Synthetic Stroke Prediction Using Machine Learning

Presented By :

Pratiksha Shinde - 202301070155
Purva Nigade - 202301070146
Amruthsvarshini Repalle - 20230107158
Aakanksha Sah - 202301070148

Problem Statement

Stroke is one of the leading causes of death and long-term disability across the world. Many patients fail to receive timely diagnosis due to lack of early detection systems.

Early prediction of stroke risk using machine learning can play a crucial role in preventive healthcare, enabling timely medical intervention and saving lives.

To develop a machine learning model capable of predicting the likelihood of stroke in an individual based on medical history and lifestyle data

Objectives

- To collect and preprocess stroke-related health data.
- To perform feature engineering and handle missing or imbalanced data.
- To train and evaluate multiple machine learning models (e.g., Random Forest, Logistic Regression).
- To compare model performance using metrics such as accuracy, recall, and ROC-AUC.
- To deploy the best-performing model for stroke prediction in a user-friendly application.

Tools & Technologies Used

Programming Language: Python

Libraries & Frameworks:

Pandas, NumPy – Data handling
Matplotlib, Seaborn – Visualization
Scikit-learn – Model building & evaluation
Imbalanced-learn (SMOTE) – Data balancing
Joblib – Model deployment

Development Environment: VS Code

Data Preprocessing

Data preprocessing steps

Split

The data was split into 80% training and 20% testing sets. This allows for model evaluation.

Missing Values

Median imputation was used to handle missing BMI values. This ensures no data is lost.

Balancing

SMOTE was used to equalize stroke vs non-stroke cases. This addresses class imbalance issues.

Encoding

Label Encoding was applied to categorical data. This converts text into numerical form.

Scaling

StandardScaler was used for numerical normalization. This brings all features to a similar scale.

Model Building

Machine Learning Models Implemented:

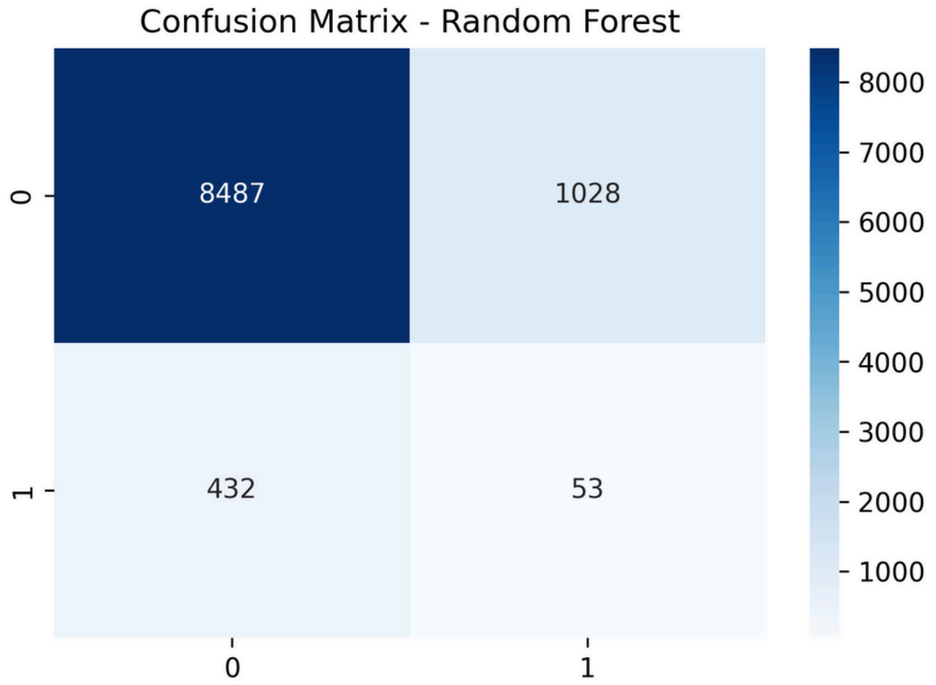
1. **Random Forest Classifier**
   - n_estimators = 200
   - class_weight = 'balanced'
   - Ensemble-based model with multiple decision trees.
2. **Logistic Regression**
   - class_weight = 'balanced'
   - Used as a baseline linear model for comparison.
   -

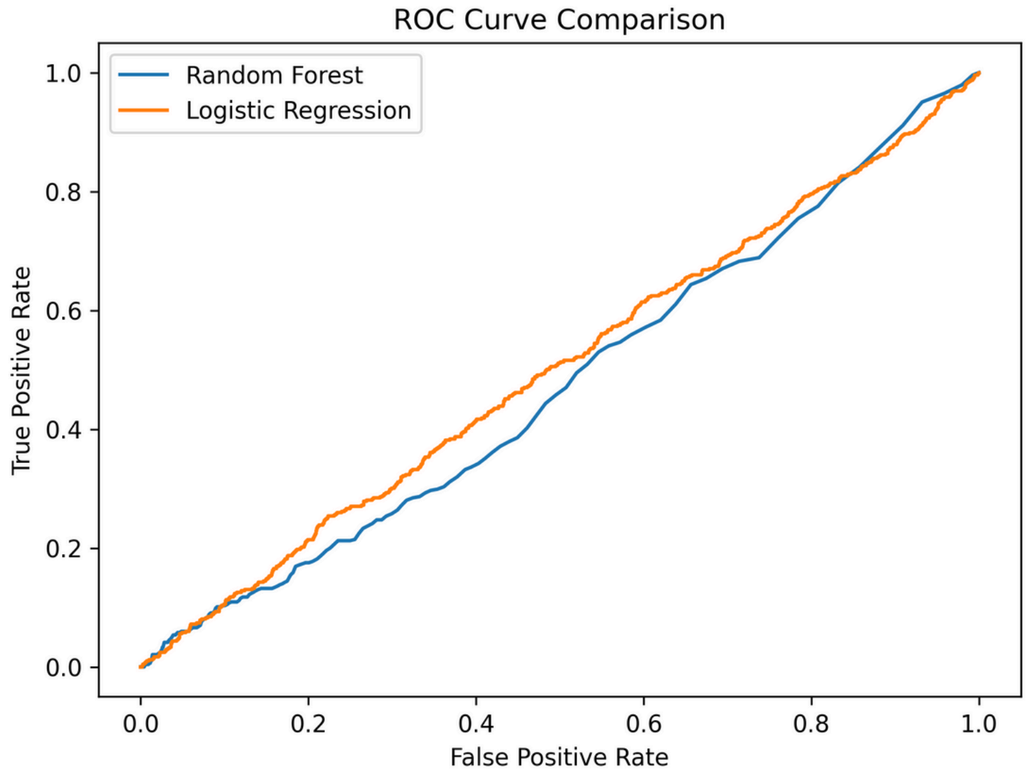
Training:
Models trained on scaled and balanced data to ensure fairness and improved accuracy.

Model Evaluation

Evaluation Metrics Used:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Score





Key Insights & Conclusion

Key Insights:

- Age, BMI, glucose level, and hypertension are major stroke predictors.
- Applying SMOTE helped balance the dataset and improved model recall.
- Random Forest provided the best balance between accuracy and interpretability.

Future Scope:

- Integrate model into a Streamlit web app for real-time use.
- Add SHAP explainability for model transparency.
- Expand dataset with clinical and genetic parameters.

Conclusion:

An efficient and accurate machine learning model was successfully developed to predict stroke risk.
This system can assist healthcare professionals in early diagnosis and preventive treatment, potentially saving lives.

Thank You