Stroke Prediction Project

Christopher Alfonso

PID: 6314071

Introduction

Strokes are ranked the second highest cause of death worldwide. Thousands of people die due to strokes each day and millions per year, according to WHO which is short for the world health organization “Annually, 15 million people worldwide suffer a stroke. Of these, 5 million die and another 5 million are left permanently disabled.” In my project I will use machine learning to predict whether or not a patient is likely to get a stroke. In the past projects and research have addressed and tackled this issue using machine learning in order to predict and possibly prevent strokes from occurring.

A challenge that previous researchers encountered is the imbalance in the distribution of participants in the stroke and non-stroke classes (Dritsas, E., & Trigka, M. 2022). This is due to there being a far greater percentage of people who have not had strokes compared to those that have. Researchers have solved this challenge by using the SMOTE (Synthetic Minority Oversampling Technique) to balance the minority class by creating synthetic examples based on existing data. This imbalance problem is also consistent with my dataset and I will have to employ a similar technique.

In my project I will be using the KNN, Logistic regression, and SVM algorithms. I will implement my system in Python using Jupyter labs, as it is one of the best and most highly recommended development environments for machine learning. In order to check the overall performance in my system I plan to compare to existing models/projects as well as the models overall performances compared to each other.

Related Works

In the Stroke Risk Prediction with Machine Learning Techniques paper, they are using a very similar data set to my own which consist of electronic health records of patients who have

both have and have not had strokes that are 18 years old and older. Other data sets used in papers such as Machine learning in action: Stroke diagnosis and outcome prediction use brain scans in order to predict Hemorrhagic strokes. But I will be mainly referring to the Stroke Risk Prediction with Machine Learning Techniques paper in this section since it is more inline with my project.

They used a total of 11 attributes 10 input being Age,Gender, Hypertension, Heart disease, ever married, work type, residence type, glucose level, BMI, and smoking status and one target attribute being “stroke” which divides into yes and no. Age having the most information gain out of all the attributes. Furthermore, they did not employ any normalization techniques, but did have the aforementioned imbalance challenged while preparing their data in which they used SMOTE.

They used a variety of machine learning models, the reasoning behind this was to test them against each other in order to find the most effective ones. All of the models they used are naive Bayes, random forest, logistic regression,KNN, decision tree, multilayer perceptron, majority voting, and stacking. In order to evaluate the performance of their data they used precision, recall, F-measure, area under curve, and accuracy. Their results are in the chart below.According to this data stacking outperformed all other methods.

	Precision	Recall	F-Measure	AUC	Accuracy
NB	0.812	0.860	0.835	0.867	0.84
LR	0.791	0.791	0.791	0.877	0.79
3-NN	0.918	0.916	0.915	0.943	0.81
SGD	0.791	0.791	0.791	0.791	0.88
DT(J48)	0.909	0.909	0.909	0.927	0.91
MLP	0.884	0.881	0.881	0.929	0.92
MVoting	0.93	0.93	0.93	0.93	0.93
RF	0.966	0.966	0.966	0.986	0.97
Stacking	0.974	0.974	0.974	0.989	0.98

My project will be conducted in a way that is very similar to this one but I would like to conduct my project in a scale that is more relevant to the things I have learned throughout this course. So I will be mainly focusing on methods like KNN and logistic regression in order to predict strokes. Ultimately, I would like to compare the results of my project with their research, as a way to evaluate my data.

Data

I will be using similar attributes to the research conducted in Stroke Risk Prediction with Machine Learning Techniques paper. For my input parameters I will use basic information such as age, gender, BMI, smoking status, cardiovascular disease, hypertension and glucose levels as my input attributes, and stoke as my target value, where it will return “1 (yes)” or “0” (no) based on how likely or unlikely someone is of having a stroke.

For my project I will be using a data set from Kaggle containing health records of patients, this dataset will be referenced below. As for preprocessing my data I will do the usual of fixing missing or incorrect and repeating values. As mentioned before, one of the challenges I will face while preprocessing data is the imbalance of the data, so I will employ SMOTE or other similar techniques in order to balance the data. I do not plan to use any normalization techniques in my project.

Implementation

Libraries Used

I used various Python libraries in order to complete my project. The first ones being OS and Pandas who were used in order to import and read the file. Pandas was also used for various other smaller scale actions like hot encoding data. Other libraries that were used were

seaborn and matplotlib. The main purpose that these libraries served was in order to be able to visually depict datasets and results using graphs, boxplots and things such as heat maps.

Libraries that were more crucial and played a bigger role in this project were imblearn and sklearn. Imblearn is a library that is mostly used and has tools in order to solve issues with imbalanced data sets. This library was used to fix the huge imbalance in the dataset I used through the SMOTE method to create new instances of data in the minority class. Sklearn was the most used library throughout this project, various modules and classes in this library allowed for the creation of the machine learning models I used, to things helpful tools when preprocessing data. It also provided tools to show the performances of the models used.

Pipeline

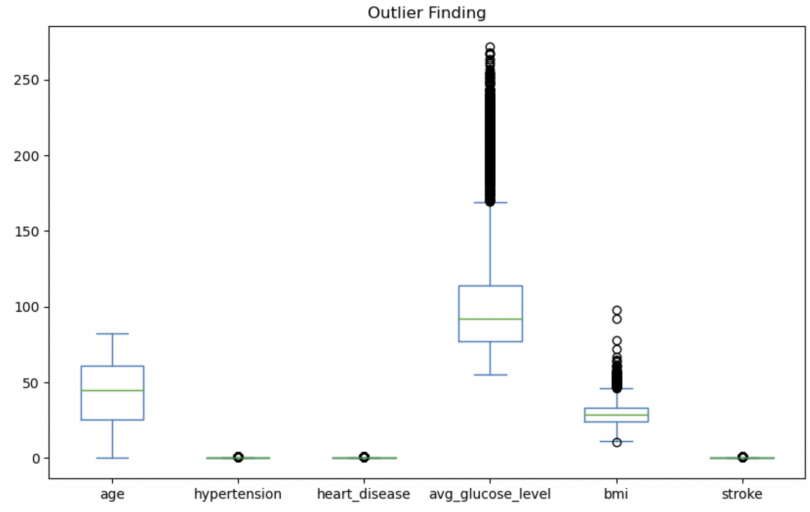
The general approach and order that this project was completed in was pretty straight forward. First I loaded in the data set, and then I did a bit of exploratory data analysis in order to familiarize myself with the data set and find issues with the data set such as things such as missing values. After this I started data manipulation/preprocessing, in this process I did things such as remove columns, fill missing values and remove outliers. Then, I did some modeling and exploring of the target feature finding a huge imbalance in the data set, during this part of the project I also did a practice referred to as “hot encoding” where I dealt with the categorical values.

After this I did more preprocessing, I fixed the imbalance issue with the SMOTE technique, split my data into training and test sets, then finally scaled my data. After all of the preprocessing and data manipulation was done I started working on the machine learning models. I implemented 3 main models, K-Nearest-Neighbors-Classifier(KNN), Logistic regression, and Support Vector Machine(SVM), which all serve as good models for classification tasks. After implementing each separate model evaluated their results through a confusion matrix and metrics such as precision, recall, F-measure, and accuracy. When I was done evaluating these

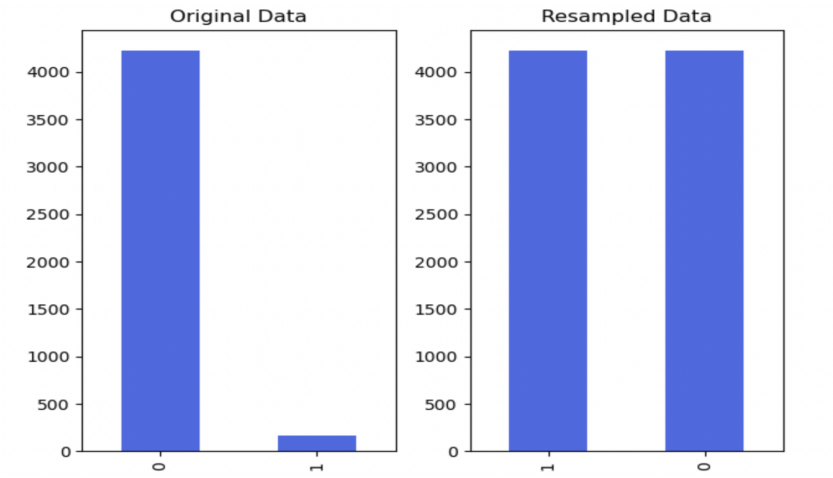
models I implemented a stacked model otherwise known as an ensemble model, here I combined the predictions of all previous models in order to try to receive more accurate results. Then I compared the results of all the models where the stacked model and KNN performed the best both having an accuracy of around 0.96.

Data Preprocessing

There was a good amount of data preprocessing done in this project as mentioned in the pipeline section. First I got removed categories that were useless or I felt do not show a direct indicator of bodily health such as 'id', 'Residence_type', 'ever_married', and 'work_type'. Then there was filling in missing values in the “bmi” categories which I did using the mean. There were outliers in both the glucose and bmi categories which I solved by dropping from the data set. Then I hot encoded the data in order to express categorical data like smoking status as binary numbers, to make it easier to process. Shown below are outliers present in their respective class



After all that I used the SMOTE or Synthetic Minority Oversampling Technique in order to balance my data set.



As shown in the graph above there was a huge imbalance in the target variable of my data set, more specifically 4226 to 165 cases which I completely balanced using the SMOTE technique which works by generating samples of the minority class.

I then split my data into the testing and training set in which I used a ratio of 75-25 percent for each respective set. Then I finally normalized the values in my dataset using StandardScaler which is a very common way to normalize data and works by normalizing each feature to have a mean of 0 and a standard deviation of 1.

Classifiers

As mentioned before there were 3 main classifiers/models that were used in this project. K-Nearest-Neighbors-Classifier(KNN), Logistic regression, and Support Vector Machine(SVM), which predictions were then combined into one stacked model. Since stroke prediction is a binary classification task I felt that these models would perform well. I tried to stay within the realm of models discussed during this course.

Briefly explaining how each of these models works: KNN is a classification algorithm that calculates the distance between two points in a feature space. Logistic Regression uses a logistic function to model the probability of a data point belonging to a specific class. SVM works by

finding the optimal hyperplane in a space to separate data points that belong to different classes, in order to maximize the distance of the margin of the hyperplane.

Performance Techniques

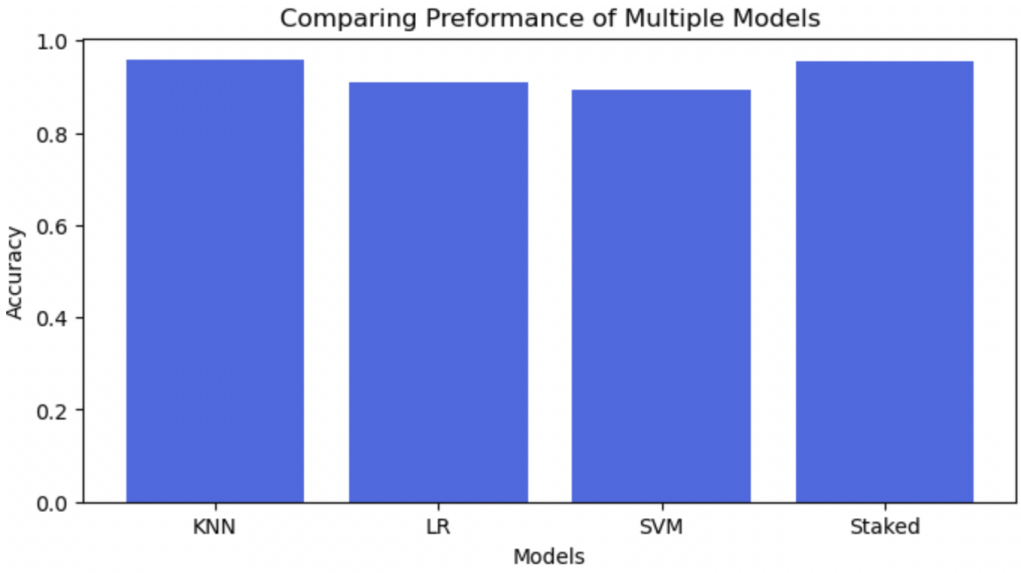
As mentioned before the performance metrics used in this project are precision, recall, F-measure, and accuracy which are all pretty straight forward and common ways to measure performance. I also created a confusion matrix for each of the models. These results were calculated using the sklearn library and modeled my data using bar graphs.

Results

Findings

In general I am quite happy with the findings of my project, overall all the the models performed well. The results of my project will display in a table below:

	Accuracy	Precision	Recal	F-1	Support
KNN	0.957	0.96	0.96	0.96	2113
LR	0.911	0.91	0.91	0.91	2113
SVM	0.894	0.91	0.90	0.89	2113
Stacked	0.956	0.96	0.96	0.96	2113



Confusion Matrix:

KNN	LR	SVM	Stacked
[[980 60] [30 1043]]	[[987 53] [135 938]]	[[1040 0] [224 849]]	[[999 41] [51 1022]]

As you can see from the table and bar graph, results across the board were all pretty favorable. The clearcut best model was the KNN model with an accuracy of 0.957 and similar numbers in other categories. Accuracy wise the worst performing model was the SVM, which still did pretty good and had an accuracy of 0.894. Furthermore, the stacked model performed very well and very similarly to the KNN model but I honestly expected it to be the best model by a landslide. In general, the results of my project show that machine learning can indeed be used for stroke prediction and possibly other medical tasks.

Comparisons to Other Works

Compared to the Stroke Risk Prediction with Machine Learning Techniques paper mentioned in the related works section my results were pretty different. In their results Logistic

regression had an accuracy of 0.79 and KNN had an accuracy of 0.81 which all fall near the bottom of their list in terms of performance, their highest performance was the stacked model with 0.98 accuracy rating which isn't that far off from my stacked model with an accuracy of 0.96. They did not use the SVM model.The differences in results could be due to many things as both projects were conducted in a vastly different manner. Although having varying results both of our projects were successful in their own right, and prove that machine learning can be used effectively in stroke prediction.

Final Discussion

Future Work

Although my project performed well, there are several improvements and additions that can be done in order to both advance and improve my work. One of the things that I would change is that I would include the other categories that I dropped in my research values that I previously thought didn't have much to do directly with bodily health if implemented could have produced vastly different results. I would have also tried other ways of encoding my categorical data.

Another change I would make is the addition of even more machine learning classification models in hopes of finding a more effective model and possibly improving the results of the stacked model. Other models that I think could have performed well in this project are Naive Bayes, decision trees and random forest.

Lastly, another way that I would like to improve this project is something that I had tried to code but wasn't very successful and didn't get very accurate results from, which is a way for a user to manually input their own information and find out whether they are at possible risk for a stroke or not. As mentioned before I attempted this in my project but wasn't getting accurate results

so I scrapped the idea, maybe in the future when I have more knowledge in machine learning and time I can implement something similar that works.

In conclusion, my system for the most part went well and performed favorably, everything worked as intended and there were no major issues. Differences in results between the work of others and my own work as stated before could be due to a plethora of reasons such as different datasets, variables, models used, ways of preprocessing and normalizing data, etc. In general, I think my project served its purpose in showing that machine learning models can be used in the medical field as a helpful tool, in this case more specifically for stroke prediction which can lead to stroke prevention.

References

Dritsas, E., & Trigka, M. (2022, June 21). *Stroke risk prediction with machine learning techniques*. Sensors (Basel, Switzerland). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9268898/#B34-sensors-22-04670

Mainali, S., Darsie, M. E., & Smetana, K. S. (2021, October 28). *Machine learning in action: Stroke diagnosis and outcome prediction*. Frontiers. https://www.frontiersin.org/articles/10.3389/fneur.2021.734345/full

World Health Organization. (n.d.). World Health Organization. https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

Data set: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset