1. Dataset Overview

The dataset is centered around the prediction and classification of stroke risk based on both demographic and clinical symptoms.

Key Components:

- **Demographics**: Age
- **Symptoms**: Binary (0 or 1) indicators for clinical symptoms such as:
    - 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
- **Target Variables:**
    - Stroke Risk (%) – a continuous variable indicating the risk percentage
    - At Risk (Binary) – derived by thresholding Stroke Risk (%) > 50%

Data Stats:

- Number of records: Varies (depending on source)

- Number of features: 20+ (including engineered features)
- Range of Age: Typically from 18 to 90+
- Stroke Risk (%): Raw values normalized to [0–100] range

---

2. Data Cleaning and Preprocessing

- Dropped duplicate records
- Inspected each feature for unique values
- Checked and handled missing values by imputation or row exclusion
- Identified and visualized outliers via boxplots and histograms

---

3. Anomaly Detection

- **Method**: Local Outlier Factor (LOF)
- **Features Evaluated**: Age, Stroke Risk (%), and all binary symptoms
- **Result**: Outliers identified and removed from modeling dataset

---

4. Exploratory Data Analysis (EDA)

- **Distribution Analysis**:
    - Histograms and KDE plots for Age, Stroke Risk (%)
- **Relationships**:
    - Scatterplots: Age vs Stroke Risk (%), colored by At Risk
    - Boxplots comparing Stroke Risk across symptoms
    - Pairplots and linear regression lines for visual correlation

---

5. Symptom Risk Analysis

- **Stacked Bar Plots**: Showing distribution of Stroke Risk grouped by symptom status

- **Chi-square Tests**: To assess dependence between symptoms and binary risk classification
- **Correlation Heatmap**: To visualize how symptoms and age relate to Stroke Risk (%)

---

6. Feature Engineering

- **Symptom Severity Score**: Sum of all 15 binary symptom indicators
- **Age Grouping**: Categorized into bins:
    - 0–30
    - 31–50
    - 51–70
    - 70+
- **One-Hot Encoding**: Applied to Age Group (with drop-first for modeling)

---

7. Outlier Correction in Stroke Risk

- Converted Stroke Risk (%) values above 100 to proper percentages
- Re-evaluated outliers and removed invalid entries

---

8. Data Balancing

- **Observed Problem**: Imbalance in At Risk classes
- **Technique Applied**:
    - Downsampling of majority class in binary classification tasks
    - Oversampling using SMOTE for other classifiers (e.g., neural network)
- **Result**: More balanced class distribution for training

---

9. Model Training and Evaluation

a. Random Forest Regressor

- Predicts Stroke Risk (%)
- Evaluated using MSE and R-squared
- Performed GridSearchCV for hyperparameter tuning (n_estimators, max_depth, min_samples_split)

b. XGBoost Classifier

- Transformed risk into classification
- Encoded labels, trained with XGBClassifier

c. Logistic Regression

- Binary classification based on Stroke Risk (%) > 50%
- Used saga solver with 500 iterations

d. Random Forest Classifier

- Used on thresholded binary risk target
- Evaluated using accuracy, classification report, ROC curve, and confusion matrix

e. MLP Neural Network

- Deep learning classifier with two hidden layers (100, 50)
- Evaluated using accuracy, classification report, and confusion matrix

---

10. Metrics & Visualization

- **Confusion Matrix & Classification Report**: Used for all classifiers
- **ROC Curve**: Evaluated model separability
- **Error Metrics (Regressors)**:
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - Mean Absolute Error (MAE)
    - R-squared
- **Predicted vs Actual Scatter Plot**: Regression diagnostic

---

11. Model Saving

- Models saved using joblib for deployment
    - stroke_prediction_model.pkl (regressor)
    - stroke_prediction_classifier.pkl (random forest classifier)
    - mlp_model.pkl (neural network)

---

12. Web API for Stroke Prediction

- **Framework**: Flask
- **Endpoints**:
    - / – Welcome page
    - /predict – Accepts JSON input, preprocesses data, returns prediction and probabilities
- **Model Loading**: Handles exceptions if model file is missing
- **Input Preprocessing**:
    - Ensures presence of required fields like Age
    - Converts symptoms to binary
    - Computes Symptom Severity Score
    - Bins Age and applies one-hot encoding
    - Ensures all features match model expectations
- **Returned Output**:
    - Prediction (0 or 1)
    - Interpretation (At Risk or Not)
    - Probabilities for both classes

---

13. Data Types and Descriptions

- **Age:**
    - Type: Numeric (Continuous)
    - Description: Represents the age of the individual in years. This is a crucial feature as age is a primary factor in stroke risk. The Age feature was used to categorize individuals into age groups.
- **Age Group:**
    - Type: Categorical (Ordinal)
    - Description: Categorical feature representing age groups derived from the "Age" column. The groups are:
        - 0–30
        - 31–50
        - 51–70
        - 70+
    - One-Hot Encoded: This column was transformed using one-hot encoding for compatibility with machine learning models.
- **Stroke Risk (%):**
    - Type: Numeric (Continuous)
    - Description: The percentage representing the risk of a stroke based on medical factors. This feature was used as the target variable in some predictive models.
- **At Risk (Binary):**
    - Type: Categorical (Binary)
    - Description: Indicates whether the individual is at risk of stroke or not, represented as 1 (at risk) and 0 (not at risk).
- **Symptoms:**
    - Type: Categorical (Binary for each symptom)
    - Description: A set of binary features representing whether an individual exhibits a certain symptom, such as chest pain, shortness of breath, irregular heartbeat, fatigue, dizziness, and more.

---

14. Descriptive Statistics of Key Columns

- **Age:**
    - Mean: Average age of individuals in the dataset.
    - Median: Middle value of age, used to check for skewness.
    - Standard Deviation: Variability in age.
- **Stroke Risk (%):**
    - Mean: The average stroke risk percentage across all individuals.
    - Min/Max: The minimum and maximum values of stroke risk percentage.
    - Skewness: To check if there are more high or low values in the risk percentages.
- **Symptoms:**
    - Proportion of 1s (presence): Percentage of individuals exhibiting a given symptom.
    - Proportion of 0s (absence): Percentage of individuals not exhibiting a symptom.

---

15. Correlation Between Variables

- The following features showed significant correlation:
    - **Age and Stroke Risk (%):** A positive correlation suggesting that older individuals tend to have higher stroke risk.
    - **Symptoms like Chest Pain and Fatigue:** Showed significant correlation with the target variable "At Risk (Binary)", indicating their relevance in predicting stroke risk.

---

16. Missing Data Analysis

- **Missing Data:** A few columns, such as some of the symptoms, had missing entries that were handled by either imputation or removal.

- **Imputation Strategy:**
    - For missing symptoms, the values were imputed with 0 (absence of symptoms).
    - For numerical columns like "Age" or "Stroke Risk (%)", missing values were handled by forward fill or mean imputation where necessary.

---

17. Data Distribution and Visualization

- **Histograms:** Visualized distributions for numerical variables like Age and Stroke Risk (%).
- **Boxplots:** Used to identify and remove outliers, especially in the "Stroke Risk (%)" column.
- **Scatter Plots:** Illustrated relationships between key variables like Age vs Stroke Risk (%).
- **Correlation Heatmap:** Showed the linear relationships between different features.

---

18. Further Insights and Key Findings

- **Imbalance in "At Risk (Binary)" classes:** The dataset initially had a class imbalance with more individuals being classified as "at risk" of stroke.
- **Solution:** Downsampling of the majority class to balance the dataset.
- **Age as a Primary Factor:** Age showed a strong correlation with stroke risk, especially for individuals above 50 years old.
- **Symptom Severity:** The symptom "Chest Pain" was highly correlated with being at risk of stroke, while symptoms like "Swelling (Edema)" and "Snoring/Sleep Apnea" had less direct correlation.

---

19. Further Exploration Opportunities

- **Feature Importance:** Using model-based feature importance, we can understand which symptoms or demographic factors contribute most to stroke risk predictions.
