JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1424

CLASSIFICATION OF LIVER DISEASE BY APPLYING RANDOM FOREST ALGORITHM AND BACKWARD ELIMINATION

Irwan Herliawan[1*]; Muhammad Iqbal[2]; Windu Gata[3]; Achmad Rifai[4]; Jajang Jaya Purnama[5]

Computer Science[1,2,3,4]
STMIK Nusa Mandiri[1,2,3,4]
www.nusamandiri.ac.id[1,2,3,4]
14002397@nusamandiri.ac.id[1*]; iqbal.mdq@nusamandiri.ac.id[2]; windu@nusamandiri.ac.id[3]
achmad.acf@nusamandiri.ac.id[4]

Computer Technology[5]
Bina Sarana Informatics University[5]
www.bsi.ac.id[5]
jajangja2412@bsi.ac.id[5]

(*) Corresponding Author

Abstract— The liver can be attacked by diseases that cause the liver cannot function normally and even cause death. The causes of liver disease vary, mostly due to viruses transmitted through the fecal-oral, parenteral, sexual, perinatal, and so on. Also, there are some liver diseases whose exact cause is unknown. The most common causes of liver disease include birth defects, viral or bacterial infections, excessive alcohol consumption, drug addiction and others. Early detection of liver or liver cancer is very important to overcome the very high risk of death caused by liver or liver cancer. This study aims to help classify liver or liver cancer based on data from routine examination results of patients summarized in the Indian Liver Data Patient (ILDP) dataset. The method used in the classification process in this research is backward elimination modeling for testing optimization and Random Forest algorithm and split validation to validate the model. The results of this study yielded 76.00% and value of AUC 0.758 results. These results indicate that the results of this study are good enough to help classify breast cancer.

Keywords: Liver, Classification, Random Forest, Backward Elimination, Split Validation

Abstrak*—Hati dapat terserang penyakit yang menyebabkan hati tidak dapat berfungsi seperti biasa dan bahkan menyebabkan kematian. Penyebab penyakit hati bervariasi, sebagian besar disebabkan oleh virus yang ditularkan melalui fecal-oral, parenteral, seksual, perinatal, dan sebagainya. Juga, ada beberapa penyakit hati yang penyebab pastinya tidak diketahui. Penyebab paling umum penyakit hati termasuk cacat lahir, infeksi virus atau bakteri, konsumsi alkohol yang berlebihan, kecanduan obat-obatan dan lainnya. Deteksi dini terhadap kanker khususnya kanker hati atau liver sangat penting dilakukan untuk menanggulangi resiko kematian sangat tinggi disebabkan oleh kanker hati atau liver. Penelitian ini bertujuan untuk membantu melakukan klasifikasi kanker hati atau liver berdasarkan data hasil pemeriksaan rutin pasien yang dirangkum dalam dataset Indian Liver Data Patient (ILDP). Metode yang digunakan dalam proses klasifikasi pada penelitian ini yaitu permodelan backward elimination untuk optimasi akurasi serta algoritma Random Forest dan validasi split validation untuk memvalidasi permodelan. Hasil dari penelitian ini menunjukkan hasil akurasi sebesar 76.00% dan nilai AUC yaitu 0.758. Hasil tersebut menunjukkan bahwa hasil penelitian ini cukup baik untuk membantu mengklasifikan kanker hati atau liver.*

Kata Kunci: *Liver, Klasifikasi, Random Forest, Backward Elimination, Split Validation*

INTRODUCTION

Health is one of the most important things in human life, this is the basis that many scientific discoveries are found in the form of medicines, medical devices, or discoveries in the health field. In the world of health one of which is discussed is a disease, since a long time ago a lot of diseases emerged whether it came from viruses, bacteria, parasites, cancer cells, and others [1]

In the human body, there are several important organs, each of which has a very

beneficial function, one of which is the heart. The liver is the largest organ in the body and has many functions for the health of the human body. The liver is the body functions to produce protein, cleanses the blood, helps metabolize protein, stores nutrients, helps the digestion process of food, produces cholesterol, produces growth hormone in children, and destroys toxins in the body so that the digestion process in the body goes perfectly. Meanwhile, according to [2] the liver is an organ in our body that has the largest size and has a very important function. The liver converts toxic substances into nutrients and then the body uses it to control hormone levels in the body. Also, the liver produces hormones and protein, controls blood sugar and helps control blood clotting, helps the formation and secretion of bile, urea synthesis, acts as cholesterol and fat metabolism, and the main function of the liver as detoxification of poisons or neutralizing poisons[3].

The liver can be attacked by diseases that cause the liver is unable to function as usual and even cause death[4]. The cause of liver disease varies, mostly caused by viruses that are transmitted in fecal-oral, parenteral, sexual, perinatal, and so on[5]. Also, there are several liver diseases whose exact cause is unknown[5]. Therefore, we must recognize the types of liver disease so that we can avoid and be able to maintain the health of our important organs. Types of liver disease include hepatitis, liver, cirrhosis, liver cancer, jaundice (jaundice), liver failure, Cholangitis, Leptospirosis, and Liver Abscess[5]. Acute liver diseases will greatly affect liver functions, these diseases can be known from clinical and physical symptoms that arise in the patient, clinical symptoms can be known from what is felt by the patient, while physical symptoms can be known from the state of the patient's body[6]. Chronic liver disease and cirrhosis are some of the liver diseases with high morbidity and mortality. Worldwide cirrhosis ranks seventh cause of death[4]. The liver is a liver disease that has been around for a long time and is quite common in society[7]. The most common causes of liver disease include birth defects, viral or bacterial infections, excessive alcohol consumption, drug addiction (especially in blood vessels), adverse reactions to various drugs (such as analgesics, anti-inflammatory drugs inflammation, some antibiotics, antifungal drugs and immune suppressants, and poor lifestyle[8].

During this time many people are unaware and it is difficult to find out someone has liver disease [9]. Almost all people experience delays in treatment because they only get checked when the liver disease is severe. Lack of public knowledge about liver disease because people have difficulty in recognizing the types of liver disease and the symptoms of liver disease which is quite a lot and there are also similarities in the symptoms of several types of liver disease[10]. Based on data from the World Health Organization (WHO), chronic hepatitis B virus is estimated to attack 350 million people in the world, especially Southeast Asia and Africa, and causes the death of 1.2 million people per year. Of that number, 15-25% who are chronically infected die from complications from cirrhosis and liver cancer [11].

the liver is cancer that occurs in liver cells (liver). The cancerous liver will enlarge like a ball in the right upper abdomen, below the diaphragm and above the abdomen. Liver cancer (Hepatocellular Carcinoma / HCC) is cancer that is often found throughout the world and is among the top 5 categories of cancer-causing death [12].

Liver disease (disorders of the liver) can originate from the most common causative factors including liver disorders that have existed since birth, the presence of viral or bacterial infections, excessive alcohol consumption, drug addiction (especially in blood vessels), adverse reactions to various drugs (such as analgesics, anti-inflammatory drugs

Of the twenty classification methods, Bayes Net, Naïve Bayes, Classification through Regression, Logistic Regression, and Random Forest is the best classification methods. For the mixed attribute dataset Naïve Bayes, Bayes Net, and Random Forest are the best classification methods. For numerical attribute dataset Regression Classification, NBTree and Multiclass Classifiers are the best methods. For the NB-Tree dataset's categorical attributes, Classification through Regression, and Bayes Net methods are the best[13].

Backward elimination is one method that has a function for optimizing the performance of a model by the way the election system works [14]. This method is usually used to improve the accuracy of the classification process, by adding this method in the data processing process is expected to obtain maximum accuracy for classifying data.

In this study, the data to be processed is Indian Liver Patient Dataset (ILPD Dataset) data obtained from the UCI machine learning repository. According to [13] The classification algorithm chosen for the classification of several liver patient datasets is the Random Forest classifier, the algorithm is evaluated based on four criteria: Accuracy, Precision, Sensitivity, and Specificity. Therefore, the data will be processed using data mining classification techniques. Random Forest algorithm. The purpose of this study is to apply the Random Forest algorithm to

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1424

classify Indian Liver Patient Datasets to produce the best data mining model.

MATERIALS AND METHODS

In this data mining research, the authors use a standard methodology called the Cross-Industry Center Process for Data Mining (CRISP-DM). The first stage of this research method is trying to determine the objectives of the research project in the formulation and definition of data mining problems. The main goal of what is to be achieved is to know the classification of liver disease. In the second stage, the writer collects, identifies, and understands the data we have and the data must also be verified for truth and reliability. The dataset used by the author is obtained from https://archive.ics.uci.edu or the UCI Repository namely Indian Liver Data Patient (ILDP). In the third stage, the data obtained from the UCI Repository the authors carry out activities such as cleaning data, reformatting the data, and so on so that the purpose is to prepare the data to be consistent following the format needed. The amount of data studied was 583 patients. On this research occasion, researchers conducted data processing using the RapidMiner software. The fourth step, the process of selecting an algorithm with optimal value parameters. It aims at computational representation of observations which are the result of searching for patterns contained in data. In this study, the chosen algorithm model is the Random Forest and Backward Elimination algorithm to improve accuracy. The fifth stage is the final process of this research is the evaluation process, where the algorithm classification process is carried out by testing the accuracy of the Random Forest and Backward Elimination algorithm by looking at the results of accuracy and AUC.

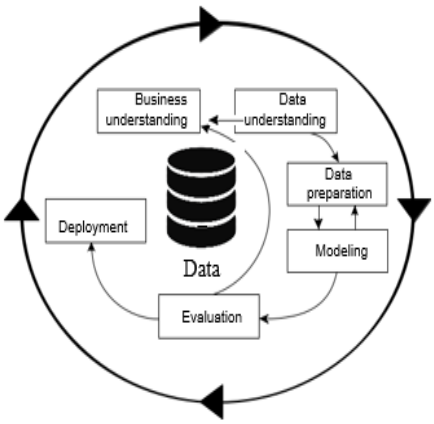

Figure 1. CRISP-DM Model

1. Business Understanding

This stage focuses on understanding the objectives of a project that are seen from a business perspective. Then change the results of Data Mining as a solution to the initial problem that has been determined

2. Data Understanding

The understanding data stage is the stage of understanding the data starting with initial data collection and starting with activities to familiarize yourself with the data, to identify data quality problems, to find the first insight into the data, or to detect interesting subsets to form hypotheses for hidden information.

3. Data Preparation

The data preparation stage is the stage for data preparation covering all activities needed to build the final dataset from the initial raw data. Data preparation tasks tend to be done repeatedly and in no specified order.

4. Modeling

The modeling stage can choose the modeling techniques provided to be applied, and also calibrate a parameter to be optimal. There is a DM problem using the same technique. Also where some techniques have a special requirement for form data

5. Evaluation

The evaluation stage is the stage to find out whether, from the perspective of the data analyst, a model that seems to be of high quality will be built at this stage of the project. In its application, the model used with evaluation before implementation can be applied to implementation. This evaluation phase will be made a decision, regarding the use of DM results..

6. Deployment

The fifth stage is the last stage but usually is not the end of a project. This model has the aim of increasing data knowledge. Arrangements must be made, and the presentation of knowledge must be arranged as well, ultimately the customer can use.

RESULTS AND DISCUSSION

1. Business Understanding

By utilizing existing data sources, it can be analyzed and predicted using data mining techniques whose business objective is to make a classification of liver disease based on optimization of Backward Elimination to predict liver disease.

2. Data Understanding

The data source used is student performance which is a dataset from the UCI Machine Learning Repositor. Collection of data collected from northeast Andhra Pradesh, India. The data set used has a record including 416 patients with liver disease, and 167 patients without liver disease. In

detail by sex, 441 male patients, and 142 female patients. Patients over the age of 89 are registered as "90" approved

Table 1. Atribut Dataset

No	Name of attribute	Description	Reference Value
1	Age	Age of the patient	Children (x<20), Young (20≤x<50), Adult (x≥50)
2	Gender	Gender of the patient	Male, Female
3	TB	Total Bilirubin	Normal (x≤1), Abnormal (x>1)
4	DB	Direct Bilirubin	Normal (x≤0,2), Abnormal (x>0,2)
5	Alkphos	Alkaline Phosphatase	low (x<30), Normal (30 ≤x≤120), high (x>120)
6	Sgpt	Alamine Aminotransferase	Normal (x<47), Abnormal (x≥47)
7	Sgot	Aspartate Aminotransferase	Normal (x<37), Abnormal (x≥37)
8	TP	Total Proteins	low (x<6), Normal (6 ≤x≤8), high (x>8)
9	ALB	Albumin	low (x<3,4), Normal (3,4 ≤x≤4,8), Tinggi (x>4,8)
10	A/G Ratio	Albumin and Globulin Ratio	Normal (x>1), Abnormal (x≤1)
11	Label	Selector field used to split the data into two sets	1 (Liver) 2 (No)

The number of attributes in the Indian Liver Patient dataset is 11. The number of classes in the dataset is 1, the Selector field used to split the data into two sets (Labels).

3. Data Preparation

The preparation of the data aims that the data source can be applied in the modeling phase then it needs to be transformed. The model to be used is Random Forest with Backward Elimination optimization.

4. Modeling

Modeling is done by developing a model that has been prepared with the Random Forest (RF) algorithm and applying Backward Elimination to improve the accuracy of liver disease prediction models in the ILDP dataset.

This modeling stage is done after the data is ready to use, wherein the previous process the data preparation is done. This modeling is done by applying Random Forest and backward elimination. The process of finding the results of this evaluation implements by getting good accuracy used by processing done with the Random Forest model and Backward Elimination to improve accuracy.

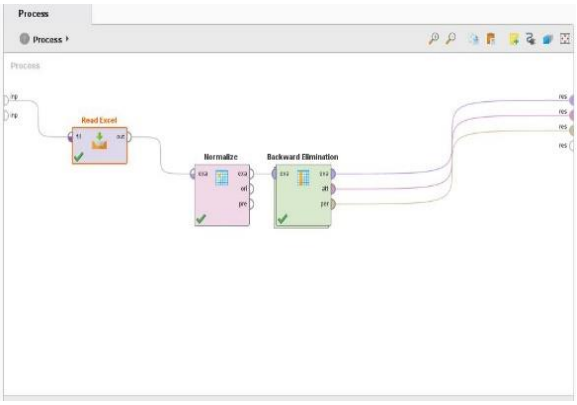

Figure 2. Testing the dataset using the Random Forest and Backward Elimination Algorithm

Based on figure 2 above explains the testing dataset process using Random Forest and Backward Elimination Algorithm. is expected to provide good results from the application of the model.

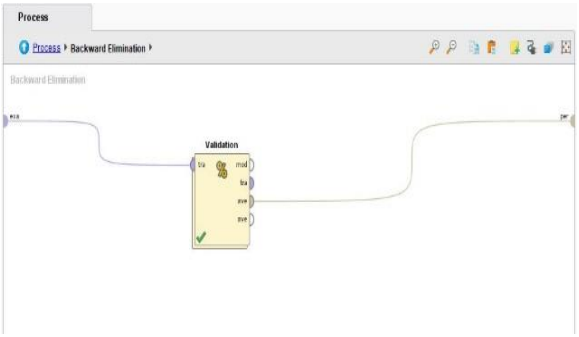

Figure 3. Testing the dataset using the Random Forest and Backward Elimination Algorithm

Figure 3 explains the calculation model using validation parameters with 10 folds. It is expected

VOL. 6. NO. 1 AUGUST 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v6i1.1424

to give good results from the application of the Random Forest and Backward Elimination algorithm models.

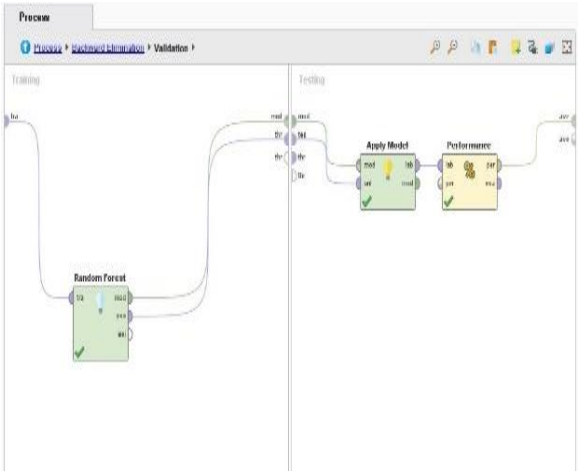

Figure 4. Testing the dataset using the Random Forest and Backward Elimination Algorithm

Figure 4 above explains the model using the Random Forest and Backward Elimination algorithm. The application of the Random Forest and Backward Elimination model is intended to improve its accuracy. The application of the model is illustrated in the following figure.

5. Evaluation

The test results using the Random Forest and Backward Elimination algorithm models, obtained confusion matrix accuracy and AUC (Area Under Curve) values as follows.

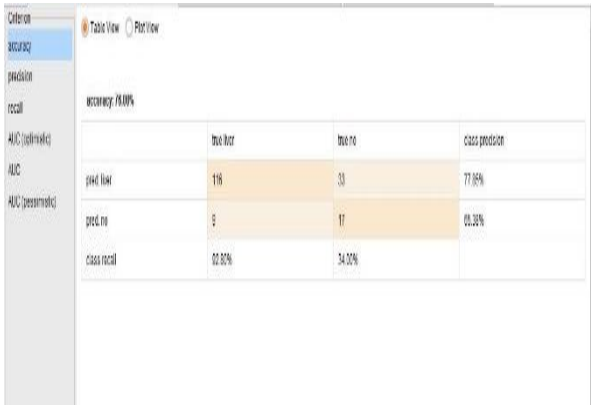

Figure 5. Testing the dataset using the Random

Forest and Backward Elimination algorithm

In figure 5. can be seen the accuracy value obtained in the modeling with the split validation method of 76.00%.

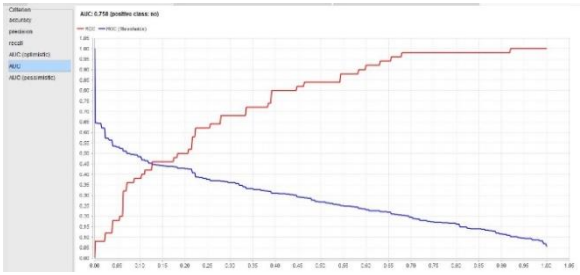

Figure 6. The results of testing the dataset using the Random Forest and Backward Elimination Algorithms

Figure 6. shows the AUC value of 0.758 and the value shows that the classification results are included in the fair classification category.

Table 2. AUC Value and Description

AUC Value	Classification
0.90 - 1.00	*excellent classification*
0.80 - 0.90	*good classification*
0.70 - 0.80	*fair classification*
0.60 - 0.70	*poor classification*
0.50 - 0.60	*failure*

Source: Annisa[13]

Based on Table 2 above, it is explained that the AUC classification value from 0.9 to 1.00 belongs to the very good classification category, the AUC classification value from 0.80 to 0.90 is included in the good classification category, the AUC classification value from 0, 60-0.70 including the poor classification category, the AUC classification value of 0.50-0.60 including the failure category.

In this study, the AUC classification obtained was 0.758, then included in the classification category of Fair Classification.

6. Deployment

From the results of modeling using the Random Forest and Backward Elimination algorithms, research results have been obtained that are good enough to predict liver disease.

CONCLUSION

Based on the results of the study, the classification of liver disease using the Random Forest and Backward Elimination algorithms on the Indian Liver Data Patient (ILDP) dataset obtained from the UCI dataset. The resulting classification model values to get the accuracy value and AUC of the algorithm used to obtain an accuracy of 76.00% with an AUC value of 0.758.

Based on these results the results are good enough to improve the prediction results in the ILDP dataset. Also, it can be concluded that processing different data such as the categorization of attributes, and differences in taking the number of samples can provide different accuracy even though using the same data and algorithm.

REFERENCE

[1] P. Handayani, E. Nurlelah, M. Raharjo, and P. M. Ramdani, "Prediksi Penyakit Liver Dengan Menggunakan Metode," *Pros. TAU SNAR-TEK Semin. Nas. Rekayasa dan Teknol.*, vol. 4, no. 1, pp. 75–80, 2019.

[2] P. Widodo, "Rule-Based Classifier untuk Mendeteksi Penyakit Liver," *Bianglala Inform.*, vol. II, no. 1, pp. 71–80, 2014.

[3] I. A. Medha and D. P. Hapsari, "Implementing Fuzzy Decision Tree for Predicting Liver Disease of Ilpd (Indian Liver Patient Dataset)," vol. 5, no. 2, 2019.

[4] D. G. Kesuma, "a Women 51 Years With Decompensated Liver Cirrhosis With Gastritis Chronic and Kidney Chronic Disease," vol. 3, no. September, pp. 151–159, 2014.

[5] R. G. Rafsanjani, N. Hidayat, and R. K. Dewi, "Diagnosis Penyakit Hati Menggunakan Metode Naive Bayes Dan Certainty Factor," vol. 2, no. 11, pp. 4478–4482, 2018.

[6] N. D. Prayoga, N. Hidayat, and R. K. Dewi, "Sistem Diagnosis Penyakit Hati Menggunakan Metode Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2666–2671, 2018.

[7] A. P. Ayudhitama *et al.*, "ANALISA 4 ALGORITMA DALAM KLASIFIKASI PENYAKIT LIVER," pp. 1–9.

[8] E. Nurlelah and M. S. Mardiyanto, "Pemilihan Atribut Pada Algoritma C4.5 Menggunakan Particle Swarm Optimization Untuk Meningkatkan Akurasi Prediksi Diagnosis Penyakit Liver," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 195–202, 2019.

[9] E. Pusporani and S. Qomariyah, "Klasifikasi Pasien Penderita Penyakit Liver dengan Pendekatan Machine Learning," vol. 2, no. March, 2019.

[10] A. R. Safutra and D. W. Prabowo, "Diagnosis Penyakit Kanker Payudara Menggunakan Metode Naive Bayes Berbasis Desktop," *J. Penelit. Dosen FIKOM*, vol. 6, no. 1, pp. 1–6, 2016.

[11] E. Rahmawati, "Analisa Komparasi Algoritma Naive Bayes Dan C4.5 Untuk Prediksi Penyakit Liver," *None*, vol. 12, no. 2, pp. 27–37, 2015.

[12] E. V. I. Noviani, K. A. Sidarto, and Y. S. Putra, "PENGELOMPOKAN PASIEN KANKER LIVER BERDASARKAN DATA EKSPRESI GEN DENGAN," pp. 1053–1062, 2012.

[13] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019.

[14] I. C. R. Drajana, "Metode Support Vector Machine Dan Forward Selection Prediksi Pembayaran Pembelian Bahan Baku Kopra," *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 116–123, 2017.

Accredited Rank 3 (Sinta 3) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti
No. 28/E/KPT/2019, September 26, 2019. Published by PPPM STMIK Nusa Mandiri