Classification and Prediction of 6-Month Mortality in Stroke Patients

Aidan Horvath & Henry Luong

Introduction:

Stroke is a common, serious medical condition with various causes, treatments, and overlap in symptoms with other medical conditions. As such, a tool for rapid assessment of patients presenting with stroke symptoms could help prioritize specific diagnostic paths or early treatments. Additionally, stroke occurs with a wide degree of severity, both in the initial event and with the possibility of recurrent episodes. Machine learning could also help predict the Bamford classification of a patient's stroke and if they are likely to experience another in the days following their first occurrence.

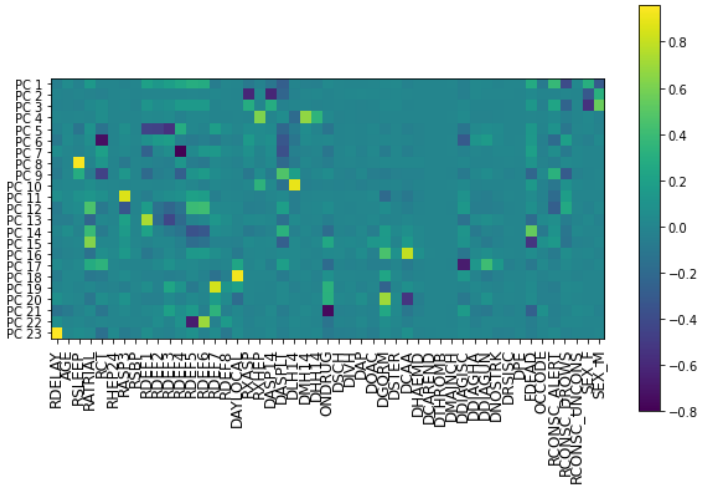
Our aim for the Milestone II project is to accurately label the different types of strokes while predicting the 6-month mortality rate in each patient. We will be performing a semi-supervised learning approach on individual patient data from the International Stroke Trial (IST) database. The international stroke trial was conducted back in 1991-1996 with 19,435 de-identified patient records. The purpose of the clinical trial was to facilitate the planning of future trials and allow for additional secondary analyses. When we initially reviewed the database, which happened to be a comma-separated values file, it included over 100 features that collected data such as demographics, vital signs, and medical history from admission to follow-up after six months.

After proceeding with the initial data exploration, the randomized controlled clinical trial results produced over 99%+ baseline and follow-up data for us to review. The goal was to one hot encode all categorical values as a numerical value. One hot encoding converts each categorical, or non-numerical, value as a new column with a 1 or 0. Although this is more advantageous than label encoding, where it has a hierarchy/rank to each numerical value, it makes the dataset more sparse. Additionally, it will replace missing data with a proxy value by performing data imputation due to rows within the features or attributes that are not filled out or are missing at random. However, most of the current methods rely on the assumption that the data are missing at random.

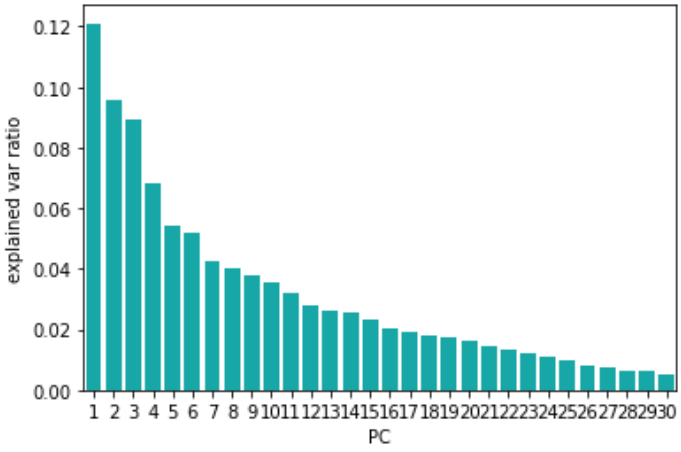
Adhering to either a mean or K-NN imputation, our evaluation metric bases on which method yielded the most accurate results from the models. Along with providing the best products, K-NN imputation uses the k-nearest neighbors to predict values of any new data points based on the proximity of every moment belonging to the training set. It is also worth considering whether or not the data should be standardized or normalized. Proceeding with K-NN imputation, the dataset will require feature scaling that normalizes and rescales the data such that all values will be in the range of 0-1.

Unsupervised Learning:

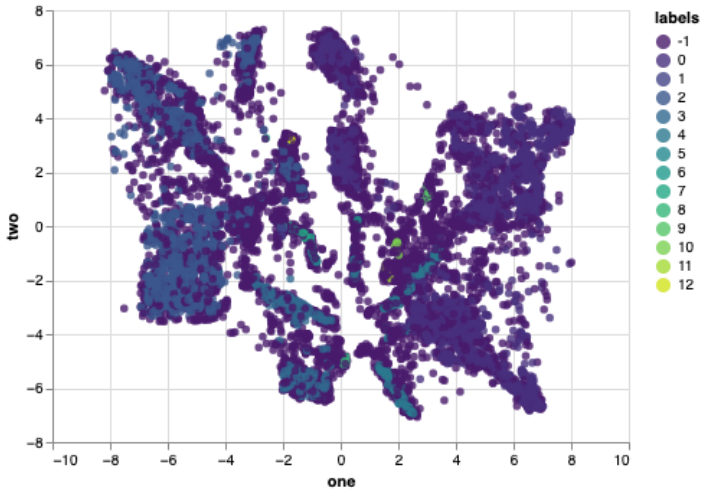
Since we are working in a high-dimensional space, we aim to identify the underlying structure and distribution of the data while reducing and extracting the features that lead us into a supervised framework for model prediction and multi-classification. For our unsupervised approach, we wanted to see if the data could group into clusters to extract the essential features, therefore reducing the feature space. Will these groups match the dataset's features, and can they be used as labels for our supervised learning classifiers? Once we had our cleaned data, we reduced the number of feature columns by about fifty components. When determining which dimensionality reduction techniques to perform, it is crucial to understand whether or not the underlying distribution is linear or non-linear.



First, we will conduct Principal Component Analysis (PCA), a linear dimension technique that compresses the feature space and describes the most variance within each component. When we ran PCA with two components, our data showed us that the total variance within them was only 20%, which is considerably lackluster. The data was then iterated through PCA to find the correct number of components that gave us at least 90% of the variance, 23. We noted that within the heatmap that certain features such as neurological deficit indicators were highly correlated. In contrast, mental state (RCONSC) and sex were inversely correlated with each other. Because the number of components was so large and the features loaded onto the components with no discernible pattern, aside from the ones stated, we concluded that dimensionality reduction might not be appropriate for this data set. Regardless, we attempted additional clustering techniques to see if they performed better.



Next, we can perform a clustering approach known as t-distributed Stochastic Neighbor Embedding (t-SNE) while overlaying the labels produced from a Density-Based Spatial Cluster of Applications with Noise (DBSCAN) model. t-SNE allows for the exploration and visualization of high-dimensional data through a non-linearity technique that retains small pairwise distances versus large pairwise distances of PCA. By reducing the features down to two components and tuning the perplexity to 50 to allow for the most significant number of nearest neighbors, we can then use DBSCAN to find core samples of high density and enlarge the clusters from them. DBSCAN's main hyperparameters consist of eps, which is the maximum distance between two points to be considered belonging to a neighborhood of the other, and the minimum number of samples in a neighbor for that to be considered a core point. Adjusting the minimum samples does not affect the algorithm as much as EPS. To find the optimal epsilon value, we calculate the distance to the nearest n, number of points. By applying the nearest neighbor algorithm, the optimal value lies at the point of maximum curvature, which is 1.4.



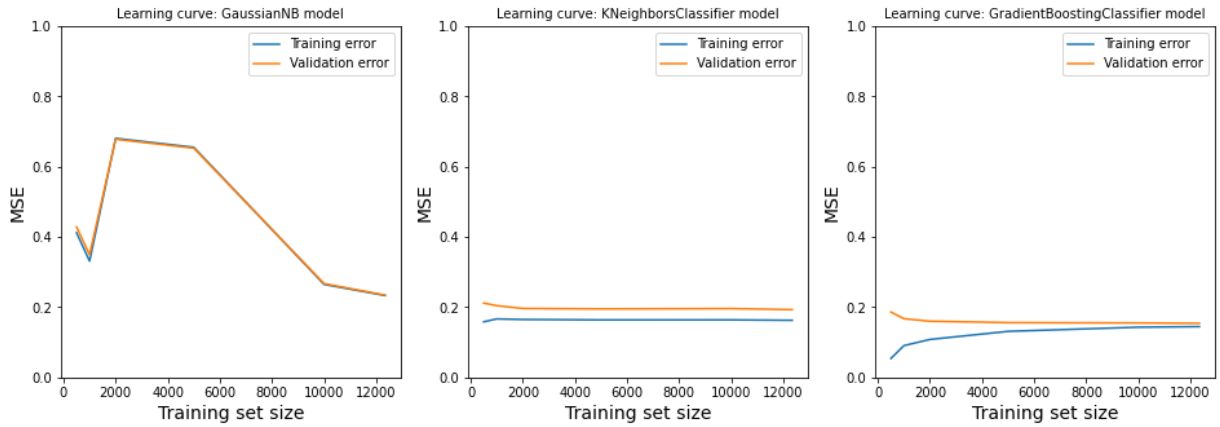
Training the unlabeled data with DBSCAN and the hyperparameters, we produced a label for these points and then projected them onto the t-SNE graph. Although the results had impressive labels, we also saw that it identified 50% of the data as noise points. This large percentage indicates the algorithm's inability to cluster the outliers into a border or core point. With the curse of dimensionality, the significant amount of columns explains why DBSCAN yielded such results. It is a bit difficult to try and interpret such high dimensional data on a 2D graph but what we see as a result is that specific labels belong to one side of the t-SNE chart compared to the other. Although the data looks to fit specific criteria on the t-SNE, we must remember not to perform DBSCAN since it does not preserve distances nor density. In doing so, we could run the risk of producing false patterns by breaking up the initially connected t-SNE. We must therefore apply the DBSCAN only to the original data and project it onto the graph.

Our results show a reduction from about 50 components down to 2 for t-SNE and 13 clusters for DBSCAN. We can see how non-linear dimensional reduction scaling techniques are a lot more effective than linear ones. Due to the reduced dimensions, we cannot identify the critical features within the approaches. To test

whether or not it will yield effective results, however, is another thing. Running the transformed results through our supervised learning approaches resulted in little change when using our evaluation metrics. The classifier allowed us to see that the features were highly correlated by looking at the VIF or variable inflation factors. VIF determined the magnitude of the correlation between the independent variables by regressing one variable against all the others. The results showed that the initial diagnosis of the event was among the highly correlated features within the dataset. Therefore, it is better to apply our full dataset into the algorithms since each attribute correlates with one another rather than keeping the transformed data through Principal Component Analysis since it did not yield improvements.

Supervised Learning:

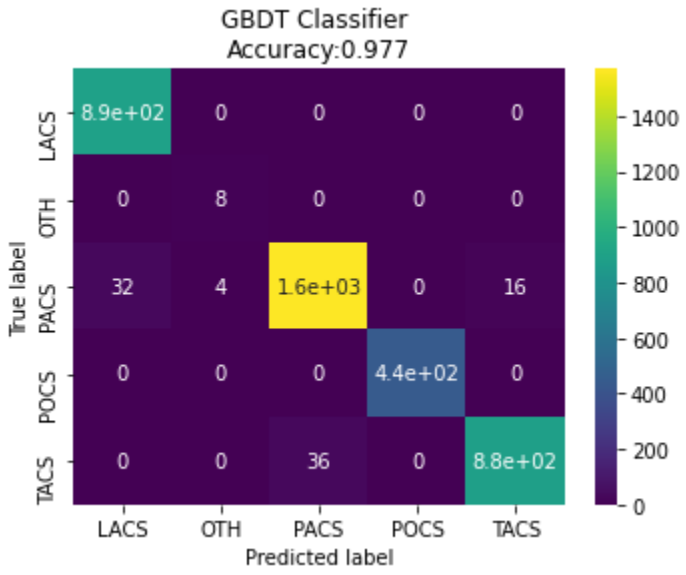
We used two supervised learning techniques, K-nearest neighbors and Gradient Boosted Decision Trees, to make predictions on our two outcomes, at six months and stroke type. Our source code began with importing our data and defining functions to split, clean, and scale the data. We then gathered information about the data, such as the number of missing data points and the degree of imbalance in some variables. Once we had code to prepare the data, we attempted to reduce its dimensions with PCA and tSNE and investigated the degree of clustering using DBSCAN. As noted in the Unsupervised Learning section, these procedures did not reveal much information about the data. Therefore, we used the scaled features in their original representations for supervised learning. The remaining source code includes the classification training and evaluation, which consist of a Naive Bayes classifier (for baseline performance), a KNN classifier, and a Gradient Boosted Decision Tree, as well as learning curves, ablation tests, and visualizations (ROC-AUC curve and confusion matrix) to analyze the model performance. We tuned our hyperparameters using a grid search for the Gradient Boosted Tree and a for loop returning the Calinski Harabasz scores for different values of K for the KNN.



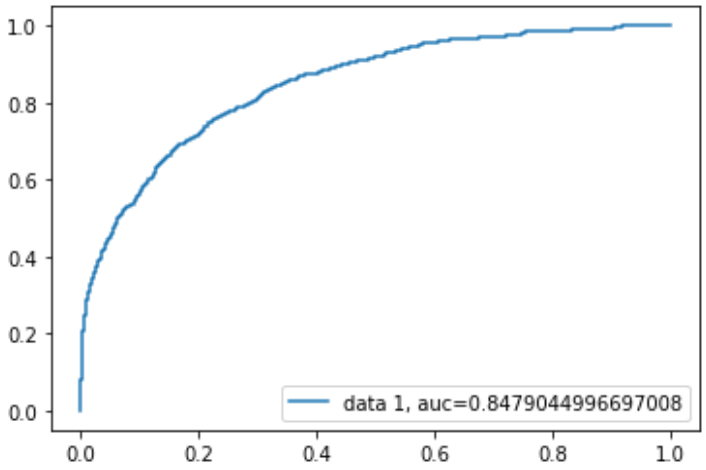
We decided to use a KNN and a Gradient Boosted Decision Tree based on the differences in our prediction cases. Death at six months is more of a similarity problem wherein a new prediction case, this is close in feature space to a neighborhood and could be said to be more likely to have the same outcome. In contrast, stroke type is more of a classification problem, and therefore, new predictions will split into different groups through successive steps. We believed they would be appropriate for these individual cases because of KNNs' distance metrics and Gradient Boosted Trees' top-down sectioning approach. Both the KNN and Gradient Boosted Tree performed well in both prediction cases. The KNN achieved a mean accuracy of 0.79 for stroke type and 0.81 for death at six months. The Gradient Boosted tree reached a mean accuracy of 0.98 for stroke type and 0.84 for death at six months. These techniques outperformed a Naive Bayes classifier, which achieved a mean accuracy of 0.72 for stroke type and 0.76 for death at six months.

Additionally, the learning curves of both models are satisfactory, with the convergence of training and validation error occurring at a mean squared error value of approximately 0.2. The difference in performance for stroke type and death between these two approaches is consistent with our justification for using KNN and Gradient Boosted Trees. The Gradient Boosted Tree performs only slightly better at predicting death at six months KNN and performs better at predicting stroke type.

Ablation studies of the models indicate that different features were essential for making predictions of either outcome. For determining whether a patient would die within six months of a stroke, two indicators stood out. These were whether the patient was discharged on long-term aspirin and the number of days they spent on trial treatment. For the prediction of stroke type, the most critical features consisted of the various deficits a patient presented, particularly dysphasia, hemianopia, and brainstem or cerebellar dysfunction signs. These results make sense. In the case of death at six months, treatments could have a causal impact on a patient's chances of dying and correlate with the severity of the stroke. For stroke type, it makes sense that different kinds of strokes would present different patterns of deficits in the patient.



We used two different visualizations to evaluate the performance of our models further. For the binary case of death at six months, we used a ROC-AUC curve, which illustrates a low level of overlap between true positives and true negatives with an AUC value of 0.85. We used a confusion matrix for the multiclass case of stroke type to show that the classifier predicted most patients correctly.



During these analyses, several prediction failures occurred. The quest for dimensionality reduction in our unsupervised approach did not provide a sufficient enough accuracy and recall score than if the original data had been cleaned and scaled. Although the unsupervised learning results could reduce the features, the amount of slight variance explained by each feature in the components was insufficient to yield improvements from the standard. Because the data had a high level of integrity and readability, the vast majority of errors were due to problems with the code. Most of the mistakes involved using the wrong X data, i.e., using a version of X_train meant for multiclass analysis in code for binary classification.

Evidently, during the ablation tests, we initially used the previously constructed model instead of a newly trained model for scoring each iteration of feature dropping, which resulted in the training data having one fewer feature than the testing data. We also wanted to note that since performing a grid search for a gradient boosted decision tree takes significant computational complexity, our hyperparameters were limited to a couple of values. In contrast, the rest of the hyperparameters were conducted via trial and error.

Discussion:

Among the patients with stroke, the classifier was able to predict with 97.7% accuracy in categorizing the main subtypes using the Bamford Stroke Classification guidelines: Total anterior circulation stroke (TACS), Partial anterior circulation stroke (PACS), Posterior circulation stroke (POCS), Lacunar stroke (LACS), and Other (OTH). In patients with an acute stroke, the Gradient Boosted Decision Tree model distinguished between life and death with 84.8% accuracy. This research may provide additional benefits to the physicians during the initial patient assessment and post-follow-up (6 months) for those who suffer from acute stroke.

As with any approach, whether unsupervised or supervised, the cause for concern arises when trying to predict stroke subtypes and death at six months simply due to racial disparities within the healthcare industry. If the underlying distribution of data is unequally represented, significant bias could ensue to the minority and may result in inaccurate predictions regarding the outcome measures. By including demographics and patient level data that equally represents the entire population, the predictive performance for each classifier should improve and allow for a generalizable model rather than exacerbating health disparities.

When extracting the most optimal features during our unsupervised investigation, our efforts resulted in a lower accuracy score during our training step in the model. This lower accuracy score is attributed to the multicollinearity in our dataset that is typically not a problem for Gradient Boosted Decision Tree models since the goal is to predict the counterfactual outcomes. In terms of our unsupervised learning approach, reducing the components generated a low proportion of variance in each one to indicate that most of the features highly correlate with each other. The t-SNE/DBSCAN overlay perfectly represents how difficult it is to reduce the components since most labels scatter in different areas, with the majority interpreted as noise points. Each feature offers valuable insight into classifying the stroke subtypes and prediction of death at six months.

Since t-SNE reduces the dimensions down to 2 components, we cannot provide additional insights into certain features and their variance within each component. However, one of the advantages of Principal Component Analysis is that it can explain the variance of each feature in the selected components. As shown in the PCA heatmap above, specific attributes are better associated and higher on the color scale than others. These features are neurological assessments that determine whether or not a particular individual is showing signs of an acute stroke (RDEF1-RDEF7). By isolating these unique features and training the classifier using just these columns, the results showed decreased accuracy compared to the entire column extracted. Future research could be conducted using different unsupervised learning approaches such as density estimation and UMAP to determine whether or not the transformed data can provide better results during the training phase of the supervised models.

During the supervised learning methods, the Gradient Boosted Decision Tree classifier provided the best results with K-Nearest Neighbors coming in second and cementing the benchmark classifier, Naive-Bayes, as last. The decision to establish a parametric machine learning algorithm as the baseline and compare it to non-parametric algorithms demonstrates the complexity and distribution of the international stroke trial dataset. Instead of assuming the functional form of the training data as linear, the models can construct a customized function and generalize the unseen data to produce the best fit.

As explained in the unsupervised approach, what made GBDT stand out among the models was its robustness to multicollinearity. It is more suitable for predictions than inferencing and analysis. As the model can construct more trees in a stepwise, albeit greedy, manner, it can ignore the redundant features while selecting the feature as a leaf for the tree at every level. Although the results produced improved results for the multi-classification, it is striking that the binary resulted at 84.8%. The product of this lower accuracy score might be due to all of the features within the dataset not having a normal distribution. This imbalance can otherwise skew the prediction, thus allowing for higher False positives and False negatives. Different modeling techniques that are robust or can control for the skewed features for binary classification of death at 6-months should be considered for future investigational purposes.

Statement of Work:

Due to medical challenges, Aidan Horvath contributed less to the source code and report compared to Henry Luong. Aidan contributed the ablation test code and the code under the heading “Aidan’s work,” which consists of exploratory analyses to evaluate model performance and optimization of KMeans clustering. Aidan also wrote the Supervised learning section of the report and contributed minor additions to the unsupervised learning section.

References

Sandercock, P.A., Niewada, M., Członkowska, A. *et al.* The International Stroke Trial database. *Trials* 12, 101 (2011). https://doi.org/10.1186/1745-6215-12-101