Stroke Prediction Data

By Shelby Belak, Christopher Birsner, Ryan Dean, Jamie Herren, and

Thomas Marianos

March 29, 2021

One of the leading causes of death in the U.S. is strokes, with an individual dying every four minutes from the medical emergency. It is also a leading cause of serious long-term disability as it reduces the mobility in more than 50 percent of survivors 65 and older. Because of how frequently strokes affect many people in the country today, it is important to look at some of the warning signs that could help catch strokes before they happen. As data scientists, we want to be able to look at the information we have in front of us and help the health care system with understanding the overall perspective of this serious health issue. That is why we are using a stroke prediction dataset to see what vital knowledge we can pull from it to help institutions like hospitals know what to look out for.

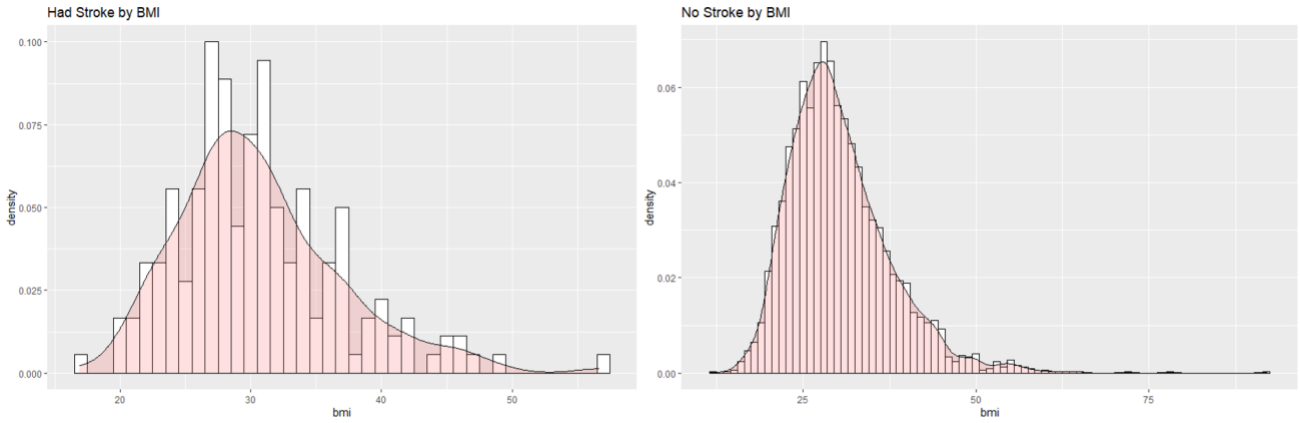
By evaluating this data, we could assist healthcare systems with helping them make business decisions in areas such as creating educational programs on strokes, investing in staff, equipment, and infrastructure for hospital-based programs, and establishing protocols for both individuals who fall into a risk group and immediate care for those who are actively experiencing symptoms. Our goal for this project was to gather and present basic descriptive information to get an initial understanding of who is most at-risk for having a stroke. We also created visualizations and mapped data to make clear the information that needed to be presented.

Age	Stats
Mean	42.87
Median	44
Max	82
Min	.08
SD	22.56
Quantile, .05	4
Quantile, .95	79
Skewness	-0.12

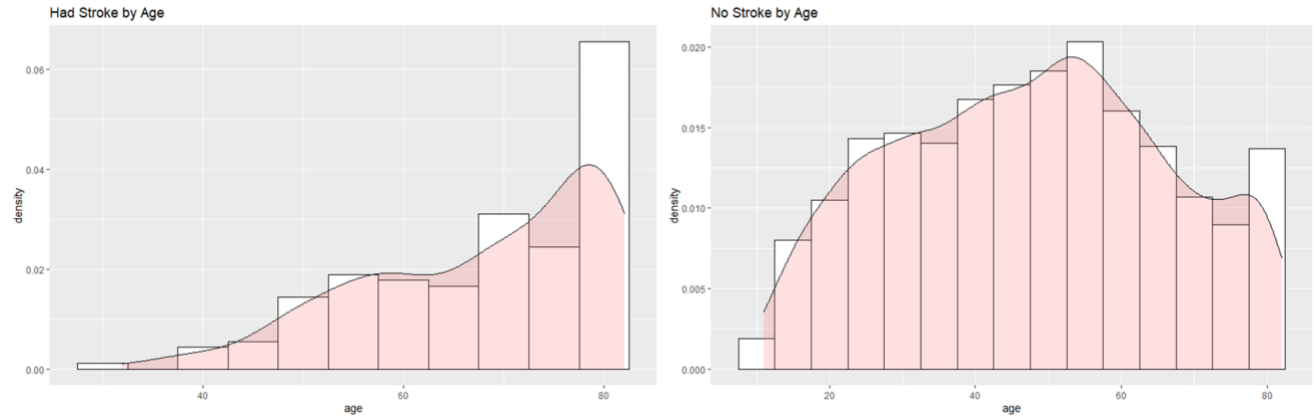
Glucose	Stats
Mean	105.3
Median	91.7
Max	271.74
Min	55.12
SD	44.42
Quantile, .05	60.61
Quantile, .95	214.64
Skewness	1.61

BMI	Stats
Mean	28.89
Median	28.1
Max	97.6
Min	10.3
SD	7.85
Quantile, .05	17.64
Quantile, .95	42.96
Skewness	1.055

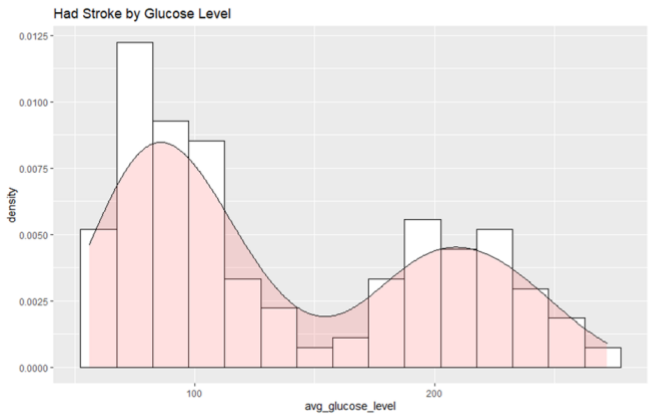
The initial data was provided by 5,111 patients who agreed to participate in this study. The team ran some initial data statistics to familiarize ourselves with the population. We looked at some general breakdowns of the variables of age, glucose, and BMI as starters. Based on those initial breakdowns, we performed some data munging to create a cleaner set. This entailed removing patient sets for those younger than 10 years of age as well as omitting any sets of values that contained unknowns. Of the subjects studied, the average age of the individuals was around 42-43, with ages ranging from under one years old to 82 years old. The average glucose level was 105.3 and the average BMI was 28.89.

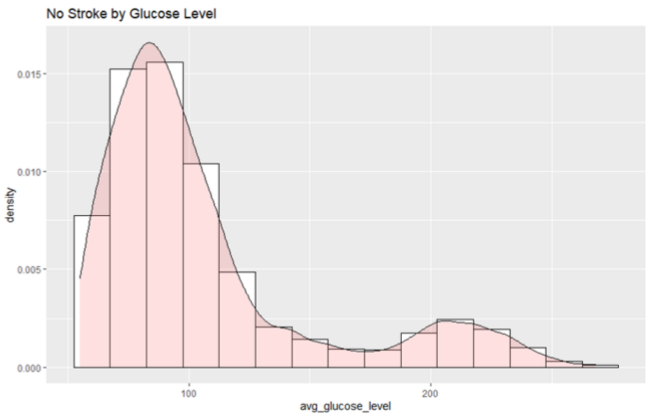


The most important focus for showcasing the results of this dataset was to create visuals that health care systems could understand when trying to understand what it is they should look for and relay to potential patients. Graphs were developed for some of these initial statistics; with an additional breakdown between patients that had/had not had strokes. This gave us additional insight as to what factors may play into potential strokes. Looking at BMI, we saw that the population that had suffered from strokes was much denser in the higher BMI range.



We also set up a histogram that shows those patients who had a stroke and a distribution and density of their age. The distribution lean heavily to the higher-age bracket, which can help us infer age is a major factor in stroke. As for the histogram that shows all patients in the dataset that did not have a stroke, it is relatively evenly distributed.





When looking at glucose levels in these patients, we saw that the graph for those who did have a stroke had a bimodal distribution. That means it is telling us that there are two different groups, which can help us infer that an individual's glucose level could be too high or too low, both of which may result in a stroke. This distribution is not nearly as bimodal for those that haven’t had a stroke but does have an obvious skew with most having a lower glucose level. Glucose level is affected by the last time an individual ate, so these results will vary by patient.

```
Call:
lm(formula = stroke ~ work_type + smoking_status + Residence_type +
    hypertension + heart_disease + gender + ever_married + bmi +
    avg_glucose_level + age, data = Stroke)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90360 -0.28386 -0.04767  0.33026  0.98307

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.2216525  0.1561841  -1.419  0.15645
work_typeGovt_job          -0.1158750  0.1561925  -0.742  0.45850
work_typeNever_worked      -0.0375996  0.2692451  -0.140  0.88899
work_typePrivate           -0.1159130  0.1481924  -0.782  0.43447
work_typeSelf-employed     -0.1535930  0.1559083  -0.985  0.32501
smoking_statusnever smoked  0.0334559  0.0422945   0.791  0.42929
smoking_statussmokes        0.0636707  0.0512919   1.241  0.21505
Residence_typeUrban         0.0228316  0.0352749   0.647  0.51776
hypertension1               0.1265957  0.0485793   2.606  0.00943 **
heart_disease1              0.1126205  0.0622397   1.809  0.07096 .
genderMale                 -0.0216844  0.0368719  -0.588  0.55672
ever_marriedYes            -0.0466269  0.0504895  -0.923  0.35618
bmi                        -0.0042612  0.0026979  -1.579  0.11484
avg_glucose_level           0.0011302  0.0003546   3.187  0.00152 **
age                         0.0118379  0.0012281   9.639  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4029 on 515 degrees of freedom
Multiple R-squared:  0.2966,    Adjusted R-squared:  0.2774
F-statistic: 15.51 on 14 and 515 DF,  p-value: < 2.2e-16
```

original	pred 0	1
0	78	9
1	32	14

The next step in our process was to create a few models to help us understand the relationship between certain attributes with whether someone will have a stroke. Our first is a linear model that tells us that the significant variables in predicting whether someone will have a stroke are hypertension, average glucose level, and age, with the most significant being age. According to the model, as all three of these variables increase, there is a more likely chance that the person will have a stroke. The Adjusted R-squared tells us that about 27 percent of potential strokes can be explained by the linear model. When we compare the predicted values with the testing data values, we can see that we have about a 69 percent accuracy when predicting whether a patient will have a stroke or not with this linear model.

```
Call:
glm(formula = stroke ~ work_type + smoking_status + Residence_type +
    hypertension + heart_disease + gender + ever_married + bmi +
    avg_glucose_level + age, family = binomial(probit), data = Stroke)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.7357  -0.3305   0.8198   2.4637

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.163583 123.119568  -0.050   0.9601
work_typeGovt_job           2.916666 123.119671   0.024   0.9811
work_typeNever_worked      -0.072085 249.245438   0.000   0.9998
work_typePrivate            2.860465 123.119561   0.023   0.9815
work_typeSelf-employed      2.727831 123.119683   0.022   0.9823
smoking_statusnever smoked  0.098464   0.153073   0.643   0.5201
smoking_statussmokes        0.241535   0.185960   1.299   0.1940
Residence_typeUrban         0.100485   0.130282   0.771   0.4405
hypertension1               0.374601   0.162518   2.305   0.0212 *
heart_disease1              0.288456   0.209728   1.375   0.1690
genderMale                 -0.079363   0.136456  -0.582   0.5608
ever_marriedYes             0.018802   0.206153   0.091   0.9273
bmi                        -0.008377   0.010140  -0.826   0.4087
avg_glucose_level           0.003302   0.001243   2.656   0.0079 **
age                         0.043066   0.004984   8.640   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 679.23  on 529  degrees of freedom
Residual deviance: 497.42  on 515  degrees of freedom
AIC: 527.42

Number of Fisher Scoring iterations: 14
```

The next model we created was a probit regression, which looks at the variables and is more sensitive to the outliers of the data. Our probit model also tells us that the significant variables in predicting whether someone will have a stroke. Once again, the variables that are the most significant are hypertension, average glucose level, and age, with the most significant being age. This is the second model to indicate to us that, as all three of these variables increase, there is a more likely chance that the person will have a stroke.

```
Call:
glm(formula = stroke ~ age + hypertension + avg_glucose_level,
    family = binomial(logit), data = Stroke)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8643  -0.7185  -0.3548   0.8550   2.4766

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.755288   0.538371 -10.690  < 2e-16 ***
age                0.072936   0.007957   9.166  < 2e-16 ***
hypertension1      0.600107   0.265524   2.260  0.02382 *
avg_glucose_level  0.005282   0.001904   2.774  0.00554 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 679.23  on 529  degrees of freedom
Residual deviance: 505.89  on 526  degrees of freedom
AIC: 513.89

Number of Fisher Scoring iterations: 5
```

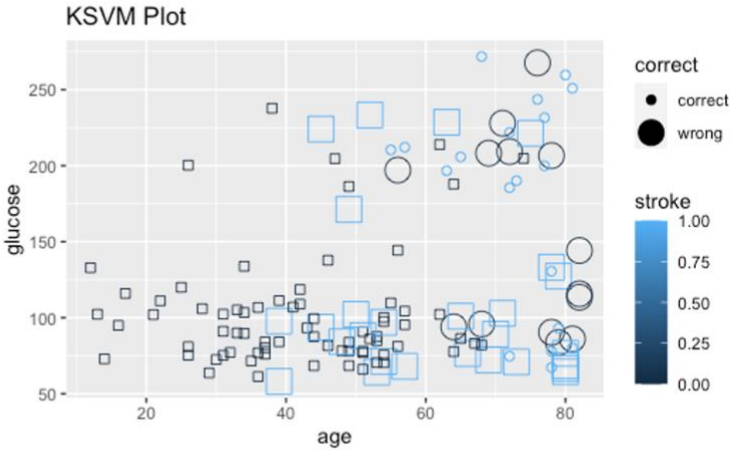
NO MULTICOLLINEARITY

```
> vif <- vif(Stroke.Logit)
> vif
              age      hypertension avg_glucose_level
         1.027651          1.026281          1.016764
```

We then looked at the logit regression, a model that looks at the variables and is less sensitive to the outliers of the data compared to the other models. We once again ran a logit model to confirm the variables with the most significant impact on whether someone has a stroke: hypertension, average glucose level, and age. We then re-ran the logit with only these significant variables and found them to be highly significant, again with age and average glucose levels appearing to have the biggest impact. By looking at each variable‘s variance inflation factor, we found that there is also no multicollinearity between the variables we kept. This tells us that the coefficients are unrelated to each other and thus are reliable.

```
Support Vector Machine object of class "ksvm"

SV type: eps-svr  (regression)
 parameter : epsilon = 0.1  cost C = 10

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.168080975220998

Number of Support Vectors : 315

Objective Function Value : -1183.089
Training error : 0.350275
Cross validation error : 0.251091
Laplace distr. width : 0.340009
```

original	pred 0	1
0	73	14
1	27	19



The last thing we wanted to do for our research is dive deep into the dataset using data mining techniques. We decided to run a KSVM model to predict the likelihood of stroke when factoring in age and glucose levels. We started out by using a training data set. The results show that the training error on the model was about 0.35, which meant the model is about 65 percent accurate. Then, we ran the same model using our testing data set and it turns out that our predictions were correct about 70 percent of the time.

Through our research and data analysis, our conclusions can help health care systems with getting important information out to their patients about the potential warning signs of a stroke, especially older patients or those with unhealthy glucose levels. This data can also be a gateway to future studies that can further assist the effort in preventing strokes before they happen.

The Code

```
#Code for cleaning the dataset

library("readxl")

library("fastDummies")

library("dplR")

library("tidyverse")

 #reading in orginal excel file

stroke <- read_excel("C:\\Syracuse DataScience\\IST 687 Applied DataScience\\Project Folder\\healthcare.xlsx")

str(stroke)

#omit all N/A's from Dataset

stroke <- na.omit(stroke)

#subsetting the age column

stroke <- subset(stroke, age>10)

#convert bmi to numeric

stroke <- stroke[- grep("N/A", stroke$bmi),]

stroke$bmi <- as.numeric(stroke$bmi)

#removing unknowns from smoking column

stroke <- stroke[- grep("Unknown", stroke$smoking_status),]
```

```
view(stroke)

#dummy variable formulas

stroke$gen_dummies <- dummy_cols(stroke$gender)

stroke$married_dummies <- dummy_cols(stroke$ever_married)

stroke$worktype_dummies <- dummy_cols(stroke$work_type)

stroke$residence_dummies <- dummy_cols(stroke$Residence_type)

stroke$smoking_dummies <- dummy_cols(stroke$smoking_status)

#removing redundant columns

stroke <-

  stroke %>%

  select(-id, -gender, -ever_married, -work_type, -Residence_type, -smoking_status)

#------------------

#Code for the descriptive statistics and graphs

library("readxl")

library("dplR")

library("tidyverse")

library("ggplot2")

library("moments")
```

```
#Data set recorded patients admitted to hospital

#reading in orginal excel file

stroke <- read_excel("C:\\Syracuse DataScience\\IST 687 Applied DataScience\\Project Folder\\healthcare.xlsx")

str(stroke)

Stroke <- stroke

#omit all N/A's from Dataset

Stroke <- na.omit(Stroke)

#subsetting the age column - Since strokes at such a young age are unlikely(dataset includes newborns)

Stroke <- subset(Stroke, age>10)

#convert bmi to numeric

Stroke <- Stroke[- grep("N/A", Stroke$bmi),]

Stroke$bmi <- as.numeric(Stroke$bmi)

#removing unknowns from smoking column

Stroke <- Stroke[- grep("Unknown", Stroke$smoking_status),]

#subsetting data set by stroke/no stroke

hadstroke <- subset(Stroke, stroke == 1)

nostroke <- subset(Stroke, stroke == 0)
```

```
#agehistograms had a stroke

hist_1 <- ggplot(hadstroke, aes(x=age)) + geom_histogram(binwidth=1)

hist_1 <- hist_1 + geom_histogram(color="black", fill="white")

hist_1

Had_Stroke.Age <- ggplot(hadstroke, aes(x=age)) + ggtitle("Had Stroke by Age") +

geom_histogram(aes(y=..density..), color="black", fill="white", binwidth = 5) +

geom_density(alpha=.2, fill="#FF6666")

Had_Stroke.Age

#age histograms no stroke

hist_2 <- ggplot(nostroke, aes(x=age)) + geom_histogram(binwidth=1)

hist_2 <- hist_2 + geom_histogram(color="black", fill="white")

No_Stroke.Age <- ggplot(nostroke, aes(x=age)) + ggtitle("No Stroke by Age") +

geom_histogram(aes(y=..density..), color="black", fill="white", binwidth = 5) +

geom_density(alpha=.2, fill="#FF6666")

No_Stroke.Age

#glucose level

hist_3 <- ggplot(hadstroke, aes(x=avg_glucose_level)) + geom_histogram(binwidth=1)

hist_3 <- hist_3 + geom_histogram(color="black", fill="white")
```

```
Had_Stroke.Glucose <- ggplot(hadstroke, aes(x=avg_glucose_level)) + ggtitle("Had Stroke by Glucose Level") + geom_histogram(aes(y=..density..), color="black", fill="white", binwidth = 15) + geom_density(alpha=.2, fill="#FF6666")

Had_Stroke.Glucose

#glucose level no stroke

hist_4 <- ggplot(nostroke, aes(x=avg_glucose_level)) + geom_histogram(binwidth=1)

hist_4 <- hist_4 + geom_histogram(color="black", fill="white")

No_Stroke.Glucose <- ggplot(nostroke, aes(x=avg_glucose_level)) + ggtitle("No Stroke by Glucose Level") + geom_histogram(aes(y=..density..), color="black", fill="white", binwidth = 15) + geom_density(alpha=.2, fill="#FF6666")

No_Stroke.Glucose

#bmi stroke

Stroke <- stroke[-grep("N/A", stroke$bmi),]

Stroke$bmi <- as.numeric(Stroke$bmi)

hist_4 <- ggplot(hadstroke, aes(x=bmi)) + geom_histogram(binwidth=1)

hist_4 <- hagehist + geom_histogram(color="black", fill="white")

Had_Stroke.BMI <- ggplot(hadstroke, aes(x=bmi)) + ggtitle("Had Stroke by BMI") + geom_histogram(aes(y=..density..), color="black", fill="white", binwidth = 1) + geom_density(alpha=.2, fill="#FF6666")

Had_Stroke.BMI
```

```
#bmi no stroke

hist_5 <- ggplot(nostroke, aes(x=bmi)) + geom_bar()

hist_5 <- hist_5 + geom_histogram(color="black", fill="white")

No_Stroke.BMI <- ggplot(nostroke, aes(x=bmi)) + ggtitle("No Stroke by BMI") +

geom_histogram(aes(y=..density..), color="black", fill="white", binwidth = 1) +

geom_density(alpha=.2, fill="#FF6666")

No_Stroke.BMI

printVecInfo <- function(x){

#summarizing the distribution

a <- mean(x)

b <- median(x)

c <- max(x)

d <- min(x)

e <- sd(x)

f <- quantile(x,0.05)

g <- quantile(x,0.95)

h <- skewness(x)

cat("mean:",a,"\nmedian:", b, "\nmax:",c, "\nmin:", d, "\nstandard dev. :", e, "\nq 5%:", f, "\nq

95%:", g, "\nskewness:", h)
```

```
}

printVecInfo(Stroke$age)

printVecInfo(Stroke$avg_glucose_level)

printVecInfo(Stroke$bmi)

#------------

#Code for each of our models

install.packages("tidyverse")

install.packages("car")

library("car")

library("readxl")

library("dplR")

library("tidyverse")

library("ggplot2")

library("moments")

library("neuralnet")

library("arulesViz")

library("kernlab")

library("e1071")
```

```
library("gridExtra")

library("caret")

library("arules")

library("cowplot")

#reading in orginal excel file

strokedesc <- read_excel("/Users/thomasmarianos/OneDrive - Syracuse University/Grad School/IST 687/Project/healthcare-dataset-stroke-data.xls")

str(stroke)

stroke <- strokedesc

#omit all N/A's from Dataset

stroke <- na.omit(stroke)

#subsetting the age column

stroke <- subset(stroke, age>10)

#convert bmi to numeric

stroke <- stroke[- grep("N/A", stroke$bmi),]

stroke$bmi <- as.numeric(stroke$bmi)

#removing unknowns from smoking column

stroke <- stroke[- grep("Unknown", stroke$smoking_status),]

#subsetting the dataset in order to generate a random sample
```

```
hadstroke <- subset(stroke, stroke == 1)

nostroke <- subset(stroke, stroke == 0)

#pulling random samples from nostroke in order to make comparable to hadstroke to run
machine learning models

SampleNoStroke <- nostroke[sample(nrow(nostroke), 350),]

#combining the two dataframes

Stroke <- rbind(SampleNoStroke, hadstroke)

Stroke <- Stroke[,-1]

#randomizing the Stroke dataframe

set.seed(45)

rows <- sample(nrow(Stroke))

Stroke <- Stroke[rows,]

#Getting Ready for Regression

#Changing Catergorical to Factors w/ levels

Stroke$gender <- factor(Stroke$gender)

Stroke$ever_married <- factor(Stroke$ever_married)

Stroke$work_type <- factor(Stroke$work_type)

Stroke$Residence_type <- factor(Stroke$Residence_type)

Stroke$smoking_status <- factor(Stroke$smoking_status)
```

```
Stroke$hypertension <- factor(Stroke$hypertension)

Stroke$heart_disease <- factor(Stroke$heart_disease)

#Changing all quantitative variables to numeric

Stroke$age <- as.numeric(Stroke$age)

Stroke$avg_glucose_level <- as.numeric(Stroke$avg_glucose_level)

Stroke$bmi <- as.numeric(Stroke$bmi)

#Train/Test Split

#train/test split

##75% of the sample size

sample_size <- floor(0.75 * nrow(Stroke))

#setting seed to reproduce partition

set.seed(123)

#subsetting training data

training_index <- sample(seq_len(nrow(Stroke)), size = sample_size)

trainingdata <- Stroke[training_index,]

train_x <- trainingdata[,(1:10)]

train_x <- as.matrix(train_x[,-1])

train_y <- trainingdata[,-(1:10)]
```

```
#subsetting testing data

testingdata <- Stroke[-training_index,]

test_x <- testingdata[,(1:10)]

test_y <- testingdata[,-(1:10)]

#LinearModel

Stroke.Linear <- lm(stroke ~ work_type + smoking_status + Residence_type +

                    hypertension + heart_disease + gender + ever_married + bmi +

                    avg_glucose_level + age, data=trainingdata)

summary(Stroke.Linear)

Pred.Linear <- predict(Stroke.Linear,testingdata)

Pred.Linear

str(Pred.Linear)

compTable.Linear <- data.frame(testingdata[,11],Pred.Linear)

colnames(compTable.Linear) <- c("test","Pred")

compTable.Linear$Pred <- ifelse(compTable.Linear$Pred<.6, 0, 1)

sqrt(mean((compTable.Linear$test-compTable.Linear$Pred)^2))

results <- table(original = compTable.Linear$test, pred = compTable.Linear$Pred)

print(results)
```

```
perc.Linear <- length(which(compTable.Linear$test ==
compTable.Linear$Pred))/dim(compTable.Linear)[1]
perc.Linear
#Probit Regression
Stroke.Probit <- glm(stroke ~ work_type + smoking_status + Residence_type +
                     hypertension + heart_disease + gender + ever_married + bmi +
                      avg_glucose_level + age, family=binomial(probit), data=Stroke)
summary(Stroke.Probit)
 #Logit Regression - Less sensitive than probit to outliers
Stroke.Logit <- glm(stroke ~ work_type + smoking_status + Residence_type +
                     hypertension + heart_disease + gender + ever_married + bmi +
                     avg_glucose_level + age, family=binomial(logit), data=Stroke)
summary(Stroke.Logit)
exp(coef(Stroke.Logit))  # Exponentiated coefficients ("odds ratios")
summary(Stroke)
Stroke.Logit <- glm(stroke ~ age + hypertension + avg_glucose_level, family=binomial(logit),
data=Stroke)
summary(Stroke.Logit)
#Variance Inflation Factor Shows no multicollinearity between variables we kept
```

```
vif <- vif(Stroke.Logit)

vif

#ksvm

Strokeksvm <- ksvm(stroke~., data = trainingdata, kernel = "rbfdot", kpar="automatic",

C=10,cross=10, prob.model=TRUE)

Strokeksvm

ksvm.pred <- predict(Strokeksvm, testingdata)

head(ksvm.pred)

#building a dataframe to compare prediction vs. actual

compare_ksvm <- data.frame(test_y, ksvm.pred)

head(compare_ksvm)

colnames(compare_ksvm) <- c("test", "pred")

compare_ksvm$pred <- ifelse(compare_ksvm$pred<.6, 0, 1)

tail(compare_ksvm)

percent_ksvm <-

length(which(compare_ksvm$test==compare_ksvm$pred))/dim(compare_ksvm)[1]

percent_ksvm

results <- table(original = compare_ksvm$test, pred = compare_ksvm$pred)

print(results)
```

```
# Plot the results

compare_ksvm$correct <- ifelse(compare_ksvm$test == compare_ksvm$pred, "correct",

"wrong")

df.ksvm <- data.frame(compare_ksvm$correct, testingdata$age, testingdata$avg_glucose_level,

testingdata$stroke, compare_ksvm$pred)

colnames(df.ksvm) <- c("correct", "age", "glucose", "stroke", "pred")

plot.ksvm <- ggplot(df.ksvm, aes(x = age,y = glucose)) +

  geom_point(aes(size = correct, color = stroke)) +

  scale_shape_identity() +

  ggtitle("KSVM Plot")

plot.ksvm

#----------------
```

References

1.Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.

2.Stroke Prediction Data. (n.d.). Kaggle. Retrieved February 11, 2021, from Stroke Prediction Dataset | Kaggle

3.Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Associationexternal icon. *Circulation.* 2020;141(9):e139–e596.