Oct 25, 2025

STROKE RISK PREDICTION

Using Machine Learning

Proposal

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

OVERVIEW

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”).

PROBLEM STATEMENT

Traditional assessments of stroke risk can be time-consuming and subjective. The goal of this project is to build a data-driven system capable of rapidly predicting stroke risk with consistent accuracy and interpretability.

OBJECTIVES

1. Build a clean, labeled dataset containing clinical and lifestyle variables relevant to stroke risk.
2. Preprocess binary and numerical features for ML model input.
3. Train and evaluate multiple algorithms (Logistic Regression, Random Forest, XGBoost).
4. Optimize model performance through hyperparameter tuning.
5. Provide visual explanations using feature importance plots.
6. Develop a predictive dashboard or web interface for real-time risk estimation.

DATA DESCRIPTION

The dataset will include binary features such as symptoms (Chest Pain, Dizziness, Sleep Apnea, etc.) and numeric attributes such as Age and Blood Pressure. The target variable will be '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)

METHODOLOGY

Step 1: Data Preprocessing

- Handle missing values using median/mode imputation.
- Encode binary and categorical features as 0/1.
- Normalize numeric data for scale consistency.
- Split data into training (70%) and testing (30%) sets.

Step 2: Model Development

- Baseline models: Logistic Regression, Decision Tree.
- Advanced models: Random Forest, XGBoost.
- Evaluate using cross-validation.

Step 3: Model Evaluation

Metrics:

- Accuracy
- Precision, Recall, F1-Score
- ROC-AUC Curve
- Calibration curve for probability prediction quality.

Step 4: Interpretability

- Use SHAP values and feature importance plots to identify which symptoms most strongly influence stroke risk.
- Generate model explanation visuals for report inclusion.

Step 5: Deployment (Optional)

- Create a web dashboard using Streamlit or Flask.
- Input form for patient data and outputs predicted stroke risk percentage and risk class.
- Visualization of top contributing symptoms.

Step 6: Tools and Technologies

- Programming: Python (Pandas, NumPy, Scikit-learn, XGBoost)
- Visualization: Seaborn, Matplotlib, Plotly
- Deployment: Streamlit or Flask
- Documentation: Jupyter Notebook, P.P / Word
- Version Control: GitHub

Step 7: Expected Outcomes

- A trained and validated ML model capable of predicting stroke risk with >85% accuracy.
- Comprehensive report detailing feature analysis, modeling process, and performance metrics.
- Visualization dashboards and interpretability plots.
- Optional deployment interface for live prediction.

Step 8: Timeline

Phase	Duration	Deliverable
Data Collection & Cleaning	1 Week	Clean dataset
EDA & Feature Engineering	1 Week	Correlation and binary analysis plots
Model Training	2 Weeks	Initial models and evaluation metrics
Model Optimization	1 Week	Final tuned model
Report & Dashboard Development	1 Week	Documentation and app prototype

Step 9: Expected Impact

This project contributes to healthcare analytics by providing a predictive framework that helps clinicians identify high-risk patients more efficiently. It also enhances students' skills in applied data science, ML model design, and healthcare data interpretation.

REFERENCES

1. World Health Organization (WHO) : Stroke Fact Sheet.
2. Kaggle : Stroke Prediction Dataset.
3. Scikitlearn Documentation.
4. Lundberg, S.M. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions.