**Methods and Results:**

The goal of this project is to explore the relationship between patient health variables and the incidence of stroke, and to build predicative models using various classification methods and examine the accuracy of these models. To do this, several supervised learning methods were considered, 1) Logistic Regression, 2) Decision Tree models, and 3) Linear Discriminant Analysis (LDA)/ Quadratic Discriminant Analysis (QDA).

LDA/QDA were determined to be not suitable for this dataset since the data has mostly categorical variables, and the continuous variables that the data does have are not normally distributed. See Figure 1.1-1.3. Thus only the Logistic Regression model and Decision Tree models were considered.

The data was collected by McKinsey & Company for their Electronic Health Record (EHR) and put on Kaggle.com by user Fedesoriano. The data consists of 12 variables for 5111 individuals. The variables include personal and medical information, including whether the individual experienced a stroke or not. One variable is a unique identifier number for each individual and is not included in analysis for this project. Another variable is the record of stroke with 1 if the patient had a stroke and 0 if not. This will be our response variable. The rest 10 variables will be analyzed as predictors and they are as listed: 1) gender: "Male", "Female" or "Other", 2) age: age of the patient in years, 3) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension, 4) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease, 5) ever\_married: "No" or "Yes", 6) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed", 7) Residence\_type: "Rural" or "Urban", 8) avg\_glucose\_level: average glucose level in blood in mg/dL, 9) bmi: body mass index, and 10) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown” (“Unknown” means n/a).

Logistic Regression.

The response or predictor variable is a categorization stroke / no-stroke, which is based on the predictor variables. A model based on all available original variables plus dummy variables was trained on the training data “train.stroke”. Summary statistics were then run to look at the possible important variables. See Figure 2.

Hypothesis testing with the hypothesis shown in Figure 3. The hypothesis testing shows that the intercept or base line is non-zero significant to 0.001. Also, the variables *age* significant to 0.001, the variables Hypertension and *avg\_glucose\_level* significant to 0.01, and the variable self\_employed, which is a dummy variable of employment type, significant to 0.05 as compared to the base line. The misclassification rate for this model tested on the testing data, “test.stroke”, is 4.8%, the misclassification rate for non-stroke cases is 0.27%, the misclassification rate for stroke cases is 98%. See confusion matrix in Figure 4.

Next model selection methods are used to try to truncate the model. Three methods were used, Best Subset, Forwards, and Backwards selection. See Figures 5, 6, 7, 8, 9, and 10 for the R^2 and BIC plots for the three methods. The R^2 plot suggests that using a model with 8 variables will perform similarly to a model with all the variables. Note that R^2 will always increase when more variables are used. The BIC values are also plotted for each selection method, but R^2 is used since BIC penalizes large number of variables and using a large number of variables in this dataset is manageable. The three model selection methods all selected the same model for 8 variables, and the misclassification rate of this model is similar to the full model and is also similarly poor at predicting positive stroke cases with overall misclassification rate of 5.28%, non-stroke misclassification rate of 0%, and stroke misclassification rate of 100%. See Figure 11.

Previously it was anticipated that misclassification rates would be high due to the small percentage of positive-stroke data compared to the entire dataset. Thus, sub-set data were created where all positive-stroke cases from the training data were kept, and an equal number of negative-stroke cases were randomly sampled from the entire pool of negative-stroke cases. This sub-set data is less than 10% of the entire dataset. Since half of the non-stroke data in the data sub-set is randomly sampled, three sub-sets were generated, and their confusion matrices were calculated. The misclassification rates, for the full model without truncating, for all three sub-sets are similar with around 27% overall misclassification rate, 16% stroke-misclassification rate, and 27% non-stroke misclassification rate. See Figure 12.

In summary, sub-setting data provided a more useful model with a much lower stroke misclassification rate but much more false-positives and higher overall classification rate.

Decision Tree.

The terminal nodes of the decision tree give a yes or no categorization based on the conditions specified by the internal nodes. All of the variables are used, and a decision tree has a built-in function of truncating the number of variables and streamlining of the significant variables. Note that although a decision tree is capable of analyzing the original fourteen variables without the need to create dummy variables, the decision tree is run on the dummy variables for better interpretability. For a plot of the model see Figure 13.

There is some variance in the nodes of the generated model, and this is due to the random 80/20 data split, where 80% of the data was randomly assigned to training and 20% of the data was randomly assigned to testing.

The confusion matrix of the decision tree is comparable to the logistic regression model generated with the full set of variables and on the full training data. The overall misclassification rate is 5.58%, the stroke misclassification rate is 100%, and the non-stroke misclassification rate is 0%. For the confusion matrix see Figure 14.

To find a better model bagging is used. Note that for random forest m = sqrt(p) is a good starting place. The errors of tree models from 0 – 500 trees are shown in Figure 15. Note that the Out-of-bag error decreases then levels out at about 20 trees, which is close to the number of variables. The misclassification rates for a bagging model is calculated to be 5.68% overall, stroke misclassification rate of 98%, and a non-stroke misclassification rate of 0.52%. See confusion matrix in Figure 16.

In summary, tree methods on the non-subset data set are comparable to the logistic regression model built on the non-subset data set.

A picture containing diagram

Description automatically generated

Figure 1.1: Histogram and QQ plot for the Age continuous variable in the data show that the data is not normally distributed. Box plot for the Age variable show that the data has a very wide range from infants to adults over 80 years old.

Diagram

Description automatically generated

Figure 1.2: Histogram (significantly skewed) and QQ plot for the Average Glucose Level continuous variable in the data show that the data is very far from being normally distributed. Box plot for the Average Glucose Level variable show that the data has many outliers towards the higher range.

Chart

Description automatically generated

Figure 1.3: Histogram and QQ plot for the BMI continuous variable in the data show that the data is not normally distributed. Box plot for the BMI variable show that the data has a significant number of outliers with extreme values that raises questions about the accuracy of the data.

A screenshot of a computer

Description automatically generated with medium confidence  
Figure 2: Summary statistics of a Logistic Regression model built on the training data from an 80/20 split. The intercept, age, hypertension condition, self\_employed individuals, and the average\_glucose\_level variables are significantly non-zero.

Text

Description automatically generated with medium confidence  
Figure 3: Hypothesis testing on the intercept. This can be generalized for the coefficients for all of the variables.

Table

Description automatically generated  
Figure 4: Confusion matrix for Logistic Regression model trained on training data from an 80/20 split. Very high number of false-negative predictions, but low false-positive predictions. The true Reference values are from the test data.

Chart, scatter chart

Description automatically generated  
Figure 5: R^2 values of different models from Best Subset Model selection show that the explanation of variance slows down drastically for each variable after 6. There is not much difference between a model using 14 variables and 8 variables.

Chart, line chart

Description automatically generated  
Figure 6: BIC values of different models from Best Subset Model selection selects a model with 6 variables as the best model. BIC has a penalty for using large number of variables in a model.

Chart, scatter chart

Description automatically generated

Figure 7: R^2 values of different models from Forwards Model selection show that the explanation of variance slows down drastically for each variable after 6. There is not much difference between a model using 14 variables and 8 variables.

Chart, line chart

Description automatically generated

Figure 8: BIC values of different models from Forwards Model selection selects a model with 6 variables as the best model. BIC has a penalty for using large number of variables in a model.

Chart, scatter chart

Description automatically generated

Figure 9: R^2 values of different models from Forwards Model selection show that the explanation of variance slows down drastically for each variable after 6. There is not much difference between a model using 14 variables and 8 variables.

Chart, line chart

Description automatically generated

Figure 10: BIC values of different models from Backwards Model selection selects a model with 6 variables as the best model. BIC has a penalty for using large number of variables in a model.

Table

Description automatically generated with low confidence

Figure 11: Confusion matrix from 8-variable model from model selection. No positive predictions were made as all are false-negative predictions, but no false-positive predictions. The true Reference values are from the test data.

Table

Description automatically generated with medium confidence

Figure 12: Confusion matrix from model built on sub-set data 1. Other confusion matrices built on sub-set data are similar. Sub-set data with an equal proportion of positive-stroke to negative-stroke cases used to train the model produced much higher accuracy as compared to 80/20 random data split.

Diagram

Description automatically generated

Figure 13: Decision tree built on all available data and dummy variables, and on non-subset data. Age and Hypertension variables were determined to be important and other variables were dropped by the model.

Calendar

Description automatically generated

Figure 14: Confusion matrix from decision tree model shows that a basic Decision Tree model is poor at predicting positive stroke cases with all true-positive cases classified as false-negative cases. The Decision Tree misclassification rate is comparable to the logistic regression model before subsetting the data.

Graphical user interface

Description automatically generated

Figure 15: Error from bagging/randomForest. The Out-of-the-bag (OOB) error decreased as the number of trees generated increased, as expected. Error also increased as the number of trees increased. Most likely due to using a non-subset data for training.

A picture containing chart

Description automatically generated

Figure 16: Confusion matrix from the bagging model where the number of trees = number of parameters show that the bagging model is poor at predicting positive stroke cases, and is worse at predicting negative stroke cases as compared to even the logistic regression model trained on the non-subset data.