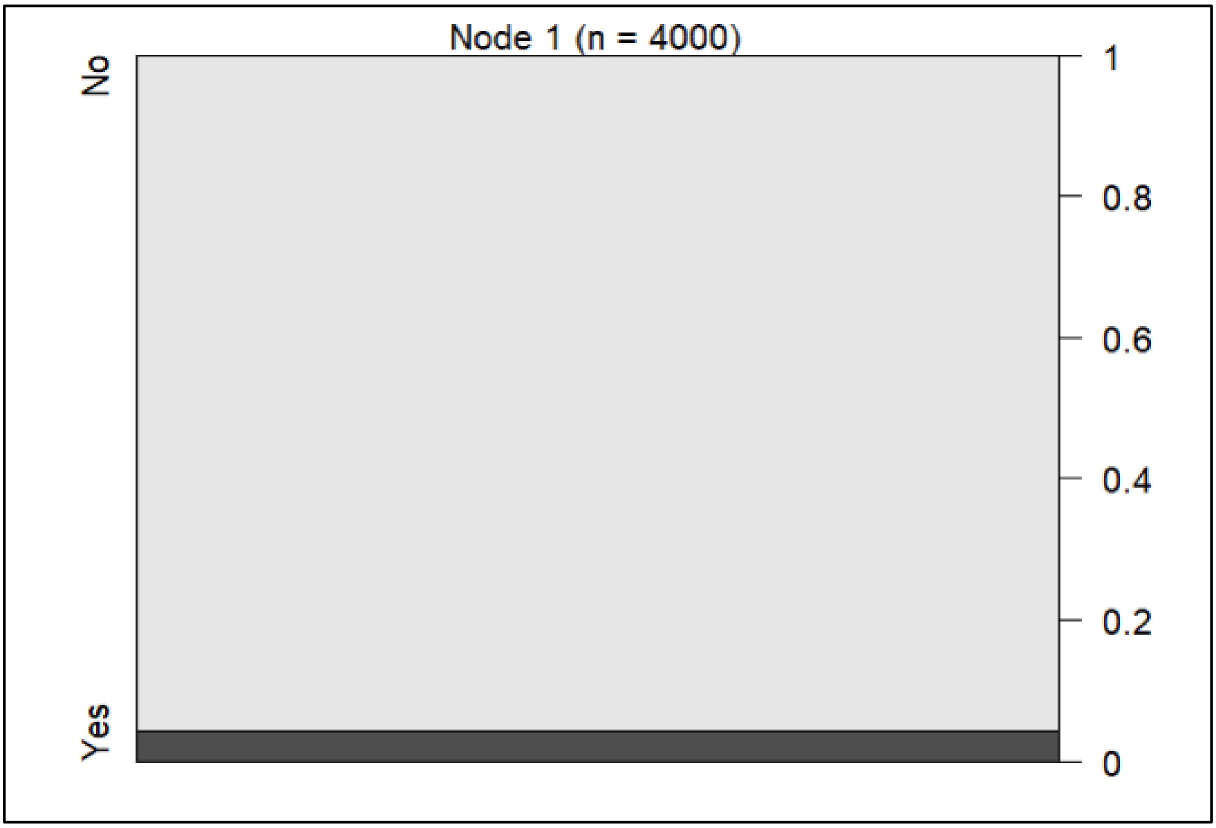

Figure 4.1: Results of Prediction using Decision Tree

There were 900 predictions which means that out of the total population, a sample consisting of 90 testable patients was available. There are two possible predicted classes: The words “Yes” indicate patients having a stroke, while the counterpart is reflected when patients do not have a stroke. From these 1636 observations, the percentage of patients who were predicted as having a stroke by classifier was recorded at 2.4% and that of those not having it stood at one-fifth (87%) while two percent erroneously labelled them for being susceptible to such attacks; thus an error rate approximately equal with what seemed reasonable based on accurate evidence available.

The C5.0 decision tree model, which has a small tightness degree 1 , showed high accuracy in the training phase with an error rate of about 4% . In the given training set, it correctly predicted ‘No’ for 3830 cases and 'Yes' for 170. But during performance on a fresh subset of 900 observations, the model always predicted 'No', implying that there were no positive predictions for stroke. This trend was further confirmed by the confusion matrix analysis on the evaluation subset (4001:It seems that this model is not very sensitive to positive cases, so there were several instances of No predicted in 861 (N$ =4900), and the ‘Yes’ prediction did not appear at all.

4.2 Bayesian Classifier

A classification method based on the Naïve Bayes algorithm is used in this project on a collection of stroke tests. Even though classification works, it can be hard to put new findings into the right classes. In training, the classifier is provided with a dataset in which each element is assigned a class name. After that, the model that was trained is used on new data that it hasn't seen before to guess class names. The confusion matrix is an important idea in classification metrics because it helps us understand how the predictions turned out.

The Naïve Bayes algorithm in the given scenario created a total of 1636 predictions to classify individuals as either having a stroke ("Yes") or not having a stroke ("No"). Among these predictions, 1667 were precise, accurately identifying 53 cases of stroke and 1477 cases of non-stroke. The confusion matrix provides a detailed breakdown of the results, showing that there were 1477 cases correctly identified as not having a stroke (true negatives), 17 cases correctly identified as having a stroke (true positives), 89 cases incorrectly identified as having a stroke (false positives), and 53 cases incorrectly identified as not having a stroke (false negatives).

```
> y_pred <- predict(classifier_stroke, newdata = test_stroke)
> cm <- table(test_stroke$stroke, y_pred)
> cm
     y_pred
        No  Yes
  No  1477   89
  Yes   53   17
> confusionMatrix(cm)
Confusion Matrix and Statistics

     y_pred
        No  Yes
  No  1477   89
  Yes   53   17

               Accuracy : 0.9132
                 95% CI : (0.8985, 0.9264)
    No Information Rate : 0.9352
    P-Value [Acc > NIR] : 0.999774

                  Kappa : 0.1493

 Mcnemar's Test P-Value : 0.003313
```

```
            Sensitivity : 0.9654
            Specificity : 0.1604
         Pos Pred Value : 0.9432
         Neg Pred Value : 0.2429
             Prevalence : 0.9352
         Detection Rate : 0.9028
   Detection Prevalence : 0.9572
      Balanced Accuracy : 0.5629

       'Positive' Class : No

> ggplot(test_stroke, aes(stroke, y_pred, color = stroke))+geom_jitter(width = 0.2, height = 0.1, size = 2)+labs
(title = "Confusion Matrix", subtitle = "Predicted VS Observed from Stroke dataset", y = "Predicted", x = "Trut
h")
> |
```

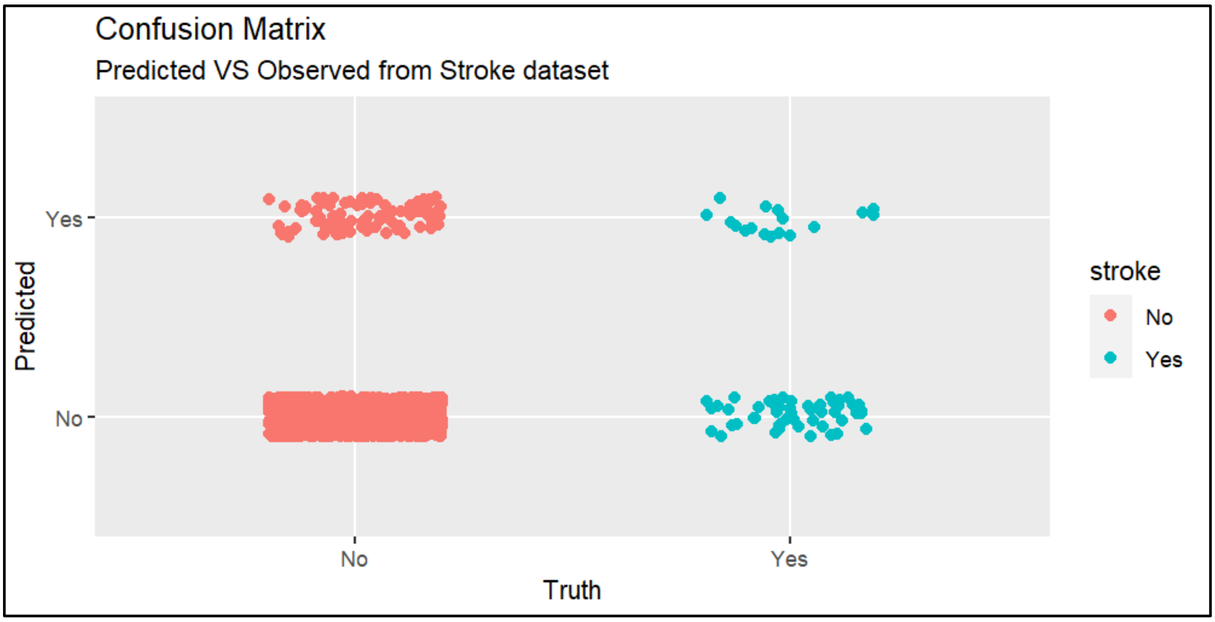

Figure 4.2: Results of Prediction using Bayesian Classifier (Naive Bayes Algorithm)

The accuracy score, which measures the model's performance, reveals that the classifier made accurate predictions around 91.32% of the time, with an error rate of 8.68%. This indicates an important level of precision in classifying the stroke attribute within the dataset, showing the efficiency of the classification model. The ggplot2 package's visual display of real and estimated strokes in Figure 4.2 improves comprehension of the classification results. Overall, the assessment metrics confirm the Naïve Bayes algorithm's accuracy and dependability for this classification task.

4.3 Partitional Clustering

In this study, a set of stroke tests was analysed using the partitional clustering method. Many characteristics, including age, BMI, average blood sugar level, cardiac disease, and hypertension, are probably included in the stroke prediction dataset. Partitional clustering of these features enables the identification of discrete clusters within the dataset based on commonalities across these variables. For example, people who share similar risk profiles or demographic traits can come together to form cohesive clusters, which can help identify common patterns that increase the risk of stroke. Partitional clustering is a technique that reveals underlying patterns in a dataset by clustering comparable data points together.

```
> str(stroke[, c("age", "hypertension", "heart_disease", "avg_glucose_level", "bm
i")])
'data.frame':   4908 obs. of  5 variables:
 $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 1 ...
 $ heart_disease    : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 2 1 1 1 2 ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi              : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
```

```
> str(stroke[, c("age", "hypertension", "heart_disease", "avg_glucose_level", "bm
i")])
'data.frame':   4908 obs. of  5 variables:
 $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension     : num  1 1 1 2 1 2 1 1 2 1 ...
 $ heart_disease    : num  2 2 1 1 1 2 1 1 1 2 ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi              : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
> stroke[, c("age", "hypertension", "heart_disease", "avg_glucose_level", "bmi")] <-
+   lapply(stroke[, c("age", "hypertension", "heart_disease", "avg_glucose_level",
"bmi")], as.numeric)
> scaled_features <- scale(stroke[, c("age", "hypertension", "heart_disease", "avg_g
lucose_level", "bmi")])
> k <- 2
> kmeans_result <- kmeans(scaled_features, centers = k)
> stroke$cluster <- kmeans_result$cluster
> table(stroke$cluster)

   1    2 
4457  451 
> cluster_means <- aggregate(. ~ cluster, data = stroke, FUN = mean)
> print(cluster_means)
  cluster       id   gender      age hypertension heart_disease ever_married
1       1 37052.37 1.406327 40.90018            1      1.041508     1.628225
2       2 37140.02 1.443459 62.32373            2      1.128603     1.895787
  work_type Residence_type avg_glucose_level      bmi smoking_status   stroke
1  3.435719       1.507516          102.7453 28.47543       2.630245 1.033431
2  3.988914       1.505543          130.5190 33.03659       2.128603 1.133038
> 
```

```
> str(scaled_features)
 num [1:4908, 1:5] 1.07 1.646 0.272 1.602 1.691 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:4908] "1" "3" "4" "5" ...
  ..$ : chr [1:5] "age" "hypertension" "heart_disease" "avg_glucose_level" ...
 - attr(*, "scaled:center")= Named num [1:5] 42.87 1.09 1.05 105.3 28.89
  ..- attr(*, "names")= chr [1:5] "age" "hypertension" "heart_disease" "avg_glucose_
level" ...
 - attr(*, "scaled:scale")= Named num [1:5] 22.556 0.289 0.217 44.426 7.854
  ..- attr(*, "names")= chr [1:5] "age" "hypertension" "heart_disease" "avg_glucose_
level" ...
> sum(is.na(scaled_features))
[1] 0
> kmeans_result <- kmeans(scaled_features, centers = k, iter.max = 100)
> kmeans_result <- kmeans(scaled_features, centers = k, nstart = 10)
> set.seed(123)
> kmeans_result <- kmeans(scaled_features, centers = k)
> library(ggplot2)
> pca_result <- prcomp(scaled_features)
> pca_data <- as.data.frame(pca_result$x[, 1:2])
> pca_data$cluster <- as.factor(kmeans_result$cluster)
> ggplot(pca_data, aes(x = PC1, y = PC2, color = cluster)) +
+   geom_point() +
+   ggtitle("K-Means Clustering") +
+   xlab("Principal Component 1") +
+   ylab("Principal Component 2") +
+   theme_minimal()
```

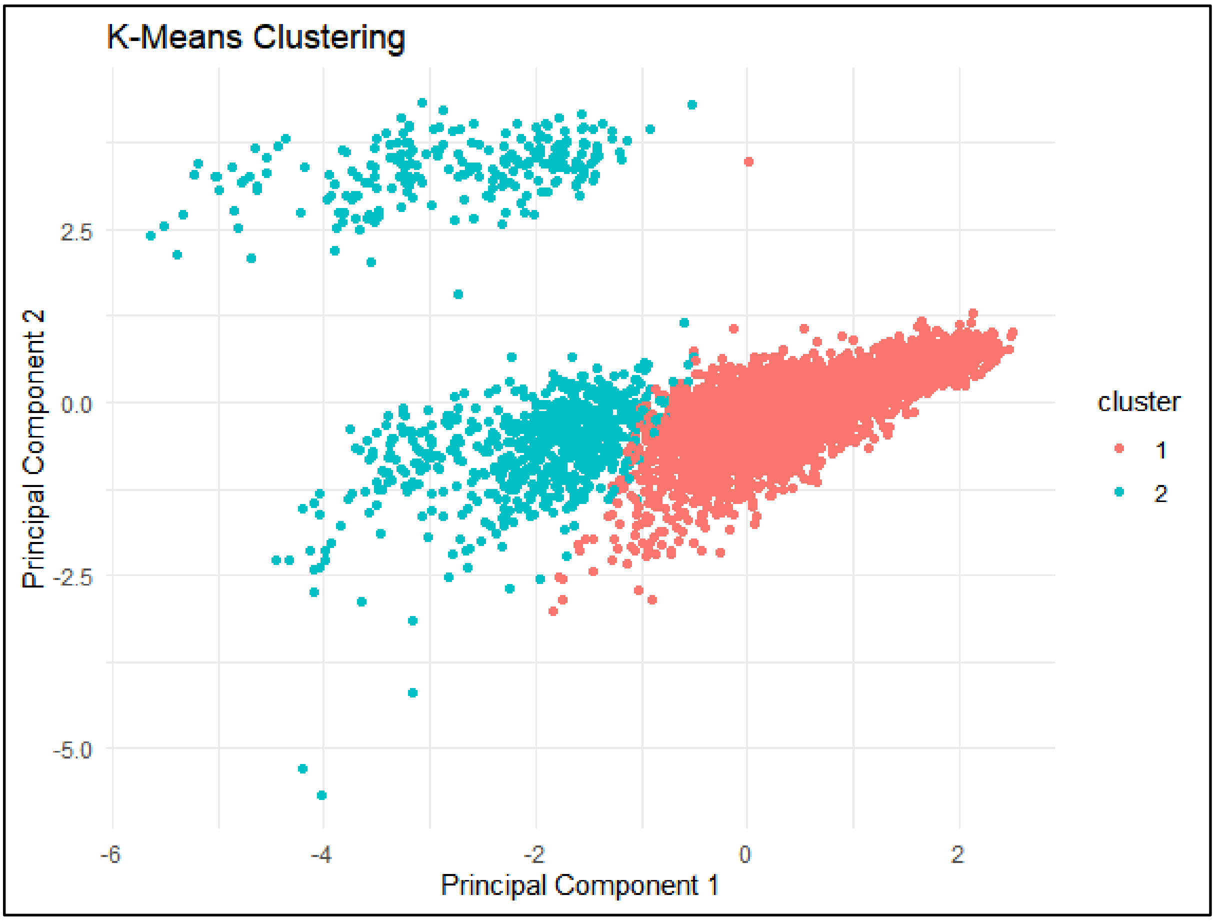

Figure 4.3: Results of Prediction using Partitional Clustering

The clustering algorithm produced a total of two predictions in this scenario: those who have not had a stroke are grouped in "1," where the answer is "No," and those who have had a stroke are grouped in "2," where the answer is "Yes." This is done in order to group people into various clusters. Of these predictions, 451 were correct in classifying data items into the appropriate groupings. The accuracy of the partitional clustering is demonstrated in detail in Figure 4.3, which also shows the number of cases that were correctly classified inside each cluster and any instances of misclassification that took place. Comprehending these outcomes is essential for assessing how well the partitional clustering technique reveals relevant patterns within the stroke test dataset.

5.0 Conclusion

In conclusion, a stroke prediction dataset was analysed using sophisticated data mining techniques, such as Partitional Clustering, Decision Trees, and Bayesian Classifiers. The objective was to create efficient models that could forecast the risk of stroke by taking into account a range of health-related characteristics, including age, gender, and lifestyle choices. As a result of the investigation of the stroke dataset, it was discovered that stroke is a serious health concern in Malaysia, which highlights the importance of developing reliable prediction models. Cleaning, converting, and examining the dataset were all part of the data preparation process. This was done to guarantee that the dataset was suitable for training and testing the models.

Other than that, there was a structured method for categorization that was offered by decision tree modelling. This method created a tree-like model that was based on features in order to anticipate the outcomes of strokes. In the evaluation, the tendency of the model to primarily forecast the absence of stroke was noted, and the model's cautious approach to finding positive cases was emphasised. The Naïve Bayes algorithm, which was utilised as a Bayesian Classifier, exhibited remarkable accuracy, accurately predicting the incidence of strokes in the majority of cases. An extremely high level of precision was found in the classification of both positive and negative cases, as demonstrated by the confusion matrix analysis. Despite the fact that it produced unique clusters, the Partitional Clustering technique demonstrated a moderate level of accuracy. It is possible that additional optimisation might be investigated in order to improve its prediction powers.

Based on the findings, it can be concluded that of the three methods, the Bayesian Classifier (also known as the Naïve Bayes Algorithm) is the most precise and dependable. It is a suggested option for stroke prediction in this dataset due to its balanced performance in correctly classifying stroke instances, simplicity, and efficiency. The Naïve Bayes algorithm is a promising tool for healthcare practitioners looking for an interpretable and effective stroke prediction solution because of its high accuracy and ability to handle categorical data.

References

Shukla, R.K. et al. (2020) 'WEB USAGE MINING-A Study of Web data pattern detecting methodologies and its applications in Data Mining', 2nd International Conference on Data, Engineering and Applications (IDEA) [Preprint]. Available at: https://doi.org/10.1109/idea49133.2020.9170690.

Fedesoriano (2021) 'Stroke Prediction Dataset', Kaggle, p. 12. Available at: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

Nurul Fatin Rakib et al. (2020) 'Preliminary Results of Hand Rehabilitation for Post Stroke Patient using Leap Motion-based Virtual Reality'. Available at: https://doi.org/10.1109/scored50371.2020.9250985.

World Health Ranking, January 2019, [online] Available: https://WWW.Worldlifeexpectancy.com/malaysia-stroke.

Intan Rahmatillah, Eriana Astuty and Ivan Diryana Sudirman (2023) 'An Improved Decision Tree Model for Forecasting Consumer Decision in a Medium Groceries Store'. Available at: https://doi.org/10.1109/iciis58898.2023.10253592.

Ain, Khusnul, et al. "Expert System for Stroke Classification Using Naive Bayes Classifier and Certainty Factor as Diagnosis Supporting Device." *Journal of Physics: Conference Series*, vol. 1445, no. 1, IOP Publishing, Jan. 2020, p. 012026. *Crossref*, https://doi.org/10.1088/1742-6596/1445/1/012026.

Kutbay, U. (2018, August 1). Partitional Clustering. *Recent Applications in Data Clustering*. https://doi.org/10.5772/intechopen.75836

UMS UNIVERSITI MALAYSIA SABAH — FACULTY OF COMPUTING & INFORMATICS / FAKULTI KOMPUTERAN DAN INFORMATIK	**BORANG DEKLARASI (BERKUMPULAN) / *DECLARATION FORM* (*GROUP*) UNTUK PENILAIAN MOD *ASYNCHRONOUS* / FOR ASYNCHRONOUS MODE ASSESSMENT**

A. MAKLUMAT PELAJAR/*STUDENT'S INFORMATION*

Nama Penuh Ketua / Leader's Full Name: [Mengikut MyKad atau paspot / *As in MyKad or Passport*]	Khoo Huang Kwang
No. Matrik Ketua / Leader's *Matric No*.:	BI20110245
No. Telefon dan e-mel ketua / Leader's Phone No. and e-mail:	01110059264 khoo_huang_bi20@iluv.ums.edu.my
Nama Kursus / *Course Name*:	Data Mining
Kod Kursus / *Course Code*:	KK04703
Jenis Penilaian / *Assessment Type* :	☑ Tugasan / *Assignment* ☐ Kuiz / *Quiz* ☐ Lain-lain / *Others* : (Nyatakan / *Specify*) ……………………………………………… ☐ Latihan Makmal / *Lab Exercise* ☐ Projek / *Project*

B. DEKLARASI/ DECLARATION

SILA BACA DENGAN TELITI SEBELUM MELENGKAPKAN BORANG INI/ *PLEASE READ CAREFULLY BEFORE COMPLETING THE FORM*

1. Saya dengan ini mengaku bahawa penilaian ini adalah hasil kerja saya sendiri sepenuhnya. Sebarang sumber maklumat tambahan, sama ada diterbitkan secara terbuka atau tidak diterbitkan, yang telah digunakan dan dirujuk dalam teks, senarai kandungan dan/atau senarai bibliografi telah diakui dengan sewajarnya dan semakan kesamaan melalui Turnitin adalah tidak melebihi 30%.
   I hereby declare that this assessment is entirely my own work. Any additional sources of information, whether publicly published or unpublished, that have been utilized and referenced within the text, contents list, and/or bibliographical list have been duly acknowledged and similarity check through Turnitin is not more than 30%
2. Kami tidak akan membenarkan sesiapa menyalin atau mengeluarkan semula karya kami tanpa kebenaran kami dengan mengambil semua langkah berjaga-jaga.
   We will not allow anyone to copy or reproduce our work without our authorization by taking necessary precautions.
3. Kami sedar dan memahami dasar Universiti mengenai plagiarisme dan amalan akademik yang baik. Kami sedar sepenuhnya bahawa sebarang pelanggaran dasar yang berkaitan akan mengakibatkan peruntukan sifar markah untuk penilaian ini dan mungkin akan mengakibatkan tindakan tatatertib.
   We are aware of and understand the University's policy on plagiarism and good academic practices. We are fully aware that any violation of the relevant policies will result in the allocation of zero marks for this assessment and may also warrant disciplinary actions.

☐ Saya ada menggunakan *AI tools* dalam penyelesaian tugasan/ penilaian ini
I have used AI tools in this assignment Assessment

C. PERAKUAN PELAJAR / *STUDENT DECLARATION*

No. K.P. / Paspot Ketua: *Leader's I.C. no. / Passport:* 980928-04-5297

Tandatangan Ketua / Leaders's *Signature*:

-- Tarikh / *Date*: 22/1/2024

Nota/ Notes

1. Pelajar adalah digalakkan untuk melampirkan bersama laporan Turnitin (Kurang 30%) bersama borang ini tertakluk kepada permintaan pensyarah masing-masing.
   Students are encouraged to attach their Turnitin reports (Less than 30%) with this form subject to the lecturer's request.

UMS
UNIVERSITI MALAYSIA SABAH

FACULTY OF COMPUTING & INFORMATICS
FAKULTI KOMPUTERAN DAN INFORMATIK

BORANG DEKLARASI (BERKUMPULAN) / *DECLARATION FORM* (*GROUP*)
ASSIGNMENT/ LAB EXERCISE/ QUIZ/ PROJECT

D. SENARAI AHLI KUMPULAN / LIST OF GROUP MEMBERS

NO.	NAMA / NAME	NO. MATRIKS / MATRICS NUMBER	NO. KAD PENGENALAN / I.C. NUMBER	TANDATANGAN / SIGNATURE
1.	Juliana Ekot	BI20110242	980713-13-5914	
2.	Emerlyn Trisha Mering ak James	BI20160322	981019-13-5396	
3.	Khoo Huang Kwang	BI20110245	980928-04-5297	
4.				
5.				
6.				
7.				
8.				
9.				
10.				