Stroke Prediction Using Machine Learning

Early Diagnosis Through Health Data Analytics

Siddharth Ranka | Sarthak Pandit | Mansi Muneshwar

Indian Institute Of Technology, Jodhpur

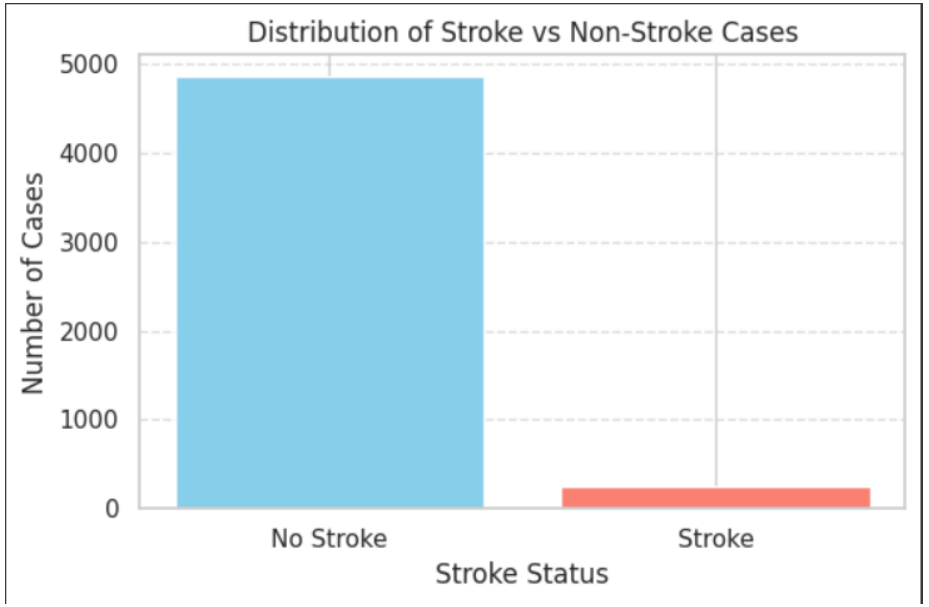
Why Stroke Prediction?

- Stroke is a leading cause of death and long-term disability.
- Early diagnosis is crucial for effective treatment.
- Traditional diagnosis can be time-consuming and prone to error.
- Can machine learning help predict stroke risk earlier?

Dataset Details

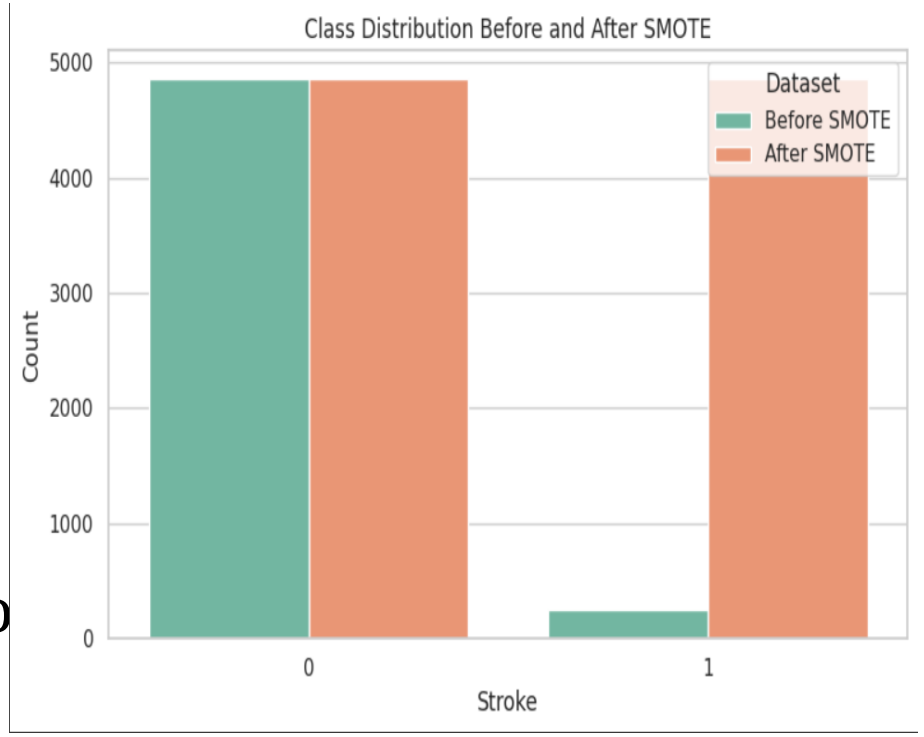
- Total records: 5,110
- Features: Age, Hypertension, Heart Disease, BMI, Glucose Level, etc.
- Issue: Highly imbalanced dataset (few stroke cases)

Feature Name	Unique Values
id	5110
gender	3
age	104
hypertension	2
heart_disease	2
ever_married	2
work_type	5
Residence_type	2
avg_glucose_level	3979
bmi	418
smoking_status	4
stroke	2



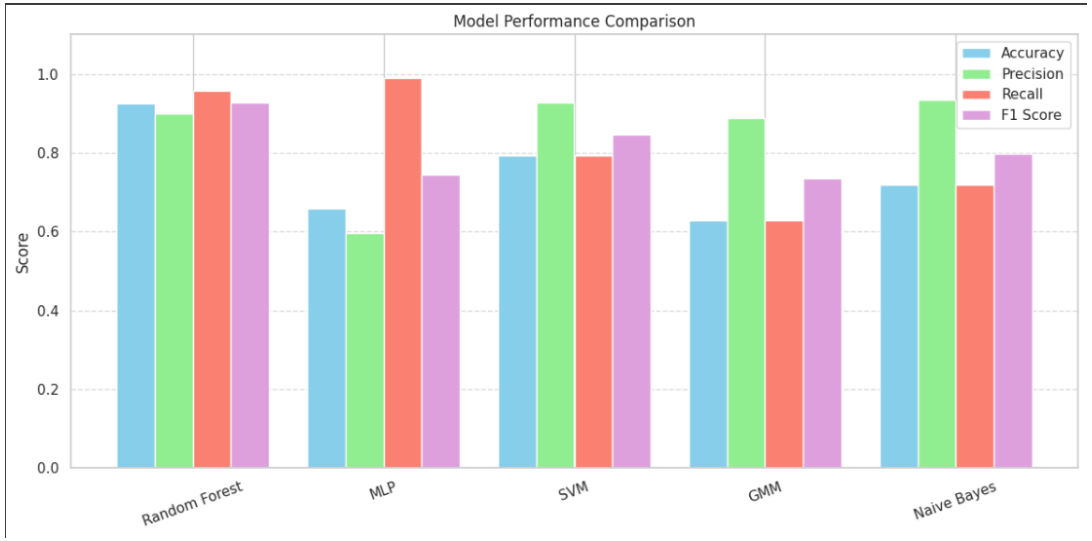
Data Preprocessing & Balancing

- Imputed Missing Values: Addressed missing data, particularly in the BMI feature, using appropriate imputation techniques.
- Encoded Categorical Variables: Converted categorical features (e.g., gender, smoking status) into numerical format using label encoding.
- Balanced the Dataset: Applied SMOTE (Synthetic Minority Oversampling Technique) to handle class imbalance between stroke and non-stroke cases.
- Normalized Feature Values: Scaled all numerical features to ensure consistent input ranges across models, improving convergence and performance.

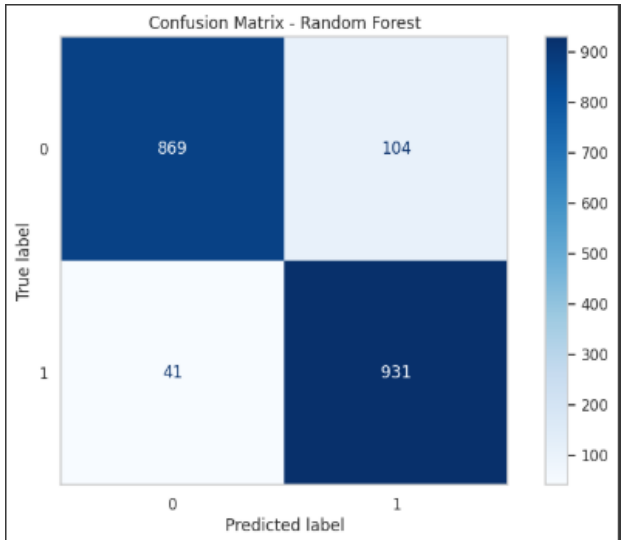


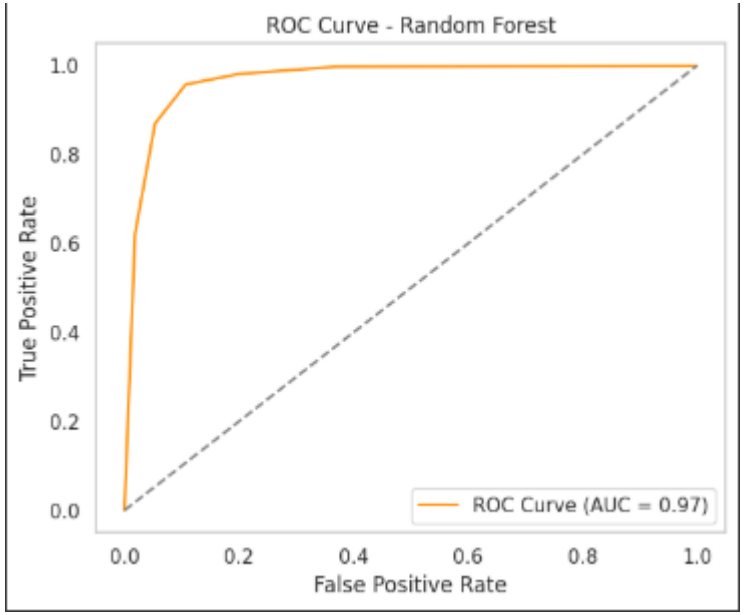
Models Trained & Performance

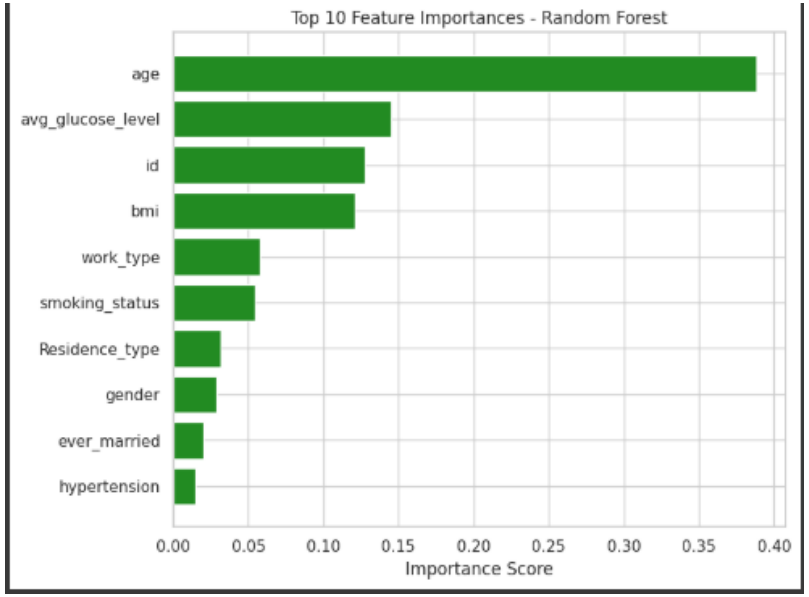
- Trained Multiple Machine Learning Models: Implemented and compared various algorithms including MLP, GMM, Support Vector Machine (SVM), Decision Tree, and Random Forest.
- Best Performing Model: The Random Forest classifier delivered the highest performance with an accuracy of 92.46%.
- Balanced Evaluation Metrics: Achieved strong results across precision, recall, and F1-score, indicating reliable and consistent model performance.



Results & Interpretation







- Confusion Matrix helps visualize model predictions: TP, FP, FN, TN
- Feature Importance shows which inputs influenced the model most
- ROC Curve and AUC reflect the model's ability to classify accurately

Conclusion & Future Work

- Machine Learning can play a significant role in early stroke risk detection.
- Among the models tested, Random Forest delivered the highest accuracy and balanced performance.
- While results are promising, there’s room for improvement and future expansion.

Future Directions:

- Incorporate more diverse and larger datasets for better generalization.
- Explore advanced models like Neural Networks for deeper insights.
- Consider deploying the model as a user-friendly web or mobile application.

THANK YOU