Stroke Risk Prediction Case Study

# Overview

Stroke is a leading cause of death and disability worldwide. This project applies decision tree modeling to predict stroke risk from patient demographics, health indicators, and lifestyle factors. By comparing two decision tree approaches, we identify key predictors of stroke and highlight how interpretable machine learning can inform preventive healthcare decisions.

# Objective

- Goal: Develop and compare decision tree models for stroke prediction.  
- Key Questions:  
 - Which factors (e.g., age, glucose levels, hypertension) most influence stroke risk?  
 - How do different tree setups affect accuracy, sensitivity, and interpretability?

# Dataset

- Source: Public stroke dataset (Excel file)  
- Target: stroke (1 = stroke, 0 = no stroke)  
- Predictors: Age, BMI, hypertension, heart disease, smoking status, glucose levels, etc.  
- Note: Only ~5% of records are stroke cases, requiring class imbalance handling.

# Approach

1️⃣ Data Preparation  
- Removed IDs, converted categorical variables to factors, and handled missing BMI values by imputing the mean.  
- Checked for class imbalance (5% stroke vs. 95% non-stroke) and planned for upsampling.  
  
2️⃣ Exploratory Analysis  
- Created a correlation heatmap showing age, glucose level, and hypertension had the strongest relationship to stroke.  
  
3️⃣ Model Development  
- Model 1: Full decision tree using all predictors.  
- Model 2: Reduced decision tree using top 6 predictors (age, glucose, ever\_married, hypertension, heart\_disease, BMI).  
- Applied 10-fold cross-validation and upsampling to address imbalance.

4️⃣ Model Evaluation & Comparison  
- Evaluated models using accuracy, sensitivity, specificity, and AUC.  
- Plotted ROC curves and visualized decision tree splits.

# Key Results

- Model 1: ~74.7% accuracy, AUC ~0.82  
- Model 2: ~74.5% accuracy, AUC ~0.83  
  
📌 Model 2 performed slightly better on AUC and is easier to interpret, making it more clinician-friendly.  
  
Top predictors for stroke risk:  
- Age  
- Glucose level  
- Hypertension  
- Heart disease  
- BMI

# Visual Summaries

- ROC Curve: Displays how well Model 2 separates stroke vs. non-stroke patients (AUC 0.83).  
- Decision Tree Plot: Visual structure highlighting age and glucose as primary decision points.  
- Correlation Heatmap: Shows relationships between health indicators and stroke.

# Insights

- Age, glucose level, and hypertension are the top drivers of stroke prediction.  
- Simplified tree (Model 2) offered higher interpretability with slightly better AUC.  
- Upsampling improved the model’s ability to catch more stroke cases.

# Next Steps

- Test Random Forest or Gradient Boosting for potential performance improvements.  
- Use SMOTE or cost-sensitive learning to further improve sensitivity.  
- Package the best model into a Shiny app for hospital and clinic use.

# Tools Used

- R (caret) – Model training, CV, and upsampling  
- rpart & rpart.plot – Decision tree modeling & visualization  
- pROC – ROC curve & AUC calculations  
- corrplot – Correlation visualization

# Conclusion

This case study demonstrates how interpretable models like decision trees can help identify at-risk patients and guide preventive health strategies. By focusing on factors like age, glucose level, and hypertension, healthcare providers can better allocate resources and intervene earlier.