SALIM HABIB UNIVERSITY

Brain Stroke Prediction System

Submitted by

ALISHA FATIMA

(F20CSC44)

Faculty of Information Technology

Department of Computer Science

Machine Learning Course Project Report Presented to

Dr. Rizwan Ahmed Khan

Faculty of Information Technology

Department of Computer Science

Salim Habib University, Karachi, Pakistan

Year 2024

Contents

1 Introduction

Brain Stroke happens when there is a blockage in the blood circulation in the brain or when a blood vessel in the brain breaks and leaks. The burst or blockage prevents blood and oxygen reaching the brain tissue. Without oxygen the tissues and cells in the brain are damaged and die in no time leading to many symptoms.

Once brain cells die, they generally do not regenerate and devastating damage may occur, sometimes resulting in physical, cognitive and mental disabilities. It is crucial that proper blood flow and oxygen be restored to the brain as soon as possible.

Worldwide, brain stroke is the second leading cause of death and third leading cause of disability. In some cases, the warning signs of a stroke can be obvious but what's going on inside the body is incredibly complex. 80 percent of strokes are preventable. But once you've had a stroke, the chances you have another one are greater.

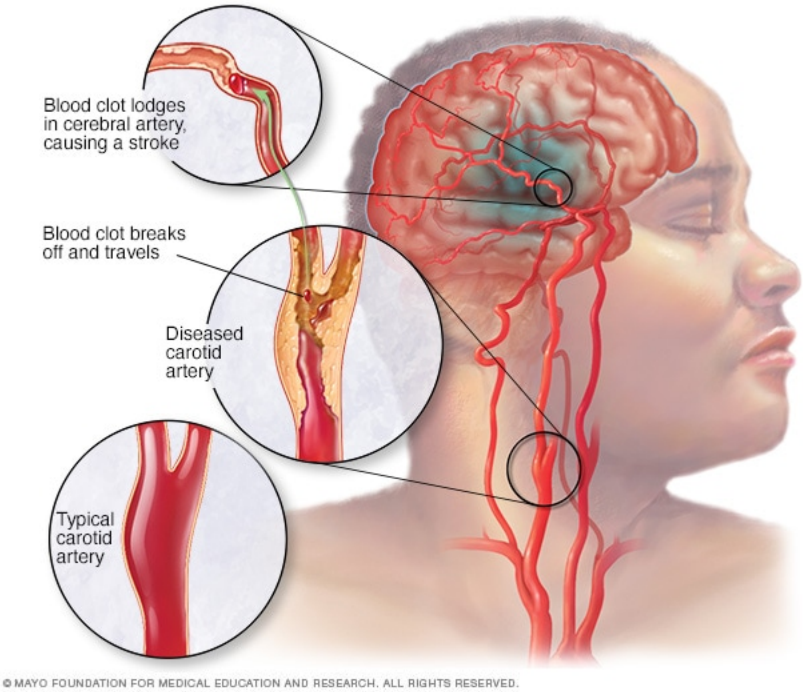

Blood clot lodges in cerebral artery, causing a stroke
Blood clot breaks off and travels
Diseased carotid artery
Typical carotid artery
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Figure 1: Brain Stroke

1.0.1 Causes of Brain Stroke

Risk factors that can contribute to the development of brain strokes include:

- High blood pressure
- High cholesterol
- Diabetes
- Smoking
- hyper tension
- heart disease
- Excessive alcohol consumption
- Family history of strokes
- Age (risk increases with age)
- Prior history of strokes or TIAs

It's important to address and manage these risk factors to reduce the likelihood of experiencing a brain stroke.

1.1 Types of Brain Strokes

1.1.1 Ischemic Stroke

Ischemic strokes occur due to a blockage or narrowing of arteries that supply blood to the brain. This blockage reduces blood flow, depriving brain cells of oxygen and nutrients. The most common cause is the formation of a blood clot within a blood vessel, often originating from plaque buildup in arteries (atherosclerosis) or a clot elsewhere in the body that travels to the brain. Ischemic strokes account for about 87% of all stroke cases.

1.1.2 Hemorrhagic Stroke

Hemorrhagic strokes result from the rupture or leakage of blood vessels in the brain, leading to bleeding within or around brain tissue. This bleeding can damage surrounding brain cells and increase pressure within the skull, causing further injury. Hemorrhagic strokes are often caused by conditions such as high blood pressure, aneurysms (weak spots in blood vessel walls), or arteriovenous malformations (abnormal tangles of blood vessels).

1.1.3 Transient Ischemic Attack (TIA)

TIAs, often referred to as "mini-strokes," are temporary episodes of reduced blood flow to the brain. Unlike full-blown strokes, TIAs typically last for a short time (usually minutes to hours) and cause temporary symptoms similar to those of a stroke. However, TIAs do not cause permanent brain damage because the blood flow is quickly restored. TIAs are often warning signs of a higher risk of future strokes and should be taken seriously as they provide an opportunity for preventive measures.

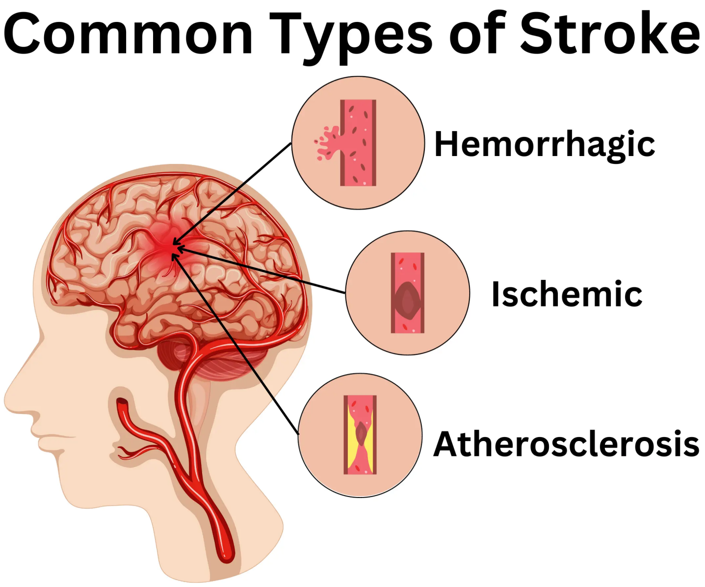


Figure 2: Types of Brain Stroke

2 Objective

The objective of this project is to analyze and predict the occurrence of brain strokes using a dataset that includes various patient attributes. By building predictive models, I aim to identify patterns and risk factors associated with strokes, ultimately assisting in prevention and early intervention strategies. This system predicts whether the Patient has a stroke or not. This System follows the following steps:

- Data Collection
- Data Pre-processing (Data Cleaning)
- Model Design
- Model Evaluation
- Comparative Analysis
- Result (Output)

3 Dataset Description

The dataset used in this project includes the following attributes:

- **id:** Unique identifier
- **gender:** Gender of the patient (Male, Female, Other)
- **age:** Age of the patient
- **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- **heart_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- **ever_married:** Yes if the patient is married, No if the patient is not married

- **work_type:** Profession of the patient (children, Govt_job, Never_worked, Private, Self-employed)

- **Residence_type:** Residence category of the patient (Rural, Urban)

- **avg_glucose_level:** Average glucose level in the blood of the patient

- **bmi:** Body Mass Index of the patient

- **smoking_status:** Smoking status of the patient (formerly smoked, never smoked, smokes, Unknown)

- **stroke:** 1 if the patient had a stroke, 0 if not

4 EDA (Exploratory Data Analysis)

The dataset contains 359 records and 11 attributes.

1. The data contains 189 records where the patients had a stroke and 170 records where they did not.

2. 200 patients are Male, 159 patients are Female.

3. 229 patients have never smoked, 70 patients formerly smoked, 39 patients currently smoke, and for 21 patients the smoking status is unknown.

4. 197 patients are married and 162 patients are not married.

5. 147 patients work in the private sector, 81 are self-employed, 61 work in the government job, 39 have never worked, and 31 are children.

6. 178 patients live in urban areas and 181 live in rural areas.

7. The average glucose level ranges from 55.42 to 271.74.

8. The BMI ranges from 15.5 to 48.9.

9. Distribution of age, average glucose level, and BMI are crucial factors for the prediction model.

10. The standard deviations for age, average glucose level, and BMI are 21.69, 58.08, and 6.79 respectively, indicating significant variability in the data.

5 Methodology

5.1 Data Pre-processing

The dataset used for this project consists of various features such as gender, age, hypertension, heart disease, marriage status, work type, residence type, average glucose level, BMI, smoking status, and stroke. The dataset was loaded and pre-processed to handle categorical variables, missing values, and to scale the numerical features.

- Categorical variables were encoded using label encoding.
- Features were standardized to have a mean of 0 and a standard deviation of 1.

5.2 Model Design

Several machine learning models were trained and evaluated to predict the occurrence of stroke based on the features in the dataset. The models include:

- Decision Tree
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Logistic Regression

5.2.1 Performance Metrics

The performance of the models was evaluated using the following metrics:

Accuracy Accuracy is the ratio of correctly predicted instances to the total instances.

Precision Precision is the number of true positive results divided by the number of all positive results, including those not identified correctly.

Recall Recall is the number of true positive results divided by the number of positives that should have been retrieved.

F1 Score The F1 Score is the harmonic mean of precision and recall, providing a single measure of a model's performance.

6 Visualization

6.1 Visualization of Target

The following graph shows the distribution of the target variable, stroke. It indicates how many patients had a stroke and how many did not.

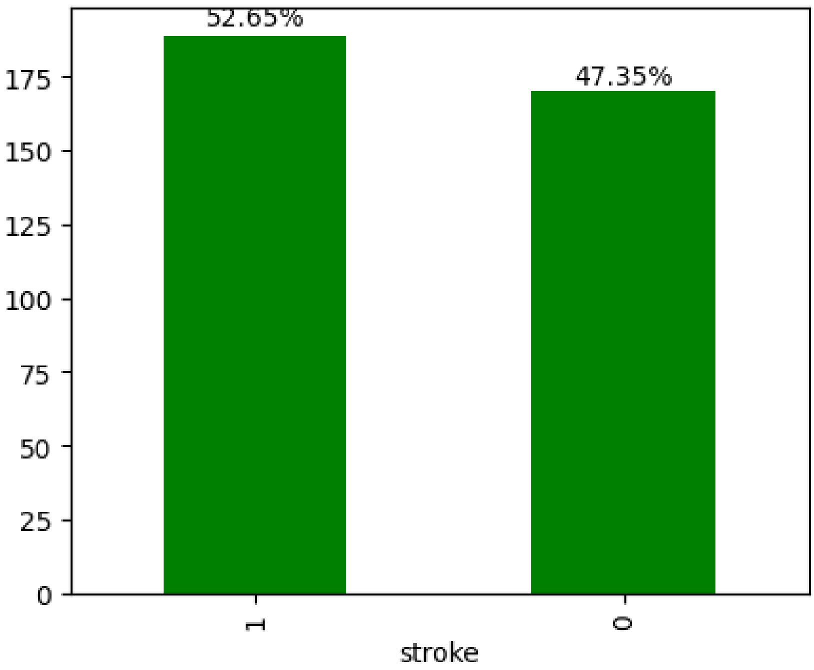


Figure 3: Distribution of Stroke

6.2 Visualization of Features

The following graphs illustrate the distributions of key features in the dataset. These visualizations help in understanding the data and identifying any patterns or anomalies.

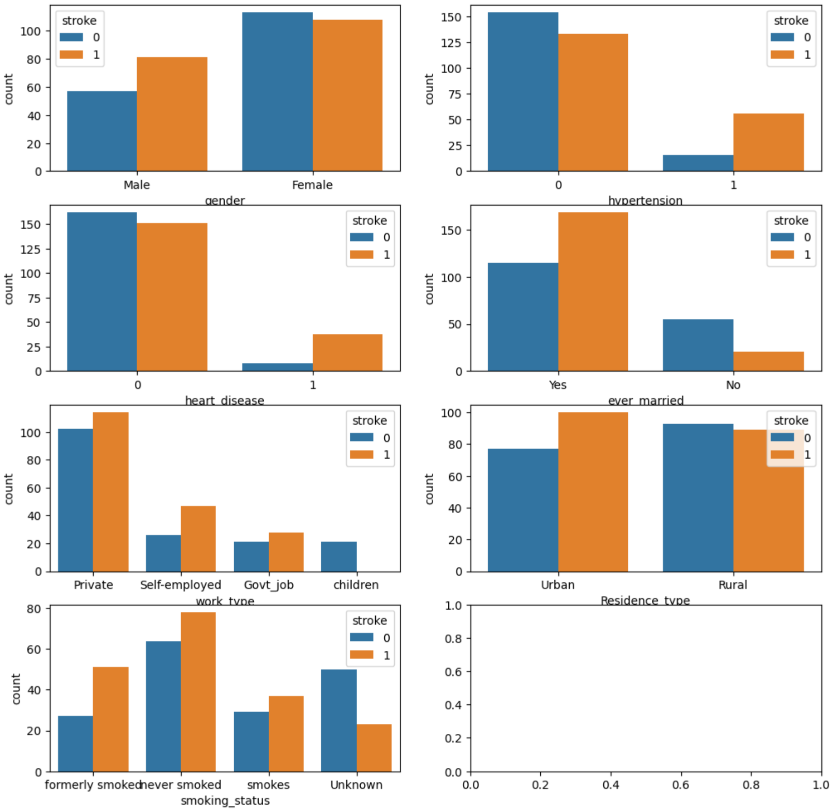


Figure 4: Distribution of Features

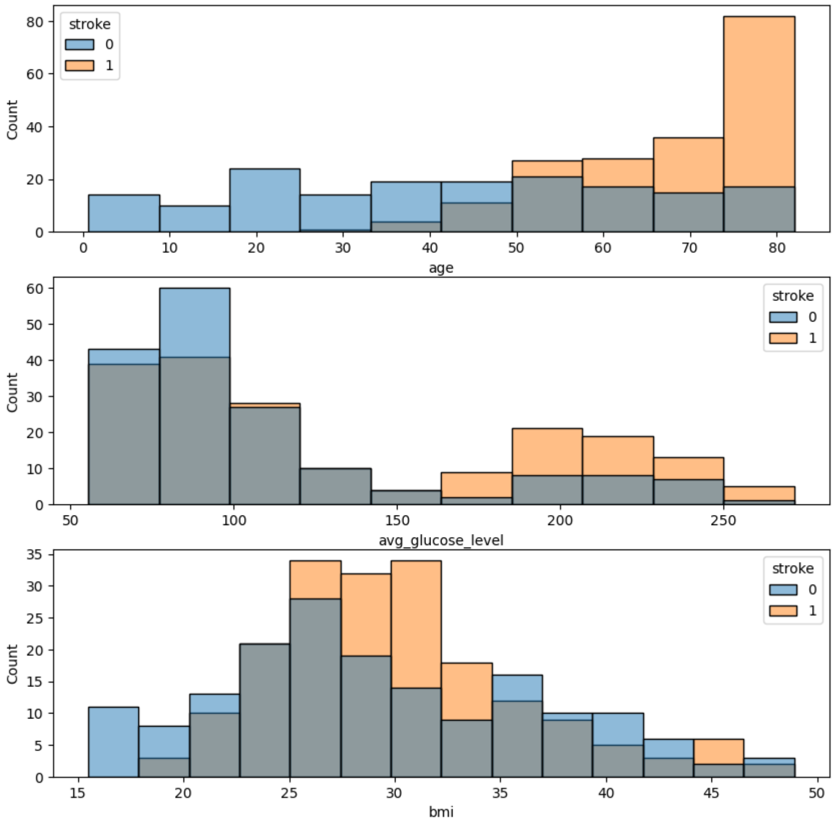


Figure 5: Distribution of Additional Features

6.3 Visualization of Outliers

The following graph shows that there is no outliers. This helps in improving the model performance as there is no extreme values that could skew the results.

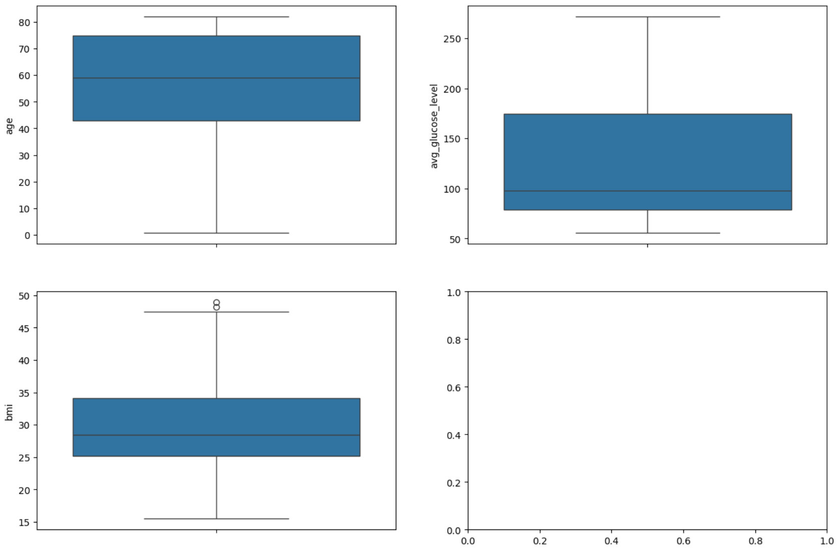


Figure 6: Data without Outliers

7 Comparative Analysis

A comparative analysis of the different models was performed based on various performance metrics such as Recall, Precision, Accuracy, and F1 Score. The following table shows the F1 Scores, Precision, Recall and Accuracy of the models:

Model	**F1 Score**
Decision Tree	0.80
KNN	0.76
SVM	0.82
Logistic Regression	0.79

Table 1: F1 Scores of different models

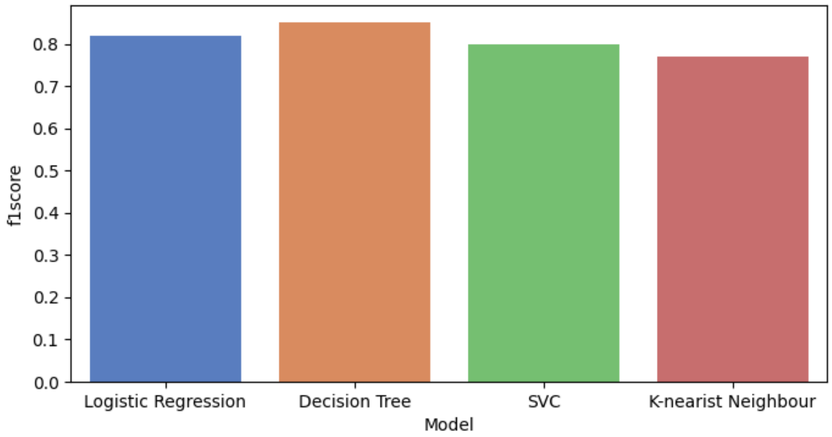


Figure 7: F1 Score Graph

Model	Precision
Decision Tree	0.82
KNN	0.77
SVM	0.81
Logistic Regression	0.80

Table 2: Precision of different models

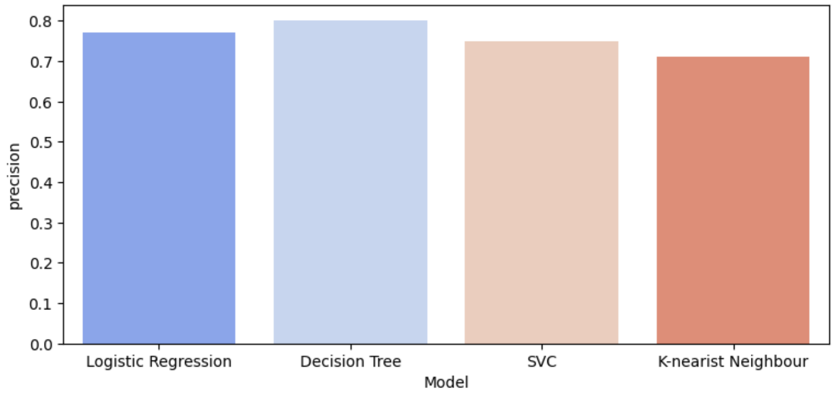


Figure 8: Precision Graph

Model	Recall
Decision Tree	0.78
KNN	0.75
SVM	0.83
Logistic Regression	0.79

Table 3: Recall of different models

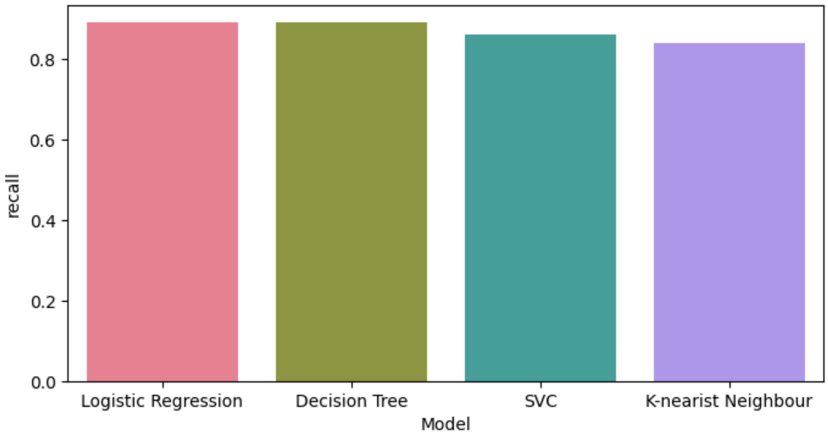


Figure 9: Recall Graph

Model	Accuracy
Decision Tree	0.79
KNN	0.76
SVM	0.82
Logistic Regression	0.78

Table 4: Accuracy of different models

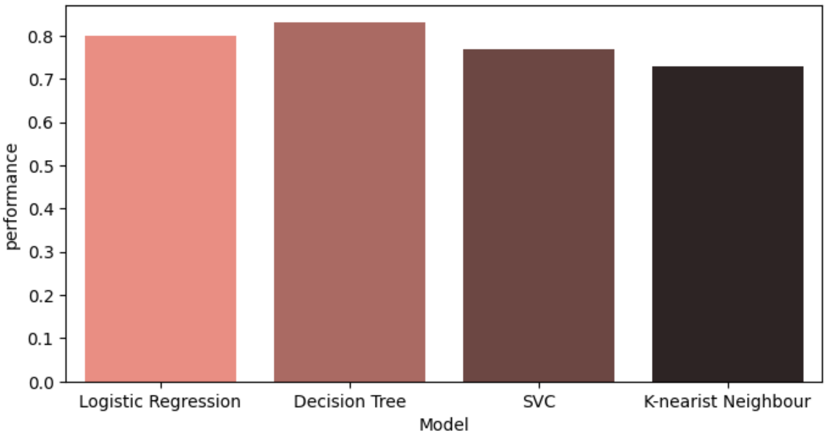


Figure 10: Accuracy Graph

8 Result

The results of the models indicate that the SVM model out perform all the model. The following table summarizes the performance metrics of all models:

Model	Recall	Precision	Accuracy	F1 Score
Decision Tree	0.78	0.82	0.79	0.80
KNN	0.75	0.77	0.76	0.76
SVM	0.83	0.81	0.82	0.82
Logistic Regression	0.79	0.80	0.78	0.79

Table 5: Performance metrics of different models

9 Graphical User Interface (GUI)

To enhance the usability of the stroke prediction system, a Graphical User Interface (GUI) is developed using Python's Tkinter library. The GUI provides an intuitive and user-friendly way for healthcare professionals and users to input patient data and receive stroke risk predictions.

9.1 GUI Design

The GUI is designed to allow users to input the required attributes, such as age, gender, hypertension status, heart disease status, etc. The interface includes input fields, dropdown menus, and buttons for submitting the data and displaying the prediction results.

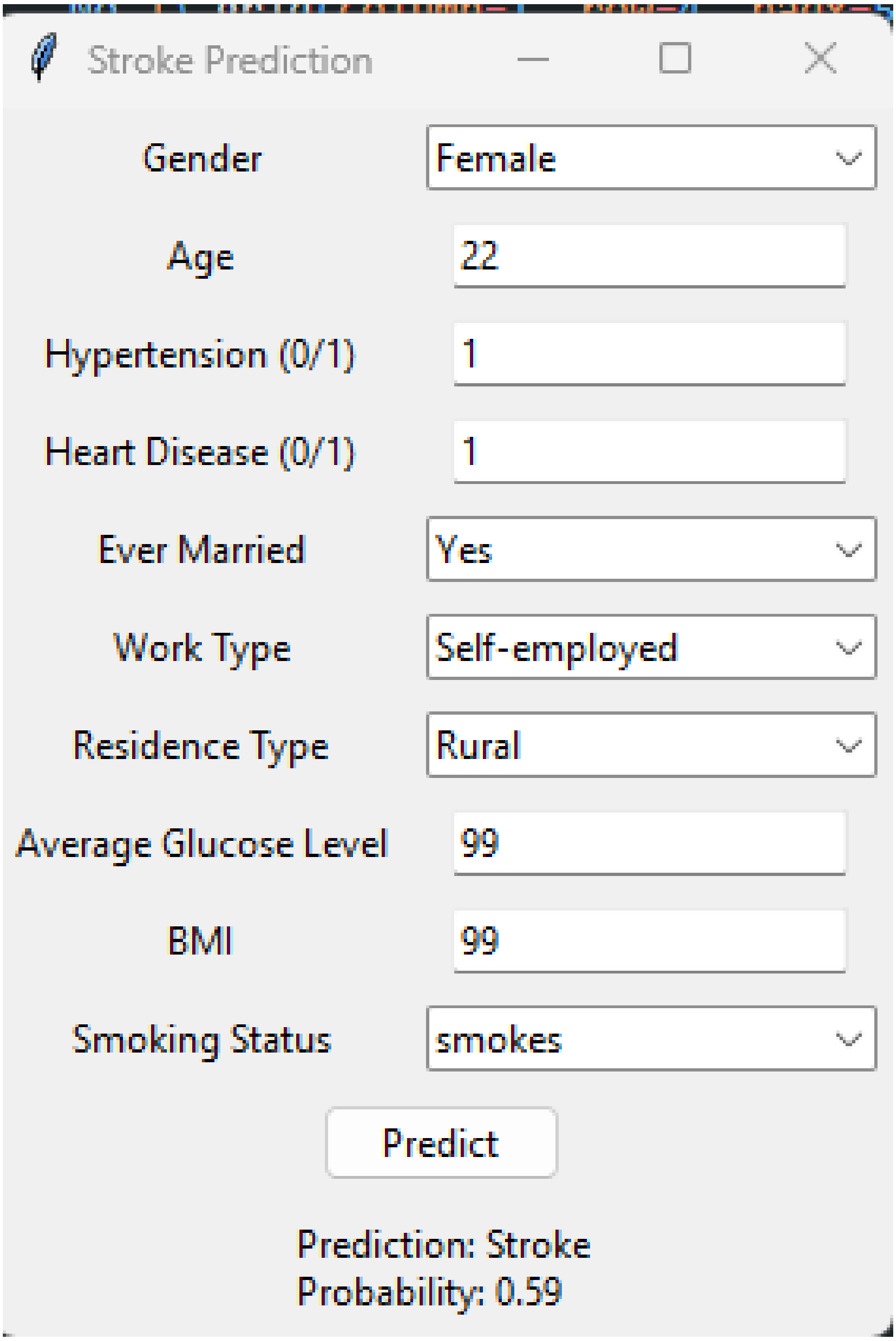
9.1.1 Features

The GUI includes features such as:

- Input validation to ensure that users provide valid data.
- Clear and user-friendly layout for easy data entry.
- Instant prediction results displayed in a message box.
- Ability to reset the input fields for new predictions.

9.2 User Experience

The GUI significantly enhances the user experience by providing an easy-to-use interface for making stroke risk predictions. Healthcare professionals can quickly input patient data and receive immediate feedback, enabling timely intervention and decision-making.



Gender: Female

Age: 22

Hypertension (0/1): 1

Heart Disease (0/1): 1

Ever Married: Yes

Work Type: Self-employed

Residence Type: Rural

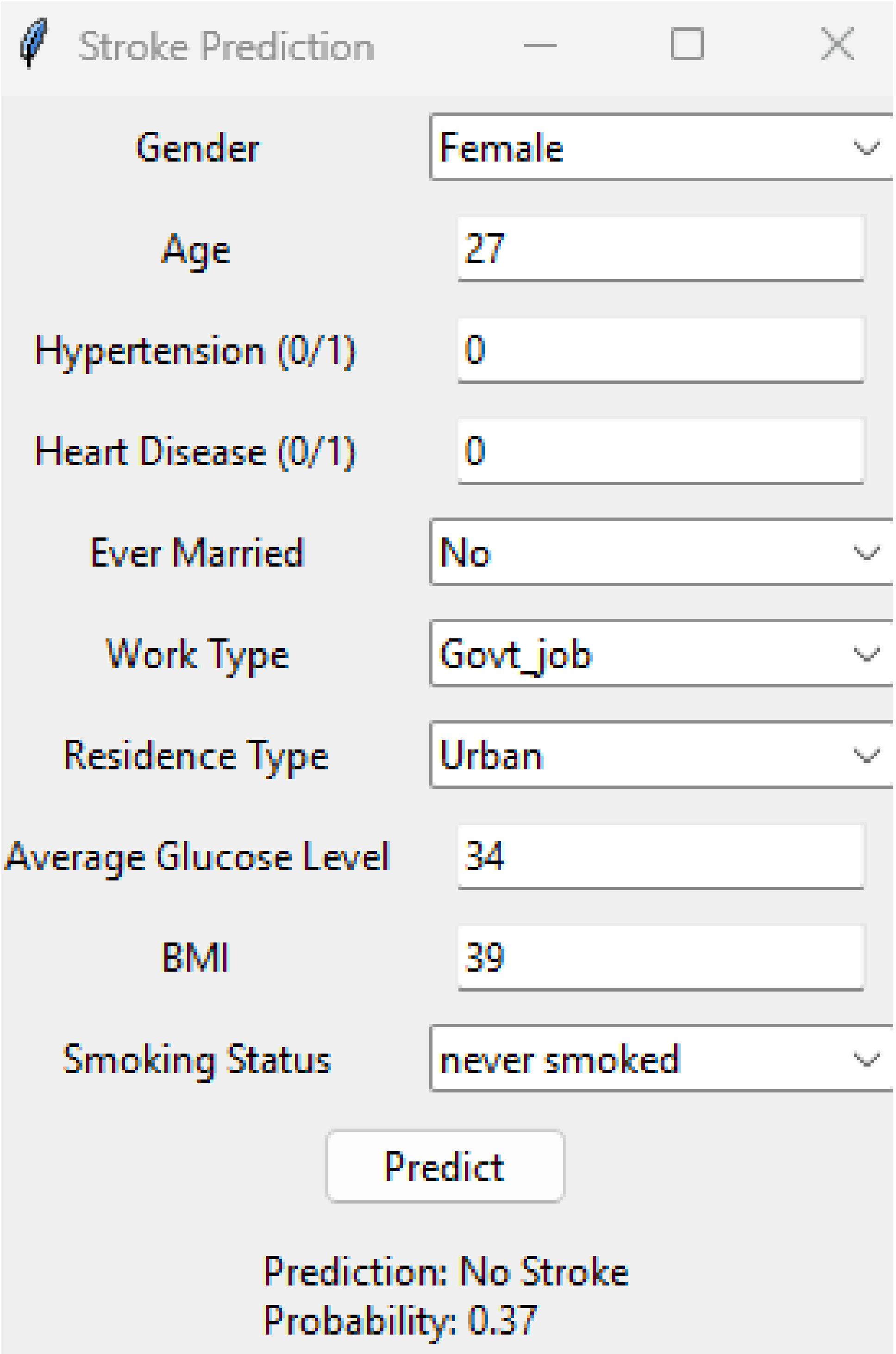
Average Glucose Level: 99

BMI: 99

Smoking Status: smokes

Prediction: Stroke
Probability: 0.59

Figure 11: Screenshot of the GUI



Gender: Female

Age: 27

Hypertension (0/1): 0

Heart Disease (0/1): 0

Ever Married: No

Work Type: Govt_job

Residence Type: Urban

Average Glucose Level: 34

BMI: 39

Smoking Status: never smoked

Prediction: No Stroke
Probability: 0.37

Figure 12: Screenshot of the GUI

10 Analysis

10.1 Support Vector Machine (SVM's Superior Performance in Stroke Prediction)

The support vector machine (SVM) algorithm emerged as the most effective model for predicting strokes within the dataset. SVM demonstrated exceptional performance attributed to its ability to effectively handle complex relationships between various patient attributes. This capability is crucial in medical datasets where the interplay between factors contributing to stroke occurrence can be intricate and non-linear. Additionally, SVM's robustness to overfitting, particularly notable in datasets with a relatively small sample size like ours, ensures better generalization to unseen data. In the context of our dataset, which includes multiple features such as age, hypertension, heart disease, and lifestyle factors, SVM's tolerance to irrelevant features and focus on informative ones further enhance predictive accuracy. Moreover, SVM's optimization objective of maximizing the margin between stroke and non-stroke instances enables it to discern subtle patterns and achieve better discrimination between classes. These factors collectively contribute to SVM's superior performance compared to other models evaluated in stroke prediction, including decision trees, k-nearest neighbors (KNN), and logistic regression.

10.2 Challenges Faced by Alternative Models in Stroke Prediction

10.2.1 Decision Tree

While Support Vector Machine (SVM) emerged as the most effective model for stroke prediction in our dataset, alternative models faced challenges that limited their performance. Decision trees, for instance, struggled with overfitting due to the complexity of the dataset. The intricate relationships between various patient attributes may have led decision trees to create overly complex decision boundaries, resulting in reduced generalization performance. Additionally, decision trees might not have effectively handled the high-dimensional feature space present in our dataset, which could have further exacerbated the risk of overfitting.

10.2.2 KNN

Similarly, K-Nearest Neighbors (KNN) encountered difficulties associated with the curse of dimensionality, especially given the high-dimensional nature of dataset. KNN relies on the proximity of data points in feature space for classification, and in datasets with a small number of instances like ours, this proximity measure may not generalize well. Moreover, KNN's computational complexity during inference could have posed challenges, particularly in scenarios where real-time predictions are required.

10.2.3 Logistic Regression

Logistic Regression, on the other hand, might have struggled to capture the non-linear relationships between features as effectively as SVM. In medical datasets with diverse patient attributes, the linear nature of logistic regression models may limit their ability to discern complex patterns. Additionally, logistic regression's performance could have been impacted by imbalanced classes in the dataset, whereas SVM's inherent ability to handle such scenarios may have provided an advantage.

10.2.4 Summary

In summary, while Decision trees, KNN, and Logistic Regression are viable machine learning approaches, they faced challenges in the dataset that limited their performance compared to SVM. Factors such as overfitting, sensitivity to high-dimensional data, and limitations in capturing non-linear relationships contributed to their comparatively lower effectiveness in stroke prediction.

11 Conclusion

In conclusion, based on the dataset, the SVM model emerged as the top performer in predicting strokes, showcasing superior accuracy and reliability. Its robustness to handle complex relationships and tolerance to irrelevant features make it a valuable tool for stroke prediction tasks.