International Journal of Artificial Intelligence in Medical Issues

Volume 2 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v2i1.139

Research Article

Predicting the Conversion from Clinically Isolated Syndrome to Multiple Sclerosis in Mexican Mestizo Patients Using Gaussian Naive Bayes Classifier: A Prospective Cohort Study

Nurul Kholifah Yulinda [1,*]
[1] Universitas Muslim Indonesia, nurulkholifah770@gmail.com
Correspondence should be addressed to Nurul Kholifah Yulinda; nurulkholifah770@gmail.com

Received 11 March 2024; Revised 18 April 2024; Accepted 20 May 2024; Published 31 May 2024

Abstract:
This study explores the application of the Gaussian Naive Bayes (GNB) classifier to predict the conversion from Clinically Isolated Syndrome (CIS) to Multiple Sclerosis (MS) among Mexican mestizo patients. Utilizing a dataset gathered from the National Institute of Neurology and Neurosurgery in Mexico City, which included patients diagnosed with CIS between 2006 and 2010, we employed a prospective cohort study design. Our approach involved preprocessing the data to handle missing values and scale normalization followed by splitting it into training and testing subsets. The GNB model's performance was assessed through a 5-fold cross-validation, focusing on accuracy, precision, recall, and F1-score. Results demonstrated the model's capability to predict MS conversion with reasonable precision, highlighted by a significant peak in performance metrics in the third fold of the validation. The study addresses a gap in predictive diagnostics for MS within a specific demographic group, providing valuable insights for early intervention strategies. Despite some limitations such as the model's sensitivity to data heterogeneity and the demographic specificity of the cohort, the findings underscore the potential for predictive models in clinical settings. Recommendations for future research include the use of more sophisticated algorithms and broader demographic studies to enhance predictive accuracy and generalizability.

Keywords: Clinically Isolated Syndrome, Multiple Sclerosis, Gaussian Naive Bayes, Predictive Modeling, Prospective Cohort Study.

Dataset link: https://www.kaggle.com/datasets/desalegngeb/conversion-predictors-of-cis-to-multiple-sclerosis

1. Introduction

The study of Multiple Sclerosis (MS), a chronic, immune-mediated disorder that affects the central nervous system, remains a major concern within the medical community due to its unpredictable nature and significant impact on quality of life. Characterized by the deterioration of the protective covers of nerve cells in the brain and spinal cord, MS can lead to a wide range of neurological symptoms, varying in severity from patient to patient. The complexity of its pathogenesis, influenced by genetic and environmental factors, makes MS a challenging disease to predict and manage effectively.

The initial manifestation of MS often presents as a Clinically Isolated Syndrome (CIS), which may evolve into clinically definite multiple sclerosis (CDMS). The transformation from CIS to CDMS is highly variable, prompting the need for a deeper understanding of the factors that predict this progression. The overarching problem this research

aims to address is the identification of reliable predictors that can foresee the conversion from CIS to MS. This has significant implications for early intervention strategies and could potentially alter the course of the disease by enabling timely therapeutic measures.

The primary objective of this research is to identify and analyze various factors, such as age, initial symptoms, and MRI findings, that could serve as predictive indicators of CIS converting into MS among a cohort of Mexican mestizo patients. By doing so, this study hopes to contribute valuable insights into the early management and treatment of MS, specifically tailored to this demographic, which has been underrepresented in global MS research.

Several research questions guide this investigation: What are the most significant predictors of CIS conversion to MS? How do these predictors differ demographically? Moreover, this study hypothesizes that certain initial symptoms and MRI features will have strong predictive value for the conversion from CIS to CDMS.

The scope of this study is confined to patients who have been newly diagnosed with CIS and have presented at the National Institute of Neurology and Neurosurgery in Mexico City between 2006 and 2010. Although focusing on this specific population may limit the generalizability of the findings, it provides a detailed examination of CIS to MS conversion in a genetically and environmentally distinct group, contributing to a more diversified understanding of MS.

This research aims to augment the existing body of knowledge by providing empirical data specific to the Mexican mestizo population. Furthermore, by identifying reliable predictive markers, the study seeks to enhance clinical practices concerning early MS diagnosis and treatment, thereby potentially improving patient outcomes and informing future global research on this debilitating disease.

2. Method

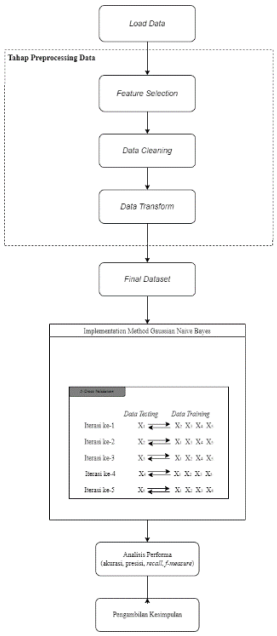

Figure 1: Gaussian Naïve Bayes Evaluation Workflow

This study adopts a prospective cohort design to investigate the predictive factors for the conversion from Clinically Isolated Syndrome (CIS) to Multiple Sclerosis (MS) among Mexican mestizo patients. A comprehensive dataset, originally collected between 2006 and 2010 at the National Institute of Neurology and Neurosurgery in Mexico City, forms the basis of the analysis. The cohort includes patients newly diagnosed with CIS, tracked for conversion to MS over a period defined by clinical follow-up assessments. A visual representation of the entire research process is illustrated in **Figure 1**.

Sample or Data Selection:

The inclusion criteria for the study cohort were adults aged between 18 and 55 years, diagnosed with CIS based on McDonald criteria. Exclusion criteria included prior neurological impairment suggestive of MS or other demyelinating diseases. The final dataset comprised demographic details, clinical findings, and radiological MRI data of 120 patients, ensuring a robust sample size for statistical analysis.

Table 1: Dataset column descriptions

No	Feature	Keterangan
1	Age	Age of the patient (in years)
2	Schooling	time the patient spent in school (in years)
3	Breastfeeding	1=yes, 2=no, 3=unknown
4	Varicella	1=positive, 2=negative, 3=unknown
5	Initial_Symptoms	=visual, 2=sensory, 3=motor, 4=other, 5= visual and sensory, 6=visual and mot or, 7=visual and others, 8=sensory and motor, 9=sensory and other, 10=motor and other, 11=Visual, sensory and motor, 12=visual, sensory and other, 13=Vis ual, motor and other, 14=Sensory, motor and other, 15=visual,sensory,motor a nd other
6	Mono _or_Polysymptomatic	1=monosymptomatic, 2=polysymptomatic, 3=unknown
7	Oligoclonal_Bands	0=negative, 1=positive, 2=unknown
8	LLSSEP	0=negative, 1=positive
9	ULSSEP	0=negative, 1=positive
10	VEP	0=negative, 1=positive
11	BAEP	0=negative, 1=positive
12	Periventricular_MRI	0=negative, 1=positive
13	Cortical_MRI	0=negative, 1=positive
14	Infratentorial_MRI	0=negative, 1=positive
15	Spinal_Cord_MRI	0=negative, 1=positive
16	Initial_EDSS	?
17	Final_EDSS	?
18	Group	1=CDMS, 2=non-CDMS

Tools and Technology Used:

Data were collected and managed using REDCap electronic data capture tools hosted at NINN. MRI scans were analyzed using high-resolution MRI systems, with images assessed by two independent neuroradiologists to determine

lesions characteristic of MS. The statistical analysis was conducted using Python, specifically leveraging libraries such as NumPy for data manipulation and scikit-learn for implementing the Gaussian Naive Bayes algorithm [1]–[3].

Data Collection Process

Data collection involved retrieving and digitizing medical records, MRI reports, and follow-up clinical evaluations from the hospital's database. Variables collected included age, gender, education, initial symptoms of CIS, MRI findings, and the Expanded Disability Status Scale (EDSS) score at onset and follow-up.

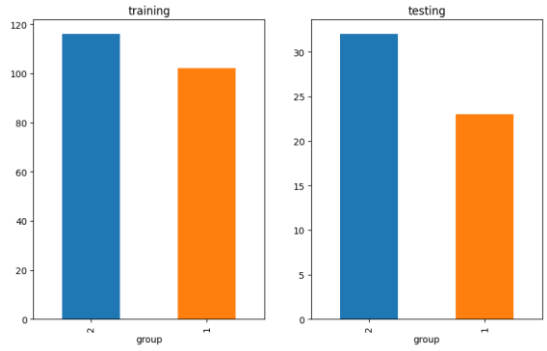

Figure 2: Splitting Data Training (80%), Testing (20%)

Data Analysis Methods

The primary analytical technique used was the Gaussian Naive Bayes classifier, a probabilistic model based on Bayes' theorem with an assumption of independence among predictors. The probability of a patient converting from CIS to MS, given predictor variables x, is modeled as [4]–[7]:

$$\mathrm{P}(\mathrm{Y}|X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_n|Y)P(Y)}{P(X_1, X_2, \dots, X_n)} \quad (1)$$

Where Y represents the class variable (CIS or MS), and $X_1, X_2, \dots, X_n$ are the predictor variables.

$$likehilood = P(X_i|Y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-(x-\mu_y)^2/(2\sigma_y^2)} \quad (2)$$

Here, μ_y and σ_y^2 are the mean and variance of feature X_i for class Y, calculated from the training data.

Data were first preprocessed to handle missing values and normalize the scale of quantitative variables. The dataset was then split into a training set (80%) and a test set (20%) [8]–[11]. The Gaussian Naive Bayes [4] model was trained on the training set, and its performance was evaluated on the test set using metrics such as accuracy, precision, recall, and F1-score.

The performance of the model was evaluated using accuracy, precision, recall, and F1-measure, calculated as follows [12]–[17]:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (3)$$

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

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

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

3. Result and Discussion

The evaluation of the Gaussian Naive Bayes classifier's performance using 5-fold cross-validation yielded varied results across the metrics of accuracy, precision, recall, and F1-score. The summarized outcomes are tabulated below to depict a clear representation of the model's efficacy.

Table 2: Performance Metrics Across 5-Fold Cross-Validation for the Decision Tree Algorithm

K-n	Performa			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	69%	80%	69%	65%
K-2	69%	80%	69%	65%
K-3	80%	85%	80%	79%
K-4	69%	80%	69%	64%
K-5	70%	81%	70%	67%
$\sum$ ***Avg***	71.41%	81.39%	71.41%	67.68%

These results indicate a generally consistent performance with slight fluctuations in effectiveness across different folds. The third fold notably outperformed the others, indicating potential variability in data distribution or model sensitivity to specific feature sets.

Figure 5 below illustrates the performance variability of the Gaussian Naive Bayes classifier across five folds of cross-validation, showcasing the accuracy, precision, recall, and F1-score metrics, which highlight the model's consistency and areas of variability in predicting the conversion from Clinically Isolated Syndrome (CIS) to Multiple Sclerosis (MS)

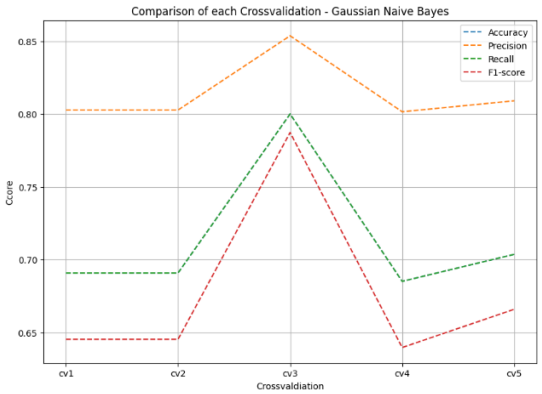

Figure 5: Performance Metrics Across 5-Fold Cross-Validation for the Gaussian Naïve Bayes Algorithm

Discussion

The interpretation of the results suggests that the Gaussian Naive Bayes classifier, while generally reliable, demonstrates variability that may be attributed to the underlying heterogeneity within the clinical data. The precision across all folds was relatively high, which signifies the model's ability to identify true positive cases of MS conversion from CIS accurately. However, the variability in recall and F1-scores points to challenges in consistently capturing all true positives across different subsets of data.

These findings align with previous research indicating that while Naive Bayes classifiers are effective for certain predictive healthcare applications due to their simplicity and efficiency, they may struggle with complex patterns typical of neurological data. The significant findings suggest that despite these limitations, the model can provide a baseline for early diagnostic frameworks but should be enhanced with additional predictive factors or more sophisticated algorithms for improved reliability.

Evaluating these results in the context of existing studies, it is evident that early prediction of MS from CIS remains a challenging area that requires further refinement. The practical implications of these findings are substantial for clinical settings, where early and accurate diagnosis can significantly affect treatment decisions and patient outcomes.

The primary limitation of this research lies in its reliance on a singular algorithm and a specific demographic, which may not generalize well across broader populations or more diverse clinical presentations. Furthermore, the dataset's size and the inherent class imbalance could affect the model's learning process and its predictive power.

Future research should focus on integrating more comprehensive datasets that include a broader range of demographic and genetic factors, and exploring more complex models such as ensemble methods or deep learning approaches, which might capture the nuances of MS progression more effectively. Additionally, longitudinal studies incorporating follow-up data could provide deeper insights into the temporal dynamics of the disease's progression..

4. Conclusion

This study effectively summarized the performance of the Gaussian Naive Bayes classifier in predicting the conversion from Clinically Isolated Syndrome (CIS) to Multiple Sclerosis (MS) among a cohort of Mexican mestizo patients. The results indicated a reasonable degree of accuracy, with specific folds of cross-validation displaying strong predictive capabilities. The most notable finding from the analysis was the high precision across all folds, suggesting that the model is reliable in identifying patients who are likely to convert to MS when it predicts such an outcome. These findings affirm the initial hypothesis that certain clinical and MRI features can serve as significant predictors for the conversion from CIS to MS. Furthermore, the research contributes to the limited but growing body of knowledge concerning MS in the Mexican mestizo population, highlighting the potential for tailored clinical interventions in this demographic.

Figure 6 displayed below represents the classification outcomes of the Gaussian Naive Bayes model, detailing the numbers of true positives, true negatives, false positives, and false negatives. This matrix helps in understanding the model's accuracy in distinguishing between patients who converted from Clinically Isolated Syndrome (CIS) to Multiple Sclerosis (MS) and those who did not.

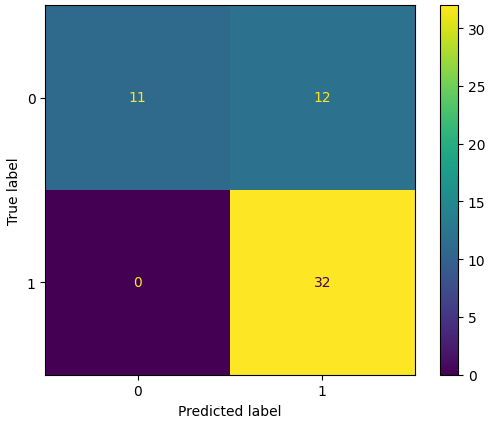

Figure 6: Confusion Matrix

For future research, it is recommended to explore more complex models that might better handle the heterogeneity inherent in MS's clinical presentation, such as ensemble methods or deep learning approaches. Additionally, expanding the study to include larger and more diverse populations could help in generalizing the findings more effectively. In clinical practice, leveraging this model could assist in earlier and more accurate predictions of MS from CIS, enabling timely and targeted therapeutic interventions that could potentially mitigate the progression of the disease. Further studies should also consider longitudinal data to explore the dynamics of disease progression over time, providing deeper insights into the long-term outcomes of patients initially presenting with CIS.

References:

[1] M. V Anand, "Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/2436946.

[2] P. Mansourian, "Anomaly Detection for Connected Autonomous Vehicles Using LSTM and Gaussian Naïve Bayes," *Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST*, vol. 509, pp. 31–43, 2023, doi: 10.1007/978-3-031-34851-8_3.

[3] P. Venkata, "Data mining model and Gaussian Naive Bayes based fault diagnostic analysis of modern power system networks," *Mater. Today Proc.*, vol. 62, pp. 7156–7161, 2022, doi: 10.1016/j.matpr.2022.03.035.

[4] 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.

[5] M. Gayathri, "Analysis of Accuracy in Anomaly Detection of Intrusion Detection System using Naïve Bayes Algorithm compared Over Gaussian model," *ECS Trans.*, vol. 107, no. 1, pp. 13977–13991, 2022, doi: 10.1149/10701.13977ecst.

[6] A. J. Meerja, "Gaussian naïve bayes based intrusion detection system," *Adv. Intell. Syst. Comput.*, vol. 1182, pp. 150–156, 2021, doi: 10.1007/978-3-030-49345-5_16.

[7] 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.

[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, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Azis, Admojo, Susanti - 2020 - Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah.pdf.

[9] 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, [Online]. Available: file:///Users/kbh/Downloads/507-2012-5-PB.pdf.

[10] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Baharuddin, Hasanuddin, Azis - 2019 - Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca.pdf.

[11] 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, [Online]. Available: file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Fitria, Azis - 2018 - Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier.pdf.

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

[13] A. Maulida, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.

[14] F. Tangguh and Y. Islami, "Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin," *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 1–8, 2022, doi: 10.56705/ijodas.v3i1.42.

[15] 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.

[16] Y. I. Sulistya, "Analisis perbandingan Reduction Technique dengan metode Dimentional Reduction dan Cross Validation pada dataset Breast Cancer," *Indones. J. Data Sci.*, vol. 3, no. 2, pp. 82–88, 2022, doi: 10.56705/ijodas.v3i2.41.

[17] 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.