STROKE PREDICTION +OPTIMISATION ALGORITHM USING MACHINE LEARNING

Akshat Khanna

Amity School of Engineering and

Technology

Amity University Noida Uttar Pradesh , India

[akshat.khanna@s.amity.edu](mailto:akshat.khanna@s.amity.edu)

***Abstract*—** Due to modern stressful conditions and deteriorating health conditions , including but not limited to hypertension, kidney and liver functioning and BMI fallacies , stroke has become one of the most prominent reasons of death, partial and complete paralysis, and lifelong disabilities. The imbalanced nature as well as the large volume of data available in this regard makes it almost impossible for early stroke prediction ad avoidance by doctors and experts alike. Given this predicament, machine learning can prove to be useful tool to make systematic predictions and aid as a decision-making tool.

Machine learning has recently found fame as an advanced diagnosis and prognosis technique. This paper aims to employ a hybrid approach in order to detect strokes using a combination of machine learning algorithms and concepts while minimising the Type 1 and Type 2 error and maximizing accuracy via hyper parameter tuning and various optimisation methods. Early detection and proper management is required to minimize the further damage in the affected area of the brain and other complication in the body parts.

In the current stressful situation, cerebral stroke has risen to be a major threat to public health as it results in severe morbidity, disability, and mortality. According to World Health Organization (WHO) in every year fifteen million people are suffering from stroke in worldwide and affected individuals are passing away every 4-5 minutes. A stroke will occur when the blood flow to various areas of the brain is disrupted or diminished, the cells in those regions do not get the nutrients and oxygen and start to die. There have been numerous efforts for early prediction as well controlling related metabolic and other key-factors to curb this condition, but they have not been substantially successful. Artificial intelligence has been making great breakthroughs in medicine and one probable action path can be to employ various machine learning algorithms to an imbalanced data set containing features such as occurrence of hypertension, body mass index level, heart disease, average glucose level, smoking status, previous stroke and age and predicting the probability of a stroke for a given individual.

***Keywords—ML algorithms, jupyter ,python, scripts.***

CHAPTER :1- INTRODUCTION

PROBLEM STATEMENT

An individual is said to be going through a stroke when the flow of blood to different regions of brain gets disrupted as well as diminished. In this case, cells are unable to get the nutrients and important substances such as oxygen which causes it to die. This is a medical emergency that needs intensive and quick care. Two of the most important factors that define between life and death are early detection of stroke as well as managing it properly. This helps to diminish the overall damage in the regions that are affected by the stroke or other dire consequences faced by other parts of body. Leading medical governing body WHO, has passed information stating about fifteen billion are suffering from this issue and out of this the amount of people actually passing away reaches about one in every four to five minutes.

Multiple medical studies have been conducted in this field and various predictors have been outlined that are statistically proven to be effective. One such study, namely, Framingham Study has marked out these parameters to be age, smoking of cigarettes or other substances, blood pressure both systolic and Di systolic, hypertension and various therapies to control it, fibration of the arteries, etc.

In related works done in this field there have been multiple approaches that have been taken in order to find out the relevant features that greatly affect the stroke prediction model and those works have used many techniques such as correlation constants and importance using various methods to achieve this.

Multiple studies have also been done in order to analyze the importance of lifestyle choices that have been made by patients and their medical records have been examined in order to find the probability of patients that can develop stroke in coming years. Apart from this, this paper focuses on the machine learning models that are now being employed in order to correctly predict that whether a patient can have the occurrence of stroke or not. In all of these related works there have been very few studies that have a holistic approach towards analyzing all of the conditions whether lifestyle or medical that are related to a particular patient and then perform the process of identifying the important features that can give us accurate stroke prediction. Hence, in this particular report we will be employing a novel method that provides a systematic examination and analysis of multiple patient records and their various medical and lifestyle features for the prediction of stroke.

In order to predict the most important features various techniques have been employed by us such as finding out the correlation between various features and using the medical data to inferred the importance of each and every feature and then brainstorming that whether this could directly affect the probability of a patient having stroke or not we have also outlined several machine learning algorithms that have been popularized in the recent times and I have given accurate results on well designed and balanced medical record.

MOTIVATION

In this report we have selected a topic that is very closely related to the current working conditions of the 21st. This is because distress in the workplaces schools universities as well as home has increased to a pitch fold where it is causing life-threatening diseases one of them which is stroke. This condition is something that is the One of the leading reasons people are losing their life the most characteristic feature of this disease is that in recent years it has escaped its usual para meters such as age or glucose levels or BMI and has now moved towards a much larger domain which has made its production much more important and difficult at the same time. Machine learning methods have now reached such a great level of efficiency that they are being used for various purposes across our lives one of them which is medical diagnosis. the tools and the methodologies that are now available to us in form of machine learning models and techniques are so wide that these models have outperformed experienced doctors and specialist when it comes to prediction of stroke in a particular patient when the medical as well as lifestyle characteristics are available to us.

So, to aid in this budding field of machine learning and medical diagnosis our paper aims to present some novel methodologies that can further improve the accuracy by employing where is optimization techniques that are available to us.

BACKGROUND INFORMATION

Background information in this particular field states that they have been multiple efforts in order to come up with a holistic plan in order to correctly diagnose the stock prediction models with greater accuracies then all the work that has been done before. Strokes caused when due to some reason oxygen supply is cut off in the brain which causes a blood clot to be formed which for the deteriorates the condition of various organs of the body as their oxygen supply is diminished and at the end the organs story begin to die beginning with the brain cells. This condition can be caused by a variety of factors that may or may not work together to increase the likelihood of a person having a stroke in coming years. Background information states that these features are sometimes easily manageable as well as diagnosable such as BMI glucose level and lifestyle choices but on the other hand these features can also be very ambiguous and very hard to manage such as creatinine level hypertension or arterial fibration.

GOALS AND OBJECTIVES

One of the major goals and objectives of this paper is coming up with a modernized and holistic solution using machine learning algorithms to correctly predict the probability of a patient having a stroke when various underlying conditions are present. in order to achieve this outcome, we will be processing our data and applying various stages of data analytics onto it so that our input data

can be deemed free of all irregularities inconsistencies false values missing values and features that may disturb our analysis. We also aim to compare the various accuracies that will be calculated

when we apply various machine learning algorithms to our refined data set. By comparing these

performance scores against each other we can then come up with a combination of machine learning algorithms as well as various optimization techniques that can be utilized in the field of medical diagnosis. This approach of combining various machine learning algorithms along with optimization techniques is gnu and presents us with a greater horizon of achieving higher accuracies and better results with the least number of false positives or false negatives false. So it can be concluded that this paper aims to build upon the already present work that has been done in this field that combines medical diagnosis along with machine learning algorithms by refining our data-processing Techniques such as feature generation and normalization and then applying the best combination of machine learning algorithms on top of this we will also be using optimization techniques that will help us move along our research outcomes.

SCOPE

The scope of our current work has far-reaching uses and utilities in Field of both machine learning and medical science especially diagnosis of complex medical issues such as stroke. As we have discussed earlier early prediction of stroke can be a major region reason which decide that whether or not a person is going to survive the chances of a brain hemorrhage or any kind of stroke. due to this reason our model focuses on the accuracy and various ways in which we can build it up so that the most accurate results can be obtained. if a model is applied to a larger data set such as a medical directory, we can confidently state that the results will be on a higher accuracy side and by the usage of those results we can start early mitigation techniques and preventive measures so that chances of stroke could be reduced and a person can lead on a healthy and productive lifestyle. Does the scope of this project is very wide and relevant to the present working and living condition that a 21st-century person faces.

ADVANTAGES

This model when given a proper refined data set has many advantages over a classic ML model in OV that promote accuracy and better prediction capabilities in order to achieve this we have used statistical probability techniques and data analysis methods in a way which works best for the kind of data set that we are dealing with. This model is also unique in the way that it handles the pre-processing part of data analysis special care has been taken to remove all the kinds of inconsistencies and incorrect values that me hamper with our further analysis. These techniques involves normalization feature generation feature scaling as well as victimization of different key features that have been included in our input data set. Apart from this this model is easily scalable which means it can be employed to a much larger data set without many changes and still give accurate results because it focuses on the features rather than the volume of data that is being input and has the unique capability of selecting attributes that are most important for our medical diagnosis.

DISADVANTAGES

If we keep apart numerous advantages that you’re models brings to the table then we can focus on a few disadvantages that are also present in our current diagnosis. These disadvantages range from being human error to incorrect information and even includes hardware limitations. Majority of the data said that has been availed to us in this particular field faces the problem of being highly imbalanced which means that most of the predictions come out as negatives rather than positives. In order to come back this problem we need to apply factorization methods that can curb this issue. Apart from this when we are dealing with a voluminous data then the machine learning part of a model requires a lot of hardware support which is incompatible with our current working personal computers or devices. Due to this we face the unique problem of hardware limitations in our current model while dealing with bigger data sets. Also some cases are exceptional in a way that it’s difficult to predict the outcome of stroke because they may be the cause of numerous ambiguous features such as hereditary issues or a particular combination of various medical and lifestyle reasons that cannot be quantized measured or analyzed. We have taken special care to the best of our abilities so that these problems and their effect can be minimized but our model is not without errors and this factor is also reflected on the accuracies that we have achieved with different machine learning algorithms.

USE OF ML AND OTHER TOOLS IN THE PROJECT

Our model makes direct use of machine learning in various ways and utilizes various techniques and algorithms that are present in the machine learning paradigm. Various python libraries used are

1. Pandas

2. Matplotlib

3. Matplotlib in-line

4. Seaborn

All of the coding has been done in python three on the Google Colab platform and utilizes cloud storage for the computation of the machine learning algorithm. Apart from this , we use multiple processes of EDA on our data set and pre process our data suck as removal of outlier removal and treatment of missing values. Practices such as label encoding and normalization. Different ML algorithms we employed are Decision Tree , Logistic Regression , KNN, Random Forest and SVM.

In order to apply all of these algorithms to our input data said we need a very fundamental understanding of how each and every algorithm works in order to do this we visited various credible sources and collected information about sin Texas formatting and hyper parameter techniques of various machine learning algorithms. All of the work that has been done has been performed on personal devices and various resources used have been listed in references below. The data said utilized has been taken from Kaggle and has been verified to be credible and reliable. Apart from this all the basic steps of machine learning model building have been utilized such a splitting the data training the data and then finally testing the data and then comparing that I could assist that we have obtained from various techniques.

CHAPTER :2- LITERATURE REVIEW

In the development of this report about 50 papers and various academic links have been visited examined and summarized in order to develop a clear understanding of the already present literature that is in main stream and is relevant to the problem statement. Most of the work that has been already done works on predictive algorithms that came to establish some sort of relationship between various risk factors and types of strokes that the patient’s face.

DATA ACQUISITION

When it comes to acquisition of the data that has been used in various papers then it can be simply generalized that all of this data either comes from medical records that have been provided by various hospitals or this data is part of a bigger medical registry that has been maintained by government of various countries most of the data scientist have used data that is validated as well as credible but one common problem that plagues all of these data sets is the highly imbalance nature that puts the entire analysis process at risk. Moreover not all of the data is complete and free of irregularities and errors due to this as well the data analysis progress slows down and the medical diagnosis cannot be done in an efficient manner. Data collection methods are often mug as third-party rather than being taken from the primary source which also creates a question on credibility of data as well as data privacy ethics such as security will you share and disclosure on how this data will be used. Almost all of the papers use Principle component analysis on these data sets which mostly contain of the status BMI smoking status glucose levels hypertension history of brain or heart related issues in family as well as common factors such as age, marriage status etc. But some studies that have been performed also included images of CT scan that had been done on patient named towards developing a neural networks that could take all of these images and then by mapping out common buttons and anomalies in the brain scan in order to compute the probability of a patient having a stroke.

CHAPTER :3- METHODOLOGY

For development of a relevant and functional model, we aim to automate the stroke prediction system by building up our application using machine learning algorithm and neural networks. we will be working with a dataset that primarily focuses on features such as alcohol consumption, family history, lifestyle, red blood cells count, smoking history, work type and other health related data which will be utilized as the input to our model. Before feeding this data, it would need to be preprocessed as well as cleaned by removing irrelevant data. Feature generation and scaling will need to be employed to increase the data set utility. After this, data will be split into training and testing modules and worked upon with machine learning algorithms. The results of this predictive procedure will be analyzed. At the end, the accuracy and efficiency of model will be calculated, and conclusions will be presented in visual format as well as via parameters such as Sensitivity, Confusion matrix and ROC Curve.

The past data sets available for demystifying this issue are highly imbalanced and would thus need to be put through appropriate processing in order to be useful. Apart from this, we would need a deep understanding of relevant programming language and all related machine learning concepts such as Naïve Bayes, Neural Networks and Data Mining as well to present a complete picture. We will be utilizing already available literature to familiarize ourselves with these terminologies. The datasets used will be acquired from registered and validated government sites to maintain the integrity of predictive analytics. The employment of these concepts will be done on platforms equipped to run Python and R.

RELATED TERMS

Exploratory Data Analysis-

Exploratory data analysis is a technique that is being used by data analyst and scientist alike in order to analyze an exam in various data sets and summarize the main characteristics results an important features. And processes such as data visualization are often used along with EDA as well. Using explode 80 data analysis we can determine what are the best ways to manipulate a data sources in order to get all of the answers that we need enhance it makes it very easy for data scientist for processes such as discovering patents and uncovering anomalies. it also helps in hypothesis testing or checking all of the assumption that we have a bother data. In general idiot is used to find out what ever details the data maybe revealing when we go beyond the modeling that is formally done all the testing that is done hypothetically hence we can say that it provides us with a better understanding of where is variables that are present in a data set and also what kind of relationships that they have between them it also helps us to deter mine various statistical techniques that we could be considering for a data and which would be prove to be appropriate. With the help of ETA we can have a deeper look into the data and this would help us from not falling for any assumptions or any biases that we might have it also helps us to identify various errors which might have escaped the initial examination of the data also the various relations that sound among the different variables help us to calculate correlation metrics which are the fundamental reason behind key behind feature generation.

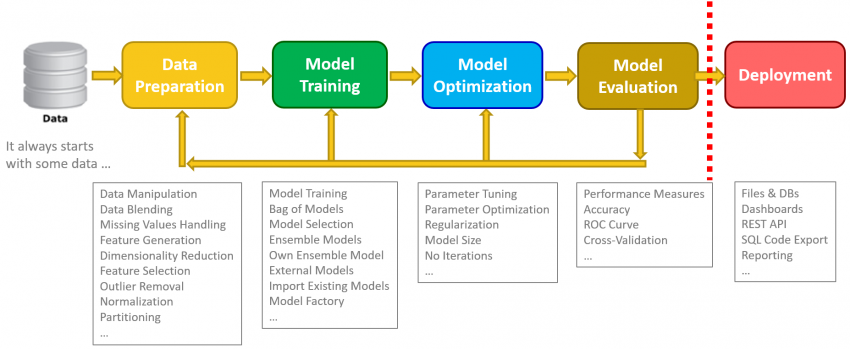


Fig-1: Steps of data analysis

MACHINE LEARNING MODELS UTILIZED

Deep learning and machine learning have both been widely adopted, yet they are also frequently misinterpreted. In this post, I'd want to take a step back, define machine learning and deep learning in simple words, go through some of the most popular machine learning algorithms, and explain how those algorithms connect to the other parts of the jigsaw of building predictive models from historical data.

Recall that a group of techniques known as machine learning are used to automatically build models out of data. The algorithms that transform a data set into a model are known as machine learning algorithms, and they are the heart of machine learning. The appropriate algorithm to use (supervised, unsupervised, classification, regression, etc.) will rely on the type of issue you're trying to solve, the computational power you have at your disposal, and the type of data you have.

In this model we have basically used 5 Machine Learning Algorithms to predict that the person can get a stroke or not according to their data.

The Algortihms which we used showed different accuracy levels of prediction of the data so when we used 5 algorithms this makes the model more precise and more efficient.

1. Decision tree
2. Logistic regression
3. KNN Algorithms
4. Random Forest
5. SVM Algorithm

Now we will knowing in brief way about the Algorithms and their accuracy levels we have also plot the graph about accuracy levels in our model .

1. DECISION TREE:-

The best and most often used technique for categorization and prediction is the decision tree. In a decision tree, which resembles a flowchart, each internal node indicates a test on an attribute, each branch shows the test's result, and each leaf node (or terminal node) has a class label.

Diagram

Description automatically generated

Fig-2:Decision tree

How do we construct a decision tree:-

It is possible to "learn" a tree by dividing the source set into subsets according to an attribute value test. Recursive partitioning is the process of repeating this operation on each derived subset. The recursion ends when the target variable in the subset at each node is equal to zero, or when splitting no longer improves the predictions. A decision tree classifier may be built without the need for subject expertise or parameter tuning, making it suited for exploratory knowledge discovery. High-dimensional data is manageable for decision trees. Decision tree classifiers often have excellent accuracy. A popular inductive strategy to learn about classification is decision tree induction.

Syntax:-

“from sklearn.tree import DecisionTreeClassifier

dt=DecisionTreeClassifier()”

Now when we applied decision tree and calculated the accuracy score of this algorithm came to be very less on comparison of other algorithms it was “0.9070450” to be precise as you can see from the snippet below:-

Graphical user interface, text, application, email

Description automatically generated

Fig-3- Accuracy of decision tree

1. LOGISTIC REGRESSION:-

The early twentieth century saw the usage of logistic regression in the biological sciences. This led to several social science applications using it. When the goal (dependent variable) is a categorical variable, logistic regression is utilised. for instance Identifying spam in emails, either (0). Whether or if the tumour is cancerous (0). Take into account a case where we must determine whether or not an email is spam. Setting up a threshold on which classification may be done is necessary if we choose to solve this problem using linear regression. For example, if the data point is categorised as non malignant but the actual class is malignant and the projected continuous value is 0.4 and the threshold value is 0.5, this might have major consequences in reality.

Chart

Description automatically generated

Fig-4: Comparison chart between Linear and Logistic regression

Logistic Regression Types:

There are just two outcomes in a binary logistic regression, such as yes or no to the question, "Is there spam?" These are the results of the chosen variable.

Multinomial Logistic Regression: The target variable contains three or more nominal categories, and that too, without ordering, such as predicting the kind of wine, the favoured superstar, or the favourite dish.

Ordinal Logistic Regression: When the target variable includes three or more ordinal categories, the categories are considered to be in a logical order. Using a scale of 1 to 5, for instance, rate any item, eatery, or movie.

We will be using Binary Logistic regression in our model and it came to be out the best predictive algorithm with higher accuracy rates .

Syntax and accuracy rates snippets from the code are below:-

Text

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Fig-5: accuracy of logistic regression

as we can see it has an accuracy rate of “0.94716242”.its pretty good!

1. KNN Algorithm:-

One of the simplest machine learning algorithms, K-Nearest Neighbor uses the supervised learning method.

The K-NN method places the new case in the category that is most similar to the available categories based on the assumption that the new instance and the data are comparable to the examples that are already accessible.

The K-NN algorithm saves all the information that is available and categorises new input based on similarity. This implies that as fresh data is generated, it may be quickly categorised into a suitable category using the K-NN method.

Chart, diagram

Description automatically generated

Fig-6: Mechanism of KNN

The following method may be used to describe how K-NN works:

Step 1: Decide which neighbour has K as their K-number.

Step 2: Calculate the Euclidean distance between K neighbours.

Step 3: Based on the determined Euclidean distance, select the K nearest neighbours.

Step 4: Count how many data points there are in each category among these k neighbours.

Step 5: Assign the new data points to the category where the neighbour count is highest.

Step 6: Our model is complete.

Syntax :- “from sklearn.neighbors import KNeighborsClassifier

knn=KNeighborsClassifier()”

accuracy level of KNN algorithm came out to be :- “0.9452054123 “ as u can see from the snippet from the code given below:-

1. RANDOM FOREST :-

An effective supervised learning method is Random Forest, a well-known machine learning algorithm. Both Classification and Regression issues in ML may be solved with it. It is based on the idea of ensemble learning, which is the act of integrating several classifiers to address a complicated issue and enhance the performance of the model.

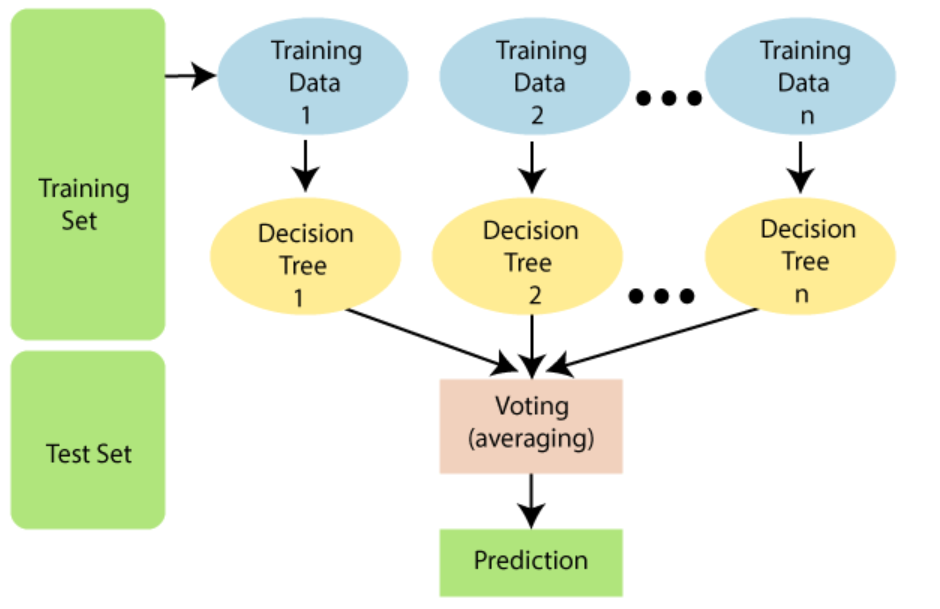


Fig-7: Mechanism of Random Forest

According to what its name implies, "Random Forest is a classifier that comprises a number of decision trees on various subsets of the provided dataset and takes the average to increase the prediction accuracy of that dataset." Instead of depending on a single decision tree, the random forest considers the forecast from each tree and guesses the result based on the predictions that have received the most votes.

Syntax :- “from sklearn.ensemble import RandomForestClassifier

rf=RandomForestClassifier() “

The accuracy of random forest Algorithm in this model came out to be :- “ 0.9471624“ below is the snippet from the code.

Graphical user interface, text, application, email

Description automatically generated

Fig-8: Accuracy of Random Forest

1. SVM ALGORITHM :-

For classification, regression, and outliers identification, support vector machines are a collection of supervised learning techniques. In machine learning, each of these tasks is typical.

By analysing millions of photos, you may use them to identify malignant cells, or you can use a well-fit regression model to forecast future travel routes.

Support vector regression (SVR), an extension of support vector classification, is one example of a specific sort of SVM you might employ for particular machine learning challenges (SVC).

The essential thing to bear in mind in this situation is that they are merely mathematical equations that have been optimised to provide you with the most accurate result in the shortest amount of time.

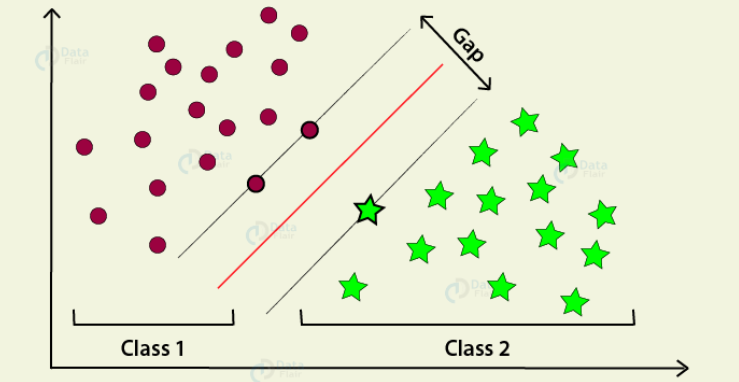


Fig-9: Working of SVM

Syntax :- “ from sklearn.svm import SVC

sv=SVC() “

Accuracy of this algorithm cameout to be around :- “0.9471624914 “ below is the snippet:-

Graphical user interface, application, Word

Description automatically generated

Fig-10:Accuracy Of SVM

Now we have also plotted a bar graph of comparing the Accuracy levels of the different algorithms we have used in our stroke prediction model so u can see the snip below it’s a bar graph generated by plot feature of python in jupyter notebook.

Bar chart

Description automatically generated with medium confidence

Fig-11: Accuracy BAR-Graph PLOT

DATASET UTILIZED

So the stroke prediction dataset is basically a large amount of data about people according to the survey about various important parameters.

This dataset plays a very important role in our stroke prediction model as this is the base on which we will decide whether the person is about to have a stroke or not .

The data we are using has 5110 rows and 12 columns pretty huge data , hence we can expect the data to be more precise and hence accuracy of prediction using Machine learning algorithms would be more!

Below is the snippet of the data:-

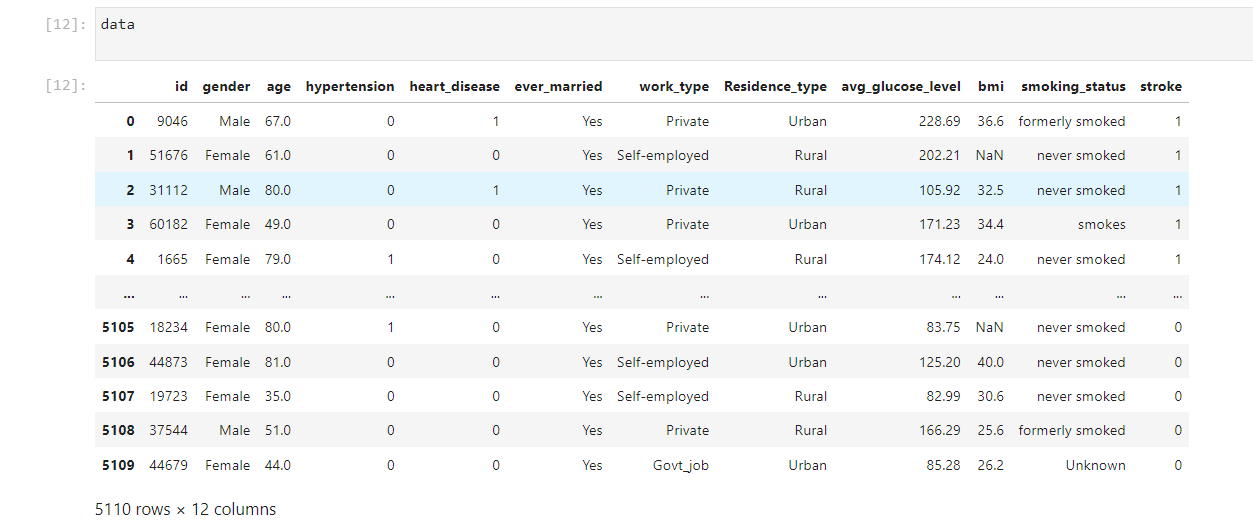


Fig-12: Snapshot of data

The parameters as we can see are:- Gender, age ,hypertension, heart disease , ever married , work type ,residence type , Bmi , smoking status, stroke.

* Now this data is not arranged so we will use various steps to prepare the data so that it can be used directly in our model.

PREPARING OF THE DATASET:-

The data has many problems so that is where is need to fix all the problems so that we can use this it into the model:-

1. There were null values encountered in some rows(we can fill null values or delete it)
2. The data has different data types like integers , alphabets , words , decimal so it cant be read by the Machine.(we will use encoding)
3. There are many useless parameters which are not required actually so we have to remove them for example :- column of id.
4. The data max and min value has too much of difference which will make traversing of data process too long and hence failure of the Model .(hence normalisation of data is required)
5. Now for the prediction process where we will apply different Algorithms of Machine learning we need to split the data in two parts. (hence we will use partiotioning of data)

REMOVING NULL VALUES:-

Now lets start with removing Null Values:-

Now as we need to check how many null values are actually present we had to use ,

“data.isnull().sum()” now here in the BMI column we saw there were 201 Null rows present. Below is the snippet:-

Chart

Description automatically generated with medium confidence

Fig-13: Null Value numbers

Now as we know 201 <<<<5110 rows so we can delete them as we can also fill them but here I decided to fill it with the value of mean . and hence we use :-“data['bmi'].fillna(data['bmi'].mean(),inplace=True)” now the data has been filled and replaced hence there are no null values found . below is the snippet:-

A picture containing table

Description automatically generated

Fig-14: Types of features

REMOVING UNWANTED COLUMNS:-

Now we will remove the unwanted columns of the data by removing them:-

Now to remove any column from the data we use:-

“data.drop('id',axis=1,inplace=True)” hence as id was present on the 1st axis it gets removed!

LABEL ENCODING:-

Now we will have to encode the dataset in order to make it readable by the machine:-

There are many ways to encode but I have used Label encoding to encode the data so that the terms get converted into different binary unique ids so its easy way to predict the data and hence easy readability and comparing gets easy.

For encoding we use

“from sklearn.preprocessing import LabelEncoder

enc=LabelEncoder()”

now as we can see the values inside Gender , residence , work status , ever married , smoking status gets changed to integers:-

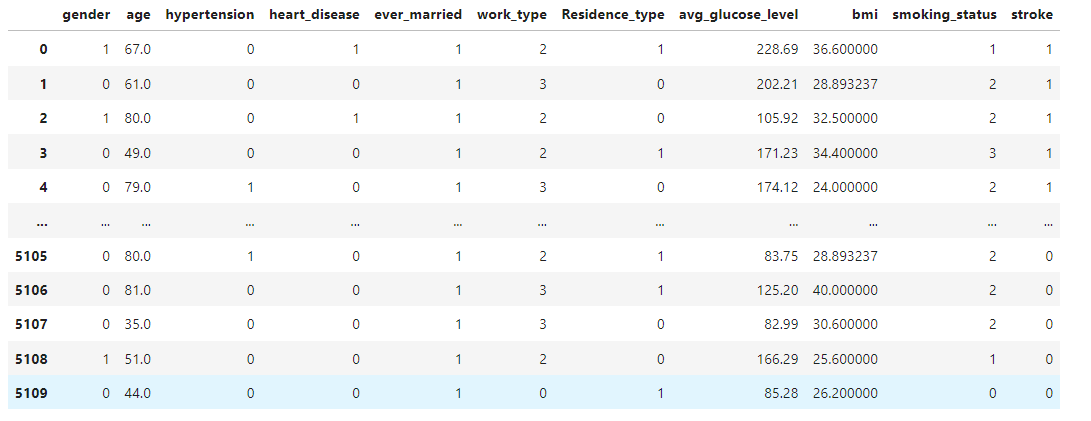


Fig-15: Encoding

As from the snippet we can see the data is encoded.

NORMALISATION OF DATA:-

Now we need to Normalize the data so it gets traversing the data from min to max value easy now see the snip below the data of min to max varies too much :-



Fig-16: Normalization

0.0800 to 82.00 too much variation present :

For normalising we will use :-

“from sklearn.preprocessing import StandardScaler

std=StandardScaler()”

this will remove this variation.

PARTITIONING OF DATA:-

To boost database management or query processing efficiency, data partitioning is a method that divides data among many tables, drives, or locations.

Now in order to use Ml Algorithms and predict the stroke we need to split the data in two parts :-

x—train/test 80/20 hence 80% of data goes in training part rest 20% in the testing part.

y-train/test this is the dataset which we will use in mapping with x-test results in order to view the accuracy of prediction.

Graphical user interface, text, application, email

Description automatically generated

Fig-17: Partitioning of data

Hence these were the steps we used in the preparation of the Dataset .

Now we are ready to apply different machine learning prediction alogorithms and check the accuracy of each in the Model.

CHAPTER :4- OPTIMISATION TECHNIQUES

The term "optimization" describes a process for determining the input variables or arguments to a function that provide the function's least or maximum output.

The most typical form of optimization issues in machine learning include continuous function optimization, where the function's input arguments are real-valued numerical values, such as floating-pointvalues.

The function's output additionally includes a real-valued assessment of the input values.

To distinguish them from functions that accept discrete variables and are referred to as combinatorial optimization issues, we may refer to problems of this kind as continuous function optimization.

There are many different kinds of optimization methods that may be applied to continuous function optimization issues, and there are probably just as many different ways to categorise and describe them.

**OPTIMISATION TECHNIQUES**:-

Graphical user interface, application, PowerPoint

Description automatically generated

Fig-18: Optimization techniques and their types

**EXHAUSTIVE SEARCH:-**

Simple is the method of exhaustive search. You would manually look for the appropriate number of clusters, for instance, if you were using the k-means approach. On the other hand, it gets too heavy and sluggish when there are hundreds of thousands of possibilities to take into consideration. As a result, in the vast majority of real-world situations, brute-force search is ineffective.

**GENETIC ALGORITHMS (simulated annealing):-**

Think about a collection of random algorithms you have at your disposal. This is who will live there. Some models are better adjusted than others when there are several models with some predetermined hyperparameters. Find them now! The accuracy of each model is first calculated. Only those that performed the best are then kept.

A second generation of models may now be created by generating some offspring with hyperparameters that are comparable to those of the best models.

In this image, the algorithm's reasoning is clear to see:

Diagram

Description automatically generated

Fig-19: Reasoning of algorithm

**GRADIENT DESCENT:-**

For model optimization and error minimization, gradient descent is the most used approach. You must go through the training dataset repeatedly while modifying the model to conduct gradient descent.

As a result of getting the minimal error feasible and increasing the model's accuracy, your objective is to minimise the cost function.

An illustration of the gradient descent algorithm's path through the variable space is shown on the graph. Pick a random location on the graph and select a direction at random to commence going. The improper path was chosen if you see that the inaccuracy is growing.

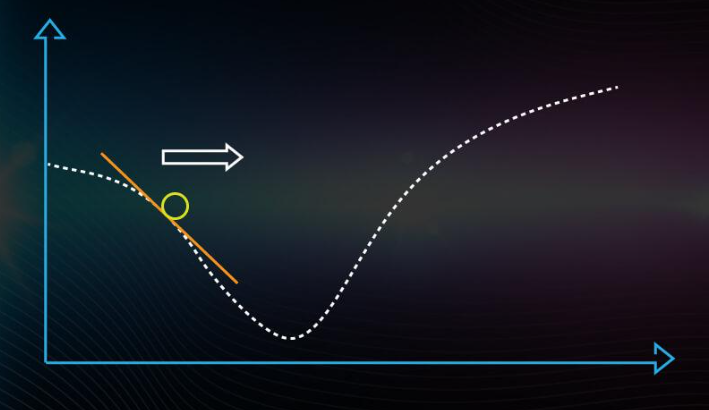


Fig-20: Graph of gradient descent

Hence , In order to achieve the best results, we compare the outcomes after each iteration while adjusting the hyperparameters. With fewer inaccuracy, we build a precise model. A model can be optimised in a variety of ways. Let's talk about the mathematics underlying the Gradient Descent and Stochastic Gradient Descent Algorithms, two crucial optimization algorithms, and how they are applied to machine learning models in this post.

CHAPTER :5- RESULTS

Following snapshots of our working model summaries the work done as well as the results that have been obtained.

Graphical user interface, application, Word

Description automatically generatedGraphical user interface, application

Description automatically generated

Graphical user interface, text, application, table, email

Description automatically generated

Table

Description automatically generated

Fig-21: Snapshot of result

CHAPTER :6- CONCLUSION AND FUTURE SCOPE

As a conclusion to a report we can easily state that machine learning plays a very important role in the medical diagnosis field and while it comes with certain challenges but if we work on the fundamentals and quality of data that is being provided to us as well as remove the hardware limitations that sometimes data scientist face then we can easily state that machine learning along with various data analysis processes can change the face of medicine as we know today. By the application of five algorithms that we have included in our paper there are many other algorithms that can also be employed in a similar way which may give similar or better results and hands there is an infinite number of constant combinations that we may employ in order to achieve greater accuracies and remove the false negatives as well as the false positives.

Machine learning in the recent times has already proved it self to be a tool that is east the process of specialists in various fields and we believe that with proper imagination an application of concepts it could do the same to the medicine as well and help people lead a better and healthier life.

By comparing various accuracies that we have obtained from machine learning algorithms we can easily see which has the song the best and which might need extra usage of accuracy boosting algorithms or other optimization techniques in order to achieve better results. We have also successfully showcased the importance of preprocessing outside input data sets by the usage of exploratory data analysis as well as normalization and visualization techniques and how they can increase our accuracy.

It has also come to our attention that various types of data sets may need different types of preprocessing and ML algorithms That has been catered to that particular kind of data set and can help us achieve better results in our medical diagnosis.

When we talk about the future scope of this particular project we can definitely work on the imbalance of the data said that we have faced which has worked significantly towards reducing or accuracy because in this way we are getting a much more higher value of type one and type two error. Such imbalances can be worked upon with LSTM techniques.

# REFERENCES

[1] Li, X., Liu, H., Du, X., Zhang, P., Hu, G., Xie, G., ... & Xie, X. (2016). Integrated machine learning approaches for predicting ischemic stroke and thromboembolism in atrial fibrillation. In AMIA Annual Symposium Proceedings (Vol. 2016, p. 799). American Medical Informatics Association.

[2] Paikaray, D., & Mehta, A. K. (2022). An extensive approach towards heart stroke prediction using machine learning with ensemble classifier. In Proceedings of the International Conference on Paradigms of Communication, Computing and Data Sciences (pp. 767-777). Springer, Singapore.

[3] Cheon, S., Kim, J., & Lim, J. (2019). The use of deep learning to predict stroke patient mortality. International journal of environmental research and public health, 16(11), 1876.

[4] Kim, J. K., Lv, Z., Park, D., & Chang, M. C. (2022). Practical machine learning model to predict the recovery of motor function in patients with stroke. European Neurology, 1-7.

[5] Heo, J., Yoo, J., Lee, H., Lee, I. H., Kim, J. S., Park, E., ... & Nam, H. S. (2022). Prediction of Hidden Coronary Artery Disease Using Machine Learning in Patients With Acute Ischemic Stroke. Neurology.

[6] Heo, J., Yoo, J., Lee, H., Lee, I. H., Kim, J. S., Park, E., ... & Nam, H. S. (2022). Prediction of Hidden Coronary Artery Disease Using Machine Learning in Patients With Acute Ischemic Stroke. Neurology.

[7] Rana, C., Chitre, N., Poyekar, B., & Bide, P. (2021, July). Stroke Prediction Using Smote-Tomek and Neural Network. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[8] <https://www.analyticsvidhya.com/blog/2021/05/how-to-create-a-stroke-prediction-model/>

[9] <https://arxiv.org/abs/2203.00497>

[10] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset