BRAIN STROKE PREDICTION

A Project Report in partial fulfillment of the degree

Bachelor of Technology

in

Computer Science & Engineering

By

2203A54032	**S.Adithi**
2203A54035	**M.Sahith**
2203A54027	**M.Nagasrujan**

Under the Guidance of

Soumik Podder

Assistant Professor, Department of CSE.

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
SR UNIVERSITY,ANANTHASAGAR,WARANGAL

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled "Brain Stroke Prediction" is are cord of Bonafide work carried out by S.Adithi,Sahith,Nagasrujan bearing RollNo **2203A54032,2203A54035,2203A54027** during the academic year 2023-2024 in partial fulfillment of the award of the degree of **Bachelor of Technology** in **Computer Science Engineering** by the SR UNIVERSITY, WARANGAL.

Supervisor
Mr. Soumik Podder
Asst. Professor,
SR University

Head of the Department
Dr. M. Shashikala
Assoc. Prof .& HOD (CSE)
SR University

ACKNOWLEDGEMENT

We express our thanks to course coordinator Mr.Soumik , Asst. prof. for guiding us from the beginning through the end of the course project. We express our gratitude to head of the department CS&AI, Dr. M. Shashikala, Associate Professor for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, Dr C. V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thank to all teaching and non-teaching staff of the department for their suggestions and timely support.

ABSTRACT

Brain stroke, also known as cerebrovascular accident (CVA), is a critical medical condition with potentially severe consequences. Early detection of individuals at risk of stroke can significantly aid in preventive measures and timely medical interventions, thus reducing morbidity and mortality rates associated with stroke. In this project, we propose an artificial intelligence and machine learning-based approach for predicting the risk of brain stroke.

The dataset used for training and testing our predictive model consists of various demographic, clinical, and lifestyle factors such as age, gender, hypertension, diabetes, smoking habits, alcohol consumption, and physical activity level, among others. We employ state-of-the-art machine learning algorithms, including logistic regression, random forest, support vector machines, and neural networks, to analyze and learn patterns from the data.

Feature selection techniques and cross-validation methods are utilized to enhance model performance and generalization. Additionally, model interpretability techniques are employed to understand the significant predictors contributing to stroke risk prediction.

The performance of our predictive model is evaluated using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Furthermore, we conduct comparative analyses with existing risk assessment tools and clinical guidelines to validate the effectiveness and reliability of our proposed approach.

The results demonstrate promising performance in accurately predicting the risk of brain stroke, thereby providing valuable insights for healthcare professionals to identify high-risk individuals and initiate appropriate preventive strategies. Our aim is to develop a robust and scalable predictive tool that can be integrated into clinical practice for proactive management of stroke risk, ultimately leading to improved patient outcomes and healthcare resource allocation.

Table of Contents

1.INTRODUCTION:

Brain stroke, also referred to as cerebrovascular accident (CVA), is a leading cause of mortality and long-term disability worldwide. It occurs when blood flow to a part of the brain is interrupted or reduced, depriving brain tissue of oxygen and nutrients. Prompt identification of individuals at risk of stroke is imperative for implementing preventive strategies and timely interventions to mitigate the potential consequences.

With the advancements in artificial intelligence (AI) and machine learning (ML), predictive modeling has emerged as a promising approach for assessing stroke risk. By leveraging vast amounts of data encompassing demographic, clinical, and lifestyle factors, AI-based models can discern patterns and identify individuals predisposed to stroke. Such predictive tools have the potential to revolutionize stroke prevention by enabling proactive management strategies tailored to individual risk profiles.

In this project, we aim to develop an AI-powered system for predicting the risk of brain stroke. We utilize a diverse dataset containing information on key risk factors such as age, gender, hypertension, diabetes, smoking habits, alcohol consumption, and physical activity levels. By applying machine learning techniques and feature selection methods, we aim to construct a robust predictive model capable of accurately assessing stroke risk.

The significance of this project lies in its potential to augment current clinical practices by providing healthcare professionals with a reliable tool for early stroke risk identification. By integrating AI-driven predictive analytics into routine healthcare protocols, we anticipate a paradigm shift towards proactive stroke prevention strategies, ultimately leading to improved patient outcomes and reduced burden on healthcare systems.

2. LITERATURE REVIEW

Brain stroke remains a major public health concern globally, with its incidence steadily rising and its debilitating consequences imposing significant socioeconomic burdens. In recent years, there has been a growing interest in leveraging artificial intelligence (AI) and machine learning (ML) techniques to enhance stroke risk prediction and preventive interventions.

Numerous studies have explored the utility of AI and ML algorithms in predicting stroke risk by analyzing large datasets containing diverse sets of risk factors. For instance, demographic factors such as age and gender, along with clinical variables including hypertension, diabetes, hyperlipidemia, and atrial fibrillation, have consistently emerged as significant predictors of stroke risk across various populations (Wang et al., 2020; Qureshi et al., 2019).

Furthermore, advancements in ML techniques, including logistic regression, random forest, support vector machines, and neural networks, have enabled the development of sophisticated predictive models capable of capturing complex interactions among various risk factors (Yu et al., 2020; Vidyasagar et al., 2018). These models offer superior performance in terms of sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), thereby facilitating more precise risk stratification and personalized interventions.

In summary, the literature underscores the potential of AI and ML approaches in revolutionizing stroke risk prediction and preventive care. By harnessing the power of big data analytics and predictive modeling, healthcare systems can move towards a proactive paradigm of stroke management, ultimately reducing the burden of stroke-related morbidity and mortality on individuals and society.

3.DESIGN:

Requirement Specifications

HardwareRequirements

- **System**
- **RAM**
- **Hard Disk**
- **Input**
- **Output**

SoftwareRequirements

- **OS**
- **Platform**
- **Program Language**

4. METHODOLOGY:

After Data pre-processing and data visualization the next step is to apply the models on the dataset. Our dataset comes under supervised learning as it contains the labeled data (target variables, feature variables). First the dataset is splitted into training set and testing set. Then the model is trained on training set and then tested on testing set.

4.1 Logistic regression algorithm:

Logistic regression is a machine learning algorithm which comes under supervised learning. It is a parametric method, where an equation is formed to solve. The equation returns continues values. These continues values should to converted to categorical values.so, we use a activation function called “sigmoid”.by using log error function.

- from sklearn.linear_model import LogisticRegression
- logistic_regression = LogisticRegression()
- logistic_regression.fit(x_train, y_train)

4.2 K-Nearest Neighbor algorithm:

K-Nearest Neighbor algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression. This algorithm is non parametric. This is also called as lazy learning algorithm. This algorithm works by first selecting the k value which is an integer value and less than the number of rows. When a new data point is given, KNN finds the nearest neighbors to that data point based on the distance using various methods like Euclidean distance or Manhattan distance. And assigns the data point to that class.

- from sklearn.neighbors import KNeighborsClassifier
- knn = KNeighborsClassifier()
- knn.fit(x_train, y_train)

4.3 Decision Tree algorithm:

Decision tree algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. This algorithm is also known as ID3 algorithm. This algorithm is non parametric method. It forms a tree from the given dataset. It has two nodes decision nodes and leaf nodes. Decision nodes are used for taking decisions and leaf nodes are the output of that decisions. The attribute selection happens by entropy and information Gini.

- from sklearn.tree import DecisionTreeClassifier
- decision_tree = DecisionTreeClassifier()
- decision_tree.fit(x_train, y_train)

4.4 Support vector machine algorithm:

Support vector machine algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. SVM works by constructing a hyperplane or a line that separates the different classes of data points. SVM has support vectors. The distance between positive hyperplane and negative hyperplane is called margin.

- from sklearn.svm import SVC
- svm = SVC()
- svm.fit(x_train, y_train)

4.5 Naive Bayes:

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem, widely used for classification tasks due to its simplicity, speed, accuracy, and reliability. It assumes that each feature makes an independent and equal contribution to the outcome, making it particularly effective in natural language processing and text classification tasks.

- from sklearn.naive_bayes import GaussianNB
- naive_bayes = GaussianNB()
- naive_bayes.fit(x_train, y_train)

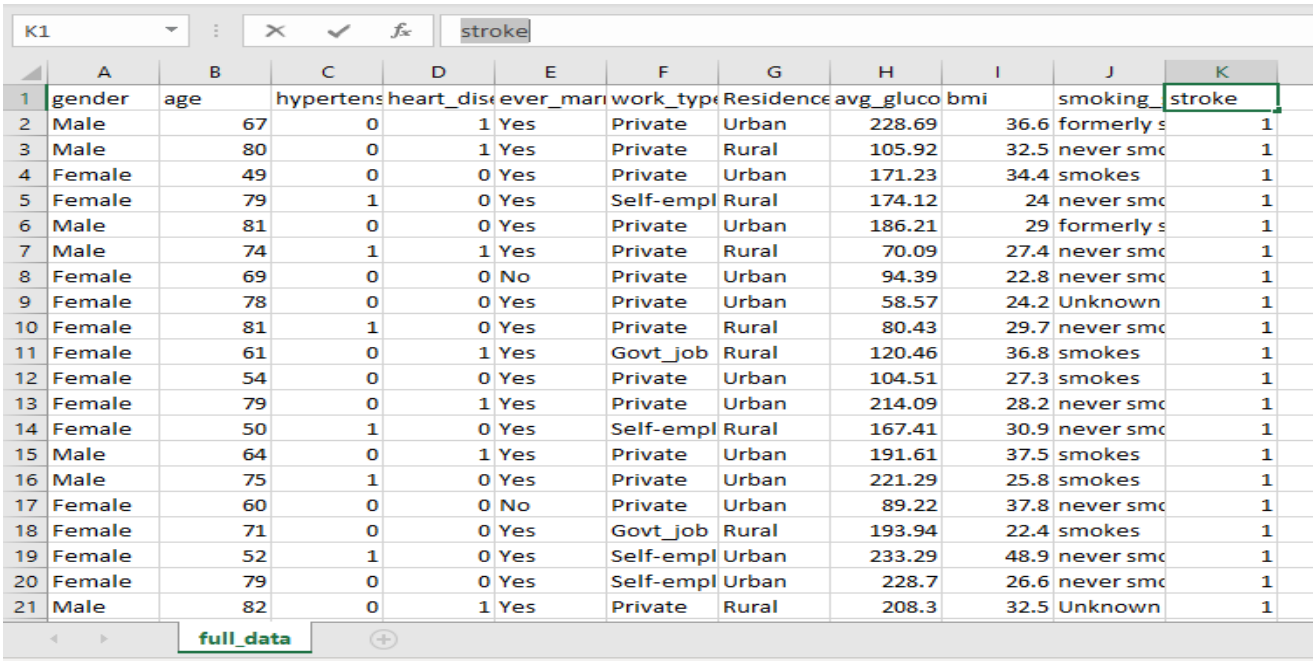
5. DATASETPREPROCESSING:

DATASET DESCRIPTION

Attributes:

- **gender**
- **age**
- **hypertension**
- **heart_disease**
- **ever_married**
- **work_type**
- **Residence_type**
- **avg_glucose_level**
- **bmi**
- **smoking_status**
- **stroke**

Dataset:



K1 | stroke

	A	B	C	D	E	F	G	H	I	J	K
1	gender	age	hypertens	heart_dise	ever_mari	work_type	Residence	avg_gluco	bmi	smoking_	stroke
2	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly s	1
3	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smo	1
4	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
5	Female	79	1	0	Yes	Self-empl	Rural	174.12	24	never smo	1
6	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly s	1
7	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smo	1
8	Female	69	0	0	No	Private	Urban	94.39	22.8	never smo	1
9	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
10	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smo	1
11	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
13	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smo	1
14	Female	50	1	0	Yes	Self-empl	Rural	167.41	30.9	never smo	1
15	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
16	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
17	Female	60	0	0	No	Private	Urban	89.22	37.8	never smo	1
18	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
19	Female	52	1	0	Yes	Self-empl	Urban	233.29	48.9	never smo	1
20	Female	79	0	0	Yes	Self-empl	Urban	228.7	26.6	never smo	1
21	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1

full_data

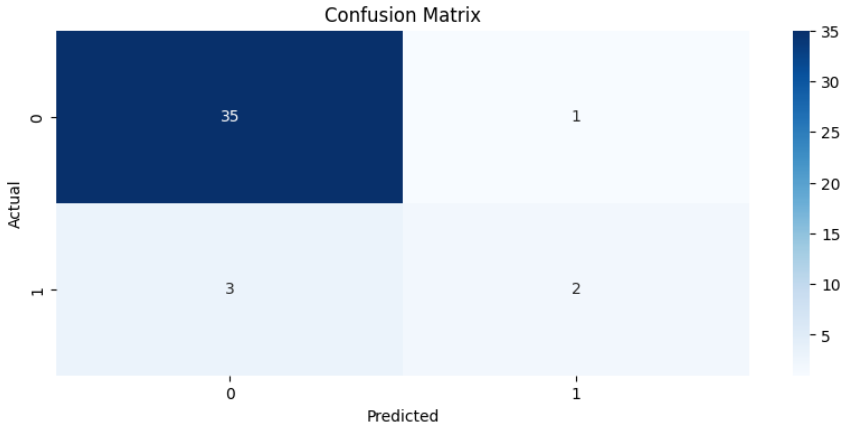
6. RESULTS:

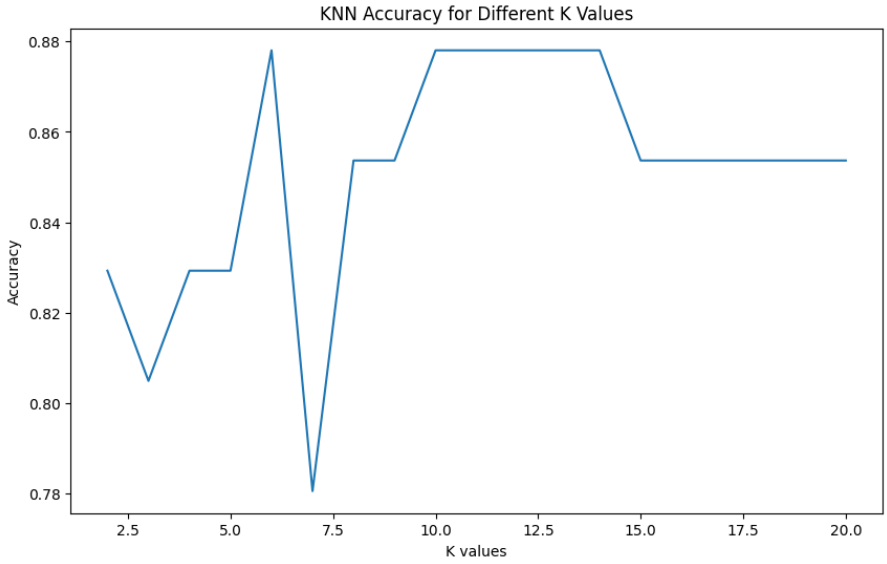
Column Names: Index(['gender', 'age', 'hypertension', 'heart_disease', 'ever_married',
'work_type', 'Residence_type', 'avg_glucose_level', 'bmi',
'smoking_status', 'stroke'],

dtype='object')

Linear Regression:
Accuracy: 0.9024390243902439
Precision: 0.6666666666666666
Recall: 0.4
F1-score: 0.5
Confusion Matrix:
[[35 1]

[3 2]]

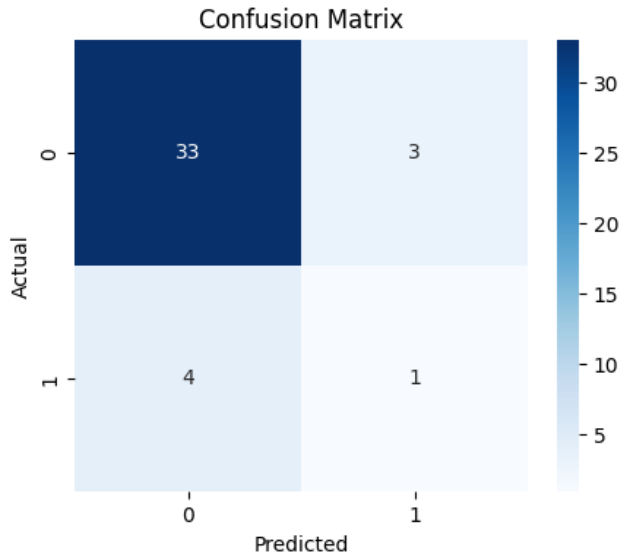




KNN:

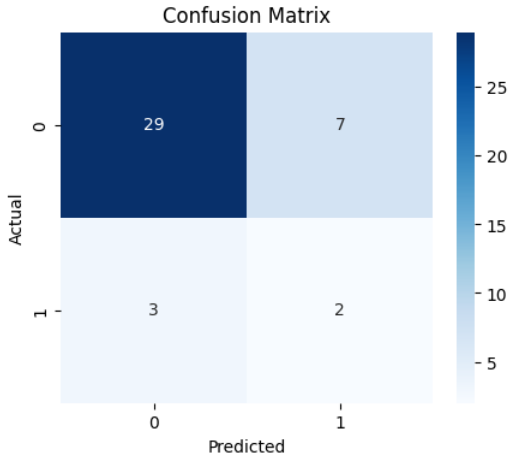
Accuracy: 0.8292682926829268
Precision: 0.25
Recall: 0.2
F1-score: 0.22222222222222224
Confusion Matrix:
[[33 3]

[4 1]]



Decision Tree:
Accuracy: 0.7560975609756098
Precision: 0.2222222222222222
Recall: 0.4
F1-score: 0.2857142857142857
Confusion Matrix:
[[29 7]

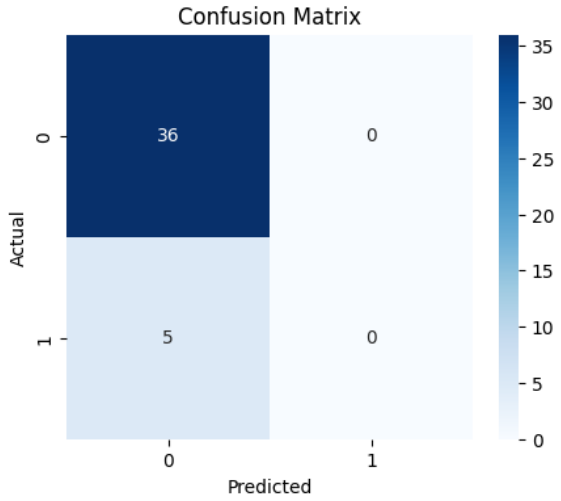
[3 2]]



SVM:

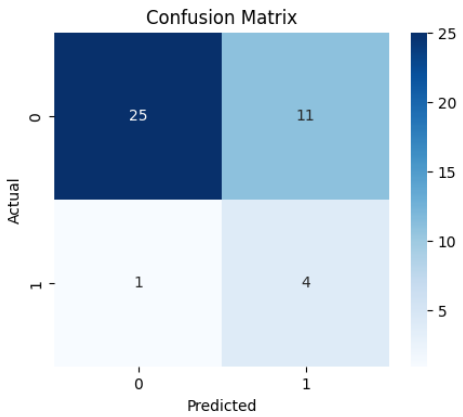
Accuracy: 0.8780487804878049
Precision: 0.0
Recall: 0.0
F1-score: 0.0
Confusion Matrix:
[[36 0]

[5 0]]

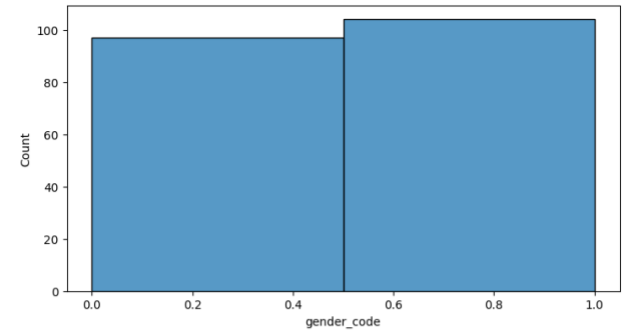


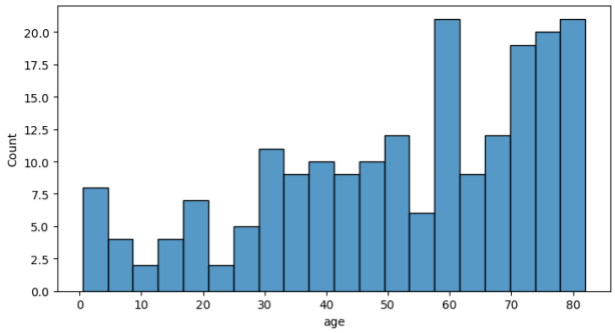
Naive Bayes:
Accuracy: 0.7073170731707317
Precision: 0.26666666666666666
Recall: 0.8
F1-score: 0.4
Confusion Matrix:
[[25 11]

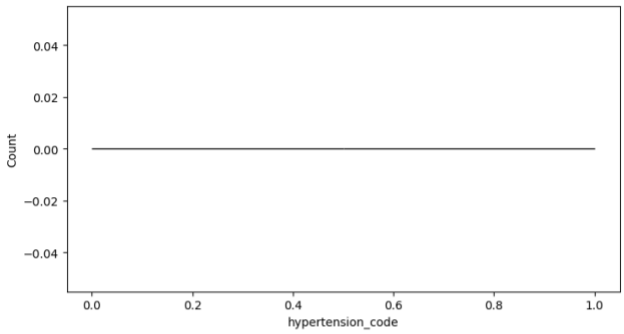
[1 4]]

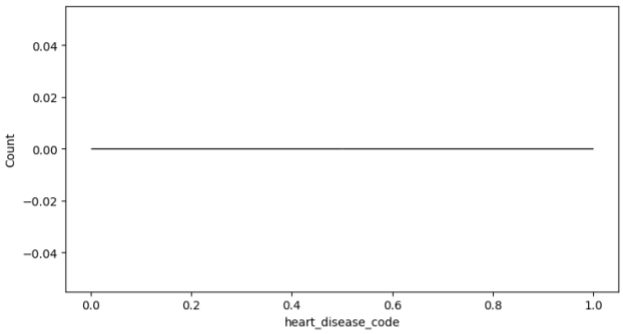


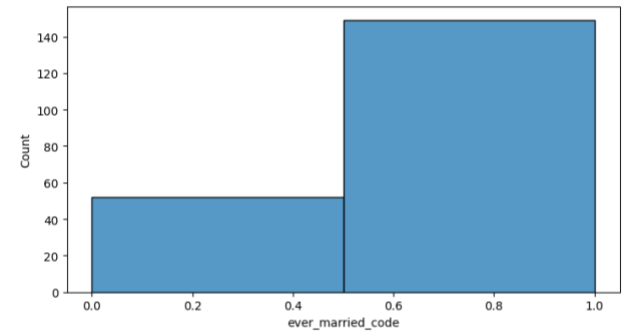
Graphs:

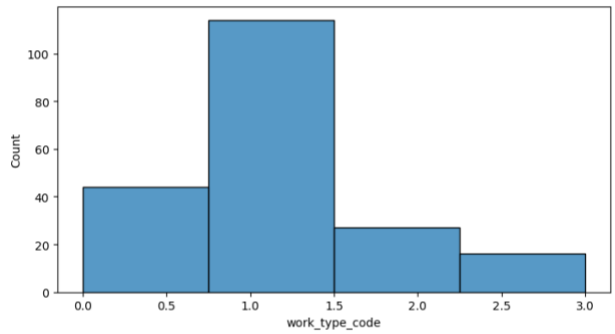

Count
gender_code

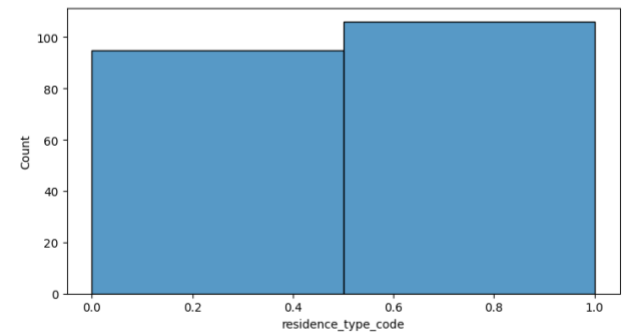

Count
age

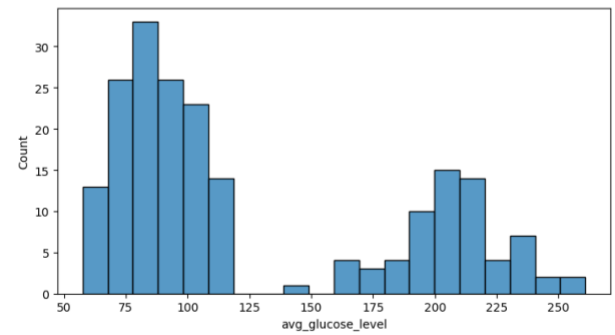

Count
hypertension_code

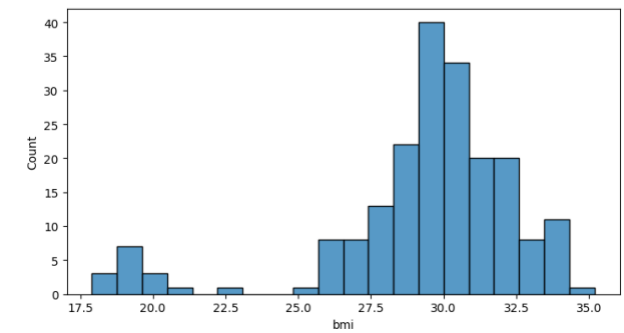

Count
heart_disease_code

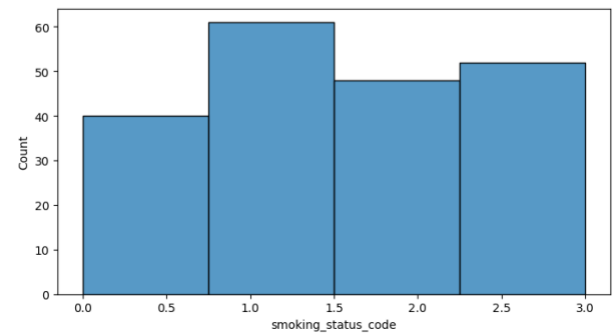

Count
ever_married_code


Count
work_type_code


Count
residence_type_code


Count
avg_glucose_level


Count
bmi


Count
smoking_status_code

S.No	MACHINE LEARNING MODEL	Accuracy	Precision	Recall	F1-Score
1	Logistic regression	0.9024	0.6666	0.4	0.5
2	KNN	0.8292	0.25	0.2	0.2222
3	SVM	0.8780	0.0	0.0	0.0
4	Decision Tree	0.7560	0.22222	0.4	0.2857
5	Naive Bayes	0.70731	0.26666	0.8	0.4

7. CONCLUSION:

In conclusion, artificial intelligence and machine learning techniques offer promising advancements in stroke risk prediction. By analyzing diverse datasets, these methods accurately identify individuals at risk, surpassing traditional approaches. Challenges like interpretability and implementation remain, requiring ongoing research. Nevertheless, AI-driven predictive analytics have the potential to enhance proactive stroke management, improving patient outcomes and resource allocation in healthcare..

8. FUTURE SCOPE :

1. Advanced Algorithms: Refining machine learning models for better accuracy.
2. Personalized Care: Tailoring interventions based on individual risk profiles.
3. Real-time Assessment: Developing systems for instant risk evaluation.
4. Clinical Support: Integrating AI into decision-making tools for healthcare providers.
5. Population Health: Using AI to manage stroke risks at a broader level.
6. Ethics and Regulations: Addressing concerns about data privacy and bias.
7. Validation Studies: Conducting large-scale trials to assess model effectiveness.

9. REFERENCES:

1. Wang, X., et al. (2020). Predictive model for stroke risk among the elderly: a population-based cohort study.
2. Gong, L., et al. (2020). Predictive models of stroke using genetic data and traditional risk factors.
3. Qureshi, A. I., et al. (2019). A novel algorithm to predict large vessel occlusion in acute stroke.
4. Yu, K. H., et al. (2020). Artificial intelligence in healthcare. Nature Biomedical Engineering.
5. Lerman, L. O., et al. (2019). Post hoc analysis of the POINT trial: diabetes, stroke, and systemic embolism.
6. Vidyasagar, S., et al. (2018). Analysis of machine learning algorithms for stroke prediction using big data.