PREDICTIVE ANALYSIS FOR STROKE PREDICTION USING 2012-2016 NHANES DATASET

DS201: FINAL PROJECT REPORT

Submitted By

Team 3

Eti Dandekar (12140650), Garima Tata (Student ID: 12140690), Kriti Gupta (Student ID: 12140940)

Course Instructor: Dr. Nitin Khanna

Department of Electrical Engineering and Computer Science

Indian Institute of Technology Bhilai

September 18, 2023

Contents

Keywords Predictive Analysis · Machine Learning · NHANES · Logistic Regression · SVC · Random Forest Classification

1 Introduction

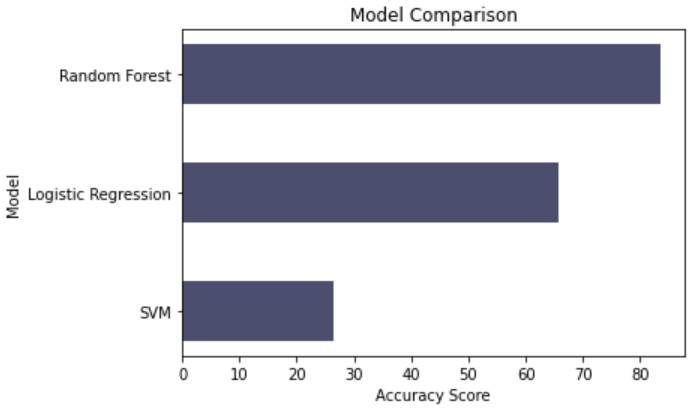
Predictive analysis aims at building an analytical model to predict a target variable.The results from these systems can be useful in marketing fraud detection and making decisions in the medical field. Data analysis and machine learning algorithms are rapidly growing areas and this article focuses on predictive analysis on healthcare data set based on Demographics, Examination data, Laboratory data, Questionnaire, Dietary data sets. We have performed this analysis on NHANES data from 2012-2016 to prepare our analytical work.
The National Health and Nutrition examination survey(NHANES) is a program conducted by the US government that aims at assessing health and nutrition status from different population groups across the US. All the files in this data set were in XPT format which we converted to .CSV format, then we cleaned and pre-processed the data. We further used Logistic Regression(LR), Random Forest Classification(RFC) and Support Vector Classification(SVC) to predict the occurence of stroke in an individual from the final dataset that we got after merging the various datasets. After performing the predictive analytical work we achieved different accuracy using each model.Based on these results we compared the three models and got our final observations.

2 Materials & Methods

For this analysis we used Both Python and R to perform various steps in order to reach to our final result. Both python and R provide important tools and machine learning libraries that support Data Mining. While performing our analysis we discovered that R is better then python for some data pre-processing phases. Since it has very good libraries for manipulating large size of data. We used R for converting XPT files to CSV files.
To convert the XPT files into CSV files we used a package from R called foreign. Then we merged all the data belonging to a single category from data files over multiple years into a total of five different files namely demographic, examination, questionnaire, labs and diet. Then we merged all these five files into a single file. After that in our next step, we excluded any rows which had null values or NA for MCQ160F, as this is the most important factor in our prediction of stroke.After this, we applied some data cleaning algorithms on our dataset. We have trained the model by splitting our data set in train and test.
We used feature selection methods like XGBoost and Upsampling with SMOTE before model training. We trained our model using Logistic Regression, Random Forest Classification and SVC and then we have performed model comparison comparing their accuracies.

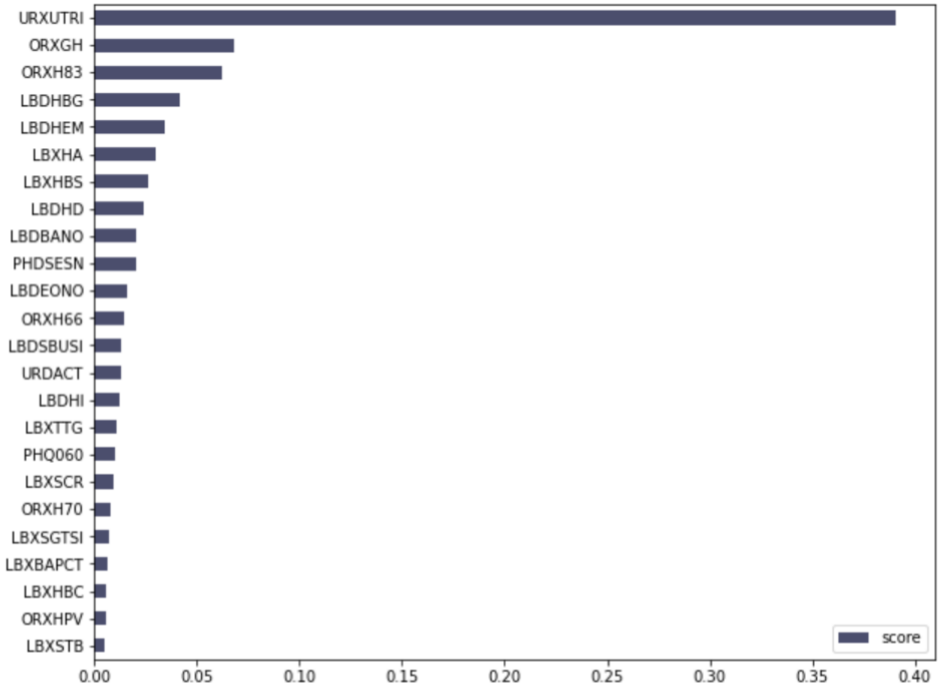
3 Results

After training our model using the three methods- Logistic Regression(LR), Random Forest Classification and SVC, our final step is to compare the three models for their accuracies which will further be helpful to us in deciding which method is the best for stroke prediction algorithm. On training our models, we found out that the accuracy score using LR is 0.6568, for RFC is 0.8352 and for SVC is 0.2636.
Thus RFC gives us the most accurate results out of all these three models.



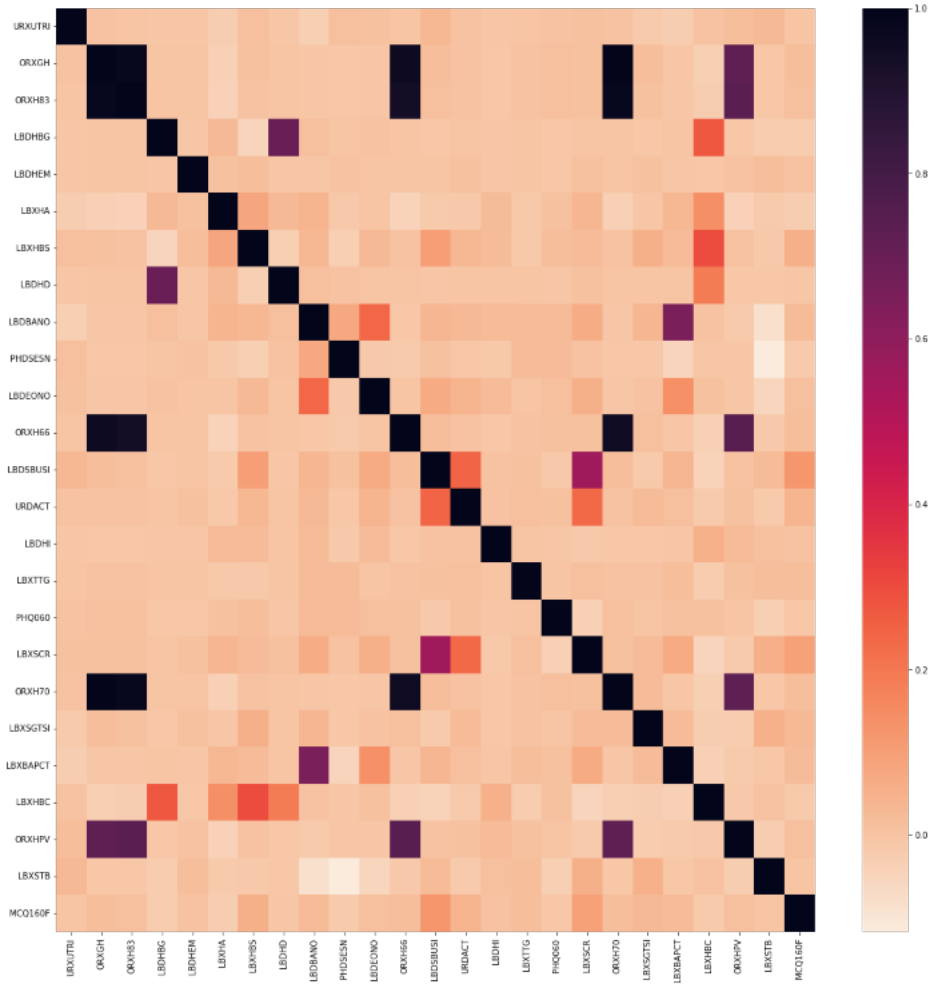
4 Discussion

To convert the XPT files into CSV files we used a package from R called foreign. then, we merged all the data belonging to a single category from data files over multiple year into a total of five different files namely demographic, examination, questionnaire, labs, diet. Then we merged all these five files into single file. After that in our next step, we excluded any rows which had null values or NA for MCQ160F, as this is the most important factor in our prediction of stroke.
For data cleaning, we excluded the non-numeric values and columns that have over 50% NaN, after this we were left with 153 columns. then we changed target variable coding from 1,2 to 0(negative), 1(positive). We found the median of each column and replaced the NaN with median.
Using XGB classifier, we performed feature selection to get the most important and relevant features for our prediction. After splitting our data set into train and test we saw that our data was very imbalanced and inconsistent as most healthcare datasets are. thus, we needed to upscale the minority classes in our data using SMOTE (Synthetic Minority Oversampling Technique) to perform further operations such as Model Training and Comparison. Just after performing feature selection using XGBClassifier, our confusion matrix has an accuracy of 96.33%. Our final confusion matrix that we get after Upsampling using SMOTE, has an accuracy of 91.54%.Then we performed feature selection with XGBoost and plotted the variables v/s their score.



According to the score, we rest the index of our data after that we merged SAS labels from the codebook. we listed out the final variables of our data set and filtered them one last time. This gave us a total of 5583 rows and 25 columns.

We plotted a heat map to check the correlation between the final variables with each other.



—— As we know that MCQ160F(has a doctor or other health professional ever told you that you had a stroke?) is very important factor in our prediction and gave Yes or No value we can not include it in the training model.So, we had to exclude it from the training model and include it in the testing model we then normalised the data set using MinMaxScaler imported from Sklearn.pre-processing.

4.1 Model training

We started model training with first method which is Logistic Regression.
From this we got the accuracy score as 0.6568.

We got the following confusion matrix.

	Predict[0]	Predict[1]
True[0]	895	461
True[1]	18	22

Then we have done Random Forest Classification and got the accuracy score as 0.8352.

We got the following confusion matrix.

	Predict[0]	Predict[1]
True[0]	1143	213
True[1]	17	23

After that we performed SVC and got the accuracy score 0.2636.

We got the following confusion matrix.

	Predict[0]	Predict[1]
True[0]	333	1023
True[1]	5	35

We finally performed model comparison for all three models and according to their accuracy score made a plot for Model v/s Accuracy score .

Conclusions

Stroke is the fifth cause of death in the USA according to a 2020 report. If such prediction systems are designed, many lives could be saved on the way.
In future, we wish to further refine our model and make it more accurate. Also, we wish to design a user interface system for stroke prediction in the future based on this project.

Acknowledgements

We would like to thank Dr. Nitin Khanna for providing us the opportunity to work on this project and the TA for the course for being open to our each and every doubt.
Also we would like to thank our seniors who helped us whenever we ran out of ideas to tackle a bug.

References

Kaggle
TowardsDataScience
AnalyticsVidhya