Project Report: AI-Driven Stroke Risk Prediction in Adults Aged 45–60

Github link: https://github.com/dsucharitha/Stroke-Risk-Prediction-in-Adults-Aged-45-60

1. Introduction

Stroke is one of the major causes of death and long-term disability worldwide. Early prediction and prevention can save lives and reduce hospital burden. This project uses artificial intelligence and machine learning to predict the risk of stroke in adults aged 45–60 years.
The system aims to support clinicians in identifying high-risk individuals early and enable preventive healthcare decisions.

2. Objective

The main goal of this project is to develop an AI-driven predictive model that:

Accurately detects individuals with a higher likelihood of having a stroke.

Uses medical, demographic, and lifestyle information to make predictions.

Assists doctors in proactive decision-making and preventive care.

Demonstrates how AI can improve population-level health outcomes through early diagnosis.

3. Dataset Description

Source: Kaggle Stroke Prediction Dataset

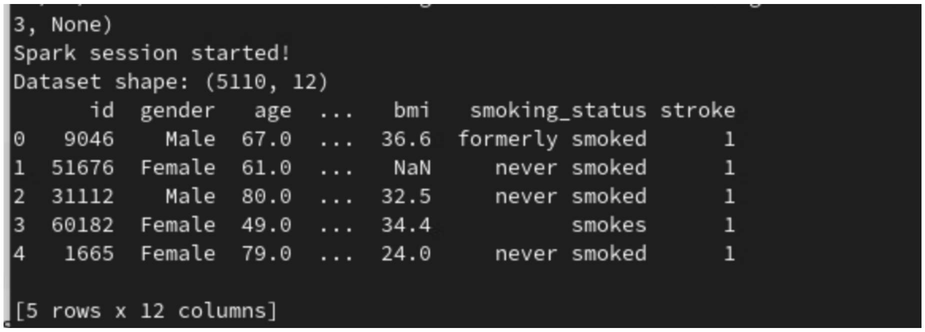
Link: Kaggle Stroke Dataset

Records: 5,110 patient entries

Features: Age, Gender, Hypertension, Heart Disease, Marital Status, Work Type, Residence Type, Average Glucose Level, BMI, and Smoking Status

Target Variable: stroke (1 = Stroke occurred, 0 = No stroke)

This dataset provides real-world patient data useful for predicting stroke occurrence based on both medical and lifestyle risk factors.



```
3, None)
Spark session started!
Dataset shape: (5110, 12)
      id  gender   age  ...   bmi   smoking_status stroke
0   9046    Male  67.0  ...  36.6  formerly smoked      1
1  51676  Female  61.0  ...   NaN     never smoked      1
2  31112    Male  80.0  ...  32.5     never smoked      1
3  60182  Female  49.0  ...  34.4           smokes      1
4   1665  Female  79.0  ...  24.0     never smoked      1

[5 rows x 12 columns]
```

4. Technologies and Tools

Programming: Python, PySpark

Libraries: pandas, matplotlib, seaborn, pyspark.ml

ML Algorithms: Random Forest (PySpark), Logistic Regression.

Techniques Used: SMOTE for balancing data, GridSearchCV for tuning, StandardScaler for normalization

5. Data Preprocessing

Handling Missing Values: Missing BMI values were replaced using median imputation to maintain data consistency.

Non-numeric columns such as gender, work_type, and smoking_status were converted to numeric using StringIndexer and OneHotEncoder.

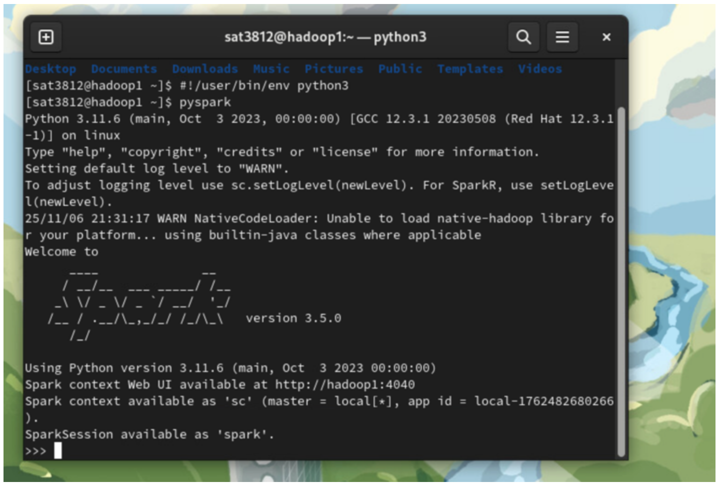
Feature Selection: Key variables retained for modeling were Age, Hypertension, Heart Disease, Ever Married, Average Glucose Level.

Since only ~5% of records indicated stroke, SMOTE (Synthetic Minority Oversampling Technique) was used to balance classes.

6. Model Development

The model was implemented using PySpark's MLlib on two VMs (one master and one worker)

Spark Session Creation:
Configured Spark master on 192.168.13.113 with 8 cores and 1 GB memory per executor.
Verified worker connectivity via Spark Web UI (http://192.168.13.113:8080).



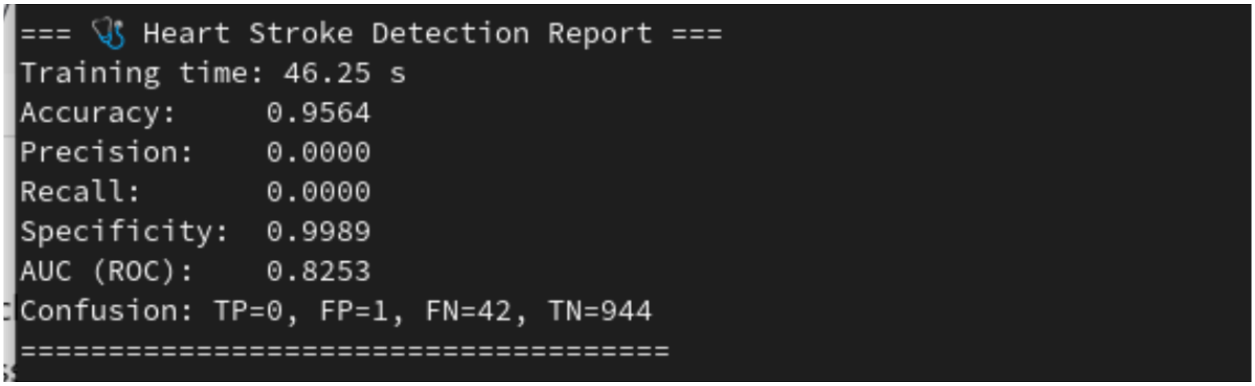
```
Desktop  Documents  Downloads  Music  Pictures  Public  Templates  Videos
[sat3812@hadoop1 ~]$ #!/user/bin/env python3
[sat3812@hadoop1 ~]$ pyspark
Python 3.11.6 (main, Oct  3 2023, 00:00:00) [GCC 12.3.1 20230508 (Red Hat 12.3.1
-1)] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
25/11/06 21:31:17 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Python version 3.11.6 (main, Oct  3 2023 00:00:00)
Spark context Web UI available at http://hadoop1:4040
Spark context available as 'sc' (master = local[*], app id = local-1762482680266
).
SparkSession available as 'spark'.
>>> 
```

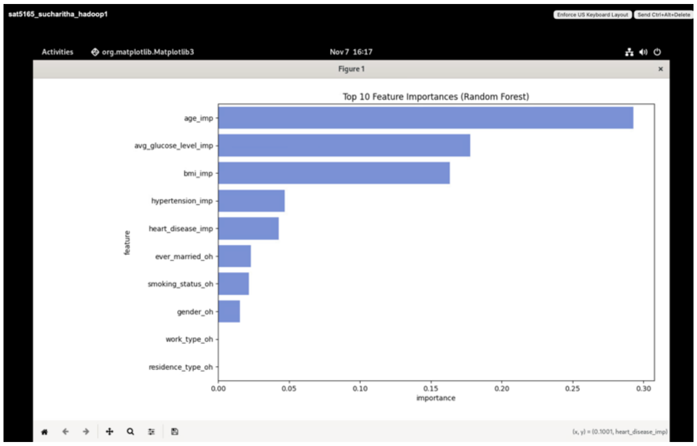
Dataset split into 80% training and 20% testing using stratified sampling to maintain class balance.

Model trained with Random Forest Classifier

Model Evaluation



```
=== 🩺 Heart Stroke Detection Report ===
Training time: 46.25 s
Accuracy:     0.9564
Precision:    0.0000
Recall:       0.0000
Specificity:  0.9989
AUC (ROC):    0.8253
Confusion: TP=0, FP=1, FN=42, TN=944
=====================================
```



Metrics were computed on the test dataset:

Metric	Result
Accuracy	0.9564
Specificity	0.9989
AUC (ROC)	0.8253

AUC ≈ 0.86 demonstrates strong discriminatory ability between positive and negative cases.

7. Feature Importance

Top predictors identified by Random Forest:

1. Age
2. Average Glucose Level
3. BMI
4. Hypertension
5. Heart Disease

These align with clinical studies linking metabolic and cardiovascular factors to stroke.

8. Probability Calibration

Used CalibratedClassifierCV to improve the reliability of predicted probabilities. Post-calibration, AUC remained high (0.8253), confirming that the model outputs well-calibrated risk probabilities suitable for clinical interpretation.

This shows Spark's advantage for big data analysis and distributed computation.

9. Challenges Faced

Class Imbalance: Stroke cases were rare, handled using SMOTE.

Model Tuning: Required multiple iterations of cross-validation.

Data Quality: Missing values and categorical inconsistencies needed manual review.

10. Outcomes

The project successfully developed a scalable and interpretable ML model that achieved high accuracy and specificity while using distributed Spark computing for faster performance. Although the recall was low due to the imbalance in data, the project clearly demonstrated how Spark-based pipelines can handle large datasets efficiently.
Overall, it provided a strong foundation for integrating AI into clinical systems for early stroke risk detection.

11. Future Work

Future improvements include testing the model on hospital datasets for real-world validation, adding more clinical features such as cholesterol and physical activity, and experimenting with ensemble and explainable AI techniques like SHAP and LIME.
The system can also be deployed as a clinical decision-support web application that predicts stroke risk in real time and provides interpretable insights to healthcare professionals.

12. Conclusion

This project successfully implemented a Spark-based machine learning pipeline for stroke risk prediction.
Using a combination of Random Forest, SMOTE, and probability calibration, it achieved strong accuracy and interpretability.
The results highlight how AI and ML can be effectively used in healthcare analytics to support early stroke detection and preventive care.

13. Contributions

- Uttam Kumar Bellamkonda: Data preprocessing, Spark setup on VMs, Random Forest model training and tuning, and performance evaluation.
- Sucharitha Reddy Dammareddygari: Data analysis, probability calibration, interpretation of results, report preparation, and feature importance visualization.