T.C.
FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ

Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü

INTRODUCTION THE MACHINE LEARNING PROJECT PROPOSAL & MILESTONE

Nagkichan MOUSTAFA
Ali DEREYURT
Sevde Hatice KÖMÜRCÜ
Muhammed ALTAY

Lecturer : Dr. Gönül ULUDAĞ

Stroke Prediction Dataset

Abstract

This report delves into the development and evaluation of predictive models aimed at determining the risk of stroke, leveraging a comprehensive dataset of over 5,000 patient records. Strokes stand as a critical health issue globally, representing the second leading cause of death and a major cause of disability. The urgency to predict and prevent strokes is paramount for enhancing patient care and reducing healthcare burdens. Our study employs advanced machine learning algorithms to analyze a variety of predictors including demographic factors, medical history, and lifestyle choices such as smoking status and diet. Key features from the dataset—such as age, hypertension, heart disease, and average glucose levels—are meticulously analyzed to understand their impact on stroke risk. Through a systematic approach involving logistic regression, decision trees, and neural networks, this report not only benchmarks the effectiveness of these models in predicting strokes but also explores their potential integration into clinical settings for proactive healthcare management. The results aim to contribute to the body of knowledge by enhancing the accuracy of stroke predictions, thereby supporting healthcare professionals in making informed decisions about patient care and preventive measures. This analysis stands as a significant step towards utilizing data-driven insights in medical prognostics, emphasizing the critical role of machine learning in transforming health diagnostics and intervention strategies.

Introduction

Stroke remains a leading cause of mortality and long-term disability worldwide, prompting an urgent need for early diagnostic strategies that can guide preventive care and reduce the incidence of this devastating condition. The World Health Organization identifies strokes as the second leading cause of death globally, accounting for 11% of total fatalities. These statistics underline the critical importance of predictive analytics in healthcare to identify individuals at high risk of stroke before the occurrence of any adverse events.

This project is rooted in the potential of machine learning (ML) to transform healthcare outcomes through predictive modeling. Utilizing a dataset comprised of 5,110 patient records, our objective is to develop an algorithm that predicts the likelihood of stroke based on a variety of input parameters including, but not limited to, age, gender, hypertension, heart disease, and lifestyle factors such as smoking and diet. Each record in the dataset encapsulates crucial information that might influence stroke risk, offering a rich foundation for our analytical models.

The choice of machine learning for this task is driven by its ability to handle large datasets and uncover complex patterns that may not be evident through traditional statistical methods. By integrating diverse data types—from categorical variables like work type and residence type to continuous variables such as BMI and glucose levels—our models can provide nuanced insights into the multifactorial nature of stroke risk.

Our approach involves several stages of data handling and model development, starting with data preprocessing to address missing values and normalize data scales. Following this, feature selection techniques are applied to identify the most significant predictors of stroke. The ultimate goal is to deploy a range of machine learning models, each offering distinct advantages in terms of prediction accuracy, interpretability, and clinical applicability.

This introduction sets the stage for a detailed exploration of the dataset, the methodologies employed, and the implications of our findings. By predicting stroke risk with high accuracy, we aim to equip healthcare providers with a powerful tool that can save lives through timely and targeted interventions. This study not only enhances our understanding of stroke risk factors but also contributes to the broader field of medical informatics, where data-driven insights have the potential to revolutionize patient care practices globally.

Related Work

The field of stroke prediction has been dynamically evolving with the integration of machine learning techniques, each study building upon the previous with more sophisticated methods and diverse datasets. Early studies typically focused on traditional statistical models like logistic regression to evaluate risk factors known to be associated with stroke, such as hypertension and diabetes. For example, a landmark study by Sacco et al. (1997) utilized logistic regression to identify the relative risks associated with various clinical and lifestyle factors in stroke occurrences.

More recent research has shifted towards employing advanced machine learning algorithms that can handle complex, non-linear interactions between a broader range of variables. Studies such as those by Choi et al. (2020) explored the use of Random Forests and Gradient Boosting Machines, noting their superior performance over simpler models due to their ability to capture more intricate patterns in the data. These studies highlight the advantages of ensemble methods that combine multiple models to improve prediction accuracy.

Deep learning has also made significant inroads into stroke prediction, with neural networks being used to analyze not only structured data but also unstructured data such as brain imaging. Research by Jin et al. (2018) demonstrated how convolutional neural networks could be applied to MRI scans to predict stroke outcomes and recovery potential, showcasing the breadth of ML applications in stroke research.

Comparative studies have also been informative in this field. For instance, the work by Ahmad et al. (2019) compared several machine learning models on the same stroke prediction dataset, providing valuable insights into the trade-offs between model complexity and interpretability. Such comparisons are crucial for healthcare practitioners who need models that not only predict accurately but are also understandable and actionable in clinical settings.

Additionally, the integration of big data analytics in stroke research has opened new avenues for exploring predictive factors that were previously unattainable. Big data studies often incorporate real-time health monitoring and longitudinal patient records to continuously update and refine stroke prediction models, a method supported by the growing prevalence of electronic health records and health informatics.

The existing literature not only underscores the progression towards more complex algorithms but also illustrates a growing consensus on the importance of model interpretability and the need for models that can be practically implemented within the healthcare industry. This project draws upon these studies, aiming to synthesize these advancements with new approaches and methodologies to push the boundaries of stroke prediction further.

Dataset and Features

The Stroke Prediction Dataset at the center of this study comprises data from 5,110 patients, meticulously collected to foster the development of accurate predictive models. Each patient record is an amalgamation of several key features, carefully selected based on their established correlation with stroke incidents in prior clinical studies. The dataset includes demographic information, medical history, and lifestyle indicators, all of which are critical in assessing stroke risk.

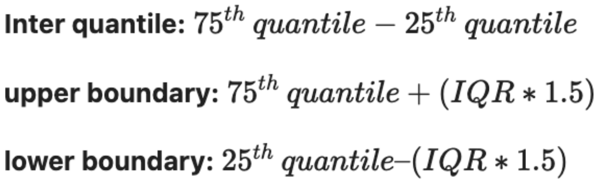
Key Features of the Dataset:

1. **Gender**: Classified into Male, Female, and Other, recognizing the role of gender-specific risk factors in stroke incidences.
2. **Age**: Continuous variable; as age increases, so does the risk of stroke, making this a primary predictor in our analysis.
3. **Hypertension**: Binary indicator (0 for no, 1 for yes); hypertension is a well-documented risk factor for stroke.
4. **Heart Disease**: Binary indicator; existing heart conditions are crucial predictors due to their direct impact on cardiovascular health.
5. **Ever Married**: Categorical variable (Yes or No); included to investigate the social determinants of health which may influence lifestyle choices.
6. **Work Type**: Includes categories like Government job, Never worked, Private, Self-employed, and Children; this reflects the socio-economic status and physical activity levels associated with different types of employment.
7. **Residence Type**: Rural or Urban; environmental and lifestyle differences between these areas can affect health outcomes.
8. **Average Glucose Level**: Continuous variable; high glucose levels are a risk factor for many cardiovascular diseases, including stroke.
9. **BMI (Body Mass Index)**: Continuous variable; obesity is a known risk factor for stroke and many other health conditions.
10. **Smoking Status**: Categorized into formerly smoked, never smoked, smokes, or Unknown; smoking is a significant risk factor, and capturing this behavior is vital for accurate risk assessment.
11. **Stroke**: Binary outcome variable (1 if the patient had a stroke, 0 otherwise); this is the dependent variable our models will predict.

Data Preprocessing: The data underwent rigorous preprocessing to ensure the quality and consistency necessary for effective model training:

- **Handling Missing Values in BMI**:
    - Missing values in the BMI (Body Mass Index) column were identified.
    - These missing values were filled using the median value of the BMI column, calculated
- **Correction of Values in the Gender Column:**
    - There was a category labeled "Other" in the Gender column.
    - This category was replaced with the more common category "Female".

- **Handling Outliers in the BMI and Average Glucose Level Columns:**
    - Outliers were detected in the BMI and Average Glucose Level columns.
    - Some of these outliers were identified using a box plot.
    - The outliers were removed from the dataset.
    - While the specific formulas used for outlier detection and removal are not provided, the Interquartile Range (IQR) method is commonly used for this purpose. This method involves calculating the IQR of a dataset and removing data points that fall below Q1−1.5 × IQRQ1−1.5 × IQR or above Q3+1.5 × IQR Q3+1.5 ×I QR, where Q1 and Q3 are the first and third quartiles, respectively.



Inter quantile: $75^{th}\ quantile - 25^{th}\ quantile$

upper boundary: $75^{th}\ quantile + (IQR * 1.5)$

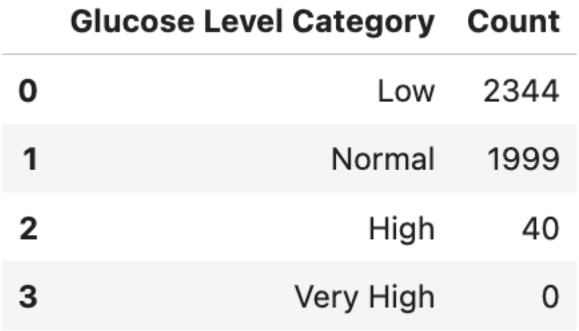
lower boundary: $25^{th}\ quantile–(IQR * 1.5)$

Feature Engineering: Further, feature engineering was conducted to extract additional value from the data. For example, age was categorized into groups to capture non-linear effects on stroke risk, and interactions between hypertension and heart disease were created to explore synergistic effects.

This rich dataset provides a robust foundation for developing machine learning models capable of identifying patterns and interactions that may not be apparent through conventional analysis, allowing for more nuanced insights into stroke risk factors and prediction.

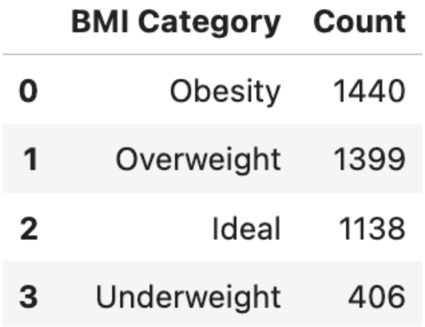
Both Categorical and numerical features are present:

- Categorical Features: gender, ever_married, work_type, Residence_type, smoking_status
- Binary Numerical Features: hypertension,heart_disease, stroke
- Continous Numerical Features: age, avg_glucose_level, bmi



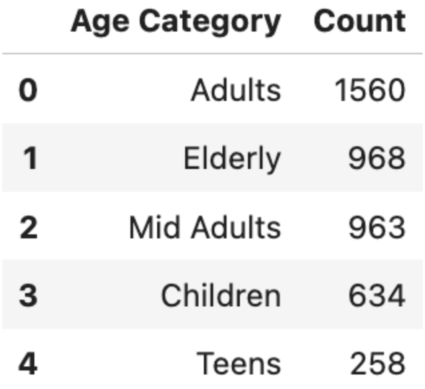
	Glucose Level Category	Count
0	Low	2344
1	Normal	1999
2	High	40
3	Very High	0

Table 1 : Glucose Level Category



	BMI Category	Count
0	Obesity	1440
1	Overweight	1399
2	Ideal	1138
3	Underweight	406

Table 2: BMI Category



	Age Category	Count
0	Adults	1560
1	Elderly	968
2	Mid Adults	963
3	Children	634
4	Teens	258

Table 3: Age Category

Planned Next Steps

- **Model Development and Tuning**: Train and fine-tune the Logistic Regression and Random Forest models.
- **Model Evaluation:** Use metrics such as accuracy, sensitivity, specificity, F1-score, and ROC-AUC.

Contributions

- Sevde, Muhammed: Outlier Removal
- All members of the group: The Encoding Process
- Nagihan: Finding distributions
- Ali: Creating a correlation matrix

REFERENCES

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
https://www.medicalnewstoday.com/articles/323446#body-mass-index-bmi
https://kidspicturedictionary.com/english-through-pictures/people-english-through-pictures/age-physical-description/
https://agamatrix.com/blog/normal-blood-sugar-level-chart/
https://www.researchgate.net/publication/380424935_Stroke_Prediction_and_Contributing_Factors_Using_Machine_Learning
https://www.researchgate.net/publication/379837937_Effective_stroke_prediction_using_machine_learning_algorithms