Capstone Proposal

Project Title

Stroke Risk Prediction Platform

Problem Statement

Stroke is a leading cause of death and disability globally. Traditional risk assessments often fail to capture complex non-linear interactions between health factors. There is a critical need for an accessible, AI-driven tool that predicts stroke risk with high accuracy and provides interpretable insights.

Objectives / KPIs

Objectives: 1. Develop a high-accuracy ML model (XGBoost). 2. Build an interactive Streamlit dashboard. 3. Implement Causal Inference (TMLE). 4. Provide model explainability (SHAP). KPIs - Model AUC-ROC > 0.85 - Dashboard Response Time < 2s - Causal Estimate p-value < 0.05

Dataset

Source: Healthcare Dataset Stroke Data (Kaggle) Size: ~5,110 records, 12 features Type: Structured Tabular Data (Demographics, Vitals, Medical History).

Methodology

1. Preprocessing: Imputation, One-Hot Encoding, SMOTE. 2. Modeling: XGBoost Classifier for risk prediction. 3. Causal Inference: TMLE using Logistic Regression & Random Forest.

Evaluation Metrics

Predictive: AUC-ROC, F1-Score, Accuracy. Causal: ATE, Standard Error, 95% CI. Explainability: SHAP values.

Tools / Libraries

Python, Pandas, NumPy, Scikit-learn, XGBoost, Statsmodels, Matplotlib, Seaborn, Streamlit.

Expected Outcome

A deployed web app where Doctors get real-time risk predictions and Patients view health profiles. Includes visual explanations of risk factors and statistical reports on hypertension impact.

Timeline / Milestones

Week 1: Data Prep & EDA Week 2: Model Training Week 3: Causal Inference Week 4: Dashboard
Week 5: Explainability Week 6: Final Polish & Docker