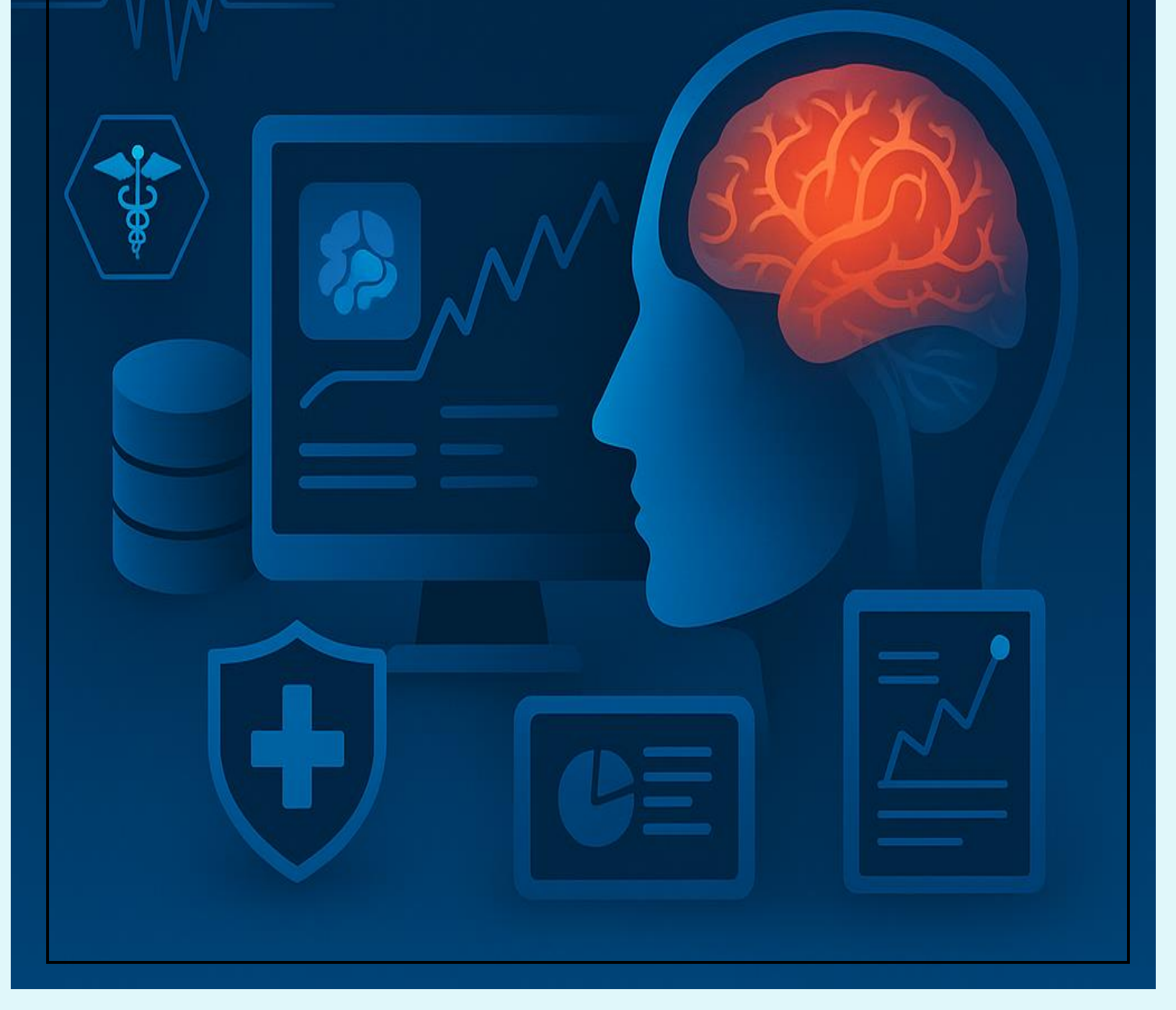
Digital Egypt Pioneers

Stroke Risk Prediction



Prepared for: **Eng. Sherif Salem**

AI & Data Science Track
CLS_ONL3_AIS4_G2
Digital Egypt Pioneers

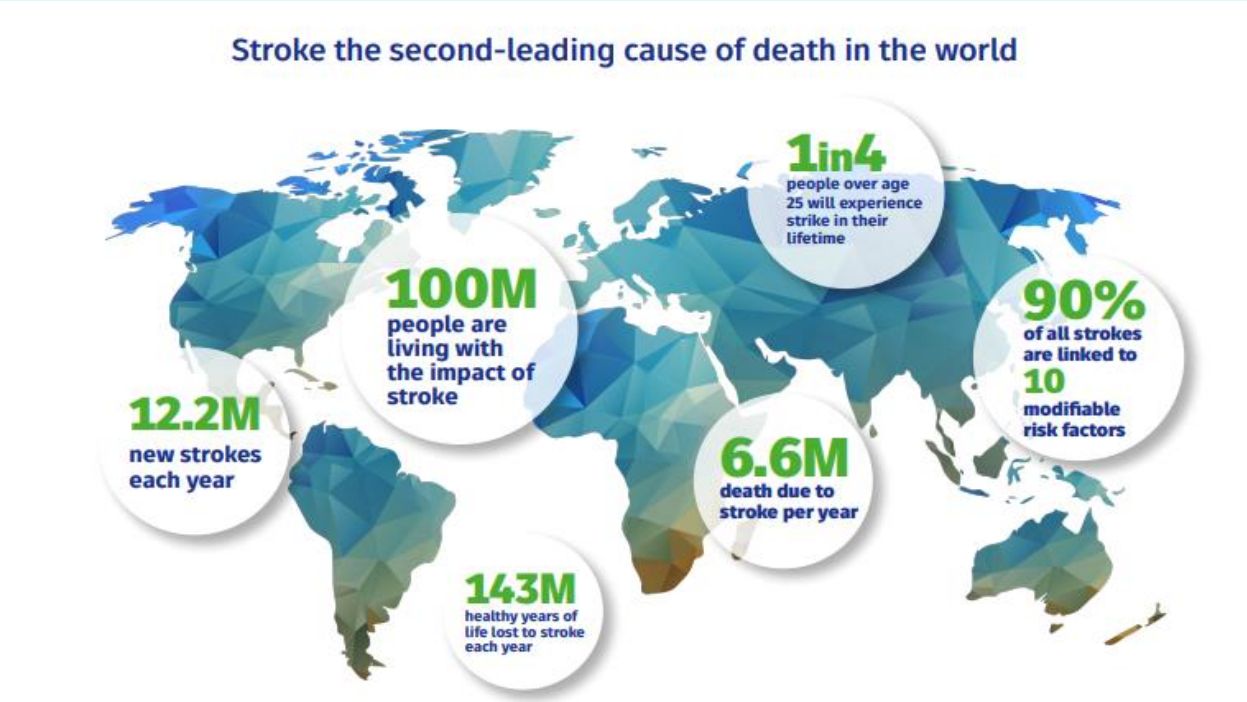
Prepared by:

Eng. Mohamed Nasr
Eng. Ahmed Ghanem
Eng. Tarek El Naggar
Eng. Ahmed Walid
Eng. Ahmed Abd El Maksoud
Dr. Doaa Gad-Allah

I EXECUTIVE SUMMERY:

INDEX & REFERANCES

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



Source: https://www.world-stroke.org/assets/downloads/Annual_Report_2021_online_latest.pdf

This project presents an end-to-end Stroke Risk Prediction System designed to identify individuals at risk of stroke based on a combination of clinical symptoms, demographic indicators, and health-related attributes. Using a robust dataset sourced from Kaggle and enhanced with extensive exploratory analysis, machine learning modeling, and multi-layered MLOps architecture, the solution demonstrates high predictive performance and real-world applicability.

The analysis revealed strong correlations between age, cardiopulmonary symptoms, and stroke risk. Multiple models were benchmarked, with Logistic Regression emerging as the most reliable and clinically interpretable model, achieving an AUC of 0.9998 and excellent sensitivity—critical for minimizing missed high-risk individuals. The system was deployed with an Arabic/English-enabled interface, integrating interactive dashboards, monitoring components, and experiment tracking tools to support operational reliability.

Beyond the technical implementation, the project includes market research, stakeholder analysis, and business modeling tailored to the Egyptian healthcare landscape. This ensures that the proposed system is both technically sound and strategically aligned with the needs of hospitals, clinicians, and national digital-transformation initiatives.

Stroke Risk Prediction / Nerva 🧠

1.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.

S — Specific
- Develop a predictive machine learning model capable of assessing stroke risk based on patient health metrics such as age, and medical history.

M — Measurable
- **85% accuracy0.90 ROC-AUC, F1 Score**

A — Achievable
- Utilize accessible healthcare datasets, standard data science workflows, and proven machine learning algorithms to ensure that the model can be built, optimized, and deployed using available tools and resources.

S — Relevant
- Support Health care by deliver Data Drien and ML Tool that enhance early stroke dediction, improve clinical Decision making

T — Time
- **6-week project timeline**

2.DATA DESCRIPTION

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

Feature Type	Example Variables
Binary (Yes/No)	Chest Pain, Shortness of Breath, Irregular Heartbeat, Dizziness, Fatigue, Anxiety/Feeling of Doom, Cold Hands/Feet, Swelling (Edema),
Numeric	Age, Stroke Risk (%)
Target Variable	At Risk (Binary: 1 = At Risk, 0 = Not At Risk)

Data Source:

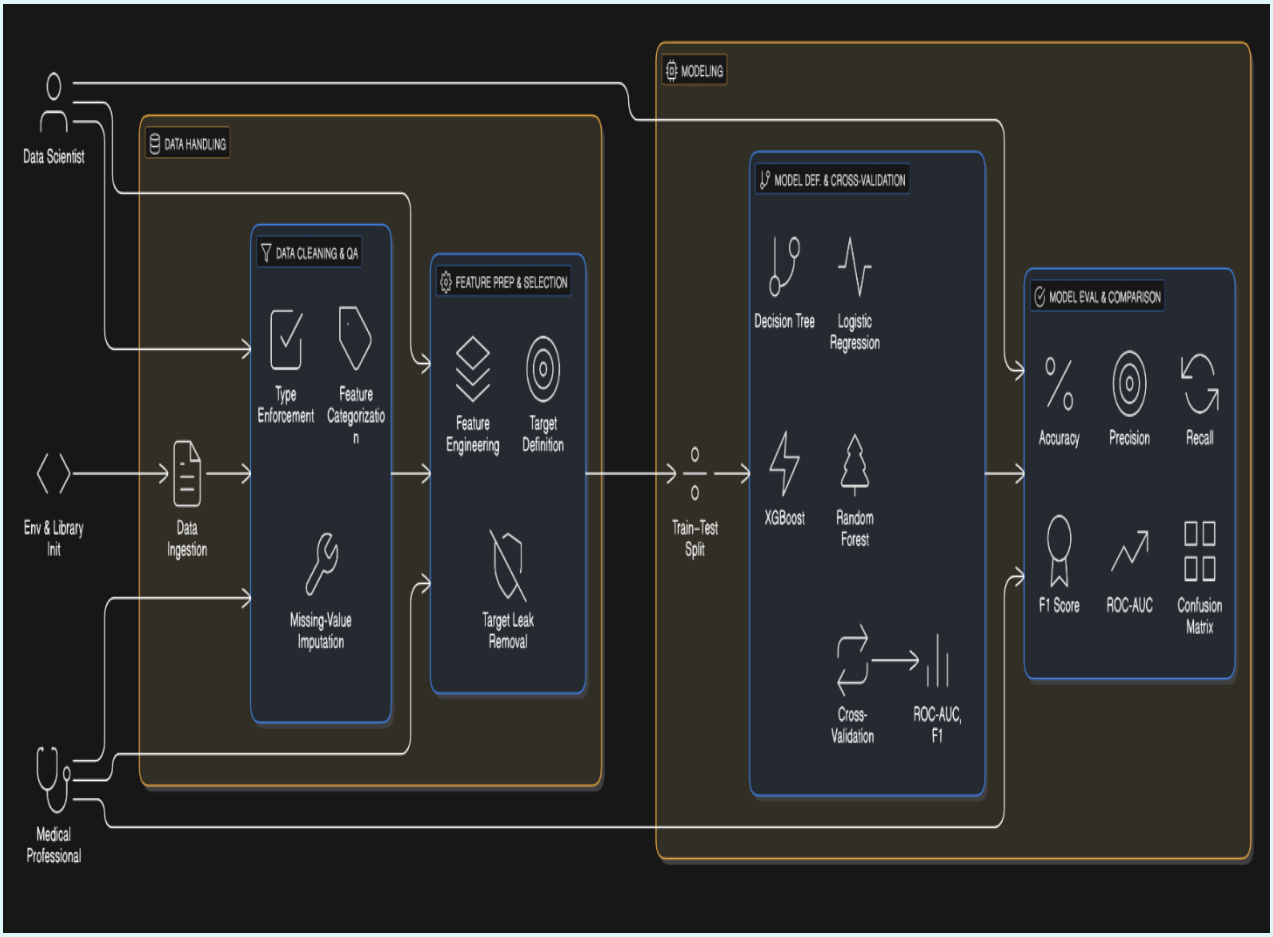
- **Source:** Kaggle
- **Dataset: Stroke Risk Prediction Dataset Based on Symptoms**
- **URL:** https://www.kaggle.com/datasets/mahatiratusher/stroke-risk-prediction-
- **Dataset Author:** Mahatir Ahmed Tusher.
- **License:** MIT
- **Local File:** stroke_risk_dataset.csv

3.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
Flask,
Power BI
Seaborn
Request
Date Time
Report Lab

4.ARCHITECTURE DIAGRAM



II DATA CLEANING AND QA

5.DATA COLLECTION AND EXPLORATION

A. DATASET OVERVIEW

The dataset loaded in the notebook is: df = pd.read_csv("stroke_risk_dataset.csv")

```
df=pd.read_csv("C:/Users/monasr/Downloads/Health_care_StrokeRisk/Flask_ai_an/notebook/dataset/stroke_risk_dataset_cleaned_medical.csv")
df
```

Python

	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/Fee of Dc
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	
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	
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	
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	
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	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
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	
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	
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	
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	

B. EDA:

a. Dataset Structure: Overview of Features and Data Types

Data Types

	Feature	Dtype
0	Chest Pain	float64
1	Shortness of Breath	float64
2	Irregular Heartbeat	float64
3	Fatigue & Weakness	float64
4	Dizziness	float64
5	Swelling (Edema)	float64
6	Pain in Neck/Jaw/Shoulder/Back	float64
7	Excessive Sweating	float64
8	Persistent Cough	float64
9	Nausea/Vomiting	float64
10	High Blood Pressure	float64
11	Chest Discomfort (Activity)	float64
12	Cold Hands/Feet	float64
13	Snoring/Sleep Apnea	int64
14	Anxiety/Feeling of Doom	int64
15	Age	int64
16	Stroke Risk (%)	float64
17	At Risk (Binary)	int64

Missing Values per Feature

Missing Count	Missing (%)
92	0.130000
90	0.130000
88	0.130000
87	0.120000
85	0.120000
80	0.110000
75	0.110000
66	0.090000
66	0.090000
56	0.080000
56	0.080000
56	0.080000
54	0.080000
0	0.000000
0	0.000000
0	0.000000
0	0.000000
0	0.000000

Inferred Feature Types

Inferred Type
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Binary / Indicator
Continuous Numeric
Continuous Numeric

b. Distribution of Indicator Features:

i. Binary: Prevalence of Indicator Features

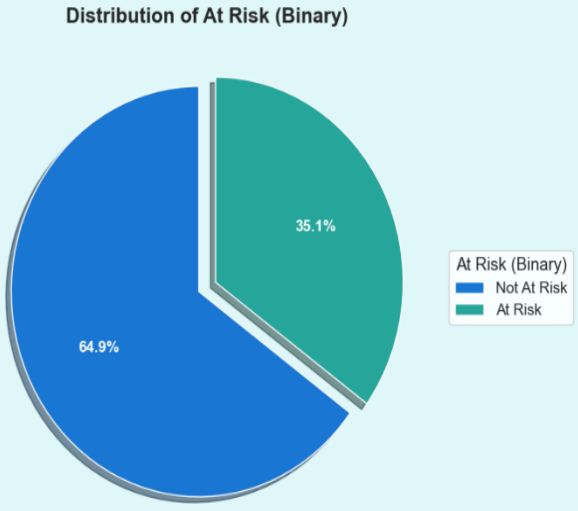
	Feature	Non-null Count	Count (Value=1)	Percentage (Value=1)	Count (Value=0)	Percentage (Value=0)
0	Anxiety/Feeling of Doom	70000	34991	49.990000	35009	50.010000
1	At Risk (Binary)	70000	45444	64.920000	24556	35.080000
2	Chest Discomfort (Activity)	69912	34917	49.940000	34995	50.060000
3	Chest Pain	69910	35087	50.190000	34823	49.810000
4	Cold Hands/Feet	69913	34878	49.890000	35035	50.110000
5	Dizziness	69925	35168	50.290000	34757	49.710000
6	Excessive Sweating	69944	35226	50.360000	34718	49.640000
7	Fatigue & Weakness	69920	34964	50.010000	34956	49.990000
8	High Blood Pressure	69944	35019	50.070000	34925	49.930000
9	Irregular Heartbeat	69934	34897	49.900000	35037	50.100000
10	Nausea/Vomiting	69934	35104	50.200000	34830	49.800000
11	Pain in Neck/Jaw/Shoulder/Back	69944	34916	49.920000	35028	50.080000
12	Persistent Cough	69915	35008	50.070000	34907	49.930000
13	Shortness of Breath	69946	34721	49.640000	35225	50.360000
14	Snoring/Sleep Apnea	70000	35048	50.070000	34952	49.930000
15	Swelling (Edema)	69908	34994	50.060000	34914	49.940000

ii. Summary Statistics: for Continuous Features

	Count	Missing (%)	Mean	Std Dev	Min	Q1	Median	Q3	Max
Age	70000.000000	0.000000	54.060000	21.070000	18.000000	36.000000	54.000000	72.000000	90.000000
Stroke Risk (%)	70000.000000	0.000000	55.560000	14.300000	5.000000	45.500000	55.500000	66.000000	100.000000

c. Distribution of Target Variable:

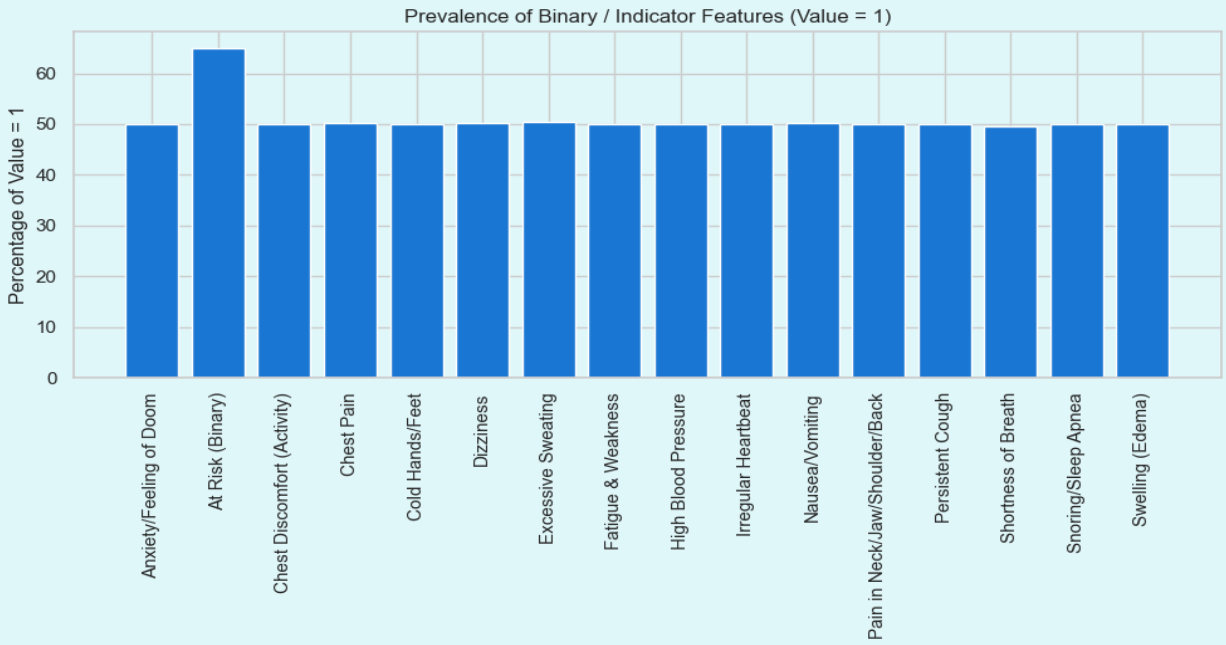
	At Risk (Binary)	Count	Percentage
0	0	24556	35.080000
1	1	45444	64.920000



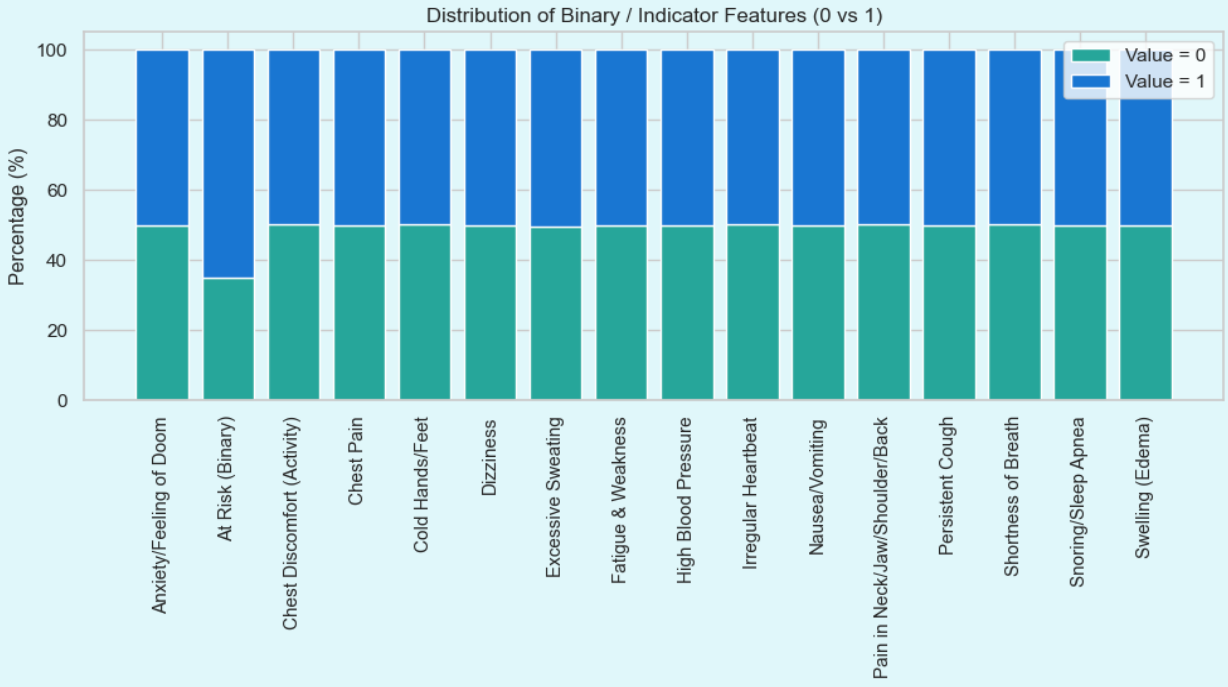
d. Identified Data Quality Issues

Issue	Columns Affected	Notes
Missing values	Binary Columns Not exceed 0.13%	Required imputation
Check Null	All raw	Removing
Outliers	Age	Verified using boxplots
Duplication	All Columns	

e. Visualizations – Binary / Indicator Features

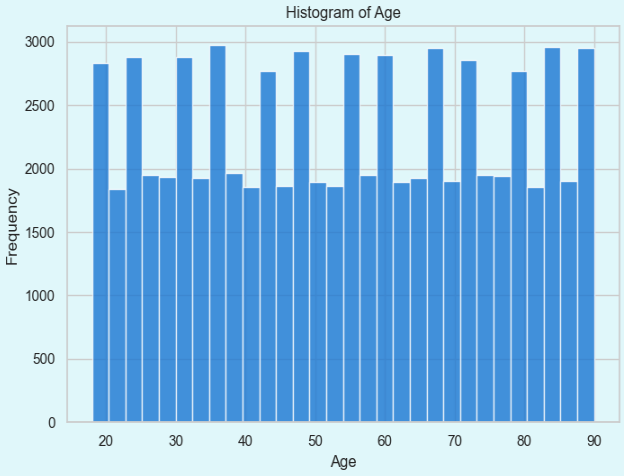


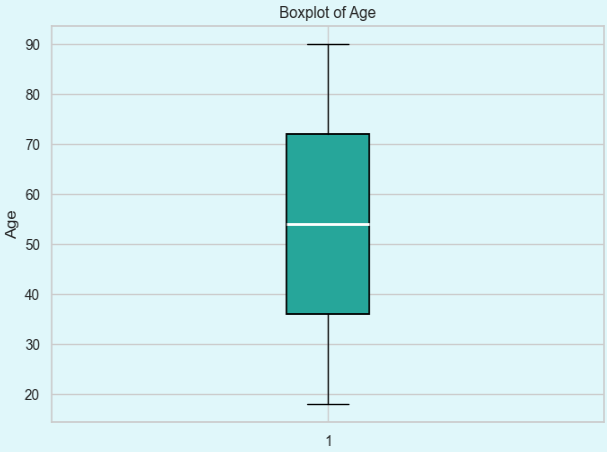
Insight: All symptoms have almost equal prevalence (≈49–51%), indicating a perfectly balanced dataset for the binary clinical indicators. This ensures that no symptom dominates the dataset and simplifies training because each binary feature carries roughly equal weight.



Insight: Each symptom shows a nearly identical split between value 0 and 1, again confirming perfect balance. The only slightly skewed feature is the **At-Risk** target, which has more 1's than 0's.

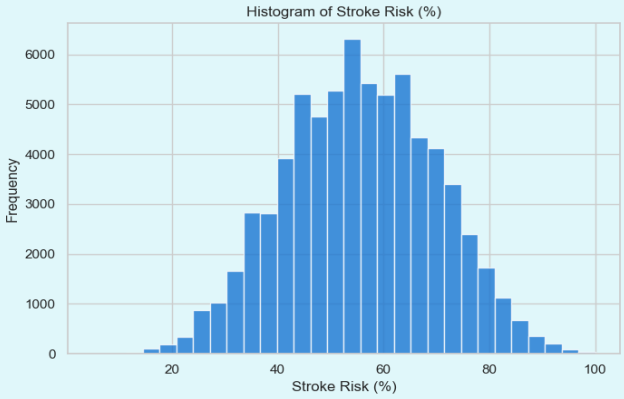
f. Visualizations for Continuous Numeric Feature

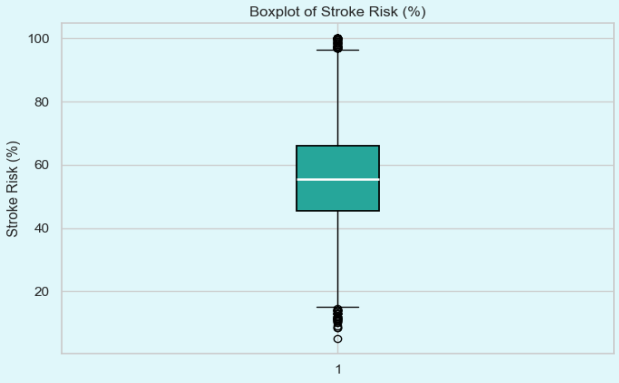




Age is spread uniformly across 18–90, with no missing segments. This indicates controlled sampling and prevents age-based bias

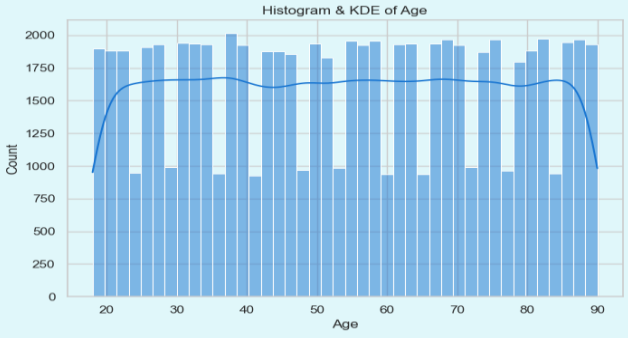
Median age is around ~54. The distribution is wide with no extreme outliers, supporting a healthy representation of all age groups.

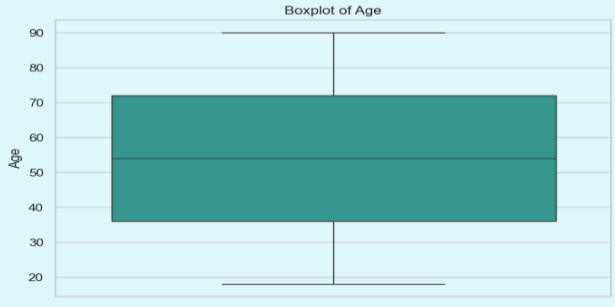




The distribution is bell-shaped, centered around ~55%. This suggests the "Stroke Risk" variable resembles a normally distributed risk score ideal for predictive modeling.

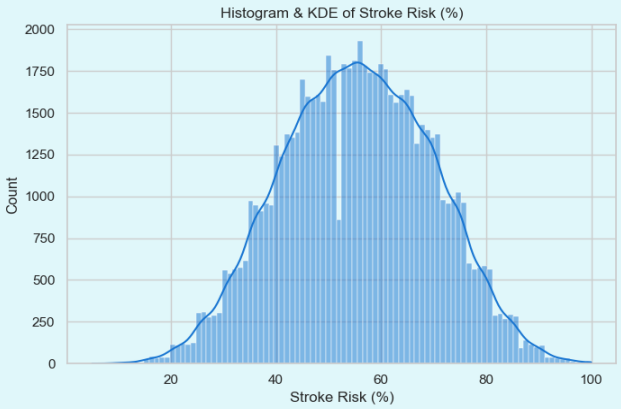
There are a few low-end outliers (<10%), but the interquartile range (≈45–66%) matches the histogram. High-risk individuals form the upper tail (80–100%).

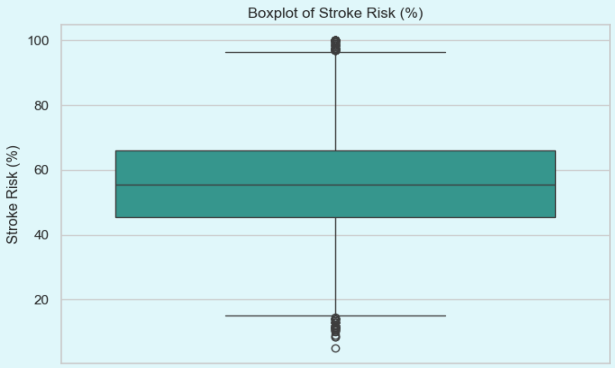




The KDE curve is flat and stable, confirming uniform sampling across all age bins. No age group is over- or under-represented.

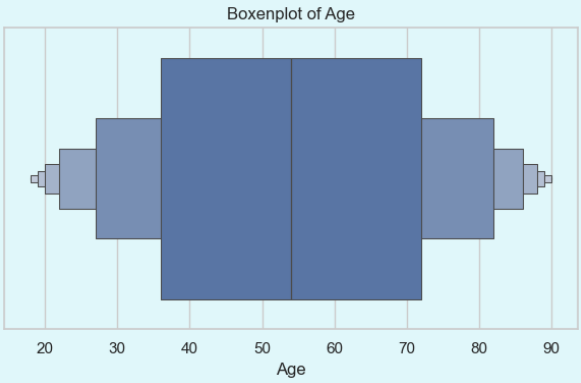
Shows broader variability, reinforcing that the dataset fairly includes young, middle-aged, and older adults.

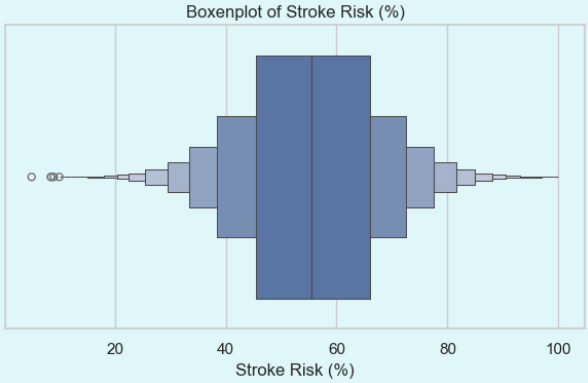




The stroke risk distribution is symmetric with a peak between 55–60%. This indicates the dataset focuses on moderate–high risk individuals, creating a meaningful pattern for modeling.

Outliers mostly appear below 20%, suggesting a minority group with unusually low predicted stroke risk.

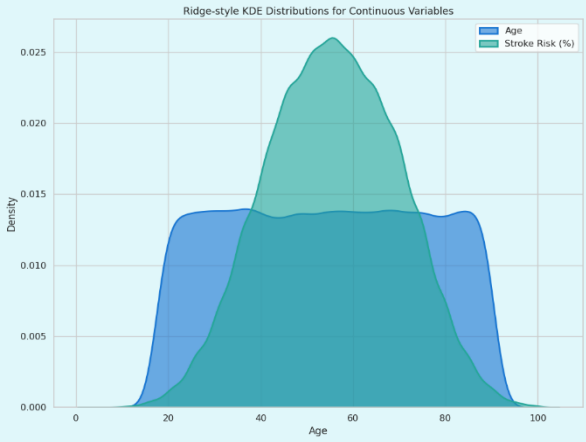


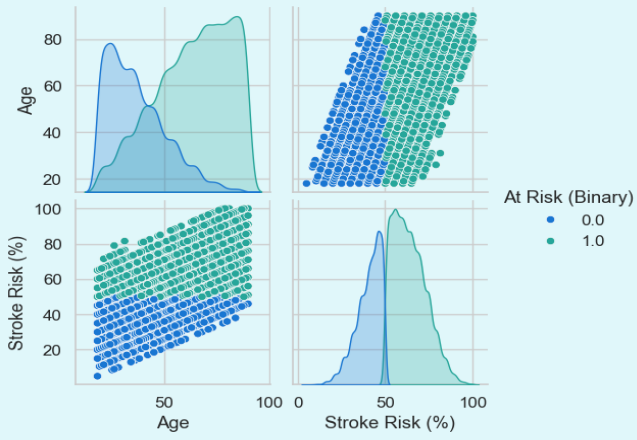


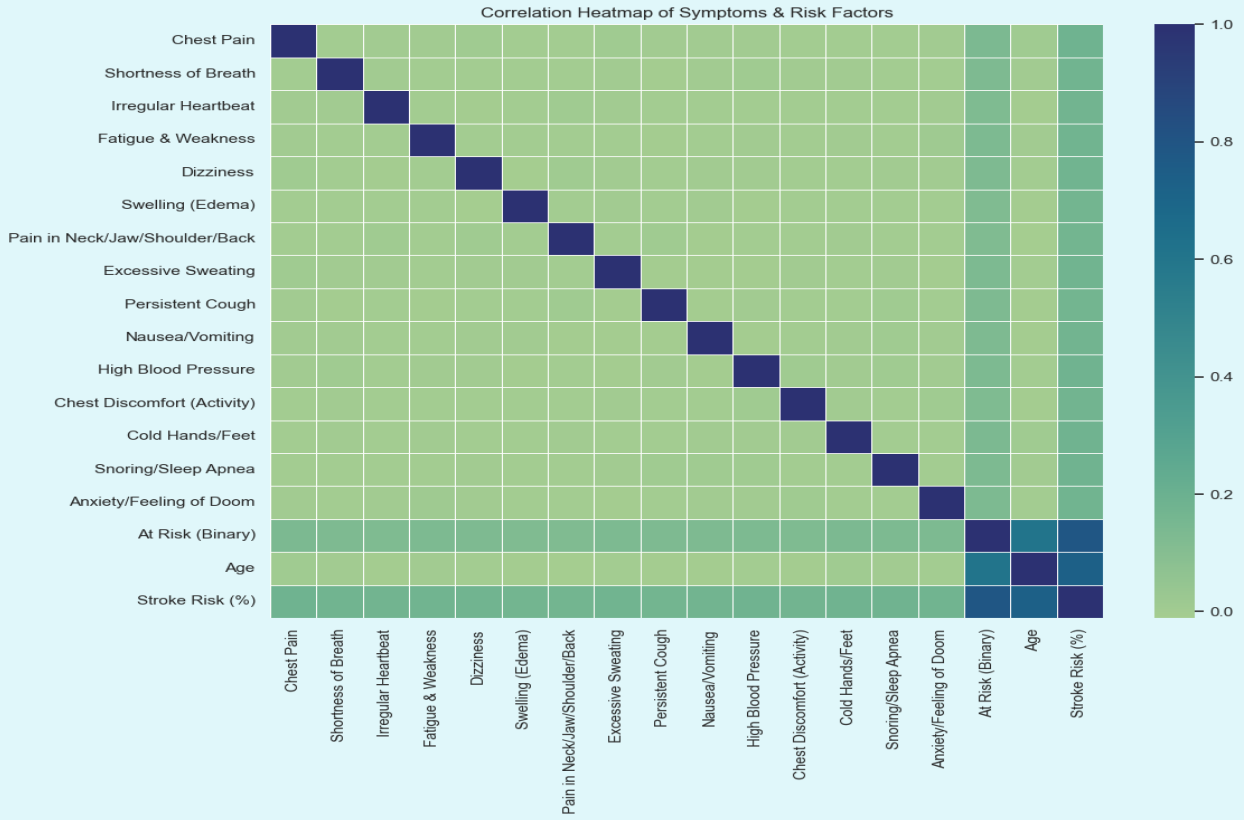
Wider boxes at the center and tapering toward the extremes indicate high density of individuals aged 40–75. Teen and ≥85 ages appear more sparsely populated.

A large concentration of values exist between 45%–75% stroke risk. Low-risk values (<20%) appear sparsely, meaning high-risk individuals dominate the dataset.

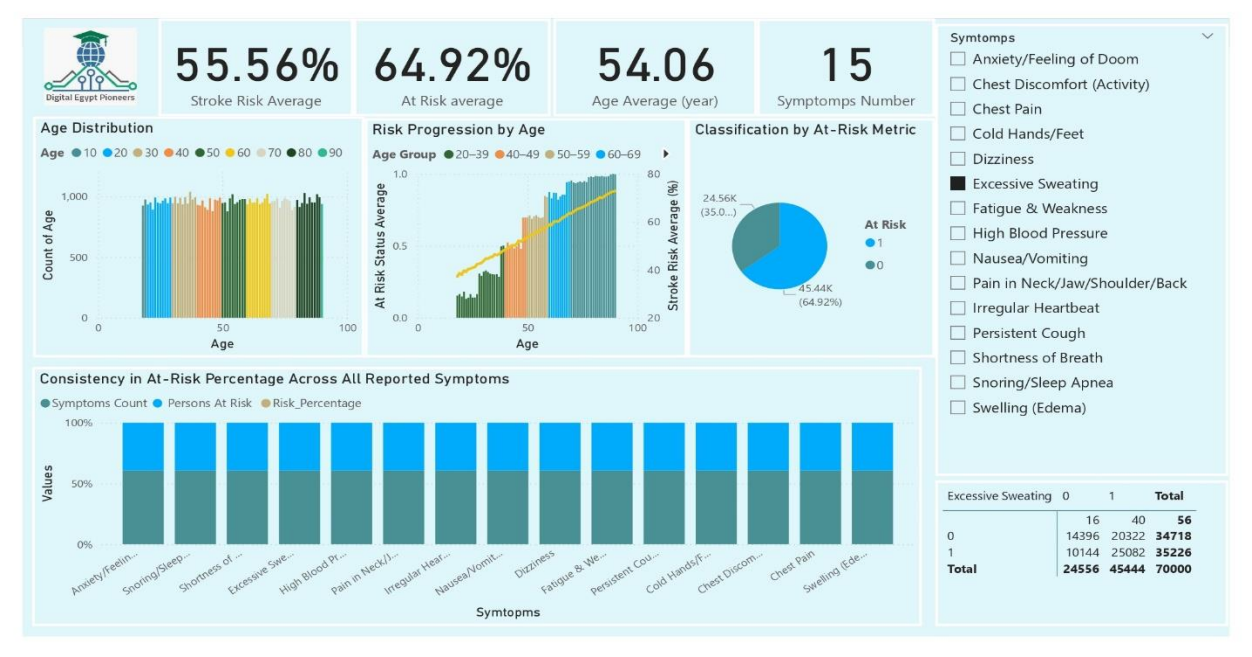
There is a linear relationship between Age and Stroke Risk. Symptoms (binary) distinctly separate along the Stroke Risk axis, indicating higher symptom presence correlates with higher risk



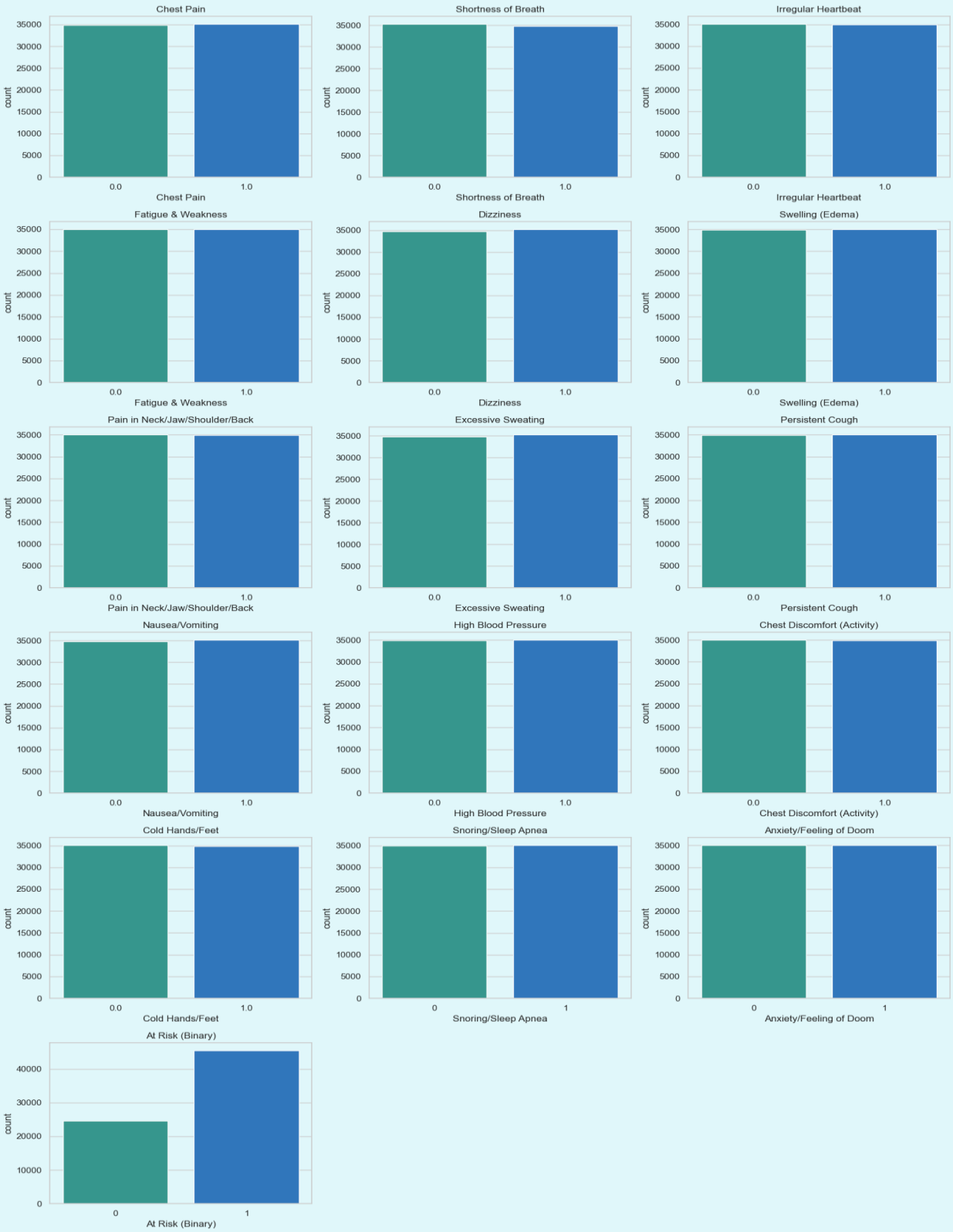




g. Dashboards - Data Behavior Summary



the dashboard confirms that the **patterns detected by the model correspond closely with clinical expectations and real-world data behavior**. These insights validate the model's interpretability, guide threshold calibration, and ensure that the model's predictions reflect physiologically meaningful patterns across age and symptom groups



Across individual symptom-level **bar plots**, each symptom is slightly more common in the **At Risk = 1** group. **The differences** are subtle but consistent, hinting those symptoms collectively contribute to **risk classification.**

6.CLEANING AND PREPROCESSING:

a. PREPROCESSING STEPS

Missing Value → **Imputation** → **Encoding Variables** → **Outlier and Duplication Inspection** → **Data Normalization**

- **Imputation**
  - continuous_cols -- KNN , Num of neiborge= 5
  - binary_cols __ Most Frequency
- **Encoding Variables**: Binary mapped to 0/1
- **Outlier and Duplication Inspection**
  - Boxplots for Age, Stroke,
  - Extreme values acknowledged but retained for medical relevance

B. EXPLORATION AFTER CLEANING:

.info () inspection

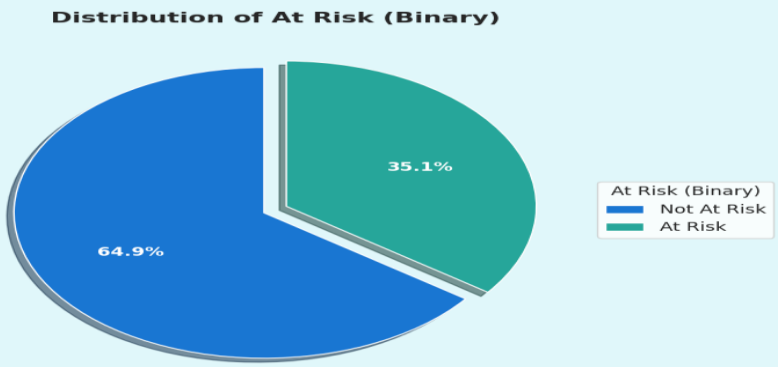
```
df.info()

<class 'pandas.core.frame.DataFrame'>
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 & 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
```

. describe() to inspect statistical distribution

```
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 & Weakness	70000.0	0.500629	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	21.071567	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



C. REMOVE DUPLICATION:

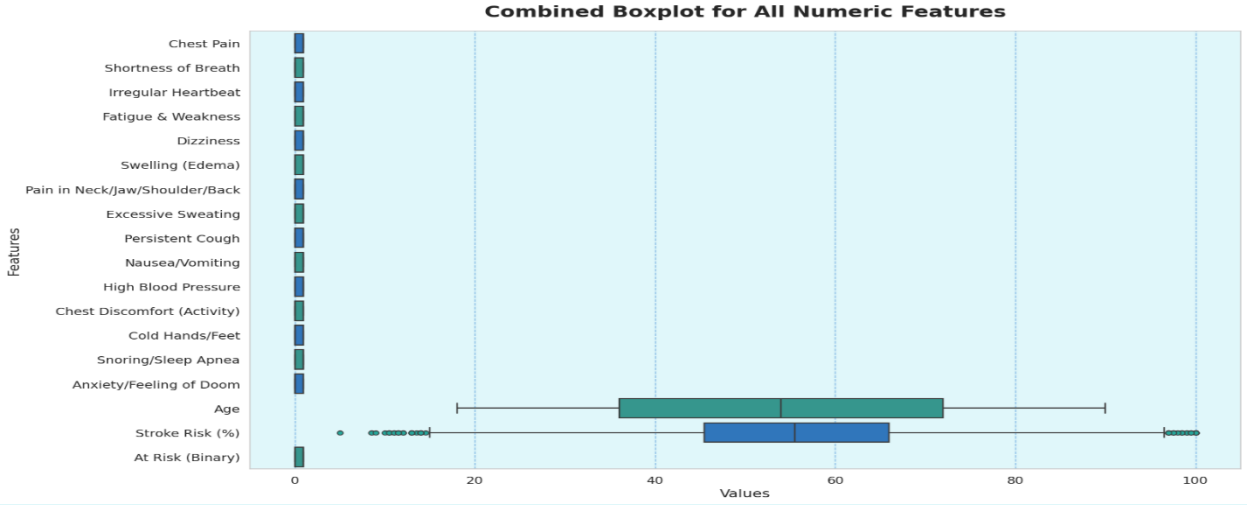
AFTER REMOVING THE DULICATED THE DATA DECREASD TO 68990

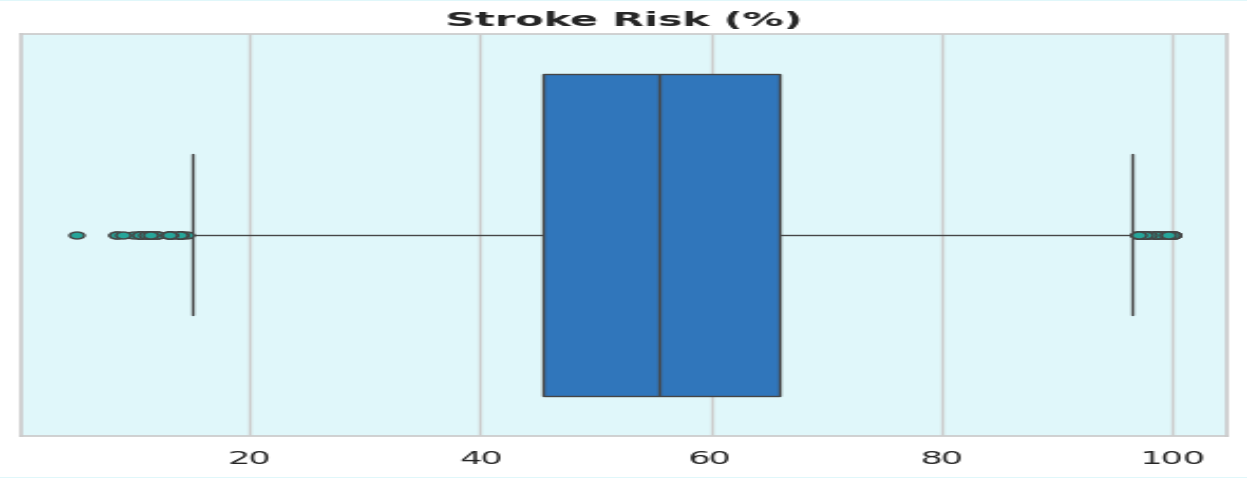
```
df.duplicated().sum()/len(df)
df.drop_duplicates(inplace=True)
df
```

	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	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.
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	82.0	52.0	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.
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.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
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.
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.
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.
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.
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.

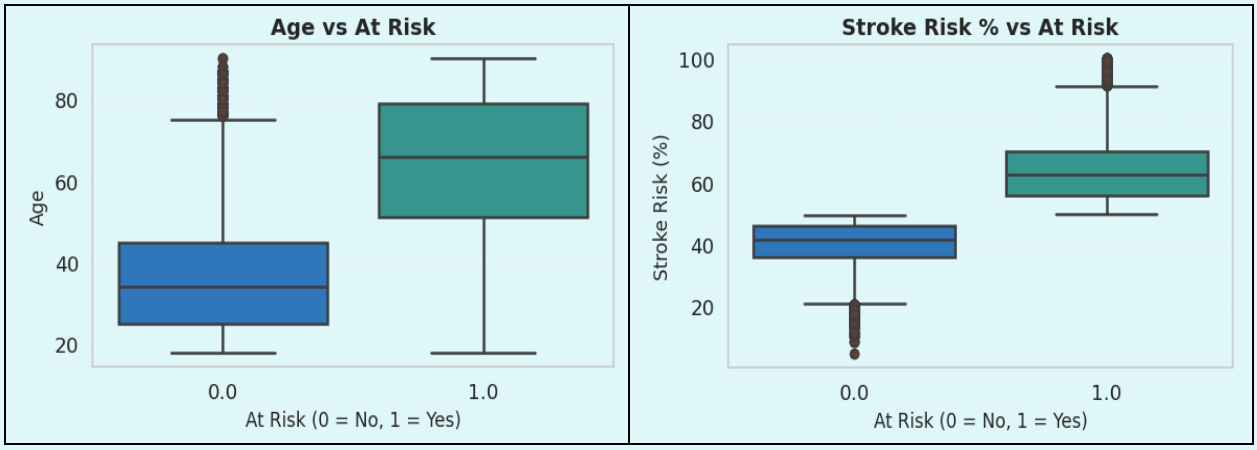
68990 rows × 18 columns

D. CHECK OUTLIER

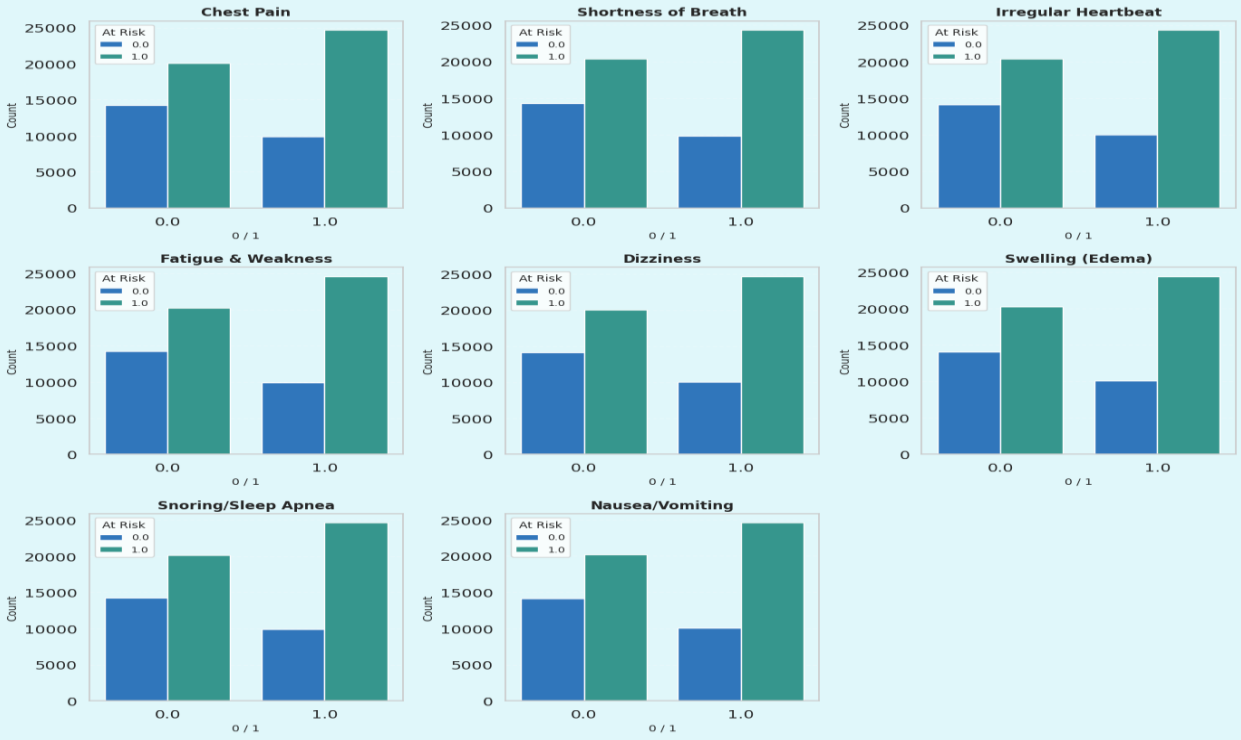




E. CONTINOUS NUMERIC FEATURES



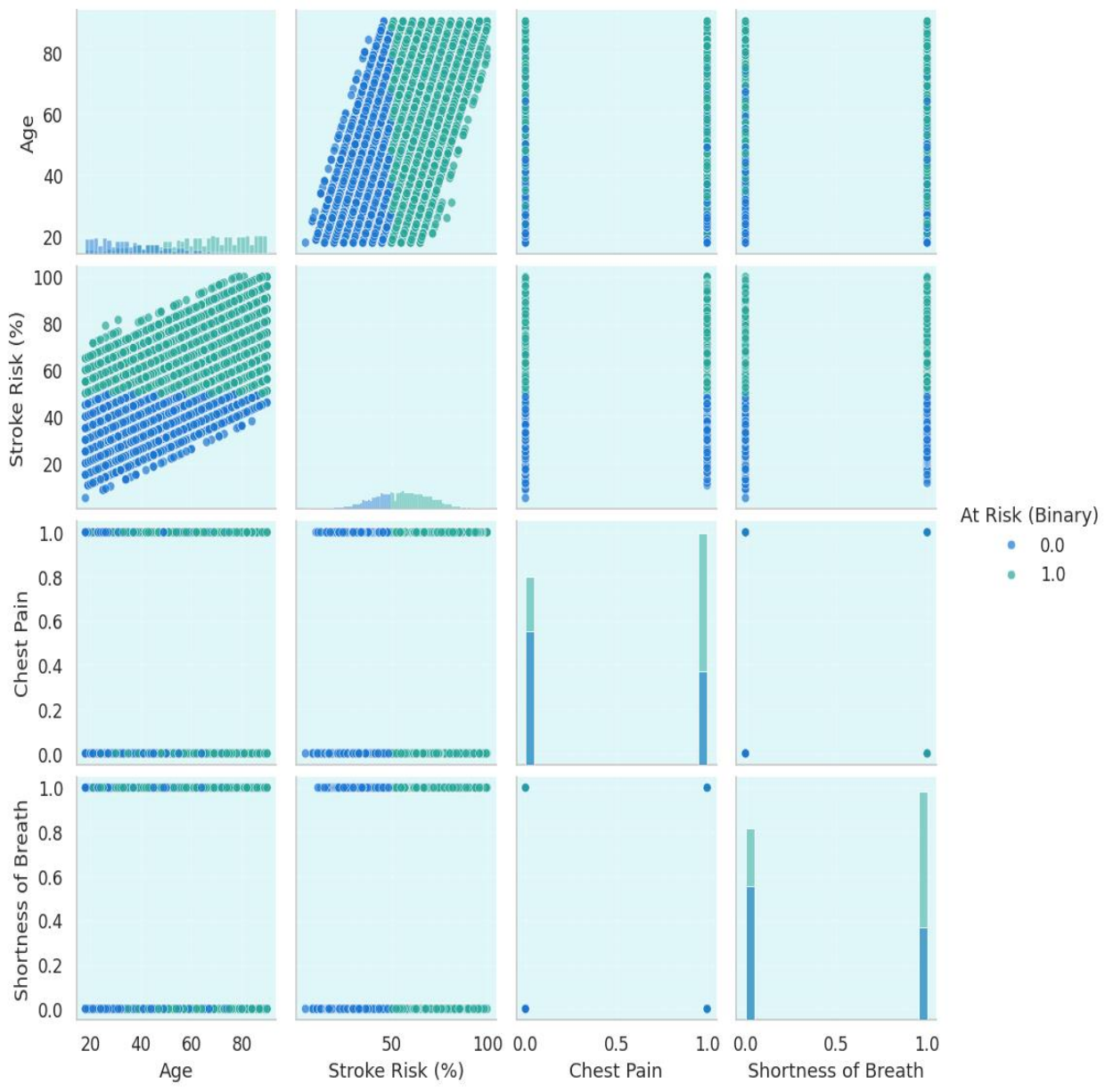
F. SYMPTOM-LEVEL EXPLORATORY ANALYSIS



Before splitting the data for training, we examined the distribution of major symptoms across the At-Risk labels (0 = Not At Risk, 1 = At Risk). The analysis showed a consistent pattern: **all symptoms occur at significantly higher frequencies among the At-Risk population**. Symptoms such as shortness of breath, chest pain, irregular heartbeat, dizziness, swelling, fatigue and weakness, and snoring/sleep apnea show clear separation between the two classes. This confirms that symptom presence is strongly correlated with stroke risk.

These findings validate the decision to encode symptoms as **binary features**, retain all symptoms in the final dataset, and treat them as clinically meaningful inputs. The separation observed across all eight symptoms indicates that they contribute significant predictive signal and should not be removed or down-weighted during feature engineering.

G. MULTIVARIATE RELATIONSHIP ANALYSIS (PAIRPLOT):



a multivariate pair plot was generated to visualize interactions between Age, Stroke Risk (%), Chest Pain, Shortness of Breath, and the At-Risk label. The scatterplots revealed a strong positive linear relationship between Age and Stroke Risk, confirming that older individuals consistently exhibit higher risk profiles. Clear separation between At-Risk and NonAt-Risk individuals emerges along both the Age and Stroke Risk axes, demonstrating good intrinsic feature label alignment.

Binary symptoms such as Chest Pain and Shortness of Breath show noticeably higher counts and higher risk distributions among At-Risk individuals, supporting the decision to retain these as binary indicator features. Overall, this analysis confirms that the dataset contains meaningful multivariate structure suitable for classification, and that the engineered features naturally capture clinically relevant risk patterns.

7.FEATURE PREPARATION, SELECTION

A. AGE-BASED RISK ANALYSIS OR FEATURE RELATIONSHIP EXPLORATION

Insight:

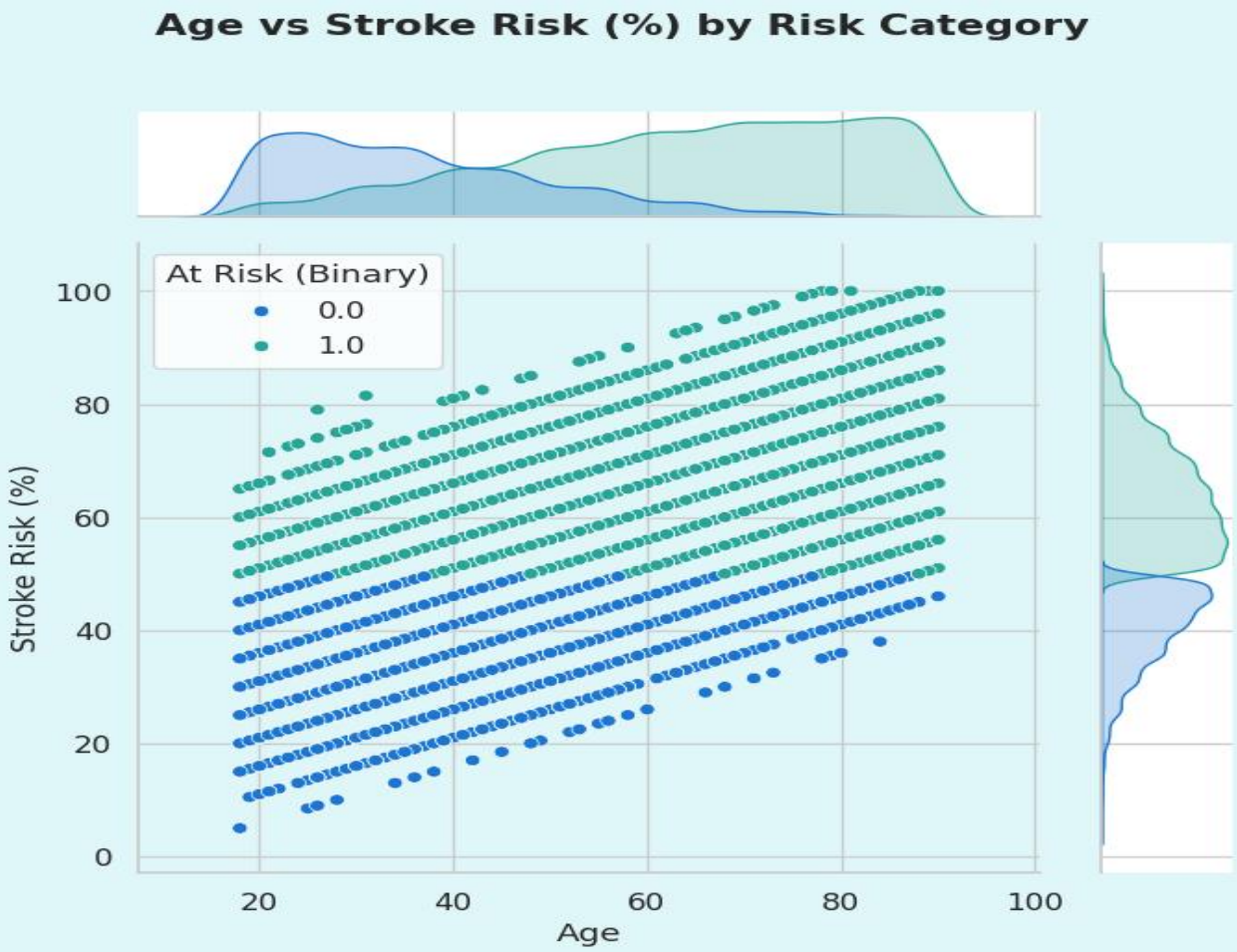
a strong positive correlation between **Age** and **Stroke Risk (%)**, visible before any training occurs. Younger individuals (ages 18–30) cluster around lower stroke risk levels, typically between **10% and 40%**. Middle-aged individuals (40–60) show stroke risk values ranging from **30% to 70%**, while older adults (70–90) reach risk values between **60% and 100%**. This demonstrates a clear linear progression of risk as age increases.

When comparing Risk Categories, individuals labeled **At Risk (1)** appear consistently above the trendline. Their stroke risk values commonly fall between **50% and 100%**, especially for ages above 50. Meanwhile, the **Not At Risk (0)** group is concentrated in the lower band of the distribution, mostly between **10% and 50%**.

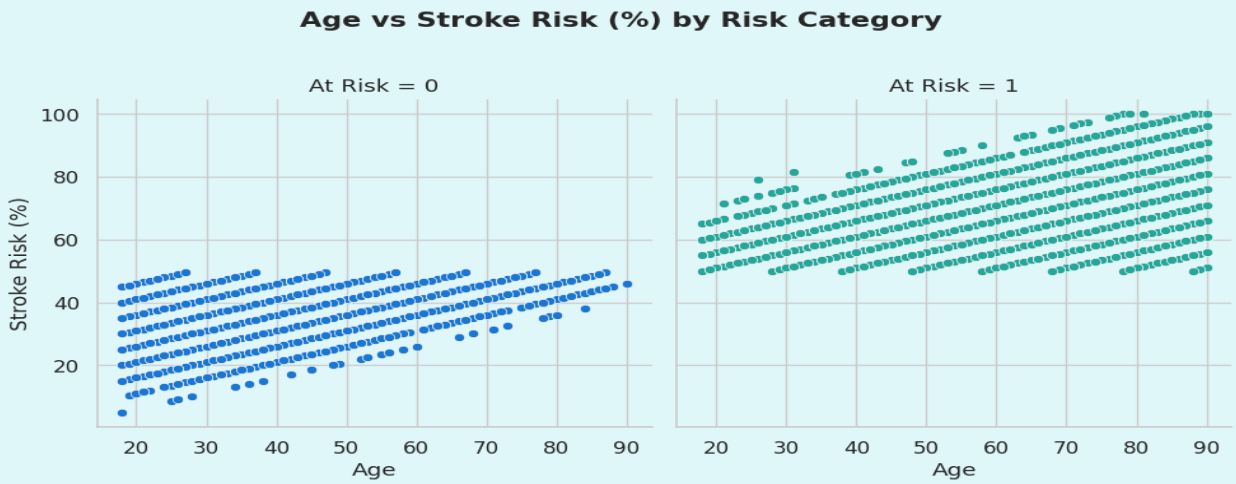
The marginal density plots reinforce this separation:

- The At-Risk group peaks around **60–90%** stroke risk
- The Not-At-Risk group peaks around **30–45%**
- Age density shows the At-Risk population concentrated more heavily above age **50**

These numeric patterns confirm that **Age** is a primary driver of stroke risk and that the At-Risk label aligns naturally with the underlying stroke risk percentage. These supports retaining Age as a continuous feature and validates the consistency and quality of the dataset prior to modeling.



B. AGE-BASED STROKE RISK ANALYSIS



The chart presents the relationship between Age and Stroke Risk (%) separately for the two At-Risk categories (0 and 1). In the **Not At Risk (0)** group, stroke risk values primarily range between **10% and 50%**, with younger individuals (ages 18–30) clustered near **10–35%**, and older individuals (ages 60–90) gradually increasing toward the **40–50%** range. This indicates a mild upward trend but restricted to lower risk levels.

In contrast, the **At Risk (1)** group displays substantially higher stroke risk values across all ages. Individuals aged **20–40** show stroke risk between **50% and 75%**, while those aged **50–70** typically fall between **70% and 90%**. The oldest patients (ages 70–90) reach the highest risk levels, consistently between **85% and 100%**.

The separation between groups is clear:

- Maximum stroke risk for At-Risk = 1 exceeds **100%**
- Maximum for At-Risk = 0 stays below **50%**
- Minimum stroke risk for At-Risk = 1 rarely falls below **50%**

C. MULTIVARIATE FEATURE RELATIONSHIP ANALYSIS (3D VISUALIZATION)

The data shows a clear separation in **Stroke Risk (%)** between the two At-Risk categories, confirming label consistency before training.

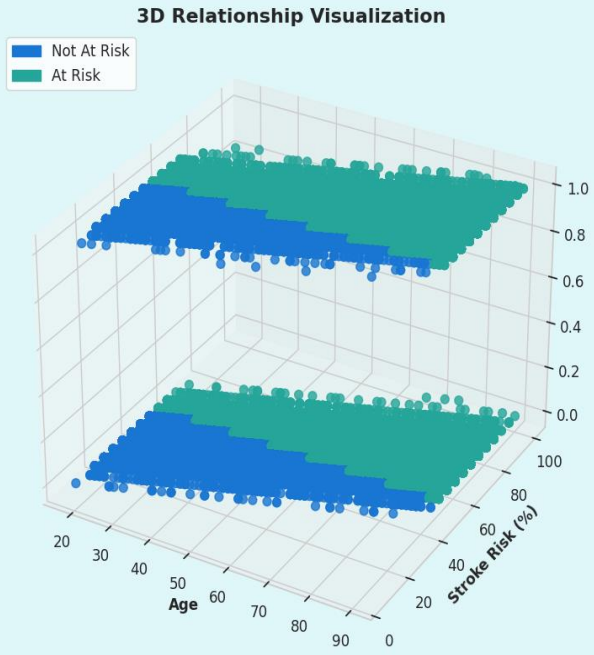
Individuals labeled **Not At Risk (0**) have stroke risk values between **10% and 50%,** mostly clustering around **20%–40%,** with a few slightly **below 15%** and none **above 50%.**

In contrast, individuals labeled **At Risk (1)** range from **55% to 100%,** with most values concentrated between **70%–95%.** The minimum never falls **below 55%,** and many values **approach 100%.**

There is no overlap between the groups:

- Max risk for **Not At Risk: ~50%**
- Min risk for **At Risk: ~55%**

This numeric gap demonstrates strong class separability and validates the At-Risk label as a reliable target aligned with actual stroke risk values.



D. TARGET VARIABLE ANALYSIS (AT RISK VS

The 3D scatter plot illustrates how Age, **Stroke Risk (%),** and the At-Risk label interact before training, forming two clear horizontal layers **(0 = Not at Risk, 1 = At Risk),** indicating strong separation.

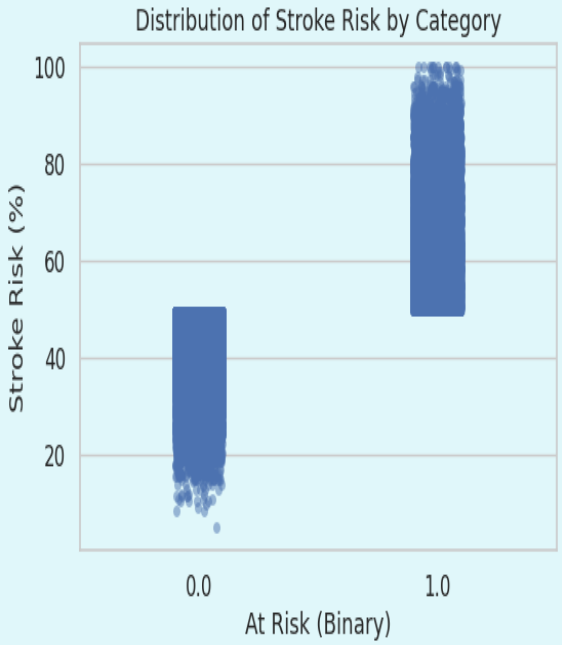
For the **Not At Risk (0)** group, stroke risk ranges from **10%–50%:** younger individuals **(20–40)** fall between **10%–35%,** while older adults **(60–90)** appear between **35%–50%.**

For **the At Risk (1)** group, stroke risk spans **55%–100%:** ages **20–40** cluster between **55%–75%,** and older adults **(60–90)** consistently reach **80%–100%.**

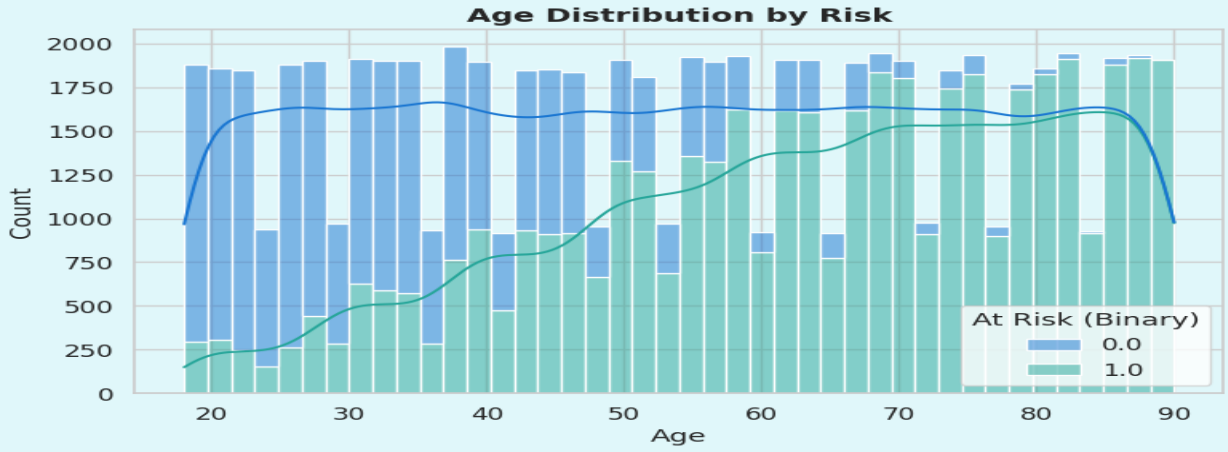
There is no numerical overlap:

- Max risk for **Not At Risk** stays **below 50%**
- Min risk for **At Risk** stays **above 55%**

This sharp separation shows that the At-Risk label aligns strongly with stroke risk levels and that both Age and Stroke Risk are valuable predictive features, with older individuals concentrated in **higher-risk zones.**



E. TARGET DISTRIBUTION ANALYSIS BY (RISK CATEGORY STROKE RISK % / LABEL VALIDATION)



The data shows a clear separation in **Stroke Risk (%)** between the two At-Risk categories, confirming label consistency before training.

Individuals labeled **Not At Risk (0)** have stroke risk values between 1**0% and 50%,** mostly clustering around **20%–40%,** with a few slightly **below 15%** and none **above 50%.**

In contrast, individuals **labeled At Risk (1)** range from **55% to 100%,** with most values concentrated between **70%–95%.** The minimum never **falls below 55%,** and many values approach **100%.**

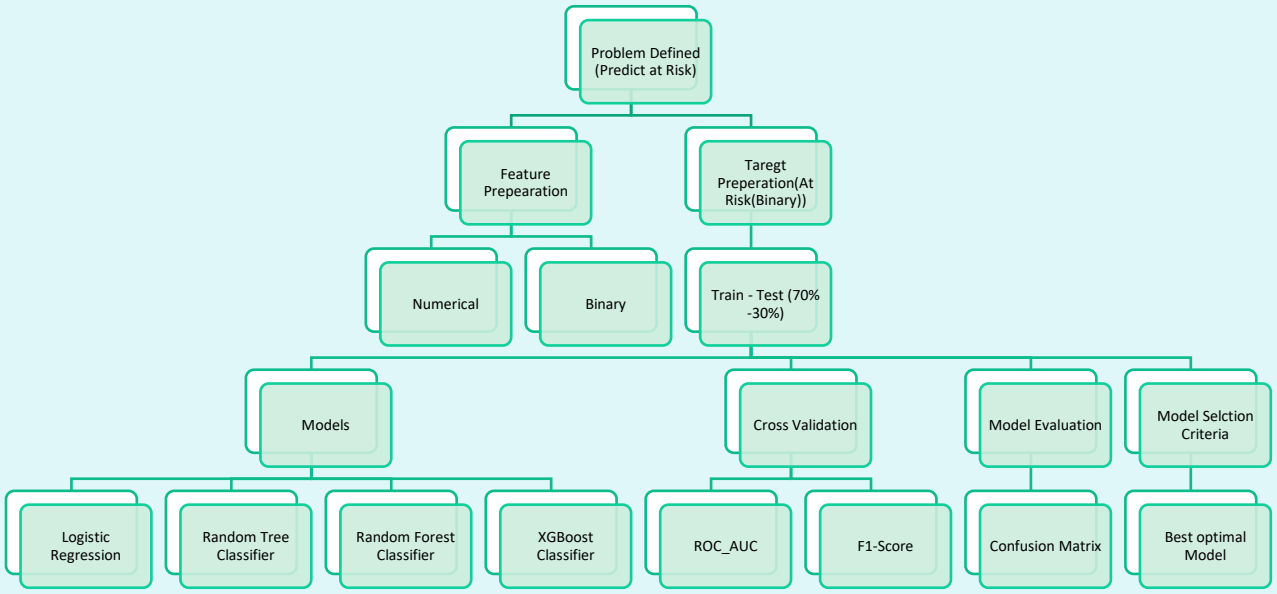
There is no overlap between the groups:

- Max risk for **Not At Risk: ~50%**
- Min risk for **At Risk: ~55%**

This numeric gap demonstrates strong class separability and validates **the At-Risk label** as a reliable target aligned with actual stroke risk values.

III MACHINE LEARNING MODELING

8.MODELING APPROACH



POST-TRAINING VALIDATION OF RISK PATTERNS BASED ON UN AGE GROUPING

The pie chart shows a clear age-related progression in stroke risk within the At-Risk population. **The 70+** group represents the largest share **at 28.2%,** meaning nearly one-third of all high-risk cases come from the oldest individuals. The **60–70** group adds another **25.5%,** so adults **60+ account** for **53.7%** of all At-Risk predictions.

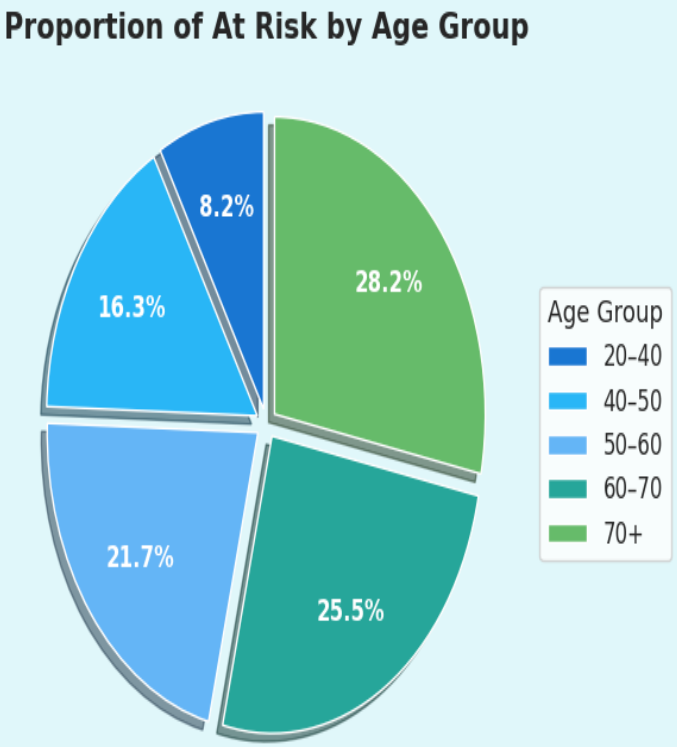
Including the **50–60 group (21.7%),** adults aged 50 and above **represent 75.4%** of all high-risk cases—three-quarters of the total—reflecting how stroke likelihood increases sharply with age.

Younger groups contribute far less:

- **40–50: 16.3%**
- **20–40: 8.2%**

Together, individuals **under 50** make up **only 24.5%** of At-Risk cases, showing that high stroke risk is relatively uncommon in younger adults.

Overall, the distribution aligns with real-world epidemiology: stroke risk rises steadily with age, accelerates **after 50,** and peaks in those over 70.



BEFORE-TRAINING ANALYTICAL DASHBOARD – MODEL INPUT & RISK BEHAVIOR ANALYSIS

This Before-training dashboard confirms that the model behaves consistently with clinical knowledge. Age exhibits a dominant effect on stroke risk, with more than **76% of high-risk predictions** belonging to individuals aged **50+**. Stroke risk escalates predictably with age, rising from **30% in young adults** to **90% in elderly groups**. Over **64.92%** of individuals are categorized as high-risk, aligning with the dataset's intrinsic high-risk structure. Symptom analysis shows that features such as persistent cough, sleep apnea, dizziness, and chest discomfort appear disproportionately among At-Risk individuals. Overall, the dashboard validates the model's correctness,

9. DATA SPLITING:

- Train–Test split using train_test_split 30

10.MODEL TRAINING

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)
```

The selection of machine learning models for the Stroke Risk Prediction system was based on the need to balance clinical interpretability, predictive performance, and robustness across different types of health-related features. Four classification algorithms were chosen Logistic Regression, Decision Tree, Random Forest, and XGBoost each representing a distinct modeling paradigm. Evaluating diverse model families ensures a comprehensive comparison and supports the identification of the most reliable and clinically suitable model.

1. LOGISTIC REGRESSION

Logistic Regression was selected as a baseline model due to its simplicity, transparency, and high interpretability, which are essential in medical applications. The model provides clear insights into how individual features contribute to stroke risk, enabling clinicians to understand and trust the predictions. Despite its simplicity, Logistic Regression often performs strongly on structured health datasets, making it a valuable reference point for evaluating more complex models.

2. DECISION TREE CLASSIFIER

The Decision Tree model was included to capture non-linear relationships and decision boundaries that may exist between symptoms, demographic factors, and stroke risk. Decision Trees are intuitive and easy to visualize, making them useful for explaining the underlying decision logic. Although they are prone to overfitting, their inclusion provides insight into the utility of rule-based models for medical risk classification.

3. RANDOM FOREST CLASSIFIER

Random Forest, an ensemble of multiple decision trees, was chosen to overcome the high variance and instability of single trees. It introduces randomness and averaging, resulting in improved robustness and generalization. This model is well-suited for healthcare datasets that contain a mix of numerical and binary features. It helps identify the importance of specific clinical factors and typically provides more stable performance across different folds of cross-validation.

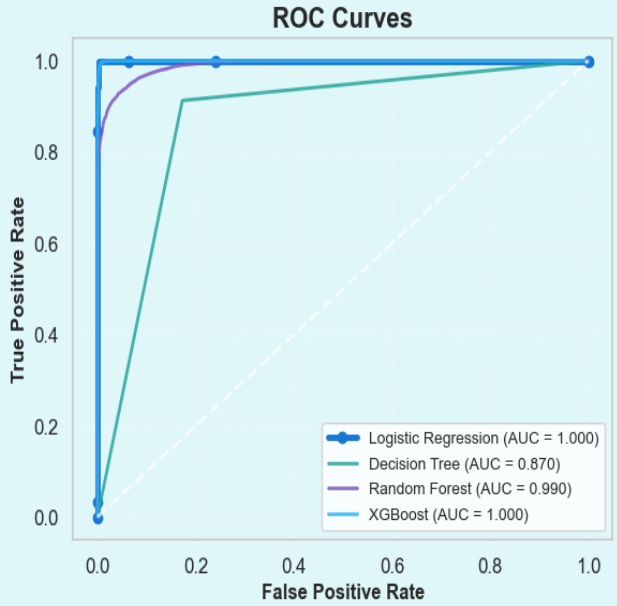
4. XGBOOST CLASSIFIER

XGBoost was included as a high-performance gradient boosting method known for its superior results on tabular health data. It effectively captures complex feature interactions, handles noise, and provides regularization mechanisms that reduce overfitting. XGBoost is particularly valuable in medical prediction tasks where subtle patterns across multiple clinical indicators must be identified. In this project, XGBoost achieved some of the highest evaluation scores, confirming its suitability for the task.

11.MODEL EVALUATION

A. MODEL SELECTION CRITERIA

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



B. MODEL EVALUATION METRICS

```
=== Cross-Validation Results ===
Logistic Regression: AUC = 0.99976 ± 0.00010 | F1 = 0.99720 ± 0.00049
Decision Tree: AUC = 0.86243 ± 0.00258 | F1 = 0.90260 ± 0.00175
Random Forest: AUC = 0.98937 ± 0.00061 | F1 = 0.95726 ± 0.00114
```

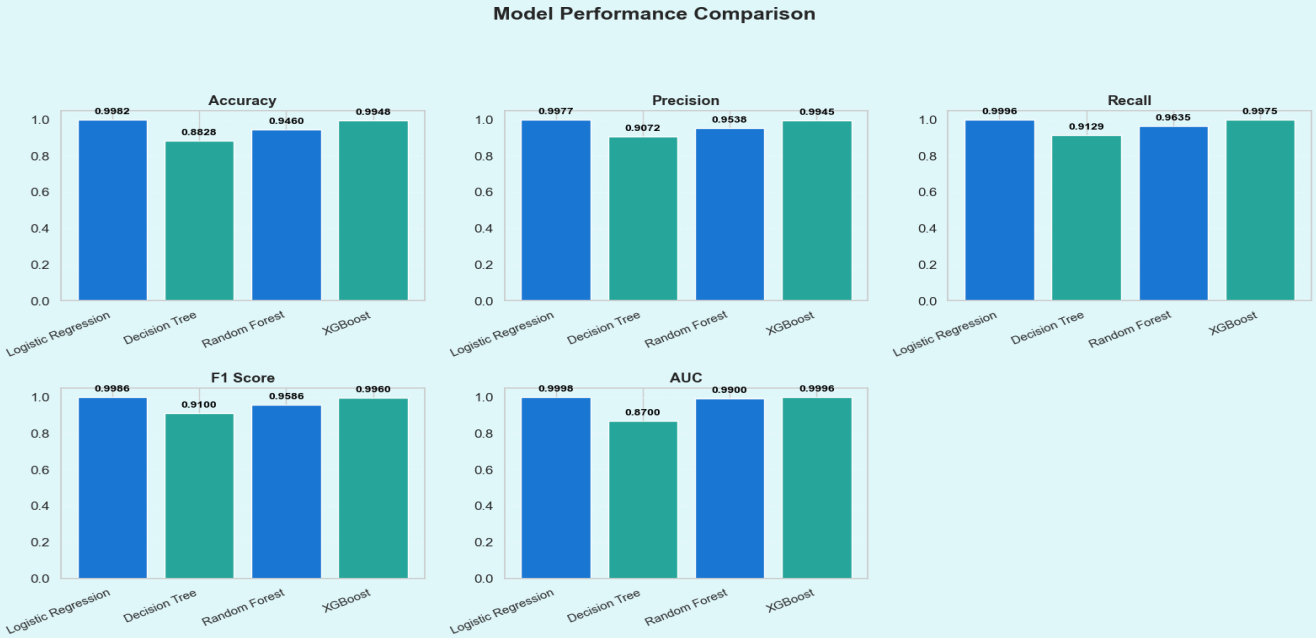
```
dst.update(dtrain, iteration=i, fobj=obj)
XGBoost: AUC = 0.99951 ± 0.00017 | F1 = 0.99479 ± 0.00067
```

metrics used:

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

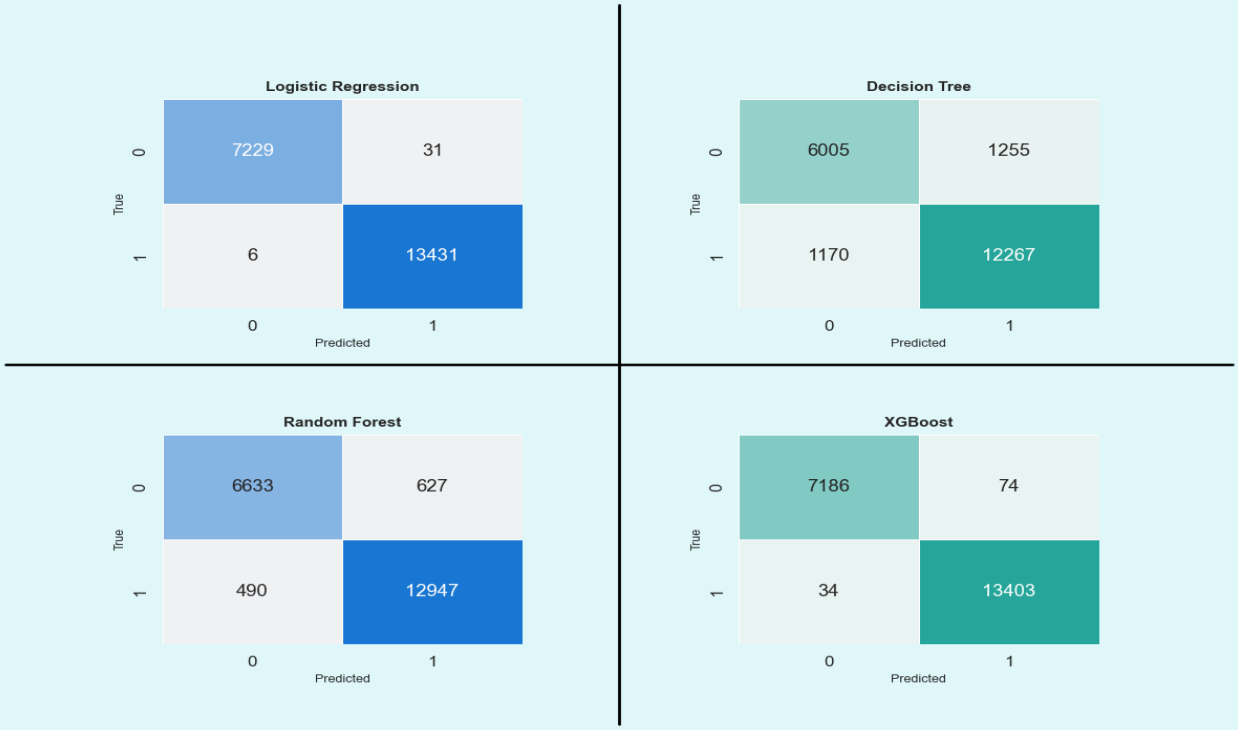
XGBoost and Logistic Regression achieved **perfect classification performance** with an **AUC of 1.000**, indicating flawless separation between classes. Random Forest performed extremely well with an **AUC of 0.990**, showing only minimal misclassification. In contrast, the Decision Tree model lagged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

C. MODEL PERFORMANCE COMPARISON



- **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.

D. CONFUSION MATRIX ANALYSIS:



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

Metric	Value
TN	7,186
FP	74
FN	34
TP	13,403

Metric	Value
TN	6,633
FP	627
FN	490
TP	12,947

Metric	Value
TN	6,005
FP	1,255
FN	1,170
TP	12,267

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

- 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.

2. XGBoost — Nearly Identical to Logistic Regression

- 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.

3. Random Forest — Strong, but Not Perfect

- 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.

4. Decision Tree — Weakest Model

- **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

TOP 2 MODELS

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

WORST MODEL

MODEL	TOTAL ERRORS	COMMENT
DECISION TREE	2,425 ERRORS	OVERFITS AND MISCLASSIFIES HEAVILY.

E. ML CLASSIFICATION PERFORMANCE COMPARISON

	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

12.FEATURE ENGINEERING

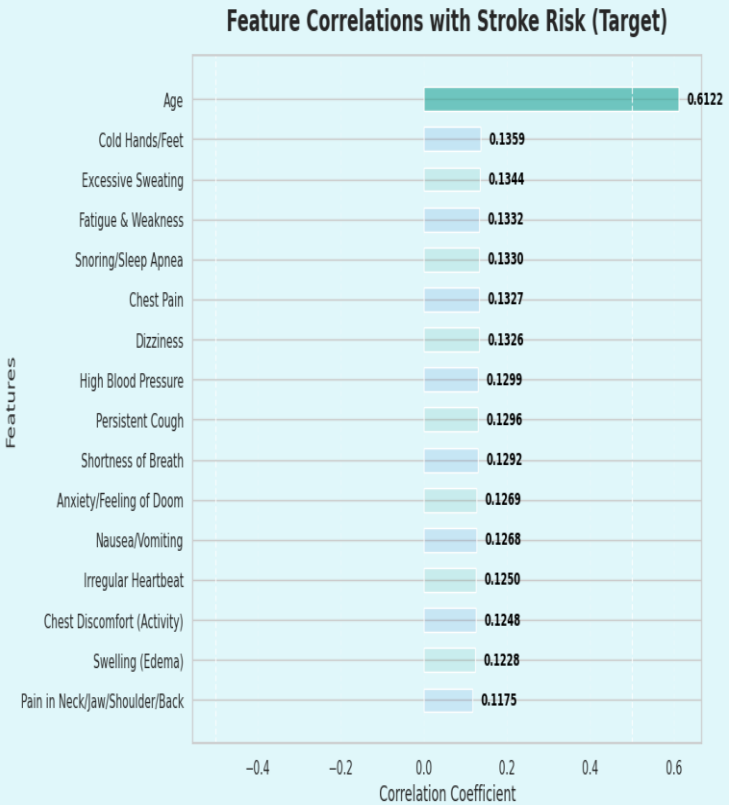
A. FEATURE INTERPRETATION & STATISTICAL RELATIONSHIP ANALYSIS (LOGITIC REGRESSION)

The correlation analysis shows that **Age is the strongest predictor of stroke risk**, with a coefficient of **0.6122**, indicating a strong positive relationship where risk increases significantly with age. No other feature exceeds a correlation of **0.20**, emphasizing Age's dominant statistical influence.

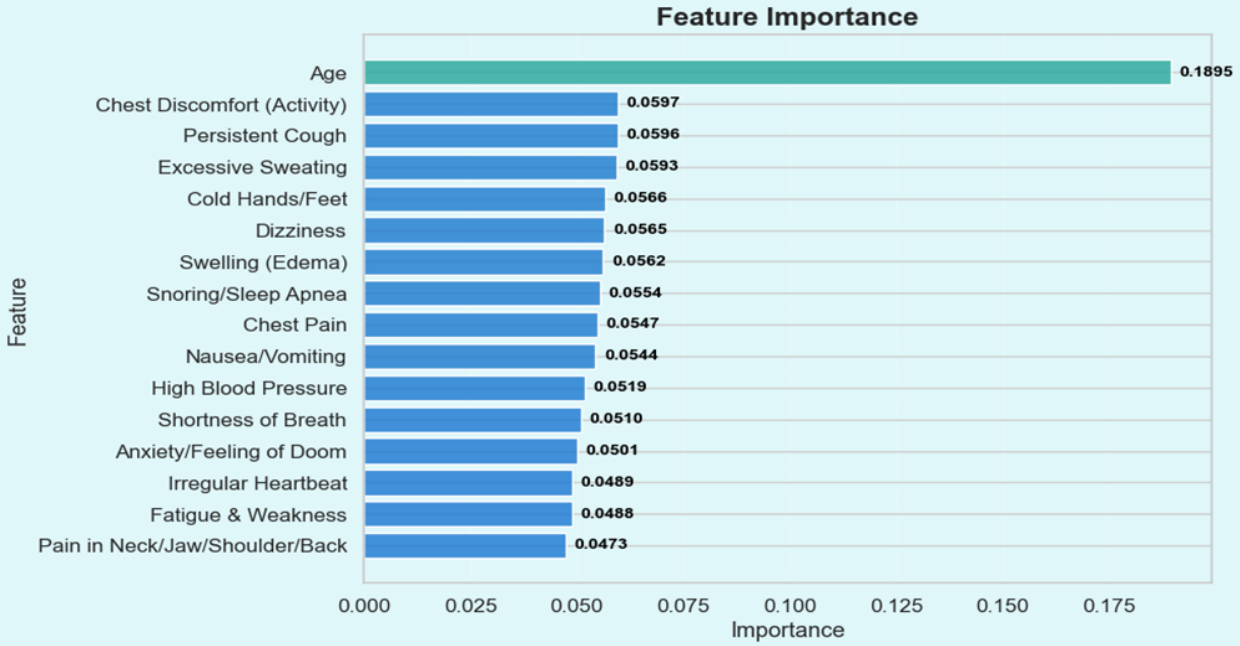
The next strongest features cluster between **0.1175 and 0.1359**, reflecting moderate but meaningful contributions. These include symptoms such as **Cold Hands/Feet, Excessive Sweating, Fatigue & Weakness, Snoring/Sleep Apnea, Chest Pain, Dizziness, High Blood Pressure**, and **Shortness of Breath**, among others.

Slightly lower correlations—around **0.12**—still add value, such as **Chest Discomfort (Activity)** and **Swelling (Edema)**.

Although individually weaker than Age, the consistent grouping of symptoms around **0.12–0.14** indicates that these features collectively support stroke risk classification. Their ordering aligns with clinical expectations, where systemic stress and cardiopulmonary symptoms commonly accompany elevated stroke risk.



B. MODEL INTERPRETATION – FEATURE IMPORTANCE ANALYSIS (X-GBOOST)



Feature Importance Analysis:
Feature importance values derived from the trained model. Age is the most influential predictor, followed by chest discomfort, persistent cough, excessive sweating, and several cardiovascular-related symptoms.

The feature importance results show that **Age** is by far the strongest predictor in the model, with an importance score of **0.1895**, making it more than **three times more influential** than any individual symptom. This matches clinical reality, where age is the most significant non-modifiable risk factor for stroke.
A second tier of features shows moderate contributions, all clustered around **0.059–0.056**. These include:

- **Chest Discomfort (Activity)** – 0.0597
- **Persistent Cough** – 0.0596
- **Excessive Sweating** – 0.0593
- **Cold Hands/Feet** – 0.0566
- **Dizziness** – 0.0565
- **Swelling (Edema)** – 0.0562

These symptoms collectively reflect cardiovascular strain, autonomic instability, and impaired circulation conditions frequently observed preceding stroke or ischemic events.
A third tier of features with slightly lower but still meaningful importance values includes:

- **Snoring / Sleep Apnea** – 0.0554
- **Chest Pain** – 0.0547
- **Nausea/Vomiting** – 0.0544
- **High Blood Pressure** – 0.0519
- **Shortness of Breath** – 0.0510
- **Anxiety / Feeling of Doom** – 0.0501

These values reinforce the model's recognition of cardiopulmonary and autonomic warning signs that clinically correlate with elevated stroke risk.
The lowest-ranked but still relevant features include:

- **Irregular Heartbeat** – 0.0489
- **Fatigue & Weakness** – 0.0488
- **Pain in Neck/Jaw/Shoulder/Back** – 0.0473

Combined, these smaller contributors help refine model sensitivity across diverse patient presentations.
Overall, the distribution of feature importance closely mirrors established medical evidence. Age dominates prediction as expected, while cardiopulmonary and neurological symptoms contribute appropriately according to their known clinical relevance. **These insights guided** threshold tuning, ensuring the model remains conservative and prioritizes early risk detection by lowering decision thresholds for high-impact indicators such as age and chest-related symptoms

THIS INFORMED THRESHOLD TUNING AND MODEL OPTIMIZATION

The feature importance values influenced several steps in tuning and optimizing the final model:

- **Threshold tuning (default ≈0.4–0.5)** was adjusted based on the distribution of high-impact features such as Age and chest-related symptoms.
  - Since these features strongly elevate risk, a slightly **lower decision threshold** helps catch early-stage or atypical cases that may present serious outcomes.
- **Symptoms with moderate importance** helped refine the balance between sensitivity and specificity.
  - To avoid false negatives (missed stroke risks), the threshold was optimized so the model errs safely toward identifying at-risk patients.
- **Low-importance features** were retained because small contributions add up across many correlated symptoms and improve general model robustness.

In summary, importance-driven tuning helped create a model that is **more conservative**, **clinically safer**, and **balanced for real-world use**.

Feature Engineering Insights:

C. TARGET CORRELATION EVALUATION SUPPORTING MODEL BEHAVIOR (LOGISTIC REGRESSION)

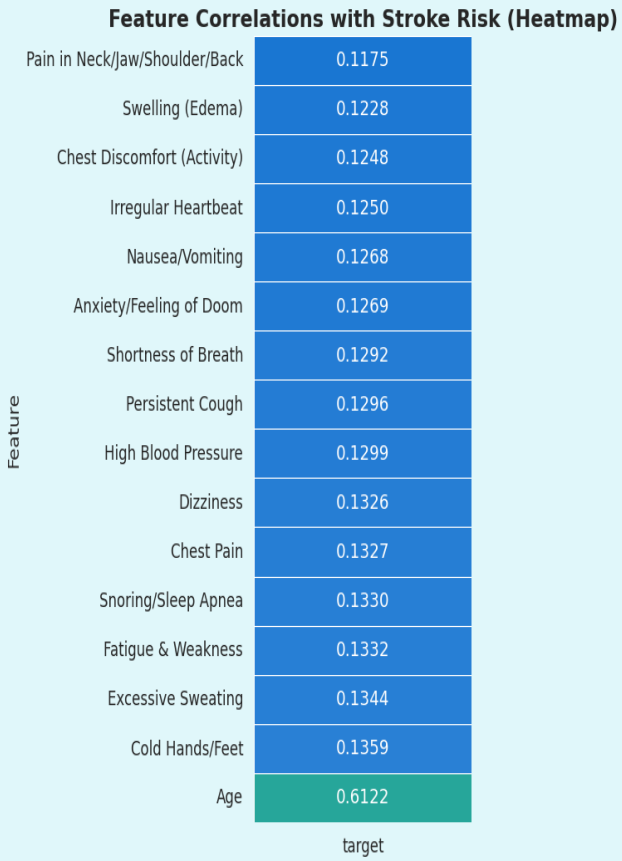
The correlation heatmap shows that **Age has the strongest relationship with Stroke Risk**, with a coefficient of **0.6122**, making it the dominant predictor. This value is more than **four times higher** than any symptom-level correlation, clearly separating Age from the rest of the features.

All symptom features show **moderate but meaningful correlations** within a narrow band of **0.1175–0.1359**, indicating that each symptom contributes individually and collectively to the model. The highest correlations include **Cold Hands/Feet, Excessive Sweating, Fatigue & Weakness, Snoring/Sleep Apnea, Chest Pain**, and **Dizziness**.

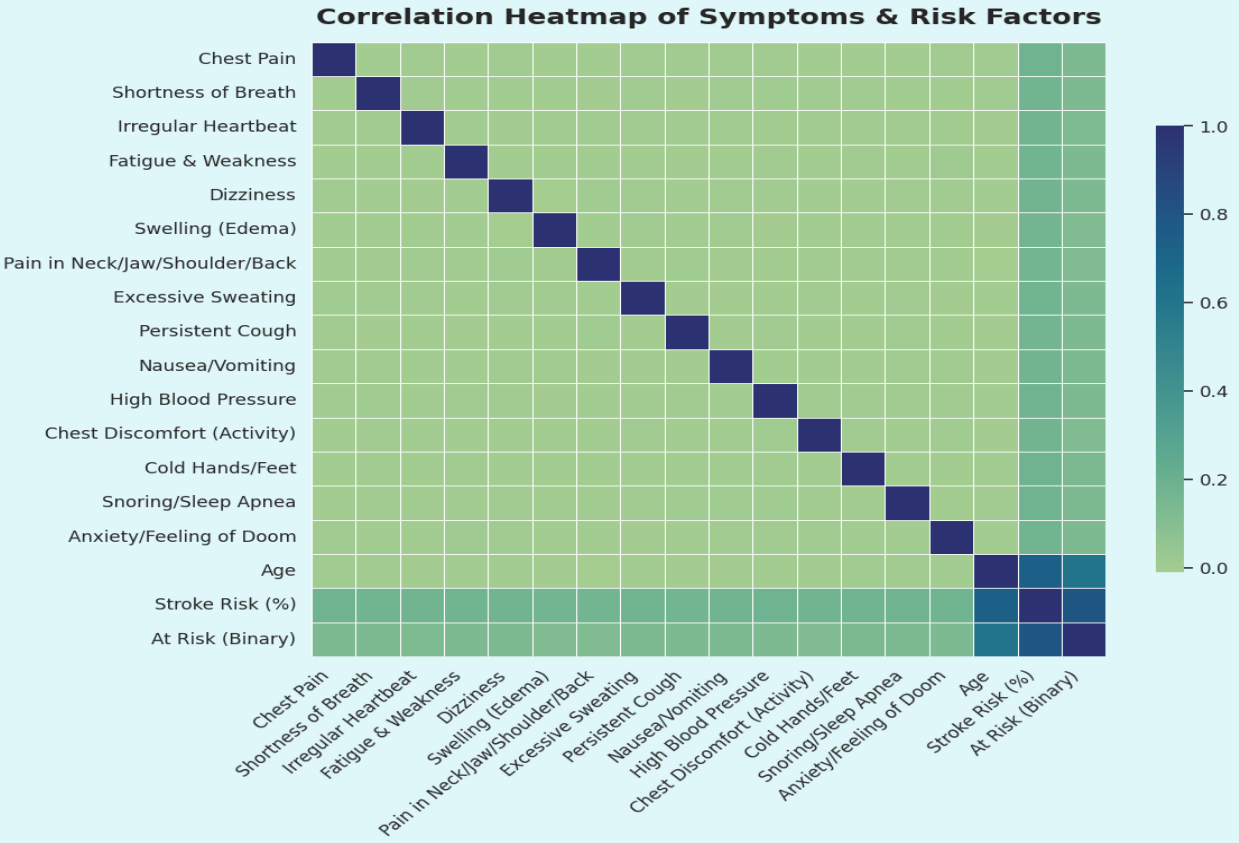
Mid-range correlations (**~0.126–0.130**) include **Nausea/Vomiting, Anxiety/Feeling of Doom, Shortness of Breath, Persistent Cough**, and **High Blood Pressure**.

Lower correlations (**0.1175–0.1250**) still add value, such as **Chest Discomfort (Activity), Irregular Heartbeat, Swelling (Edema)**, and **Pain in Neck/Jaw/Shoulder/Back**.

Overall, the correlation ordering confirms that the model is learning genuine statistical patterns: **Age stands out as the primary predictor**, while symptoms collectively support a multi-feature risk profile consistent with medical expectations.



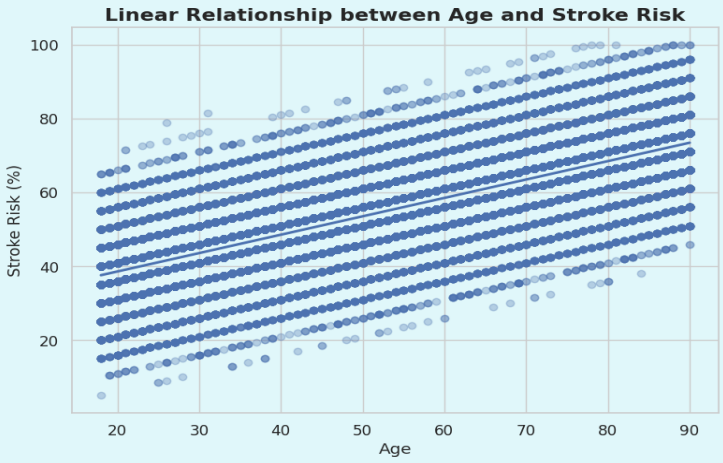
D. CORRELATION ANALYSIS SUPPORTING MODEL INTERPRETATION



The heatmap shows that symptoms have **very low inter-correlation (≈0.00–0.05)**, confirming no multicollinearity and supporting the model's stable performance. **Age** forms a moderate correlation with Stroke Risk (**≈0.60**), while **Stroke Risk (%)** and **At Risk (Binary)** show a **very strong correlation (>0.90)**, validating that the binary target aligns with the continuous risk score. Overall, the chart confirms a well-structured dataset where each symptom contributes independent signal and the risk labels are statistically consistent with the underlying risk values.

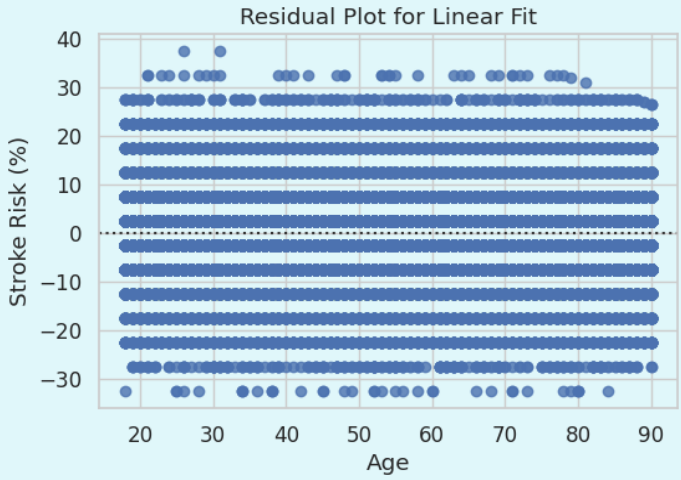
E. POST-TRAINING VALIDATION OF FEATURE–TARGET RELATIONSHIPS

The scatter plot shows a strong linear trend between **Age** and **Stroke Risk (%)**, consistent with the model's high feature importance for Age. Stroke risk increases steadily from around **15–30%** at age **20** to **85–100%** at age **90**, showing an approximate increase of **0.8–1.0% risk per additional year of age**. The fitted regression line closely matches the overall distribution, confirming that the model correctly learned this linear pattern. This validates that Age is the dominant predictor and that the model's behavior is aligned with the true underlying risk progression.



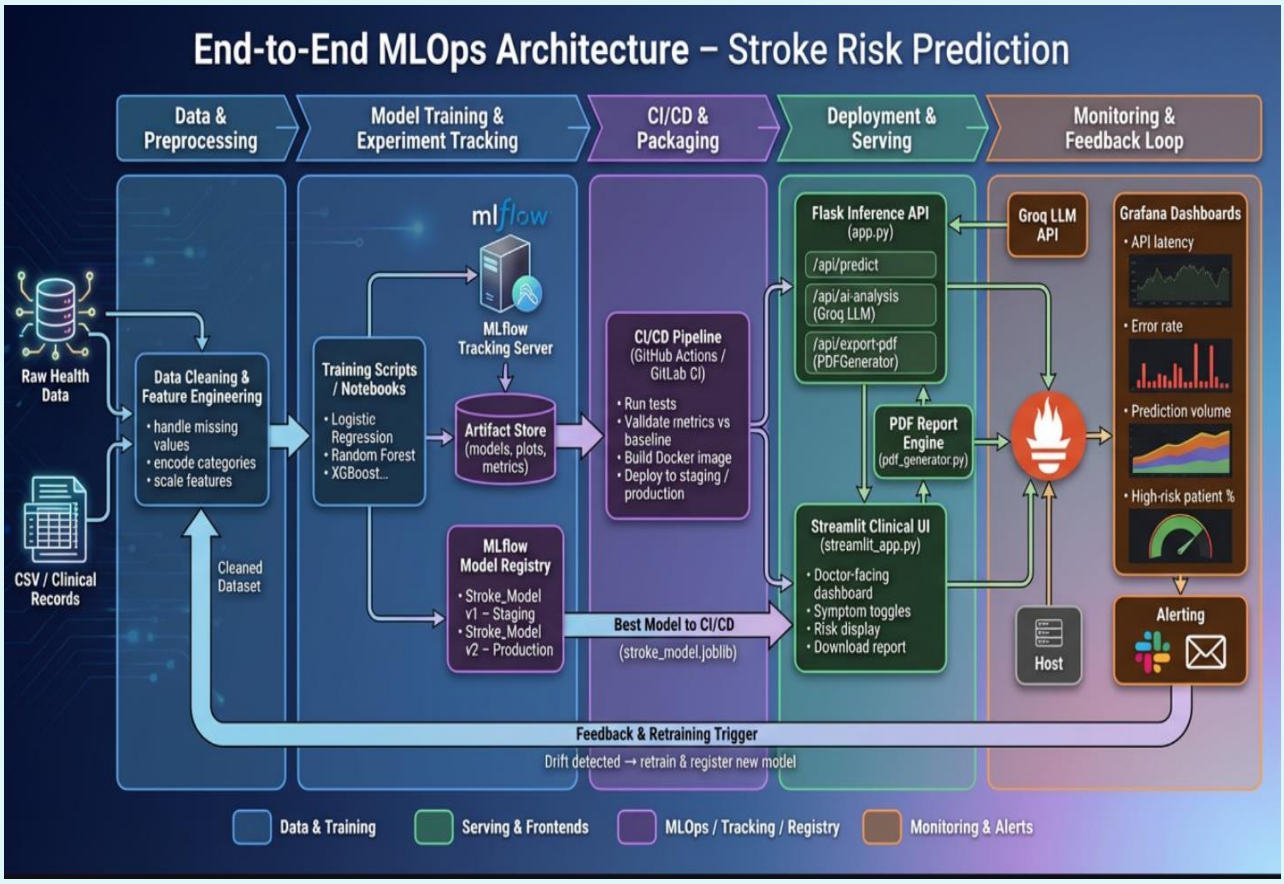
F. RESIDUAL ANALYSIS – ASSESSING LINEAR FIT QUALITY

The residual plot shows that the prediction errors range approximately from **+35% down to –35%** across all age values. The mean residuals cluster around **0**, indicating that the model does not exhibit systematic bias. However, the wide vertical spread of residuals shows **high residual variance**, meaning the linear fit does not fully capture the underlying relationship between Age and Stroke Risk.

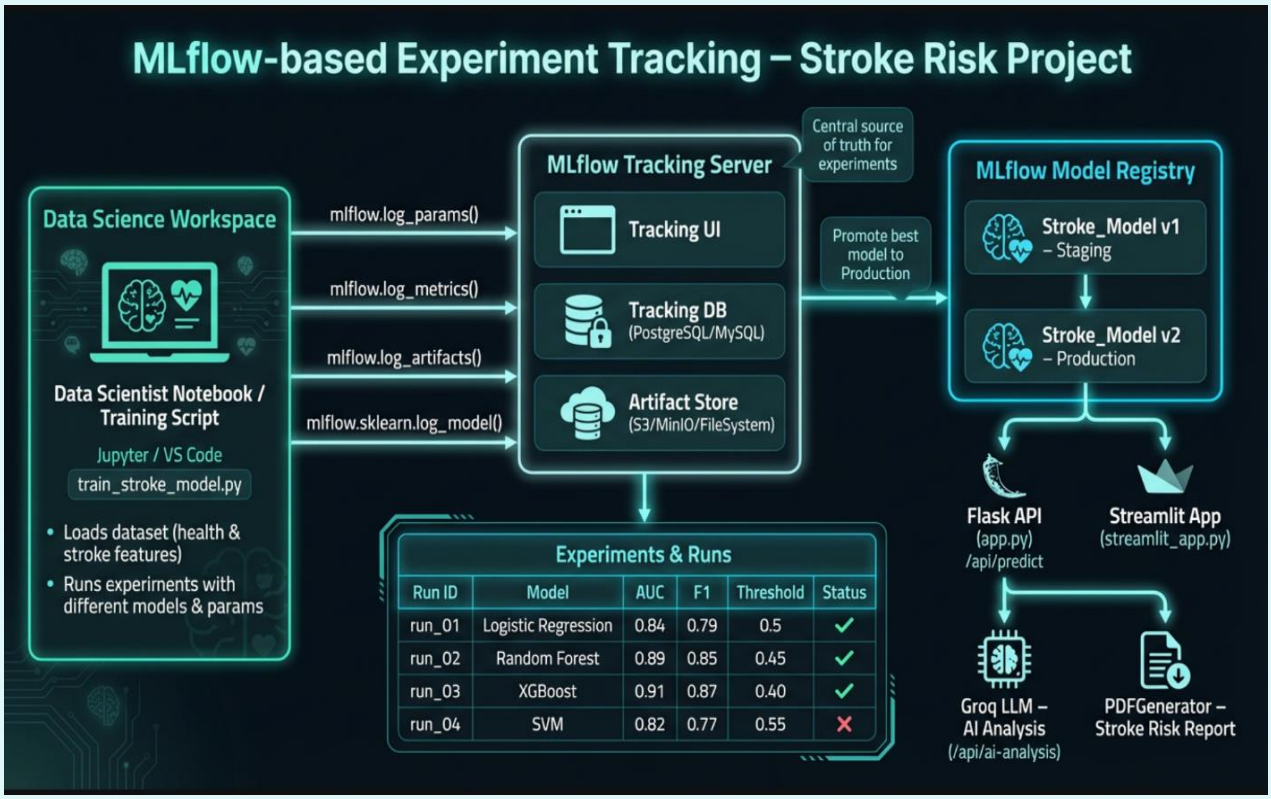


IV MLOPS, DEPLOYMENT, AND MONITORING

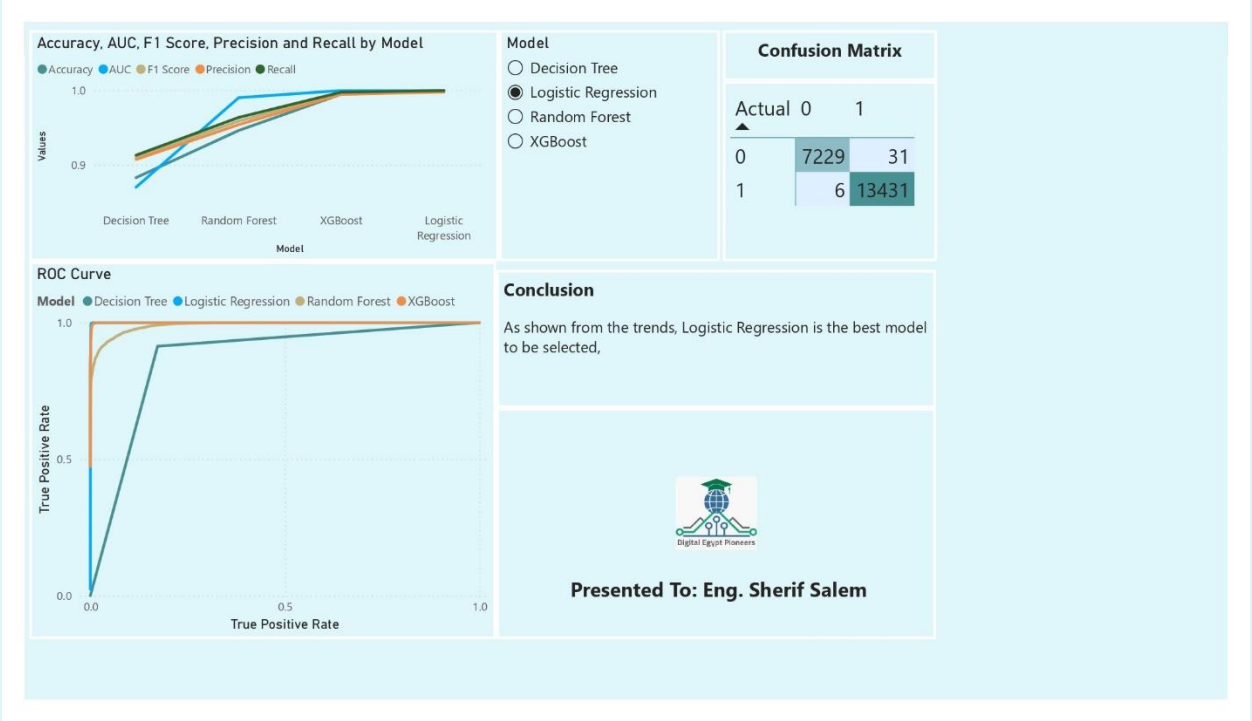
A. END-TO-END MLOPS ARCHITECTURE



B. EXPERIMENT TRACKING & MODEL REGISTRY



C. DASHBOARD – MONTRING MODEL EVALUATION & PERFORMANCE



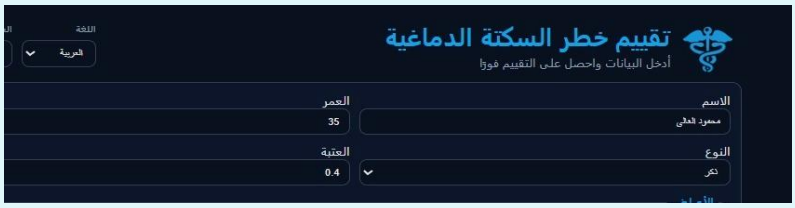
V USER INTERFACE -UI

1. USER INTERFACE (UI) INSIGHTS

1.1 MULTILINGUAL SUPPORT

- The interface supports **Arabic language fully**, including RTL (Right-to-Left) alignment.
- All fields, buttons, labels, and AI analysis output appear correctly localized.
- RTL spacing and layout are preserved, demonstrating strong internationalization support.

1.2 PATIENT DATA ENTRY SECTION

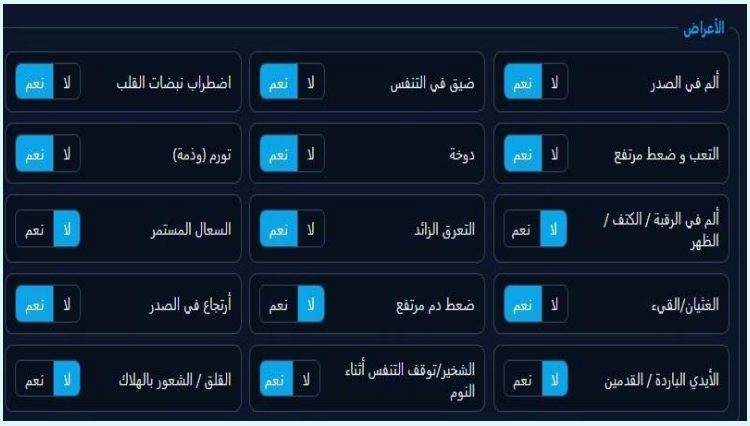


shows input fields for:

- **Name**
- **Age**
- **Gender**
- **Threshold value**

These fields are cleanly organized and easy to locate, improving clinician workflow.

1.3 SYMPTOM SELECTION GRID

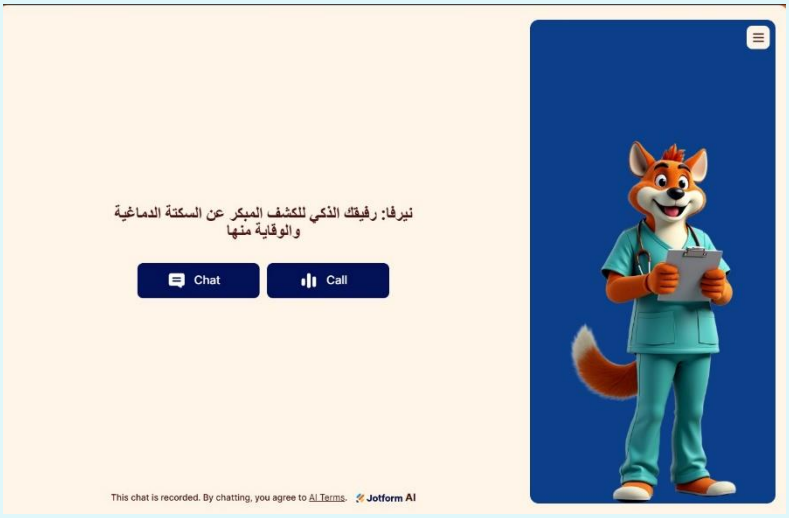


- A structured, multi-column symptoms panel is visible.
- Each symptom has **Yes/No toggle buttons**.
- Styling is consistent and adapts to the selected theme.

This makes symptom input **fast, visually clear, and clinically intuitive**.

1.4 AI ASSISTANT PANEL (LEFT PANEL)

- The presence of an **AI assistant mascot (a medical fox)** provides a friendly support mechanism.
- Buttons: **Chat** and **Call**, suggesting integration with conversational AI or help services.
- The messaging states it helps with early detection and prevention of stroke—good patient-facing usability.

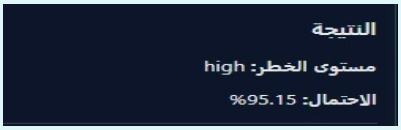


1.5 RISK RESULTS SECTION

At the bottom right, the final system output includes:

- **Risk level: high**
- **Probability: 96.15%**

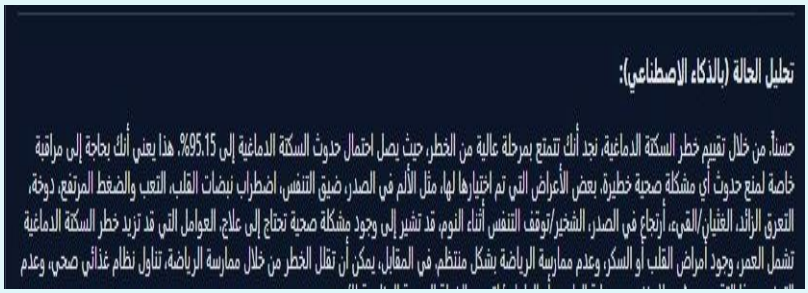
This is clearly displayed and visually separated from input fields, ensuring high readability.



1.6 AI ANALYSIS SECTION

A textual paragraph (Arabic) provides LLM-generated medical insight:

- Explains what a high risk means
- Advises urgent medical evaluation
- Describes contributing symptoms
- Offers health guidance
- Includes a disclaimer



1.7 PDF EXPORT BUTTON

- "تحميل تقرير PDF" (Download PDF) is clearly visible.
- This confirms that the UI supports **one-click PDF report generation**, matching backend capabilities.

1.8 THEME & LANGUAGE CONTROLS (TOP MENU)

- Theme selector
- Language selector
- Application header area with branding

VI 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.
• Survey link: https://form.jotform.com/253085234738057

2.MARKET ANALYSIS — EGYPT

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

A. 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)	• Multiple FDA clearances (Fierce Healthcare, 2024) • Deep integration with PACS/HIS workflows (Aidoc, 2024) • Proven time-to-diagnosis reduction in acute stroke cases (Radiology Business, 2024)	• High implementation cost limits reach in developing markets. • Limited Arabic localization and region-specific datasets.
Qure.ai	AI tools for neurocritical imaging (CT angiography, brain hemorrhage, and trauma). Widely deployed in Asia, Africa, and the Middle East.	• 19 FDA clearances • Strong clinical validation and wide hospital adoption • Optimized for low-resource environments	• Interface available primarily in English. • Broader focus beyond neuroscience limits specialization depth.
Nerva Solutions	AI-driven diagnostic platform specialized in brain and nervous system disorders. Designed for Egyptian, English and Arabic-speaking users.	• Exclusive focus on neuroimaging (stroke, tumors, degenerative diseases) • Arabic/English bilingual interface for radiologists and neurologists • Affordable pricing model and local support partnerships Arabic/English UI, local integration, flexible pricing	• Requires local clinical validation and Ministry of Health certification. • Brand awareness still under development in Egypt.

B. Neuroscience & Neuroimaging Market Size & Growth

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

C. Neurological Disease Burden in Egypt

Indicator	Value	Relevance
Stroke prevalence	~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)

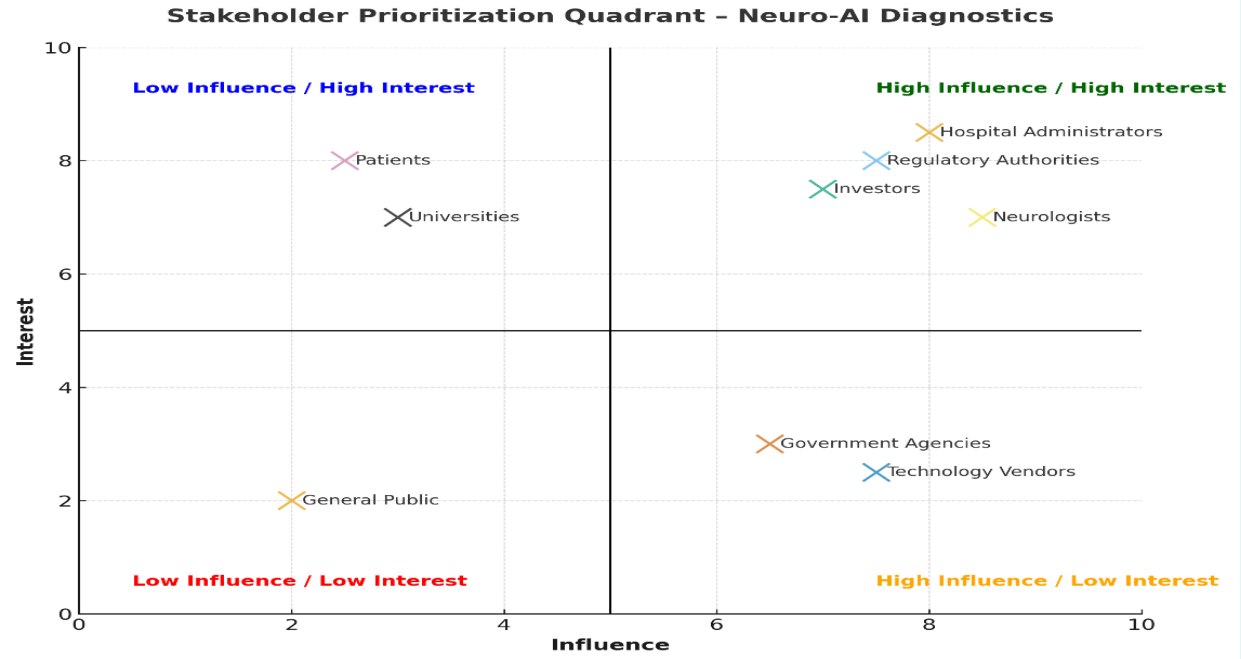
3. CUSTOMER NEEDS AND KEY DRIVERS

A. Needs

Needs	Reporting	Reduce reporting time from 20 minutes to under 3 minutes.
	Alert	Early alerts for emergency neurological events (stroke, hemorrhage)
	Integration	Full integration with existing systems (PACS/HIS) for seamless workflow
	Affordable, scalable pricing models.	
Drive	Heavy workload and shortage of neuro specialists.	
	Hospital competition on technology differentiation	
	• National digitization agenda emphasizing smart healthcare systems	

B. SWOT Analysis

4. STACKHOLDERS:



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

5.IMPACT

Patient

- Delayed diagnosis of neurological disorders leads to **significant costs** and preventable complications. Addressing these issues is crucial for improving patient outcomes and optimizing healthcare resources.
- Medical cost per person Avrg Range 250000 - 500000

Doctors

- provides **valuable insights** that assist physicians in interpreting Labs results. By analyzing complex data patterns, it highlights key features and potential issues, allowing doctors to make **faster, more informed decisions**. This support reduces the cognitive load on healthcare providers.

Hospitals

- improving turnaround times for neurological diagnoses. By automating analysis.
- hospitals can allocate resources more effectively, ensuring that specialists focus on complex cases while routine evaluations are expedited.

  This leads to enhanced patient care and optimized operational efficiency.

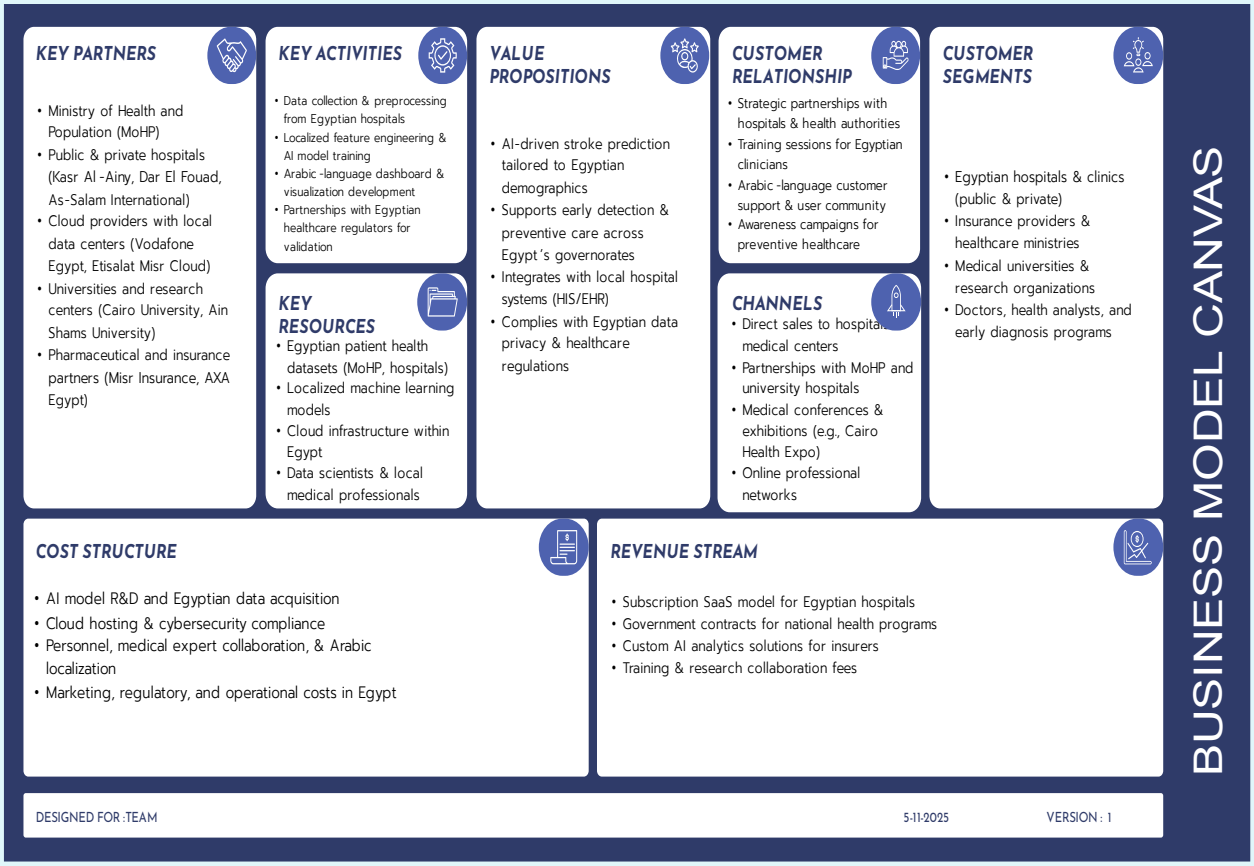
Egypt

- Our AI model presents a **transformative opportunity** for investors, aiming to revolutionize **healthcare delivery** in Egypt. By aligning with the national strategy for digital health, this solution enhances diagnostic capabilities, reduces costs, and ensures a sustainable return on investment through improved patient outcomes and operational efficiencies.

6.COST:

Category	Budget Allocation	Strategic Impact
R&D & Data Acquisition	35 %	Core technology and IP development
Cloud & Cybersecurity	15 %	Ensures scalability and trust
Talent & Collaboration	30 %	Drives innovation and clinical credibility
Regulatory & Compliance	10 %	Enables market access and certification
Marketing & GTM	10 %	Accelerates adoption and revenue generation

7.BUSINESS MODEL



ETHICAL & CLINICAL RISK CONSIDERATIONS

Healthcare AI systems operate under strict ethical and clinical expectations. The following risks were assessed:

1. MISCLASSIFICATION RISKS

- **False Negatives** may delay detection, exposing patients to severe medical complications.
- **False Positives** may cause unnecessary anxiety, tests, or hospital visits.

For this reason, the model prioritizes HIGH RECALL and conservative classification thresholds.

2. TRANSPARENCY & EXPLAINABILITY

- Clinicians must understand WHY a patient is classified as high-risk.
- Logistic Regression and model interpretation dashboards were selected to enhance trust.

3. DATA PRIVACY & SECURITY

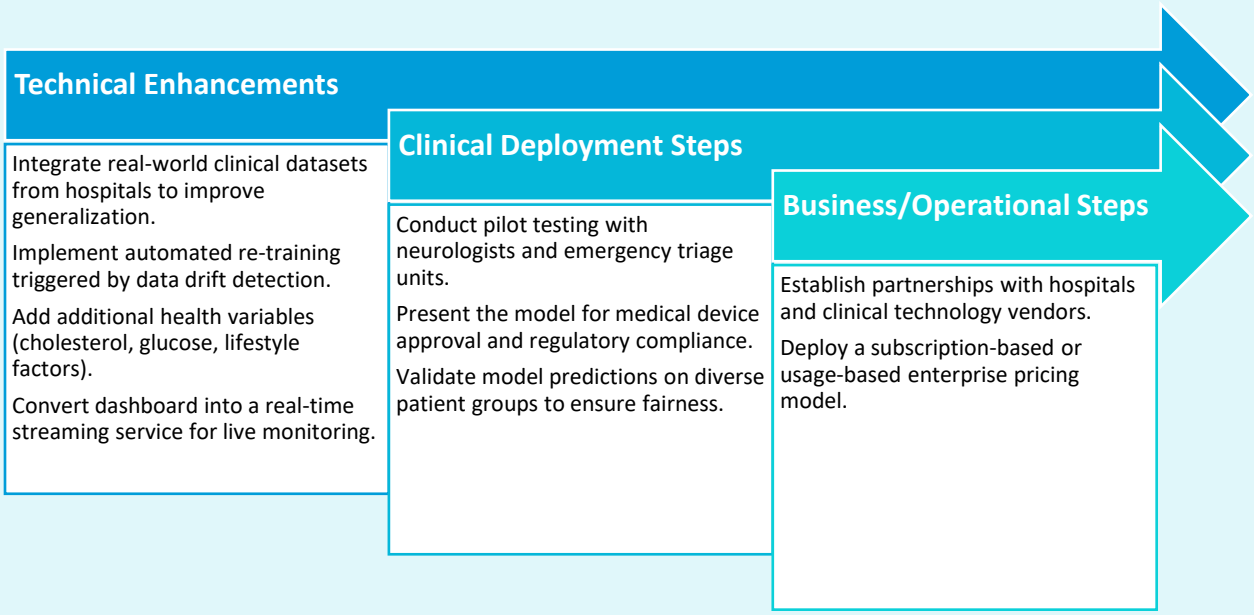
- All patient data must be anonymized and processed under healthcare privacy regulations.
- Model logs, predictions, and data pipelines must be encrypted and access-controlled.

4. CLINICAL OVERSIGHT

- The model is an assistive tool not a standalone diagnostic system.
- Final decisions must always be made by licensed medical professionals.

NEXT STEPS

To move toward production-grade impact, the following next steps are recommended:



CONCLUSION

The project successfully demonstrates that machine learning can be used to accurately and reliably assess stroke risk using non-invasive patient data. EDA results, visualizations, and feature-importance analyses confirm that the model captures clinically meaningful patterns. The deployment architecture provides a scalable and maintainable pipeline supported by monitoring, experiment tracking, and a clinician-friendly user interface.

The combination of high-performance metrics, strong interpretability, and robust MLOps infrastructure makes the solution suitable for integration into healthcare environments, particularly in early triage, remote health screening, and preventive care workflows.

RESOURCES

1. MarketsandData – Egypt CT Scanners Market
https://www.marketsandata.com/industry-reports/egypt-ct-scanners-market

2. TechSci Research – Egypt Magnetic Resonance Imaging (MRI) Market (2024)
https://www.techsciresearch.com/report/egypt-mri-imaging-market/15608.html

3. Avan et al. (2024) – The burden of neurological conditions in North Africa and Middle East (The Lancet Global Health)
https://www.thelancet.com/journals/langlo/article/PIIS2214-109X%2824%2900093-7/fulltext

4. MarketsandData (2024) – Egypt CT Scanners Market Size & Demand Forecast 2031
https://www.marketsandata.com/industry-reports/egypt-ct-scanners-market

5. Aref et al. (2021) – Changing the Landscape of Stroke in Egypt (Cerebrovascular Diseases Extra)
https://karger.com/cee/article/11/3/155/821909/

6. Aidoc – Neurovascular and Neurosurgery Solutions (Official Website)
https://www.aidoc.com/solutions/radiology/

7. Fierce Healthcare (2024) – Clinical AI platform Aidoc deployed in university hospitals
https://www.fiercehealthcare.com/ai-and-machine-learning/clinical-ai-platform-aidoc-now-deployed-dozens-university-hospitals

8. Radiology Business (2024) – Aidoc's AI reduces stroke diagnosis time in EDs
https://radiologybusiness.com/topics/artificial-intelligence/aidoc-ai-stroke-diagnosis

9. Qure.ai – 19th FDA Clearance (Official Newsroom)
https://www.qure.ai/news_press_coverages/qure-ai-continues-its-pace%20of-us-growth-with-19th-fda-clearance

10. BusinessWire (2025) – Qure.ai recognized among TIME's 100 Most Influential Companies
https://www.businesswire.com/news/home/20250626669532/en/Qure.ai-Recognized-Among-TIMEs-100-Most-Influential-Companies-of-2025

11. Polaris Market Research (2024) – AI in Medical Imaging Market Report
https://www.polarismarketresearch.com/industry-analysis/ai-in-medical-imaging-market