Predicting Stroke Risk with Machine Learning Using Health and Lifestyle Data

Contents

[Abstract 2](#_Toc183283808)

[Acknowledgements 2](#_Toc183283809)

[Chapter 1: Introduction 2](#_Toc183283810)

[I. Overview 2](#_Toc183283811)

[II. Problem Statement 2](#_Toc183283812)

[III. Research Questions 2](#_Toc183283813)

[Chapter 2: Background 2](#_Toc183283814)

[I. Dataset 2](#_Toc183283815)

[II. Literature Review 2](#_Toc183283816)

[III. Research Gap 2](#_Toc183283817)

[IV. Algorithms 2](#_Toc183283818)

[a. TFIDF 2](#_Toc183283819)

[b. Count Vectorizer 2](#_Toc183283820)

[c. Logistic Regression 2](#_Toc183283821)

[d. SVM 2](#_Toc183283822)

[e. XGBoost 2](#_Toc183283823)

[f. Decision Tree 2](#_Toc183283824)

[g. Artificial Neural Network 2](#_Toc183283825)

[Chapter 3: Methodology 2](#_Toc183283826)

[I. Tools and Techniques 2](#_Toc183283827)

[II. EDA and Visualization 2](#_Toc183283828)

[Chapter 4: Results and Conclusion 2](#_Toc183283829)

[I. Critical Discussion 2](#_Toc183283830)

[II. Conclusion 2](#_Toc183283831)

[III. Future Work 2](#_Toc183283832)

[Chapter 5: Legal, Ethical and Professional Issues 2](#_Toc183283833)

[References 2](#_Toc183283834)

[Appendices 2](#_Toc183283835)

# Abstract

# Acknowledgements

# Chapter 1: Introduction

## Overview

A stroke is also one of the leading causes of death and a leading cause of long-term disability worldwide (*Stroke statistics*, no date). Knowledge of such risk factors as high pressure, heart issues, and behaviours are important so that they can be addressed early enough (Holland, 2018). Machine learning thus offers a sensible means of processing big medical data and forecasting the propensity for a stroke, which may help with prevention. This project will use patient profile characteristics, genetic and clinical parameters to generate accurate prediction models and explore the effects of hypertension and heart disease on stroke chances. Thirdly, it will establish whether experiencing both significantly raises the likelihood of stroke, which shall be useful in determining appropriate future treatments.

## Problem Statement

## Research Questions

1. How significantly do hypertension and heart disease influence stroke prediction?
2. Do patients with both hypertension and heart disease have a higher likelihood of stroke compared to those with only one or neither condition?

## Objectives

* To assess the effect of health attributes (like hypertension and heart disease) in the prediction of stroke using the several learning models; Logistic regression, Decision tree, Random Forest, XGBoost and Artificial Neural Networks.
* To evaluate the degree of significance of each variable that has been selected, statistical models including feature importance, ANOVA, and correlation analysis will be used.
* The goal of this system would be to create a preprocessing pipeline and clean the dataset in addition to creating features and training models.
* To make predictions with high accuracy using metrics Such as, Accuracy, F1-Score, Precision & Recall.

# Chapter 2: Background

## Literature Review

## Research Gap

## Algorithms

### Logistic Regression

### Decision Tree

### Random Forest

### XGBoost

### Neural Network

# Chapter 3: Methodology

## Dataset

The data in the dataset covers the patient features include the patient’s age, gender, the diseases that have ever had, hypertension, heart disease, smoking, the type of work, etc. This resulted in 12 columns as well as 5110 records (“Stroke Prediction Dataset,” 2021). The dependent variable is ‘stroke,’ where 1 represents a patient diagnosed of having a stroke and 0 otherwise.

## Tools and Techniques

## EDA and Visualization

# Chapter 4: Results and Conclusion

## Critical Discussion

## Conclusion

## Future Work

# Chapter 5: Legal, Ethical and Professional Issues

# References

# Appendices