STAT 652 Final Project

Dhruv Patel (dnp5@sfu.ca) & Rahil Lavkumar Balar (rlb4@sfu.ca)

30/11/2021

Introduction

A stroke is a life-threatening medical disorder. It's a serious ailment, but if caught early enough, we may save a person's life and provide excellent care. There are several variables that might cause strokes, and we will attempt to investigate a few of them in this research using statistical learning in R. This is an analysis report of the Stroke Prediction Dataset. The dataset has 5110 rows and 12 columns. In this dataset, Our goal is to determine if a patient is at risk for stroke based on various factors such as gender, age, illness, and smoking status, etc.

Predicting Stroke (Classification Analysis)

1 Data

1.1 Data Loading

The first step is to load the appropriate dataset in R Studio environment using the read.csv command.

```
data <- read.csv("/Users/dhruv/Desktop/Docs/STAT_652/Project3/healthcare-dataset-stroke-data.csv", strin
summary(data)
```

```
##        id           gender          age         hypertension    
##  Min.   :   67   Female:2994   Min.   : 0.08   Min.   :0.00000  
##  1st Qu.:17741   Male  :2115   1st Qu.:25.00   1st Qu.:0.00000  
##  Median :36932   Other :   1   Median :45.00   Median :0.00000  
##  Mean   :36518                 Mean   :43.23   Mean   :0.09746  
##  3rd Qu.:54682                 3rd Qu.:61.00   3rd Qu.:0.00000  
##  Max.   :72940                 Max.   :82.00   Max.   :1.00000  
##                                                                 
##  heart_disease     ever_married          work_type    Residence_type
##  Min.   :0.00000   No :1757     children     : 687   Rural:2514    
##  1st Qu.:0.00000   Yes:3353     Govt_job     : 657   Urban:2596    
##  Median :0.00000                Never_worked :  22                 
##  Mean   :0.05401                Private      :2925                 
##  3rd Qu.:0.00000                Self-employed: 819                 
##  Max.   :1.00000                                                   
##                                                                    
##  avg_glucose_level      bmi                smoking_status     stroke       
##  Min.   : 55.12    N/A    : 201   formerly smoked: 885   Min.   :0.00000  
##  1st Qu.: 77.25    28.7   :  41   never smoked   :1892   1st Qu.:0.00000  
##  Median : 91.89    28.4   :  38   smokes         : 789   Median :0.00000  
##  Mean   :106.15    26.1   :  37   Unknown        :1544   Mean   :0.04873
##  3rd Qu.:114.09    26.7   :  37                          3rd Qu.:0.00000
##  Max.   :271.74    27.6   :  37                          Max.   :1.00000
##                    (Other):4719
```

From the summary of the Stroke Prediction Data, we can observe that there are total 12 columns. In addition, the summary shows that there are no NA or missing values in any columns except the bmi column.

1.2 Exploratory Data Analysis

Missing Data:

To start with our data analysis, we notied that in the bmi column there are 201 entries with value “N/A”. We can replace that with the actual NA values. As we have the luxury of data so we could simply remove the NA bmi records.

```
data[data=="N/A"] <- NA
data <- na.omit(data)
```

ID:

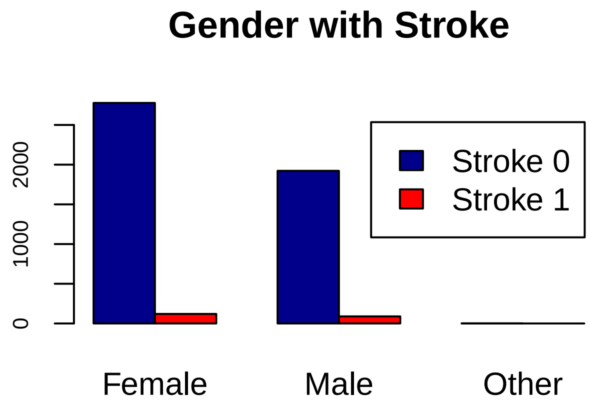
The column id is not needed for prediction as it just states the unique id of each patient, so we can remove the id column from our data.

```
data <- data[2:12]
```

Gender:

The gender of the patient is indicated by this feature. Let’s look at how gender influences stroke rates and compare stroke rates by gender.

```
##   data$gender    n
## 1      Female 2897
## 2        Male 2011
## 3       Other    1
```



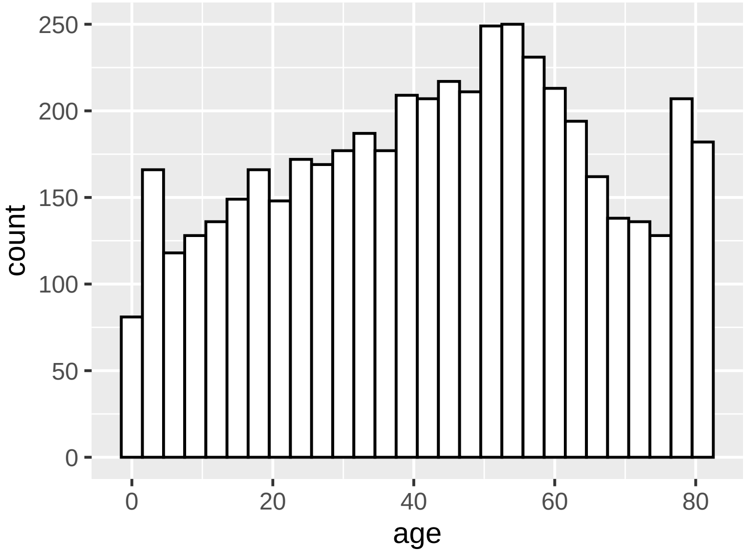
The dataset appears to be unbalanced. As far as we can tell, there isn't much of a difference in stroke rates between men and women. Also there is only one value of other type gender, since it is not helping in predicting stroke (as it is only one data sample) we can remove the record having other as gender type.

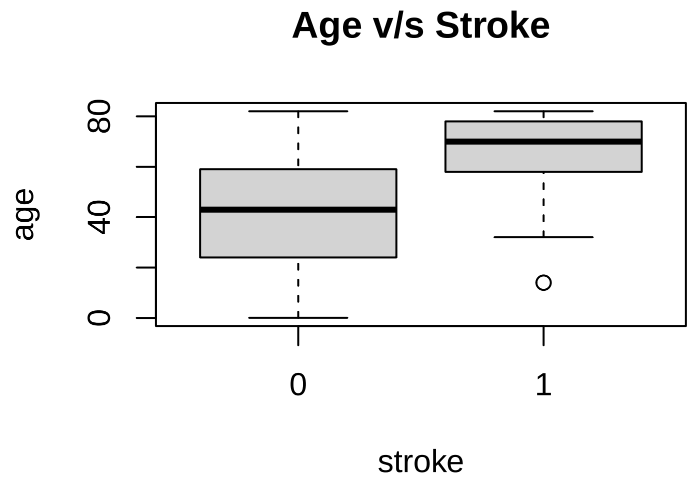
```
data <- data[!(data$gender=="Other"),]
```

Age:

Analytically, age would play a huge role in deciding whether the person will suffer from stroke or not. Let us find how many unique values of age are there and what is the age distribution.

```
## [1] 104
```



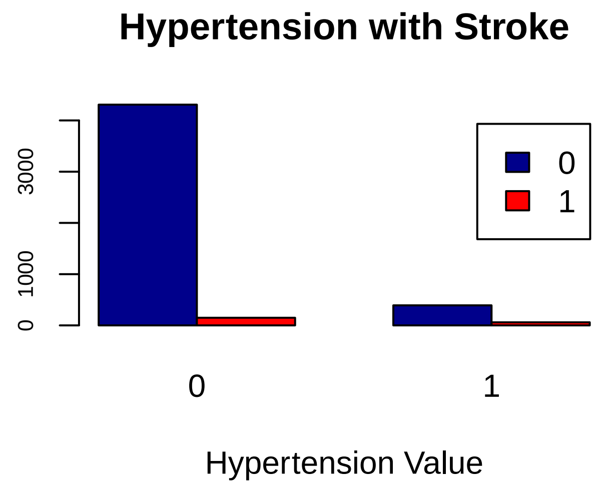


As expected, we can see the mean age for people suffering from stroke is higher for older people than the younger ones.

Hypertension:

Hypertension is a condition in which a person's blood pressure is abnormally high. A stroke may occur as a result of hypertension which makes it one of the deciding factor to predict whether a person will suffer from stroke or not.

```
##   data$hypertension    n
## 1                 0 4457
## 2                 1  451
```

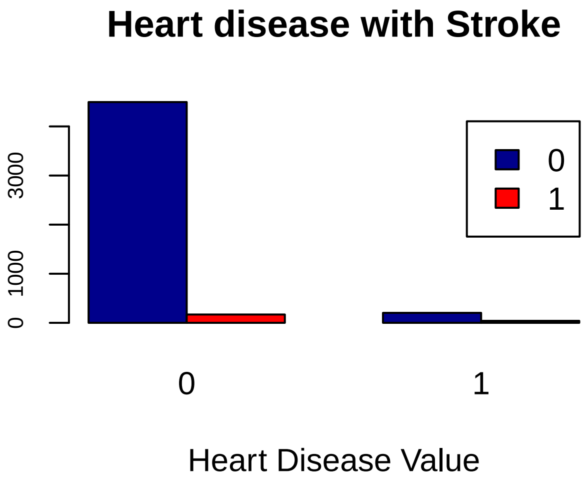


As we can see in our data plot, hypertension is uncommon in young adults but widespread in the elderly. A stroke can be caused by hypertension but the picture for hypertension is not as clear as it appears based on our statistics because our data provides very little information about hypertensive individuals.

Heart Disease:

People suffering from heart disease are more vulnerable towards stroke, if proper care is not taken. Let us find out how heart_disease and stroke are related

```
##   data$heart_disease    n
## 1                  0 4665
## 2                  1  243
```

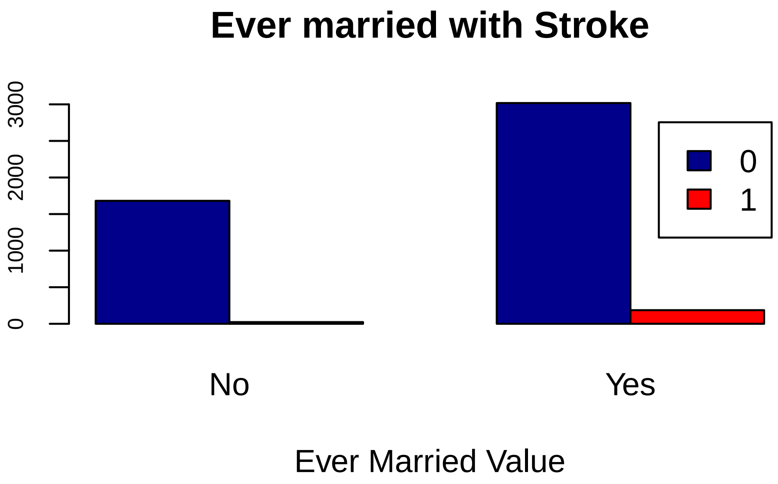


The data is quite imbalanced for this feature, making it harder for us to assert a concrete assumption. But we think, heart disease would not play a deciding factor for stroke.

Ever Married:

This feature lets us know whether the patient was married or not. Let us check whether this feature affects the probability of having stroke.

```
##   data$ever_married    n
## 1               Yes 3204
## 2                No 1704
```

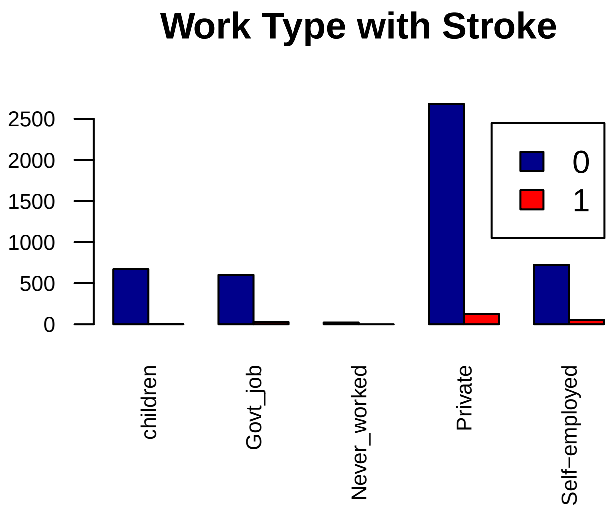


Statistically speaking, we can depict from the graphs that people who are married have a high chance of suffering from stroke.

Work Type:

This attribute contains information regarding the patient's work. Various kinds of work present different issues and obstacles, which might lead to feelings of enthusiasm, thrill, tension, and so on. Stress is bad for the health, so let's explore how it affects the chances of getting a stroke.

```
##   data$work_type    n
## 1        Private 2810
## 2  Self-employed  775
## 3       children  671
## 4       Govt_job  630
## 5   Never_worked   22
```

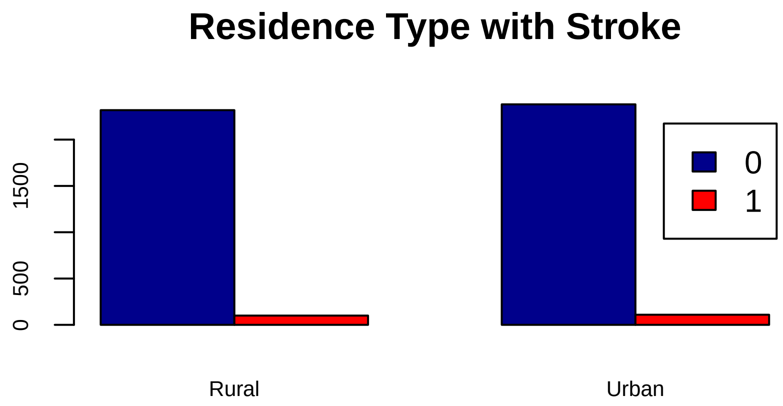


We can observe form the data that the people who are working in the private sector have the highest amount of risk of getting a stroke, whereas people who have never worked have the least amount of risk. Other categories have some what similar amount of risk of getting a stroke.

Residence Type:

This attribute lets us know whether the person lives in an Urban residence or a Rural residence.

```
##   data$Residence_type    n
## 1               Urban 2490
## 2               Rural 2418
```

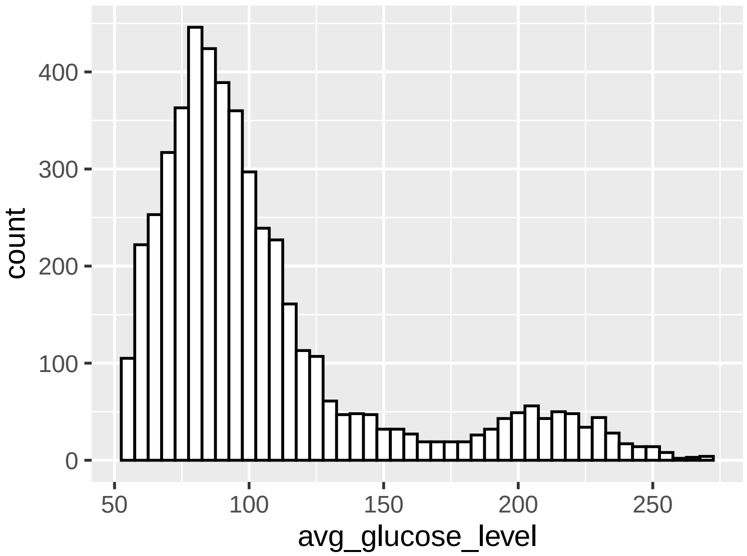


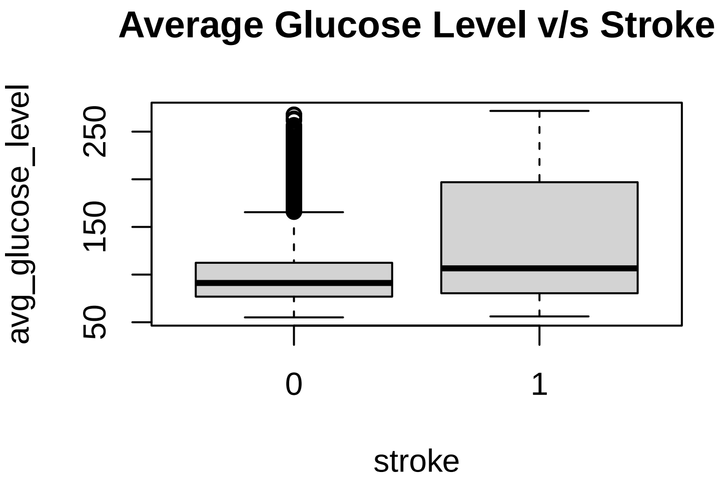
After observing the plots we can see this attribute is showing similar trends in both Urban and Rural values, so we can disregard this variable when we predict the chances of Stroke.

Average Glucose Level:

This attribute lets us know about the average glucose level of the patient's body. Let us find how many unique values of average glucose level are there and what is the average glucose level distribution.

```
## [1] 3851
```



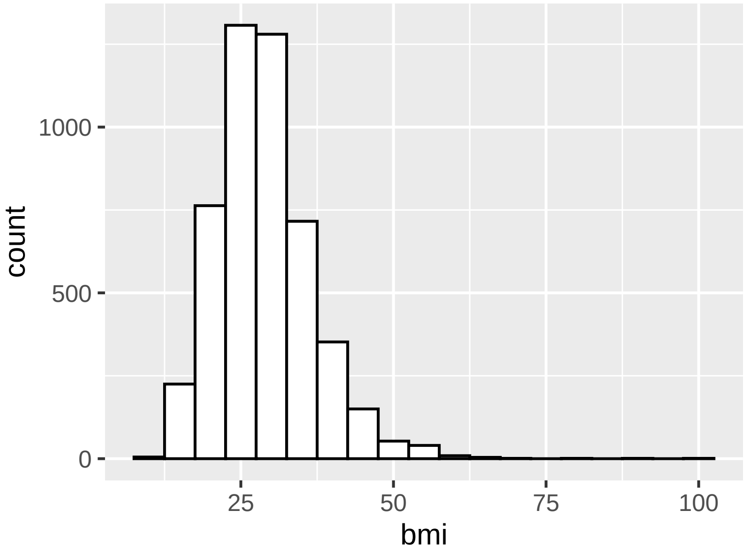


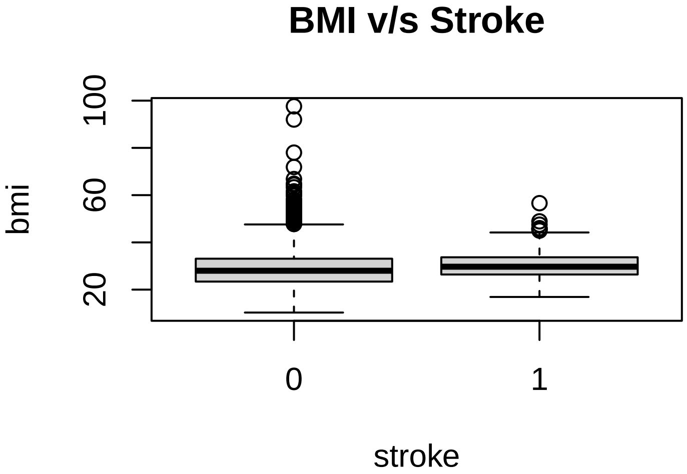
We can see from the graph above that persons who have a stroke have a glucose level of more than 100. Although there are some evident outliers in people who have not had a stroke, which assures that these are authentic data.

BMI:

The BMI is a calculation based on a person's weight and height. The BMI is calculated by dividing the body mass by the square of the body height.

```
## [1] 418
```



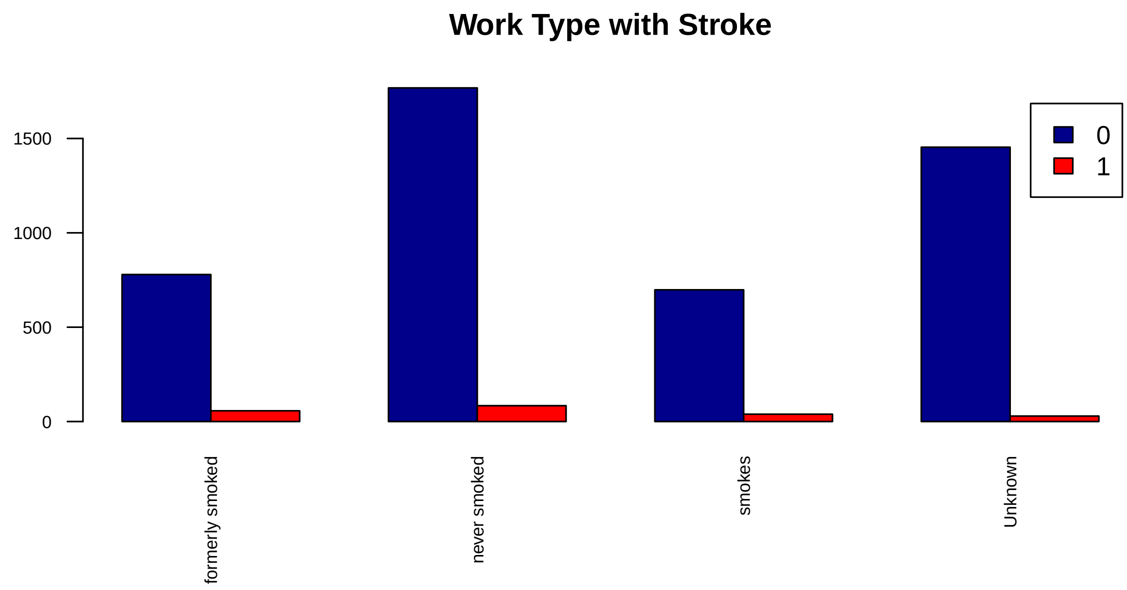


The boxplots does not provide a concrete evidence that can differentiate the patient's bmi relation with the stroke prediction, as both boxplots have similar trend.

Smoking Status:

We do have the prior knowledge that smoking can be a deciding factor for predicting that the patient will suffer from stroke or not. As, smokers tend to have a larger risk of experiencing stroke than a non-smoker patient.

```
##   data$smoking_status    n
## 1        never smoked 1852
## 2             Unknown 1483
## 3     formerly smoked  836
## 4              smokes  737
```



After observing the plots, there is no discernible amount of difference between the various smoking_status categories.

Stroke:

This is our target variable, which determines whether the patient will not suffer or will suffer from stroke. 0 means not affecting from stroke and 1 means affecting from stroke.

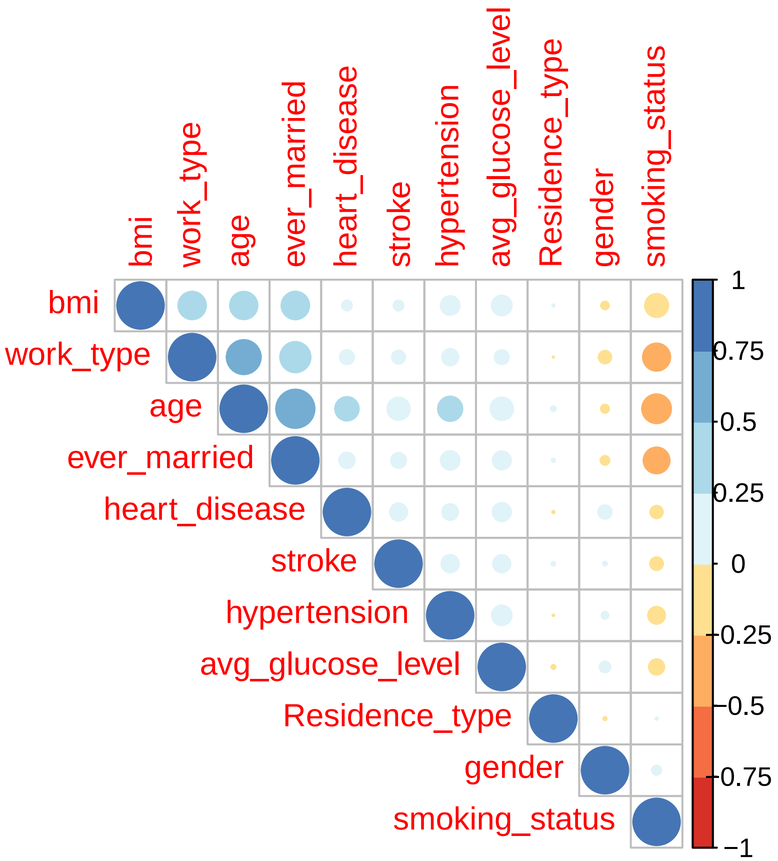
```
##   data$stroke    n
## 1           0 4699
## 2           1  209
```

1.3 Feature Engineering

Intuitively we now how the individual factors affect the probability of a person suffering from stroke or not but statistically we would like to find the correlation between various explanatory variables with themselves and with stroke, so we used corrplot() function to plot out the different correlations within the variables, and find out the important variables. Then by implementing various ML algorithms we can prove our hypothesis and check whether our results justify the correlation plot.

```
data$gender <- as.numeric(factor(data$gender))
data$ever_married <- as.numeric(factor(data$ever_married))
data$work_type <- as.numeric(factor(data$work_type))
data$Residence_type <- as.numeric(factor(data$Residence_type))
data$bmi<- as.numeric(factor(data$bmi))
data$smoking_status <- as.numeric(factor(data$smoking_status))
data$heart_disease <- as.numeric(factor(data$heart_disease))
data$hypertension <- as.numeric(factor(data$hypertension))
```

```
corrplot(cor(data),type="upper", order="hclust", col=brewer.pal(n=8, name="RdYlBu"))
```



As from the above correlation plot we can observe that, stroke and age have the highest amount of positive correlation i.e. as age increases chances of suffering from stroke increases whereas, stroke and smoking status have the highest amount of negative correlation i.e. the chances of stroke decreases if the patient never smokes.

1.4 Handling Categorical Variables

Now, we should convert the numeric variables into factor, as that is how we treat categorical variables in R. ML algorithms do not understand categorical values.

```
data$gender <- factor(data$gender)
data$ever_married <- factor(data$ever_married)
data$work_type <- factor(data$work_type)
data$Residence_type <- factor(data$Residence_type)
data$smoking_status <- factor(data$smoking_status)
data$heart_disease <- factor(data$heart_disease)
data$hypertension <- factor(data$hypertension)
data$stroke <- factor(data$stroke)
```

1.5 Data Splitting

Here we are splitting our data into two parts i.e. training and testing set. We are implementing a 75:25 split where 75% of the data is put into training set and the rest 25% is put into testing set. We are keeping our seed as 2021.

```
n = nrow(data)
set.seed(2021)
new.order = sample.int(n)
size.train = floor(n*0.75)
ind.train = new.order[1:size.train]
ind.test = new.order[(size.train + 1):n]
data.train = data[ind.train, ]
data.test = data[ind.test, ]
```

2 Methods

2.1 Logistic Regression

```
logreg1=glm(stroke~., data=data.train, family = "binomial")
summary(logreg1)
```

```
##
## Call:
## glm(formula = stroke ~ ., family = "binomial", data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1063  -0.2822  -0.1509  -0.0791   3.4491
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.155e+00  1.063e+00  -6.734 1.65e-11 ***
## gender2           -1.477e-01  1.838e-01  -0.804 0.421425
## age                7.125e-02  7.427e-03   9.593  < 2e-16 ***
## hypertension2      5.681e-01  2.049e-01   2.773 0.005553 **
## heart_disease2     3.534e-01  2.461e-01   1.436 0.151011
## ever_married2     -2.111e-01  2.905e-01  -0.727 0.467417
## work_type2        -7.483e-01  1.153e+00  -0.649 0.516179
## work_type3        -1.110e+01  5.786e+02  -0.019 0.984690
## work_type4        -6.761e-01  1.131e+00  -0.598 0.549901
## work_type5        -8.836e-01  1.155e+00  -0.765 0.444219
## Residence_type2    5.721e-02  1.774e-01   0.323 0.747052
## avg_glucose_level  5.430e-03  1.517e-03   3.579 0.000344 ***
## bmi                2.711e-04  1.463e-03   0.185 0.852924
## smoking_status2    2.442e-02  2.233e-01   0.109 0.912907
## smoking_status3    2.813e-01  2.765e-01   1.018 0.308866
## smoking_status4   -1.521e-01  2.889e-01  -0.526 0.598655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1247.57  on 3680  degrees of freedom
## Residual deviance:  986.67  on 3665  degrees of freedom
## AIC: 1018.7
##
```

```
## Number of Fisher Scoring iterations: 15
```

We can observe that the p value for age, hypertension, and average__glucose__level are highly significant and they play a higher role in predicting stroke than other variables.

Now, let us try another logistics regression model, with the variables classified as important by the first logistics regression model. Then compare the two models by doing ANOVA Test.

```
logreg2=glm(stroke~age+avg_glucose_level+hypertension, data=data.train, family = "binomial")
summary(logreg2)
```

```
##
## Call:
## glm(formula = stroke ~ age + avg_glucose_level + hypertension,
##     family = "binomial", data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0287  -0.2882  -0.1566  -0.0755   3.6208
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.835489   0.453437 -17.280  < 2e-16 ***
## age                0.068719   0.006458  10.641  < 2e-16 ***
## avg_glucose_level  0.005522   0.001457   3.790  0.00015 ***
## hypertension2      0.585286   0.201093   2.911  0.00361 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1247.57  on 3680  degrees of freedom
## Residual deviance:  993.49  on 3677  degrees of freedom
## AIC: 1001.5
##
## Number of Fisher Scoring iterations: 7
```

```
anova(logreg2,logreg1,test='LR')
```

```
## Analysis of Deviance Table
##
## Model 1: stroke ~ age + avg_glucose_level + hypertension
## Model 2: stroke ~ gender + age + hypertension + heart_disease + ever_married +
##     work_type + Residence_type + avg_glucose_level + bmi + smoking_status
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3677     993.49
## 2      3665     986.67 12   6.8228   0.8691
```

We can observe that the model having the subset of variables is better than the complex model having all the variables. As, the P value for the complex model is high, and it is common practice that if a simple model is more computationally efficient or equivalently efficient, we should select that model.

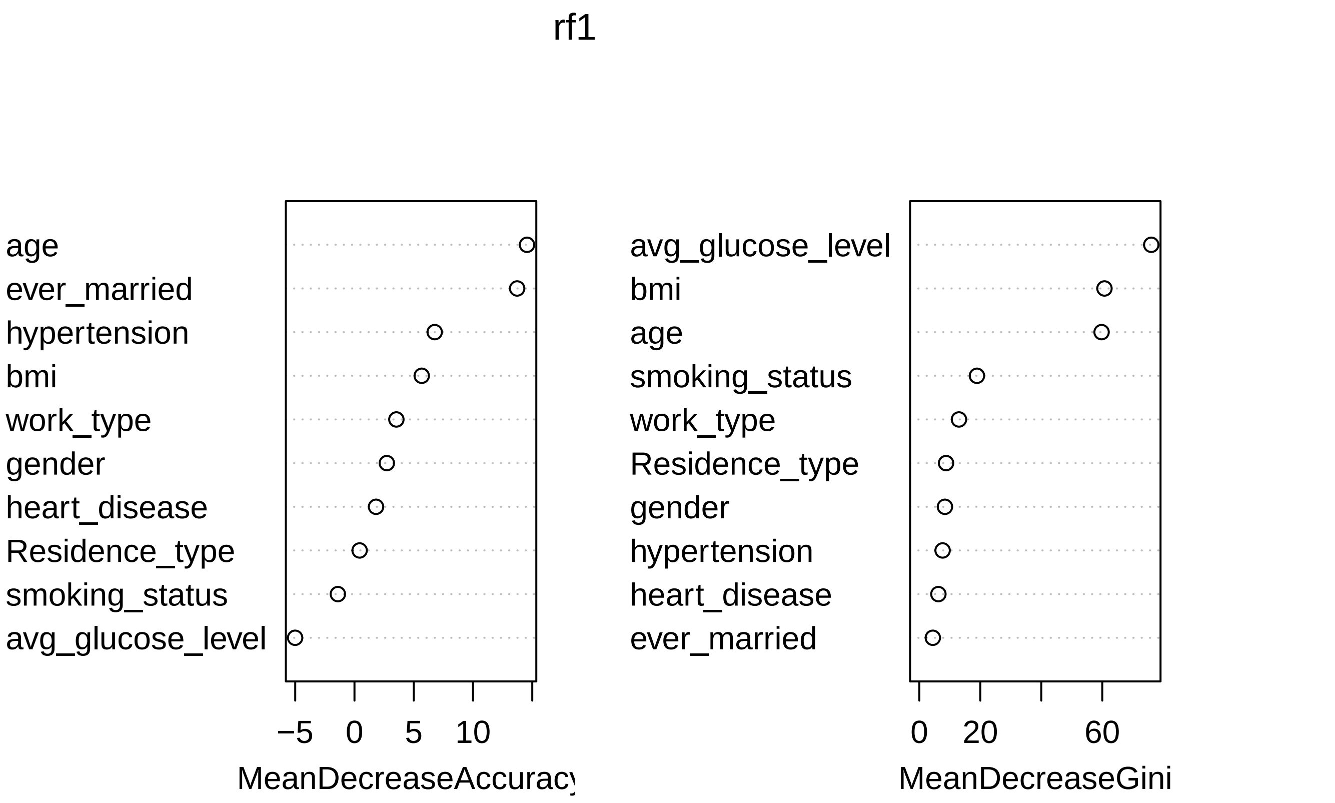
2.2 Random Forest

```
rf1 = randomForest(stroke~., data.train, importance=TRUE, proximity = TRUE);

#Confusion Matrix
rf1
```

```
##
## Call:
##  randomForest(formula = stroke ~ ., data = data.train, importance = TRUE,      proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 4.07%
## Confusion matrix:
##      0 1  class.error
## 0 3531 1 0.0002831257
## 1  149 0 1.0000000000
```

```
varImpPlot(rf1)
```



Here, we are observing the Mean Decrease in Accuracy method to find the variable importance in Random Forest.

We can see age, ever_married, hypertension, and bmi are the most important variables in predicting the stroke.

```
#Training Error and Confusion Matrix
H=predict(rf1, data.train)
mean(H != data.train$stroke)
```

```
## [1] 0.0005433306
```

```
table(H, data.train$stroke)
```

```
##
## H      0    1
##   0 3532    2
##   1    0  147
```

```
#Test Error and Confusion Matrix
H=predict(rf1, data.test)
mean(H != data.test$stroke)
```

```
## [1] 0.04971475
```

```
table(H, data.test$stroke)
```

```
##
## H      0    1
##   0 1166   60
##   1    1    0
```

By looking at the confusion matrix we can tell the random forest model performs very well for the training data. But for testing data we can tell the model does not do a good job, as number of false negatives are way more which hampers our accuracy. Let us try different parameters for mtry and nidesize and the select the optimum combination and try to fit the random forest.

Optimizing RF

```
NT=1:10
MT=1:10
NDT=1:50
MinError=1
MinNT=0
minNDT=0
minMT=0

for(nt in NT){
  for(mt in MT){
    for(ndt in NDT){
      rf2 = randomForest(stroke~., type="classification", data.train, ntree=nt, mtry=mt, nodesize=ndt,
      H=predict(rf2, data.train)
```

```
   Error=mean(H != data.train$stroke)
   if(Error < MinError){
    MinError=Error
    minNT=nt
    minNDT=ndt
    minMT=mt
    #print(c("NT=",nt," MT=",mt," NDT=",ndt," minE=",MinError))
   }
  }
 }

}
print(c("NT=",minNT," MT=",minMT," NDT=",minNDT," minE=",MinError))
```

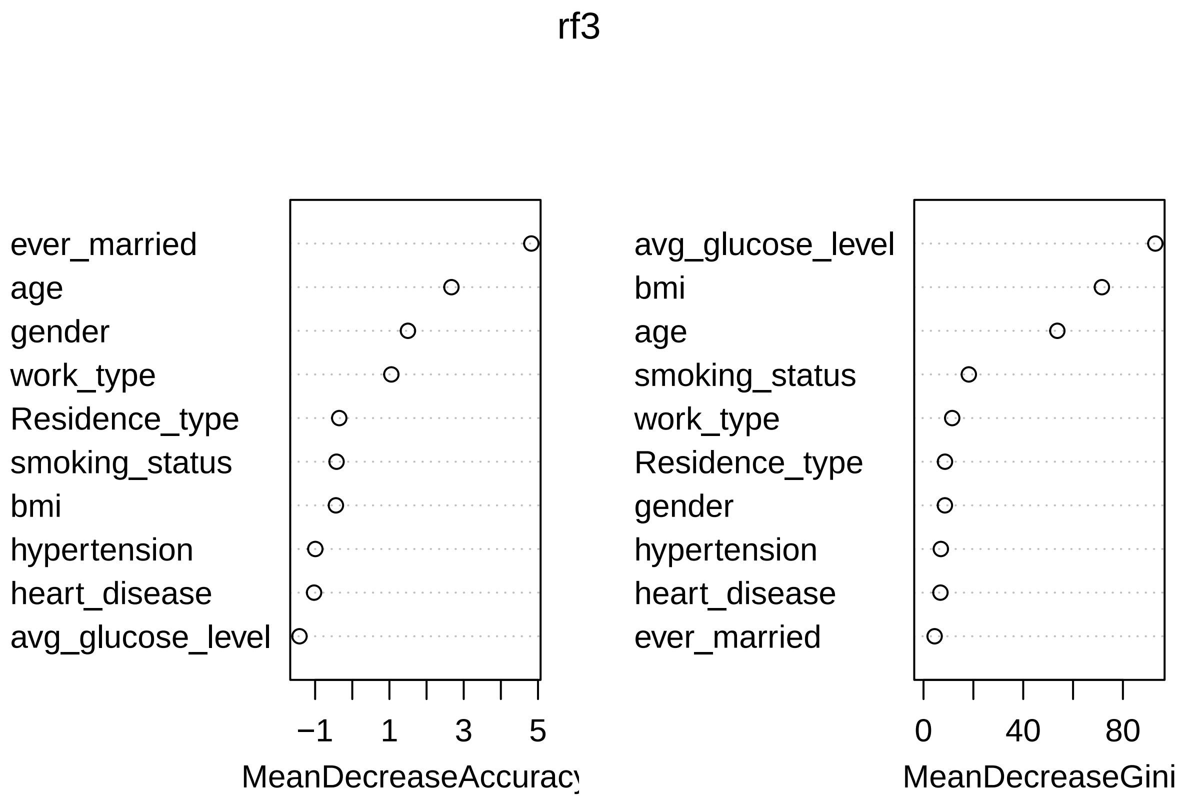
```
## [1] "NT="                 "9"                   " MT="
## [4] "10"                  " NDT="               "1"
## [7] " minE="              "0.00380331431676175"
```

```
rf3 = randomForest(stroke~., type="classification", data.train, ntree=minNT, mtry=minMT, nodesize=minND

importance(rf3)
```

```
##                            0          1 MeanDecreaseAccuracy MeanDecreaseGini
## gender             1.2518368  0.1464756            1.4980113         8.585638
## age                1.5274517  4.3506019            2.6689755        53.694903
## hypertension      -1.5204171  0.8097672           -0.9946378         6.959610
## heart_disease     -1.0928831 -0.3410342           -1.0285175         6.793541
## ever_married       4.9578090 -2.0392456            4.8200208         4.463466
## work_type          0.8217926  0.9542695            1.0517364        11.513508
## Residence_type    -0.1422969 -0.4387491           -0.3506602         8.610625
## avg_glucose_level -2.2874884  2.1064854           -1.4204275        93.005115
## bmi               -0.8613293  2.1977628           -0.4407555        71.565990
## smoking_status    -0.2619711 -0.5563273           -0.4232517        18.179272
```

```
varImpPlot(rf3)
```



As expected we got age, ever_married as the important variables but surprisingly we got gender and work_type as the important variables as well. Let us find the confusion matrix for this random forest fit on training and testing data.

```
#Training Error and Confusion Matrix
H=predict(rf3, data.train)
mean(H != data.train$stroke)
```

```
## [1] 0.00407498
```

```
table(H, data.train$stroke)
```

```
##
## H      0    1
##   0 3532   15
##   1    0  134
```

```
#Test Error and Confusion Matrix
H=predict(rf3, data.test)
mean(H != data.test$stroke)
```

```
## [1] 0.05378973
```

```
table(H, data.test$stroke)
```

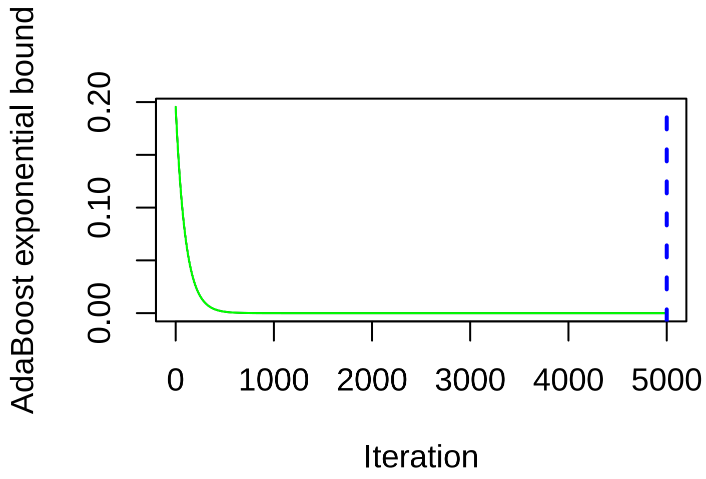
```
##
## H       0    1
##    0 1160   59
##    1    7    1
```

We can see the improvement of results compared to the previous random forest implementation but still the accuracy is not getting increased too much. Let us try Boosting.

2.3 Boosting

```
gbm.CVA <- gbm(stroke~., data=data.train, distribution = 'adaboost', n.trees = 5000, shrinkage = 0.01,

perf_gbm1 = gbm.perf(gbm.CVA, method="cv")
```



```
summary(gbm.CVA)
```

```
##                                   var      rel.inf
## age                               age 7.125743e+01
## avg_glucose_level avg_glucose_level 2.267121e+01
## smoking_status         smoking_status 2.245676e+00
## bmi                               bmi 1.651843e+00
## hypertension             hypertension 8.619417e-01
## heart_disease           heart_disease 7.406397e-01
## work_type                   work_type 5.097032e-01
## ever_married             ever_married 6.155014e-02
## Residence_type         Residence_type 2.351112e-06
## gender                         gender 2.539075e-09
```

In boosting, we can concur our results that age, hypertension, ever_married are the important variables in predicting stroke.

Let us check the training and testing accuracy on the boosting model.

```r
## Training error
y1=data.train$stroke
pred1gbm <- predict(gbm.CVA,newdata = data.train, n.trees=perf_gbm1, type="response")
y1hat <- ifelse(pred1gbm < 0.5, 0, 1)
sum(y1hat != y1)/length(y1)
```

```
## [1] 0.9595219
```

```r
## Testing error
y2=data.test$stroke
y2hat <- ifelse(predict(gbm.CVA,newdata = data.test,
n.trees=perf_gbm1, type="response") < 0.5, 0, 1)
mean(y2hat != y2)
```

```
## [1] 0.9511002
```

We can observe a good training accuracy of 96% and testing accuracy of 95%. Which means the gradient boosting model predicts whether a person suffers from stroke or not almost 95 times in 100 samples.

3 Conclusion & Discussion

We looked at some of the elements that might lead to stroke in our project. Age was significantly associated, followed by hypertension, heart_disease, avg_glucose_level, and whether or not they had ever married. There were certain outliers in the prediction too. As, even though a person's BMI is high, he or she will not have a stroke if they are young and have no heart problems. So boosting works very well in such kind of scenarios as it makes smaller simpler trees which relate the explanatory variables in a more granular level, which benefits in prediction. Furthermore we can try using more sophisticated models, such as deep neural networks, to see if it improves our results. Finally, all of these arguments are predicated on the dataset we were given. The accuracy of this model will alter when additional data becomes available in the future, so we may need to fine-tune it later.