"Education for Knowledge, Science and Culture"

-Shikhanmaharshi Dr. Bapuji Salunkhe

(स्वायत्त) कोल्हापूर

VIVEKANAND COLLEGE, KOLHAPUR
(Empowered Autonomous)

DEPARTMENT OF STATISTICS

A PROJECT REPORT

on

"IDENTIFYING KEY RISK FACTORS AFFECTING BRAIN STROKE"

Submitted by

1) Ms. Janhavi Shrinivas Mane.
2) Ms. Krushna Chandravadan Petkar.
3) Ms. Pratiksha Tanaji Shinde.

in partial fulfillment for the award of

the degree of

BACHELOR OF SCIENCE

in

STATISTICS

2024-25

Education for Knowledge, Science and Culture"

-Shikhanmaharshi Dr. Bapuji Salunkhe

(स्वायत्त) कोल्हापूर

VIVEKANAND COLLEGE, KOLHAPUR
(Empowered Autonomous)

DEPARTMENT OF STATISTICS

This is to Certify that,

Sr. No.	Name	Roll No.
1	Ms. Janhavi Shrinivas Mane.	8269
2	Ms. Krushna Chandravadan Petkar.	8272
3	Ms. Pratiksha Tanaji Shinde.	8274

have satisfactorily completed the project work **"IDENTIFYING KEY RISK FACTORS AFFECTING BRAIN STROKE"** as prescribed by ***Vivekanand College, Kolhapur (Empowered Autonomous)*** in partial fulfilment for **B. Sc. III** skill enhancement course in **STATISTICS**, in the academic year **2024-25**.

This project has been completed under our guidance and supervision. To the best of our knowledge and belief, the matter presented in this project report is original and has not been submitted elsewhere for any other purpose.

29.3.25 | 4.4.25 | 29.3.25
Project Guide | **Examiner** | **Head**

(Mrs. V. C. Shinde) | | **(Mrs. V. C. Shinde)**

HEAD
DEPARTMENT OF STATISTICS
VIVEKANAND COLLEGE, KOLHAPUR
(EMPOWERED AUTONOMOUS)

ACKNOWLEDGEMENT

Gratitude is the hardest emotion to express and often doesn't find adequate words to convey that entire one feels.We are extremely thankful to **Dr. R.R.Kumbhar**, Principal of Vivekanand College, Kolhapur (Empowered Autonomous) for giving permission for performing this project. It is our foremost duty to express our deep sense of gratitude and respect to the HOD Associate Prof. **V. C. Shinde** madam and Assistant Prof. **A. B. Bhosale** Sir for their up-lifting tendency and inspiring us for making of this project work completely and successfully.

Last but not the least we are thankful to our colleagues and those helped us directly or indirectly throughout this project work.

Your Sincerely,
B.Sc. III
Department of Statistics

INDEX

INTRODUCTION

Stroke is a serious and life-threatening medical condition that ranks as the second leading cause of death and a primary cause of disability globally. Every year, millions of people suffer from strokes, which can result in permanent neurological damage, affecting an individual's ability to speak, move, and perform everyday activities. The consequences of stroke are not only limited to physical impairment but extend to long-term cognitive and emotional challenges, thereby severely affecting the quality of life. Additionally, stroke places a substantial burden on healthcare systems due to the costs of treatment, rehabilitation, and long-term care.

Understanding the key risk factors that contribute to the occurrence and severity of stroke is crucial for reducing its impact. Stroke is a multi-factorial disease, influenced by a combination of genetic, environmental, and behavioral factors. There are two main categories of risk factors: modifiable and non-modifiable. Non-modifiable risk factors, such as advancing age, gender, family history, and genetic predisposition, cannot be changed but can help in identifying individuals at higher risk. On the other hand, modifiable risk factors—such as hypertension, smoking, obesity, physical inactivity, high cholesterol, and diabetes—are lifestyle-related and can be controlled or mitigated with proper intervention.

Recent advancements in medical research and public health efforts have shed light on the significance of these risk factors in stroke prevention and management. However, despite the growing body of knowledge, many individuals remain unaware of how their behaviors and health conditions can predispose them to stroke. Furthermore, the interaction between multiple risk factors complicates the ability to predict and prevent stroke with accuracy, making it imperative to continuously evaluate and understand these factors.

The primary aim of this project is to identify, analyze, and assess the most significant risk factors contributing to brain stroke. By reviewing both scientific literature and clinical data, this study seeks to identify high-risk populations and emphasize the importance of early intervention strategies. In doing so, the project will help inform public health campaigns, medical guidelines, and personalized healthcare approaches for stroke prevention. Ultimately, the goal is to provide a comprehensive understanding of how lifestyle modifications, medical management, and awareness initiatives can reduce the incidence of stroke, improve outcomes, and enhance public health.

OBJECTIVES

- To identify, assess, and analyse key risk factors affecting brain stroke, with a focus on prevention, early detection, and public awareness.
- To investigate the impact of lifestyle factors such as age, gender, heart disease, hypertension, smoking status, residential type, glucose level and work type on the risk of stroke, using data from patient histories and lifestyle surveys.
- To develop a regression model to classify the stroke occurrence.

DATA DESCRIPTION

❖ Data source:

We collected secondary data for this project and the dataset is taken from the Kaggle wesite : "https://www.kaggle.com" . The dataset includes 4980 people's data records with 11 attributes including sociodemographic, demographic and health related attributes as given below. The dataset contains 10 independent features and a dependent feature (target variable) "Stroke".

❖ Variables Description:

1. gender: "Male", "Female" or "Other"
2. age: age of the patient
3. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
4. heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
5. ever_married: "No" or "Yes"
6. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
7. residence_type: "Rural" or "Urban"
8. glucose_level: glucose level in blood
9. bmi: body mass index
10. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
11. stroke: 1 if the patient had a stroke or 0 if not

STATISTICAL TOOLS & TECHNIQUES

❖ Statistical Tools :

1) MS-Excel
2) Python
3) R-Software

❖ Statistical Techniques :

1) Exploratory data analysis:
   Histograms, bar charts, and pie charts.

2) Inferential Statistics:
   Chi-square test for goodness of fit and 2x2 contingency table.

3) Logistic regression

DATA PREPROCESSING

The data preprocessing is undertaken in the following steps: -

1. **Data Understanding:**

For the prediction of brain stroke based on the variables, variables are described before analyzing them for understanding them clearly. i.e, Information about number of observations, the data type of each variable.

2. **Data Cleaning:**

In this step we check that is there any missing values & duplicate values in the data-set or not. We observed that there are no any missing as well as duplicate values in the data set.

3. **Dropping irrelevant features:**

Variables which have no impact on the topic of discussion shall be considered as irrelevant. Keeping such variables may sometimes affect the performance of classifiers. While considering the brain stroke data set, the variables named ever married has nothing to do with the prediction. So, these features have been neglected manually in this study.

4. **Encoding categorical data:**

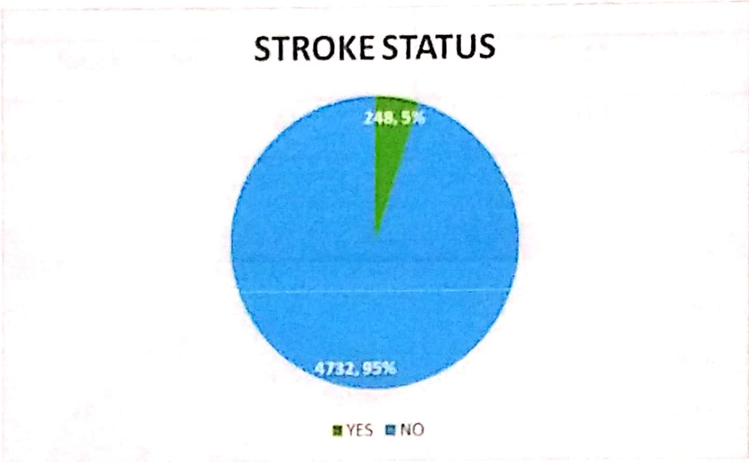
Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different model.

EXPLORATORY DATA ANALYSIS (EDA)

To see the distribution and relationship between the variables performed the Exploratory Data Analysis.

- **PIE DIAGRAM OF STROKE STATUS :**

STROKE	NO. OF PEOPLE
YES	248
NO	4732

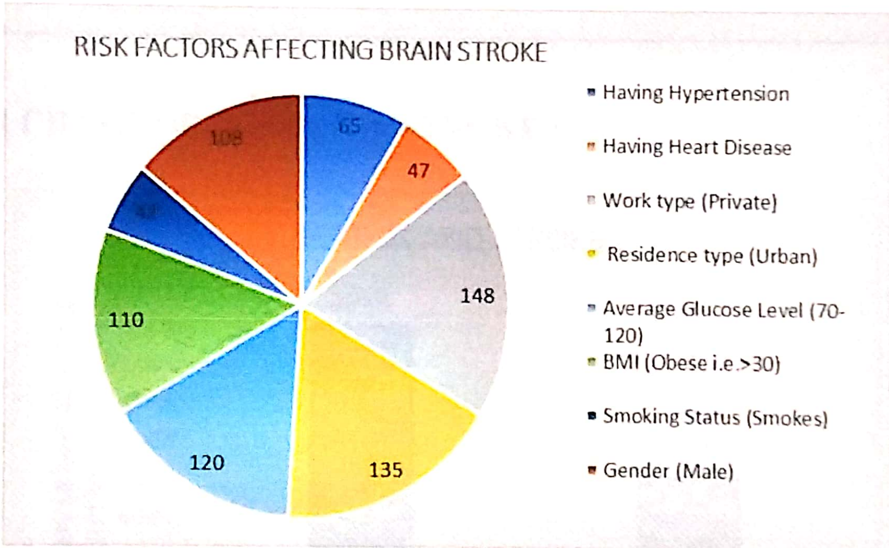


Conclusion :

The pie diagram shows that 5% of the population show stroke occurrence whereas rest of the population have no stroke occurrence.

- **PIE DIAGRAM FOR RISK FACTORS AFFECTING BRAIN STROKE :**

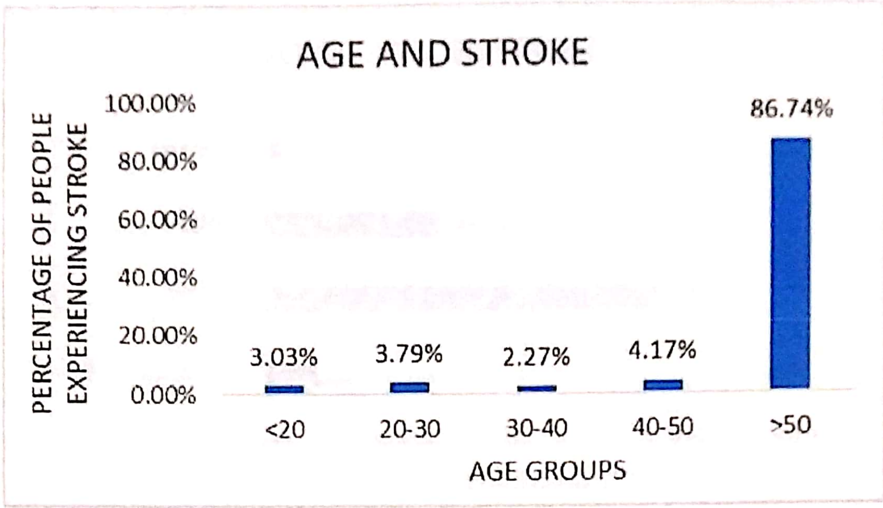
FACTORS	No. of people getting stroke
Having Hypertension	65
Having Heart Disease	47
Work type (Private)	148
Residence type (Urban)	135
Average Glucose Level (70-120)	120
BMI (Obese i.e.>30)	110
Smoking Status (Smokes)	42
Gender (Male)	108



Conclusion :

Out of the above 8 factors, work type factor has the most effect on brain stroke. Especially, private employees have the highest risk of stroke. Whereas, the pie diagram indicates that smoking status factor has the lowest effect on brain stroke.

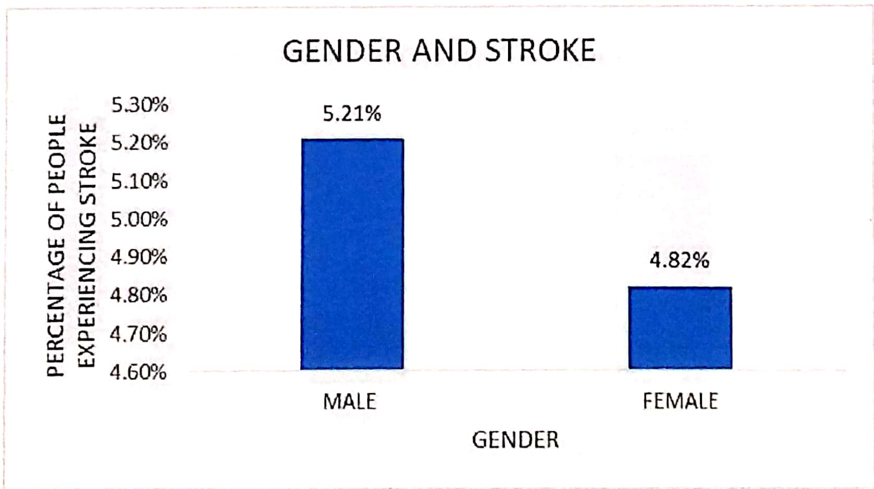
- **BAR CHART OF AGE & STROKE :**



Conclusion :

This graph indicates that the people above the age 50 have 86.74% risk of getting stroke. The people below the age 20 have 3.03% risk of getting stroke. Hence, risk of stroke increases significantly with age. While in younger age groups, the risk is generally low.

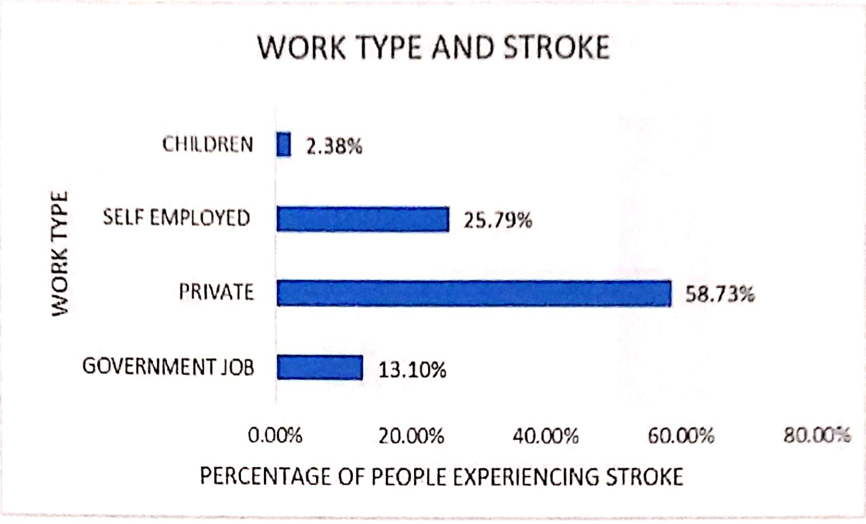
- **BAR CHART OF GENDER & STROKE :**



Conclusion :

This graph shows that the percentage of males affected by brain stroke is 5.21% .Whereas, the percentage of females is 4.82%.

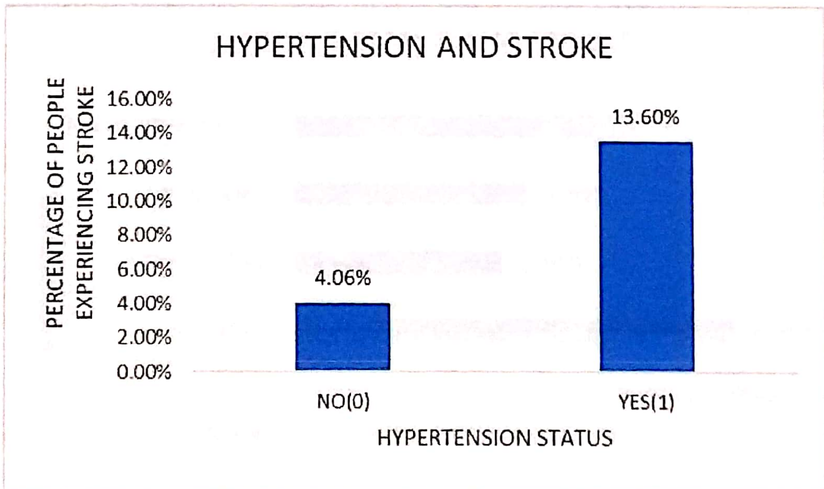
- **BAR CHART OF WORK TYPE & STROKE :**



Conclusion :

Private sector employees have 58.73% risk of experiencing stroke. Whereas, government employees have 13.10% stroke risk.

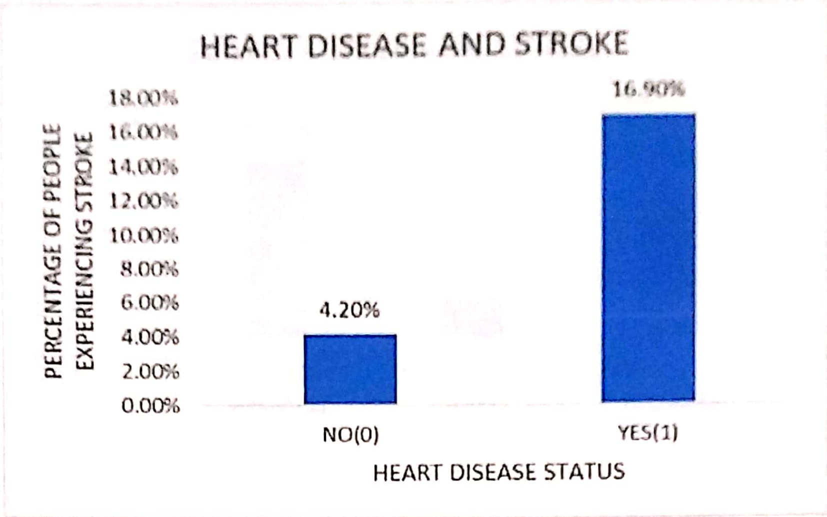
- **BAR CHART OF HYPERTENSION & STROKE :**



Conclusion :

Individuals with hypertension have a higher percentage of stroke occurence of 13.60%. As a result, hypertension significantly increases the risk of stroke.

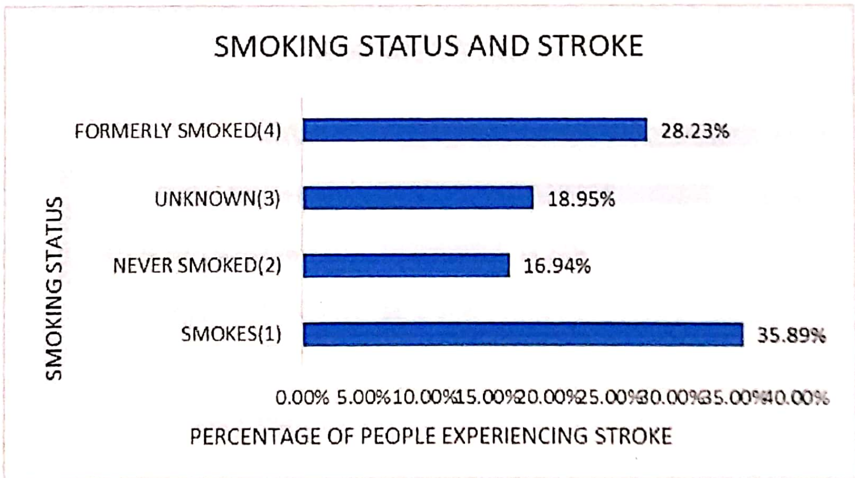
- **BAR CHART OF HEART DISEASE & STROKE :**



Conclusion :

Individuals with heart disease have a higher percentage of stroke occurrence i.e,16.9% . Heart disease significantly increases the risk of stroke.

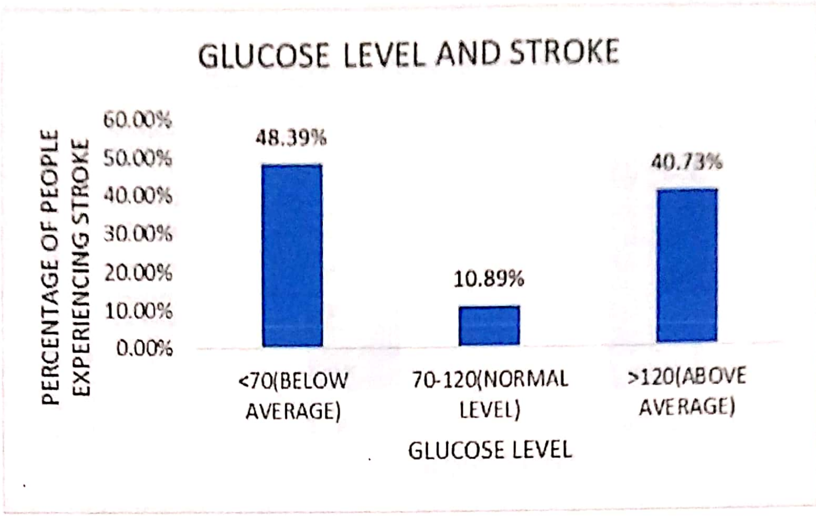
- **BAR CHART OF SMOKING STATUS & STROKE :**



Conclusion :

The largest proportion of stroke cases comes from those who currently smoke with the risk of 35.89%. Formerly smoked cases also showcase 28.23% stroke risk. This indicates that smoking is a significant risk factor for stroke.

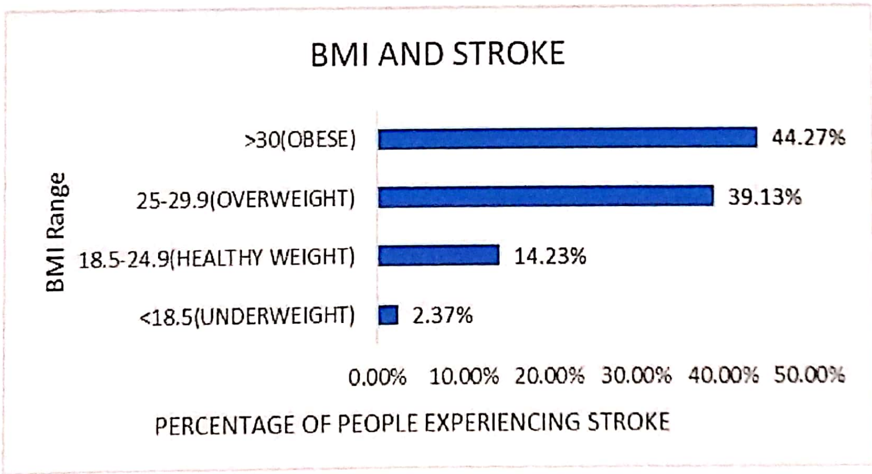
- **BAR CHART OF GLUCOSE LEVEL & STROKE:**



Conclusion :

People having low or high glucose level than standard range show 48.39% and 40.73% risk of stroke respectively.

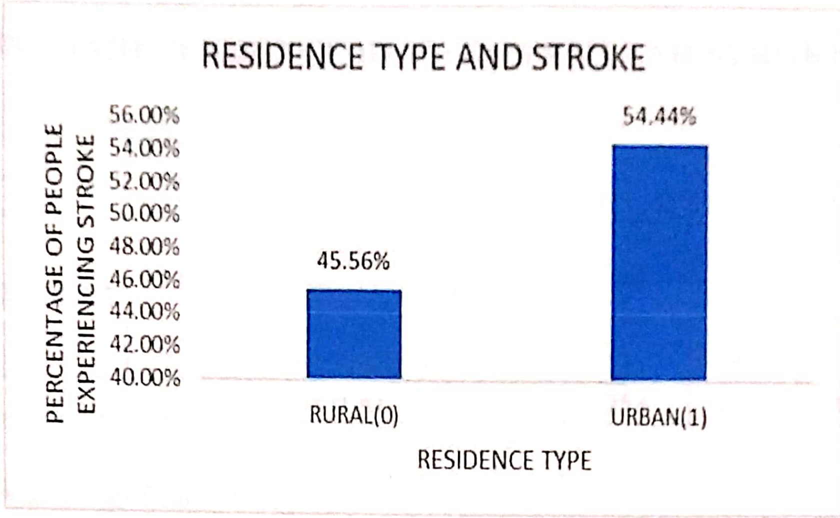
- **BAR CHART OF BMI & STROKE :**



Conclusion :

The risk of stroke for people having obesity is 44.27% and for people being overweight is 39.13%. This indicates that there is a strong association between obesity and stroke risk. Risk of stroke increases with increase in BMI.

- **BAR CHART OF RESIDENCE TYPE & STROKE :**



Conclusion :

People living in rural areas have 45.56% risk of stroke. On the other hand, people living in urban areas have 54.44% risk of stroke. Comparatively, there is more stroke risk for people living in urban areas than rural areas but the difference is not significantly high.

HYPOTHESIS TESTING

➢ TESTING INDEPENDENCE BETWEEN BMI AND STROKE :

		STROKE		
		YES (1)	NO(0)	TOTAL
BMI GROUPS	<18.5 (UNDERWEIGHT)	6	313	319
	18.5-24.9 (HEALTHY WEIGHT)	36	1210	1246
	25-29.9 (OVERWEIGHT)	99	1391	1490
	>30 (OBESE)	112	1813	1925
	TOTAL	253	4727	4980

(With level of significance 0.05)

➢ HYPOTHESIS :

Ho : Stroke is independent of BMI.

V/S

H1 : Stroke is dependent on BMI.

Under Ho,

CHI-SQ CAL : 26.9910

CHI-SQ TAB : 7.8147

CHI-SQ CAL > CHI-SQ TAB

➢ DECISION :

Stroke is dependent on BMI.

➢ TESTING INDEPENDENCE BETWEEN GENDER AND STROKE :

		GENDER		
		MALE	FEMALE	TOTAL
STROKE STATUS	YES(1)	108	140	248
	NO(0)	1966	2766	4732
	TOTAL	2074	2906	4980

(With level of significance 0.05)

➢ HYPOTHESIS :

Ho : Stroke is independent of gender.

V/S

H1 : Stroke is dependent on gender.

Under Ho,

CHI-SQ CAL : 0.2803

CHI-SQ TAB : 3.8415

CHI-SQ CAL < CHI-SQ TAB

➢ DECISION :

Stroke is independent of gender.

➢ TESTING INDEPENDENCE BETWEEN RESIDENCE TYPE AND STROKE :

		RESIDENCE TYPE		
		RURAL(0)	URBAN(1)	TOTAL
STROKE STATUS	YES(1)	113	135	248
	NO(0)	2335	2397	4732
	TOTAL	2448	2532	4980

(With level of significance 0.05)

➢ HYPOTHESIS :

Ho : Stroke is independent of residence type.

V/S

H1 : Stroke is dependent on residence type.

Under Ho,

CHI-SQ CAL : 0.1331

CHI-SQ TAB : 3.8415

CHI-SQ CAL < CHI-SQ TAB

➢ DECISION :

Stroke is independent of residence type.

➤ TESTING INDEPENDENCE BETWEEN SMOKING STATUS AND STROKE :

		STROKE STATUS		
		YES(1)	NO(0)	TOTAL
SMOKING STATUS	SMOKED(1)	89	1749	1838
	NEVER SMOKED(2)	42	734	776
	UNKNOWN(3)	47	1452	1499
	FORMERLY SMOKED(4)	70	797	867
	TOTAL	248	4732	4980

(With level of significance 0.05)

➤ HYPOTHESIS :

Ho : Stroke is independent of smoking status.

V/S

H1 : Stroke is dependent on smoking status.

Under Ho,

CHI-SQ CAL : 28.6964

CHI-SQ TAB : 7.8147

CHI-SQ CAL > CHI-SQ TAB

➤ DECISION :

Stroke is dependent on smoking status.

➤ TESTING INDEPENDENCE BETWEEN AGE AND STROKE :

		STROKE STATUS		
		YES(1)	NO(0)	TOTAL
AGE GROUPS	<20	8	920	928
	20 - 30	10	527	537
	30 - 40	6	634	640
	40 - 50	11	691	702
	>50	229	1944	2173
	TOTAL	264	4716	4980

(With level of significance 0.05)

➤ HYPOTHESIS :

Ho : Stroke is independent of age.

V/S

H1 : Stroke is dependent on age.

Under Ho,

CHI-SQ CAL : 211.5785

CHI-SQ TAB : 9.4877

CHI-SQ CAL > CHI-SQ TAB

➤ DECISION :

Stroke is dependent on age.

- **TESTING INDEPENDENCE BETWEEN GLUCOSE LEVEL AND STROKE :**

		STROKE STATUS		
		YES(1)	NO(0)	TOTAL
GLUCOSE LEVEL	<70 (BELOW AVERAGE)	120	3046	3166
	70 - 120 (NORMAL LEVEL)	27	706	733
	>120 (ABOVE AVERAGE)	101	980	1081
	TOTAL	248	4732	4980

(With level of significance 0.05)

- **HYPOTHESIS :**

Ho : Stroke is independent of glucose level.

V/S

H1 : Stroke is dependent on glucose level.

Under Ho,

CHI-SQ CAL : 55.5652

CHI-SQ TAB : 5.9915

CHI-SQ CAL > CHI-SQ TAB

- **DECISION :**

Stroke is dependent on glucose level.

➢ TESTING INDEPENDENCE BETWEEN WORK TYPE AND STROKE :

		STROKE STATUS		
		YES(1)	NO(0)	TOTAL
WORK TYPE	GOVERNMENT JOB (1)	33	611	644
	PRIVATE (2)	148	2712	2860
	SELF EMPLOYED (3)	65	739	804
	CHILDREN (4)	6	666	672
	TOTAL	252	4728	4980

(With level of significance 0.05)

➢ HYPOTHESIS :

Ho : Stroke is independent of work type.

V/S

H1 : Stroke is dependent on work type.

Under Ho,

CHI-SQ CAL : 39.6837

CHI-SQ TAB : 7.8147

CHI-SQ CAL > CHI-SQ TAB

➢ DECISION :

Stroke is dependent on work type.

➢ TESTING INDEPENDENCE BETWEEN HYPERTENSION AND STROKE STATUS :

		HYPERTENSION		
		NO(0)	YES(1)	TOTAL
STROKE STATUS	NO(0)	4319	413	4732
	YES(1)	183	65	248
	TOTAL	4502	478	4980

(With level of significance 0.05)

➢ HYPOTHESIS :

Ho : Stroke is independent of hypertension.

V/S

H1 : Stroke is dependent on hypertension.

Under Ho,

CHI-SQ CAL : 82.9981

CHI-SQ TAB : 3.8415

CHI-SQ CAL > CHI-SQ TAB

➢ DECISION :

Stroke is dependent on hypertension.

➤ TESTING INDEPENDENCE BETWEEN HEART DISEASEAND STROKE :

		HEART DISEASE		
		NO(0)	YES(1)	TOTAL
STROKE	NO(0)	4504	228	4732
	YES(1)	201	47	248
	TOTAL	4705	275	4980

(With level of significance 0.05)

➤ HYPOTHESIS :

Ho : Stroke is independent of heart disease.

V/S

H1 : Stroke is dependent on heart disease.

Under Ho,

CHI-SQ CAL : 90.2244

CHI-SQ TAB : 3.8415

CHI-SQ CAL > CHI-SQ TAB

➤ DECISION :

Stroke is dependent on heart disease.

LOGISTIC REGRESSION MODEL

- **Data Balancing :**

The original stroke dataset exhibited significant class imbalance, with the number of non-stroke cases greatly outnumbering stroke cases. To address this issue and improve model generalization, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE synthetically generates new samples for the minority class by interpolating between existing examples. This ensured that the model did not become biased toward the majority class and could effectively learn the patterns associated with stroke cases. After resampling, the dataset was split into training and testing subsets. Additionally, logistic regression was implemented with class_weight='balanced' to further mitigate any residual imbalance and reinforce fair treatment of both classes during model training. This approach helped achieve a balanced classification performance with an accuracy of 80.3%.

- **Train-Test Split:**

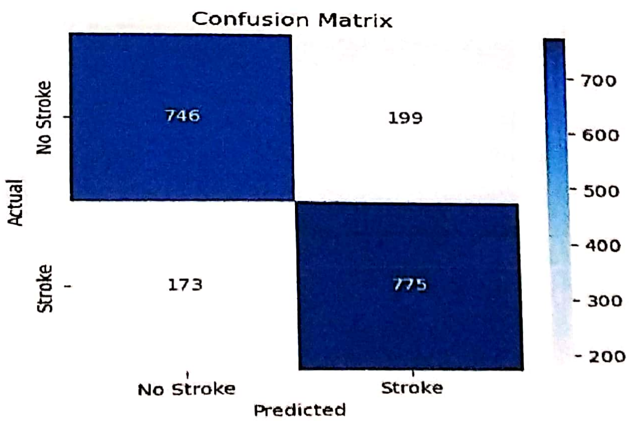
To evaluate the performance of our model, the dataset obtained from Kaggle was divided into training and testing sets. We used an 80:20 ratio, where 80% of the data was used to train the model and the remaining 20% was reserved for testing. This split ensures that the model is trained on a substantial portion of the data while still allowing for an unbiased evaluation on unseen data. The split was performed using the train_test_split function from the scikit-learn library, with a fixed random state to ensure reproducibility of results.

- **Logistic Regression model :**

Output:

Accuracy: 0.803

	precision	recall	f1-score	support
0	0.81	0.79	0.8	945
1	0.8	0.82	0.81	948



- **Model Results :**

The model shows good overall performance with an accuracy of approximately 80%, correctly predicting most stroke and non-stroke cases. However, it still misclassifies 173 actual stroke cases and 199 non-stroke cases, indicating room for improvement, especially in reducing false negatives.

CONCLUSION

1. The pie diagram shows that 5% of the population show stroke occurrence whereas rest of the population have no stroke occurrence.

2. The Age & Stroke graph indicates that the people above the age 50 have 86.74% risk of getting stroke. the people below the age 20 have 3.03% risk of getting stroke. Hence, risk of stroke increases significantly with age. While in younger age groups, the risk is generally low.

3. The Gender & Stroke graph shows that the percentage of males affected by brain stroke is 5.21% .Whereas, the percentage of females is 4.82%.

4. Private sector employees have 58.73% risk of experiencing stroke.Whereas, government employees have 13.10% stroke risk.

5. Individuals with hypertension have a higher percentage of stroke occurence of 13.60%. As a result, hypertension significantly increases the risk of stroke.

6. Individuals with heart disease have a higher percentage of stroke occurrence i.e,16.9% . Hence, heart disease significantly increases the risk of stroke.

7. The largest proportion of stroke cases comes from those who currently smoke with the risk of 35.89%. Formerly smoked cases also showcase high stroke risk i.e, 28.23%. This indicates that smoking is a significant risk factor for stroke.

8. People having low or high glucose level than standard range show 48.39% and 40.73% risk of stroke respectively.

9. The risk of stroke for people having obesity is 44.27% and for people being overweight is 39.13%. This indicates that there is a strong association between obesity and stroke risk. Risk of stroke increases with increase in BMI.

10. People living in rural areas have 45.56% risk of stroke. On the other hand, people living in urban areas have 54.44% risk of stroke. Comparatively, there is more stroke risk for people living in urban areas than rural areas but the difference is not significantly high.

11. Stroke is independent of gender and residence type and stroke is dependent on hypertension, heart disease, age, BMI, glucose level, smoking status and work type.

12. Out of the above 8 factors, work type factor has the most effect on brain stroke. Especially, private employees have the highest risk of stroke. Whereas, the pie diagram indicates that smoking status factor has the lowest effect on brain stroke.

13. The regression model shows good overall performance with an accuracy of approximately 80%, correctly predicting most stroke and non-stroke cases. However, it still misclassifies 173 actual stroke cases and 199 non-stroke cases, indicating room for improvement, especially in reducing false negatives.

REFERENCES

1) S.P.Gupta :"Statistical methods", 46th Revised Edition (2021), Sultan Chand & Sons publications.

2) C.R.Kothari : "Research methodology : Methods and techniques", 5th Revised Edition (2023), New Age International Publishers.

3) S.C.Gupta and V.K.Kapoor : "Fundamentals of Mathematical Statistics", 12th Edition (2020), Sultan Chand & Sons publications.

4) Paul Newbold, William L.Carroll, and Betty Thorne : "Statistics for Business and Economics", 10th Edition (2023), Pearson publications.

5) Gopal K. Kanji : "100 Statistical Tests", 3rd Edition (2006), New Delhi sage publications.