



Implementation Of C5.0 Classification And Support Vector Machine Algorithm With Correlation-Based Feature Selection In Application Liver Disease

Andi Nur Rachman^a, Cecep Muhamad Sidiq^{a,*}, Muhammad Hanif Insani^b

^a Department of Information System, Siliwangi Universty, Tasikmalaya, Indonesia

^bDepartment of Informatics, Siliwangi Universty, Tasikmalaya, Indonesia

Corresponding author: cecepmuhamad@unsil.ac.id

Abstract— Liver disease is a general term that refers to a number of disorders or problems that affect the liver. The liver is an important organ in the human body and has many diverse functions, including food processing, protein production, toxin removal, and energy storage. Therefore, when the liver experiences disorders or disease, it can have a serious impact on the overall health and function of the body. Liver disease is a significant global health problem. Early detection as well as classification of liver disease can provide valuable guidance for effective treatment. Based on the problems above, the aim of this research is to create a liver disease classification model using C5.0 and Support Vector Machine with Radial Basis Function (RBF) and Sigmoid kernels. With data obtained from the liver disease dataset. The two methods will be compared and we will find out which one produces the best results. The method used is also optimized with CFS (Correlation Based Feature Selection) feature selection. The results of the classification process, namely the C5.0 Model and Support Vector Machine (RBF) with CFS have a similar accuracy of 76%, while the Support Vector Machine (Sigmoid) has an accuracy of 70%, without feature selection the C5.0 algorithm has an accuracy of 66% , Support Vector Machine between RBF and sigmoid kernels has an accuracy of 69% and 55%.

Keywords— Liver; C5.0; Support Vector Machine; Radial Basis Function (RBF); Sigmoid; CFS (Correlation Based Future Selection)

Manuscript received 1 Mei. 2024; revised 9 Mei 2024; accepted 21 Mei 2024. Date of publication 21 Mei 2024. International Journal of Applied Information and Informatic is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.



I. INTRODUCTION

The liver is one of the important organs in the human body. The main function of the liver is to destroy toxins that enter the body so that it can help the digestive process, help blood clot, store nutrients, produce protein, bile, hormones and energy[1]. This liver function is very influential in fighting disease viruses and keeping the blood clean, if liver function is disturbed then your body's health will also be affected. According to USC (University of Southern California) cases of liver disease increased again by around 30 percent in hospital admissions during the recent pandemic. (Costello et al., 2021)[2]. According to WHO (World Health Organization) data, in 2013, liver disease suffered by Indonesian citizens reached 28 million and continued to increase until 2019[3]. There are also diseases that attack the liver, one of which is Hepatitis, especially Hepatitis B, from WHO (World Health Organization) data.) which has attacked 350 million people, especially Southeast Asia and Africa,

causing the deaths of around 1.2 million people per year[3]. One of the things that causes the increase in liver disease is because people are not aware or are not aware that they have liver disease and are not rushed to the hospital, causing death due to delays in treatment[4].

Liver failure, also known as hepatic failure, is a condition in which the liver no longer functions properly. This includes its ability to process toxins and produce essential substances for the body. Liver failure can manifest as either acute or chronic. Acute liver failure occurs suddenly and can lead to serious liver damage, posing a life-threatening risk[5]. The causes of acute liver failure are quite diverse and may include hepatitis, drug toxicity, viral infections, or complications arising from other illnesses[6]. Conversely, chronic liver failure develops gradually, often progressing over many years. Chronic liver failure is a condition that tends to receive less awareness from the general public. The causes of chronic liver failure are also varied and may involve excessive alcohol

consumption, non-alcoholic fatty liver disease (NAFLD), and chronic hepatitis[7].

Seeing several things from the background above, this research was carried out by implementing a classification to help diagnose whether a patient was diagnosed positively with liver disease or not by examining blood counts, increased liver enzymes, increased GGT (gamma glutamyl transferase) and ALP (alkaline phosphatase), increased bilirubin, and low albumin levels. In the classification process for initial treatment of liver disease in this study, the C5.0 Algorithm classification method was used. The final result of the C5.0 algorithm is a decision tree that can be used to predict target values or classes from new data[8]. After carrying out the classification process using this method, it is continued with the accuracy method whose value is used to determine the level of closeness between the actual value and the predicted value[9]. Followed by the precision method to determine the accuracy of the answer from the system according to what is requested by the user. After obtaining the evaluation results, it measures whether the algorithm and feature selection used are effective in obtaining high accuracy and precision values.

II. MATERIALS AND METHOD

This classification research method consists of several stages starting from the literature study stage followed by data collection, data preprocessing such as searching for missing values and performing feature selection Correlatio-based feature selection, Data split, Implementation of the C5.0 algorithm classification and Support Vector Machine including design system and evaluation of results or conclusions.

Research related to CFS[10]–[13] Which basically explains The dataset used was feature selected using the correlation-based method where the correlation test was carried out by calculating and comparing the level of correlation between features and their classes as well as features with other features. Correlation-based feature selection (CFS) is a feature selection method that uses the correlation between features and target classes to determine which features should be retained in a model. CFS works by calculating the correlation coefficient between each feature and class, and selecting the subset of features with the highest correlation[10].

The dataset used is divided into training data and testing data. To divide training data and testing data, four percentages are taken, 80%:20, 90%:10%, 70%:30%, 60%:40% to find the highest accuracy value to be used in future applications.

Research related to C5.0 [8], [9], [14], [15] and related to Support Vektor Machine [13], [16], [17] that use algorithms for classification after the data has been selected and divided, the data is then processed using the C5.0 algorithm and Support Vector Machine accompanied by feature selection. For research on support vector machines using the RBF Kernel and Sigmoid Kernel, where the related research is [18]–[21], The kernel is the choice used in many cases because of its ability to handle complex non-linear fragmentation.

Model performance measurement is carried out using a confusion matrix and 10-fold cross validation. Confusion Matrix helps to understand how well a classification model

predicts correctly, while K-Fold Cross-Validation helps measure the overall performance of the model by avoiding bias caused by random separation of test and training data[22], [23].

The method used for application development in this research is using one of the SDLC methods, namely Waterfall. System development with Waterfall is recommended to be carried out systematically and sequentially[24]. The Waterfall method is considered the most appropriate in this research because it has a stable and clear system integrity, and allows only small changes to be expected during the development cycle[25].

III. RESULT AND DISCUSSION

A. Resource Data

The data source for this research is secondary data which comes from one of the trusted websites in providing and analyzing datasets, namely Kaggle.com. By comparison, the number of positive liver data was 414 patients and the number of negative liver data was 169 patients.

TABLE I
LIVER DATA PATIENT

No	feature	Data type
1	Age	Numeric
2	Gender	Category
3	Total_Bilirubin	Numeric
4	Direct_Bilirubin	Numeric
5	Alkaline_Phosphotase	Numeric
6	Alamine_Aminotransferase	Numeric
7	Aspartate_Aminotransferase	Numeric
8	Total_Protiens	Numeric
9	Albumin	Numeric
10	Albumin_and_Globulin_Ratio	Numeric
11	Dataset (<i>Class</i>)	Numeric

).

B. Preprocessing

The first preprocessing stage is looking for missing values, that is, searching for empty data (null). Four null data were found in this research data, and null values were searched for by calculating the average value of the attributes that had null values[26].

0	Age	583 non-null	int64
1	Gender	583 non-null	object
2	Total_Bilirubin	583 non-null	float64
3	Direct_Bilirubin	583 non-null	float64
4	Alkaline_Phosphotase	583 non-null	int64
5	Alamine_Aminotransferase	583 non-null	int64
6	Aspartate_Aminotransferase	583 non-null	int64
7	Total_Protiens	583 non-null	float64
8	Albumin	583 non-null	float64
9	Albumin_and_Globulin_Ratio	579 non-null	float64
10	Dataset	583 non-null	int64

Fig. 1 Missing Value Data

In the image there is non-null data that does not match (Missing Value) in the Albumin and Globulin Ratio attributes,

so the solution is applied with Mean/Median Imputation, namely replacing the missing values with the average or median value from the same column. This method is suitable for numerical data such as the database used in this research[27].

Features were selected using the CFS (feature-based correlation) method. For this study, 3 features with the highest correlation were taken. Feature selection is carried out with the help of the WEKA application. The results of the feature correlation with its class are as follows:

```
Ranked attributes:
 0.1623 10 Albumin_and_Globulin_Ratio
 0.1614 9 Albumin
 0.035 8 Total_Protiens
-0.0824 2 Gender
-0.1374 1 Age
-0.1519 7 Aspartate_Aminotransferase
-0.1634 6 Alamine_Aminotransferase
-0.1849 5 Alkaline_Phosphotase
-0.2202 3 Total_Bilirubin
-0.246 4 Direct_Bilirubin
```

Fig. 2 CFS Results On Weka

From the experiments and table above, the correlation for each attribute is attached. Where the highest results are in the A&G Ratio, Albumin and Total Protein attributes. These 3 attributes will be used in implementing the algorithm.

C. Implementation C5.0

The algorithm implementation in this research uses the C5.0 algorithm and Support Vector Machine (SVM). The Support Vector Machine uses two kernels, namely RBF and sigmoid.

Based on the results of the CFS feature selection, the attributes used are Total Protein, Albumin, Albumin and Globulin Ratio, and Age. For manual calculations of the C5.0 Algorithm, a Gain ratio is required which will be the root node of the tree. But to make it shorter, author can use help from python library.

This stage seeks the best accuracy and precision results from the C5.0 algorithm with CFS feature selection and without CFS feature selection. The following are the results of the evaluation of the C5.0 method using four Data split trials using the Python programming language:

TABLE II
C5.0 DATA SPLIT EXPERIMENT RESULT

Data split	Feature Selection	Accuracy	Precision
9:1	CFS	0.74576	0.44444
	No CFS	0.75271	0.50000
8:2	CFS	0.76068	0.54846
	No CFS	0.66667	0.35294
7:3	CFS	0.77143	0.45454
	No CFS	0.62857	0.32500
6:4	CFS	0.74359	0.43750
	No CFS	0.72650	0.31256

From the table above, it can be seen that the average results of Data split experiments with feature selection have

increased, where the highest results are in the 8:2 Data split experiment with feature selection, achieving an accuracy of 0.76068 and also having the highest precision with a value of 0.54846. Meanwhile, the highest result in the experiment without feature selection was the 9:1 split with a result of 0.75271 with a precision value of 0.4444. From the experiment above, it can be seen that CFS feature selection has a quite good result on the C5.0 algorithm

D. Support Vector Machine

In this ILPD Support Vector Machine classification research, two experimental kernels were used, namely the Radial Basis Function (RBF) and Sigmoid kernels to compare the best performance. In function analysis with RBF and sigmoid kernels, Cost (C) and Gamma (γ) parameters are needed. Then, to find the best parameters, a trial and error process was carried out by comparing the accuracy values of several parameters C and γ [17].

TABLE III
TRIAL AND ERROR TESTING DATASET KERNEL RBF

Parameter	Accuracy				
	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
C=1	0.74359	0.74359	0.73504	0.73504	0.74359
C=3	0.73504	0.74359	0.75890	0.71795	0.70940
C=5	0.73504	0.75890	0.74359	0.70940	0.68376
C=10	0.73504	0.56410	0.75214	0.70940	0.64957

Dari tabel pengujian uji coba Parameter di atas, dapat dilihat bahwa hasil akurasi tertinggi terjadi pada C = 3 dan $\gamma = 3$, serta pada C = 5 dan $\gamma = 2$. Oleh karena itu, C dan gamma yang dipilih untuk pengujian adalah C = 3 dan $\gamma = 3$, dengan nilai akurasi sebesar 0.75890.

TABLE IV
TRIAL AND ERROR TESTING DATASET KERNEL SIGMOID

Parameter	Accuracy				
	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
C=1	0.67521	0.57265	0.58120	0.57265	0.70085
C=3	0.68376	0.56410	0.58120	0.58120	0.70085
C=5	0.67521	0.56410	0.58120	0.57265	0.70085
C=10	0.68376	0.56410	0.58120	0.57265	0.70085

From the Trial and Error testing table for the Sigmoid kernel, the highest accuracy results are found in the gamma parameter 5. However, the highest results from the Sigmoid kernel are still lower than the RBF kernel, for further experiments it is taken from one of them, namely C=1 and $\gamma=5$ with testing value 0.70085. After C and gamma have been determined, then the selected C and gamma are searched for the best results from the four data splits

TABLE V
SVM DATA SPLIT EXPERIMENT RESULT

Data split	Kernel	Accuracy	Precision
60:40	RBF	0.70085	0.37037
	Sigmoid	0.58547	0.21667
70:30	RBF	0.70857	0.40000
	Sigmoid	0.60000	0.27451
80:20	RBF	0.76068	0.75000
	Sigmoid	0.70085	0.41379
90:10	RBF	0.76271	0.65000
	Sigmoid	0.55932	0.16667

From the table above it is known that the average results of Data split experiments with the RBF kernel have higher accuracy values, where the highest results are in the 9:1 and 8:2 Data split experiments with the Radial Basis Function kernel which have good results with a difference in value. which is not much different, namely between 0.76271 and 0.76068. But it has quite different precision, namely between 0.75000 and 0.65000, so it could be said that the 80:20 data split is better here.

Meanwhile, the highest result in the experiment with the Sigmoid kernel was the 9:1 data split with a result of 0.75271 with a precision value of 0.4444. From the experiments above, it can be seen that the RBF Kernel processing has better results compared to the Sigmoid Kernel.

E. Confusion Matrix

Confusion matrix is necessary after evaluating model performance as it provides deeper insight into how the model behaves in various aspects of its classification. This helps make more informed decisions in optimizing or improving models, especially in situations where accurate classification is critical[28]. To get the confusion matrix results the author uses Python with the help of several libraries such as Scikit-Learn

```

Program Jurnal
from sklearn.metrics import confusion_matrix
# Evaluate the model on the test set
y_pred = model.predict(x_test)

# calculate confusion matrix
cm = confusion_matrix(y_test, y_pred)

print("Confusion Matrix:")

```

This is the result of the confusion matrix from the model used in this research

Algorithm	Feature Selection	Accuracy	Precision	Recall	F-1 Score
C5.0	CFS	0,76068	0,93181	0,78846	0,854165
	No CFS	0,66666	0,75	0,79518	0,771929
SVM (RBF)	CFS	0,76068	0,98852	0,76106	0,859998
	No CFS	0,6923	0,82758	0,77419	0,800022
SVM (Sigmoid)	CFS	0,70085	0,80459	0,78545	0,799998
	No CFS	0,55555	0,62068	0,739726	0,674994

Fig. 3 Confusion Matrix Result

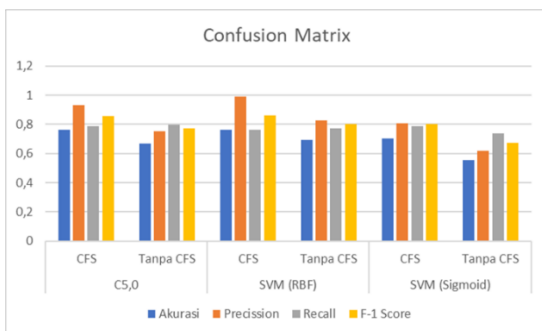


Fig. 4 Confusion Matrix Result Diagram

The figure 3 is a comparison of several model experiments with and without CFS feature selection. Models that use CFS feature selection have higher averages in their evaluations, including Accuracy, Precision, Recall and F-1 Score. Where the C5.0 and SVM (RBF) models have an average accuracy

that is not much different at 76%, and also have an average precision and recall (F-1 Score) that is not much different at 85%. But because This kind of detection system prioritizes accuracy and recall, the C5.0 model is better in classifying liver disease[16].

F. Time Consumption

Consumption time is the time required to build a model for an experiment[29]. Consumption time is calculated in units of seconds. To measure consumption time, the execution time of the training part of the model was measured using `time.time()` in python.

Program listing records the processing time of the algorithm

```

Program Jurnal
# Start measuring time
start_time = time.time()

# Train the model
.....

# stop measuring time
end_time = time.time()

# calculate and print the elapsed time
elapsed_time = end_time - start_time
print(f'Training time: {elapsed_time:.5f} seconds')

```

The initial time is measured before training begins, and the final time is measured after training is completed. By calculating the time difference you can estimate the time required to train the model for each type of model. The following is a comparison of consumption time between tests of each algorithm.

TABLE VI
TIME CONSUMPTION

Model Pengujian	Seleksi Fitur	Waktu Konsumsi
C5.0	CFS	0.00337
	No CFS	0.00448
SVM (RBF)	CFS	0.00939
	No CFS	0.05723
SVM (Sigmoid)	CFS	0.00942
	No CFS	0.01878

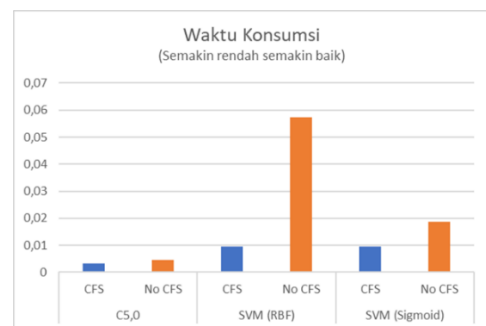


Fig. 5 Time Consumption Diagram

Based on the table VI and after conducting three tests with an 80% data split, it is evident that the C5.0 algorithm boasts the shortest processing time, clocking in at just 0.003 seconds. In contrast, SVM with RBF and Sigmoid Kernels consistently falls within a narrow time range, with processing times ranging from 0.009 to 0.01 seconds after several iterations. This indicates that C5.0 generally excels in swiftly training models on small to medium-sized datasets, but it may

experience a slowdown when dealing with very large datasets[30].

G. Application

The system in this research was developed using the Python programming language and utilized several libraries such as PyQt5 and sklearn within the PyCharm application.

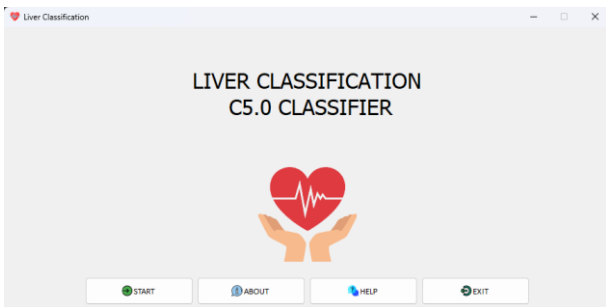


Fig. 6 Home Page

On the menu screen, there are four button options. First, the "START" button is used to initiate the data input process. Second, the "ABOUT" button provides information about the researcher and details about the classification application. The third button is "HELP," which opens a user guide explaining how to use the classification application. Lastly, there is the "EXIT" button used to exit the application.

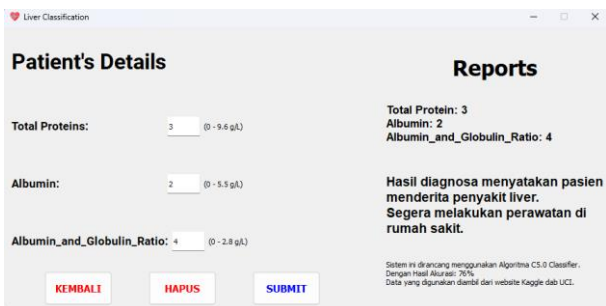


Fig. 7 Positive Classification Results

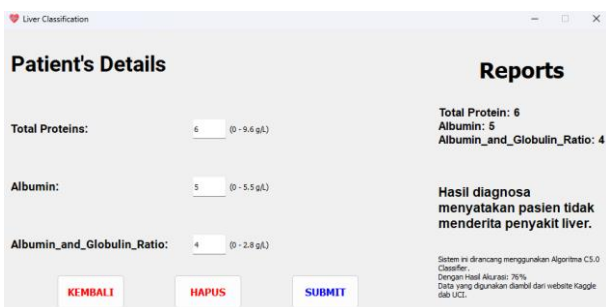


Fig. 8 Negative Classification Results

The screen shown in the image is used by the user to input medical information. The form includes the selected feature attributes, namely Total Protein, Albumin, and the Albumin-to-Globulin Ratio. In the images, there are diagnosis result pages displaying the outcome based on the medical information provided by the user. In the "reports" section, it will indicate whether the diagnosis result suggests a positive or negative liver disease, in accordance with the model's calculations.

IV. CONCLUSION

The validation test results of the C5.0 model and Support Vector Machine with and without feature selection have different accuracy values. Model validation with feature selection produces better accuracy. The C5.0 model and the Support Vector Machine (RBF) with feature selection have similar accuracy, namely 76%, while the Support Vector Machine (Sigmoid) has an accuracy of 70%. If without feature selection the C5.0 algorithm has an accuracy of 66%, the Support Vector Machine between the RBF and sigmoid kernels has an accuracy of 69% and 55%. So it is concluded that CFS feature selection can improve accuracy results in classifying this system and C5.0 and SVM are quite good in classifying this liver dataset. However, in terms of time required, the C5.0 algorithm is slightly more effective in classifying this research.

With this liver disease classification system, it is hoped that it will help users find out whether they have the potential to experience liver disease or not. This has great benefits in preventing potential negative impacts on their liver health. This system opens up the opportunity for users to be more aware of the risk of developing liver disease, allowing for better preventive measures for the health of their liver.

In future research, the C5.0 and SVM algorithms can be developed by adding certain parameters in the process, and also adding other feature selections to find more optimal performance results. Or use unsupervised learning algorithms to classify this Liver dataset.

REFERENCES

- [1] C. Yustitya Gobel, "SISTEM PAKAR PENYAKIT LIVER MENGGUNAKAN K-NEAREST NEIGHBORS ALGORITM BERBASIS WEBSITE," 2018.
- [2] A. Siburian and C. F. Chang, "CHRONIC LIVER DISEASE LOWERING PHYSICAL AND MENTAL HEALTH DIMENSIONS," *Jurnal Keperawatan Indonesia*, vol. 24, no. 3, pp. 149–156, Nov. 2021, doi: 10.7454/JKI.V24I3.1240.
- [3] S. Cheemerla and M. Balakrishnan, "Global Epidemiology of Chronic Liver Disease," *Clinical Liver Disease*, vol. 17, no. 5. John Wiley and Sons Inc, pp. 365–370, May 01, 2021. doi: 10.1002/cld.1061.
- [4] E. Costello *et al.*, "Exposure to per-and Polyfluoroalkyl Substances and Markers of Liver Injury: A Systematic Review and Meta-Analysis," *Environ Health Perspect*, vol. 130, no. 4, Apr. 2022, doi: 10.1289/EHP10092.
- [5] E. Patimah *et al.*, *Klasifikasi Penyakit Liver dengan Menggunakan Metode Decision Tree*. 2021.
- [6] A. Farida Muti and C. Anindya, "Analysis of Potential Drug Interactions in Liver Cirrhosis Patients," *Galenika Journal of Pharmacy (e-Journal)*, vol. 7, no. 1, pp. 18–29, 2021, doi: 10.22487/j24428744.2021.v7.i1.
- [7] J. Ilmiah Matematika, S. Sofia Nurwananda, and R. Sulaiman, "MATHunesa Tahun 2022 APLIKASI HIMPUNAN FUZZY INTUISIONISTIK DALAM DIAGNOSA PENYAKIT HEPATITIS

- MENGGUNAKAN EXTENDED HAUSDORFF DISTANCE”.
- [8] N. Benediktus and R. S. Oetama, “Algoritma Klasifikasi Decision Tree C5.0 untuk Memprediksi Performa Akademik Siswa,” *14 ULTIMATICS*, vol. XII, no. 1, 2020, [Online]. Available: <https://www.kaggle.com/aljarah/xAPI-Edu-Data>
- [9] P. W. Kastawan, D. M. Wiharta, and M. Sudarma, “Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil,” *Majalah Ilmiah Teknologi Elektro*, vol. 17, no. 3, p. 371, Dec. 2018, doi: 10.24843/mite.2018.v17i03.p11.
- [10] D. Pramadhana, “Klasifikasi Penyakit Diabetes Menggunakan Metode CFS dan ROS dengan Algoritma J48 Berbasis Adaboost,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 5, no. 1, pp. 89–98, Jun. 2021, doi: 10.29408/edumatic.v5i1.3336.
- [11] D. Pramadhana, “Edumatic: Jurnal Pendidikan Informatika Klasifikasi Penyakit Diabetes Menggunakan Metode CFS dan ROS dengan Algoritma J48 Berbasis Adaboost,” vol. 5, no. 1, 2021, doi: 10.29408/edumatic.v5i1.3336.
- [12] Y. Priantama, T. Azhima, and Y. Siswa, “OPTIMASI CORRELATION-BASED FEATURE SELECTION UNTUK PERBAIKAN AKURASI RANDOM FOREST CLASSIFIER DALAM PREDIKSI PERFORMA AKADEMIK MAHASISWA,” *Jurnal Informatika dan Komputer*, vol. 6, no. 2, pp. 251–260, 2022.
- [13] W. Widayani and H. Harliana, “Analisis Support Vector Machine Untuk Pemberian Rekomendasi Penundaan Biaya Kuliah Mahasiswa,” *Jurnal Sains dan Informatika*, vol. 7, no. 1, pp. 20–27, Jun. 2021, doi: 10.34128/jsi.v7i1.268.
- [14] N. Benediktus and R. S. Oetama, “Algoritma Klasifikasi Decision Tree C5.0 untuk Memprediksi Performa Akademik Siswa,” *14 ULTIMATICS*, vol. XII, no. 1, 2020, [Online]. Available: <https://www.kaggle.com/aljarah/xAPI-Edu-Data>
- [15] R. N. Amalda, N. Millah, and I. Fitria, “IMPLEMENTASI ALGORITMA C5.0 DALAM MENGANALISA KELAYAKAN PENERIMA KERINGANAN UKT MAHASISWA ITK,” *Teorema: Teori dan Riset Matematika*, vol. 7, no. 1, p. 101, Mar. 2022, doi: 10.25157/teorema.v7i1.6692.
- [16] W. Hidayat, M. Ardiansyah, and A. Setyanto, “Pengaruh Algoritma ADASYN dan SMOTE terhadap Performa Support Vector Machine pada Ketidakseimbangan Dataset Airbnb,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 5, no. 1, pp. 11–20, Jun. 2021, doi: 10.29408/edumatic.v5i1.3125.
- [17] M. Khubaib Tamami and I. Kharisudin, “Indonesian Journal of Mathematics and Natural Sciences Komparasi Metode Support Vector Machine dan Naive Bayes Classifier untuk Pemodelan Kualitas Pengajuan Kredit.” [Online]. Available: <http://journal.unnes.ac.id/nju/index.php/JM>
- [18] I. A. Muis and M. Affandes, “Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet,” *Jurnal Sains, Teknologi dan Industri*, vol. 12, no. 2, pp. 189–197, 2015, [Online]. Available: <http://ejournal.uin-suska.ac.id/index.php/sitekin>
- [19] R. Indraswari and A. Z. Arifin, “RBF KERNEL OPTIMIZATION METHOD WITH PARTICLE SWARM OPTIMIZATION ON SVM USING THE ANALYSIS OF INPUT DATA’S MOVEMENT,” *Jurnal Ilmu Komputer dan Informasi*, vol. 10, no. 1, p. 36, Feb. 2017, doi: 10.21609/jiki.v10i1.410.
- [20] J. Mase, M. T. Furqon, and B. Rahayudi, “Penerapan Algoritme Support Vector Machine (SVM) Pada Pengklasifikasian Penyakit Kucing,” 2018. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [21] T. Meisya, P. Aulia, N. Arifin, and R. Mayasari, “PERBANDINGAN KERNEL SUPPORT VECTOR MACHINE (SVM) DALAM PENERAPAN ANALISIS SENTIMEN VAKSINISASI COVID-19,” [Online]. Available: <https://doi.org/10.31598>
- [22] S. Widodo, H. Brawijaya, and S. Samudi, “Stratified K-fold cross validation optimization on machine learning for prediction,” *Sinkron*, vol. 7, no. 4, pp. 2407–2414, Oct. 2022, doi: 10.33395/sinkron.v7i4.11792.
- [23] I. Widhi Saputro and B. Wulan Sari, “Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa Naïve Bayes Algorithm Performance Test for Student Study Prediction,” *Citec Journal*, vol. 6, no. 1, 2019.
- [24] W. Steven Dharmawan *et al.*, “Penerapan Metode SDLC Waterfall Dalam Perancangan Sistem Informasi Administrasi Keuangan Berbasis Desktop,” vol. VI, no. 2, 2018.
- [25] M. Diah Larasati and I. Satriadi, “Model Waterfall Untuk Pengembangan Sistem Informasi Pengolahan Nilai Pada SMP Kartika XI-3 Jakarta Timur”, doi: 10.31294/jtk.v4i2.
- [26] D. B. Setyohadi and F. A. Kristiawan, “PREPROCESSING ITERATIVE PARTITIONING FILTER ALGORITHM,” 2017.
- [27] R. P. Padang, “IMPLEMENTASI DATA MINING ALGORITMA C5.0 DALAM MEMPREDIKSI PENERIMAAN CLEANING SERVICE (CS) PADA PT ISS INDONESIA MEDAN,” 2019.
- [28] D. Normawati and S. A. Prayogi, “Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter,” 2021.
- [29] J. A. Firdaus, A. Setia Budi, and E. Setiawan, “ANALISIS PERFORMA ALGORITMA MACHINE LEARNING PADA PERANGKAT EMBEDDED ATMEGA328P,” vol. 10, no. 2, pp. 245–254, 2023, doi: 10.25126/jtiik.2023106196.
- [30] R. N. Amalda, N. Millah, and I. Fitria, “IMPLEMENTASI ALGORITMA C5.0 DALAM MENGANALISA KELAYAKAN PENERIMA KERINGANAN UKT MAHASISWA ITK,” *Teorema: Teori dan Riset Matematika*, vol. 7, no. 1, p. 101, Mar. 2022, doi: 10.25157/teorema.v7i1.6692.