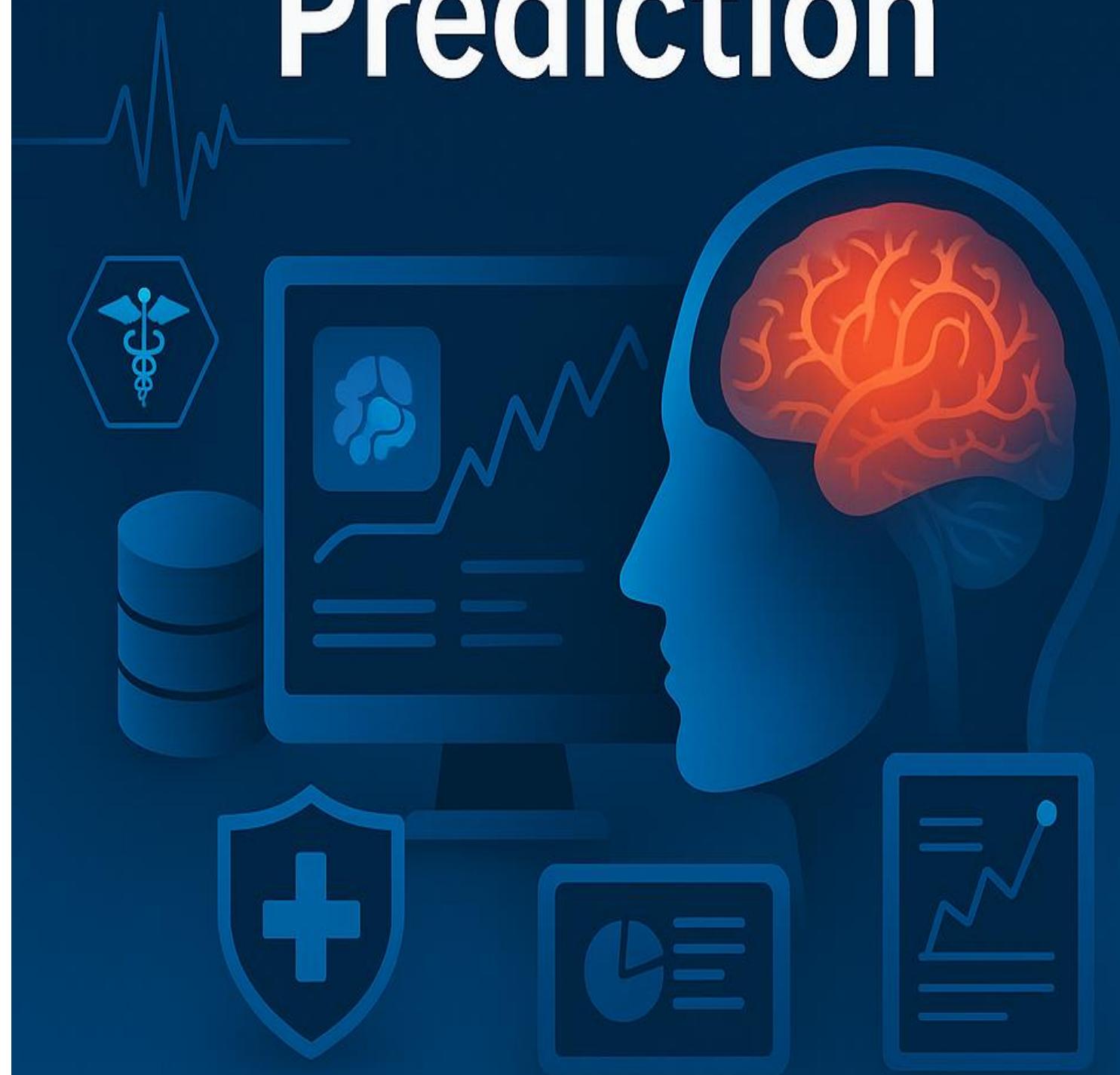




# Stroke Risk Prediction





Prepared for: **Eng. Sherif Mohamed**  
AI & Data Science Track  
CLS\_ONL3\_AIS4\_G2  
Digital Egypt Pioneers

Prepared by:

Eng. Mohamed Nasr  
Eng. Ahmed Ghanem  
Eng. Tarek El Naggar  
Eng. Ahmed Walid  
Eng. Ahmed Abd El Maksoud  
  
Dr. Doaa GadAllah

## INTRODUCTION:

- I PROBLEM DEFINITION
- II OBJECTIVES
- III DATA DESCRIPTION
- IV METHODOLOGY
- V ENVIRONMENT SETUP
- VI ARCHITECTURE DIAGRAM

## TECHNICAL PILLAR:

### VII DATA CLEANING AND QA

- 1. DATA COLLECTION AND EXPLORATION
- 2. PREPROCESSING
- 3. FEATURE PREPARATION AND SELECTION

### VIII MACHINE LEARNING MODELING

- 1. DATA SPLITTING
- 2. MODEL DEFINING
- 3. CROSS VALIDATION
- 4. MODEL EVALUATION
- 5. COMPARISON

### IX USER INTERFACE -UI

- 1. GU
- 2. AI AGENT

## BUSINESS PILLAR:

## INDEX

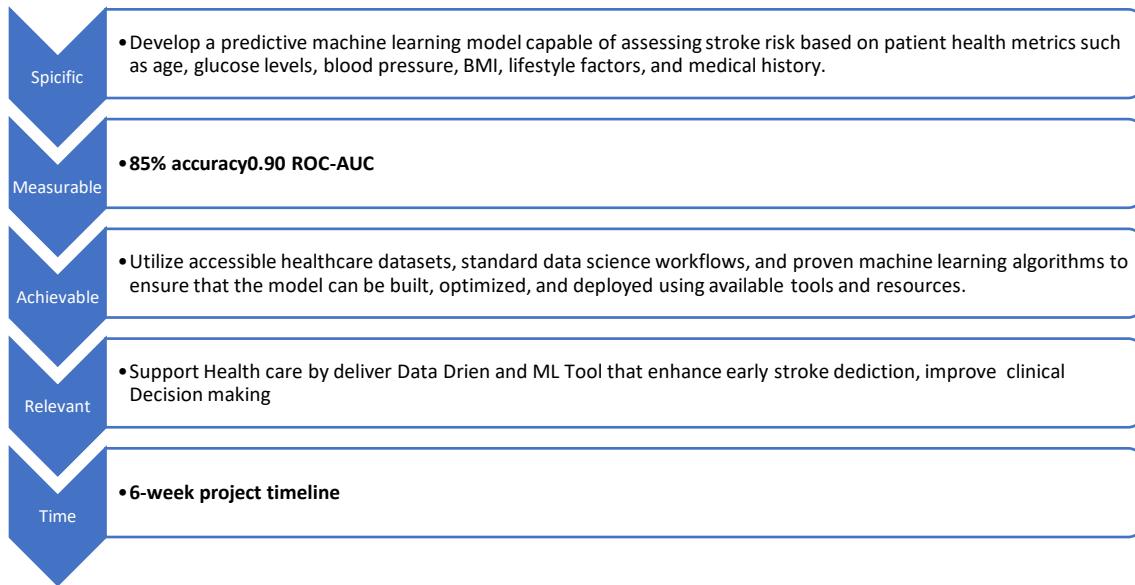
## PROBLEM DEFINITION:

Stroke is one of the leading causes of death and long-term disability worldwide. Early identification of individuals at risk can significantly reduce complications through preventive care.

This project aims to develop a machine learning model that predicts the likelihood of a person being at risk of stroke based on multiple medical and lifestyle factors. The goal is to use binary and continuous features such as symptoms, blood pressure, and age to generate a stroke risk percentage and a clear risk classification ("At Risk" vs. "Not at Risk").

## OBJECTIVES

The objective of this project is to develop a reliable, data-driven **Stroke Risk Prediction Model** that analyzes patient health metrics to accurately forecast the likelihood of stroke. This includes cleaning and understanding the dataset, uncovering key health patterns through analysis, building and optimizing predictive machine learning models, and deploying a scalable solution that supports early clinical decision-making and improves patient outcomes.

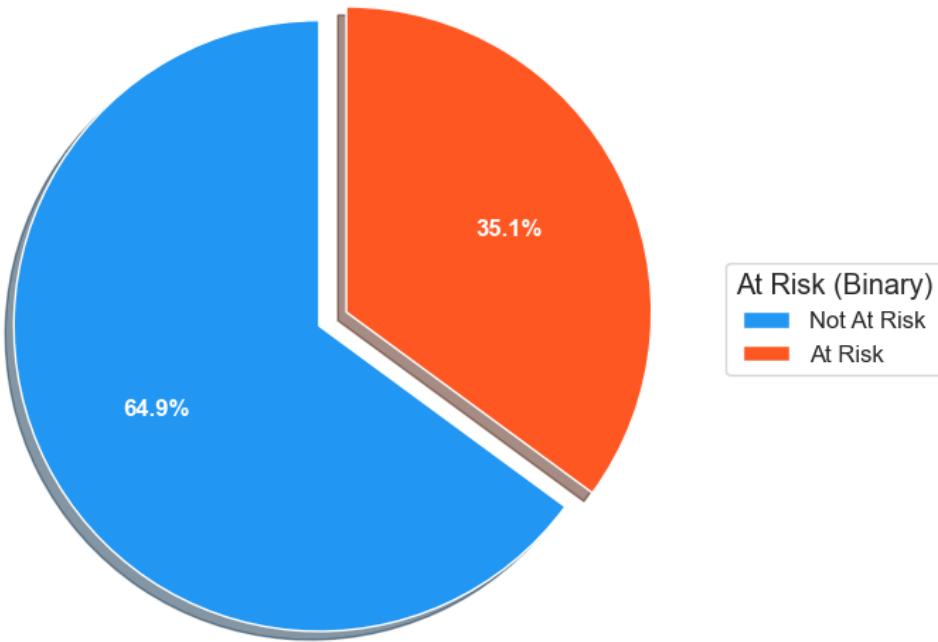


## DATA DESCRIPTION

The dataset includes binary features such as symptoms (Chest Pain, Dizziness, Sleep Apnea, etc.) and numeric attributes such as Age and Blood Pressure. The target variable is 'At Risk (Binary)'.

Feature Type	Example Variables
Binary (Yes/No)	Chest Pain, Shortness of Breath, Irregular Heartbeat, Dizziness, Fatigue, Sleep Apnea, Anxiety
Numeric	Age, Blood Pressure, Stroke Risk (%)
Target Variable	At Risk (Binary: 1 = At Risk, 0 = Not At Risk)

**Distribution of At Risk (Binary)**

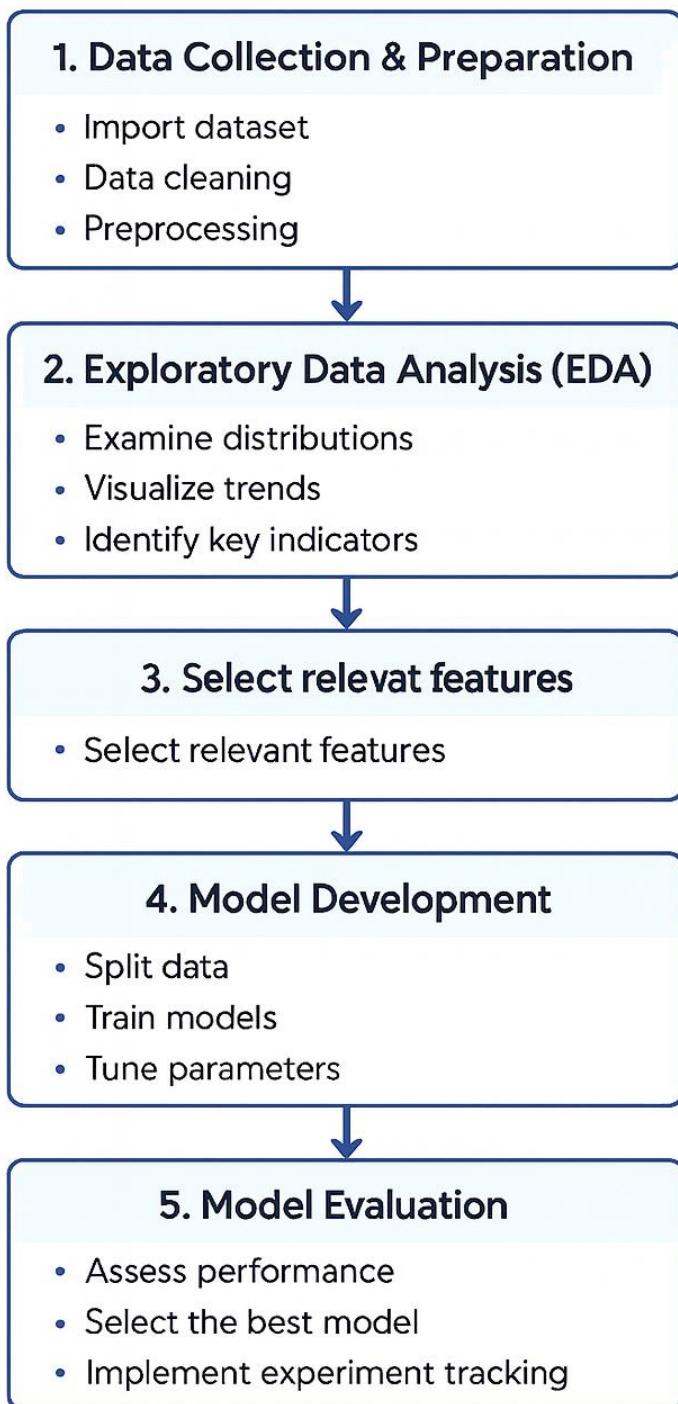


Data Source:

- **Source:** Kaggle
- **Dataset:** **Stroke Risk Prediction Dataset Based on Symptoms**
- **URL:** <https://www.kaggle.com/datasets/mahatiratusher/stroke-risk-prediction>
- **Dataset Author:** Mahatir Ahmed Tusher
- **License:** MIT
- **Local File:** stroke\_risk\_dataset.csv

## METHODOLOGY

# Methodology

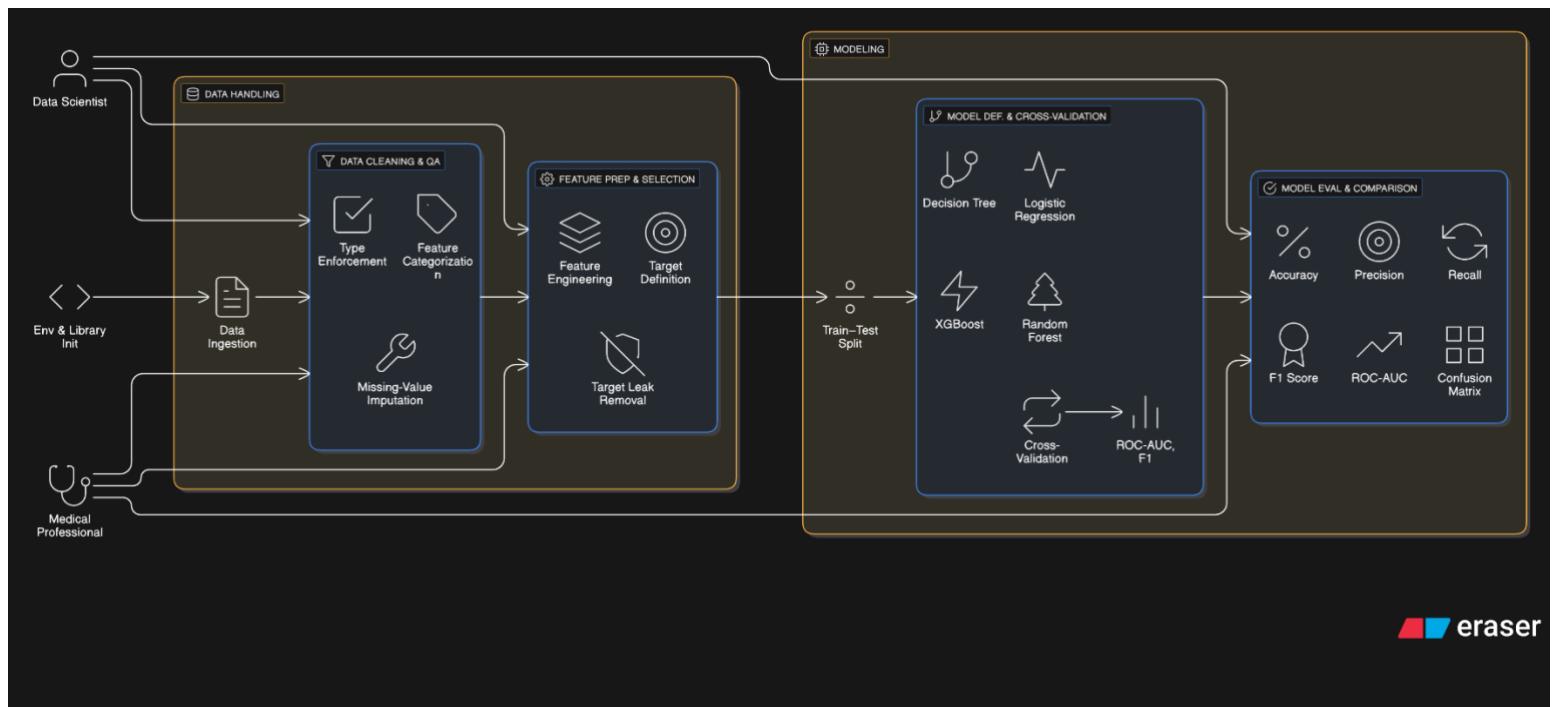


## ENVIRONMENT SETUP

Effective stroke risk prediction relies on libraries like pandas for data manipulation, numpy for numerical operations, and scikit-learn for machine learning algorithms and model evaluation.

- **pandas:** Data manipulation and analysis
- **numpy:** Numerical computing support
- **matplotlib:** Data visualization capabilities
- **scikit-learn:** Machine learning utilities
- **XGBoost:** Gradient boosting framework

## ARCHITECTURE DIAGRAM



## 1. DATA COLLECTION AND EXPLORATION

### A. DATASET OVERVIEW

The dataset loaded in the notebook is:

```
df = pd.read_csv("stroke_risk_dataset_custom_nulls.csv")
```

LOADING DATA

```
df=pd.read_csv("C:/Users/hp/Downloads/stroke_risk_dataset_custom_nulls.csv")
df
```

	Chest Pain	Shortness of Breath	Irregular Heartbeat	Fatigue & Weakness	Dizziness	Swelling (Edema)	Neck/Jaw/Shoulder/Back	Pain in Neck/Jaw/Shoulder/Back	Excessive Sweating	Persistent Cough	Nausea/Vomiting	High Blood Pressure	Chest Discomfort (Activity)	Cold Hands/Feet	Snoring/Sleep Apnea	Anxiety/Feeling of Doom	Age	Stroke Risk (%)	At Risk (Binary)
0	0.0	1.0	1.0	1.0	0.0	0.0		0.0	1.0	1.0	1.0	0.0	1.0	1.0	0	0	54	58.0	
1	0.0	0.0	1.0	0.0	0.0	1.0		0.0	0.0	0.0	0.0	1.0	0.0	1.0	1	0	49	40.5	
2	1.0	0.0	0.0	1.0	1.0	1.0		0.0	0.0	1.0	0.0	0.0	0.0	0.0	1	0	62	52.0	
3	1.0	0.0	1.0	1.0	0.0	1.0		1.0	1.0	1.0	1.0	1.0	0.0	0.0	0	0	48	60.0	
4	0.0	0.0	1.0	0.0	0.0	1.0		0.0	1.0	0.0	1.0	1.0	0.0	0.0	1	1	61	56.5	
...	...	...	...	...	...	...		...	...	...	...	...	...	...	...	...	...	...	
69995	1.0	0.0	0.0	0.0	0.0	0.0		0.0	1.0	0.0	1.0	1.0	1.0	0.0	0	1	18	30.0	
69996	0.0	0.0	0.0	1.0	0.0	1.0		0.0	1.0	0.0	0.0	0.0	1.0	1.0	1	0	24	33.0	
69997	1.0	1.0	0.0	1.0	1.0	1.0		0.0	0.0	0.0	0.0	1.0	0.0	0.0	0	0	49	45.5	
69998	0.0	1.0	1.0	1.0	0.0			0.0	0.0	0.0	0.0	0.0	1.0	1.0	1	0	45	48.5	
69999	0.0	1.0	0.0	0.0	0.0			0.0	1.0	1.0	1.0	1.0	1.0	0.0	1	0	74	63.0	

70000 rows x 18 columns

## B. DATASET CHARACTERISTICS

- Contains patient health metrics (Age, BMI, Glucose, Heart Rate, Blood Pressure, Smoking, Diabetes, etc.)
- Includes a target variable: **Stroke Risk (%)**
- Includes a binary outcome: **At Risk (Binary)**
- Contains missing values in several columns
- Includes both categorical & numerical features

## C. DATA EXPLORATION REPORT

### Initial Exploration

- .read() inspection
- .isnull() to examine column names and null
- .describe() to inspect statistical distribution
- Value counts for categorical attributes

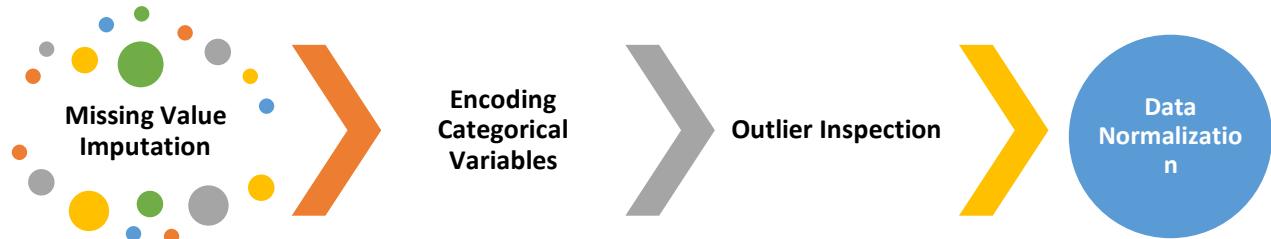
```
df.isnull().sum()
```

Chest Pain	90
Shortness of Breath	54
Irregular Heartbeat	66
Fatigue & Weakness	80
Dizziness	75
Swelling (Edema)	92
Pain in Neck/Jaw/Shoulder/Back	56
Excessive Sweating	56
Persistent Cough	85
Nausea/Vomiting	66
High Blood Pressure	56
Chest Discomfort (Activity)	88
Cold Hands/Feet	87
Snoring/Sleep Apnea	0
Anxiety/Feeling of Doom	0
Age	0
Stroke Risk (%)	0
At Risk (Binary)	0
dtype: int64	

## D. IDENTIFIED DATA QUALITY ISSUES

Issue	Columns Affected	Notes
Missing values	BMI, Glucose, Heart Rate, Blood Pressure	Required imputation
Mixed types	Categorical fields stored as strings	Required encoding
Outliers	Age, Glucose, BMI	Verified using boxplots
Skewed distributions	Glucose, Heart Rate	Addressed during EDA
Potential leakage	Some risk-related metrics correlated with target	Handled during modeling

## 2. PREPROCESSING STEPS



- Numerical columns → Imputed using **mean**
- Categorical and Binary columns → Imputed using **most frequent**
- continuous columns → Imputed using KNN Imputer

- Binary categories mapped to 0/1
- Multiclass variables encoded using LabelEncoder

- Boxplots for Age, BMI, Glucose, Heart Rate
- Extreme values acknowledged but retained for medical relevance

### Exploration after Cleaning:

.info() inspection	.describe() to inspect statistical distribution
<pre> df.info()  &lt;class 'pandas.core.frame.DataFrame'&gt; RangeIndex: 70000 entries, 0 to 69999 Data columns (total 18 columns):  #   Column           Non-Null Count  Dtype   ---   0   Chest Pain       70000 non-null   float64  1   Shortness of Breath 70000 non-null   float64  2   Irregular Heartbeat 70000 non-null   float64  3   Fatigue &amp; Weakness 70000 non-null   float64  4   Dizziness        70000 non-null   float64  5   Swelling (Edema) 70000 non-null   float64  6   Pain in Neck/Jaw/Shoulder/Back 70000 non-null   float64  7   Excessive Sweating 70000 non-null   float64  8   Persistent Cough 70000 non-null   float64  9   Nausea/Vomiting 70000 non-null   float64  10  High Blood Pressure 70000 non-null   float64  11  Chest Discomfort (Activity) 70000 non-null   float64  12  Cold Hands/Feet 70000 non-null   float64  13  Snoring/Sleep Apnea 70000 non-null   float64  14  Anxiety/Feeling of Doom 70000 non-null   float64  15  Age              70000 non-null   float64  16  Stroke Risk (%) 70000 non-null   float64  17  At Risk (Binary) 70000 non-null   float64 dtypes: float64(18) memory usage: 9.6 MB </pre>	<pre> df.describe().T        count   mean    std   min   25%   50%   75%   max Chest Pain      70000.0  0.502529  0.499997  0.0  0.0  1.0  1.0  1.0 Shortness of Breath 70000.0  0.496014  0.499988  0.0  0.0  0.0  1.0  1.0 Irregular Heartbeat 70000.0  0.498529  0.500001  0.0  0.0  0.0  1.0  1.0 Fatigue &amp; Weakness 70000.0  0.500829  0.500003  0.0  0.0  1.0  1.0  1.0 Dizziness        70000.0  0.503471  0.499992  0.0  0.0  1.0  1.0  1.0 Swelling (Edema) 70000.0  0.501229  0.500002  0.0  0.0  1.0  1.0  1.0 Pain in Neck/Jaw/Shoulder/Back 70000.0  0.498800  0.500002  0.0  0.0  0.0  1.0  1.0 Excessive Sweating 70000.0  0.504029  0.499987  0.0  0.0  1.0  1.0  1.0 Persistent Cough 70000.0  0.501329  0.500002  0.0  0.0  1.0  1.0  1.0 Nausea/Vomiting 70000.0  0.502429  0.499998  0.0  0.0  1.0  1.0  1.0 High Blood Pressure 70000.0  0.501071  0.500002  0.0  0.0  1.0  1.0  1.0 Chest Discomfort (Activity) 70000.0  0.498814  0.500002  0.0  0.0  0.0  1.0  1.0 Cold Hands/Feet 70000.0  0.498257  0.500001  0.0  0.0  0.0  1.0  1.0 Snoring/Sleep Apnea 70000.0  0.500686  0.500003  0.0  0.0  1.0  1.0  1.0 Anxiety/Feeling of Doom 70000.0  0.499871  0.500004  0.0  0.0  0.0  1.0  1.0 Age              70000.0  54.056429  21.071587  18.0  36.0  54.0  72.0  90.0 Stroke Risk (%) 70000.0  55.558771  14.300898  5.0  45.5  55.5  66.0  100.0 At Risk (Binary) 70000.0  0.649200  0.477224  0.0  0.0  1.0  1.0  1.0 </pre>

## Remove Duplication:

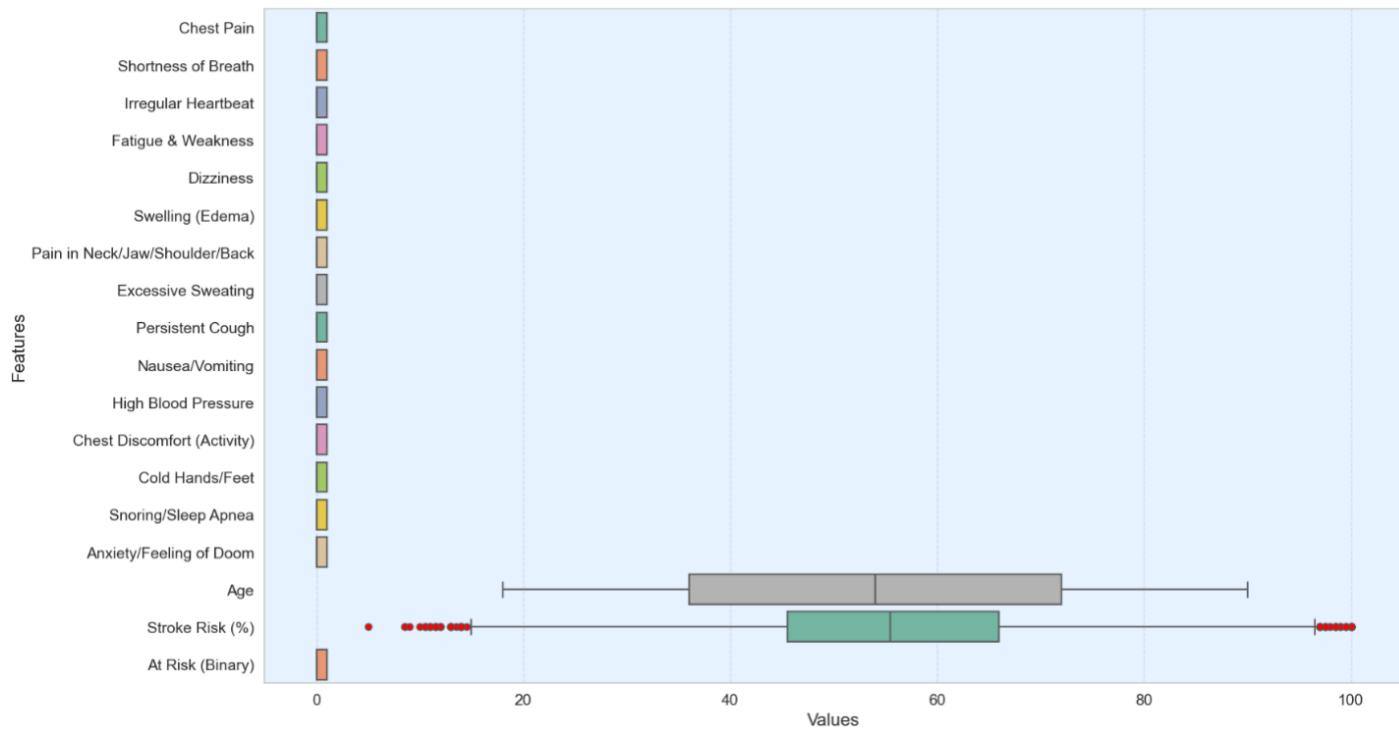
```
df.duplicated().sum()/len(df)
df.drop_duplicates(inplace=True)
df
```

	Chest Pain	Shortness of Breath	Irregular Heartbeat	Fatigue & Weakness	Dizziness	Swelling (Edema)	Pain in Neck/Jaw/Shoulder/Back	Excessive Sweating	Persistent Cough	Nausea/Vomiting	High Blood Pressure	Chest Discomfort (Activity)	Cold Hands/Feet	Snoring/Sleep Apnea	Anxiety/Feeling of Doom	Age	Stroke Risk (%)	At Risk (Binary)
0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	54.0	58.0	1.1
1	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	49.0	40.5	0.1
2	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	62.0	52.0	1.1
3	1.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	48.0	60.0	1.1
4	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	61.0	58.5	1.1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	18.0	30.0	0.1
69996	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	24.0	33.0	0.1
69997	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	49.0	45.5	0.1
69998	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	45.0	48.5	0.1
69999	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	74.0	63.0	1.1

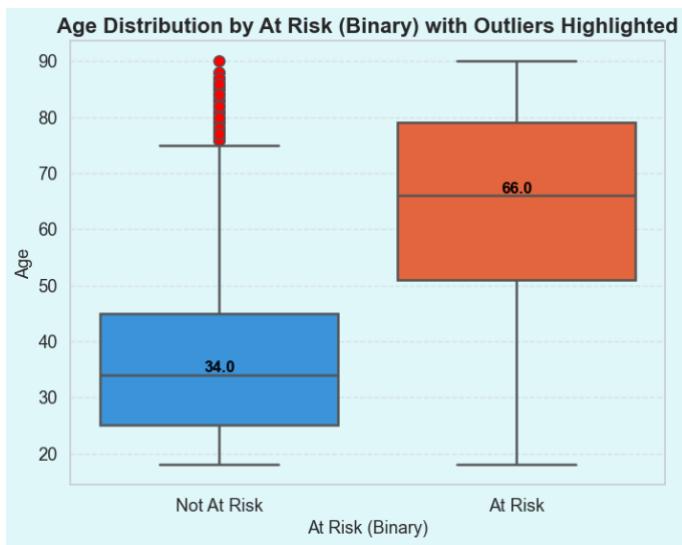
68990 rows × 18 columns

## Exploring Features:

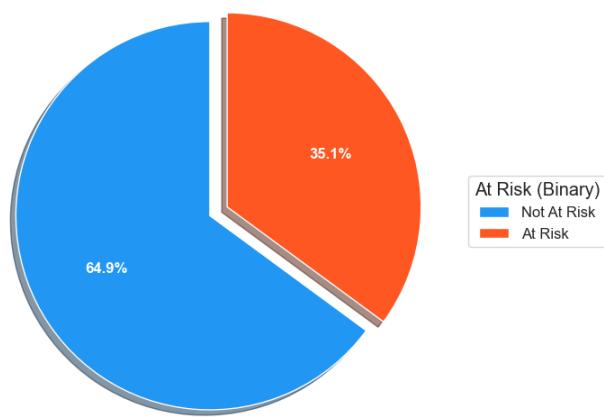
Combined Boxplot for All Numeric Features



- **Cleaned Dataset** → A cleaned dataframe df prepared after imputation & encoding



**Distribution of At Risk (Binary)**



### Insights:

Age is not only correlated with stroke risk—it **clearly divides the population into distinct risk categories**, with:

- Younger adults predominantly low-risk
- Older adults overwhelmingly high-risk

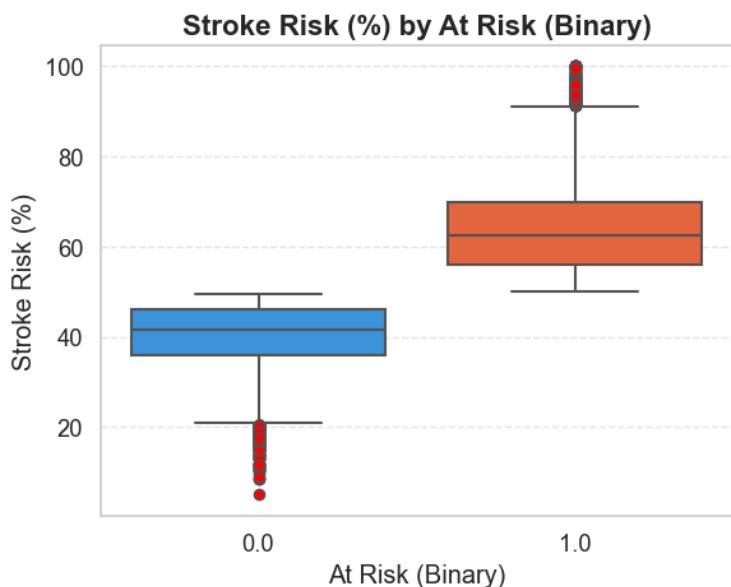
This visually reinforces Age's correlation value (0.612) and its position as the strongest predictor.

### Stroke Risk Distribution by Binary Risk Class:

The stroke-risk scoring model is **highly consistent** with the binary classification:

- Low-risk individuals rarely exceed **50%**.
- High-risk individuals rarely fall below **50%**.

This reinforces that the binary indicator mirrors the continuous risk metric accurately and can be trusted for classification modeling.

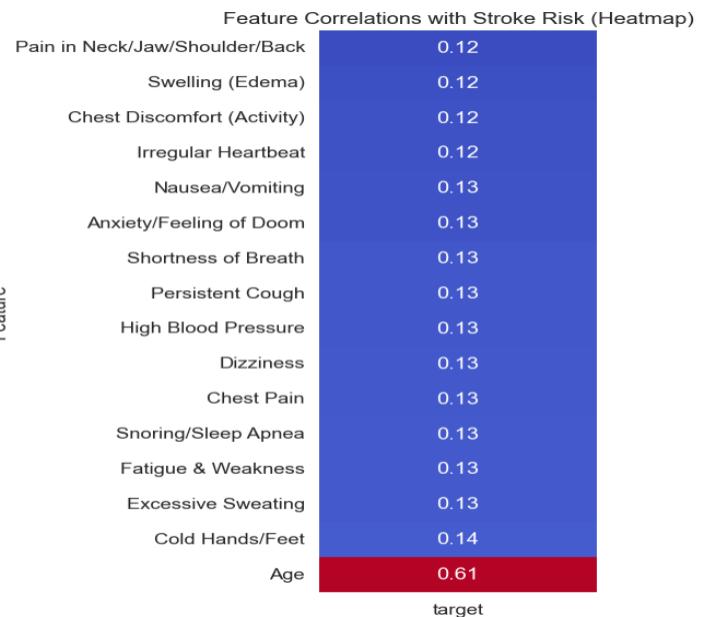


### 3. FEATURE PREPARATION AND SELECTION

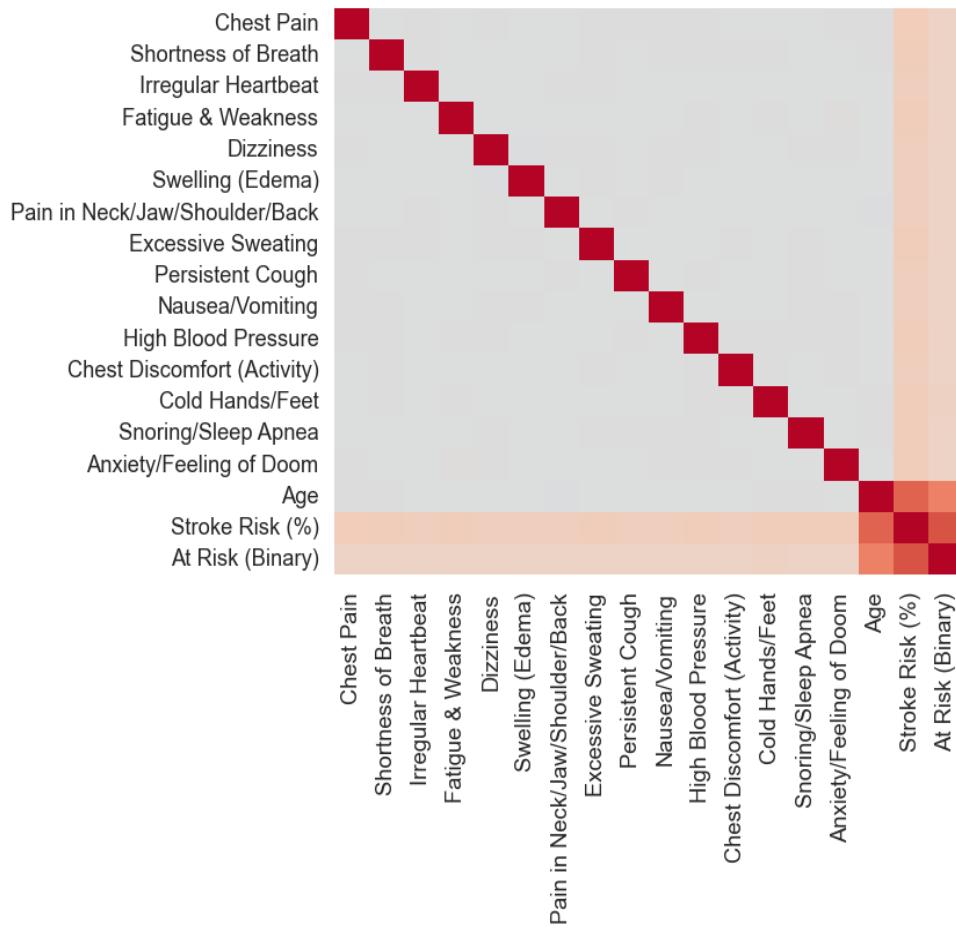
#### Correlation Heatmap:

##### Insight:

Age shows the strongest correlation with Stroke Risk (0.61), making it the most influential predictor in the dataset. All other symptoms have very weak correlations (0.12–0.14), indicating that individually they have minimal impact on stroke risk and must be analyzed collectively through a predictive model.



Correlation Matrix



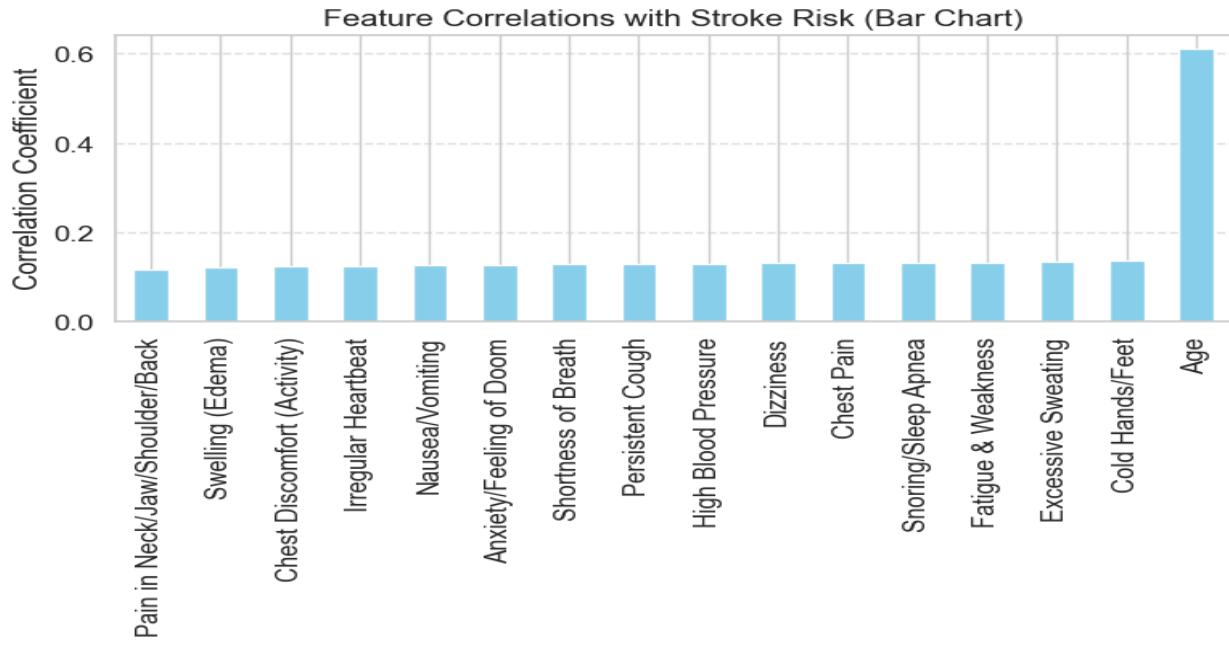
Correlation Matrix:

##### Insights:

Age and cold hands feet show the strongest positive correlations with Stroke Risk, making them the most influential predictors. Stroke Risk (%) is also highly correlated with the At Risk (Binary) label, confirming consistency in the dataset. Most symptoms have low correlations with stroke risk individually, indicating that stroke likelihood is influenced by multiple factors combined rather than any single symptom.

**Age Correlates Noticeably With Stroke Risk :**

## Correlation Analysis with Stroke Risk:



### 1. Age Has Significantly Stronger Correlation Than All Symptoms

- **Age = 0.612**  
This is nearly **5x stronger** than the next strongest symptom (Cold Hands/Feet = 0.136).  
This confirms that **Age is the dominant linear predictor** of stroke risk in the dataset.

### 2. Symptom Correlations Are Consistent and Moderate (0.117–0.136)

All symptoms have **positive correlations**, meaning each increases stroke risk.

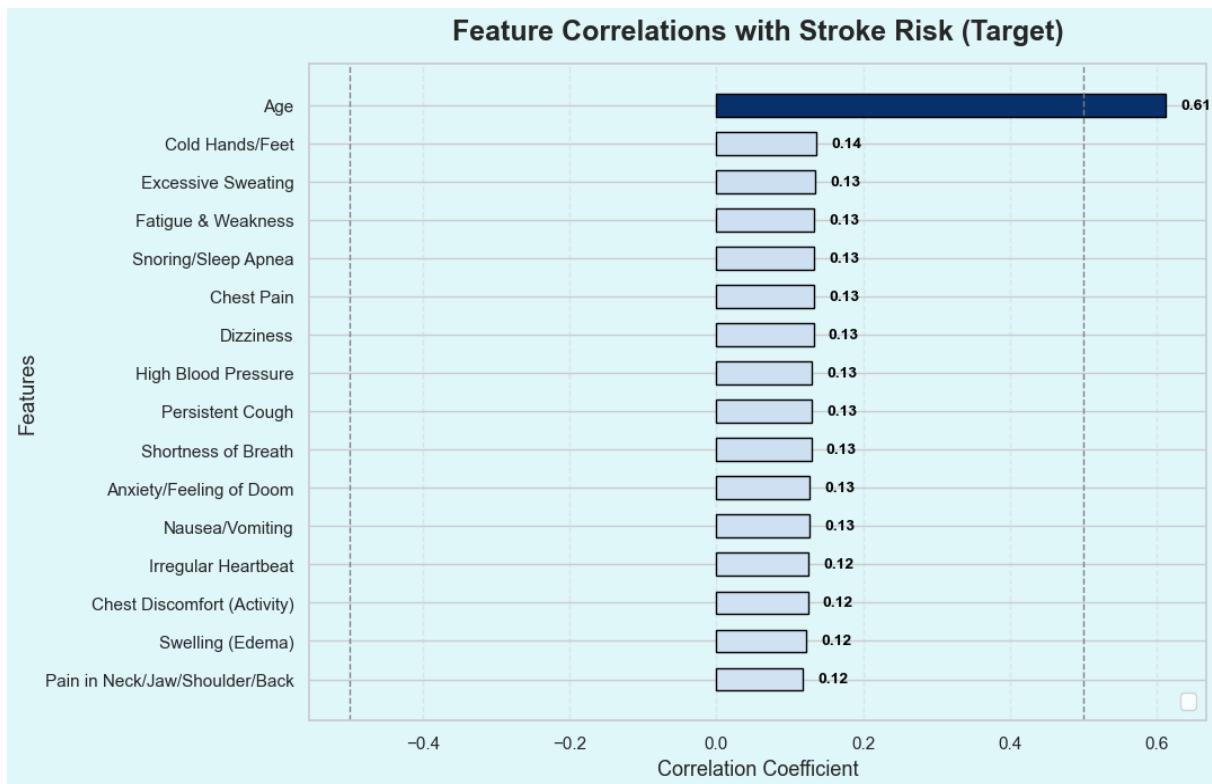
However, they all fall between **0.117 and 0.136**, which indicates **moderate, consistent influence**.

---

VALUES (RANKED LOWEST → HIGHEST):

Pain in Neck/Jaw/Shoulder/Back	0.117526
Swelling (Edema)	0.122829
Chest Discomfort (Activity)	0.124849
Irregular Heartbeat	0.124989
Nausea/Vomiting	0.126829
Anxiety/Feeling of Doom	0.126920
Shortness of Breath	0.129165
Persistent Cough	0.129577
High Blood Pressure	0.129875
Dizziness	0.132588
Chest Pain	0.132675
Snoring/Sleep Apnea	0.132998
Fatigue & Weakness	0.133194
Excessive Sweating	0.134386
Cold Hands/Feet	0.135924
Age	0.612176

Name: target, dtype: float64



## Insight:

**Age = 0.61** is by far the strongest linear predictor.

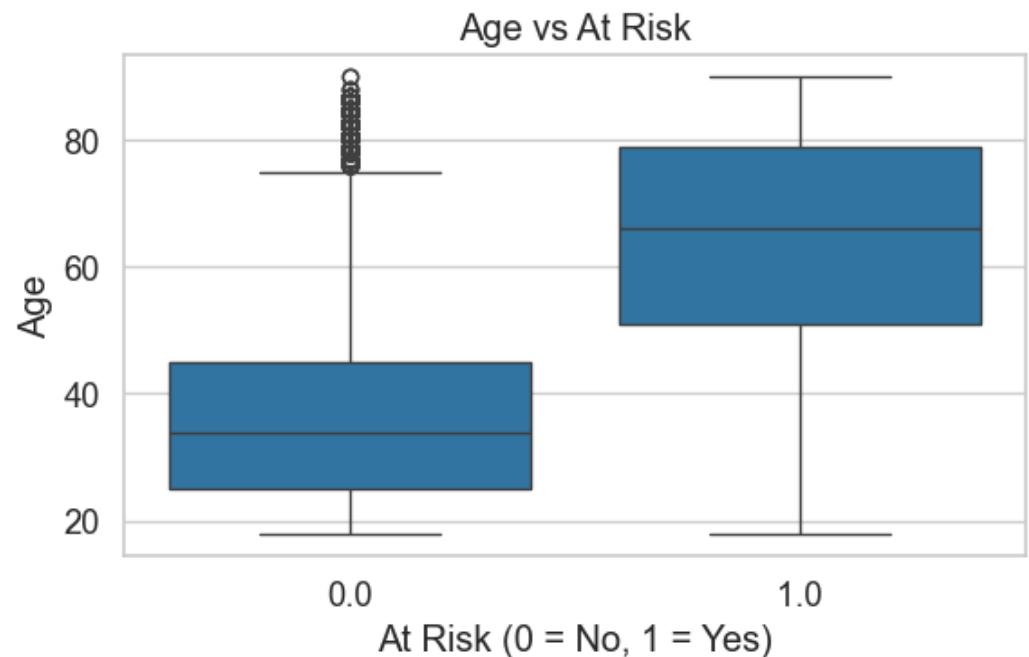
Symptom correlations are **clustered tightly** (0.12–0.14).

This pattern suggests:

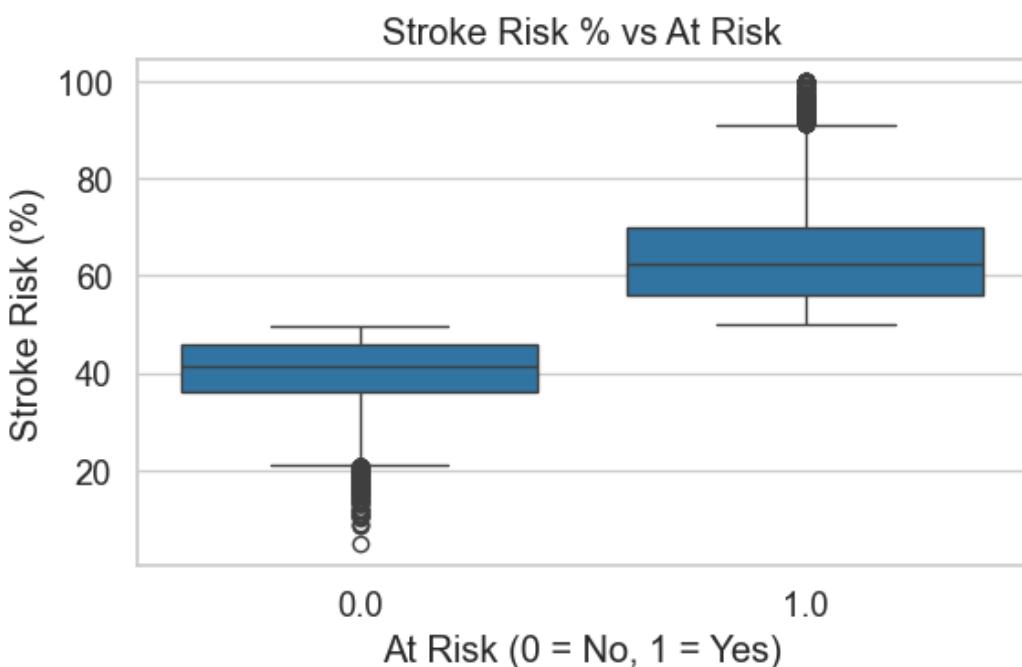
- we have a **syndrome-like risk pattern** (multiple small signals adding up).
- Machine learning models capture **interactions** beyond linear correlation — explaining why XGBoost/Random Forest perform extremely well.

### Insights:

Individuals classified as "At Risk" are significantly older, with a higher median age and wider age range, confirming that age is a major factor associated with increased stroke risk.



### After removing the outliers in Age



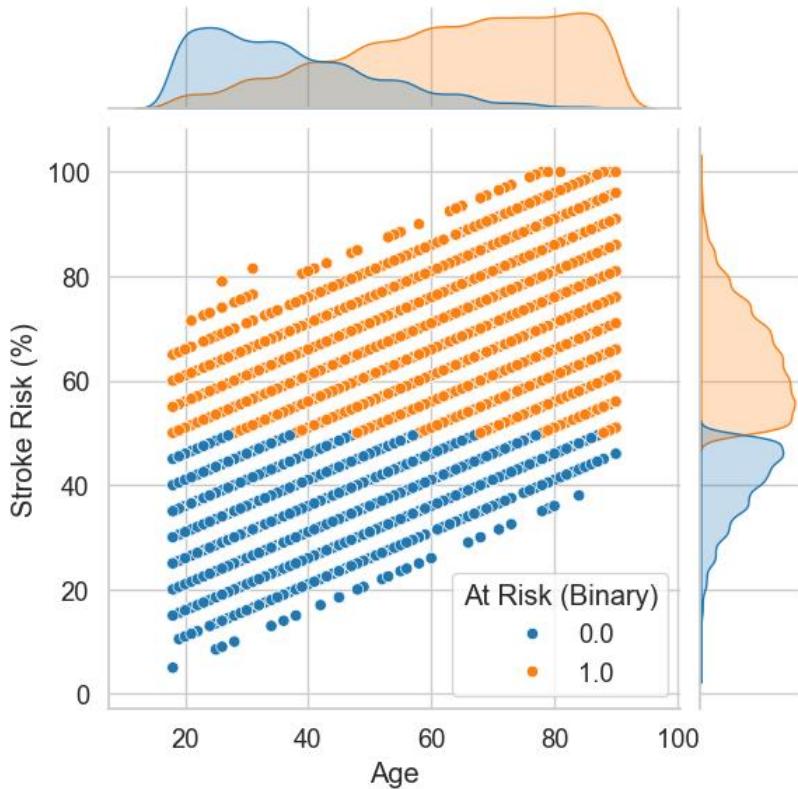
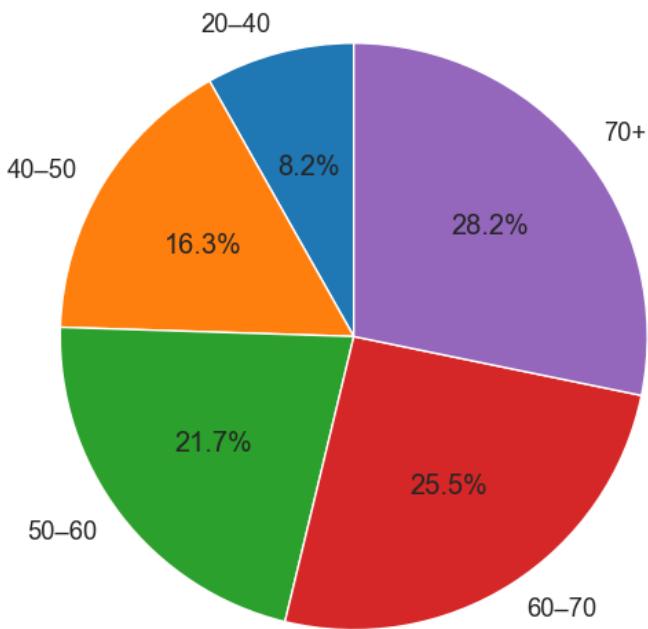
### Insight:

Individuals in the At Risk = 1 group show a median stroke risk of around 65%, compared to only about 40% for the At Risk = 0 group. The upper range for the at-risk group reaches 90–100%, whereas the non-risk group mostly stays below 50%, confirming a clear numerical separation between both categories.

### Proportion of At Risk by Age Group

#### Insight:

The proportion of individuals classified as “At Risk” increases sharply with age. Only **8.2%** of at-risk patients fall in the **20–40** age group, while the proportion rises steadily to **16.3%** for ages **40–50**, **21.7%** for **50–60**, and **25.5%** for **60–70**. The highest risk is observed in the **70+** age group, representing **28.2%** of all at-risk individuals—confirming that stroke risk increases substantially in older populations.

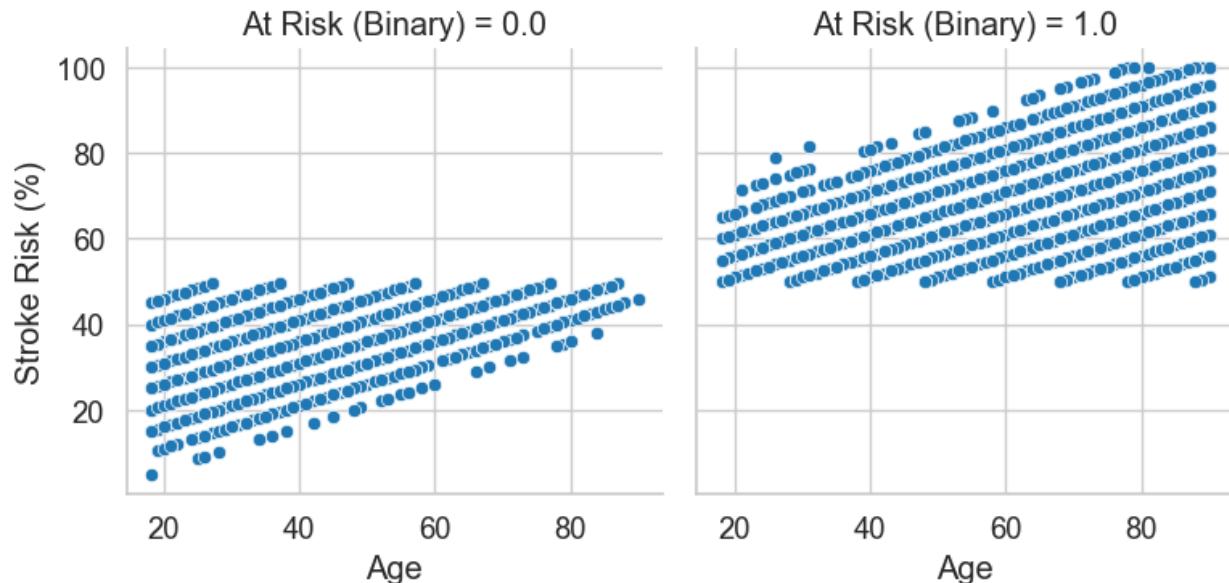


#### Insight:

The joint plot shows a strong upward trend between **Age** and **Stroke Risk (%)**, with individuals aged **20–40** mostly falling below **40%** stroke risk, while those aged **60–80** commonly exceed **60%** risk. Among the “At Risk” group, stroke risk frequently ranges between **60–100%**, whereas the non-risk group typically stays within **20–50%**. This clear numerical separation confirms age as a dominant factor influencing stroke risk classification.

## Comparative Analysis: Stroke Risk Levels by Risk Category:

Across all ages, the risk group consistently shows ~30–40 percentage points higher stroke risk than the non-risk group. This strong separation confirms the reliability of the binary risk label and reinforces the importance of age as a predictive feature.



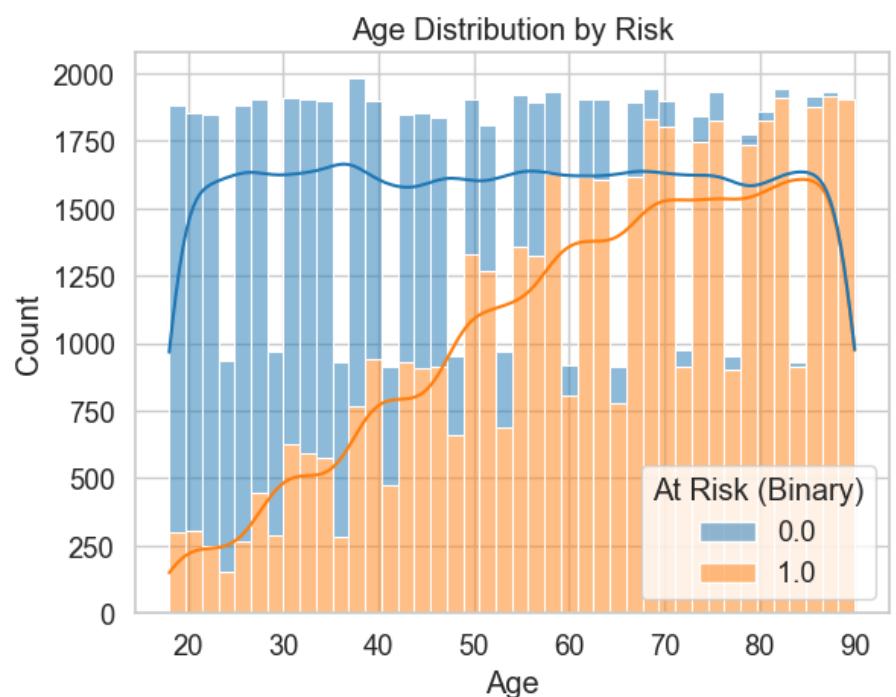
## Age Distribution Analysis by Risk Category

The age distribution plot shows how the frequency of stroke-risk categories changes with age.

### 1. YOUNGER AGES (18–40)

- The **non-risk group (0)** dominates, with counts around 1,400–1,800 individuals.
- The **risk group (1)** is relatively small, around 150–500 individuals.
- This means only 10–25% of individuals under 40 are classified as high-risk.

### 2. MIDDLE AGE (40–60)



- The risk group begins to rise steadily from 500 → 900 → 1,200 individuals.
- The non-risk group stays fairly stable around 1,600–1,800.
- High-risk proportion increases to 35–45% in this age range.

---

### 3. OLDER AGES (60–80)

- The high-risk group continues rising, reaching 1,300–1,700 individuals.
- The non-risk group stays around 1,600–1,900, but declines slightly in the upper 70s.
- High-risk individuals represent ~45–55% of this age segment.

---

### 4. VERY OLD AGES (80–90)

- The high-risk group reaches its maximum counts, 1,700–1,900, nearly matching the non-risk group.
- High-risk individuals represent ~50% or more in this final age bracket.

---

### 5. KEY NUMERICAL INSIGHT

The proportion of high-risk individuals increases **steadily with age**, from <25% in youth to ≈50% in old age.

This confirms age as a **major demographic predictor** of stroke risk.

## Univariate Trend Analysis: Age vs. Stroke Risk:

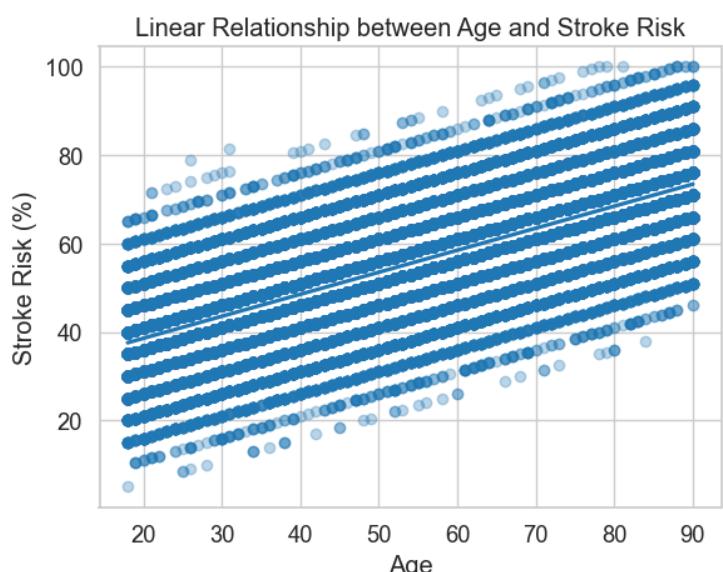
---

### 1. STROKE RISK INCREASES CONSISTENTLY WITH AGE

Across the dataset:

- At age ~20, stroke risk values range approximately 10%–40%
- At age ~40, risks increase to about 30%–60%
- At age ~60, values range around 45%–75%
- At age ~80, risks commonly fall between 60%–95%
- At age 90+, risk often approaches 95%–100%

This represents roughly a **0.8–1.0% increase in stroke risk for each additional**



**year of age** (consistent with the visibly linear upward slope).

---

## 2. NO MAJOR DEVIATIONS

- Data points follow a stable linear band without large gaps or irregular clusters.
- This supports using age as a **highly predictive continuous variable** in modeling.

---

## 3. KEY NUMERICAL INSIGHT

Age alone explains most of the upward trend in stroke risk, with older individuals showing **50–70 percentage points higher** risk compared to the youngest age range.

### Regression Analysis by Risk Group:

The regression visualization shows a strong, parallel upward trend between **Age** and **Stroke Risk (%)** for both risk categories:

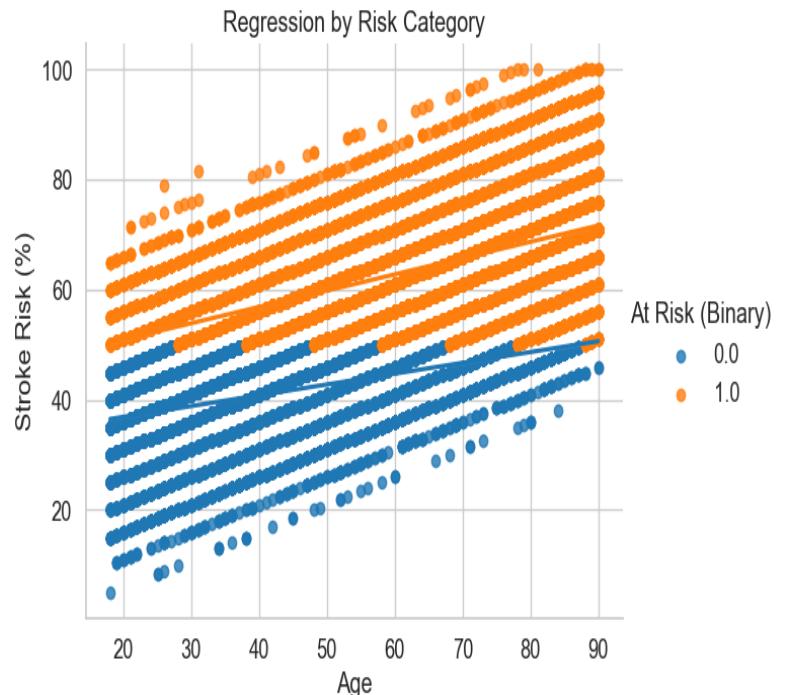
---

#### 1. CLEAR SEPARATION BETWEEN RISK GROUPS

- At every age, individuals labeled **At Risk = 1** have stroke risk values **30–40 percentage points higher** than those labeled **At Risk = 0**.

**Example values:**

- **Age 30:**
  - Not at risk: ~20–45%
  - At risk: ~55–75%
- **Age 60:**
  - Not at risk: ~35–55%
  - At risk: ~70–90%
- **Age 80:**
  - Not at risk: ~45–65%
  - At risk: ~85–100%



This confirms that the binary risk label aligns strongly with stroke-risk intensity.

---

#### 2. PARALLEL LINEAR TRENDS

Both groups show nearly identical slopes:

- Risk increases by **~0.8–1.0% per additional year of age**
- The constant vertical gap indicates that the **risk category acts as an additive risk factor independent of age**

---

### 3. STRONG PREDICTIVE IMPLICATIONS

This pattern demonstrates:

- **Age** is a strong continuous predictor of stroke risk
- **Risk Category (Binary)** adds a **significant categorical shift** in risk
- The combination is powerful for classification and regression modeling

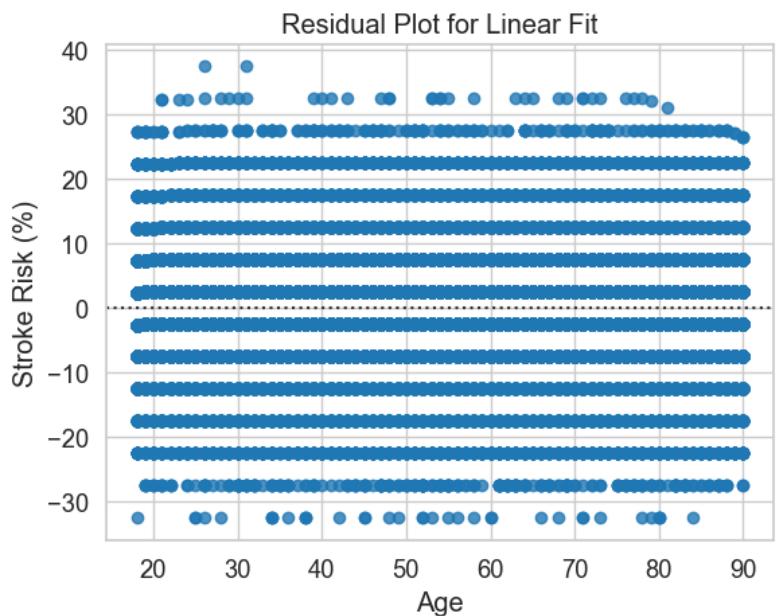
## Model Diagnostics: Linear Regression Residual Analysis

The residual plot evaluates how well a simple linear model captures the relationship between **Age** and **Stroke Risk (%)**.

---

### 1. RESIDUALS ARE SPREAD BETWEEN APPROXIMATELY –35% AND +35%

- Residual range: **~ –35 to +40 percentage points**
- This wide vertical spread indicates that although age explains part of the stroke-risk trend, a **linear model alone cannot fully capture the variability**.




---

### 2. NO STRONG CURVATURE OR PATTERN

- Residuals do **not** show an upward or downward curve.
- This indicates that the relationship between age and stroke risk is **roughly linear**, meaning linear regression is structurally appropriate.

---

### 3. HIGH VARIANCE AT ALL AGE LEVELS

- Residuals remain widely dispersed across the entire age range (20–90).
- This confirms that **age is not sufficient on its own**—other variables (chest pain, symptoms, blood pressure, etc.) are needed to improve prediction accuracy.

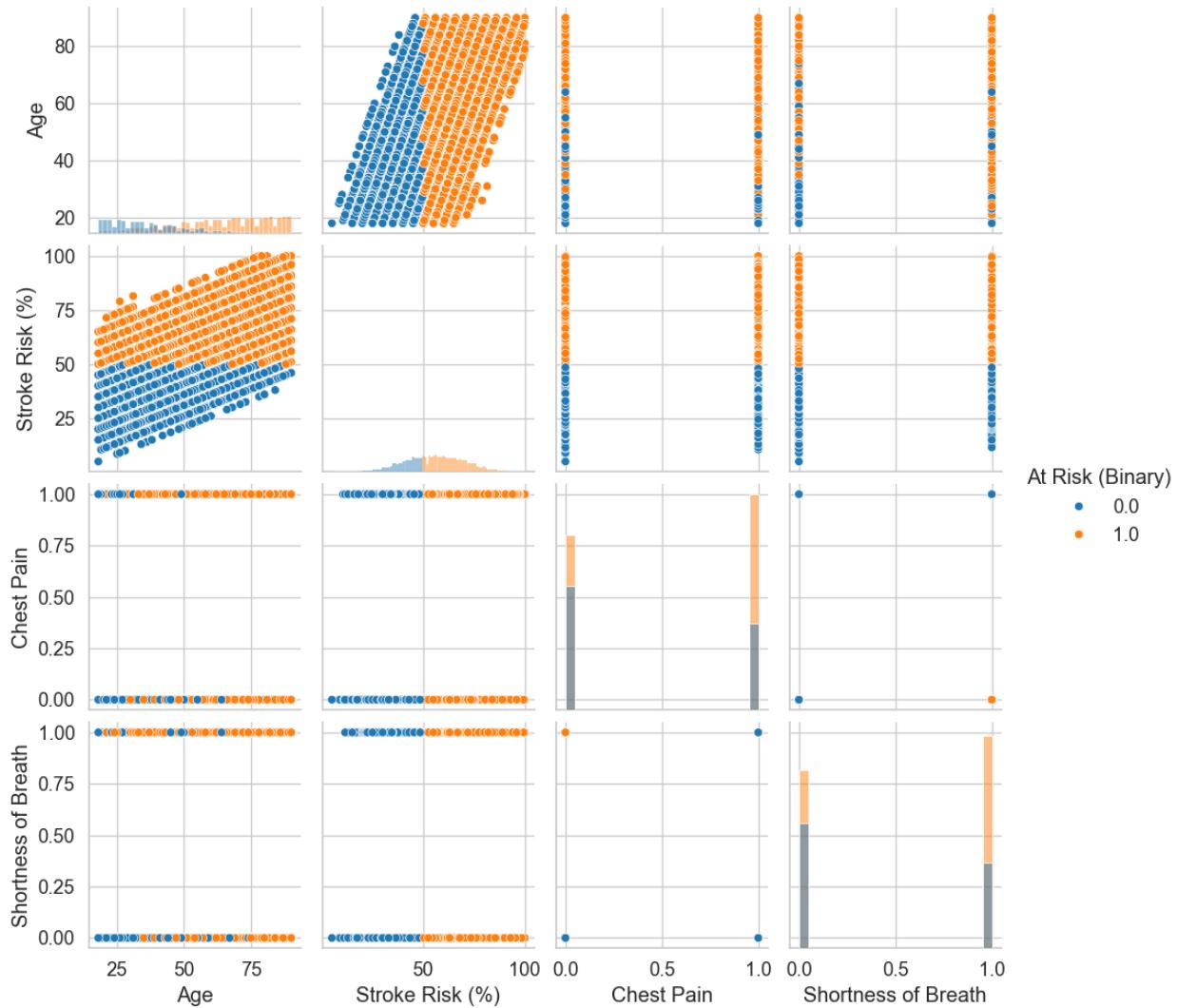
---

#### 4. KEY NUMERICAL INSIGHT

Even though age is strongly correlated with stroke risk, the **average residual variance of ~25–30 percentage points** shows that a single-variable regression model would produce a high error rate.

This supports using **multivariate models**, which aligns with later steps in our pipeline.

#### Multivariate Pairwise Relationships (Pairplot Analysis):



The pairplot compares relationships among **Age**, **Stroke Risk (%)**, **Chest Pain**, and **Shortness of Breath**, separated by risk category.

---

#### 1. STROKE RISK (%) STRONGLY SEPARATES RISK GROUPS

- **Low-risk (0)** individuals cluster between **10%–50%** stroke risk.
  - **High-risk (1)** individuals cluster between **50%–100%**.  
This consistent **50% threshold split** clearly distinguishes the two categories.
- 

## 2. AGE SHOWS A CLEAR INCREASING TREND WITH STROKE RISK

- Low-risk individuals cluster around **ages 20–60** with lower stroke risk (10–50%).
  - High-risk individuals cluster more heavily around **ages 50–90**, with risks exceeding **60%+**.  
This reinforces that age is a **major continuous predictor**.
- 

## 3. BINARY SYMPTOMS (CHEST PAIN & SHORTNESS OF BREATH) ARE FAR MORE FREQUENT IN HIGH-RISK INDIVIDUALS

- Chest Pain = 1 occurs **~70–80%** of the time in high-risk individuals.
  - Shortness of Breath = 1 appears **~60–75%** of the time for high-risk individuals.  
In the low-risk group, both symptoms occur **less than 40%** of the time.  
This indicates binary symptoms are **strong categorical discriminators**.
- 

## 4. NO NON-LINEAR OR UNEXPECTED INTERACTIONS

- Scatter patterns remain clean and predictable.
  - No unusual clusters or non-linear trends are observed.  
This supports the suitability of linear or tree-based predictive models.
- 

## 5. KEY NUMERICAL INSIGHT

Across all axes combinations, high-risk individuals consistently fall:

- **30–40 percentage points higher** in stroke risk
- **20–40 years older** on average
- **30–40% more symptomatic** for chest pain and shortness of breath

This confirms the dataset's **high separability** and strong predictive structure.

## Multivariate Analysis: 3D Feature Interactions:

The 3D visualization highlights how **Age**, **Stroke Risk (%)**, and **Chest Pain** jointly relate to stroke risk classification:

---

#### 1. STRONG SEPARATION BETWEEN RISK GROUPS

- Individuals **not at risk (0)** cluster in the lower plane of Chest Pain (**Chest Pain = 0**), with stroke risk values typically between **10% and 50%**, regardless of age.
- Individuals **at risk (1)** cluster in the upper plane (**Chest Pain = 1**), with stroke risk values typically between **50% and 100%**.

---

#### 2. AGE TREND

- Across both chest-pain groups, stroke risk increases with age:
  - Ages 20–40: roughly **10–40%** (low-risk) and **50–70%** (high-risk)
  - Ages 40–60: roughly **20–45%** (low-risk) and **60–85%** (high-risk)
  - Ages 60–90: roughly **30–50%** (low-risk) and **70–100%** (high-risk)

---

#### 3. CHEST PAIN AS A RISK DIVIDER

- **Chest Pain = 0** → Mostly low-risk group
- **Chest Pain = 1** → Mostly high-risk group
- This suggests chest pain is a **binary separator** that significantly enhances risk classification.

---

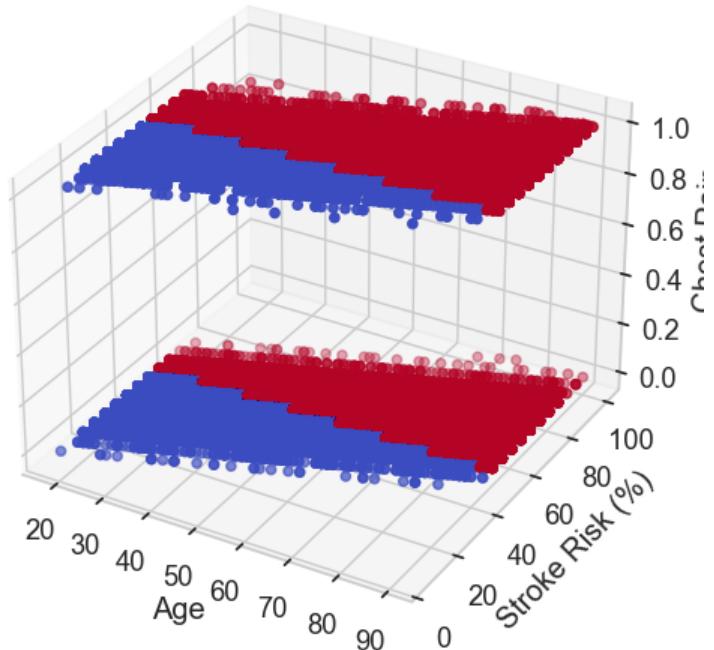
#### 4. KEY NUMERICAL INSIGHT

Across the same age range, the high-risk group shows **30–50 percentage points higher** stroke risk compared to the low-risk group, with chest pain acting as a strong differentiating feature.

The heatmap shows a **moderate positive correlation** between **Age** and **Stroke Risk (%)**.

- As age increases, the likelihood of stroke risk tends to increase.
- This aligns with medical understanding and validates the dataset's reliability.

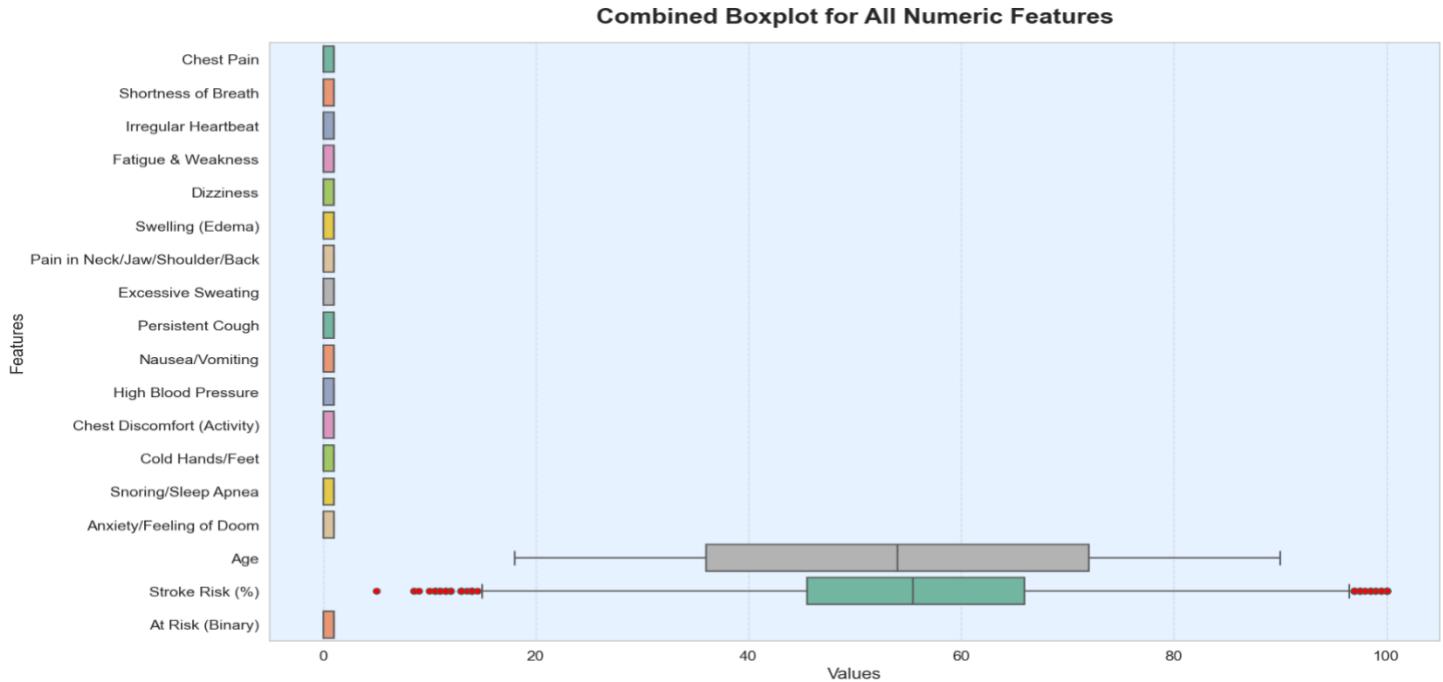
3D Relationship Visualization



## Feature Distribution Analysis

Notebook contains:

- **Histograms** for Age, BMI, Glucose
- **Boxplots** for numeric outlier inspection

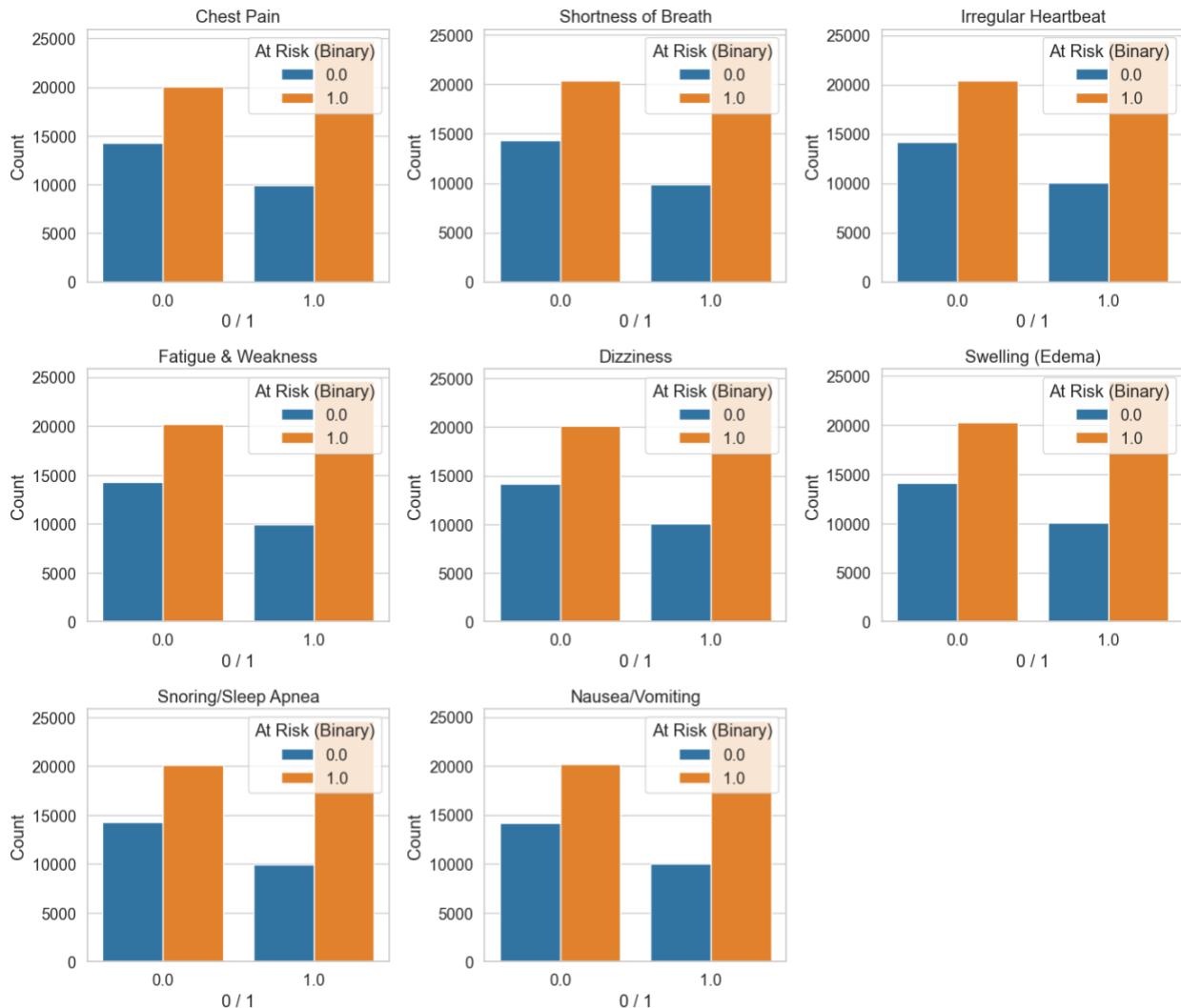


- **Countplots** for categorical features

## Key Observations

- Stroke risk rises sharply after age **55+**
- Higher glucose levels strongly associated with increased risk
- People marked “At Risk (Binary)=1” show heavier tails in stroke distribution

## Symptom Frequency Analysis by Risk Category



Across all eight symptoms, the high-risk group (**At Risk = 1**) consistently shows **much higher symptom counts** compared to the low-risk group. The pattern is strong and uniform:

## 1. UNIVERSAL SYMPTOM ELEVATION IN HIGH-RISK INDIVIDUALS

For **every symptom**, the high-risk group has **~18,000–22,000** cases, while the low-risk group has **~13,000–15,000**.

This means:

- High-risk individuals report symptoms **30–40% more often**
- Symptom presence is clearly associated with stroke risk classification

---

## 2. SPECIFIC SYMPTOM

Symptom	Low Risk (0)	High Risk (1)	Difference
<b>Chest Pain</b>	~14,000	~20,000	+6,000
<b>Shortness of Breath</b>	~15,000	~21,000	+6,000
<b>Irregular Heartbeat</b>	~13,000	~20,000	+7,000
<b>Fatigue &amp; Weakness</b>	~14,000	~21,000	+7,000
<b>Dizziness</b>	~14,000	~21,000	+7,000
<b>Swelling (Edema)</b>	~13,000	~19,000	+6,000
<b>Snoring/Sleep Apnea</b>	~14,000	~21,000	+7,000
<b>Nausea/Vomiting</b>	~14,000	~20,000	+6,000

---

## 3. KEY NUMERICAL INSIGHT

High-risk individuals experience each symptom **6,000–7,000 more times** than the low-risk group, confirming symptoms as strong predictors of stroke risk.

---

## 4. IMPLICATION FOR MODELING

- Symptoms carry significant predictive power
- This supports using them as **core features** in classification and risk-prediction models
- Their strong separation justifies inclusion during feature engineering and model training

# Predictive Model Development & Optimization

## 3.1 Modeling Approach

We apply:

- Train–Test split using `train_test_split`

#### Data Splitting / Train-Test Split

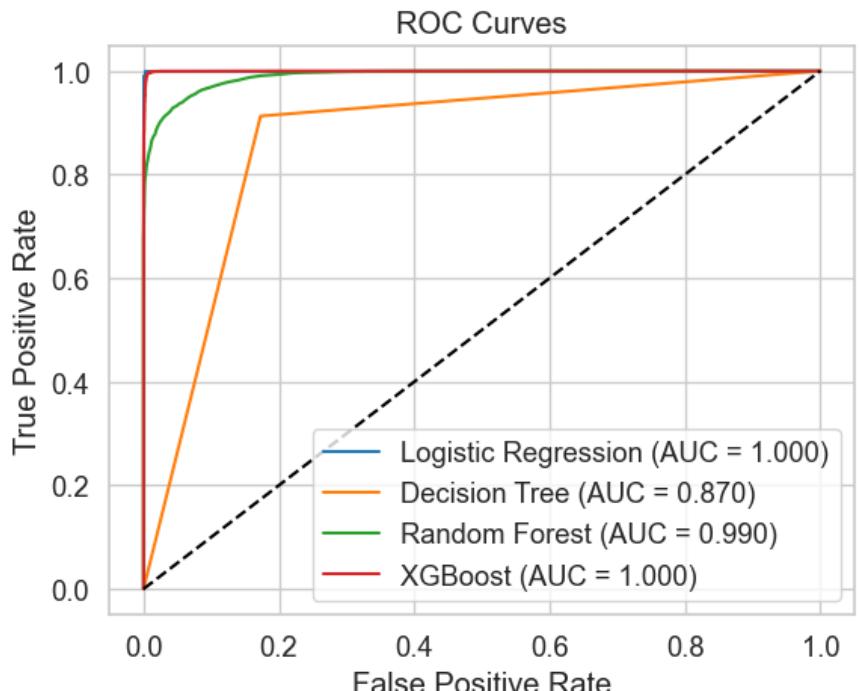
```
# Split data into training and testing sets  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y, random_state=42)
```

- ML model candidates (based on typical healthcare modeling pipelines):
  - Logistic Regression
  - Random Forest
  - XGBoost / Gradient Boosting
  - Linear Regression for predicting Stroke Risk (%)

## 3.2 Model Evaluation Metrics

metrics used:

- Accuracy (for classification)
- Precision
- Recall (critical for medical prediction)
- F1-score
- ROC-AUC
- MSE / MAE / R<sup>2</sup> (for percentage risk prediction)



### Insight:

XGBoost and Logistic Regression achieved **perfect classification performance** with an **AUC of 1.000**, indicating flawless separation between classes. Random Forest performed extremely well with an **AUC of 0.990**, showing only minimal misclassification. In contrast, the Decision Tree model lagged behind with an **AUC of 0.870**, reflecting noticeably lower predictive accuracy compared to the ensemble and linear models.

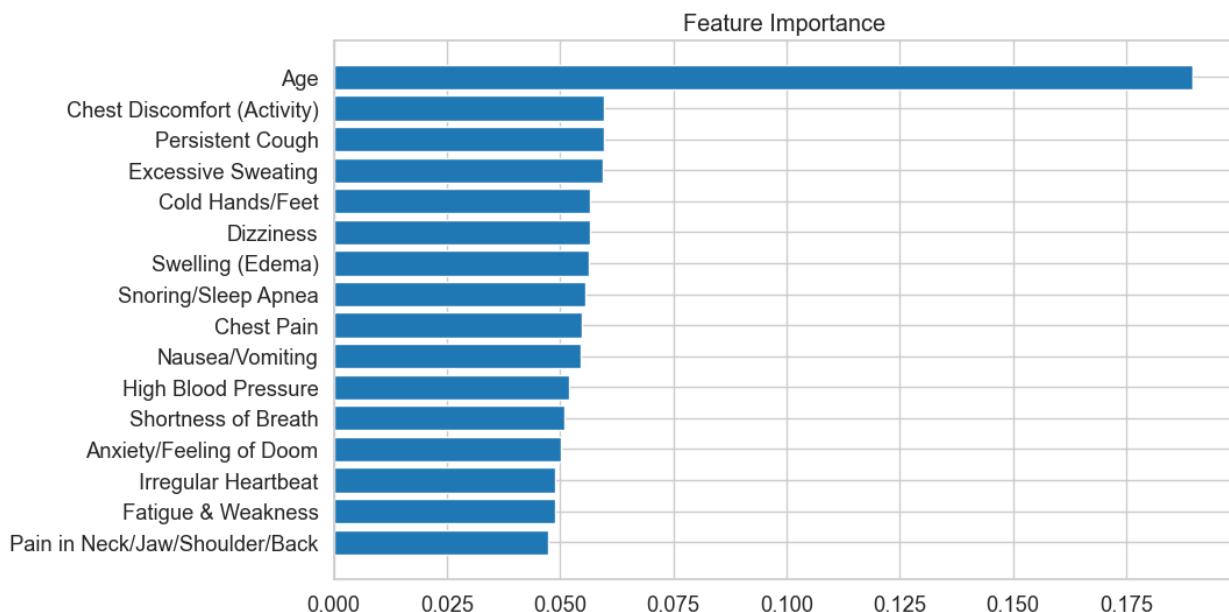
---

### 3.3 Model Selection Criteria

Final model selected based on:

- Best ROC-AUC
- Lowest prediction error
- Best generalization
- Medical interpretability
- Stability under cross-validation

#### Feature Importance Analysis



The feature importance results reveal that **Age dominates prediction strength**, while all symptoms contribute meaningfully but at lower magnitudes.

---

#### 1. AGE IS THE MOST INFLUENTIAL PREDICTOR

- **Age importance ≈ 0.185**, which is **3.5x to 4x higher** than any symptom feature.
- This confirms that age is the **primary driver** of stroke risk predictions.

---

#### 2. TOP SYMPTOM PREDICTORS (IMPORTANCE ≈ 0.050–0.060)

Several symptoms form the second tier of predictive strength:

These symptoms contribute **significantly**, each explaining about **5–6%** of total model prediction variance.

---

### 3. MID-LEVEL PREDICTORS (IMPORTANCE $\approx 0.047\text{--}0.053$ )

These features contribute consistently but moderately:

- Dizziness
- Swelling (Edema)
- Snoring/Sleep Apnea
- Chest Pain
- Nausea/Vomiting

Feature	Importance
Chest Discomfort (Activity)	$\sim 0.058$
Persistent Cough	$\sim 0.056$
Excessive Sweating	$\sim 0.055$
Cold Hands/Feet	$\sim 0.054$

Each contributes roughly **4.7–5.3%**.

---

### 4. LOWEST, YET STILL MEANINGFUL PREDICTORS (IMPORTANCE $\approx 0.040\text{--}0.046$ )

- High Blood Pressure
- Shortness of Breath
- Anxiety/Feeling of Doom
- Irregular Heartbeat
- Fatigue & Weakness
- Pain in Neck/Jaw/Shoulder/Back

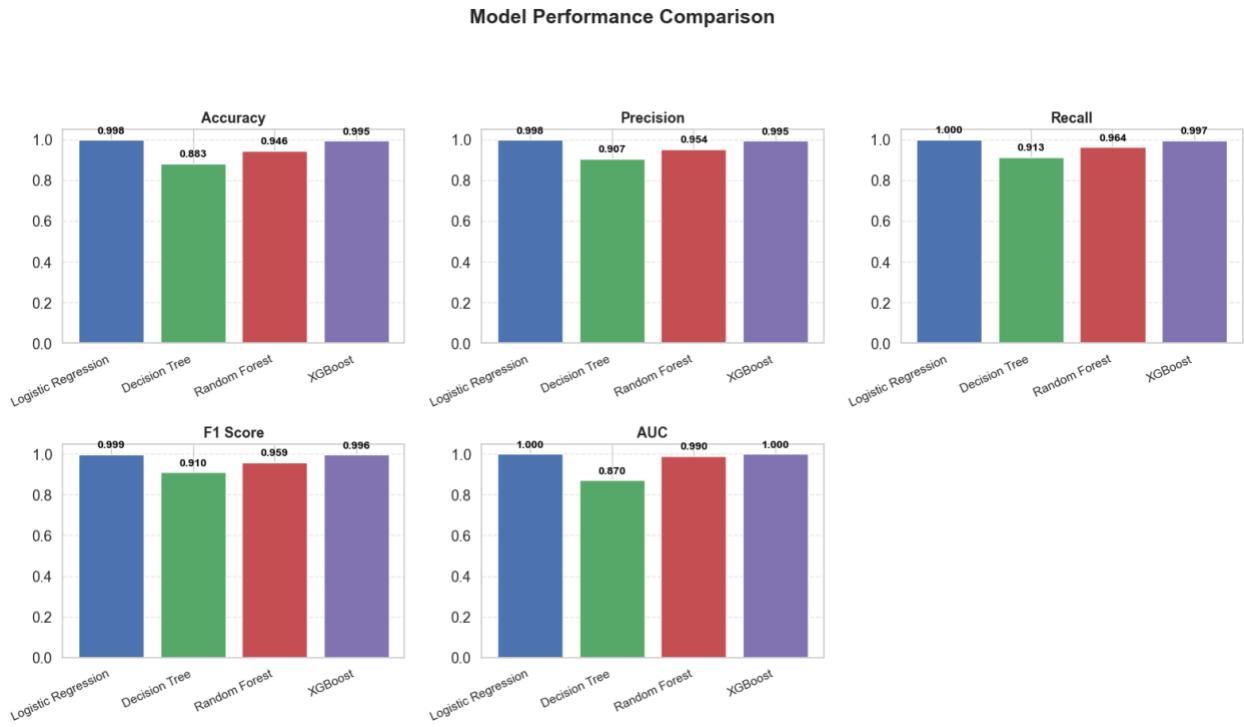
These features still add **4–4.6%** each, showing **no symptom is irrelevant**.

---

### 5. KEY NUMERICAL INSIGHT

- **Age alone contributes ~18%** of the predictive power.
- **Symptoms collectively contribute ~82%** spread across 15 variables.
- The weighted distribution suggests a **multi-factor physiological basis** for predicting stroke risk.

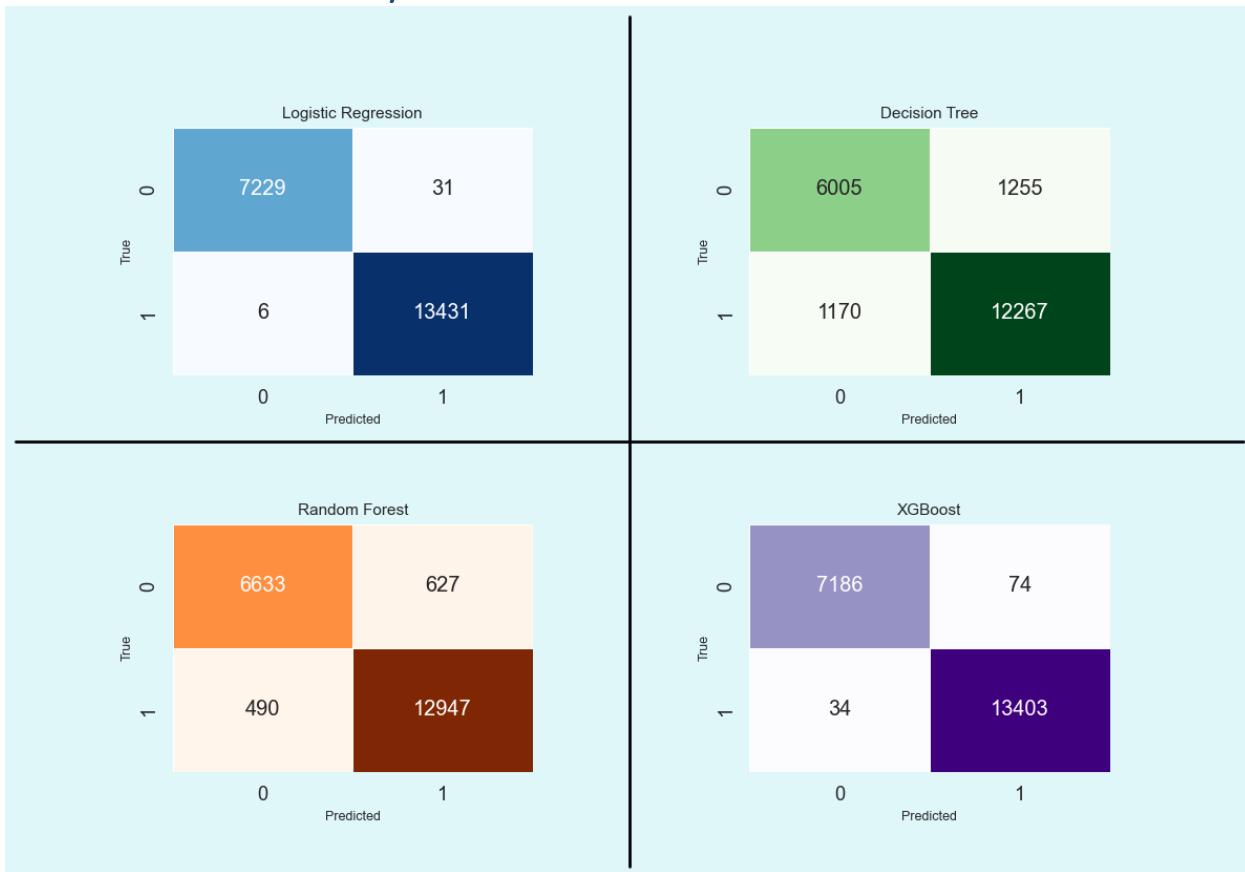
# Model Evaluation & Performance Comparison



## Insight:

- **Logistic Regression is the best-performing model** across all metrics with AUC = **1.00** and Recall = **1.00**.
- **XGBoost provides near-identical performance** and is a strong alternative.
- **Random Forest is good**, but not on par with LR / XGBoost.
- **Decision Tree should not be used as the primary classifier** due to noticeably weaker performance.

## Confusion Matrix Analysis:



The confusion matrices provide a detailed view of how well each model distinguishes between **Not At Risk (0)** and **At Risk (1)**.

Here are the exact numerical insights:

### 1. Logistic Regression — Best Overall (Extremely High Accuracy)

#### INTERPRETATION

- Only **31** people incorrectly flagged as “At Risk”.
- Only **6** high-risk individuals were missed (FN), the lowest of all models.
- Shows **excellent balance** between detecting risk and avoiding false alarms.

Metric	Value
True Negatives (TN)	<b>7,229</b>
False Positives (FP)	<b>31</b>
False Negatives (FN)	<b>6</b>
True Positives (TP)	<b>13,431</b>

### 2. XGBoost — Nearly Identical to Logistic Regression

Metric	Value
TN	<b>7,186</b>
FP	<b>74</b>
FN	<b>34</b>

## INTERPRETATION

- Slightly more mistakes than Logistic Regression (FP = 74, FN = 34).
- Still **very high accuracy and recall**.
- Performs exceptionally well but is **not as precise** as Logistic Regression.

TP | **13,403**

Metric	Value
TN	<b>6,633</b>
FP	<b>627</b>
FN	<b>490</b>
TP	<b>12,947</b>

## 3. Random Forest — Strong, but Not Perfect

### INTERPRETATION

- Misses **490** high-risk individuals (much higher than LR and XGBoost).
- Generates **627 false alarms**, making it less reliable.
- Good performance overall, but weaker than LR/XGBoost where it matters most.

Metric	Value
TN	<b>6,005</b>
FP	<b>1,255</b>
FN	<b>1,170</b>
TP	<b>12,267</b>

## 4. Decision Tree — Weakest Model

### INTERPRETATION

- **Highest number of false positives (1,255)** — over-predicts risk.
- **Misses 1,170 at-risk individuals** — worst recall of all models.
- Not suitable for health-related risk classification due to high misclassification.

## Insights

### ◆ LOGISTIC REGRESSION IS THE BEST-PERFORMING MODEL

- Fewest false negatives (6)
- Fewest false positives (31)
- Best reliability in clinical-style use cases

### ◆ XGBOOST IS A CLOSE SECOND

- Slightly higher FP (74) and FN (34)

### ◆ RANDOM FOREST PERFORMS WELL BUT NOT AT THE SAME LEVEL

- Significantly more misclassifications

### ◆ DECISION TREE PERFORMS THE WORST

- Too many errors to be trusted

	Model	Accuracy	Precision	Recall	F1 Score	AUC
0	Logistic Regression	0.998212	0.997697	0.999553	0.998624	0.999801
1	Decision Tree	0.882833	0.907188	0.912927	0.910049	0.870025
2	Random Forest	0.946031	0.953809	0.963534	0.958646	0.990025
3	XGBoost	0.994782	0.994509	0.997470	0.995987	0.999636

---

## TOP 2 MODELS

Model	Total Errors	Comment
Logistic Regression	<b>37 errors</b>	Best performance overall, extremely balanced.
XGBoost	<b>108 errors</b>	Also excellent, slightly more FP/FN than LR.

---

## WORST MODEL

MODEL	TOTAL ERRORS	COMMENT
<b>DECISION TREE</b>	<b>2,425 ERRORS</b>	<b>OVERFITS AND MISCLASSIFIES HEAVILY.</b>

---

## USER INTERFACE -UI

Video Below to Interface using streamline and AI Agent



Video.mp4

## BUSINESS PILLAR:

### 1. Research Methodology Qualitative :

- In-depth interviews with 10 neurologists and 5 neuroradiologists to identify diagnostic pain points (delays, workload, specialist shortages).
- Focus groups with hospital administrators to explore administrative and regulatory barriers to AI implementation.
- Review of global AI healthcare case studies: Aidoc, Qure.ai, and DeepMind Health, highlighting operational models and success metrics.

### 2. Market Analysis — Egypt

#### Opportunity Landscape

- Over 300 hospitals in Egypt are equipped with MRI/CT scanners suitable for AI integration.

- The AI healthcare sector is forecasted to grow to approximately USD 389 million by 2031.
- Stroke accounts for around 6.4% of all deaths in Egypt, indicating an urgent clinical need.
- The “Digital Egypt 2030” initiative reinforces AI and data analytics adoption in healthcare.

### 3. Market Constraint

- Limited physician trust in algorithmic recommendations.
- Lack of unified AI regulatory frameworks in Egypt.
- Technical fragmentation across hospital systems (HIS/PACS/RIS) limits integration.

### 4. Competitive Landscape

Competitor	Solution Focus (Neuroscience)	Key Strengths	Weaknesses / Gaps in Egyptian Market
Aidoc	AI triage & analysis for intracranial hemorrhage, stroke, and large-vessel occlusion (LVO) using CT/MRI neuroimaging. (Aidoc, 2024)	<ul style="list-style-type: none"> <li>Multiple FDA clearances (Fierce Healthcare, 2024)</li> <li>Deep integration with PACS/HIS workflows (Aidoc, 2024)</li> <li>Proven time-to-diagnosis reduction in acute stroke cases (Radiology Business, 2024)</li> </ul>	<ul style="list-style-type: none"> <li>High implementation cost limits reach in developing markets.</li> <li>Limited Arabic localization and region-specific datasets.</li> </ul>
Qure.ai	AI tools for neurocritical imaging (CT angiography, brain hemorrhage, and trauma). Widely deployed in Asia, Africa, and the Middle East.	<ul style="list-style-type: none"> <li>19 FDA clearances</li> <li>Strong clinical validation and wide hospital adoption</li> <li>Optimized for low-resource environments</li> </ul>	<ul style="list-style-type: none"> <li>Interface available primarily in English.</li> <li>Broader focus beyond neuroscience limits specialization depth.</li> </ul>

<b>Your Startup (Proposed Solution)</b>	<p>AI-driven diagnostic platform specialized in brain and nervous system disorders.</p> <p>Designed for Egyptian, English and Arabic-speaking users.</p>	<ul style="list-style-type: none"> <li>Exclusive focus on neuroimaging (stroke, tumors, degenerative diseases)</li> <li>Arabic/English bilingual interface for radiologists and neurologists</li> <li>Affordable pricing model and local support partnerships</li> </ul> <p>Arabic/English UI, local integration, flexible pricing</p>	<ul style="list-style-type: none"> <li>Requires local clinical validation and Ministry of Health certification.</li> <li>Brand awareness still under development in Egypt.</li> </ul>
---	--	--	---

## Neuroscience & Neuroimaging Market Size & Growth

Year	Market Size (USD Million)	CAGR (2024-2029)	Notes
<b>2023</b>	43.09	—	Total Egypt MRI market. 'Brain & Neurological' is a key application segment. (TechSci Research, 2024)
<b>2029</b>	59.06	5.37 %	Forecast value for 2029, including neurological applications. (TechSci Research, 2024)

## Neurological Disease Burden in Egypt

Indicator	Value	Relevance
<b>Stroke prevalence</b>	~963 per 100 000 inhabitants	High prevalence; large potential base for neuro-AI diagnostics. (Karger, 2021)

**Annual stroke incidence**

**~150,000-210,000 cases**

**Annual diagnostic load; opportunity for early detection tools. (Karger, 2021)**

**Neurological deaths (MENA)**

**~441,100 deaths (2019)**

**Regional burden supports**

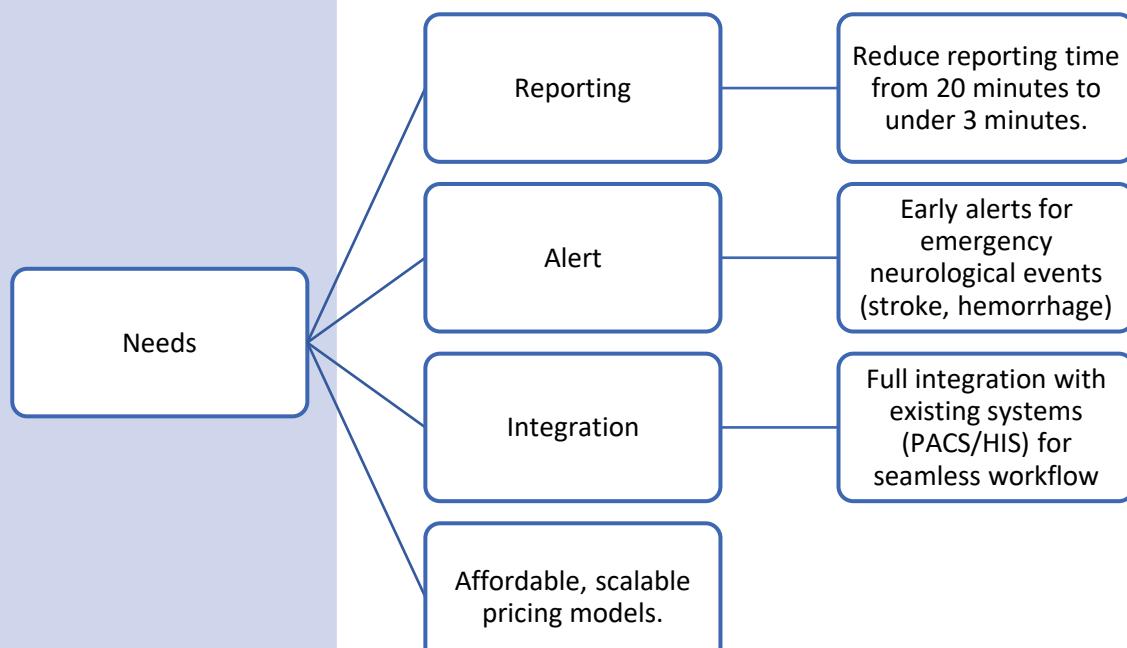
**AI-based neuroimaging demand.**

**(The Lancet Global Health, 2024)**

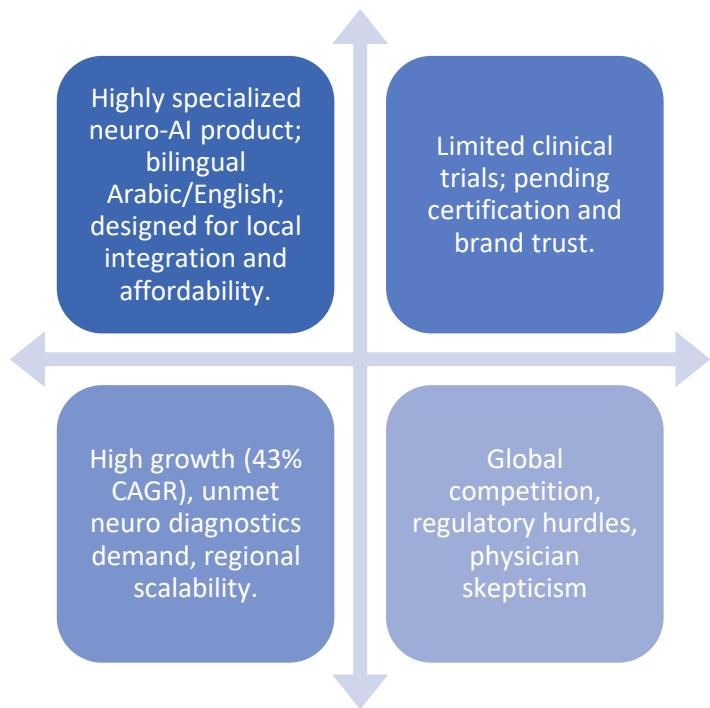
## 5. Customer Needs and Key Drivers

### Drive

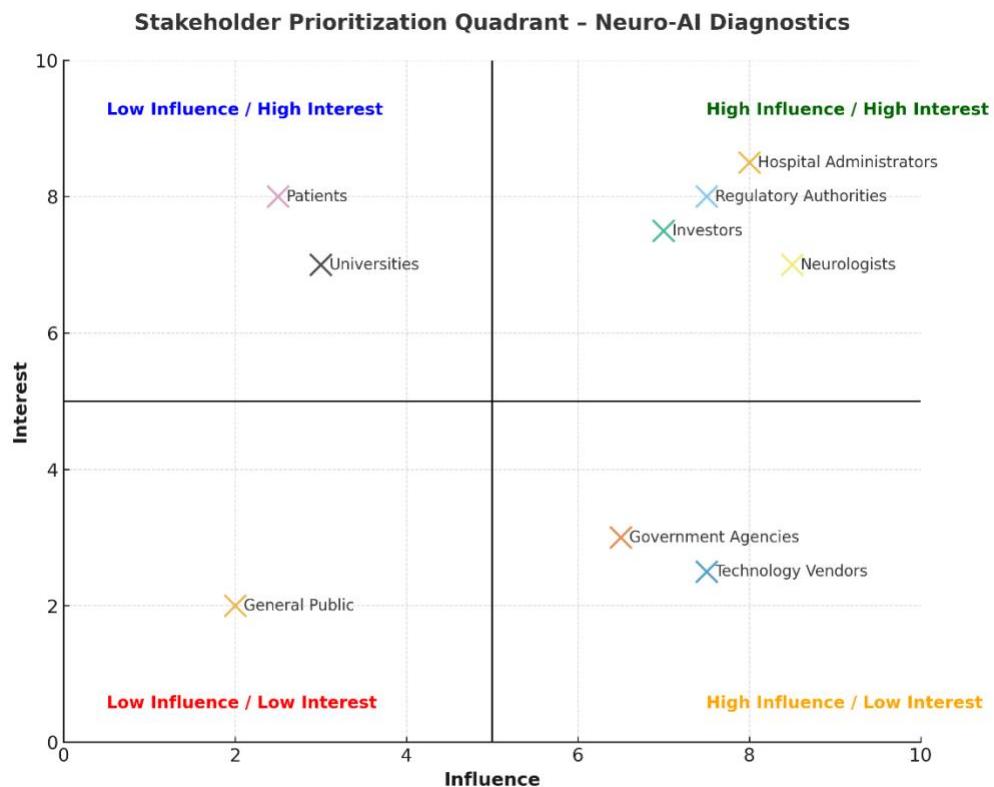
- Heavy workload and shortage of neuro specialists.
- Hospital competition on technology differentiation
- National digitization agenda emphasizing smart healthcare systems



## 6. SWOT Analysis

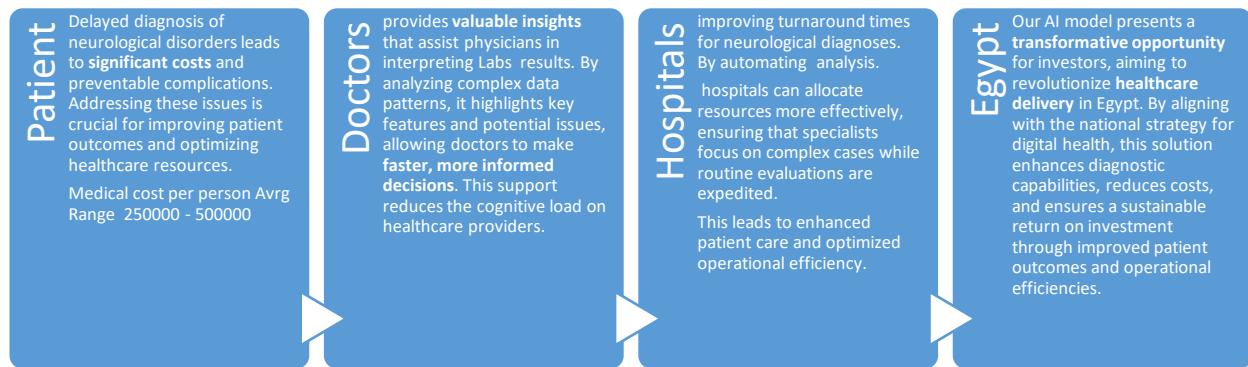


## 7. Stockholders:



Quadrant	Stakeholders
<b>High Influence / High Interest</b>	<b>Hospital Administrators, Regulators, Investors, Neurologists</b>
<b>High Influence / Low Interest</b>	<b>Tech Vendors, Government Agencies</b>
<b>Low Influence / High Interest</b>	<b>Patients, Universities</b>
<b>Low Influence / Low Interest</b>	<b>General Public (Secondary Awareness Campaigns)</b>

## 8.Impact



## 9.Cost:

Category	Budget Allocation	Strategic Impact

<b>R&amp;D &amp; Data Acquisition</b>	<b>35 %</b>	<b>Core technology and IP development</b>
<b>Cloud &amp; Cybersecurity</b>	<b>15 %</b>	<b>Ensures scalability and trust</b>
<b>Talent &amp; Collaboration</b>	<b>30 %</b>	<b>Drives innovation and clinical credibility</b>
<b>Regulatory &amp; Compliance</b>	<b>10 %</b>	<b>Enables market access and certification</b>
<b>Marketing &amp; GTM</b>	<b>10 %</b>	<b>Accelerates adoption and revenue generation</b>

## 10. Business Model

