MASTER OF SCIENCE IN DATA ANALYTICS

B9DA103 - Data Mining

"Stroke Prediction"

Submitted by

Mohanraj Jayakumar – 10577457

Group Members

Sharvari Devdas Narvekar - 10576559

Vinay Ramasare Kurmi – 10576078

Supervised by

Terri Hoare

Acknowledgement

We take this opportunity to show my gratitude by thanking my Lecturer Ms. Terri Hoare for all her efforts and continuous showering of knowledge on the course module - B9DA103 - Data Mining. The online sessions have been more interactive and interesting since from the day one and all her ideas about this subject helped us in making this report successful. Furthermore, the notes provided in the Moodle was so advantageous and handy for all the works which is done here to complete the “Stroke Prediction” report using data mining technology.

Business Understanding – Stage 1

Background

The initial stage of understanding the background of business objectives of the process. Data mining now and then called as Knowledge Discovery in Databases are the way toward dissecting information from various data sets. What's more, converting into valuable data that can be utilized to foresee the datasets and extract knowledge to take or settle on choices in future exchanges. This study is about the stroke prediction which is a life-threatening illness that has been positioned second driving reason for death all over the world. We are going to use the CRISP-DM methodology for understanding the business objectives. The abbreviation of CRISP-DM-Cross Industry Standard Process for Data Mining. The CRISP-DM system gives an organized way to deal with arranging a data mining project

Introduction

A stroke is serious disease that happens when the blood supply to a piece of the mind is cut off, loss of motion, abrupt torment in chest, discourse weakness, loss of memory and thinking capacity, trance like state, or passing. And in most cases, it is recognized for clinical blunder happens due to expiry of meds, inaccurate medications, mistaken measurements, and treatment given to some unacceptable patient. Stroke influences the individual on any age, and it is urgent treatment is fundamental [2]. The sooner an individual gets treatment for a stroke, the less harm is probably going to occur. Therefore, we decided to predict the strokes on humans based on the given data and prediction accuracy makes the change in health domains.

Keywords: CRISP-DM, Stroke, Prediction, Data mining classification

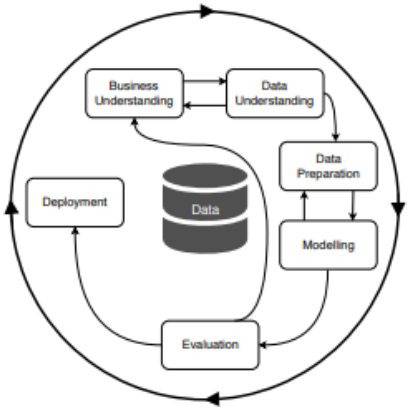


Figure 1: CRISP-DM process model [1]

The above diagram is the process of preparing the model for any business ideology and these loops are followed and executed in our reports.

Aim of the Research

This dataset is utilized to foresee whether a patient is probably going to get stroke dependent on the information boundaries like sex, age, different illnesses, and smoking status. Each line in the information gives relevant data about the patient [3]. The principal goals of this examination are twofold

- Use information mining strategies to anticipate patient in danger of creating stroke.
- Find the patient with who has higher opportunities to predict stroke.

Predictor Variables

The variables that going to be implemented for prediction and which navigates us to choose which model is best comparing with other models. The variables are provided below,
Gender, age, hypertension, heart disease, marital status, wort type, residence type, glucose level, bmi, smoking status

Target Variable

In every case we must fix the targetable variable to predict the final output, in this paper we are considering the "Stroke" attribute for predicting the result

Data Understanding – Stage 2

Data Collection

The second phase of the CRISP-DM measure the data format and data resource. This underlying assortment incorporates in information stacking for data understanding. For instance, if we utilize a particular apparatus for data understanding, it suits well to stack our information into the device. If there is a chance of using two different datasets, we need to think about how and when you will coordinate and relate the sources. Since we use single dataset with multiple related variable RapidMiner found to be the best one for execution.

- Datasets consist of 12 columns and 5111 rows.
- We used the dataset provided by Kaggle
  https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

This pie chart describes the gender type, there is an inflation in male with 59% and deflation of percentage as 41 for female.

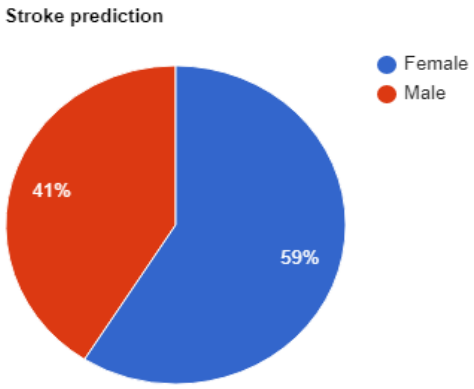


Figure 2: Prediction Rate between genders

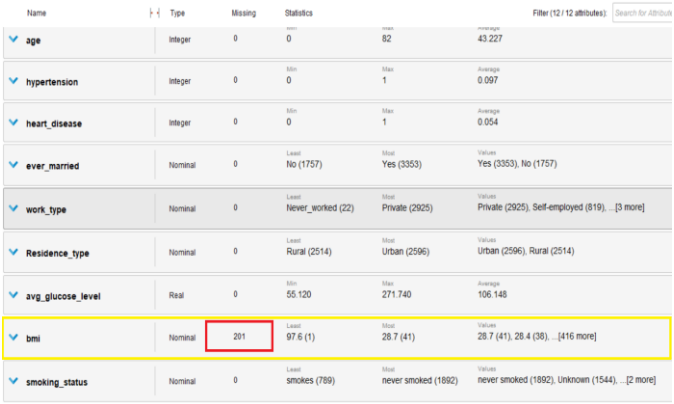
Data Exploration

This explains the results of our data exploration of initial findings and initial hypothesis of each row and columns. They are detailed exploration of patients
1) Id: Unique Values
2) Gender: Male, Female and Other
3) Age: Patient age
4) Hypertension: 0 = Doesn't have hypertension, 1= Have hypertension
5) Heart disease: 0 = doesn't have any heart diseases, 1 = have a heart disease
6) Ever married: No or Yes
7) Work type: Govt job, Never worked, Private or Self-employed
8) Residence type: Rural or Urban
9) Avg glucose level: Average glucose level in blood
10) BMI: Body Mass Index
11) Smoking Status: formerly smoked, never smoked, smokes or Unknown
12) Stroke: 1 = Yes or 0 = No
Note: "Unknown" in smoking_status means that the information is unavailable for the patient

Data Quality Report

The quality of the report is considered by the following attributes they are accuracy, clarity, consistency, transparency, unambiguity, and language. All the following needs are satisfying the methods which we are processing but there are few problems we faced during the data understanding they are mentioned below.

1. The values in the BMI columns found missing since there are no data of height and weight to measure BMI.
2. we found that there 201 null values in this column and we excluded all these rows.



Name	Type	Missing	Statistics		
age	Integer	0	0	82	43.227
hypertension	Integer	0	0	1	0.097
heart_disease	Integer	0	0	1	0.054
ever_married	Nominal	0	No (1757)	Yes (3353)	Yes (3353), No (1757)
work_type	Nominal	0	Never_worked (22)	Private (2925)	Private (2925), Self-employed (819), ...[3 more]
Residence_type	Nominal	0	Rural (2514)	Urban (2596)	Urban (2596), Rural (2514)
avg_glucose_level	Real	0	55.120	271.740	106.148
bmi	Nominal	201	97.6 (1)	28.7 (41)	28.7 (41), 28.4 (38), ...[416 more]
smoking_status	Nominal	0	smokes (789)	never smoked (1892)	never smoked (1892), Unknown (1544), ...[2 more]

Figure 3: Initial Findings

Data Preparation – Stage 3

The data which we are using is continuous and discrete type. All the above operations are performed using the RapidMiner. The below picture explains the process of data preparation to use all the classification and decide which model is the best.

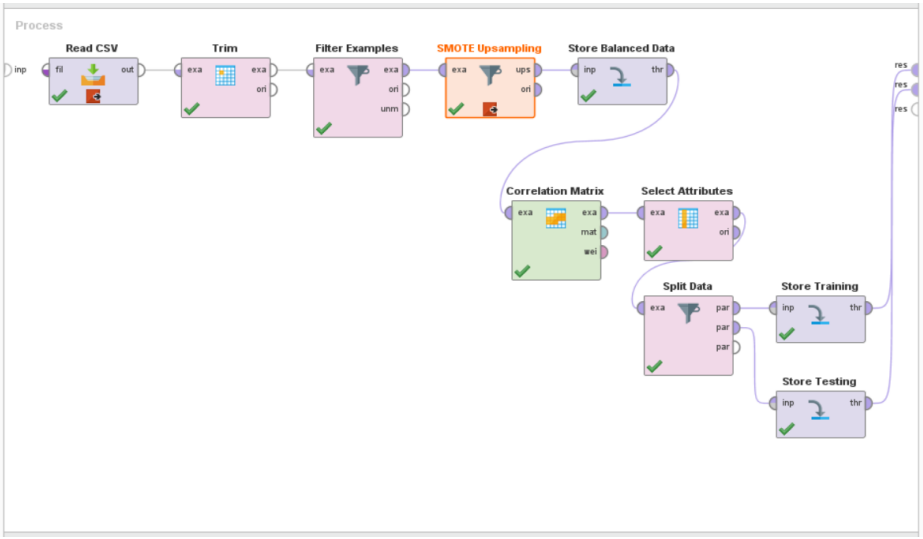


Figure 4: Data preparation loops

Selection

In this we have decide the final values which have to be performed for analyzing the process and to make decision for satisfying the goals. Therefore, the dataset will be uploaded using the command 'Read CSV'.

Trimming

The Trim administrator makes new properties from the selected nominal attributes by eliminating driving and following spaces from the nominal values.

Filtering

It is an important factor for measurement of noise or errors in the data and limits their usefulness in practice. In this, we picked BMI data and utilized that subset for filtering.

Sampling

We found the unbalanced sampling in the statistical way of binomial model, so we decided to over sample the technique to make the model stable and balance to get the quality of model by performing SMOTE Up [5]. The below pictures describes before and after the sampling process.

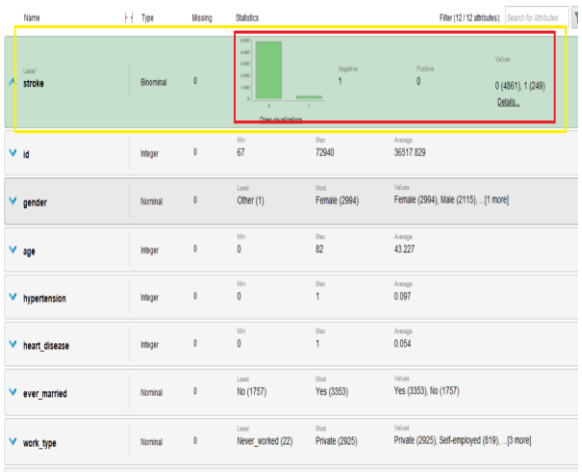


Figure 5: Before Sampling

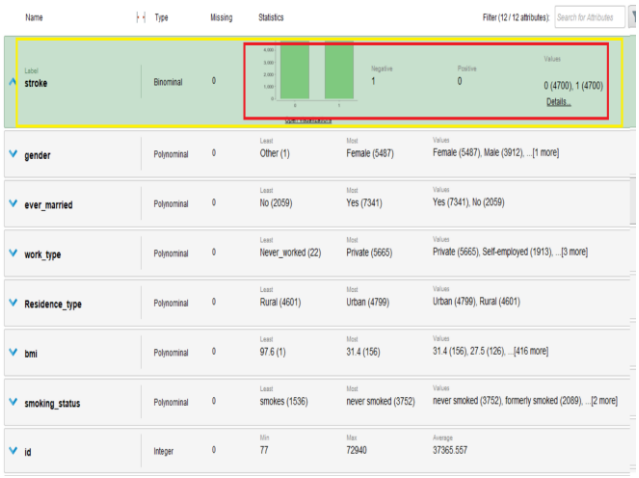


Figure 6: After Sampling

Splitting

These datasets are stored locally in RapidMiner for proposing the training set and testing set and moreover they are divided separately to build the model, and to evaluate the model [3].

Data Modelling – Stage 4

Select Technique

The tool we selected to perform the model was Rapid Miner, and we used the technique for the project is Auto Model Function. Since the data is small auto model function make quick prediction on the data set. Thus, it provides a proper vision for the business purpose.

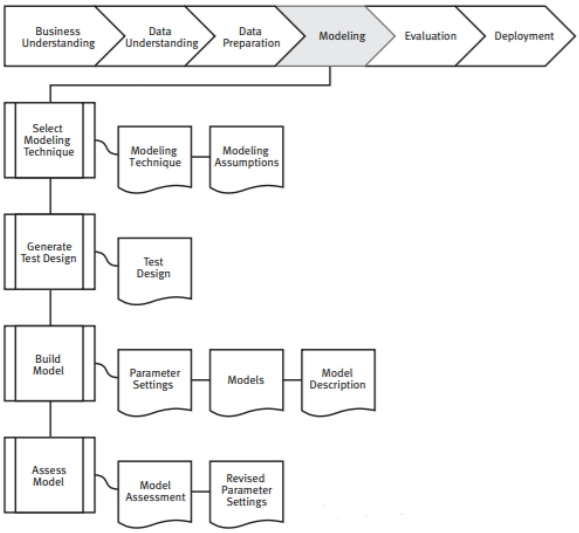


Figure 7: Modelling [6]

Generate Test

Every model requires a testing to verify whether the performance of the model is giving accuracy on the data which we going to predict, so we retrieved the stroke data to apply for checking the output.

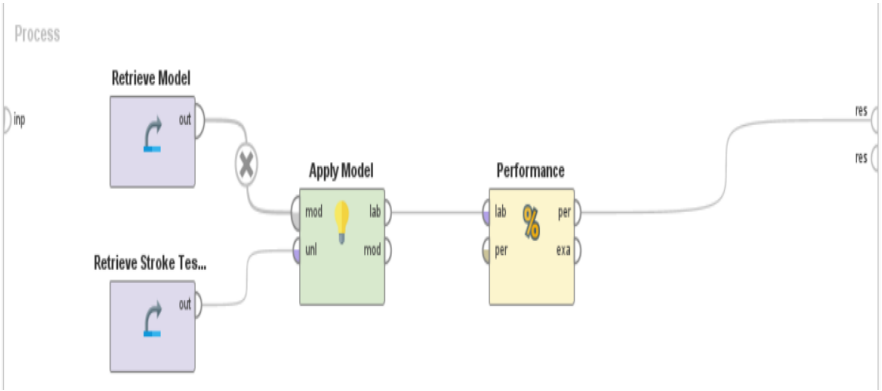


Figure 8: Testing process on selected data

Build Technique

In Rapid Miner all the techniques will be available to classify the dataset, so we implemented using all the algorithms to check which model performs well. And there is option to select which model we need, and we can exclude another model which we don't need.

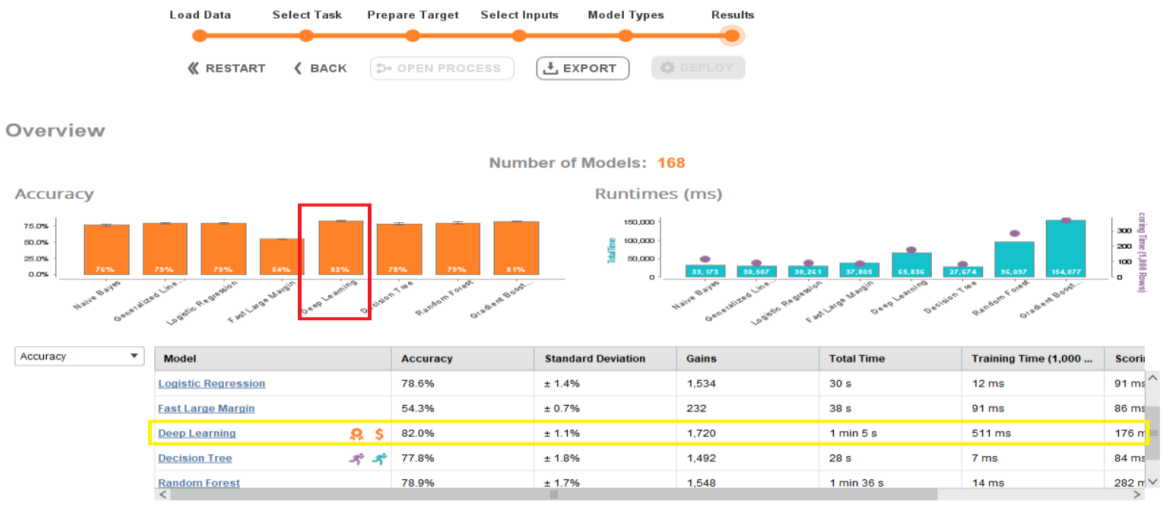


Figure 9: Auto Model Overview

The above picture explains that all the models have the average rate of accuracy at 70% and we finally decided that Deep Learning method for predicting the stroke have the most accuracy level compared with other techniques.

Auto Model Performances

The below picture represents the criterion table of each algorithm. In this table it measures the targeted values and output the summary for predicted values of the stroke dataset. This table can be used to analyze which algorithm can be practiced for finding the best result.

Criterion	Naïve Bayes	Generalized Linear Model	Logistic Regression	Fast Large Margin	Deep Learning	Decision Tree	Random Forest	Gradient Boosted Trees
AUC	82.80%	78.70%	78.60%	54.30%	82.00%	77.80%	78.90%	81.30%
Accuracy	75.60%	21.30%	21.40%	45.70%	18.00%	22.20%	21.10%	18.70%
Classification Error	24.40%	86.10%	86.10%	54.30%	89.50%	80.10%	85.10%	90.40%
F Measure	75.50%	78.20%	78.10%	54.40%	80.50%	72.30%	73.70%	85.20%
Precision	75.80%	79.70%	79.60%	53.80%	84.60%	90.30%	89.80%	75.70%
Recall	75.20%	78.90%	78.80%	54.00%	82.50%	80.30%	81.00%	80.20%
Sensitivity	75.20%	79.70%	79.60%	53.80%	84.60%	90.30%	89.80%	75.70%
Specificity	76.00%	77.70%	77.60%	54.90%	79.50%	65.30%	67.90%	86.80%

Figure 10: Performance Table

Confusion Matrix

This table is for deep learning model and it is predicted based on storke data to identify the confusion matrix. And there we found the type I error and type II error for the dataset.

Confusion Matrix

	true 0	true 1	class precision
pred. 0	1067	206	83.82%
pred. 1	276	1136	80.45%
class recall	79.45%	84.65%	

Figure 11: Confusion Matrix Table

Evaluation - Stage 5

In this project we evaluated that deep learning algorithms findings and accuracy are comparatively have the higher level of prediction, and it is found to be good way of resulting in business outputs. The business requirements and the key factors of the model satisfied in each level. The total runtime to predict the outcome was rapid and accurate to decide the classification. In every project there will be some deficiency in finding the clear view in our case we had issue with few missing data.[6] Therefore, we excluded all the missing values and provided a good data for the modelling process. Now we are efficient to deploy the deep learning method to successfully run this project

Future Improvement

Missing data can be retrieved and filtered properly to improve the accuracy of prediction. This helps in improving the betterment of human beings.

Deployment – Stage 6

The stroke prediction model can be developed as a web application or mobile application and that can be utilized in healthcare system or clinics. This facility is to discover the patient stroke possibilities and it can be likewise utilized by an individual to discover their stroke possibilities dependent on current wellbeing boundaries.

References

[1] L. C.-O. C. F. J. `. e. H. ´. a.-O. M. K. L. M. J. R. ´. ´.-Q. a. P. F. Fernando Mart´ınez-Plumed, "CRISP-DM Twenty Years Later:From Data Mining Processesto Data Science Trajectories," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING ,* p. 14, 2019.

[2] P. N. A.Sudha, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods," *International Journal of Computer Applications,* p. 6, 2012.

[3] R. A. Ohoud Almadani, "Prediction of Stroke using Data Mining Classification Techniques," *International Journal of Advanced Computer Science and Applications,* p. 5, 2018.

[4] R. A. M. T. F. S. A. M. F. J. K. N. T. Leila Amini, "Prediction and Control of Stroke by Data Mining," *Int J Prev Med 2013;Suppl 2: S245-9. ,* 2013.

[5] S. A. S. A. A. B. T. A. Hosam Alhakami, "A Hybrid Efficient Data Analytics Framework for Stroke Prediction," *International Journal of Computer Science and Network Security,* p. 11, 2020.

[6] J. C. (. R. K. (. T. K. (. T. R. (. C. S. (. a. R. W. Pete Chapman (NCR), "CRISP-DM 1.0 - Step-by-step data mining guide," *The CRISP-DM consortium,* p. 75, 2000.