Possibility of Stroke

The Team Behind the Project

Sara Ali Mahmoud Ibrahim

Mariam Gamal Askr

Basmala Ahmed

Engy Ahmed

Problem Statement

Stroke is a leading cause of death and long-term disability globally. Early identification of individuals at high risk is crucial for timely medical intervention. However, predicting stroke risk remains a complex challenge due to the multifactorial nature of the condition.

Goal: To develop a machine learning model that predicts the likelihood of a person having a stroke based on health-related attributes, enabling early diagnosis and preventive care.

Data Overview

Dataset Source

- Kaggle- Stroke Risk Prediction Dataset

🧬 Dataset Summary

- Total Records: ~7,000
- Total Features: 18
  - 15 Symptom-based binary predictors
  - 1 Demographic feature
  - 2 Target variables

Data Overview - Feature Categories

1 Symptom Features (Binary Predictors)

Each symptom is marked as 1 (present) or 0 (absent):

- 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

2 Demographic Feature

- Age — a continuous numerical variable; stroke risk increases with age

3 Target Variables

- At Risk (Binary): 1 = At risk of stroke, 0 = Not at risk
- Stroke Risk (%): A continuous value representing probability of stroke (0–100%)

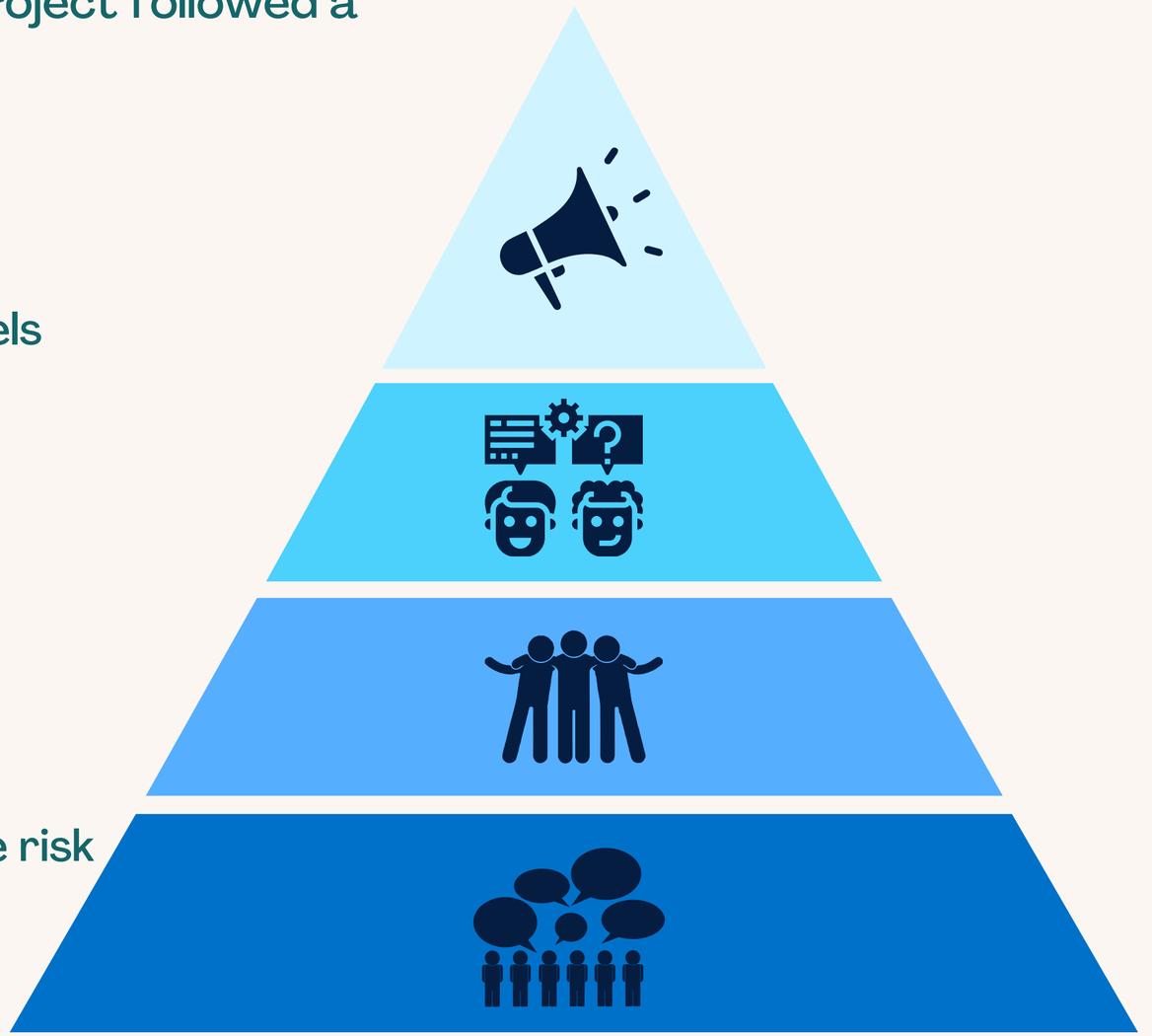
Project Process

To develop a reliable stroke risk prediction model, the project followed a structured data science workflow:

1. Data Collection
   - Obtained from Kaggle: Stroke Risk Prediction Dataset
   - Contains symptom indicators, age, and stroke risk labels

2. Data Cleaning & Preprocessing
   - Handled duplicates
   - Verified data types and consistency
   - Prepared categorical and numerical data for modelin

3. Exploratory Data Analysis (EDA)
   - Visualized relationships between symptoms and stroke risk
   - Identified patterns, trends, and outliers
   - Assessed class imbalance and feature distributions



Project Process - Cont'd

4. Feature Selection
   - Evaluated feature importance using correlation and model-based methods
   - Selected the most predictive features to improve performance

5. Model Building
   - Trained various machine learning models
   - Tuned hyperparameters for better accuracy

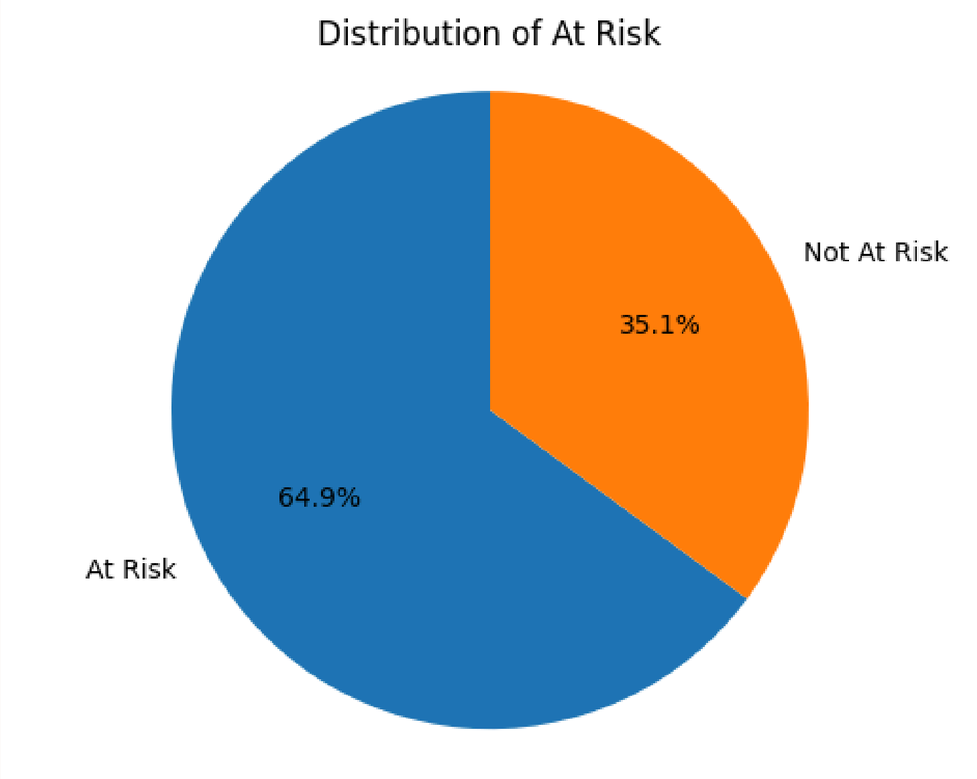
6. Model Evaluation
   - Assessed models using metrics like Accuracy, Precision, Recall, F1-Score, and AUC
   - Selected the best-performing model for final predictions

7. Interpretation & Deployment Considerations
   - Interpreted model outputs to understand key stroke indicators
   - Considered real-world use and clinical implications

Exploratory Data Analysis (EDA)

Distribution of Stroke Risk

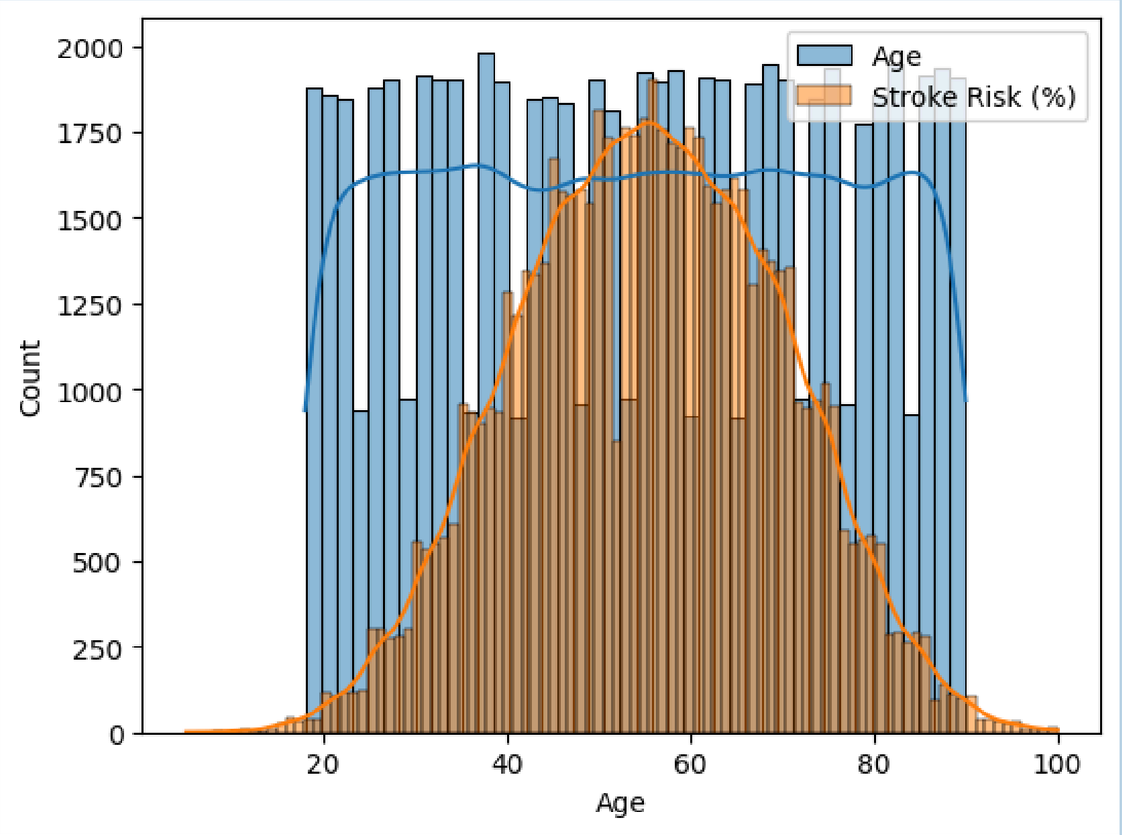
- The pie chart illustrates the distribution of individuals categorized as "At Risk" vs. "Not At Risk" of stroke.
- 64.9% of the individuals are labeled as "At Risk", while 35.1% are "Not At Risk".
- This relatively imbalanced dataset indicates a potential bias that needs to be addressed during model training.
- Techniques like resampling or class weighting are necessary to prevent the model from favoring the majority class.



EDA-Cont'd

Age vs Stroke Risk Distribution

- The histogram visualizes the distribution of individuals' ages (blue bars) and their associated stroke risk percentages (orange bars).
- Most individuals in the dataset are aged between 20 and 80, with a slight concentration around middle and older ages.
- The Stroke Risk (%) (orange curve) shows a bell-shaped pattern, peaking around ages 50–70, indicating that stroke risk tends to increase with age and is highest in older adults.
- This indicates that age is a significant factor in predicting stroke risk and should be treated as a key input feature in the model.



EDA-Cont'd

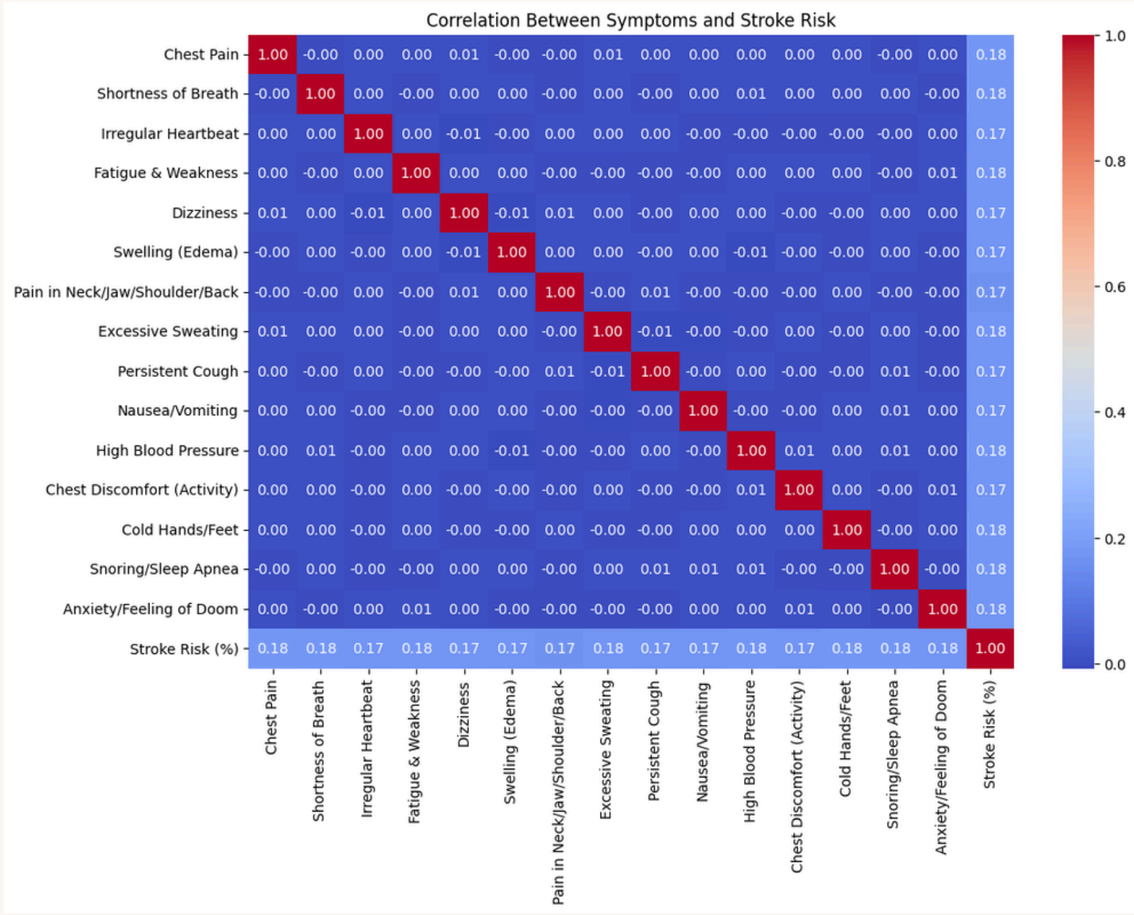
Statistical Significance of Symptoms

- Chi-Squared Tests: All symptoms have p-value ≈ 0
- Each symptom is significantly associated with stroke risk
- Supports the idea that every symptom adds value

- Most symptoms have a low positive correlation (~0.17–0.18) with stroke risk.

- This suggests each symptom individually provides a small but measurable signal.

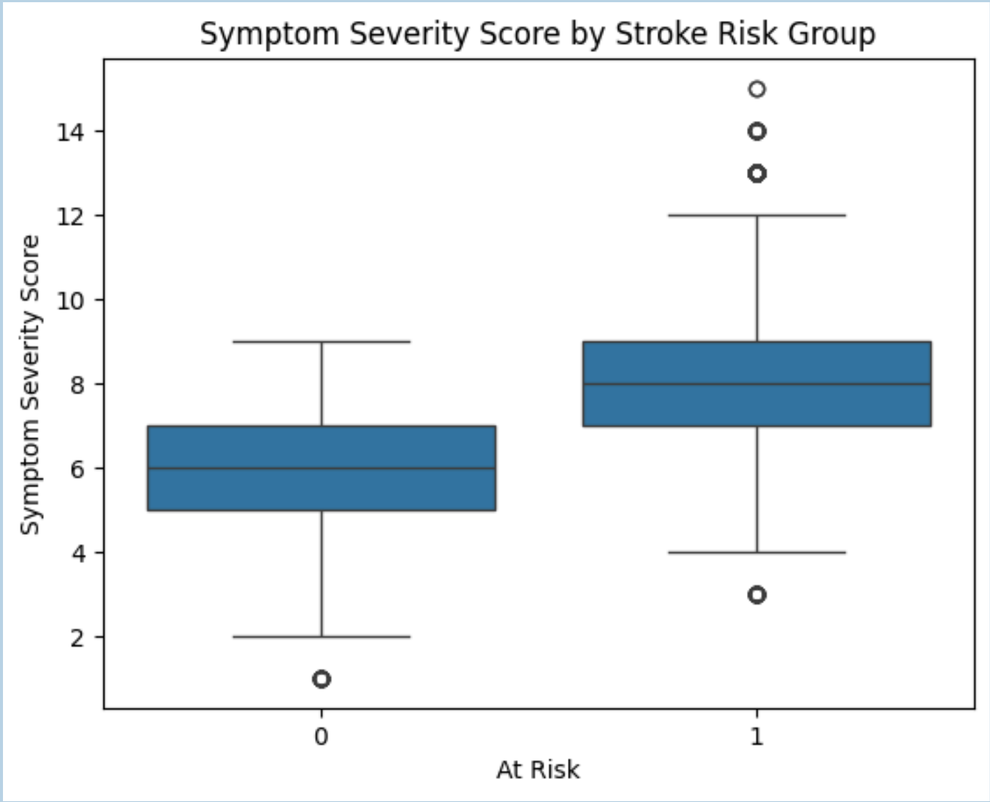
- Symptom-to-symptom correlations are mostly near zero, indicating low multicollinearity — symptoms are not redundant.



EDA-Cont'd

Higher Symptom Severity in At-Risk Individuals

- This plot shows total symptom count per individual (i.e., how many symptoms are present).
- Individuals at risk of stroke tend to have higher symptom severity scores.
- Median score for Not At Risk ≈ 6, Median for At Risk ≈ 8–9
- Indicates that stroke risk increases with the number of symptoms present, not just any single symptom.



Machine Learning Overview

Goal: Predict stroke risk based on patient features using supervised learning techniques.

Approach: Multiple models were trained and evaluated to identify the most effective approach, targeting both regression (exact risk %) and classification (high-risk vs low-risk).

Model Building

Models Used:

- Logistic Regression for binary classification (At Risk / Not At Risk)
- Random Forest Regressor for estimating exact stroke risk (%)
- MLP Neural Network to boost classification performance

⚙️ Preprocessing Steps:

- Removed duplicates
- Normalized risk scores & handled outliers
- Engineered features: Symptom Severity Score & Age Group
- Applied one-hot encoding for categorical variables
- Used data balancing techniques to address class imbalance

Regression Model–Random Forest

Model: Random Forest Regressor

Steps:

- Features: All except Stroke Risk (%)
- Target: Stroke Risk (%) (continuous values)
- Train-test split: 80/20
- Evaluation Metrics: MSE, R^2

Results:

- Initial model:
  - Mean Squared Error: 0.00
  - R-squared: 1.00 (Potential overfitting)
- After tuning with GridSearchCV:
  - Best Parameters: n_estimators=200, max_depth=30, min_samples_split=2
  - Performance dropped (MSE ~3202.85, R^2 = -52849.44), suggesting regression may not suit this target format

Classification Models

Binary Classification:

Converted Stroke Risk (%) into binary:

- 1 if risk > 50%
- 0 otherwise

Models Used:

- Random Forest Classifier
- Logistic Regression
- XGBoost Classifier

Best Performing Classifier–Random Forest

Model: RandomForestClassifier (n_estimators=100, max_depth=20)

Results:

- Accuracy: 1.00
- Precision/Recall/F1-score: 1.00 for both classes
- Confusion Matrix: [[4935 27]
  [0 8834]]
- ROC AUC: ~1.00

Interpretation:

- Model was able to perfectly distinguish between low and high-risk cases on the test set.

Final Evaluation

Evaluation Metrics (Best Classifier):

- MSE: 0.00
- RMSE: 0.04
- MAE: 0.00
- R^2 Score: 0.99

API Integration & Future Work

Purpose: Build a simple interface for making stroke risk predictions using the ML model.

✔ Completed:

- Flask API (/predict) handles JSON input, applies preprocessing, and returns:
  - Binary prediction (At Risk / Not At Risk)
  - Class probabilities
- HTML front-end served via Flask + ngrok for easy access
- Automatic handling of feature engineering (e.g., Symptom Severity Score, one-hot encoding)

🛠 Future Enhancements:

- Deploy to Production: Use platforms like Render or AWS with Gunicorn & Nginx
- Add Model Explainability: Integrate SHAP or LIME to show which symptoms drive predictions
- User Feedback Loop: Let users verify predictions and help improve model accuracy
- Improve Accessibility: Support multi-language inputs or batch (CSV) submissions
- Monitor Usage: Implement request logging and simple analytics for insights

Conclusion

The machine learning component of the project demonstrated that classification was a more effective approach than regression for predicting stroke risk. While initial regression results appeared promising, the models failed to generalize well after tuning. By reframing the problem as a binary classification task, we achieved significantly better and more reliable results. The Random Forest Classifier, in particular, stood out for its accuracy, robustness, and ability to handle class imbalance.

1. Robust Model Performance
   Random Forest Classifier achieved nearly perfect accuracy, precision, recall, and AUC, even after balancing the dataset with SMOTE

2. Model Interpretability
   Random forests allow for feature importance analysis, which can support clinical insights and explainability in real-world applications.

3. Ready for Deployment
   The final model has been saved (stroke_prediction_classifier.pkl) and is ready to be integrated into a decision-support system or further validated.

Thank you very much!