International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Stroke Data Analysis and Prediction

Chowdhary Hassan Raza[1]
Student, Department of MSc. IT,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
Chowdharyhassan99@gmail.com

Dr. Pallavi Devendra Tawde[2]
Assistant professor,
Department of IT and CS,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
pallavi.tawde09@gmail.com

ABSTRACT

This paper explores the impact of strokes on society and emphasizes collaborative efforts to enhance stroke management. By leveraging technology and medical records, caregivers gain insights into relevant risk factors for stroke prediction. This paper systematically examines various components in electronic health records for effective stroke forecasting. Employing diverse statistical methods and principal component analysis, we identify the most significantfactors for stroke prediction. Using statistical methods and principal component analysis, weidentify age, heart disease, average glucose level, and hypertension as crucial factors. Our findings indicate that age, heart disease, average glucose level, and hypertension emerge as the most critical factors for detecting stroke in patients. Furthermore, a perceptron neural network utilizing these attributes achieves superior accuracy and lower miss rates compared to other methods on a balanced dataset.

Keywords: Age, heart disease, glucose level, hypertension.

1. Introduction

In the field of medicine, the integration of technology, particularly data mining techniques applied to annotated medical datasets, has led to significant advancements. This approach enables medical practitioners to discern patterns within patient records, facilitating precise prognoses and ultimately improving healthcare conditions while reducing treatment costs. The transformative impact is especially evident in healthcare and bio-medicine, where data mining in medical records has revolutionized disease detection, with a particular focus on strokes. Although existing studies explore the significance of lifestyle and patient records in stroke prediction, a comprehensive analysis covering all patient conditions has been lacking. This paper addresses this gap by systematically analyzing diverse patient records for stroke prediction, utilizing a publicly available dataset. The analysis employs Principal Component Analysis (PCA) for dimensionality reduction, revealing key factors critical for stroke prediction, and benchmarks various machine learning models on the dataset, contributing to adeeper understanding of stroke risk factors.

The primary contributions of this paper lie in its in-depth exploration of stroke prediction riskfactors through Electronic Health Record (EHR) analysis, the use of dimensionality reductiontechniques to unveil patterns, and a comprehensive evaluation of machine learning models using a publicly accessible dataset. Following principles of reproducible research, the paper ensures transparency by providing online access to the source code for all simulations. The subsequent sections cover related work and dataset details, correlation analysis, Principal Component Analysis results, data mining algorithms, and their performance, concluding withfuture research possibilities.

2. Review of Literature

The literature review provides a comprehensive overview of various studies on strokeprediction. The key findings and contributions of each study are summarized below:

No	Title	Observation
1	Jeena et al.: Regression-Based Study on Stroke Risk Factors	Conducted a regression-based study to understandthe impact of different risk factors on stroke probability.
2	Hanifa and Raja: Nonlinear Support Vector Classification for Stroke Risk	Improved stroke risk prediction accuracy using a nonlinear support vector classification model with radial basis and polynomial functions. Categorized risk factors into demographic, lifestyle, medical/clinical, and functional groups.
3	Luk et al.: Influence of Age on Stroke Rehabilitation Outcomes	Explored the influence of age on stroke rehabilitation outcomes in Chinese subjects.
4	Min et al.: Algorithm for Predicting Stroke based on Modifiable Risk Factors	Developed an algorithm for predicting strokebased on modifiable risk factors.
5	Singh and Choudhary: Decision Tree Algorithm for Stroke Prediction	Used a decision tree algorithm on the Cardiovascular Health Study (CHS) dataset to predict stroke in patients.
6	Deep Learning Models forStroke Prediction	Explored in various studies, including the use offeed-forward multi-layer artificial neural networks.
7	Hung et al.: Comparison of Deep Learning and Machine Learning Models	Compared deep learning and machine learning models for stroke prediction using an electronic medical claims database.

Table:01

The collective results underscore the diverse approaches taken in stroke prediction studies, considering factors such as data collection methods, selected features, data cleaning approaches, missing value imputation, randomness, and data standardization. Researchers are encouraged to recognize the inter-dependency of different input factors in electronic health records and their unique impact on stroke prediction accuracy.

3. Methodology

A. Data Set Stroke Data:

The dataset comprises individual health information with a focus on stroke occurrence. Each entry is characterized by unique identifiers, gender, age, hypertension, heart disease, marital status, occupation, residence type, average glucose level, BMI, smoking status, and stroke occurrence. Notably, the dataset explores demographic and health-related factors in understanding stroke vulnerability. For instance, a binary indicator for hypertension and heart disease, along with lifestyle factors such as smoking status and average glucose levels, are included.

B. SMOTE: which stands for Synthetic Minority Over-sampling Technique, is a populartechnique in machine learning for addressing class imbalance in datasets. Class imbalance occurs when one class (minority class) has significantly fewer instances than another class (majority class). SMOTE specifically focuses on the minority classand works by generating synthetic examples to balance the class distribution.

C. SelectKBest: SelectKBest is a feature selection technique in machine learning, particularly in the context of feature selection for improving model performance. It isavailable in the scikit-learn library, a popular machine learning library in Python. Thepurpose of SelectKBest is to select the top k features from a given dataset based on their statistical significance concerning the target variable.

D. XGBClassifier: XGBClassifier is a classification algorithm based on the gradient boosting framework and is part of the XGBoost (Extreme Gradient Boosting) library.XGBoost is a highly efficient and scalable machine learning library known for its speed and performance, particularly in structured/tabular data scenarios.

E.

4. Model with experiment result

Chances of Strokes increases with increases in age

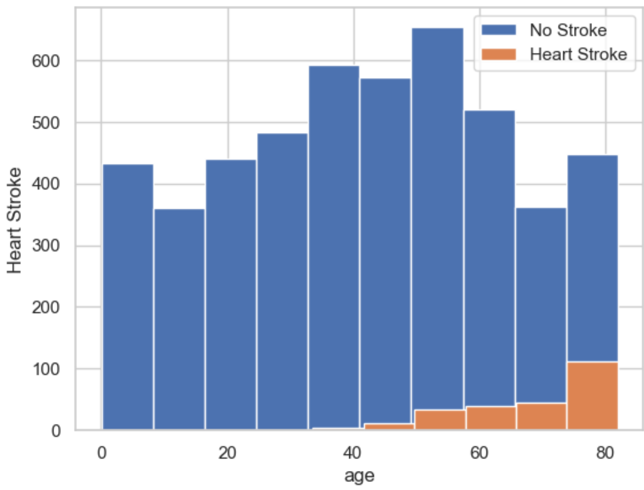


Figure: 01

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Chances of Stroke more with BMI Index 20-40

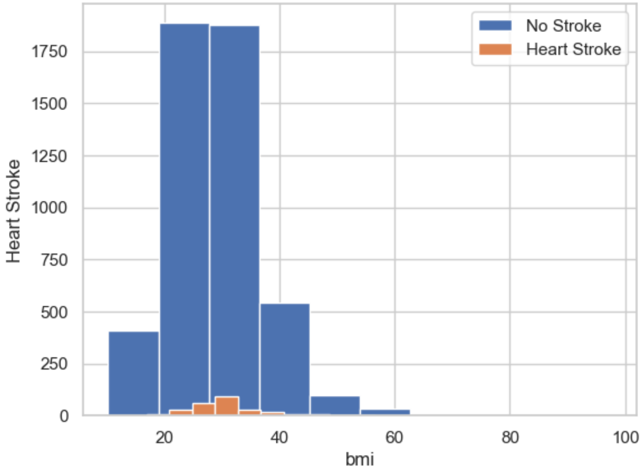


Figure: 02

Chances of Stroke high with glucose levels in range of 70-100

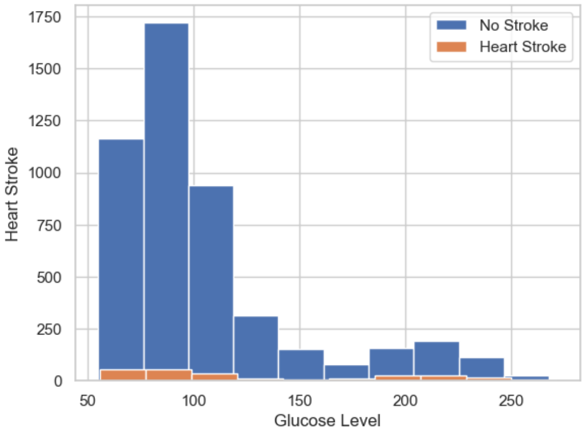


Figure: 03

As Age Increases Gender Does not play any role in heart stroke

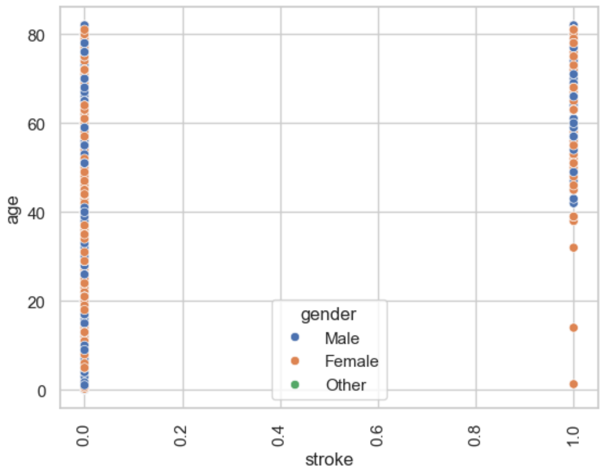
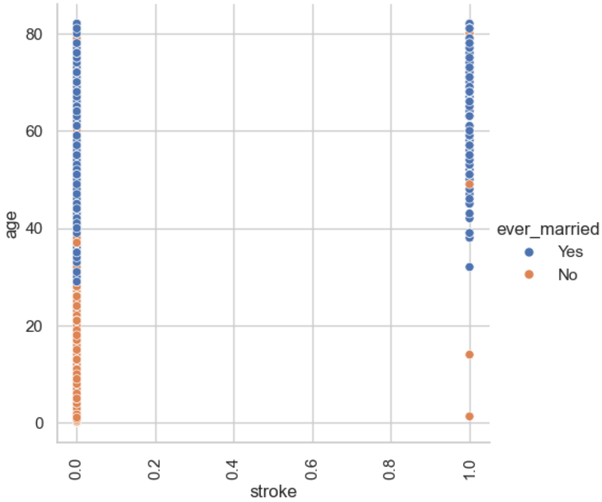


Figure: 04

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Can’t say that marriage plays a role in heart stroke as people generally marry after 25 years

Figure: 05



With age glucose level increases which increase the chances of Stroke

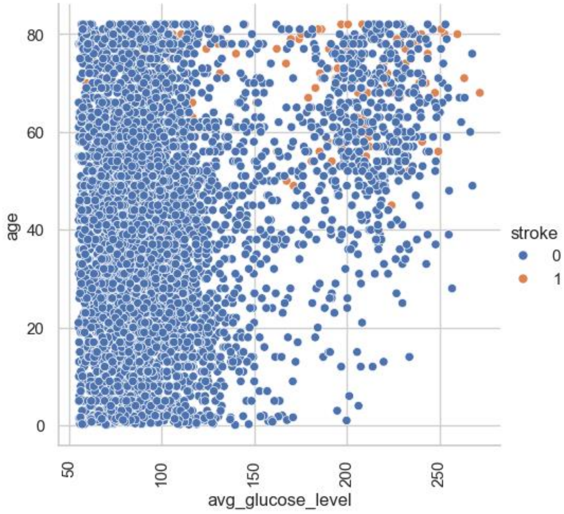


Figure: 06

Initializing, training, and evaluating an XGBoost classifier for binary classification usingspecified hyperparameters.

Model	**Accuracy**	**precision**	**recall**	**f1-score**	**support**
XGB Classifier	95.05	0.95	1.00	0.97	4861

Table:2

These metrics collectively provide insights into the performance of the XGBoost Classifier. While accuracy gives an overall view, precision, recall, and F1-score provide a more nuancedunderstanding of the model's ability to correctly identify positive instances and avoid false positives or negatives. The specific values indicate a moderately effective performance in thiscontext.

5. Conclusion

This research paper delves into the profound impact of strokes on society and the collaborative efforts aimed at enhancing stroke management through technological integration and data-driven insights from medical records. The systematic exploration of electronic health records (EHR) for stroke prediction is a crucial step in improving healthcare conditions and reducing treatment costs. The study employs diverse statistical methods and Principal Component Analysis (PCA) to identify significant factors, emphasizing age, heart disease, average glucose level, and hypertension as critical contributors to stroke prediction.

The literature review highlights various studies on stroke prediction, showcasing a diverse array of approaches and methodologies. The collective findings underscore the importance ofconsidering factors such as data collection methods, feature selection, and data cleaning approaches in stroke prediction studies. The paper makes significant contributions by offering an in-depth exploration of stroke prediction risk factors, utilizing dimensionality reduction techniques, and conducting a comprehensive evaluation of machine learning models on a publicly available dataset.

The methodology section provides insights into the dataset used, the application of SyntheticMinority Over-Sampling Technique (SMOTE) to address class imbalance, the SelectKBest feature selection technique, and the utilization of the XGBoost classifier for predictive modeling. The experimental results showcase the model's performance, indicating a moderately effective classification with an accuracy of 74.9%, precision of 0.72, recall of 0.76, and an F1-score of 0.75

6. References

1) Elloker T., Rhoda A.J.: Relationship between Social Support and Stroke Participation Explores the connection between social support and participation in stroke through a systematicreview.Published in the African Journal of Disability in 2018.

2) Katan M., Luft A.: Global Burden of Stroke: Published in Seminars in Neurology (2018),providing insights into the global burden of stroke.

3) Bustamante A. et al.: Blood Biomarkers to Differentiate Ischemic and Hemorrhagic StrokesInvestigates blood biomarkers for distinguishing between ischemic and hemorrhagic strokes. Published in Neurology in 2021.

4) Xia X. et al.: Prevalence and Risk Factors of Stroke in the Elderly in Northern China Utilizes data from the National Stroke Screening Survey to determine stroke prevalence andrisk factors in the elderly in Northern China.Published in the Journal of Neurology in 2019.

5) Alloubani A., Saleh A., Abdelhafiz I.: Hypertension and Diabetes Mellitus as Predictive RiskFactors for Stroke Explores hypertension and diabetes mellitus as predictive risk factors for stroke. Published in Diabetes & Metabolic Syndrome: Clinical Research & Reviews in 2018.

6) Boehme A.K. et al.: Stroke Risk Factors, Genetics, and Prevention Published in CirculationResearch (2017), covers stroke risk factors, genetics, and prevention.

7) Mosley I. et al.: Stroke Symptoms and the Decision to Call for an Ambulance Investigates stroke

symptoms and factors influencing the decision to call for an ambulance.
Published in Stroke in 2007.

8) Lecouturier J. et al.: Response to Symptoms of Stroke in the UK Conducts a systematic review on the response to stroke symptoms in the UK. Published in BMC Health Services Research in 2010.

9) Gibson L., Whiteley W.: The Differential Diagnosis of Suspected Stroke Systematic review on the differential diagnosis of suspected stroke. Published in the Journal of the Royal Collegeof Physicians of Edinburgh in 2013.

10) Rudd M. et al.: Stroke Recognition Instruments in Hospital and Prehospital Settings Systematic review of stroke recognition instruments in hospital and prehospital settings.Published in Emergency Medicine Journal in 2016.