JOURNAL MEDICAL INFORMATICS TECHNOLOGY
medinftech.org ISSN 2988-7003 (online)

# Hepatitis Prediction Using K-NN, Naive Bayes, Support Vector Machine, Multilayer Perceptron and Random Forest, Gradient Boosting, K-Means

*Heru Dwi Saputra [1],Ade Irfan Efendi [2], Edwin Rudini [3], Dwiza Riana[4], and Alya Shafira Hewiz[5]*

[1,2,3,4] *Universitas Nusa Mandiri, Indonesia*
[5] *Universitas Airlangga, Indonesia*



ARTICLE HISTORY
Received: 17 November 23
Final Revision: 12 December 23
Accepted: 30 December 23
Online Publication: 31 December 23

KEYWORDS
Measuring Accuracy, Precision, Recall, ROC, Best Score

CORRESPONDING AUTHOR
author_correspondent@mail.com

DOI
10.37034/medinftech.v1i1.1

**A B S T R A C T**

Hepatitis is a serious disease that causes death throughout the world. It is responsible for inflammation in the human liver. If we manage to detect this life-threatening disease early, we can save many lives from it. In this research paper, we predict hepatitis disease using data mining techniques. We have attempted to propose a feasible approach to improve the performance of our prediction models in our research. We address the problem of missing values in the dataset by replacing them with the mean value. Nine algorithms were applied to the hepatitis disease dataset to calculate prediction accuracy. We measure accuracy, precision, recall, ROC and best score, and we compare them with random search hyperparameter tuning. It is hoped that by using them we will find the optimal combination of hyperparameters to improve the performance of machine learning models which helps us compare the performance of classification models.

## 1. Introduction

Hepatitis is a disease defined as inflammation of the liver and is most often caused by viral infections, resulting in 1.5 million deaths worldwide each year [1]. Viral hepatitis has become a major threat to human health in recent decades, with a wide variety of hepatitis-associated viruses [2]. Medical diagnosis is an important and complex task that requires accurate identification. It plays an important role in diagnosing the disease at the right time and early stages of recovery. The liver is an important organ in the human body, and hepatitis is a serious disease that affects its function.

The main factor that causes liver inflammation is the presence of viruses in life [3]. Classification algorithms can help medical professionals in diagnosing diseases. A classification algorithm will be applied to predict patient data for hepatitis [4], [5]. Determining the diagnosis of hepatitis is a challenging task for doctors because many factors need to be considered and analyzed [6]. The healthcare industry collects information from various clinical reports and diagnostic test results to identify dataset class labels by observing invisible patterns and correlated features in the dataset [7]. Both hidden and correlated patterns help distinguish between those who have hepatitis and those who do not. Predicting the survival of hepatitis patients is a challenging task in the early stages due to interdependent features. Therefore, models can be developed to predict the survival of hepatitis patients [8]. Data mining refers to the extraction or "mining" of knowledge from large amounts of data. Data mining has been widely used in bioinformatics to analyze biomedical data. Data mining algorithms can be used efficiently for prediction and classification of interrelated data. The use of data in the health care industry is very important to assist in reliable early disease detection and improve the quality of health services [9].

## 2. Research Method

This research aims to improve the accuracy of predictions used in data mining algorithms. The datasets used in prediction models must be more precise and accurate. The collected data set may contain irrelevant or missing values. To ensure that the data mining process produces the best results in terms of accuracy, it must be managed effectively with the framework presented in Figure 1.

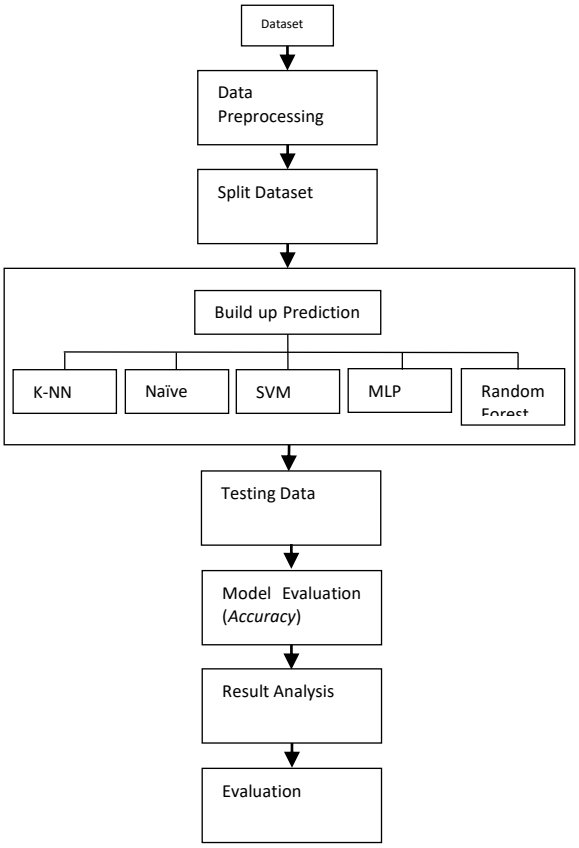


Figure 1. Architecture of Methods Used

### 2.1. Attribute Identification

The amount of data is 155 samples and 20 features with classes indicating whether the prediction is "yes" or "no" for survival, the dataset is taken from the UCI Machine Learning Repository. The dataset consists of six multi-valued characteristics and 14 nominal attributes. The characteristics listed are the most common in the dataset used and are presented in Table 1.

Table 1. Atribut Himpunan Data

| Attribute | Value |
|---|---|
| Age | Numeric |
| Sex | Male (1), Female (2) |
| Steroid | No (1), Yes (2) |
| Antivirals | No (1), Yes (2) |
| Fatigue | No (1), Yes (2) |
| Malaise | No (1), Yes (2) |
| Anorexia | No (1), Yes (2) |
| Liver Big | No (1), Yes (2) |
| Liver Firm | No (1), Yes (2) |
| Spleen palpable | No (1), Yes (2) |
| Spiders | No (1), Yes (2) |
| Ascites | No (1), Yes (2) |
| Varises | No (1), Yes (2) |
| Bilirubin | 0.6, 0.7, 0.8, 1, 1.2, 2.00, 3.00 |
| Alk_phosphate | 40, 70,100, 130, 160, 200, 250 |
| Sgot | 16, 50, 100, 140, 200, 400, 500 |
| Albumin | 2.7, 3, 3.3, 3.8, 4, 4.4, 4.7 |
| Protime | 10, 20, 30, 40, 50, 60, 70, 80, 90 |
| Histology | No (1), Yes (2) |
| Class | Die (1), Live (2) |

### 2.2. Naive Bayes

Naïve Bayes is used for classification and is based on Bayes' theorem. It is very easy to build this classifier model. We can determine the probability of an event occurring given the probability of another event that has occurred before, using the help of the Bayes Hypothesis [10]. The posterior probability value is calculated using Equation (1).

$$P(C|X) = \frac{P(C)P.P(X|C)}{P(X)} \quad (1)$$

Where X is Attribute, C is Class, and P(C|X) is The probability that C given X.

### 2.3. Random Forest

The Random Forest algorithm is a machine learning algorithm that is very popular for classification and regression purposes. In this study, we use it for classification purposes. It works in three processes. In the first process during the learning phase, a Decision Tree is generated from a number of trees. In the second process, for each dataset, the tree used to make the decision in the previous step predicts the class name. In the final step, which is the third process, the correct class name is assigned to the dataset based on the majority of each data present in the dataset encountered in step 3 [11]. Comparing different types of supervised machine learning algorithms for predicting heart disease is the focus of this paper. In this paper, the info-gain feature selection technique is applied to improve the accuracy of the classification model. The best results were obtained from Logistic Regression with an accuracy value of 92.76% [8].

### 2.4. K-Nearest Neighbors (KNN)

KNN has three stages in the classification process of this classifier. In step 1, it calculates the K value. In step 2, it sorts and calculates the distance between all training data for each test sample. In step 3 a majority voting approach is used to assign class names to the test sample data [12]. Calculating the Euclidean distance is presented in Equation (2).

$$D_e = \sqrt{\sum_{i=1}^{n}(ai - bi)^2} \quad (2)$$

### 2.5. Support Vector Machine (SVM)

SVM is considered to be a good classifier in terms of accuracy and generalization ability, but its limitation lies in its higher training time. Therefore, to overcome this, various feature selection techniques have been developed that can be integrated with SVM to achieve better results with smaller dimensional data [13].

### 2.6. *Multilayer* Perceptron (MLP)

Multilayer Perceptron (MLP) can be considered as an effective model due to its compact structure and adaptive mechanisms. Especially for Medical IoT (MoT) based data, this data usually consists of complex

features with a very large volume, such as highly correlated relationships between features or biases among relationships [14], [15].

2.7. Data Cleaning and Feature Selection

Data sets retrieved from the UCI repository may have characteristics of duplicates and missing values. Missing values can be handled in one of two ways: either by removing them or by replacing them with a mean, maximum, or minimum value as a replacement. Data quantity can also be reduced by removing missing values in the data set, but this will decrease the prediction accuracy. Therefore, zero values were used as substitutes for these missing values, which had only a minor impact on data quality. Feature selection can be done using feature weighting. Using RapidMiner, attributes were weighted by replacing missing values with the mean. The dataset used in this research is hepatitis patient data, consisting of 155 examples of patient data categorized into 20 attributes (features), including labels. This data is taken from the UCI Machine Learning Repository. The dataset used has 20 attributes (including labels), namely Class, Age, Gender, Steroid, Antivirus, Fatigue, Malaise, Anorexia, Large Liver, Firm Liver, Palpable Spleen, Spider, Ascites, Varices, Bilirubin, Alk phosphate, Got, Albumin, Protime and Histology. The two classes in this dataset are defined using the parameters "Live" and "Die," which classify the survival and death of hepatitis patients based on their condition. In the initial experiment, data preprocessing will be carried out by replacing missing data with the average value. After the process of replacing missing values in this section, we classify hepatitis patient data using the K Nearest Neighbor, Guassian Naive Bayes, Logistic Regression, Neural Networks, Support Vector Machine, Decision Tree, Random Forrest, AdaBoost, Gradient Boosting methods and compare using hyperparameter tuning random search method with attribute values is presented in Table 2.

Table 2. Atribut values

| No | Atribut | Missing Value |
|---|---|---|
| 1 | Age | 0 |
| 2 | Sex | 0 |
| 3 | Steroid | 1 |
| 4 | Antivirals | 0 |
| 5 | Fatigue | 1 |
| 6 | Malaise | 1 |
| 7 | Anorexia | 1 |
| 8 | Liver Big | 10 |
| 9 | Liver Firm | 11 |
| 10 | Spleen Palpable | 5 |
| 11 | Spiders | 5 |
| 12 | Ascites | 5 |
| 13 | Varises | 5 |
| 14 | Bilirubin | 6 |
| 15 | Alk_phosphate | 29 |
| 16 | Sgot | 4 |
| 17 | Albumin | 16 |
| 18 | Protime | 67 |
| 19 | Histology | 0 |
| 20 | Class | 0 |

## 3. Result and Discussion

The nine classification algorithms implemented to determine the value that achieves the highest accuracy are then compared with a classification algorithm that uses hyperparameter tuning random search to find the optimal combination using hyperparameter tuning random search to randomly determine a combination of values from a predetermined search space. In the random search process, we trained and tested the model based on several possible combinations of hyperparameter tuning, then obtained a comparison of the process before and after using hyperparameter tuning which is presented in Table 3.

Table 3. Classification Uses 20 Features with Missing Values Replaced by The Average and Divided by Data Split for Testing

| Classification Algorithm | Accuracy | Precision | Recall | F1-Score | Roc Area |
|---|---|---|---|---|---|
| K-Nearest Neighbor | 77.42 | 77.42 | 100.00 | 0.87 | 0.50 |
| Guassian Naive Bayes | 77.42 | 86.96 | 83.33 | 0.85 | 0.70 |
| Logistic Regression | 70.97 | 80.00 | 83.33 | 0.82 | 0.55 |
| Neural Networks | 70.97 | 80.00 | 83.33 | 0.82 | 0.55 |
| Support Vector Machine | 67.74 | 81.82 | 75.00 | 0.78 | 0.58 |
| Decision Tree | 67.74 | 79.17 | 79.17 | 0.79 | 0.50 |
| Random Forrest | 74.19 | 80.77 | 87.50 | 0.84 | 0.50 |
| AdaBoost | 87.10 | 85.71 | 100.00 | 0.92 | 0.71 |
| Gradient Boosting | 77.42 | 77.42 | 100.00 | 0.87 | 0.50 |

Table 3 presents the classification results using 20 features with missing values replaced by the mean and divided by the data split for testing. The accuracy percentage shows the performance of each algorithm in classifying data. The results show that Ada Boosting achieved the highest accuracy of 87.10%. From the results above, we then compared them with the results using hypertuning random search parameters with the same data and the same data distribution presented in Table 4.

Table 4. Classification Uses 20 Features with Missing Values Replaced by The Average and Uses Hyperparameter Tuning Random Serch

| Classification Algorithm | Accuracy RS | Precision RS | Recall RS | F1-Score RS | Best Score (%) |
|---|---|---|---|---|---|
| K-Nearest Neighbor | 74.19 | 95.83 | 76.67 | 0.85 | 82.82 |
| Guassian Naive Bayes | 74.19 | 87.50 | 80.77 | 0.84 | 95.82 |
| Logistic Regression | 77.42 | 91.67 | 81.48 | 0.86 | 89.32 |
| Neural Networks | 83.87 | 91.67 | 88.00 | 0.89 | 91.98 |
| Support Vector Machine | 77.42 | 83.33 | 86.96 | 0.85 | 88.67 |
| Decision Tree | 74.19 | 91.67 | 78.57 | 0.84 | 87.42 |
| Random Forrest | 77.42 | 95.83 | 79.31 | 0.86 | 88.67 |
| AdaBoost | 74.19 | 87.50 | 80.77 | 0.84 | 91.56 |
| Gradient Boosting | 74.19 | 91.67 | 78.57 | 0.84 | 90.59 |

Table 4 presents the classification results using RS hyperparameter tuning to optimize the best search by producing the best performance. From the results above, it was found that Hyperparameter Tuning RS in the Random Search algorithm for neural networks succeeded in increasing model accuracy by 83.87%, achieving the best score of 91.98%. This means that a model that has been well optimized is able to achieve higher accuracy than the model before hyperparameter tuning was carried out. The results are presented in Figure 2, Figure 3, Figure 4, and Figure 5.

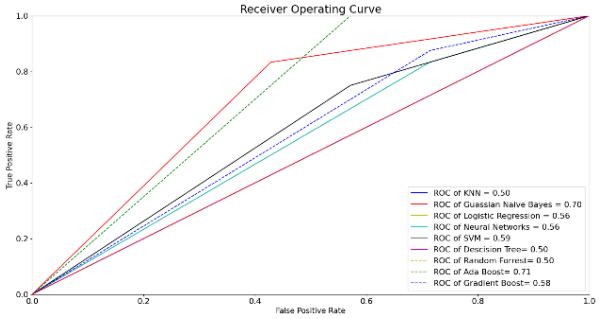


Figure 2. Before using Hypertuning Research Random Parameters

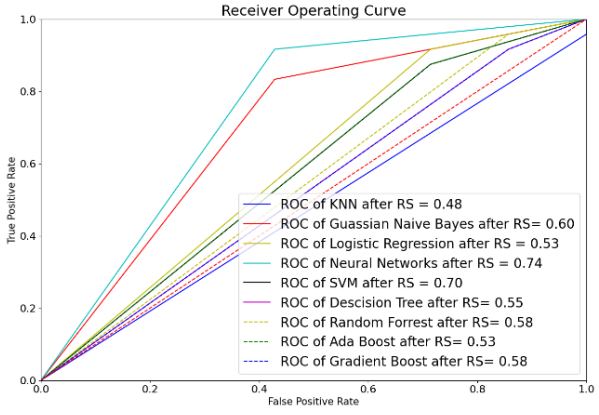


Figure 3. After using Random Research Parameter Hypertuning

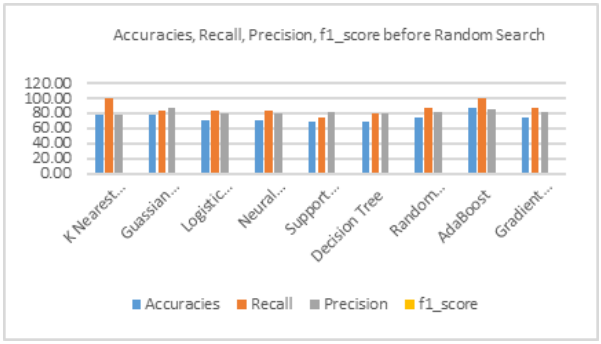


Figure 4 Accuracies, Recall, Precision, f1_score before Random Search.

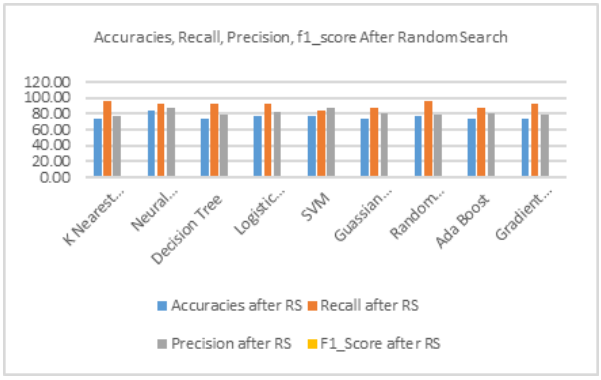


Figure 5. Accuracies, Recall, Precision, f1_score after Random Search

## 4. Conclusion

In our research, we have felt the great importance of dealing with datasets with missing values and also decisiveness in feature selection to improve the accuracy of classification models. to get the best classifier, we have made a comparison between our classification models before and after using RS Tuning hyperparameters. To apply a machine learning model to this problem, hyperparameters must be set to handle a particular data set. Hyperparameters are used in ML models to get the best hyperparameters. The dataset we use is very small and it can be seen that the randomly selected set is very limited in representing the dataset. It is best to use a dataset with a large capacity because hyperparameters are very effective in optimizing the ML model that will be used. In future research, it is hoped that large datasets will be used so that comparisons with these hyperparameters are more optimal with better feature selection.

## References


[1] T. I. Trishna, S. U. Emon, R. R. Ema, G. I. H. Sajal, S. Kundu, and T. Islam, "Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier," in 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019, Institute of Electrical and Electronics Engineers Inc., Jul. 2019. doi: 10.1109/ICCCNT45670.2019.8944455.

[2] A. Rasche, A. L. Sander, V. M. Corman, and J. F. Drexler, "Evolutionary biology of human hepatitis viruses," Journal of Hepatology, vol. 70, no. 3. Elsevier B.V., pp. 501–520, Mar. 01, 2019. doi: 10.1016/j.jhep.2018.11.010.

[3] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification," 2011.

[4] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052068.

[5] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification," 2011.

[6] M. J. Nayeem, S. Rana, F. Alam, and M. A. Rahman, "Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest," in 2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings, Institute of Electrical and Electronics Engineers Inc., Feb. 2021, pp. 280–284. doi: 10.1109/ICICT4SD50815.2021.9397013.

[7] N. Komal Kumar and D. Vigneswari, "Hepatitis- infectious disease prediction using classification algorithms," Res J Pharm Technol, vol. 12, no. 8, pp. 3720–3725, Aug. 2019, doi: 10.5958/0974-360X.2019.00636X.

[8] S. Hashem et al., "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," IEEE/ACM Trans Comput Biol Bioinform, vol. 15, no. 3, pp. 861–868, May 2018, doi: 10.1109/TCBB.2017.2690848.

[9] N. Nahar and F. Ara, "Liver Disease Prediction by Using Different Decision Tree Techniques," International Journal of Data Mining & Knowledge Management Process, vol. 8, no. 2, pp. 01–09, Mar. 2018, doi: 10.5121/ijdkp.2018.8201.

[10] S. M. M. Hasan et al., "Comparative Analysis of Classification Approaches for Heart Disease Prediction Data Security View project A Cryptographic Algorithm Based on ASCII and Number System Conversions along with a Cyclic Mathematical Function View project Comparative Analysis of Classification Approaches for Heart Disease Prediction." [Online]. Available: https://www.researchgate.net/publication/334929171

[11] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052068.

[12] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," SN Appl Sci, vol. 1, no. 12, Dec. 2019, doi: 10.1007/s42452-019-1356-9.

[13] K. S. Sahoo et al., "An Evolutionary SVM Model for DDOS Attack Detection in Software Defined Networks," IEEE Access, vol. 8, pp. 132502–132513, 2020, doi: 10.1109/ACCESS.2020.3009733.

[14] S. J. Lee et al., "A dimension-reduction based multilayer perception method for supporting the medical decision making," Pattern Recognit Lett, vol. 131, pp. 15–22, Mar. 2020, doi: 10.1016/j.patrec.2019.11.026.

[15] S. M. M. Hasan et al., "Comparative Analysis of Classification Approaches for Heart Disease Prediction Data Security View project A Cryptographic Algorithm Based on ASCII and Number System Conversions along with a Cyclic Mathematical Function View project Comparative Analysis of Classification Approaches for Heart Disease Prediction." [Online]. Available: https://www.researchgate.net/publication/334929171