Stroke Prediction Using Machine Learning

Executive Summary

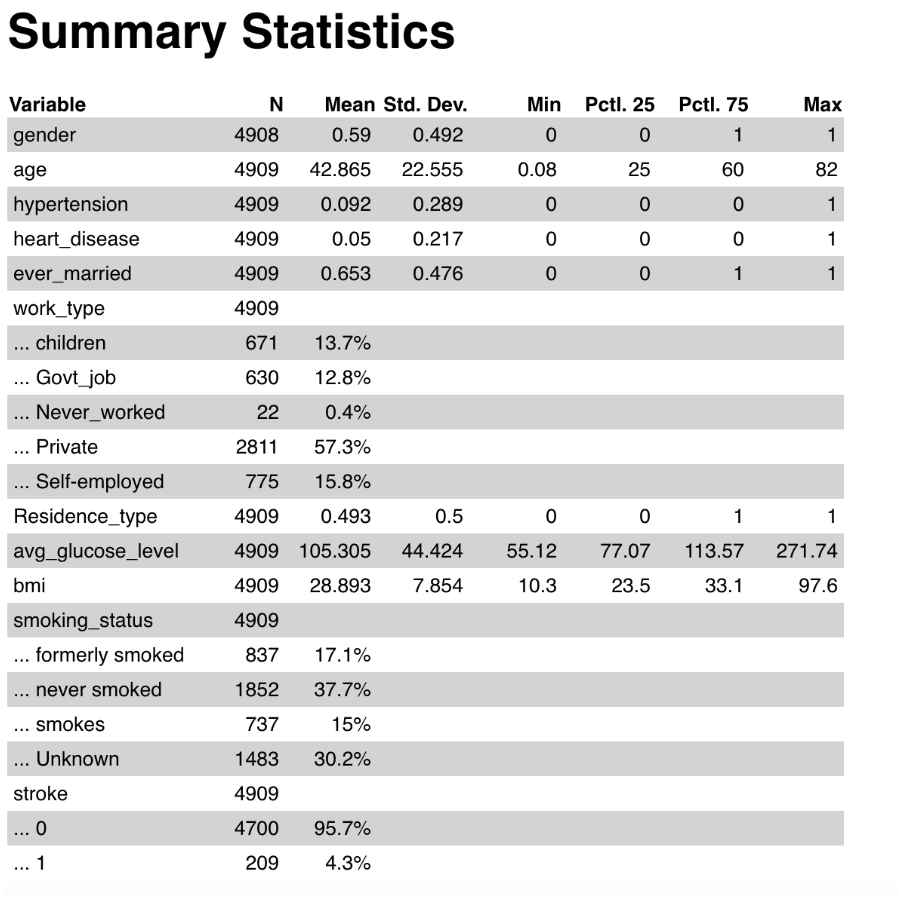
Strokes are the leading cause of disability and death, with approximately 5.5 million people dying every year. There is a need for better management and patient care due to the adverse effects of a stroke. With the increased usage of machine learning in the medical field, healthcare professionals can create opportunities for a better way of patient care and management. By gathering information through large databases through data exploration I can examine the relationships between variables through data mining. In this case, I used the two on the patient's health care records (EHR), I can create models that, in turn, help with early diagnosis. Using various techniques and predictive models along with clustering, I evaluate which risk factors from those recorded in one's EHR can help us create an accurate model for stroke prediction. I created different models to finalize that age, average glucose level, hypertension, and heart disease are all indicators in someone's EHR that are correlated with strokes. Research shows that most of the variables above are caused by an unhealthy lifestyle that includes higher weight and stress. I created a GLM, a decision tree, and a Random Forest model and conducted multiple tests isolating different variables in order to create the most accurate predictive model I could. Our analysis found that our Random Forest out-of-sample model had the highest accuracy rate of 83.6% and is, therefore, our most successful predictive model. I were able to conclude, using a balanced dataset, that an unhealthy lifestyle, overeating, smoking, and our relationships can, ultimately, lead to strokes.

Introduction

I have seen the tremendous help data, and data analytics can provide to businesses in many industries. What if I brought these skills to the healthcare industry to evaluate trends in datasets that could potentially save lives? With our data set, I can use descriptive and predictive analytics to identify trends in one’s lifestyle and past medical history to identify the key risk factors for strokes by simply analyzing their electronic health record. By doing this, I are allowing healthcare practitioners to evaluate diseases differently and are creating opportunities for discovering the onset of disease early on. In this paper, I examine the dataset, identify our key risk factors, and build predictive models using various strategies like GLM, decision trees, and random forest to conclude on the best predictive model for strokes, along with discovering how a healthy lifestyle can affect the chances of one having a stroke.

Description of data

I used a data set of electronic health records released by McKinsey & Company that is available on Kaggle. Our input or x-variables include age, gender, work type, residency type, marital status, hypertension, heart disease, average glucose level, and BMI. Our output is a binary column that tells whether that person has had a stroke or not. The dataset includes 5110 data points and, as I mentioned, 12 attributes.



Summary Statistics

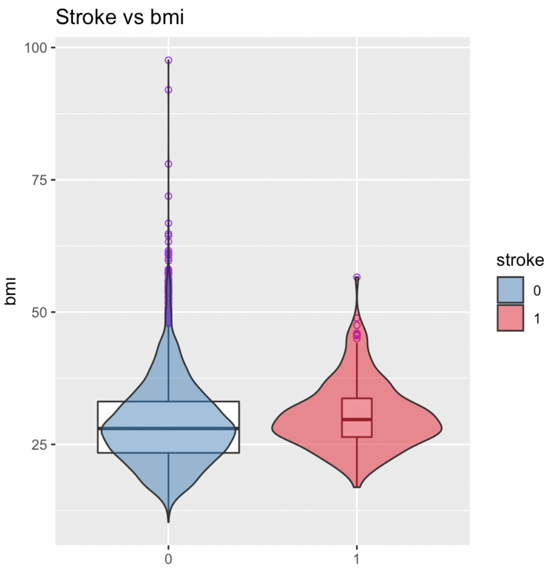
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
gender	4908	0.59	0.492	0	0	1	1
age	4909	42.865	22.555	0.08	25	60	82
hypertension	4909	0.092	0.289	0	0	0	1
heart_disease	4909	0.05	0.217	0	0	0	1
ever_married	4909	0.653	0.476	0	0	1	1
work_type	4909						
... children	671	13.7%					
... Govt_job	630	12.8%					
... Never_worked	22	0.4%					
... Private	2811	57.3%					
... Self-employed	775	15.8%					
Residence_type	4909	0.493	0.5	0	0	1	1
avg_glucose_level	4909	105.305	44.424	55.12	77.07	113.57	271.74
bmi	4909	28.893	7.854	10.3	23.5	33.1	97.6
smoking_status	4909						
... formerly smoked	837	17.1%					
... never smoked	1852	37.7%					
... smokes	737	15%					
... Unknown	1483	30.2%					
stroke	4909						
... 0	4700	95.7%					
... 1	209	4.3%					

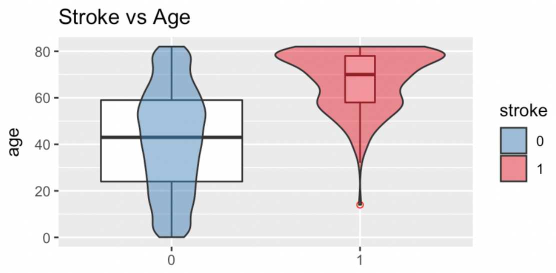
Exploratory Data Analysis

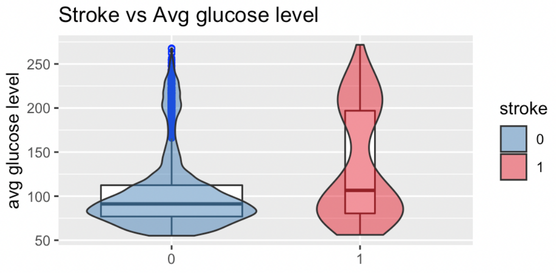
Univariate Analysis

Our next step was understanding the data by looking into each variable separately against a stroke to evaluate the relationships between each further. Figure 2 includes the graphs of age, BMI, and average glucose level against if one has had a stroke or not. On our left side, we have BMI vs. Stroke. As we can tell, while the boxplot value is similar, the range of patients with BMI not inside the central area is more than those who have not had a stroke. On the other hand, those with a stroke all have very similar levels of BMI with little outliers. This tells us that if one has a BMI in the range of roughly 40-50, they have a higher risk of getting a stroke.

In the top right corner, we see Age vs. Stroke. As we see, the difference in y between the two boxplots is staggering and has made us come to the apparent conclusion that those above 45 have a much higher risk of having a stroke compared to younger ones. And finally, on the bottom, we see the Average Glucose level vs. Stroke. Based on the shape of the pink figure, I can evaluate that those with higher glucose levels ranging above 100 are most at risk for a stroke.

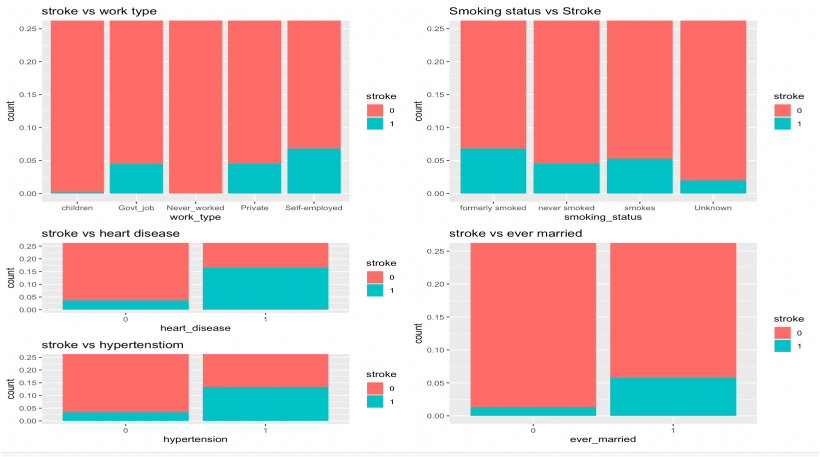






The following figure inspects the majority of our variables in relation to stroke. While I know our 4 variables that have the most substantial linear relationship with stroke, these graphs will help us compare the count and the ratio of each for every category between 0 and 1. This will let us know if any other variables may have an effect on having a stroke.

As I examine our graphs I can infer that the sections that promote an unhealthy lifestyle or have had incidents of medical stress have a high count.
From our first graph which compares the type of work vs. stroke, I can infer that those who work have a higher chance of stroke. Similarly, I can see those who were married have a higher chance of stroke as well. This is essential to note as I evaluate how stress and lifestyle may contribute to the likelihood of a stroke. Another unhealthy habit that can promote the likelihood of a stroke is smoking. As we can see, those who have formally smoked have the highest value of those who have had a stroke, followed by those who currently smoke. The final two graphs below show heart disease vs. stroke and hypertension vs. stroke. We can see a very high count in both graphs meaning that if one has hypertension or heart disease, they are more likely to have a stroke.

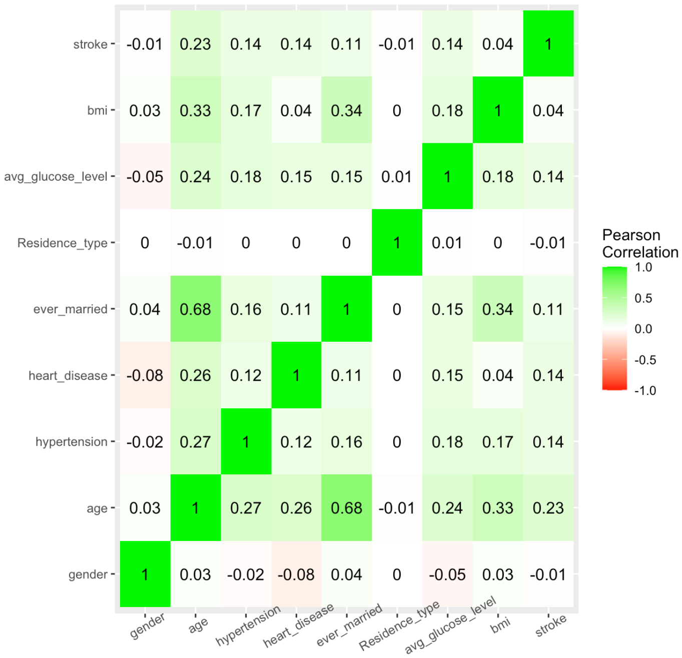


Analyzing the variables presented in the EHR

This section will analyze the relationships between the variables of the electronic health records by using correlation analysis. This will be useful to us as if features have a high correlation, I know that they relate to strokes similarly. I will also be able to look at the relationships between all the x-variables and the stroke itself to see if I have multicollinearity and determine the issues it may present.

Bivariate Analysis

As shown in Fig 1, average glucose level, heart disease, age, and hypertension had the highest correlation with stroke. I can see that age has the highest correlation, as indicated by the size and color of the circle, and is followed by hypertension, heart disease, and the average glucose level. Research shows that hypertension, heart disease, and the average glucose level all rise with age. I also know that BMI is a significant cause of hypertension, smoking can cause heart disease, and stress can often be linked to a higher glucose level. Based on this correlation plot alone, I can definitively say that age, average glucose level, hypertension, and heart disease have a significant linear relationship with heart disease. Smoking, Stress, and BMI have a substantial relationship with the 4 variables leading to strokes, reinforcing our thesis that a healthy lifestyle can prevent strokes.



	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke
stroke	-0.01	0.23	0.14	0.14	0.11	-0.01	0.14	0.04	1
bmi	0.03	0.33	0.17	0.04	0.34	0	0.18	1	0.04
avg_glucose_level	-0.05	0.24	0.18	0.15	0.15	0.01	1	0.18	0.14
Residence_type	0	-0.01	0	0	0	1	0.01	0	-0.01
ever_married	0.04	0.68	0.16	0.11	1	0	0.15	0.34	0.11
heart_disease	-0.08	0.26	0.12	1	0.11	0	0.15	0.04	0.14
hypertension	-0.02	0.27	1	0.12	0.16	0	0.18	0.17	0.14
age	0.03	1	0.27	0.26	0.68	-0.01	0.24	0.33	0.23
gender	1	0.03	-0.02	-0.08	0.04	0	-0.05	0.03	-0.01

Based on our EDA I am able to infer two things:

1. The most critical variables in relation to stroke

    a. What variables may cause these variables to become risk factors

2. Variables show proof that one's relationships and unhealthy lifestyle can cause a stroke.

The risk factors are age, hypertension, heart disease, and the average glucose level. The variables that cause these variables to become risks are smoking status and BMI. Along with that stress is another factor which I will look more into which occurs through variables work and marital status.

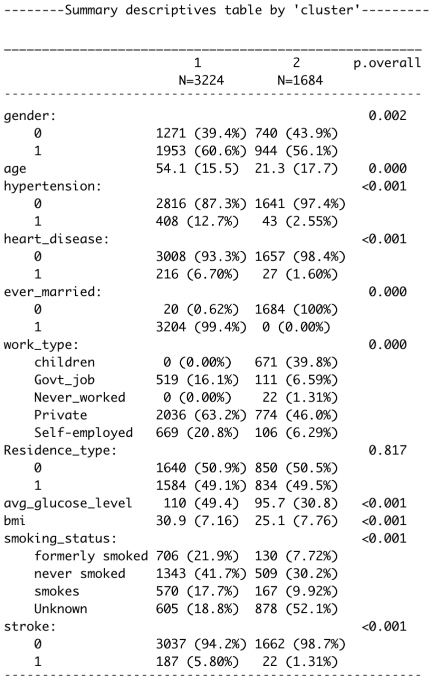
<u>Clusters</u>

Unsupervised learning can help us analyze our datasets by finding hidden patterns and data groupings. By using this method I am able to gain a deeper understanding of patient profiles and understand what can make them at high risk for a stroke.

Our final step in evaluating what variables were significant was creating clusters. K-means clustering works by partitioning n observations into k clusters, each of which belongs to the cluster with the nearest mean. As we can see below, the two clusters are grouped around the two age groups, 54 and 21, separating the older and younger patients.

I notice right away residence type's p-value is too high and, therefore, not significant. I can also conclude that gender is irrelevant as not much changes between the values.

When I look at hypertension, heart disease, and stroke, I notice the majority of their values under 1, meaning those who have had a stroke, hypertension, or heart disease are under the cluster with the high age group. This lets us know that age has a strong and notable relationship with the probability of having a stroke. I also see that the BMI and average glucose levels of those in the older age group with more strokes are higher than the younger group's BMI and AGL. With the cluster, I can solidify age's correlation with not only the risk of having a stroke but also the effect it has on other variables. The cluster helps us conclude that all the variables mentioned above have an effect on the likelihood of having a stroke. Therefore I can conclude that heart disease, hypertension, age and AGL are the key risk factors along with a high BMI, which as we know can cause hypertension.



```
--------Summary descriptives table by 'cluster'---------

_______________________________________________________
                         1            2        p.overall
                      N=3224       N=1684
-------------------------------------------------------
gender:                                          0.002
    0                1271 (39.4%) 740 (43.9%)
    1                1953 (60.6%) 944 (56.1%)
age                  54.1 (15.5)  21.3 (17.7)    0.000
hypertension:                                   <0.001
    0                2816 (87.3%) 1641 (97.4%)
    1                408 (12.7%)   43 (2.55%)
heart_disease:                                  <0.001
    0                3008 (93.3%) 1657 (98.4%)
    1                216 (6.70%)   27 (1.60%)
ever_married:                                    0.000
    0                 20 (0.62%)  1684 (100%)
    1                3204 (99.4%)  0 (0.00%)
work_type:                                       0.000
    children          0 (0.00%)   671 (39.8%)
    Govt_job         519 (16.1%)  111 (6.59%)
    Never_worked      0 (0.00%)    22 (1.31%)
    Private          2036 (63.2%) 774 (46.0%)
    Self-employed    669 (20.8%)  106 (6.29%)
Residence_type:                                  0.817
    0                1640 (50.9%) 850 (50.5%)
    1                1584 (49.1%) 834 (49.5%)
avg_glucose_level     110 (49.4)  95.7 (30.8)   <0.001
bmi                  30.9 (7.16)  25.1 (7.76)   <0.001
smoking_status:                                 <0.001
    formerly smoked  706 (21.9%)  130 (7.72%)
    never smoked     1343 (41.7%) 509 (30.2%)
    smokes           570 (17.7%)  167 (9.92%)
    Unknown          605 (18.8%)  878 (52.1%)
stroke:                                         <0.001
    0                3037 (94.2%) 1662 (98.7%)
    1                187 (5.80%)   22 (1.31%)
-------------------------------------------------------
```

What does the cluster tell us?

- As I stated in our bivariate analysis heart disease, hypertension, age and AGL are the key risk factors
    - BMI can cause hypertension and since they are related I excluded it from our top 4

How do other variables relate to heart disease, hypertension, age and AGL

The graph on the right compares smoking status vs. smoke. As we can see, those who have formally smoked have the highest value of those who have had a stroke, followed by those who currently smoke. This is important to note because, as I stated, smoking can cause heart disease, and this is proof of how smoking and leading an unhealthy lifestyle can cause a stroke.

Below I evaluate ever-married vs. stroke to see how stress and relationships can cause strokes. As we can see, those who an under the cluster with the high age group. I also see that the BMI and average glucose level of those in the older age group with more strokes are higher compared to the younger group's BMI and AGL.

With the cluster, I are able to solidify age's correlation with not only the risk of having a stroke but also the effect it has on other variables. I am also able to strengthen our claims regarding the variables heart disease, hypertension, BMI, and AGL regarding how one's lifestyle can prevent the likelihood of having a stroke.

Predictive Models

I created three predictive models to understand what risk factors in a patient's health record can be a predictor for strokes.

(i)GLM

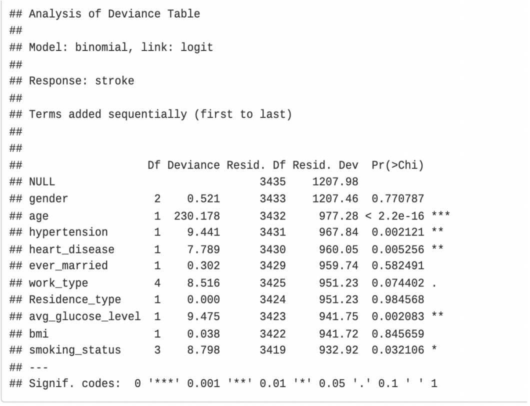
(ii)Random forest

(iv)Decision tree

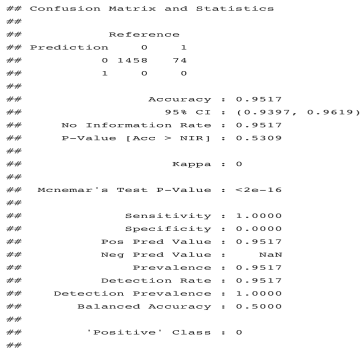
GLM

Generalized Linear Models are a flexible version of logistic regression. GLM models allow us to build a linear relationship between the output and the variables, even though the relationship between the two is not linear.

I started by splitting our dataset into two sections (70-30): testing and training. Then I created the GLM model using all the variables. Due to how unbalanced our dataset is, I can not use this model and must create a balanced dataset before I continue. Our balanced data set was created using undersampling where I reduced the dataset down to roughly 400 points in order to make sure both 0 and 1, were equal.

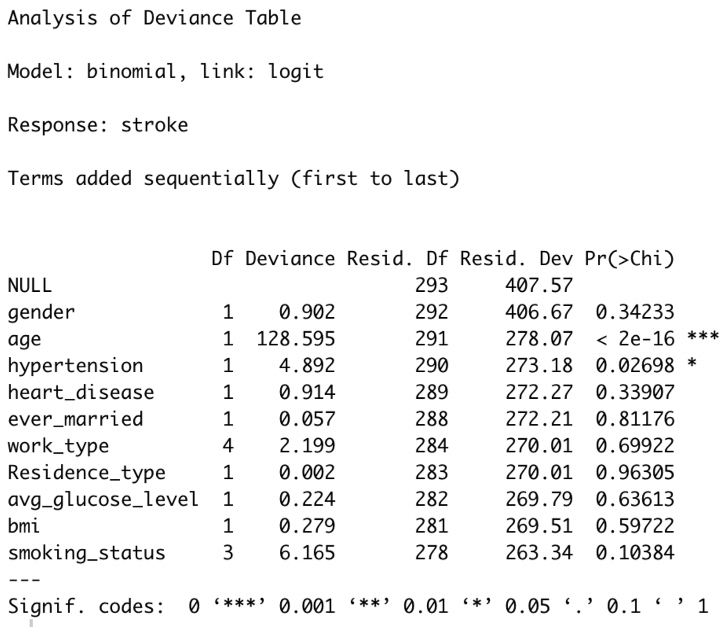


```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: stroke
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                               3435    1207.98
## gender             2    0.521      3433    1207.46  0.770787
## age                1  230.178      3432     977.28 < 2.2e-16 ***
## hypertension       1    9.441      3431     967.84  0.002121 **
## heart_disease      1    7.789      3430     960.05  0.005256 **
## ever_married       1    0.302      3429     959.74  0.582491
## work_type          4    8.516      3425     951.23  0.074402 .
## Residence_type     1    0.000      3424     951.23  0.984568
## avg_glucose_level  1    9.475      3423     941.75  0.002083 **
## bmi                1    0.038      3422     941.72  0.845659
## smoking_status     3    8.798      3419     932.92  0.032106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1458   74
##          1    0    0
##
##                Accuracy : 0.9517
##                  95% CI : (0.9397, 0.9619)
##     No Information Rate : 0.9517
##     P-Value [Acc > NIR] : 0.5309
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9517
##          Neg Pred Value :    NaN
##              Prevalence : 0.9517
##          Detection Rate : 0.9517
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

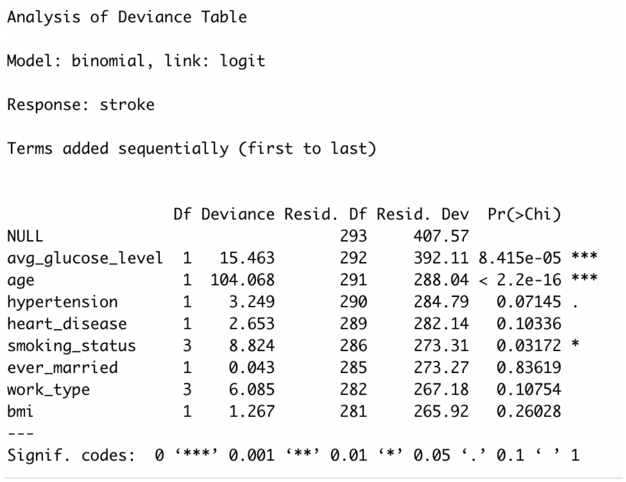
As we can see through our Anova table that was created using our balanced data set, two of our 4 risk factors, age, and hypertension, are marked as significant in this model, which includes all the variables. With our new, balanced model, I was able to achieve an accuracy rate of roughly **79.25%**.



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: stroke

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev Pr(>Chi)    
NULL                                293     407.57             
gender             1    0.902       292     406.67  0.34233    
age                1  128.595       291     278.07  < 2e-16 ***
hypertension       1    4.892       290     273.18  0.02698 *  
heart_disease      1    0.914       289     272.27  0.33907    
ever_married       1    0.057       288     272.21  0.81176    
work_type          4    2.199       284     270.01  0.69922    
Residence_type     1    0.002       283     270.01  0.96305    
avg_glucose_level  1    0.224       282     269.79  0.63613    
bmi                1    0.279       281     269.51  0.59722    
smoking_status     3    6.165       278     263.34  0.10384    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

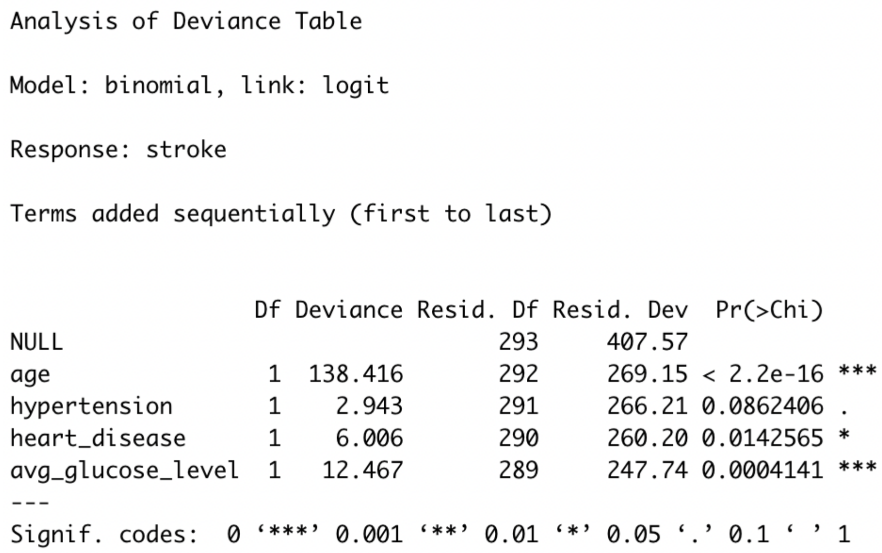
Since I want to test what variables will increase the accuracy rate, I once again run this model with all the variables, excluding gender and residence type, as they are the only ones to show little to no significance.

With our new model, I achieved an accuracy rate of roughly **79.59%**, slightly higher than our previous model, which means I chose to drop the variables bringing the accuracy rate down. I also noticed that in this model, age, average glucose level, and, interestingly enough, smoking status had significance.



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: stroke

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                293     407.57              
avg_glucose_level  1   15.463       292     392.11 8.415e-05 ***
age                1  104.068       291     288.04 < 2.2e-16 ***
hypertension       1    3.249       290     284.79   0.07145 .  
heart_disease      1    2.653       289     282.14   0.10336    
smoking_status     3    8.824       286     273.31   0.03172 *  
ever_married       1    0.043       285     273.27   0.83619    
work_type          3    6.085       282     267.18   0.10754    
bmi                1    1.267       281     265.92   0.26028    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our final model was to take our 4 most significant risk factors and rerun the model. By doing this, I achieved our highest accuracy rate of **79.97%.**



```
Analysis of Deviance Table

Model: binomial, link: logit

Response: stroke

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                               293     407.57              
age               1  138.416       292     269.15 < 2.2e-16 ***
hypertension      1    2.943       291     266.21 0.0862406 .  
heart_disease     1    6.006       290     260.20 0.0142565 *  
avg_glucose_level 1   12.467       289     247.74 0.0004141 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

So what can we conclude from this?

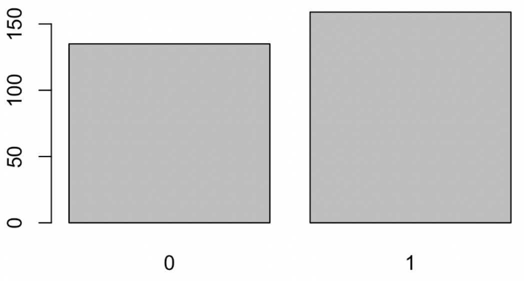
1. Age, hypertension, heart disease, and average glucose level prove to be the leading risk factors of stroke
2. Smoking status, BMI, marital status, and work type improved the accuracy rate and inadvertently affected our risk factors.

As I continue to evaluate our risk factors, I learn more about how they are impacted and what I can do to prevent strokes through our daily routines.

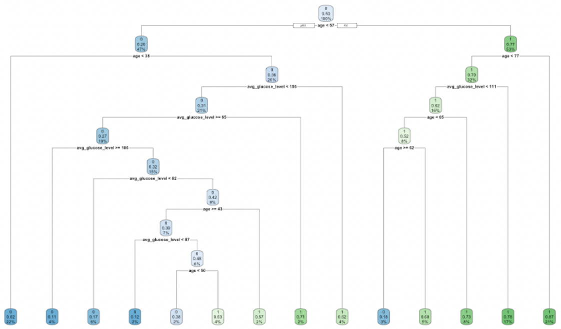
Decision Tree

Decision trees are, in my opinion, one of the best models as the concerns we generally have in regression and rf models are addressed. In this model, the data is segmented according to specific parameters. The decision tree has 2 nodes. The decision node is where the data is divided, and the leaf node is the result.

I examined the model with the top 8 variables for our first decision tree. This model provided us with an accuracy rate of **81.22%.** The model started with age, which has proven time and time over again to be our most vital correlated variable. As I follow the DT, I clearly see how the variables are connected and how each effects the other. An important observation is that the decision tree confirmed all of our theories for each variable except for heart disease. For example, I stated that if one has been married, they may have a higher chase of a stroke due to age and stress which is proven on the right side of the decision tree. As I follow all our other variables, I gain confirmation of the information I learned through our previous models. On the surface, I may be able to recognize a couple of risk factors quickly, but as I dig deeper, I realize how all these variables lead and relate to each other.



For our final model, I wanted to narrow down our variables to provide a tangible solution and conclusion. I used age, hypertension, heart disease, and average glucose levels, all of which were proven risk factors earlier on. With this model, I achieved an accuracy rate of **83.22%.** Due to the reduced number of variables, the decision tree for this model was volleying between age and average glucose level. This is due to the fact that it does not have a strong relationship with hypertension and heart disease and only average glucose levels.



What did our decision trees tell us?

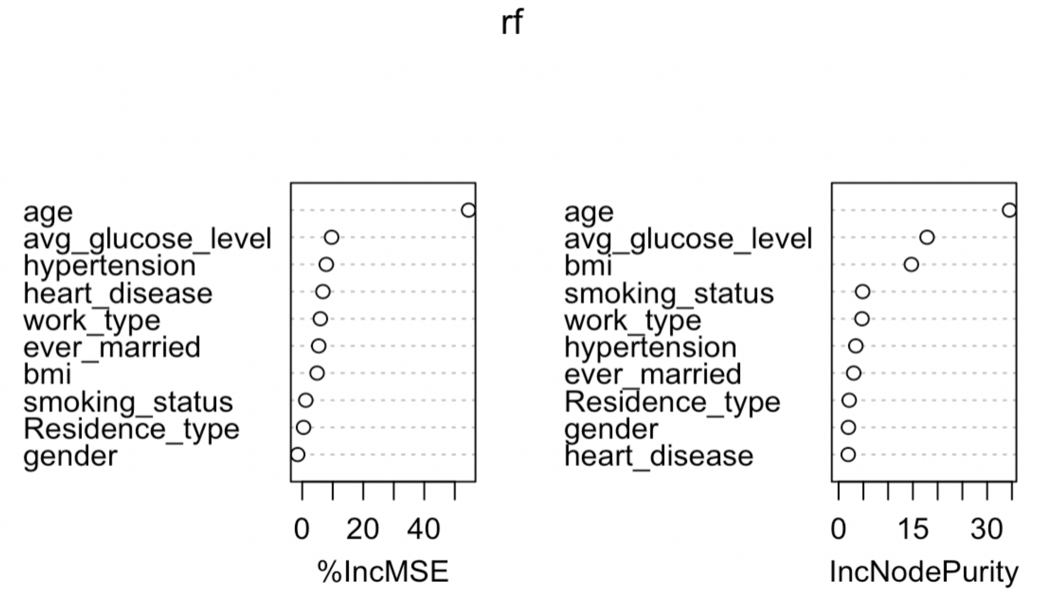
1. The DT visually showed us how each variable could lead to a stroke
   a. I knew our risk factors could, but now I know the boundaries around which BMI, work type, marital status, and smoking status can as well.

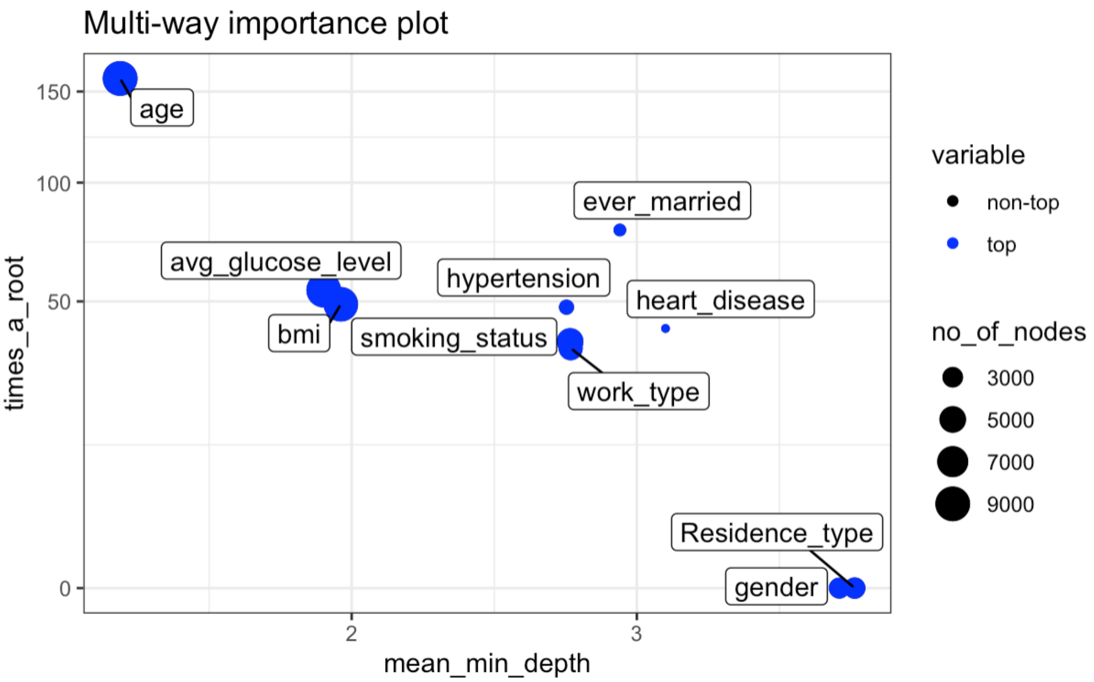
Random Forest

Random forests are numerous independent decision trees that are trained separately on a random sample of data. The decision tree outputs are collected during training, as that is when the trees are created. The final model made by RF is created by each Decision Tree voting for one of the two outputs. Then the final prediction is determined by which output has the most votes.

Our first Random Forest included all the variables to give us a benchmark from the accuracy rate I need to improve on by isolating specific variables. With our first RF, I received an accuracy rate of **70.97%.**

The figure below shows two graphs. The %INCMse shows how much our accuracy rate would decrease by removing that variable. As we can see, our top 4 risk factors lead the way for importance to the accuracy rate, followed by the 4 variables that affect our risk factors and shape our lifestyle. IncNodePurity shows the importance of each e variable and is visualized in the two graphs below. Since my goal is accuracy I will be following the IncMSE graph. Now that I have successfully narrowed down our 8 variables in importance through two models, I can continue to create more RF's to improve our accuracy rate.





Our next model has our top 8 variables. As we can see from the plot above, removing any variables would reduce our accuracy in this model. With this model, I achieved an accuracy rate of **89.92%.** While the accuracy rate was to our standards, the large number of variables did not help us specify what risk factors I had to focus on.

For a more concise model, I took the top 4 variables, also known as our risk factors, and made them into an RF model. I achieved an accuracy rate of **83.6%** with just age, hypertension, heart disease, and average glucose level. This means that just these 4 variables make up 83.6% of our accuracy while the other 4 make up roughly 6 percent.

What can we conclude from our Random Forest Models?

1. Through both models, gender and residence type are of no significance or help to our stroke prediction model
2. Age, hypertension, heart disease, and average glucose level are our leading risk factors
3. Our highest accuracy rate was achieved with our risk factors and BMI, work type, marital status, and smoking status, which means they have some significance
    a. Since BMI, work type, marital status, and smoking status are significant to the risk factors rather than the actual stroke variable, their relationship at a first glance shows up as weak

Conclusion

In this paper, I presented a detailed analysis of the variables in electronic health records to predict strokes. By using correlation models, logistic models, Random Forest, and Decision trees, I discovered that age, hypertension, average glucose level, and heart disease are the attributes that can indicate a stroke. Using these models, I was also able to conclude that BMI, marital and smoking status, and work type have relationships with our risk factors and, therefore, indirectly affect the likelihood of a stroke.

The three machine learning algorithms were implemented on a set of different features. I found that Random Forest had the highest accuracy rate, with both 8 variables and 4. The model of 4 is our best model as it has a high accuracy of 83.6% while consisting of just four variables. This is essential to note as medical professionals can predict the likelihood of a stroke occurring with just 4 variables that are noted on every patients electronic health record.

Our most important discovery was the relationships between BMI, marital and smoking status, and work type vs. risk factors. Research has shown that BMI accounts for 65-75% of risk of hypertension. Research also shows that smoking and BMI can put one at risk for heart disease by 25-30%. Last but not least stress, which can often come from work and ones relationships raises ones glucose level due to hormones that are released. By looking further than the risk factors first presented to us I was able to find the causes for them within our own dataset and realize that we can prevent diseases by changing our behavior today. By creating and maintaining a healthy lifestyle free of smoking and overeating, along with surrounding yourself with stress-free relationships, you can reduce the likelihood of having a stroke today.