Jurnal Pendidikan Matematika (Kudus)
P-ISSN 2615-3939 | E-ISSN 2723-1186
https://journal.iainkudus.ac.id/index.php/jmtk
DOI: http://dx.doi.org/10.21043/jpmk.v5i2.17052
Volume 5, Number 2, December 2022, pp. 237-254

Classification of Hepatitis Patients Using Logistic Regression and Support Vector Machines Methods

Diana Nurlaily[1*], Yoga Prastya Irfandi[2], Noviyanti Santoso[3], Siti Qomariyah[4], Dandy Wibowo[5]
[1]*Statistics Department, Institut Teknologi Kalimantan, Balikpapan, Indonesia*
[2]*Statistics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[3]*Graduate School of Science and Technology, Kumamoto University, Kumamoto, Japan*
[4]*Mathematics Education Department, Institut Agama Islam Negeri Kudus, Kudus, Indonesia*
[5]*Institute of Computer Science, University of Silesia, Sosnowiec, Poland*
** Correspondence: sitiqomariyah@iainkudus.ac.id*

Abstract

Hepatitis is an inflammatory disease of the liver. The virus often causes hepatitis and it becomes the number one world health problem. From 2019 to 2020, there were 1.5 million new cases of hepatitis B and C infection per year. WHO (World Health Organization) aims to eliminate hepatitis by 2030. Based on this problem, it is necessary to classify which health indicators may be vulnerable to the survival of hepatitis patients. This research aims to obtain the best method for classifying hepatitis patients by comparing the logistic regression method and SVM (Support Vector Machines). The classification using logistic regression and SVM is the suitable alternative for this case because the response category is binary data. This research is quantitative research and the researcher uses the hepatitis data set obtained from the UCI repository learning machine. The hepatitis data set contains 19 predictive variables (6 continuous and 13 discrete variables). The patients are divided into two groups, living, and dead patients' groups. The results show that the best accuracy value produced by using the logistic regression method is 79.3%, and by using the SVM method is 81.94%. Thus, the best classification result for the hepatitis data set is the holdout stratified SVM method using Kernel radians with an accuracy value of 81.94%. This result indicates that the holdout stratified SVM method using Kernel radians can classify hepatitis patients' data.

Keywords: Hepatitis; Logistic Regression; Support Vector Machines

Abstrak

Hepatitis adalah penyakit peradangan pada hati. Hepatitis sering disebabkan oleh virus. Hepatitis termasuk masalah kesehatan dunia. Tahun 2019 sampai dengan 2020, terdapat 1,5 juta kasus baru infeksi hepatitis B dan C per tahun. WHO (World Health Organization) bertujuan untuk menghilangkan penyakit hepatitis pada tahun

2030. Berpondasikan masalah tersebut, perlu adanya pengklasifikasian untuk mengetahui indikator kesehatan mana yang mungkin rentan terhadap kelangsungan hidup pasien hepatitis. Tujuan penelitian ini untuk mendapatkan metode terbaik dalam mengklasifikasikan pasien hepatitis dengan cara membandingkan metode regresi logistik dan SVM (Support Vector Machines). Klasifikasi menggunakan regresi logistik dan SVM merupakan alternatif yang tepat untuk kasus ini, karena kategori respon adalah data biner. Penelitian ini merupakan penelitian kuantitatif. Penelitian ini menggunakan dataset hepatitis yang diperoleh dari UCI machine learning repository. Kumpulan data hepatitis berisi 19 variabel prediksi (6 variabel kontinu dan 13 variabel diskrit). Pasien dibagi menjadi dua kelas yaitu hidup dan mati. Hasil penelitian menunjukkan bahwa nilai akurasi terbaik yang dihasilkan metode regresi logistik adalah 79.3% sementara menggunakan metode SVM adalah 81.94%. Jadi hasil klasifikasi terbaik untuk dataset hepatitis adalah metode SVM holdout stratified menggunakan kernel radian dengan akurasi sebesar 81,94%. Hasil ini mengindikasikan bahwa metode SVM holdout stratified menggunakan kernel radian dapat digunakan untuk mengklasifikasikan data pasien hepatitis.

Kata Kunci: Hepatitis, Regresi Logistik, *Support Vector Machines*

Introduction

Hepatitis is a liver inflammation that is often caused by a virus. However, hepatitis can also be caused by other factors such as autoimmune, alcohol, and drugs (Martinez, 1996). Hepatitis is a worldwide health problem. In 2019-2020, there were 1.5 million new cases per year of hepatitis B and C infection (WHO, 2021). The hepatitis virus is one thing that causes high morbidity and mortality in humans (Zuckerman & Baron, 1996).

The 63rd World Health Assembly (WHA) 2010 was held at the United Nations Geneve Assembly Hall, Switzerland. One of the resolutions is all countries around the world conduct a comprehensive hepatitis prevention, starting from prevention to treatment (Kemenkes, 2022). WHO (World Health Organization) aims to eliminate hepatitis by 2030. Therefore, there are several targets have to be achieved by all countries to reach the goal. These targets include reducing new hepatitis B and C infections by 90%, reducing the occurrence of deaths due to hepatitis, liver cirrhosis and cancer by 65%, ensuring that at least 90% of people with hepatitis B and C are diagnosed, and at least 80% the detected people get proper care and treatment (WHO, 2022).

Hepatitis patients generally do not realize that they are being infected. It is difficult to recognize because the symptoms are not immediately felt before their severe condition (Chow & Chow, 2006). A blood test is used to determine the patients' condition by focusing on the value of some indicators of hepatitis disease. Preventive actions can be conducted by recognizing the pattern of patients from

blood test results and the patients' physical condition to maximize treatment measures (García, Luengo, & Herrera, 2015).

In this research, the researcher classifies hepatitis patients' data. The response variable in the hepatitis data is categorical data. The response is divided into living and dead patient groups. Hence, this problem can be solved using the classification methods. They are conventional statistical methods and learning machine methods to classify hepatitis data. The differences between traditional and learning machine methods are the learning machine is data-driven and the conventional statistical way is model-driven (Ley et al., 2022).

One of the algorithms used in the conventional statistical method is logistic regression (Samosir, Wilandari, & Yasin, 2015; Ley et al., 2022). Logistic regression is an efficient and straightforward linear and binary classification method. Logistic regression is also flexible and easily used function (Edgar & Manz, 2017). If the dependent variable is a categorical variable and the independent variable is either a continuous or categorical variable, we can use logistic regression (Park, 2013).

The machine learning method studies the pattern of the dataset and it can improve data experiences automatically (Sarker, 2021). The learning Machine method comprises various algorithms appropriate for any dataset (Sarker, 2021). One of the popular machine learning algorithms is the Support Vector Machine (SVM) (Shihong, Ping, & Peiyi, 2003). SVM is widely used in many different research fields such as health (Kistenev, Vrazhnov, Shnaider, & Zuhayri, 2022; Ikerionwu et al., 2022), business (Kumbure, Lohrmann, Luukka, & Porras, 2022), economics (Raman, 2022), geophysics (K C, Bhusal, Gautam, & Rupakhety, 2022), government (Qomariyah, Iriawan, & Fithriasari, 2019) and other researches. The advantage of SVM is this algorithm can be used in complex datasets, linear, and nonlinear data (Huang et al., 2018).

Nurlaily, Irhamah, Purnami, and Kuswanto (2019) examined the classification of unbalanced microarray data and found that the use of SVM-ACO-SMOTE provides better performance than without using SMOTE. Wilkinson, Mamas, and Kontopantelis (2022) used simulated data and stated that the logistic regression method was superior to the propensity score for extensive data.

The accuracy of health data classification is very essential to avoid misdiagnosis. Then, through this research the researcher compares logistic regression and SVM methods. This research also aims to see the higher accuracy in the hepatitis dataset using classical and learning machine methods. Not only logistic regression and SVM methods but also the researcher compares training and testing

data separation method. Moreover, the classification model using all variables and reduced variables is also compared in this research.

Several studies have compared the logistic regression and SVM methods. (Novianti & Purnami, 2012) have reached the diagnostic analysis of breast cancer patients using logistic regression and SVM methods, where it was found that the SVM method provides a higher accuracy value than logistic regression method. (Utami, 2018) conducted research to compare SVM classification and logistic regression methods performance on graduation punctuality problem of FMIPA UNTAD students. The result was that the misclassification using SVM method was smaller than the classification using logistic regression method. In 2021, Hasanah and Widjarnarko (2021) compared the propensity score matching-support vector machine method and propensity score matching-binary logistic regression methods on HIV/AIDS cases and found that the number of bias was reduced using the Propensity Score Matching SVM method and after the matching process, the result was smaller than the result using regression logistics method.

The classification would be very essential to determine which health indicators are likely to be susceptible to the survival of hepatitis patients. The classification using a Support Vector Machine and logistic regression methods is an appropriate alternative way for this case since the response categories are binary data (Chen, Lu, Yang, & Li, 2004). Thus, the researcher in this research compares the logistic regression and SVM methods to get which one is the best method is in classifying hepatitis patients.

Method

In this research, the researcher uses the quantitative approach for classification analysis. In this research researcher uses SVM and Logistic Regression methods to analyze hepatitis patients data set. The flow chart for analyzing the hepatitis patient's data set is seen in Figure 1.

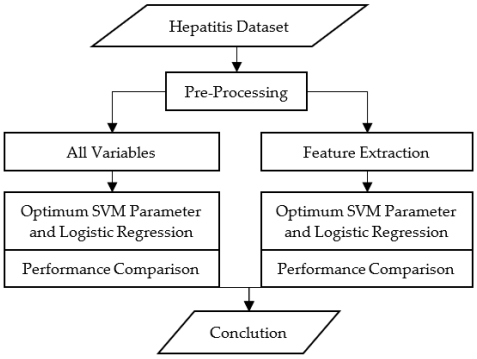

Figure 1. Flow Chart

Data Set Description

The hepatitis patient dataset obtained from the UCI Learning Machine Repository is used in this research. The data set contains 19 prediction variables. The patients are divided into two groups, living patients and dead patient's groups.

Preprocessing Data Set

Data preparation is a technique of initializing data correctly to serve as input for a specific DM algorithm. It changes prior useless data into new data that fit~~s~~ a collecting data process. The steps to the data preparation are data cleaning, transformation, normalization, integration, imputation of the missing data, and noise identification. The data reduction method can get a reduced representation from the original dataset. The new data set is much smaller in volume and tries to maintain most of the integrity of the original data. The aim is to provide the collection process with a mechanism to generate the same outcome when applied over reduced data simultaneously when collection process becomes efficient. Data reduction technique~~s~~ includes feature selection and extraction. Principal Component Analysis is a data reduction method. The main idea is to find a set of linear transformations of the initial variables that can explain most of the variance with a smaller number of variables. Therefore, it looks for k n orthogonal vectors with n dimensions that best represent the data, where k value is less than or equal to n value $(k \leq n)$. The basic procedure is as follows:

a. Inputting data normalization.
b. Computing k orthonormal vectors to supply a premise for the normalized input information.
c. Ordering the principal components according to the power given by the associated eigenvalues.
d. Reducing data by removing weaker and low-variance components.
e. The final output from PCA is a new set of variables representing the original data set.

Feature selection is the stage of selecting the optimal subset of features based on specific criteria. Feature selection aims to identify essential elements in the data set and discard others as redundant or irrelevant (Batista, Prati, & Monard, 2004).

Feature Extraction

Predictor variables in this problem have 19 variables. Hence, it is necessary to make a variable reduction with feature extraction using principal component analysis.

Building A Model using Logistic Regression and SVM methods

Predicting classes or groups models for hepatitis patients categorized into living patients or dead patients group using logistic regression and SVM methods. In each analysis, the data are divided into training and testing data. Cross-validation tests used in this analysis are holdout stratified, holdout non-stratified, K fold stratified, and K fold non-stratified. Two models are built: a model with reduced variables and a model with all predictor variables. The goal is to compare which method is better.

Logistic regression is a data analysis technique to explain the response variable with one or more dependent variables. In binary logistic regression, the dependent variable is grouped into two categories, namely 0 and 1; the dependent variable follows the Bernoulli distribution with the probability function as follows $f(y_i) = \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}$, with y_i =0, 1. The logistic regression model is as below

$$\pi(\mathrm{x}) = \frac{\exp\left(\beta_0 + \beta_1 \mathrm{x}_1 + \cdots + \beta_\mathrm{p} \mathrm{x}_\mathrm{p}\right)}{1 + \exp\left(\beta_0 + \beta_1 \mathrm{x}_1 + \cdots + \beta_\mathrm{p} \mathrm{x}_\mathrm{p}\right)} \qquad (1)$$

The equation (1) transformed becomes logit transformation as $\pi(\mathrm{x})$ to get g(x) linear parameters for more straightforward regression parameter estimation, which follow Equation (2) (Gail, Krickeberg, Samet, Tsiatis, & Wong, 2012).

$$g(\mathrm{x}) = \ln\left[\frac{\exp\left(\beta_0 + \beta_1 \mathrm{x}_1 + \cdots + \beta_\mathrm{p} \mathrm{x}_\mathrm{p}\right)}{1 + \exp\left(\beta_0 + \beta_1 \mathrm{x}_1 + \cdots + \beta_\mathrm{p} \mathrm{x}_\mathrm{p}\right)}\right] \qquad (2)$$

The maximum Likelihood Estimator (MLE) is used to estimate the parameters contained in the logistic regression model. This method is predicted by maximizing likelihood function. The raised likelihood function is

$$L(\beta) = \sum_{j=0}^{p}\left[\sum_{i=1}^{n} y_i x_{ij}\right]\beta_j - \sum_{i=1}^{n} \ln\left[1 + \exp\sum_{j=0}^{p} \beta_j \right. \quad (3)$$

g(x) is defined p again then equated to zero, but this method often obtains implicit results. Therefore, the Newton-Rhapson cycle strategy is utilized to

maximize the probability of work. Parameter testing in logistic regression is carried out simultaneously and/or individually. The statistics test used in the simultaneous test is the G test or the likelihood ratio test. At the same time, the statistics test on the partial test is the Wald test. One of the measures used to explain the coefficient of the predictor variable is the Odds ratio (Gail, Krickeberg, Samet, Tsiatis, & Wong, 2012). The odds ratio displays the proportion of the likelihood that an event will occur or not occur. If the odds ratio is < 1, there is always a negative correlation between the independent and dependent variables. If the odds ratio is > 1, then there is a positive correlation between the predictor variable and the response variable. The statistic test used for the fit test model is the Hosmer - Lemeshow Test (C).

> *SVM is a new machine learning technique proposed by Vapnik et al. SVM is a technique for classifying linear and nonlinear data. This method uses nonlinear mapping to transform data into higher dimensions [3]. The thought of classification with SVM is to discover the finest separator (hyperplane) between two classes of information. SVM can be utilized to fathom issues with tall dimensional information and little preparing tests (Pal & Mather, 2005).*

The principle of SVM is to classify linear data, but it can also be used for non-linear data by combining the kernel's trick concepts in a high-dimensional feature space. In linear classification, SVM can be separated into two types, specifically straightly distinct and directly indivisible. Figure 2 shows an illustration of the linear classification of the SVM method. Generally, the data in the real life are rarely linear, primarily non-linear. Nonlinear data are often found in real problems; SVM is modified by incorporating Kernel functions from nonlinear functions.

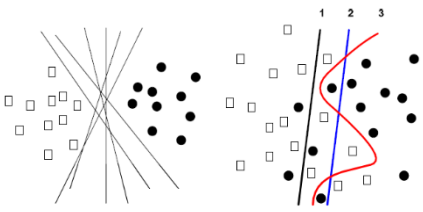

Figure 2. Linearly Separable (Left) and Linearly Nonseparable (Right)

The SVM (non-linear separable) optimization equation can be written as the following equation (Chen, Lu, Yang, & Li, 2004).

$$\max_{\alpha} L_D = \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \quad (4)$$

with constraint,

$$\sum_{i=1}^{n} \alpha_i y_i, 0 \leq \alpha_i \leq C, i = 1, 2, \dots, n, \alpha_i \geq 0 \quad (5)$$

The value of the classification function (score) can be formulated as Equation (García, Luengo, & Herrera, 2015).

$$f(x_i) = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{\alpha}_i y_i K(x_i, x_j) + b\right) \quad (6)$$

Commonly used kernel functions are linear kernel, polynomial kernel, radial basis function, sigmoid kernel. The function in Equation (6) can be changed into the following decision function in Equation (7).

$$g(x_i) = \left(\sum_{i=1}^{n} \hat{\alpha}_i y_i K(x_i, x_j) + b\right) \quad (7)$$

The parameters of γ and C are kernel parameters, α_i is the Lagrange Multiplier, which is zero or positive ($\alpha_i \geq 0$) and parameter C as a penalty due to errors in the classification, and the value is determined by user. The requirement of a function to be a kernel function follows Mercer's theorem, which states that the resulting kernel matrix must be semi-definite positive. The standard kernel function used in the SVM method is

a. Linear

b. $K(x_i, x_j) = \mathrm{x_i^T x_j}$

c. Polynomial

d. $K(\mathrm{x_i}, \mathrm{x_j}) = \left(\gamma \mathrm{x_i^T x_j} + r\right)^p, \gamma > 0$

e. Radial Basis Function (RBF)

f. $K(\mathrm{x_i}, \mathrm{x}) = \exp(-\gamma \|\mathrm{x} - \mathrm{x_i}\|^2)$

The RBF kernel function is recommended for the first test. The RBF kernel has the same results as the linear kernel on certain parameters. It has properties like the sigmoid kernel function with certain parameters and a small range of values [0, 1](Hsu, Chang, & Lin, 2003). Some of the advantages of SVM can be used for both classification and regression. It has special generalization ability, especially for problem of sample with small size, SVM has no trouble of local minimum and high dimensional problems (Chen, Lu, Yang, & Li, 2004).

Prediction Class of Testing Data Set

It is conducted based on the model that has been formed from training data and the prediction class of hepatitis patients in testing data.

Calculating the Accuracy

The values contained in the confusion matrices are used to measure the classification performance. The confusion matrix (see Table 1) contains actual and

predicted classifications from the prediction result on testing data, the accuracy is calculated by using confusion matrix [8]

Table 1. Confusion Matrix

Observation		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Accuracy} = \frac{\text{true positive} + \text{true negatif}}{\text{total observation}} \qquad (8)$$

Results

Preprocessing Hepatitis Data Set

The hepatitis patient's data consists 19 predictor variables, 6 of them are continuous and 13 are discrete. From the 19 variables, 15 variables have missing values as follow (see Table 2):

Table 2. The Missing Value Variables

Variable	The Amount of Missing Value
Steroid	1
Fatigue	1
Malaise	1
Anorexia	1
Liver Big	10
Liver Firm	11
Spleen Palpable	5
Spiders	5
Ascites	5
Varices	5
Bilirubin	6
Alk Phosphate	29
SGOT	4
Albumin	16
Protime	67

Categorical variables that have missing values are imputed with the pmm method while for continuous variables are assigned using the mean method. After

the missing value problem overcomes, the predictor variables are reduced because they have a large number. It aims to get a good model with a smaller volume of variables. The method used to reduce variables is Principal Component Analysis (PCA). The first step is to determine the number of components used based on the eigenvalues.

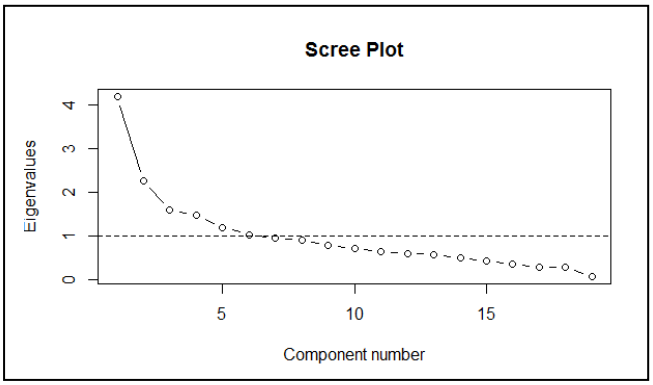

Figure 3. Scree Plot for Determine the Number of Principal Component

Figure 3 shows that the number of components having eigenvalue of more than 1 is five components, but the 6th component is located exactly in the eigenvalue 1. If the component used is five, then the cumulative value explained by that component is 0.89. Moreover, if the number of components used is 6, having cumulative value which can be explained by 0.90. Hence, the number of features selected is five because there is no significant difference between the cumulative value of 5 members and six components. Here are the details and their variables:

Table 3. Component PCA

Component	Variable
Component 1	Fatigue, Spleen Palpable, Spiders, Ascites, Varices, Bilirubin, Albumin, Protime
Component 2	Steroid, Histology
Component 3	Liver Big, Liver Firm
Component 4	Sex, Anorexia
Component 5	Age, Antivirals, Alk. Phosphate, SGOT

Classification of Hepatitis Patients using Logistic Regression

In the classification of hepatitis patients' data using logistic regression, data is separated into training and testing. The data used are the original data, which have been reduced using PCA compared to the better. The methods used to separate data are holdout and k fold-cv. In each method, the data are separated into stratified and non-stratified. Based on training data, a binary linear regression model was built for holdout stratified, non-stratified, k fold cv-stratified and k fold cv-non stratified.

Table 4. Accuracy Binary Logreg For Reduced Variables

Cross Validation	Accuracy (%)					
	Iteration					Mean
	1	2	3	4	5	
Holdout Stratified	0.633	0.767	0.733	0.8	0.7	0.727
Holdout Non-Stratified	0.806	0.709	0.870	0.806	0.774	0.793
K fold CV Stratified	0.8	0.806	0.7	0.733	0.806	0.769
K Fold CV Non-Stratified	0.935	0.935	0.774	0.580	0.580	0.761

The model that has been formed conducted predictions on training data and calculated the value of accuracy, specificity, and sensitivity. Here is the value of the training data prediction accuracy, specificity and sensitivity with reduction data. Table 4 shows that the most fantastic accuracy value is obtained from K Fold CV non-stratified (93,5%). Based on the accuracy average of 5 repetitions, the highest value is 79,3% from the holdout non-stratified, and the smallest is from the holdout stratified with a value of 72,7%. Then, based on the accuracy average of the logistic regression method with reduced variables, holdout non- stratified has the highest accuracy.

The following is the value of the logistic regression model based on accuracy from training and applied to prediction testing data using data with all variables without feature extraction. Table 5 shows that the most outstanding accuracy value is obtained from the holdout stratified (83,3%). On accuracy average, the highest value is 72,7% from the holdout stratified, and the smallest is from the holdout non-stratified with a value of 69,7%. Therefore, the holdout stratified has the highest accuracy based on the accuracy average of the logistic regression method for data with all variables.

Table 5. Accuracy Binary Logreg for Data with All Variables

Cross Validation	Accuracy (%)					
	Iteration					Mean
	1	2	3	4	5	
Stratified	0.733	0.667	0.833	0.7	0.7	0.727
Holdout Non-Stratified	0.61	0.709	0.709	0.742	0.71	0.697
K-Fold Stratified	0.75	0.656	0.633	0.677	0.87	0.717
K-Fold Non-Stratified	0.774	0903	0.774	0.613	0.516	0.716

Classifying Hepatitis Patients using SVM

SVM is one of the methods that can be implemented for classification. There are two classification scenarios. The first grouping uses all the variables, while the second one uses the feature extraction result in the previous discussion. To select

the optimum SVM parameter, researcher uses the accuracy value. The best classification has the most considerable accuracy value. The accuracy of SVM using some kernel and combined with cross-validation method for data with reduced variables is in Table 6.

Table 6. Accuracy SVM For Reduced Variables

Cross Validation	Kernel	Accuracy (%) Iteration 1	2	3	4	5	Mean
Holdout Stratified	Linear	0,806	0,806	0,806	0,806	0,806	80,65%
Holdout Random	Linear	0,903	0,742	0,710	0,774	0,871	80,00%
K-Fold Stratified	Linear	0,806	0,774	0,806	0,774	0,806	79,35%
K-Fold Unstratified	Linear	0,935	0,935	0,774	0,710	0,613	79,35%
Holdout Stratified	Polynomial	0,806	0,806	0,806	0,806	0,839	**81,29%**
Holdout Random	Polynomial	0,903	0,742	0,710	0,774	0,871	80,00%
K-Fold Stratified	Polynomial	0,806	0,774	0,806	0,645	0,581	72,26%
K-Fold Unstratified	Polynomial	0,903	0,935	0,774	0,710	0,613	78,71%
Holdout Stratified	Radian	0,806	0,806	0,806	0,806	0,806	80,65%
Holdout Random	Radian	0,903	0,742	0,710	0,774	0,871	80,00%
K-Fold Stratified	Radian	0,806	0,774	0,806	0,452	0,387	64,52%
K-Fold Unstratified	Radian	0,935	0,935	0,774	0,710	0,631	79,35%

In Table 6, it is known that the highest accuracy average value is obtained from holdout stratified using a polynomial kernel with 81,29%. The smallest is from K Fold stratified using radian kernel with a value of 64,52%. Hence, based on the accuracy average of the SVM method for data with reduced variables, holdout stratified using polynomial kernel has the highest accuracy. The value of the accuracy-based SVM utilizing some seed combined with the cross-validation method for data with all variables is in Table 7.

Table 7 shows that the highest accuracy average value is obtained from holdout stratified using radian kernel with the value of 81,94%. The smallest is from K Fold stratified and K Fold non-stratified using a linear kernel with a value of 67,74%. Therefore, based on the accuracy average of the SVM method for data with reduced variables, holdout stratified using radian kernel has the highest accuracy.

Table 7. The SVM Accuracy For Data With All Variables

Cross Validation	Kernel	Accuracy (%) Iteration 1	2	3	4	5	Mean
Holdout Stratified	Linear	0,710	0,710	0,645	0,774	0,806	72,90%
Holdout Random	Linear	0,677	0,742	0,871	0,903	0,677	77,42%
K-Fold Stratified	Linear	0,677	0,677	0,677	0,677	0,677	67,74%
K-Fold Unstratified	Linear	0,677	0,677	0,677	0,677	0,677	67,74%
Holdout Stratified	Polynomial	0,710	0,613	0,806	0,774	0,806	74,19%
Holdout Random	Polynomial	0,742	0,645	0,710	0,774	0,774	72,90%
K-Fold Stratified	Polynomial	0,774	0,774	0,774	0,774	0,774	77,42%
K-Fold Unstratified	Polynomial	0,774	0,774	0,774	0,774	0,774	77,42%
Holdout Stratified	Radian	0,806	0,806	0,839	0,806	0,839	**81,94%**
Holdout Random	Radian	0,701	0,742	0,871	0,839	0,774	78,71%
K-Fold Stratified	Radian	0,774	0,774	0,774	0,774	0,774	77,42%
K-Fold Unstratified	Radian	0,774	0,774	0,774	0,774	0,774	77,42%

Based on the accuracy of each ethos, the highest accuracy is chosen to determine which method is better. Table 7 It is found in Table 7 that the analysis with the reduced data is better than the data with all variables. The SVM classification method is better than logistic regression one. SVM method with radial kernel has the highest accuracy of 81,94% than others. Here is the comparison table of accuracy (see Table 8).

Table 8. Comparison of Accuracy

Data	Method	Cross Validation	Accuracy
Logistic Regression	All Variables	Holdout Non-Stratified	72,7%
	Reduced Variables	Holdout Stratified	79,3%
SVM	All Variables	Holdout Stratified Radian	81,29%
	Reduced Variables	Holdout Stratified Polynomial	81,94%

Discussion

In this research, SVM gives better accuracy than logistic regression. The result of this analysis is in line with the findings of Narayan (2021) and Y. Huang et al. (2016) that the performance of SVM is better than logistic regression. Hidayat, Ruldeviyani, and Aditama (2022) also uses logistic regression and SVM as a classifier in the sentiment analysis of Twitter data and the result is SVM method classifier produces excellent results. This result analysis is same with Panesar, D'Souza, Yeh, and Miranda (2019) research. That research compares logistic regression and machine learning. Then the result shows that SVM is better than an artificial neural network, decision trees, and logistic regression.

In this analysis, SVM gives better results than logistic regression because the hepatitis dataset is complex data. It is in line with Huang et al. (2018) that SVM is good for complex data, and not only linear but also nonlinear dataset. But logistic regression is used for the linear dataset. The difference between the results of this research with the previous research is the variable selection carried out in this research. Moreover, the researcher also chooses the kernel that gives the highest accuracy value to the SVM method.

Conclusion

The classification with feature extraction is better than with the original data. SVM and logistic regression methods are used in this research to classify the hepatitis patient's dataset. The best accuracy average of the SVM method is 81.29%. The best accuracy average of logistic regression is 81.94. Hence, the researcher concludes that the classification using a Support Vector Machine obtains higher accuracy than Logistic regression classification. SVM's best accuracy with linear, polynomial, and radian kernels are 77.42%, 77.42%, and 81.94%. The accuracy using SVM with the radial kernel is higher than the linear and polynomial kernel. In conclusion, SVM holdout stratified method using the radian kernel is the best

classification method in this hepatitis patient's dataset having accuracy of 81.94%. The limitation in this research is if the researcher uses different data, the accuracy of the results may change. In addition, the SVM kernel method that provides the highest accuracy is, the other kernel not a polynomial.

References

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data.

Chen, N., Lu, W., Yang, J., & Li, G. (2004). Support Vector Machine in Chemistry. Support Vector Machine in Chemistry. https://doi.org/10.1142/5589.

Chow, J. H., & Chow, C. (2006). The Encyclopedia of Hepatitis and Other Liver Diseases. 372. https://books.google.com/books/about/The_Encyclopedia_of_Hepatitis_and_Other.html?id=HfPU99jIfboC.

Edgar, T. W., & Manz, D. O. (2017). Exploratory Study. Research Methods for Cyber Security, 95–130. https://doi.org/10.1016/B978-0-12-805349-2.00004-2.

Gail, M., Krickeberg, K., Samet, J. M., Tsiatis, A., & Wong, W. (2012). Logistic Regression: A Self-learning Text, Third Edition (Statistics in the Health Sciences). http://www.springer.com/series/2848.

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Cham, Switzerland: Springer International Publishing.

Hasanah, S., & Widjarnarko O. B. (2021). Perbandingan Metode Propensity Score Matching-Support Vector Machine dan Propensity Score Matching-Regresi Logistik Biner pada Kasus HIV/AIDS. *Jurnal Ilmiah Matematika dan Ilmu Pengetahuan Alam, 18*(1). https://doi.org/10.31851/sainmatika.v18i1.4925.

Hidayat, T. H. J., Ruldeviyani, Y., & Aditama, A. R. (2022). Sentiment Analysis of Twitter Data Related to Rinca Island development Using Doc2Vec and SVM and Logistic Regression as Classifier. Elsevier. https://www.sciencedirect.com/science/article/pii/S187705092102411X.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw/~cjlin.

Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics and Proteomics*, *15*(1), 41–51. International Institute of Anticancer Research. https://doi.org/10.21873/cgp.20063.

Huang, Y., Zhang, L., Lian, G., Zhan, R., Xu, R., Huang, Y., Mitra, B., Wu, J., & Luo, G. (2016). A Novel Mathematical Model to Predict Prognosis of Burnt Patients Based on Logistic Regression and Support Vector Machine. *Burns, 42*(2), 291-299. https://www.sciencedirect.com/science/article/pii/S0305417915002338.

Ikerionwu, C., Ugwuishiwu, C., Okpala, I., James, I., Okoronkwo, M., Nnadi, C., Orji, U., Ebem, D., & Ike, A. (2022). Application of Machine and Deep Learning

Algorithms in Optical Microscopic Detection of Plasmodium: A malaria diagnostic tool for the future. *Photodiagnosis and Photodynamic Therapy, 40*, 103198. https://doi.org/10.1016/J.PDPDT.2022.103198.

K C, S., Bhusal, A., Gautam, D., & Rupakhety, R. (2022). Earthquake Damage and Rehabilitation Intervention Prediction Using Machine Learning. *Engineering Failure Analysis, 144*, 106949. https://doi.org/10.1016/J.ENGFAILANAL.2022.106949.

Kemenkes. (2022). Hepatitis Can't Wait. http://P2p.Kemkes.Go.Id/Hepatitis-Cant-Wait/.

Kistenev, Y. v., Vrazhnov, D. A., Shnaider, E. E., & Zuhayri, H. (2022). Predictive Models for COVID-19 Detection Using Routine Blood Tests and Machine Learning. *Heliyon, 8*(10), e11185. https://doi.org/10.1016/J.HELIYON.2022.E11185.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine Learning Techniques and Data for Stock Market Forecasting: A literature Review. *Expert Systems with Applications, 197*, 116659. https://doi.org/10.1016/J.ESWA.2022.116659.

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., & Tischer, T. (2022). Machine Learning and Conventional Statistics: Making Sense of the Differences. *Knee Surgery, Sports Traumatology, Arthroscopy, 30*(3), 753–757. https://doi.org/10.1007/s00167-022-06896-6.

Martinez, A. J. (1996). Medical Microbiology. Medical Microbiology, 4th Edition, 1–9. https://www.ncbi.nlm.nih.gov/books/NBK7627/.

Narayan, Y. (2021). Direct Comparison of SVM and LR Classifier for SEMG Signal Classification Using TFD Features. *Materials Today: Proceedings, 45*, 3543-3546. https://www.sciencedirect.com/science/article/pii/S2214785320406972.

Novianti, F. A., & Purnami, S. W. (2012). Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi Fourina Ayu Novianti dan Santi Wulan Purnami. *Jurnal Sains dan Seni ITS, 1*(1), D147-D152. https://doi.org/10.12962/j23373520.v1i1.1937.

Nurlaily, D., Irhamah, Purnami, S. W., & Kuswanto, H. (2019). Support Vector Machine for Imbalanced Microarray Dataset Classification Using Ant Colony Optimization and Genetic Algorithm. *AIP Conference Proceedings, 2194*(1), 020076. AIP Publishing LLC. https://doi.org/10.1063/1.5139808.

Pal, M., & Mather, P. M. (2005). Support Vector Machines for Classification in Remote Sensing. *International Journal of Remote Sensing, 26*(5), 1007–1011. https://doi.org/10.1080/01431160512331314083.

Panesar, S. S., D'Souza, R. N., Yeh, F.-C., & Miranda, J. C. F. (2019). Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in A Small, Heterogeneous Glioma Database. *World neurosurgery: 10*(2), 100012. https://www.sciencedirect.com/science/article/pii/S2590139719300432.

Park, H. A. (2013). An Introduction to Logistic Regression: from Basic Concepts to Interpretation with Particular Attention to Nursing Domain. Journal of Korean

Academy of Nursing, 43(2), 154–164. https://doi.org/10.4040/jkan.2013.43.2.154.

Qomariyah, S., Iriawan, N., & Fithriasari, K. (2019). Topic Modeling Twitter Data Using Latent Dirichlet Allocation and Latent Semantic Analysis. *AIP Conference Proceedings, 2194*(1), 020093. https://doi.org/10.1063/1.5139825.

Raman, G. (2022). Identifying Extra-Large Pore Structures in Zeolites with A Machine Learning Approach and Its Deployment into Production. *Microporous and Mesoporous Materials*, *112362*. https://doi.org/10.1016/J.MICROMESO.2022.112362.

Samosir, R. O., Wilandari, Y., & Yasin, H. (2015). Perbandingan Metode Klasifikasi Regresi Logistik Biner dan Radial Basis Function Network pada Berat Bayi Lahir Rendah (Studi Kasus: Puskesmas Pamenang Kota Jambi). (Doctoral Dissertation, FSM Universitas Diponegoro). http://ejournal-s1.undip.ac.id/index.php/gaussian.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science 2*(3). Springer. https://doi.org/10.1007/s42979-021-00592-x.

Shihong, Y., Ping, L., & Peiyi, H. (2003). SVM Classification: Its Contents and Challenges. *Applied Mathematics-A Journal of Chinese Universities, 18*(3), 332–342. https://doi.org/10.1007/S11766-003-0059-5.

Utami, I. T. (2018). Perbandingan Kinerja Klasifikasi Support Vector Machine (SVM). Dan Regresi Logistik Biner Dalam Mengklasifikasikan Ketepatan Waktu Kelulusan Mahasiswa Fmipa Untad. *Jurnal Ilmiah Matematika dan Terapan, 15*(2), 256-267.

WHO. (2021). Global Progress Report on HIV, Viral Hepatitis and Sexually Transmitted Infections, 2021-Data slides.

WHO. (2022). World Hepatitis Day 2022. https://www.Who.Int/Indonesia/News/Campaign/World-Hepatitis-Day/2022.

Wilkinson, J., Mamas, M. A., & Kontopantelis, E. (2022). Logistic Regression Frequently Outperformed Propensity Score Methods, Especially for Large Datasets: A Simulation Study. *Journal of Clinical Epidemiology*, *152*, 176–184. https://doi.org/10.1016/J.JCLINEPI.2022.09.009.

Zuckerman, A. J., & Baron, S. (1996). Hepatitis Viruses - PubMed. Retrieved November 21, 2022, from https://pubmed.ncbi.nlm.nih.gov/21413272/.

This page is intentionally left blank