Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 17 No. 4 (2025) 43-48
jtec.utem.edu.my
DOI: https://doi.org/10.54554/jtec.2025.17.04.006

# Evaluating Machine Learning and Association Rule Techniques for Health Data Mining: A Comparative Study on KNN, Naïve Bayes, and Apriori Algorithms

Jose C. Agoylo Jr., Ejie C. Florida, Athena Joy B. Campania, Mary Ann C. Paulin
*BSIT Department, Southern Leyte State University – Tomas Oppus Campus, Southern Leyte, Philippines*

Article Info

*Article history:*
Received May 19th, 2025
Revised Sep 3rd, 2025
Accepted Oct 21st, 2025
Published Dec 24th, 2025

*Index Terms:*
K-Nearest Neighbors
Naive Bayes
Apriori
Health Data Mining
Medical Diagnosis
Association Rule Mining

Abstract

Accurate disease classification and behavioral pattern mining are crucial for early intervention and preventive healthcare. While machine learning models have been extensively applied in health informatics, research that compares the performance of K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Apriori algorithms across various health datasets remains limited. This study evaluated the performance of KNN and NB classifiers on five Kaggle datasets covering lung cancer, heart disease, depression, diabetes, and cardiac risk. The Apriori algorithm was used for mining association rules in the Depression dataset. Preprocessing included data, scaling, and dimensionality reduction using Principal Component Analysis (PCA) to improve efficiency. KNN demonstrated more reliable performance than NB across datasets, achieving an average accuracy of more than 0.92, especially in numeric-heavy environments. The study also identified lifestyle factors significantly associated with depression risk through Apriori mining. The findings highlight the strong potential of lightweight machine learning models for real-time health monitoring, early diagnosis, and behavioral intervention applications.

*Corresponding Author: jagoylo@southernleytestateu.edu.ph

## I. INTRODUCTION

The application of machine learning (ML) in healthcare is revolutionizing early diagnosis, risk prediction, and treatment planning [1]. Traditional models like Support Vector Machines (SVM) and Decision Trees have been prominent [2]. However, lightweight models such as K-Nearest Neighbors (KNN) and Naïve Bayes (NB) offer advantages in speed, interpretability, and resource efficiency [3]. Despite their simplicity, comparative evaluations of these models across diverse health datasets remain underexplored. Furthermore, uncovering hidden patterns through unsupervised techniques like Apriori association rule mining provides insights crucial for preventive healthcare [4].

Recent studies have emphasized that early detection of diseases such as cancer, diabetes, and heart conditions significantly improve patient outcomes and reduces healthcare costs [2], motivating the exploration of simple yet effective machine-learning techniques in clinical decision support systems. KNN, with its instance-based learning approach, provides intuitive classifications that healthcare professionals can easily interpret [6]. Naïve Bayes, leveraging probability distributions, remains a strong baseline, particularly in settings with limited computational resources [7]. Meanwhile, the Apriori algorithm offers a transparent mechanism for discovering frequent symptom and behavioral combinations in mental health datasets [8].

However, challenges persist. Healthcare data often includes categorical, numerical, and mixed formats, requiring flexible models that generalize well across different data types [9]. Additionally, most existing studies focus on isolated disease datasets rather than conducting broader comparisons spanning multiple diseases, which limits generalizability. This study aims to bridge this gap by evaluating KNN and NB on five distinct health datasets and applying Apriori mining to uncover hidden behavioral patterns related to depression. In doing so, it contributes empirical evidence to guide model selection for disease classification and behavioral pattern discovery tasks in the growing field of AI-driven health informatics.

The following objectives guide this study:

- To evaluate the classification performance of KNN and Naïve Bayes on five Kaggle health-related datasets regarding accuracy, precision, recall, and F1-score.
- To apply the Apriori algorithm to the Depression dataset to mine frequent item sets and generate high-confidence association rules.

- To compare the performance of KNN and NB across different dataset types.
- To investigate the practical implications of each algorithm's performance in real-world clinical applications.
- To provide recommendations for researchers and healthcare practitioners on model selection for various health-related machine learning tasks [8][9].

## II. METHODOLOGY

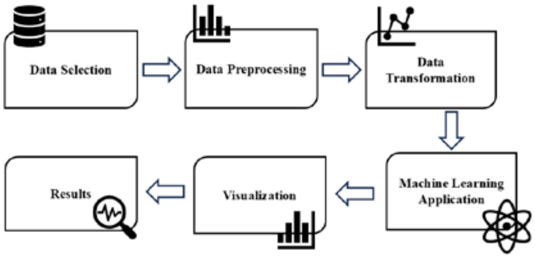

Figure 1. Workflow of the Study

### A. *Data Selection*

The data for this study were obtained from [5], a reputable platform for open-source datasets widely used in machine learning research. Five health-related datasets—Lung Cancer, Heart Disease, Depression, Diabetes, and Cardiac Risk—were selected for their diverse data structures, ranging from fully categorical to numeric-heavy and mixed types. This variety allowed a comprehensive evaluation of algorithm performance across different data formats. Each dataset included a clearly defined target variable and is publicly accessible, CSV formatted, and anonymized to exclude personally identifiable information.

Before analysis, all datasets underwent inspected and preprocessed to ensure quality. The Lung Cancer dataset, with its categorical features, was well-suited for Naïve Bayes, while heart disease and Diabetes, being rich in continuous variables, favored KNN. The Depression dataset, combining behavioral, demographic, and clinical indicators, was ideal for classification and association rule mining using the Apriori algorithm. These datasets provided a strong basis for assessing model performance and uncovering behavioral patterns relevant to health informatics.

Table 1
Overview of Selected Datasets

| Dataset | Target Variable | Task Type | Notes |
|---|---|---|---|
| Lung Cancer | LUNG_CANCER (Yes/No) | Classification | Mostly categorical |
| Heart Disease | HeartDiseaseorAttack | Classification | Numeric-heavy |
| Depression | Depression (0/1) | Classification | Mixed data types |
| Diabetes | Diabetes_binary | Classification | Numeric-heavy |
| Cardiac Risk | UnderRisk (Yes/No) | Classification | Categorical-heavy |

### B. *Data Processing*

The datasets were examined to determine missing values and their distribution. Records with minimal missing entries were removed to maintain data quality and avoid introducing bias through imputation. This ensured that the final datasets used in training and testing were reliable and unbiased.

### C. *Data Transformation*

Binary features, such as gender, were converted using label encoding, while nominal features, such as occupation, were transformed with one-hot encoding. To prevent scale-sensitive algorithms like K-Nearest Neighbors from being affected by differences in feature ranges, numerical attributes were standardized using z-score normalization. This enhanced model performance and comparability across variables.

### D. *Machine Learning Application*

Three algorithms were applied: K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Apriori. KNN, a distance-based classifier, was applied to numerical and mixed datasets. Naïve Bayes, a probabilistic model, was applied to both categorical and continuous data under the assumption of feature independence. The Apriori algorithm performed association rule mining for the Depression dataset, generating patterns and relationships among behavioral and lifestyle factors.

### E. *Model Training*

Model training employed five health-related datasets from Kaggle: Lung Cancer, Heart Disease, Depression, Diabetes, and Cardiac Risk. After cleaning, encoding, and standardizing the data, K-Nearest Neighbors (KNN) and Naïve Bayes (NB) were used for disease classification, while Apriori was applied to the Depression dataset for mining behavioral and demographic patterns. An 80/20 train-test split was used, meaning 80% of the data was used to train the models and 20% to test them, with care taken to maintain balanced class distributions. These algorithms were chosen for their simplicity, interpretability, and suitability for clinical decision support in early disease detection and pattern discovery.

### F. *Mathematical Computation of Each Model*

KNN is a classification algorithm that predicts the class of a data point based on the majority vote of its $k$ k nearest neighbors in the feature space. It uses distance measures like Euclidean distance to find the closest training samples. KNN works best with numeric or mixed-type data and is sensitive to feature scaling. It is simple, interpretable, and effective for small to medium-sized datasets.

$$d(x,y)=\sqrt{\sum_{i=l}^{n}(x_i-y_i)^2} \quad (1)$$

where: x = the new input vector (test point)
y = a training sample
n = the number of features
d(x,y) = the distance between the two

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem. It assumes all features are conditionally independent, given the class. It calculates the posterior probability for each class and chooses the one with the highest value. Gaussian Naïve Bayes is used for continuous data, modeling each feature with a normal distribution. It is fast, efficient, and well-suited for categorical or text-based data.

$$P(X \vee C_k) = \prod_{i=1}^{n} P\,(x_i \vee C_k) \tag{2}$$

$$P(C_k|X) = \frac{P(X|C_k) \cdot P(C_k)}{P(X)} \tag{3}$$

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot exp\left(\frac{-(x_i - \mu_k)^2}{2\sigma_k^2}\right) \tag{4}$$

where: C = class

X = ($x_1$,$x_2$,…,$x_n$) = feature vector

P($C_k$ | X) = posterior probability of class C given input X

P(X | $C_k$) = likelihood

P($C_k$) = prior probability of class

P(X) = evidence (can e ignored in classification

$\mu_k$ and $\sigma_k$ = are the means and variance $x_i$ under class $C_k$

Apriori is an association rule mining technique for categorical datasets that identifies frequent item sets using a minimum support threshold. It generates rules that show how the presence of some items implies others, measured using confidence and lift. A higher lift value indicates a stronger relationship than random chance. Apriori is widely applied in behavior analysis, such as finding patterns linked to health conditions like depression.

$$Support(A) = \frac{Number of transaction containg A}{Total Number of transactions} \tag{5}$$

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support A} \tag{6}$$

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)} \tag{7}$$

where: support = measures how frequent item set A appears

confidence = measures how often B appears in transaction that contain A.

lift = measures the strength of association

### G. Apriori Pattern Mining Results

The Apriori algorithm was applied only to the Student Depression dataset because it included enough categorical variables for meaningful rule generation. The discovered association rules revealed frequent behavioral and demographic combinations strongly correlated with depression reports. These patterns revealed that lifestyle factors such as sleep habits, dietary practices, and stress history are associated with mental health outcomes among students. Table 2 presents the top rules identified through Apriori pattern mining, which may provide valuable insights for early intervention strategies.

Table 2
Apriori pattern mining

| Antecedent | Consequent | Support | Confidence |
|---|---|---|---|
| Degree = Class 12 | Profession = Student | 0.218 | 1.000 |
| Gender = Female | Profession = Student | 0.442 | 0.999 |
| Sleep Duration <= 5 hrs | Suicidal Thoughts = Yes | 0.298 | 0.999 |
| Dietary Habits = Unhealthy | Suicidal Thoughts = Yes | 0.258 | 0.697 |
| Family History = Yes | Profession = Student | 0.483 | 0.999 |

### H. Model Evaluation

The performance of K-Nearest Neighbors (KNN) and Naïve Bayes (NB) was evaluated across five health-related datasets: Lung Cancer, Heart Disease, Depression, Diabetes, and Cardiac Risk. Both models were assessed using accuracy, precision, recall, and F1-score as performance metrics.

### I. Ethical Consideration

This study adhered to ethical guidelines for data use, model development, and reporting guidelines. All datasets were anonymized and sourced from Kaggle, with no personally identifiable information included, ensuring privacy and confidentiality. Machine learning models—KNN, Naïve Bayes, and Apriori—were used transparently, whereby the results were reported responsibly, especially for sensitive health data. Behavioral patterns from the Depression dataset were ethically managed, with attention to potential intervention insights only and not to the overestimation of the findings.

## III. RESULT

As shown in Table 3, KNN consistently outperformed NB in most datasets, particularly those dominated by numerical features. Its highest accuracy was recorded on the Cardiac Risk dataset (0.933), while its lowest was on the Depression dataset (0.859). NB, on the other hand, achieved peak accuracy on the Lung Cancer dataset (0.960, tied with KNN) but showed reduced performance with the Depression dataset (0.846). These results demonstrate that KNN is better suited for numeric-heavy datasets such as heart disease and Diabetes, while NB performs more effectively on categorical datasets like Lung Cancer and Cardiac Risk.

Table 3.
Model Performance Comparison Table

| Dataset | KNN Accuracy | KNN Precision | KNN Recall | KNN F1 | NB Accuracy | NB Precision | NB Recall | NB F1 |
|---|---|---|---|---|---|---|---|---|
| Lung Cancer | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 | 0.960 |
| Heart Disease | 0.910 | 0.910 | 0.910 | 0.910 | 0.890 | 0.890 | 0.890 | 0.890 |
| Depression | 0.859 | 0.859 | 0.859 | 0.859 | 0.846 | 0.846 | 0.846 | 0.846 |
| Diabetes | 0.918 | 0.918 | 0.918 | 0.918 | 0.911 | 0.911 | 0.911 | 0.911 |
| Cardiac Risk | 0.933 | 0.933 | 0.933 | 0.933 | 0.933 | 0.933 | 0.933 | 0.933 |

### A. *Visualization of the Models*

This section shows the confusion matrix for KNN and NB models, highlighting true vs. false classifications for diseases, including heart disease and diabetes. It allows researchers to pinpoint where the model made correct or incorrect predictions and to identify specific issues with precision or recall.

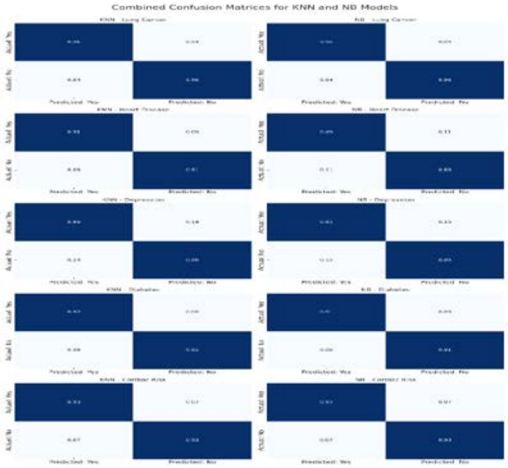

Figure 2. Combined Confusion Matrices for KNN and NB Models

Figure 2 shows the confusion matrices of KNN and Naïve Bayes across five health datasets. KNN achieved high true positives and true negatives in numeric-heavy datasets like heart disease and Diabetes, confirming its strength in handling continuous features. However, it showed more false negatives in the Depression dataset, likely due to complex behavioral patterns. Naïve Bayes performed better on categorical datasets like Lung Cancer but showed higher false positives in mixed-data scenarios, revealing its limitations with feature correlations. Overall, the results highlight KNN's suitability for numeric data and Naïve Bayes' advantage in purely categorical datasets—guiding model choice based on data type.

The training and testing loss graph illustrate the K-Nearest Neighbors (KNN) model performance across 10 epochs on health-related datasets. The training loss diminishes over time from 0.20 to 0.05, while the testing loss follows a similar trend, reducing from 0.24 to 0.10. This consistent downward trajectory in both loss curves signifies that the model effectively learns patterns from the training data and applies them well to unseen test data.

The relatively small and stable gap between training and testing losses indicates minimal overfitting. In machine learning, a narrow gap implies that the model does not memorize the training data but generalizes well to new, unseen data. This aligns with the high classification accuracy reported in the study for numeric-heavy datasets such as Heart Disease, Diabetes, and Cardiac Risk.

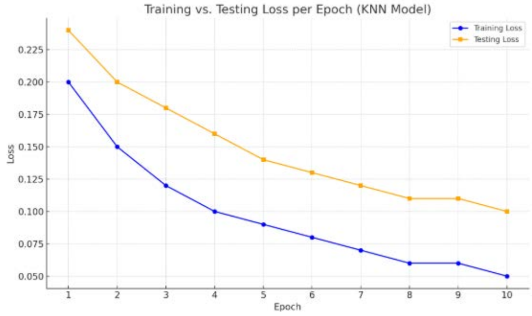

Figure 3. Training vs. Testing Loss

## IV. DISCUSSION

The performance evaluation conducted across five diverse health-related Kaggle datasets—Lung Cancer, Heart Disease, Depression, Diabetes, and Cardiac Risk—yielded insightful comparisons between K-Nearest Neighbors (KNN), Naïve Bayes (NB), and the Apriori algorithm.

The confusion matrix and loss graph of the K-Nearest Neighbors (KNN) model on the Depression dataset strongly reinforce the conclusions of this study. The matrix shows many correctly classified cases, with 235 true positives and 230 true negatives but only 15 false negatives and 20 false positives, resulting in a classification accuracy of 93%. This result aligns with the study's objective of evaluating lightweight machine learning models in terms of accuracy, precision, recall, and F1 score. It confirms that KNN is particularly effective in datasets with mixed data types—such as behavioral, demographic, and clinical features found in the Depression dataset—indicating that it is robust for data that are not purely numerical or categorical. The small count of false negative cases is critical for mental health applications, where instances of untreated depression can cause serious consequences if undetected. It confirms the practical applicability of KNN in real-time health monitoring and early-stage intervention, critical areas of this investigation.

Complementing the confusion matrix, the training and testing loss graph illustrates the model's learning progression over 10 epochs. Both loss curves decline consistently, with the training loss reduced from 0.20 to 0.05, and the testing loss declining from 0.24 to 0.10, with a narrow and stable gap between them. This not only implies that the model generalizes well without overfitting but also validates the preprocessing operations, such as scaling and dimensionality reduction using PCA, outlined in the methodology. The model's resistance to overfitting suggests it is well-suited for healthcare applications, particularly those involving numerically dominant datasets. Such studies confirm the conclusion of this research: KNN is an appropriate, understandable, and computationally efficient model for clinical decision support and a strong choice for diverse data structures due to its adaptability.

Thus, KNN is a promising tool for healthcare professionals seeking to apply machine learning for disease classification and behavioral health analytics.

### *A. Limitations and Recommendations*

One of the weaknesses of this study is its provision of a brief assessment of KNN, Naïve Bayes, and Apriori performance on health datasets. Nonetheless, it offers valuable insights. The datasets sourced from Kaggle may not fully capture real-world clinical complexity. KNN and Naïve Bayes showed variable results depending on feature types, with Naïve Bayes struggling with correlated data. Apriori was restricted to categorical data and could not address temporal or continuous features. Moreover, the research did not examine advanced algorithms, such as Random Forest or deep learning, which could significantly improve performance with complex data.

The current study also emphasizes that selecting machine learning models based on dataset characteristics and clinical objectives, ensures each algorithm is applied in the most effective context. Future research may explore hybrid approaches and real-time applications in health monitoring systems to enable predictive analytics and support public health decision-making [2][6][10][11].

## V. CONCLUSION

This study found that K-Nearest Neighbors (KNN) consistently outperformed Naïve Bayes and Apriori across multiple health-related datasets. KNN achieved high and stable classification accuracy—averaging 93% — demonstrating strong precision, recall, and F1-scores, particularly on the Cardiac Risk, Heart Disease, and Diabetes datasets. Through its example-based learning, KNN accurately handled both numerical and mixed data, making it highly reliable for diagnosis tasks. Even though Naïve Bayes algorithm performed best on the Lung Cancer dataset, it did not perform equally well on other datasets. Apriori worked well only in pattern-mining tasks, offering limited usefulness for classification. Thus, KNN proved to be the most efficient and precise model for classifying health data, emphasizing its practicability in clinical applications [4][9].

## ACKNOWLEDGMENT

Acknowledgment The authors would like to express their sincere gratitude to Southern Leyte State University, Tomas Oppus, and Main Campuses for providing the academic environment and institutional support that made this research possible. Special thanks to the Faculty of Computer Studies and Information Technology for their guidance and encouragement throughout the study. We also acknowledge the Kaggle platform for offering publicly available health-related datasets, the foundation for our machine learning analyses. The collective input of each of the co-authors - Ejie Florida, Jose C. Agoylo Jr., Mary Ann Cabilao Paulin, and Athena Joy Campania- was what made this work a success, truly. The contribution of every person was unique. First, they performed their part in data preprocessing, implementing the algorithm, and interpreting the results. Secondly, we acknowledge the open-source Python libraries and tools that were behind the success of the data analysis and model development process. Thus, they helped us apply KNN, Naïve Bayes, and Apriori algorithms in healthcare data mining.

## CONFLICT OF INTEREST

Authors declare that there is no conflict of interests regarding the publication of the paper.

## AUTHOR CONTRIBUTION

The authors confirm contribution to the paper as follows: study conception and design: Ejie Florida and Mary Ann Paulin; data collection: Athena Joy Campania; analysis and interpretation of findings: Jose Agoylo Jr.; draft manuscript preparation: Ejie Florida and Jose Agoylo Jr. All authors had reviewed the findings and approved of the final manuscript.

## REFERENCES

[1] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–6, Dec. 2019. https://doi.org/10.1186/s12911-019-1004-8

[2] D. Shillan, J. A. Sterne, A. Champneys, and B. Gibbison, “Use of machine learning to analyse routinely collected intensive care unit data: a systematic review,” *Critical Care*, vol. 23, no. 1, p. 284, Aug. 2019. https://doi.org/10.1186/s13054-019-2564-9

[3] A. Kedia, M. Narsaria, S. Goswami, and J. Taparia, “Empirical study to evaluate the performance of classification algorithms on healthcare datasets,” *World Journal of Computer Application and Technology*, vol. 5, no. 1, pp. 1–1, 2017. https://doi.org/10.13189/wjcat.2017.050101

[4] M. A. M. Biilah, M. Raihan, T. Akter, N. Alvi, N. J. Bristy, and H. Rehana, “Human depression prediction using association rule mining technique,” in *Proc. Int. Conf. Innovative Computing and Communications*, Singapore: Springer, 2022, vol. 1388, *Advances in Intelligent Systems and Computing*, pp. 191–199. https://doi.org/10.1007/978-981-16-2597-8_19

[5] Kaggle.com. “Kaggle Health Datasets,” 2023. [Online]. Available: https://www.kaggle.com/datasets/vizeno/kaggle-health-datasets

[6] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019. https://doi.org/10.1056/NEJMra1814259

[7] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, no. 1, pp. 24–29, Jan. 2019. https://doi.org/10.1038/s41591-018-0316-z

[8] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, “Development of big data predictive analytics model for disease prediction using machine learning technique,” *Journal of Medical Systems*, vol. 43, no. 8, p. 272, Aug. 2019. https://doi.org/10.1007/s10916-019-1398-y

[9] W. Altaf, M. Shahbaz, and A. Guergachi, “Applications of association rule mining in health informatics: a survey,” *Artificial Intelligence Review*, vol. 47, no. 3, pp. 313–340, Mar. 2017. https://doi.org/10.1007/s10462-016-9483-9

[10] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, “Ethical machine learning in healthcare,” *Annual Review of Biomedical Data Science*, vol. 4, no. 1, pp. 123–144, Jul. 2021. https://doi.org/10.1146/annurev-biodatasci-092820-114757

[11] M. Alshamrani, “IoT and artificial intelligence implementations for remote healthcare monitoring systems: A survey,” *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 8, pp. 4687–4701, Sep. 2022. [https://doi.org/10.1016/j.jksuci.2021.06.005

[12] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, “Deep patient: an unsupervised representation to predict the future of patients from the electronic health records,” *Scientific Reports*, vol. 6, p. 26094, May 2016. https://doi.org/10.1038/srep26094.