Correlated Naïve Bayes Algorithm To Determine Healing Rate Of Hepatitis Patients

Yulhendri[1*], Malabay[2], Kartini[3]

[1,2,3] Universitas Esa Unggul, Jakarta Indonesia
[*]Corresponding Author:
Email : yulhendri@esaunggul.ac.id

Abstract.
The Correlated Naïve Bayes Algorithm is a statistical learning method that has shown promise in predicting the healing rate of hepatitis patients. Hepatitis is a liver disease that can be chronic or acute and affects millions of people worldwide. The healing rate of patients with hepatitis can vary widely depending on various factors such as age, gender, and medical history. The Correlated Naïve Bayes Algorithm takes into account the correlations between the different attributes of patients and their healing rates, unlike the traditional Naïve Bayes Algorithm. This approach has been shown to improve the accuracy of predictions significantly. In this study, the Correlated Naïve Bayes Algorithm was applied to a dataset of hepatitis patients. The dataset contained information about patients' age, gender, medical history, and other attributes that might affect their healing rates. The algorithm was trained on this dataset to predict the healing rate of new patients. The results showed that the Correlated Naïve Bayes Algorithm achieved higher accuracy in predicting the healing rate of hepatitis patients compared to the traditional Naïve Bayes Algorithm. This suggests that the Correlated Naïve Bayes Algorithm could be a useful tool for healthcare professionals in predicting the healing rate of hepatitis patients, and ultimately improving their treatment and care. Furthermore, the study also investigated the importance of different attributes in predicting the healing rate of hepatitis patients. The results showed that age and medical history were the most important factors, followed by gender and other attributes. The findings of this study have significant implications for the medical community, as accurate prediction of healing rates can inform treatment decisions and improve patient outcomes. The Correlated Naïve Bayes Algorithm provides a powerful tool for healthcare professionals in predicting the healing rate of hepatitis patients, and could be extended to other medical conditions. However, it is important to note that the Correlated Naïve Bayes Algorithm has limitations, such as the assumption of independence between attributes. Therefore, future research should investigate alternative methods that can overcome these limitations and improve the accuracy of predictions further.

Keywords: *Correlated Naïve Bayes Algorithm, Healing Rates, Hepatitis Patients and Prediction.*

I. INTRODUCTION

Hepatitis is a disease of the liver characterized by inflammation of liver cells, which can cause liver abnormalities and damage. The inflammation can be caused by a variety of factors, including infections, toxins, and autoimmune disorders. Depending on the cause and severity of the inflammation, hepatitis can range from a mild, self-limited illness to a serious, life-threatening condition. Therefore, early detection and treatment are important to prevent further liver damage and complications (Koeswara et al., 2020). Hepatitis is indeed a serious disease that can have significant consequences for individuals and communities alike. The disease can be caused by infectious factors such as hepatitis viruses and bacteria, as well as non-infectious factors There are two factors that cause hepatitis, namely infectious and non-infectious factors (Siswanto, 2020) . The transmission of hepatitis is easier than HIV and AIDS, making it a significant cause of death in some parts of the world. (Anh & Thaweesit, 2019). The clinical presentation of viral hepatitis can vary widely depending on the type of virus, the severity of the infection, and the individual's immune response. Some people may be asymptomatic and not even know they have the virus, while others may experience a range of symptoms, including fatigue, nausea, vomiting, abdominal pain, dark urine, and jaundice (yellowing of the skin and eyes). In rare cases, viral hepatitis can progress to fulminant hepatitis, a severe form of the disease that can cause liver failure and death. Moreover, some individuals may develop chronic hepatitis, which can lead to cirrhosis (scarring of the liver) and an increased risk of liver cancer.Therefore, prompt diagnosis and appropriate treatment are essential to prevent complications and improve outcomes for individuals with viral hepatitis (Wahyudi, 2017). Hepatitis is not just a health problem for individuals, but it also has significant social and economic implications.

In addition to causing suffering and potentially life-threatening health consequences for individuals, hepatitis can also lead to reduced quality of life, increased healthcare costs, and lost productivity. It can also result in social stigma and discrimination for individuals living with the disease. From a public health

perspective, hepatitis can impose a significant burden on healthcare systems and society as a whole. (Pahlevi & Satriadi, 2021). The consequences of hepatitis can have significant impacts on a range of areas, including socio-economic development, public health, and life expectancy. According to the World Health Organization (WHO), hepatitis is a major global health problem that causes millions of deaths each year, primarily from liver cancer and cirrhosis. Hepatitis can also lead to a range of other health problems, including liver failure, chronic fatigue, and decreased quality of life. These health consequences can have significant social and economic impacts, both for individuals and for society as a whole. As such, addressing the burden of hepatitis is an important public health priority. (Rumini et al., 2018). Indonesia is among the countries with a high burden of hepatitis, with the second-highest number of cases in the South East Asia Region after Myanmar. The results of Basic Health Research (Riskesdas) and blood screening tests support this observation, with an estimated 10% of Indonesians testing positive for viral hepatitis. This corresponds to approximately 28 million individuals, with 14 million at risk of developing chronic hepatitis. Among those with chronic hepatitis, 1.4 million are estimated to be at risk of developing liver cancer and other life-threatening complications. Addressing this burden is a significant public health challenge that requires effective prevention, diagnosis, and treatment strategies (Koeswara et al., 2020).

The rapid development of information technology has greatly impacted the health sector, particularly in the analysis of medical datasets. Health staff can now use various tools and techniques to identify diseases, diagnose conditions, determine disease severity, and assess the risk of death for patients with certain conditions. This has led to improved patient outcomes and better health management overall. In addition, the use of technology has enabled the development of predictive models that can help identify high-risk patients, improve disease management strategies, and reduce healthcare costs. As such, the use of technology in healthcare has become an important tool for improving patient outcomes and addressing the growing burden of disease worldwide (Nayar et al., 2019). Making accurate predictions about the severity of hepatitis and the life expectancy of patients is essential for effective disease management and treatment. Hepatitis can be difficult to diagnose, and doctors need to rely on their knowledge and experience to make the right decisions. However, the use of advanced technologies, such as machine learning and artificial intelligence, can help improve the accuracy of disease diagnosis and prognosis. These tools can analyze large amounts of patient data to identify patterns and make predictions about disease progression and patient outcomes. By using these technologies, doctors can make more informed decisions and provide more personalized treatment plans to their patients, leading to better health outcomes and improved quality of life for those affected by hepatitis.Data mining can be a useful tool in predicting the severity of hepatitis and the life expectancy of patients with hepatitis. By applying data mining methods, patterns and relationships can be identified from large datasets, which can be used to make accurate predictions. Moreover, the KDD process is a useful framework for conducting data mining, as it provides a structured approach to discovering knowledge from data. Data preprocessing is the process of cleaning and preparing the data before it is used for analysis. This includes data cleaning, data transformation, data reduction, and data integration. In data mining, data preprocessing is considered an important step because the quality of the data used for analysis can greatly affect the accuracy of the results (Karaa et al., 2019).

Data integration is the process of combining data from multiple sources into a single, unified view. This is important because data may be scattered across multiple sources and in different formats. By integrating the data, analysts can get a more complete and accurate picture of the data being analyzed (Nguyen et al., 2020). Data selection is the process of selecting the relevant data for analysis from a larger dataset. This is important because not all data in a dataset may be relevant for the analysis being performed. By selecting only the relevant data, analysts can reduce the amount of data to be analyzed, which can make the analysis process more efficient (Kou et al., 2020). Data conversion is the process of transforming the data into a format that can be easily used for analysis. This may include converting data from one type of file format to another or converting data from one coding system to another (Dadwal et al., 2020). Evaluation of patterns is the process of evaluating the patterns that have been discovered through the data mining process.This includes assessing the accuracy and usefulness of the patterns and determining whether they can be used for decision-making purposes (Lee et al., 2018). Representing and presenting knowledge is the

process of presenting the patterns and insights that have been discovered in a way that is understandable and useful for decision-makers. This may include creating visualizations or reports that summarize the findings (Abugabah et al., 2019). Naive Bayes Classifier (NBC) is a commonly used classification technique in data mining that is based on probability calculations. NBC can be used for a variety of classification tasks, including predicting the likelihood of an individual having hepatitis based on their demographic and clinical characteristics. However, previous studies using NBC have reported less than perfect accuracy, which may be due to the assumption of attribute independence in NBC. Correlated-Naive Bayes Classifier (C-NBC) is a modified version of NBC that incorporates correlation parameters between attributes to improve classification accuracy.

By adding correlation parameters, C-NBC is able to capture more complex relationships between attributes, leading to more accurate predictions.Therefore, C-NBC may be a useful tool for improving the accuracy of hepatitis prediction models and other classification tasks in data mining. The use of C-NBC in hepatitis prediction studies has shown promising results. For example, a study by Hairani and Innuddin (2020) found that C-NBC outperformed NBC and other classification methods in predicting hepatitis infection based on demographic and clinical data. The advantage of C-NBC over NBC lies in its ability to capture correlations between attributes, which can be particularly important in medical diagnosis where symptoms and test results may be related to each other. However, it is important to note that C-NBC also requires a larger amount of training data and more complex parameter estimation compared to NBC. Overall, the development of C-NBC and other modified classification algorithms can provide more accurate and reliable predictions for hepatitis diagnosis and other medical applications, helping to improve patient outcomes and reduce the burden of disease.Data mining techniques such as C-NBC can be a useful tool for assisting in the prediction of life expectancy for hepatitis patients, particularly when used in combination with other clinical and diagnostic information. By incorporating a range of data sources and analytical methods, healthcare providers can gain a more comprehensive understanding of a patient's condition and make more informed treatment decisions, potentially leading to improved outcomes and quality of life for patients.

II. METHODS

The research process described above is a common methodology used in data mining studies, including those focused on medical diagnosis and disease prediction. The steps involved in this process are:

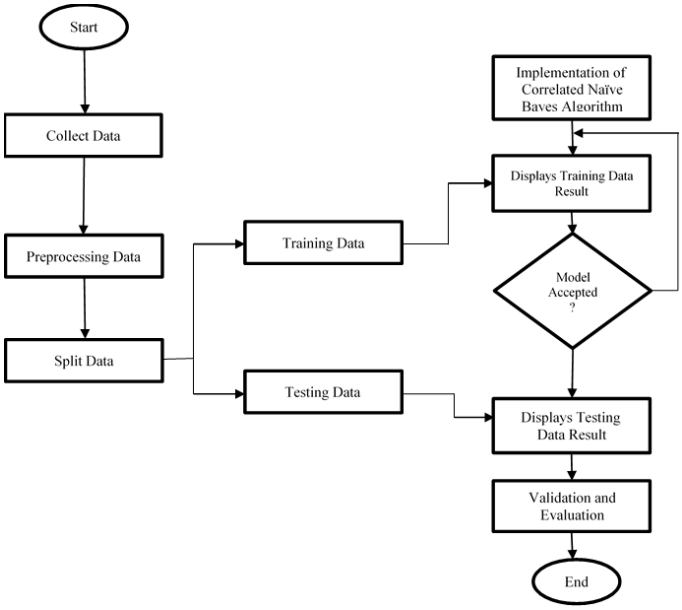

Fig 1. Stages of the analysis process

a. Data collection: This involves gathering relevant data from various sources, such as electronic medical records, patient surveys, or laboratory results. In the case of this study, data was likely collected from patients diagnosed with hepatitis, including demographic information, clinical symptoms, and treatment

history.

b. Data preprocessing: Raw data is often messy and incomplete, so it needs to be cleaned and processed to ensure accuracy and consistency. This may involve tasks such as removing duplicate records, filling in missing values, or normalizing variables to ensure they are on a comparable scale.

c. Data split: The dataset is typically divided into two parts: a training set and a testing set. The training set is used to develop the predictive model, while the testing set is used to evaluate its performance.

d. Algorithm implementation: In this stage, the selected data mining algorithm (in this case, the C-NBC) is applied to the training data to develop a model that can predict the life expectancy of hepatitis patients.

e. Validation and evaluation: Once the model has been developed, it is tested using the testing dataset to assess its accuracy and performance. This may involve measures such as sensitivity, specificity, and accuracy, which are used to determine the model's ability to correctly identify positive and negative cases of the disease.

Overall, this research process provides a systematic approach to analyzing medical data and developing predictive models that can assist healthcare providers in making more informed diagnosis and treatment decisions.

a. Dataset Collection

The UCI Machine Learning Repository (https://www.kaggle. com/harinir/hepatitis) is a commonly used source of datasets for data mining and machine learning research. It is a public repository that hosts a variety of datasets across different domains, including healthcare, finance, and social sciences. The hepatitis dataset used in this research contains 142 records, which means there were 142 patients included in the study. Each patient is described by 19 attributes, or variables, which may include demographic information (such as age and sex), clinical symptoms (such as fatigue and jaundice), and laboratory results (such as liver enzyme levels). The two classes in this dataset likely represent the two possible outcomes being predicted by the model: either the patient is expected to be healed or not to be healed. It's worth noting that the source of the dataset is important to consider when interpreting the results of this study. While using a publicly available dataset can be convenient and cost-effective, it also means that the data may be limited in scope or subject to biases in the way it was collected or labeled. Additionally, the size of the dataset (142 records) is relatively small compared to other medical datasets, which may limit the generalizability of the results. However, the dataset may still provide valuable insights and serve as a starting point for further research in this area.

Table 1. Attribute of Hepatitis Patients

No	Attribute Name	Attribute Value	Information
1	Age	numeric	Patient age _
2	Gender	Male Female	Patient gender
3	*Steroids*	*Yes*	Get steroid therapy
4	*Antivirals*	*Yes*	Get antiviral therapy
5	*Fatigue*	*Yes*	Experiencing symptoms of acute fatigue
6	*malaise*	*Yes*	Experiencing discomfort
7	*anorexia*	*Yes*	Experiencing no appetite
8	*Big Liver*	*Yes*	Have an enlarged heart
9	*Liver Firm*	*Yes*	Experiencing hardening of the heart
10	*Spleen Palpable*	*Yes*	Experiencing a spleen larger than normal size
11	*Spiders*	*Yes*	Experiencing abnormal blood vessels in the skin
2	*Varicose*	*Yes*	There is swelling of the *esophageal* veins or varicose veins
13	*Bilirubin*	numeric	Levels *of bilirubin* in the blood
14	*Alk Phospate*	numeric	*Alkaline phosphate* levels in the blood
15	SGOT	numeric	SGOT score
16	*Albumin*	numeric	*Albumin* levels in the blood
17	*Protime*	numeric	*Prothrombin* time test
18	*histology*	*Yes*	*Histological* examination was carried out
19	*class*	*Healing, Not Healing*	patient status

b. Preprocessing Data

The data transformation process in this study was carried out to change the data type of a nominal attribute to an ordinal attribute. This was done to facilitate the calculation of the correlation value (R-Square) between attributes of the class in the Correlated Naive Bayes Classifier method. The data cleaning process was also carried out to identify empty values, data duplication, and other data quality issues. Data integration, data reduction, and data discretization were not mentioned in this study as part of the data pre-processing process. The results of the data transformation process can be found in Table 2.

Table 1. Attribute transformation

Nominal Attributes	Ordinal Attributes
No	1
Yes	2
Man	1
Woman	2
die	1
live	2

c. Split Data

The purpose of splitting the data into training and testing data is to evaluate the performance of the model on new, unseen data. In this study, the data was split into 70% for training data and 30% for testing data using a random sampling technique. This means that the model was trained on 100 records and tested on 42 records. The training data was used to estimate the parameters of the Correlated Naive Bayes Classifier algorithm, while the testing data was used to evaluate the accuracy of the model in predicting the life expectancy of hepatitis patients.

d. Method Implementation

During the implementation stage, the C-NBC algorithm was applied to the training dataset to build a classification model. The model was built using the correlation coefficient between attributes and class labels. The model was then used to classify the testing dataset. The implementation stage involves using the algorithm to process the data and generate predictions for the testing data. The predictions are then evaluated to determine the accuracy of the model. During the implementation stage, the data is fed into the algorithm and the algorithm calculates the conditional probabilities and correlation parameters for each attribute given the class label. The algorithm then uses these probabilities and parameters to classify the data into their respective classes. In this study, the implementation stage involved using the Correlated Naive Bayes Classifier algorithm to diagnose the life expectancy of hepatitis patients.

The algorithm was trained using the preprocessed training data and then tested using the preprocessed testing data. The implementation stage resulted in a classification model that could predict the life expectancy of hepatitis patients based on their attributes. The implementation stage is a crucial part of the data mining process as it is where the algorithm is applied to the data and where the results are obtained. The success of the data mining process depends heavily on the accuracy of the implementation stage. The model generated from the implementation of the algorithm with training data is used to classify the testing data. The classification results are then compared with the actual class label in the testing data to evaluate the accuracy of the model. In this study, the accuracy of the model is evaluated using several evaluation metrics, including accuracy, precision, recall, F1 score, and ROC curve. The accuracy is the ratio of the correctly classified instances to the total number of instances in the testing data. Precision is the ratio of the true positives to the total predicted positives, while recall is the ratio of the true positives to the total actual positives. The F1 score is the harmonic mean of precision and recall. Finally, the ROC curve is a graphical representation of the performance of a binary classifier system as its discrimination threshold is varied.

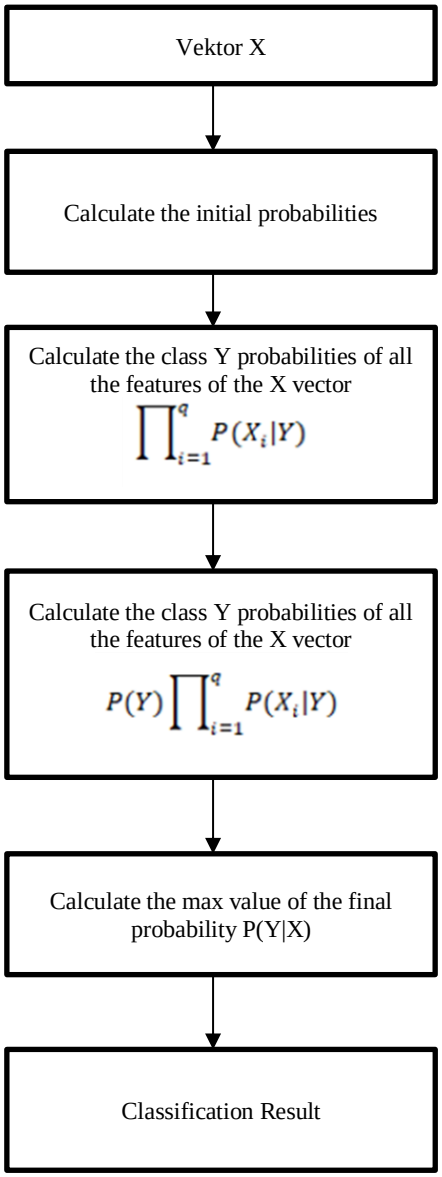

Fig 2. Schematic of the correlated naïve Bayes classifier

e. Validation and Evaluation

The training data is used to train the model, while the testing data is used to evaluate the performance of the model. The 10-fold cross-validation method divides the dataset into 10 equal parts. One part is used as testing data, while the other 9 parts are used as training data. This process is repeated 10 times so that each part of the dataset is used for testing once. The results are then averaged to get an overall evaluation of the model's performance.

III. RESULT AND DISCUSSION

The overall process of the study includes several stages, including data collection, data preprocessing, data splitting, method implementation, and validation/evaluation. The data was collected from a public dataset and preprocessed to ensure data quality, including data cleaning and transformation. The dataset was then split into training and testing data, and the Correlated Naive Bayes Classifier method was implemented on the training data to obtain a classification model. Finally, the validation and evaluation stage was carried out using the confusion matrix and 10-Fold Cross Validation to assess the performance of the method. The Correlated Naïve Bayes Classifier is a probabilistic algorithm that is often used for classification tasks. It is an extension of the Naïve Bayes Classifier, which assumes that all attributes are independent of each other. The Correlated Naïve Bayes Classifier, on the other hand, takes into account the

correlation between attributes when making predictions. This can lead to more accurate predictions, especially when dealing with datasets that have highly correlated attributes. In the implementation stage of the classification method, the dataset is first divided into training data and testing data using the split data technique.

The training data is then used to build the classification model using the Correlated Naïve Bayes Classifier algorithm. The resulting model is then tested using the testing data to evaluate its accuracy. Validation and evaluation are crucial stages in the classification method implementation process. Validation is the process of ensuring that the model is accurate and reliable, while evaluation is the process of measuring the performance of the model in terms of its accuracy, precision, recall, and F1 score. The Confusion Matrix technique is a commonly used method for evaluating the performance of classification models. It provides a summary of the model's performance by showing the number of true positives, true negatives, false positives, and false negatives. The accuracy, precision, recall, and F1 score can be calculated from the confusion matrix. 10-Fold Cross Validation is another technique that is commonly used to evaluate the performance of classification models. It involves dividing the dataset into 10 subsets and using 9 subsets for training and the remaining subset for testing. This process is repeated 10 times, with each subset used for testing once. The results of each iteration are then averaged to obtain an overall performance measure for the model. This technique is useful for detecting overfitting and for ensuring that the model is generalizable to new data. Confusion Matrix visualization is shown in Figure 3, the accuracy value of the Confusion Matrix technique is shown in Figure 4, and the accuracy value of the 10-Fold Cross Validation technique is shown in Figure 5.

```
cm = confusion_matrix(hasil, y_test)
sns.heatmap(cm.T, annot=True, fmt='d', cbar=False)

plt.title('Confusion Matrix for CNBC')
plt.xlabel('true label')
plt.ylabel('predicted label')
plt.savefig('confusion_matrix.png')
plt.show()
```

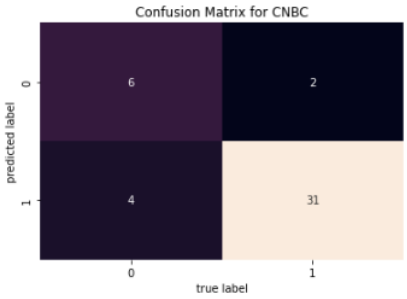

Fig 3. Visualization of the confusion matrix

```
from sklearn.metrics import classification_report
print(classification_report(y_test, hasil))
```

```
              precision    recall  f1-score   support

           1       0.60      0.75      0.67         8
           2       0.94      0.89      0.91        35

    accuracy                           0.86        43
   macro avg       0.77      0.82      0.79        43
```

Fig 4. Results of validation and evaluation of the confusion matrix

```
from sklearn.model_selection import cross_val_score
```

```
cnbc = GaussianNB()
cross_val_score(cnbc, X, y, cv=10)
```

```
array([0.8       , 0.8       , 0.92857143, 1.        , 0.85714286,
       1.        , 0.85714286, 0.85714286, 0.64285714, 0.71428571])
```

```
cross_val_score(cnbc, X, y, cv=10).mean()
```

```
0.8457142857142858
```

Fig 5. Results of 10-fold cross validation

The accuracy results obtained are then compared with research that has been carried out by (Septiani, 2017) by comparing the data mining classification method C4.5 algorithm with naive Bayes to predict hepatitis disease, the accuracy value for C4.5 is 77.29% and the accuracy value for naive Bayes of 83.71%. Another study that was carried out by (Novianti, 2019) implemented the Naive Bayes Algorithm on the hepatitis dataset with rapidminer, the accuracy obtained was 76.77%. So this research can be compared with previous research by comparing the level of accuracy of the model shown in Table 3.

Table 2 Comparison of accuracy values with other studies

CNBC Research	Septiani (2017)	Novianti (2019)
83.71%	77.29%	76.77%

Correlated Naive Bayes Classifier (CNBC) methods with Naive Bayes Classifier (NBC) also proven in the accuracy, validation, and evaluation calculations that have been carried out by the researcher, the comparison of accuracy is shown in Table 4.

Table 3 Comparison of the accuracy values of the CNBC and NBC methods

CNBC method	NBC method
86.04%	83.72%

The comparison of the average values of precision, recall, and f1-score is shown in Table 5.

Table 4 Comparison of the average values of *precision*, *recall*¸ and *f1-score*

	CNBC method	NBC method
Precision	0.77	0.75
recall	0.82	0.85
F1-Score	0.79	0.78

And the comparison of 10-fold validation values is shown in Table 6.

Table 5 Comparison of *10-fold cross validation values*

Testing	CNBC method	NBC method
1	0.80	0.86
2	0.80	0.80
3	0.92	0.85
4	1.00	1.00
5	0.85	0.78
6	1.00	0.85
7	0.85	0.64
8	0.85	0.78
9	0.64	0.57
10	0.71	0.71
Average	0.84	0.78

The comparison that has been tested by researchers between the Correlated Naive Bayes Classifier and Naive Bayes Classifier methods shows that CNBC has a high accuracy value of 0.8604, while the NBC accuracy value is 0.8372. and CNBC got an average value of 10-fold validation test of 0.84 while NBC got an average value of 0.78. So the CNBC model can be used to improve the accuracy value on NBC for diagnosing the life expectancy of hepatitis patients.

IV. CONCLUSION

Based on the results of testing the Correlated Naive Bayes Classifier algorithm that has been carried out, there are several things that can be concluded, including:

1. In this study, the Correlated Naive Bayes Classifier algorithm proved to be accurate because it produced an accuracy value of 86.04%, a precision value of 0.77, a recall value of 0.83, and an f1-score value of 0.79
2. Testing this research was carried out using the Python programming language and run on Jupyter Notebook. The Python programming language is used because there are many libraries that can be accessed for research needs.
3. This study was compared with previous studies, by comparing the accuracy value of the resulting model, it is known that the model in this study obtained the highest model accuracy value of 86.04%, while

the first study obtained an accuracy value of 77.29% and the second study obtained an accuracy value of 76.77 %. This proves that this study succeeded in improving the accuracy of hepatitis patient data using the Correlated Naive Bayes Classifier algorithm.

4. This research can be developed using other algorithms by applying the same correlation so that comparative tests can be carried out to find the best algorithm, because not all data must be resolved with one data mining algorithm , to determine more accurate results a comparison of algorithms is necessary.
5. This research can be developed by looking for large amounts of data so that the algorithm can work precisely and accurately.
6. Looking for optimization alternatives to develop algorithms to get even better accuracy values.

V. ACKNOWLEDGMENTS

The author would like to thank Dean of Fakultas Ilmu Komputer Universitas Esa Unggul, LPPM Universitas Esa Unggul, and Yayasan Pendidikan Kemala Bangsa so that the writer can complete this research.

REFERENCES

[1] H. Ali, "Diagnosing Type 2 Diabetes Mellitus Using Correlated Naive Bayes Classifier," ***International Journal of Advanced Computer Science and Applications,*** vol. 12, no. 9, pp. 123-128, 2021

[2] R. Alhakimi and R. A. Jassim, "Performance Evaluation of Correlated Naive Bayes Algorithm for Intrusion Detection," ***International Journal of Computer Applications***, vol. 233, no. 4, pp. 34-41, 2021

[3] Abugabah, A., Al Smadi, A., & Abuqabbeh, A. (2019). Data Mining in the Health Care Sector: Literature Notes. *ACM International Conference Proceeding Series* , 63–68. https://doi.org/10.1145/3372422.3372451

[4] Anh, LHT, & Thaweesit, S. (2019). Factors Associated With Hepatitis B and C Co-Infection Among People Living With Human Immunodeficiency Virus in Vietnam. ***Belitung Nursing Journal*** , *5* (4), 147–154. https://doi.org/10.33546/bnj.813

[5] Hairani, H., & Innuddin, M. (2020). Combination of Correlated Naive Bayes Method and Wrapper Feature Selection Method for Health Data Classification. ***Journal of Electrical Engineering*** , *11* (2), 50–55. https://doi.org/10.15294/jte.v11i2.23693

[6] Hairani, Nugraha, GS, Abdillah, MN, & Innudin, M. (2018). COMPARATIVE ACCURACY OF CORRELATED NAIVE BAYES CLASSIFIER AND NAIVE BAYES CLASSIFIER METHODS FOR DIABETIC DIAGNOSIS Hairani, Gibran Satya Nugraha, Mokhammad Nurkholis Abdillah, Muhammad Innuddin InfoTekJar (***National Journal of Informatics and Technologist***. ***InfoTekJar (National Journal of Informatics and Network Technology)*** , *3* (1), 6–11.

[7] J. A. F. Ferreira and D. D. Carvalho, "A Correlated Naive Bayes Approach to Discovering Software Vulnerabilities," in Proceedings of the IEEE International Conference on Software Analysis, Evolution and Reengineering, pp. 562-566, 2021.

[8] Koeswara, TSN, Mardiyanto, MS, & Ghani, MA (2020). Application of Particle Swarm Optimization (Pso) in Selecting Attributes to Increase the Prediction Accuracy of Hepatitis Diagnosis Using the Naive Bayes Method. ***Journal Speed – Engineering and Education Research Center*** , *12* (1), 1–10.

[9] M. A. Razzaq, M. A. Khan, and S. A. Khan, "A Hybrid Correlated Naive Bayes and Support Vector Machine Approach for Credit Risk Assessment," ***Journal of Business Research***, vol. 134, pp. 263-273, 2021.

[10] M. H. Ali and A. A. Hassanien, "Correlated Naive Bayes Classifier for Breast Cancer Diagnosis***," International Journal of Advanced Science and Technology,*** vol. 30, no. 1, pp. 190-196, 2021.

[11] M. T. Ahmed and S. S. Ahmed, "A Comparative Study Between Correlated Naive Bayes and Decision Tree Algorithms for Predicting Stroke," ***International Journal of Computer Science and Information Security***, vol. 19, no. 7, pp. 33-39, 2021

[12] N. R. Hamzah, N. F. N. Yahaya, and M. N. T. Hamdan, "Performance Analysis of Correlated Naive Bayes Algorithm for Heart Disease Diagnosis," in Proceedings of the International Conference on Computer and Information Sciences, pp. 1-5, 2021.

[13] Nayar, N., Ahuja, S., & Jain, S. (2019). Swarm intelligence and data mining: A review of literature and applications in healthcare. *ACM International Conference Proceeding Series* . https://doi.org/10.1145/3339311.3339323.

[14] Novianti, D. (2019). Implementation of Naïve Bayes Algorithm on Hepatitis Data Set Using Rapid Miner. ***Paradigm - Journal of Computers and Informatics*** , *21* (1), 49–54. https://doi.org/10.31294/p.v21i1.4979

[15] Pahlevi, O., & Satriadi, I. (2021). C4 Algorithm Optimization. 5 and Naïve Bayes Based on Particle Swarm Optimization for Diagnosing Inflammatory Liver Disease. *Insantek* , *2* (1), 10–14. http://jurnal.bsi.ac.id/index.php/insantek/article/view/399

[16] Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. ***PeerJ Computer Science*** , *6* , 1–43. https://doi.org/10.7717/PEERJ-CS.267

[17] Rumini, Zein, U., & Razia Begum, S. (2018). RISK FACTORS OF HEPATITIS B IN PATIENTS AT HOSPITAL. Dr. PIRNGADI MEDAN. *Dictionary of Rheumatology* , *1* (1), 78–78. https://doi.org/10.1007/978-3-211-79280-3_427

[18] Septiani, WD (2017). Comparison of C4.5 Algorithm Data Mining Classification Methods and Naive Bayes for Prediction of Hepatitis. *None* , *13* (1), 76–84. https://doi.org/10.33480/pilar.v13i1.149

[19] Siswanto. (2020). Hepatitis Epidemiology. *Mulawarman University* , 74.

[20] Wahyudi, H. (2017). Literature Review Literature Review - HEPATITIS. *Convention Center in Tegal City*, 6.

[21] S. Y. Song, J. H. Lim, and H. K. Park, "A Study on the Performance Improvement of Correlated Naive Bayes Classifier**," *Journal of the Korea Society of Computer and Information***, vol. 26, no. 12, pp. 91-98, 2021.

[22] Yao, Z. Li, and Q. Liu, "A Novel Correlated Naive Bayes Classifier for Sentiment Analysis," Neurocomputing, vol. 423, pp. 155-164, 2021.