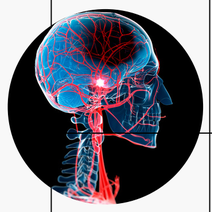
PREDICTING THE RISK OF GETTING A STROKE USING A MACHINE LEARNING TOOl

MADE BY MARK AGYAPONG



INTRODUCTION

Over the years, there have been many ways used in predicting different health outcomes. The introduction of machine learning in healthcare provides a more accurate means of measuring these outcomes using artificial intelligence. Stroke is a disease that has constantly caused a lot of deaths and chronic illness causing governments billions of dollars. For this reason, there have been numerous studies done in predicting stroke. The main goal of the research is to explore and predict stroke occurrences in the dataset using machine learning tool Python. Python, which is a collection of machine learning algorithms is a very efficient tool that is used for data mining activities (Coursera, 2022). It is important to note that the dataset being used has previously been analyzed using machine learning by several researchers.

OBJECTIVE

The main goal of the research is to predict the risk of a person getting a stroke using a machine learning tool. Using different portions and attributes of the dataset under study, this will train them to determine which of the attributes are strongly associated with an increased risk of getting a stroke. These will be done by classifying and clustering different aspects of the dataset sets and also most importantly using graphical representations to have a clear overview of both increased and decreased risks. Therefore, the project objectives formulated are:

- To determine the risk factors associated with stroke using Python algorithms.
- Compare the accuracy of risks level with or without feature selection using the attributes.

METHODOLOGY

The chosen programming language for this thesis is Python. The goal is to utilize machine learning algorithms to predict the likelihood of stroke occurrence. The predicted outcomes will be visually presented in graphs to facilitate clear interpretation. The machine learning that will be used in thesis will be:

- Decision tree
- Naive bayes
- Linear regression
- Support Vector Machine

Dataset on stroke information was taken from Kaggle, a specialized online platform that is built for data scientists and machine learning lovers. The dataset used contains 5110 rows with 12 attributes.

- Patient ID
- Gender
- Age
- Hypertenstion
- Heart disease
- Marital status
- Work type
- Residence type

DATA PREPROCESSING

The process of converting data to a common format to enable users to process and analyze it.

- Data cleaning - Removing all rows with missing values
- Data standardization - Converting integer values to categorical or binanary
- Data catergorisation - Converting continuous into categorical data

Image 1. *Raw data of stroke dataset*

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1

Image 2. *After data preprocessing*

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	1	3	1	228.69	36.6	3	1
1	51676	Female	61.0	0	0	1	3	0	202.21	NaN	2	1
2	31112	Male	80.0	0	1	1	3	0	105.92	32.5	2	1
3	60182	Female	49.0	0	0	1	3	1	171.23	34.4	1	1

RESEARCH / FINDINGS

This section presents the results of the exploratory data analysis (EDA). The algorithms will be used after the EDA.

CORRELATION MATRIX

A correlation matrix created to show the correrlation coeffiectuens between attributes. This will idenify any strong correlated attributes. .

AGE

The correlation matrix shows a strong 23% correlation between age and stroke, suggesting that older people are more likely to experience a stroke.

HYPERTENSION

The correlation matrix shows an 14% correlation between hypertension and stroke, indicating that hypertension plays a role in increasing the risk of stroke.

HEART DISEASE

The correlation matrix indicates an 14% correlation between heart disease and stroke, suggesting that heart disease is a factor that may increase the risk of stroke.

CONCLUSION

The main goal of this research is to predict the risk of stroke using Python algorithms. The dataset contains 12 attributes that underwent preprocessing steps, including data cleaning, standardization, and categorization, before conducting exploratory data analysis (EDA). The EDA revealed age, hypertension, and heart disease as the features with the strongest correlation to stroke. The final phase of the analysis involves training and testing the Python algorithm to predict stroke risk.

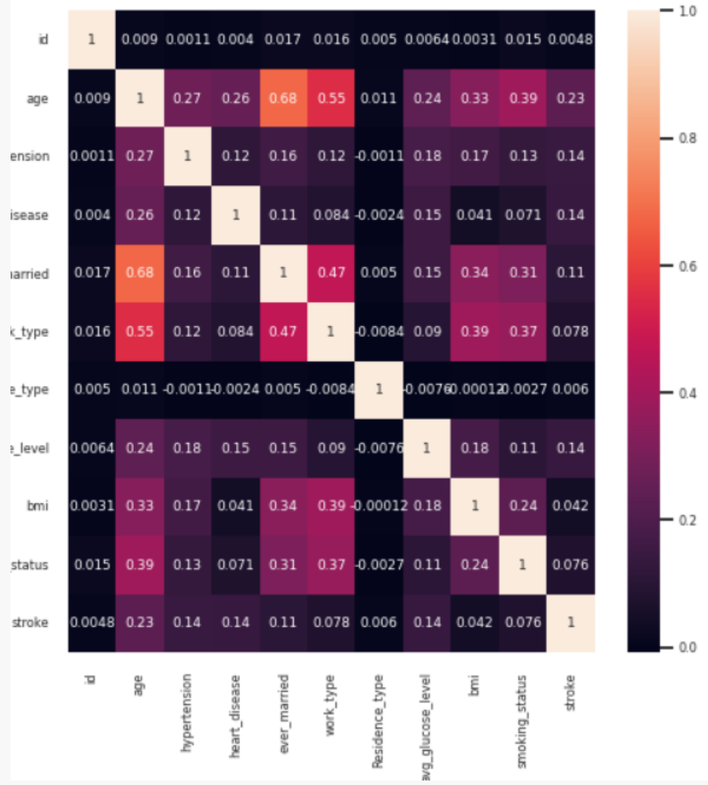


Image 3. *Correlation matrix of stroke dataset*