STROKE PREDICTION USING PATIENT HEALTH RECORDS

Team Name: **Stream Savants**

Authors: Aele Santhosh, Ratul Khan, Balla Malleswara Rao, Abdul Hafiz

Email: santhoshaele@iisc.ac.in, ratulkhan@iisc.ac.in, malleswarar1@iisc.ac.in, abdulhafiz@iisc.ac.in

PROBLEM DEFINITION

Stroke is the second leading cause of death worldwide and a major contributor to long-term disability. Its prevention hinges on the early identification of modifiable risk factors like hypertension, diabetes (reflected in glucose levels), smoking, obesity (BMI), and heart disease. Current clinical practices often rely on manual assessment, which can lead to late diagnosis and emergency intervention.

MOTIVATION

Stroke imposes a massive economic and human burden in India, with 7.7 million cases annually.
Up to 80% of strokes are preventable through timely intervention.
Each case costs ₹3–5 lakhs to treat, while prevention is ten times cheaper.
Machine learning on patient data enables scalable, proactive stroke prevention.

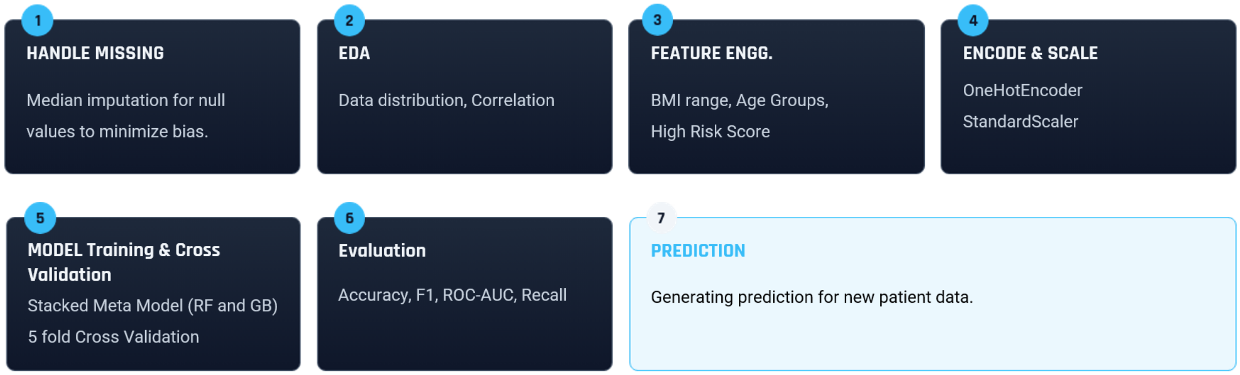
DESIGN GOALS

- **Build an Accurate Model:** Develop a classifier that achieves high performance metrics.
- **Ensure Clinical Interpretability:** Provide clear feature importance rankings so healthcare professionals can understand why a patient is flagged as high-risk.
- **Handle Imbalanced Data:** Achieve >85% Recall to minimize false negatives (missed cases).

Data Overview

- **Demographics:** Age, Gender, Marital Status, Work Type, Residence Type
- **Vitals:** BMI, Glucose Level, Hypertension
- **Risk Indicators:** Heart Disease, Smoking Status

METHODOLOGY



1 HANDLE MISSING

Median imputation for null values to minimize bias.

2 EDA

Data distribution, Correlation

3 FEATURE ENGG.

BMI range, Age Groups, High Risk Score

4 ENCODE & SCALE

OneHotEncoder
StandardScaler

5 MODEL Training & Cross Validation

Stacked Meta Model (RF and GB)
5 fold Cross Validation

6 Evaluation

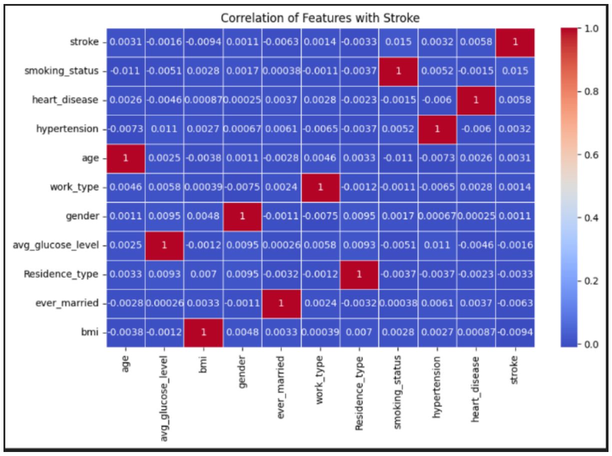
Accuracy, F1, ROC-AUC, Recall

7 PREDICTION

Generating prediction for new patient data.

- Handle Missing: Clean the data by filling in any missing values.
- EDA: Identify patterns, distributions and data quality issues.
- Feature Engineering: Create better input signals from the raw data.
- Encode & Scale: OneHotEncoding for categorical columns and turn all inputs into comparable numeric values.
- Model Training & Cross Validation: Train the model and test it several times for fairness using 5-K fold cross validation.
- Evaluation: Measure how well the model is performing.
- Prediction: Use the final model to predict stroke risk for each new patient.

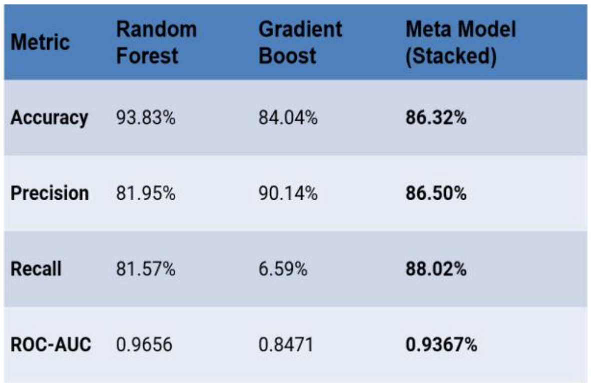
Exploratory Data Analysis



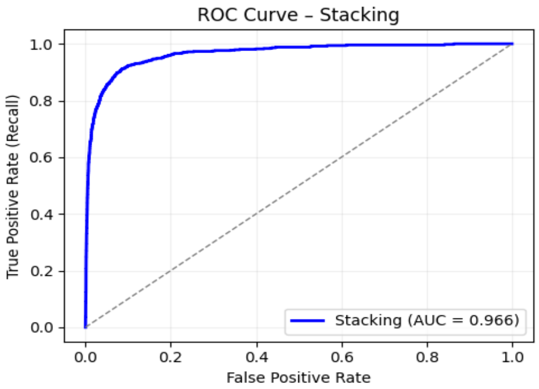
- Correlation is very low suggesting non linearity between features.
- Dataset is evenly distributed across gender, work_type, Residence_type and smoking_status, age, average_glucose_level and BMI.

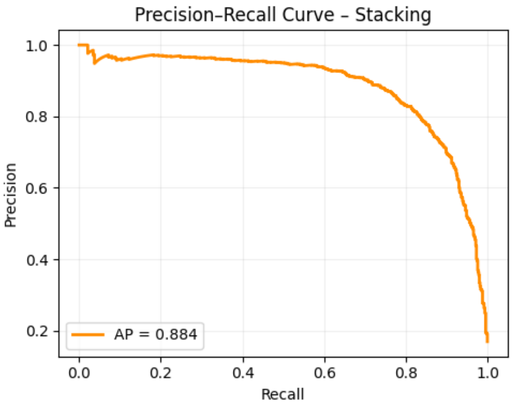
- The stroke and non-stroke distribution across gender, age, average_glucose_level and BMI is also very similar.
- There are not many outliers in the dataset. Data is very evenly distributed. So, there is no need for any normalization.

MACHINE LEARNING MODEL & EVALUATION METRICS



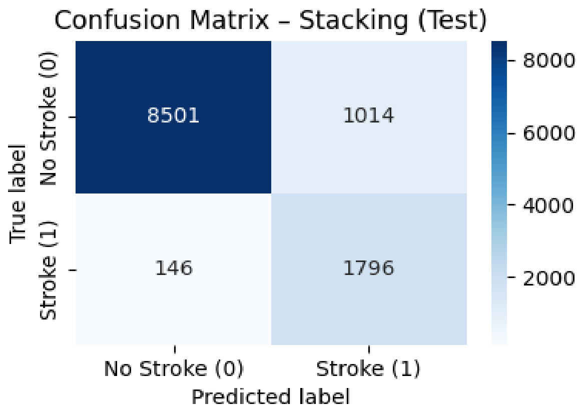
Metric	Random Forest	Gradient Boost	Meta Model (Stacked)
Accuracy	93.83%	84.04%	**86.32%**
Precision	81.95%	90.14%	**86.50%**
Recall	81.57%	6.59%	**88.02%**
ROC-AUC	0.9656	0.8471	**0.9367%**

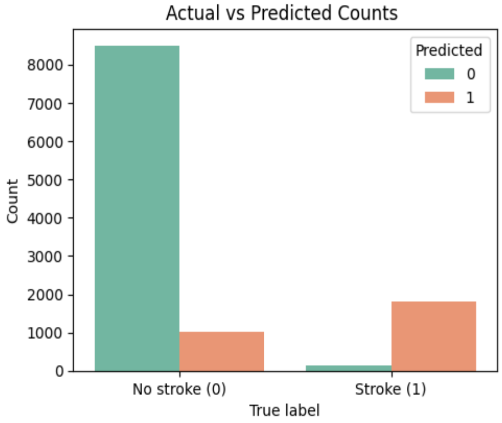


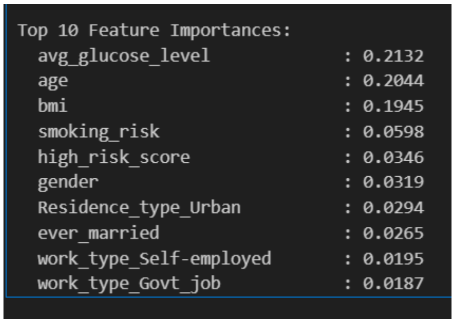


- Random Forest – High Accuracy but lower precision and recall.
- Gradient Boost – High Precision but low recall
- Stacked Meta Model – High Accuracy, Precision and Recall

RESULTS & ANALYSIS







```
Top 10 Feature Importances:
  avg_glucose_level            : 0.2132
  age                          : 0.2044
  bmi                          : 0.1945
  smoking_risk                 : 0.0598
  high_risk_score              : 0.0346
  gender                       : 0.0319
  Residence_type_Urban         : 0.0294
  ever_married                 : 0.0265
  work_type_Self-employed      : 0.0195
  work_type_Govt_job           : 0.0187
```

- Stroke detection (1) Accuracy: 92.5%, Recall rate: 88%.
- Feature importance: avg_glucose_level, Age, BMI, smoking risk.

LIMITATIONS

Challenges included ensuring scalability, and maintaining interpretability. Training data lacks more complex inputs like imaging or genetic data and current model does not incorporate temporal health changes over time.

SUMMARY

The system meets scalability, accuracy, and interpretability goals using big data and ML frameworks.

FUTURE EXTENSIONS

- Support for multiple hospitals and regions.
- Explore deep learning models for sequential health data.
- Develop dashboards for clinicians.