**PROJECT DOCUMENTATION**

**Course: Business Analytics With R**

**Topic: Prediction of heart stroke occurrence and the factors influencing it**

**1.EXCECUTIVE SUMMARY:**

**a. Business objective:**

For hospital management, the implementation of heart stroke prediction models can contribute to various business objectives, enhancing patient care, resource allocation, and overall hospital efficiency. They can determine which patients are at risk for a heart attack so that prompt medical attention can be provided. Contribution to community health initiatives and preventive healthcare programs can also be done with the help of the models which we create because we will get to know the potential factors responsible for a heart stroke.

**b. Data:**

The data on which we worked is taken from the Kaggle website in the csv file format. The data is in structured form but contain some NA values. Here is the link to the dataset:

[**https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data**](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data)

**c. Findings:**

The key findings from heart stroke prediction models that we have created are that they can provide valuable insights into the factors influencing the likelihood of a stroke and contribute to better understanding, prevention, and management of cardiovascular health. From our models, we concluded that age, body-mass index and smoking are the key factors in the occurrence of heart stroke.

**d. Recommendations:**

A hospital's management team may make strategic decisions to improve patient care, resource allocation, and overall healthcare efficiency by following recommendations based on heart stroke prediction models. To identify patients at risk of heart stroke, the hospital management team can use the predictive models to develop and implement early warning systems. This makes prompt medical care and proactive interventions possible. To make sure healthcare personnel are prepared to understand and act upon the model's predictions, they can fund training courses for them.

**e. Benefits:**

The project encourages the hospital to adopt a culture of data-driven decision-making. The predictive model's insights can be used by hospital management to guide strategic choices, resource allocation, and quality-improvement programmes. Implementing a heart stroke prediction project demonstrates the hospital's commitment to technological innovation in healthcare. This can enhance the hospital's reputation and position it as a leader in adopting advanced analytics for patient care.

**2.BACKGROUND/CONTEXT**

**a. Domain:**

The domain of a heart stroke prevention project typically falls within the broader fields of healthcare, medical research, and public health. Heart stroke prevention projects have a public health dimension, as they aim to reduce the overall incidence of strokes in a population. Public health professionals work on initiatives such as public awareness campaigns, policy advocacy, and community-based programs to promote healthy lifestyles and reduce stroke risk factors. This project particularly involves medical research to better understand the underlying causes of strokes and identify novel risk factors. Researchers may conduct clinical trials, epidemiological studies, and genetic research to advance our knowledge of stroke prevention.

**b. Brief description of scenario:**

The goal of this scenario is to identify individuals who are at higher risk of experiencing a heart stroke so that preventive measures and interventions can be taken to reduce the risk and potentially save lives. It combines healthcare data, few prediction models, and predictive analytics to assist healthcare professionals in making informed decisions and individuals in managing their health proactively. Identifying patients at risk of heart stroke can enable early intervention and timely medical treatment. This will mitigate the hospital's risk and liability associated with stroke-related incidents.

**c. Decision of interest:**

The main goal of the project is to assess the risk of stroke in individuals. This involves developing predictive models that consider various factors such as age, gender, medical history, lifestyle, and more. Relevant features that are strongly correlated with stroke risk are identified, which may include age, gender, glucose levels, heart disease, body mass index (BMI), and more. Proper preventive measures can be taken in the prevention of heart stroke if we are able to create a nearly accurate model.

d. **Decision maker:**

The primary decision maker for this project would be hospital’s management team and the board of directors. The team would have a look at the models which are created and observe the trends and relations of heart stroke due to different factors. They would provide the necessary inputs to the board of directors from the predictions who in turn plan proper preventive measures that can be taken to minimize the emergency admissions to the hospital. They could help the society by reducing the amount of heart stroke rates as time progresses.

**3.BUSINESS UNDERSTANDING:**

**a. Business objective:**

The objective of the predictive model we develop is to assess the risk of heart stroke in an individual based on their medical and lifestyle data. It is also to promote a data-driven approach to healthcare decision-making, helping providers prioritize patients for preventive interventions. In addition to these, the main goal is to engage patients in their own healthcare by providing them with personalized risk assessments and recommendations, fostering better patient compliance and self-management. This would help in gaining the trust and confidence of the patients. Implementing a heart stroke prediction project demonstrates the hospital's commitment to technological innovation in healthcare. This can enhance the hospital's reputation and position it as a leader in adopting advanced analytics for patient care.

**b. Situation assessment:**

A situation assessment for a heart stroke prediction project involves analyzing the current

situation, understanding the context, and identifying key factors that will influence the model’s success. By conducting a thorough situation assessment, we can gain a comprehensive understanding of the landscape in which the heart stroke prediction project will operate, helping them make informed decisions and develop a successful implementation plan. When using logistic regression or decision trees for predicting the occurrence of heart strokes through data mining, several key goals should be established to guide the data analysis and model development process. These goals help ensure that the predictive model is effective, interpretable, and valuable for healthcare professionals and decision-makers. All the goals are described in the business objective above.

**c. Data mining goals:**

The project’s data mining goal includes identifying patterns for each factor in the dataset like age, BMI, gender, hypertension, heart disease etc. After identifying the critical elements impacting the decision, we will be developing models with high predictive accuracy. The primary goal is to correctly classify individuals as either at risk of a heart stroke or not, minimizing false positives and false negatives. We will use validation methods like cross-validation to make sure the model works effectively with unseen data. The objective is to develop a model that applies well to new cases.

**4.DATA UNDERSTANDING:**

**a. Data requirements:**

To train and test the dataset, the data must be comprehensive. Each patient's information, including identity information like gender, and age as well as health information like blood pressure, heart disease, average glucose level, and BMI, is required. A certain amount of personal information, such as smoking status, marital status, and employment type, should also be included in the data. We will select the model attributes that are contributing to the stroke and focus on those for our forecasts.

**b. Describe data:**

The data which we had taken is structured and is in the csv format.

The data consists of following columns :

1) id: It is a unique identifier for each patient

2) gender: "Male", "Female" or "Other"

3) age: It shows the age of the patient.

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension.

5) heart disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever married: Marital status of the patient: "No" or "Yes"

7) work type: This variable depicts the employment type of the patient. The column shows as children or Govt job or Never worked or Private or Self-employed

8) Residence type: It is the area of residence of the patient whether he lives in urban area or a rural area.

9) avg\_glucose\_level: It shows the average glucose level in blood.

10) bmi: It showcases the body mass index of the patient.

11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown2) stroke: 1 if the patient had a stroke or 0 if not.

\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

**c. Sources:**

The dataset is taken from the Kaggle website.

[**https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data**](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data)

**d. Quality:**

Data quality is very important for any forecast model. The dataset is of high quality. There are 12 columns and 5110 row values in it. The data is free from typographical errors and other mistakes. The data has no missing values. The gender column contains one “other” value which can be removed easily. The dataset consists of few outliers, and they should be removed to build an accurate model. Other than this, the data is accurate, consistent and timely.

**5. Data Preparation:**

**a. Data selection:**

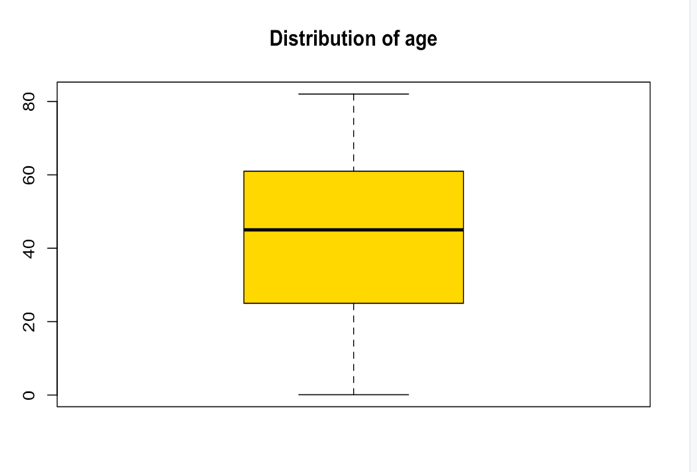
The id column is not necessary for classifying heart stroke occurrences because it does not play a role in heart stroke prediction. Thus, that column will be removed from the dataset. The entire dataset appears the same aside from this removal.

**b. Data cleaning:**

Missing values are checked with the help of is.na (). There are no missing values in the dataset. In the gender column, there is a single row item with the value “other” rather than Male or Female. So, we will drop that row item.

In the BMI column, there are many N/A values. That is why we are unable to find missing values using is.na () function. As there are many N/A values, we will replace them with the mean value of BMI. This had become a problem to me because NA values are in the form of N/A, and we simply can’t use na. omit (). So, I first converted N/A to NA using numeric function and then I replaced the NA values with the mean of the bmi column.

Now, we will check the numerical variables for any outliers in the data. In our dataset there are three numerical variables, and we check for the outlier’s using boxplot.

 A diagram of a level of glucose

Description automatically generated

A diagram of a mass index

Description automatically generated

By looking at the above boxplots, we can clearly interpret that average glucose level and BMI have outliers. So, we need to do treat these outliers. As Age column don’t contain any outliers, the column would remain the same. But for the other two columns we will replace outliers with NA values and then eliminate the row values in which NA has occurred. So, our dataset now has changed to 4390 row values from 5110 values.

**c. Prepare data:**

For the graphical analysis of the dataset, it is better to convert all the binary variables to “Yes” or “No”. So, I converted the columns hypertension, heart\_disease and stroke as “Yes” or “No”. This is done using the factor function. Also, as said previously id column is removed and made as NULL because id column is not required in predicting the heart stroke.

**6. Modeling-Building models:**

**a. Describe the data in detail:**

The dataset consists of the columns stroke, gender, age, hypertension, heart\_disease, ever married, work type, Residence type, avg\_glucose\_level, bmi and smoking\_status.

1.Stroke: It is our outcome variable and contain 0’s and 1’s.

2. gender: After data cleaning, it consists of values “Male” and “Female”.

3.age: It shows the age of the individual.

4. Hypertension: It tells whether a person has hypertension or not.

5. heart disease: “No” if the patient doesn't have any heart diseases, “Yes” if the patient has a heart disease

6.ever married : It consists of values “Yes” or “No”

7.work\_type: The work type shows the type of work each individual is doing. The type of work consists of children, government job, never worked, private and self-employed.

8.Residence\_type: It shows whether an individual stay in rural or urban area.

9.avg\_glucose\_level: It shows average glucose level in the blood

10.bmi: It shows the body mass index of the individual.

10.smoking\_status: It shows whether the person smokes, formerly smoked, never smoked or unknown.

Here are some graphs from the dataset:

A chart of a number of green bars

Description automatically generated A chart of a glucose level

Description automatically generated

A blue rectangular shapes with red edges

Description automatically generated A graph of a graph of a graph

Description automatically generated with medium confidence

A graph of a heart disease

Description automatically generatedA couple of purple squares

Description automatically generated

A graph of blue and red lines

Description automatically generated

A graph of different colored squares

Description automatically generated with medium confidenceA graph of smoking status

Description automatically generated

So graphical representations are for each column and the graphs clearly distinct each attribute with the heart stroke occurrence. In this way, we can get the interpretation of our data more clearly.

**b. What type of decision making model(s) is appropriate for the decision-making tasks?**

The decision making tasks for this project are to identify the factors that influence the heart stroke in an individual and the models which predict this accurately. This will help in taking preventive measures and warning the patients about risk off each factors. The following models can be used for the decision making tasks:

1. Logistic regression: Logistic regression is a classification model that can be used to predict the probability of an event occurring which in our case is a heart stroke.

2. Decision trees: A decision tree is a classification model that can be used to identify the rules that distinguish between the possibility of occurrence of the heart stroke . It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

**c. Provide rationale for choice of model(s):**

Logistic regression and decision trees are used in the project because they are the best ones to use for classification problems and they are also widely used for problems in which finding the influencing factors are required especially in our case like heart stroke classification. They are really easy to interpret specially speaking about the decision trees.

**d. Detail model development and output :**

The model development is generally started with dividing the dataset into two subsets which are training set and validation set. The model is trained using the data from training set .These created models are then evaluated using the data from validation set. The output of these logistic regression model and decision tree is the probability of our input factors belonging to class 1.i.e Belonging to the stroke occurence in the individual.

**7. Model evaluation:**

Model evaluation in general means evaluating the performance of the model on the validation dataset. The model is evaluated using evaluation metrics like:

1.Accuracy: It is the percentage of the predictions that the model predicts correctly.

2.Sensitivity: It is the metric that evaluates a model's ability to predict true positives of each available category.

3.Specificity: It is the metric that evaluates a model’s ability to predict true negatives of each available category.

To calculate these values accurately, the data should be divided in a way that all types of values are present in both training and validation set. This had become a problem in my dataset because all the 1’s in the values of the outcome variable are present in the first 25% of the data. If I divide the dataset in the format of first 75% training and next 25% validation set the performance measures would not be accurate. I tried it and I got sensitivity as 0 and specificity as 100. So, I did the data partition using createDataPartition function present in the caret package and I got the accurate values.

Also

For the decision trees, even graphs are drawn like ROC plot, lift chart and decile chart.

After looking at the output from both the models, the decision tree proves to be the better one because of high accuracy, and specificity.

Here are the values which I got for the models:

For best logistic regression model:

Accuracy: 89.425706

Sensitivity: 60.526316

Specificity: 90.462701

For the decision tree:

Accuracy: 90.34

Sensitivity: 41.463

Specificity: 92.235

Often sensitivity and specificity are inversely related. The balance between sensitivity and specificity depends on the specific goals and consequences of false positives and false negatives in the context of the application.

Sensitivity is less in our models because we got an imbalanced dataset. Our dataset consists of very less 1 values in the stroke column. It consists of only 250 one values out of nearly 5000 values.

**8. Discussion:**

**a. Based on the Model, what would your decision/recommendation be? Why?:**

Based on the results from the model, the hospital management team can now tell the important factors that are resulting the heart stroke In an individual. From the decision tree and logistic regression we can say that age, avg\_glucose\_level and bmi are the primary factors in the occurence of heart stroke . If these three factors are more in an individual, the chance of getting an heart stroke would be more. Even smoking\_status can be an a big factor in the case of people with more age.

**b. What are the limitations of the Model you have used?**

The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. The limitation of the decision tree is that uncertain values can lead to complex calculations and often even get uncertain outcomes. Also for these models, they are dependent on the data that they are trained on. If this trained data is incomplete or even inaccurate, then the model will not make accurate predictions. Also, the model may not generalize to the data which is completely different from the training data.

**c. What cognitive biases would you expect (most likely) to influence the decision-making process? How does decision support mitigate some/all of these?**

Human decision-making is often subject to various cognitive biases, which are systematic patterns of deviation from norm or rationality in judgment. Here are a few common cognitive biases that can influence the decision-making process:

**1.Confirmation Bias:** This bias involves giving preference to information that confirms one's pre-existing beliefs or values.

**2.Overconfidence Bias:** People tend to overestimate their own abilities, knowledge, or the accuracy of their beliefs. This bias can lead to overly optimistic assessments of a situation and suboptimal decision-making.

**3.Anchoring Bias:** This bias occurs when individuals rely too heavily on the first piece of information encountered (the "anchor") when making decisions. Subsequent information may not be given enough weight.

**4.Availability Bias:** Decision-makers give undue importance to information readily available to them, often from recent or memorable events. This bias can lead to overlooking less salient but equally relevant information.

**d. What enhancements would you aim for to enable better decision support for this task?**

The first enhancement would be definitely to introduce new features that are more directly related to cardiovascular health, such as a combined measure of cholesterol and blood pressure, or features derived from medical literature.

The next one would be to experiment with different machine learning algorithms beyond logistic regression, decision trees such as random forests, gradient boosting, or support vector machines, to identify the model that best fits the data.

The other one would be if available, consider incorporating time series data to capture trends and changes over time, as cardiovascular health may evolve gradually.

Finally, the last one is to educate healthcare professionals on the strengths and limitations of the model, promoting awareness of its role as an aid in decision-making rather than a replacement for clinical judgment.