BAREKENG: Journal of Mathematics and Its Applications
June 2026 Volume 20 Issue 2 Page 1125–1138
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss2pp1125-1138

TRADITIONAL LOGISTIC REGRESSION AND MACHINE LEARNING APPROACHES OF SOCIODEMOGRAPHIC AND ANTHROPOMETRIC FACTORS INFLUENCING HYPERTENSION IN ATHLETES

A'yunin Sofro [1*], Asri Maharani [2], Mutia Eva Mustafidah [3], Khusnia Nurul Khikmah [4], Affiati Oktaviarina [5], Danang Ariyanto [6]

[1,5,6]*Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya*
[3]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya*
Jln. Ketintang, Surabaya, Jawa Timur, 60213, Indonesia

[2]*Mental Health Research Group, Division of Nursing, Midwifery and Social Work, School of Health Sciences, University of Manchester and Manchester Academic Health Science Centre (MAHSC)*
Oxford Rd, Manchester, M13 9PL, United Kingdom

[4]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Palangka Raya, Jln. Yos Sudarso, Palangka Raya, Kalimantan Tengah, 74874, Indonesia,*

*Corresponding author's e-mail: * **ayuninsofro@unesa.ac.id***

Article Info

Article History:

Received: 30th April 2025
Revised: 1st July 2025
Accepted: 13th September 2025
Available online: 18th January 2026

Keywords:

Athletes;
Binary logistic regression;
Hypertension;
Machine learning;
Random effects.

ABSTRACT

The type and intensity of exercise performed by athletes play an important role in affecting blood pressure stability, putting them at risk of developing hypertension. Hypertension, or high blood pressure, is a medical condition in which the blood pressure in the arteries rises above normal limits. Hypertension in athletes becomes an essential factor in real cases if not detected early. Therefore, this study aims to model and analyse the sociodemographic and anthropometric factors that influence the incidence of hypertension. The data used in this study are primary data from 200 athlete selection participants at the University of Surabaya and the Indonesian National Sports Committee (INSC) of East Java. This research method proposes to compare the traditional approach with machine learning to prove the accuracy comparison of the model's goodness, where both approaches are proposed by considering the novelty proposed through the machine learning approach but still maximizing the traditional approach. The proposed methods are binary logistic regression, binary logistic regression with the addition of random effects, highly randomized tree, and support vector classification. The binary logistic regression model is better than the binary logistic regression model with random effects, random trees, and support vector classification because the accuracy, sensitivity, specificity, and F1-score value (68.5%, 69%, 68%, and 68.8%) is highest than the others. Other results showed that the waist circumference variable, the father's occupation variable, and the salary variable significantly affected hypertension at the 5% significance level.

How to cite this article:

A. Sofro, A. Maharani, M. E. Mustafidah, K. N. Khikmah, A. Oktaviarina and D. Ariyanto., "TRADITIONAL LOGISTIC REGRESSION AND MACHINE LEARNING APPROACHES OF SOCIODEMOGRAPHIC AND ANTHROPOMETRIC FACTORS INFLUENCING HYPERTENSION IN ATHLETES," *BAREKENG: J. Math. & App.*, vol. 20, no. 2, pp. 1125-1138, Jun, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Hypertension, or high blood pressure, is a chronic medical condition characterized by consistently elevated blood pressure in the arteries beyond normal levels [1]. This condition is one of the leading causes of morbidity and mortality worldwide, affecting approximately 25% of the global population. Hypertension is considered a catastrophic disease due to its grave and potentially fatal consequences if left untreated [2]. Persistent high blood pressure can damage blood vessels and vital organs such as the heart, brain, kidneys, and eyes. Consequently, hypertension increases the risk of heart disease, stroke, kidney failure, and other life-threatening complications. Furthermore, hypertension often presents no early symptoms, meaning many individuals remain unaware of their condition until severe complications arise [3]. In athletes, hypertension has significant implications for both performance and overall health. The type and intensity of physical activity can influence blood pressure stability, thereby increasing the risk of hypertension [4]. While exercise generally benefits cardiovascular health, extreme or uncontrolled physical activity may have adverse effects. However, research on hypertension among athletes is still limited, despite the unique characteristics of this population that differentiate them from the general population. Sociodemographic factors, such as gender, and anthropometric factors, such as body mass index (BMI) and height, are known to have a significant relationship with hypertension [5]. Research by [6] indicates that men tend to have a higher risk of hypertension compared to women. Meanwhile, other studies show that increased body weight and BMI are directly associated with the prevalence of hypertension [7]. However, there are several research gaps in the context of hypertension among athletes. First, most previous studies have focused on the general population without accounting for the unique characteristics of athletes, such as cardiovascular adaptations resulting from intensive training. Second, the analytical models used in earlier research, such as conventional logistic regression, often fail to consider correlations between individuals, resulting in analyses that do not fully capture the realities of heterogeneous populations like athletes. Third, studies exploring how the interaction between sociodemographic and anthropometric factors influences hypertension in athletes are scarce.

This study employs binary logistic regression, a method used to describe the relationship between a response variable and one or more predictor variables [8]. This approach is further developed by incorporating random effects, resulting in binary logistic regression with random effects. The inclusion of random effects addresses the issue of inter-individual correlations commonly found in binary logistic regression. This approach allows for a more in-depth analysis by modeling the relationship between hypertension as the response variable and sociodemographic and anthropometric factors as predictor variables while accounting for individual-level variability. Thus, this model is expected to provide more accurate and relevant estimates than conventional methods. Furthermore, to enhance the robustness of the analysis, this study compares the performance of various machine learning models in predicting hypertension among athletes. In addition to binary logistic regression and binary logistic regression with random effects, the study integrates ensemble methods such as Extremely Randomized Trees (Extra Trees) and Support Vector Classification (SVC) [9]. Extra Trees is an ensemble algorithm that builds multiple decision trees with high levels of randomization in feature selection and split points, enabling the model to handle complex and heterogeneous data variability [10]. Meanwhile, SVC is a classification algorithm that effectively separates data into distinct categories by finding an optimal hyperplane [11], [12]. The best model selection is conducted by comparing the actual values of hypertension status with the predicted values from each classification model, obtained using the Extra Trees and SVC methods, ensuring that the most accurate and reliable classification model is chosen [8].

This study aimed to predict hypertension influenced by sociodemographic and anthropometric factors in athletes using two traditional method approaches and machine learning methods: binary logistic regression, binary logistic regression with random effects, Extra Trees, and SVC. The findings of this study are intended to gain new insights into the accuracy of predictions, especially cases of hypertension, against the proposed method. In addition, this study is intended to provide health practitioners and coaches with knowledge of the conditions of athletes, identify risk factors for hypertension in athletes, and design more effective interventions.

2. RESEARCH METHODS

2.1 Research Data

The type of research conducted is quantitative research. The research data comprises primary data from 200 athlete selection participants at the University of Surabaya and the Indonesian National Sports Committee (INSC) of East Java. Data collection in the selection process was carried out in two stages. In the first stage, observations were made regarding the sociodemographic and anthropometric factors of the athletes. The second stage involved measuring hypertension, which was assessed by checking the blood pressure of the athletes. If the blood pressure exceeds 120/80 mmHg, the athlete is indicated to have hypertension. In this study, the response variable ($\boldsymbol{y}$) is athlete hypertension, which is categorical, consisting of two categories: $\boldsymbol{y}$= 1 indicating hypertension and $\boldsymbol{y} = \boldsymbol{0}$ indicating no hypertension. The predictor variables ($\boldsymbol{x}$) include 12 categorical variables consisting of sociodemographic and anthropometric factors. Meanwhile, the random effects categorize athletes into two groups: 1 for athletic athletes and 0 for non-athletic athletes. Observation of sociodemographic factors involves collecting data on individuals' or groups' social and demographic characteristics. The observed factors include age (categorized as 1 for < 21 years and 0 for ≥ 21 years), gender (categorized as 1 for female and 0 for male), father's education (categorized as 1 for school and 0 for college), mother's education (categorized as 1 for school and 0 for college), father's occupation (categorized as 1 for formal and 0 for informal), mother's occupation (categorized as 1 for formal and 0 for informal), and parents' salary, which is divided into two categories: salary 1 (1 for < 3 million and 0 for 3-6 million) and salary 2 (1 for < 3 million and 0 for >6 million). Survey methods, interviews, direct observation, and secondary data analysis were utilized to collect this data. Observing anthropometric factors aims to understand body proportions and physical characteristics that affect athlete performance in specific sports. The observed factors included height (categorized as 1 for < 170 cm and 0 for ≥170 cm), weight (categorized as 1 for < 60 kg and 0 for ≥ 60 kg), body mass index (categorized as 1 for < 25 and 0 for ≥25), and waist circumference (categorized as 1 for < 85 cm and 0 for ≥ 85 cm).

2.2 Research Design

The results of parameter estimation in binary logistic regression with random effects for athletes, using Maximum Likelihood Estimation (MLE). According to the research problem being investigated, this study employs binary logistic regression and binary logistic regression with random effects to model hypertension and analyze the factors influencing hypertension in athletes. This method can be used to identify the effects of predictor variables on the response variable [8]. The parameters of both models are obtained by estimating the parameters using Maximum Likelihood Estimation (MLE). The next step involves conducting significance tests for the parameters, performed simultaneously using the G and partially using the Wald tests. Then, the best model was selected by comparing the accuracy value. Additionally, machine learning methods, as previously explained, will be applied to enhance the analysis further and improve model prediction accuracy. These methods include ensemble techniques such as Extremely Randomized Trees (Extra Trees) and Support Vector Classification (SVC), which will be utilized to compare the performance of different classification models in predicting hypertension among athletes.

2.3 Binary Logistic Regression

Binary logistic regression a statistical method used to model the relationship between one or more predictor variables and a response variable. The response variable is denoted by y, and the predictor variable is denoted by x. If y falls into two categories, for example, "1" for success and "0" for failure, then the variable y follows a Binomial distribution, with the probability function given as follows.

$$P(y_i|x_i) = \big(P(y_i = 1|x_i)\big)^{y_i}\big(P(y_i = 0|x_i)\big)^{1-y_i}, \tag{1}$$

where $y_i = 0; 1, P(y_i = 1|x_i)$ is the probability of success, and $P(y_i = 0|x_i)$ is the probability of failure, where $i = 1,2, \dots, n_j$.

The binary logistic regression model can be expressed by the following equation.

$$P(y_i = 1|x_i) = \frac{\exp\big(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}\big)}{1 + \exp\big(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}\big)}. \tag{2}$$

By applying the logit transformation from Eq. (2), the binary logistic regression model can be expressed as follows:

$$\log\left(\frac{P(y_i=1|x_i)}{P(y_i=0|x_i)}\right)=\beta_0+\sum_{k=1}^{p}\beta_k x_{ki}. \tag{3}$$

Next, we will explain about parameter estimation using Maximum Likelihood Estimation (MLE). Maximum Likelihood Estimation (MLE) is a parameter estimation method used to estimate parameters in models with known distributions. In this case, the distribution used is the binomial distribution. The likelihood function of the logistic regression model with random effects can be expressed as follows [13].

$$\begin{aligned} L(\beta) &= \Pi_{i=1}^{n} P(y_i|x_i). \\ &= \Pi_{i=1}^{n}\big(P(y_i=1|x_i)\big)^{y_i}\big(P(y_i=0|x_i)\big)^{1-y_i}. \end{aligned} \tag{4}$$

Using MLE, Eq. (4) will be transformed into the log-likelihood function as follows [14].

$$\log L\,(\beta)=\sum_{k=1}^{p}\log\left(\sum_{i=1}^{n}y_i\,x_{ki}\right)\beta_k-\sum_{i=1}^{n}\ln\left(1+\exp\left(\beta_0+\sum_{k=1}^{p}\beta_k x_{ki}\right)\right), \tag{5}$$

with:

$$P(y_i|x_i)=\left(\frac{\exp\left(\beta_0+\sum_{k=1}^{p}\beta_k x_{ki}\right)}{1+\exp\left(\beta_0+\sum_{k=1}^{p}\beta_k x_{ki}\right)}\right)^{y_i}\left(\frac{1}{1+\exp\left(\beta_0+\sum_{k=1}^{p}\beta_k x_{ki}\right)}\right)^{1-y_i}. \tag{6}$$

The function obtained is in implicit form, so it cannot be solved in a simple form. To address this issue, the numerical method, specifically the Newton-Raphson method, is used to estimate the parameters.

2.4 Binary Logistic Regression with Random Effects

Binary logistic regression with random effects is an extension of ordinary binary logistic regression. This model involves a binary or dichotomous response variable, along with the inclusion of random effects. The addition of random effects addresses issues in binary logistic regression where correlations between individuals may exist. A dichotomous variable is one that has only two possible values, such as success and failure. The response variable is denoted by y, and the predictor variable is denoted by x. If y falls into two categories, for example, "1" for success and "0" for failure, then the variable y follows a Binomial distribution, with the probability function given as follows:

$$P\left(y_{ji}|x_{ji},u_j\right)=\left(P\left(y_{ji}=1|x_{ji},u_j\right)\right)^{y_{ji}}\left(P\left(y_{ji}=0|x_{ji},u_j\right)\right)^{1-y_{ji}}, \tag{7}$$

where $y_{ji}=0;1, P\left(y_{ji}=1|x_{ji},u_j\right)$ is the probability of success, and $P\left(y_{ji}=0|x_{ji},u_j\right)$ is the probability of failure, with $j=1,2,\dots,J$ and $i=1,2,\dots,n_j$.

The binary logistic regression model with random effects can be expressed by the following Eq. (8) [15],[16].

$$P\left(y_{ji}=1|x_{ji},u_j\right)=\frac{\exp\left(\beta_0+\sum_{k=1}^{p}\beta_k x_{kji}+u_j\right)}{1+\exp\left(\beta_0+\sum_{k=1}^{p}\beta_k x_{kji}+u_j\right)}, \tag{8}$$

where u_j represents the random effect with $u_j\sim N(0,\sigma^2)$. By applying the logit transformation from Eq. (8), the binary logistic regression model with random effects can be expressed as follows.

$$\log\left(\frac{P\left(y_{ji}=1|x_{ji},u_j\right)}{P\left(y_{ji}=0|x_{ji},u_j\right)}\right)=\beta_0+\sum_{k=1}^{p}\beta_k x_{kji}+u_j. \tag{9}$$

Maximum Likelihood Estimation (MLE) is a parameter estimation method used to estimate parameters in models with known distributions. In this case, the distribution used is the binomial distribution. The likelihood function of the logistic regression model with random effects can be expressed as follows.

$$L(\beta)=\Pi_{j=1}^{J}\int\left(\Pi_{i=1}^{n_j}P\left(y_{ji}|x_{ji},u_j\right)\right)f\left(u_j|\sigma^2\right)du_j. \tag{10}$$

Using MLE, Eq. (10) will be transformed into the log-likelihood function as follows.

$$\log L(\beta) = \sum_{j=1}^{J} \log\left(\int \Pi_{i=1}^{n_j} P(y_{ji}|x_{ji}, u_j) f(u_j|\sigma^2) du_j \right), \tag{11}$$

with,

$$P(y_{ji}|x_{ji}, u_j) = \left(\frac{\exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_{kji} + u_j)}{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_{kji} + u_j)} \right)^{y_{ji}} \left(\frac{1}{1 + \exp(\beta_0 + \sum_{k=1}^{p} \beta_k x_{kji} + u_j)} \right)^{1-y_{ji}}. \tag{12}$$

Since the integral in this likelihood function does not have a simple analytical solution, we employ a numerical approach to calculate the integral, using Gauss-Hermite quadrature.

2.5 Parameter Significance Testing

The binary logistic regression with and without the random effect model must be tested to determine whether the predictor variables significantly affect the response variable collectively [17], with the following hypothesis: $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ and H_1: at least $\beta_k \neq 0$ with $k = 1,2,\ldots,p$. The test statistic used is the G test, as follows.

$$G = \mathrm{U}(\tilde{\beta}_0)^T \mathbf{I^{11}}(\tilde{\beta}_0) \mathrm{U}(\tilde{\beta}_0), \tag{13}$$

with $\mathrm{U}(\beta) = \frac{\partial \ell}{\partial \beta}$ is score function, while $\tilde{\beta}_0$ is the maximum likelihood of the logistic regression model under H_0, $G \sim \chi^2_q$, and $\mathbf{I^{11}}$ is submatrix of the inverse Fisher information for β_1.

The likelihood ratio test statistic follows the Chi-Square distribution, denoted as χ^2 (Chi-Square). The criteria for rejecting H_0 if the value of $G > X^2_{\alpha,p}$ or the value of $p - value < \alpha$, indicating that at least one predictor variable has a significant effect on the response variable, where p is the number of predictor variables. Subsequently, partial testing was conducted to assess the influence of each predictor variable on the response variable, with the following hypothesis: $H_0: \mathrm{L}\beta_k = 0$ and $H_1: \mathrm{L}\beta_k \neq 0$, $k = 1,2,\ldots,p$ with $\mathbf{L}$ is a contrast matrix. The test statistic used is the Wald test as follows.

$$W = (\mathbf{L}\hat{\boldsymbol{\beta}})^T \left[\mathbf{L}\widehat{\boldsymbol{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{L}^T \right]^{-1} (\mathbf{L}\hat{\boldsymbol{\beta}}), \tag{14}$$

with $W \sim \chi^2_q$, where q is rank of $\mathbf{L}$.

The testing criteria state that H_0 should be rejected if $|W| > Z_{\frac{\alpha}{2}}$ or $p - value < \alpha$. This indicates that a predictor variable has a significant effect on the response variable.

2.6 Extremely Randomized Trees (Extra Trees)

Extra trees are a development of the random forest method where the selection of features and threshold values are chosen randomly to obtain a very high tree diversity. Thus, this method allows for regression and classification with its accuracy and high tree diversity and can prevent overfitting, which is also robust to noise. This approach uses an ensemble strategy to classify data, which results in a extremely randomized tree that is a union of single trees with a high degree of randomization. In addition, the feature selection and point assignment for node splitting is based on the randomization technique. Thus, the training data creates each extremely randomized tree to reduce errors [10].

In general, extra trees have the following stages of analysis using training data.

1. Stages of selecting the best split.
   a. Randomly select m features.
   b. Randomly selecting k-cut points.
   c. Determining the best splitting criteria.
   d. Repeat steps 1 to 3 until the stopping criterion is reached, resulting in the prediction of a single tree.
2. Step a is repeated until m trees are formed.
3. Combine the estimation results of each classification tree using the majority votes.

2.7 Support Vector Classification (SVC)

SVC was first introduced in the 1990s to perform regression and outlier detection. However, this method of development can classify high-dimensional data well. In addition, according to previous research by [12], this method is robust to overfitting and effective on non-linear data. Other research on IOP shows that the classification results on YG data with SVC, namely IO%, are accurate. In general, SVC has an analysis stage that uses training data using a kernel technique that replaces the dot product of two vectors in the input space with a function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle \mathcal{F}$ is an implicit mapping from the input space to a high-dimensional feature space for each vector $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and $\phi: \mathcal{X} \to \mathcal{F}$. If α_j is a Lagrange multiplier with absolute value represented by y_j, $y_j = 1$, and $\boldsymbol{x_j}$ is on the positive side of h and the path w and -1 if on the other [18]. Then, the SVC classification function is defined as follows.

$$f(x) = sign\left(\sum_j \alpha_j y_j k\left(x_j, x\right) + b\right). \tag{15}$$

Another comparison approach proposed in this research is the novel classification approach with machine learning extra trees and SVC. Consideration of the novelty of the approach in classifying proposed by this research is aimed at obtaining the approach with the best accuracy. In general, the two approaches the first step in the analysis is to perform initialization parameters having a specification of n-estimators 100 and max-depth 3 for extra trees and for SVC based on radial basis function kernel to map the data to a higher feature space according to the characteristics of the data. The complexity of the model in the radial basis function kernel is symbolized by γ, which is defined as:

$$\gamma = \frac{1}{n\,features * X.var}, \tag{16}$$

where $X.var$ represents the variance of the input features. This adaptive setting ensures that γ is scaled according to the dataset characteristics, preventing extreme values that may lead to overfitting (when γ is too large) or underfitting (when γ is too small) [10].

2.8 Accuracy

Evaluating the performance of a classification model is a critical step in ensuring its effectiveness and reliability. Among various metrics, accuracy is one of the most commonly used measures, where accuracy is defined as the proportion of correct predictions (both positive and negative) made by the model out of the total number of predictions. Accuracy is calculated based on the confusion matrix. The formula is the ratio between the number of correct predictions and the total number of predictions. Mathematically, the accuracy formula is as follows.

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN}, \tag{17}$$

where TP (True Positives) is the number of positive instances correctly predicted by the model, TN (True Negatives) is the number of negative instances correctly predicted by the model. FP (False Positives) is the number of negative instances incorrectly predicted as positive (Type I error). The last, FN (False Negatives), is the number of positive instances incorrectly predicted as negative (Type II error).

3. RESULTS AND DISCUSSION

To understand the characteristics of the data, descriptive statistical analysis is required. The aim is to provide an overview of the data distribution for each variable. In this study, the response variable analyzed is hypertension, and the data representation is as follows.

Table 1. Percentage of Predictor Variables

Parameter	Hypertension (%)	No Hypertension (%)
Hypertension	34.5	65.5

Based on Table 1, there are only two categories of the response variable. A total of 34.5% falls into category 1, representing individuals with hypertension, while 65.5 % falls into Category 0, representing individuals without hypertension. This study uses 12 predictor variables, which are explained as follows:

Table 2. Percentage of Predictor Variables

Parameters	Category	Hypertension (%)	No Hypertension (%)
Height	< 170 cm	17	30.5
	≥ 170 cm	17.5	35
Weight	< 60 kg	16.5	30.5
	≥ 60 kg	11.5	30
Body Mass Index	< 25	11.5	14
	≥ 25	23	51
Waist Circumference	< 85 cm	15.5	17
	≥ 85 cm	19	48.5
Age	< 21 years	22.5	40
	≥ 21 years	12	25.5
Gender	Female	15.5	36
	Male	19	29.5
Father's Occupation	Formal	14.5	33.5
	Informal	20	32
Mother's Occupation	Formal	14	19.5
	Informal	20	46
Father's Education	School	14.5	24.5
	Collage	20	41
Mother's Education	School	14	25
	Collage	20.5	40.5
Salary_1	< 3 milion	15.5	19
	3-6 milion	27	38.5
Salary_2	< 3 milion	14.5	20
	> 6 milion	16.5	49.0

Based on Table 2, height is divided into two categories: 1 for height below 170 cm and 0 for height of 170 cm or above. Among athletes, 17.0% and 17.5% suffer from hypertension, while 30.5% and 35.0% do not have hypertension. Weight is also categorized, with 1 representing weight below 60 kg and 0 representing 60 kg or more. In this category, 18.0% and 16.5% of athletes suffer from hypertension, while 35.5% and 30.0% do not. Body Mass Index (BMI) is calculated as a person's weight in kilograms divided by the square of their height in meters and is divided into two categories: 1 for a BMI below 25 and 0 for a BMI of 25 or more. Among athletes, 11.5% and 23.0% have hypertension, while 14.5% and 51.0% do not. Waist circumference is also split into two categories: 1 for waist circumference under 85 cm and 0 for 85 cm or more, where 15.5% and 19.0% of athletes suffer from hypertension, while 17.0% and 48.5% do not. Age is categorized into 1 for those under 21 and 0 for those 21 or older. In this grouping, 22.5% and 12.0% of athletes suffer from hypertension, while 40.0% and 25.5% do not.

Gender is divided into 1 for females and 0 for males, with 15.5% and 19.0% of athletes having hypertension and 36.0% and 29.5% without. Father's occupation is categorized as 1 for formal jobs and 0 for informal jobs, where 14.5% and 20.0% of athletes suffer from hypertension, while 33.5% and 32.0% do not. Mother's occupation is similarly divided: 1 for formal and 0 for informal. Here, 14.0% and 20.5% of athletes suffer from hypertension, while 19.5% and 46.0% do not. Father's education is categorized as 1 for schooling and 0 for university, where 14.5% and 20.0% of athletes suffer from hypertension, while 24.5% and 41.0% do not. Similarly, mothers' education is split into 1 for schooling and 0 for university, with 14.0% and 20.5% of athletes suffering from hypertension and 25.0% and 40.5% not affected. For parental income, the category "salary_1" is split as 1 for below 3 million and 0 for 3-6 million. Here, 15.5% and 27.0% of athletes suffer from hypertension, while 19.0% and 38.5% do not. The "salary_2" category is 1 for below 3 million and 0 for parental income of 6 million or more, where 14.5% and 16.5% of athletes have hypertension, and 20.0% and 49.0% do not. The minimum, maximum, mean, variance and standard deviation values of the predictor variables are presented in the descriptive statistical analysis table for the predictor variables as follows.

Table 3. Descriptive Statistics of Predictor Variables

Parameter	Min	Max	Mean	Variance	Std. Dev
Height	0	1	0.475	0.2506	0.5006
Weight	0	1	0.535	0.25	0.5
Body Mass Index	0	1	0.26	0.1933	0.4397
Waist Circumference	0	1	0.325	0.2204	0.4695
Age	0	1	0.625	0.2355	0.4853
Gender	0	1	0.31	0.251	0.501
Father's Occupation	0	1	0.48	0.2508	0.5008
Mother's Occupation	0	1	0.335	0.2238	0.4731
Father's Education	0	1	0.39	0.239	0.4889
Mother's Education	0	1	0.39	0.239	0.4889
Salary_1	0	1	0.425	0.2456	0.4955
Salary_2	0	1	0.31	0.2149	0.4636

Source: R output

Table 3 shows that all minimum values are 0, and the maximum values are 1. Height has an average value of 0.475 and a variance of 0.2506. Weight has an average value of 0.535 and a variance of 0.2500. Body Mass Index has an average of 0.26 and a variance of 0.1933. waist circumference has an average of 0.325 and a variance of 0.2204. Age has an average of 0.625 and a variance of 0.2355. Gender has an average of 0.515 and a variance of 0.2510. Father's occupation has an average of 0.48 and a variance of 0.2508. Mother's occupation has an average of 0.335 and a variance of 0.2238. Both the father's and mother's education have the same average and variance of 0.39 and 0.2390, respectively. Salary 1 has an average of 0.425 and a variance of 0.2456. Salary 2 has an average of 0.31 and a variance of 0.2149.

The results of parameter estimation in binary logistic regression, using Maximum Likelihood Estimation (MLE), are presented in Table 4.

Table 4. Descriptive Statistics of Predictor Variables

Parameter	Estimate	Std. Error	Z-value	P-value
Intercept	-1.2322	0.4405	-2.797	0.0052
Height	-0.1098	0.3659	-0.3	0.7642
Weight	-0.0678	0.3521	-0.193	0.8473
Body Mass Index	0.2887	0.3983	0.725	0.4684
Waist Circumference	0.7372	0.3631	2.03	0.0423
Age	0.4605	0.3521	1.308	0.1909
Gender	-0.3877	0.3503	-1.107	0.2685
Father's Occupation	-0.7029	0.3333	-2.109	0.0350
Mother's Occupation	0.3115	0.3614	0.862	0.3888
Father's Education	0.1800	0.3337	0.539	0.5897
Mother's Education	-0.0730	0.3651	-0.2	0.8416
Salary_1	0.3870	0.3486	1.11	0.2669
Salary_2	0.7532	0.3763	2.001	0.0454
G test = 40.072				

Source: R output

Based on Table 4, binary logistic regression model using a log link function as follows:

$$\log\left(\frac{P(y_i=1|x_i)}{P(y_i=0|x_i)}\right) = -1.23216 - 0.10978x_{1i} - 0.06780x_{2i} + 0.28874x_{3i} + 0.73718x_{4i} + 0.46051x_{5i} - 0.38766x_{6i} - 0.70289x_{7i} + 0.31146x_{8i} + 0.17998x_{9i} - 0.07296x_{10i} + 0.38703x_{11i} + 0.75315x_{12i}.$$

Based on Table 4, the results of the simultaneous parameter significance test of binary logistic regression indicate that the G test obtained a result of 40.072, greater than the critical value of Chi-square, 21.026, with α = 0.05 and p = 12. In addition, the p-value obtained is 0.0002486, which is smaller than α = 0.05. Therefore, the decision is to reject H_0, which means that at least one predictor variable affects the response variable.

Meanwhile, based on the results of the partial parameter significance test of binary logistic regression shown in Table 4 above, the statistical value of the Wald test for the waist circumference variable is 2.030,

the father's occupation variable is -2.109, and the salary_2 variable is 2.001. These results are greater than the critical value of $Z_{\frac{\alpha}{2}}$ with α = 0.05.

Additionally, the p-values for the waist circumference variable are 0.04233, for the father's occupation variable is 0.03497, and for the salary_2 variable is 0.04535; all these results are less than α = 0.05. Therefore, the decision is to reject H_0, which means it can be concluded that the waist circumference variable, father's occupation, and salary_2 are factors that have a significant effect on hypertension.

The results of parameter estimation in binary logistic regression with athlete random effects, using Maximum Likelihood Estimation (MLE), are presented in Table 5.

Table 5. Results of Binary Logistic Regression Parameter Estimation with Athletes Random Effects

Parameter	Estimate	Std. Error	Z-value	P-value
Intercept	-1.2299	0.4427	-2.7784	0.00546
Height	-0.1115	0.3663	-0.3043	0.76088
Weight	-0.0664	0.3524	-0.1885	0.85049
Body Mass Index	0.2869	0.3987	0.7196	0.47176
Waist Circumference	0.7369	0.3633	2.0285	0.04251
Age	0.4607	0.3522	1.3083	0.19077
Gender	-0.3866	0.3506	-1.1027	0.27014
Father's Occupation	-0.7023	0.3335	-2.1057	0.03523
Mother's Occupation	0.311	0.3616	0.8601	0.38972
Father's Education	0.1795	0.3338	0.5377	0.59078
Mother's Education	-0.0732	0.3652	-0.2005	0.84112
Salary_1	0.3864	0.3487	1.1082	0.26778
Salary_2	0.7525	0.3765	1.9986	0.04565
G test = 39.993				

Source: R output

Based on Table 5, binary logistic regression model with athletes' random effects using a log link function as follows:

$$\log\left(\frac{P(y_{ji}=1|x_{ji},u_j)}{P(y_{ji}=0|x_{ji},u_j)}\right) = -1.2299 - 0.1115x_{1(1)i} - 0.0664x_{2(1)i} + 0.2869x_{3(1)i} + 0.7369x_{4(1)i7} + 0.4607x_{5(1)i} - 0.3866x_{6(1)i} - 0.7023x_{7(1)i} + 0.3110x_{8(1)i} + 0.1795x_{9(1)i} - 0.0732x_{10(1)i} + 0.3864x_{11(1)i} + 0.7525x_{12(1)i} + u_j$$

$u_j \sim N(0, \sigma^2)$.

with value of σ^2 = 0.002390.

Based on Table 5, the results of the simultaneous parameter significance test of binary logistic regression with random effects indicate that the G test obtained a result of 39.993, greater than the critical value of Chi-square, 21.026, with α = 0.05 and p = 12. In addition, the p-value obtained is 0.0002486, which is smaller than α = 0.05. Therefore, the decision is to reject H_0, which means that at least one predictor variable affects the response variable.

Meanwhile, based on the results of the partial parameter significance test of binary logistic regression with random effects shown in Table 5 above, the statistical value of the Wald test for the waist circumference variable is 2.0285, the father's occupation variable is -2.1057, and the salary_2 variable is 1.9986. These results are greater than the critical value of $Z_{\frac{\alpha}{2}}$ with α = 0.05. Additionally, the p-values for the waist circumference variable are 0.04251, the father's occupation variable is 0.03523, and the salary_2 variable is 0.04565; all these results are less than α = 0.05. Therefore, the decision is to reject H_0, which means it can be concluded that the waist circumference variable, father's occupation, and salary_2 are factors that have a significant effect on hypertension.

After conducting the classification process and evaluating the model's performance, the accuracy of the analysis results is determined. The evaluation compares the predicted and actual classifications in the test dataset. The accuracy measurement reflects the model's ability to classify instances correctly and indicates its effectiveness. The accuracy of the analysis results is shown in Table 6.

Table 6. Accuracy of Model

Model	Accuracy	Sensitivity	Specificity	F1-Score
Binary Logistic Regression	68.5%	69%	68%	68.8%
Extra Trees	67.5%	67%	68%	67.3%
Binary Logistic Regression with Random Effect	63.5%	65%	62%	63.8%
SVC	60.0%	61%	61%	60.5%

Source: Python output

A higher accuracy value indicates that the model is improving. The accuracy value of the binary logistic regression model is higher than that of the binary logistic regression model with random effects. Thus, as shown in Table 6, the binary logistic regression model is better than the binary logistic regression model with random effects.

Based on the results obtained, the best model was identified as binary logistic regression because the accuracy, sensitivity, specificity, and F1-score value of the binary logistic regression model (68.5%, 69%, 68%, and 68.8%) is higher than the other approach. Factors influencing hypertension include waist circumference, father's occupation, and parental income. Athletes with a waist circumference of 85 cm or more have a higher risk of hypertension, consistent with research by [19] using multivariate logistic regression. Waist circumference is an indicator of excess abdominal fat, which is linked to increased hypertension risk because abdominal fat is metabolically active and can cause inflammation and insulin resistance, thereby raising the risk of hypertension. Additionally, athletes whose fathers work in the informal sector also have a higher risk of hypertension, in line with findings by [20] using multivariate linear regression. A father's occupation can influence a child's hypertension risk through psychosocial and lifestyle effects within the family. For example, high-stress, demanding jobs can increase tension at home and set examples of unhealthy habits, such as poor sleep or diet, which may raise hypertension risk in children. Economic instability linked to particular jobs can worsen family stress, adding to hypertension risk in children. This study also found that athletes with parents earning less than 3 million have a higher hypertension risk, consistent with findings by [21]. Parental income indirectly affects hypertension risk in children through environmental, lifestyle, and mental health factors. Children from low-income families often have limited access to nutritious food, healthcare, and environments that support physical activity. Financial instability can increase family stress, affecting the mental health of children and raising their risk of chronic diseases, including hypertension, as they grow. Chronic stress exposure and poor diet are key factors in increasing blood pressure in these children. Next, the odds ratio is used to determine the magnitude of the influence of each significant predictor variable on the response variable. Athletes with a waist circumference of $\geq$85 cm have a 2.09 times greater risk of experiencing hypertension compared to athletes with a waist circumference of $\leq$ 85 cm. Additionally, athletes whose fathers work in the formal sector have a 0.495 times lower risk of hypertension compared to those whose fathers work in the informal sector. Meanwhile, athletes with parents earning less than 3 million have a 2.124 times greater risk of hypertension compared to those whose parents earn more than 6 million.

By using a different approach than previous studies, the results of this study align with existing literature but provide new insights into how sociodemographic and anthropometric factors influence hypertension risk among athletes. The relationship between abdominal obesity (measured by waist circumference) and hypertension is well-established in the general population, and this study further highlights its relevance in the athlete population. The metabolic effects of excess abdominal fat, including increased inflammation and insulin resistance, likely play an important role in the development of hypertension in athletes. This underscores the importance of monitoring waist circumference as part of routine health checks for athletes, particularly those undergoing intensive training regimens. The impact of a father's occupation and family income on hypertension risk reveals a complex interaction between social determinants of health. Psychosocial stress, unhealthy lifestyle habits, and economic instability experienced within the family can increase the risk of hypertension. These findings emphasize the need for a holistic approach to athlete health, which considers physical training and the broader social and economic factors that influence their well-being.

Our findings are consistent with previous studies that have identified a significant relationship between waist circumference, parental income, and hypertension risk. However, this study is one of the few to specifically examine these factors in athletes, a population with unique health dynamics. Unlike earlier research focused on the general population, our study considers the impact of intensive training and physical stress on cardiovascular health in athletes. Including sociodemographic factors such as the father's occupation

and parental income provides new perspectives, emphasizing the role of family environment and socioeconomic status in shaping health outcomes. Nevertheless, several research gaps need to be addressed. Future studies could explore additional factors that may influence hypertension risk in athletes, such as genetic predisposition, mental health factors like stress, and training-related variables. Furthermore, examining the interaction between physical activity levels and sociodemographic and anthropometric factors could provide a deeper understanding of how lifestyle factors mediate hypertension risk in this population. Longitudinal studies also help determine the causality of these factors and their long-term effects on athletes' cardiovascular health.

Referring to the accuracy of the prediction results carried out by this study, the binary logistic regression approach method is the method with the best accuracy, which is in accordance with previous research on [22], which predicts flood problems where the prediction strength is based on the highest accuracy value compared to other proposed approach methods such as support vector classifier, k-nearest neighbors, and decision tree classifier. The study's results, which are also in line with the findings, are in the [23] study, which discusses the problem of customer churn. The study found that the prediction accuracy of the binary logistic regression method is the highest compared to other approaches such as extremely randomized trees, adaboost, k-nearest neighbors, random forests, and support vector classifiers.

The findings of this study can provide significant benefits to the development of knowledge and the practice of athlete health. First, this research can be used to design more effective hypertension prevention programs by considering factors such as waist circumference, father's occupation, and family income. Second, coaches and sports professionals can identify high-risk athletes and provide healthier training and lifestyle strategies. Finally, this study opens up opportunities for further research into the relationship between socioeconomic factors, lifestyle, and cardiovascular health in athletes, as well as the importance of comprehensive health monitoring.

4. CONCLUSION

Based on the analysis and discussion in this study, where the four methods proposed by this study, it can be concluded that the binary logistic regression model is better than other approaches (binary logistic regression with the addition of random effects, extremely randomized trees, and support vector classification) this measure of goodness is based on the resulting accuracy, sensitivity, specificity, and F1-score value where this method has the highest value. In addition, the best approximation method found that waist circumference, father's occupation, and salary were significant factors influencing hypertension in athletes at the 5% significance level. Thus, these findings suggest that sociodemographic and anthropometric factors should be considered in assessing the risk of hypertension in athletes. These insights may provide a basis for more effective interventions to reduce the risk of hypertension and improve cardiovascular health among athletes. Based on the results, concrete solutions by tailored training regimens, personalized nutritional plans, and regular blood pressure monitoring are recommended to prevent hypertension in athletes.

Author Contributions

A'yunin Sofro: Conceptualization, Data Curation, Funding Acquisition, Investigation, Project Administration, Resources, Writing - Original Draft. Asri Maharani: Supervision, Validation, Writing - Review and Editing. Mutia Eva Mustafidah: Formal Analysis, Writing - Original Draft. Khusnia Nurul Khikmah: Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft. Affiati Oktaviarina: Validation; Writing - Review and Editing. Danang Ariyanto: Validation; Writing - Review and Editing. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research has been funded by the Directorate General of Higher Education, Ministry of Higher Education, Science and Technology.

Acknowledgment

The authors express their gratitude to the Directorate General of Higher Education, Ministry of Higher Education, Science and Technology, for the basic research grant that provided funding to support this research.

Declarations

The authors declare no competing interest.

Declaration of Generative AI and AI-assisted Technologies

The authors declare that no generative AI or AI-assisted technologies were used in the preparation of this manuscript, including for writing, editing, data analysis, or the creation of tables and figures.

REFERENCES

[1] F. D. Fuchs and P. K. Whelton, "HIGH BLOOD PRESSURE AND CARDIOVASCULAR DISEASE," *Hypertension*, vol. 75, no. 2, pp. 285–292, 2020. doi: https://doi.org/10.1161/hypertensionaha.119.14240

[2] E. de P. F. Resende, J. J. L. Guerra, and B. L. Miller, "HEALTH AND SOCIOECONOMIC INEQUITIES AS CONTRIBUTORS TO BRAIN HEALTH," *JAMA Neurol*, vol. 76, no. 6, pp. 633–634, 2019.doi: https://doi.org/10.1001/jamaneurol.2019.0362.

[3] R. Brathwaite, E. Hutchinson, M. McKee, B. Palafox, and D. Balabanova, "THE LONG AND WINDING ROAD: A SYSTEMATIC LITERATURE REVIEW CONCEPTUALISING PATHWAYS FOR HYPERTENSION CARE AND CONTROL IN LOW-AND MIDDLE-INCOME COUNTRIES," *Int J Health Policy Manag*, vol. 11, no. 3, p. 257, 2020. doi: https://doi.org/10.34172/ijhpm.2020.105.

[4] V. Schweiger, D. Niederseer, C. Schmied, C. Attenhofer-Jost, and S. Caselli, "ATHLETES AND HYPERTENSION," *Curr Cardiol Rep*, vol. 23, pp. 1–11, 2021. doi: https://doi.org/10.1007/s11886-021-01608-x

[5] C. E. Chukwu, O. A. T. Ebuehi, J. N. A. Ajuluchukwu, and A. H. S. Olashore, "ANTHROPOMETRIC, SOCIO-DEMOGRAPHIC AND BIOCHEMICAL RISK FACTORS OF HYPERTENSION IN LAGOS, NIGERIA," *Alexandria Journal of Medicine*, vol. 57, no. 1, pp. 44–51, 2021. doi: https://doi.org/10.1080/20905068.2021.1874626

[6] S. Kawasoe *et al.*, "ASSOCIATION BETWEEN ANTHROPOMETRIC INDICES AND 5-YEAR HYPERTENSION INCIDENCE IN THE GENERAL JAPANESE POPULATION," *Hypertension Research*, vol. 47, no. 4, pp. 867–876, 2024. doi: https://doi.org/10.1038/s41440-023-01505-6 .

[7] B. M. Heo and K. H. Ryu, "PREDICTION OF PREHYPERTENISON AND HYPERTENSION BASED ON ANTHROPOMETRY, BLOOD PARAMETERS, AND SPIROMETRY," *Int J Environ Res Public Health*, vol. 15, no. 11, p. 2571, 2018. doi: https://doi.org/10.3390/ijerph15112571

[8] J. M. Hilbe, *PRACTICAL GUIDE TO LOGISTIC REGRESSION*. CRC Press, Taylor & Francis Group Boca Raton, USA, 2016. doi: https://doi.org/10.1201/b18678

[9] D. A. Pisner and D. M. Schnyer, "CHAPTER 6 - SUPPORT VECTOR MACHINE," in *Machine Learning*, A. Mechelli and S. Vieira, Eds., Academic Press, 2020, pp. 101–121. doi: https://doi.org/10.1016/B978-0-12-815739-8.00006-7

[10] K. N. Khikmah, B. Sartono, B. Susetyo, and G. A. Dito, "PERFORMANCE COMPARATIVE STUDY OF MACHINE LEARNING CLASSIFICATION ALGORITHMS FOR FOOD INSECURITY EXPERIENCE BY HOUSEHOLDS IN WEST JAVA," *Jurnal Online Informatika*, vol. 9, no. 1, pp. 128–137, 2024. doi: https://doi.org/10.15575/join.v9i1.1012

[11] T. V. Rampisela and Z. Rustam, "CLASSIFICATION OF SCHIZOPHRENIA DATA USING SUPPORT VECTOR MACHINE (SVM)," in *Journal of Physics: Conference Series*, IOP Publishing, 2018, p. 012044. doi: https://doi.org/10.1088/1742-6596/1108/1/012044

[12] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A COMPREHENSIVE SURVEY ON SUPPORT VECTOR MACHINE CLASSIFICATION: APPLICATIONS, CHALLENGES AND TRENDS," *Neurocomputing*, vol. 408, pp. 189–215, 2020. doi: https://doi.org/10.1016/j.neucom.2019.10.118

[13] J. K. Harris, "PRIMER ON BINARY LOGISTIC REGRESSION," *Fam Med Community Health*, vol. 9, no. Suppl 1, p. e001290, 2021. doi: https://doi.org/10.1136/fmch-2021-001290

[14] K. N. Khikmah, I. Indahwati, A. Fitrianto, E. Erfiani, and R. Amelia, "BACKWARDS STEPWISE BINARY LOGISTIC REGRESSION FOR DETERMINATION POPULATION GROWTH RATE FACTOR IN JAVA ISLAND," *Jambura Journal of Mathematics*, vol. 4, no. 2, pp. 177–187, 2022. doi: https://doi.org/10.34312/jjom.v4i2.13529

[15] A. Zaidi and A. S. M. Al Luhayb, "TWO STATISTICAL APPROACHES TO JUSTIFY THE USE OF THE LOGISTIC FUNCTION IN BINARY LOGISTIC REGRESSION," *Math Probl Eng*, vol. 2023, no. 1, p. 5525675, 2023. doi: https://doi.org/10.1155/2023/5525675

[16] S. Muff, J. Signer, and J. Fieberg, "ACCOUNTING FOR INDIVIDUAL-SPECIFIC VARIATION IN HABITAT-SELECTION STUDIES: EFFICIENT ESTIMATION OF MIXED-EFFECTS MODELS USING BAYESIAN OR FREQUENTIST COMPUTATION," *Journal of Animal Ecology*, vol. 89, no. 1, pp. 80–92, 2020. doi: https://doi.org/10.1111/1365-2656.13087

[17] A. Agresti, *FOUNDATIONS OF LINEAR AND GENERALIZED LINEAR MODELS*. John Wiley & Sons, 2015.

[18] C. Avcı, M. Budak, N. Yağmur, and F. Balçık, "COMPARISON BETWEEN RANDOM FOREST AND SUPPORT VECTOR MACHINE ALGORITHMS FOR LULC CLASSIFICATION," *International Journal of Engineering and Geosciences*, vol. 8, no. 1, pp. 1–10, 2023. doi: https://doi.org/10.26833/ijeg.987605

[19] J. Y. Sun *et al*, "ASSOCIATION BETWEEN WAIST CIRCUMFERENCE AND THE PREVALENCE OF (PRE) HYPERTENSION AMONG 27,894 US ADULTS," *Front Cardiovasc Med*, vol. 8, p. 717257, 2021. doi: https://doi.org/10.3389/fcvm.2021.717257

[20] A. D. Murray, C. J. McNeil, S. Salarirad, L. J. Whalley, and R. T. Staff, "EARLY LIFE SOCIOECONOMIC CIRCUMSTANCE AND LATE LIFE BRAIN HYPERINTENSITIES–A POPULATION BASED COHORT STUDY," *PLoS One*, vol. 9, no. 2, p. e88969, 2014. doi: https://doi.org/10.1371/journal.pone.0088969

[21] R. G. Victor *et al*, "A CLUSTER-RANDOMIZED TRIAL OF BLOOD-PRESSURE REDUCTION IN BLACK BARBERSHOPS," *New England Journal of Medicine*, vol. 378, no. 14, pp. 1291–1301, 2018. doi: https://doi.org/10.1056/NEJMoa1717250

[22] M. M. A. Syeed *et al*, "FLOOD PREDICTION USING MACHINE LEARNING MODELS," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, IEEE, 2022, pp. 1–6. doi: https://doi.org/10.1109/HORA55278.2022.9800023

[23] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "CUSTOMER CHURN PREDICTION SYSTEM: A MACHINE LEARNING APPROACH," *Computing*, vol. 104, no. 2, pp. 271–294, 2022. doi: https://doi.org/10.1007/s00607-021-00908-y