Prediction of Stroke in Sub-Saharan Africa/United States

Group 16: Bruce Wen, Deirdre Edward, Isabelle Russo, Lynnette Jiang

Introduction

The incidence of stroke is rapidly increasing in low-income countries, particularly in Sub-Saharan Africa which has the “highest incidence, prevalence, and fatality” rates from stroke. However, few have studied the impact of risk factors specific for this region.

Based on more than 2,800 patients’ data from an unprecedented study on stroke in Nigeria and Ghana by the Stroke Investigative Research and Education Network (SIREN), we isolated what factors are most significant in the region. These included: hypertension, height, diabetes mellitus, stress, obesity, and heart problems. Keeping these factors in mind, we designed several algorithms to predict the incidence of stroke in Ghana and Nigeria, contrasting our results with the United States’ risk factors.

Aims

- Test Different Machine Learning Classification algorithms to predict stroke accurately
- Select best algorithm through training/testing on SIREN dataset to predict stroke in Sub-Saharan Africa
- Visualize, geographically, publicly available stroke data and significant risk factors in the United States
- Compare trained algorithms between datasets from the United States and Sub-Saharan Africa

Methodology

Datasets Utilized

CDC Stroke Mortality Data:

- Stroke Mortality among US adults above 35 years old, 2015-2017, used to generate US stroke maps

SIREN Stroke Data:

- Stroke data and 16 predictor variables for 2800+ patients in Nigeria and Ghana

Kaggle US Stroke Data:

- Stroke data and 10 predictor variables for ~43,000 patients in US

Data Cleaning Flow

Filter by County, Overall Stratification, and remove null values → Plot Geo-Scatters of Stroke Cases via plot.ly library

Convert .sav to .csv format → Drop rows with missing values and binarize categorical variables

Binarize categorical variables → Standardize continuous variables

Classification Algorithms, Optimization, and Evaluation

Classification Algorithms

- Logistic Regression
- Classification and Regression Trees (CART Decision Trees)
- K Nearest Neighbors

Optimization Process

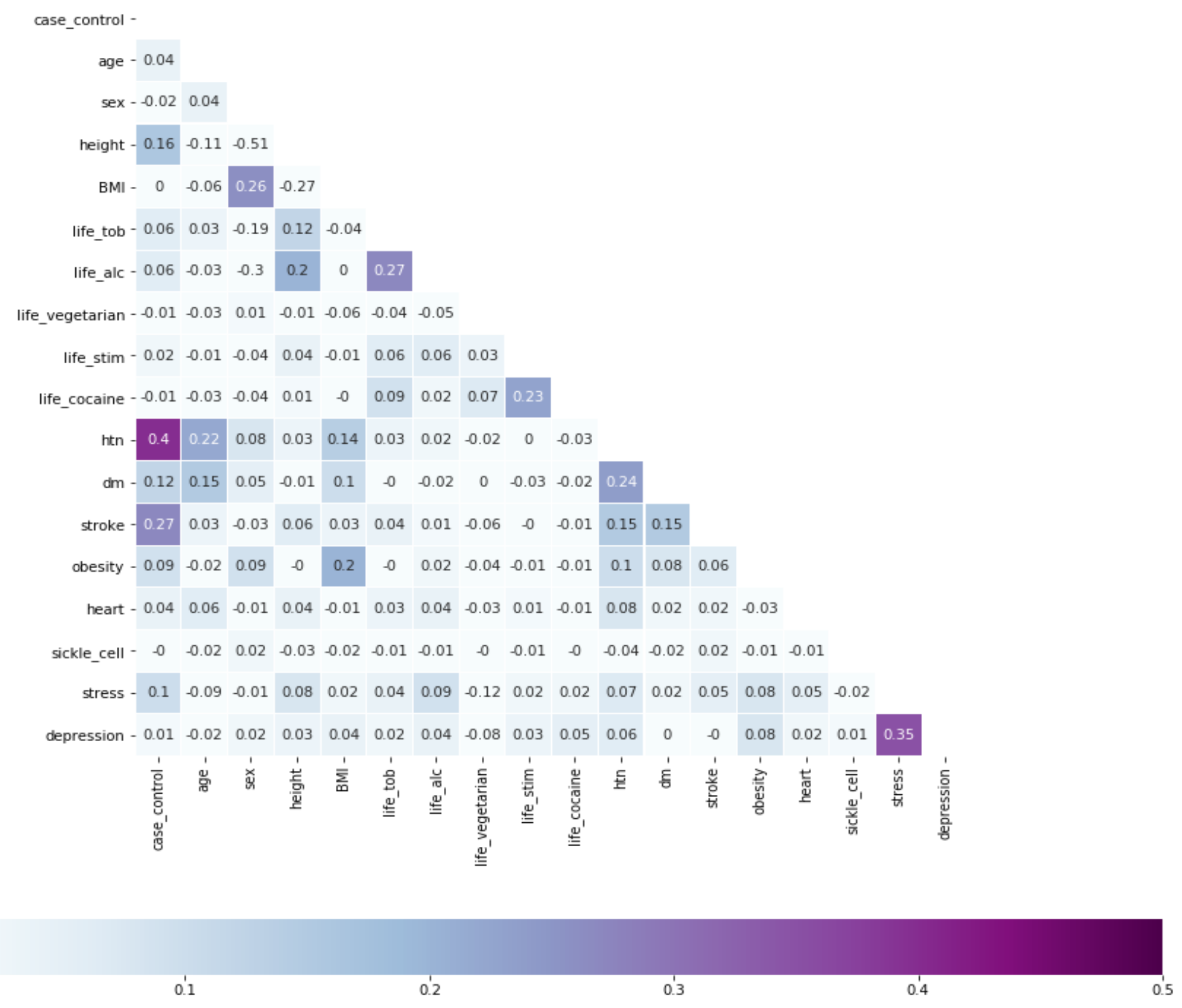
- 80% Training, 20% Testing
- Select variables from PCC table *OR* Loop through *Power Set of variables*

Evaluation of Algorithms

- Accuracy Rate
- Rate of True Positive and True Negative (*Confusion Matrix*)
- Receiver Operating Characteristic (*ROC*) Curve

Results

Pearson Correlation Coefficient Table

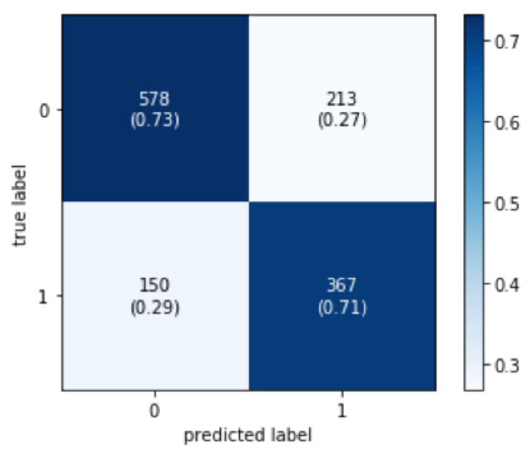


	case_control	age	sex	height	BMI	life_tob	life_alc	life_vegetarian	life_stim	life_cocaine	htn	dm	stroke	obesity	heart	sickle_cell	stress	depression
case_control																		
age	0.04																	
sex	-0.02	0.04																
height	0.16	-0.11	-0.51															
BMI	0	-0.06	0.26	-0.27														
life_tob	0.06	0.03	-0.19	0.12	-0.04													
life_alc	0.06	-0.03	-0.3	0.2	0	0.27												
life_vegetarian	-0.01	-0.03	0.01	-0.01	-0.06	-0.04	-0.05											
life_stim	0.02	-0.01	-0.04	0.04	-0.01	0.06	0.06	0.03										
life_cocaine	-0.01	-0.03	-0.04	0.01	-0	0.09	0.02	0.07	0.23									
htn	0.4	0.22	0.08	0.03	0.14	0.03	0.02	-0.02	0	-0.03								
dm	0.12	0.15	0.05	-0.01	0.1	-0	-0.02	0	-0.03	-0.02	0.24							
stroke	0.27	0.03	-0.03	0.06	0.03	0.04	0.01	-0.06	-0	-0.01	0.15	0.15						
obesity	0.09	-0.02	0.09	-0	0.2	-0	0.02	-0.04	-0.01	-0.01	0.1	0.08	0.06					
heart	0.04	0.06	-0.01	0.04	-0.01	0.03	0.04	-0.03	0.01	-0.01	0.08	0.02	0.02	-0.03				
sickle_cell	-0	-0.02	0.02	-0.03	-0.02	-0.01	-0.01	-0	-0.01	-0	-0.04	-0.02	0.02	-0.01	-0.01			
stress	0.1	-0.09	-0.01	0.08	0.02	0.04	0.09	-0.12	0.02	0.02	0.07	0.02	0.05	0.08	0.05	-0.02		
depression	0.01	-0.02	0.02	0.03	0.04	0.02	0.04	-0.08	0.03	0.05	0.06	0	-0	0.08	0.02	0.01	0.35	

Machine Learning Algorithms on SIREN dataset

Logistic Regression

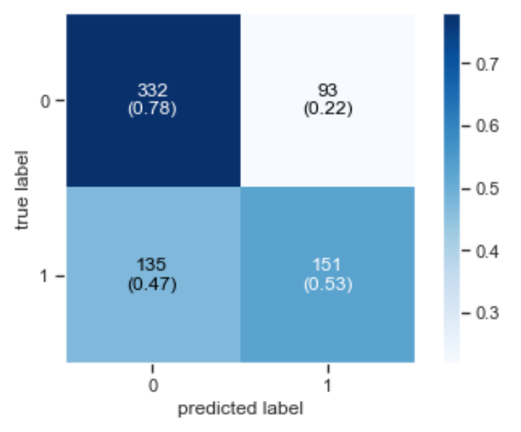
Confusion Matrix for selected variables: Hypertension, height, history of stroke, stress, diabetes



Mean cross validation score: 0.7002
Score without cv: 0.7156

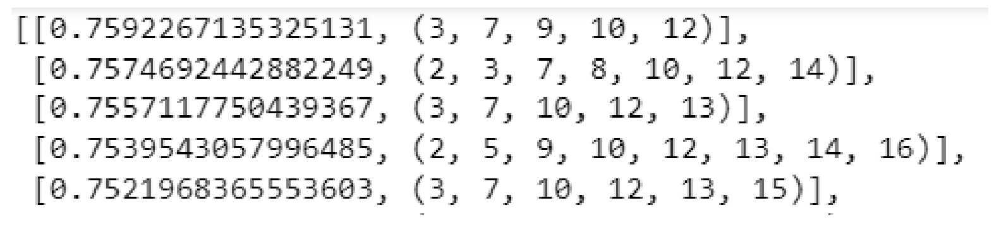
K Nearest Neighbors

Confusion Matrix for selected variables: Hypertension, height, history of stroke, stress, diabetes, obesity, tobacco usage, heart disease



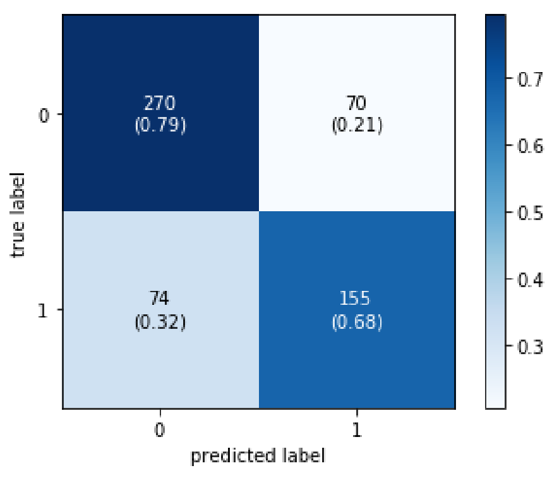
Mean cross validation score: 0.6225
Score without cv: 0.7467

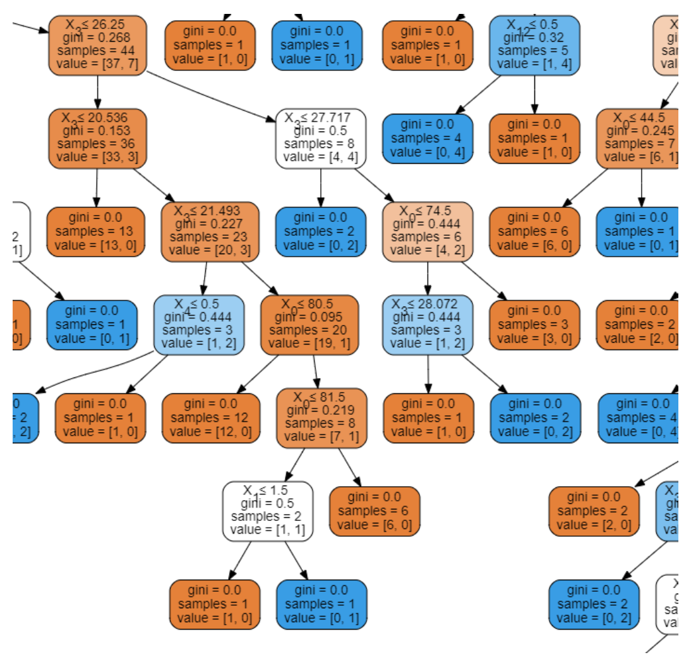
Decision Tree



```
[[0.7592267135325131, (3, 7, 9, 10, 12)],
 [0.7574692442882249, (2, 3, 7, 8, 10, 12, 14)],
 [0.7557117750439367, (3, 7, 10, 12, 13)],
 [0.7539543057996485, (2, 5, 9, 10, 12, 13, 14, 16)],
 [0.7521968365553603, (3, 7, 10, 12, 13, 15)],
```

Top 5 Accuracy Rates among Variable Combos

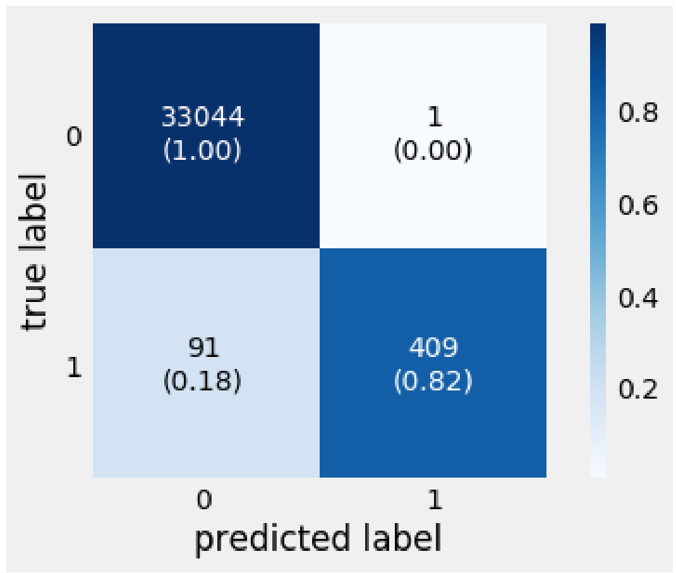




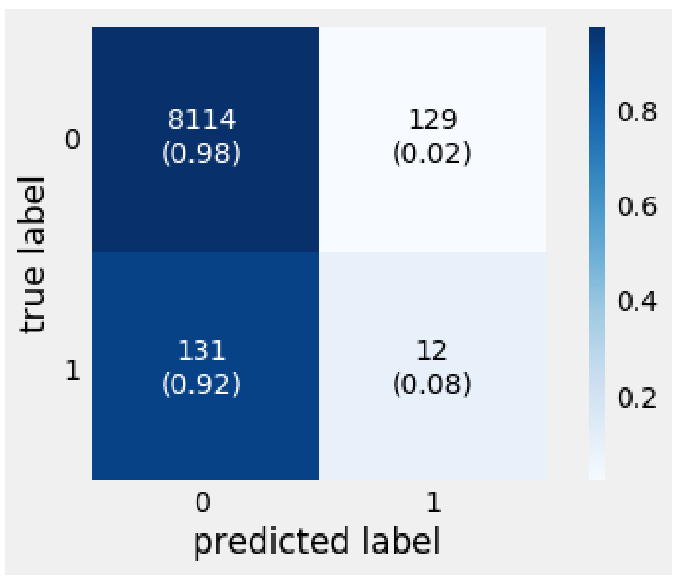
Confusion Matrix for selected variables: *Sex, height, vegetarian, stimulant, cocaine, hypertension, past instance of stroke, heart disease*

Mean cross validation score: 0.7035
Score without cv: 0.7879

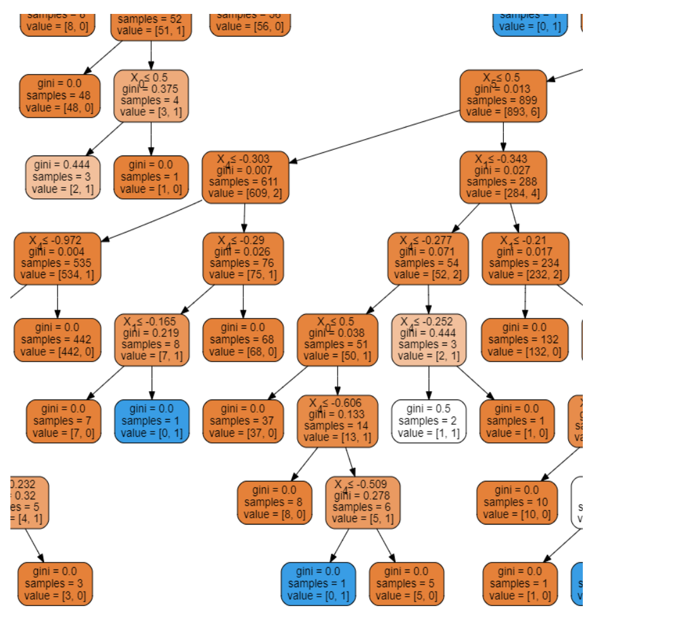
CART Decision Tree on US Kaggle dataset



Confusion Matrix on Training Dataset. Selected Variables: Gender, Age, Hypertension, Heart Disease, BMI, tobacco user



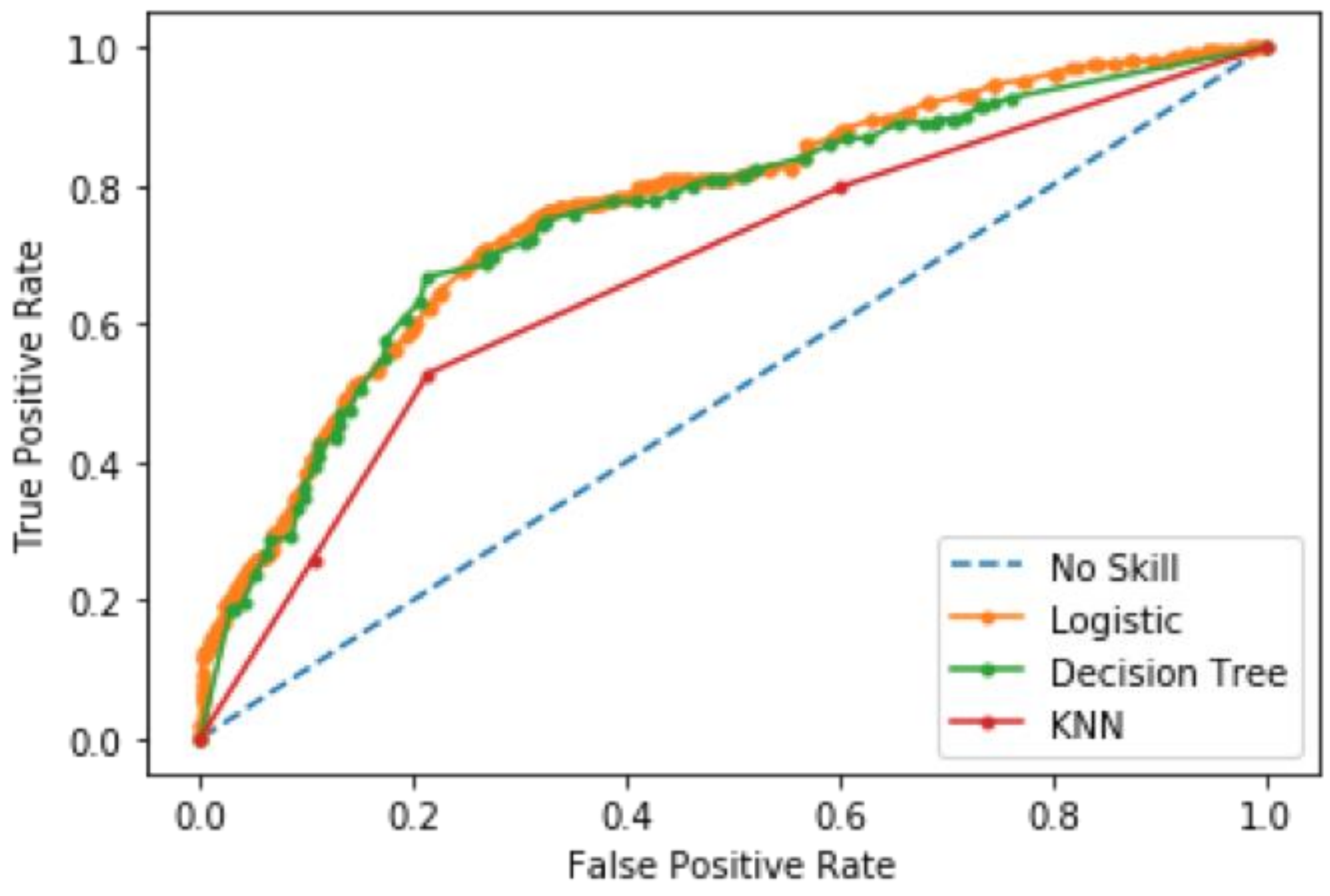
Confusion Matrix on Testing Dataset. Same 6 variables



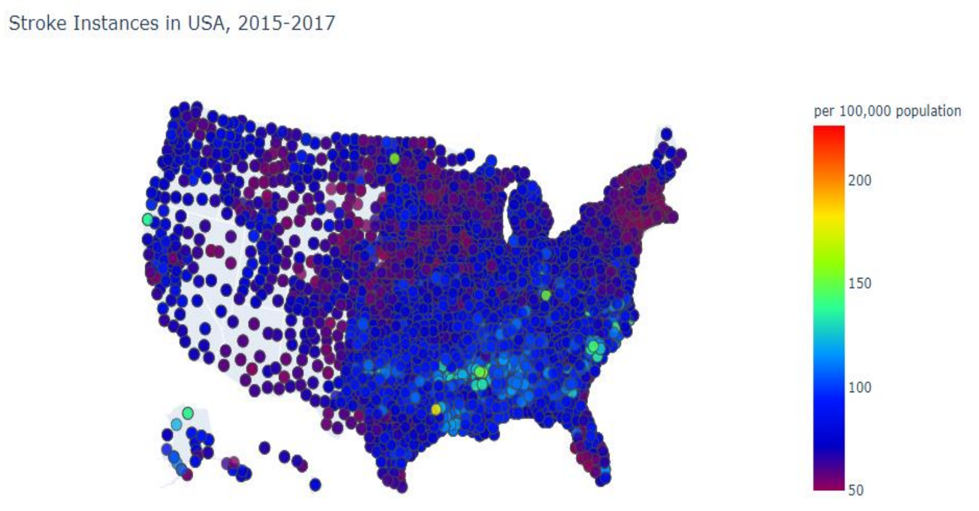
Decision Tree (segment)

Mean cross validation score: 0.9634
Score without cv: 0.999

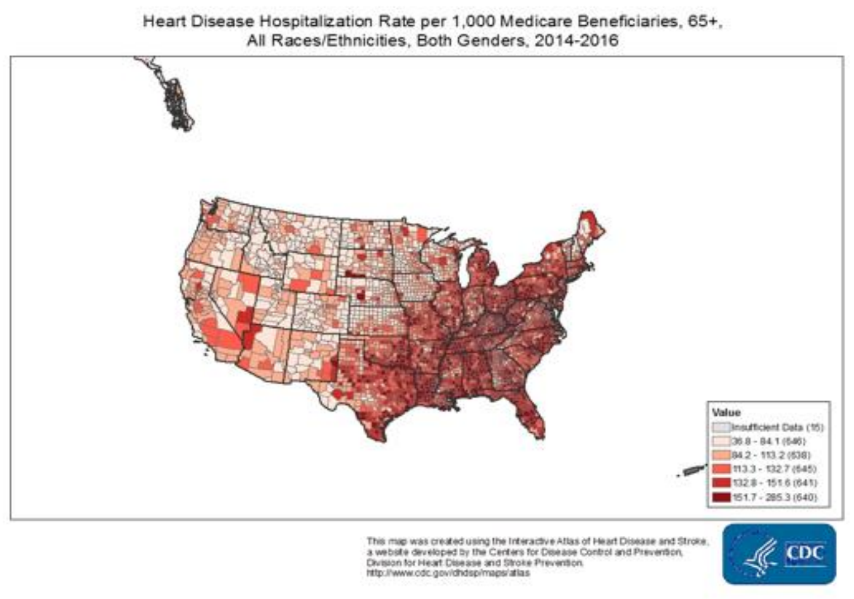
ROC Curve: Compare Optimized Algorithms (SIREN data)

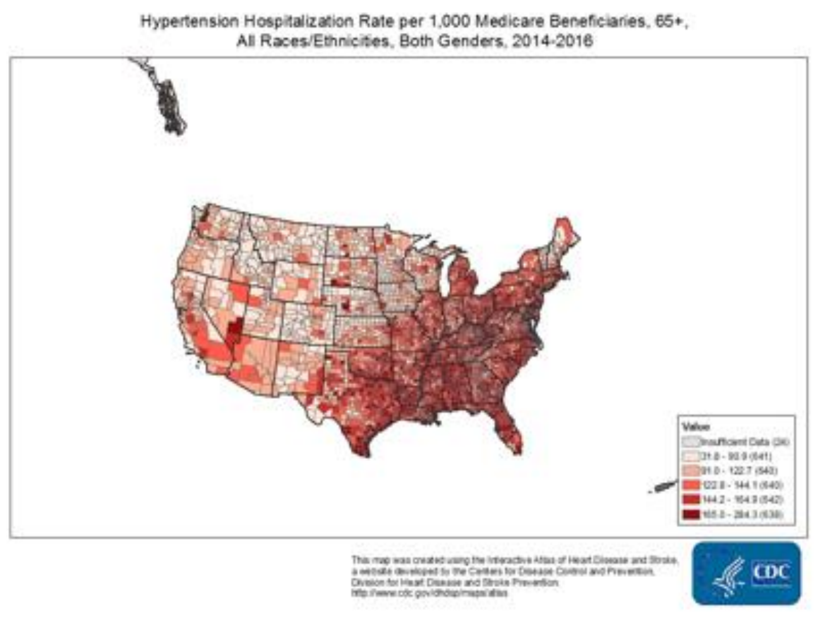


US Stroke and Related Factors Maps



Plotted via *plot.ly* library. Data Source: CDC


Heart Disease Hospitalization Rate per 1,000 Medicare Beneficiaries, 65+, All Races/Ethnicities, Both Genders, 2014-2016
Value
Insufficient Data (15)
36.8 - 84.1 (646)
84.2 - 113.2 (638)
113.3 - 132.7 (645)
132.8 - 151.6 (641)
151.7 - 285.3 (640)
This map was created using the Interactive Atlas of Heart Disease and Stroke, a website developed by the Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention. http://www.cdc.gov/dhdsp/maps/atlas
CDC


Hypertension Hospitalization Rate per 1,000 Medicare Beneficiaries, 65+, All Races/Ethnicities, Both Genders, 2014-2016
Value
Insufficient Data (34)
31.0 - 90.9 (641)
91.0 - 122.7 (640)
122.8 - 144.1 (640)
144.2 - 164.9 (642)
165.0 - 284.3 (638)
This map was created using the Interactive Atlas of Heart Disease and Stroke, a website developed by the Centers for Disease Control and Prevention, Division for Heart Disease and Stroke Prevention. http://www.cdc.gov/dhdsp/maps/atlas
CDC

Discussion and Conclusion

- 3 Machine Learning Algorithms were implemented, and optimized.
- Algorithms were compared by accuracy rates (with/without Cross Validation), Confusion Matrix and ROC Curve.
- CART Decision Tree and Logistic Regression both produced high True Positive and True Negative Rates, along with relatively high accuracy
- Based on analysis, some important factors affecting stroke rates are: hypertension, history of stroke, height, heart disease, obesity
- Relationships between some factors and stroke instances in the USA are visualized

Future Work

- Note the low True Positive rate in the algorithm on US Kaggle dataset. The likely reason is the low number of stroke cases in the dataset relative to non-stroke cases. Changing the proportion could help increase the True Positive prediction accuracy. This could involve comparing data from SIREN dataset.
- PCC may not be optimal for categorical variables. Attempt Spearman’s rank correlation coefficients.
- Visualize African stroke distribution based on geography and time, when data is available.

References

Owolabi, M., Sarfo, F., Akinyemi, R., Gebregziabher, M., Akpa, O., Akpalu, A., … Ovbiagele, B. (2018). Dominant modifiable risk factors for stroke in Ghana and Nigeria (SIREN): a case-control study. *The Lancet Global Health*, *6*(4). doi: 10.1016/S2214-109X(18)30002-0

Sanuade, O. A., Dodoo, F. N.-A., Koram, K., & Aikins, A. D.-G. (2019). Prevalence and correlates of stroke among older adults in Ghana: Evidence from the Study on Global AGEing and adult health (SAGE). *Plos One*, *14*(3). doi: 10.1371/journal.pone.0212623

Yan, L. L., Li, C., Chen, J., Miranda, J. J., Luo, R., Bettger, J., … Wu, Y. (2016). Prevention, management, and rehabilitation of stroke in low- and middle-income countries. *ENeurologicalSci*, *2*, 21–30. doi: 10.1016/j.ensci.2016.02.011