كلية الهندسة
جامعة المنصورة

قسم هندسة الحاسبات ونظم التحكم
Computers and Control Systems Engineering Dept.

Project Report: Predictive Model for Stroke Risk Assessment

Supervised by: Dr / Amira Yassien
Eng / Ahmed Elhelow
Eng / Fatma Gamal

Team members

no.	الاسم	السكشن
1	احمد محمد احمد محمد احمد شتيه	2
2	زياد محمد السعيد الميمي	3
3	عبدالحليم اسامه محمد	4
4	محمد فتحي قطب عبدالمجيد	6
5	محمود عبدالرحمن ابراهيم	6
6	سعد احمد عبدالغفار	3
7	انس محمد السعيد عبدالغني	2
8	محمد محمود حامد قميحه	6

Table of Contents

1. Introduction
2. Dataset
3. Data Preprocessing
4. Feature Engineering
5. Model Selection and Training
6. Model Deployment
7. Full project
8. Conclusion

Introduction

The goal of this project was to construct a predictive model capable of assessing an individual's likelihood of experiencing a stroke based on a variety of demographic and lifestyle factors. Leveraging machine learning techniques, the model was designed to analyze input data and provide accurate predictions regarding stroke risk.

Dataset

For this project we are using Cerebral Stroke Prediction-Imbalanced Dataset from Kaggle. The dataset consists of 12 features including the target column which is imbalanced.
columns are

- ID
- Gender
- Age
- Hypertension
- Heart disease
- Ever married
- Work type
- Residence type
- Avg glucose level
- Bmi
- Smoking status
- stroke

Data Preprocessing

The dataset utilized for this project was characterized by its cleanliness, with minimal missing values. However, to ensure the integrity of the data, a robust preprocessing phase was conducted. Numerical feature containing missing values was imputed using the KNN imputer, a method chosen for its effectiveness in handling missing data by computing the nearest neighbors of each sample. Categorical feature with missing values was strategically assigned a new category labeled as "unknown" to maintain the dataset's comprehensiveness while accounting for missing information.

During the exploratory data analysis (EDA) phase, it became evident that certain features exhibited skewness. To address this issue and improve the model's performance, a Yeo-Johnson power transformation was applied to these skewed features. This transformation not only helped normalize the data but also enhanced the model's ability to capture underlying patterns accurately. Furthermore, the target column displayed class imbalance, a common challenge in binary classification tasks. To rectify this imbalance, the ADASYN algorithm was employed to oversample the minority class, thereby ensuring that the model learned from a more balanced dataset and made fair predictions for both classes.

Feature Engineering

A critical aspect of the preprocessing phase was feature engineering, which aimed to extract meaningful insights from the raw data and enhance the predictive power of the model. To streamline this process, function transformers were developed to execute specific transformations for numerical and categorical features independently. These transformers encapsulated tasks such as scaling numerical features and encoding categorical variables, thereby simplifying the overall preprocessing workflow.

Subsequently, these feature transformers were integrated into separate pipelines—one for numerical and another for categorical columns—using the Pipeline module from scikit-learn. These pipelines were then unified using ColumnTransformer, facilitating the seamless execution of all preprocessing steps in a coherent manner. By encapsulating feature engineering within these pipelines, the preprocessing phase was automated, ensuring consistency and reproducibility across different datasets and experiments.

Model Selection and Training

With the preprocessed data in hand, the next step involved selecting and training appropriate machine learning models. A variety of algorithms were considered and evaluated based on their performance metrics and suitability for the task. Through rigorous experimentation and comparative analysis, it was determined that Logistic Regression and GradientBoostingClassifier exhibited the most promising results in terms of predictive accuracy and generalization capability.

Once the models were identified, they underwent a fine-tuning process to optimize their hyperparameters and further enhance their performance. This optimization was carried out using RandomizedSearchCV, a technique that efficiently explores the hyperparameter space and identifies the optimal configuration through randomized search and cross-validation. By fine-tuning the models, their predictive capabilities were maximized, resulting in more robust and reliable stroke risk assessments.

Model Deployment

The culmination of this project was the deployment of the selected model, GradientBoostingClassifier, to real-world applications. To ensure widespread accessibility and usability, the model was deployed through two distinct platforms—an Android application developed using the Flutter framework and a web page created with Flask, a lightweight web framework for Python.

Through these deployment channels, users could input their demographic and lifestyle information and receive real-time predictions regarding their stroke risk. This seamless integration of the predictive model into everyday technologies enhances its practical utility and empowers individuals to make informed decisions about their health and well-being.

Full project

In this **GitHub repository**, You will find

- the code used for model training and deployment
- code used for developing web page using flask
- code used for developing android app using flutter
- screenshots from web page and android app

Conclusion

In summary, this project successfully developed a predictive model for assessing stroke risk based on demographic and lifestyle factors. By leveraging advanced machine learning techniques, including robust preprocessing, feature engineering, and model selection, a highly accurate and reliable predictive model was constructed. The deployment of this model through both mobile and web platforms signifies a significant step towards democratizing healthcare and promoting proactive health management among individuals.