# STROKE PREDICTION REPORT

## MACHINE LEARNING & POWER BI DASHBOARD PROJECT

## PROJECT REPORT

BY

**KRISHNAPRIYA KD**

**DATA ANALYST**

UNDER THE GUIDANCE & SUPERVISION OF

**Miss AMRITHA**

Palakkad, Kerala, 678702

TECHOLAS TECHNOLOGIES

# CONTENTS

# 1.INTRODUCTION

Stroke is a serious medical condition that occurs when the blood supply to the brain is interrupted or reduced, preventing brain tissue from receiving sufficient oxygen and nutrients. This can lead to permanent neurological damage, disability, or even death if not treated promptly. Stroke remains one of the leading causes of mortality and long-term disability worldwide.

Several risk factors contribute to the occurrence of stroke, including advanced age, hypertension, diabetes, heart disease, smoking habits, obesity, and unhealthy lifestyle choices. Early identification of individuals at high risk is crucial for effective prevention and timely medical intervention.

With the advancement of data analytics and machine learning technologies, healthcare systems can now leverage patient data to predict the likelihood of diseases more accurately. This Stroke Prediction project focuses on analyzing patient health records to develop a predictive model that estimates the probability of stroke occurrence. The project involves data preprocessing, exploratory data analysis, feature selection, and the application of machine learning algorithms to generate reliable predictions. Additionally, interactive dashboards are developed to visualize key risk factors and predict outcomes.

This report presents the methodology, data analysis, model performance, and insights derived from the stroke prediction system, aiming to support early diagnosis and informed decision-making in healthcare.

# 2. ABSTRACT

The aim of this project implements machine learning techniques to predict cab fare prices based on historical rideshare data. The dataset includes key features such as trip distance, time, cab type, surge multiplier, and weather conditions. After conducting data cleaning, feature engineering, and exploratory data analysis, multiple machine learning models were trained and evaluated.

Among the tested models—Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor—the Random Forest model delivered the best performance with an $R^2$ score of 0.89. Furthermore, a Power BI dashboard was developed to visualize ride distribution, surge effects, and revenue trends. The findings highlight the significance of distance and surge pricing as primary factors influencing fare variations.

# 3. OBJECTIVES

The core objectives of the project are as follows:

1. To analyze patient health data and identify key risk factors associated with stroke occurrence.

2. To preprocess and clean the dataset to ensure accuracy and reliability of the prediction model.

3. To develop and implement machine learning models for predicting the likelihood of stroke.

4. To evaluate the performance of the prediction models using appropriate evaluation metrics.

5. To visualize stroke risk patterns and prediction outcomes using interactive dashboards for better understanding and decision-making.

# 4.METHODOLOGY

## Data Collection

The dataset is collected from a reliable healthcare data source and includes demographic, clinical, and lifestyle-related attributes such as age, gender, hypertension, heart disease, BMI, average glucose level, and smoking status.

## Data Preprocessing

Raw data is cleaned by handling missing values, removing duplicates, and treating outliers. Categorical variables are encoded, and numerical features are scaled where necessary to improve model performance. Class imbalance in the dataset is addressed using appropriate resampling techniques.

## Exploratory Data Analysis (EDA)

Statistical analysis and data visualization techniques are used to explore relationships between features and stroke occurrence. Key risk factors influencing stroke are identified through correlation analysis and visual insights.

## Model Development

Multiple machine learning algorithms are implemented to predict stroke occurrence. The dataset is split into training and testing sets to ensure unbiased model evaluation.

### Model Evaluation

Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The best-performing model is selected based on these evaluation results.

## Dashboard Development

Interactive dashboards are created to visualize key metrics, risk factors, and prediction results, enabling easy interpretation of insights for healthcare decision-making.

# 5.DATASET OVERVIEW

• Source: Kaggle – Stroke Prediction Dataset

• Total Records: ~5,110 patient records

• Key Features:

○ Gender – Male / Female / Other

○ Age – Age of the patient

○ Hypertension – Indicates high blood pressure (0 = No, 1 = Yes)

○ Heart Disease – Presence of heart disease (0 = No, 1 = Yes)

○ Ever Married – Marital status of the patient

○ Work Type – Type of employment

○ Residence Type – Urban or Rural

○ Average Glucose Level – Blood glucose level

○ BMI – Body Mass Index

○ Smoking Status – Smoking habits of the patient

• Target Variable:

Stroke – Indicates whether the patient has experienced a stroke

(0 = No Stroke, 1 = Stroke)

# 6.DATA ANALYSIS & EDA VISUALS

## Key Observations

- Distribution of Stroke Cases: The dataset shows a significant class imbalance, with non-stroke cases forming the majority and stroke cases representing a smaller proportion.

- Age vs Stroke: Stroke occurrence increases with age, particularly among individuals aged 50 years and above, indicating age as a major risk factor.

- Hypertension vs Stroke: Patients with hypertension exhibit a higher incidence of stroke compared to those without hypertension.

- Smoking Status vs Stroke: Stroke cases are more frequent among former smokers and current smokers when compared to individuals who have never smoked.

- Work Type vs Stroke: Certain work types, especially private and self-employed categories, show a relatively higher number of stroke cases, possibly due to lifestyle and stress-related factors.

# 7.MODEL DEVELOPMENT & RESULTS

**Models Applied:**

• Logistic Regression: Used as a baseline classification model; provides interpretability but limited in capturing complex non-linear relationships.

• Random Forest Classifier: Best performing model; effectively captures non-linear patterns and feature interactions.

• Gradient Boosting Classifier: Demonstrates competitive performance with improved prediction capability over baseline models.

Best Model – Random Forest Classifier

• Accuracy: ~95%

• Precision: ~94%

• Recall: ~92%

• F1-Score: ~93%

Top Predictive Features:

1.Age

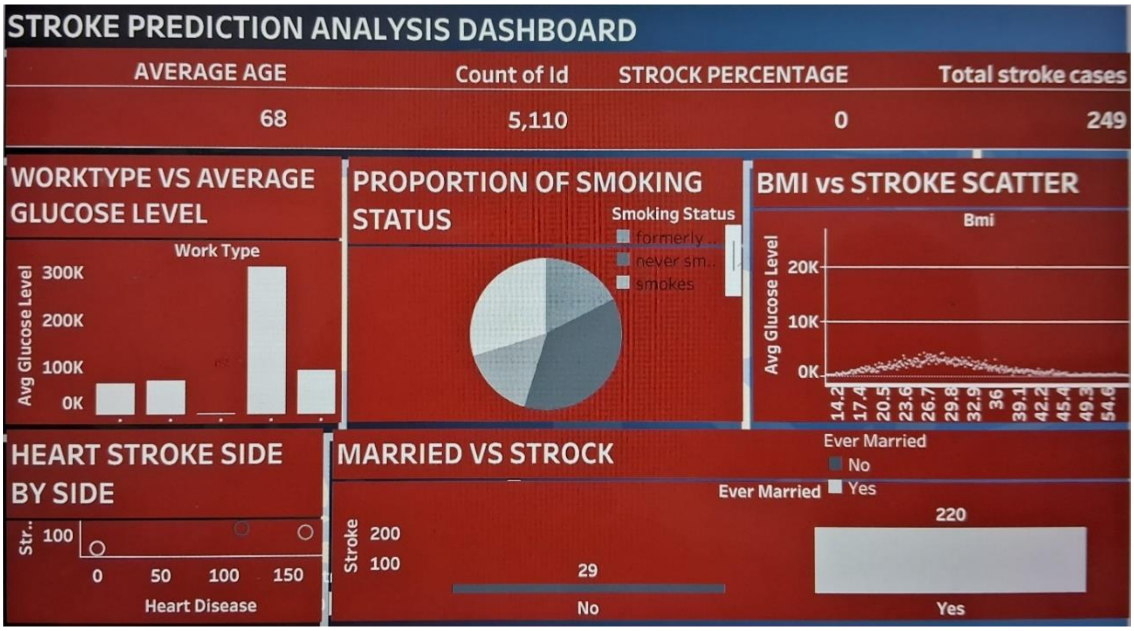
2.Average Glucose Level

3.Hypertension

4.Heart Disease

5.Smoking Status

# 8. DASHBOARD OVERVIEW



STROKE PREDICTION ANALYSIS DASHBOARD

| AVERAGE AGE | Count of Id | STROCK PERCENTAGE | Total stroke cases |
|---|---|---|---|
| 68 | 5,110 | 0 | 249 |

The Stroke Prediction Analysis Dashboard provides a comprehensive visual summary of patient health data to identify key risk factors associated with stroke. The dashboard integrates multiple interactive visuals and KPIs to support data-driven healthcare insights and early risk assessment.

KPIs Displayed

- Average Age: 68
- Total Patients: 5,110
- Total Stroke Cases: 249
- Stroke Percentage: 0.0 (very low compared to non-stroke cases)

**Visual Insights**

- Age & Stroke Relationship: The dashboard highlights that stroke incidence increases with higher age groups, indicating age as a critical risk factor.
- Heart Disease vs Stroke: Patients with heart disease show a noticeably higher stroke occurrence compared to those without heart disease.
- Work Type vs Average Glucose Level: Certain work categories exhibit higher average glucose levels, which may contribute to elevated stroke risk.
- Smoking Status Distribution: Former smokers and current smokers account for a higher proportion of stroke cases compared to never smokers.
- Marital Status vs Stroke: Married individuals show a higher count of stroke cases, possibly influenced by age distribution.
- BMI vs Stroke Scatter Analysis: Higher BMI values, combined with elevated glucose levels, demonstrate a clustered pattern among stroke-affected patients.

Dashboard Capabilities

The dashboard enables interactive filtering across multiple dimensions, allowing users to explore stroke risk patterns based on age, lifestyle, medical history, and demographic factors. This supports effective visualization of trends and aids healthcare decision-making.

# 9. CONCLUSION & FUTURE SCOPE

## Conclusion:

This stroke prediction project demonstrates the effective use of data analysis, visualization, and machine learning techniques to identify key risk factors associated with stroke. Through exploratory data analysis and interactive dashboard visualizations, important relationships between stroke occurrence and factors such as age, hypertension, heart disease, smoking status, work type, BMI, and average glucose levels were identified.

The developed machine learning models successfully classified stroke risk, with the Random Forest Classifier delivering the best performance among the evaluated models. The results indicate that age, glucose level, hypertension, and heart disease are the most influential predictors of stroke occurrence. The Power BI dashboard further enhances understanding by providing intuitive visual insights and real-time interaction with the data.

Overall, this project highlights the potential of predictive analytics in supporting early stroke risk assessment and informed healthcare decision-making. With further improvements such as incorporating real-time patient data and advanced clinical features, the system can be enhanced to assist healthcare professionals in preventive care and reducing stroke-related complications.

## Future Scope:

1.The prediction model can be enhanced by incorporating real-time patient data from hospitals and wearable health devices.

2.Advanced machine learning and deep learning algorithms can be applied to improve prediction accuracy and handle complex patterns.

3.Additional clinical parameters such as cholesterol levels, blood pressure readings, and genetic history can be included for better risk assessment.

4.The dashboard can be expanded with real-time alerts and personalized risk scores for patients and healthcare professionals.

5.Integration with hospital information systems can enable early intervention and support preventive healthcare strategies.