Project Report

Data 611 – Predictive Analytics

Project 2 – G1

Stroke Prediction

1st March, 2022

Deshant Sachdeva (30150728)
Ketan Bassi (30146366)

Introduction

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Almost 30% of adults younger than 45 don't know the five most common symptoms of a stroke, according to new research. At the same time, stroke is on the rise in that age group.
Each year, 10% to 15% of the nearly 795,000 people in the United States who have a stroke are between the ages of 18 and 45. And despite a decline in the general population, stroke rates – and hospitalizations for it – have increased by more than 40% among younger adults in the past several decades.

Importance and the goal

Through this dataset we will be analysing what are the key factors responsible to cause stroke in an individual. If we are able to make predictions and analyse the factors that cause stroke it will help doctors in analysing and diagnosing patients more effectively.

Formalizing problem into data mining task

This information will be used to create a Logical regression model that can be used by doctors to make patients aware on what precautions they can take to minimize their chances of stroke.

Dataset and its components

Dataset is taken from Kaggle and contains a total of 5110 rows and 12 columns. This dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

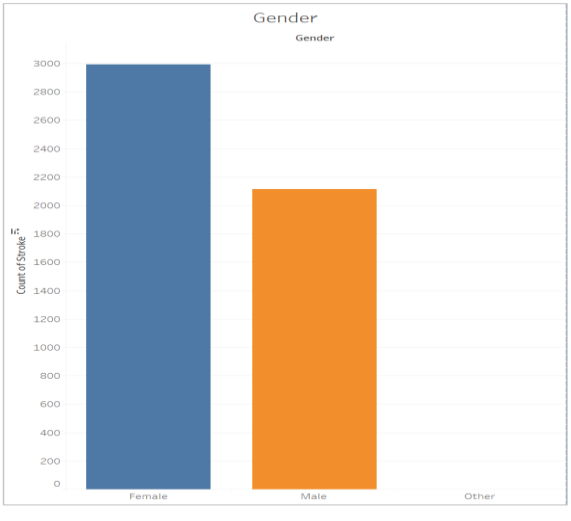
Attribute Information

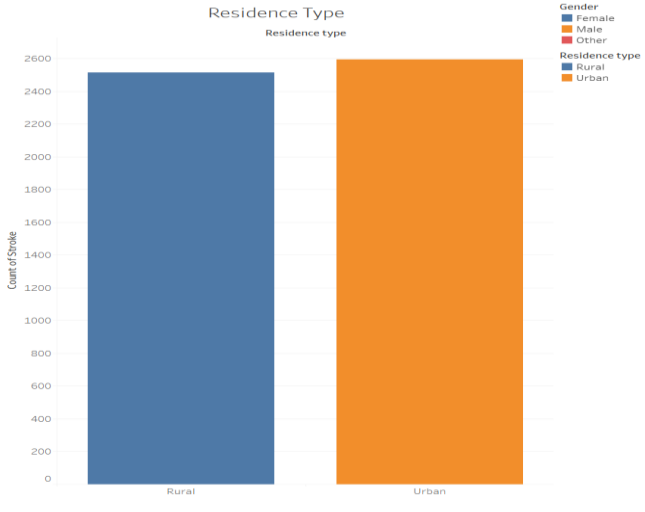
- **id:** unique identifier
- **gender:** "Male", "Female" or "Other"
- **age:** age of the patient
- **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever_married:** "No" or "Yes"
- **work_type:** "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- **Residence_type: "**Rural" or "Urban"
- **avg_glucose_level:** average glucose level in blood
- **bmi:** body mass index
- **smoking_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"*
- **stroke:** 1 if the patient had a stroke or 0 if not

id	gender	age	hypertensi	heart_dise	ever_marri	work_type	Residence_	avg_glucos	bmi	smoking_s	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly sr	1
51676	Female	61	0	0	Yes	Self-emplo	Rural	202.21	N/A	never smol	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smol	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-emplo	Rural	174.12	24	never smol	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly sr	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smol	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smol	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smol	1

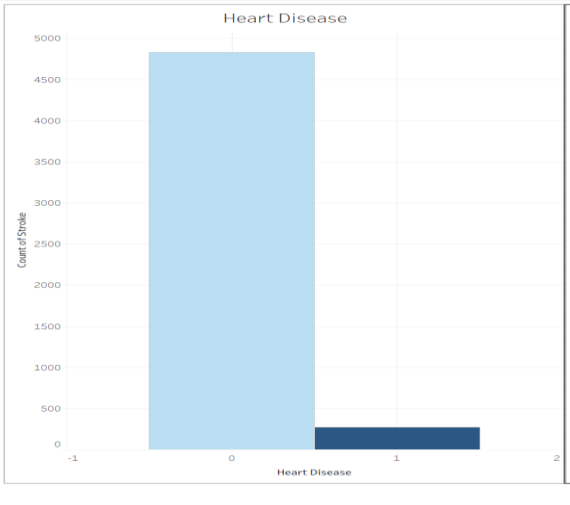
Exploratory data analysis

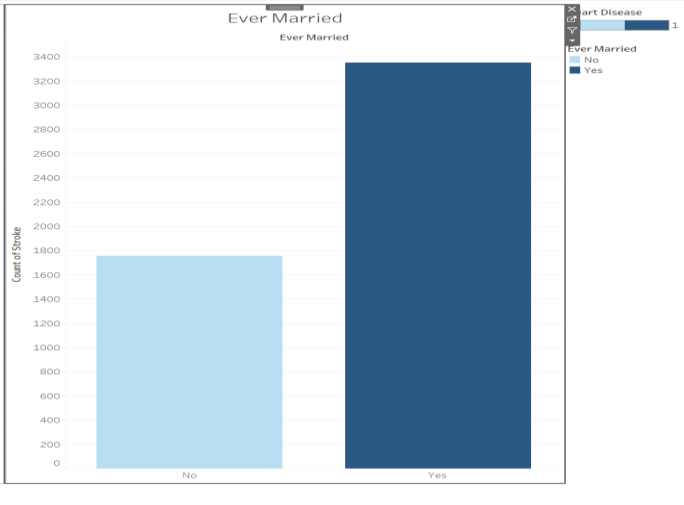
- From the below 2 plots we can see that the number of females in the data set is higher compared to the number of males in the data set while the number of people that live in the Rural are compared to the Urban area is almost similar.



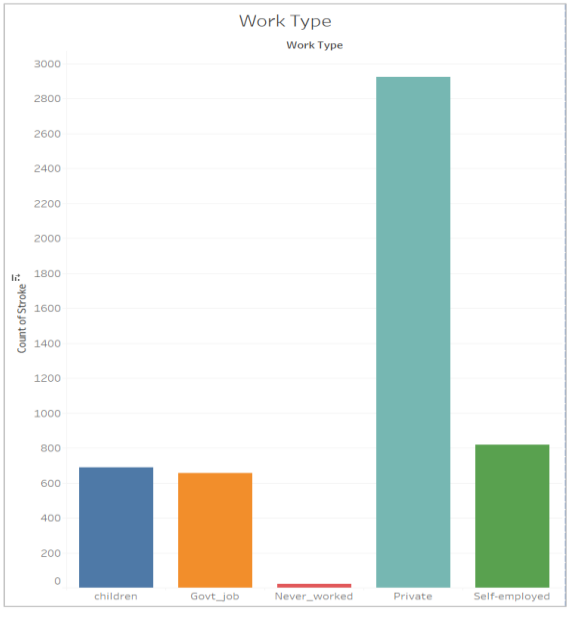


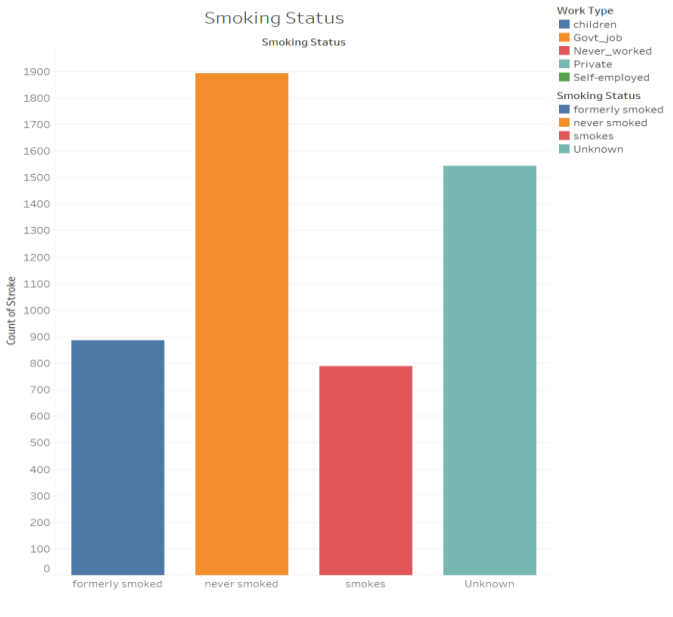
- In the following 2 plots, on the left we can see that most of the people in the data do not have a heart disease. And the plot on the left showcases the split between the number of people who were ever married or not



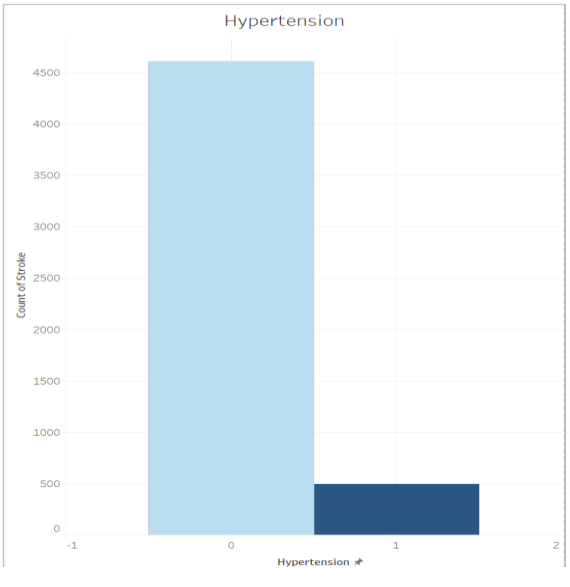


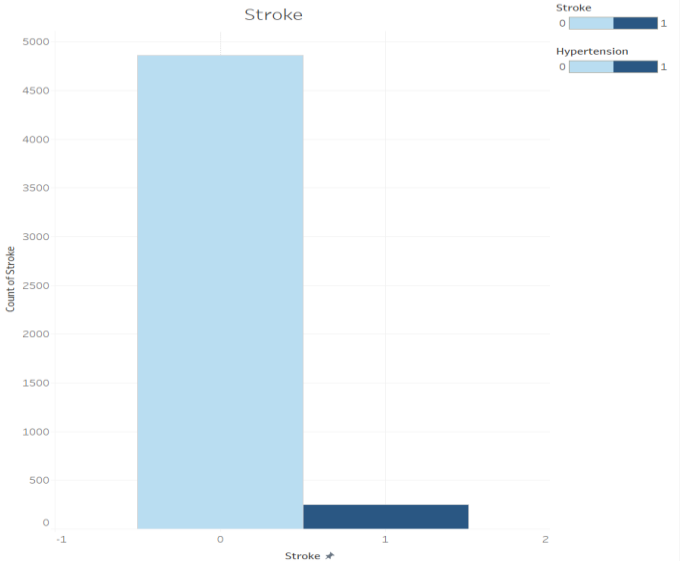
- In the next 2 plots we can see the number of peoples working in different work types like govt. job, private, self employed etc. From the plot we observe that majority of the people work in the private sector. And the smoking status plot on the right shows whether the person formerly smoked, they currently smoke, or they never smoked.





- Lastly, we found out that most people in the data set have neither of hypertension (left plot) or stroke (right plot)

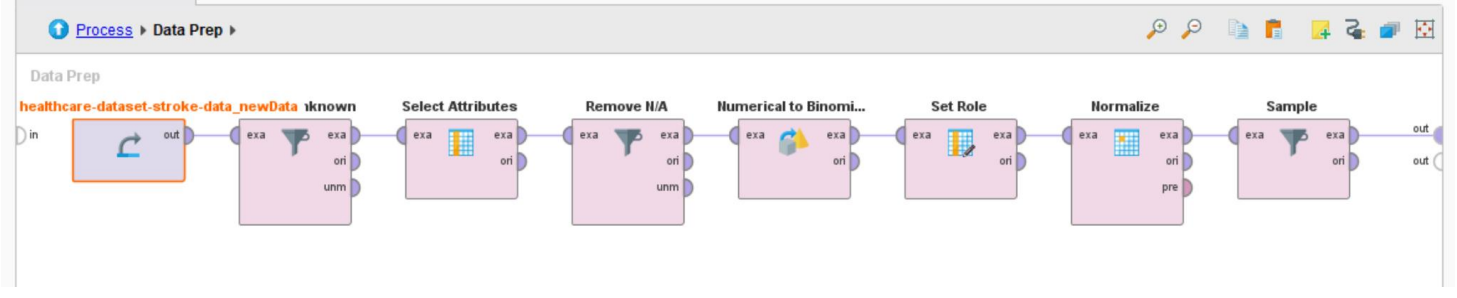




Data Wrangling

Data wrangling is an important part before we perform any data analysis over the data. We did the following data wrangling and cleaning onto our dataset so that we can obtain the desired and correct model output in the end.

- Removed non-essential column like **id**
- Removed rows having N/A values under BMI column
- Removed rows containing Unknown value under smoking status column
- Converted datatype of the column stroke to appropriate type in order to perform logistic regression
- Performed sampling onto the dataset to avoid underfitting or overfitting of our model.



Data Analysis using Logistic Regression and K-NN model

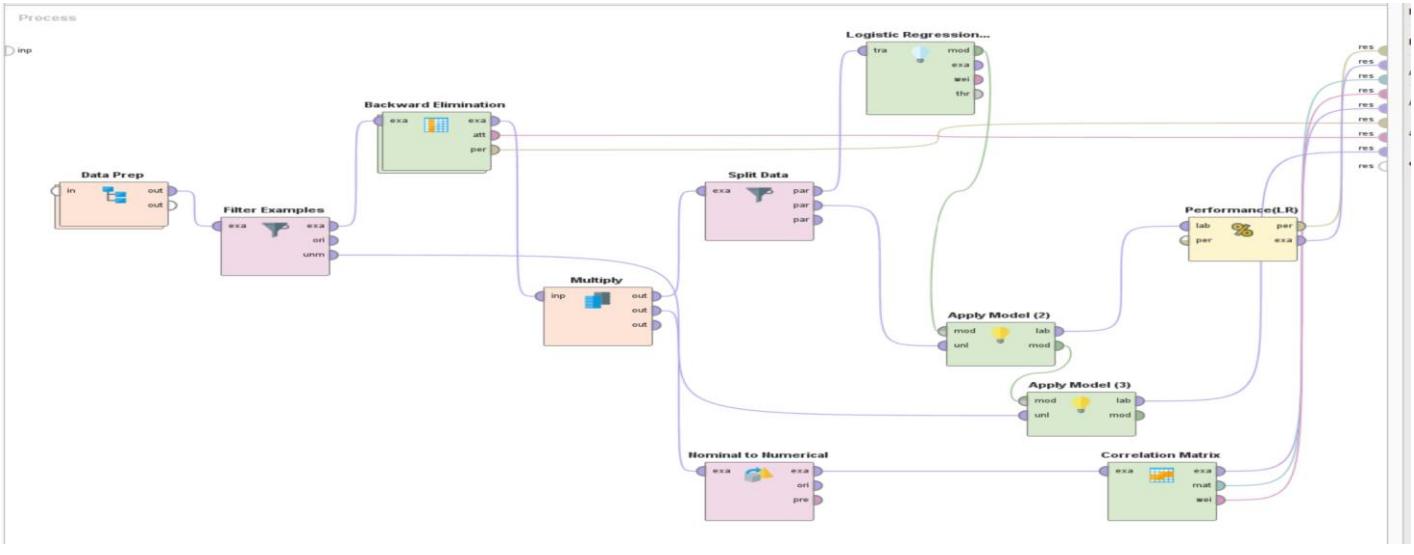
Logistic Regression

Before performing the logistic regression, we performed backward elimination in order to remove non-significant variables from the data on which we will be performing logistic regression. After performing backward elimination, the variable BMI was removed from the end result.

attribute	weight
age	1
hypertension	1
heart_disease	1
avg_glucose_level	1
gender	1
ever_married	1
work_type	1
Residence_type	1
bmi	0
smoking_status	1

After performing the backward elimination, we performed logistic regression on to the final dataset which was split into 60% training data and 40% testing data.

The end model looks like below,



Below results were received after executing the model,

1. At 5% significant level variables included are avg_glucose_level, age, and heart_disease.
2. At 7% significant level in addition to above we have work type as govt job and self-employed, hypertension as significant factors that are responsible for stroke.

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
work_type.Self-employed	-0.469	-0.469	0.255	-1.838	0.066
work_type.Govt_job	-0.688	-0.688	0.372	-1.847	0.065
work_type.children	-6.607	-6.607	102.501	-0.064	0.949
work_type.Never_worked	-6.997	-6.997	218.595	-0.032	0.974
smoking_status.never smoked	0.012	0.012	0.247	0.050	0.960
smoking_status.smokes	0.262	0.262	0.298	0.877	0.381
smoking_status.Unknown	0	0	?	?	?
gender.Female	0.005	0.005	0.222	0.021	0.983
gender.Other	-7.804	-7.804	663.889	-0.012	0.991
Residence_type.Rural	0.031	0.031	0.212	0.148	0.883
ever_married.No	-0.136	-0.136	0.364	-0.373	0.709
age	1.293	1.286	0.164	7.868	0.000
hypertension	0.143	0.143	0.078	1.836	0.066
heart_disease	0.141	0.142	0.067	2.107	0.035
avg_glucose_level	0.236	0.231	0.084	2.804	0.005
Intercept	-3.619	-3.628	0.300	-12.065	0

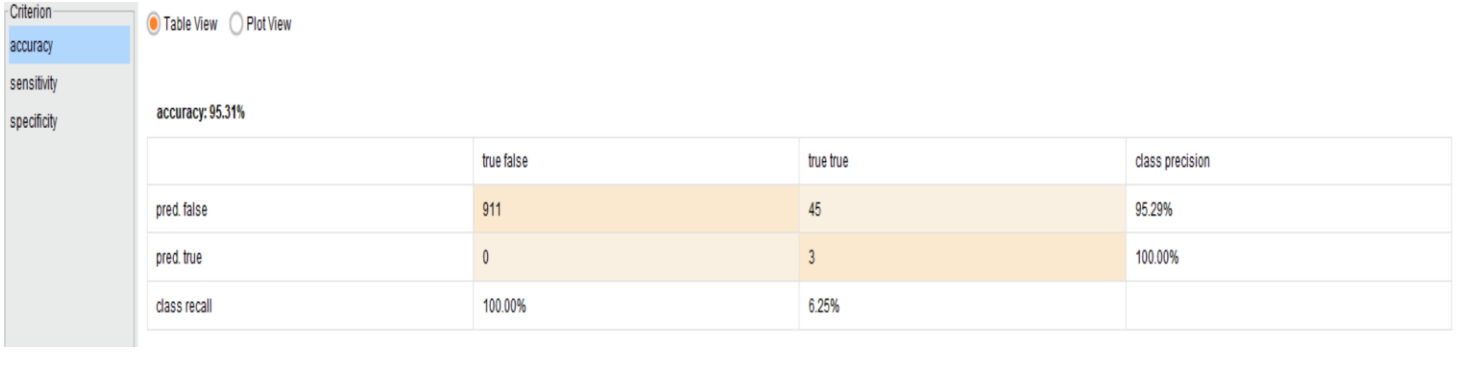
Correlation Matrix

Factors like Age, smoking-formerly or smokes, job type being self employed or govt. job are more correlated to stroke.

Attribut...	gender ...	gender ...	ever_m...	work_ty...	work_ty...	work_ty...	work_ty...	Residen...	smokin...	smokin...	smokin...	age	hyperte...	heart_d...	avg_glu...
gender =...	1	-0.014	-0.018	-0.014	0.009	0.032	-0.004	0.013	0.071	0.035	?	0.045	0.038	0.102	0.070
gender =...	-0.014	1	0.030	-0.008	-0.007	-0.002	-0.001	0.017	0.030	-0.009	?	-0.021	-0.006	-0.004	0.013
ever_ma...	-0.018	0.030	1	-0.132	-0.065	0.252	0.114	0.010	-0.105	-0.018	?	-0.523	-0.117	-0.077	-0.119
work_typ...	-0.014	-0.008	-0.132	1	-0.199	-0.067	-0.030	-0.007	0.076	-0.039	?	0.296	0.089	0.045	0.063
work_typ...	0.009	-0.007	-0.065	-0.199	1	-0.060	-0.027	-0.009	0.001	0.005	?	0.052	0.007	-0.007	0.007
work_typ...	0.032	-0.002	0.252	-0.067	-0.060	1	-0.009	-0.006	-0.022	-0.064	?	-0.274	-0.052	-0.036	-0.020
work_typ...	-0.004	-0.001	0.114	-0.030	-0.027	-0.009	1	-0.035	-0.036	-0.034	?	-0.109	-0.024	-0.016	-0.014
Residen...	0.013	0.017	0.010	-0.007	-0.009	-0.006	-0.035	1	-0.005	-0.036	?	-0.016	0.003	-0.010	0.012
smoking...	0.071	0.030	-0.105	0.076	0.001	-0.022	-0.036	-0.005	1	-0.298	?	0.189	0.022	0.056	0.051
smoking...	0.035	-0.009	-0.018	-0.039	0.005	-0.064	-0.034	-0.036	-0.298	1	?	-0.046	-0.013	0.032	-0.020
smoking...	?	?	?	?	?	?	?	?	?	?	1	?	?	?	?
age	0.045	-0.021	-0.523	0.296	0.052	-0.274	-0.109	-0.016	0.189	-0.046	?	1	0.267	0.260	0.234
hyperten...	0.038	-0.006	-0.117	0.089	0.007	-0.052	-0.024	0.003	0.022	-0.013	?	0.267	1	0.112	0.169
heart_di...	0.102	-0.004	-0.077	0.045	-0.007	-0.036	-0.016	-0.010	0.056	0.032	?	0.260	0.112	1	0.143
avg_gluc...	0.070	0.013	-0.119	0.063	0.007	-0.020	-0.014	0.012	0.051	-0.020	?	0.234	0.169	0.143	1

Correlation Parameters

- Accuracy: 95.31%
- Sensitivity: 6.25%
- Specificity: 100%



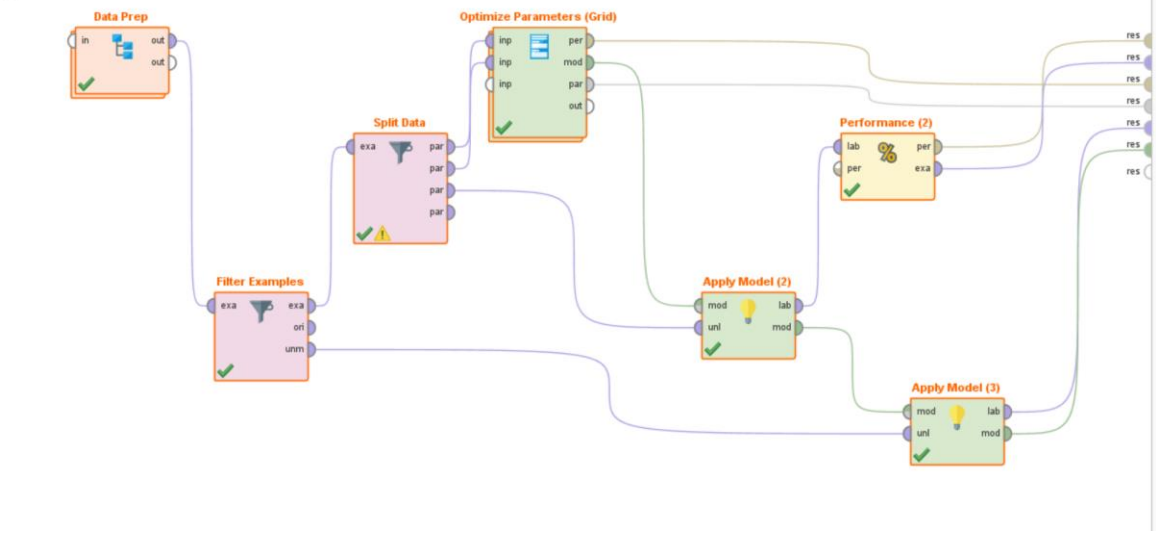
K-NN model

For K-NN model we partition the data into training (60%), validation (30%) and test (10%) sets using “Split Data” operator. The model will be trained using the training data. We choose the best value for k using performance on the validation set and evaluate the performance of the best model picked on the test set. Use Shuffled Sampling and the local random seed value of 2021.

We used the Optimize Parameters (Grid) operator, to try different values for k. We will pick the best one based on performance on the validation set. The first partition is passed for training while the second one is passed as validation set to evaluate performance. We will use the last partition later to evaluate the performance of the best model picked on a completely unseen data.

We used Performance (Classification) in this case because we have a classification task. Last time we used Performance (Regression), which is used for continuous response variable Y. We can use Performance (Regression) for k-NN if we are doing prediction instead of classification.

The model port of the Optimize Parameters operator provides the best model chosen (i.e., the best value for k). We can now apply this model to the third partition (i.e., the test set) to see the performance on unseen data. Again, we will use performance classification and note that you can pick any performance measure that you are interested in.



ParameterSet

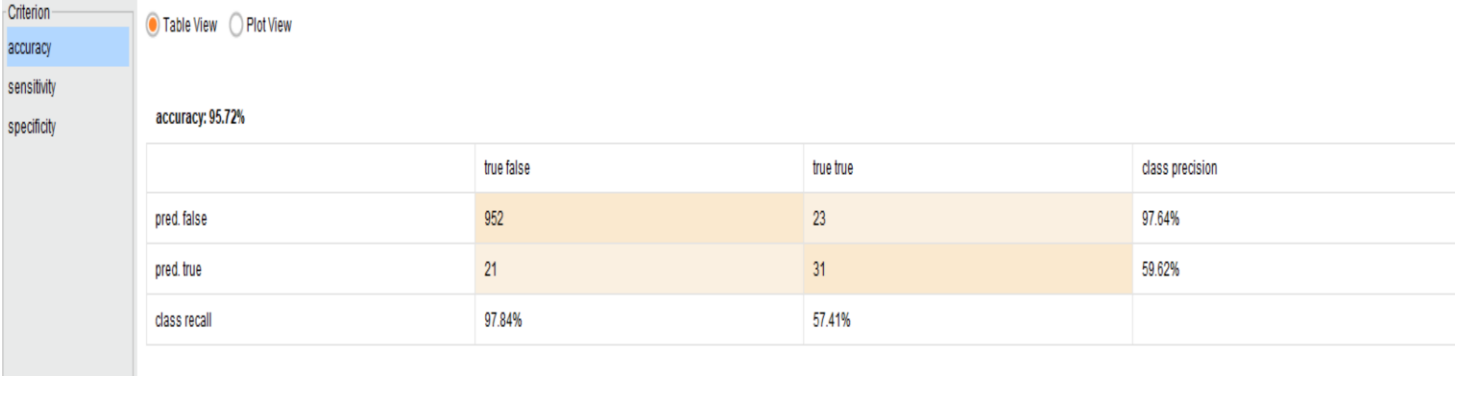
```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 95.72%
ConfusionMatrix:
True:   false   true
false:  952     23
true:   21      31
-----sensitivity: 57.41% (positive class: true)
ConfusionMatrix:
True:   false   true
false:  952     23
true:   21      31
-----specificity: 97.84% (positive class: true)
ConfusionMatrix:
True:   false   true
false:  952     23
true:   21      31
]
k-NN.k  = 1
```

From the results the best value of k turns out to be 1.

Correlation Parameters

- Accuracy: 95.72%
- Sensitivity: 57.41%
- Specificity: 97.84%



Comparison K-NN and Logistic Model

Correlation Parameters

	Accuracy	Sensitivity	Specificity
Logistic Regression	95.31%	6.25%	100%
K-NN	95.72%	57.41%	97.84%

As we can see that the sensitivity of K-NN model is much higher and should be a recommended model for our further analysis.

Prediction Analysis:

Data tested to be predicted is as below,

id	gender	age	hypertensi	heart_dise	ever_marri	work_type	Residence_	avg_glucos	bmi	smoking_s	stroke
1	Male	70	1	1	Yes	Govt_job	Urban	210	35	smokes	

Result from Logistic Regression

Row No.	stroke	prediction(s...	confidence(f...	confidence(t...	age	hypertension	heart_disea...	avg_glucos...	gender	ever_married	work_type	Residence_t...	bmi	smoking_st...
1	?	false	0.720	0.280	1.132	2.716	3.943	2.130	Male	Yes	Govt_job	Urban	35	smokes

Result from K-NN model

Row No.	stroke	prediction(s...	confidence(f...	confidence(t...	age	hypertension	heart_disea...	avg_glucos...	gender	ever_married	work_type	Residence_t...	bmi	smoking_st...
1	?	true	0	1	1.132	2.716	3.943	2.130	Male	Yes	Govt_job	Urban	35	smokes

Insights and Findings

The sensitivity obtained from logistic regression is at 6.25% which is very low. While the accuracy of model is very high with logistic regression at 95.31% but it should not be the criteria to conclude that we should recommend the model or not. Sensitivity and specificity are more of the high set of criteria's when it comes to predicting stroke as we need to maximize the number of True positives is this case.

If we consider the K-NN model the specificity is good at 57.41%. This model also gives us a slight better accuracy than logistic regression at 95.72%. Considering these parameters and the results we can tell from the data given that a person is likely to cause stroke which is correctly predicted by K-NN model and not by Logistic regression. Hence, we conclude that K-NN model should be a recommended model for the further analysis of stroke data.

Challenges and limitations

- Performed data cleaning in order to receive better results.
- There is a strong class imbalance at the level of the target. The percentage of unaffected patients exceeds 90%. This problem can be addressed with class balancing using a few resampling techniques (under sampling, oversampling, or both).
- We have performed sampling in both of our models that led us to obtaining better results.

Recommendations and Findings

- Models tells us that the variables age and **avg_glucose_level** influence whether a patient has a stroke or not. Hence these factors should be given more attention while diagnosing stroke.
- The **bmi** variable has no dependency on the stroke variable.
- The age variable is strongly correlated with certain categorical variables. In particular the **work_type**, **ever_married** and **smoking_status** variables.
- We will retain the hypertension and **heart_disease** variables which are clinically very interesting for our analyses.

References

- Fedesoriano, (2021-01-26). Stroke Prediction Dataset. kaggle, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
- CDC (May 25, 2021), Stroke Facts, https://www.cdc.gov/stroke/facts.htm