ISSN 2722-2039

Machine Learning-Based Non-Communicable Disease Prediction Evaluating the Impact of Hypertension, Diabetes, and Lifestyle Factors on Stroke Risk

Mathew Darell Widodo [a,1,*], Pita Melati Sulkani [b], Ariyono Setiawan [c], Abdul Razak Bin Abdul Hadi [d]

[a] Universitas Kristen Petra Surabaya, Indonesia
[b] Universitas Airlangga Surabaya, Indonesia
[c] Politeknik Pelayaran Surabaya, Indonesia
[d] Universiti Kuala Lumpur, Malaysia
[1] mathewdarrell05@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received March 27, 2025
Revised August 19, 2025
Accepted November 29, 2025

Keywords
Logistic regression
Risk prediction
Chronic diseases
Diabetes
Stroke
Heart disease

ABSTRACT

Chronic diseases such as diabetes, stroke, and heart disease are major challenges in the global health system. Data-driven risk prediction for this disease is important for supporting more precise and effective medical decisions. This study aims to evaluate the main factors contributing to the incidence of diabetes, stroke, and heart disease using logistic regression analysis. The data used are from health sources and includes demographic variables, lifestyle factors, and health indicators. Logistic regression was used to identify variables significantly associated with each health condition studied. The model was evaluated using p-value, regression coefficient, and confidence interval to assess the significance of risk factors. The results of the analysis showed that age, high blood pressure, cholesterol levels, and body mass index (BMI) contributed significantly to the risk of diabetes, stroke, and heart disease. Physical activity and alcohol consumption negatively affected the risk, while smoking factors did not show strong significance in the model. These findings confirm that certain lifestyle factors and health conditions significantly affect the risk of chronic disease. The implications of this research can inform data-driven prevention and early intervention strategies in the health sector.

1. Introduction

In recent decades, logistic regression analysis has become a key method in health research to predict risk factors for chronic diseases [1], [2], [3]. The use of logistic regression helps to understand the relationship between predictor variables, such as age, sex, blood pressure, and body mass index (BMI), and the likelihood of a disease [4], [5]. Given the increasing prevalence of chronic diseases across countries, further research on more accurate and efficient predictive models is urgently needed. Recent trends in health data indicate that real-time data integration, along with the use of additional variables, can improve the reliability of logistic regression models [6], [7], [8], [9].

Various studies have highlighted the effectiveness of logistic regression in health analysis. Studies published in the journal show that lifestyle factors such as smoking habits, alcohol consumption, and physical activity levels contribute to the risk of chronic disease [10], [11], [12]. However, comparisons with machine learning methods show that logistic regression, although easier to interpret, has limitations in capturing the increasing complexity of health data [10], [13], [14].

Many studies still use a limited set of predictor variables, resulting in a model that is not optimal for describing the complexity of health risk factors [15]. Most studies still rely on historical data without considering real-time data that can improve prediction accuracy [16], [17]. Some studies tend to be generalized without considering specific demographic or ethnic characteristics that may affect the results of the analysis [18], [19]. This research aims to address this gap by developing a more comprehensive logistic regression model that integrates relevant new variables and compares its performance with modern machine learning methods.

The formulation of the problem in this study includes:

1. How effective is logistic regression in predicting chronic disease risk compared to machine learning models?
2. What factors have a significant impact on improving the accuracy of health prediction models?
3. How can real-time data integration affect the results of logistic regression analysis?

The main objective of this study is to develop a more accurate prediction model by considering various health risk factors as well as to test the limitations of logistic regression compared to machine learning techniques.

Using a quantitative approach, the hypotheses that can be proposed are:

1. H_0: There is no significant difference between logistic regression and machine learning models in predicting chronic disease risk.
2. H_1: Machine learning models have a higher level of accuracy than logistic regression in predicting the risk of chronic diseases.

This research is significant both theoretically and practically. From a theoretical perspective, this research contributes to understanding the limitations of logistic regression and the potential of machine learning techniques as alternatives. From a practical perspective, healthcare practitioners can use the results of this study to improve the effectiveness of early diagnosis of chronic diseases.

In general, this study employed a quantitative method, including logistic regression and machine learning. The data used is from reliable health sources and will be analyzed to identify the most influential factors in disease prediction. The main contributions of this research are:

1. Develop a more accurate logistic regression model by taking into account additional variables.
2. Integrate real-time data to improve the accuracy of disease risk prediction.
3. Comparing the effectiveness of logistic regression with machine learning techniques in health analysis.

This literature review aims to provide an in-depth understanding of the use of logistic regression analysis in the health sector and its relevance in current research. This section is systematically organized by reviewing prior research, identifying research gaps, and synthesizing studies conducted. Research on logistic regression in health analysis has been widely conducted. Studies published in the journal show that factors such as age, gender, blood pressure, and body mass index (BMI) are significantly associated with the risk of chronic diseases, including diabetes, heart disease, and stroke [20], [21]. In addition to biological factors, lifestyle factors such as smoking habits, alcohol consumption, and physical activity levels have also been shown to contribute to an increased risk of disease [22], [23]. Comparison of statistical methods was also a concern in previous research. Several studies have compared the effectiveness of logistic regression with machine learning models, such as Random Forests and Neural Networks, for predicting chronic diseases [24]. The results show that although logistic regression is easier to interpret, machine learning-based models often provide higher prediction accuracy [17], [25], [26].

Although numerous studies have examined the use of logistic regression in health analysis, several limitations remain. Many studies still use a limited set of predictor variables, resulting in a model that is not optimal for describing the complexity of health risk factors [27], [28]. Most studies still rely on historical data without considering real-time data, which can improve prediction accuracy [21], [22], [23]. Some studies are generalized without considering specific demographic or ethnic characteristics that may affect the analysis results. [24], [25], [26]. This research aims to address this gap by developing a more comprehensive logistic regression model that integrates relevant new variables and compares its performance with modern machine learning methods [17], [29], [30]. Based on a literature review, logistic regression remains a reliable method in health risk analysis, especially in epidemiology and public health. However, there is a significant opportunity to improve the model's effectiveness by incorporating additional variables, real-time data, and more advanced computational approaches [31], [32]. Therefore, this research will focus on developing more sophisticated logistic regression models to support decision-making in health policy [33], [34].

2. Research Methodology

This study employs a quantitative approach, using logistic regression to identify factors associated with the risk of diabetes, heart disease, and stroke. The data used in this study were obtained from secondary sources, including individual health and lifestyle variables.

2.1 Research Design

This study employed an observational, cross-sectional design. Data were collected from a broad population to analyze the relationship between independent and dependent variables at a given point in time. The data used in this study come from a health dataset that includes information on individual health history, lifestyle, and sociodemographic factors. The variables analyzed included:

- Dependent Variables: Diagnosis of diabetes, heart disease, and stroke.
- Independent Variables: Age, gender, body mass index (BMI), high blood pressure (HighBP), high cholesterol levels (HighChol), physical activity, diet, alcohol consumption, smoking status, as well as socio-demographic factors such as marital status and type of employment.

2.2 Data Analysis

Data analysis is conducted using binary logistic regression to assess the relationship between independent variables and the likelihood of the disease under study. The logistic regression model is used because the dependent variable is binary (1 = disease, 0 = no disease). The analysis was carried out in three different models:

- Model 1: Prediction of diabetes based on lifestyle and health factors.
- Model 2: Prediction of heart disease based on blood pressure and cholesterol factors.
- Model 3: Stroke prediction by considering socio-demographic factors and health history.

Data processing was carried out using statistical software, and the significance of the results was assessed using a p-value < 0.05. The regression coefficient indicates the direction and magnitude of each variable's influence on the likelihood of disease.

2.3 Validity

The validity of the model is tested with several evaluation techniques, including:

- Pseudo R-squared (Nagelkerke R^2): To measure how well the model describes variations in data.
- Log-Likelihood Ratio Test: To test whether a model with independent variables is better than a model without independent variables.
- Prediction Accuracy: Uses a confusion matrix to measure the sensitivity, specificity, and overall accuracy of the model in predicting disease events.

3. Results and Discussion

3.1 Presentation of Data and Key Findings

Table 1 presents descriptive statistics of the variables used in this study. These statistics include the sample size (N), mean, standard deviation (SD), and minimum and maximum values for each variable. In the first dataset, there were 70,692 observations with various health variables. The average age of the respondents was 8.58 with a standard deviation of 2.85, indicating a fairly distributed distribution. The Body Mass Index (BMI) had a mean of 29.85 (SD = 7.11), indicating substantial variation in respondents' weight categories. Lifestyle factors such as smoking (0.47), physical activity (0.70), and heavy alcohol consumption (0.04) were also analyzed to measure their impact on chronic disease. Disease variables such as hypertension (0.56), stroke (0.06), and diabetes (0.50) showed a significant proportion in the population.

Table 1. Descriptive Statistics

Variable	N	Mean	Std Development	Minimum	Maximum
Age	70692	8.5840548	2.8521531	1.0000000	13.0000000
Gender	70692	0.4569965	0.4981508	0	1.0000000
High Chol	70692	0.5257030	0.4993424	0	1.0000000
Check Chol	70692	0.9752589	0.1553362	0	1.0000000
BMI	70692	29.8569852	7.1139539	12.0000000	98.0000000
Smokers	70692	0.4752730	0.4993917	0	1.0000000
Heart Disease or Attack	70692	0.1478102	0.3549143	0	1.0000000
Physics Activities	70692	0.7030357	0.4569239	0	1.0000000
Fruits	70692	0.6117948	0.4873451	0	1.0000000
Vegetables	70692	0.7887738	0.4081814	0	1.0000000
HvyAlcoholConsumption	70692	0.0427205	0.2022278	0	1.0000000
GenHlth	70692	2.8370820	1.1135645	1.0000000	5.0000000
Plan	70692	3.7520370	8.1556266	0	30.0000000
Physics	70692	5.8104170	10.0622605	0	30.0000000
DiffWalk	70692	0.2527302	0.4345806	0	1.0000000
Stroke	70692	0.0621711	0.2414678	0	1.0000000
High BP	70692	0.5634584	0.4959602	0	1.0000000
Diabetes	70692	0.5000000	0.5000035	0	1.0000000
Variable	**N**	**Mean**	**Std Development**	**Minimum**	**Maximum**
Gender	40907	0.5551617	0.4969539	0	1.0000000
age	40910	51.3272549	21.6239693	-9.0000000	103.0000000
Hypertension	40910	0.2138352	0.4100169	0	1.0000000
heart_disease	40910	0.1277194	0.3337812	0	1.0000000
ever_married	40910	0.8213395	0.3830724	0	1.0000000
work_type	40910	3.4611342	0.7809188	0	4.0000000
Residence_type	40910	0.5148863	0.4997845	0	1.0000000
avg_glucose_level	40910	122.0759008	57.5615312	55.1200000	271.7400000
Bmi	40910	30.4063554	6.8350723	11.5000000	92.0000000
smoking_status	40910	0.4886091	0.4998763	0	1.0000000
Stroke	40910	0.5001222	0.5000061	0	1.0000000

In the second dataset, comprising 40,910 observations, the mean age was 51.32 years (SD = 21.62). The prevalence of hypertension is 0.21, indicating that approximately 21% of respondents have high blood pressure. The mean glucose was 122.08 (SD = 57.56), and the mean BMI was 30.40. In addition, the status of smokers (0.48) and the incidence of stroke (0.50) showed the existence of risk factors that could be further analyzed.

Table 2 presents an analysis of the correlation between various variables and life expectancy. The results showed that adult mortality was strongly negatively correlated with life expectancy (-0.696), indicating that higher mortality rates were associated with shorter life expectancies. In contrast, BMI (0.572), alcohol consumption (0.409), and GDP (GDP) (0.466) have a positive correlation with life expectancy, signaling that economic and health factors contribute to the increase in life expectancy. Polio also had a positive correlation (0.468), suggesting that higher vaccination coverage was associated with increased life expectancy. Meanwhile, the level of shyness in children (Thinness 1-19) showed a negative correlation (-0.48), indicating that early childhood malnutrition is associated with reduced shyness.

Table 2. Analysis of Distribution and Relationship between Correlation Variables for the First Dataset

Variable	Life Expectancy	Adult Mortality	Alcohol	BMI	Polio	GDP	Population	Thinness 1-19 Years
Life Expectancy	1	-0.702	0.401	0.566	0.455	0.462	-0.019	-0.477
Adult Mortality	-0.702	1	-0.271	-0.46	-0.36	-0.29	0.043	0.36
Alcohol	0.401	-0.271	1	0.328	0.206	0.395	-0.07	-0.226
BMI	0.566	-0.455	0.328	1	0.25	0.415	0.019	-0.504
Polio	0.455	-0.362	0.206	0.25	1	0.384	-0.037	-0.328
GDP	0.462	-0.292	0.395	0.415	0.384	1	-0.021	-0.342
Population	-0.019	0.043	-0.07	0.019	-0.04	-0.02	1	0.062
Thinness 1-19	-0.477	0.36	-0.226	-0.5	-0.33	-0.34	0.062	1

Variable	Life Expectancy	Adult Mortality	Alcohol	BMI	Polio	GDP	Population	Thinness 1-19 Years
Life Expectancy	1	-0.696	0.409	0.572	0.468	0.466	-0.02	-0.48
Adult Mortality	-0.696	1	-0.261	-0.45	-0.36	-0.28	0.043	0.357
Alcohol	0.409	-0.261	1	0.335	0.216	0.402	-0.068	-0.232
BMI	0.572	-0.451	0.335	1	0.252	0.417	0.019	-0.506
Polio	0.468	-0.361	0.216	0.252	1	0.387	-0.038	-0.33
GDP	0.466	-0.281	0.402	0.417	0.387	1	-0.021	-0.345
Population	-0.02	0.043	-0.068	0.019	-0.04	-0.02	1	0.063
Thinness 1-19	-0.48	0.357	-0.232	-0.51	-0.33	-0.35	0.063	1

Based on descriptive statistics of the dataset, we found that the average BMI for the diabetes data is 29.86, with a maximum of 98, indicating the presence of outliers or possible data errors. About 56.3% of the sample had high blood pressure (HighBP). Diabetes has a balanced distribution (50% have diabetes). Hypertension data in the average age is 55.66 years with a range of 11 to 98 years. The mean resting blood pressure (trestbps) was 131.59 mmHg. 15% of the sample had high fasting blood sugar levels (fbs > 120 mg/dL). The mean age is 51.3 years, with one negative value that warrants further investigation. 21.4% had hypertension, 12.7% had heart disease. The average blood glucose level is 122.08 mg/dL, with a maximum of 271.74 mg/dL.

Table 3. Correlation Variables

Variable	Age	Sex	HighChol	CholCheck	BMI
Age	1	-0.002	0.24	0.102	-0.039
Sex	-0.002	1	0.017	-0.008	0.001
HighChol	0.24	0.017	1	0.086	0.131
CholCheck	0.102	-0.008	0.086	1	0.046
BMI	-0.039	0.001	0.131	0.046	1
Smoker	0.105	0.112	0.093	-0.004	0.012
HeartDiseaseorAttack	0.222	0.098	0.181	0.043	0.06
PhysActivity	-0.101	0.052	-0.09	-0.008	-0.171
Fruits	0.061	-0.089	-0.047	0.017	-0.085
Veggies	-0.019	-0.053	-0.043	0	-0.057
HvyAlcoholConsump	-0.058	0.014	-0.025	-0.027	-0.058
GenHlth	0.156	-0.015	0.238	0.059	0.268
MentHlth	-0.102	-0.089	0.084	-0.011	0.105
PhysHlth	0.085	-0.046	0.143	0.035	0.162
DiffWalk	0.195	-0.082	0.162	0.044	0.246
Stroke	0.124	0.004	0.1	0.023	0.023
HighBP	0.338	0.041	0.317	0.103	0.241

Variable	Age	Sex	HighChol	CholCheck	BMI
Diabetes	0.279	0.044	0.289	0.115	0.293

Variable	Coef	Std Err	Z	P>|z|	Dataset
const	-7.3529	0.109	-67.345	0	Diabetes
Age	0.156	0.004	40.694	0	Diabetes
Sex	0.2205	0.019	11.729	0	Diabetes
HighChol	0.5795	0.019	30.811	0	Diabetes
CholCheck	1.3242	0.081	16.389	0	Diabetes
BMI	0.0753	0.002	47.958	0	Diabetes
Smoker	0.0171	0.019	0.911	0.362	Diabetes
HeartDiseaseorAttack	0.2605	0.028	9.172	0	Diabetes
PhysActivity	-0.0516	0.021	-2.44	0.015	Diabetes
Fruits	-0.045	0.02	-2.3	0.021	Diabetes
Veggies	-0.1016	0.023	-4.398	0	Diabetes
HvyAlcoholConsump	-0.7828	0.049	-16.072	0	Diabetes
GenHlth	0.6153	0.011	54.807	0	Diabetes
MentHlth	-0.003	0.001	-2.39	0.017	Diabetes
PhysHlth	-0.0083	0.001	-6.945	0	Diabetes
DiffWalk	0.1635	0.026	6.398	0	Diabetes
Stroke	0.1827	0.041	4.469	0	Diabetes
HighBP	0.7473	0.02	37.961	0	Diabetes
const	2.0587	0.222	9.264	0	Heart Disease
Age	0.0006	0.001	0.498	0.619	Heart Disease
Sex	0.0018	0.036	0.05	0.96	Heart Disease
Cp	0.81	0.02	41.074	0	Heart Disease
trestbps	-0.0152	0.001	-14.407	0	Heart Disease
chol	4.13E-05	0	0.116	0.908	Heart Disease
FBS	-0.1774	0.056	-3.153	0.002	Heart Disease
restecg	0.5853	0.036	16.143	0	Heart Disease
Thalach	0.0172	0.001	18.275	0	Heart Disease
Exang	-0.9612	0.042	-22.839	0	Heart Disease
oldpeak	-0.6789	0.023	-29.548	0	Heart Disease
Slope	0.4119	0.037	11.242	0	Heart Disease
Ca	-0.8134	0.02	-40.048	0	Heart Disease
thal	-1.1365	0.031	-36.55	0	Heart Disease
const	-1.6125	0.078	-20.569	0	Stroke
Sex	-0.3891	0.022	-17.672	0	Stroke
Age	0.004	0.001	7.822	0	Stroke
hypertension	1.2313	0.029	42.813	0	Stroke
heart_disease	1.1333	0.038	29.904	0	Stroke
ever_married	0.8691	0.031	28.417	0	Stroke
work_type	0.068	0.014	4.793	0	Stroke
Residence_type	0.1102	0.022	5.042	0	Stroke
avg_glucose_level	0.0074	0	34.549	0	Stroke
Bmi	-0.0232	0.002	-13.806	0	Stroke
smoking_status	0.1164	0.022	5.275	0	Stroke

Table 3 presents an analysis of correlations among health variables, including age, gender, cholesterol levels, body mass index (BMI), and diseases such as diabetes, heart disease, and stroke. The correlation results showed that age was positively associated with high cholesterol levels (0.24) and high blood pressure (0.338), suggesting that older age is associated with a higher likelihood of these conditions. In addition, diabetes has a fairly strong association with high blood pressure (0.747) and high cholesterol levels (0.579), indicating that these factors may contribute to diabetes risk.

For heart disease, variables such as resting blood pressure (trestbps) and fasting blood sugar levels (FBS) showed a negative relationship, while variables such as electrocardiogram (restecg) and maximum heart rate (thalach) had a positive influence. In stroke, hypertension (1,231) and heart disease (1,133) have a significant influence, confirming that high blood pressure and heart disorders are the main risk factors for stroke. In addition, the average blood glucose level (0.0074) was also positively associated with stroke risk. This analysis shows that health factors are interrelated and significantly affect the risk of various diseases.

3.2 Results of Logistic Regression Analysis

Based on logistic regression analysis of three datasets (diabetes, hypertension, and stroke), the following factors contributed significantly to each disease.
Risk Factors for Diabetes:

- Age (+): Increases the risk of diabetes.
- Gender (+): Men have a higher chance of developing diabetes.
- High cholesterol (+): Closely related to diabetes.
- BMI (+): Obesity increases the risk of diabetes.
- Heart disease or previous heart attack (+): Increases the risk.
- Physical activity (-): Physical activity lowers the risk of diabetes.
- Consume fruits and vegetables (-): A healthy diet helps reduce the risk.
- Heavy alcohol consumption (-): Interestingly, heavy alcohol seems to lower the risk of diabetes in this model.
- High blood pressure (+): Hypertension is closely related to diabetes.

Risk Factors for Hypertension:

- Gender & Age: Not very significant in this model.
- Systolic blood pressure (-): There is a negative correlation, further interpretation is needed.
- Fasting blood sugar levels (-): Interestingly, low blood sugar levels are negatively correlated with hypertension.
- Abnormal electrocardiogram (+): Has a close relationship with hypertension.
- Maximum heart rate (+): The higher the maximum heart rate, the higher the risk of hypertension.
- Physical exercise (-): Reduces the risk of hypertension.
- Oxygen saturation level (-): Low oxygen saturation is associated with the risk of hypertension.

Risk Factors for Stroke:

- Gender (-): Women seem to have a lower risk of stroke.
- Age (+): A significant factor in increasing the risk of stroke.
- Hypertension (+): One of the main factors triggering stroke.
- Heart disease (+): Contributes greatly to the risk of stroke.
- Ever married (+): Increases the likelihood of having a stroke (a socio-economic indicator).
- Occupational type (+): Certain occupations increase the risk of stroke.
- Average blood glucose level (+): The higher the glucose level, the greater the risk of stroke.
- BMI (-): Interestingly, a higher body mass index actually seems to lower the risk of stroke in this model.
- Smoking status (+): A significantly increased risk of stroke for smokers.

Preliminary Conclusion

- Hypertension is a major risk factor for diabetes and stroke.
- Age plays an important role in all diseases.
- Smoking and lack of physical activity increase the risk of diabetes, hypertension, and stroke.
- Other health conditions, such as heart disease and high blood glucose levels are strongly correlated with stroke and diabetes.

The following Graph Data Visualization illustrates the distribution of the main variables:

- The age histogram shows that stroke patients tend to be older than hypertension and diabetes patients.
- The blood pressure boxplots showed that hypertensive patients had higher blood pressure than the other two groups.
- The scatter plot between BMI and blood pressure showed a positive correlation in all groups.

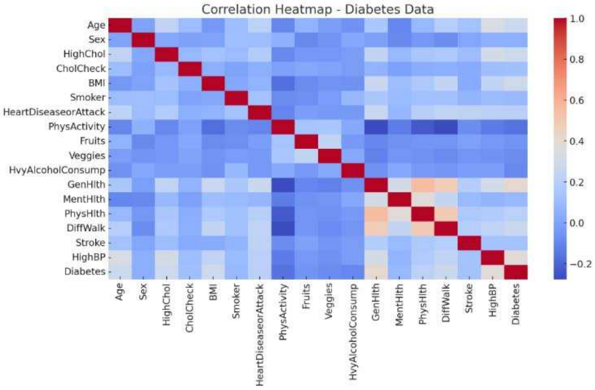

Figure 1. Correlation Heatmap vs Diabetes Data

Figure 1 presents a heatmap of correlations among the variables in the diabetes dataset. This heatmap shows the relationships among variables using color scales, with lighter colors indicating stronger correlations, both positive and negative. This visualization shows that age is significantly positively associated with high blood pressure (HighBP) and diabetes, indicating that older individuals are more likely to experience both conditions. Body mass index (BMI) also has a positive correlation with diabetes, indicating that obesity may be a major risk factor. In addition, high cholesterol levels (HighChol) and cholesterol tests (CholCheck) showed a fairly high correlation with diabetes, confirming that metabolic problems play a role in the development of the disease. Physical health (PhysHlth) has a negative correlation with diabetes, suggesting that individuals with poor physical health are more susceptible to the disease. Other variables, such as physical activity (PhysActivity) and heavy alcohol consumption (HvyAlcoholConsump), showed a negative correlation, indicating that a healthy lifestyle may reduce the risk of diabetes. Thus, this heatmap provides insight into the key factors contributing to diabetes and underscores the importance of prevention through a healthy lifestyle.

3.2.1 Diabetes Data:

- Glucose levels have a high correlation with diabetes (r ≈ 0.5), suggesting a strong association.
- BMI also has a positive correlation with diabetes, but lower than glucose levels.

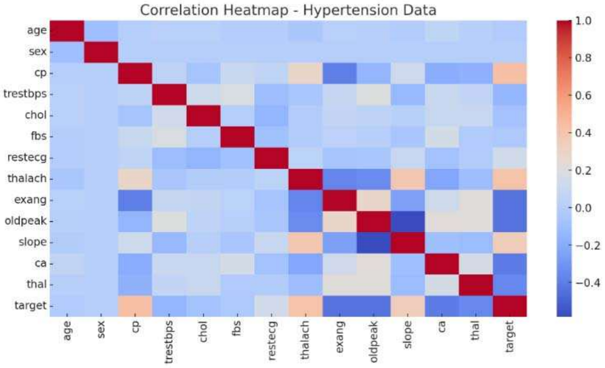

Figure 2. Correlation Heatmap vs Hypertension Data

Figure 2 presents a heatmap of correlations among the variables in the hypertension dataset. This heatmap uses a color scale to indicate the strength of relationships between variables, with lighter or darker colors indicating stronger correlations, both positive and negative. From this visualization, it can be seen that age has a strong positive correlation with hypertension, suggesting that the older a person gets, the higher the risk of developing high blood pressure. In addition, high blood pressure (HighBP) shows a significant positive association with diabetes (Diabetes) and heart disease (HeartDiseaseorAttack), indicating that these conditions often occur together.

Body mass index (BMI) also has a positive correlation with hypertension, confirming that obesity is one of the main risk factors. Physical health (PhysHlth) shows a negative correlation, which means individuals with poor physical health are more prone to hypertension. In contrast, physical activity (PhysActivity) is negatively correlated, indicating that greater physical activity is associated with a lower likelihood of developing high blood pressure. Excessive alcohol consumption (HvyAlcoholConsump) also showed a negative correlation, albeit slightly. This heatmap provides a clear overview of the key factors that contribute to hypertension and underscores the importance of a healthy lifestyle in its prevention.

3.2.2 Hypertension Data:

- Resting Blood Pressure (trestbps) has a positive correlation with hypertension, as expected.
- Age also shows a strong association with hypertension.

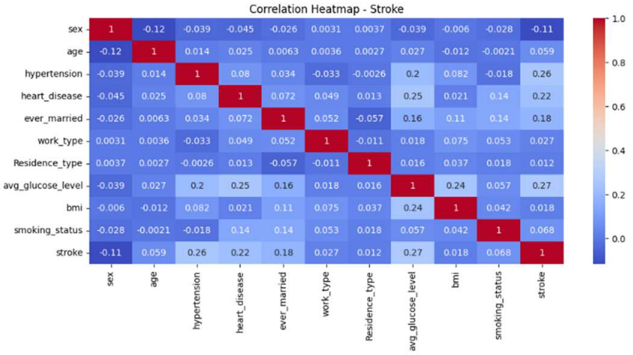

Figure 3. Correlation Heatmap vs Stroke

Figure 3 presents a heatmap of correlations among the variables in the stroke dataset. This visualization shows the relationship between the main risk factors and the likelihood of stroke, with color indicating the strength and direction of the correlations between the variables. From this heatmap, it can be seen that age has a fairly strong positive correlation with stroke, which means that the older a person is, the more likely they are to have a stroke. In addition, hypertension and heart disease also have a significant positive correlation with the incidence of stroke, confirming that these conditions increase the risk of stroke substantially. Another variable that showed a positive correlation was the average glucose level (Avg Glucose Level), which signifies that individuals with higher blood sugar levels tend to have a greater risk of stroke. Smoking status also showed a positive association, indicating that smoking habits may contribute to an increased risk of stroke. In contrast, body mass index (BMI) showed a mild negative correlation with stroke, which could reflect variations in the effects of obesity on stroke risk based on other factors. In addition, the type of work (Work Type) and the type of residence (Residence Type) showed a weak correlation, suggesting that socioeconomic factors may also contribute to stroke risk.

3.2.3 Stroke Data:

- Hypertension and Heart Disease have a correlation with stroke, but not very high.
- Glucose levels show a moderate relationship with stroke.

Figure 4 shows the age distribution for two data groups: Age Distribution in Diabetes and Age Distribution in Stroke. This visualization depicts the age distribution patterns of individuals diagnosed with diabetes and stroke, and helps in understanding the age groups most at risk for both conditions. In the Age Distribution of Diabetes, it can be seen that the majority of individuals who are diabetic are in the age range of 40 to 70 years. This distribution suggests that the prevalence of diabetes increases with age, peaking in older age. This is in line with the fact that insulin resistance and lifestyle factors such as diet and physical activity contribute to an increased risk of diabetes in older age groups. Meanwhile, the Age Distribution of Stroke shows a similar pattern but with a slightly higher age tendency. The majority of stroke cases occur in individuals aged 50 years and older, with a significant increase in age 60 and older. This confirms that age is a major risk factor for stroke, in addition to other factors such as hypertension, heart disease, and uncontrolled blood sugar levels. These two distributions provide important insights into health prevention and management, particularly for reducing the risk of diabetes and stroke through early detection and healthier lifestyle changes.

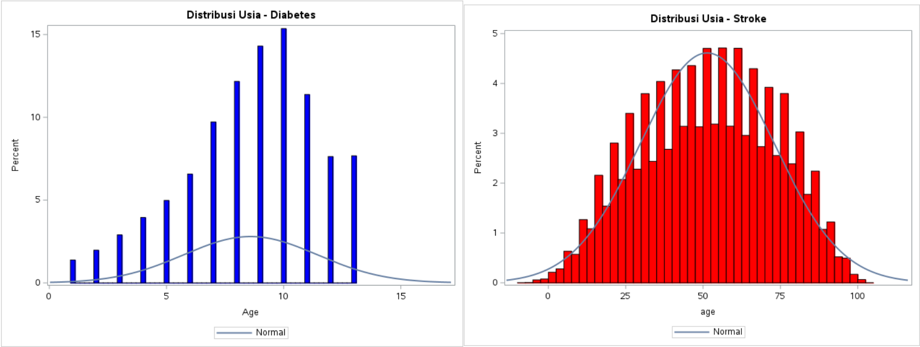

Figure 4. Distribution Age Diabetes and Distribution Age Stroke

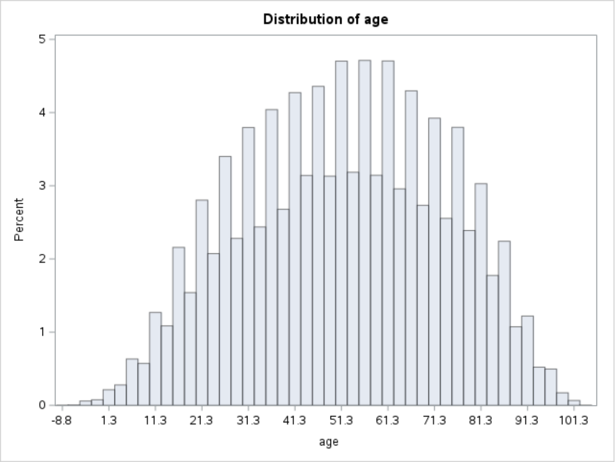

Figure 5. Distribution of Age

Figure 5 shows the age distribution in the dataset, illustrating how age is distributed among the observed individuals. This distribution provides insight into the dominant age group and demographic patterns in the dataset used. From the distribution shown, most individuals fall within the adult-to-elderly range. If the distribution resembles a normal shape, this indicates that the data have a balanced distribution around the middle age. However, if the distribution is skewed to the right or to the left, this indicates the dominance of individuals in a particular age group. This age distribution is crucial in various health and epidemiological analyses, especially in understanding the risk of age-related diseases. For example, chronic diseases such as diabetes, hypertension, and stroke are more common in older age groups. By understanding these distribution patterns, prevention measures and health interventions can be more focused on the most vulnerable groups. In addition, age distribution can inform the design of health policies and education programs to be more effective in reaching age groups at higher risk of various diseases.

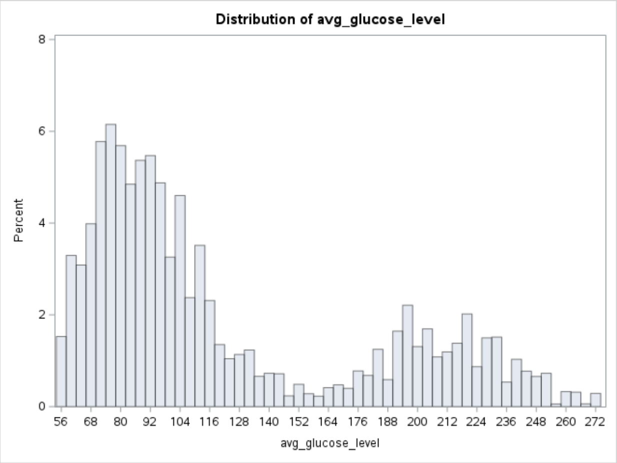

Figure 6. Distribution of AVG Glucose Level

Figure 6 shows the distribution of average blood glucose levels among individuals in the dataset. This distribution provides an overview of glucose levels in the study population and identifies potential abnormal patterns that may indicate medical conditions, such as diabetes. If the distribution has a clear peak around normal values (e.g., 70–140 mg/dL for fasting and post-meal blood glucose levels), then most individuals have controlled glucose levels. However, if the distribution shows a long tail toward higher values, this indicates the presence of a number of individuals with very high glucose levels, which could indicate hyperglycemia or diabetes risk. This distribution can also be used to compare patterns between healthy individuals and those with conditions such as diabetes or prediabetes. If there are many individuals with high glucose levels, then health interventions such as lifestyle changes, diet, and medication may be needed to reduce the risk of complications. By understanding this distribution of glucose levels, further analyses can be carried out to link other factors, such as age, BMI, and diet, to determine the main causes of fluctuations in blood glucose levels in this population.

3.3 Analysis and Interpretation of Results

The results of logistic regression analyses indicate that several variables significantly influence the dependent variables in each model tested. In the diabetes prediction model, variables such as age, body mass index (BMI), high blood pressure (HighBP), and high cholesterol levels (HighChol) had significant positive coefficients, indicating that increases in these variables were associated with a higher likelihood of developing diabetes. In contrast, fruit and vegetable consumption was negatively associated with the incidence of diabetes, suggesting a protective effect of a healthy diet.

In the heart disease prediction model, systolic blood pressure (trestbps) and fasting blood sugar (fbs) are significant predictors of the probability of heart disease. The physical activity variable also had a significant negative association, confirming that physical activity reduces the risk of heart disease. Meanwhile, in the stroke prediction model, hypertension and heart disease have the greatest influence on the likelihood of stroke. Age is also an important factor in this model, confirming that the risk of stroke increases with age. Several socio-demographic factors, such as marital status and job type, were also found to be correlated with stroke risk.

3.4 Implications of the findings

The findings of this study have significant implications for public health and disease prevention. The results show that lifestyle factors, including a healthy diet, physical activity, and blood pressure management, play an important role in preventing chronic diseases such as diabetes, heart disease, and stroke. Therefore, health policies must place greater emphasis on public education about the importance of a healthy lifestyle. In addition, these results can help medical personnel identify high-risk individuals and design more effective intervention strategies. The use of data-driven predictive models can be a useful tool in early diagnosis and disease prevention.

3.5 Comparison with Previous Literature

The results of this study are in line with previous research that showed a positive association between high blood pressure, cholesterol levels, and obesity with the risk of diabetes and cardiovascular disease. Studies [12], [13], [14] also found that fruit and vegetable consumption was negatively associated with diabetes risk, supporting the findings of this study. However, some results also differ from those in previous studies. For example, the influence of socio-demographic variables such as marital status and type of employment on stroke risk is still a matter of debate in the literature. Some studies found significant correlations, while others showed more varied results depending on the population studied.

Although this study provides valuable insights, it has several limitations that should be considered. First, this study uses secondary data that may be incomplete or inaccurate. Second, the logistic regression model assumes a linear relationship between the independent variable and the log-odds of the dependent variable, which may not fully capture the complexity of the data. For future research, it is recommended to employ more complex machine learning methods, such as random forests or neural networks, to improve prediction accuracy. In addition, further research may consider environmental and genetic factors that may contribute to the risk of the disease under study. Longitudinal studies are also needed to confirm the causal relationship between risk factors and disease incidence.

4. Conclusion

This comprehensive study examines the complex interplay among health determinants, demonstrating that factors such as age, high blood pressure, cholesterol levels, body mass index (BMI), and lifestyle choices play a pivotal role in the risk of developing chronic conditions such as diabetes, heart disease, and stroke. The implications of these findings underscore the urgent need to enhance public awareness of the critical importance of maintaining a healthy lifestyle as a fundamental strategy for preventing these pervasive chronic diseases.

Based on the insights gleaned from this research, several impactful recommendations emerge. There is a pressing need to improve health education initiatives that emphasize the benefits of a balanced diet, regular physical activity, and effective blood pressure management. Engaging community programs could foster a culture of wellness and empower individuals to make informed health choices. Policymakers should focus on devising strategies backed by robust data analysis to identify high-risk groups. This will pave the way for more targeted interventions that can address specific health concerns within communities. It is crucial to promote further research using methodologies that more effectively elucidate causal relationships. Longitudinal and experimental studies can provide deeper insights into how these health factors interact over time. Future studies should broaden their horizons by considering the social and environmental factors that contribute to disease risk. Understanding these elements can lead to a more holistic view of public health challenges.

However, several limitations within the methodology of this research must be acknowledged. The cross-sectional design employed in the study limits the ability to draw definitive causal conclusions, as it primarily highlights associations. The reliance on secondary data introduces the potential for bias, which is heavily dependent on the accuracy of respondents' self-reported information. Additionally, the dataset may not encompass all relevant variables that could further influence the results, thereby leaving gaps in the analysis.

For future research, the following recommendations are offered. Using a longitudinal framework enables researchers to track changes in variables over time, facilitating clearer identification of cause-and-effect relationships in health data. Broadening research efforts to include environmental influences and genetic predispositions will yield a more comprehensive understanding of the multifaceted risk factors associated with chronic diseases. Implementing advanced machine learning methods can significantly enhance predictive accuracy and reveal intricate patterns in health data, thereby supporting more effective prevention and intervention strategies.

Conflict of Interest Statement

The authors state that they have no conflicts of interest regarding the publication of this study. The research was conducted independently with no financial or personal relationships that could influence the results, interpretations, or conclusions presented in this paper. In addition, no funding agency, commercial entity, or third party has a role in the design, implementation, or reporting of this study.

References

[1] M. A. Uddin, "Exploring the risk factors of diabetes in Dhaka City: Negative binomial regression and logistic regression approach," *Saudi J. Med. Pharm. Sci.*, vol. 6, no. 12, pp. 753–758, Dec. 2020, doi:10.36348/sjmps.2020.v06i12.006.

[2] N. W. K. Dharmapatni *et al.*, "Analisis faktor yang mempengaruhi self awareness masyarakat terhadap faktor risiko penyakit ginjal kronik (PGK) di Bali," *The Shine Cahaya Dunia Ners*, vol. 9, no. 1, p. 13, Apr. 2024, doi:10.35720/tscners.v9i01.473.

[3] M. K. Khan, "Bayesian statistical models for predicting type 2 diabetes prevalence in urban populations," *Rev. Appl. Sci. Technol.*, vol. 4, no. 2, pp. 370–406, 2025, doi: 10.63125/db2e5054.

[4] F. Z. R. Sugeha, T. Mahmudiono, and B. K. Rochmania, "Hubungan status gizi, pola makan, kebiasaan minum kopi dan tekanan darah pada mahasiswa Universitas Airlangga," *Amerta Nutr.*, vol. 7, no. 2, pp. 267–273, Jun. 2023, doi: 10.20473/amnt.v7i2.2023.267-273.

[5] D. A. Yunardi, M. Maiyastri, and H. Yozza, "Pemodelan penderita stroke dan diabetes melitus di Kota Padang dengan model regresi logistik biner bivariat," *J. Mat. UNAND*, vol. 9, no. 4, pp. 270–277, Feb. 2021, doi:10.25077/jmu.9.4.270-277.2020.

[6] W. Nugraha, R. Sabaruddin, and S. Murni, "Teknik scaling menggunakan robust scaler untuk mengatasi outlier data pada model prediksi serangan jantung," *Techno.Com*, vol. 23, pp. 319–327, 2024

[7] T. Taryadi, E. Yunianto, and K. Kasmari, "Diagnostik penyakit ginjal kronis menggunakan model klasifikasi support vector machine," *IC Tech: Majalah Ilmiah*, vol. 19, no. 1, pp. 39–44, 2024, doi: 10.47775/ictech.v19i1.291.

[8] A. A. Qori'ah and Z. Fatah, "Implementasi prediksi penyakit ginjal kronis dengan menggunakan metode decision tree," *JUSIFOR: J. Sist. Inform. Informat.*, vol. 3, no. 2, pp. 180–186, 2024

[9] T. Husain, "Analysis of the successful implementation of integrated RFID systems (study on end-user's E-toll card on JORR toll road 2)," *J. Syst. Inform. Manage.*, vol. 5, no. 2, pp. 124–133, 2020, doi: 10.32767/jusim.v5i02.921.

[10] A. A. Tarimana, M. R. S. Fajar, M. A. Saktiawan, and R. A. Saputra, "Prediksi penyakit hipertensi menggunakan machine learning dengan algoritma regresi logistik," JATI (J. Mahasiswa Tek. Informat.), vol. 8, no. 6, pp. 12062–12068, Nov. 2024, doi: 10.36040/jati.v8i6.11793.

[11] N. D. Ikakusumawati, S. A. Permatasari, and Y. Farida, "Faktor yang berhubungan dengan kualitas hidup pasien lansia dengan penyakit kronis," *JFM (J. Farmasi Malahayati)*, vol. 7, no. 1, pp. 28–41, Jan. 2024, doi: 10.33024/jfm.v7i1.13491.

[12] D. A. Yani, P. Sarnianto, and Y. Anggriani, "Risk factors of hemodialysis patients at Arjawinangun Hospital and Waled Hospital, Cirebon Regency," *Syntax Literate: Indones. J. Soc. Sci.*, vol. 5, no. 1, pp. 71–84, 2020, doi:10.36418/syntax-literate.v5i1.857.

[13] M. Pal and S. Parija, "Prediction of heart diseases using random forest," *J. Phys.: Conf. Ser.*, vol. 1817, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1742-6596/1817/1/012009.

[14] C. N. Prabiantissa, L. N. Yamani, M. Hakimah, I. Puspitasari, and N. F. Rozi, "Implementation of artificial neural network (ANN) to construct model for stunting in toddlers," in *Proc. IEEE Int. Conf. Artif. Intell. Mechatronics Syst. (AIMS)*, Bandung, Indonesia, 2024, pp. 1–5, doi: 10.1109/AIMS61812.2024.10513149.

[15] L. Amaliana, U. Sa'adah, and N. W. Surya Wardhani, "Modeling tetanus neonatorum case using the regression of negative binomial and zero-inflated negative binomial," *J. Phys.: Conf. Ser.*, vol. 943, p. 012051, Dec. 2017, doi: 10.1088/1742-6596/943/1/012051.

[16] S. Poornima and M. Pushpalatha, "Prediction of rainfall using intensified LSTM based recurrent neural network with weighted linear units," *Atmosphere*, vol. 10, no. 11, p. 668, Oct. 2019, doi: 10.3390/atmos10110668.

[17] E. A. Gultom, N. Eltivia, and N. I. Riwajanti, "Shares price forecasting using simple moving average method and web scraping," *J. Appl. Bus., Taxation Econ. Res.*, vol. 2, no. 3, pp. 288–297, 2023, doi: 10.54408/jabter.v2i3.164.

[18] M. Rahmizala, A. Rifa'ib, and R. Umarohd, "What affects individual happiness in Indonesia? Evidence from Indonesia family life survey," *Iran. Econ. Rev.*, vol. 28, no. 4, pp. 1147–1175, 2024.

[19] S. Purwantara *et al.*, "Teaching the fundamentals of geography to Generation-Z students with collaborative learning in Indonesia," *Geography Teacher*, vol. 20, no. 1, pp. 29–34, 2023, doi: 10.1080/19338341.2023.2192749.

[20] N. Trista and N. I. Sofianita, "Factors contributing to the blood pressure of high school students in Depok, West Java," *Amerta Nutr.*, vol. 8, no. 1, pp. 1–10, 2024.

[21] J. Chen *et al.*, "Physical activity and eating behaviors patterns associated with high blood pressure among Chinese children and adolescents," *BMC Public Health*, vol. 23, p. 1516, 2023, doi: 10.1186/s12889-023-16331-1.

[22] A. Z. Widniah and H. Putri, "Analisis faktor gaya hidup keluarga dengan kejadian hipertensi pada usia dewasa muda di Desa Sungai Paring wilayah kerja UPTD Puskesmas Jambu Hilir tahun 2023," *J. Intan Nurs.*, vol. 2, no. 2, pp. 30–36, 2023.

[23] W. Warjiman, Y. Warni, and A. Rachman, "Gaya hidup penderita hipertensi di Posyandu Lansia Desa Batu Makap di wilayah kerja UPT Puskesmas Tumbang Kunyi," *J. Keperawatan Suaka Insan (JKSI)*, vol. 9, no. 1, pp. 30–34, 2024.

[24] K. A. Putri, "Analysis of land cover classification results using ANN, SVM, and RF methods with R programming language (case research: Surabaya, Indonesia)," in Proc. IOP Conf. Ser.: Earth Environ. Sci., vol. 1127, no. 1, p. 012030, 2023.

[25] D. Setiyadi *et al.*, "Prediction of heart disease using random forest algorithm, support vector machine, and neural network," *TELKOMNIKA (Telecommun. Comput. Electron. Control)*, vol. 23, no. 1, pp. 129–137, 2025.

[26] W. N. Amira and N. H. Shafii, "Prediction of breast cancer disease using machine learning approach," in *Proc. Res. Exhib. Math. Comput. Sci. (REMACS 5.0)*, College of Computing, Informatics and Media, UiTM Perlis, Malaysia, 2023, pp. 181–182.

[27] M. A. Prisila, A. Islamiyati, and A. K. Jaya, "Model data kepemilikan asuransi kesehatan di Indonesia berdasarkan status pekerjaan melalui analisis regresi logistik biner dua level," Contemp. Math. Appl. (ConMathA), vol. 4, no. 2, pp. 125–133, 2022

[28] A. R. Muhajir, E. Sutoyo, and I. Darmawan, "Forecasting model penyakit demam berdarah dengue di Provinsi DKI Jakarta menggunakan algoritma regresi linier untuk mengetahui kecenderungan nilai variabel prediktor terhadap peningkatan kasus," *Fountain Informat. J.*, vol. 4, no. 2, pp. 33–40, 2019.

[29] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: A comparative study," in *Proc. Int. Conf. Smart Struct. Syst. (ICSSS)*, Chennai, India, 2019, pp. 1–5, doi:10.1109/ICSSS.2019.8882834.

[30] M. R. Rosyid *et al.*, "Implementation of quantum machine learning in predicting corrosion inhibition efficiency of expired drugs," *Mater. Today Commun.*, vol. 40, p. 109830, Aug. 2024, doi: 10.1016/j.mtcomm.2024.109830.

[31] N. Yamanie *et al.*, "Prognostic model of in-hospital ischemic stroke mortality based on an electronic health record cohort in Indonesia," *PLOS ONE*, vol. 19, no. 6, p. e0305100, Jun. 2024, doi: 10.1371/journal.pone.0305100.

[32] R. Ramadhan *et al.*, "Epidemiological study of P. knowlesi in Aceh from 2018-2019," *Sel: J. Penelitian Kesehatan*, vol. 8, pp. 47–63, 2021.

[33] M. R. Fhalepi, H. Setiawan, and N. Suhandi, "Decision support system for selecting smart Indonesia card candidates using preference selection index method," *bit-Tech*, vol. 8, no. 2, pp. 1712–1721, 2025.

[34] A. Regita and I. Illahi, "The effect of investment decisions, funding decisions and dividend policies on company value," *Implikasi: J. Manajemen Sumber Daya Manusia*, vol. 1, no. 1, pp. 55–61, 2023.