Stroke Prediction

Jay Patel, Joey Peric, Cristian Gallegos

1. Problem Statement

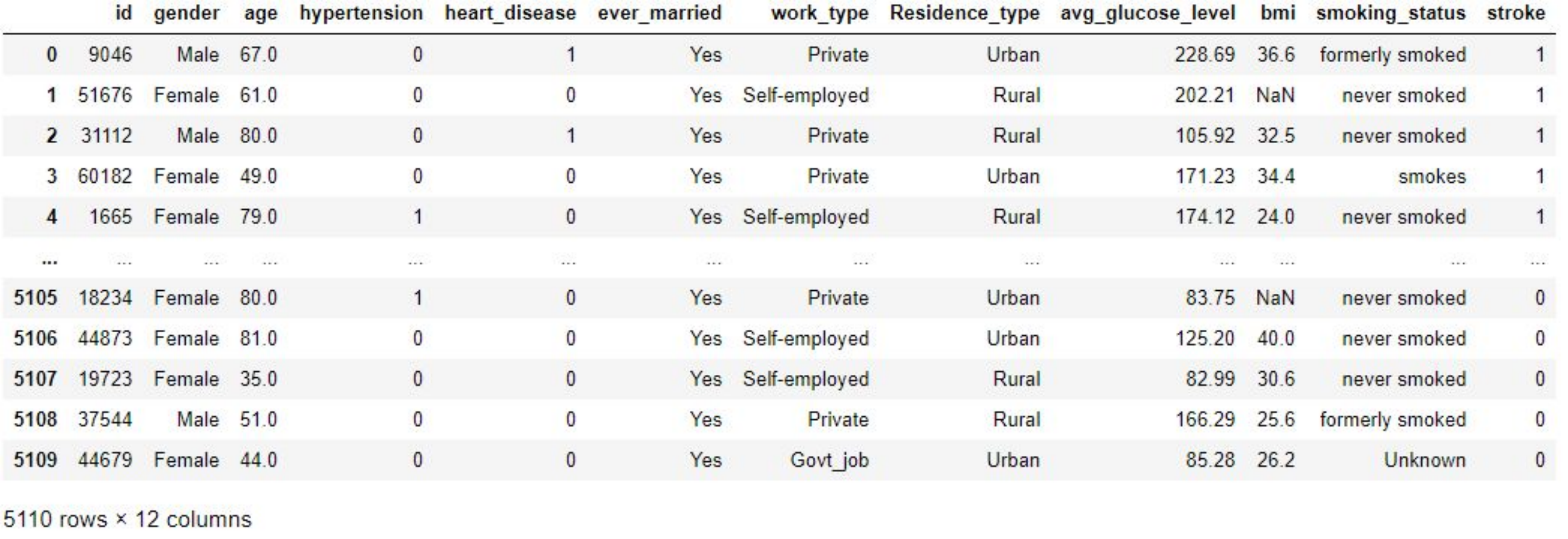
Predict strokes based upon multiple factors:

- Gender, Age, Heart disease, Marital status, Glucose Levels, BMI, Smoking Status, and Hypertension

Determine which predictor variables have a large impact on whether or not someone gets a stroke.

2. Data Sources

- The dataset that we used is Stroke Prediction from Kaggle.
- 5110 rows x 12 columns

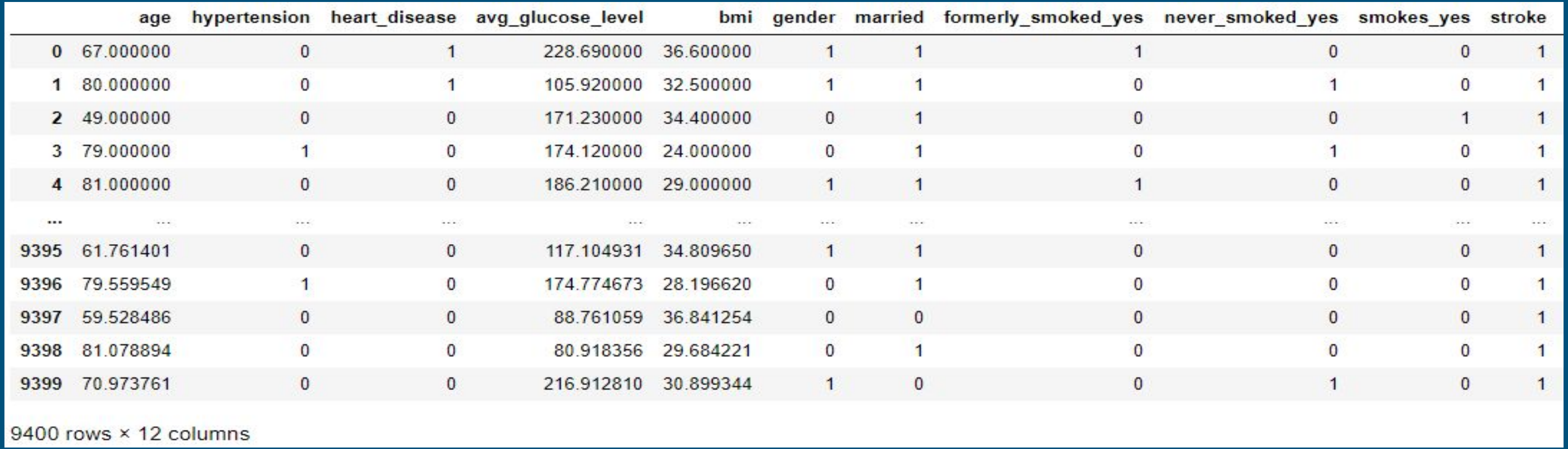


	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

3. Data Science Solution

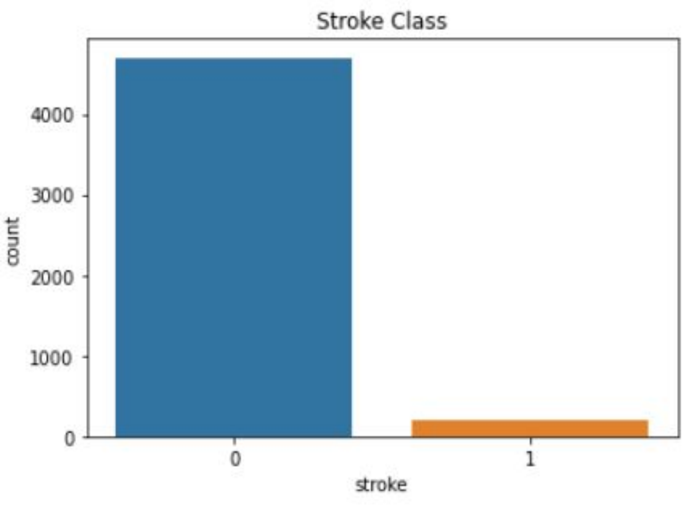
1. Used dummy variables for gender, marriage status, and smoking status.
2. Removed unneeded columns, id and residence type.
3. Removed observations that had a few null BMI.
4. Synthesized observations for stroke class.



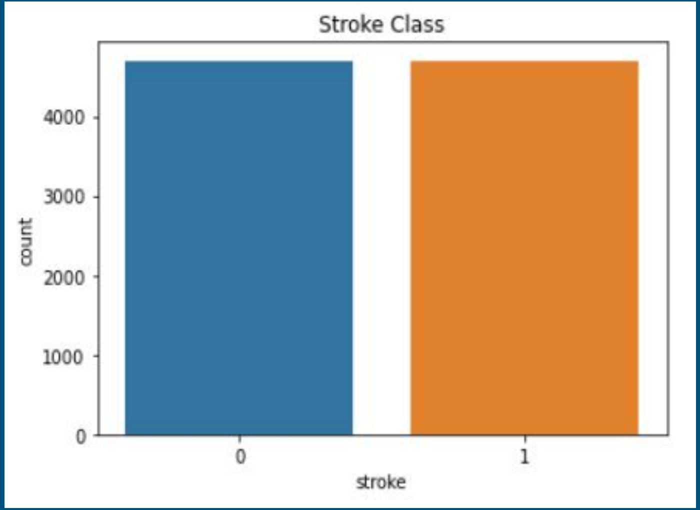
	age	hypertension	heart_disease	avg_glucose_level	bmi	gender	married	formerly_smoked_yes	never_smoked_yes	smokes_yes	stroke
0	67.000000	0	1	228.690000	36.600000	1	1	1	0	0	1
1	80.000000	0	1	105.920000	32.500000	1	1	0	1	0	1
2	49.000000	0	0	171.230000	34.400000	0	1	0	0	1	1
3	79.000000	1	0	174.120000	24.000000	0	1	0	1	0	1
4	81.000000	0	0	186.210000	29.000000	1	1	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...
9395	61.761401	0	0	117.104931	34.809650	1	1	0	0	0	1
9396	79.559549	1	0	174.774673	28.196620	0	1	0	0	0	1
9397	59.528486	0	0	88.761059	36.841254	0	0	0	0	0	1
9398	81.078894	0	0	80.918356	29.684221	0	1	0	0	0	1
9399	70.973761	0	0	216.912810	30.899344	1	0	0	1	0	1

9400 rows × 12 columns

3. Data Science Solution Continued...



SMOTE



3. Data Science Solution Continued...

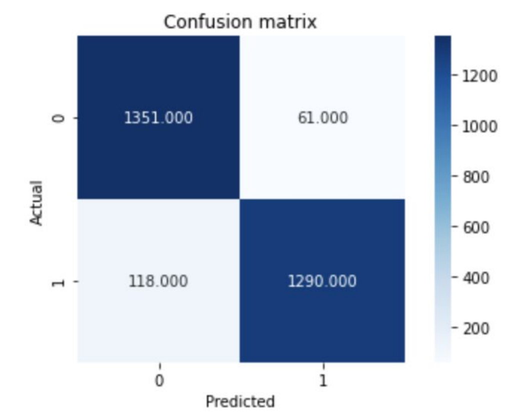
Classification and Clustering Techniques:

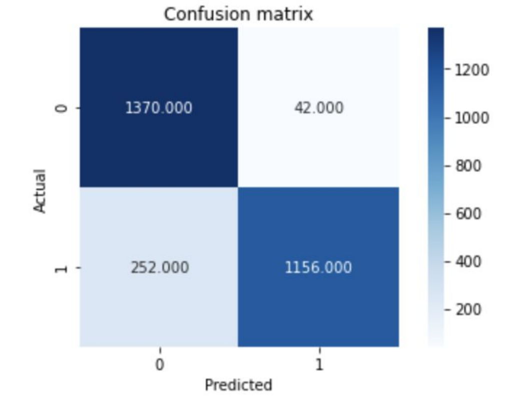
- Decision Trees
    - Entropy & Gini
- K-nearest Neighbors
    - N_neighbors = 4 & N_neighbors = 2
- Hierarchical
    - Complete Linkage & Single Linkage.
- K-means
    - 10 Initial Centroids & 20 Initial Centroids

Variable Combinations:

- Age and avg_glucose_level
- Married, Gender, Age
- Currently smoking, hypertension, heart_disease
- Heart disease and BMI
- All variables

4. Results (Best Classification Models)





Decision Tree with entropy

Predictor Variables: Marriage Status, Gender, Age

Accuracy: 93%, Error: 6%, Precision: 95%, Recall: 92%, F-1 Score: 93%

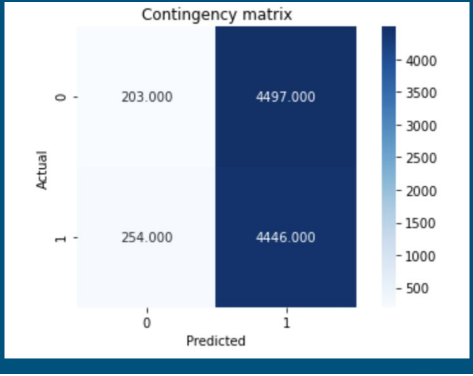
K-nearest neighbors with n_neighbors value of 2

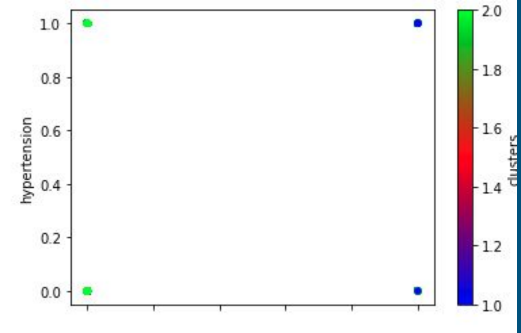
Predictor Variables: Marriage Status, Gender, Age

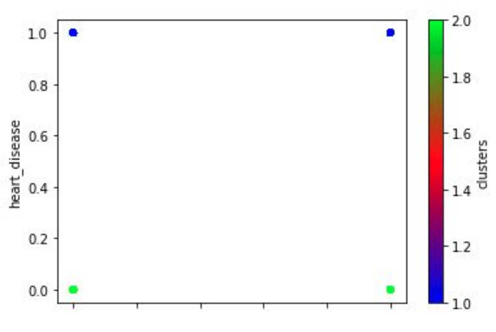
Accuracy: 89% , Error: 10%, Precision: 96%, Recall: 82%, F-1 Score: 88%

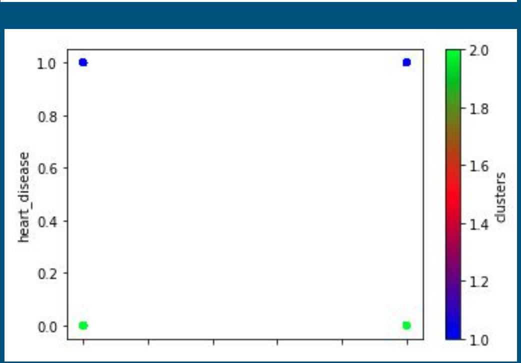
Results (Best Hierarchical Model)

Predictor Variables: current smoker, hypertension, heart disease.



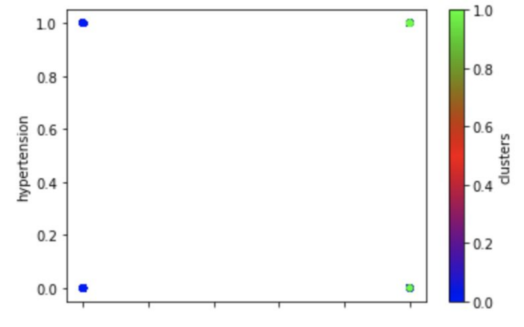


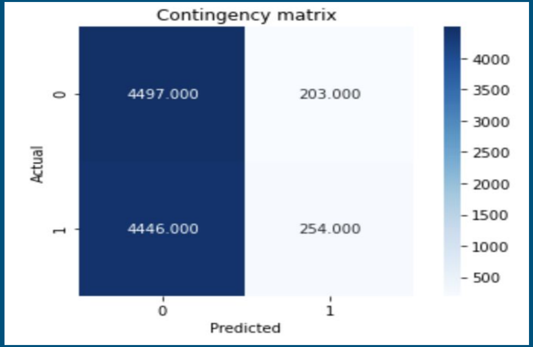




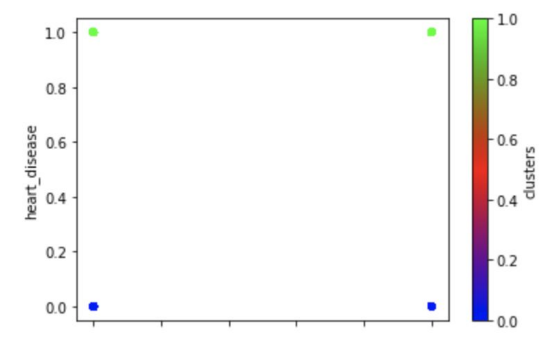
Results (Best K-Means Model)

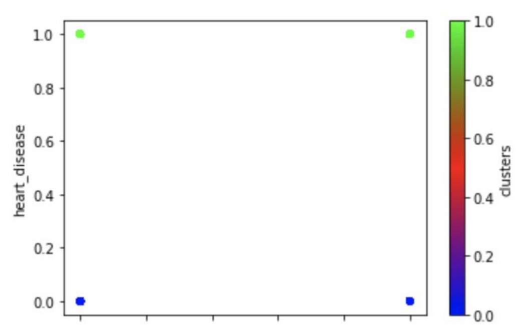
Predictor Variables: current smoker, hypertension, heart disease.




Contingency matrix

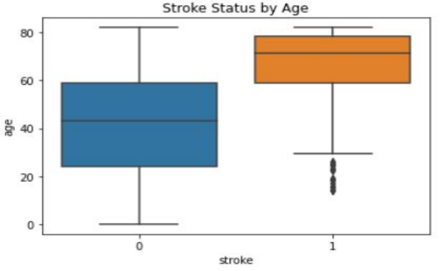
Actual \ Predicted	0	1
0	4497.000	203.000
1	4446.000	254.000

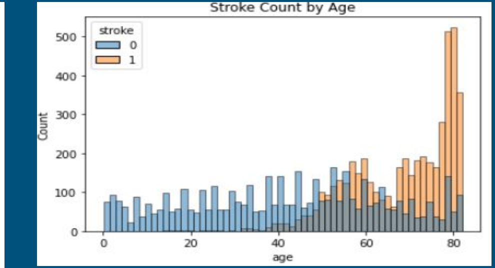


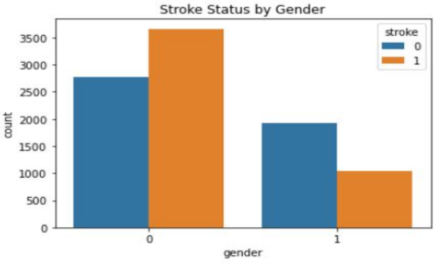


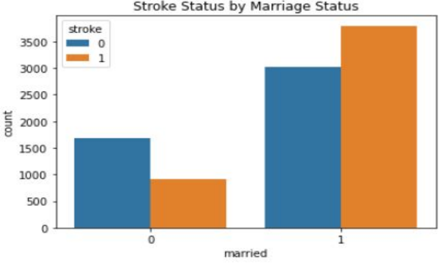
5. Conclusions

- Best overall model is Decision Tree with entropy.
- Clustering models overall do not model the data well.
- Age, Gender, and Marriage Status predict stroke the best using this dataset.
- Strokes are likely at age 50+
- Women are more likely to get strokes.
- People who are married are more likely to get strokes.


Stroke Status by Age
age
stroke


Stroke Count by Age
stroke
Count
age


Stroke Status by Gender
stroke
count
gender


Stroke Status by Marriage Status
stroke
count
married