Jeremy Schwartzbord
Data Science (SPOC, S22)

Project Report: Stroke Prediction Model
DSTI Python Machine Learning Labs

Objectives

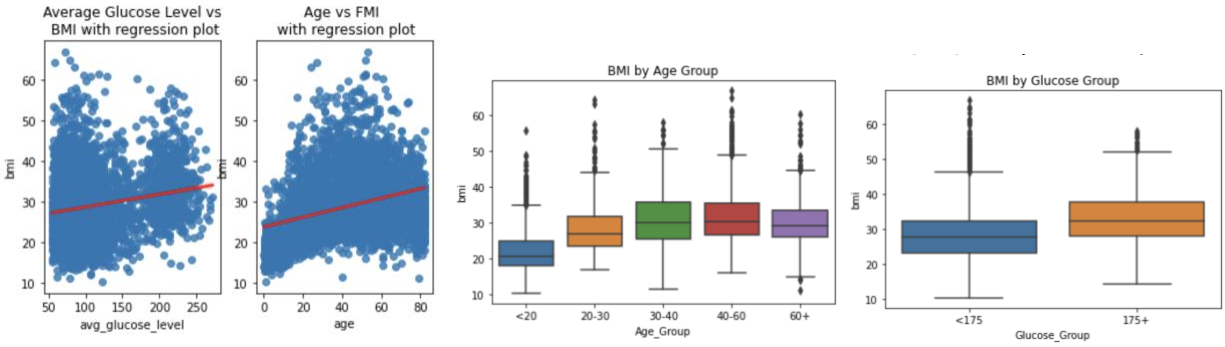
The objective of this project was to train a machine learning model to predict whether a patient had a stroke or not, using a data set of 5110 patients. Each patient represented an observation with variables such as stroke (yes/no), as well as demographic variables (i.e., gender, age), lifestyle (i.e., smoking) and health history (i.e., hypertension, BMI, glucose, etc.) that could be used to predict stroke. The complete methods of this project included an exploratory data analysis, feature engineering and selection for a model, model training, and model evaluation.

Approach

The entire project was completed using Python libraries, including numpy, pandas, matplotlib, seaborn and Scipy, and the full analysis can be viewed in the Jupyter Notebook "Stroke_Detection_JRS.ipynb." The original data set included 11 columns total: the target variable (stroke) and 10 potential explanatory variables.

Data Cleaning

Data cleaning was necessary for several reasons and involved three key steps.
First, values for the variable Body Mass Index (BMI) were missing among 201 participants of the study. There were too many participants with missing values for BMI to simply eliminate them, so expected BMI was estimated based on two associated factors: average Glucose Level (linear regression p-value <0.001) and Age (p-value <0.001). Other variables known in biomedical literature to be associated with BMI, such as gender and community, did not appear associated with BMI in this data set. The relationship between age and BMI was not clearly monotonic, because BMI increased with age up to 40 years, leveled, then decreased at 60 years and older. Therefore BMI for participants with missing data was estimated based on average BMI by sub-groups set in this analysis for age (<20, 20-30, 30-40, 40-60, 60+ years) and glucose level (<175 mg/dL, >175mg/dL). The average BMI for each cross-tabulated sub-group (for instance 20-30 years old with a glucose level <175 mg/dL) was subsequently filled into the data frame.



Second, some columns were numeric (e.g., age, hypertension, heart_disease, avg_glucose_level, bmi, and stroke), while others were not (gender, ever_married, work_type,

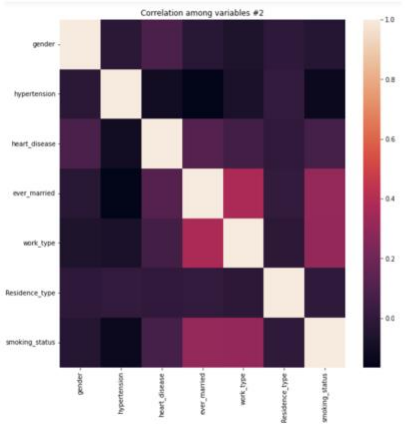
Residence_type, smoking_status). All variables needed to be numeric for the predictive model, and this was resolved with feature engineering, where categorical variables were transformed to numeric ones. Binary variables such as gender and residence type were simply encoded as Boolean variables in a single column, whereas variables with more than two categories (ever_married, work_type, smoking_status) were encoded using a one-hot encoding method, where each category was represented as a column with a boolean value, and the first category was removed to prevent linear independence among columns.

Lastly, data cleaning included removal of one rare observation where gender was "other." It was not possible to discern the reason of this category. Did the individual identify as non-binary or decline to share gender information in the study? Was the data entry an error? At any rate, it was a rare observation and therefore removed from the machine learning model.

Exploratory Data Analysis

Characteristics about the study population were considered when developing the predictive model. Most importantly, stroke occurred in 4.9% of the population and, fortunately for patients, was rare. Thus the predictive models would have much less information about individuals who had a stroke than those who did not. With respect to demographic details the average age of participants was 43 years old. Concerning health history, 9,7% had hypertension, 5,4% had heart disease, the average glucose level was 106 mg/dL, and average body mass index (BMI) was 28,9 kg/sq.m. Men (n = 2115) and women (n=2994) were similarly represented, as were urban (n = 2596) and rural (n = 2513) residents. The majority (65%) of participants were married. Smoking status, however, was not completely understood and labeled unknown among 1544 (30%) participants. Can we consider it missing data? Should our machine learning model estimate smoking history based on demographic factors of study participants?

Not all types of workers were represented in this study. Over half of participants worked in the private sector. The percent of them that were self-employed, minors, or government-employed were 16%, 13%, and 13%, respectively. Less than 1% never worked. Consequently, our machine learning model may be more accurate about participants who worked in the private sector, but less accurate for those with other work types.



work type	Govt_job	Never_worked	Private	Self-employed	children
smoking status					
Unknown	0.185693	0.363636	0.218878	0.190476	0.899563
formerly smoked	0.202435	0.000000	0.181259	0.253968	0.018923
never smoked	0.429224	0.636364	0.414159	0.404151	0.078603
smokes	0.182648	0.000000	0.185705	0.151404	0.002911

Marriage history	No	Yes
Work type		
Govt_job	0.066629	0.161050
Never_worked	0.012528	0.000000
Private	0.466970	0.627498
Self-employed	0.062642	0.211452
children	0.391230	0.000000

Marriage history	No	Yes
smoking status		
Unknown	0.513098	0.191769
formerly smoked	0.083144	0.220101
never smoked	0.301822	0.406203
smokes	0.101936	0.181927

Heat maps demonstrated high correlations, compared to other variables, between marriage history (ever_maried) and smoking status, work type and smoking status, and marriage history and work type (coefficient of determination approximately 0.4, see figure above). One possible reason could be that the study included children (13%). The tables below illustrate how patients were distributed across subcategories of work_type, smoking, and marriage history and could explain why these variables were correlated. Clearly, lifestyle factors such as marriage, smoking, and work type all would be influenced by whether a person was a child or an adult and therefore would not be completely independent. Therefore, linear regression-based predictive models should not include all three variables in order to avoid linear dependence among features of the predictive model.

Model Generation

Patients were split into a model training and model testing group by 80% and 20%, respectively. Model performance was evaluated based on the precision (tp/(tp+fp), recall (tp/(tp+fn), and f1-score (harmonic mean between precision and recall) determined in the classification report. The following hypotheses were tested in model generation:

Hypothesis objective:	**Null Hypothesis**	**Alternative Hypothesis**
Model type: • logistic regression (LR) • decision tree (DT) • random forest (RF) • neural network (NN)	Model performance is the same across all models	Model performance depends on model type
Patient sampling: • Inclusion of all observation (n=5109) • Under-sampling of “no stroke” patients (n = 250, randomly selected with all stroke patients) with all stroke patients (n = 249)	Patient sampling does not affect model performance	Patient sampling affects model performance
Variable selection : • full data frame (DF: all variables after cleaning and feature engineering) • DF without ever_married, work_type • DF without smoking and work_type • DF without ever_married and smoking	Model performance is the same regardless of variable selection	Model performance depends on variable selection

Model Validation and Prediction Accuracy

Generation of all four models using the complete data frame (DF_1, in Jupyter Notebook “Stroke_Detection_JRS.ipynb”) after cleaning revealed consistently acceptable precision of 0.95 when predicting no-stroke and consistently acceptable recall (>0.95) for predicting no-stroke in the test data set (sample size n = 1022). The precision of predicting stroke was highly variable (LR: 1, DT: 0.13, RF: 0.0, NN: 0), and the recall for predicting stroke was consistently very bad (LR: 0.02, DT: 0.11, RF: 0.0, NN: 0). Patient sampling, where the data set was reduced to 250 no-stroke patients plus the 249 stroke patients (DF_sample1 in Jupyter Notebook), showed a clear improvement in stroke prediction precision (LR: 0.74, DT: 0.68, RF: 0.73, NN: 78) and recall (LR: 0.82, DT: 0.69, RF: 0.71, NN: 0.76) across all four models. This came to a cost in average overall precision (0.69-0.77). These results confirmed the second alternative hypothesis and demonstrated that our models were not able to predict stroke effectively when it was a rare

event; when the proportion of stroke victims was relatively higher in model training, prediction accuracy improved. Indeed a stroke incidence rate of 4.9%, even in a data set of over 5000 observations was insufficient data for the predictive models chosen. Note that prediction remained poor even with variable selection when all observations (n=5109) were included. Lastly, variable selection was tested using the data set where stroke patients were under-sampled. Average F1 accuracy across all four models and variable selection groups was:

Model Type	All variables	without ever_married, work_type	without smoking and work_type	without ever_married and smoking	Min-Max
LR	0.77	0.76	0.75	0.76	0.75-0.77
DT	0.69	0.71	0.70	0.68	0.68-0.71
RF	0.75	0.78	0.77	0.72	0.72-0.78
NN	0.73	0.76	0.77	0.73	0.73-0.77

In summary, all four prediction models demonstrated an average F1 accuracy between 0.68 and 0.78; the decision tree model should not be selected because it had lower accuracy than all other model types. The best model performance (0.78) was observed with the random forest when ever_married and work_type were excluded. The most consistent, high performing model was the logistic regression with an accuracy of 0.75 to 0.77. Model performance changed little based on the variable selection group in this analysis, so we cannot reject the null hypothesis concerning variable selection.

Discussion

After data set cleaning and feature engineering, stroke prediction models were developed given a data set of over 5000 observations with comparable accuracy across distinct model types (logistic regression, decision tree, random forests, neural networks). Some variable selection produced minor performance differences, but the largest performance increase occurred when no-stroke patients were under-sampled. With average F1 accuracy between 0.68 and 0.78 across all models, improved accuracy would likely be desired in a clinical context due to the serious health consequences of stroke. The present project was not exhaustive and can foresee additional techniques worth pursuing for improved prediction model performance. First, the sampling procedure applied here could be replaced with a more robust sampling algorithm, such as Synthetic Minority Oversampling Technique (SMOTE),[1] which has been shown to be effective in pre-processing unbalanced date. Second, a more robust variable selection method could have been used as well. One interesting approach would be to select explanatory variables with the Lasso algorithm[2] and develop the final model with the classical logistic regression. Lastly, the neural network and random forest could potentially be improved with additional tuning. Future study to improve these models would indeed be well worth the effort for the advancement of patient and improved health outcomes.

[1] https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.SMOTE.html
[2] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html