Stroke Prediction Using Machine Learning

By: David Obele & Zachery Davis

Group 10

THE WILLIAM STATES LEE COLLEGE OF ENGINEERING

Why Stroke Prediction Matters

- Stroke is the 2nd leading cause of death globally (WHO).
- Responsible for ~11% of global deaths.
- Project Goal: Develop a machine learning model to predict stroke occurrences using health and demographic features.

Key Questions We Aim to Answer

- What factors significantly contribute to stroke risk?
- Which individuals are most at risk?
- Which machine learning model performs best for this data?

Data Collection & Pre-processing

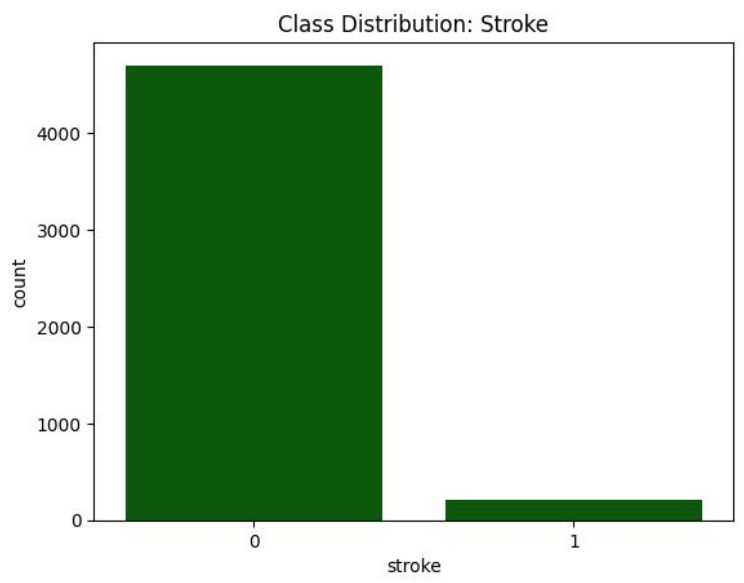
- Dataset: Publicly available from Kaggle (5,110 samples, 11 features + target variable).
- Features: Gender, Age, Hypertension, Heart Disease, Marital Status, Work Type, Residence Type, Avg. Glucose Level, BMI, Smoking Status.

Pre-processing

- Data Cleaning: Drop data for missing BMI values.
- Encoding: One-hot encoding for categorical variables.
- Normalization: Scaling age, glucose level, and BMI.
- Class Imbalance Handling: Applied SMOTE for oversampling.

Insights from Data Analysis

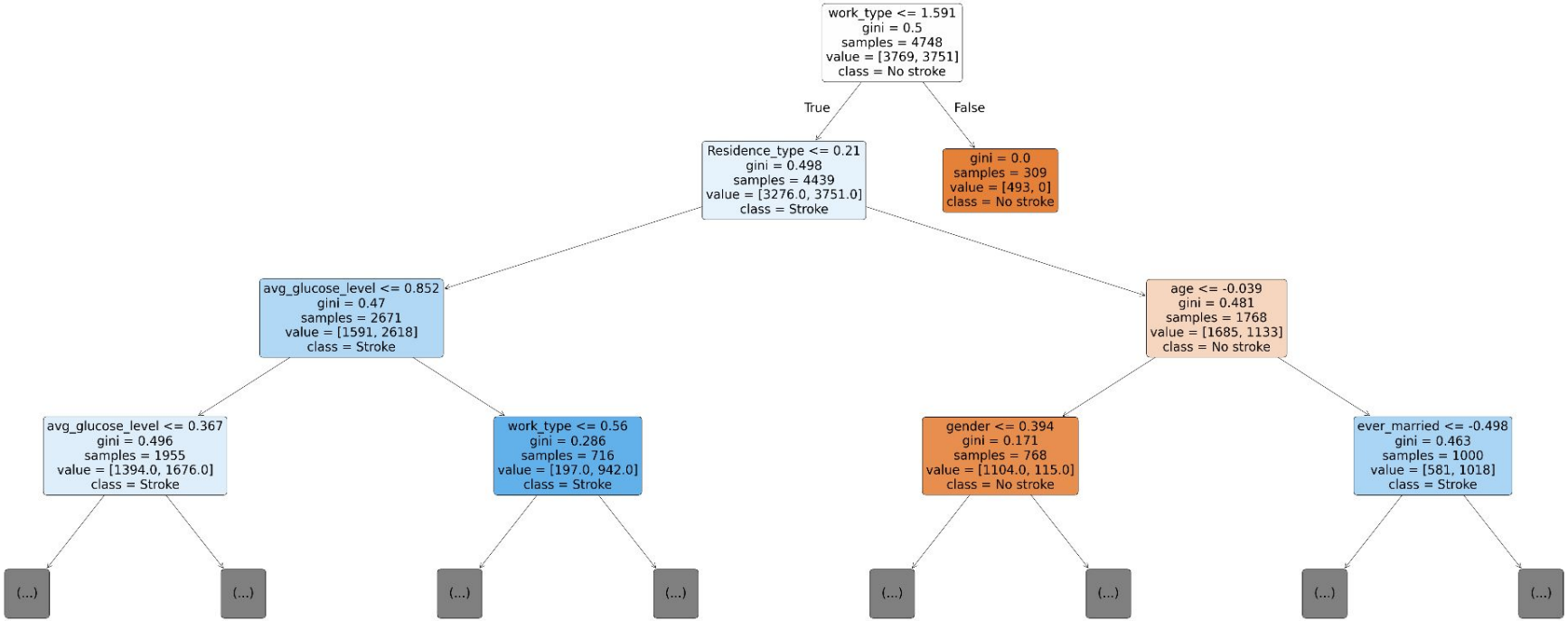
- Average Age: 43.2 years
- Stroke Incidence: ~4.9%
- Visualizations of Data:



Machine Learning Models Explored

- Algorithms Tested: Logistic Regression and Random Forest
- Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, ROC AUC.
- Best Model Performance:
  - Random Forest
  - Accuracy: ~95%
  - Balanced Precision and Recall.

Random Forest

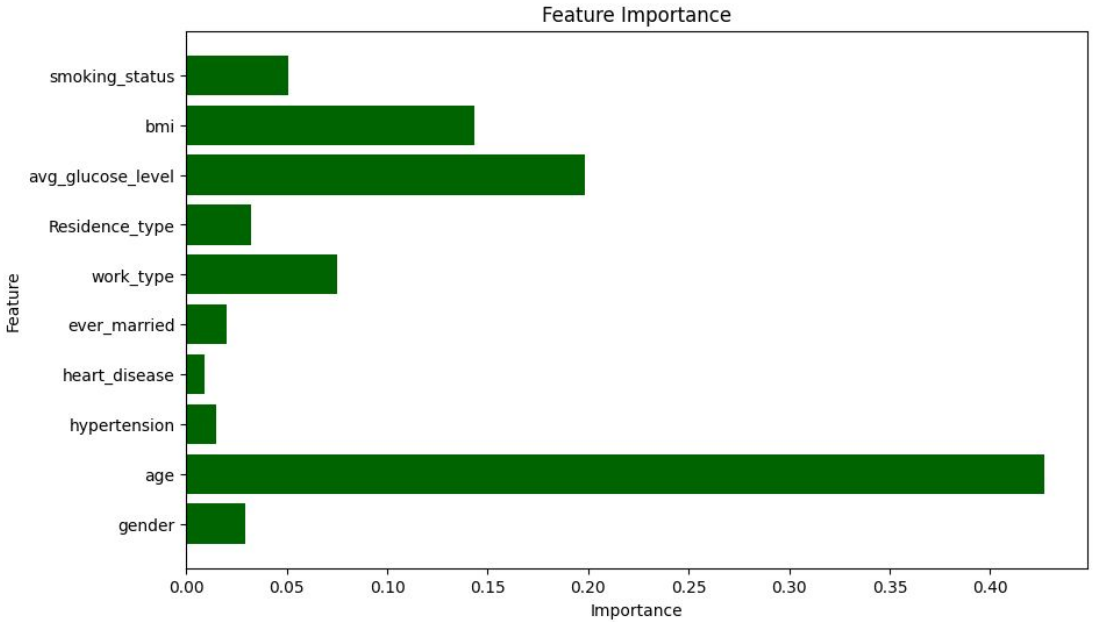


Model Evaluation Summary

- Key Metrics Achieved:
  - High Accuracy (~95%)
  - Strong Precision and Recall
  - ROC AUC: Demonstrated good model discrimination.
- Cross-validation: Model generalizes well on unseen data.

Key Factors Identified

- Most Influential Features:
  - Age
  - Average Glucose Level
  - BMI



- Feature Engineering & Hyperparameter Tuning: Enhanced performance.

Final Takeaways

- Successes:
  - Accurate stroke prediction model.
  - Addressed initial questions with clear findings.
- Learnings:
  - Importance of data preprocessing.
  - Need for handling class imbalance.
  - Value of hyperparameter tuning.

Next Steps

- Expand dataset with more samples.
- Include additional health metrics.
- Explore deep learning models.
- Conduct more model evaluations with updated data.

Considerations for Real-World Impact

- Positive Impacts:
  - Improves healthcare decisions.
  - Proactive stroke prevention.
- Ethical Concerns:
  - Data privacy.
  - Bias in datasets.
  - Over-reliance on automated systems.
- Solution: Transparency, fairness evaluations, and regular updates.

Thank You for Your Attention!