

# **Digital Egypt Pioneers**

## **Healthcare Predictive Analytics Project (WEQAYA)**

**Submitted By:**

**Abdelrahman Aboraya**

**Ahmed Mohammed Adel**

**Alaa Mohammed Rehab**

**Ayman Ashraf**

**Hagar Ibrahim Elemam**

**Mazin Atif**

**Supervised By**

**Eng: Basma Reda**

# Table of Contents

---

## **1. Introduction**

---

1.1. Project Overview

---

## **2. Problem Statement**

## **3. Project Motivation**

## **4. Objectives**

## **5. Why This Project Is Valuable and Unique**

## **6. Methodology**

---

6.1 Data Collection

6.2 Data Exploration

6.3 Data preprocessing

6.3 Exploratory Data Analysis (EDA)

6.4 Model Selection and Training

---

## **7. Description of Models**

---

7.1. Hypertension Prediction Model

---

7.1.1 Problem Statement & Solution

7.1.2 Introduction & Model Flow

7.1.3 Methodology

---

7.2. Stroke Prediction Model

---

7.2.1 Problem Statement & Solution

7.2.2 Introduction & Model Flow

7.2.3 Methodology

---

7.3. Diabetes Prediction Model

---

7.3.1 Problem Statement & Solution

7.3.2 Introduction & Model Flow

---

7.3.3 Methodology

---

**8. MLOPS, Deployment, and Monitoring**

**9. Future Enhancement**

**10. Conclusion**

**11. References**

# Healthcare Predictive Analytics Project

---

## 1. Project Overview

The Healthcare Prediction System is a machine learning-based solution designed to assist in the early detection and risk prediction of common yet critical health conditions:

**Hypertension, Stroke, and Diabetes.** The system leverages three specialized predictive models, each trained to identify potential risks based on patient data. By providing timely and accurate predictions, this project aims to support individuals and healthcare professionals in making proactive and informed decisions.

## 2. Problem Statement

Chronic diseases such as hypertension, stroke, and diabetes are prevalent and pose significant challenges to public health worldwide. Early detection can substantially reduce complications and mortality rates. However, the absence of efficient predictive tools often delays diagnosis, leading to severe health consequences and increased healthcare costs.

The primary problem addressed by this project is the **lack of accessible, data-driven tools for predicting chronic disease risks**. This gap leaves many individuals unaware of their health status until symptoms become severe.

## 3. Project Motivation

This project is driven by the goal of enhancing preventive healthcare using machine learning. By analyzing individual health indicators, the system can predict the likelihood of developing one of the three conditions. Early prediction enables timely medical consultation, lifestyle adjustments, and more effective health management.

## 4. Objectives

1. Early Risk Prediction: Develop machine learning models to predict the likelihood of hypertension, stroke, and diabetes.

2. **Improved Healthcare Accessibility:** Create a tool that empowers users to monitor their health risk using personal data.
3. **Data-Driven Insights:** Leverage patient data to generate reliable predictions and actionable insights.
4. **Scalable and Adaptable:** Design a flexible framework that can integrate additional models and diseases.

## **5. Why This Project Is Valuable and Unique**

1. **Multiple Condition Coverage:** Unlike many existing models that focus on a single condition, this project targets three major health risks within a unified framework.
2. **Data-Driven Decision Support:** It transforms raw patient data into actionable insights using trained ML models, helping bridge the gap between data and clinical decision-making.
3. **Scalability and Accessibility:** The system is designed to be integrated into healthcare applications or mobile platforms, making it usable by a wide range of users.
4. **Customizable and Expandable:** The architecture allows for future integration of more models and health indicators.

## **6. Methodology**

### **1. Data Collection**

The project uses healthcare datasets containing patient information such as age, BMI, blood pressure, glucose levels, physical activity, and medical history. The primary datasets used include:

- **Hypertension Dataset:** Contains blood pressure readings and related factors.
- **Stroke Dataset:** Includes demographic, clinical, and lifestyle attributes.
- **Diabetes Dataset:** Features blood sugar levels, insulin intake, and lifestyle habits.

### **2. Data Preprocessing**

- **Data Cleaning:** Handling missing values and outliers.
- **Feature Engineering:** Extracting relevant features to enhance model accuracy.
- **Data Normalization:** Scaling data for consistent model performance.

- Train-Test Split: Dividing data to ensure unbiased evaluation.

### 3. Model Selection and Training

Three separate models were developed, one for each health condition:

- Hypertension Prediction Model (Random Forest)
- Stroke Prediction Model (Random Forest)
- Diabetes Prediction Model (Random Forest)

#### Model Training and Tuning

- Training Algorithms: Models were trained using machine learning techniques suitable for classification tasks.
- Hyperparameter Tuning: Optimized using Grid Search and Cross-Validation.
- Performance Metrics: Accuracy, Precision, Recall, and F1-Score were used to evaluate the models.

## 7.Models Description

---

### a. Hypertension Prediction Model:

#### 1. Problem Definition

##### **Problem:**

Hypertension, commonly known as high blood pressure, is a chronic medical condition in which the blood pressure in the arteries is persistently elevated. It significantly increases the risk of heart disease, stroke, and kidney failure. Often asymptomatic, hypertension can go undiagnosed for years, leading to severe health complications if not detected and managed early.

The challenge lies in predicting hypertension using available health data, as many individuals may not exhibit obvious symptoms despite being at risk. Therefore, an intelligent and data-driven model is required to predict the likelihood of hypertension based on patient-specific factors.

**Solution:**

The **Hypertension Prediction Model** leverages machine learning techniques to analyze patient data, including age, BMI, blood pressure readings, smoking status, physical activity level, and medical history. The chosen algorithm for this model is **Random Forest**, known for its robustness in classification tasks and ability to handle diverse features.

The model training process involves data preprocessing, including cleaning, normalization, and feature selection. The model is then trained on a labeled dataset, with hyperparameters optimized to improve accuracy. Performance is evaluated using metrics such as accuracy, precision, recall, and F1-score to ensure the model's reliability in real-world applications.

By accurately predicting hypertension risk, the model assists healthcare professionals in making data-driven decisions, enabling early interventions and personalized health recommendations.

## **2. Introduction & Model Flow**

This project follows a structured data science workflow to address the problem of hypertension prediction using machine learning. The flow consists of several key stages: problem identification, data collection, data exploration, preprocessing, exploratory data analysis (EDA), model development, evaluation, and final deployment.

**The project flow is summarized below:**

1. **Problem Identification** – Understanding the importance of hypertension prediction.
2. **Data Collection** – Gathering stroke-related health datasets.
3. **Data Exploration** – Analyzing dataset structure, missing values, and distributions.
4. **Data Preprocessing** – Handling missing values, duplicates, outliers, and encoding categorical data.
5. **Exploratory Data Analysis (EDA)** – Visualizing relationships between features and stroke risk.
6. **Model Development** – Training and evaluating multiple ML models.
7. **Best Model Selection** – Choosing the most accurate model for stroke prediction.

## Flowchart

[Problem] → [Data Collection] → [Data Exploration] ↓ [Data Preprocessing] → [EDA & Visualization] ↓ [Model Training] → [Model Evaluation] → [Best Model Selection]

## 3. Methodology

### 3.1 Data Collection

Datasets Used:

Primary Dataset: hypertension.csv (26,083 samples, 14 features)

#### Feature Description:

Feature	Description	Data Type
age	patient's age (in years)	Numerical
sex	patient's gender (1: male; 0: female)	Binary
cp	Chest pain type: 0: asymptomatic 1: typical angina 2: atypical angina 3: non-anginal pain	Numerical
trestbps	Resting blood pressure (in mm Hg)	Numerical
chol	Serum cholesterol in mg/dl	Numerical
fbs	if the patient's fasting blood sugar > 120 mg/dl (1: yes; 0: no)	Binary
restecg	Resting ECG results: 0: normal 1: ST-T wave abnormality (T wave inversions and/or ST elevation or depression	Binary
thalach	Maximum heart rate achieved.	Numerical
exang	Exercise induced angina (1: yes; 0: no)	Binary
oldpeak	ST depression induced by exercise relative to rest.	Numerical



<b>slope</b>	he slope of the peak exercise ST segment: 0: upsloping 1: flat 2: downsloping	Numerical
<b>ca</b>	Number of major vessels (0–3) colored by flourosopy	Numerical
<b>thal</b>	3: Normal; 6: Fixed defect; 7: Reversible defect	Numerical
<b>target</b>	Whether the patient has hypertension (1) or not (0)	Binary

## 3.2 Data Exploration

### 1. Initial Data Inspection

- **Dataset Dimensions:**  
The dataset contains 26,083 rows and 14 columns.
- **Data Types:**  
A mix of numerical and categorical data types is present. For example:
  - age, chol, trestbps, etc., are numerical.
  - sex is treated as categorical but encoded numerically (0/1).
- **Missing Values:**
  - The **sex** column contains **25** missing values.
  - All other columns are complete.
- **Duplicate Records:**
  - No duplicate rows were detected in the dataset.

### 2. Statistical Summary

- **Age Range:**
  - Minimum: 11 years, Maximum: 98 years, Mean: ~55.7 years.

- **Cholesterol (chol) Range:**
  - Min: 126, Max: 564, Mean: ~246.2 (extreme values suggest possible outliers).
- **Resting Blood Pressure (trestbps) Range:**
  - Min: 94, Max: 200, Mean: ~131.6.
- **Maximum Heart Rate Achieved (thalach) Range:**
  - Min: 71, Max: 202, Mean: ~149.7.

### 3. Categorical Data Analysis

- **Categorical Columns Identified:**
  - sex, cp (chest pain type), fbs (fasting blood sugar), restecg, exang, slope, thal, target.
- **Target Variable (target):**
  - Binary classification:
    - **0 = No hypertension disease**
    - **1 = Presence of hypertension disease**
  - Distribution is **not highly imbalanced** (approx. **54.7%** positive cases).

### Key Summary Statistics

- **Average Age:** ~55.7 years
- **Gender Encoding (0/1):** Data appears roughly balanced, pending detailed count.
- **Average Cholesterol:** ~246 mg/dL
- **Average Maximum Heart Rate (thalach):** ~149.7 bpm
- **Outliers:** Detected in chol, oldpeak, and possibly ca.

### 3.3 Data Preprocessing

#### 1. Handling Missing Values

- 25 missing values in the sex column were identified and removed.

#### 2. Removing Duplicates

- No duplicate records were found in the dataset.

#### 3. Fixing Data Errors

- No rows had negative age values, so no rows were removed for this reason.

#### 4. Handling Unknown Values

- No 'unknown' entries were found in any categorical column.

#### 5. Encoding Categorical Variables

- No object-type (string) columns were found in the cleaned dataset, so no encoding was applied.

#### 6. Outlier Detection

- Extreme values ('trestbps', 'chol', 'thalach', 'oldpeak') may require further investigation.

**Final Dataset :** After cleaning, the final dataset contains 26,058 records and 14 columns.

### 3.4 Exploratory Data Analysis (EDA)

#### Visualizations & Insights

##### 1. Relationships with Hypertension Disease Risk

**The following features show notable relationships with heart disease (target = 1):**

- **Chest Pain Type (cp):** Strong positive correlation. Patients with atypical or asymptomatic chest pain types are more likely to have heart disease.
- **Max Heart Rate (thalach):** Higher max heart rate is more common in healthy individuals (inverse relationship with heart disease).

- **Exercise-induced Angina (exang) and ST Depression (oldpeak):** Higher values indicate greater heart disease risk.
- **Number of Major Vessels (ca) and Thalassemia (thal)** also strongly contribute to disease prediction.

## 2. Work-Type Analogue Analysis (using cp as a lifestyle indicator)

- **Highest Heart Disease Prevalence:** Chest pain type 0 (typical angina) and 3 (asymptomatic).
- This implies that pain symptoms and their absence could be informative of underlying cardiovascular conditions.

## 3. Correlation Analysis

- **Strong Positive Correlation:** cp, thalach (with no disease).
- **Strong Negative Correlation:** oldpeak, exang, ca, and thal (associated with disease).
- **Weak/Neutral Correlation:** age, sex, chol, fbs.

## Advanced Data Analysis:

### 1. T-Tests for Continuous Variables:

Variable	t-stat	p-value	Insight
Age	-3.75	< 0.001	Patients with heart disease tend to be older.
Cholesterol	-13.54	< 0.001	Cholesterol levels vary significantly.
Max Heart Rate	73.59	< 0.001	Healthy group tends to have higher max heart rate.

2. Chi-Squared Tests for Categorical Variables:

Variable	$\chi^2$ -stat	p-value	Significance & Insight
Sex	0.00	1.00	No statistically significant difference.
Fasting Blood Sugar (fbs)	30.01	< 0.001	Statistically significant difference.
Resting ECG (restecg)	841.37	< 0.001	Strong association with target.
Exercise Angina (exang)	5024.81	< 0.001	Very strong relationship with heart disease.
Slope	4111.85	< 0.001	Highly significant factor.
Number of Vessels (ca)	6172.09	< 0.001	One of the most predictive features.
Thalassemia (thal)	7604.72	< 0.001	Very strong association.

Conclusion

This EDA reveals crucial factors contributing to heart disease risk. Variables like chest pain type, ST depression, exercise-induced angina, thalassemia, and number of major vessels are statistically and clinically relevant. These insights support targeted diagnostics and preventative strategies.

3.5 Model Development

Models Tested:

Model	Accuracy	Precision	Recall
Logistic Regression	87%	84%	93%
SVM	99%	98%	99%
KNN	100%	100%	100%
ANN	100%	100%	100%
Random Forest	100%	100%	100%

## 3.6 Best Model Selection

**Best Model:** Random Forest

**Reason:** Highest accuracy (100%) and recall (100%) for hypertension detection.

### Feature Importance:

Chest pain type (cp) emerges as the strongest predictor of heart disease in the model, followed closely by thalassemia (thal), maximum heart rate achieved (thalach), and number of major vessels colored by fluoroscopy (ca). Other moderately important features include ST depression (oldpeak), exercise-induced angina (exang), and cholesterol level (chol). In contrast, features like age, sex, and fasting blood sugar (fbs) show relatively lower importance in the model.

---

## b. Stroke Prediction Model:

### 1. Problem Definition

#### Problem:

Stroke is a leading cause of death and disability worldwide. Early prediction of stroke risk can significantly improve patient outcomes by enabling timely medical intervention. However, diagnosing stroke risk manually is challenging due to the complexity of contributing factors such as age, hypertension, heart disease, and lifestyle habits.

#### Solution:

This project aims to develop a machine learning model that predicts the likelihood of a stroke based on patient health data. By analyzing features such as age, BMI, glucose levels, and smoking habits, the model will help healthcare professionals identify high-risk individuals for preventive care.

### 2. Introduction & Project Flow

This project follows a structured data science workflow to address the problem of stroke prediction using machine learning. The flow consists of several key stages: problem identification, data collection, data exploration, preprocessing, exploratory data analysis (EDA), model development, evaluation, and final deployment.

The project flow is summarized below:

- 1. **Problem Identification** – Understanding the importance of stroke prediction.
- 2. **Data Collection** – Gathering stroke-related health datasets.
- 3. **Data Exploration** – Analyzing dataset structure, missing values, and distributions.
- 4. **Data Preprocessing** – Handling missing values, duplicates, outliers, and encoding categorical data.
- 5. **Exploratory Data Analysis (EDA)** – Visualizing relationships between features and stroke risk.
- 6. **Model Development** – Training and evaluating multiple ML models.
- 7. **Best Model Selection** – Choosing the most accurate model for stroke prediction.

Flowchart

[Problem] → [Data Collection] → [Data Exploration] ↓ [Data Preprocessing] → [EDA & Visualization] ↓ [Model Training] → [Model Evaluation] → [Best Model Selection]

3. Methodology

3.1 Data Collection

Datasets Used:

Primary Dataset: stroke\_data(40k,11).csv (40,000 samples, 11 features)

Secondary Dataset: healthcare-dataset-stroke-data.csv (5110 samples, 11 features)

Feature Description:

Feature	Description	Data (1) Type	Data (2) Type
Sex	Patient's gender (1: Male, 0: Female)	Binary	object
age	Patient's age (in years)	Numerical	Numerical
hypertension	Hypertension history (1: Yes, 0: No)	Binary	Binary
heart disease	Heart disease history (1: Yes, 0: No)	Binary	Binary

<b>ever married</b>	Marital status (1: Married, 0: Not married)	Binary	object
<b>work type</b>	Occupation type (0: Never worked, 1: Children, 2: Govt job, 3: Self-employed, 4: Private)	Numerical	object
<b>Residence type</b>	Living area (1: Urban, 0: Rural)	Binary	Binary
<b>Avg glucose level</b>	Average blood sugar level	Numerical	Numerical
<b>bmi</b>	Body Mass Index	Numerical	Numerical
<b>smoking status</b>	Smoking habit (1: Smokes, 0: Never smoked)	Binary	object
<b>stroke</b>	Stroke occurrence (1: Yes, 0: No)	Binary (Target)	Binary (Target)

### Data Merging:

Transformed categorical data into numerical values in df2 to match df1's structure, then combined both datasets after ensuring consistent

## 3.2 Data Exploration

### 1. Initial Data Inspection:

Checked dataset dimensions.

Verified data types.

Identified missing values — Missing Values: 201 entries in bmi.

Detected duplicates — Duplicates: 245 rows.



## 2. Statistical Summary:

Used `df.describe()` for numerical features:

Age Range: -9 to 103 years (negative age indicates data error).

BMI Range: 11.5 to 92 (extreme outliers present)

## 3. Categorical Data Analysis:

Value counts for gender, hypertension, heart disease, smoking status, etc.

Target variable "stroke" shows class imbalance (46.4% stroke cases)

### Key Statistics

- The dataset has a balanced distribution of genders (54.3% male, 45.7% female).
- The average age of patients is 51 years.
- 20.6% of patients have hypertension, and 12.2% have heart disease.
- 81.6% of patients were married.
- The average glucose level is 120.98 mg/dL, and the average BMI is 30.4 (indicating obesity in many cases).
- 46.4% of patients had a stroke (target variable).

## 3.3 Data Preprocessing

### 1. Handling Missing Values

- 201 missing values in the bmi column were dropped.
- 3 missing values in the sex column were removed.

### 2. Removing Duplicates

- 245 duplicate records were identified and removed to avoid bias.

### 3. Fixing Data Errors:

Removed rows with negative age (`df = df[df['age'] >= 0]`).

#### **4. Handling Unknown Values**

- Removed rows with 'unknown' in sex and smoking\_status.

#### **5. Encoding Categorical Variables**

- Categorical columns (sex, ever married, work type, smoking status, Residence type) were converted to numerical values for modelling.

#### **6. Outlier Detection**

- Negative age values (e.g., -9) were dropped as they are biologically implausible
- Extreme BMI values (e.g., 92) may require further investigation.

#### **7. Feature Engineering**

- Normalized avg\_glucose\_level and bmi to ensure consistent scaling.
- One-hot encoding for categorical variables like (work type).

The final dataset contains 44,087 records after cleaning.

### **3.4 Exploratory Data Analysis (EDA)**

#### **Visualizations & Insights:**

##### **1. Relationships with Stroke Risk (Logistic Regression Plots):**

- Hypertension: Positive correlation – higher hypertension increases stroke probability.
- Heart Disease: Positive correlation – presence of heart disease raises stroke risk.
- Avg. Glucose Level: Positive correlation – elevated glucose levels increase stroke likelihood.

##### **2. Work Type Analysis (Histogram):**

- Highest Stroke Cases: Self-employed and private sectors.
- Possible Reason: Work-related stress or lifestyle factors.

3. Correlation Analysis

- Strong Positive Correlation: age, hypertension, heart\_disease, avg\_glucose\_level, and bmi were positively correlated with stroke.
- Weak Correlation: Residence\_type and smoking\_status had minimal impact.

Advanced Data Analysis:

1. T-Tests for Continuous Variables:

- Age: Stroke patients were significantly older (t-stat = 15.04, p < 0.001).
- Avg. Glucose Level: Higher in stroke patients (t-stat = 57.01, p < 0.001).
- BMI: Stroke group had different (often higher) BMI (t-stat = 6.02, p < 0.001).

2. Chi-Squared Tests for Categorical Variables:

- Strongest Associations: Hypertension ( $\chi^2 = 2971.92$ ) and heart disease ( $\chi^2 = 2284.64$ ).
- Other Significant Factors: Smoking status, work type, and marital status (all p < 0.05).
- Weakest but Significant: Residence type ( $\chi^2 = 5.69$ , p = 0.017).
- All categorical variables in the study are statistically correlated with stroke, as all p-values are less than 0.05, indicating a significant relationship.

**Conclusion: The EDA highlights actionable insights for stroke prevention, emphasizing medical and lifestyle interventions.**

3.5 Model Development

Models Tested:

Model	Accuracy	Precision	Recall
Logistic Regression	68%	69%	65%
Random Forest	99%	99%	99%
SVM	78%	77%	75%

KNeighborsClassifier	89%	83%	88%
----------------------	-----	-----	-----

### 3.6 Best Model Selection

Best Model: Random Forest

Reason: Highest accuracy (99.7%) and recall (99.6%) for stroke detection.

**Feature Importance:**

average glucose level emerges as the strongest predictor of stroke, followed by average BMI and hypertension, while factors like residence type and sex show relatively lower importance in the model.

---

## c. Diabetes Prediction Model:

### 1. Problem Definition

**Problem:**

Diabetes is a chronic health condition that affects millions globally, leading to severe complications such as cardiovascular disease, nerve damage, and kidney failure. Early prediction of diabetes can help in timely intervention, thereby preventing severe health outcomes.

**Solution:**

This project aims to develop a machine learning model to predict the likelihood of diabetes based on patient health data. By analyzing features such as BMI, blood pressure, cholesterol levels, physical activity, and smoking habits, the model will assist healthcare professionals in identifying high-risk individuals for preventive care.

### 2. Introduction & Project Flow

This project follows a step-by-step data science process to build a machine learning model that can predict a person's risk of having diabetes. The goal is to use health and lifestyle data to support early detection and help guide preventive care.

**Here's a breakdown of the main steps in the project:**

1-Problem Understanding – Looking at why early diabetes detection is important and how it can help improve health outcomes.

2-Data Collection – Using a large health dataset that includes information from over 250,000 individuals.

3-Data Exploration – Getting familiar with the dataset’s structure, checking for missing values, and understanding the types of features it contains.

4-Data Preprocessing – Cleaning the data by handling missing values, removing duplicates, dealing with outliers, and converting any categorical data if needed.

5-Exploratory Data Analysis (EDA) – Creating visualizations and summaries to better understand the relationships between features and diabetes risk..

6-Model Development – Training different machine learning models to predict diabetes status (non-diabetic, prediabetic, or diabetic).

7- Model Evaluation & Selection – Comparing models and selecting the one that performs best.

### 3. Methodology

#### 3.1 Data Collection

Datasets Used:

For this project, we used a large and comprehensive health dataset focused on diabetes-related risk factors. The dataset includes 253,680 records and 22 features, offering a wide range of health and demographic information.

Feature	Description
Diabetes_012	Diabetes status (0: No, 1: Prediabetes, 2: Diabetes)
HighBP	High Blood Pressure (1: Yes, 0: No)
HighChol	High Cholesterol (1: Yes, 0: No)
CholCheck	Cholesterol Check (1: Yes, 0: No)
BMI	Body Mass Index
Smoker	Smoking Status (1: Smoker, 0: Non-smoker)

<b>Stroke</b>	Stroke history (1: Yes, 0: No)
<b>HeartDiseaseorAttack</b>	Heart Disease or Attack (1: Yes, 0: No)
<b>PhysActivity</b>	Physical Activity (1: Yes, 0: No)
<b>Fruits</b>	Fruit Consumption (1: Yes, 0: No)
<b>Veggies</b>	Vegetable Consumption (1: Yes, 0: No)
<b>HvyAlcoholConsump</b>	Heavy Alcohol Consumption (1: Yes, 0: No)
<b>AnyHealthcare</b>	Access to Healthcare (1: Yes, 0: No)
<b>NoDocbcCost</b>	Could not afford a doctor (1: Yes, 0: No)
<b>GenHlth</b>	General Health (1-5, 1: Excellent, 5: Poor)
<b>MentHlth</b>	Mental Health (0-30 days of poor mental health)
<b>PhysHlth</b>	Physical Health (0-30 days of poor physical health)
<b>DiffWalk</b>	Difficulty Walking (1: Yes, 0: No)
<b>Sex</b>	Gender (1: Male, 0: Female)
<b>Age</b>	Age Category (1-13, representing age groups)
<b>Education</b>	Education Level (1-6, 1: No Schooling, 6: College Graduate)
<b>Income</b>	Income Level (1-8, 1: Lowest, 8: Highest)

### 3.2 Data Exploration

#### Key Steps:

##### 1. Initial Data Inspection:

- Checked dataset dimensions.
- Verified data types

- Detected duplicates – Duplicates: 23899 rows.

## **2. Statistical Summary:**

- Used **df.describe()** for numerical features:
  - Age Range: 1 to 13
  - BMI Range: 12 to 98
  - Mental health: 1 to 30 days

## **3. Categorical Data Analysis:**

- Analyzed value counts for Sex, HighBP, HighChol, Smoker, etc.
- Target variable 'Diabetes\_012' shows class imbalance (29.7% diabetes cases).

## **Key Statistics:**

- The dataset has a balanced distribution of genders (44% male, 56% female).
- The average BMI is 28.4, indicating a tendency towards overweight/obesity.
- 43% of patients have hypertension, and 42% have high cholesterol.
- The average age category is 8, representing middle-aged adults.

## **3.3 Data Preprocessing**

### **1. Handling Missing Values**

- No missing values

### **2. Removing Duplicates**

- 23899 duplicate records were identified and removed.

### **3. Outlier Detection**

- Extreme BMI values (e.g., above 50) were further investigated.
- Extreme PhysHlth values (e.g., close to 30 days of poor physical health)
- Extreme MentHlth values (e.g., close to 30 days of poor mental health)

### **4. Handling Imbalanced Data**

- The target variable Diabetes\_012 showed significant class imbalance:
  - 0 (No Diabetes): 190,055 samples
  - 2 (Diabetes): 35,097 samples
  - 1 (Prediabetes): 4,629 samples
- To address this imbalance, both oversampling (minority classes) and undersampling (majority class) techniques were applied.

- Among the techniques evaluated for handling imbalanced data, oversampling proved to be the most effective approach.
- This ensured the model was not biased toward the majority class and improved its ability to detect diabetic and prediabetic cases accurately.

### 3.4 Exploratory Data Analysis (EDA)

#### Visualizations & Insights:

##### 1. Relationships with Diabetes Risk:

- HighBP: Positive correlation – higher blood pressure increases diabetes risk.
- HighChol: Positive correlation – higher cholesterol increases diabetes risk.
- BMI: Positive correlation – higher BMI is associated with increased diabetes risk.

##### 2. Correlation Analysis:

- Strong Positive Correlation: BMI, HighBP, HighChol, and Age show strong correlation with diabetes.
- Weak Correlation: Fruits and Veggies consumption showed minimal impact.

### 3.5 Model Development

The following machine learning models were tested for diabetes prediction, along with their respective performance metrics:

Model	Accuracy	Precision	Recall
Logistic Regression	51%	50%	51%
Random Forest	93%	93%	93%
Xgboost	63%	63%	63%
MLP	53%	52%	51%

### 3.6 Best Model Selection

Best Model: Random Forest

- Highest accuracy (93%) and recall (93%) for diabetes prediction.



**Feature Importance:**

- BMI, HighBP, and HighChol are the top predictors of diabetes.
  - Less important features include Fruits and Veggies consumption.
- 

## 8. MLOPS, Deployment, and Monitoring

### Deployment Hypertension, Stroke, and Diabetes Prediction Models:

We have successfully deployed machine learning models for predicting the risk of Hypertension, Stroke, and Diabetes using **Streamlit**, an open-source Python library that allows for the rapid development and deployment of interactive web applications. Each model has its own dedicated **Streamlit app** with an intuitive UI for user inputs, visualization of risk factors, and real-time prediction results.

#### 1. Hypertension Model Deployment :

The deployed application serves to predict the likelihood of a patient having hypertension based on several health-related features using a pre-trained machine learning model.

#### Key Features

##### 1. User Interface:

- Built using **Streamlit** for quick interaction and ease of use.
- Title, header image, and description to introduce the purpose of the app.

##### 2. Data Visualization:

- Sidebar option to select and view risk factors such as:
  - Chest Pain Type
  - Serum Cholesterol
  - Maximum Heart Rate
  - ST Depression
- Visuals rendered using **Plotly**, **Seaborn**, and **Matplotlib**.

### 3. User Inputs:

- Form-based input collection with widgets like slider and selectbox.
- Inputs include age, gender, cholesterol level, ECG results, heart rate, and more.
- Converted into numerical format suitable for model inference.

### 4. Model Integration:

- Preprocessing using a **Standard Scaler** (scaler.pkl).
- Prediction using a trained **classification model** (model.pkl).
- Output includes both class prediction (High Risk or Low Risk) and probability score.

### 5. Dynamic Feedback:

- Displays probability and classification outcome.
- Shows appropriate warning or success message with a relevant image.
- Guidance tips for users based on the prediction outcome.

### 6. Performance Optimization:

- Caching of visualization functions using @st.cache\_resource to improve app responsiveness.

## 2. Stroke Prediction Model:

### Providing insights into medical decision-making.

1. Prioritizing glucose and blood pressure control: The results indicate that average glucose level and hypertension are the strongest predictors of stroke, emphasizing the importance of regular monitoring and management of these factors in high-risk patients.
2. Focusing on patients with heart disease: Individuals with a history of heart disease are significantly more likely to experience a stroke, making cardiovascular conditions a critical risk factor in stroke risk assessment.
3. Targeting older age groups for preventive care: Stroke incidence increases substantially after the age of 60, with a clear peak between 80–100 years,

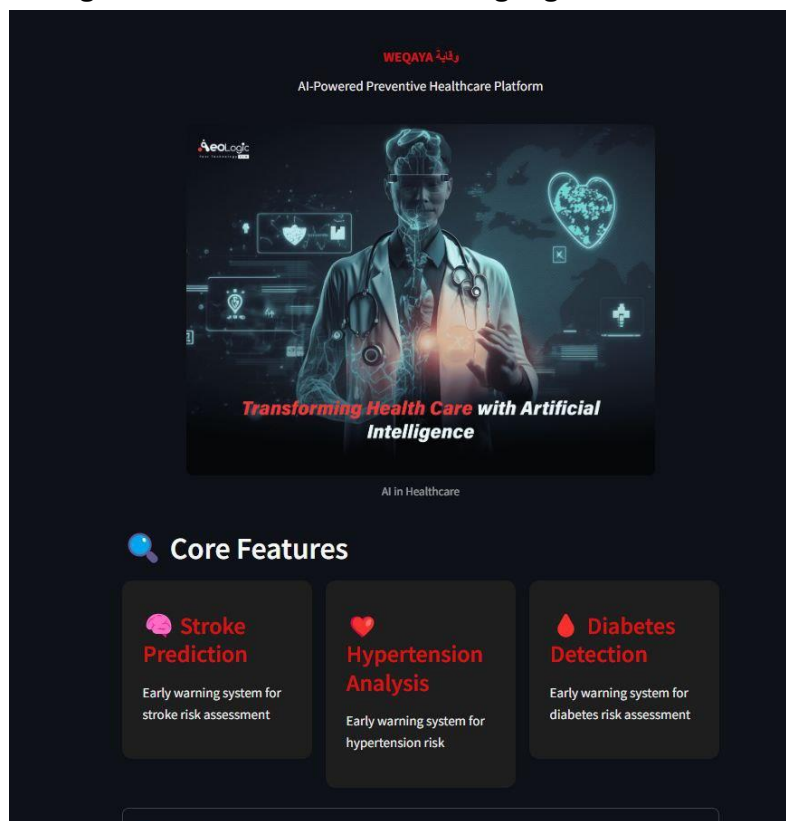
highlighting the need for early screening and continuous monitoring in elderly populations.

4. Considering occupational stress as a potential risk: Higher stroke rates were observed among individuals working in the private and self-employed sectors, suggesting that work-related stress and lifestyle factors may contribute to stroke risk and warrant preventive interventions.
5. Integrating multiple factors for accurate prediction: Combining variables such as age, glucose level, hypertension, and heart disease provides a more reliable assessment of stroke risk than evaluating factors individually, supporting the use of multivariate predictive models to enhance clinical decision-making.

### 3.Diabetes Prediction Model:

#### Deployment Plan:

- Built a Streamlit web app for diabetes risk prediction.
- Utilizes Random Forest for real-time predictions based on user inputs.
- Integrated data visualizations to highlight risk factors.



## Stroke Prediction System



Real-time visualization of a deep brain stroke — emphasizing the urgency of early detection and prevention.

### Enter Patient Information

Select Gender

male

Select Hypertension

Not Had hypertension

Select Ever Married

No

Select Heart Disease

Not Had heart\_disease

Select Smoking Status

Never smoke

Select Age

50

Select BMI

20

Select Avg. Glucose Level

150

1

100

10

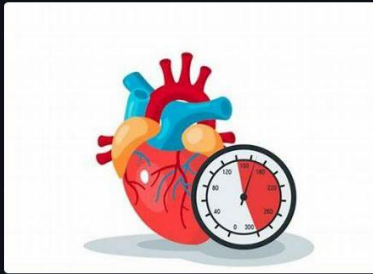
35

50

270

predict

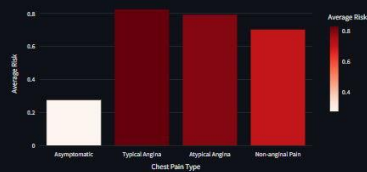
## Hypertension Prediction System

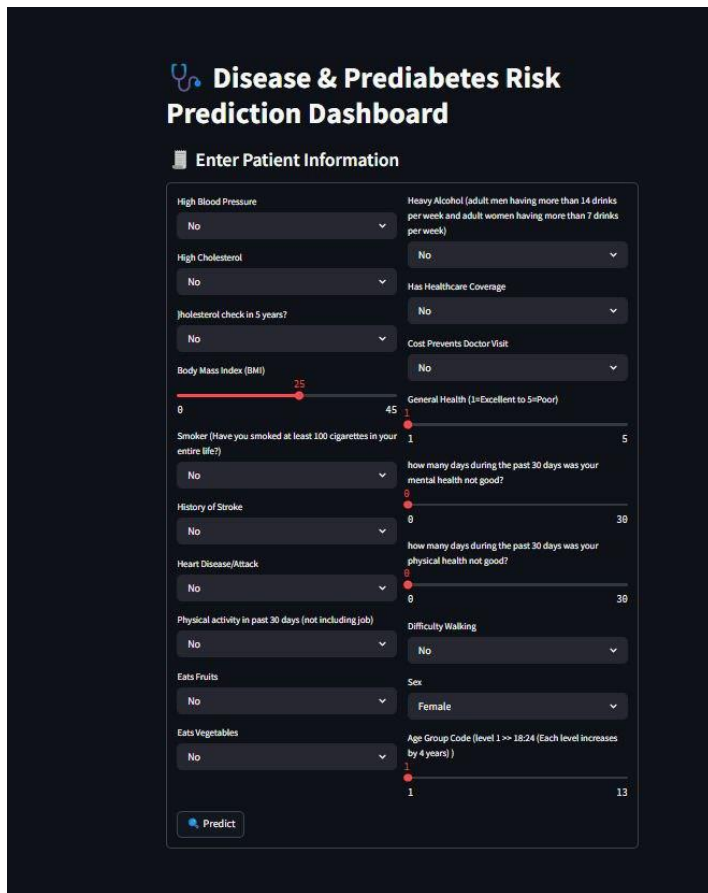


Early detection and prevention of hypertension is key to a healthier life.

Welcome to the Hypertension Prediction System! This app uses machine learning to predict the likelihood of hypertension based on various health factors. Please enter your information below and get insights about your health risk.

### Average Hypertension Risk by Chest Pain Type





**Disease & Prediabetes Risk Prediction Dashboard**

**Enter Patient Information**

High Blood Pressure No	Heavy Alcohol (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) No
High Cholesterol No	Has Healthcare Coverage No
Cholesterol check in 5 years? No	Cost Prevents Doctor Visit No
Body Mass Index (BMI) 25	General Health (1=Excellent to 5=Poor) 1
Smoker (Have you smoked at least 100 cigarettes in your entire life?) No	how many days during the past 30 days was your mental health not good? 0
History of Stroke No	how many days during the past 30 days was your physical health not good? 0
Heart Disease/Attack No	Difficulty Walking No
Physical activity in past 30 days (not including job) No	Sex Female
Eats Fruits No	Age Group Code (level 1 == 18-24 (Each level increases by 4 years) ) 1
Eats Vegetables No	

**Predict**

## Next Step:

## 2.Mlflow for tracking model performance

### Hypertension Predictive Model

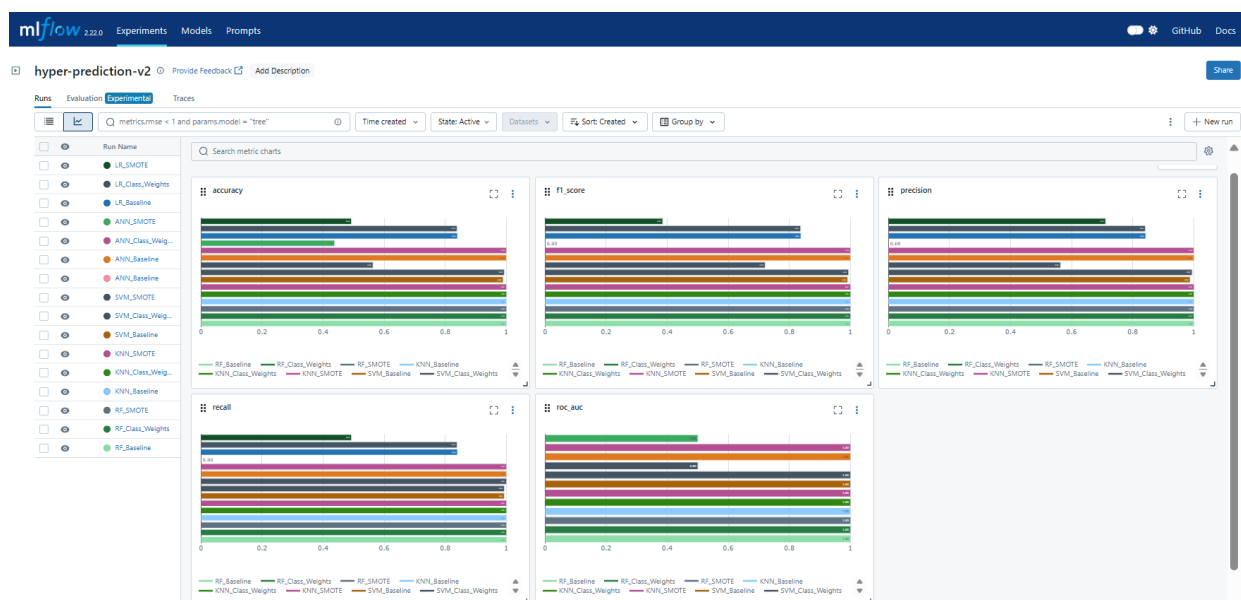
the performance of four machine learning models (Random Forest, Logistic Regression, SVM, ANN, and KNN ) applied to predict Hypertension risk, evaluated using MLflow for tracking experiments (model performance).

All four models were trained and evaluated using three distinct strategies to address class imbalance:

1. Baseline: Original imbalanced data (no adjustments).
2. Class Weights: Algorithmic weighting of minority class during training.
3. SMOTE: Synthetic Minority Oversampling (balanced dataset generation).

## Key findings:

- Best Model: Random Forest, KNN (SMOTE & Baseline) achieved 100% accuracy and 100% AUC, demonstrating superior predictive power.
- Top Alternatives:
  - SVM (SMOTE & Baseline & Class Weight) : 99% accuracy, 98% AUC (balanced performance).
  - ANN (Class Weights & BaseLine): 99% accuracy, high recall (99.4%) but lower precision (98%).
- Weakest Model: Logistic Regression underperformed (49% accuracy), likely due to sensitivity to imbalanced data.



hyper-prediction-v2

[Provide Feedback](#)
[Add Description](#)

Share

Runs

Experimental

Traces

metrics.rmse < 1 and params.model = "tree"

Time created

State: Active

Datasets

Sort: Created

Columns

Group by

+ New run

	Run Name	Created	Duration	Metrics				Parameters							
				accuracy	f1_score	precision	recall	roc_auc	C	hidden_layer_size	kernel	max_depth	max_iter	n_estimators	n_neighbors
<input type="checkbox"/>	LR_SMOTE	31 minutes ago	9.1s	0.49202453...	0.38467183...	0.71075354...	0.49202453...	-	2	-	-	-	300	-	-
<input type="checkbox"/>	LR_Class_Weights	31 minutes ago	9.2s	0.83824130...	0.83637512...	0.84081214...	0.83824130...	-	0.5	-	-	-	100	-	-
<input type="checkbox"/>	LR_Baseline	32 minutes ago	11.1s	0.83926380...	0.83734943...	0.84205993...	0.83926380...	-	1.0	-	-	-	100	-	-
<input type="checkbox"/>	ANN_SMOTE	34 minutes ago	11.4s	0.43701431...	0	0	0	0.50015119...	-	(100, 50)	-	-	500	-	-
<input type="checkbox"/>	ANN_Class_Weights	35 minutes ago	48.6s	0.99979550...	0.99981841...	0.99963689...	1	1	-	(100, 50)	-	-	500	-	-
<input type="checkbox"/>	ANN_Baseline	36 minutes ago	54.2s	0.99979550...	0.99981841...	0.99963689...	1	1	-	(100, 50)	-	-	500	-	-
<input type="checkbox"/>	ANN_Baseline	37 minutes ago	-	-	-	-	-	-	-	-	-	-	-	-	-
<input type="checkbox"/>	SVM_SMOTE	44 minutes ago	48.1s	0.56298568...	0.72039774...	0.56298568...	1	0.5	1.0	-	rbf	-	-	-	-
<input type="checkbox"/>	SVM_Class_Weights	45 minutes ago	42.0s	0.99284253...	0.99363983...	0.99418181...	0.99309843...	0.99987727...	1.0	-	rbf	-	-	-	-
<input type="checkbox"/>	SVM_Baseline	46 minutes ago	38.2s	0.98916155...	0.99040028...	0.98771676...	0.99309843...	0.99988101...	1.0	-	rbf	-	-	-	-
<input type="checkbox"/>	KNN_SMOTE	52 minutes ago	10.4s	1	1	1	1	1	-	-	-	-	-	-	15
<input type="checkbox"/>	KNN_Class_Weights	52 minutes ago	10.4s	1	1	1	1	1	-	-	-	-	-	-	15
<input type="checkbox"/>	KNN_Baseline	52 minutes ago	11.9s	1	1	1	1	1	-	-	-	-	-	-	15
<input type="checkbox"/>	RF_SMOTE	1 hour ago	10.3s	1	1	1	1	0.99999999...	-	-	-	12	-	150	-
<input type="checkbox"/>	RF_Class_Weights	1 hour ago	9.8s	1	1	1	1	1	-	-	-	12	-	150	-
<input type="checkbox"/>	RF_Baseline	1 hour ago	11.8s	1	1	1	1	1	-	-	-	12	-	150	-

## Stroke Predictive Model

the performance of four machine learning models (Random Forest, XGBoost, SVM, and KNN) applied to predict stroke risk, evaluated using MLflow for tracking experiments (model performance).

All four models were trained and evaluated using three distinct strategies to address class imbalance:

4. Baseline: Original imbalanced data (no adjustments).
5. Class Weights: Algorithmic weighting of minority class during training.
6. SMOTE: Synthetic Minority Oversampling (balanced dataset generation).

Key findings:

- **Best Model:** XGBoost (SMOTE & Baseline) achieved 97.5% accuracy and 99.6% AUC, demonstrating superior predictive power.
- **Top Alternatives:**
  - Random Forest (Class Weights): 94.3% accuracy, 98% AUC (balanced performance).

- KNN (Class Weights): 86.7% accuracy, high recall (96.4%) but lower precision (79.5%).
- Weakest Model: SVM underperformed (77% accuracy), likely due to sensitivity to imbalanced data.

**stroke-prediction-v2** [Provide Feedback](#) [Add Description](#)

Runs Evaluation **Experimental** Traces

metrics.rmse < 1 and params.model = "tree" Time created State: Active Datasets

Metrics						
Run Name	Created	accuracy	f1_score	precision	recall	roc_auc
SVM_SMOTE	1 hour ago	0.77026208...	0.75799888...	0.73999065...	0.77690546...	0.85592224...
SVM_Class_Weights	1 hour ago	0.76905014...	0.75986453...	0.73279659...	0.78900883...	0.85492393...
SVM_Baseline	1 hour ago	0.76783820...	0.74176425...	0.76490008...	0.71998691...	0.85513082...
KNN_SMOTE	2 hours ago	0.85820330...	0.86377528...	0.77805453...	0.97072293...	0.95586684...
KNN_Class_Weights	2 hours ago	0.86789880...	0.87108219...	0.79471270...	0.96368989...	0.95547670...
KNN_Baseline	2 hours ago	0.86789880...	0.87108219...	0.79471270...	0.96368989...	0.95547670...
XGB_SMOTE	2 hours ago	0.97485229...	0.97356687...	0.94849519...	1	0.99644358...
XGB_Class_Weights	2 hours ago	0.97015603...	0.96874504...	0.94054220...	0.99869152...	0.99597986...
XGB_Baseline	2 hours ago	0.97500378...	0.97365479...	0.95102932...	0.99738305...	0.99620227...
RF_SMOTE	2 hours ago	0.93713073...	0.93123446...	0.94358629...	0.91920183...	0.98680190...
RF_Class_Weights	2 hours ago	0.94341766...	0.93892568...	0.93869543...	0.93915603...	0.98966210...
RF_Baseline	2 hours ago	0.93069231...	0.92222694...	0.96000707...	0.88730781...	0.98959113...

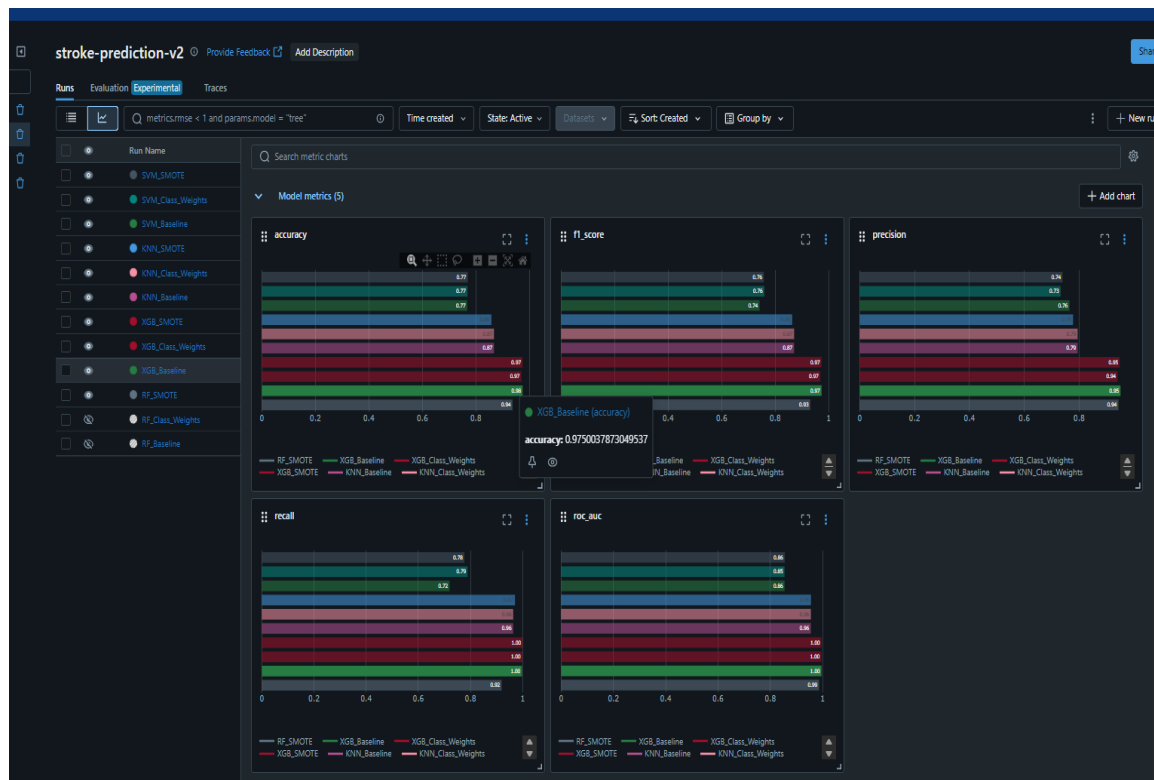
## Strategy Comparison

- **SMOTE**: Best for **XGBoost/RF** (improves recall without sacrificing precision).
- **Class Weights**: Effective for **KNN/RF**, but less impactful for **XGBoost**.
- **Baseline**: Only viable for **XGBoost**, suggesting inherent robustness to imbalance.

## Critical Observations from MLflow

- **XGBoost's Consistency**: Performed well across all strategies (accuracy range: **97.0–97.5%**).
- **SVM's Limitations**: Low metrics suggest need for hyperparameter tuning or kernel adjustments.
- **KNN's Trade-off**: High recall (**96.4%**) but low precision (**79.5%**) risks false positives.





## 1. Deploy as a web app for doctors.

Integrate with electronic health records (EHR) for real-time predictions.

Stroke Prediction System – Deployment Report:

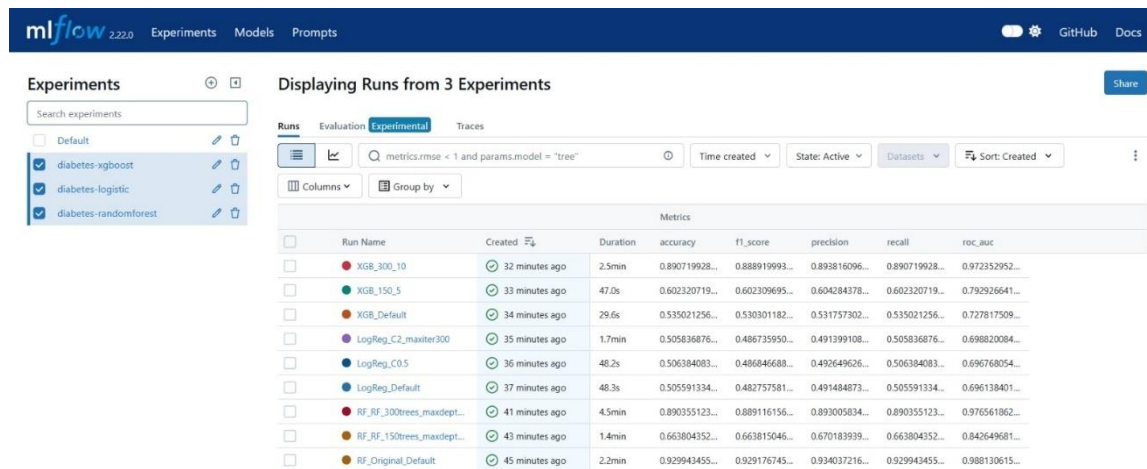
- Built with **Streamlit** to deploy an interactive web app for stroke risk prediction.
- Uses a trained **Random Forest Classifier** to predict stroke based on patient inputs like age, BMI, glucose level, hypertension, and more.
- Provides **dynamic data visualizations** using Plotly and Seaborn to show relationships between risk factors and stroke occurrence.
- Includes a **form-based user interface** for input and displays personalized prediction results with probability and advice.
- The system also offers **educational messages** and **visual alerts** to guide users after prediction.

## Diabetes Predictive Model

This MLflow experiment tracked the performance of three machine learning models—XGBoost, Random Forest, and Logistic Regression—for predicting diabetes risk. Each model was evaluated under different configurations and hyperparameters using multiple performance metrics.

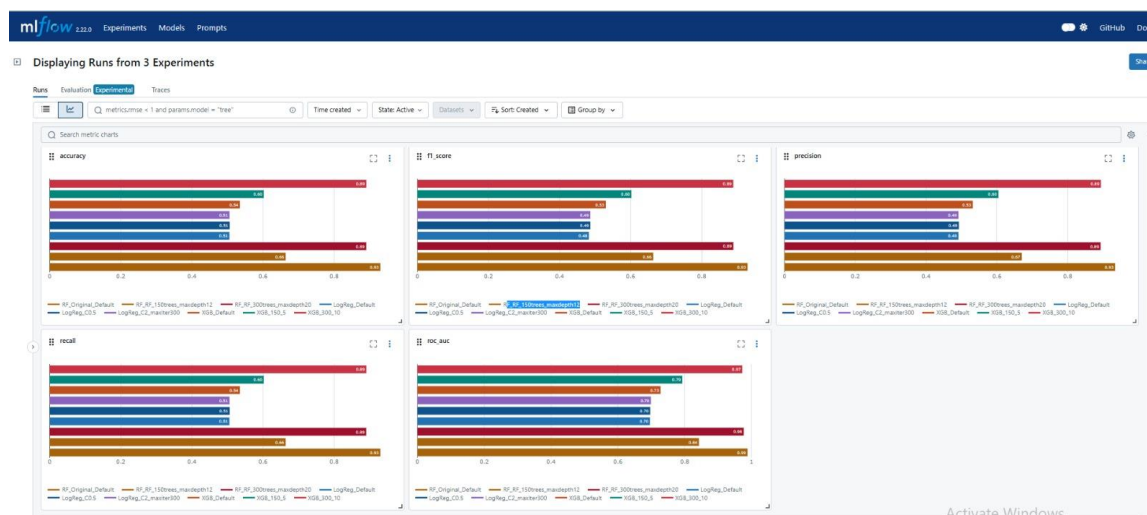
### Key Findings:

- **Best Model:**  
Random Forest (RF\_Original\_Default) delivered the highest overall performance with:
  - Accuracy: 92.99%
  - F1 Score: 0.929
  - Precision: 0.934
  - Recall: 0.929
  - AUC: 0.988
- **Top Alternative:**  
XGBoost (XGB\_300\_10) also showed strong results with:
  - Accuracy: 89.07%
  - F1 Score: 0.889
  - Precision: 0.893
  - Recall: 0.890
  - AUC: 0.972
- **Logistic Regression (LogReg\_C0.5 & LogReg\_C2\_maxiter300):**
  - Lower metrics across all categories.
  - Accuracy ranged from 50.5% to 51%, with AUC under 0.70.
  - Indicates limited ability to capture complex patterns in the data compared to ensemble models.



## Critical Observations from MLflow:

- **Random Forest's Robustness:** Performed consistently well across all tested configurations, especially in recall and AUC.
- **XGBoost's Efficiency:** Delivered near-top results with significantly shorter training time compared to Random Forest.
- **Logistic Regression's Simplicity:** Fast to train but significantly underperforms in non-linearly separable data.



## 2. Deploy as a web app for doctors.

Integrate with electronic health records (EHR) for real-time predictions.

### Diabetes Prediction System – Deployment Report:

Built with Streamlit to deploy an interactive web app for Diabetes risk prediction.

Uses a trained Random Forest Classifier to predict Diabetes based on patient inputs like Age, BMI, High Blood Pressure, High Cholesterol, and more.

Provides dynamic data visualizations using Plotly and Seaborn to show relationships between risk factors and Diabetes occurrence.

Includes a form-based user interface for input and displays personalized prediction results with probability and advice.

## 9. Refrences

Database URL : [Diabetes, Hypertension and Stroke Prediction](#)

Repository URL : [Weqaaya/DEPI\\_Graduation\\_Project: AI-powered predictive system for chronic diseases \(Hypertension, Stroke, Diabetes\) using real healthcare data. Includes ML models, Streamlit interface.](#)

Streamlit UI URL : <https://health-care-f2arfvp3dmjkcmlmmgpu4.streamlit.app/>