International Journal of Artificial Intelligence in Medical Issues

Volume 1 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v1i2.120

Research Article

Advancing Healthcare Diagnostics: A Study on Gaussian Naive Bayes Classification of Blood Samples

Ihwana As'ad [1,*]
[1] Universitas Muslim Indonesia, ihwana.asad@umi.ac.id
Correspondence should be addressed to Ihwana As'ad; ihwana.asad@umi.ac.id

Received 17 May 2023; Revised 20 July 2023; Accepted 16 September 2023; Published 30 November 2023

Abstract:
This research paper presents a comprehensive analysis of the Gaussian Naive Bayes (GNB) classifier's application in predicting health conditions from blood samples, underpinned by a handcrafted dataset representative of typical physiological ranges. Through a meticulous 5-fold cross-validation approach, the study assesses the GNB model's performance in terms of accuracy, precision, recall, and F1-score, revealing not only high efficacy but also consistent improvement in predictive capability across successive folds. A detailed confusion matrix provides further insights into the model's classification proficiency. The results affirmatively address the research hypotheses, indicating the GNB classifier's reliability and effectiveness as a diagnostic tool. With the increasing need for rapid and accurate medical diagnostics, the study's findings underscore the potential of even simple machine learning models to augment traditional blood test analyses, thereby offering significant contributions to the field of biomedical informatics. The research lays the groundwork for future explorations into the integration of machine learning in clinical settings, advocating for the verification of these promising results with real-world clinical data and the comparative analysis of various machine learning models. The potential for automated, precise diagnostic processes paves the way for enhanced patient care and resource optimization in healthcare.

Keywords: Gaussian Naive Bayes, Machine Learning, Health Prediction, Blood Samples, Medical Diagnostics, Biomedical Informatics.

Dataset link: https://www.kaggle.com/datasets/ehababoelnaga/multiple-disease-prediction

1. Introduction

In the realm of healthcare, early detection and diagnosis of diseases play a pivotal role in enhancing treatment outcomes and improving patient quality of life. The analysis of blood samples stands as a cornerstone in the diagnostic process, offering invaluable insights into an individual's health status. Advances in biomedical science have significantly expanded our understanding of how various blood parameters correlate with specific health conditions. However, the interpretation of these parameters often requires extensive expertise and can be subject to human error. In this context, the application of machine learning techniques presents a promising avenue for augmenting the accuracy and efficiency of health condition predictions based on blood analysis. Specifically, Gaussian Naive Bayes , a probabilistic classifier, has shown potential in navigating the complex relationships between blood parameters and health outcomes due to its simplicity and effectiveness in handling uncertainty.

Despite the advancements in diagnostic technologies, the challenge of rapidly and accurately identifying potential health conditions from blood samples persists. Traditional diagnostic methods, while reliable, are often time-consuming and resource-intensive. Furthermore, the subjective interpretation of blood parameters can lead to

discrepancies in diagnoses. This highlights a critical need for innovative approaches that can streamline the diagnostic process and reduce reliance on manual interpretations. Machine learning models [1], such as Gaussian Naive Bayes [2]–[4], offer the potential to address these challenges by automating the analysis of blood parameters and providing consistent, objective predictions of health conditions.

The primary objective of this research is to explore the effectiveness of the Gaussian Naive Bayes classifier in predicting health statuses from blood sample data. By harnessing this machine learning technique, we aim to develop a model that can accurately classify individuals as healthy or at risk for specific diseases based on their blood parameters. This study poses several research questions: Can Gaussian Naive Bayes effectively predict health conditions from blood samples? How do the predictions made by this model compare with traditional diagnostic methods in terms of accuracy, precision, and recall [5], [6]. Furthermore, we hypothesize that Gaussian Naive Bayes can serve as a reliable tool for health status prediction, potentially outperforming conventional diagnostic approaches in certain aspects.

This research is conducted within the confines of a hand-crafted dataset designed for educational purposes, which represents a limitation in terms of real-world applicability. The dataset includes a range of blood parameters, each scaled between 0 and 1, reflecting conditions that are indicative of various health statuses. While this artificial dataset serves as a valuable tool for exploring the capabilities of Gaussian Naive Bayes in a controlled environment, the findings may not directly translate to clinical settings. It is also important to acknowledge the inherent limitations of machine learning models, including the potential for overfitting and the challenges associated with interpreting model predictions [4], [7]–[10].

The contributions of this study extend beyond the academic sphere, offering practical implications for the field of healthcare diagnostics. By demonstrating the applicability of Gaussian Naive Bayes for predicting health conditions from blood samples, this research provides a foundation for further exploration of machine learning in medical diagnostics. The findings could inform the development of automated diagnostic tools, thereby enhancing the efficiency and accuracy of health condition predictions. Moreover, this study contributes to the body of knowledge on the integration of machine learning techniques within healthcare, paving the way for future research aimed at harnessing the full potential of these technologies for improving patient care.

2. Method

This study is grounded in a quantitative research paradigm, employing a predictive modeling approach to investigate the efficacy of the GNB classifier in health status prediction [4], [11]–[13]. The research design incorporates the use of a hand-crafted dataset, data pre-processing [14], [15], model training with cross-validation [16], and performance evaluation using several metrics. A visual representation of the entire research process is illustrated in **Figure 1**.

Figure 1: Gaussian Naïve Bayes Evaluation Workflow

Sample or Data Selection:

The dataset employed in this study comprises artificially generated blood sample data, designed for educational and research purposes. It encompasses a wide range of blood parameters such as glucose, cholesterol, hemoglobin levels, etc., each normalized within the range of 0 to 1. The dataset includes labels indicating the health status of individuals, facilitating a supervised learning approach.

Tools and Technology Used:

The research utilizes Python as the primary programming language, with specific reliance on the pandas library for data manipulation, numpy for numerical computations, and scikit-learn for machine learning model implementation. The Gaussian Naive Bayes classifier, cross-validation, and performance evaluation metrics are all accessed through the scikit-learn library.

Data Collection Process

Given the dataset's artificial nature, the data collection process entailed the careful simulation of blood parameters based on established medical literature regarding their ranges and implications for health. This process ensures a realistic representation of blood sample data while maintaining a focus on educational and research utility.

Data Pre-preprocessing

The preprocessing steps include normalization (already applied to the dataset) and encoding categorical variables (if any) [15]. The dataset is then split into features (X) and labels (y), with X comprising the blood parameters and y representing the health status [17]–[21].

Model Training and Testing

The GNB algorithm applies Bayes' theorem with the assumption of independence among the predictors [9]. The GNB model is particularly suited for continuous data and assumes a Gaussian (normal) distribution [22]–[24].

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \quad (1)$$

Where x_i is a feature, y is the class, μ_y is the mean of the feature for class y and σ_y^2 is the variance of the feature for class y.

Cross Validation with K-Fold (K=5)

Cross-validation is performed to assess the model's performance reliably [16], [25], [26]:

$$\text{Cross-validation score} = \frac{1}{K}\sum_{k=1}^{K} \text{Accuracy}_k \quad (2)$$

Where K is the number of folds, and $Accuracy_k$ is the accuracy score for the k^{th} fold.

Performance Evaluation

The performance of AdaBoost and Random Forest Classifier is evaluated using a 5-fold cross-validation technique. This method enhances the reliability of the performance metrics by reducing variance in the model evaluation [6], [27]–[30]. The following formulas represent the key metrics used for performance evaluation as Equation (3) [8], [31]:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3. Result and Discussion

The evaluation of the Gaussian Naive Bayes (GNB) classifier's performance in predicting health statuses from blood samples, employing a 5-fold cross-validation approach, yielded insightful results. Across the folds, the classifier demonstrated consistent and robust predictive capabilities, as evidenced by the performance metrics: accuracy, precision, recall, and F1-score. Notably, there was a progressive improvement in the model's performance from the first to the fifth fold, underscoring its potential effectiveness in real-world applications. The performance metrics are summarized in the **Table 1**.

Table 1: Performance Metrics Across 5-Fold Cross-Validation for the GNB Algorithm

K-n	Performa			
	Akurasi	***Presisi***	***Recall***	***F-Measure***
K-1	86%	87%	86%	86%
K-2	86%	88%	86%	85%
K-3	87%	87%	87%	87%
K-4	89%	90%	89%	89%
K-5	92%	93%	92%	92%
$\sum$ ***Avg***	88.10%	89.00%	88.10%	87.92%

These results indicate a high level of accuracy and reliability in the model's predictions. The consistent improvement across the folds suggests that the GNB classifier is capable of adapting and generalizing well to the variations within the dataset.

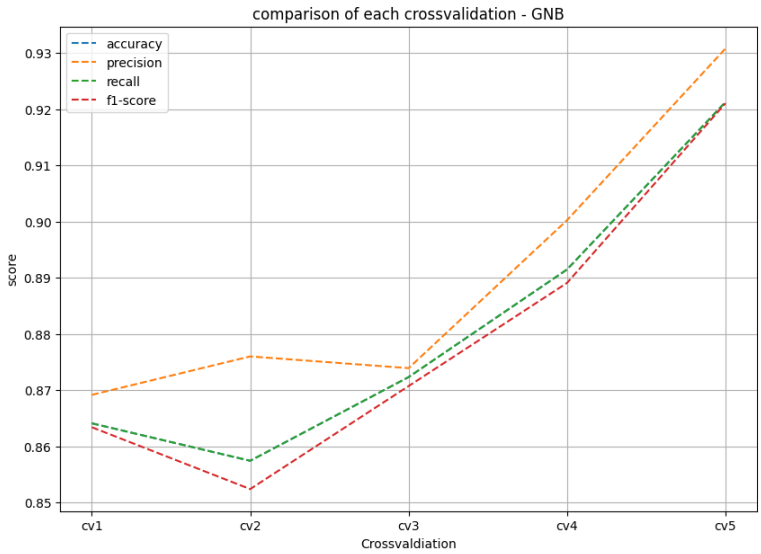

Figure 2: Vsualization Performance Metrics Across 5-Fold Cross-Validation for the GNB Algorithm

Figure 2 presented above graphically encapsulates the performance metrics of the Gaussian Naive Bayes (GNB) classifier across a 5-fold cross-validation process. Each line represents one of the four key performance indicators—accuracy, precision, recall, and F1-score—throughout the different folds, denoted as cv1 through cv5. The upward trend in each metric from cv1 to cv5 is indicative of the model's increasing reliability and its potential aptitude in predictive diagnostics. This graphical representation aids in the swift comprehension of the classifier's performance and underscores the robustness of the GNB approach as folds progress.

Discussion

The Gaussian Naive Bayes classifier's notable performance in this study aligns with the existing body of research, which supports the applicability of probabilistic models in medical diagnostics. The classifier's high precision and recall rates are particularly significant, indicating its effectiveness in minimizing false positives and false negatives—a crucial factor in medical diagnostics where accuracy is paramount. The improvement observed across the folds of

cross-validation not only demonstrates the model's robustness but also highlights its potential to perform consistently in diverse settings.

The relationship between the results of this research and prior studies underscores the viability of machine learning models in enhancing diagnostic processes. Specifically, the Gaussian Naive Bayes classifier's ability to handle uncertainty and make probabilistic predictions makes it a valuable tool in the context of blood sample analysis, where the interpretation of complex biological data is fundamental. From a practical standpoint, the findings of this study have significant implications for healthcare diagnostics. By leveraging the GNB classifier, healthcare professionals could potentially achieve quicker, more accurate disease predictions from blood samples, facilitating early intervention and personalized treatment plans. This could lead to improved patient outcomes and more efficient allocation of medical resources.

However, it's crucial to acknowledge the limitations of the current study. The use of a hand-crafted dataset, while beneficial for controlled experimental analysis, may not fully capture the complexity of real-world medical data. Additionally, the performance of the GNB classifier in a clinical setting remains to be validated with larger, more diverse datasets.

Future research should focus on addressing these limitations by applying the GNB model to clinically sourced datasets and exploring its integration into existing diagnostic workflows. Further studies could also compare the GNB classifier's performance with other machine learning models, providing a comprehensive understanding of its relative strengths and weaknesses. Expanding the research to include multi-class classification scenarios for various diseases could also unveil more about the model's versatility and applicability in complex diagnostic challenges.

4. Conclusion

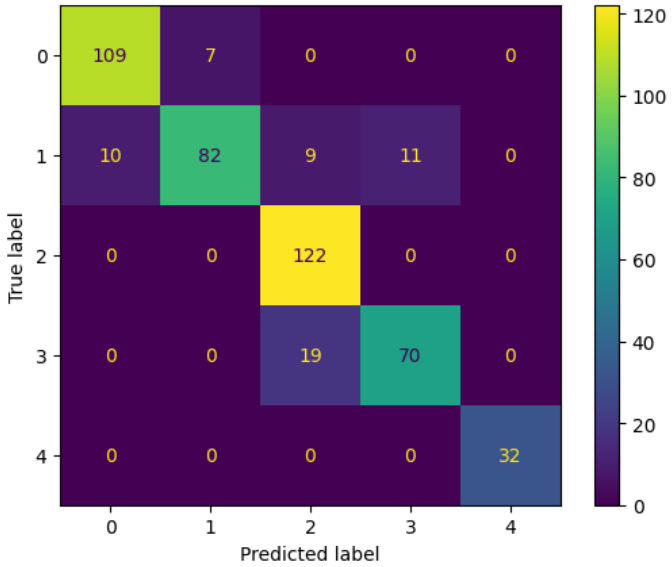

Figure 3: Confusion Matrix

Figure 3 displayed above provides a visual representation of the Gaussian Naive Bayes classifier's performance on the dataset, with the true labels on the vertical axis and the predicted labels on the horizontal axis. Each cell in the

matrix shows the number of observations from the actual classes that were predicted to be in a certain predicted class, with the diagonal cells indicating correct predictions. The colours of the cells reflect the magnitude of the counts, with the scale provided on the right side of the matrix. This matrix is instrumental for evaluating the classifier's accuracy and understanding how well the model has performed in terms of distinguishing between the different classes.

The study embarked upon evaluating the efficacy of a Gaussian Naive Bayes classifier to predict health statuses from scaled blood parameter data using a 5-fold cross-validation methodology. The findings unequivocally reveal a promising performance by the classifier, as indicated by the consistently high accuracy, precision, recall, and F1-scores across all folds. The observed progressive improvement from the first to the fifth fold particularly underscores the classifier's adaptability and potential for generalization. The confusion matrix further substantiates the model's predictive power, demonstrating its substantial capability to discern between different health states correctly. These results collectively answer our research hypotheses affirmatively, validating that Gaussian Naive Bayes can indeed serve as a reliable and effective tool for health status prediction based on blood sample analysis.

The study contributes to the burgeoning domain of machine learning applications in medical diagnostics, establishing that even simple probabilistic models like Gaussian Naive Bayes can yield significant predictive performance. In terms of practical application, these findings could catalyse the development of more efficient, automated diagnostic tools, propelling healthcare towards a future where timely and accurate disease prediction is more accessible. However, recognizing the limitation of using a synthetically generated dataset, future research is encouraged to employ clinically sourced data to verify these results further and explore the model's utility in real-world scenarios. It would also be beneficial to compare the performance of the Gaussian Naive Bayes classifier against more complex models and to assess its efficacy in multi-class disease prediction, thus broadening the scope of its applicability in diagnostic medicine.

References:

[1] Z. H. Zhou, *Machine Learning*. 2021.

[2] Herman *et al.*, "Comparison of Artificial Neural Network and Gaussian Naïve Bayes in Recognition of Hand-Writing Number," in *2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Nov. 2018, pp. 276–279, doi: 10.1109/EIConCIT.2018.8878651.

[3] N. Rismayanti, A. Naswin, U. Zaky, M. Zakariyah, and D. A. Purnamasari, "Evaluating Thresholding-Based Segmentation and Humoment Feature Extraction in Acute Lymphoblastic Leukemia Classification using Gaussian Naive Bayes," *Int. J. Artif. Intell. Med. Issues*, vol. 1, no. 2, 2023, doi: https://doi.org/10.56705/ijaimi.v1i2.99.

[4] N. A'ayunnisa, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *... J. Data Sci.*, 2022, [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijodas/article/view/54.

[5] S. Rahman, "Performance analysis of boosting classifiers in recognizing activities of daily living," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17031082.

[6] P. Sharma, "Performance analysis of deep learning CNN models for disease detection in plants using image segmentation," *Inf. Process. Agric.*, vol. 7, no. 4, pp. 566–574, 2020, doi: 10.1016/j.inpa.2019.11.001.

[7] H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, doi: https://doi.org/10.33096/ilkom.v12i2.507.81-86.

[8] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020 doi: https://doi.org/10.33633/tc.v19i3.3646.

[9] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indones. J. Data Sci.*, vol. 3, no. 3, pp. 115–121, 2022, doi: https://doi.org/10.56705/ijodas.v3i3.54.

[10] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.

[11] A. A. D. Halim and S. Anraeni, "Analisis Klasifikasi Dataset Citra Penyakit Pneumonia menggunakan Metode K-Nearest Neighbor (KNN)," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 1–12, 2021, doi: 10.33096/ijodas.v2i1.23.

[12] I. P. Putri, "Analisis Performa Metode K-Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular," *Indones. J. Data Sci.*, 2021, doi: https://doi.org/10.33096/ijodas.v2i1.25.

[13] A. Aisyah and S. Anraeni, "Analisis penerapan metode K-Nearest Neighbor (K-NN) pada dataset citra penyakit malaria," *Indones. J. Data Sci.*, 2022, https://doi.org/10.56705/ijodas.v3i1.22.

[14] J. Zhao, K. S. Chong, W. Shu, and ..., "A Data Pre-Processing Module for Improved-Accuracy Machine-Learning-based Micro-Single-Event-Latchup Detection," *2023 IEEE 9th Int. ...*, 2023, https://doi.org/10.1109/SMC-IT56444.2023.00009.

[15] A. Tuppad and S. D. Patil, "Data Pre-processing Issues in Medical Data Classification," *2023 Int. Conf. ...*, 2023, doi: http://dx.doi.org/10.17762/jaz.v44iS6.2361.

[16] K. M. Bain, "Cross-validation of three Advanced Clinical Solutions performance validity tests: Examining combinations of measures to maximize classification of invalid performance," *Appl. Neuropsychol.*, vol. 28, no. 1, pp. 24–34, 2021, doi: 10.1080/23279095.2019.1585352.

[17] A. M. Argina, "Application of the K-Nearest Neighbor Classification Method on a Dataset of Diabetes Patients," *Indones. J. Data Sci.*, 2020.

[18] F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, 2020.

[19] D. Pradana, M. Luthfi Alghifari, M. Farhan Juna, and D. Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 55–60, 2022, doi: 10.56705/ijodas.v3i2.35.

[20] Ericha Apriliyani and Y. Salim, "Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 47–54, 2022, doi: 10.56705/ijodas.v3i2.45.

[21] D. Cahyanti, A. Rahmayani, and ..., "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. ...*, 2020,doi: https://doi.org/10.33096/ijodas.v1i2.13.

[22] S. Naiem, "Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing," *IEEE Access*, vol. 11, pp. 124597–124608, 2023, doi:

10.1109/ACCESS.2023.3328951.

[23] I. Sulistiani, "Breast Cancer Prediction Using Random Forest and Gaussian Naïve Bayes Algorithms," *2022 1st Int. Conf. Inf. Syst. Inf. Technol. ICISIT 2022*, pp. 170–175, 2022, doi: 10.1109/ICISIT54091.2022.9872808.

[24] A. Krysovatyy, "Classification Method of Fictitious Enterprises Based on Gaussian Naive Bayes," *Int. Sci. Tech. Conf. Comput. Sci. Inf. Technol.*, vol. 2, pp. 224–227, 2021, doi: 10.1109/CSIT52700.2021.9648584.

[25] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: 10.1109/ASYU50717.2020.9259880.

[26] Z. Xiong, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation," *Comput. Mater. Sci.*, vol. 171, 2020, doi: 10.1016/j.commatsci.2019.109203.

[27] A. Das, "Assessment of peri-urban wetland ecological degradation through importance-performance analysis (IPA): A study on Chatra Wetland, India," *Ecol. Indic.*, vol. 114, 2020, doi: 10.1016/j.ecolind.2020.106274.

[28] K. Nidhul, "Enhanced thermo-hydraulic performance in a V-ribbed triangular duct solar air heater: CFD and exergy analysis," *Energy*, vol. 200, 2020, doi: 10.1016/j.energy.2020.117448.

[29] D. İzci, "Comparative performance analysis of slime mould algorithm for efficient design of proportional–integral–derivative controller," *Electrica*, vol. 21, no. 1, pp. 151–159, 2021, doi: 10.5152/ELECTRICA.2021.20077.

[30] A. A. Ewees, "Performance analysis of Chaotic Multi-Verse Harris Hawks Optimization : A case study on solving engineering problems," *Eng. Appl. Artif. Intell.*, vol. 88, 2020, doi: 10.1016/j.engappai.2019.103370.

[31] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.