Stroke Prediction Using Machine Learning

Alexander Spivey#1

CSE 398 - Data Mining,

Lehigh University, Pennsylvania, United States

1alexander.a.spivey@gmail.com

1,2aas423@lehigh.edu

***Abstract* — The following research conducted below tested a multitude of 11 models within Python and a multitude of models in Weka for the use of Stroke prediction. The provided dataset only has 11 attributes with 5,112 instances. The overall class distribution is heavily skewed, with only 5% being stroke cases. As such, SMOTE was implemented for both the training and testing set. The paper concludes that due to the dataset's unrealistic data distribution and lack of instances, the model to use should be based on the actual testing data distribution present.**

***Keywords* — Stroke, SMOTE, Machine Learning, Weka, Python, Gradient Boosting, SVM, Decision Trees, Catboost, XGB**

I. Introduction

The Stroke Prediction Dataset provided by fedesoriano on Kaggle contains 12 columns: id, gender, age, hypertension, heart disease, ever married, work_type, residence type, avg glucose level, BMI, smoking_status, and stroke [1]. This dataset was given a 10.0 score in terms of usability and contained a total of 5110 different instances. While this dataset may be considered smaller due to its limited number of entries, due to this dataset's relationship to users private medical data, if not handled right, could be considered a violation in regards to HIPAA and could lead to other ethical violations. Because of this and the confidentiality behind the dataset's source, having 5,000 instances is sufficient. Another reason for using this dataset would be to conduct exploratory data analysis on the unique trends between attribute pairs. Due to the mathematical and graphical properties needed to conduct this type of analysis, Weka, Python, Jupyter Notebook will be the primary coding platforms, with Weka being used for baseline testing scores and Python/Jupyter for the implementation of advanced models and plotting graphical relationships within the dataset. Using the 11 attributes provided, with the exclusion of stroke, multiple machine learning models will be implemented and hyper-tuned to solve the binary classification problem at hand. The end of the research aims to provide the "best" model and classification significantly higher than the baseline, 95% for the dataset being used.

II. Survey Of Current Field's Standings

As of the statistics accumulated during 2019, stroke is the five leading cause of death for citizens within the United States of America; someone has a stroke every 40 seconds and every 4 minutes and dies from it [2, 3]. With stroke being such a heavy toll on America's populus, countless numbers of research and medical studies have been conducted in the past decade in hopes of identifying the best predictors along with their corresponding machine learning model. However, as every dataset differs, the best model also differs. By first analyzing the previous works within the field, this project aims to converge the best ideas while avoiding mistakes others have made.

One of the original models for analyzing the survival time of patients on one or model attributes was the Cox Proportional-Hazards Model introduced during the 1970s, which acts similar to a Logistic Regression model that is modified to work with respect to time [4]. One prominent issue with this model is its extreme reliance on high-quality attributes [5]. Study [6] from Stanford University, written in 2010, points out the aforementioned issue and mentions an overall lack of machine learning usage within the field, with the only study at the time being misguided due to a lack of pre-selected attributes. The team believed that it was possible to "present(s) an integrated machine learning approach for stroke risk prediction" using SVM on the selected feature vectors [6]. This original founding paper within the field goes on to present that their model outperformed that of the regularized Cox model and the regularized Logistic Regression model on the CHS dataset. They go on to mention that their "method can be used for identifying potential risk factors for diseases without performing clinical trials" [6].

While the previous paper's methodology outperformed that of logistic regression, other studies have shown that logistic regression and other variations such as gradient boosting can significantly improve the classification. For example, study [7] aims to identify thousands at high risk for stroke within China. The team used the stroke screening and intervention program declared in 2009 by the Ministry of Health of China to cause interventions with the high-risk populus, specifically above the age of 40 [8]. Stroke risk factors used within the program share some common grounds with the dataset being used for this project; risk factors include "hypertension, diabetes, dyslipidemia, smoking, lack of exercise, overweight and family history of stroke" [7, 8]. Their overall methodology incorporated SMOTE on the imbalanced dataset, increasing the minority class and the use of undersampling to level out the class distribution. Afterwhich, the Weka package was used to implement models such as "logistic regression, Naïve Bayesian, Bayesian network, decision tree, neural network, random forest, bagged decision tree, voting and boosting with decision trees to classify stroke risk levels" [7]. The resulting best recall model was a boosting model in combination with a decision tree (99.94%), with the best precision model being a random forest (97.33%).

In the introduction of this paper, it was briefly mentioned that the dataset of choosing only contained 5,000 instances, while considered small in size compared to a typical dataset, is actually on par and sometimes larger than other medical datasets. In study [9], the research team chose to work with 421 males and 382 females, with only a total of 73 stroke cases. However, what it did lack in instances, the dataset made up for in terms of attributes, totalling up to 229 characteristics for each of the participants with a heavy focus on hypertension, smoking, diabetes, and hyperlipidemia. While the previous studies focused more primarily on the fundamental and historically tested methods for classification, this study chose to branch out to more modern solutions, such as the gradient-boosting tree algorithm, CatBoost. This model was implemented with the assistance of Shapley to calculate each attribute's overall contribution to labeling. While the research results are a bit underwhelming due to the pre-existing knowledge of strokes, such as the attributes of age, pre-existing symptoms, and current NIHSS score are the most relevant factors. However, the model did score a 0.89 area under ROC, highlighting the prominence of using different methodologies than typical.

Other such methods could include the implementation of a Deep Neural Network. While there are multiple studies to support the use of neural networks [10, 11], each consensus that while they may outperform other methods, they tend to be a black box of sorts and may be difficult to diagnose or understand the resulting products. Both studies [10, 11] suggest using them, but not entirely on their own. Instead, use them in relation to other methods, such as logistic regression and Catboost, to compare results and build the understanding of neural networks of the success of the older and simpler methods. While both studies do go on to show their exemplary results of using neural networks, they tend to require tenfold data per degree of freedom. However, study [10] mentions that DNN scored on par with GDBT with fewer amounts of patient data, and study [11] demonstrated once again that previous stroke severity and age are the fundamental attributes behind the binary classification hand.

III. Medical Attribute Analysis

As mentioned previously, the dataset on hand consists of 11 attributes, with six being medical more inclined: hypertension, heart disease, average glucose level, BMI, and smoking status. Due to the confidentiality behind the dataset and the general lack of time frame, it is hard to determine if any of these instances within the dataset are all first-time strokes or not, and other such time-sensitive based analysis and questions. However, general research into each of these attributes will give a better underlying understanding of the problem at hand, which can assist with the diagnosis of the 'black box dilemma.'

Hypertension, also known as high blood pressure, may result in health problems due to the long-term damage from the force of the blood pumping against artery walls. This condition relies on two primary conditions, how much blood the subject's heart pumps and how narrow the arteries are. The two numbers given at the end of each blood pressure test are systolic and diastolic. The first number measures the pressure within the arteries while the heart is beating, whereas the second measures in between each pump. Hypertension is mentioned so frequently as one of the major classifying factors due to the possible blood vessel damage within the brain. If hypertension ever gets too high, it may cause a narrow vessel to rupture or break. However, it may also cause a clot to form due to the overabundance of blood but limited passage, in turn, a stroke to occur. Hypertension can also arise from a few of the earlier mentioned attributes, such as smoking and Body Mass Index. When using tobacco, it is chemicals that can drastically increase blood pressure for hours and damage the inner linings of the arteries, weakening them over time [12].

As for BMI's effect on hypertension, the more a person weighs, the more oxygen and blood are needed to flow throughout the system. As the number of blood cells increases, so does the overall pressure. As higher BMI is associated with those who are less active, those who exercise less tend to have higher heart rates. The higher the rate, the harder the heart is pumping. Another prominent factor associated with BMI is the overindulgence of food and/or drinks [12]. Foods highly concentrated with sodium can cause cells to retain water for more extended periods, increasing overall pressure. As for drinking, alcohol can damage the heart and liver. Having just one drink can affect the subject's hypertension for hours [12].

The last medical attribute of focus is average glucose level, which is heavily associated with BMI as higher levels lead to diabetes. Foods highly concentrated in sugar can make blood vessels harden, causing fatty deposits, which in turn can lead to strokes [13]. After a stroke, the blood glucose level tends to be higher than average, with extremely high levels that can be a determining factor of lesions, leading to higher risks of mortality and less likelihood of complete recovery [14]. While this paper does not aim to cover all medical symptoms and treatments, it does aim to inform readers of the risks of strokes and underlying factors that cause them. Another would be to inform the reader of ways to change their own lives if they see possible risks within their own.

As always, regardless of the context given, drinking, smoking, and overindulgence of anything good can also lead to health complications. Be responsible and stay safe. While there is a lot a person can do about not drinking alcohol and smoking tobacco, sometimes the options for nutrition come more limited. However, nutrition is vital for preventative purposes, with a Cretan Mediterranean diet being the best overall. Such a diet consists of "beneficial oils, whole grains, fruits, and vegetables and low in cholesterol and animal fat, (which) has been shown to reduce stroke and myocardial infarction by 60% in 4 years compared with the American Heart Association diet" [15]. Other beneficial activities include mediating, decreasing stress, exercising, strengthening the heart and arteries, and adequate care and rest for the body. If possible, constant communication with physicians will entail the best possible outcome.

IV. Weka Implementation & Results

As mentioned in both the introduction and one of the previous works, Weka will be used for baseline testing and overall comparison against Python's machine learning capabilities. The classifiers tested in Weka included: OneRule, RandomForest, J48, RandomTree, Multilayer Perceptron, and Naive Bayes. To load the dataset into Weka, the null values have to be either replaced or modified such that all null values are represented as '?'. Conducting exploratory data analysis found 201 missing values within the BMI attribute. To fix this issue, values were predicted using a Random Forest Regressor and filled in (this methodology will be explained later). In doing so, all null values were replaced. The only remaining step before training required the conversion of the class label column from 'numerical' to 'nominal.'

Before looking at each model's accuracy and results, it is crucial to calculate the dataset's null accuracy score, better known as the ZeroRule classifier in Weka. The model's premise is to predict the most significant class within the dataset consistently. As such, for the current stroke dataset, the null accuracy is 95.1272%, with it accurately predicting ‘0’ or no stroke for 4861 instances (or the number of no stroke instances within the dataset). This score will be used as the baseline for future model testing.

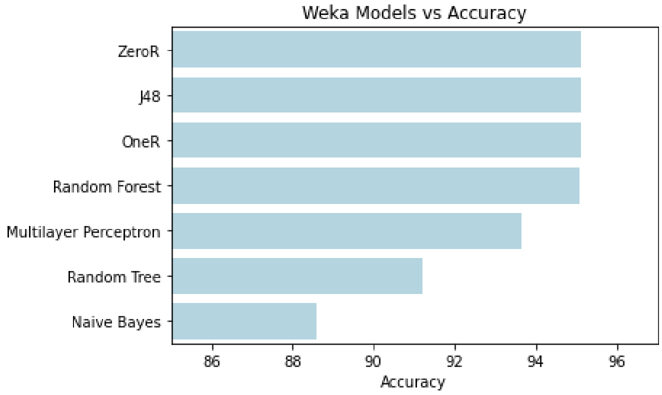

Fig. 1: Weka Models implemented vs resulting accuracy

The above chart displays the results of each model, along with their accuracy. In the end, no model scored above the baseline, with J48, Random Tree, and OneRule duplicating the results. However, taking a deeper look at the confusion matrix of those who did not perform on par with the baseline shows some redeemable marks. For those performed on par, the models built themselves to essentially only predict no, which would always result in the baseline accuracy. As for those who did not, they could predict strokes but often had cases of false positives and negatives. With false positives being ‘harmless’ compared to false negatives, the best model was Naive Bayes, predicting 80 stroke cases, 414 false positives, and 169 false negatives. The only comparable model was MLP, which accurately classified 31 stroke cases, 107 false positives, and 218 false negatives.

V. Python/Jupyter Exploratory Data Analysis

For the dataset to be initially loaded, Panda is used to read the .csv file in as a DataFrame. Conducting initial data checking for null values, the aforementioned 201 missing values within BMI re-appears. A further analysis into the class division shows again that the data is heavily imbalanced.

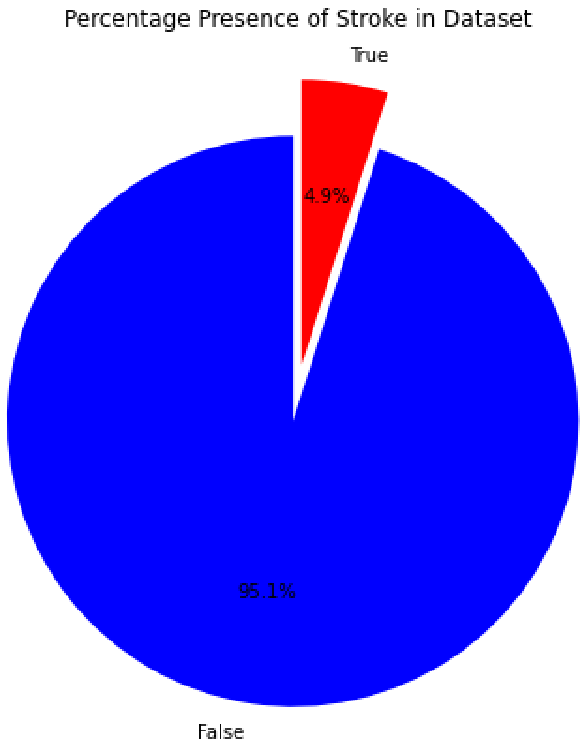

Fig. 2: Percentage of stroke cases present within the dataset

Furthermore, there is an unequal distribution of males (41%) to females (59%). Due to this, 56.6% of all stroke cases are female. When analyzing the entries based on gender, age, and stroke, the jump in stroke risk for females typically occurs when they turn 75-79, whereas, for males, it occurs when they turn 55-59 and 75-79. However, the overall distribution of stroke victims based on gender is almost equal, suggesting that gender is not as impactful on stroke risk as some of the other attributes provided.

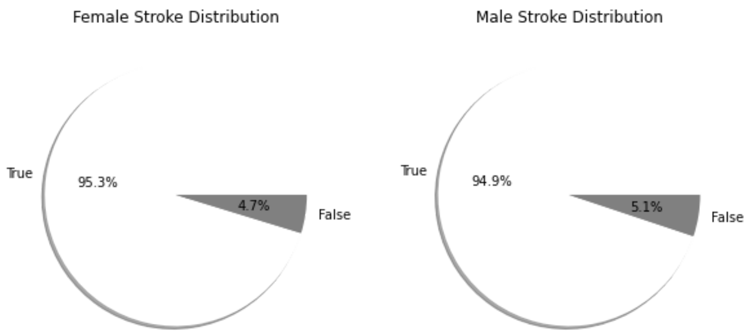

Fig. 3: Percentage of females with stroke vs percentage of males with strokes

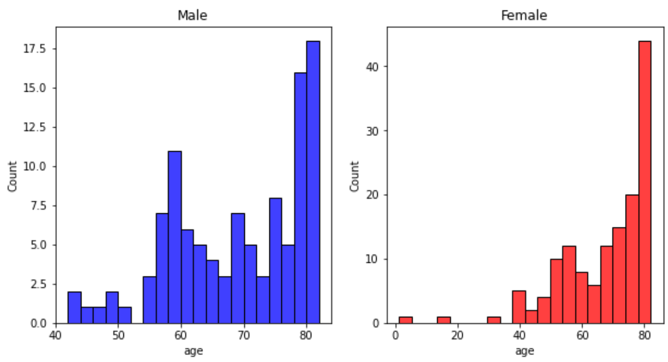

Fig. 4: Age distribution of stroke victims

Continuing this analysis on each attribute, smoking is modified such that all ‘previously smoked’ is changed to ‘smokes,’ just for graphic data purposes. In doing so, it becomes visually apparent that those who smoke/smoked are at higher risk of having a stroke. Smokers only make up 32% of the dataset and 45% of all stroke cases.

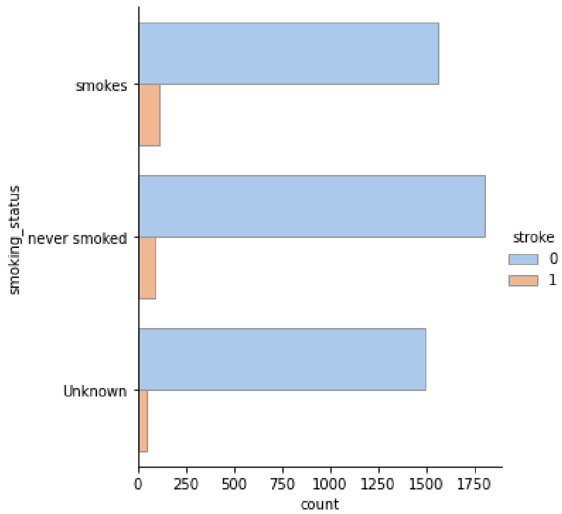

Fig. 5: Barplot displaying stroke instances with their respective smoking status

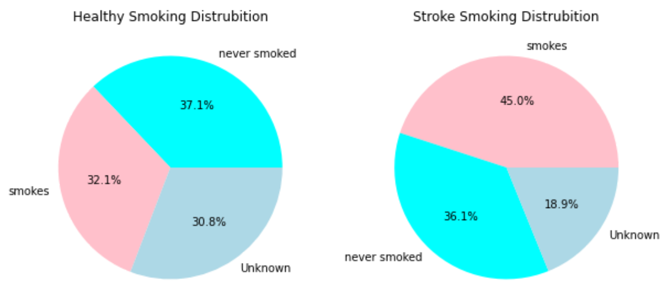

Fig. 6: Non-stoke smoking distribution and stroke smoking distribution

When analyzing hypertension, the same consensus from smoking is applicable, but to much less of a degree. Hypertension only affects around 600 people within the dataset and those who are affected are only at a 15% higher risk, but this may be due to a combination of attributes.

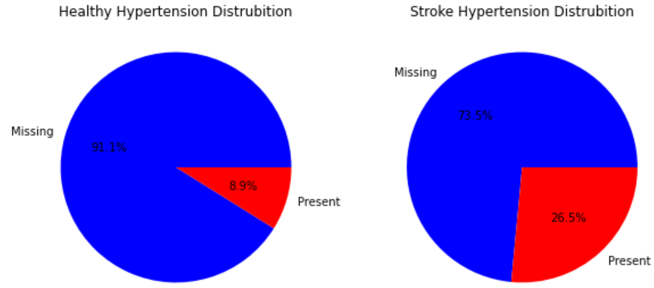

Fig. 7: Distribution of hypertension within the dataset

Similarly enough, heart disease has about the same data distribution as hypertension, a total of less than 600 cases, with those affected only being 13% more at risk. As such, the main attributes that have high correlation with strokes are smoking and age, with the other attributes having some correspondence.

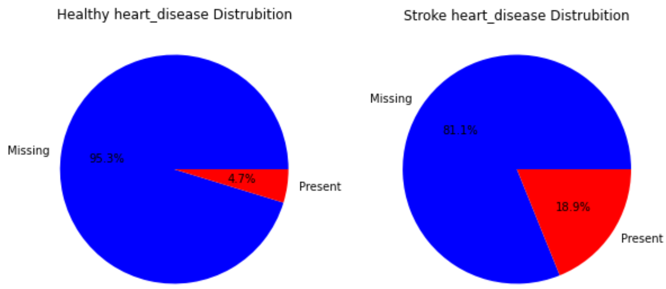

Fig. 8: Distribution of heart disease within the dataset

VI. Python/Jupyter Data Preprocessing

The only attribute remaining to analyze is BMI, but as previously mentioned, it lacked values for some instances. To counteract this issue, the mean of the attribute could be filled in for the null indexes, but in doing so, it could harm further development down the line (trash in, trash out).

While researching, a more favorable method was found for datasets with lesser instances. To fix the issue of nulls, a Random Forest Regressor could be used to predict the BMI values off available attributes. The attributes of age, gender, smoking status, hypertension, and heart disease were fed to the model [16]. Other applicable models for filling the null values include Decision Trees, K Nearest

Neighbor, or K-means. The resulting distribution, seen below, still follows the rules of Gaussian distribution.

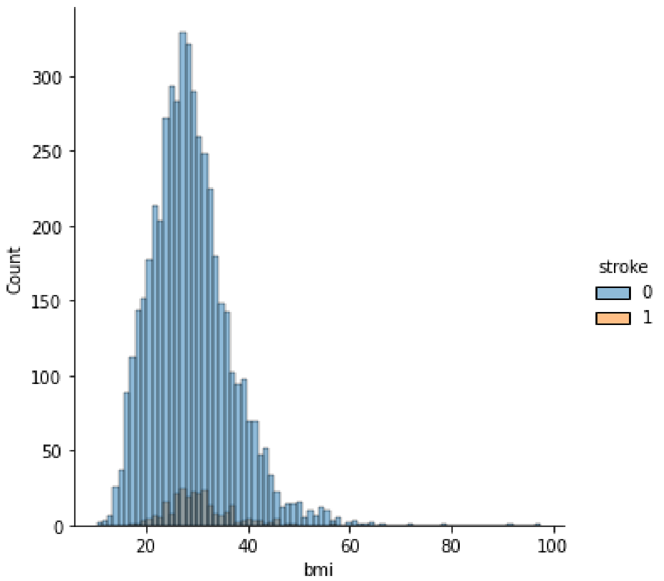

Fig. 9 : Distribution of BMI after filling null values

To feed the data into the training models, it was imperative to transform gender, residence type, work type, smoking status, and ever married using a label encoder such that each instance's attribute value was numerical. Additionally, gender, ever married, and smoking status was transformed one final time using a one hot encoder, so each value had its own respective column. Afterward, the stroke column was popped off the data frame and made as to the Y data, with the remaining attributes of the data frame (minus ID) made as to the X data. The resulting X and Y were transformed into the training and testing set, with an 80:20 ratio split.

The resulting graphics from the exploratory analysis show that the data is extremely imbalanced and requires some more preprocessing before moving on. The SMOTE algorithm was employed to increase the amount of data instances in the minority case to counteract the uneven distribution. This is done by analyzing the minority class and generating new instances based upon what was initially present. Doing so prevents overfitting to some degree, as they are not copied instances, and increases the original training instances size from 4088 to 7786 instances.

The same process is then applied to the testing data. While this is unconventional, models will be tested on both the SMOTE testing data and the original testing data. The resulting two scores provided will dictate how the model will work in terms of a more realistic data distribution (as the training data distribution is structured similarly to that of smote testing data), allowing a closer analysis on the resulting false positive and false negatives, and in terms of actual testing distribution provided from the original dataset. The SMOTE testing set has 1936 instances, whereas the original has only 1022 instances.

VII. Python/Jupyter Base Model Implementation

After the preparatory steps have been concluded for the data, model implementation can finally start. To solve the binary classification problem on hand, the following 11 models will be first implemented barebones and then compared: Logistic Regression, Bernoulli Naive Bayes, Decision Tree, Random Forest, KNeighbors, SVM, XGBoost, AdaBoostClassifier, Cat Boost, XGBClassifier, and MLPClassifier.

All models were appended to a list and then iterated through while calculating metric scores such as accuracy, cross-validation accuracy, precision, recall, f1-score, and ROC AUC to train and test each model. The resulting DataFrame below shows the models ranked based on the accuracy (the prediction accuracy score on the original testing set).

However, deciding the best model with the highest accuracy is not the best metric in the present case. Any model that scored above 90% never predicted more than five-stroke cases on the original dataset, instead of gaining that accuracy entirely off predicting the null class.

	Model	Accuracy	K-Fold Mean Accuracy	Std.Deviation	ROC_AUC	Precision	Recall	F1 Score
6	XGBoost	94.716243	96.687583	7.289349	0.517485	0.500000	0.037037	0.068966
3	RandomForest	94.520548	97.098531	6.869592	0.516452	0.333333	0.037037	0.066667
8	Cat Boost	94.520548	96.854546	7.168889	0.525195	0.375000	0.055556	0.096774
9	XGBClassifier	94.520548	96.841726	7.171412	0.525195	0.375000	0.055556	0.096774
7	AdaBoostClassifier	93.542074	95.904115	7.074843	0.511287	0.125000	0.037037	0.057143
2	DecisionTree	91.291585	94.503945	5.417977	0.525635	0.111111	0.092593	0.101010
10	MLPClassifier	90.998043	93.001046	3.450708	0.576542	0.183333	0.203704	0.192982
1	BernoulliNB	84.637965	86.771931	3.998354	0.656623	0.158940	0.444444	0.234146
4	KNeighbors	80.332681	89.031254	1.208642	0.651381	0.130653	0.481481	0.205534
0	Logistic Regression	77.201566	78.538285	1.311287	0.739765	0.149020	0.703704	0.245955
5	SVM	72.211350	78.088826	1.471652	0.730908	0.129032	0.740741	0.219780

Fig. 10: Models and their respective metric scores

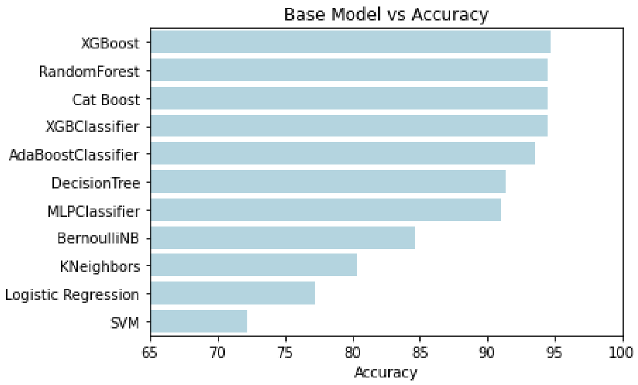

Fig. 11: Base models vs testing accuracy

The best models, in terms of classifying the most stroke cases (minority class recall), were SVM and Logistic Regression, which happened to be the worst models in terms of accuracy; due to these models predicting countless cases of false positives. This is why the aforementioned testing set has been duplicated and modified such that there is also a SMOTE testing set that has been normalized in respect to the SMOTE training set that all models were trained upon.

```
SVM :          XGBoost :
[[698 270]     [[966   2]
 [ 14  40]]     [ 52   2]]
```

Fig. 12: Comparison of the worst (left) and best (right) model's confusion matrix

VIII. Python/Jupyter Model Hypertuning

Moving forward, Random Forest, XGBoost (also known as GradientBoosting), Adaboost, XGB, and Decision Tree Classifier were tested upon both testing sets for a multitude of hyperparameters, such as changing the model's criterion, learning rate, and the number of estimators. The resulting best model in training accuracy and initial testing set accuracy was a Random Forest classifier. The best model in terms of SMOTE testing accuracy was Adaboost and XGB.

To my dismay, all models created could not predict strokes accurately within the original testing set. As mentioned, since the training set had been normalized, it makes each model difficult to predict strokes within the original testing set. To fully gauge and compare the accuracy of each model, SMOTE testing set accuracy will be the primary metric for comparison. Due to this, Adaboost is the best model, hitting a highest of 94% accuracy, accurately classifying 865 stroke SMOTE instances, 103 false negatives, and nine false positives. This is way higher than the SMOTE testing null accuracy, which is only around 50%, due to the normalized class distribution

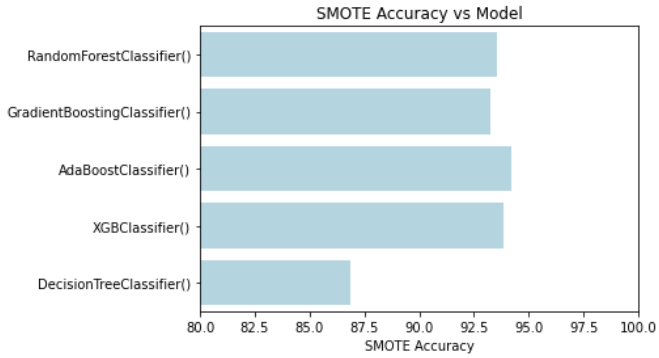

Fig. 13: Hypertuned models vs SMOTE testing accuracy

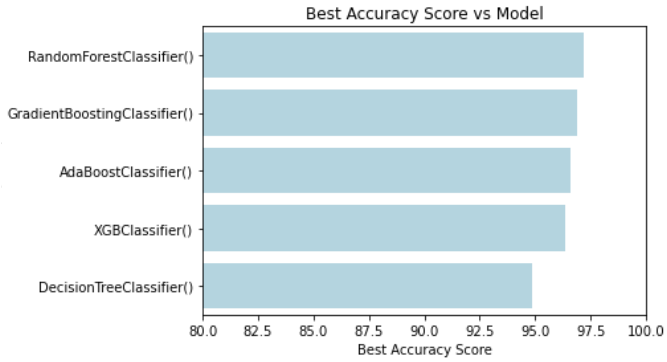

Fig. 14: Hypertuned models vs best accuracy

However more research was conducted into what parameters could be tuned for each model to increase the recall. After testing a multitude of parameters, it became clear that the only model with a tunable parameter to increase recall was XBGClassifier. The remaining models were all built to focus on accuracy, regardless of what metric it was tested upon. By tinkering with XGB's scale_pos_weight parameter, the resulting XBG model was able to predict the following:

```
[[697 271]
 [ 96 872]]
[[697 271]
 [ 21  33]]
```

Fig. 15: Resulting confusion matrix on the SMOTE testing set (top) and original testing set (bottom).

This resulting model predicted the most number of stroke cases accurately and least cases of false negatives when comparing it on both testing sets.

IX. Concluding Thoughts

The main issue encountered during model implementation was the lack of data for the models to train on. To fix the issues, SMOTE was used on the training set and the testing set for metric comparison purposes. The resulting best model for SMOTE training accuracy was Random Forest Classifier (97%) which scored above the baseline null accuracy. The resulting best model for SMOTE and initial testing was using XGBclassifier, hyper tuned using scale_pos_weight. As mentioned above, the best model in terms of recall on the original dataset would still be the SVM, which has not been tuned (to prevent accuracy overfitting). While the result did not conclude a single model for use, it gave us a multitude of options to use as the original dataset provided was unrealistic in the distribution before pre-processing. When the dataset is normally distributed, the use of Random Forest Classifier and XGB would be highly valued. In contrast, when the data is imbalanced and skewed, the use of SVM, Logistic Regression, and XGB would be highly recommended.

Finally, as mentioned, Weka was used for baseline testing and comparison against Python's machine learning library. In the end, Weka was outshined by Python's machine learning and graphics library in all cases. If additional work were to be conducted on the project, the main focus would be the integration of another dataset for more data and discovering other possible models to train upon that learn with respect to recall. In the end, all project proposals were met to their best extent due to Random Forest testing higher than baseline, and final models were chosen and suggested based on what data distribution is present.

X. References

[1] fedesoriano. 2021. Stroke Prediction Dataset - 11 clinical features for predicting stroke events. Retrieved from October 25, 2021 from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[2] CDC/National Center for Health Statistics. 2021. Leading Causes of Death. (October 2021). Retrieved November 10, 2021 from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

[3] National Center for Chronic Disease Prevention and Division for Stroke Prevention. 2021. Stroke Facts. (May 2021). Retrieved November 10, 2021 from https://www.cdc.gov/stroke/facts.htm

[4] N/A. Cox Proportional-Hazards Model. Retrieved November 10, 2021 from http://www.sthda.com/english/wiki/cox-proportional-hazards-model

[5] J. Goeman. l1 penalized estimation in the Cox proportional hazards model. Biometrical Journal, 52(1):70–84, 2009.

[6] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, and Honglak Lee. 2010. An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). Association for Computing Machinery, New York, NY, USA, 183–192. DOI:https://doi.org/10.1145/1835804.1835830

[7] Xuemeng Li, Di Bian, Jinghui Yu, Mei Li, and Dongsheng Zhao. 2019. Using machine learning models to improve stroke risk level classification methods of China National Stroke Screening. (December 2019). Retrieved November 10, 2021 from https://doi.org/10.1186/s12911-019-0998-2

[8] Yu J, Mao H, Li M, et al. CSDC — A nationwide screening platform for stroke control and prevention in China. In: Proceedings of the 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC 16); 2016. p. 2974.

[9] Yan Chu et al. 2021. Machine learning based models to predict functional improvement in stroke patients during inpatient rehabilitation. (September 2021). Retrieved November 15, 2021 from https://www.sciencedirect.com/science/article/abs/pii/S0003999321011576

[10] C. Hung, W. Chen, P. Lai, C. Lin and C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 3110-3113, doi: 10.1109/EMBC.2017.8037515.

[11] Zihni E, Madai VI, Livne M, Galinovic I, Khalil AA, Fiebach JB, et al. (2020) Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. PLoS ONE 15(4): e0231166. https://doi.org/10.1371/journal.pone.0231166

[12] Mayo Clinic Staff. (2021, July 01). High blood pressure (hypertension). Retrieved December 04, 2021, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410

[13] Diabetes. (2021, September 17). Retrieved December 04, 2021, from https://www.stroke.org.uk/what-is-stroke/are-you-at-risk-of-stroke/diabetes

[14] Olsen TS. Blood glucose in acute stroke. Expert Rev Neurother. 2009 Mar;9(3):409-19. doi: 10.1586/14737175.9.3.409. PMID: 19271949.

[15] Spence JD. Nutrition and Risk of Stroke. Nutrients. 2019 Mar 17;11(3):647. doi: 10.3390/nu11030647. PMID: 30884883; PMCID: PMC6470893.

[16] Factors affecting weight & health. (n.d.). Retrieved December 04, 2021, from https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/factors-affecting-weight-health