Literature Review #1: “Predicting Stroke from Electronic Health Records"

The central piece of literature for our project is “Predicting Stroke from Electronic Health Records" written by Nwosu et al.. It was first published in 2019 and presented at the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Science. When creating this article the researchers wanted to go beyond the standard analysis of predicting stroke occurrence from patient medical records and analyze the interdependence of commonly collected risk factor data in order to quantify the impact of these risk factors. To achieve this goal they not only supplied visualization of the risk factors but employed the use of three machine learning algorithms (Decision Tree, Random Forest, and Neural Network) in order to create useful benchmarks for the average accuracy of predictors.

Before the analysis could begin the researchers needed to choose an appropriate data set that allowed them to perform the various analyses they hoped to perform. They decided on a set of electronic health records distributed by the reputable electronic record publishing group McKinsey & Company. The record consisted of 12 attributes of pertinent medical and demographic information for 29,072 patients. Since it is so necessary to keep medical data private to ensure the anonymity of the patients, not much background information is given on the population the data is sourced from. The most that is known is that data collected was recent to the time of its release so the results drawn from its analysis are not entirely outdated. The 12 attributes provided in the data set consist of 11 input variables and the single response or output variable which is whether or not the patient had a stroke. The 11 input variables include: patient identifier, gender, age, hypertension (binary: suffering from or not suffering from), heart disease (binary: suffering from or not suffering from), marital status, type of occupation, type of residence (urban/ rural), average glucose level, body mass index, and the patient’s smoking status (Nwosu). Our group is very familiar with this dataset because it is the same one that we decided to use to build our predictor. In accordance with the aforementioned need for anonymity the patient identifier attribute was left out of analysis, so the researchers decided to move forward with 10 input attributes at this stage. We found this article when trying to learn more about the origins of the dataset and when we saw how they used predictors in a similar fashion it seemed like a good idea to modify our project a bit in order to make comparisons between our predictors and the ones created by the researchers.

The first step for analyzing the risk factors present in the dataset was to perform a Principal Component Analysis. This would provide information for which attributes were responsible for the majority of the variance in the response variable by using three plots (scree plot, biplot, and subspace representation). With this information the researchers can decide which attributes are the most crucial to building their model in order to optimize its performance. To showcase and explain the variance a scree plot was used. The results of the variance analysis were that the 9 of the 10 attributes or principal components explained 96.06% of the output’s variance (Nwosu). This means that accuracy would decrease by 4% if one of the components was excluded. What was more interesting was that the first two components explained 31.4% of the variance. Due to this large difference, the researchers were leaning towards the idea that all attributes should be used when building the models because there was only an increase in optimization if at least two components were removed which would vastly lessen the integrity of their model. To truly ensure that all 10 attributes were necessary to uphold the integrity of their model’s results the researchers decided to determine the specific contributions of each variable through a biplot. The results of the biplot outlined that both the age and the marital status of the patient contributed the most to the two principal components while the residence type had no contribution. Another figure was also created which shows a projection of the subspace representation in respect to the two primary components for qualitative and binary attributes. Essentially all of the data for the binary attributes were adequately explained by the two primary components. The same was not true for the qualitative attributes which had a significant number of observations positioned towards the center of the plot which represents a lack of representation in the two primary contributions. These results were in line with the scree plot so the researchers came to the consensus that all ten attributes were to be included while building their algorithms.

One more major point needed to be taken into consideration before the models could finally be built and the relative results could be compared. One downside to the dataset chosen is that of the 29,072 patients only 548 of them had a stroke. This means that only 1.9% of the data present was indicative of risk factors of stroke and the other 98.1% of the data was for healthy individuals (in respect to having a stroke). If a predictor was built using the entirety of the dataset it would not truly be accurate in directly identifying risk of stroke. What the researchers decided to do in order to combat the inherent imbalance in the dataset was to perform random downsampling of the majority class (in this case the patients who did not have a stroke). This would achieve far more balance with the minority class (in this case the patients who did have a stroke). The balanced data that would be fed into the machine learning algorithms would then only include 1,096 patients which would be the entire 548 patients of the minority class and a random sample of size 548 patients from the majority class. From the new total data size, 70% was used to train the model and the other 30% was used to assess the prediction accuracy of each respective model (Nwosu). Since we are using this same dataset the issues identified by the researchers for the majority and minority class are of great importance to us. For the model that we will produce we will need to perform some similar action to balance the two classes and because random downsampling is explained in the code they provided it makes the most sense for us to also take this approach.

To assist in finding less biased average values for benchmark performance the researcher's performed 1000 accuracy tests for each machine learning algorithm and did a different random downsamplings for each repetition. The results of these trials showed that the neural network model had the highest accuracy with 75.02%, next was the random forest model with 74.53% accuracy, and the least accurate was the decision tree model with 74.31% accuracy. The researchers reasoned that the neural network was the most accurate because of the three models it works the best with multivariable input variables (Nwosu). It is important to underscore that the difference between the three was quite small (not even a full percent difference) and when the researchers plotted the distribution of each model's accuracy across the 1000 trials the results for all three appeared even more similar. As previously mentioned the researcher's goal was to just provide a benchmark for future work with new models and were not attempting to identify which of the three models should be used preferentially based on higher predictor accuracy.

Overall, the paper provided key information for the interdependence of the risk factor attributes provided in the data set when they performed a PCA analysis. The results showed that the attributes were not highly correlated so the number of features for building the model cannot be simplified without losing a significant portion of the valuable explanatory information. With the total number of attributes ironed out, the researchers could build their decision tree, random forest, and neural network models. The most accurate was the neural network due in large part to the fact that it is a multi-layered perception model. The three did have comparable accuracies so the researcher’s showed that all three could be viable for future work in the field while using their results as a benchmark. The researchers also discussed their plans for future work in the space but their ideas were out of the scope of our project and our class as a whole. Our group decided to use the models created in this research study for a variety of reasons. The most obvious reason being that the article uses the same dataset that we planned to use for our own model, which would make comparison more standardized. Another reason was that their code was readily available (linked on the article itself) so we could easily reproduce it and feed it into a bias analyzing tool such as Aequitas in order to assess the fairness for the protected attributes present in the dataset and compare it to the Aequitas assessment of our own predictor.