AJSTME, Volume. 12, Issue 2. 34-43, March, 2026. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Modelling Hypertension Risk with Logistic Regression and Naïve Bayes Using Health and Demographic Metrics

E.G. Sokomba and E.S. Oguntade
Department of Statistics, Faculty of Science, University of Abuja, Nigeria
Corresponding Author: eg.sokomba@gmail.com

Abstract

Hypertension(HTN) has been identified as one of the significant risk factors of morbidities and mortalities globally, particularly in low-middle income countries like Nigeria. Hence supervised machine learning models were applied to predict HTN using prevalent health data. A retrospective observational design was employed and appropriate secondary data extracted from the National Health and Nutrition Examination Survey 2017-2018 dataset sourced from Kaggle repository. After data pre-processing, 3439 complete-cases were used with 11 predictors. Both Logistic Regression and Naïve Bayes classifiers were used as classification algorithms. Chi-square tests were employed to determine associations and receiver operating characteristic (ROC) curve for evaluation of models.Logistic Regression produced 94.6% accuracy with AUC of 0.940 whereas Naïve Bayes classifier recorded 94.1% accuracy with AUC of 0.945.Sex showed no impact and the relationship between systolic/diastolic blood pressure and HTN was high. Both models were of immense use in prediction of HTN.

Keywords: Logistic Regression, Machine Learning, Naïve Bayes, HTN, Predictive Modelling, SBP, DBP ROC, AUC, Risk Factors, Nigeria.

Introduction

Hypertension (HTN) is one of the most prevalent health problems and is a significant problem in public health globally. Metabolic syndrome has multiple factors with a systolic hypertension of 140 mmHg, and or diastolic hypertension of 90 mmHg (Williams et al., 2018; Unger et al., 2020). The exact cause of HTN is unknown. However, risk factors including modifiable (dietary, e.g., high salt intake, low physical activity and obesity) and non-modifiable (e.g. Genetics) are important factors in the development of hypertension (Mills et al., 2020). HTN has become a leading cause of cardiovascular disease (CVD), and one of the biggest contributors to morbidity and mortality in the world (Mills et al., 2020). Although the disease was thought to occur mainly in adults, recent studies indicate that the disease is becoming more widespread in all age groups in the world.The worldwide prevalence of HTN was predicted to affect 1.39 billion people globally, particularly in low and middle income countries (LMICs) (World Health Organization, 2023) where major contributors to the rise of HTN include rapid urbanization, dietary changes, and more sedentary lifestyles (Adeloye et al., 2015; Boateng and Ampofo, 2023). Moreover, in low-resource areas, a lack of access to health care, inadequate screening initiatives, and low awareness levels only compound the problem, thus adding to the economic and health burden of HTN in these areas, where health systems are already stretched thin. Due to the complexity and multi-factorial causes of HTN, a thorough understanding of determinants of HTN prevalence is required (WHO, 2023). These types of insights are crucial in creating effective prevention and control strategies. Despite a vast number of studies investigating the risk factors for HTN, the use of machine learning (ML) techniques in studying these risk factors is less explored and primarily done in Europe and North America with few studies in Sub-Saharan Africa (Islam et al., 2022). ML methods can be used to model the relationships between multiple risk factors, which could be non-linear, and hence they could give more comprehensive predictive insights.

Review of the Literature and Research Gaps

HTN is still one of the world's leading risk factor of cardiovascular disease and early mortality especially among low- and middle-income countries (Mills et al., 2020; WHO, 2023). Several studies report an association between factors such as age, obesity,

physical inactivity, bad diet and markers of increased BP (systolic BP, SBP; diastolic BP, DBP) to HTN (Williams et al., 2018; Unger et al., 2020). A rise in hypertension prevalence in Sub Saharan Africa is also related to factors such as rapid urbanization, changing life style and lack of adequate access to medical care (Adeloye et al., 2015; Boateng and Ampofo, 2023). Conventional statistical methods such as logistic regression were commonly used in HTN research and are known to be useful and interpretable for binary classification problems (Hosmer et al., 2013). But standard statistical methods may be less effective in dealing with the intertwining of multiple clinical variables that are also non-linear in nature. So the use of machine learning (ML) is becoming increasingly researched for predicting and classifying diseases. The value of ML algorithms in predicting hypertension and associated cardiovascular disorders has been shown in recent research. In studies concerning HTN it was found that machine learning (ML) techniques Naïve Bayes, Random Forest, Support Vector Machines (SVM) and Logistic Regression (LR) were extremely effective at prediction (Islam et al, 2022). A study by James et al. (2021), stated that machine learning techniques can be useful in identifying patterns in massive data sets that might not be evident initially which can lead to more accurate predictions. Of these, Naïve Bayes is noted for its simplicity, low computational cost and good classification performance, while Logistic Regression is still widely used for its interpretability. Despite these development, there remain many gaps in the literature. Firstly, there were many research studies that focused on traditional statistical approaches without trying to explore the different machine learning techniques. There has been limited research that has directly compared the prediction power of Logistic Regression and Naïve Bayes for the classification of HTN using simple routinely collected health indicators. Third, very few machine learning methods are used to predict HTN in Sub Saharan Africa and Nigeria, where the prevalence of HTN is on the rise. In addition, some prior studies were specifically about the prevalence estimation and not predictive modelling and classification of risk. Thus, the aim of this research is to address these gaps by applying and comparing two classification models for the prediction of HTN using Health and Demographic Metrics from the NHANES dataset: Logistic Regression and Naïve Bayes model. Additionally, the study will use the chi-square analysis to test associations that may exist between selected variables and HTN status. It is an important study, given the growing application of machine learning methods in public health research, and serves as evidence that could assist in strengthening early identification and screening efforts for HTN.

Methods

Study Design

The study design is a retrospective observational study with a selected dataset available on Kaggle to explore the prevalence and risk factors for HTN. Retrospective studies use existing data to detect patterns, correlations and trends in the data without actively intervening in the data. These studies are appropriate for epidemiological studies. A descriptive analytical method is used for the risk factor distribution and the accuracy of the classification. Predictive models are improved by use of logistic regression and Naïve Bayes machine learning algorithms; more insight into factors associated with HTN was obtained by using the Naïve Bayes algorithm. By combining machine learning with chi square analysis, the study's cross-sectional design allows for an in-depth assessment of HTN at a single time, which can be helpful for public health planning. The results will inform research on the HTN and public health interventions, based on these data, in Nigeria and other parts of the world.

Data Collection

The data collection process is very significant as the quality of the data gathered in this phase will impact the outcome of the analysis. Data collected to be used for this study was secondary data obtained from the chosen kaggle data set of 4,494 data samples and 13 columns (https://www.kaggle.com/datasets/rileyzurrin/national-health-and-nutrition-exam-survey-2017-2018). The variables used in this study consist of range of variables related to demographic characteristics and other HTN related variables. The variables included in the study are: Sex, Age, High-Density Lipoprotein Cholesterol, Total Cholesterol, Triglycerides level, Glycated Hemoglobin, Body Mass Index, Systolic Blood Pressure, Diastolic Blood Pressure, Alanine Aminotransferase level, Creatinine level.

A Binary Logistic Regression Model

Logistic Regression is a statistical analysis technique used for prediction of a binary outcome (usually represented as 0/1) from previous observations of a data set. Logistic regression is a model that uses one or more existing independent variables to predict an outcome (dependent variable). Logistic regression is not truly a

AJSTME, Volume. 12, Issue 2. 34-43, March, 2026. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

regression in the sense of predicting continuous values, but a classification model due to its name "regression. HTN probability was estimated using logistic regression. Logistic regression uses the logistic (sigmoid) function, a transformation of the linear combination of the input variables into a probability ranging from 0 to 1 (Hosmer, Lemeshow, and Sturdivant, 2013; James, Witten, Hastie, and Tibshirani, 2021). Here's the general form:

$$\text{Log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

Where: **p** is the probability of HTN occurrence.
x_i Represents the independent variables.
β_i Are the coefficients.

Variable Description and Model Specification for a Logistic Regression

Based on the available data set on HTN, the model variables were defined and the corresponding logistic regression model is stated in equation (2). The dependent variable is the subjects' HTN Status (y), which is a binary outcome variable coded as 1 or 0. The hypertensive is coded as 1, while non-hypertensive is coded as 0. The independent variables are as stated in Table 2.

Exploring these variables, the model equation for this logistic regression model is represented by equation (1).

$$\text{logit}(P(Y=1) = \beta_0 + \beta_1 X_{sex} + \beta_2 X_{age} + \beta_3 X_{\text{high-density lipoprotein cholesterol}} + \beta_4 X_{\text{total cholesterol}} + \beta_5 X_{\text{triglycerides level}} + \beta_6 X_{\text{glycated hemoglobin}} + \beta_7 X_{\text{body mass index}} + \beta_8 X_{\text{systolic blood pressure}} + \beta_9 X_{\text{diastolic blood pressure}} + \beta_{10} X_{\text{alanine aminotransferase level}} + \beta_{11} X_{\text{creatinine level}} + \text{et} \quad (2)$$

Where:
Y is the dependent variable (HTN status: 0 or 1).
β_0 Is the intercept of the model.
$\beta_1, \beta_2, \ldots \beta_{11}$ Are the coefficients for the respective independent variables.
et Is the error term.

Gaussian Naïve Bayes Model

Naïve Bayes is a set of simple and powerful probabilistic classifiers based on Bayes' Theorem. Naïve Bayes may be suitable for classification problems where the attributes are considered conditionally independent, given a class label. It is simple, yet it is efficient and easy to interpret, and it is often used for text classification, medical diagnosis and filtering spam. The fundamental concept underlying Naïve Bayes is Bayes' Theorem, a way to revise beliefs based on new information.

Given a class **C** and a set of features X= $(x_1, x_2, \ldots, x_n)$, Bayes' theorem states:

$$P(C/X) = \frac{P(X/C)P(C)}{P(X)} \quad (3)$$

Where:

P(C/X) is the posterior probability, the probability that instance X belongs to class C given the observed features.
P(X/C) is the likelihood, which represents how probable the features X are for a given class C
P(C) is the prior probability of class C (before observing the data).
P(X) is the evidence, which ensures that the result is a valid probability distribution by normalizing the probabilities across all possible classes.

Classification Rule

A classification rule was used to determine the class by the highest probability. This approach enabled the classification of each observation according to the event with the highest probability of occurrence according to the model's probability assessment. Naive Bayes makes an assumption of Gaussian distribution for continuous features (columns) provided the class, i.e., $P(x_i/C) \sim$ N (μc,$\delta^2 c$) where μc, and $\delta^2 c$ are the mean and variance of the feature values for class C.

$$\hat{C} = \arg\max_{c} P(C) \prod_{i=1}^{n} P(x_i/C)$$

Where:

$\hat{c}$ Is the predicted class.
P(C) is the prior probability of class C.
$P(x_i)$ Is the likelihood of feature x_i given class C.

Thus, the classifier assigns the class C that maximizes the posterior probability P(C/X).

Let X = (x_1, x_2) be a person's SBP and DBP.
Let Y be the class label:
Y= 0 means non-hypertensive, Y= 1 means hypertensive.

Model Evaluation Metrics:

The evaluation of a model is an important aspect in the development of a machine learning model as it helps in

determining the most suitable model for the set of data used to predict the future. The performance of the machine learning algorithms cannot be measured on the trained set to prevent overfitting, rather the performance can be measured on the test set. Based on the ML algorithms that are used, (e.g., regression and classification), numerous evaluation metrics are considered to evaluate the performance of ML algorithms. It's generally advisable to use a variety of evaluation metrics to gauge performance of a single ML algorithm as certain ML algorithms might achieve good performance based on one metric while poor performance based on the other; so we can evaluate the correct and optimal performance by using multiple evaluation metrics. Evaluation of ML algorithm performances (classifers) in this research were used to measure the performance of the ML algorithms. Machine learning algorithms were trained using the training set and the performance of each machine learning algorithm were measured on the test set to accomplish this. The performance of binary ML classifiers was measured on the evaluation metrics for the sake of this study.

The Feature Importance (Logistic Regression)

According to the feature importance plot shown in Figure 1, Systolic Blood Pressure (SBP) is the most important feature to determine the HTN status, having a much higher importance score than Diastolic Blood Pressure (DBP). SBP has a strong impact on the risk of having HTN, and DBP, though important, is less influential. This supports the medical evidence which shows that SBP is a better marker of cardiovascular risk. The model highlights the importance of high SBP in the diagnosis and risk stratification of HTN.

Results and Discussion

Descriptive Statistics

Table 1: Descriptive Statistics (Continuous variables)

	Age	HDL_c	TC	Trig	HbA1_c	BMI	SBP	DBP	ALT	creatinine
N	3439	3439	3439	3439	3439	3439	3439	3439	3439	3439
Mean	62.3	1.28	4.74	1.59	6.84	31.5	131	69.9	23.2	79.3
Median	64	1.22	4.65	1.42	6.5	30.6	129	70	21	76
Standard deviation	13.2	0.331	1.04	0.779	1.25	6.42	18.9	12	9.48	22.9
Minimum	22	0.44	1.97	0.248	3.5	15.4	76	34	4	27
Maximum	85	2.25	7.99	4.22	11	51	188	107	55	153

The Table 1 presents descriptive statistics for 3,439 individuals, summarizing key health-related continuous variables.

Age: The average age is 62.3 years with median 64 years, suggesting that the population is predominantly older age. Age range was from 22 to 85 years with SD = 13.2 years (moderate difference between subjects). The mean HDL cholesterol (HDL-c) is 1.28 mmol/L, total cholesterol (TC) is 4.74 mmol/L, with some variation (SD 0.779), indicating that there is a mixture of normal and elevated levels. Triglyceride (Trig) levels vary, with a mean of 1.59 mmol/L and some variation (SD 0.779), suggesting a mix of normal and elevated levels. Glycemic Control (HbA1c) is 6.84%, meaning that many may have prediabetes or diabetes, since HbA1c of 6.84% or above is considered abnormal. The summary statistics indicate that the average BMI of the population is 31.5 kg/m2 with the highest being 51.0 kg/m2, indicating high prevalence of overweight and obesity. Blood Pressure (SBP, DBP): The average systolic blood pressure (SBP) is 131 mmHg while the average diastolic blood pressure (DBP) is 69.9 mmHg. The median values (129/70 mmHg) are near normal, but the elevated maximum SBP (188 mmHg) indicates that there may be some individuals with HTN. Liver Function (ALT) and Kidney Function (Creatinine) showed that the mean ALT is 23.2 U/L, which is in the normal range, however, some had an increase>40 U/L, which is an indication of liver disease. The mean creatinine level is 79.3 µmol/L, which is generally normal renal function, although the highest value

measured was 153 µmol/L indicating that some may have kidney impairment.

Correlation and Significance level Between HTN and Model Variables

Table 2: Significance of Variables with Hypertensive Status

Variable	Correlation (r)	Significance Level	Interpretation
Age	0.197	*** ($p < 0.001$)	Highly significant positive relationship
HDL_c	0.045	** ($p < 0.01$)	Significant but very weak positive relationship
TC	0.045	** ($p < 0.01$)	Significant but very weak positive relationship
Trig	0.022	Not significant	No meaningful relationship
HbA1c	0.009	Not significant	No meaningful relationship
BMI	-0.032	Not significant	No meaningful relationship
SBP	0.782	*** ($p < 0.001$)	Extremely strong and highly significant
DBP	0.287	*** ($p < 0.001$)	Moderate and highly significant
ALT	-0.044	* ($p < 0.05$)	Weak but statistically significant negative relationship
Creatinine	0.060	*** ($p < 0.001$)	Weak but highly significant positive relationship
Hypertensive_status	**1.000**	—	Perfect self-correlation (reference variable)

Table 2 is the matrix of correlations showing the linear relationship between the study variables and HTN status. Levels of significance were $p<0.05$, $p<0.01$, and $p<0.001$. Results from this matrix demonstrate significant linear correlations between a number of the study variables and HTN status, but the magnitude of such correlations varied widely. SBP was most strongly positively correlated with HTN status ($r=0.782$, $p<0.001$) and was the most important variable in the data set. Likewise, there was a moderate positive correlation ($r = 0.287$, $p < 0.001$) with DBP. The results are not surprising as both SBP and DBP are important clinical markers that are utilized for HTN diagnosis (WHO, 2021; International Society of HTN, 2020). Age also had a strong positive correlation with HTN ($r = 0.197$, $p < 0.001$), which means that HTN is more likely to occur with aging. This is consistent with epidemiology data available at the time, which shows that age is a risk factor for high blood pressure (Franklin SS et al., 1997; Williams B et al., 2018). Other markers such as HDL-c and TC showed statistically significant, but very modest positive associations ($r = 0.045$, $p < 0.01$, for both). However, the coefficients are small, so from these data, the contribution to prediction of HTN was not practically relevant. This finding is in line with the study that suggests that lipid parameters may have indirect or weak linkage with the blood pressure parameters compared with the primary hemodynamic factors (Grundy SM et al., 1999). Aminotransferase (ALT) and creatinine, on the other hand, were moderately associated with HTN status but statistically significant. In both cases, the correlation was weak and not statistically significant, as ALT was weakly negatively correlated with the variable ($r = -0.044$, $p < 0.05$), and the same for creatinine ($r = 0.060$, $p < 0.001$). These results are statistically significant but the effect sizes are very small and unlikely to be useful as biomarkers on their own for HTN. A previous study also confirmed that liver enzyme and renal biomarker were more useful as markers of comorbid conditions than as primary markers of HTN (Chobanian AV et al., 2003). In addition, there was no statistically significant association between Trig, HbA1c, BMI and HTN status ($p>0.05$). This implies that, in this dataset these variables have no meaningful linear relationship with HTN. It should be noted, however, that non-correlation does not mean there is no association, since there may be non-linear relationships or interactions (Hosmer DW et al., 2013). In general, the analysis shows that some of the variables are statistically significant but only SBP and DBP have great practical importance. This separation highlights the importance of taking into

account not only statistical significance but also the size of the effects when defining meaningful predictors. The results of these analyses serve as the basis for the next phase of predictive modelling, with Logistic Regression and Naïve Bayes, which are expected to be able to explain model performance more effectively than other models if there are more significant associations between the variables (James G et al., 2021).

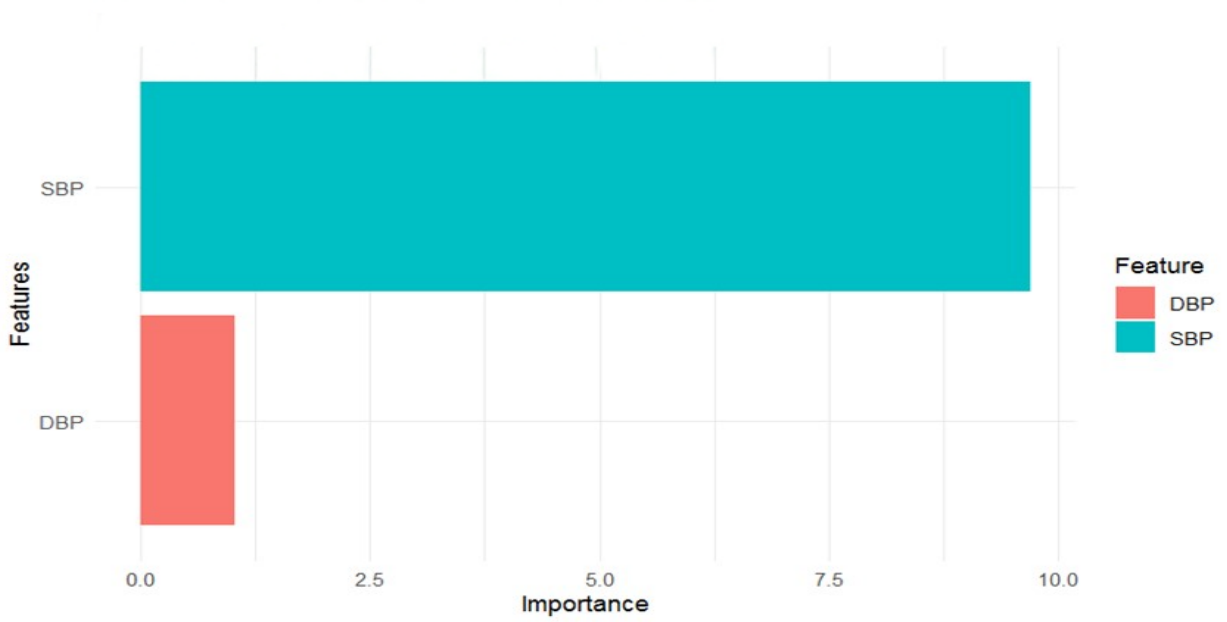
Table 3: **Chi-Square Test for Independence between Sex and Hypertension Status (Age >60 years)**

Sex	Hypertensive	Non-Hypertensive	Total	X^2 value	P-value (Two tailed)	P-value (One tailed)
Female	403	582	985			
Male	380	742	1,122			
Total	783	1,324	2,107	3.45	0.063	0.0315

A

Chi-Square Test for Independence was performed to test the association between sex and prevalence of HTN for those older than age 60 years (Table 3). Data was analyzed using a contingency table summarizing the distribution of HTN status across sex in which 2,107 people participated. Of these, 403 were considered hypertensive and 582 were considered non-hypertensive. Likewise, the number of males who were hypertensive was 380 and the number of males who were non-hypertensive was 742. Result of the test was a Chi-Square value (X^2) of 3.45 and one degree of freedom. The two-tailed P-value (0.063) showed that there was no significant association at the 5% level of significance. The one-tailed p-value (0.0315) however, gave moderate support for the hypothesis that females older than 60 years were more likely to have HTN than male patients. The association was weak, although statistically significant in the one-tailed test, in terms of Cramér's V (0.12). These findings indicate that sex is a potentially important determinant of the prevalence of HTN, but other factors are probably more important. Finally, it can be concluded that females older than 60 years had a higher prevalence of hypertension than males. But, the small effect size suggests that the practical significance of this difference is small.

Figure 1: The feature importance for the logistic regression.



Logistic Regression Model Estimate

The logistic regression model gives an estimate of the probability of an individual being hypertensive (Hypertensive_status = 1) given their SBP and DBP. It provides a model of the relationship between these predictors and the log odds of being hypertensive as follows:

$$\text{Log}\left(\frac{p(hypertensive)}{1-p(hypertensive)}\right) = -63.847 + 0.423 * SBP_{systolic} + 0.081 * DBP_{diastolic}$$

Table 4: Logistic Regression Model Estimates for HTN Prediction

Predictor	Estimate (β)	Odds Ratio (95% CI)	P-value
Intercept	-63.847	–	0.001
SBP	0.423	1.527	<0.001
DBP	0.081	1.084	<0.001

To explore how the SBP and DBP affect the risk of HTN, Logistic regression was performed and the outcome was showed in Table 4. SBP and DBP were found to be important predictors of HTN. For SBP, it was found to have a significant positive effect on HTN, the odds of being HTN would increase by 52.7% for a unit (1mmHg) increase in SBP (=0.423; odds ratio=1.527; p<0.001). As for DBP, the result was showed to have a significant positive association with the outcome (=0.081; odds ratio=1.084; p<0.001) and there was a statistically significant association between the two when SBP was controlled for. The intercept was negative (β = -63.847) indicating a low base level probability of HTN when predictor variables are controlled. In conclusion, the results suggest that both SBP and DBP are important and independent risk factors for HTN and that there is a stronger relationship between SBP and HTN.

AJSTME, Volume. 12, Issue 2. 34-43, March, 2026. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Normal Distribution Modelling of Blood Pressure by Hypertensive Status

Table 5: Estimated Parameters of the Gaussian Model

Predictor	Class	Mean	SD
SBP	0	120.7945	11.24603
SBP	1	155.2164	15.84139
DBP	0	67.94049	10.45921
DBP	1	75.30569	16.33763

Tables 5 shows the parameters of the Gaussian (Normal distribution) of two predictors SBP and DBP for two classes 0 (non-hypertensive) and 1 (hypertensive), in terms of mean and standard deviation (SD). They are usually employed in Gaussian Naïve Bayes classification where there is a normal distribution for each feature for each class. SBP: In the case of class 0 (non-hypertensive), the mean SBP was about 120.79 mmHg with a standard deviation of 11.25 mmHg. The mean SBP for class 1 (hypertensive) is significantly higher at 155.22 mmHg with a SD of 15.84 mmHg. This means that people with hypertension have significantly higher blood pressures of systolic. Class 0: Mean DBP is 67.94 mmHg and Standard Deviation of DBP is 10.46 mmHg. The mean DBP for class 1 is 75.31 mmHg with a larger standard deviation (16.34 mmHg). Again this is because on average, people who are hypertensive have higher diastolic pressure. These parameters indicate that both SBP and DBP are valid parameters to differentiate between hypertensives and non-hypertensives. The high difference of means between the two classes is evidence for their discriminative power for classification with the Gaussian Naïve Bayes model. The class that is most likely to have occurred is chosen as the prediction. The Gaussian Naive Bayes model is efficient and is very good even when simplifying assumptions are used. It categorizes people by calculating the likelihood of being a hypertensive or non-hypertensive according to likelihood of their SBP and DBP being in that category.

Table 6: Comparison of Evaluation Metrics for Two Machine Learning Models

Model	Accuracy	Recall	F1-Score	Kappa	AUC
Logistic Regression	0.9461	0.9265	0.9097	0.8714	0.940
Naïve Bayes	0.9407	0.9559	0.9043	0.8616	0.945

To compare Logistic Regression and Naïve Bayes machine learning models, the performance measures including accuracy, recall, F1-score, Kappa and AUC (Area under the Curve) were utilized as in Table 6.

I. **Accuracy:** Logistic Regression closely matches with Naïve Bayes with accuracy of 94.6% and 94.1% respectively. This means both models are performing the same with Logistic Regression predicting the correct classifications slightly more accurately.
II. **Recall:** Naïve Bayes has a higher recall of 95.6% while Logistic Regression has a recall of 92.7%. This means that Naïve Bayes can be better at identifying all relevant positive instances with fewer false negatives, is better at detecting the positive class, and it is more effective at detecting the positive class.
III. **F1-Score:** Logistic Regression has a higher F1-score of 91.0%, compared to 90.4% for Naïve Bayes. Both models achieve high performance, but Logistic Regression achieves better performance in terms of precision/recall balance and a higher F1-scores.
IV. **Kappa**: The Naive Bayes has a higher Kappa score of 0.8616 and Kappa: Logistic Regression a score of 0.8714. The Kappa statistic gives the degree of agreement between the model prediction and the actual label, and there is significant agreement between the two models. Logistic Regression, however, has slightly better agreement.
V. **AUC (Area Under the Curve)**: Naïve Bayes (0.945) outperforms Logistic Regression (0.940), suggesting Naïve Bayes has a slightly better classification ability on different thresholds.

To sum up, both models exhibit excellent performance. Logistic Regression has the highest accuracy, F1-score and Kappa values, while Naïve Bayes shows the best recall and AUC values. Either model may be appropriate depending on the specific objectives of the task (some prefer to focus on recall over precision and vice versa), but Logistic Regression is somewhat better overall for the two models.

Discussion of Findings

ML techniques (Logistic Regression and Gaussian Nave Bayes) and Chi-square were used to predict HTN. HTN prevalence in the study population was found to be significant. This is consistent with other reports published both globally and regionally that indicated HTN as a major public health concern particularly in LMCs. (Mills et al., 2018; WHO, 2023) Logistic regression result revealed that SBP and DBP both were significantly predictive of HTN with SBP having higher magnitude of effect (= 0.423; odds ratio=1.527; $p<0.001$) compared to DBP (= 0.081; odds ratio=1.084; $p<0.001$), suggesting that increased values of SBP has higher contribution in influencing probability of developing HTN. This is similar to previous clinical study which indicated importance of SBP for predicting cardiovascular risk factors more prominently among the elderly population (Franklin et al., (1997); Williams et al., (2018)). The negative intercept reflects the low probability of HTN at low blood pressure values, highlighting the role of blood pressure values in determining HTN classification. The mean value of DBP and SBP correctly classified between the hypertensive and non-hypertensive groups in the Gaussian Naïve Bayes model. The non-hypertensive persons had significantly lower SBP (120.79 mmHg) and DBP (67.94 mmHg) compared to the hypertensive (155.22 mmHg and 75.31 mmHg, respectively). These differences suggest that blood pressure measurements are effective for determining the status of HTN and warrant their inclusion as significant predictors in classification models (Chobanian et al., (2003); Unger et al., (2020)). There was no statistically significant sex association with HTN for the population ≥ 60 years old ($p = 0.063$) by the chi-square test. Results from the one-tailed test, however, showed a moderate association suggesting a slightly higher prevalence of HTN in females compared to males in this age group. This result is however with a small effect size (Cramér's $V = 0.12$) which suggests that sex is not a strong predictor of HTN and other factors like lifestyle, comorbidities and family history may be more influential (Adeloye et al., (2015); Boateng et al., (2023)). The prediction accuracy of both Logistic Regression and Naïve Bayes was good, at 94%. Logistic Regression also had better accuracy (94.6%), F1 Score (91.0%)) and Kappa (0.8714) than Naïve Bayes, indicating a better balance of precision and recall, and higher agreement with the observed results. Naïve Bayes had higher recall (95.6%) and AUC (0.945), which means that it is more accurate in identifying hypertensive cases as well as in classification between classes at thresholds. In spite of Logistic Regression being a more well-rounded and interpretable model, the results indicate that Naïve Bayes could be better for scenarios where there is a significant focus on avoiding false negatives (Islam et al., (2022); James et al., (2021). Overall results of the present study are in line with previous studies and highlight the importance of systolic and diastolic blood pressure as critical factors for prediction of HTN. Furthermore the study indicated the capacity of machine learning models in accurate prediction of HTN and their potential to be implemented in the Clinical Decision Support System (CDSS) (Islam et al., (2022).

Conclusion

This study provided valuable insights on risk factors present in the data set contributing to HTN. SBP and DBP were found to be both independent significant risk factors for HTN. It has been observed that SBP has a more significant correlation with the risk of developing HTN compared to DBP. The comparison among machine learning models established the suitability of the Logistic Regression model
and Naïve Bayes model performed well in terms of prediction. Logistic Regression achieved slightly better accuracy, F1-score and Kappa values, while Naïve Bayes achieved higher recall scores, and marginally higher Area under the Curve (AUC). The results demonstrated that both models were effective for prediction of HTN, and depending on the purpose of the analysis, the model to be used should be selected. Additionally, the chi-square revealed that sex was not a significant determinant for HTN in those older than 60 years, but there was a slight association. This discovery highlights the polyetiology of HTN and the importance of exploring a broader spectrum of risk factors when evaluating and treating HTN.

Recommendations

Based on the results of this study, it can be recommended to monitor blood pressure regularly, especially in older people to help identify and treat HTN

early. Emphasizing healthy lifestyle habits (decrease sodium, do lots of exercise, eat a healthy diet, manage weight properly) as important ways to prevent and control HTN should be a focus.

References

1. Adeloye, D., Basquill, C., Aderemi, A. V., Thompson, J. Y., & Obi, F. A. (2015). An estimate of the prevalence of hypertension in Africa: A systematic review and meta-analysis. *PLoS ONE, 10*(7), e0129177. https://doi.org/10.1371/journal.pone.0129177
2. Boateng, E. B., & Ampofo, A. G. (2023). A glimpse into the future: Modelling global prevalence of hypertension. *BMC Public Health, 23*, 1906. https://doi.org/10.1186/s12889-023-16662-z
3. Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., Jr., Jones, D. W., Materson, B. J., Oparil, S., Wright, J. T., Jr., & Roccella, E. J. (2003). The seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension, 42*(6), 1206–1252. https://doi.org/10.1161/01.HYP.0000107251.49515.c2
4. Franklin, S. S., Gustin, W., Wong, N. D., Larson, M. G., Weber, M. A., Kannel, W. B., & Levy, D. (1997). Hemodynamic patterns of age-related changes in blood pressure. *Circulation, 96*(1), 308–315. https://doi.org/10.1161/01.CIR.96.1.308
5. Grundy, S. M., Pasternak, R., Greenland, P., Smith, S., Jr., & Fuster, V. (1999). Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations. *Circulation, 100*(13), 1481–1492. https://doi.org/10.1161/01.CIR.100.13.1481
6. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
7. Islam, M. R., Bhuiyan, M. A. H., Rahman, M. M., Rahman, M. S., & Rahman, M. A. (2022). Machine learning approaches for predicting hypertension and identifying its risk factors: A systematic review. *Journal of Human Hypertension, 36*(10), 803–815. https://doi.org/10.1038/s41371-021-00588-9
8. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer. https://www.statlearning.com/
9. Mills, K. T., Stefanescu, A., & He, J. (2020). The global epidemiology of hypertension. *Nature Reviews Nephrology, 16*(4), 223–237. https://doi.org/10.1038/s41581-019-0244-2
10. Unger, T., Borghi, C., Charchar, F., Khan, N. A., Poulter, N. R., Prabhakaran, D., Ramirez, A., Schlaich, M., Stergiou, G. S., Tomaszewski, M., & Wainford, R. D. (2020). 2020 International Society of Hypertension global hypertension practice guidelines. *Hypertension, 75*(6), 1334–1357. https://doi.org/10.1161/HYPERTENSIONAHA.120.15026
11. Williams, B., Mancia, G., Spiering, W., Agabiti Rosei, E., Azizi, M., Burnier, M., Clement, D. L., Coca, A., de Simone, G., Dominiczak, A., & Kahan, T. (2018). 2018 ESC/ESH guidelines for the management of arterial hypertension. *European Heart Journal, 39*(33), 3021–3104. https://doi.org/10.1093/eurheartj/ehy339
12. World Health Organization. (2021). *Guideline for the pharmacological treatment of hypertension in adults*. https://www.who.int/publications/i/item/9789240033986
13. World Health Organization. (2023). *Global report on hypertension: The race against a silent killer*. https://www.who.int/publications/i/item/9789240081062