Predictive Modeling for Stroke Risk Identification: A Comparative Study of Classification Algorithms

Omkar Nagarkar

Abstract

This research explores the application of various classification algorithms to develop a predictive model for identifying individuals at risk of having a stroke. The study follows the Knowledge Discovery in Databases (KDD) methodology, encompassing data understanding, preparation, modeling, evaluation, and deployment. The models' performances are compared based on accuracy, precision, recall, F1 score, and ROC-AUC score. The findings reveal key charac- teristics associated with a higher stroke risk and highlight the potential of predictive modeling in healthcare.

1 Introduction

Stroke is a leading cause of death and disability worldwide. Early identification of individuals at risk is crucial for timely intervention and prevention. This study aims to develop a predictive model utilizing various classification algorithms and compare their performances in classifying stroke risk based on medical and demographic features.

2 Methodology

The study follows the KDD methodology, com-prising the following phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

3 Data Understanding andPreparation

The dataset includes several medical predictor variables and one target variable indicating stroke occurrences. The dataset exhibited class imbalance, addressed through oversampling of the minority class. Data preprocessing involved handling missing values, encoding categorical variables, and scaling numerical features.

4 Modeling

Several classification models were trained and evaluated, including:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)
- Gradient Boosting Classifier

5 Results

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC Score
Logistic Regression	79.13%	76.60%	83.85%	80.06%	79.13%
Decision Tree	97.74%	95.67%	100.00%	97.79%	97.74%
Random Forest	99.07%	98.18%	100.00%	99.08%	99.07%
SVM	84.88%	79.95%	93.11%	86.03%	84.89%
Gradient Boosting	86.22%	81.83%	93.11%	87.10%	86.22%

Table 1: Performance Metrics Comparison of Classification Models

6 Discussion

The Random Forest Classifier emerged as the best-performing model, demonstrating near-perfect scores across all evaluation metrics. Through the analysis of feature importance and data exploration, the study identified several key characteristics associated with a higher risk of having a stroke:

- **Age**: Older individuals were observed to be at a significantly higher risk.
- **Hypertension**: Individuals with a history of hypertension exhibited an increased risk.
- **Heart Disease**: The presence of heart disease was identified as a significant risk factor.
- **Married Status**: Being married was associated with a higher risk.
- **Work Type - Self-Employed**: Individuals who were self-employed faced a higher risk.
- **Residence Type - Urban**: Urban residents showed a slightly higher risk.
- **Average Glucose Level**: Elevated average glucose levels were associated with an increased risk.
- **BMI**: Individuals with higher body mass index (BMI) values faced increased risk.
- **Smoking Status**: Former smokers and current smokers were identified as higher-risk groups.

7 Conclusion

This study underscores the significance of predictive modeling in healthcare for early risk identification, enabling preventive measures and improved health outcomes. The Random Forest model, due to its superior performance, is recommended for deployment in real-world applications for stroke risk classification.

8 Future Work

- Exploration of additional features and data sources to enhance model performance.
- Implementation of model fine-tuning and hyperparameter optimization.
- Real-world deployment and continuous monitoring of the model.
- Exploration of deep learning approaches for further improvement in predictive accuracy.

9 References

1. Breiman, L. (2001). Random Forests. Ma-chine Learning, 45(1), 5-32.
2. Cortes, C., & Vapnik, V. (1995). Support- Vector Networks. Machine Learning, 20(3), 273-297.
3. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Ma- chine. Annals of Statistics, 29(5), 1189- 1232.

4. Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. Wiley-Interscience.
5. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106.