Stroke Prediction Using Machine Learning Techniques

Momin Shahzad 2020098
Suliman Bek 2000819
ANAS Aljanabi 2019965
School of Engineering and Natural Sciences
Bahcesehir University

Abstract

This research focused on the analysis of a Kaggle dataset containing health records of 20,000 patients with 12 attributes, some of the data are generated synthetically. We used LightGBM and XGBoost algorithms to predict the likelihood of a stroke based on these attributes. Our results revealed that Age, BMI and Glucose level are among the most important variables that significantly influenced stroke prediction. Our models achieved an ROC AUC score of 0.888. These findings underscore the potential for machine learning in healthcare, specifically in aiding early detection of critical conditions such as strokes.

1 Introduction

Stroke is a serious medical disorder caused by ruptured blood vessels or decreased blood supply to the brain. Unhealthy habits raise the risk of stroke, highlighting the importance of a healthy lifestyle and early recognition. With an increased synergy the healthcare system has difficulties in providing quality services and correct diagnosis. Technology and medical records can help enhance managing patients. The relationship of risk factors in patient health records is critical for stroke prediction. This study employs data mining and machine learning to identify and predict strokes in order to provide appropriate care while avoiding significant effects.

1.1 Literature Survey

The role of data mining and machine learning in healthcare has been a subject of investigation in numerous studies. Machine learning techniques have been used extensively in recent years for predicting various medical conditions. Stroke prediction, in particular, has been a focus due to the severity of the disease and its associated risk factors. Weng et al. [1] applied machine learning methods to predict stroke in a study of over 378,256 patients using electronic health records from two large UK hospitals. Their study focused on the comparison of machine learning methods to standard prediction models. Similarly, our study employs machine learning techniques but on a dataset acquired from Kaggle, with a focus on comparing different machine learning algorithms for stroke prediction. Kamal, Ullah, and Ahmad [2] developed a stroke prediction model using a Naïve Bayes classifier. They reported an accuracy of 73.6, which is lower than the accuracy rate we achieved. Our study includes the use of the Naive Bayes model but also explores other methods like Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, and Support Vector Classification. Liaw et al. [3] utilized Decision Tree, Naive Bayes, and k-Nearest Neighbor algorithms to predict the risk of stroke in patients. They achieved an accuracy of 85, which aligns with our findings, as we also used these algorithms and attained a similar accuracy. In conclusion, our study aligns with and builds upon the methodologies and findings of previous research in stroke prediction using machine learning techniques. Our study differs in the range of machine learning methods applied, the level of accuracy attained, and the preprocessing techniques employed.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 SYSTEM METHODOLOGY

For implementation a Kaggle datasets was used. The dataset was preprocessed before modelling. Preprocessing included dealing with missing values, uneven data, and label encoding. The models used were LightGBM and XGBoost. ROC AUC score was used to compare these models, these models were tuned to the best hyper parameters and put together in an ensemble.

2.1 Dataset & Data Loading and Pre-processing

Kaggle provided the dataset for stroke prediction. This dataset contains 20,414 rows and 12 columns. The important attributes of the column are: id, gender, age, hypertension, heart disease, ever married, worktype, residence type, avg.glucose level, bmi, and smoking status. The target variable 'stroke' has a value of '1' or '0', 1 if the patient had a stroke and 0 if not. This dataset is extremely unbalanced since the likelihood of a '0' in the target stroke column stroke surpasses the possibility of a '1'. Only 249 rows have the value '1' in the stroke column, while 4861 rows have the value '0', this is in the real and synthetic data. To help improve accuracy, data preprocessing is used to balance the data. The data is loaded using the load_data function, which reads CSV files and concatenates them into a single Data Frame. The preprocess function is then applied to the data to perform various preprocessing steps. The preprocess function handles data cleaning and creating some features useful in plotting. Categorical variables are encoded using one-hot encoding, and some numerical features are transformed or binned into categories for plotting.

2.2 EDA and Modeling

To look into this dataset, several visualization techniques were used. The "plot_var_stroke" function was implemented to show the occurrence of strokes in both genders (Male Female). Bar plots were created to demonstrate the stroke distributions which are stroke rate among different features. Kdeplot (KDE) plots are used for displaying the age distribution based on marriage status and smoking status. After loading the data and exploring it, we do two things before modelling first, Feature Engineering, we create some features that might help the models using the "generate_features" function which is defined to generate additional features based on the existing variables in the dataset. New features created include 'age/bmi', 'age*bmi', 'bmi/prime', 'obesity', and 'blood_heart'. Secondly, the "preprocess" function is defined to preprocess the data before model training. One-hot encoding is applied to categorical variables, the function returns the preprocessed train dataset. If a test dataset is provided, it is also preprocessed and returned. The preprocessed train dataset is split into features (X) and target variable (y). To address class imbalance, the upsample function performs upsampling of the minority class (stroke) using resampling techniques. The LightGBM classifier is trained on the upsampled data and used to predict stroke probabilities for the original data. The area under the receiver operating characteristic curve (ROC AUC) is computed as an evaluation metric on the train set.

3 Results

3.1 EDA and Modelling

The results of the Exploritoray Data Analysis where very insightful in this section we are going to show the plots and the insights they provide: First we take a look at the target variable.

Fig 1. Shows the count of the target stroke variable, we can see that the dataset is very unbalanced with very little patients who had stroke. Next we check the relationship between the features or the independent variables and the target. We start with the age variable, we split the continous age variable into 4 categories to plot a bar plot as seen in Fig 2.

The Gender plot in Fig 3 shows slightly higher stroke rate for Males than Females but its not important. The Hypertension and Heart Disease Plots shows that people with Hypertension or Heart Disease have significantly higher risks of stroke so these might be important variables.

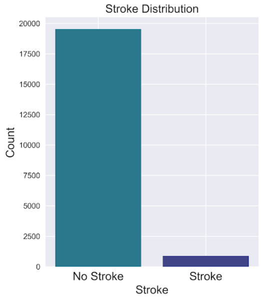


Figure 1

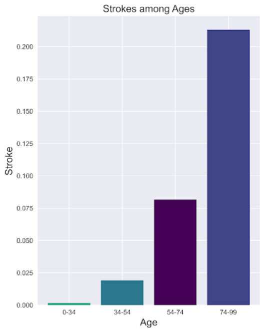
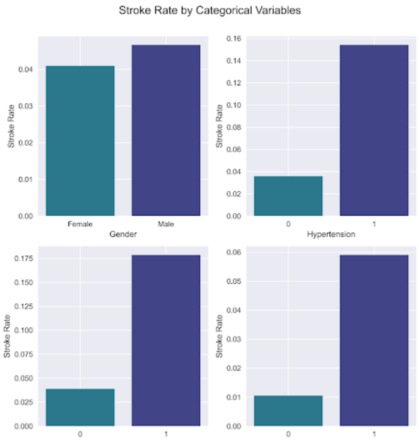


Figure 2

The ever_married shows that people who have reported being married before have much higher stroke rate, so it looks that this feature is important, however this variable should not have this much of an effect so it is likely that there is an underlying cause, after plotting the age distribution of the two columns we can see whats the actual cause, which is that the people who're married are older on average than people not married as shown in Fig 4. We can look at the ever_married feature for each age group to get a better understanding of the effect of the feature, Fig 5.



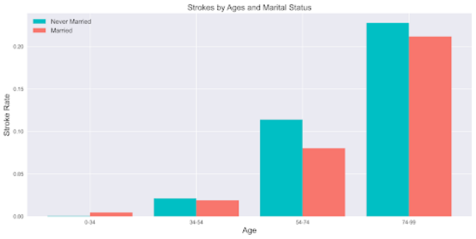


Figure 4

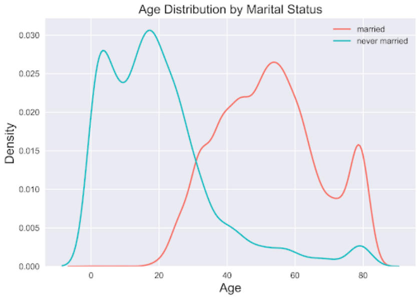


Figure 5

So by checking the plots we can see that the actual reason for the high stroke rate in married people is because married people have higher average age so they have higher stroke rate since age is a strong predictor of stroke. In Fig 8 we plot the glucose level and the bmi index features. Glucose level and weight seem to be important predictors, with higher weight and higher glucose levels related with higher stroke rates The baseline LightGBM model gets a score of 0.84, then after feature engineering and hyper parameter optimization for the two ensemble models we get a score of 0.88, in Fig 9 we can see the feature importances of the ensemble model.

4 Discussion

The findings of this study demonstrate the effectiveness of data mining and machine learning techniques in predicting strokes based on the analyzed dataset. The identified predictors, such as age, hypertension, heart disease, smoking status, glucose levels, and BMI, provide valuable insights for

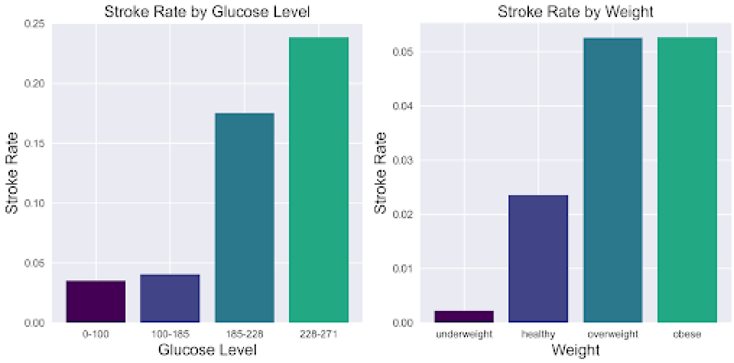


Figure 8

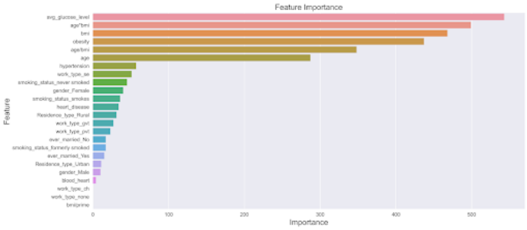


Figure 9

stroke risk assessment. However, there are still potential avenues for improvement and future research. One possible approach could be to explore more advanced machine learning algorithms or ensemble methods to further enhance the accuracy of stroke prediction models. Additionally, incorporating additional relevant features or exploring the use of advanced imaging techniques could potentially improve the predictive power of the models. Moreover, conducting larger-scale studies with more diverse and representative datasets would help validate and generalize the findings. Furthermore, integrating real-time patient monitoring data and leveraging artificial intelligence techniques could enable the development of personalized stroke prediction models for early detection and intervention. Overall, further research in this area holds great potential for advancing stroke prediction capabilities and improving healthcare outcomes.

5 References

1. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, 12(4), e0174944.
2. Kamal, A. H. M., Ullah, S. Z., & Ahmad, M. (2016). Stroke prediction using machine learning algorithm in cloud-based electronic health records system. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 7(4).
3. Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.
4. Khan, Y. U., Farooque, M., Mehmood, A., & Khalid, S. (2019). An effective framework for predicting the human stroke disease. *Soft Computing*, 23(10), 3287-3299.

6 Contribution

Student Anas Aljanaby wrote the code for the data analysis, student Momin Shahzad wrote the code for the modelling, then Student Suliman Bek prepared the report and the latex formatting.