BRAIN STROKE PREDICTOR

E R Shyama

28-10-2023

Problem Statement

A stroke happens when a section of the brain experiences a reduction in blood flow, which results in the dysfunction of the body part that those brain cells control. This decrease in blood supply may be ischemic or haemorrhagic due to inadequate blood flow or bleeding into the brain tissue. A stroke is a medical emergency due to the potential for death or permanent disability.

Stroke significantly burdens both people and national healthcare systems. Stroke risk factors that may be altered include hypertension, diabetes, and cardiovascular disease, atrial fibrillation, abnormal glucose metabolism, and lifestyle risk factors. Therefore, the goal of our study is to correctly predict the stroke utilising large existing data sets based on potentially modifiable risk components using machine learning approaches.

Numerous apps currently in use implement brain stroke prediction utilising naive bayes and decision tree algorithms. It becomes challenging for the decision tree to decide on the proper threshold to divide the data points into distinct nodes in the case of continuous characteristics. Naive Bayes bases its usefulness in real-world use situations on the supposition that all features are independent. This approach encounters the "zero-frequency problem," where it gives a categorical variable with zero probability if its category was not present in the training dataset but was present in the test data set. Different machine learning models, like Logistic Regression, K-Nearest Neighbours, AdaBoost Classifier, XGBoost Classifier, and Random Forest Classifier, are used in the proposed solution. These models can be used to predict stroke and may be utilised by doctors to do so in the real world.

Market/Customer Need Assessment

With 400–800 strokes per 100,000, 15 million new acute strokes annually, 28,500,000 disability adjusted life years, and 28–30-day case fatality rates ranging from 17% to 35%, stroke is the second greatest cause of death and adult disability globally. With the number of fatalities from heart disease and stroke expected to rise to five million in 2020 from three million in 1998, the burden of stroke is anticipated to get worse. This will happen as a result of the ongoing demographic and health changes that will lead to an increase in the risk factors for vascular disease and the senior population. 85% of all stroke deaths worldwide occur in developing nations. The effects of stroke on society and the economy are significant. According to estimates, In the United States of America (USA), the cost of stroke was anticipated to be as high as $49 billion in 2002, whilst expenditures upon discharge were estimated to total of around 3 billion Euros in France.

Target Specification and Characterization

The proposed system can be used by clinics and doctors in more analysis on a patient so that he/she can be given more medical attention. The suggested system operates as a machine that aids in diagnosis and supports forecasts. This software is designed to be used by hospitals and other healthcare organisations so that they can give their patients preventive care by anticipating the possibility of brain stroke.

External Search

- About Stroke in Wikipedia - https://en.wikipedia.org/wiki/Stroke
- "Mayo Clinic" Stroke - Symptoms and causes - https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113
- https://www.jetir.org/papers/JETIR2204518.pdf
- https://towardsdatascience.com
- https://www.foreseemed.com/blog/machine-learning-in-healthcare
- Stroke prediction dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
-

Data description:

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: 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: "No" or "Yes"
7) work_type: "children", "Govt_jov", "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 if the patient had a stroke or 0 if not

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

Applicable Regulations

- Data Protection and Privacy Rules
- License for open-source codes that might be used in the model implementation
- Government Rules and Regulations

Business Model

Stroke cannot be easily predicted by evaluating doctor's diagnostic data. By predicting the likelihood of a stroke occurring in patients, an AI-based tool enables neurologists and cardiologists to give better treatment to their patients. Patients who sought treatment for a stroke as soon as possible had a good chance of continuing to be healthy for years to come.

A smartphone app powered by AI might be released on the market to help medical professionals. This opens up the possibility of marketing a service-based business model where customers subscribe to apps by paying monthly fees.

Concept Generation

When the brain's blood supply is cut off or there is unexpected brain haemorrhage, a stroke may result. There are two distinct stroke kinds. An ischemic stroke is one when there is a blockage of blood flow to the brain. Blood cannot provide the brain with nutrition and oxygen. Brain cells start to die within minutes of being deprived of oxygen and nutrition. Haemorrhagic strokes are caused by abrupt bleeding in the brain and are the most common type of stroke. Blood leakage causes pressure on brain cells, which harms them.

Strokes that are ischemic (blood arteries are blocked) account for little under 90% of cases, whereas haemorrhagic strokes account for the remaining 10%. Based on the location of the blockage or bleeding in the brain, strokes are further categorised.

A stroke is an urgent medical matter. A stroke may result in permanent brain damage, chronic disability, or even fatality. Mild weakness, paralysis, or numbness on one side of the body or face can all be symptoms of a stroke. Other symptoms might be a sudden, strong headache, abrupt weakness, difficulty seeing, difficulty speaking or comprehending speech, and difficulty looking.

We must thus utilise machine learning to create the model based on the aforementioned challenge. The study of computer algorithms that develop automatically via usage and learning from data is known as machine learning (ML). It is considered to be a component of artificial intelligence. Without being expressly taught to do so, machine learning algorithms create a model using sample data, sometimes referred to as training data, in order to make predictions or judgements

Code Implementation

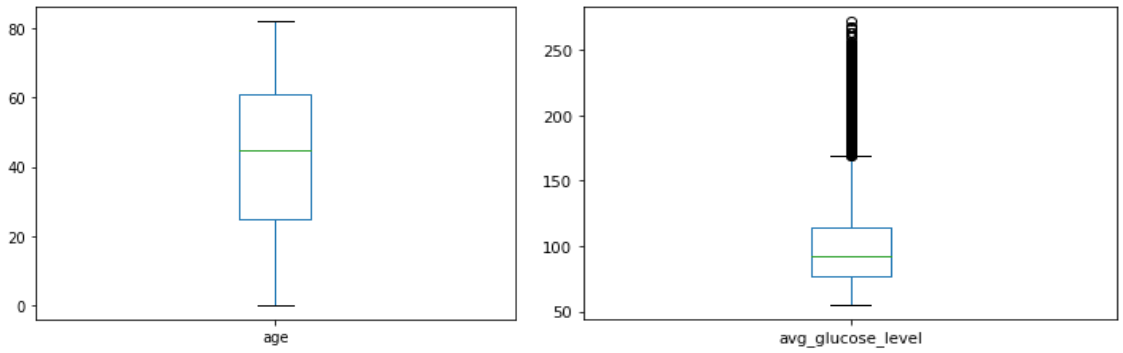
GitHub link:

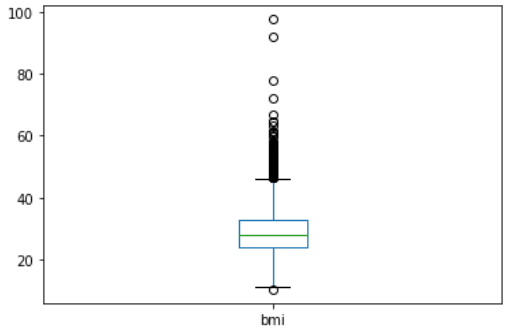
https://github.com/shyamaer/Feynn-Labs/tree/main/Brain%20Stroke%20Pediction

```
# Checking for missing value
df.isnull().any()
```

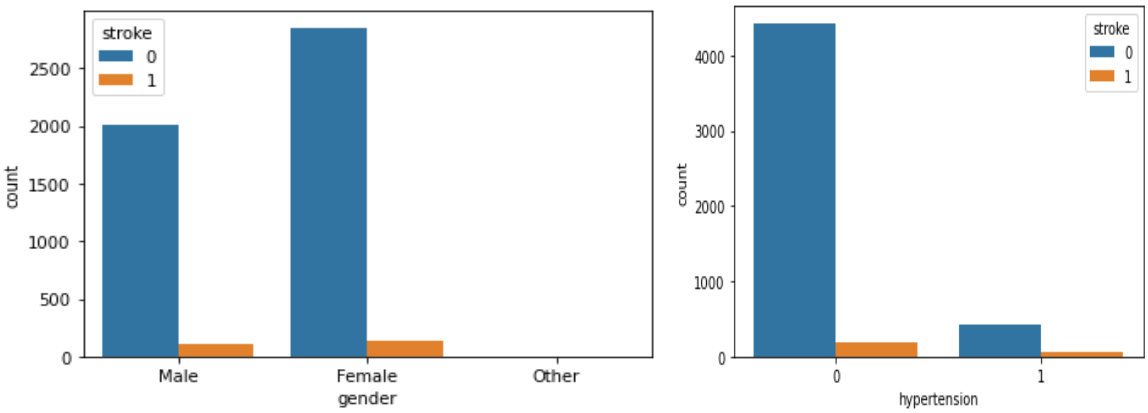
```
gender              False
age                 False
hypertension        False
heart_disease       False
ever_married        False
work_type           False
Residence_type      False
avg_glucose_level   False
bmi                  True
smoking_status      False
stroke              False
dtype: bool
```

bmi column has null values. This can be treated by imputing median value.





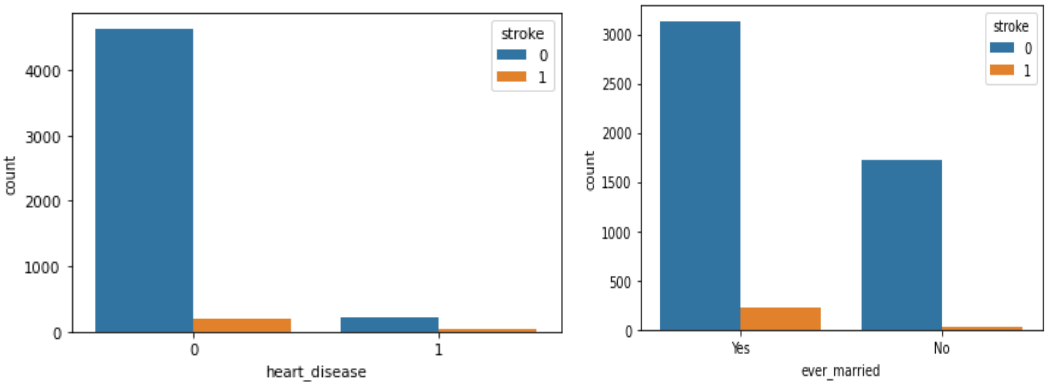
The columns ‘bmi’ and avg_glucose_level has outliers in it. This could be the result of both high or low BMI and glucose levels in individuals. The age column does not have any outlier.



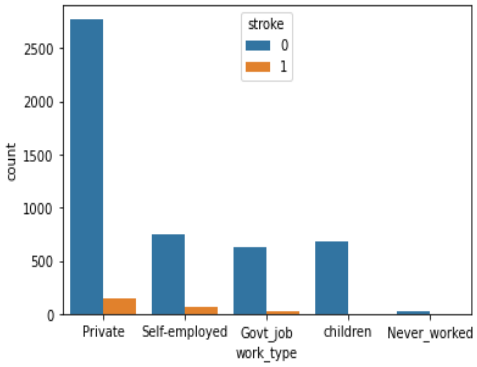
The first figure shows count of stroke in male, female and other category people.

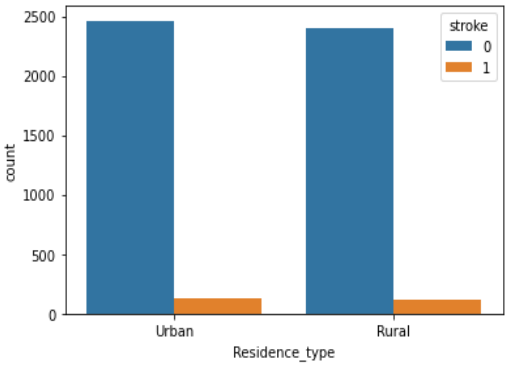
Male and female stroke rates are about equal, while female stroke rates are higher.

The number of strokes in persons with and without hypertension is depicted in the second image. Data indicates that stroke occurs in individuals without hypertension.

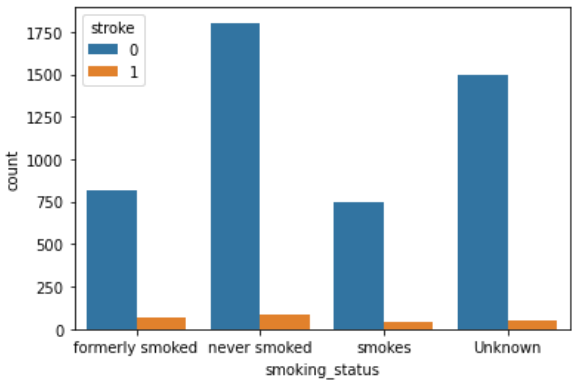


From the first figure people who do not have heart disease are more likely to have strokes and from second figure marriage is associated with a higher risk of stroke.



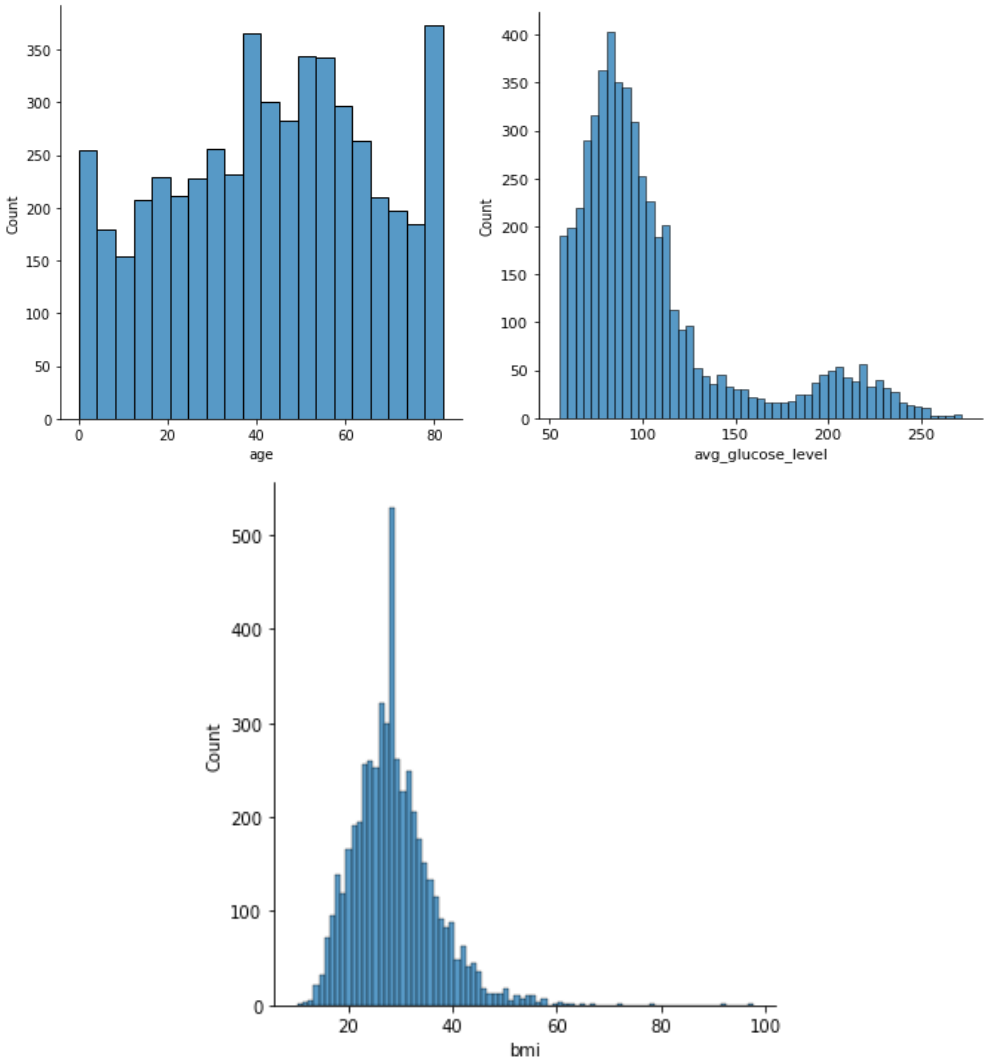


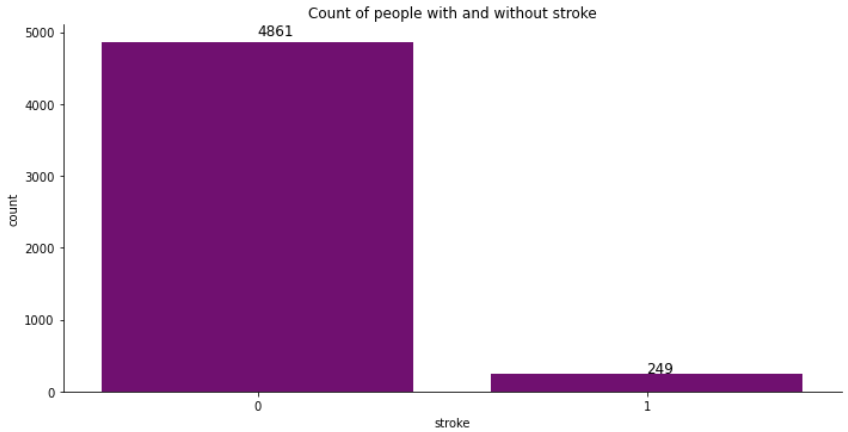
People who work in the private sector and reside in cities have a higher risk of stroke.



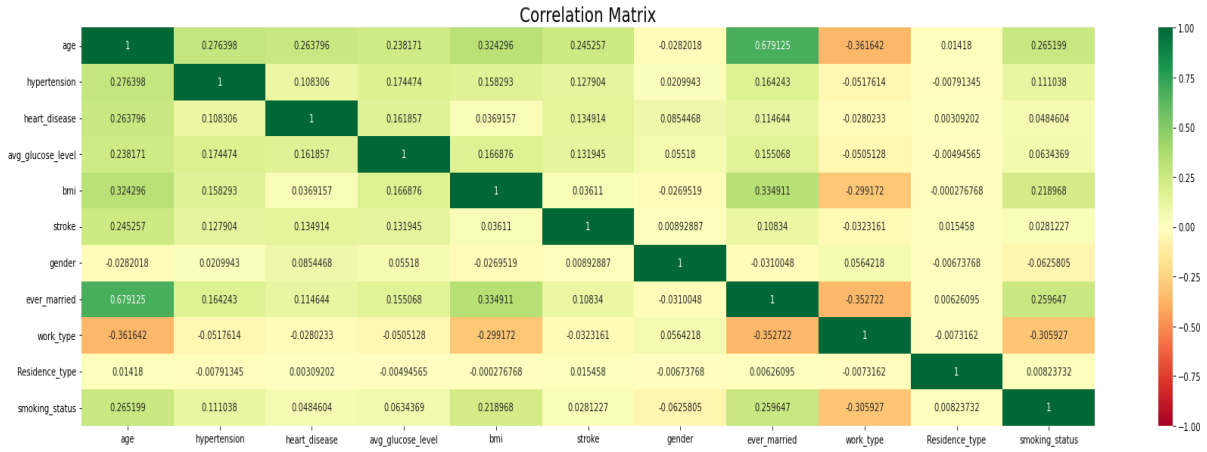
People who have never smoked and those who have smoked in the past are at risk for stroke, according to the findings.

The following pictures display the number of stroke victims at various ages together with their typical blood sugar levels and BMIs.





According to the statistics, the above figure displays the number of persons who have had and have not had a stroke. The dataset exhibits imbalance.



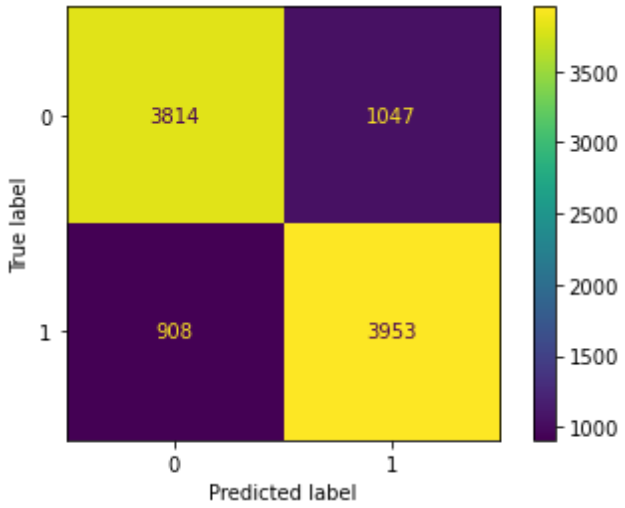
There is a slight correlation among some variables.

```
# balancing dataset
smote=SMOTE()
X,y=smote.fit_resample(X,y)
```

As the dataset is imbalanced SMOTE (Synthetic Minority Oversampling Technique) has been used to balanced our data.

Various classification models have been used.

Confusion matrix from Logistic Regression:



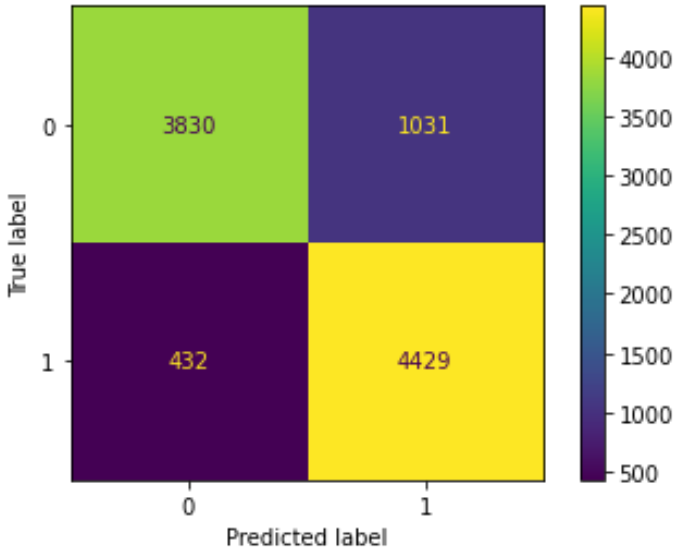

```
# Logistic Regression

from sklearn import metrics
from sklearn.metrics import confusion_matrix
print("Accuracy:",metrics.accuracy_score(y_test, y_pred_logreg))
print("Precision",metrics.precision_score(y_test,y_pred_logreg))
print("Recall",metrics.recall_score(y_test,y_pred_logreg))
print("f1_score",metrics.f1_score(y_test,y_pred_logreg))
```

```
Accuracy: 0.7984368572603867
Precision 0.7932421560740145
Recall 0.8088597210828548
f1_score 0.8009748172217709
```

The above figure shows the metrics calculated for Linear Regression Model.

Confusion Matrix from Random Forest Classifier



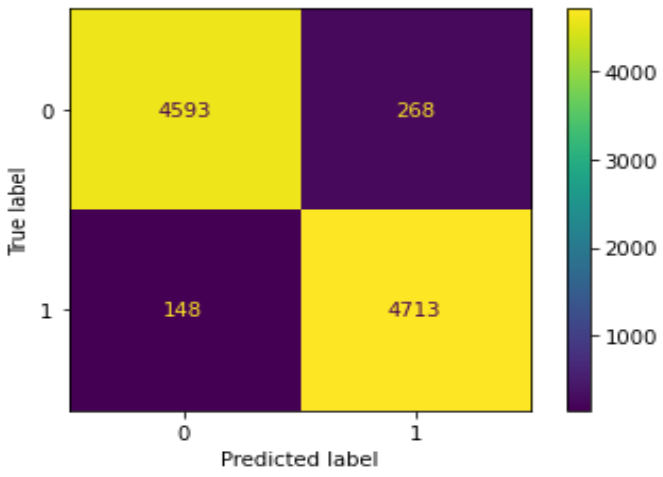
Metric values calculated are:

```
# Random Forest

print("Accuracy:",metrics.accuracy_score(y_test, y_pred_rf))
print("Precision",metrics.precision_score(y_test,y_pred_rf))
print("Recall",metrics.recall_score(y_test,y_pred_rf))
print("f1_score",metrics.f1_score(y_test,y_pred_rf))
```

```
Accuracy: 0.8445084327437269
Precision 0.8044895003620565
Recall 0.911402789171452
f1_score 0.8546153846153846
```

Confusion Metrics for Gradient Boost Classifier is:



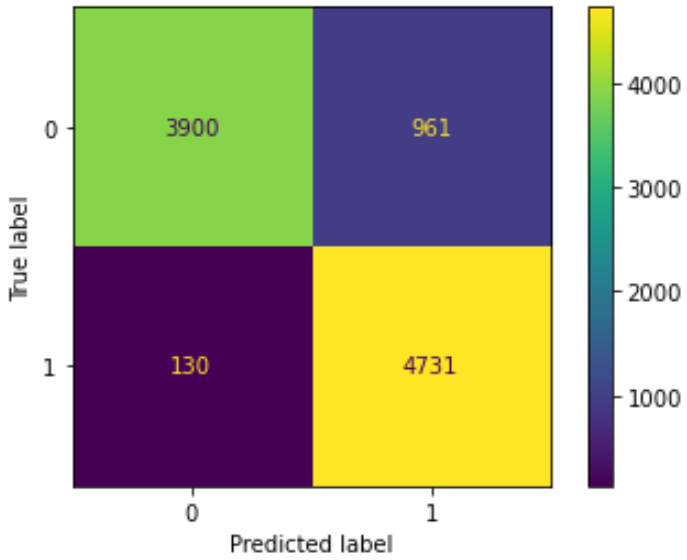
Metric values calculated are:

```
# Gradient Boosting

print("Accuracy:",metrics.accuracy_score(y_test, y_pred_gbm))
print("Precision",metrics.precision_score(y_test,y_pred_gbm))
print("Recall",metrics.recall_score(y_test,y_pred_gbm))
print("f1_score",metrics.f1_score(y_test,y_pred_gbm))
```

```
Accuracy: 0.9341834635952283
Precision 0.9219123505976096
Recall 0.9491386382280558
f1_score 0.9353274050121261
```

Confusion Metrics for K-Nearest Neighbour classification



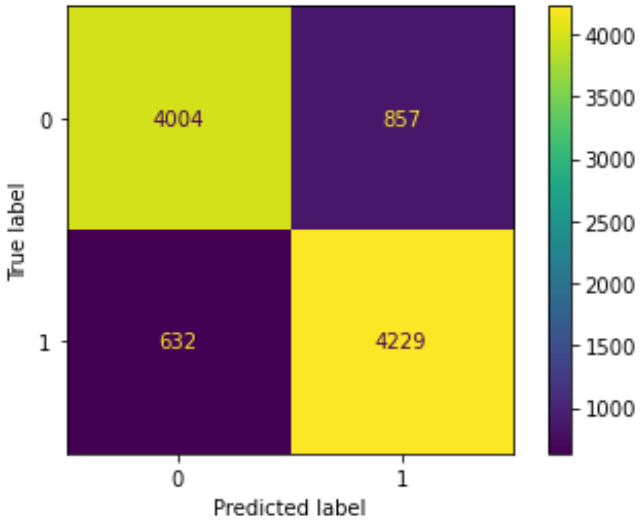
Metrics value calculated are:

```
# K-Nearest Neighbours

from sklearn import metrics
from sklearn.metrics import confusion_matrix
print("Accuracy:",metrics.accuracy_score(y_test, y_pred_knn))
print("Precision",metrics.precision_score(y_test,y_pred_knn))
print("Recall",metrics.recall_score(y_test,y_pred_knn))
print("f1_score",metrics.f1_score(y_test,y_pred_knn))
```

```
Accuracy: 0.8654874537227478
Precision 0.8046448087431693
Recall 0.9663658736669402
f1_score 0.8781215057771151
```

Confusion Metrics for Ada Boost Classifier



Metrics values are:

```
# Ada Boost

print("Accuracy:",metrics.accuracy_score(y_test, y_pred_adb))
print("Precision",metrics.precision_score(y_test,y_pred_adb))
print("Recall",metrics.recall_score(y_test,y_pred_adb))
print("f1_score",metrics.f1_score(y_test,y_pred_adb))
```

```
Accuracy: 0.8436857260386672
Precision 0.8259518259518259
Recall 0.8720262510254306
f1_score 0.8483639265762171
```

Gradient Boosting offers the best accuracy.

Concept Development

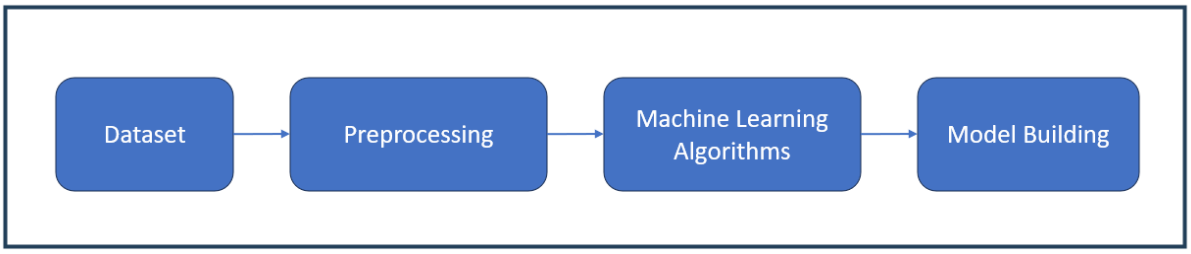
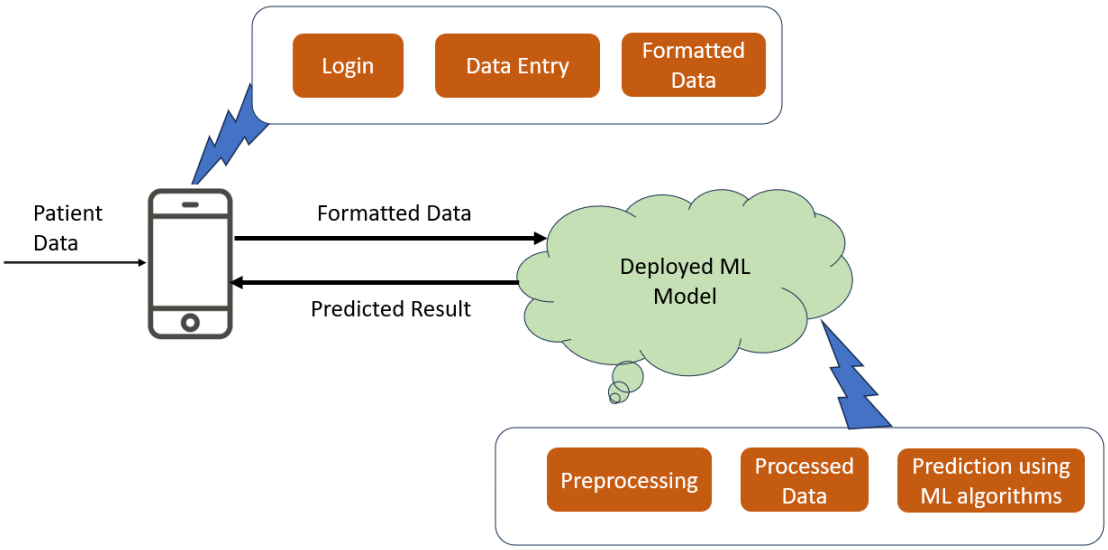


Fig : - Block diagram for the proposed system

Python is the programming language used by the Jupyter, which is where machine learning algorithms are learned and performed. A stroke dataset is first prepared. This dataset contains details of patients about hypertension, heart disease, type of employment and domicile, body mass index, average blood sugar level, smoker status, family history of stroke, blood pressure, and cholesterol levels. The dataset is made up of information gathered from individuals with a range of medical issues, from people of all ages, and from both genders.

The Jupyter environment imports this dataset. On the dataset, preprocessing is carried out. Boxplot diagrams are used to identify outliers and treat them. Statistical techniques like mean, median, and mode are used to identify and deal with missing values. The process of transforming categorical information into numerical representation is known as label encoding or one hot encoding. The dataset is then divided into training and test data. The Python has a number of libraries for training machine learning algorithms, including seaborn, scikit learn, and pandas. The model is then trained and tested using a variety of classification techniques, including Logistic Regression, K-Nearest Neighbours, AdaBoost Classifier, XGBoost Classifier, and Random Forest Classifier. Extensive testing is done on the chosen model and the algorithm that offers the best accuracy is chosen.

Final Product Prototype



The hospital uses an android application as part of the system. The patient's information is gathered by the hospital and translated to a JSON file format. All of the fundamental elements needed by the model will be included in JSON format, together with the metadata (patient name, hospital name, phone number). After then, it gets moved to AWS Cloud. The pre-processing takes place here. Following preprocessing, the processed data is fed to the ML models in order to make the prediction. The application receives and displays the projected outcome.

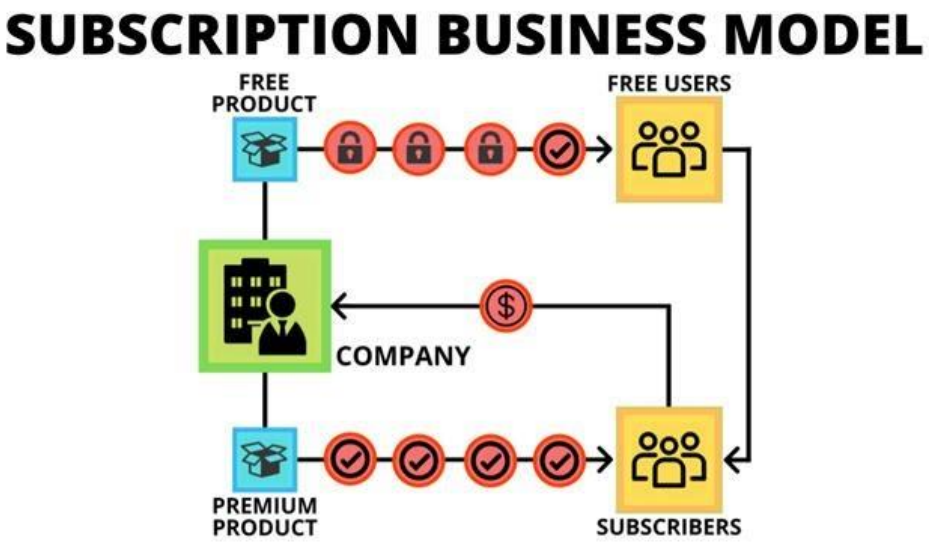
Conclusion

Researchers are creating models that are more and more accurate as datasets grow and improve in quality. ML will likely replace our neighborhood doctors in the future decades, which is fairly exciting even if we might not witness AI doing the role of a doctor today. The majority of ML models still lack sufficient data and are biased, so there is still a long way to go. Machine learning can be trained just as well as doctor prognosis, and prognosis is not an additional cost. The actual implementation is to be done to have clear understanding on the performance of the solution.

Business Modelling

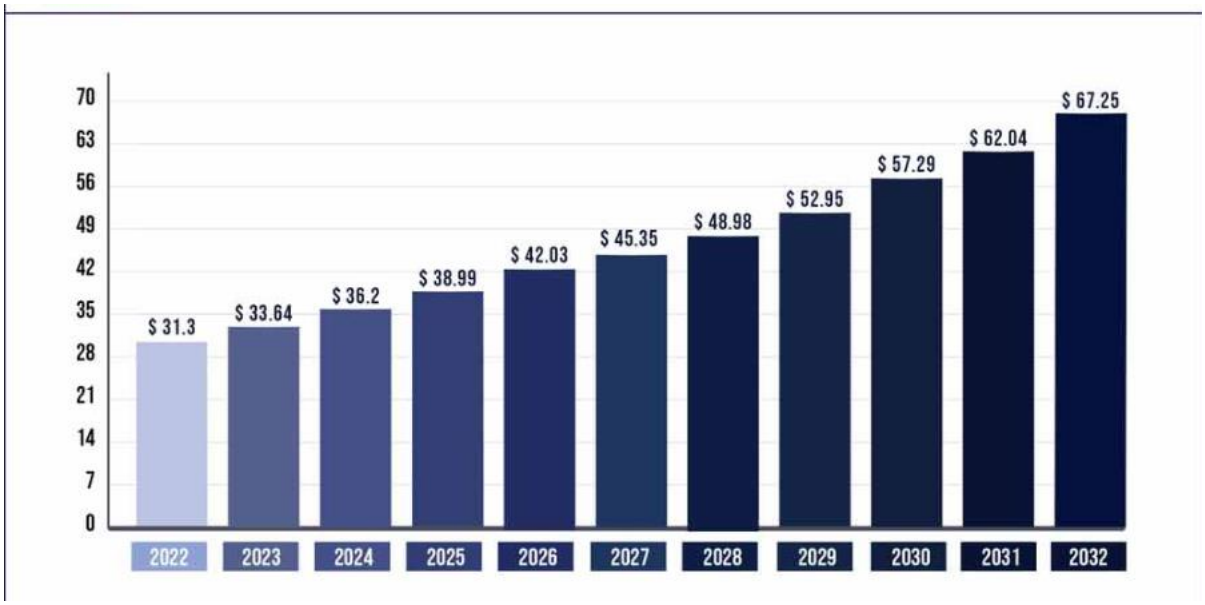
Each business model consists of three components: creating and producing the product, identifying the ideal clientele, arranging for client payment, and determining how the organisation will make money.

The Subscription Business Model is the one that may be applied in this situation. Customers pay a certain amount of money at predetermined intervals to access the company's product or service under the subscription business model. Because of its price structure, which involves charging clients a regular fee to access a product or service, the subscription model is utilised by businesses.



Financial Equation

Research indicates that the market for stroke is expected to expand between 2022 and 2032. The key drivers driving the market's expansion are the rising number of stroke-prone individuals, rising patient awareness, and rising healthcare costs.



The cost of the product comprises expenses for internet access, servers, software, research and development cost, maintenance cost, application launching, advertisement expenses and labour costs associated with data collection, ML modelling and product development.

So, it would be advisable to price our service at a subscription cost of about Rs.10000 per month. The duration of developing model would take 2.5 to 3 years as it is a medical application which deals with a critical condition.

$$y = 10000 * x(t) - (s_{ml} + e_{mt})$$

y = total profit

x(t) = total rate as a function of time

s_{ml} = fraction of development expense component.

e_{mt} = monthly maintenance expense component