PROJECT REPORT INTRODUCTION TO DATA SCIENCE FALL 2024

Heart Stroke Prediction

Urwa Khalid[1], Areeba Noor[1], and Mariyam[1]

[1]Department of Computer Science, National University of Computer and Emerging Sciences, Lahore

November 27, 2024

Abstract

Heart strokes are among the leading causes of mortality globally, responsible for one death every minute. With modern healthcare facing the challenge of processing vast amounts of data, machine learning (ML) provides an effective means to automate and improve early prediction. This study uses three machine learning algorithms—Random Forest, Decision Trees, and Support Vector Machines (SVM)—to predict heart stroke risk based on patient health data.

The dataset includes demographic, health, and lifestyle features and was subjected to extensive preprocessing, including outlier removal and feature encoding. Each model's performance was evaluated on metrics like accuracy, precision, recall, and F1-score. Results showed that the Random Forest algorithm performed the best, achieving an accuracy of 95% and an F1-score of 0.96. This study highlights the potential of ML in automating heart stroke prediction, providing tools for early detection and improved patient care.

1). Introduction

Heart strokes occur when the heart's blood supply is disrupted, causing cell death and often leading to severe health consequences or death. According to the World Health Organization, heart strokes are one of the top causes of death globally causing 11% of deaths, claiming millions of lives annually. Identifying individuals at risk early can significantly reduce mortality and morbidity by allowing timely interventions.

Machine learning has become an essential tool in healthcare for its ability to analyze large datasets and uncover hidden patterns. This study aims to develop an automated system for predicting heart strokes using ML algorithms. By analyzing patient data such as age, blood glucose levels, hypertension status, and lifestyle habits, we trained and evaluated three ML models: Random Forest, Decision Trees, and SVM. The ultimate goal was to identify the most effective algorithm for heart stroke prediction and provide actionable insights for clinicians.

2).Literature Survey

Several studies have applied machine learning to heart stroke prediction:

- Govindarajan et al. used text mining and ML classifiers to classify heart stroke diseases, achieving 95% accuracy using SGD.
- Amini et al. classified 50 risk factors for stroke using the C4.5 Decision Tree and KNN, reporting accuracies of 95% and 94%, respectively.
- Cheng et al. analyzed stroke data using ANN models, achieving accuracies of 79% and 95%.
- Cheon et al. employed deep neural networks and PCA on 15,099 patient records to predict stroke outcomes, achieving an AUC of 83%.
- Singh et al. used neural networks with the CHS dataset for stroke prediction, achieving 97%

accuracy.

These studies underline the efficacy of ML in stroke prediction and set the groundwork for our approach.

3). Methodology and Experimentation

Dataset and Preprocessing

The dataset contains 5,110 entries with 12 attributes, including demographic, health, and lifestyle factors. Preprocessing steps included:

1. Data Cleaning:
    - Outliers in `avg_glucose_level` and `bmi` were identified and removed using IQR, reducing the dataset to 4,391 records.
    - Missing values in `bmi` were imputed.
2. Exploratory Data Analysis (EDA):
    - Relationships between variables were analyzed through visualization techniques to identify trends and key predictors.
    - Features such as age, hypertension, and average glucose levels were found to have good correlations with the target variable (`stroke`).
3. Feature Engineering:
    - Categorical variables (e.g., `gender`, `work_type`) were encoded using one-hot encoding.
    - Continuous variables were normalized to improve model convergence.

Model Training

Three ML models were trained on both the original cleaned (unsampled) data and SMOTE-balanced data to address class imbalance:

- Random Forest: Combines multiple decision trees to enhance accuracy and reduce overfitting.
- Decision Trees: Splits data into branches based on feature thresholds.
- SVM: Identifies the optimal hyperplane to separate stroke and non-stroke cases.

SMOTE was applied with random seeds of 10 and 42 to generate balanced datasets. Evaluation metrics included accuracy, precision, recall, F1-score, log loss, and ROC AUC.

4). Results & Discussion

i).Comparison of Results

The models were evaluated under three conditions: unsampled data, SMOTE (random_state=10), and SMOTE (random_state=42).

Without Sampling:

- Random Forest achieved the highest accuracy (96%) but failed to predict the minority class effectively (F1-score: 0.0).
- Decision Trees and SVM struggled with similar issues, highlighting the need for sampling techniques.

With SMOTE (random_state=10):

- Random Forest improved significantly, with an accuracy of 94.2% and an F1-score of 0.94.
- Decision Trees and SVM also performed better but remained less effective than Random Forest.

With SMOTE (random_state=42):

- Random Forest delivered the best overall performance:
  - Accuracy: 95.8%
  - Precision: 93.9%
  - Recall: 97.9%
  - F1-Score: 0.96
  - ROC AUC: 0.991
- Decision Trees and SVM also showed improvements, but Random Forest consistently outperformed them.

Feature Importance and Insights

1. Age, hypertension, and avg_glucose_level were the most significant predictors of heart stroke risk.
2. SMOTE effectively addressed the class imbalance, allowing all models to better predict minority cases.

Sampling Method	**Model**	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**Log Loss**	**ROC AUC**
Without Sampling	Random Forest	96.05	0.000	0.000	0.000	0.263	0.781
	Decision Tree	92.94	0.113	0.115	0.114	2.543	0.539
	SVM	96.05	0.000	0.000	0.000	0.164	0.604
SMOTE (random_state=10)	Random Forest	94.28	0.914	0.977	0.945	0.192	0.985
	Decision Tree	89.04	0.866	0.924	0.894	3.951	0.890
	SVM	87.46	0.828	0.946	0.883	0.280	0.949
SMOTE (random_state=42)	Random Forest	**95.82**	**0.939**	**0.979**	**0.959**	0.159	**0.991**
	Decision Tree	92.07	0.912	0.931	0.921	2.857	0.921
	SVM	89.71	0.851	0.962	0.903	0.246	0.961

ii).Challenges

1. Imbalanced Dataset:
   - The original dataset's class imbalance limited model performance on minority cases.
2. Data Quality:
   - Ambiguity in features like `smoking_status` introduced noise into the models.

5).Conclusion

This study demonstrated the application of machine learning in predicting heart stroke risk, with Random Forest emerging as the best-performing model. Using SMOTE (random_state=42), Random Forest achieved an accuracy of 95.8% and an F1-score of 0.96, highlighting its robustness and ability to handle imbalanced data effectively.

Future Directions:

1. Enhancing Data:
   - Include additional features such as cholesterol levels, family history, and physical activity for improved predictions.
   - Expand the dataset for better generalization.
2. Algorithm Exploration:
   - Investigate advanced models like Gradient Boosting or deep learning architectures for enhanced accuracy.
3. Real-World Applications:
   - Develop an interactive tool integrating these models to assist healthcare providers in real-time heart stroke risk assessment.

This work contributes to the growing body of research on ML-driven healthcare solutions, offering a promising approach to early heart stroke detection and prevention.