Are you in danger of stroke? An insight into the leading causes.

Anjali Bansal, Pallabi Deb, Musthan M., Dr. Goutam Chakraborty, Dr. Miriam McGaugh
Oklahoma State University

ABSTRACT

Stroke aka cerebrovascular accident is one of the leading cause of death and a significant root of disability. According to World Health Organization, 15 million people suffer stroke worldwide each year. Of these, 5 million die and another 5 million are permanently disabled. Currently it is the fifth leading cause of death in U.S. and the most important cause of disability. The question is- *What leads to STROKE?*

A stroke is usually the result of lack of blood supply to the brain either from interruption of flow or reduction, thereby, depriving the brain tissue of oxygen and necessary nutrients. Brain cells begin to die within minutes of the event, and can lead to permanent disability. Hypertension plays a significant role in the occurrence of stroke. Hypertension weakens arterial walls in the brain that can lead to a rupture resulting in hemorrhagic stroke. This paper gives a detailed insight about the stroke occurrence in United States. This project also attempts to study the association between different cardiovascular diseases and stroke and to understand the importance of hypertension in bringing about a stroke.

The data was obtained from the Center for Health Systems Innovation at Oklahoma State University. Data regarding patient demographics, patient disease status, hospital encounters and the treatment were examined and predictive model was built using SAS Enterprise Miner. The results indicated that stroke was more prevalent in males with age group greater than 55 years. Hypertension and coronary atherosclerosis contribute significantly to the stroke development.

1. INTRODUCTION

Stroke is a medical emergency which can occur due to insufficient blood supply to the brain either due to insufficient flow or due to a complete lack of flow. Within minutes, the brain cells begin to die leading to permanent disability and if not treated early can even lead to death. The statistics show that someone in USA is having a stroke encounter every 40 seconds and someone dies of stroke every 3 minutes 45 seconds. Currently, stroke is the fifth leading cause of death in United States leading to 1 of every 19 deaths killing nearly 133,000 people a year. Even if one survives the deathly toll of stroke, yet it leads to long term disability such as paralysis, memory loss and emotional problems. The statistics also tell that each year, about 795,000 people experience a new or recurrent stroke. Coronary Heart Disease is the leading cause of death in USA (43.8 percent) followed by stroke (16.8 percent) making stroke the second leading global cause of death behind heart disease. Cardiovascular Diseases and stroke accounted for 14% of total health expenditures in 2013-2014. It is frequently seen that there is a relationship between cardiovascular diseases and stroke. This project aimed to study the relationship between cardiovascular diseases and stroke and to understand the role of different cardiovascular diseases in bringing about stroke with particular emphasis on the role of hypertension.

2. DATA BACKGROUND

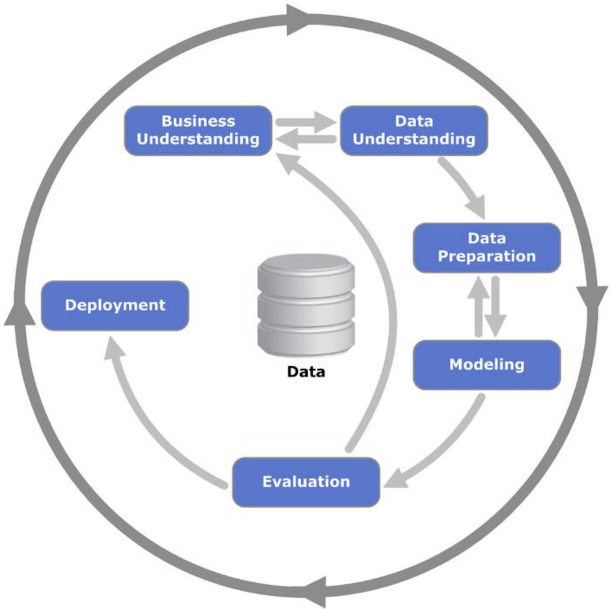
The data was obtained from Center for Health Systems Innovation. It is a department in OSU which is transforming healthcare through creativity, innovation and entrepreneurship. Three different RPT files were obtained from CHSI.

The files contained detailed information on patients' demographics, hospital encounters, clinical conditions and laboratory values from the year 2012-2014. For study purposes, the area was localized to South Atlantic region (includes FL, GA, SC, NC, VA, WV, MD, and DE) of USA and age group between 30-90 years. The files were first imported in excel and thereby in SAS Enterprise Guide. Different variables were studied in each dataset and relevant variables were selected.

3. DATA PREPARATION

The data was handled according to the principles of CRISP-DM

CRISP-DM stands for Cross-Industry Standard Process for Data Mining which provides a structured approach for planning a data mining project. The process includes the following steps:



CRISP-DM Methodology

a. Business Objective

This project was aimed at predicting the occurrence of stroke or not in patients already suffering from cardiovascular diseases. Further goal was to study the relative importance of each cardiovascular disease in bringing about stroke. Once the risk factors are identified, we could target these risk factors for prevention and hence, occurrence of stroke.

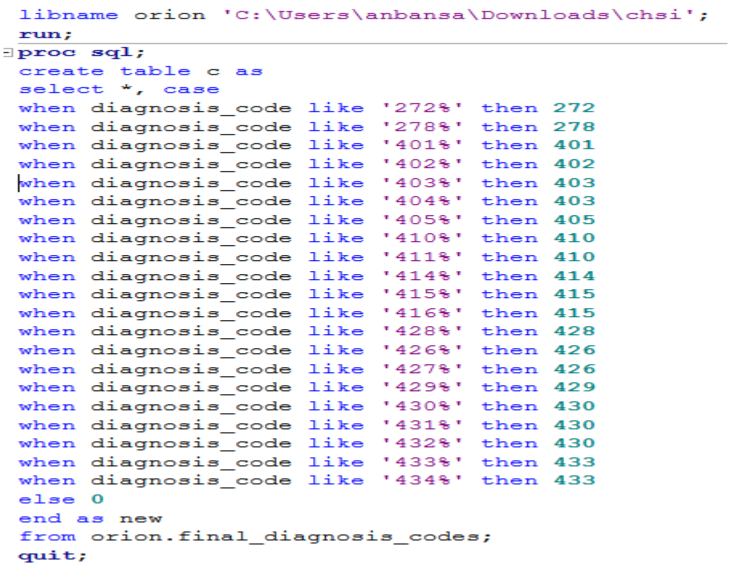
b. Understanding the data

There were three different RPT files referring to the patient's data. The clinical file gave details about clinical conditions of each patient. The demographics file showed the demographic features of each patient. The laboratory data file gave information about the laboratory tests performed over the span of 3 years. The relevant variables were selected from each data file.

c. Data Cleaning & Manipulation

There were about 2 million records in clinical condition file and a subset was created according to the clinical codes pertaining to cardiovascular diseases. The following ICD-9 codes were selected:

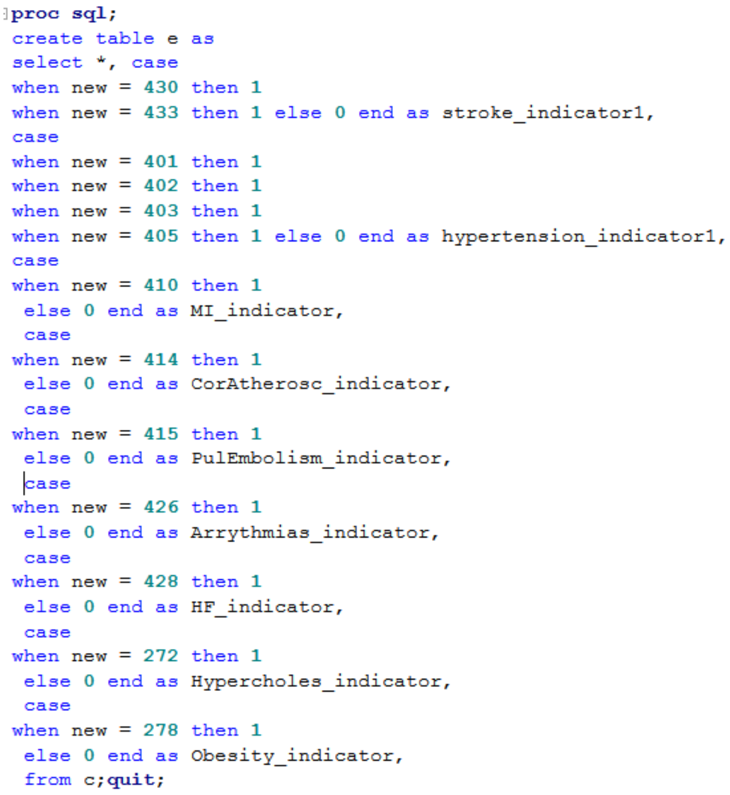
272(Hypercholesterolemia), 250 (Diabetes), 278(Obesity),401-405 (Hypertension and related disorders),410-411 (Myocardial Infarction and related conditions), 414 (Coronary Atherosclerosis), 415-416(Pulmonary Embolism and related conditions),426-427(Arrhythmias and related conditions),428(Heart Failure and related conditions),430-434(Stroke-hemorrhagic and ischemic). The clinical codes were combined into a single code if they were indicating similar diseases. For e.g. Clinical codes such as 401 and 402 are combined into a single clinical code of 401 indicating hypertension. This was done to reduce the dimensionality. Patient_sk was selected as unique id for each patient. A patient had several hospital encounters over three years so each patient had a unique id but several encounter_ids.



```
libname orion 'C:\Users\anbansa\Downloads\chsi';
run;
proc sql;
create table c as
select *, case
when diagnosis_code like '272%' then 272
when diagnosis_code like '278%' then 278
when diagnosis_code like '401%' then 401
when diagnosis_code like '402%' then 402
when diagnosis_code like '403%' then 403
when diagnosis_code like '404%' then 403
when diagnosis_code like '405%' then 405
when diagnosis_code like '410%' then 410
when diagnosis_code like '411%' then 410
when diagnosis_code like '414%' then 414
when diagnosis_code like '415%' then 415
when diagnosis_code like '416%' then 415
when diagnosis_code like '428%' then 428
when diagnosis_code like '426%' then 426
when diagnosis_code like '427%' then 426
when diagnosis_code like '429%' then 429
when diagnosis_code like '430%' then 430
when diagnosis_code like '431%' then 430
when diagnosis_code like '432%' then 430
when diagnosis_code like '433%' then 433
when diagnosis_code like '434%' then 433
else 0
end as new
from orion.final_diagnosis_codes;
quit;
```

New Variable creation

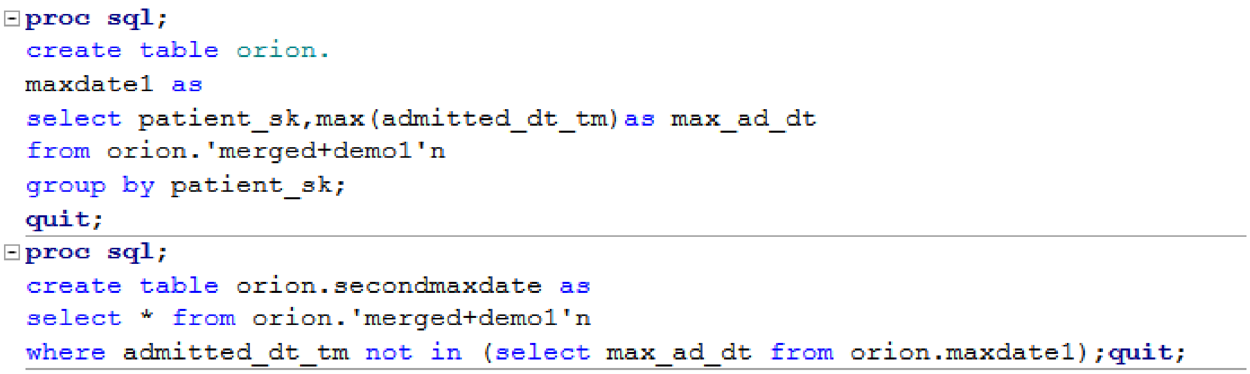
An indicator binary variable (1/0) was created for each diseased condition showing the presence or absence of a disease. So a clinical file dataset was created indicating the disease codes and the indicator variables showing the presence or absence of a disease for each patient.



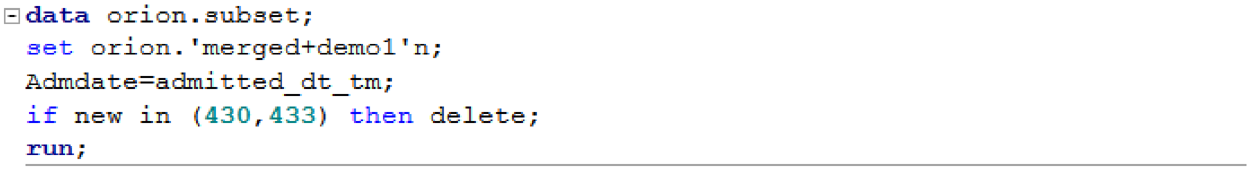
```
proc sql;
create table e as
select *, case
when new = 430 then 1
when new = 433 then 1 else 0 end as stroke_indicator1,
case
when new = 401 then 1
when new = 402 then 1
when new = 403 then 1
when new = 405 then 1 else 0 end as hypertension_indicator1,
case
when new = 410 then 1
 else 0 end as MI_indicator,
 case
when new = 414 then 1
 else 0 end as CorAtherosc_indicator,
 case
when new = 415 then 1
 else 0 end as PulEmbolism_indicator,
 case
when new = 426 then 1
 else 0 end as Arrythmias_indicator,
 case
when new = 428 then 1
 else 0 end as HF_indicator,
 case
when new = 272 then 1
 else 0 end as Hypercholes_indicator,
 case
when new = 278 then 1
 else 0 end as Obesity_indicator,
 from c;quit;
```

The demographics file (around 2 million records) was joined with the clinical table file on the common variable of patient_sk to create a dataset named (merged_demo1). Each patient had numerous encounters over three years with different encounter_ids. The encounter_id was sorted according to the admission date and time. For study purposes, the encounter immediately preceding the stroke encounter in stroke patients or second last encounter in non-stroke patients was considered. This means that the last encounter for each patient was dropped in consideration. A final data (A) set containing clinical condition, indicator variables and demographics was created. The patients with clinical code of 430,433(stroke) were separated from this dataset (A).

Two separate datasets were therefore created from (A)-(1) stroke dataset having stroke codes 430 and 433 and included patients having stroke (2) A second dataset named -subset1 having all codes except stroke codes. This dataset included patients who did not develop stroke and were given a stroke indicator of 0.

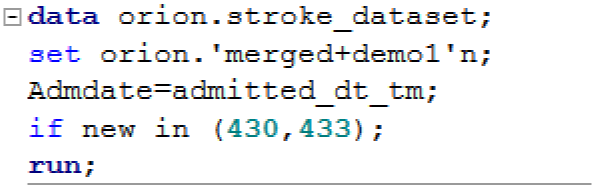


```
proc sql;
create table orion.
maxdate1 as
select patient_sk,max(admitted_dt_tm)as max_ad_dt
from orion.'merged+demo1'n
group by patient_sk;
quit;
proc sql;
create table orion.secondmaxdate as
select * from orion.'merged+demo1'n
where admitted_dt_tm not in (select max_ad_dt from orion.maxdate1);quit;
```

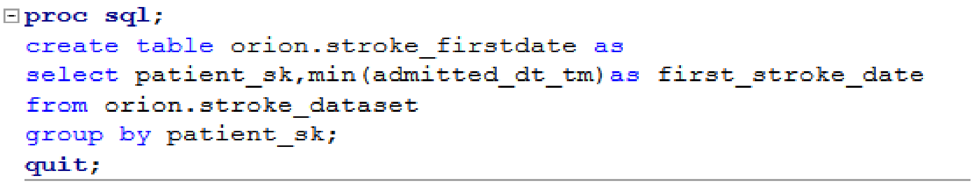


```
data orion.subset;
set orion.'merged+demo1'n;
Admdate=admitted_dt_tm;
if new in (430,433) then delete;
run;
```

The **stroke dataset** had the stroke codes (430,433) and included a total of 1741 patients, who had one or more than one episode of stroke.

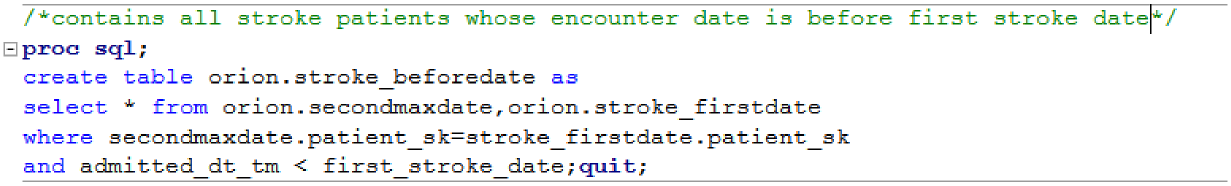


```
data orion.stroke_dataset;
set orion.'merged+demo1'n;
Admdate=admitted_dt_tm;
if new in (430,433);
run;
```

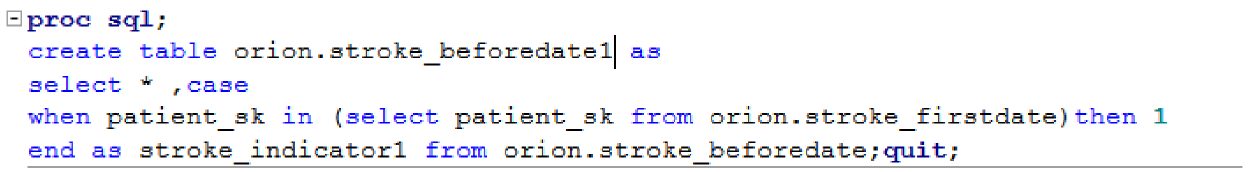


```
proc sql;
create table orion.stroke_firstdate as
select patient_sk,min(admitted_dt_tm)as first_stroke_date
from orion.stroke_dataset
group by patient_sk;
quit;
```

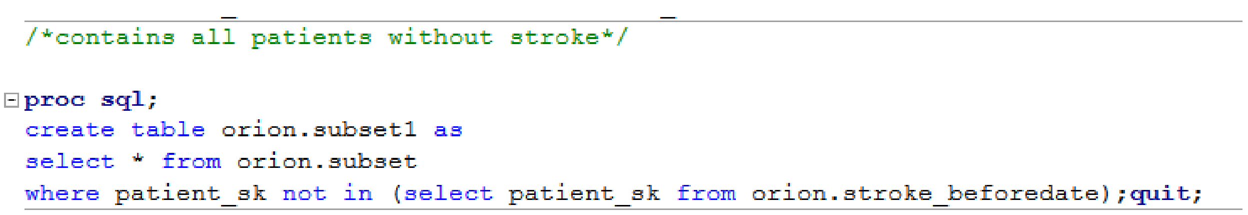
The patients in the stroke dataset were studied and it was found that 626 patients had previous hospital encounters before first episode of stroke. These patients were suffering from cardiovascular diseases and potentially developed stroke in the future, therefore, they were given a stroke indicator of 1 for consideration in predictive modelling. With the below query, all hospital encounters before first stroke encounter were selected.



```
/*contains all stroke patients whose encounter date is before first stroke date*/
proc sql;
create table orion.stroke_beforedate as
select * from orion.secondmaxdate,orion.stroke_firstdate
where secondmaxdate.patient_sk=stroke_firstdate.patient_sk
and admitted_dt_tm < first_stroke_date;quit;
```



```
proc sql;
 create table orion.stroke_beforedate1 as
 select * ,case
 when patient_sk in (select patient_sk from orion.stroke_firstdate)then 1
 end as stroke_indicator1 from orion.stroke_beforedate;quit;
```

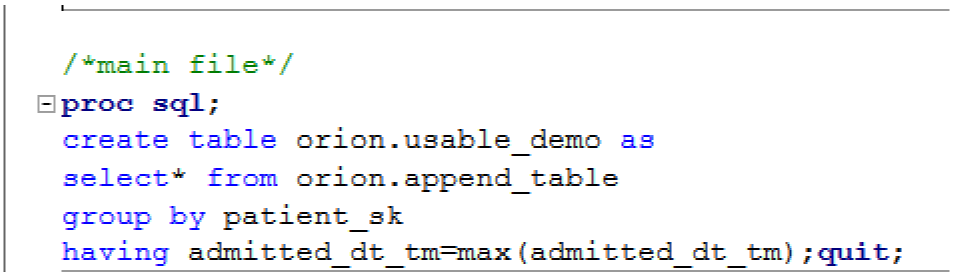


```
/*contains all patients without stroke*/

proc sql;
 create table orion.subset1 as
 select * from orion.subset
 where patient_sk not in (select patient_sk from orion.stroke_beforedate);quit;
```

So datasets under consideration were the final two data sets-

Subset1-has patients who did not have stroke with stroke indicator as 0 and stroke_beforedate1 (626 patients) has stroke patients with stroke indicator as 1. Both of these datasets were appended together to create a final dataset Append_table. The observations with encounter_id having maximum admitted date and time was selected from this append_table and a new dataset was created named usable_demo which was considered for predictive modeling purposes. This data set had 34572 observations and included 626 patients who potentially developed stroke in future. It was further merged with laboratory tests files.



```
/*main file*/
proc sql;
 create table orion.usable_demo as
 select* from orion.append_table
 group by patient_sk
 having admitted_dt_tm=max(admitted_dt_tm);quit;
```

The main laboratory file included encounters, lab tests and lab values. It was subset into different files according to the laboratory tests. For e.g. a separate file was created from the laboratory data including only glucose test. Similarly, a lab file referring to urea test and white blood cell tests, neutrophils and red blood cells were created separately. These files had encounter_ids along with test names and numeric values.

The separate tables of glucose, urea and wbc, neut, red blood cells were merged with the usable_demo dataset on the variable encounter_id.

The dataset for predictive modelling contained 34572 observations. Out of these total, 33946 patients did not have any previous episode of stroke and were given stroke indicator of 0. Rest 626 patients had one or more episode of stroke in the future and were given stroke indicator of 1. As this was a case of rare target event (target~2%), over sampling (50:50) was performed in predictive modelling with equal number of stroke and non-stroke patients. Therefore, the final dataset had a total of 1252 observations and 21 variables and was used for predictive modeling to predict the occurrence of stroke or not.

d. Modeling

In this phase we use various modeling tools such as SAS Enterprise Miner, SAS Enterprise Guide to build models on the cleaned data.

e. Evaluation

In this phase we compare results to the business objective to decide whether we have achieved our objective or not.

f. Deployment

In this phase we deploy the models created in the previous phases then they are applied to business operations for many purposes, including prediction or identification of key situations.

4. DATA DICTIONARY

Variable	**Data Type**	**Description**
Patient_SK	ID	Unique identifier for each patient
Age_in_years	Interval	Indicates the age of the patient at second last encounter
Marital_status	Nominal	Indicates marital status
Urban_Rural	Nominal	Indicates patient belong to rural/urban
Gender	Nominal	Indicates Male or female patient
Race	Nominal	Indicates patient belongs to which race
Hypertension_Indicator	Binary	Indicates hypertension is present or not
HF_indicator	Binary	Indicates presence/absence of heart failure

Coratherosc_indicator	Binary	Indicates presence/absence of coronary atherosclerosis(Fatty deposition in heart arteries)
Arrythmias_Indicator	Binary	Indicates presence/absence of arrhythmias
Pulembolism_indicator	Binary	Indicates presence/absence of pulmonary embolism
MI_indicator	Binary	Indicates occurrence of Myocardial Infarction
Hypercholes_indicator	Binary	Indicates high cholesterol level or not
Diabetes_indicator	Binary	Indicates presence or absence of diabetes
Stroke_indicator1	Binary(Target)	Indicates occurrence of stroke or not
Obesity_indicator	Binary	Indicates presence/absence of obesity
glucose	Interval	Indicates avg glucose level in the encounter under study
urea	Interval	Indicates Avg urea level in the encounter under study
red	Interval	Indicates Avg red blood cell counts in the encounter under study
wbc	Interval	Indicates Avg white blood cell counts in the encounter under study
neut	Interval	Indicates Avg neutrophil counts in the encounter under study

5. DATA INSIGHTS

The stroke dataset having patients with all stroke encounters were studied to understand the demographics associated with stroke. The highlights of the descriptive analysis are:

A. It is seen that female have a higher occurrence of stroke as compared to males

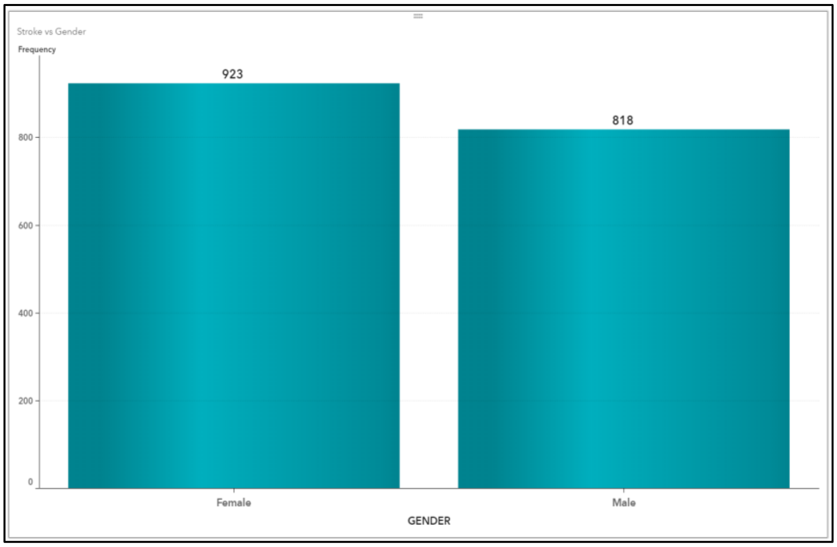


Figure 1: Stroke encounters vs Gender

B. Caucasian race people had higher stroke encounters followed by African American people.

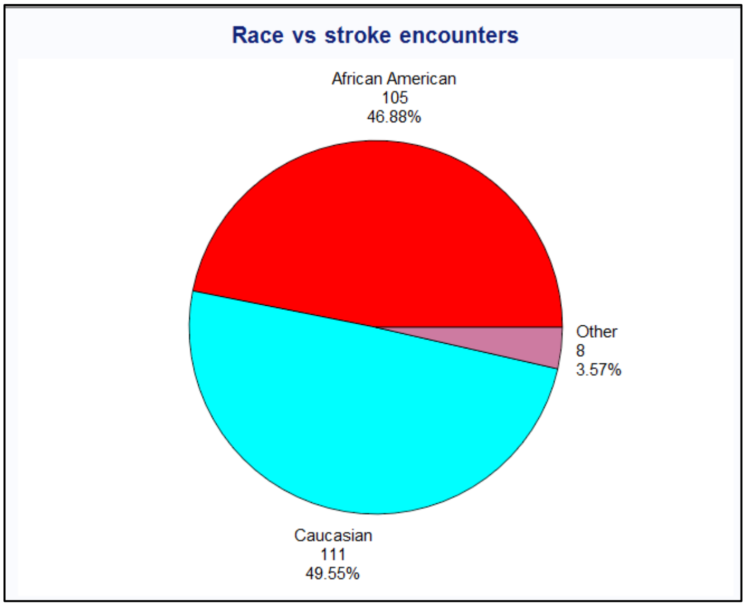


Figure 2: Stroke encounters in different Races

C. Stroke category 433(Ischemic stroke) is more common than 430 (Hemorrhagic stroke). Ischemic stroke mostly occurs due to occlusion or narrowing of brain arteries.

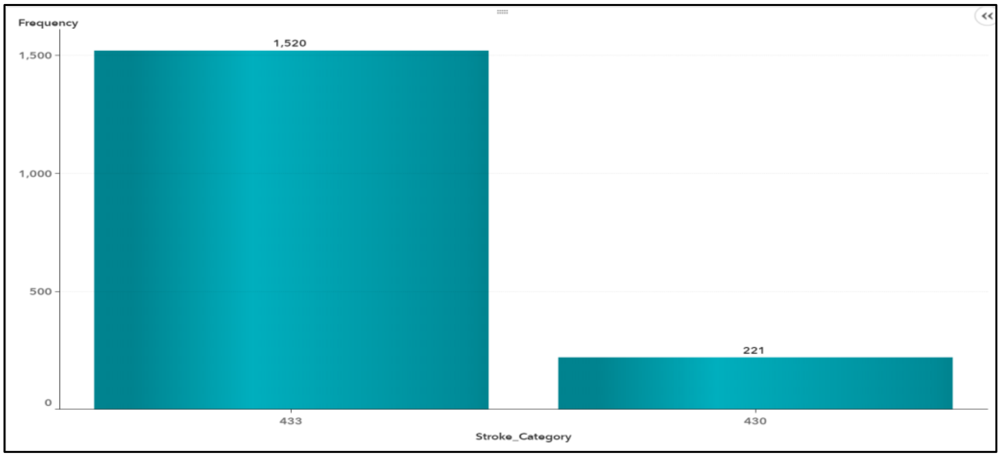


Figure 3: Graph showing frequency of different type of stroke

D. Single individuals (Either widowed/Divorced/Single) together show a higher occurrence of stroke as compared to married persons.

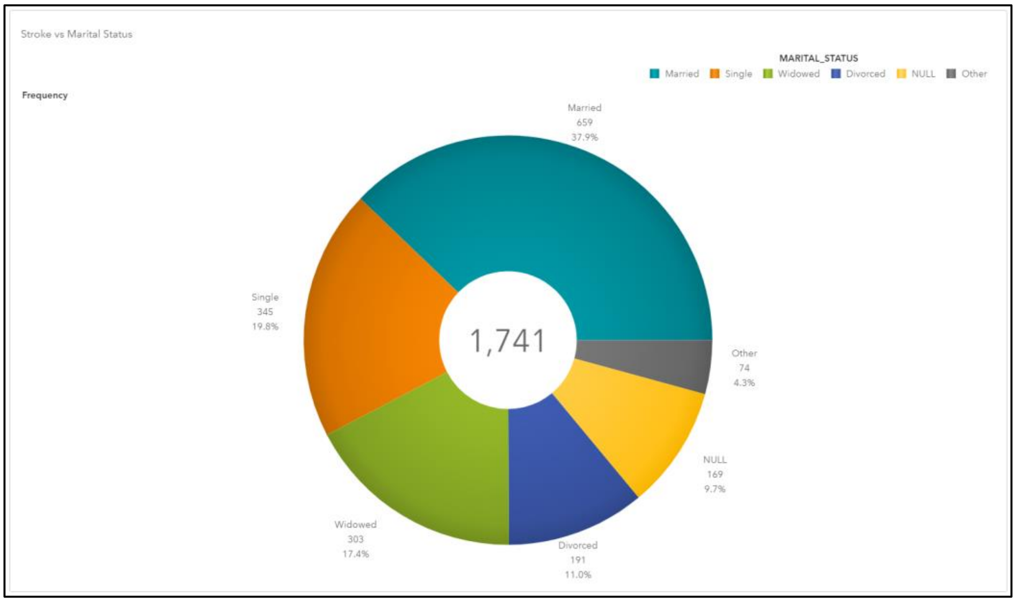


Figure 4: Stroke encounters vs marital status

E. Stroke encounters were found to be more prevalent in urban population as compared to rural population.

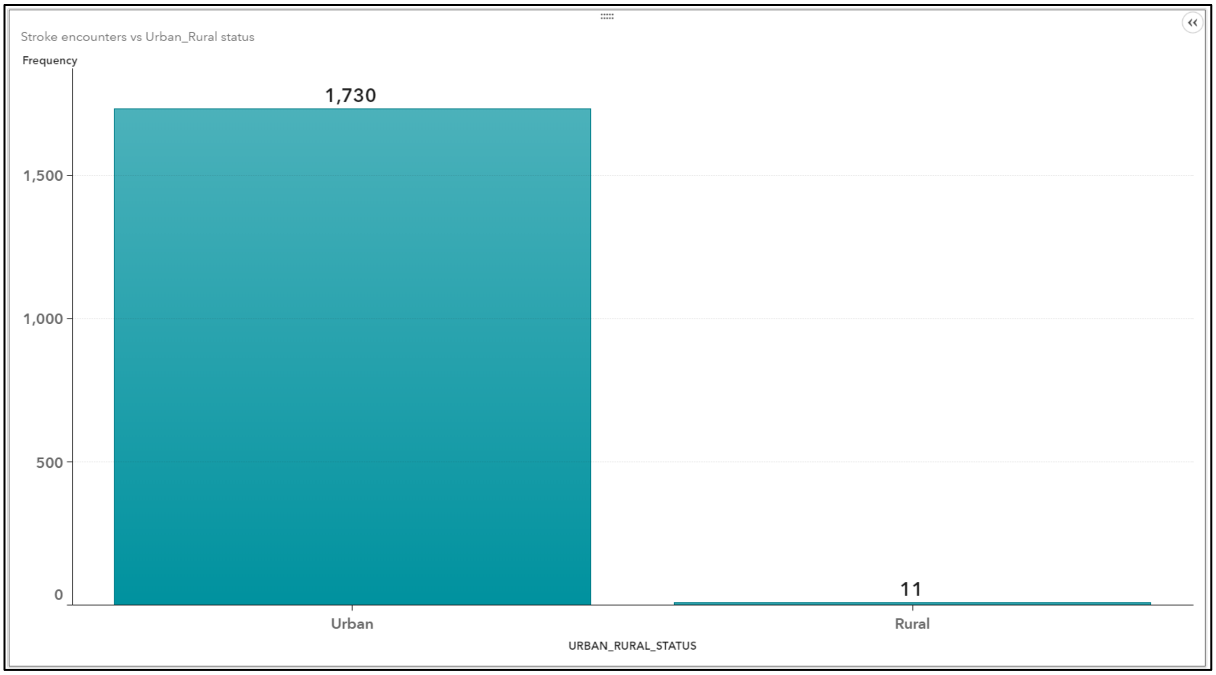


Figure 5: Stroke encounters vs Urban/Rural status

F. The mean age is higher (67.45 years) for 433-Ischaemic stroke as compared to 430-Haemorrhagic stroke where mean age is 64.3 years.

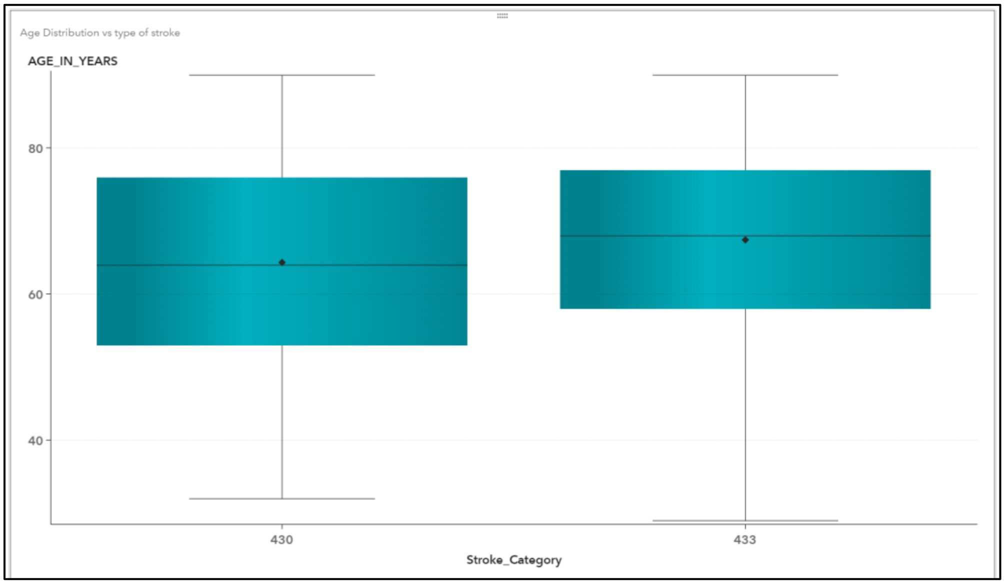


Figure 6: Comparison of mean age in different types of stroke

G. The number of stroke encounters are higher for older patients above 55 years.

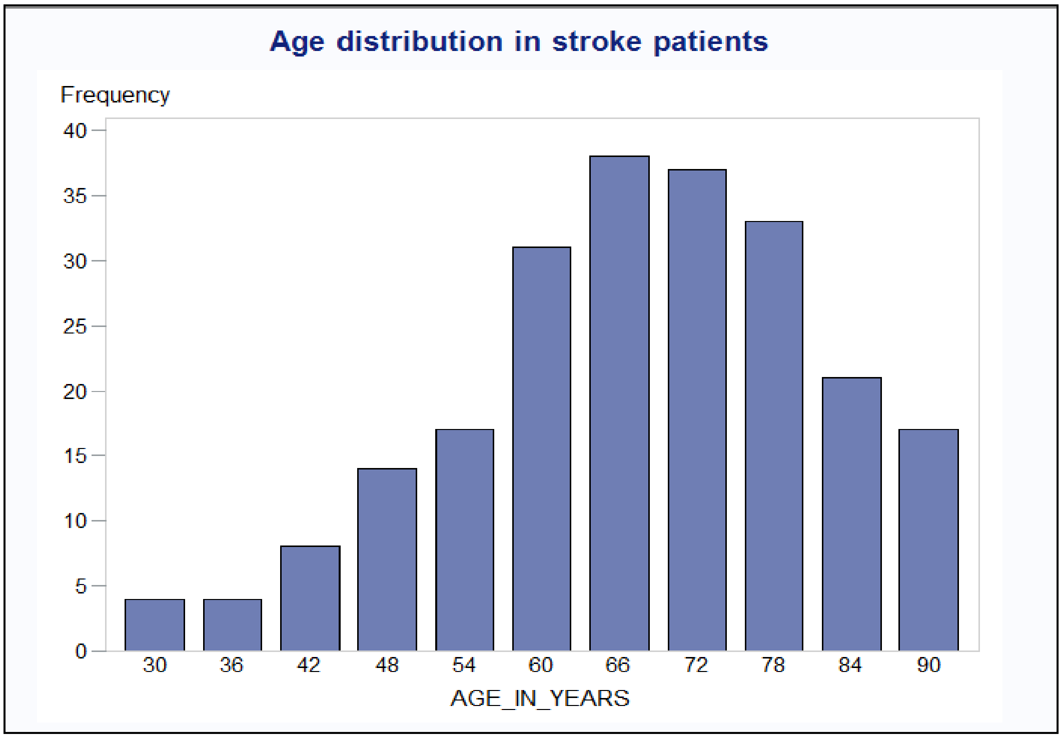


Figure 7: Age distribution over stroke encounters

6. MODEL BUILDING

The predictive model was built in SAS Enterprise miner to predict the occurrence of stroke or not, taking into account the patient's' demographics and presence/absence of other cardiovascular diseases. As the stroke indicator(target) is ~2% of total observations, this is a case of rare event target and over sampling is performed.

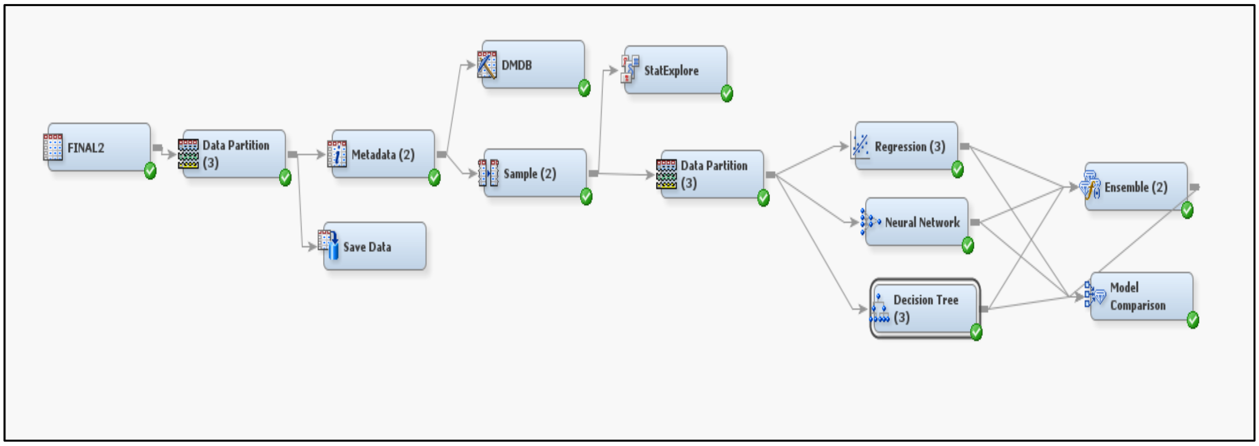
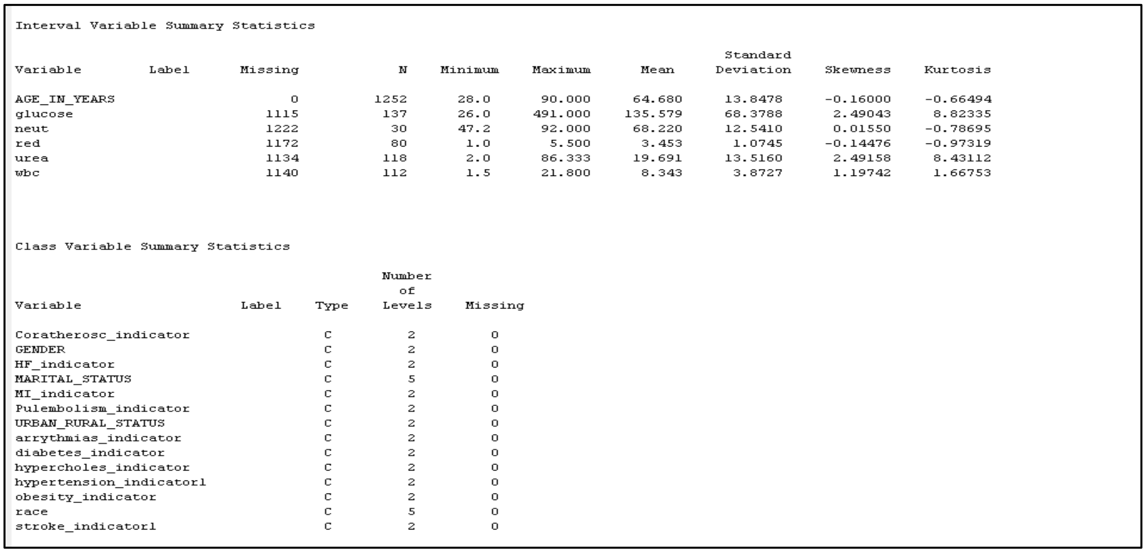


Figure 8: SAS EM Process Flow Diagram

The results of the DMDB node are:



```
Interval Variable Summary Statistics

                                                                                 Standard
Variable        Label       Missing           N    Minimum    Maximum       Mean   Deviation    Skewness    Kurtosis

AGE_IN_YEARS                      0        1252       28.0     90.000     64.680     13.8478    -0.16000    -0.66494
glucose                        1115         137       26.0    491.000    135.579     68.3788     2.49043     8.82335
neut                           1222          30       47.2     92.000     68.220     12.5410     0.01550    -0.78695
red                            1172          80        1.0      5.500      3.453      1.0745    -0.14476    -0.97319
urea                           1134         118        2.0     86.333     19.691     13.5160     2.49158     8.43112
wbc                            1140         112        1.5     21.800      8.343      3.8727     1.19742     1.66753




Class Variable Summary Statistics

                                              Number
                                                of
Variable                    Label     Type    Levels    Missing

Coratherosc_indicator                  C         2         0
GENDER                                 C         2         0
HF_indicator                           C         2         0
MARITAL_STATUS                         C         5         0
MI_indicator                           C         2         0
Pulembolism_indicator                  C         2         0
URBAN_RURAL_STATUS                     C         2         0
arrythmias_indicator                   C         2         0
diabetes_indicator                     C         2         0
hypercholes_indicator                  C         2         0
hypertension_indicator1                C         2         0
obesity_indicator                      C         2         0
race                                   C         5         0
stroke_indicator1                      C         2         0
```

Figure 9: Results from DMDB node

As greater than 50% of the lab values are missing so they are given a rejected role in further analysis. Imputation was not performed as it would lead to bias in view of large number of missing values.

The variables' worth was studied from stat explore node

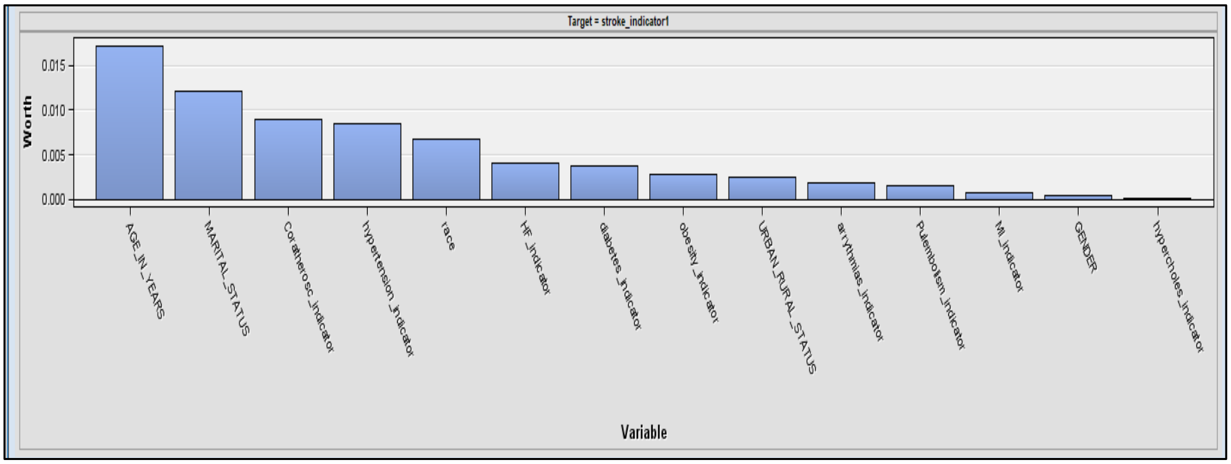


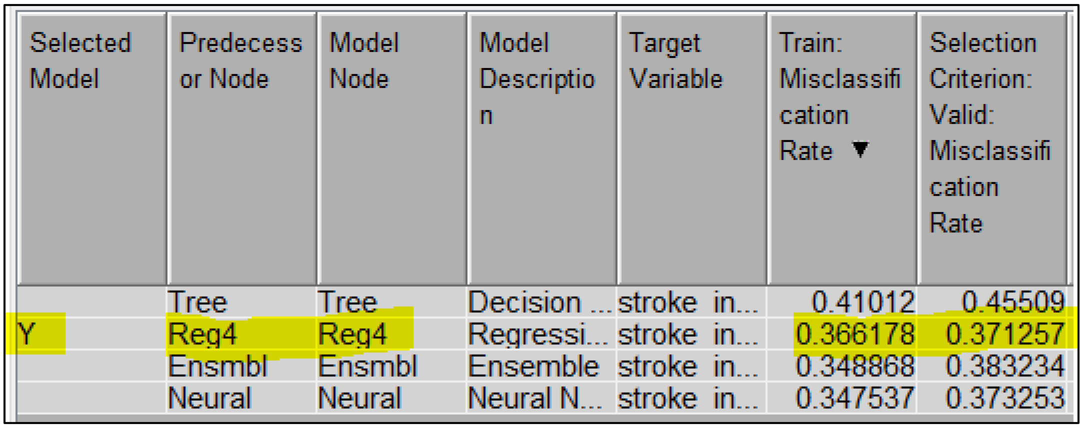
Figure 10: Results from STAT Explore node

It was seen that age plays the most important role in the occurrence of stroke followed by marital status. Hypertension and coronary atherosclerosis are the two most important cardiovascular disorders associated with stroke.

The dataset was further partitioned into 60% training and 40% validation for model building purposes and to obtain proper model assessment. Regression model, Decision tree, Neural

Network and Ensemble models were built to predict the occurrence of stroke. Misclassification rate was used as the model selection criteria.

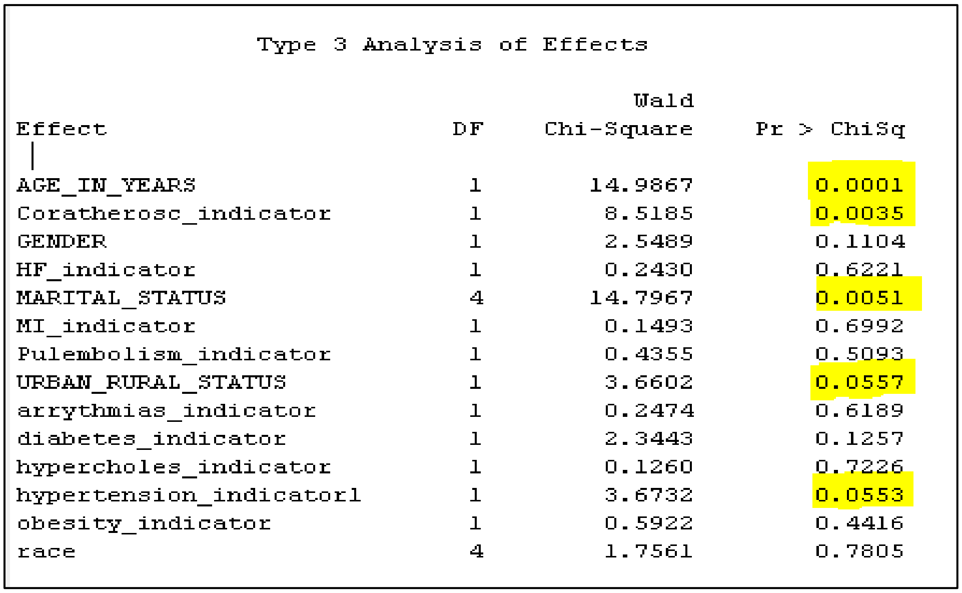
The results from the different models were studied and it was seen that logistic regression model with default settings gave the best results. As the odds ratio results could also be visualized with regression model, therefore it was selected as the model of choice.



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Misclassification Rate ▼	Selection Criterion: Valid: Misclassification Rate
	Tree	Tree	Decision ...	stroke in...	0.41012	0.45509
Y	Reg4	Reg4	Regressi...	stroke in...	0.366178	0.371257
	Ensmbl	Ensmbl	Ensemble	stroke in...	0.348868	0.383234
	Neural	Neural	Neural N...	stroke in...	0.347537	0.373253

Table 1: Model comparison and assessment results

The significant variables were age of patient, coronary atherosclerosis, marital status, urban/rural status and hypertension.



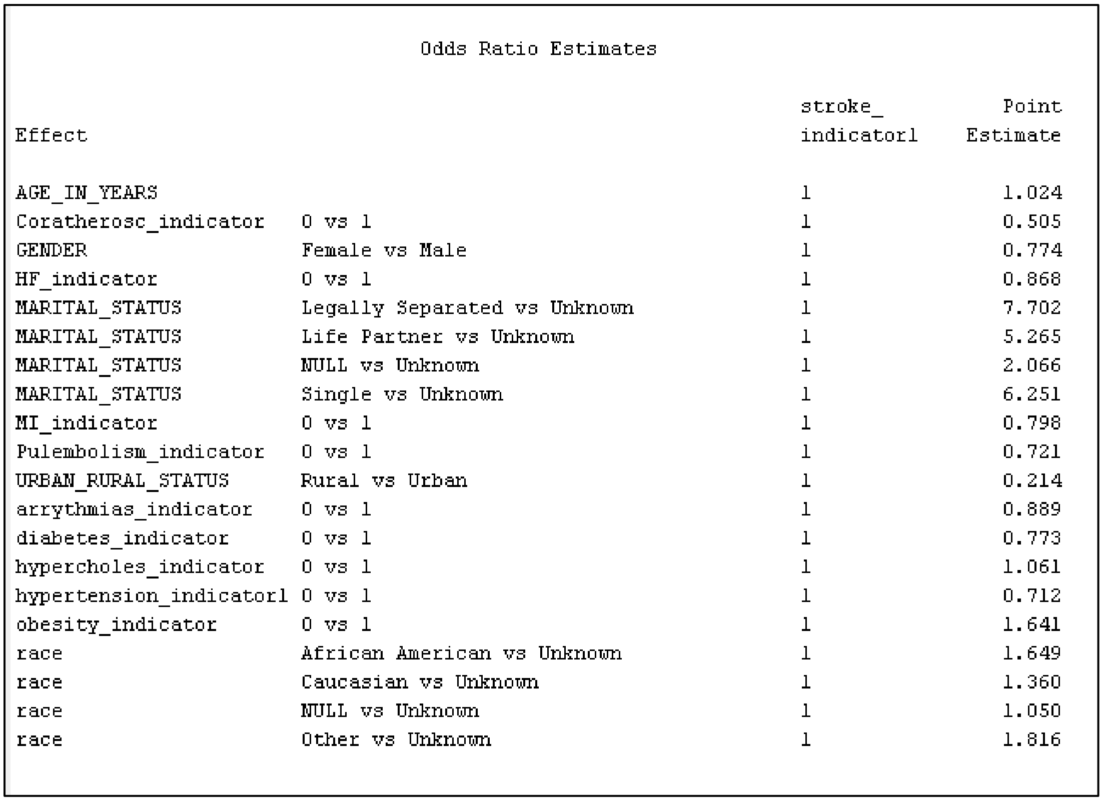
```
                    Type 3 Analysis of Effects

                                                Wald
Effect                            DF      Chi-Square      Pr > ChiSq

AGE_IN_YEARS                       1         14.9867          0.0001
Coratherosc_indicator              1          8.5185          0.0035
GENDER                             1          2.5489          0.1104
HF_indicator                       1          0.2430          0.6221
MARITAL_STATUS                     4         14.7967          0.0051
MI_indicator                       1          0.1493          0.6992
Pulembolism_indicator              1          0.4355          0.5093
URBAN_RURAL_STATUS                 1          3.6602          0.0557
arrythmias_indicator               1          0.2474          0.6189
diabetes_indicator                 1          2.3443          0.1257
hypercholes_indicator              1          0.1260          0.7226
hypertension_indicator1            1          3.6732          0.0553
obesity_indicator                  1          0.5922          0.4416
race                               4          1.7561          0.7805
```

Figure 11: Results of logistic regression model showing significant variables

Odds ratio was analyzed to study the effect of different variables



```
                              Odds Ratio Estimates

                                                                 stroke_         Point
Effect                                                           indicator1    Estimate

AGE_IN_YEARS                                                     1               1.024
Coratherosc_indicator   0 vs 1                                   1               0.505
GENDER                  Female vs Male                           1               0.774
HF_indicator            0 vs 1                                   1               0.868
MARITAL_STATUS          Legally Separated vs Unknown             1               7.702
MARITAL_STATUS          Life Partner vs Unknown                  1               5.265
MARITAL_STATUS          NULL vs Unknown                          1               2.066
MARITAL_STATUS          Single vs Unknown                        1               6.251
MI_indicator            0 vs 1                                   1               0.798
Pulembolism_indicator   0 vs 1                                   1               0.721
URBAN_RURAL_STATUS      Rural vs Urban                           1               0.214
arrythmias_indicator    0 vs 1                                   1               0.889
diabetes_indicator      0 vs 1                                   1               0.773
hypercholes_indicator   0 vs 1                                   1               1.061
hypertension_indicator1 0 vs 1                                   1               0.712
obesity_indicator       0 vs 1                                   1               1.641
race                    African American vs Unknown              1               1.649
race                    Caucasian vs Unknown                     1               1.360
race                    NULL vs Unknown                          1               1.050
race                    Other vs Unknown                         1               1.816
```

Figure 12: Odds ratio estimates of different variables

7. RESULTS

1. Males have 30% more chance of having stroke as compared to females.
2. Stroke occurrence is found at a higher rate in single/legally separated individuals as compared to people with life partners.
3. The rate of stroke encounters is 5 times higher in urban population as compared to rural.
4. The descriptive statistics showed that Caucasians outnumbered African American people in having stroke occurrences.
5. The predictive analytics model showed that African American people are 21% more prone to the development of stroke as compared to Caucasians.
6. Ischemic stroke is much more prevalent as compared to hemorrhagic stroke
7. The mean age of ischemic stroke is much more than hemorrhagic stroke
8. Amongst the cardiovascular diseases, it was seen that hypertension plays a significant role in the occurrence of stroke. People with hypertension have approximately 40% more chance of having stroke as compared to non-hypertensive patients.
9. The risk of stroke increases 2 times in persons with coronary atherosclerosis. People with co-occurrence of diabetes have 1.4 times more risk of stroke.

10. Presence of pulmonary embolism increases the stroke risk by 40% and arrhythmias increase the stroke risk by 14%.

8. CONCLUSIONS

1. Stress came out to be an important factor contributing towards stroke. Stressful conditions such as legally separated/single individuals or people living in urban areas are more prone to stroke.
2. Ischemic stroke which is due to lack/reduced blood supply is the more prevalent form of stroke. Reduced blood supply can be due to atherosclerosis of brain vessels or an embolus lodging itself into the vessels.
3. The results show that a co-existent cardiovascular disorder markedly increases the risk of stroke.
4. Coronary atherosclerosis and hypertension came out be the most significant cardiovascular risk factors for stroke.
5. Age plays a major role in the occurrence of stroke. As the age advances, the risk of stroke also increases.
6. Males with advanced age and suffering from hypertension or coronary atherosclerosis are most vulnerable to stroke.

9. LIMITATIONS OF STUDY

The study was localized in the South-Atlantic region of United States. It is not reflective of the overall picture in the whole country. The study was also limited to the role of few cardiovascular disorders in bringing about stroke and had not taken into account other diseases such as brain diseases or any other organ dysfunction, which could also have a potential role in bringing about stroke.

10. FURTHER SCOPE

There were more than 50% missing laboratory values in the dataset, which if present would have improved the model assessment criteria. Other variables such as smoking history, alcohol intake, body mass index can also be included to study their impact in bringing about stroke.

11. ACKNOWLEDGEMENTS

I thank Dr. Goutam Chakraborty, SAS® Professor of Marketing Analytics and Dr. Miriam McGaugh, Clinical Professor at Oklahoma State University for their constant support and guidance throughout this project.

This work was conducted with data from the Cerner Corporation's Health Facts database of electronic medical records provided by the Oklahoma State University Center for Health Systems Innovation (CHSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Cerner Corporation.

12.REFERENCES

1. Stroke-In Wikipedia-The free Encyclopedia from: https://en.wikipedia.org/wiki/Stroke

2. Stroke health center: https://www.webmd.com/stroke/default.htm

3. Stroke: Causes, symptoms, diagnosis, and treatment:
https://www.medicalnewstoday.com/articles/7624.php

4. Stroke | CVA | Cerebrovascular Accident | MedlinePlus:
https://medlineplus.gov/stroke.html

5. Stroke Information | cdc.gov:
https://www.cdc.gov/stroke/index.htm

6. What is stroke? | Stroke.org:
http://www.stroke.org/understand-stroke

7. How High Blood Pressure Can Lead to Stroke
https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-stroke

8. Heart Disease and Stroke
https://www.webmd.com/heart-disease/stroke-types#1

CONTACT INFORMATION

Your comments, feedback and questions are valued and encouraged. You can contact the author at:

Anjali Bansal
Oklahoma State University, Stillwater OK
Email: anjali.bansal@okstate.edu
Phone no.: 405-762-1777