CMSC/STAT 119 Final Project: Prediction of Stroke in Sub-Saharan Africa/United States

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

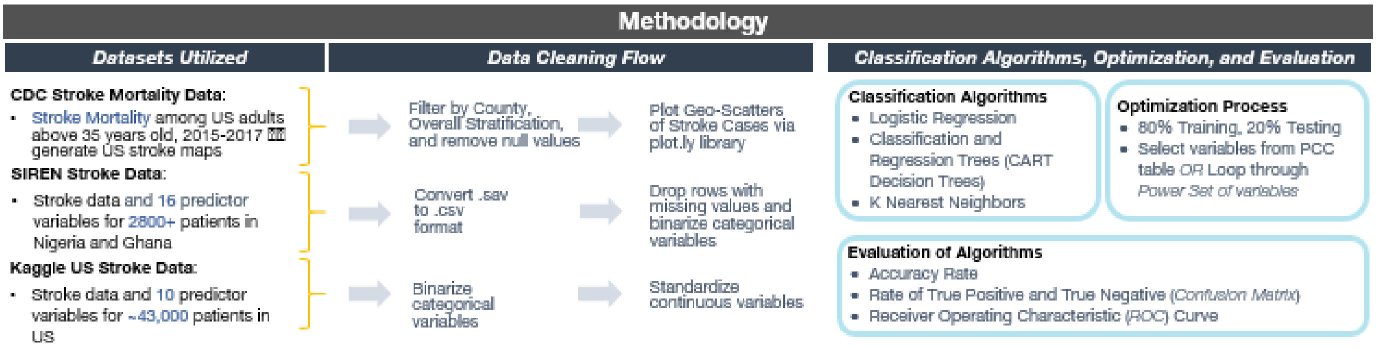
Introduction

While stroke incidence is decreasing in high-income countries, it is on the rise in low and middle-income countries. [4] Sub-Saharan Africa has the "highest incidence, prevalence, and fatality" from stroke as compared to other regions. [2] However, few have studied risk factors particular to the region. We have obtained the relevant stroke and risk factors data for 2800+ patients in Nigeria and Ghana based on data collected from Stroke Investigative Research and Education Network (SIREN). This is a non-public, research dataset based on Deirdre Edward's research with Dr. Steffen Sammet from UChicago Medicine. The size of this dataset is unprecedented for stroke patients from Africa. Contrasting our findings with an American dataset, we isolated factors that were important in the African context. These included: hypertension, height, diabetes mellitus, stress, obesity, and heart problems. In this project, we seek to identify the most important risk factors in Ghana and Nigeria in order to construct algorithms that can predict stroke incidence in the region.

In low-to-medium-income countries (LMICs), stroke has important social and economic consequences. The increasing burden of stroke calls for better methods of tracking current trends, particularly in LMICs. [4]
The **aims** of this project are to *1) Implement and evaluate different machine learning classification algorithms to predict stroke accurately, 2) Select the best algorithm through training/testing on SIREN dataset to predict stroke in Sub-Saharan Africa, 3) Visualize geographically, publicly available stroke data and significant risk factors in the United States, and 4) Compare trained algorithms between datasets from the United States and Sub-Saharan Africa.*

Methodology

Fig 1: Methodology Flow



Methodology

Datasets Utilized

CDC Stroke Mortality Data:
- Stroke Mortality among US adults above 35 years old, 2015-2017 □ 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

Fig. 1 highlights the workflow of the project. *3 datasets* were utilized in this project.

The **first** is CDC Stroke Mortality Data from 2015-2017, which was obtained at https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Stroke-Mortality-Data-Among-US-Adults-35-by-State-/v246-z5tb. This dataset, *Stroke_Mortality_Data_Among_US_Adults__35___by_State_Territory_and_County___2015-2017.csv* includes columns with geographic information and stroke mortality counts. Data cleaning was done through filtering out State level data such that only county level data remains. Null values were removed using *df = df.loc[df['Data_Value'].notnull()]*. The data was used to generate geo-scatter plots in Figures 12 and 13 below.

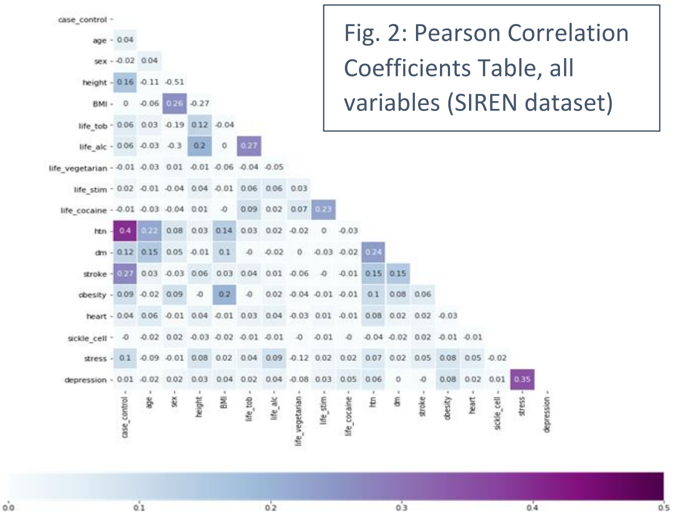
Second, the SIREN stroke dataset forms the bulk of this project's analysis. The SIREN dataset is based on a study that involved patients from 15 sites in Nigeria and Ghana. It includes *1133 stroke patients and 1709 control cases, for a total of 2842 rows*. There are *16 columns* in the dataset detailing the risk factors: age, sex, height, BMI, tobacco usage, alcohol usage, vegetarian, stimulants usage, cocaine usage, hypertension, diabetes mellitus, past instance of stroke, obesity, heart disease, sickle cell, stress, and depression. The protocol of the SIREN multicenter study has been detailed previously by Akpalu et al. The relative balanced number of stroke to control cases allow for our machine learning algorithms to develop higher True Positive rates as compared to the American data from Kaggle.

Third, a sample of *~43,000* USA patients with/without stroke is obtained from Kaggle https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data. *train_2v.csv* contains stroke data and *10 predictor variables*, namely, gender, age, hypertension, heart disease, marriage status, work type, residence type (urban/rural), average glucose level, BMI, and smoking status. This was the most complete and detailed dataset our group was able to procure, but some limitations remain: the predictor variables do not entirely match with the SIREN dataset, and there are much fewer stroke cases than control cases. This limits our capability to evaluate the generalization of the predictions based on the SIREN dataset to the Kaggle dataset. Nevertheless, it is worthy of analysis in this project's context.

Data Cleaning: All 3 datasets required substantial data cleaning for the purposes of our analyses. The CDC stroke mortality data was filtered by county level, overall stratification, and removing all state level data since they overlapped with the county data. The SIREN stroke data was converted from .sav to .csv format, and binarized. Finally, the Kaggle US Stroke data categorical variables were binarized and the continuous variables were standardized. Certain columns, including residence type and work type were dropped because of the lack of confidence in associating these variables with relevant risk factors such as stress.
Classification Algorithms: We select 3 classification algorithms. Besides K Nearest Neighbors and Logistic Regression, classification and regression tree (CART) classification tree algorithm was utilized, because of the high number of categorical variables present in both SIREN dataset and US Kaggle dataset. CART decision trees are generally a favorable classification algorithm for cases where a large

number of categorical variables are involved. For all analysis, we use an **80% : 20% training : testing split**. To ***evaluate*** the algorithms and decide the risk factors most associated with stroke, we generated the accuracy rates, confusion matrix, and ROC curves for each test.

SIREN (Sub-Saharan Africa) Dataset Results



	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	

Fig. 2: Pearson Correlation Coefficients Table, all variables (SIREN dataset)

Fig. 2 depicts the Pearson Correlation Coefficients (PCC) table, calculated for each of the 16 variables against the other 15 variables. The generation of the PCC table allows us to **note which variables may be particularly good predictors of stroke in our subsequent machine learning algorithms optimization**. Of particular focus is the first column "*case_control*" and its correlation with the other 15 variables. We note some of the highest correlation coefficients in this column: **hypertension at 0.4, past incidence of stroke at 0.27, height at 0.16**, and other factors such as diabetes mellitus and obesity. These principle risk factors are a focus for our subsequent machine learning algorithms. Notably, height has positive correlation with stroke, 0.16. This *contradicts literature studies*, which usually demonstrate the correlation to be negative [5] . One reason is Sub-Saharan Africans are different from North American/European test subjects. Yet, it could also possibly be because of biased sampling in the 1100 stroke/1700 control cases. Nevertheless, this is worthy of further investigation and future study.

Algorithm 1: Logistic Regression

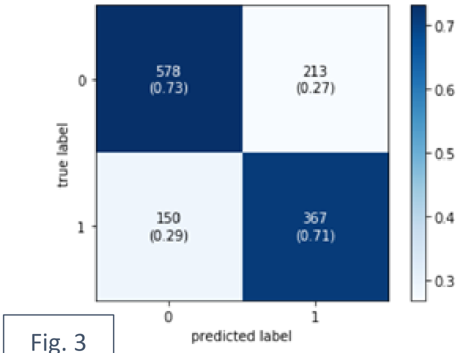


Fig. 3

A logistic regression was conducted using the *scikitlearn.linear_model LogisticRegression* library. Through testing of the top correlated variables, we generate the highest accuracy rate using the following 5 selected variables: **hypertension, height, history of stroke, stress, and diabetes**. Furthermore, the mean cross validation score conducted with **5-fold** cross validation is 0.7002, while the score without cross validation is 0.7156. These scores were generated using the *sklearn.model_selection cross_val_score* library.

A confusion matrix is generated for this optimal combination of predictors, using the *mlxtend.plotting plot_confusion_matrix* library. We observe true negative, false positive, false negative, true positive rates of 0.73, 0.27, 0.29 and 0.71 respectively. These results are significantly improved from a random coin toss, which will generate 0.5 for all.

Algorithm 2: K Nearest Neighbors

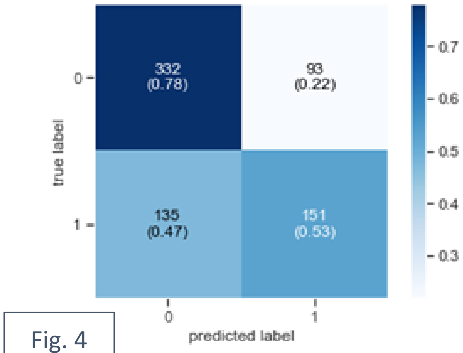


Fig. 4

We next applied the K-Nearest Neighbors algorithm to predict the incidence of stroke in Ghana and Nigeria. Based on our correlation coefficient table, we selected variables with high associations with stroke. Next, we standardized all variables in order to give equal weighting to each. Because many variables were categorical, we transformed them into dummy variables, allowing us to use a Euclidean distance. In order to optimize the algorithm, we tried many combinations of variables as well as odd numbers for K between 1 and 7. Relying on the variables of hypertension, height, history of stroke, stress, diabetes mellitus, obesity, tobacco usage, and heart disease, we got our highest mean cross validation score of 0.6225. Number of neighbors K = 5. A confusion matrix is plotted as was done in algorithm 1. We observe optimal **true negative, false positive, false negative, true positive rates of 0.78, 0.22, 0.47 and 0.53 respectively**.

Algorithm 3: CART Decision Tree

The CART decision tree builds a classification model in the form of a tree structure with decision nodes and leaf nodes. At each leaf node, we produce a Gini impurity value. The ending leaf of each branch has a Gini index of zero. Decision trees are generally good for categorical variables, which suits the SIREN dataset well since it contains 13 categorical variables out of 16 variables. The Decision Tree algorithm is implemented using the *sklearn 'tree'* library. A blind run of the decision tree algorithm including all 16 variables, trained with the training dataset, produced an accuracy rate of ~0.65 on the testing dataset. Due to the intrinsic suitability of the CART decision tree algorithm in classifying from categorical variables, we conduct more thorough analysis for this algorithm to determine the subset of variables that predict stroke with the highest accuracy. A power set of the 16 risk variables is generated using the *itertools* library, importing *chain, combinations*. The result is a list containing 65,535 tuples, representing all subsets of the 16 variables. The tree classifier is then implemented for each of these subsets, and the top 5 prediction accuracies along with the columns selected are generated in Table 1.

Prediction Accuracy	Columns Selected in Algorithm
0.7592	Height, Vegetarian, Cocaine User, Hypertension, Past Stroke
0.7574	Sex, Height, Vegetarian, Stimulant User, Hypertension, Past Stroke, Sickle Cell
0.7557	Height, Vegetarian, Hypertension, Past Stroke, Obesity
0.7540	Sex, Tobacco User, Cocaine User, Hypertension, Past Stroke, Obesity, Sickle Cell, Depression
0.7522	Height, Vegetarian, Hypertension, Past Stroke, Obesity, Stress

Table 1

Table 1 is insightful as it demonstrates some of the top variables for predicting stroke, another way to evaluate which factors are especially relevant besides the PCC table generated in Fig. 2. The top accuracy rate is 0.7592.
For a visual understanding of the plotted decision tree, Fig. 6 shows a zoomed-in segment of the tree, while Fig. 5 shows the entire decision tree. The trees are generated using the *graphviz* library.

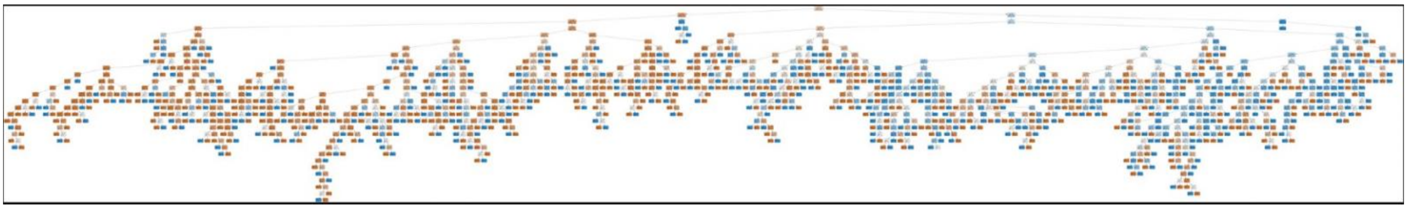


Fig. 5

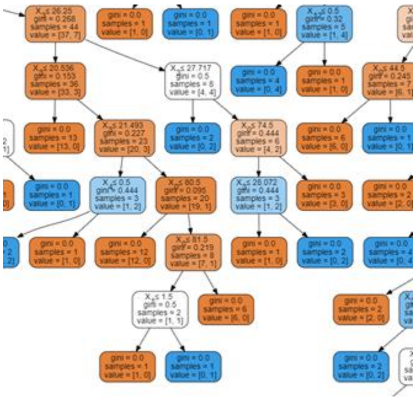


Fig.6

Finally, using the columns that generated the accuracy rate of 0.7592 in the testing dataset, we plot the confusion matrix in Fig. 7. The optimal **true negative, false positive, false negative, true positive rates of 0.79, 0.21, 0.32 and 0.68** respectively. The obtained true negative is highest among all algorithms, at **0.79**.

We now provide a comparison of all 3 algorithms. The mean cross validation and score without cross validation, generated with *scikit learn* library and inbuilt function for each algorithm, is shown in Table 2. From Table 2, the CART decision tree is shown to generate the most accurate prediction for stroke.

The **confusion matrices** generated in Fig. 3, 4, and 7 are especially important in our comparison. Fig. 3, or the confusion matrix generated by the optimized Logistic Regression algorithm, gives a false negative rate of 0.29 (lowest among all algorithms). Comparatively, the confusion matrix generated in Fig. 7 by the optimized Decision Tree algorithm gives a false positive rate of 0.21 (lowest among all algorithms).

The **ROC curve**, generated in Fig. 8, is an easy way to visualize the performance of each algorithm. Logistic Regression and Decision Tree are relatively close, while the optimized kNN algorithm performs significantly worse.

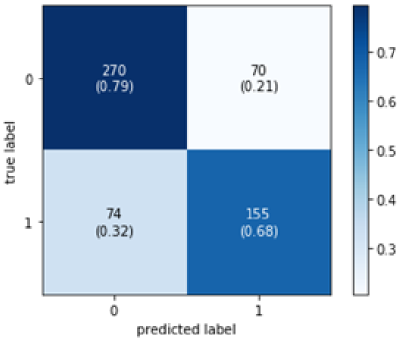


Fig. 7

Comparison and Discussion of the 3 Implemented Algorithms

Given the similarity in performance between the optimized logistic regression algorithm and the decision tree algorithm, we have to consider the **tradeoffs** in using either algorithm. Based on the higher false positive rates for the logistic regression algorithm, there is higher **Type I error**, which could mean *more wasted resources in treating and hospitalizing these patients who do not actually have stroke*. If the decision tree algorithm were to be used instead, there is higher false negative rate of 0.32 compared to 0.29 in logistic regression, which means higher **Type II error**, which could mean *less identification of patients who actually have stroke*. Decision on usage of algorithms should be based on public health care needs.

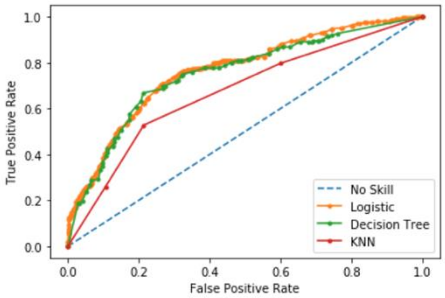


Fig. 8

	Logistic Regression	Decision Tree	K Nearest Neighbors
Mean cross validation score	*0.7002*	*0.7035*	0.6225
Score without CV	0.7156	*0.7879*	0.7467

Table 2

US Kaggle Dataset Results

Fig. 11

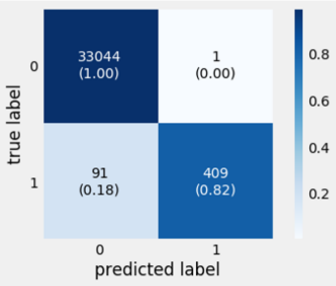


Fig. 9

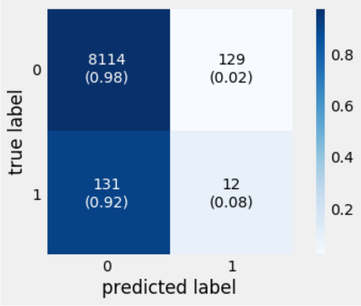


Fig. 10

We next conduct similar analysis, using decision tree algorithm (plotted tree in Fig. 11), on the US Kaggle dataset. The optimal combination with highest accuracy rate is the incorporation of the 6 variables: gender, age, hypertension, heart disease, BMI, and tobacco user. This generated a mean cross validation score of 0.9634. The confusion matrices for training and testing dataset (split in 80% : 20% as for all datasets) are plotted in Fig. 9 and 10 respectively. Interestingly, we note a **comparatively low true positive/high false negative** in Fig. 10, the testing dataset, compared to the training dataset. This is despite a very high cross validation score demonstrating the accuracy of the algorithm. A reason for the low true positive is the lack of stroke cases relative to control cases in the dataset, skewing the machine learning algorithm against predicting for stroke. To resolve this in future studies, a few methods are proposed: biased resampling, or 'matching' stroke data from other datasets to complement this dataset.

US Stroke Map Visualization and Related Factors

In this section, we attempt to **visualize** stroke patients and related factors to complement this study. Unfortunately, no African geospatial data was publicly available, so we plotted only the US stroke map. Fig. 12 and Fig. 13 were generated using the *plot.ly* library. Fig. 12 filters only stroke densities 50/100,000 and above, while Fig. 13 filters only stroke densities of 100/100,000 and above. Note that stroke cases are prevalent especially in Southeast regions of USA. These geo-scatter plots are placed beside Heart Disease geographical distribution (Fig. 15) and Hypertension geographical distribution (Fig. 16). These figures are obtained from CDC. Interestingly, we observe **very similar spatial distributions as from our statistical analyses of the significant risk factor variables**. The southern region incidence of stroke geographically corresponds to the risk factors.

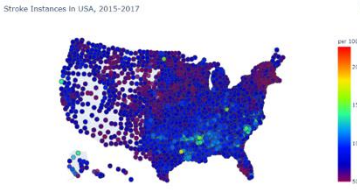


Fig. 12

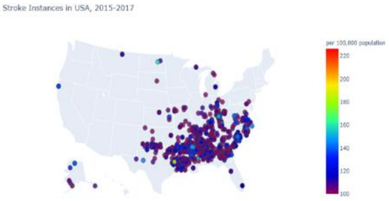


Fig. 13

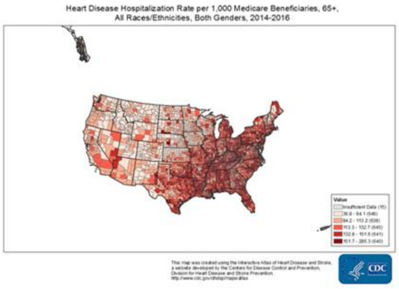

Heart Disease Hospitalization Rate per 1,000 Medicare Beneficiaries, 65+,
All Races/Ethnicities, Both Genders, 2014-2016
Value
CDC

Fig. 14

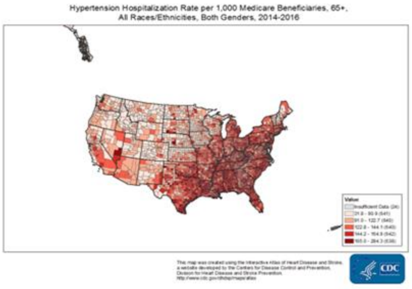

Hypertension Hospitalization Rate per 1,000 Medicare Beneficiaries, 65+,
All Races/Ethnicities, Both Genders, 2014-2016
Value
CDC

Fig. 15

Conclusion and Future Work

3 Machine Learning Algorithms (Logistic Regression, K Nearest Neighbors, CART Decision Tree) were implemented, and optimized. The 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. There is room for further evaluation of both algorithms because of the **relative differences in false positive and false negative rates for both**. Based on analysis, some important factors affecting stroke rates are **hypertension, history of stroke, height, heart disease, and obesity**. Lastly, relationships between some factors and stroke instances in the USA **are visualized**.
For *future work*, note the low True Positive/ high False Negative rate in the testing dataset (Fig. 10) algorithm on US Kaggle dataset. The likely reason is the low number of stroke cases in the dataset relative to non-stroke cases, as discussed. Changing the proportion could help increase the True Positive prediction accuracy. This could involve **biased resampling, or incorporation of data from other datasets to supplement this dataset**. There is room for further research, upon obtaining the correct data, to determine the extent of **ability to generalize** the predictor variables from Africa to the United States. **Height** as a factor, for example has shown a positive correlation in the PCC table (Fig. 2) for the African SIREN dataset but is traditionally negative correlated. Further study can be done on this front to investigate. Finally, **PCC may not be the most optimal for categorical variables. We suggest attempting Spearman's rank correlation coefficients**. When the data is available, it will also be insightful to visualize African stroke distribution based on geography and time and compare the maps against geospatial predictor variables.

Group Member Contributions: All members contributed to the entire data science process and preparation of poster/report.

References

[1]Akpalu, Albert, Fred Stephen Sarfo, Bruce Ovbiagele, Rufus Akinyemi, Mulugeta Gebregziabher, Reginald Obiako, Lukman Owolabi, et al. 2015. "Phenotyping Stroke in Sub-Saharan Africa: Stroke Investigative Research and Education Network (SIREN) Phenomics Protocol." *Neuroepidemiology* 45 (2): 73–82.
[2]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
[3]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
[4]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
[5]Njølstad, Inger, et al. "Body Height, Cardiovascular Risk Factors, and Risk of Stroke in Middle-Aged Men and Women." Circulation, vol. 94, no. 11, 1996, pp. 2877–2882., doi:10.1161/01.cir.94.11.2877.