PREDICTING A STROKE

Ivan Rivera, Seve Silvestre, Jonathan Shaklan, Rylan Mukssood

ABOUT THE DATASET

- 11 variables
- 4,909 rows
- 7 categorical variables
- Female heavy

```
   gender          age        hypertension heart_disease ever_married         work_type
 Female:2897   Min.   : 0.08   No :4458     No :4666      No :1705     children     : 671
 Male  :2011   1st Qu.:25.00   Yes: 451     Yes: 243      Yes:3204     Govt_job     : 630
 Other :   1   Median :44.00                                           Never_worked :  22
               Mean   :42.87                                           Private      :2811
               3rd Qu.:60.00                                           Self-employed: 775
               Max.   :82.00
 Residence_type avg_glucose_level      bmi                smoking_status stroke
 Rural:2419     Min.   : 55.12    Min.   :10.30   formerly smoked: 837   0:4700
 Urban:2490     1st Qu.: 77.07    1st Qu.:23.50   never smoked   :1852   1: 209
                Median : 91.68    Median :28.10   smokes         : 737
                Mean   :105.31    Mean   :28.89   Unknown        :1483
                3rd Qu.:113.57    3rd Qu.:33.10
                Max.   :271.74    Max.   :97.60
```

VARIABLES:

```
Rows: 4,909
Columns: 11
$ gender            <fct> Male, Male, Female, Female, Male, Male, Female, Female, Female, Female, Fem…
$ age               <dbl> 67, 80, 49, 79, 81, 74, 69, 78, 81, 61, 54, 79, 50, 64, 75, 60, 71, 52, 79,…
$ hypertension      <fct> No, No, No, Yes, No, Yes, No, No, Yes, No, No, No, Yes, No, Yes, No, No, Ye…
$ heart_disease     <fct> Yes, Yes, No, No, No, Yes, No, No, No, Yes, No, Yes, No, Yes, No, No, No, N…
$ ever_married      <fct> Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, N…
$ work_type         <fct> Private, Private, Private, Self-employed, Private, Private, Private, Privat…
$ Residence_type    <fct> Urban, Rural, Urban, Rural, Urban, Rural, Urban, Urban, Rural, Rural, Urban…
$ avg_glucose_level <dbl> 228.69, 105.92, 171.23, 174.12, 186.21, 70.09, 94.39, 58.57, 80.43, 120.46,…
$ bmi               <dbl> 36.6, 32.5, 34.4, 24.0, 29.0, 27.4, 22.8, 24.2, 29.7, 36.8, 27.3, 28.2, 30.…
$ smoking_status    <fct> formerly smoked, never smoked, smokes, never smoked, formerly smoked, never…
$ stroke            <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
```

- **gender:** "Male", "Female" or "Other"
- **age:** age of the patient
- **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever_married:** "No" or "Yes"
- **work_type:** "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- **Residence_type:** "Rural" or "Urban"
- **avg_glucose_level:** average glucose level in blood
- **bmi:** body mass index
- **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"

“

CAN WE PREDICT WHETHER A PERSON WILL HAVE A STROKE BASED ON THE ALL SELECTED VARIABLES?

MOTIVATION

- 10,000,000 people per year experience long term damage from strokes
- Prediction can be crucial in recognition and prevention
- Healthcare professionals can treat patients before the event of a stroke

ANALYSIS

Logistic Regression:

- To predict the odds of “stroke” occurring, we used a logistic regression after cleaning our dataset and adding a variable “stroke_numeric” that would contain the same data as “stroke” but as a numeric data type
- Set a threshold of 0.09 where anything greater receives a value of 1 for “stroke” and 0 otherwise

KMeans Clustering:

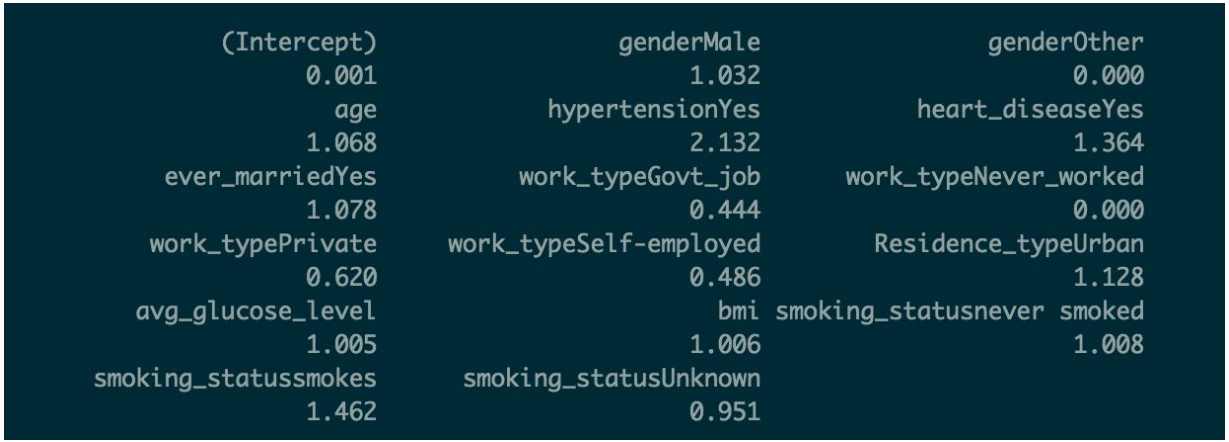
- Performed a K Means clustering algorithm using ‘euclidean’ as our distance metric. The idea here was for us to find distinct groups based on the variables “age”, “avg_glucose_level”, and “bmi” (continuous variables)
- Used a random sample of 1000 rows.

Decision Tree:

- The decision tree set boundaries as to what different criterias being meant may affect somebody’s chances of having a stroke.
- Indicates which coefficients are most effective in determining somebody’s likelihood of of having a stroke which in this case was age and average glucose levels

MODEL 1: LOGISTIC REGRESSION

- Performed a logistic regression model to find out which variable has the highest correlation to the chance of a stroke.



(Intercept)	genderMale	genderOther
0.001	1.032	0.000
age	hypertensionYes	heart_diseaseYes
1.068	2.132	1.364
ever_marriedYes	work_typeGovt_job	work_typeNever_worked
1.078	0.444	0.000
work_typePrivate	work_typeSelf-employed	Residence_typeUrban
0.620	0.486	1.128
avg_glucose_level	bmi	smoking_statusnever smoked
1.005	1.006	1.008
smoking_statussmokes	smoking_statusUnknown	
1.462	0.951	

MODEL 1: LOGISTIC REGRESSION

Training Confusion Matrix

Prediction	Truth 0	1
0	3051	64
1	474	93

Testing Confusion Matrix

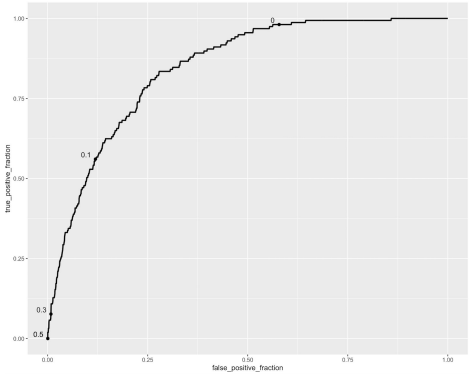
Prediction	Truth 0	1
0	1015	14
1	160	38

```
[1] "Training Accuracy: 0.854427"
[1] "Training Sensitivity: 0.598726"
[1] "Training Specificity: 0.865816"
```

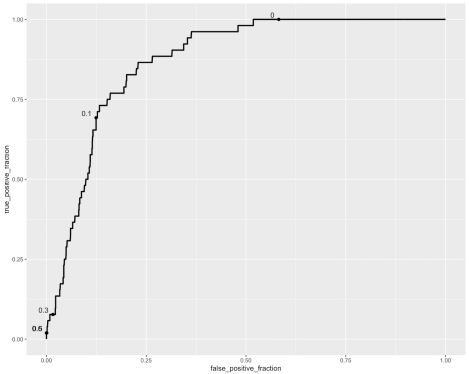
```
[1] "Testing Accuracy: 0.852486"
[1] "Testing Sensitivity: 0.711538"
[1] "Testing Specificity: 0.858723"
```

MODEL 1: LOGISTIC REGRESSION

Training ROC



Testing ROC



Training AUC

AUC
<dbl>
0.8449492

Testing AUC

AUC
<dbl>
0.8720376

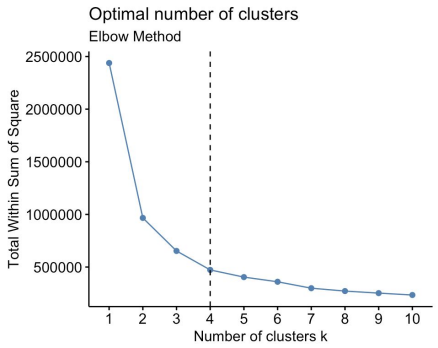
MODEL 1: EVALUATION AND RECOMMENDATION

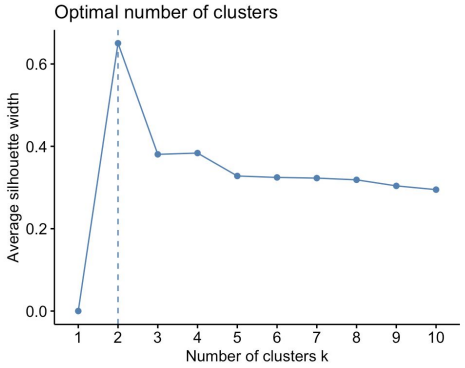
- Confusion Matrix:
  - More false negatives than false positives
- Model:
  - Fairly high testing accuracy with 85%
- Recommendation:
  - As patients are falsely informed that they will not experience a stroke, they should be receive early treatment.

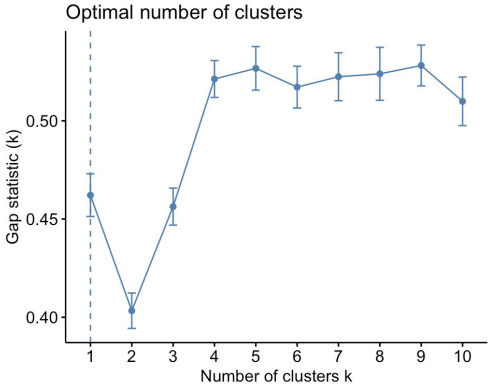
	Truth	
Prediction	0	1
0	1015	14
1	160	38

```
[1] "Testing Accuracy: 0.852486"
[1] "Testing Sensitivity: 0.711538"
[1] "Testing Specificity: 0.858723"
```

MODEL 2: CLUSTERING







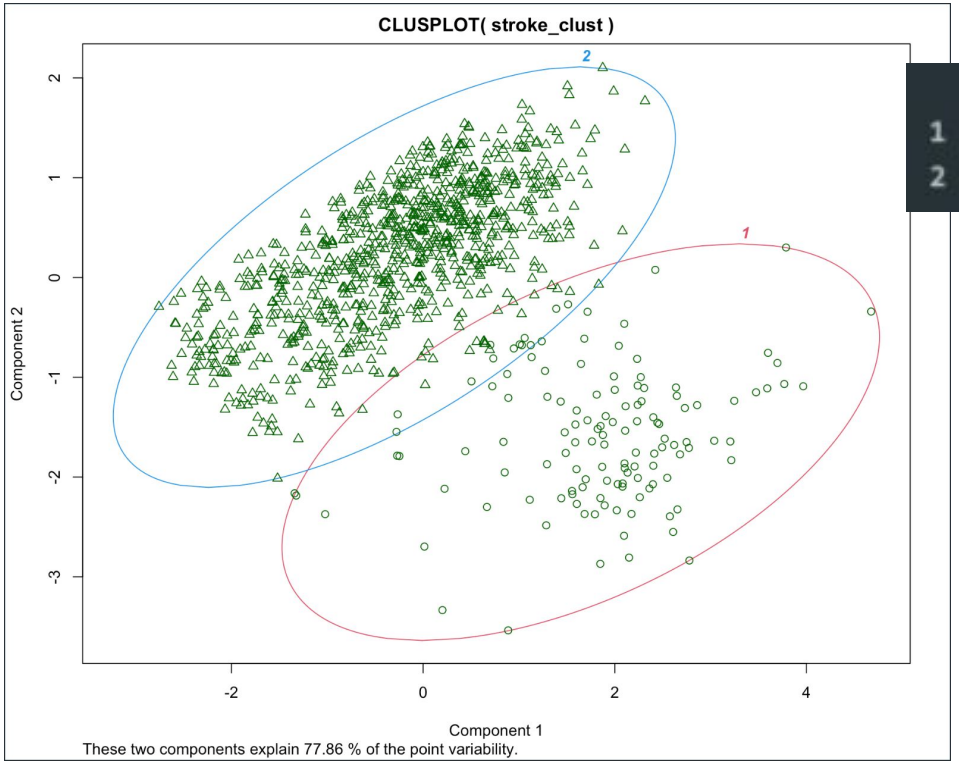
MODEL 2: CLUSTERING

```
* Among all indices:
* 10 proposed 2 as the best number of clusters
* 8 proposed 3 as the best number of clusters
* 4 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 6 as the best number of clusters

                 ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

MODEL 2: CLUSTERING

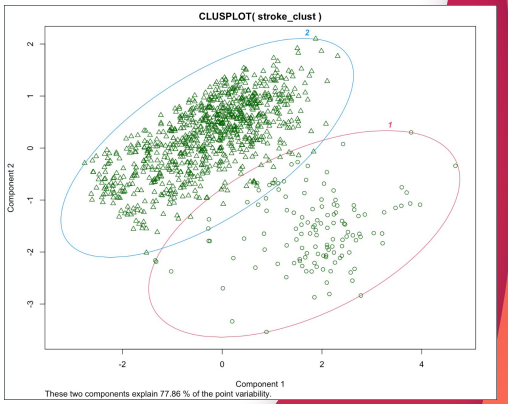


	age	avg_glucose_level	bmi
1	56.29085	202.81209	33.01240
2	40.77350	89.52545	28.26648

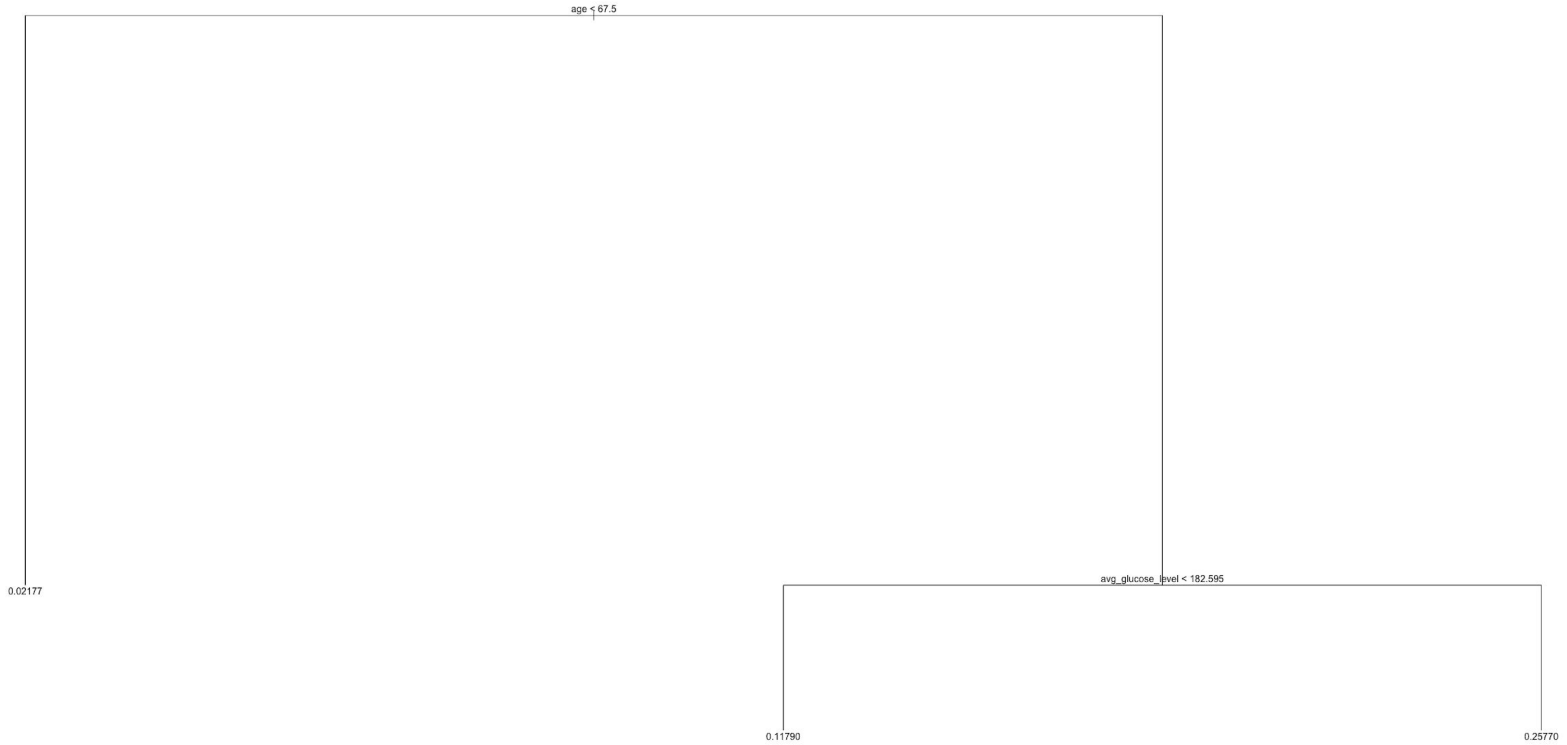
MODEL 2: EVALUATION AND RECOMMENDATION

- 2 Distinct Clusters:
  - Cluster 1: Older Adults - Obese (Diabetes)
  - Cluster 2: Middle Aged Adults - Overweight
- Model: Only used 20% of the whole data set, but the 2 clusters represent 78% of point variability
- Recommendation: Patients in either cluster should be taken with more precautionary care/early treatment to prevent a future stroke

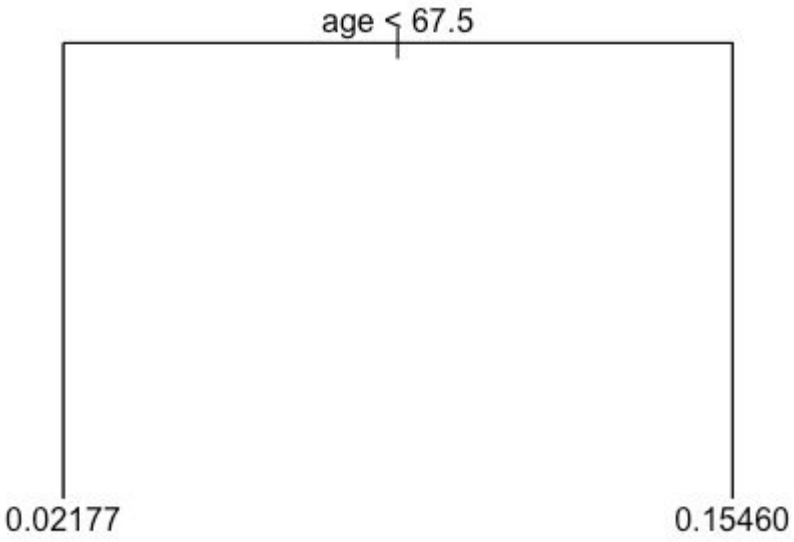
	age	avg_glucose_level	bmi
1	56.29085	202.81209	33.01240
2	40.77350	89.52545	28.26648



MODEL 3: DECISION TREE



MODEL 3: DECISION TREE



MODEL 3: EVALUATION AND RECOMMENDATION

- With a decision tree model, we are able to pinpoint different ‘splits’ of decisions or outcomes based on our data set.
- Model: In the model, we decided to input all of the variables but it only returned age and average glucose levels which is a strong indication that these two coefficients are the most influential in determining the possibility of somebody suffering from a stroke.
- Recommendation: If a patient that is over 67.5 years of age arrives with high glucose levels then it is strongly recommended that medical personnel do screenings in order to isolate symptoms that could result in a stroke.

COMPARISON OF PERFORMANCE

Logistic Regression:

- Accuracy: 85% for both training and testing
- Specificity is high and sensitivity is low

KMeans Clustering:

- 2 Clusters creating representing 78% of point variability
  - Overweight Middle Aged Adults
  - Obese Older Aged Adults
- Used a random sample of 1000 data points

Decision Tree:

- Unpruned Highest Probability (25.77%):
  - Age > 67.5
  - Average Glucose Level: 182.5
- Pruned Highest Probability (15.46%):
  - Age > 67.5

CONCLUSION

- Succeeded in creating multiple models to predict whether a person will have a stroke based on all variables
  - Best: Logistic Regression
    - Allowed us to compare coefficients between all variables to having a stroke
    - Allowed us to create an accurate predictive model to test if a person will have a stroke based on all variables
  - Worst: Decision Tree
    - Gave us an unpruned small tree that branched from age to average glucose level
    - Gave us an even smaller pruned tree only focusing on age
- In the future:
  - Remove variables that don't have a large impact and focus solely on the most impactful variables to create a more accurate model

THANK YOU FOR LISTENING!