Journal of Software Engineering, Information and Communication Technology (SEICT)

Journal homepage: https://ejournal.upi.edu/index.php/SEICT

Comparison of Machine Learning Algorithms in the Role of Hepatitis Patient Disease Classification

Daud Fernando[1], Faris Huwaidi[2], Muhammad Hafidz Ananto[3], dan Sahrial Pramadya[4]

[1]Software Engineering
Univesitas Pendidikan Indonesia
Bandung, Indonesia
Correspondence: E-mail: daudfernando@upi.edu

ABSTRACT

Hepatitis is one of the diseases with the highest patient percentage. About a third of the world's population is afflicted with hepatitis. In several cases, patients show symptoms, while in others, patients show no symptoms. Hepatitis is commonly caused by hepatitis A, B, C, D or E virus and yellow fever virus (YFV). Hepatitis can be detected through blood tests. From the blood sample, we could extract information like Alanine Transferase (ALT), bilirubin, creatine, Alkaline Phosphatase (ALP), Aspartate Aminotransferase (AST) and Gamma Glutamyl Transferase (GGT) levels, levels of these compound will be able to determine whether the patient is afflicted or not. Machine learning can be applied to help process the information to raise the effectiveness of information processing. Several algorithms like support vector machine (SVM), decision tree, K-Nearest Neighbor (KNN), Random Forest and X-Gradient Boost (XGBoost) can be used to process hepatitis data. This research is aimed to determine which algorithm has the highest accuracy in diagnosing hepatitis

ARTICLE INFO

Article History:
Submitted/Received 02 Oct 2023
First Revised 24 Oct 2023
Accepted 05 Nov 2023
First Available online 01 Dec 2023
Publication Date 01 Dec 2023

Keyword:
Hepatitis, Compound, Algorithm

1. INTRODUCTION

Hepatitis is a disease that attacks a vital organ, namely the liver. The liver is an organ that processes nutrients, filters blood, neutralizes toxins or dangerous substances, and fights infections. If inflammation occurs in the liver, liver function can be disrupted. Heavy alcohol use, toxins, certain drugs, autoimmune diseases, and certain medical conditions can cause hepatitis. The general types of hepatitis are hepatitis A, hepatitis B, hepatitis C, hepatitis D, hepatitis E, and Yellow Fever Virus (YFV) (M. Jefferies, et al., 2018).

Detecting Hepatitis can be done by detecting the levels of several compounds in the body. Increased levels of bilirubin and creatinine are positive indicators that someone has hepatitis, increased levels of Gamma Glutamyl Transferase (GGT) indicate that there is damage to the liver, which could be due to hepatitis, increased levels of alkaline phosphatase (ALP), alanine transferase (ALT) and aspartate aminotransferase (AST) is also a positive indicator of hepatitis. Changes in levels of these compounds can be an indication of whether a person is positive for hepatitis or vice versa. The process of detecting hepatitis can be done through machine learning.

Machine learning is a branch of artificial intelligence and computer science that uses data and algorithms to imitate how humans learn things. The process of machine learning learning something is called the learning process. The learning process is related to input such as training data, test data, and algorithms to play a role in decision-making and predictions. The learning process iteratively will gradually increase accuracy, resulting in more accurate predictions or decision making (I. C. Education, 2020). Machine learning can be applied in predicting patient hepatitis to maintain the accuracy of the patient's medical records.

The learning process in machine learning uses different models depending on the algorithm used. The algorithm applied will also vary depending on the case and function. Algorithms used in machine learning for the prediction and classification process include Support Vector Machine (SVM), Decision Tree, and K-Nearest-Neighbor (KNN). On this occasion, researchers compared six algorithms for each model to obtain the most accurate algorithm for the hepatitis disease classification process.

2. METHOD

The process carried out in this research consists of several stages. The research stages carried out are shown in the figure below.

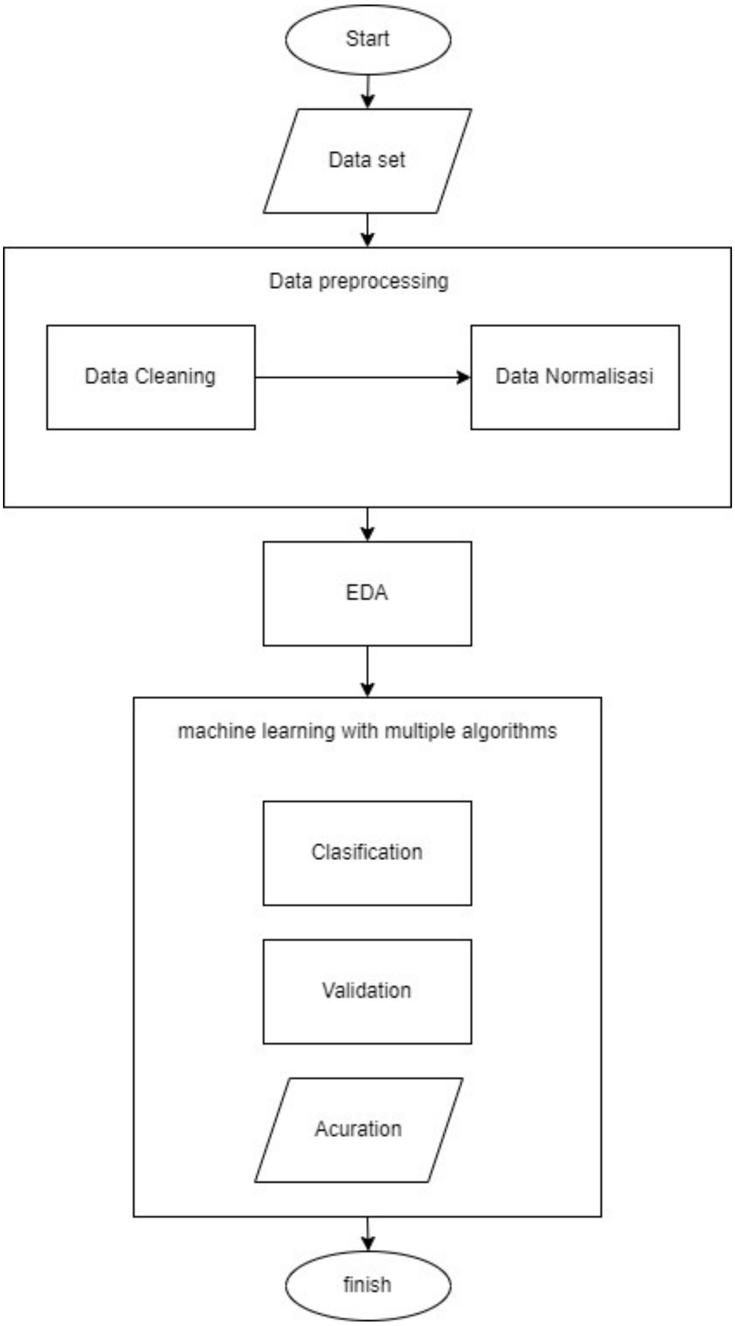


Figure 1. Research Flow Chart

Levels of compounds that are indicators of hepatitis are needed to create a model for detecting hepatitis. The data is also used to train the model to obtain the highest accuracy. Before creating the model, the data needs to go through an Exploratory Data Analysis (EDA) process to analyze the distribution of the data, a data pre-processing process to ensure the data can be used and/or improve the quality of the data (Admin, 2020) as well as data separation to divide the data into training and testing data for training the model. After all these processes have been passed, modeling can be done.

Exploratory Data Analysis (EDA) is used to determine the distribution of data in the dataset used. The dataset used consists of 12 categories of hepatitis indicators. EDA is used to determine the data distribution in each category and find relationships between these categories. This process follows three stages: univariate analysis to find the numerical distribution, bivariate analysis to find the distribution of one category with other categories, and multivariate analysis to find category relationships.

The stage after EDA is data pre-processing to ensure the dataset is ready for use and to improve the data quality. To eliminate anomalies in the dataset, pre-processing is carried out to handle these anomalies. Anomalies in the data are missing or empty data (missing values), duplicates and outliers. This stage will also transform the data to reduce the standard deviation so that the data range is not too far. After transformation, the Synthetic Minority Over-Sampling (SMOTE) technique balances the data.

Models are created based on different algorithms. The aim of using different algorithms is to find an algorithm that produces a model with the highest level of accuracy. In this case, the algorithms used are regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), Random Forest and x-gradient boost. Of all these algorithms, the algorithm with the highest accuracy will be taken to be implemented into the website application.

3. Results and Discussion

3.1. Exploratory Data Analysis

Exploratory data analysis (EDA) is a strategy for analyzing data to find hidden information that can provide new insights for a basis for decision-making [4]. Therefore, it is very important to carry out the EDA stage before modifying the existing data set to find hidden information in the original data. Implementing EDA will be divided into three main stages including:

1) Univariate Analysis

One variable is analyzed to determine the distribution of the numerical data. Does it tend to bias or normal distribution? In the picture on the next page, it can be seen that several variables have a positive or right curve, namely the variables ALP, ALT, AST, BIL, CREA, and GGT. For variables with a normal distribution, they are Unnamed: 0, Age, ALB, CHE, and CHOL. Then, the variable that has a negative or left curve is the PROT variable.

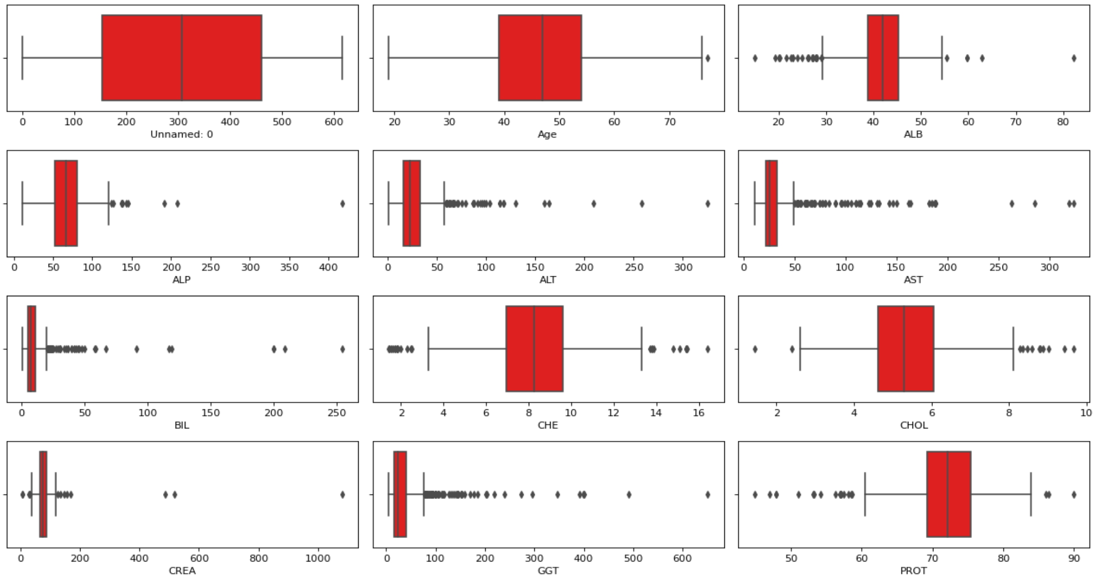


Figure 2. Univariate Analysis

DOI: https://doi.org/10.17509/seict.v2i1.34253
p- ISSN 2774-1656 e- ISSN 2774-1699

In the analysis of one variable, outlier data can also be found, which will affect the decline in model performance. So, to anticipate this, outlier data can be removed using the z-distribution technique. For example, the CREA variable has one data value of more than 1000, even though the data distribution is generally 0 – 200. Removing outliers using the z-distribution technique makes this data highly likely to be deleted because it is outside the normal distribution range.

2) Bivariate Analysis

Analysis of two variables will show how the value of a particular variable is distributed over other variables. In this case, the classes in the data set will be seen to see how large their distribution is in the Sex variable, shown in the image below. This analysis shows that most of the existing data set consists of patients who do not have hepatitis in both male (m) and female (f) genders. Apart from that, the majority of people with hepatitis are male.

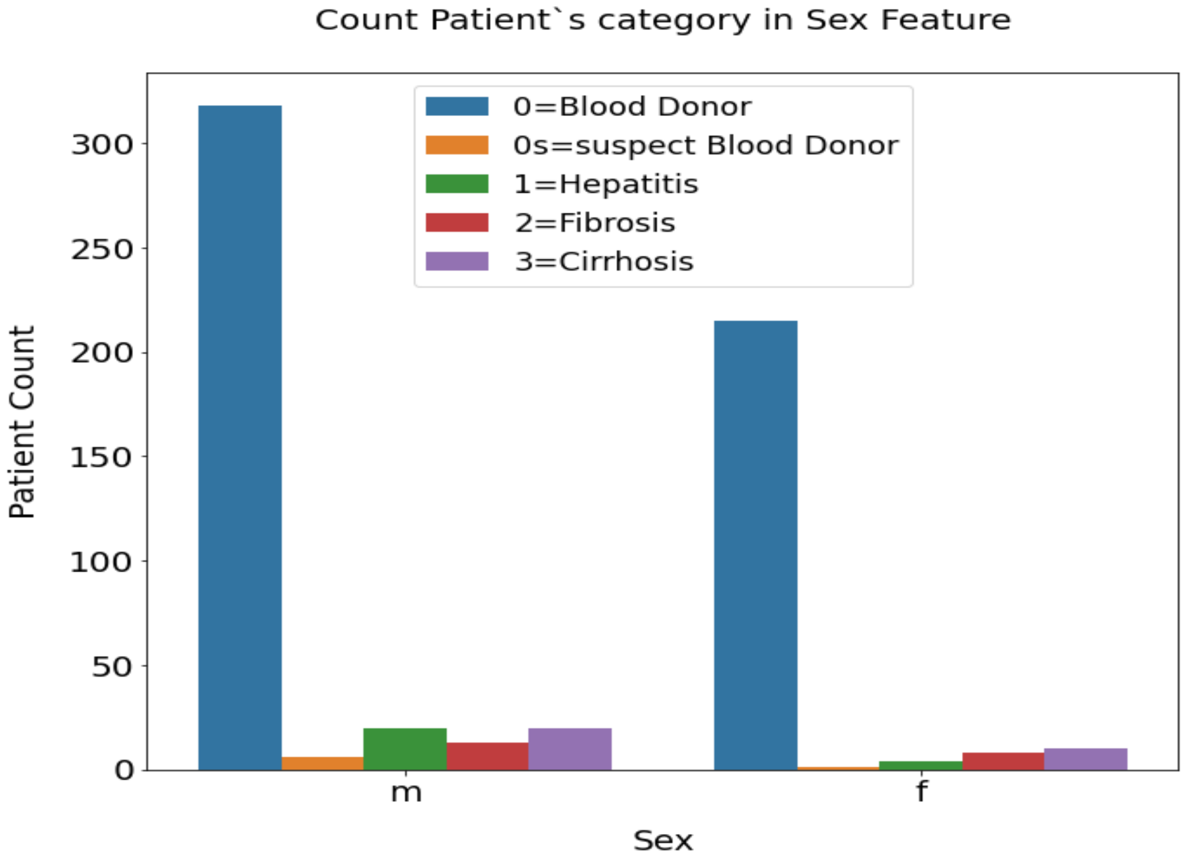
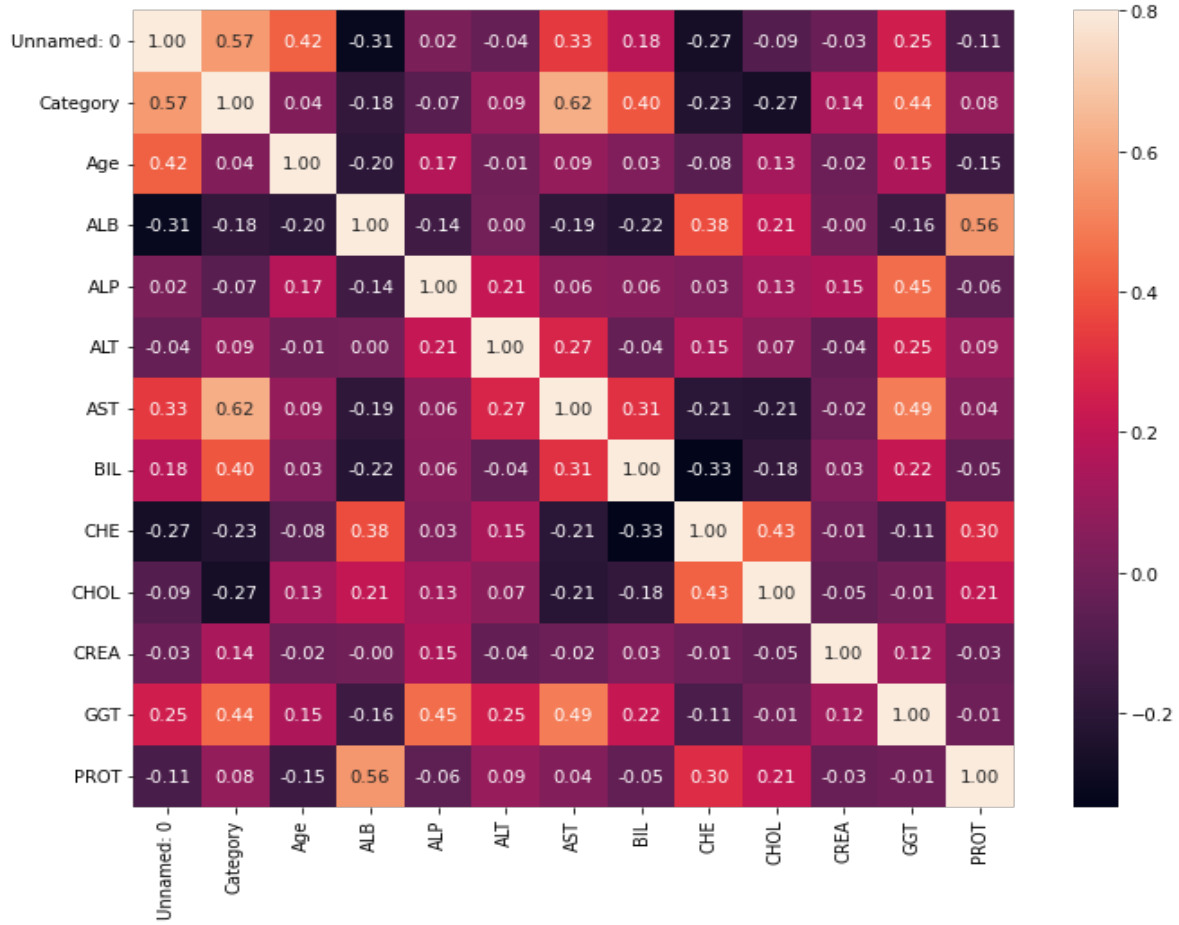


Figure 3. Bivariate Analysis

3) Multivariate Analysis

Multivariable analysis was carried out as a basis for machine learning modeling analysis. Multivariable analysis can determine which variables are most related to classes in a data set. The image on the next page shows the relationship between variables and their classes. The range of connectedness values given is -1 to 1. Where the value -1 means the connectedness is inversely proportional, but if the value is 1, the connectedness is directly proportional. This is different from the value 0, which means there is no connection between the two variables. Based on this picture, the Category class (determining whether someone has hepatitis) is closely related to the variables AST, GGT, and BIL, with connectedness values sequentially 0.62, 0.44, and 0.40.

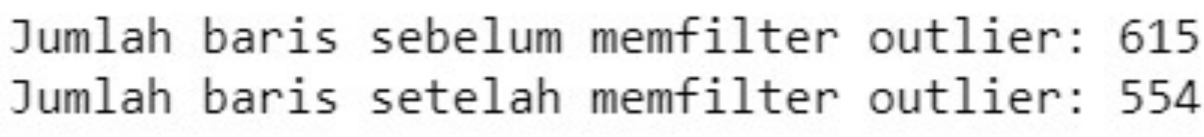


	Unnamed: 0	Category	Age	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
Unnamed: 0	1.00	0.57	0.42	-0.31	0.02	-0.04	0.33	0.18	-0.27	-0.09	-0.03	0.25	-0.11
Category	0.57	1.00	0.04	-0.18	-0.07	0.09	0.62	0.40	-0.23	-0.27	0.14	0.44	0.08
Age	0.42	0.04	1.00	-0.20	0.17	-0.01	0.09	0.03	-0.08	0.13	-0.02	0.15	-0.15
ALB	-0.31	-0.18	-0.20	1.00	-0.14	0.00	-0.19	-0.22	0.38	0.21	-0.00	-0.16	0.56
ALP	0.02	-0.07	0.17	-0.14	1.00	0.21	0.06	0.06	0.03	0.13	0.15	0.45	-0.06
ALT	-0.04	0.09	-0.01	0.00	0.21	1.00	0.27	-0.04	0.15	0.07	-0.04	0.25	0.09
AST	0.33	0.62	0.09	-0.19	0.06	0.27	1.00	0.31	-0.21	-0.21	-0.02	0.49	0.04
BIL	0.18	0.40	0.03	-0.22	0.06	-0.04	0.31	1.00	-0.33	-0.18	0.03	0.22	-0.05
CHE	-0.27	-0.23	-0.08	0.38	0.03	0.15	-0.21	-0.33	1.00	0.43	-0.01	-0.11	0.30
CHOL	-0.09	-0.27	0.13	0.21	0.13	0.07	-0.21	-0.18	0.43	1.00	-0.05	-0.01	0.21
CREA	-0.03	0.14	-0.02	-0.00	0.15	-0.04	-0.02	0.03	-0.01	-0.05	1.00	0.12	-0.03
GGT	0.25	0.44	0.15	-0.16	0.45	0.25	0.49	0.22	-0.11	-0.01	0.12	1.00	-0.01
PROT	-0.11	0.08	-0.15	0.56	-0.06	0.09	0.04	-0.05	0.30	0.21	-0.03	-0.01	1.00

Figure 4. Multivariate Analysis

3.2 Data Preprocessing

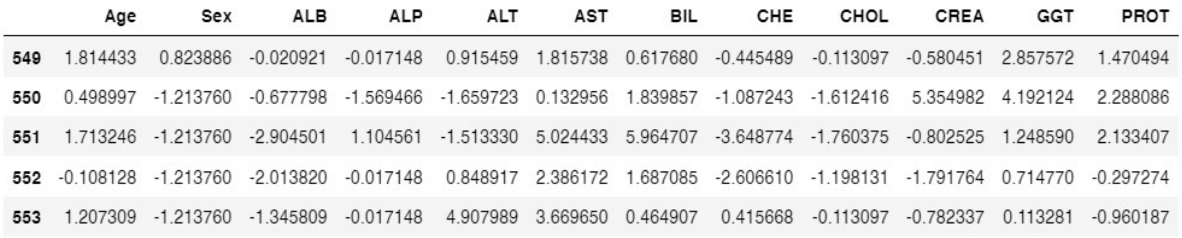
The data preprocessing consists of three main stages: dealing with outlier values, data transformation, and class balancing. Based on the normal distribution of data on the z-score value, outlier data that is outside 99.7% of the data distribution will be deleted and ultimately produce data of:



```
Jumlah baris sebelum memfilter outlier: 615
Jumlah baris setelah memfilter outlier: 554
```

Figure 5. Dealing with Outlier Data

The next preprocessing is to transform the data distribution using the StandardScaler module in the sklearn.preprocessing package. This transformation will create distribution values that are not too far apart or make the standard deviation values between numerical variables smaller. Ultimately, the data set values owned have the data distribution in the image below.



	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
549	1.814433	0.823886	-0.020921	-0.017148	0.915459	1.815738	0.617680	-0.445489	-0.113097	-0.580451	2.857572	1.470494
550	0.498997	-1.213760	-0.677798	-1.569466	-1.659723	0.132956	1.839857	-1.087243	-1.612416	5.354982	4.192124	2.288086
551	1.713246	-1.213760	-2.904501	1.104561	-1.513330	5.024433	5.964707	-3.648774	-1.760375	-0.802525	1.248590	2.133407
552	-0.108128	-1.213760	-2.013820	-0.017148	0.848917	2.386172	1.687085	-2.606610	-1.198131	-1.791764	0.714770	-0.297274
553	1.207309	-1.213760	-1.345809	-0.017148	4.907989	3.669650	0.464907	0.415668	-0.113097	-0.782337	0.113281	-0.960187

Figure 6. Data Transformation

Then, after transforming the existing data set, especially in classes, an oddity occurs that can affect the machine learning model, namely that there is an imbalance between each value, as shown in the image below:

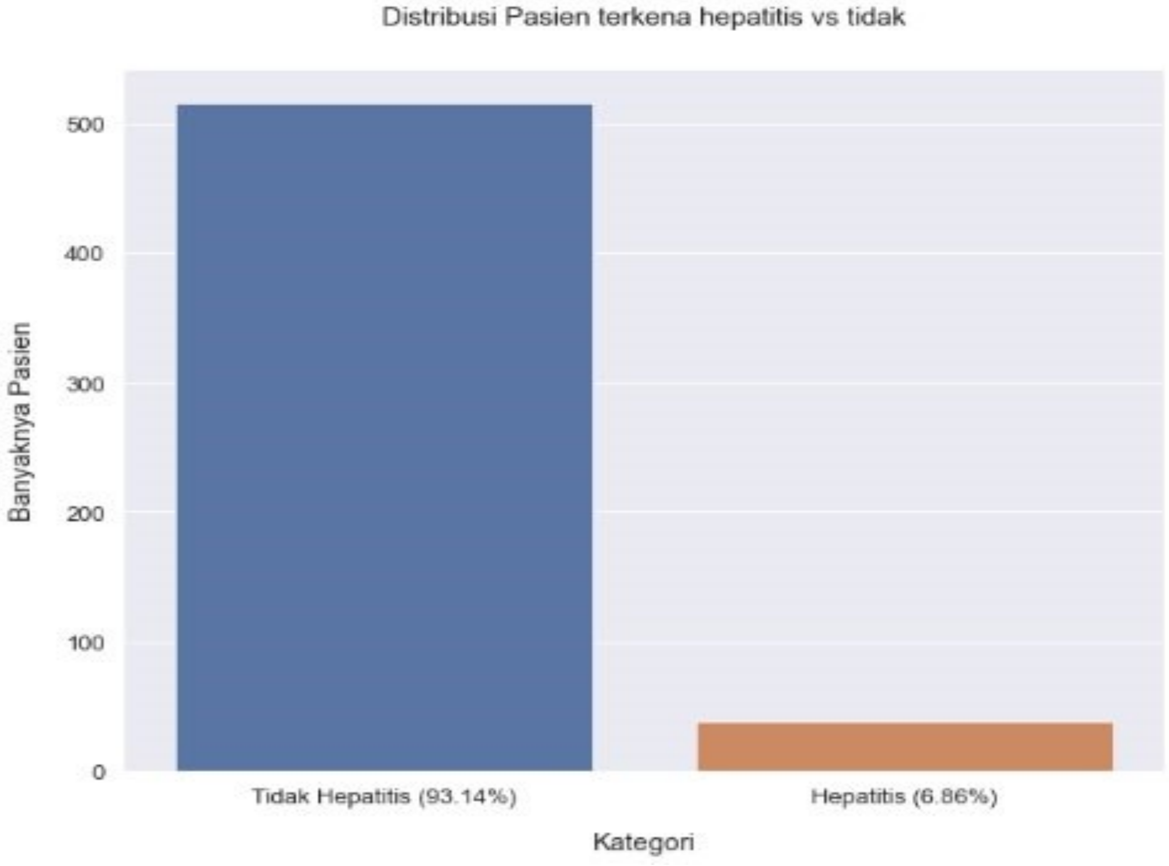


Figure 7. Imbalanced Class

To overcome this, you can use Synthetic Minority Over-Sampling (SMOTE). Based on previous research (A. Nikmatul Kasanah and U. Pujianto,2017), the SMOTE technique can improve the performance of machine learning modeling by up to 6.67%. Therefore, based on the imblearn.over_sampling package in the SMOTE module, the data set that was initially unbalanced becomes balanced as follows:

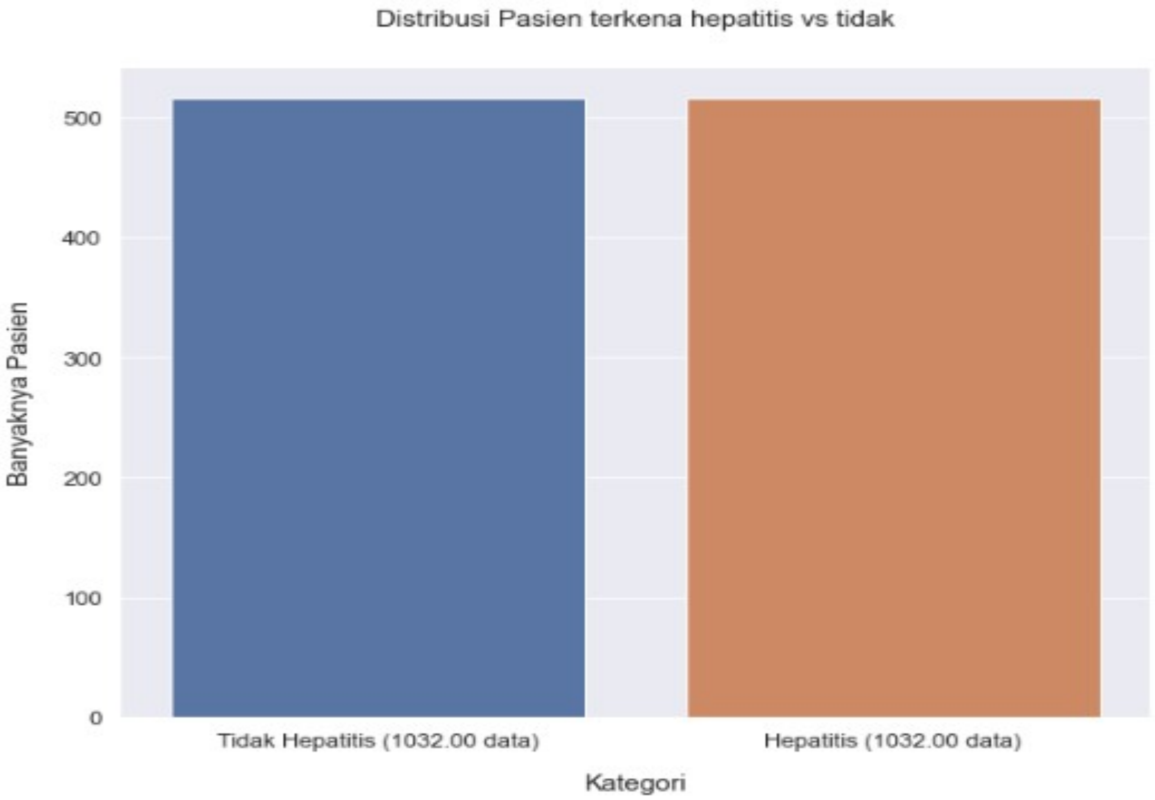


Figure 8. Balanced Class

3.3 Machine Learning Model Building

A machine learning model can be formed when the data set is ready and clean to produce optimal performance. The terminology "garbage in, garbage out" represents the modeling. The modeling carried out is in the form of categorizing data into which class type, making the model formed by applying classification. Before forming, the model will be split into two types of data sets: training data and test data. The percentages include 80% and 20% of the total data set. This is based on previous research that obtained optimal classification results using the Naïve Bayes algorithm (A. Byna and M. Basit, 2020). In its implementation, the train_test_split module is available in the sklearn.model_selection package. In this way, the owned data set is divided into two types.

The algorithm used in this research uses Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, and X-Gradient Boost (XGBoost). These algorithms are suitable for implementing data classification and have produced optimal performance from several studies conducted by (A. Handayani, et al., 2017) and also by (F. Rachman and S. W. Purnama, 2012). Both yielded accuracy above 80% in classification case studies. Therefore, this research will determine which algorithm provides the most optimal performance in hepatitis classification case studies.

Creating each algorithm begins with importing modules from the Scikit Learn library. After that, a new model object will be created, and training will be carried out using the training data. After that, the model will be used to predict test data, and evaluation will be carried out based on the prediction results.

3.4 Evaluating Machine Learning Models

The metrics used in evaluating machine learning models are the confusion matrix and the AUC score. The results can be seen in the table on the next page.

Table 1. Comparative Evaluation of Classification Algorithms

	Accuracy	Precision	**Recall**	F1-Score	AUC
Logistic Regression	0.932	0.937	0.918	0.927	0.930
Support Vector Machine	0.995	0.989	**1.000**	0.949	1.000
Decision Tree	0.966	0.959	0.969	0.964	0.970
K-Nearest Neighbor	0.966	0.933	**1.000**	0.965	0.970
Random Forest	0.985	0.979	0.989	0.984	0.990
XGBoost	0.975	0.969	0.979	0.974	0.980

In evaluating a machine learning model, especially in the case of classification, the Recall metric is the main basis for selecting the best model. This is because the Recall metric focuses on the level of success in classifying data. Therefore, the greater the Recall value, the better a model classifies existing data categories (N. A. Setifani, 2020). Of the six models that created the highest Recall value was held by the Support Vector Machine (SVM) algorithm and also the K-Nearest Neighbor.

However, it is necessary to evaluate the two models further because there is a high possibility that they are overfitting the training data. The second metric to use is the accuracy of the two models, both on the training and test data. The accuracy values for both can be seen in the table below.

Table 2. Comparison of Algorithm Evaluation with Best Recall

	Training Data Accuracy	Test Data Accuracy
Support Vector Machine	0.993	**0.995**
K-Nearest Neighbor	0.986	0.966

As a consideration between the two, the accuracy value of test data is the main determining metric for determining the best model to use. This is because the accuracy value of the test data is an overall classification test carried out on data without classes or labels, which is very suitable for later use. After all, the user will enter data as variables without labels (D. Carty, 2021) . The following are the confusion matrix results from the Support Vector Machine algorithm model.

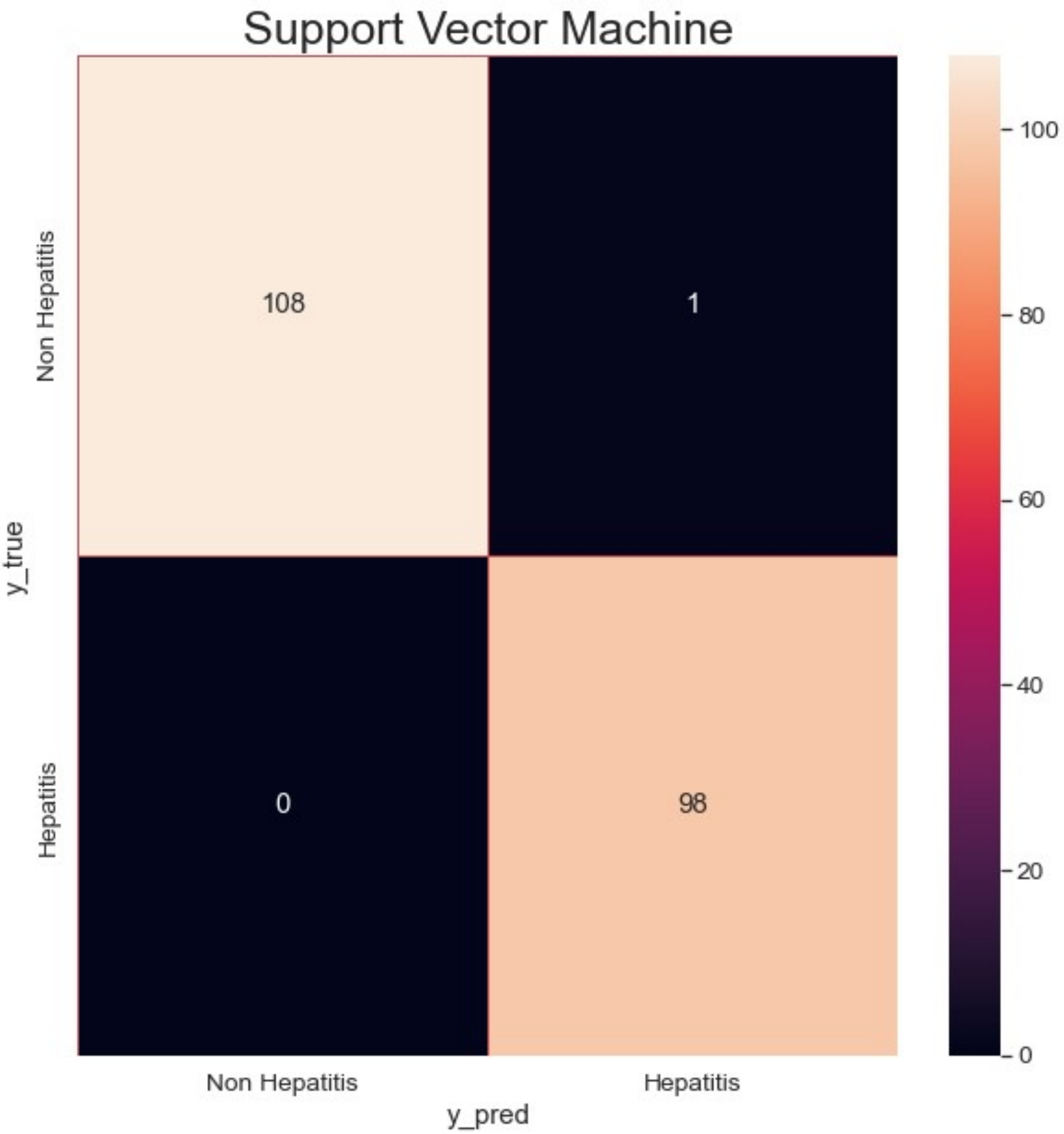


Figure 9. Support Vector Machine Confusion Matrix Model

From the confusion matrix obtained, the model only had one error regarding the data on patients who did not have hepatitis. Still, when classifying these patients, they were categorized as hepatitis patients. This is, of course, still within reasonable limits because errors are only found in False Positive data.

4. CONCLUSION

Based on the discussion regarding the comparison of the Logistic Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, and X-Gradient Boost

(XGBoost) algorithms on the hepatitis dataset obtained from the UCI Machine Learning Repository, it can be concluded that the Support Vector Machine algorithm more accurate in classifying hepatitis when compared with five other algorithms. The Support Vector Machine algorithm obtains a Recall value of 100% together with the K-Nearest Neighbor, where the Recall value is the main basis for classifying data because it focuses on the success rate of classifying the data. The next stage is testing accuracy on training and test data, and the Support Vector Machine algorithm can outperform the K-Nearest Neighbor by obtaining a higher accuracy value. In the overall confusion matrix results, the Support Vector Machine algorithm only got one error found in the False Positive data.

7. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

8. REFERENCES

M. Jefferies, B. Rauff, H. Rashid, T. Lam, and S. Rafiq, "Update on global epidemiology of viral hepatitis and preventive strategies," WJCC, vol. 6, no. 13, pp. 589–599, 2018.

I. C. Education, "What is Machine Learning?," ibm.com, 2020. https://www.ibm.com/cloud/learn/machine-learning.

Admin, "Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data," tableau.com, 2020. https://www.tableau.com/learn/articles/what-is-data-cleaning.

E. D. Wahyuni, A. A. Arifiyanti, and M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," Pros. Nas. Rekayasa Teknol. Ind. dan Inf. XIV Tahun 2019, vol. 2019, no. November, pp. 263–269, 2019, [Online]. Available: http://journal.itny.ac.id/index.php/ReTII.

A. Nikmatul Kasanah and U. Pujianto, "Terakreditasi SINTA Peringkat 2 Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 1, no. 3, pp. 196–201, 2017.

A. Byna and M. Basit, "Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes," J. Sisfokom (Sistem Inf. dan Komputer), vol. 9, no. 3, pp. 407–411, 2020, doi: 10.32736/sisfokom.v9i3.1023.

A. Handayani, A. Jamal, and A. A. Septiandri, "Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara," vol. 6, no. 4, pp. 394–403, 2017.

F. Rachman and S. W. Purnama, "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)," J. Sains Dan Seni Its, vol. 1, no. 1, pp. 130–135, 2012.

N. A. Setifani, D. N. Fitriana, and A. Yusuf, “Perbandingan Algoritma Naïve Bayes, Svm, Dan Decision Tree Untuk Klasifikasi Sms Spam,” JUSIM (Jurnal Sist. Inf. Musirawas), vol. 5, no. 02, pp. 153–160, 2020, doi: 10.32767/jusim.v5i02.956.

D. Carty, “Training Data vs. Validation Data vs. Test Data for ML Algorithms,” applause.com, 2021. https://www.applause.com/blog/training-data-validation-data-vs-test-data.

DOI: https://doi.org/10.17509/seict.v2i1.34253
p- ISSN 2774-1656 e- ISSN 2774-1699