STROKE PREDICTION USING MACHINE LEARNING

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

##### 

##### 

##### 

##### 

##### **STROKE PREDICTION USING MACHINE LEARNING**

##### **A Major Project Report**

***in partial fulfilment for the award of the degree***

***of***

##### **BACHELOR’s OF TECHNOLOGY**

***In***

**COMPUTER SCIENCE ENGINEERING**

###### **by**

##### **MAYANK SINGH, PRITIKA RANA and SHIVAM RATHI**

**Guided by**

**Dr SUNITA TIWARI**

**Assistant. Professor**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**G.B. PANT GOVT. ENGINEERING COLLEGE, NEW DELHI.**

(AFFILIATED TO GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY, DELHI)

**MAY, 2021**

**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the Major Project entitled **“Stroke Prediction Using Machine Learning”** in partial fulfilment for the award of the Degree of Bachelor of Technology in Computer Science and Engineering / Information Technology affiliated to **Guru Gobind Singh Indraprastha University, New Delhi** and submitted to the Department of Computer Science and Engineering G. B. Pant Govt. Engineering College is an authentic record of my work carried out during a period from March 2021 to May 2021. The matter represented in this report has not been submitted by me for the award of any other degree of this or any other institute/university.

**Date: - 10-05-2021 Name and Roll No:**

**Mayank Singh (00420907218)**

**Pritika Rana (02420902717)**

**Shivam Rathi (41420902717)**

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

#### *Signature of Supervisor*

#### **Date: - Name & Designation**

**TABLE OF CONTENTS**

**CHAPTER NO. TITLE PAGE NO.**

**LIST OF FIGURES i**

**ACKNOWLEDGMENT iii**

**ABSTRACT iv**

**CHAPTER 1. INTRODUCTION 1**

1.1 OVERVIEW 1

1.2 MOTIVATION 1

1.3 PROBLEM STATEMENT 1

1.4 SCOPE OF PROJECT 1

1.5 ACCESS TO USERS 2

**CHAPTER 2. RELATED WORKS 3**

2.1 INTRODUCTION 3

2.2 EXISTING SYSTEM 3

2.3 PROPOSED SYSTEM 4

2.3.1 IMAGE FEATURE EXTRACTION 4

2.3.2 THE LANGUAGE MODEL 5

2.3.3 FITTING THE MODEL 6

2.3.4 CAPTION GENERATION 6

2.3.5 SUGGESTION API 6

**CHAPTER 3. PROBLEM DESCRIPTION & SPECIFICATION 7**

3.1 PROBLEM OVERVIEW 7

3.2 SPECIFICATION 7

3.3 REQUIREMENTS 8

3.3.1 SOFTWARE REQUIREMENTS 8

3.3.2 HARDWARE REQUIREMENTS 8

3.3.3 OTHER REQUIREMENTS 9

**CHAPTER 4. SYSTEM DESIGN 10**

4.1 INTRODUCTION 10

4.1.1 DATA SOURCES 10

4.1.2 CONVOLUTION NEURAL NETWORK 10

4.1.3 RECURRENT NEURAL NETWORK 11

4.2 ARCHITECTURE DIAGRAM 11

4.3 UML DIAGRAM 12

4.4 DATA FLOW DIAGRAM 14

4.5 FLOW CHART DIAGRAM14

**CHAPTER 5. IMPLEMENTATION AND RESULT 16**

5.1 INTRODUCTION 16

5.2 MODEL SELECTION 16

5.3 IMAGE PREPROCESSING 17

5.4 TEXT PREPROCESSING 19

5.5 CREATING VOCABULARY 21

5.6 PREPARING THE MODEL 23

5.7 CONFIGURING THE MODEL 25

5.8 TRAINING AND SAVING THE MODEL 28

5.9 PREDICTING THE CAPTIONS 29

5.10 WEB PAGE DEVELOPMENT 31

5.11 API INTEGRATION 32

5.12 PROVIDING USER WITH CAPTIONS32

**CHAPTER 6. CONCLUSION AND FUTURE SCOPE 34**

6.1 CONCLUSION 34

6.2 FUTURE SCOPE34

**CHAPTER 7. REFERENCES 36**

**APPENDICES 39**

**APPENDIX 1: DETAILED TEST STRATEGY AND TEST CASES 39**

**APPENDIX 2: USER GUIDE 40**

**List of Figures**

**Figure No. Figure Title Page No.**

1 Flow Diagram of the project 8

2 Architecture of the image captioning model 11

3 Illustration image of functioning of the captioning model 12

4 Use Case Diagram for the captioning system 12

5 DFD Level 0 13

6 DFD Level 1 13

7 DFD Level 2 14

8 State Chart Diagram of the system 14

9 Flow chart for the caption generation process 15

10 Code Snippet showing the selection and adjustment of the model 16

11 Output representing different layers of the selected model 17

12 Code snippet showing preprocessing of images 17

13 A sample picture fetched from the dataset 18

14 Code snippet - Extracting feature map from passed images 18

15 Output representing the process of feature map creations 19

16 Code snippet showing text preprocessing 20

17 Sample captions fetched from Flickr8k.token.txt 20

18 Code snippet showing the process of caption dictionary creation 21

19 Sample view of caption dictionary 21

20 Code snippet showing Vocabulary Creation Process 22

21 Sample output showing the count of different words 22

22 Sample output showing the word dictionary 23

23 Sample output showing the integer form captions 23

24 Code snippet showing the max length of the caption present 24

25 Output showing the length of the largest caption in our collection 24

26 Code snippet showing the input preparation for the model 24

27 Code snippet showing the conversion to NumPy array form 25

28 Sample output showing the shapes of input 25

29 Configuring process of the image and language model 26

30 Output showing the image sequential model 26

31 Output showing the language sequential model 27

32 Output showing the combined mode 27

33 Detailed layout of the model along with shapes of inputs 28

34 Code snippet showing the training process 29

35 Sample output of the training process 29

36 Sample output of the inverse dictionary 29

37 Code snippet showing the image preprocessing for testing 30

38 Code snippet showing the prediction process and results 30

39 Some sample results generated by the model 31

40 Layout of the Home page of the Web App 33

41 Sample final results on the Web Platform 33

A1 Sample image showing loss and accuracy obtained by the model 38

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to my supervisor’s **Dr Sunita Tiwari** & **Dr Anu Saini** for providing their invaluable guidance, comments and suggestions throughout the project. I would specially thank Dr Sunita Tiwari for constantly motivating me to work harder and suggesting me different ways to tackle the problems encountered.

Also, I would like to thank my team members for their assistance in the project. I would like to express my gratitude for the Department of Computer Science Engineering, Gobind Ballabh Pant Government Engineering college to provide me with the opportunity to work on this project and to all the teachers and mentors guiding along the way.

* Mayank Singh (00420907218)
* Pritika Rana (02420902717)
* Shivam Rathi (41420902717)

**ABSTRACT**

The project Stroke Prediction Using Machine Learning provides user an ability to check the chances of stroke to him/her, based on different parameters asked to him/her. The user interacts with the project with the help of a Web Platform which contains a form having different key parameters which constitutes in stroke. On providing these inputs the model predicts the chances of stroke to him/her.

Stroke is a deadly disease and a leading cause of death. In most of the case the patient suffers grave consequences, along with the patient the family and relatives also suffer a lot. But studies suggest that 80 percent of stroke can be pre-diagnosed and, in that scenario, patient can be saved. There are some key parameters which contributes to chances of stroke. There parameter includes age, gender, hypertension, blood pressure, heart disease, average glucose level, BMI, smoking habits, work type, place of residence, marital status, previous history of stroke or heart attack etc. of the patient. As a patient is mostly well aware of these factors, he/she can get a diagnosis in the case of suspicion about chances of stroke. A lot of machine learning methods are developed and are currently under research to predict the chances of stroke based on different parameters.

In this project we have taken 10 different machine learning classifiers and tested them on our dataset “Stroke dataset” for accuracy. Then selected the best one for prediction. We have also performed k-fold validation and hyper-parameter optimization on the model. Along with that we have tried to identify the best features for the classifiers. The inputs gathered from user via the web platform is provided to the model using the flask backend system and the saved model is used for predicting the output which is displayed to the user on the result page.

The project provides user the facility to check the chances of stroke to them based on inputs features at the choice of location and time through their web enabled devices. The project will be hosted online on web app or mobile app and can be used by user from anywhere. Along with this our objective is also to compare various machine learning algorithm and learn how they perform on the dataset for predicting the stroke risk. What are differences among different model and also understand which classifiers performs best and what may be the reasons behind it.

**CHAPTER-1**

**INTRODUCTION**

**1.1 OVERVIEW**

# **Stroke: -**

A **stroke** is a [medical condition](https://en.wikipedia.org/wiki/Disease) in which poor [blood flow](https://en.wikipedia.org/wiki/Cerebral_circulation) to the [brain](https://en.wikipedia.org/wiki/Brain) causes [cell death](https://en.wikipedia.org/wiki/Cell_death). There are two main types of stroke: [ischemic](https://en.wikipedia.org/wiki/Brain_ischemia), due to lack of blood flow, and [hemorrhagic](https://en.wikipedia.org/wiki/Intracranial_hemorrhage), due to [bleeding](https://en.wikipedia.org/wiki/Bleeding). Both cause parts of the brain to stop functioning properly. Signs and symptoms of a stroke may include an [inability to move or feel](https://en.wikipedia.org/wiki/Hemiplegia) on one side of the body, [problems understanding](https://en.wikipedia.org/wiki/Receptive_aphasia) or [speaking](https://en.wikipedia.org/wiki/Expressive_aphasia), [dizziness](https://en.wikipedia.org/wiki/Dizziness), or [loss of vision to one side](https://en.wikipedia.org/wiki/Homonymous_hemianopsia). Signs and symptoms often appear soon after the stroke has occurred. If symptoms last less than one or two hours, the stroke is a [transient ischemic attack](https://en.wikipedia.org/wiki/Transient_ischemic_attack) (TIA), also called a mini-stroke. A [hemorrhagic stroke](https://en.wikipedia.org/wiki/Subarachnoid_hemorrhage) may also be associated with a [severe headache](https://en.wikipedia.org/wiki/Thunderclap_headache). The symptoms of a stroke can be permanent. Long-term complications may include [pneumonia](https://en.wikipedia.org/wiki/Pneumonia) and [loss of bladder control](https://en.wikipedia.org/wiki/Urinary_incontinence).

[](https://en.wikipedia.org/wiki/File:MCA_Territory_Infarct.svg)

Figure 1 [CT scan](https://en.wikipedia.org/wiki/CT_scan) of the brain showing a prior right-sided [ischemic](https://en.wikipedia.org/wiki/Ischemic) stroke from blockage of an artery. Changes on a CT may not be visible early on.

The main [risk factor](https://en.wikipedia.org/wiki/Risk_factor) for stroke is [high blood pressure](https://en.wikipedia.org/wiki/Hypertension). Other risk factors include [tobacco smoking](https://en.wikipedia.org/wiki/Tobacco_smoking), [obesity](https://en.wikipedia.org/wiki/Obesity), [high blood cholesterol](https://en.wikipedia.org/wiki/Hypercholesterolemia), [diabetes mellitus](https://en.wikipedia.org/wiki/Diabetes_mellitus), a previous TIA, [end-stage kidney disease](https://en.wikipedia.org/wiki/End-stage_kidney_disease), and [atrial fibrillation](https://en.wikipedia.org/wiki/Atrial_fibrillation).

An ischemic stroke is typically caused by blockage of a blood vessel, though there are also less common causes.

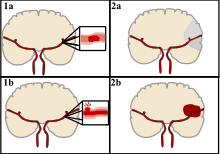
A hemorrhagic stroke is caused by either [bleeding directly into the brain](https://en.wikipedia.org/wiki/Intracerebral_hemorrhage) or into the [space](https://en.wikipedia.org/wiki/Subarachnoid_hemorrhage) between the [brain's membranes](https://en.wikipedia.org/wiki/Meninges). Bleeding may occur due to a ruptured [brain aneurysm](https://en.wikipedia.org/wiki/Intracranial_aneurysm). Diagnosis is typically based on a [physical exam](https://en.wikipedia.org/wiki/Physical_exam) and supported by [medical imaging](https://en.wikipedia.org/wiki/Medical_imaging) such as a [CT scan](https://en.wikipedia.org/wiki/CT_scan) or [MRI scan](https://en.wikipedia.org/wiki/Magnetic_resonance_imaging). A CT scan can rule out bleeding, but may not necessarily rule out ischemia, which early on typically does not show up on a CT scan. Other tests such as an [electrocardiogram](https://en.wikipedia.org/wiki/Electrocardiogram) (ECG) and [blood tests](https://en.wikipedia.org/wiki/Blood_test) are done to determine risk factors and rule out other possible causes. [Low blood sugar](https://en.wikipedia.org/wiki/Low_blood_sugar) may cause similar symptoms.

Prevention includes decreasing risk factors, [surgery to open up the arteries to the brain](https://en.wikipedia.org/wiki/Carotid_endarterectomy) in those with problematic [carotid narrowing](https://en.wikipedia.org/wiki/Carotid_stenosis), and [warfarin](https://en.wikipedia.org/wiki/Warfarin) in people with [atrial fibrillation](https://en.wikipedia.org/wiki/Atrial_fibrillation). [Aspirin](https://en.wikipedia.org/wiki/Aspirin) or [statins](https://en.wikipedia.org/wiki/Statin) may be recommended by physicians for prevention. A stroke or TIA often requires emergency care.

**An ischemic stroke, if detected within three to four and half hours, may be treatable with a** [**medication**](https://en.wikipedia.org/wiki/Thrombolytic_drug) **that can** [**break down the clot**](https://en.wikipedia.org/wiki/Thrombolysis)**.** Some hemorrhagic strokes benefit from [surgery](https://en.wikipedia.org/wiki/Neurosurgery). Treatment to attempt recovery of lost function is called [stroke rehabilitation](https://en.wikipedia.org/wiki/Stroke_rehabilitation), and ideally takes place in a stroke unit; however, these are not available in much of the world.

In 2013, approximately 6.9 million people had an ischemic stroke and 3.4 million people had a hemorrhagic stroke. In 2015, there were about 42.4 million people who had previously had a stroke and were still alive. Between 1990 and 2010 the number of strokes which occurred each year decreased by approximately 10% in the [developed world](https://en.wikipedia.org/wiki/Developed_world) and increased by 10% in the developing world. In 2015, stroke was the second most [frequent cause of death](https://en.wikipedia.org/wiki/List_of_causes_of_death_by_rate) after [coronary artery disease](https://en.wikipedia.org/wiki/Coronary_artery_disease), accounting for 6.3 million deaths (11% of the total). About 3.0 million deaths resulted from ischemic stroke while 3.3 million deaths resulted from hemorrhagic stroke. About half of people who have had a stroke live less than one year. Overall, two thirds of strokes occurred in those over 65 years old.

## **Classification**

[](https://en.wikipedia.org/wiki/File:Ischemic_Stroke.svg)

There are two main categories of strokes. Ischemic (top), typically caused by a blood clot in an artery (1a) resulting in brain death to the affected area (2a). Hemorrhagic (bottom), caused by blood leaking into or around the brain from a ruptured blood vessel (1b) allowing blood to pool in the affected area (2b) thus increasing the pressure on the brain.

[](https://en.wikipedia.org/wiki/File:MCA-Stroke-Brain-Human-2.JPG)

A slice of brain from the autopsy of a person who had an acute [middle cerebral artery (MCA)](https://en.wikipedia.org/wiki/Middle_cerebral_artery) stroke

Strokes can be classified into two major categories: [ischemic](https://en.wikipedia.org/wiki/Ischemia) and [hemorrhagic](https://en.wikipedia.org/wiki/Bleeding). Ischemic strokes are caused by [interruption of the blood supply](https://en.wikipedia.org/wiki/Perfusion#Malperfusion) to the brain, while hemorrhagic strokes result from the rupture of a [blood vessel](https://en.wikipedia.org/wiki/Blood_vessel) or an [abnormal vascular structure](https://en.wikipedia.org/wiki/Cerebral_arteriovenous_malformation). About 87% of strokes are ischemic, the rest being hemorrhagic. Bleeding can develop inside areas of ischemia, a condition known as "hemorrhagic transformation." It is unknown how many hemorrhagic strokes actually start as ischemic strokes.

### **Definition**

In the 1970s the [World Health Organization](https://en.wikipedia.org/wiki/World_Health_Organization) defined stroke as a "neurological deficit of cerebrovascular cause that persists beyond 24 hours or is interrupted by death within 24 hours" although the word "stroke" is centuries old. This definition was supposed to reflect the reversibility of tissue damage and was devised for the purpose, with the time frame of 24 hours being chosen arbitrarily. The 24-hour limit divides stroke from [transient ischemic attack](https://en.wikipedia.org/wiki/Transient_ischemic_attack), which is a related syndrome of stroke symptoms that resolve completely within 24 hours. With the availability of treatments that can reduce stroke severity when given early, many now prefer alternative terminology, such as brain attack and acute ischemic cerebrovascular syndrome (modeled after [heart attack](https://en.wikipedia.org/wiki/Myocardial_infarction) and [acute coronary syndrome](https://en.wikipedia.org/wiki/Acute_coronary_syndrome), respectively), to reflect the urgency of stroke symptoms and the need to act swiftly.

### **Ischemic**

In an ischemic stroke, blood supply to part of the brain is decreased, leading to dysfunction of the brain tissue in that area. There are four reasons why this might happen:

1. [Thrombosis](https://en.wikipedia.org/wiki/Thrombosis) (obstruction of a blood vessel by a blood clot forming locally)
2. [Embolism](https://en.wikipedia.org/wiki/Embolism) (obstruction due to an [embolus](https://en.wikipedia.org/wiki/Embolus) from elsewhere in the body),
3. Systemic hypoperfusion (general decrease in blood supply, e.g., in [shock](https://en.wikipedia.org/wiki/Shock_(circulatory)))
4. [Cerebral venous sinus thrombosis](https://en.wikipedia.org/wiki/Cerebral_venous_sinus_thrombosis).

A stroke without an obvious explanation is termed [cryptogenic](https://en.wikipedia.org/wiki/Idiopathic_disease) (of unknown origin); this constitutes 30–40% of all ischemic strokes.

### **Hemorrhagic**

[](https://en.wikipedia.org/wiki/File:Parachemableedwithedema.png)

[CT scan](https://en.wikipedia.org/wiki/CT_scan) of an intraparenchymal bleed (bottom arrow) with surrounding edema (top arrow)

There are two main types of hemorrhagic stroke:

* [Intracerebral hemorrhage](https://en.wikipedia.org/wiki/Intracerebral_hemorrhage), which is basically bleeding within the brain itself (when an artery in the brain bursts, flooding the surrounding tissue with blood), due to either [intraparenchymal hemorrhage](https://en.wikipedia.org/wiki/Intraparenchymal_hemorrhage) (bleeding within the brain tissue) or [intraventricular hemorrhage](https://en.wikipedia.org/wiki/Intraventricular_hemorrhage) (bleeding within the brain's [ventricular system](https://en.wikipedia.org/wiki/Ventricular_system)).
* [Subarachnoid hemorrhage](https://en.wikipedia.org/wiki/Subarachnoid_hemorrhage), which is basically bleeding that occurs outside of the brain tissue but still within the skull, and precisely between the [arachnoid mater](https://en.wikipedia.org/wiki/Arachnoid_mater) and [pia mater](https://en.wikipedia.org/wiki/Pia_mater) (the delicate *innermost* layer of the three layers of the [meninges](https://en.wikipedia.org/wiki/Meninges) that surround the brain).

The above two main types of hemorrhagic stroke are also two different forms of [intracranial hemorrhage](https://en.wikipedia.org/wiki/Intracranial_hemorrhage), which is the accumulation of blood anywhere within the [cranial vault](https://en.wikipedia.org/wiki/Cranial_vault); but the other forms of intracranial hemorrhage, such as [epidural hematoma](https://en.wikipedia.org/wiki/Epidural_hematoma) (bleeding between the skull and the [dura mater](https://en.wikipedia.org/wiki/Dura_mater), which is the thick *outermost* layer of the meninges that surround the brain) and [subdural hematoma](https://en.wikipedia.org/wiki/Subdural_hematoma) (bleeding in the [subdural space](https://en.wikipedia.org/wiki/Subdural_space)), are not considered "hemorrhagic strokes".

Hemorrhagic strokes may occur on the background of alterations to the blood vessels in the brain, such as [cerebral amyloid angiopathy](https://en.wikipedia.org/wiki/Cerebral_amyloid_angiopathy), [cerebral arteriovenous malformation](https://en.wikipedia.org/wiki/Cerebral_arteriovenous_malformation) and an [intracranial aneurysm](https://en.wikipedia.org/wiki/Intracranial_aneurysm), which can cause intraparenchymal or subarachnoid hemorrhage.[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

In addition to neurological impairment, hemorrhagic strokes usually cause specific symptoms (for instance, subarachnoid hemorrhage classically causes a severe [headache](https://en.wikipedia.org/wiki/Headache) known as a [thunderclap headache](https://en.wikipedia.org/wiki/Thunderclap_headache)) or reveal evidence of a previous [head injury](https://en.wikipedia.org/wiki/Head_injury).

## **Signs and symptoms**

Stroke symptoms typically start suddenly, over seconds to minutes, and in most cases do not progress further. The symptoms depend on the area of the brain affected. The more extensive the area of the brain affected, the more functions that are likely to be lost. Some forms of stroke can cause additional symptoms. For example, in intracranial hemorrhage, the affected area may compress other structures. Most forms of stroke are not associated with a [headache](https://en.wikipedia.org/wiki/Headache), apart from subarachnoid hemorrhage and cerebral venous thrombosis and occasionally intracerebral hemorrhage.

### **Early recognition**

Various systems have been proposed to increase recognition of stroke. Different findings are able to predict the presence or absence of stroke to different degrees. **Sudden-onset face weakness, arm drift (i.e., if a person, when asked to raise both arms, involuntarily lets one arm drift downward) and abnormal speech are the findings most likely to lead to the correct identification of a case of stroke, increasing the likelihood by 5.5 when at least one of these is present**. Similarly, when all three of these are absent, the likelihood of stroke is decreased (– [likelihood ratio](https://en.wikipedia.org/wiki/Likelihood_ratios_in_diagnostic_testing) of 0.39). While these findings are not perfect for diagnosing stroke, the fact that they can be evaluated relatively rapidly and easily make them very valuable in the acute setting.

In most cases, the symptoms affect only one side of the body (unilateral). Depending on the part of the brain affected, the defect in the brain is *usually* on the [opposite side](https://en.wikipedia.org/wiki/Contralateral) of the body. However, since these pathways also travel in the [spinal cord](https://en.wikipedia.org/wiki/Spinal_cord) and any lesion there can also produce these symptoms, the presence of any one of these symptoms does not necessarily indicate a stroke. In addition to the above CNS pathways, the [*brainstem*](https://en.wikipedia.org/wiki/Brainstem) gives rise to most of the twelve [cranial nerves](https://en.wikipedia.org/wiki/Cranial_nerves). A [brainstem stroke](https://en.wikipedia.org/wiki/Brainstem_stroke_syndrome) affecting the brainstem and brain, therefore, can produce symptoms relating to deficits in these cranial nerves:[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

* altered smell, taste, hearing, or vision (total or partial)
* drooping of eyelid ([ptosis](https://en.wikipedia.org/wiki/Ptosis_(eyelid))) and weakness of [ocular muscles](https://en.wikipedia.org/wiki/Extraocular_muscles)
* decreased reflexes: gag, swallow, pupil reactivity to light
* decreased sensation and muscle weakness of the face
* [balance problems](https://en.wikipedia.org/wiki/Balance_disorder) and [nystagmus](https://en.wikipedia.org/wiki/Nystagmus)
* altered breathing and heart rate
* weakness in [sternocleidomastoid muscle](https://en.wikipedia.org/wiki/Sternocleidomastoid_muscle) with inability to turn head to one side
* weakness in tongue (inability to stick out the tongue or move it from side to side)

If the [*cerebral cortex*](https://en.wikipedia.org/wiki/Cerebral_cortex) is involved, the CNS pathways can again be affected, but also can produce the following symptoms:

* [aphasia](https://en.wikipedia.org/wiki/Aphasia) (difficulty with verbal expression, auditory comprehension, [reading](https://en.wikipedia.org/wiki/Alexia_(condition)) and [writing](https://en.wikipedia.org/wiki/Agraphia); [Broca's](https://en.wikipedia.org/wiki/Broca's_area) or [Wernicke's area](https://en.wikipedia.org/wiki/Wernicke's_area) typically involved)
* [dysarthria](https://en.wikipedia.org/wiki/Dysarthria) ([motor speech disorder](https://en.wikipedia.org/wiki/Motor_speech_disorders) resulting from neurological injury)
* [apraxia](https://en.wikipedia.org/wiki/Apraxia) (altered voluntary movements)
* [visual field](https://en.wikipedia.org/wiki/Visual_field) defect
* memory deficits (involvement of [temporal lobe](https://en.wikipedia.org/wiki/Temporal_lobe))
* [hemineglect](https://en.wikipedia.org/wiki/Hemineglect) (involvement of [parietal lobe](https://en.wikipedia.org/wiki/Parietal_lobe))
* disorganized thinking, confusion, [hypersexual](https://en.wikipedia.org/wiki/Hypersexual) gestures (with involvement of frontal lobe)
* [lack of insight](https://en.wikipedia.org/wiki/Anosognosia) of his or her, usually stroke-related, disability

If the [*cerebellum*](https://en.wikipedia.org/wiki/Cerebellum) is involved, [ataxia](https://en.wikipedia.org/wiki/Ataxia#Cerebellar) might be present and this includes:

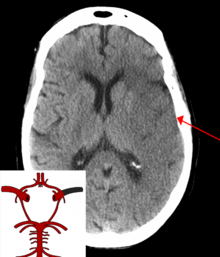
* altered walking [gait](https://en.wikipedia.org/wiki/Gait_abnormality)
* altered [movement coordination](https://en.wikipedia.org/wiki/Motor_coordination)
* [vertigo](https://en.wikipedia.org/wiki/Vertigo_(medical)) and or disequilibrium

### **Associated symptoms**

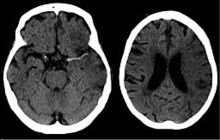
[Loss of consciousness](https://en.wikipedia.org/wiki/Unconsciousness), headache, and vomiting usually occur more often in hemorrhagic stroke than in thrombosis because of the increased [intracranial pressure](https://en.wikipedia.org/wiki/Intracranial_pressure) from the leaking blood compressing the brain.

If symptoms are maximal at onset, the cause is more likely to be a subarachnoid hemorrhage or an embolic stroke.

## **Diagnosis**

[](https://en.wikipedia.org/wiki/File:StrokeMCA_overlay.png)

A CT showing early signs of a middle cerebral artery stroke with loss of definition of the gyri and grey white boundary

[](https://en.wikipedia.org/wiki/File:Dens_media_sign_mit_Mediainfarkt_-_CCT_001.jpg)

Dens media sign in a patient with middle cerebral artery infarction shown on the left. Right image after 7 hours.

Stroke is diagnosed through several techniques: a neurological examination (such as the [NIHSS](https://en.wikipedia.org/wiki/NIHSS)), CT scans (most often without contrast enhancements) or [MRI scans](https://en.wikipedia.org/wiki/MRI_scan), [Doppler ultrasound](https://en.wikipedia.org/wiki/Doppler_ultrasound), and [arteriography](https://en.wikipedia.org/wiki/Arteriography). The diagnosis of stroke itself is clinical, with assistance from the imaging techniques. Imaging techniques also assist in determining the subtypes and cause of stroke. There is yet no commonly used [blood test](https://en.wikipedia.org/wiki/Blood_test) for the stroke diagnosis itself, though blood tests may be of help in finding out the likely cause of stroke.

### **Physical examination**

A [physical examination](https://en.wikipedia.org/wiki/Physical_examination), including taking a [medical history](https://en.wikipedia.org/wiki/Medical_history) of the symptoms and a neurological status, helps giving an evaluation of the location and severity of a stroke. It can give a standard score on e.g., the [NIH stroke scale](https://en.wikipedia.org/wiki/NIH_stroke_scale).

### **Imaging**

For diagnosing ischemic (blockage) stroke in the emergency setting:

* CT scans (*without* contrast enhancements)

[sensitivity](https://en.wikipedia.org/wiki/Sensitivity_(tests))= 16% (less than 10% within first 3 hours of symptom onset)

[specificity](https://en.wikipedia.org/wiki/Specificity_(tests))= 96%

* MRI scan

sensitivity= 83%

specificity= 98%

For diagnosing hemorrhagic stroke in the emergency setting:

* CT scans (*without* contrast enhancements)

sensitivity= 89%

specificity= 100%

* MRI scan

sensitivity= 81%

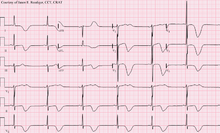
specificity= 100%

For detecting chronic hemorrhages, MRI scan is more sensitive.

For the assessment of stable stroke, nuclear medicine scans SPECT and PET/CT may be helpful. SPECT documents cerebral blood flow and PET with FDG isotope the metabolic activity of the neurons.

CT scans may not detect an ischemic stroke, especially if it is small, of recent onset, or in the brainstem or cerebellum areas. A CT scan is more to *rule out* certain stroke mimics and detect bleeding.

### Underlying cause

[](https://en.wikipedia.org/wiki/File:Left_MCA_Stroke.png)

12-lead ECG of a patient with a stroke, showing large deeply inverted [T-waves](https://en.wikipedia.org/wiki/T_wave). Various ECG changes may occur in people with strokes and other brain disorders.

### **Risk factors**

The most important modifiable risk factors for stroke are **high blood pressure** and **atrial fibrillation** although the size of the effect is small with 833 people have to be treated for 1 year to prevent one stroke. Other modifiable risk factors include **high blood cholesterol levels**, [**diabetes mellitus**](https://en.wikipedia.org/wiki/Diabetes_mellitus)**,** [**end-stage kidney disease**](https://en.wikipedia.org/wiki/End-stage_kidney_disease)**, cigarette smoking** (active and passive), **heavy** [**alcohol**](https://en.wikipedia.org/wiki/Alcohol_consumption_and_health) use**, drug use**, **lack of** [**physical activity**](https://en.wikipedia.org/wiki/Physical_activity)**,** [**obesity**](https://en.wikipedia.org/wiki/Obesity)**, processed** [**red meat**](https://en.wikipedia.org/wiki/Red_meat) **consumption, and unhealthy diet**. **Smoking just one cigarette per day increases the risk more than 30%.** Alcohol use could predispose to ischemic stroke, as well as intracerebral and subarachnoid hemorrhage via multiple mechanisms (for example, via hypertension, atrial fibrillation, rebound [thrombocytosis](https://en.wikipedia.org/wiki/Thrombocytosis) and [platelet aggregation](https://en.wikipedia.org/wiki/Platelet_aggregation) and [clotting](https://en.wikipedia.org/wiki/Clotting) disturbances). Drugs, most commonly amphetamines and cocaine, can induce stroke through damage to the blood vessels in the brain and acute hypertension. [Migraine](https://en.wikipedia.org/wiki/Migraine) with [aura](https://en.wikipedia.org/wiki/Aura_(symptom)) doubles a person's risk for ischemic stroke. Untreated, [celiac disease](https://en.wikipedia.org/wiki/Celiac_disease) regardless of the presence of symptoms can be an underlying cause of stroke, both in children and adults.

**High levels of physical activity reduce the risk of stroke by about 26%**. There is a lack of high-quality studies looking at promotional efforts to improve lifestyle factors. Nonetheless, given the large body of circumstantial evidence, best medical management for stroke includes advice on diet, exercise, smoking and alcohol use. Medication is the most common method of stroke prevention; [carotid endarterectomy](https://en.wikipedia.org/wiki/Carotid_endarterectomy) can be a useful surgical method of preventing stroke.

#### **Blood pressure**

[**High blood pressure**](https://en.wikipedia.org/wiki/Hypertension) **accounts for 35–50% of stroke risk**. Blood pressure reduction of 10 mmHg systolic or 5 mmHg diastolic reduces the risk of stroke by ~40%. Lowering blood pressure has been conclusively shown to prevent both ischemic and hemorrhagic strokes. It is equally important in secondary prevention.

#### **Blood lipids**

**High cholesterol levels** have been inconsistently associated with (ischemic) stroke. [**Statins**](https://en.wikipedia.org/wiki/Statins) **have been shown to reduce the risk of stroke by about 15%.**

#### **Diabetes mellitus**

[**Diabetes mellitus**](https://en.wikipedia.org/wiki/Diabetes_mellitus) **increases the risk of stroke by 2 to 3 times.** While intensive blood sugar control has been shown to reduce small blood vessel complications such as [kidney damage](https://en.wikipedia.org/wiki/Diabetic_nephropathy) and [damage to the retina of the eye](https://en.wikipedia.org/wiki/Diabetic_retinopathy) it has not been shown to reduce large blood vessel complications such as stroke.

#### **Diet**

Nutrition, specifically the [Mediterranean-style diet](https://en.wikipedia.org/wiki/Mediterranean_diet), has the potential for decreasing the risk of having a stroke by more than half.

### **Women**

A number of specific recommendations have been made for women including taking aspirin after the 11th week of pregnancy if there is a history of previous chronic high blood pressure and taking blood pressure medications during pregnancy if the blood pressure is greater than 150 mmHg systolic or greater than 100 mmHg diastolic. In those who have previously had [preeclampsia](https://en.wikipedia.org/wiki/Preeclampsia) other risk factors should be treated more aggressively.

### **Previous stroke or TIA**

Keeping blood pressure below 140/90 mmHg is recommended. Anticoagulation can prevent recurrent ischemic strokes. **Among people with nonvalvular atrial fibrillation, anticoagulation can reduce stroke by 60% while antiplatelet agents can reduce stroke by 20%**. However, a recent meta-analysis suggests harm from anticoagulation started early after an embolic stroke. **Stroke prevention treatment for atrial fibrillation is determined according to the** [**CHA2DS2–VASc score**](https://en.wikipedia.org/wiki/CHA2DS2%E2%80%93VASc_score). The most widely used anticoagulant to prevent thromboembolic stroke in people with nonvalvular atrial fibrillation is the oral agent [warfarin](https://en.wikipedia.org/wiki/Warfarin) while a number of newer agents including [dabigatran](https://en.wikipedia.org/wiki/Dabigatran) are alternatives which do not require [prothrombin time](https://en.wikipedia.org/wiki/Prothrombin_time) monitoring.

### **Stroke unit**

Ideally, people who have had a stroke are admitted to a "stroke unit", a ward or dedicated area in a hospital staffed by nurses and therapists with experience in stroke treatment. It has been shown that people admitted to a stroke unit have a higher chance of surviving than those admitted elsewhere in hospital, even if they are being cared for by doctors without experience in stroke. [Nursing care](https://en.wikipedia.org/wiki/Nursing_care) is fundamental in maintaining [skin care](https://en.wikipedia.org/wiki/Skin), feeding, hydration, positioning, and monitoring [vital signs](https://en.wikipedia.org/wiki/Vital_signs) such as temperature, pulse, and blood pressure.

### **Rehabilitation**

[Stroke rehabilitation](https://en.wikipedia.org/wiki/Stroke_rehabilitation) is the process by which those with disabling strokes undergo treatment to help them return to normal life as much as possible by regaining and relearning the skills of everyday living. It also aims to help the survivor understand and adapt to difficulties, prevent secondary complications, and educate family members to play a supporting role. Stroke rehabilitation should begin almost immediately with a multidisciplinary approach. The rehabilitation team may involve physicians trained in rehabilitation medicine, [neurologists](https://en.wikipedia.org/wiki/Neurologist), [clinical pharmacists](https://en.wikipedia.org/wiki/Clinical_pharmacist), nursing staff, [physiotherapists](https://en.wikipedia.org/wiki/Physiotherapist), [occupational therapists](https://en.wikipedia.org/wiki/Occupational_therapist), [speech-language pathologists](https://en.wikipedia.org/wiki/Speech-language_pathology), and [orthotists](https://en.wikipedia.org/wiki/Orthotist). Some teams may also include [psychologists](https://en.wikipedia.org/wiki/Psychologists) and [social workers](https://en.wikipedia.org/wiki/Social_work), since at least one-third of affected people manifests [post stroke depression](https://en.wikipedia.org/wiki/Post_stroke_depression). Validated instruments such as the [Barthel scale](https://en.wikipedia.org/wiki/Barthel_scale) may be used to assess the likelihood of a person who has had a stroke being able to manage at home with or without support subsequent to discharge from a hospital.

[Stroke rehabilitation](https://en.wikipedia.org/wiki/Stroke_rehabilitation) should be started as quickly as possible and can last anywhere from a few days to over a year. Most return of function is seen in the first few months, and then improvement falls off with the "window" considered officially by [U.S. state](https://en.wikipedia.org/wiki/U.S._state) rehabilitation units and others to be closed after six months, with little chance of further improvement. However, some people have reported that they continue to improve for years, regaining and strengthening abilities like writing, walking, running, and talking. Daily rehabilitation exercises should continue to be part of the daily routine for people who have had a stroke. Complete recovery is unusual but not impossible and most people will improve to some extent: proper diet and exercise are known to help the brain to recover.

#### **Physical and occupational therapy**

Physical and occupational therapy have overlapping areas of expertise; however, physical therapy focuses on joint range of motion and strength by performing exercises and relearning functional tasks such as bed mobility, transferring, walking and other gross motor functions. Physiotherapists can also work with people who have had a stroke to improve awareness and use of the [hemiplegic](https://en.wikipedia.org/wiki/Hemiplegic) side. Rehabilitation involves working on the ability to produce strong movements or the ability to perform tasks using normal patterns. Emphasis is often concentrated on functional tasks and people's goals. One example physiotherapists employ to promote [motor learning](https://en.wikipedia.org/wiki/Motor_learning) involves [constraint-induced movement therapy](https://en.wikipedia.org/wiki/Constraint-induced_movement_therapy). Through continuous practice the person relearns to use and adapt the hemiplegic limb during functional activities to create lasting permanent changes. Physical therapy is effective for recovery of function and mobility after stroke. Occupational therapy is involved in training to help relearn everyday activities known as the [activities of daily living](https://en.wikipedia.org/wiki/Activities_of_daily_living) (ADLs) such as eating, drinking, dressing, bathing, cooking, [reading](https://en.wikipedia.org/wiki/Alexia_(condition)) and [writing](https://en.wikipedia.org/wiki/Agraphia), and toileting. Approaches to helping people with urinary incontinence include physical therapy, cognitive therapy, and specialized interventions with experienced medical professionals, however, it is not clear how effective these approaches are at improving urinary incontinence following a stroke.

Treatment of spasticity related to stroke often involves early mobilizations, commonly performed by a physiotherapist, combined with elongation of spastic muscles and sustained stretching through various different positions. Gaining initial improvement in range of motion is often achieved through rhythmic rotational patterns associated with the affected limb. After full range has been achieved by the therapist, the limb should be positioned in the lengthened positions to prevent against further contractures, skin breakdown, and disuse of the limb with the use of splints or other tools to stabilize the joint. Cold in the form of ice wraps or ice packs have been proven to briefly reduce spasticity by temporarily dampening neural firing rates. [Electrical stimulation](https://en.wikipedia.org/wiki/Sensory_stimulation_therapy#Coactivation) to the antagonist muscles or vibrations has also been used with some success. Physical therapy is sometimes suggested for people who experience sexual dysfunction following a stroke.

#### **Speech and language therapy**

[Speech and language therapy](https://en.wikipedia.org/wiki/Speech_and_language_therapy) is appropriate for people with the speech production disorders: [dysarthria](https://en.wikipedia.org/wiki/Dysarthria) and [apraxia of speech](https://en.wikipedia.org/wiki/Apraxia_of_speech), [aphasia](https://en.wikipedia.org/wiki/Aphasia), cognitive-communication impairments, and [problems with swallowing](https://en.wikipedia.org/wiki/Dysphagia). Speech and language therapy for aphasia following stroke compared to no therapy improves functional communication, reading, writing and expressive language. There may be benefit in high intensity and high doses over a longer period, but these higher intensity doses may not be acceptable to everyone.

People who have had a stroke may have particular problems, such as [dysphagia](https://en.wikipedia.org/wiki/Dysphagia), which can cause swallowed material to pass into the lungs and cause [aspiration pneumonia](https://en.wikipedia.org/wiki/Aspiration_pneumonia). The condition may improve with time, but in the interim, a [nasogastric tube](https://en.wikipedia.org/wiki/Nasogastric_intubation) may be inserted, enabling liquid food to be given directly into the stomach. If swallowing is still deemed unsafe, then a [percutaneous endoscopic gastrostomy](https://en.wikipedia.org/wiki/Percutaneous_endoscopic_gastrostomy) (PEG) tube is passed and this can remain indefinitely. Swallowing therapy has mixed results as of 2018.

#### **Devices**

Often, [assistive technology](https://en.wikipedia.org/wiki/Assistive_technology) such as [wheelchairs](https://en.wikipedia.org/wiki/Wheelchairs), walkers and canes may be beneficial. Many mobility problems can be improved by the use of [ankle foot orthoses](https://en.wikipedia.org/wiki/Orthotics#Ankle-foot_orthosis_(AFO)).

#### **Physical fitness**

A stroke can also reduce people's general fitness. Reduced fitness can reduce capacity for rehabilitation as well as general health. Physical exercises as part of a rehabilitation program following a stroke appear safe. Cardiorespiratory fitness training that involves walking in rehabilitation can improve speed, tolerance and independence during walking, and may improve balance. There are inadequate long-term data about the effects of exercise and training on death, dependence and disability after a stroke. The future areas of research may concentrate on the optimal exercise prescription and long-term health benefits of exercise. The effect of physical training on cognition also may be studied further.

The ability to walk independently in their community, indoors or outdoors, is important following stroke. Although no negative effects have been reported, it is unclear if outcomes can improve with these walking programs when compared to usual treatment.

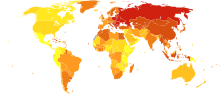
## **Prognosis**

Disability affects 75% of stroke survivors enough to decrease their ability to work. Stroke can affect people physically, mentally, emotionally, or a combination of the three. The results of stroke vary widely depending on size and location of the lesion.

### **Physical effects**

Some of the physical disabilities that can result from stroke include muscle weakness, numbness, [pressure sores](https://en.wikipedia.org/wiki/Pressure_sore), [pneumonia](https://en.wikipedia.org/wiki/Pneumonia), [incontinence](https://en.wikipedia.org/wiki/Urinary_incontinence), [apraxia](https://en.wikipedia.org/wiki/Apraxia) (inability to perform learned movements), difficulties carrying out [daily activities](https://en.wikipedia.org/wiki/Activities_of_daily_living), appetite loss, [speech loss](https://en.wikipedia.org/wiki/Aphasia), [vision loss](https://en.wikipedia.org/wiki/Vision_loss) and [pain](https://en.wikipedia.org/wiki/Pain). If the stroke is severe enough, or in a certain location such as parts of the brainstem, [coma](https://en.wikipedia.org/wiki/Coma) or death can result. Up to 10% of people following a stroke develop [seizures](https://en.wikipedia.org/wiki/Seizure), most commonly in the week subsequent to the event; the severity of the stroke increases the likelihood of a seizure. An estimated 15% of people experience urinary incontinence for more than a year following a stroke. 50% of people have a decline in sexual function ([sexual dysfunction](https://en.wikipedia.org/wiki/Sexual_dysfunction)) following a stroke.

## **Epidemiology**

[](https://en.wikipedia.org/wiki/File:Stroke_world_map-Deaths_per_million_persons-WHO2012.svg)

Stroke deaths per million persons in 2012

  58–316

  317–417

  418–466

  467–518

  519–575

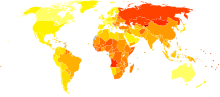
  576–640

  641–771

  772–974

  975-1,683

  1,684–3,477

[](https://en.wikipedia.org/wiki/File:Cerebrovascular_disease_world_map_-_DALY_-_WHO2004.svg)

[Disability-adjusted life year](https://en.wikipedia.org/wiki/Disability-adjusted_life_year) for cerebral vascular disease per 100,000 inhabitants in 2004.[[208]](https://en.wikipedia.org/wiki/Stroke#cite_note-208)

|  |  |
| --- | --- |
| no data    <250    250–425    425–600    600–775    775–950    950–1125 | 1125–1300    1300–1475    1475–1650    1650–1825    1825–2000    >2000 |

Stroke was the second most frequent cause of death worldwide in 2011, accounting for 6.2 million deaths (~11% of the total). Approximately 17 million people had a stroke in 2010 and 33 million people have previously had a stroke and were still alive. Between 1990 and 2010 the number of strokes decreased by approximately 10% in the developed world and increased by 10% in the developing world. Overall, two-thirds of strokes occurred in those over 65 years old. South Asians are at particularly high risk of stroke, accounting for 40% of global stroke deaths.

**The risk of stroke** [**increases exponentially**](https://en.wikipedia.org/wiki/Exponential_growth) **from 30 years of age, and the cause varies by age.** Advanced age is one of the most significant stroke risk factors. **95% of strokes occur in people age 45 and older, and two-thirds of strokes occur in those over the age of 65**. A person's risk of dying if he or she does have a stroke also increases with age. However, stroke can occur at any age, including in childhood.

Family members may have a genetic tendency for stroke or share a lifestyle that contributes to stroke. Higher levels of [Von Willebrand factor](https://en.wikipedia.org/wiki/Von_Willebrand_factor) are more common amongst people who have had ischemic stroke for the first time. The results of this study found that the only significant genetic factor was the person's [blood type](https://en.wikipedia.org/wiki/Blood_type). Having had a stroke in the past greatly increases one's risk of future strokes.

**Men are 25% more likely to suffer strokes than women, yet 60% of deaths from stroke occur in women**. Since women live longer, they are older on average when they have their strokes and thus more often killed. Some risk factors for stroke apply only to women. Primary among these are pregnancy, childbirth, [menopause](https://en.wikipedia.org/wiki/Menopause), and the treatment thereof ([HRT](https://en.wikipedia.org/wiki/Hormone_replacement_therapy_(menopause))).

**1.2 MOTIVATION**

The stroke prediction model provides the chances of stroke to the user. This is done by collecting the inputs from user from the web platform and use the trained model to predict the chance of stroke to the user. It is easy to use and always available and can be accessed from anywhere and anytime through the web enabled device. Whenever a person feels symptoms of stroke like loss in vision in one eye, inability to move or feel on one side of the body, problem in understanding or speaking, dizziness these could be symptoms of stroke. There are some key risk factors like blood pressure, smoking habits, high blood cholesterol, obesity, diabetes, kidney problem etc. which are often know to user. In this case he/she can do a self-diagnosis test for chances for stroke risk to him/her and then consult a doctor. Timely detection can help to a large extend and can even save life in some instances. The easy to access app will provide ease to user. Also, the project will help us understand the capability and factors useful for machine learning classifiers for predicting such results.

**1.3 PROBLEM STATEMENT**

To compare various machine learning classifiers based on accuracy and select the best one to build a system for predicting stroke risk.

**1.4 SCOPE OF PROJECT**

This project provides user will a easy to access app which they can use to check the chances of stroke to them. The aim of the project is to make such app available to everyone. This project also plans to learn which machine learning classifiers are best suited for this kind of job and understand the reason behind it.

**1.5 ACCESS TO USERS**

The project aims to host the web app or mobile app for the users making it easily available for everyone at all time. The interface aims to be very basic and user friendly. The major objective is this app can be used by users on phone instantaneously and does not require any expertise.

**CHAPTER - 2**

**RELATED WORK**

**2.1 INTRODUCTION**

In this section, we will look into some of the earlier works presented in the direction of this project. There has been an extensive amount of work that has already happened in the field of image recognition and natural language processing. While preparing this project we have gathered ideas from different research materials and models. Here, we are going to enlist them.

**2.2 EXISTING SYSTEM**

Lately a lot of research has been done on automatic image captioning. The research can be briefly categorized into three different categories including the template-based approaches, retrieval-based approaches, and novel image caption generation approaches.

The template-based approach is aimed at generating captions by using fixed templates with a number of blank slots, in which way different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled. For example, Farhadi use a triplet of scene elements to fill the template slots for generating image captions. Li et al extract the phrases related to detected objects, attributes, and their relationships for this purpose. Kulkarni et al. adopt a conditional random field (CRF) method to infer the objects, attributes, and prepositions before filling in the gaps. Template-based methods can generate grammatically correct captions. However, templates are predefined and length of captions cannot be variable.

The retrieval-based approach tries to generate description for an image by selecting the most semantically similar sentences from sentence pool or directly copying sentences from other visually similar images. For example, Gong et al. utilize stacked auxiliary embedding method to generate image descriptions from millions of weakly annotated images. Ordonez find similar images in the Flickr database and return the descriptions of these retrieved images to query based on millions of images and their corresponding descriptions. Sun use semantic similarity and visual similarity scores to cluster similar terms and images together first and then retrieve the caption of target image from captions of similar images in the same cluster. Hodosh establish a ranking-based framework to treat sentence-based image description as the task of ranking a set of captions for each test image. These methods generate general and syntactically correct captions. However, it is difficult for them to generate image-specific and semantically correct captions.

Different from the mentioned two categories, novel caption generation approaches mainly use deep learning and machine learning to generate the new captions. A general implementation of this method is to analyze the visual content of the image first and then generate image captions from the visual content using a language model. For instance, Vinyals et al. use CNN as an encoder for image classification and LSTM as a decoder to generate sentence for the description. The main drawbacks of the work are the quick model overfitting, so they use the heavy and expensive GoogLeNet with 22 hidden layers and the absence of attention layer that significantly improved the description accuracy. Karpathy investigate the possibility of generating an image description in natural language. Their approach uses image datasets and their description in natural language and seeks an intermodal correspondence between words from the description and visual data. The first model aligns the fragments of sentences to the visual areas, then forms a single description by multimodal embedding. This description is treated as learning data for a second model of a recurrent neural network that learned to a generate caption. Xu et al. use a convolutional neural network to extract feature maps and LSTM to describe the input image, by processing already extracted feature maps. The limitation of this work is the using of obsolete and expensive Oxford VGGnet, where the quality of image classification is low in the modern CNN . Some researchers have put their attention on classification as Yu. who propose a SVM classification-based two-side cross-domain algorithm by inferring intrinsic user and item features (CTSIF-SVMs), a two-side cross-domain algorithm with expanding user and item features via the latent factor space of auxiliary domains (TSEUIF).

**2.3 PROPOSED SYSTEM**

Our proposed model, for automatic image captioning, is based on ResNet50 and LSTM. The ultimate purpose of this model is to generate the proper description for the given images. To do so, the model is designed with an encoder-decoder architecture based on CNN and RNN. In particular, to extract visual features, we use the ResNet50 network as the encoder to generate a one-dimensional vector representation of the input images. After that, to generate the description sentences, we adopt the LSTM as the language model for the decoder to decode the vector into a sentence.

**2.3.1. Image Feature Extraction**

For image feature extraction we used CNN, RestNet50, which is a deep network that has 50 layers. The image feature extractor needs an image of 224x224x3 size. The model uses ResNet50 pretrained on ImageNet dataset where the features of the image are extracted just before the last layer of classification. Another dense layer is added and converted to get a vector of length 2048.

The model takes an image and produces a caption, encoded as a sequence of 1-K coded words.



where K is the size of the dictionary and *c* is the caption length. We use CNN in particular, ResNet50, to obtain set annotation vectors like the feature vectors. The extractor produces L-vectors, all of which is a D-dimensional representation of the corresponding part of an image.

##### **2.3.2. The Language Model**

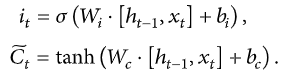
In our design, we adopt LSTM as our language model to generate proper caption based on the input vector from the ResNet50 output.



where the output vector of the previous cell ht-1 with the new element of the sequence xt is concatenated and passed as one vector through the layer with the sigmoid activation function.



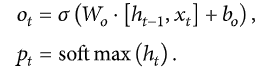
Two created vectors are used to update the state from Ct-1 to Ct. To do this, we multiply the past state by ƒt to “forget” the data recognized as unnecessary in the previous step, then add .



The input gate must determine what values will be updated, and the tanh layer creates a vector of new candidates for  , and values can be added to the cell state.



The obtained values of Ct and ht are transmitted to the neural network input at a time *t*+1.



The multiplicative filters allow to effectively train LSTM, as they are good to prevent the exploding and vanishing gradients. Nonlinearity is provided by the sigmoid 𝝈(.) and the hyperbolic tangent h(). In the last equation, is fed to the softmax function to calculate the probability distribution pt over all the words. This function is calculated and optimized on the entire training dataset. The word with maximum probability is selected at each time step and fed into the next time step input to generate a full sentence.

**2.3.3. Fitting the Model**

After building the model, the model is fit using the training dataset. The model is made to run for 50 epochs and the best model is chosen among the 50 epochs by computing loss function on Flickr8k development dataset. The model with the lowest loss function is chosen for generating captions.

**2.3.4. Caption Generation**

To test our trained model, we input an image to the model. Next the image is fed into the feature extractor to recognize what all objects and scenes are depicted in the image, after resizing it. The process of caption generation is done using the RNN trained model. Then for that image, sequentially, word-by-word the caption is generated by selecting the word with maximum weight for the image at that particular iteration. The indexed word is converted to word and then appended into the final caption. When a tag is detected or the size of the caption reaches 40, the final caption is generated and printed along with the input image.

##### **2.3.5. Suggestion API**

Finally, in the web app, the description string is then passed to API which will suggest a few caption options to the user for selecting their preferred one.

**CHAPTER - 3**

**PROBLEM DESCRIPTION AND SPECIFICATION**

**3.1 PROBLEM DESCRIPTION**

The problem introduces a captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The image captioning task generalizes object detection when the descriptions consist of a single word.

**3.2 SPECIFICATION**

Image caption generator is a task that involves computer vision and natural language processing concepts to recognize the context of an image and describe them in a natural language like English.

The aim of image captioning is to automatically describe an image with one or more natural language sentences. This is a problem that integrates computer vision and natural language processing, so its main challenges arise from the need of translating between two distinct, but usually paired modalities. First, it is necessary to detect objects on the scene and determine the relationships between them and then, express the image content correctly with properly formed sentences. The generated description is still much different from the way people describe images because people rely on common sense and experience, point out important details and ignore objects and relationships that they imply. Moreover, they often use imagination to make descriptions vivid and interesting. The process flow diagram below in Figure 1 outlines the working of the project.

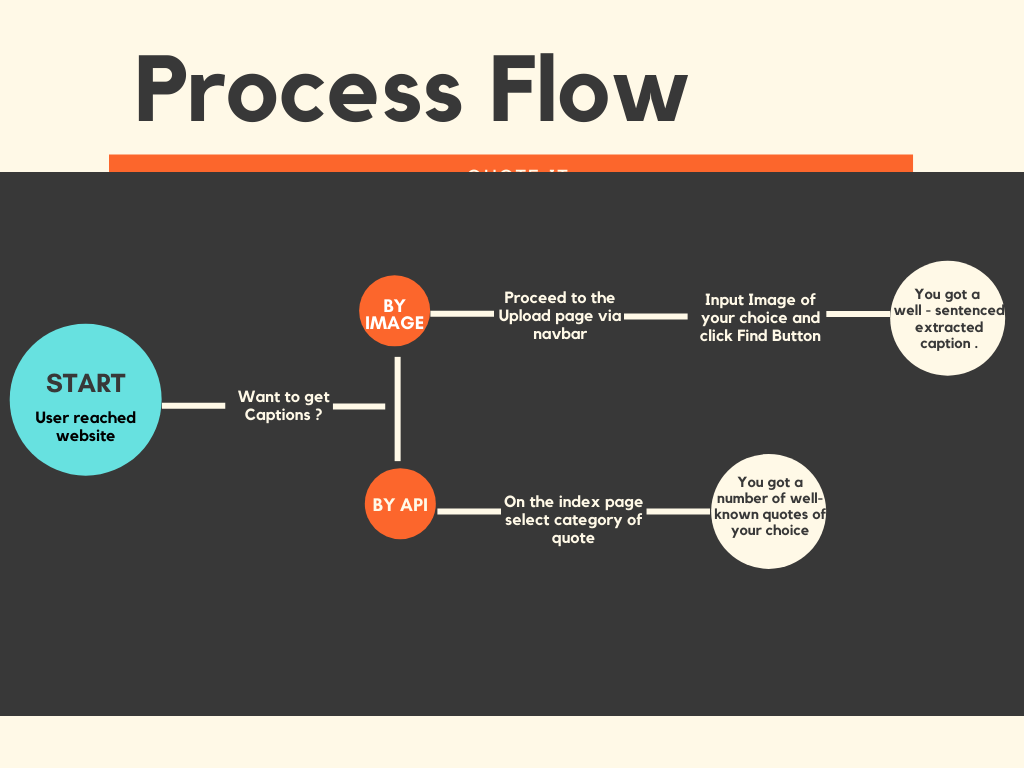


Figure 1: Process Flow Diagram

**3.3 REQUIREMENTS**

**3.3.1 SOFTWARE REQUIREMENTS**

* Python Version 3.6+
* IDE: JupyterLab or PyCharm
* Frameworks:-
  + Flask
  + Bootstrap
* Python Libraries:-
  + Scikit-learn
  + Seaborn
  + Numpy: 1.18.5
  + Pandas: 1.1.14

**3.3.2 HARDWARE REQUIREMENTS**

* Processor:- Intel i3+ or AMD A6+
* RAM:- 2GB/+
* Hard Disk Space:- 2GB

**3.3.3 OTHER REQUIREMENTS**

* Stroke Prediction Dataset
* For training the model - Kaggle Notebook/Google Colab/JupyterLab should be preferred (Recommended)

**CHAPTER - 4**

**SYSTEM DESIGN**

**4.1 INTRODUCTION**

According to World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of the total deaths. Through this project we will compare various machine learning algorithms based on their accuracy of stroke prediction and build the system for the same.

The model will consist of a front-end GUI for collecting data. Flask file to connect the model to backend model. For the backend model we will be comparing 10 different ML algorithms for their accuracy. Select the best one and perform K-fold Cross validation and hyper-parameter optimization on it.

**4.1.1. Data Sources**

We had selected “Stroke-prediction dataset” obtained from Kaggle as our dataset. This dataset contains of 5110 records having 12 attributes (id, age, gender, hypertension, heart disease, marital status, residence type, work type, average glucose level, BMI, smoking status and stroke occurrence). The below image provides the info about the dataset.

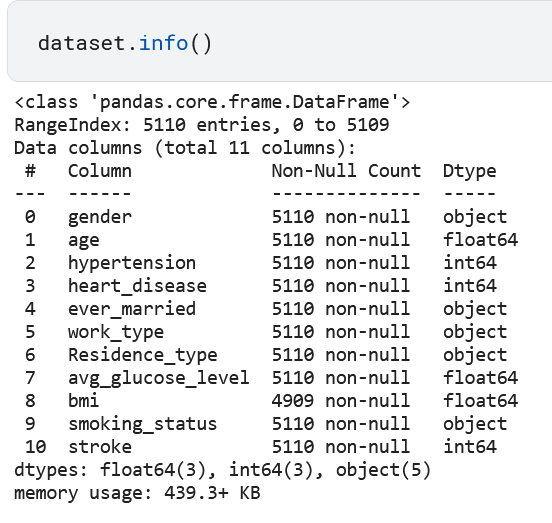


Figure: Output showing the dataset information

**4.1.2. MACHINE LEARNING ALGORITHMS USED IN THE PROJECT**

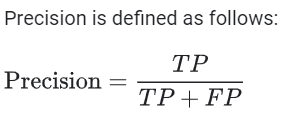
In this project we are going to train our dataset on 10 different machine learning algorithms and based on their performance we will select the best model and further use it in building the stroke prediction system. The 10 machine learning algorithms used in this project are: -

1. Gaussian Naive Bayes
2. Bernoulli naïve Bayes
3. Logistic Regression
4. Support Vector Machines
5. Random Forest
6. Decision Tree classifier
7. K-Nearest Neighbor
8. Gradient Boosting
9. Stochastic Gradient Descent
10. Neural nets

**4.1.3. MODEL PARAMETERS**

A model parameter is a configuration variable that is internal to the model and whose value can be estimated from data. The various parameters we computed for the machine learning algorithms we covered are: -

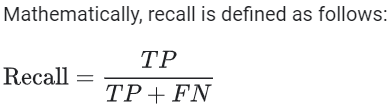
* 1. Accuracy- The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points.
  2. Precision- Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances.



Where, TP is True positives

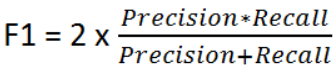
FP is False Negatives

* 1. Recall- Recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.



* 1. F1 Score- In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test.

The formula for F1 Score is-



* 1. Specificity- Specificity (True Negative rate) measures the proportion of negatives that are correctly identified (i.e., the proportion of those who do not have the condition (unaffected) who are correctly identified as not having the condition).

Mathematically specificity is given by-



**4.2 ARCHITECTURE DIAGRAM**

The stroke predicition system using machine learning predicts the risk of stroke to the user based on the various symptom information the user gives, such as age, gender, marital status, hypertension, prior heart disease and smoking status. The architecture of the stroke risk prediction using machine learning consists of the dataset through which we will compare the symptoms of the user. Model is trained using the dataset. The classification algorithms process the data and predicts the risk of stroke.

The diagram below explains about the system in perception of overview of the system.

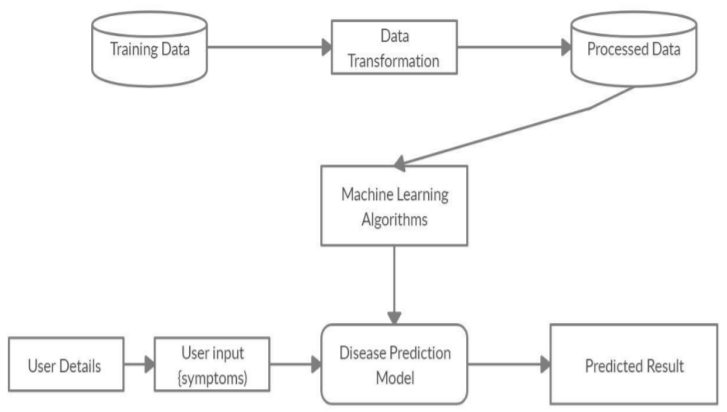
****

Figure 2: Architecture of the Stroke prediction model

**4.3 UNIFIED MODELING LANGUAGE(UML)**

**4.3.1 USE CASE DIAGRAM**

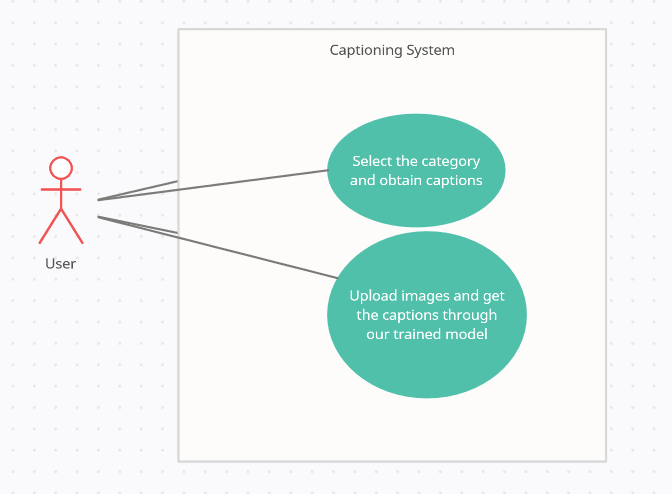
****

Figure 4: Use Case Diagram for the captioning system

Figure 4 shows that there are basically two use-cases available for the user. When the user visits the web site they can choose:

1. Any category available on the home page to obtain the collection of images of that category and then select and copy the caption of their choice.
2. User can navigate to the Upload section and then upload the image from their device and then by the help of the image caption generator model running in the background he/she can obtain the caption for the provided image.

**4.3.2 DATA FLOW DIAGRAM**

**4.3.2.1 Data Flow Diagram Level 0**

Figure 5: DFD level 0

In Figure 5 the level 0 DFD shows the top-level view of the data flow. It shows that initially when the user interacts with the GUI, they have two option. The first one to choose a caption of their favourite category and the second option is to provide an image to the model for generation caption.

**4.3.2.2 Data Flow Diagram Level 1**

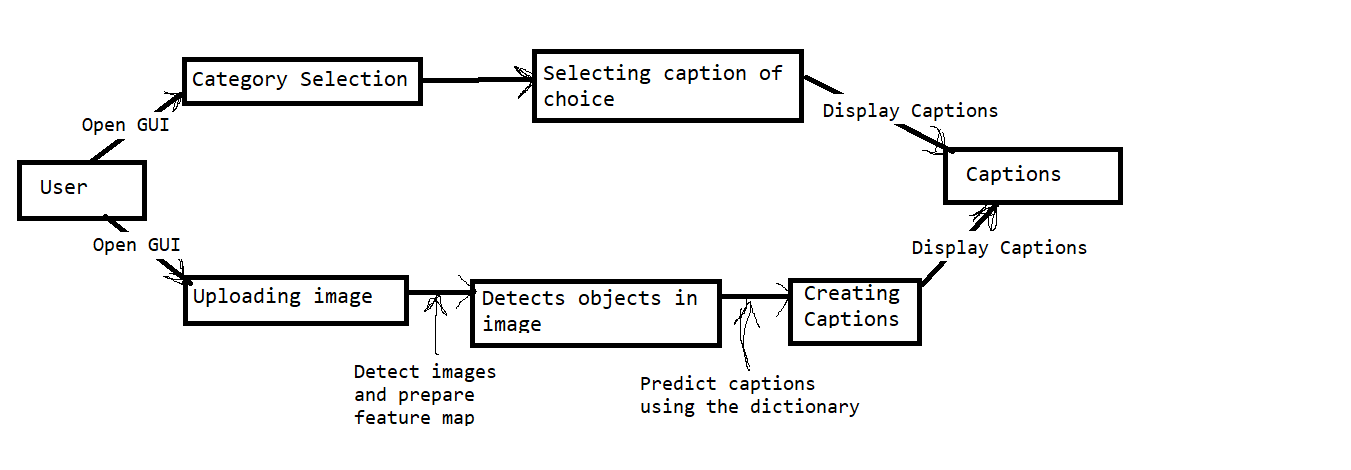
****

Figure 6: DFD level 1

In figure 6 we see the medium level view of data flow. It shows how the input passed goes through a channel of different segments. In our case when a user uploads the image it passes through the CNN model and the features of the image generated are then passed to Language model which gives the captions to the GUI for display.

**4.3.2.3 Data Flow Diagram Level 2**

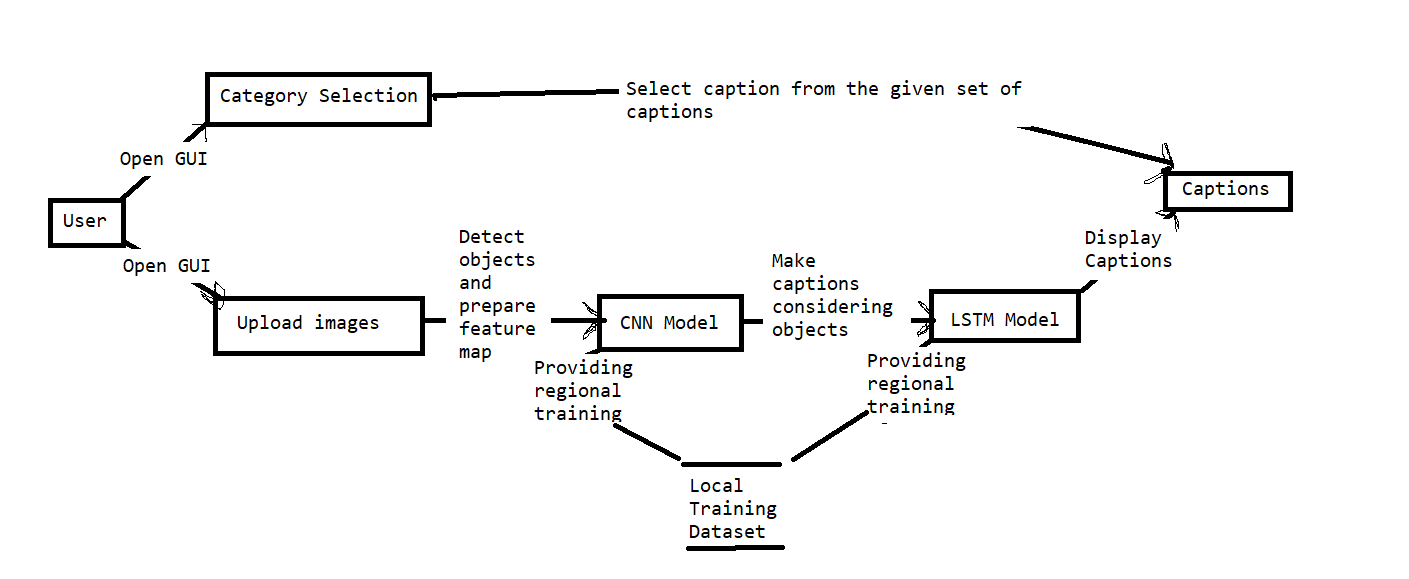
****

Figure 7: DFD level 2

In figure 7 explains the functioning of the model in great depth. The first option remains the same, but the second option shows the usage of a regional dataset to train the model. This improves the functioning of the model by training the model on our custom dataset.

**4.4 STATE CHART DIAGRAM**

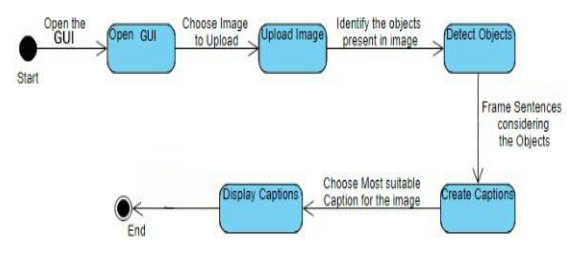
****

Figure 8: State Chart Diagram of the system

Figure 8 shows the state of the program at every deciding state and its output which are fed to the next state as input. The program basically contains five states as stated in the figure. The GUI and uploading of the image are handled by the flask library. The object detection phase uses the CNN model to detect the objects in the figure and then the create caption state uses the LSTM model to generate the captions.

**4.5 FLOW CHART FOR CAPTION GENERATION PROCESS**

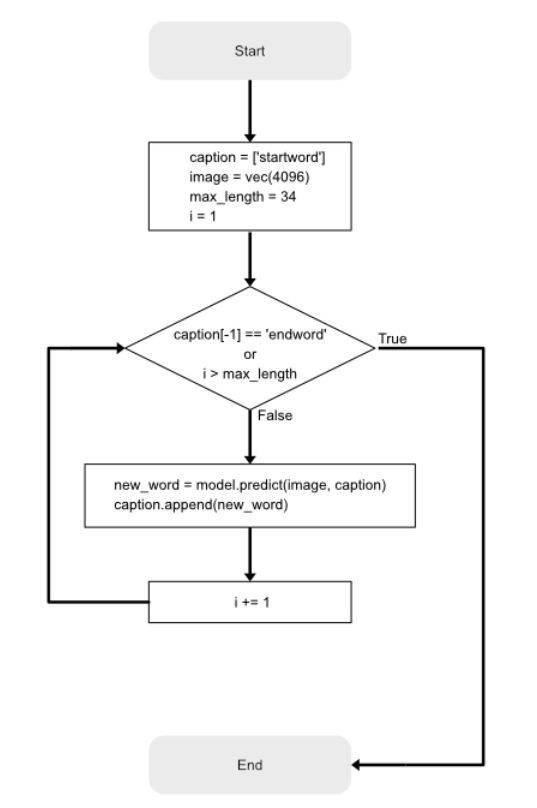
****

Figure 9: Flow chart for the caption generation process

Figure 9 tells us about the functioning of the caption prediction task. Once the image feature of the passed test image is extracted. The trained model tries to predict the next word in the caption by using the Dense layer and agrmax function. The inverse dictionary is used for obtaining the word corresponding to the integer provided. When the “endword” is passed to the model, it stops and the final caption is given in output.

**CHAPTER - 5**

**IMPLEMENTATION AND RESULT**

**5.1 OVERVIEW**

In this section, we are going to cover all the steps we have taken while implementing the program. The whole program consists of different sections giving a brief about the program implementation and their respective outputs.

**5.2 IMPORTING THE NECESSARY LIBRARIES AND DATASET**

…………Write a brief

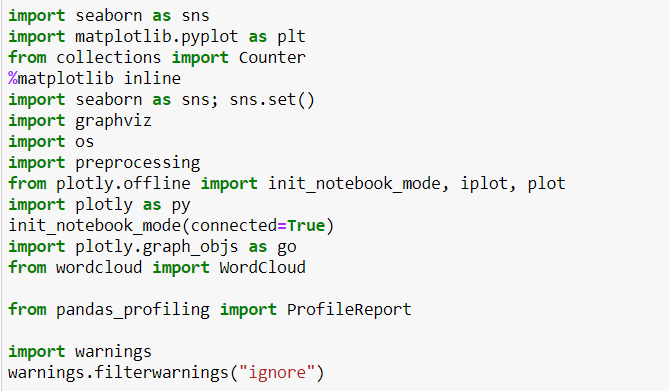


Figure: Code snippet showing the import of libraries used in the project



Figure: Code snippet to read the dataset

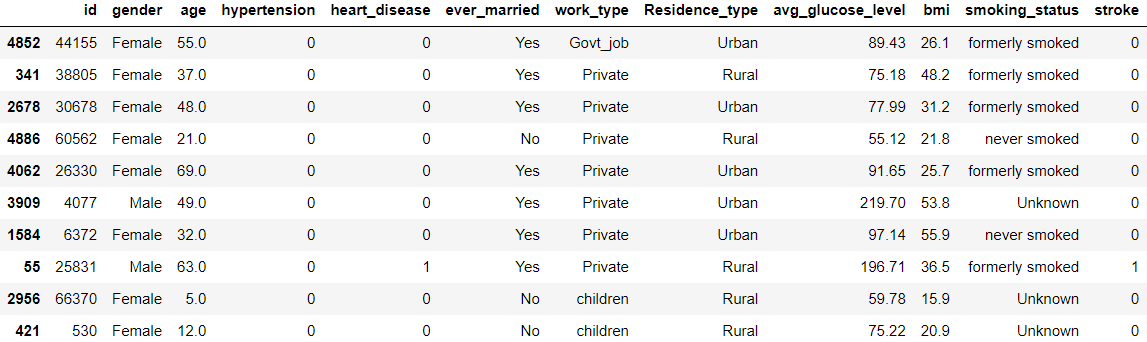


Figure: 10 Sample data from the dataset

We’ll drop the ‘id’ column as it may cause unwanted correlation.

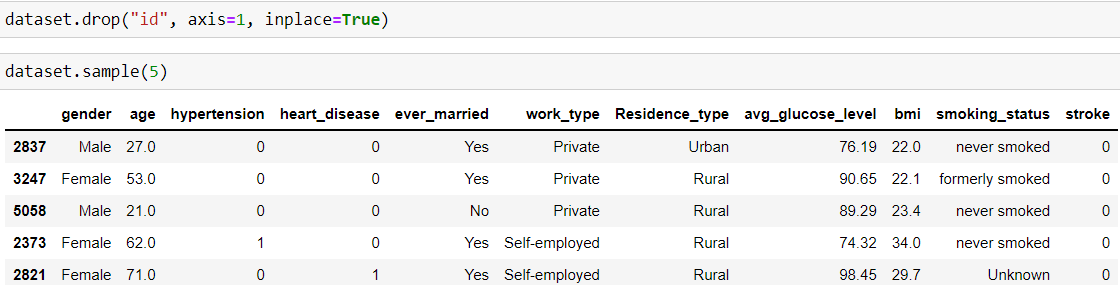


Figure- Code snippet and resultant table after dropping the id column

**5.3 BASIC DATA ANALYSIS**

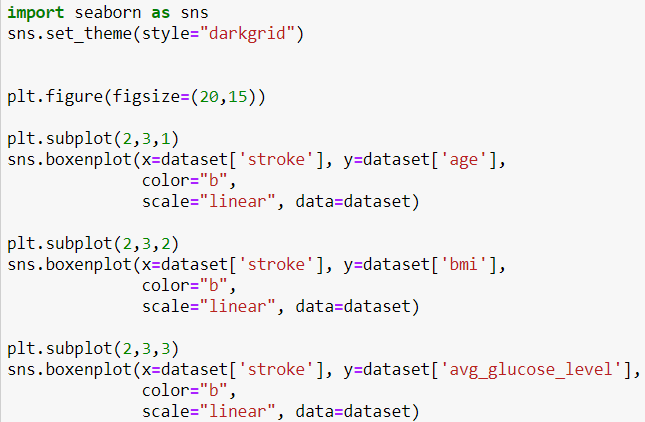
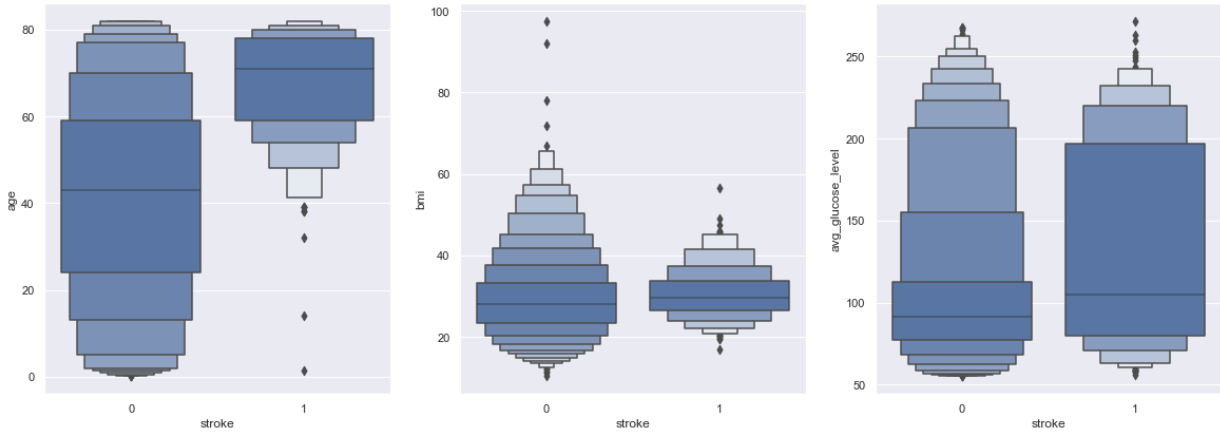




Figure: Code snippet to visualize the dataset



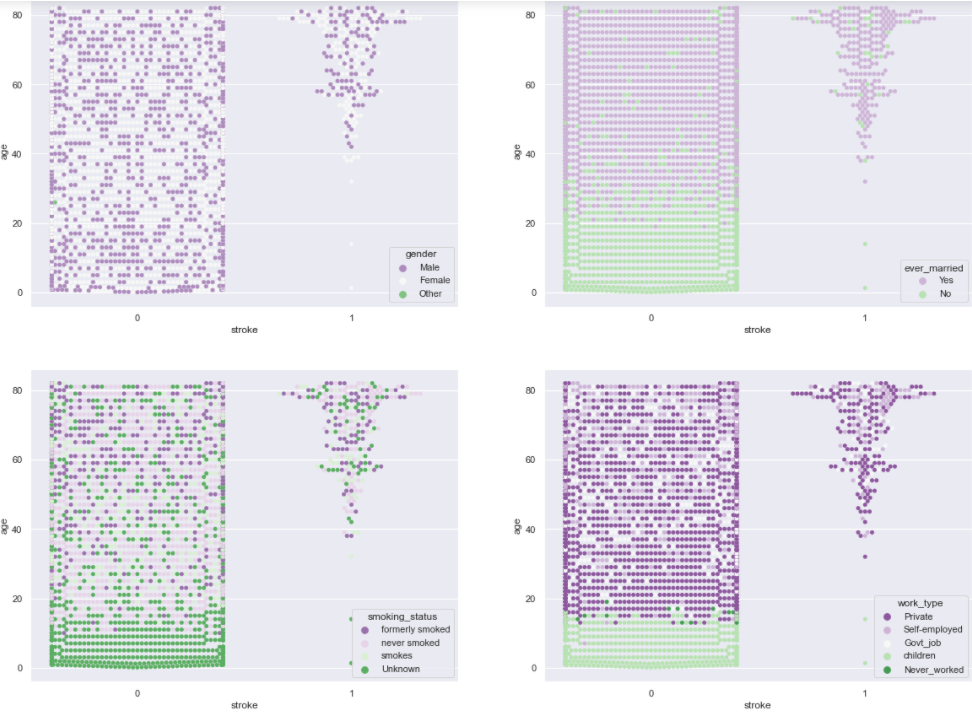


Figure 13: Visual representation of dataset

**5.4 PANDAS PROFILING**

Pandas profiling is a useful library that generates interactive reports about the data. With using this library, we can see types of data, distribution of data and various statistical information. This tool has many features for data preparing. Pandas Profiling includes graphics about specific feature and correlation maps too.



Figure : Code snippet showing pandas profiling

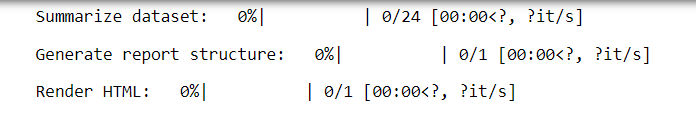


Figure : Pandas profile report

**5.5 CORRELATION**

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

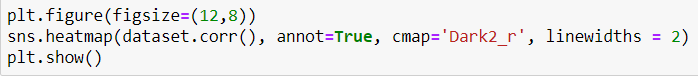


Figure 20: Code snippet to show correlation

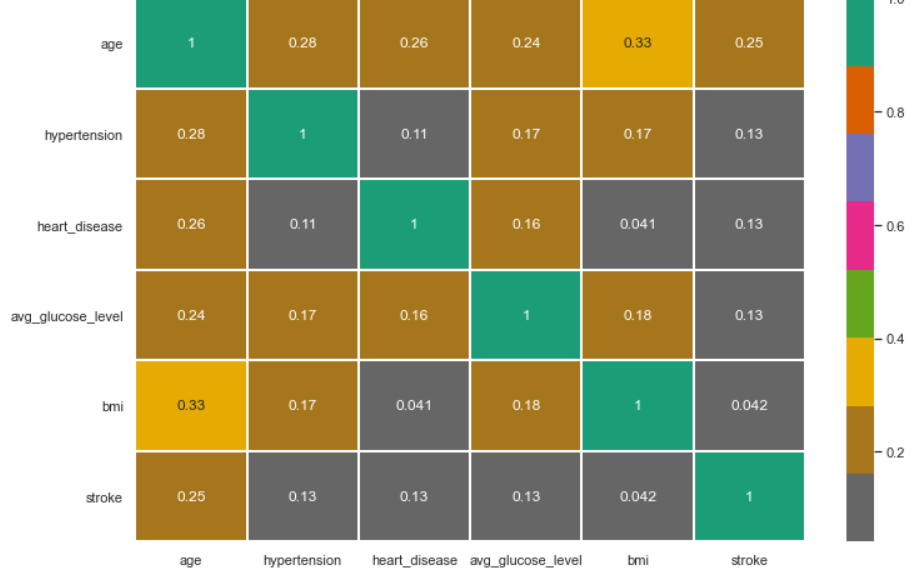
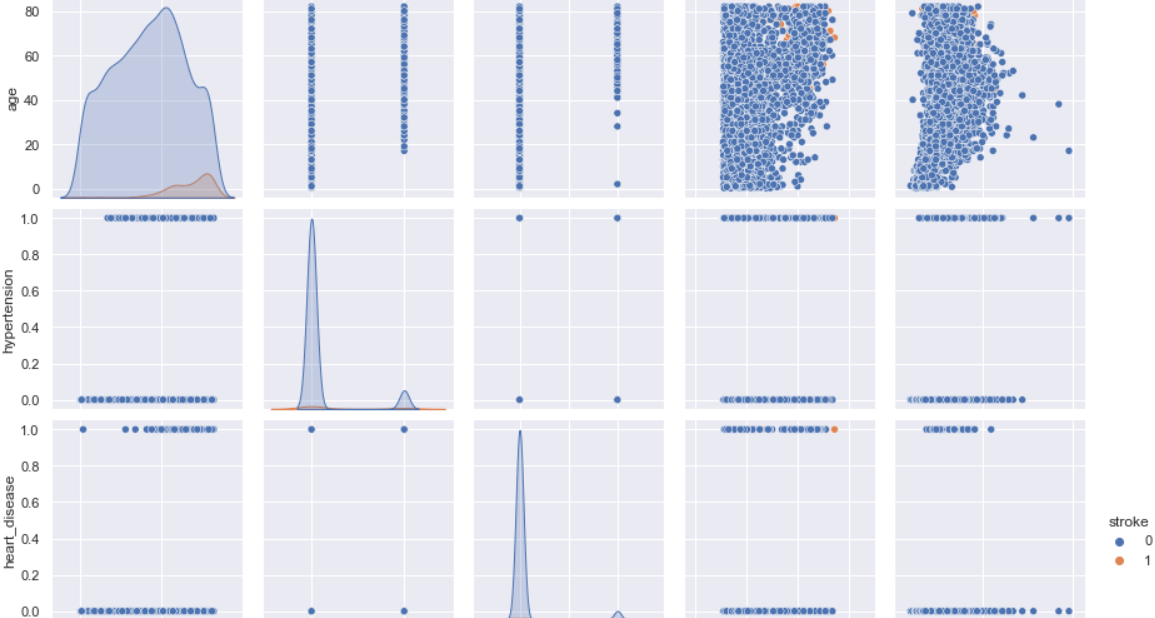


Figure 21: Sample output showing the heatmap of correlation



Figure 22: Code snippet for…..



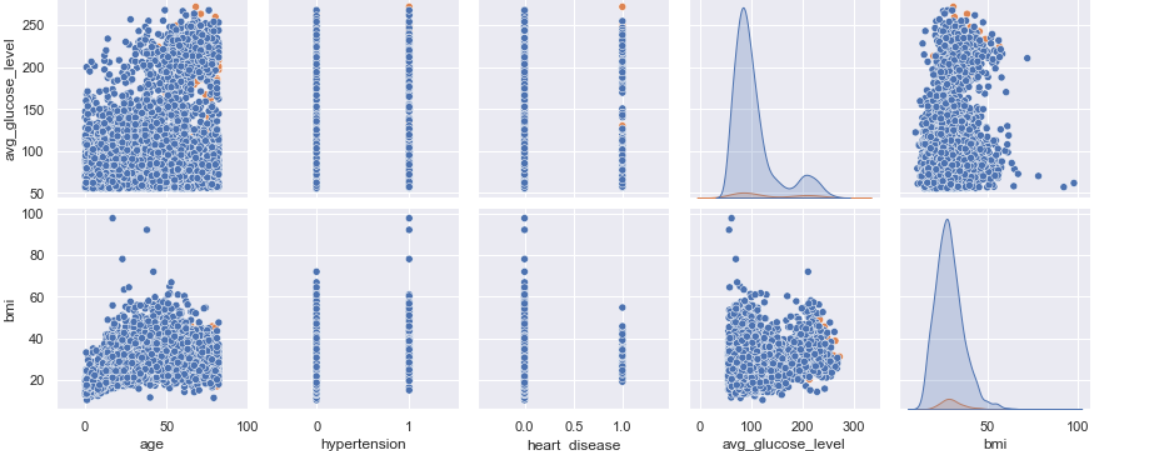


Figure 23: Sample output showing the…..

**5.6 ANAMOLY DETECTION**

Anomaly is one that differs / deviates significantly from other observations in the same sample. An anomaly detection pattern produces two different results. The first is a categorical tag for whether the observation is abnormal or not; the second is a score or trust value. Score carries more information than the label. Because it also tells us how abnormal the observation is. The tag just tells you if it's abnormal. While labeling is more common in supervised methods, the score is more common in unsupervised and semi supervised methods.

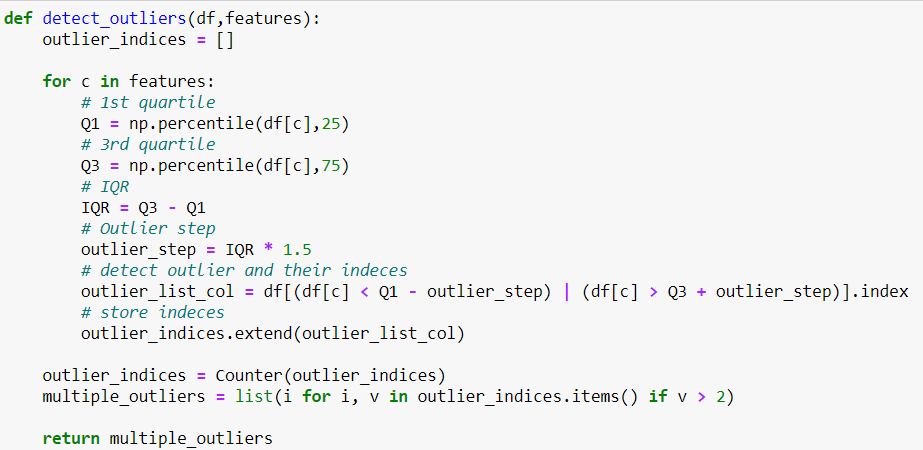


Figure 24: Code snippet showing the detection of outliers

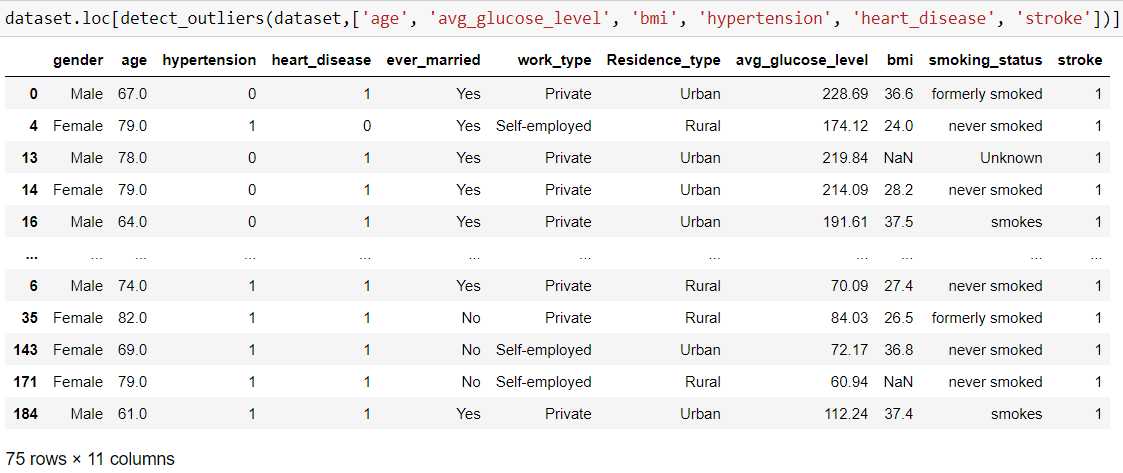


Figure 25: Output showing the detection of outliers

dataset = dataset.drop(detect\_outliers(dataset,['age', 'avg\_glucose\_level', 'bmi', 'hypertension', 'heart\_disease', 'stroke']),axis = 0).reset\_index(drop = True)

Figure 26: Code snippet to drop the dataset containing outliers

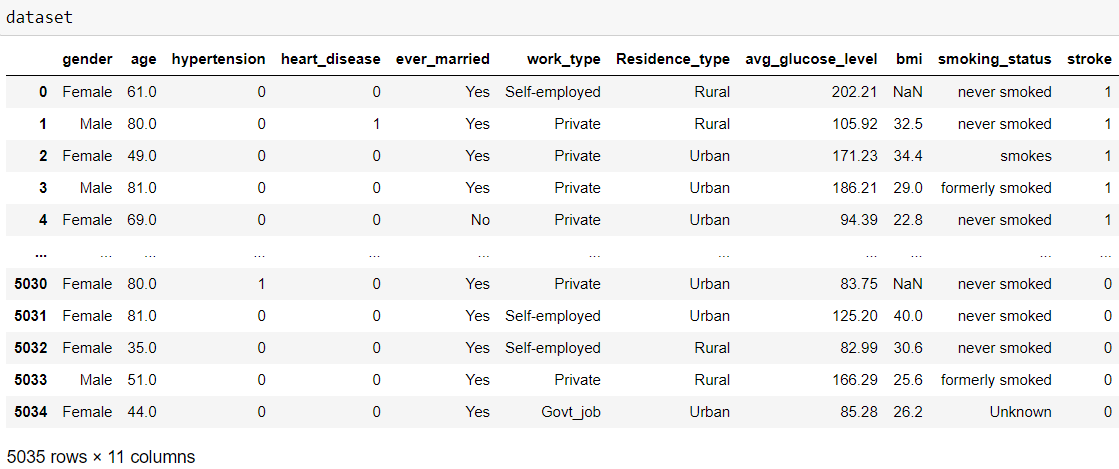


Figure 27: Resultant table after anomaly detection and rectification

**5.7 HANDLING MISSING VALUES**

We have 201 null values in total. After anomaly detection it reduced to 193.



Figure 29: Code snippet to check for missing bmi values.

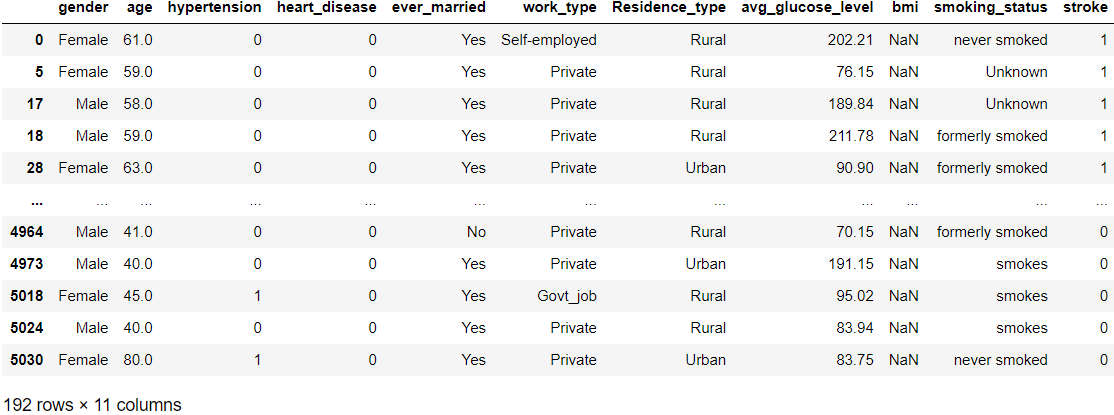


Figure 30: Output showing the rows where BMI value is null.

How to handle missing values?

We selected difference of BMI will be between gender

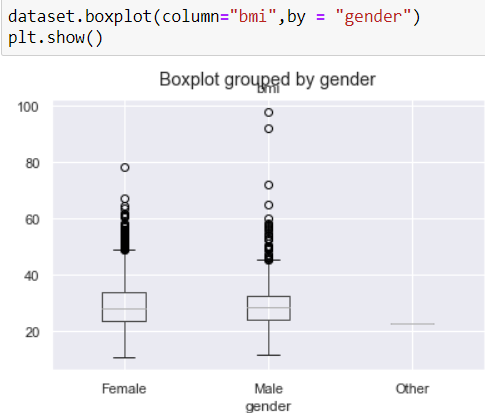


Figure 31: Output showing the boxplot grouped by gender

We get different BMI averages for women and men, although not very large. So we assigned the total BMI mean as there are very few examples for the Others gender.

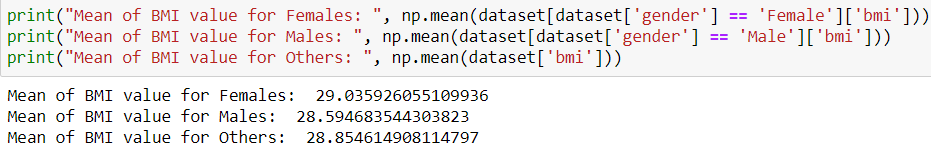


Figure: Code snippet and output showing the mean BMI based on gender

We fill the BMI field with null values to 0.



Figure 32: Code snippet and output showing that none of the field contains null or missing values

**5.8 ENCODING**

****

Figure 34: Code snippet for encoding

First we will handle categorial values.

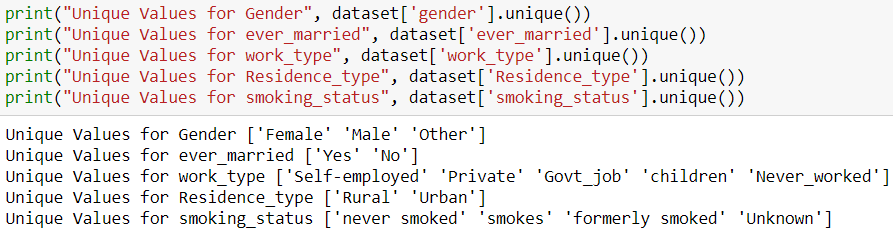


Figure 35: Code snippet and output for handling categorial values

**5.8.1 LABEL ENCODING**

Label Encoding is an encoding technique for handling categorical variables. In this technique, each data is assigned a unique integer.

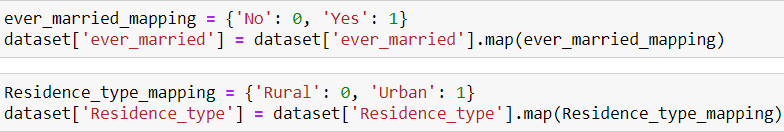


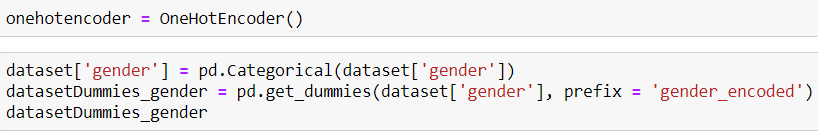
Figure 36: Code snippet to implement label encoding

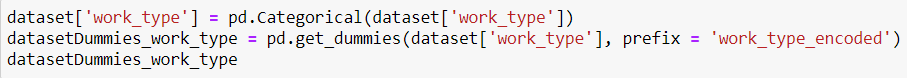
**5.8.2 ONE HOT ENCODING**

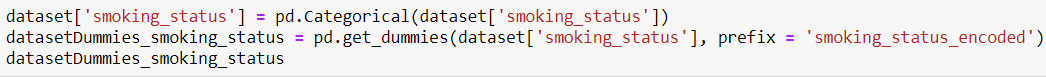
One Hot Encoding is the binary representation of categorical variables. This process requires categorical values to be mapped to integer values first. Next, each integer value is represented as a binary vector with all values zero except the integer index marked with 1.

One Hot Encoding makes the representation of categorical data more expressive and easier. Many machine learning algorithms cannot work directly with categorical data, so categories must be converted to numbers. This operation is required for input and output variables that are categorical.

Here, we converted categorical data to the binary values. This operation increases the accuracy.







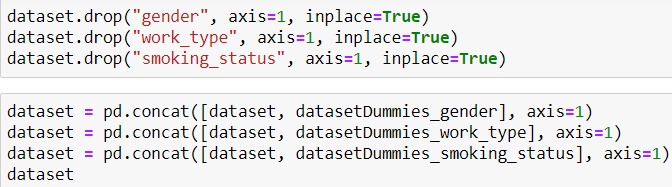


Figure: Code snippet for implementing one hot encoding

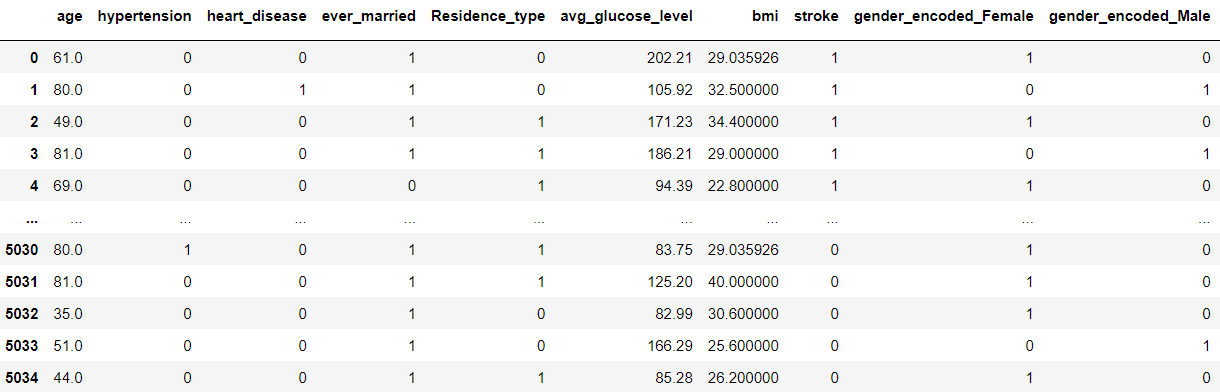


Figure: Sample output showing the table doesn’t contain categorial variables.

Dataset is now ready for implementing machine learning algorithms.

**5.9 TRAINING AND TESTING OF DATA**

Here we split the dataset in the ratio……



Figure 37: Code snippet showing the dataset is split for the purpose of training, testing and validation.

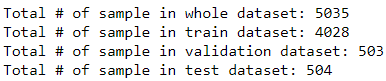


Figure 38: Output showing the amount of data used for training, testing and validation

**5.9.2 STANDARDIZATION**

Standardization is a method in which the mean value is 0 and the standard deviation is 1, and the distribution approaches the normal. The formula is as follows, we subtract the average value from the value we have, then divide it by the variance value.

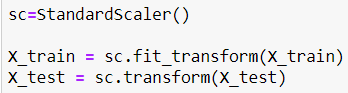


Figure: Code snippet to implement Standardization

**5.9.3 SAVING THE SCALER OBJECT**

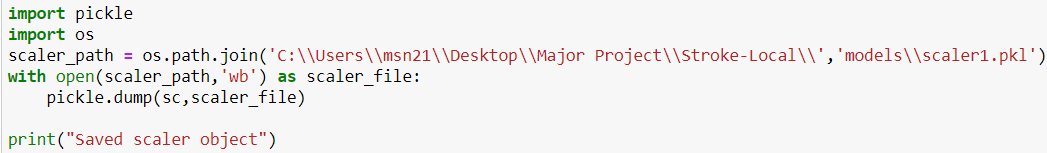


Figure: Code snippet for saving the scaler object

**5.10 COMPUTING TRAINING SCORE OF MODELS**

The following are the ML algorithms that will apply to dataset. Results will contain train-validation-test scores, confusion matrix, statistical information and classification reports for each algorithm.



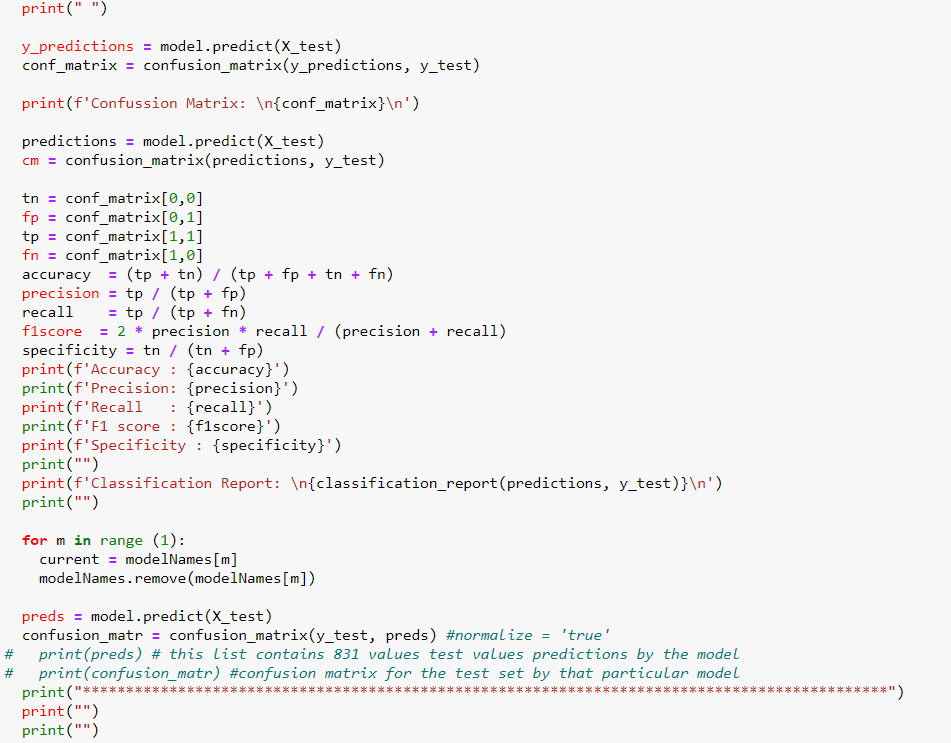
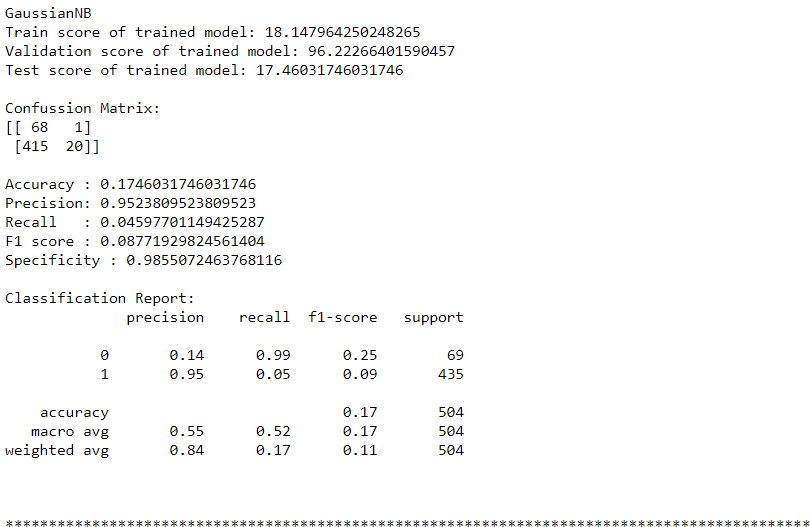
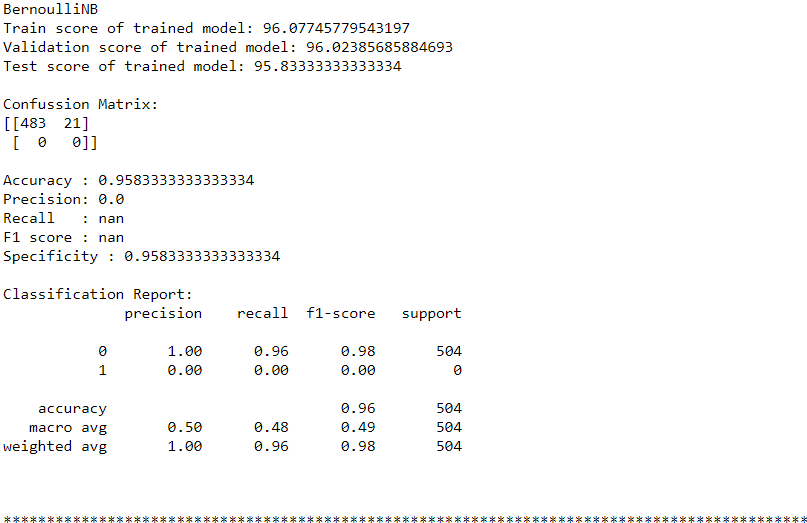
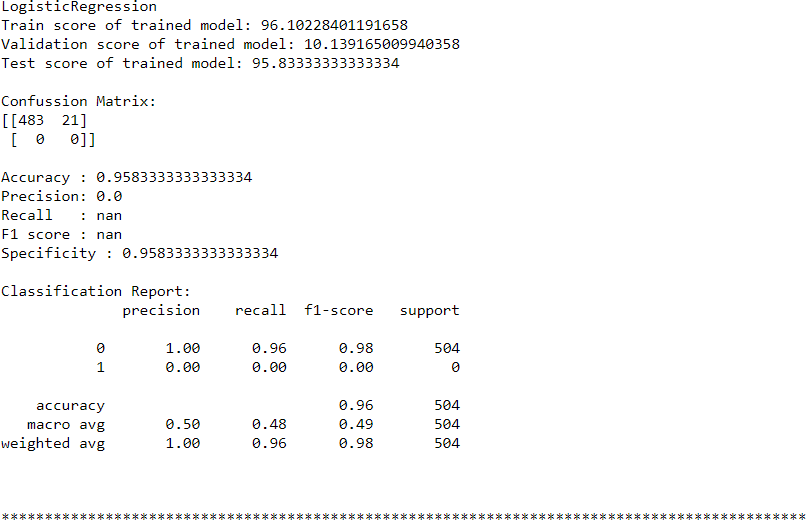
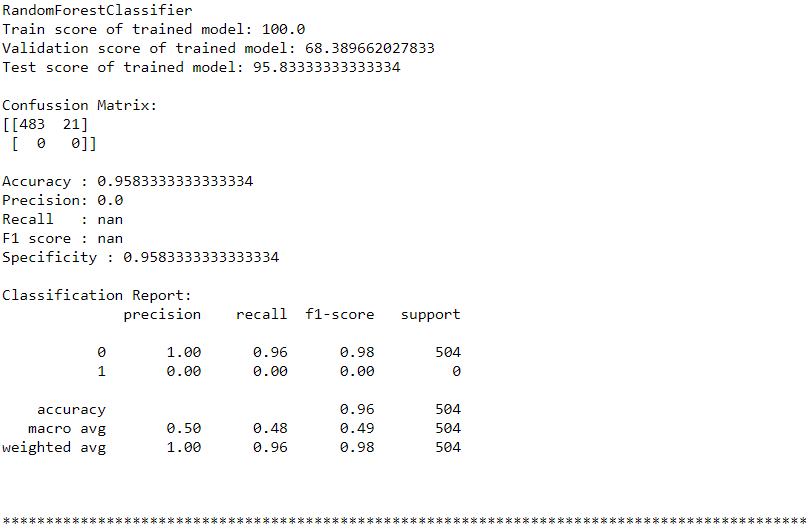


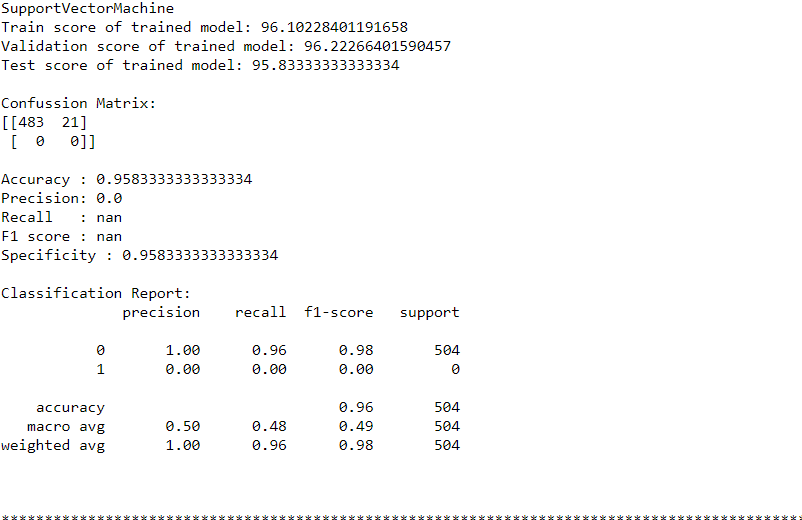
Figure: Code snippet for computing accuracy, precision, recall, F1 score, specificity and confusion matrix of various ML algorithms

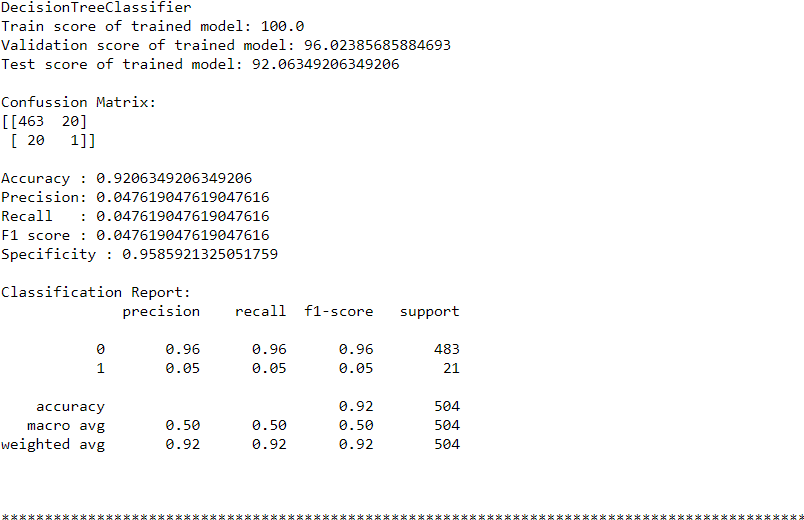


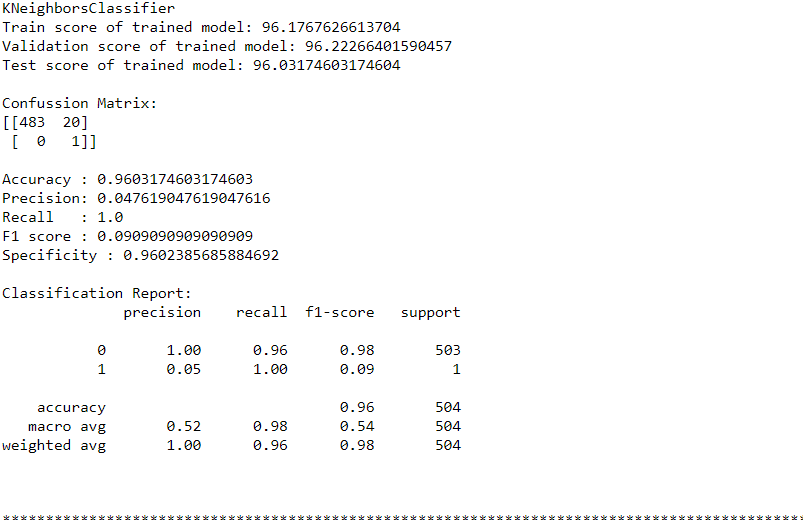


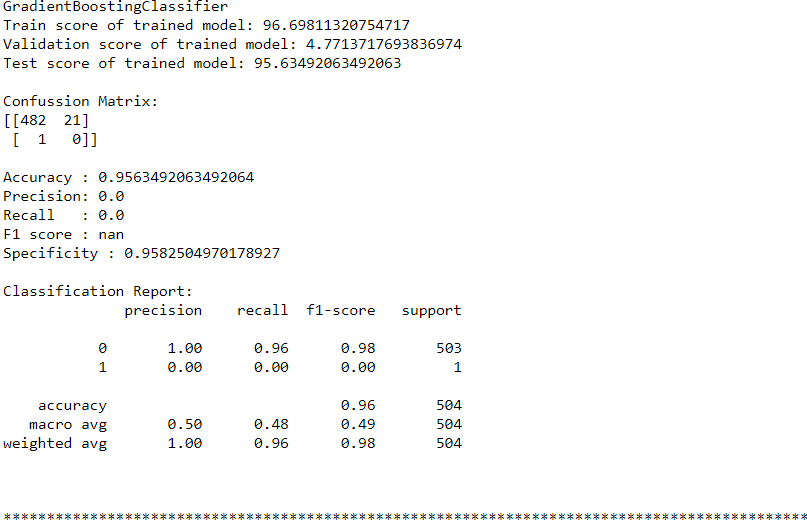


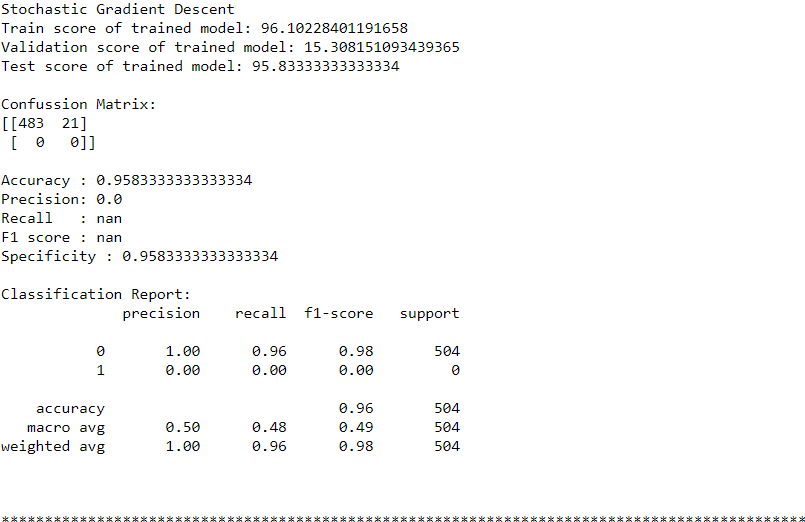












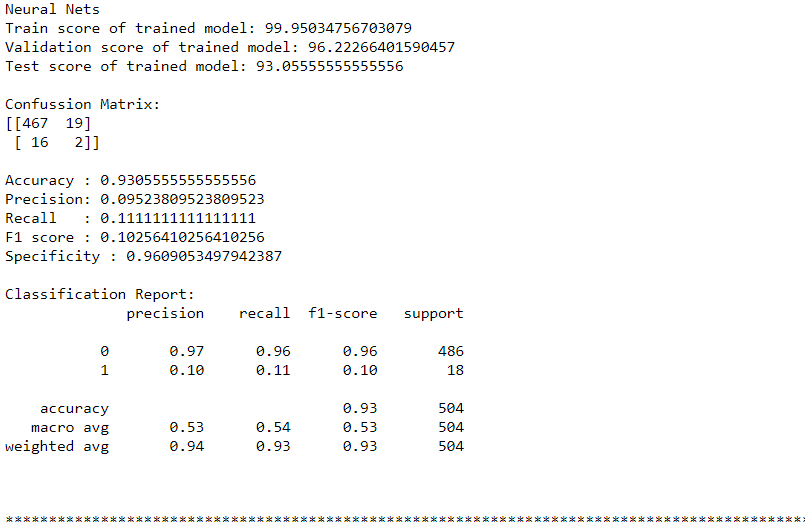


Figure: Output containing accuracy, precision, recall, F1 score, specificity and confusion matrix of various ML algorithms



Figure: Code snippet to graphically visualize the train-validation-test score of models



Figure: Graphical output of train-validation- test score of models

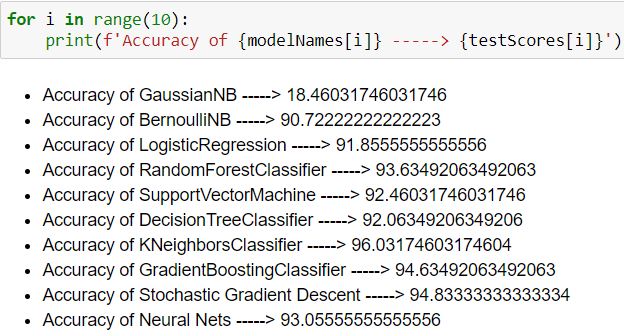


Figure: Code snippet and output showing the accuracy of each model

We observed that KneighborsClassifier has the highest accuracy of 96.03174603174604

**5.11 MODEL SELECTION AND OPTIMIZATION**

We evaluated these models according to their accuracies, precision, recall, F1 score, specificity and created their confusion matrix. Selecting model based on parameter accuracy, Best algorithm is KNN with highest accuracy of 96.03%. Now we will make k-Fold Cross Validation and Hyper-Parameter Optimization for KNN algorithm**.**

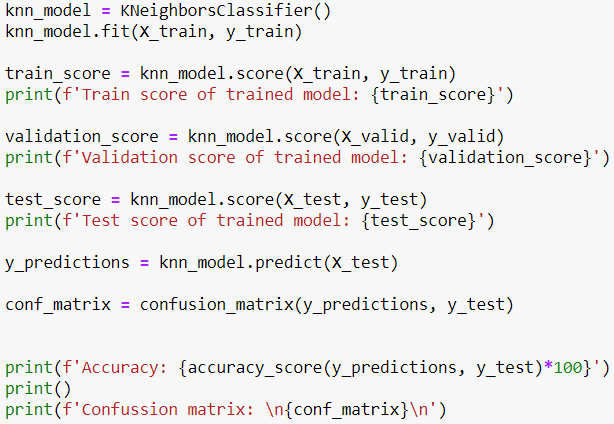
****

Figure: Code snippet for model selection and training

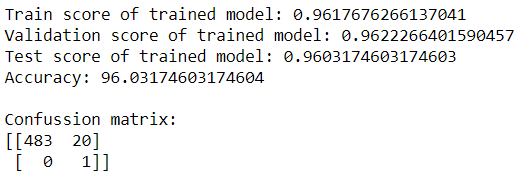
****

Figure: KNeighborsClassifier score along with accuracy and confusion matrix

**5.11.1 SAVING THE BEST MODEL**

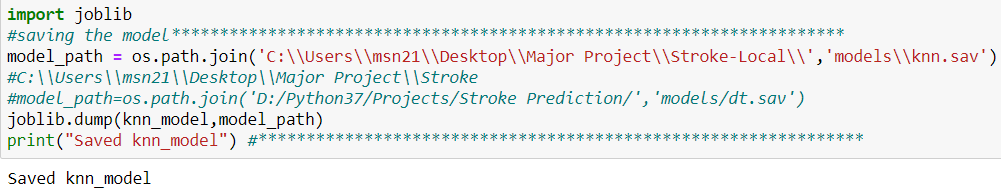


Figure: Code snippet for saving the KNN Model

**5.11.2 MODEL OPTIMIZATION: K-FOLD CROSS VALIDATION**

Cross Validation will enable us to see whether we are facing an overfitting problem and also to see the quality of our model. Thus, it will enable us to test the performance of our model before encountering high error rates in the test data set that we have not seen yet. It is a method that is frequently used because it is easy to apply.

**cv = 10 means k = 10 for KNN.**

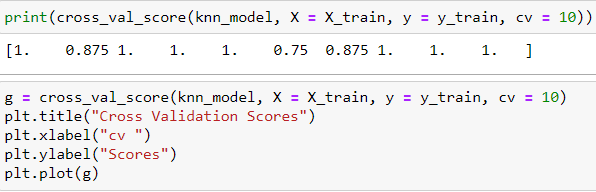


Figure: Code snippet to implement K-Fold Cross validation

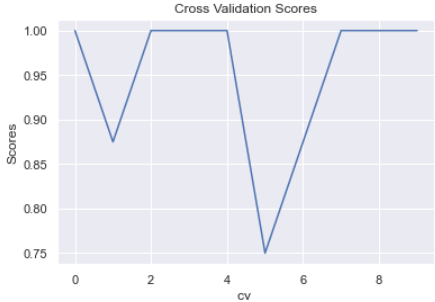


Figure: Graphical output after implementing K-Fold Cross validation

MEAN ABSOLUTE ERROR

In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement.



Figure: Code snippet to compute MAE after applying K-Fold Cross Validation



Figure: Output showing MAE and std

**5.11.3 MODEL OPTIMIZATION: HYPERPARAMETER OPTIMIZATION**

Unlike parameters, hyperparameters are not learned during training the model. They are determined by the data scientist before the modeling phase. For example, KNN algorithm, which is one of the non-parametric classification algorithms, makes classification by looking at the nearest k neighbors to the desired value. Here, the k number (n\_neighbors:) and the distance metric (metric:) to be used are the hyperparameters that should be specified by the data scientist before the modeling, which increases the performance of the model.

**Hyperparameter optimization** is the process of finding the most suitable hyperparameter combination according to the success metric specified for a machine learning algorithm.

Given that there are dozens of hyperparameters for a machine learning algorithm and dozens of values these hyperparameters can take, it's clear how difficult it will be to try all combinations one by one and pick the best combination. For this reason, different methods have been developed for hyperparameter optimization. GridSearcCV and RandomizedSearchCV are among these methods.

### **GridSearchCV:** For the hyperparameters and their values that are desired to be tested in the model, a separate model is established with all combinations and the most successful hyperparameter set is determined according to the specified metric.

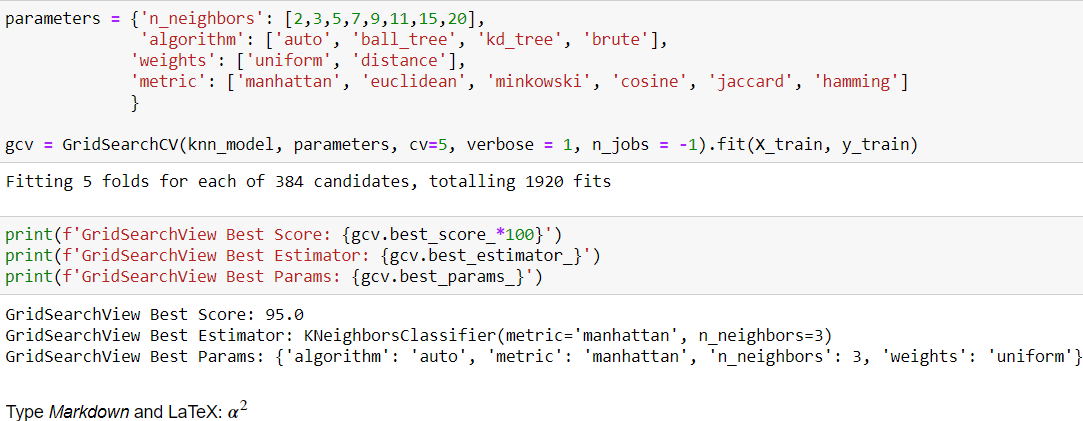


Figure: Code snippet and sample output after applying GridSearchCV

### **RandomizedSearchCV:** A set of hyperparameters is randomly selected and tested by cross-validation and the model set up. These steps continue until the specified calculation time limit or the number of iterations is reached.



Figure: Code snippet and sample output after applying RandomizedSearchCV

5.11.4 BEST FEATURE SELECTION



Figure: Code snippet for best feature selections

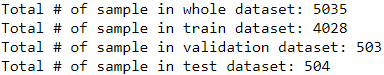


Figure: Sample output

**5.12 WEB PAGE DEVELOPMENT**

For the purpose of providing interfaces to users, we needed to develop a platform. Thus we chose web development for this. In this project, we created a web app using Python and it’s web framework flask, which reduces development time and allows us to build faster and smarter. For the Frontend, we have created a simple webpage which consist of a form which is to be filled by the user.

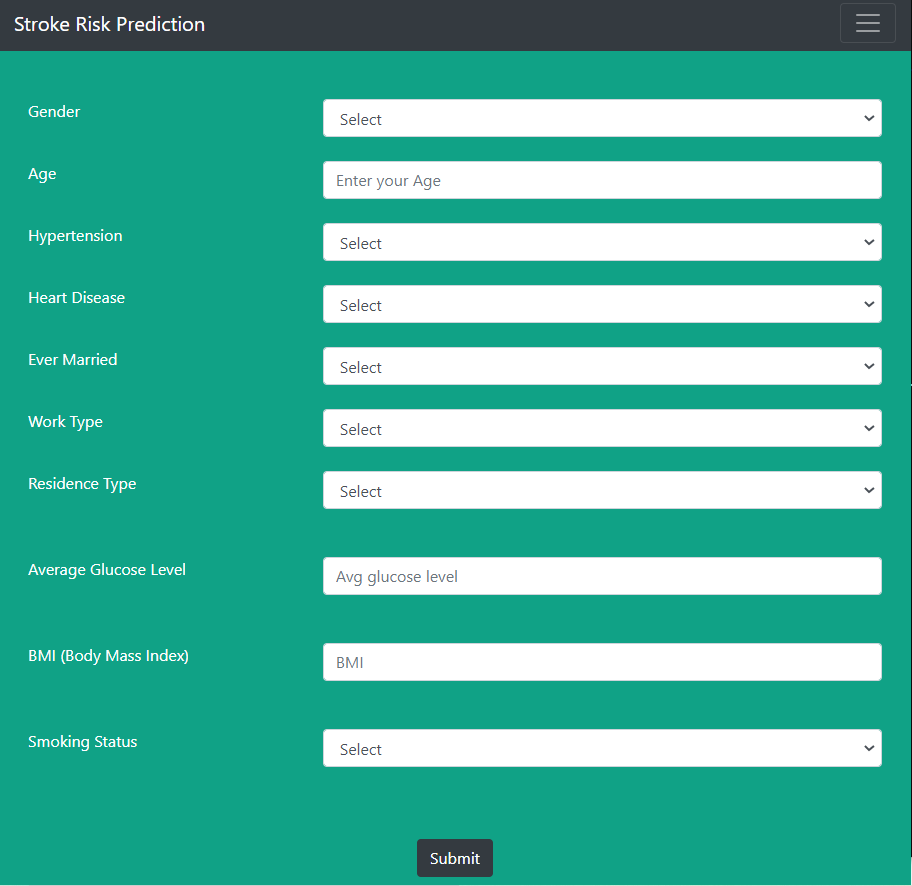


Figure: Layout of the webpage (User form)

**5.13 CONNECTING MODEL WITH FRONTEND AND PYTHON CODE**

The backend and frontend both work together to serve a single goal. It’s pretty helpful to keep it in mind at all times. They are made, so a user can access them.

WORKING OF WEBSITE

The user points their browser to one of your website’s URLs and waits for the browser to render the page. The user sees a useful and usable page. The user interacts with the page.

Thus till now, our website was working as static. In order to connect it to our working python code, we used flask i.e, micro-framework of python. In this, we created several routes using route decorator and thus it helped us in hosting it to a local server - http://localhost:7000/

**5.14 PREDICTING THE RISK OF STROKE**

The index() function is decorated with @app.route so that it is invoked when the browser sends a POST request. Using the request module of Flask it creates an object of the file and saves it in local storage. At the same time, it sends that filled parameters to python code by calling its function which in response returns the output string, which is then displayed on the webpage.

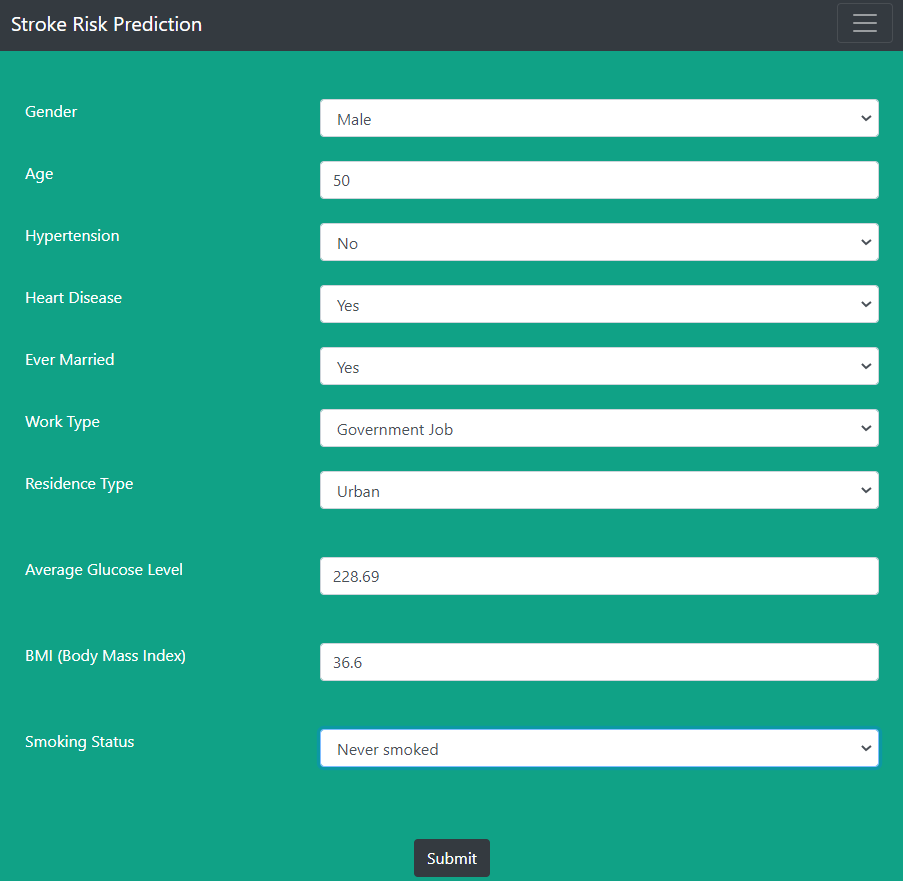


Figure 40: Sample input data filled in the Form

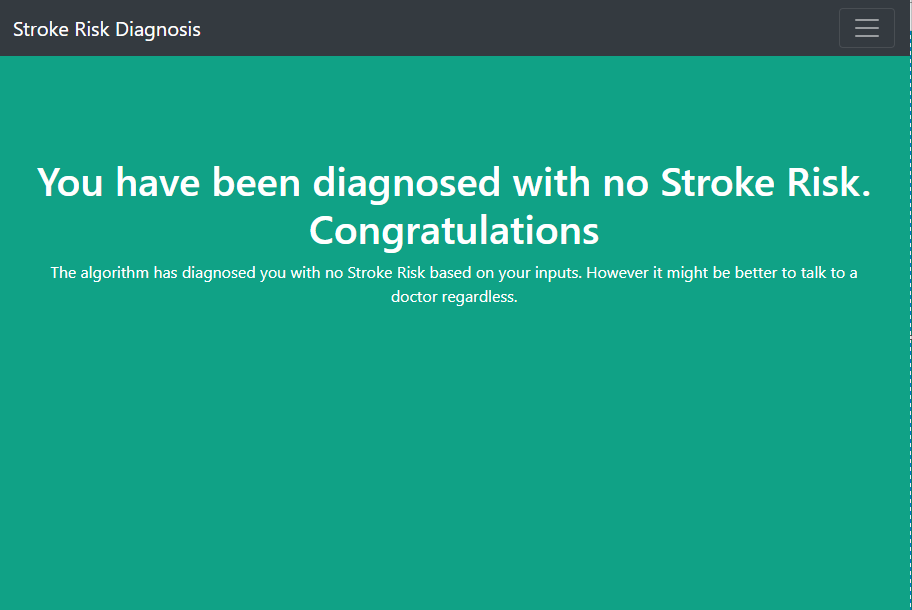


Figure 41: Sample results displayed at the Web Platform when using the model for prediction

Figure shows the front page of the Web App as visible to the user. Here he/she can fill the form based on their medical history and click on the submit button. The model functions in the background and provide the stroke risk which is shown on the Web App as shown in figure. Flask library is used to fetch the result predicted by the model and display it on the Web Page.

**CHAPTER - 6**

**CONCLUSIONS AND FUTURE SCOPE**

**6.1 CONCLUSIONS**

Through this project, we were able to build a system that can be used by user from the timing, location of his/her choice and check for chances of stroke to him. We compared 10 different machine learning classifiers based on accuracy on our dataset, used the best one and performed k-fold validation and hyper-parameter optimization on it too. We were able to understand the different and working of different classifiers and reason behind there particular behavior on our dataset. We are ultimately able to achieve our objective of learning about various machine learning classifier, comparing them and using the best one to build a system which can be used by anyone from anywhere to test the chances of stroke to him/her.

**6.2 FUTURE SCOPE**

There is a huge feature approach for this model some of which are enlisted below:

* The model is currently trained and tested for accuracy on a small dataset of approx. 5k records. We believe training it on any other dataset, which is bigger in size would enhance or prediction capability even more.
* Right now, the user is only asked for details in the form. We plan to enhance the web platform to display more information and statistics about the stroke.
* An android app for same purpose can be really useful and handy, we will this could be a potential future scope.
* Right now, the model uses 10 parameters for predicting the chances of stroke, adding more crucial parameters which play role in stroke and its prediction could be a really nice enhancement to the model.
* Adding more information about what stroke actually is, it causes, symptoms and other related and relevant information on the web page is also a nice and useful scope.
* Adding functionality to fetch all health record data and then use patient history for predicting chances of stroke could be a super huge enhancement.

**CHAPTER - 7**

**REFERENCES**

**Datasets:**

[1] “Dataset: Cardiovascular health study (CHS).” https://biolincc.nhlbi.nih.gov/studies/chs/. Accessed: 2016-05-8.

[2] "Stroke Prediction Dataset - (fedesoriano)." https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

**Model Classifiers: -**

[3] Scikit-learn Supervised Models - https://scikit-learn.org/stable/supervised\_learning.html

[4] GaussianNB - UPDATING FORMULAE AND A PAIR-WISE ALGORITHM FOR COMPUTING SAMPLE VARIANCES - Tony F.Chan, Gene H. Golub, Randall J. LeVeque:-Dept. Of CS, Standord NOV,1979 {<https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html> }

[5] MultinomialNB - {<https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html> } {<https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html> }

[6] BernoulliNB - {<https://nlp.stanford.edu/IR-book/html/htmledition/the-bernoulli-model-1.html> } {<https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html> }

[7] CategoricalNB - {<https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.CategoricalNB.html> }

[8] LinearRegression - {<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> }

[9] KNeighborsClassifier - {<https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm> } {<https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html> }

[10] Logistic Regression - Minimizing Finite Sums with the Stochastic AverageGradient - Mark Schmidt, Nicolas Le Roux, Francis Bach - 10 May 2016 - {<https://hal.inria.fr/hal-00860051v2> }

[11] Logistic Regression - {<https://en.wikipedia.org/wiki/Logistic_regression> } {<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html> }

[12] Logistic Regression - SAGA:A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives - Aaron Defazio, Francis Bach, Simon Lacoste-Julien - 16 Dec 2014 - {arXiv:1407.0202v3 [cs.LG], <https://arxiv.org/pdf/1407.0202.pdf> }

[13] Logistic Regression - Dual Coordinate Descent Methods for Logistic Regression and Maximum Entropy Models - Hsiang-Fu Yu, Fang-Lan Huang, Chih-JenLin - {<https://www.csie.ntu.edu.tw/~cjlin/papers/maxent_dual.pdf> }

[14] Decision Tree Classifier - Classification and Regression Trees Software and New Developments - Francesco Mola - 1998 -{<https://sci-hub.se/10.1007/978-3-642-72253-0_42> }

[15] Decision Tree Classifier - {<https://en.wikipedia.org/wiki/Decision_tree_learning> } {<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> }

[16] Decision Tree Classifier - The Elements of statistical Learning Ebook- {<https://web.stanford.edu/~hastie/Papers/ESLII.pdf> }

[17] Decision Tree Classifier - Understanding Decision Tree Classifier Blog - Tarun Gupta - Oct 2020 - {<https://towardsdatascience.com/understanding-decision-tree-classifier-7366224e033b> }

[18] Support Vector Classification(SVC) - {<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> }

[19] SVC - LIBSVM: A Library for Support Vector Machines - Chih-Chung Chang and Chih-Jen Lin - Jan 2021 - {<https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf> }

[20] SVM - Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods - by John C. Platt - 1999 - {<http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639> }

[21] SVM - Support Vector Machines with Scikit-learn Blog- Avinash Navlani - Dec 2019 - {<https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python>

[22] KMeans-{<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html> }

[23] KMeans Clustering - {<https://en.wikipedia.org/wiki/K-means_clustering> }

[24] KMeans - K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks - Imad Dabbura - Sept 2018 - {<https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a> }

[25] Random Forest Classifier - {<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> }

[26] Random Forest Classifier - Random Forest - LEO BREIMAN - 2001 - {<https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf> }

[27] Random Forest Classifier - {<https://en.wikipedia.org/wiki/Random_forest> }

[28] Random Forest Classifier - Understanding Random Forest - How the Algorithm Works and Why It Is So Effective Blog - Tony Yiu- Jun 2019 -{<https://towardsdatascience.com/understanding-random-forest-58381e0602d2> }

[29] Random Forest Classifier - Understanding Random Forests Classifiers in Python Blog - Avinash Navlani - May 2018 - {<https://www.datacamp.com/community/tutorials/random-forests-classifier-python> }

[30] Gradient Boosting Classifier - {<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html> }

[31] StandardScaler - {<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html> }

[32] XGBClassifier - How to Develop Your First XGBoost Model in Python Blog - Jason Brownlee - Aug 2016{<https://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/> }

[33] k-Fold Cross Validation - A Gentle Introduction to k-fold Cross-Validation Blog - Jason Brownlee - May 2018 - {<https://machinelearningmastery.com/k-fold-cross-validation/> }

[34] k-Fold Cross Validation - Cross-validation: evaluating estimator performance - {<https://scikit-learn.org/stable/modules/cross_validation.html> }

[35] k-Fold Cross Validation - Permutation Tests for Studying Classifier Performance - Markus Ojala, Gemma C. Garriga - Jun 2010 -{<https://www.jmlr.org/papers/volume11/ojala10a/ojala10a.pdf> }

[36] k-Fold Cross Validation - {<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html> }

[37] k-Fold Cross Validation - Why and how to Cross Validate a Model? Importance and types of Cross-validation techniques Blog- Sanjay M- Nov 2018 - {<https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f> }

[38] Hyper-Parameter Optimization - {<https://en.wikipedia.org/wiki/Hyperparameter_optimization> }

[39] Hyper-Parameter Optimization - An introduction on how to fine-tune Machine and Deep Learning models using techniques such as: Random Search, Automated Hyperparameter Tuning and Artificial Neural Networks Tuning - Pier Paolo Ippolito - Sept,2019 - {<https://towardsdatascience.com/hyperparameters-optimization-526348bb8e2d> }

[40] Hyper-Parameter Optimization - Random Search for Hyper-Parameter Optimization - James Bergstra, Yoshua Bengio - 2012 - {<https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a> }

[41] Hyper-Parameter Optimization - Algorithms for Hyper-Parameter Optimization - James Bergstra,Yoshua Bengio,Remi Bardenet,Balazs Kegl - Nov,2011 - {<https://hal.inria.fr/hal-00642998/file/draft1.pdf> }

[42] Hyper-Parameter Optimization - A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring - Yufei Xiaa,Chuanzhe Liua,Yu Ying Lib,Nana Liua - Feb,2017 - {<https://sci-hub.se/10.1016/j.eswa.2017.02.017> }

[43] GridSearchCV - {<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html> }

[44] RandomizedSearchCV -{<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html> }

[45] GridSearchCV - Machine Learning: GridSearchCV & RandomizedSearchCV - Hyperparameter tuning with GridSearchCV and RandomizedSearchCV Blog - Papa Moryba Kouate - Sept,2020 - {<https://towardsdatascience.com/machine-learning-gridsearchcv-randomizedsearchcv-d36b89231b10>}

[46] Interpretable Classifiers Using Rules And Bayesian Analysis: Building A Better Stroke Prediction Model - Benjamin Letham, Cynthia Rudin, Tyler H. Mccormick And David Madigan - 2015

[47] Comparison Of Different Machine Learning Approaches To Model Stroke Subtype Classification And Risk Prediction - Luis Garcia-Terriza, Jose L. Risco-Martin, Jose L. Ayala, Gemma Reig Rosello, Juan M. Camarasaltas - April 2019

[48] Prediction Models Aided Postoperative Decision Making Based on Neural Network and Support Vector Machines - Ibrahim Abuaqel, Abdullah Alsaif, Mohammed Almulhim, Mohammed Aljallal, Musab Alghuraibi, Mujahid Alzahrani, Abdulrahman AlHarby, Sunday O. Olatunji. - 2017

[49] Prediction of Stroke Using Deep Learning Model - Pattanapong Chantamit-o-pas and Madhu Goyal - 2017

[50] Classification of stroke disease using machine learning algorithms - Priya Govindarajan, Ravichandran Kattur Soundarapandian, Amir H. Gandomi, Rizwan Patan, Premaladha Jayaraman, Ramachandran Manikandan - Nov 2018

[51] An Integrated Machine Learning Approach to Stroke Prediction - Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, Honglak Lee - July 2017

**APPENDICES**

**APPENDIX 1**

**1.1 DETAILED TEST STRATEGY AND TEST CASES**

Our project is based on medical data collected from different hospitals. The testing and validation of model is performed using the scikit-learn functions.

Figure: Image of histogram showing the comparison of train, test and validation score.

About the testing in real life scenario, it is to be done on field.

**APPENDIX 2**

**2.1 USER GUIDE**

The user needs to follow the below-mentioned procedure for using the model to predict the chances of stroke to him/her.

Step 1: Visit to the webpage using “https://strokemsp.herokuapp.com”

Step 2: Fill in all the details asked.

Step 3: Click on submit button and wait for a moment for the model to predict.

Step 4: Within a few moments the model will predict the results which is being displayed to you on a new web page.

Step 5: If the model predicts you to have stroke risk, we recommend you to pay a visit to a doctor.

Step 6: Done. If the model predicts you having no risk of stroke; and still you feel any of the symptoms we recommend you to talk to a doctor.

\*\*EOF\*\*