Contents lists available at https://citedness.com/index.php/jdsi

Data Science Insights

Journal Page is available to https://citedness.com/index.php/jdsi

Research article

Drug Classification using Machine Learning Algorithms

Hengky Fernando

Jurusan Teknik Informatika, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Riau, Indonesia
email: [1]hengky.fernando@student.pelitaindonesia.ac.id

ARTICLE INFO

Article history:
Received 14 December 2024
Revised 01 January 2024
Accepted 31 January 2024
Available online 15 February 2025

Keywords:
Drug management
Random Forest
Cross-validation
Drug type grouping
Machine learning algorithms

Please cite this article in IEEE style as:
H. fernando and Abdullah, "Drug Classification Using Machine Learning Algorithms ", *Data Science Insights*, vol. 3, no. 1, pp. 1–9, Feb. 2025.

ABSTRACT

The right selection of drugs is a crucial factor in the treatment of various diseases to ensure the effectiveness of therapy and avoid risks that can worsen the patient's condition. This study aims to develop a machine learning-based prediction model to classify the appropriate type of drug based on patient characteristics. Several machine learning algorithms are tested to determine the most optimal model. The results of the analysis show that the Random Forest algorithm provides the best performance with the highest level of accuracy in predicting the right type of drug. Thus, the Random Forest-based model is recommended to be implemented as a decision support tool in the selection of drug therapies that is more accurate and efficient.

Correspondence:
Hengky Fernando
Jurusan Teknik Informatika, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Riau, Indonesia
hengky.fernando@student.pelitaindonesia.ac.id

1. Introduction

Choosing drugs that suit the patient's condition is one of the main challenges in the medical world. This decision is very important because it is directly related to the effectiveness of treatment [1] , patient compliance, and the reduction of the risk of side effects. However, this process is often complex because it involves various factors such as age, medical history, clinical diagnosis, as well as the individual's biological response to a particular drug. Inaccuracies in drug selection can lead to less effective or even harmful treatment for patients [2] .

To overcome this challenge, machine learning technology [3] has become one of the promising solutions. Machine learning enables in-depth analysis of medical data to identify patterns, make predictions, and provide more accurate [4] data-driven recommendations. In the context of drug classification, machine learning algorithms [5] can be used to predict the type of drug that best suits the patient's condition based on available medical data.

Some machine learning algorithms that are often used for classification tasks, such as Naïve Bayes [6] , k-Nearest Neighbors [7] , Random Tree [8] , Deep Learning [9] , and Random Forest [10] , offer different approaches to analyzing data. Each algorithm has advantages and disadvantages depending on the complexity of the data, the number of features, and the need for analysis.

This machine learning algorithm aims to explain the importance of selecting drugs that are suitable for the patient's condition in the medical world and the challenges faced in the process. Additionally, this introduction outlines the role of machine learning technology as a potential solution to improve the accuracy and effectiveness of treatment through in-depth analysis of medical data. By comparing the performance of various algorithms [11] , the most effective method in predicting and classifying drugs based on the available datasets can be determined, The proper implementation of this technology also has the potential to lower overall healthcare costs by minimizing trial-and-error in drug selection.

Several previous studies have examined the use of machine learning in predicting drug selection that suits the patient's condition. A study implemented the Random Forest [10] algorithm to predict the right type of drug as well as identify the key factors that influenced the decision. In addition, other approaches have developed techniques to improve the accuracy of classification models by combining clustering [12] and logistic [13] regression methods, which show promising results. On the other hand, genetic algorithms are also used to improve the machine [14] learning process with a more adaptive approach, so that it is able to optimize prediction accuracy. These approaches make an important contribution to the development of data-driven predictive technologies to support decision-making in the medical world.

2. Research Methodology

In the field of Data Science [15] , there are a number of important procedures that must be implemented to ensure that the data produced has a high level of accuracy, so as to be able to support the credibility and validity of research. Quality data will be a strong foundation in supporting various aspects of research, as well as making a significant contribution to future decision-making. The procedures carried out by the author in processing data include:

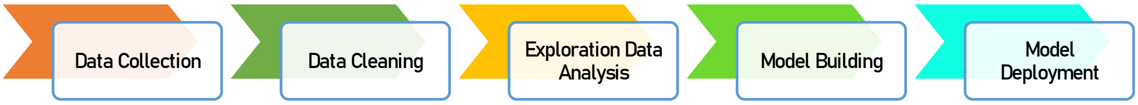


Figure 1 Data Science Process

The Figure 1 above explain the stages of implementing the data model, starting from collection, cleaning, analysis, development, to model implementation. This process is carried out meticulously to produce relevant solutions and support research.

2.1 Data Collection

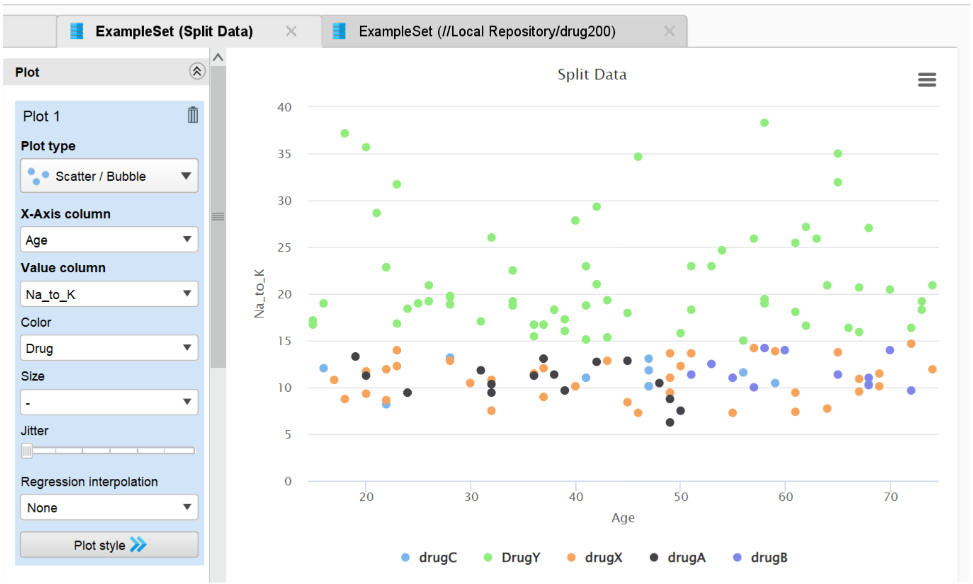
The first step is to collect data from various sources, which can be numbers, text, images, and more. Data can also be obtained through direct surveys. It is important to ensure that the data collected is accurate, complete, and trustworthy. In this study, the data used was obtained from the Kaggle website (www.kaggle.com). Kaggle is an online platform and community focused on data science and machine learning. Since its inception in 2010, Kaggle has become a premier place for data scientists, data analysts, and learners looking to improve their skills, collaborate, and stay up-to-date with the latest developments in the field. The data available on the site is quite accurate and complete, depending on the research topic you want to study

2.2 Data cleaning

The data that has been collected generally needs to be screened to remove irrelevant information from the dataset. At this stage, cleanup is carried out to overcome errors, redundancy, and data inconsistencies. This process is essential to ensure that the data to be analyzed is accurate and reliable. Data cleaning is done using Microsoft Excel software. At this stage, problematic data is selected and the format is standardized so that the results obtained from data processing are more accurate.

2.3 Exploratory Data Analysis

Once the data is completed through the cleanup stage, data scientists begin to apply various statistical methods and visualization techniques to find hidden patterns, trends, and relationships in the dataset. At this stage, it is important to remain objective and not rush into concluding results based solely on what is seen. Data scientists must evaluate various possibilities and ensure the analysis is supported by evidence relevant to the research objectives.



Plot type: Scatter / Bubble
X-Axis column: Age
Value column: Na_to_K
Color: Drug
Size: -
Jitter
Regression interpolation: None

Figure 2 Data Visualization

In Figure 2, this visualization shows the relationship between age, sodium to potassium ratio (Na_to_K), and the type of drug prescribed (drugA, drugB, drugC, drugX, drugY). Each type of drug has a different distribution pattern based on age and Na_to_K ratio, which can be used for predictive analysis in drug selection.

Data visualization at this stage is carried out using RapidMiner to simplify the analysis process. RapidMiner offers the advantage of high flexibility in processing and visualizing data from various sources. This allows researchers to conduct a more holistic and in-depth analysis. With a wide range of visualization features provided, from simple graphs to complex visualizations, Tableau supports more effective and comprehensive data exploration.

2.4 Model Building

With the insights gained from the previous data analysis process, the next step is to build a model that can be used to make predictions, groupings, or classifications of new data with a high degree of accuracy. The process of building this model requires not only technical skills, but also critical thinking and a deep understanding of the data being analyzed as well as relevant algorithms. The selection of the right algorithm depends largely on the nature of the data, the purpose of the research, and the complexity of the problem at hand.

At this stage, the model construction is carried out using the RapidMiner software. RapidMiner is one of the popular data analytics platforms due to its ability to provide intuitive tools and environments for data processing, exploration, and application of machine learning algorithms. With an easy-to-use drag-and-drop interface, RapidMiner allows data scientists and researchers to implement a variety of analysis models without having to write complex code.

In addition, RapidMiner supports the application of various machine learning algorithms, such as Naïve Bayes, Random Forest, Random Tree, Deep Learning, to other improvement algorithms, which can help evaluate which algorithm is most suitable to achieve research goals. RapidMiner also comes with model performance evaluation tools, such as confusion, accuracy, precision, and recall matrices, making it easy to compare the effectiveness of different algorithms.

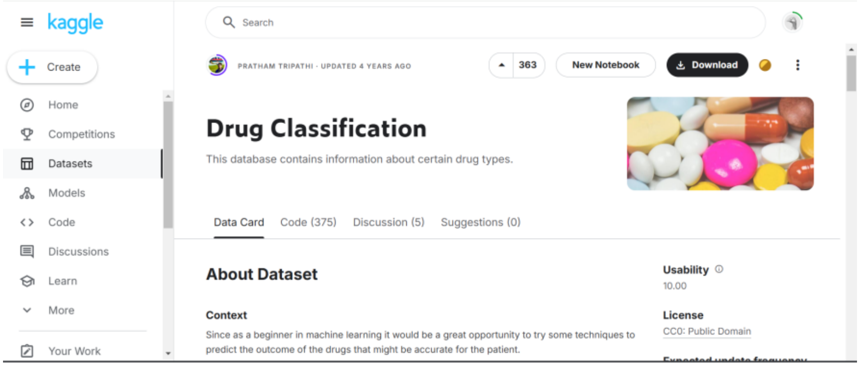
2.5 Model Deployment

The model that has been successfully built will be applied in various aspects relevant to the research objectives, such as sales strategies, customer retention efforts, and others. By carrying out this entire process carefully and critically, data scientists can contribute to solving various problems in the world, especially in achieving the results of the research that has been carried out.

3 Results and Discussion

In this section, the results will be given according to the stages carried out:

3.1 Data Collection



PRATHAM TRIPATHI · UPDATED 4 YEARS AGO

Drug Classification

This database contains information about certain drug types.

About Dataset

Context

Since as a beginner in machine learning it would be a great opportunity to try some techniques to predict the outcome of the drugs that might be accurate for the patient.

Usability 10.00

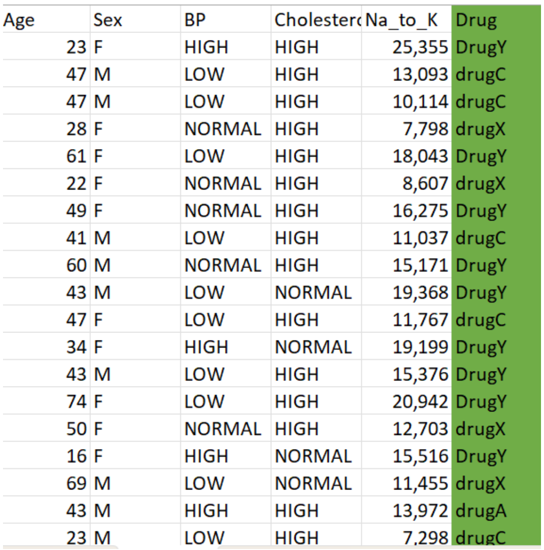
License CC0: Public Domain

Figure 3 Kaggle Website

The dataset used in this study contains customer data related to drug use that is stored to support behavioral predictions in machine learning (Figure 3). This study aims to analyze and classify the results of drug use, in order to predict the level of accuracy and effectiveness for patients.

3.2 Data Collection Results

The following is a snippet of the dataset. We obtain this data set from the website (https://www.kaggle.com/).



Age	Sex	BP	Cholester	Na_to_K	Drug
23	F	HIGH	HIGH	25,355	DrugY
47	M	LOW	HIGH	13,093	drugC
47	M	LOW	HIGH	10,114	drugC
28	F	NORMAL	HIGH	7,798	drugX
61	F	LOW	HIGH	18,043	DrugY
22	F	NORMAL	HIGH	8,607	drugX
49	F	NORMAL	HIGH	16,275	DrugY
41	M	LOW	HIGH	11,037	drugC
60	M	NORMAL	HIGH	15,171	DrugY
43	M	LOW	NORMAL	19,368	DrugY
47	F	LOW	HIGH	11,767	drugC
34	F	HIGH	NORMAL	19,199	DrugY
43	M	LOW	HIGH	15,376	DrugY
74	F	LOW	HIGH	20,942	DrugY
50	F	NORMAL	HIGH	12,703	drugX
16	F	HIGH	NORMAL	15,516	DrugY
69	M	LOW	NORMAL	11,455	drugX
43	M	HIGH	HIGH	13,972	drugA
23	M	LOW	HIGH	7,298	drugC

Figure 4 Dataset

From the dataset (Figure 4), the data is divided by a ratio of 20:20:20:20:20 based on the type of drug, DrugY, DrugX, DrugC, DrugA, DrugB.

3.3 Data Model Results

In this section, the results of the performance evaluation of five machine learning algorithms used in drug data classification will be presented, namely Naïve Bayes, k-Nearest Neighbors (KNN), Random Forest, Random Tree, and Deep Learning. Each algorithm was chosen because it has a unique approach to analyzing medical data and the potential to provide accurate results. The analysis was carried out to understand the advantages and disadvantages of each algorithm in handling the available datasets. The following are the results of the evaluation and discussion of the application of each of these algorithms.

3.3.1. Naïve Bayes

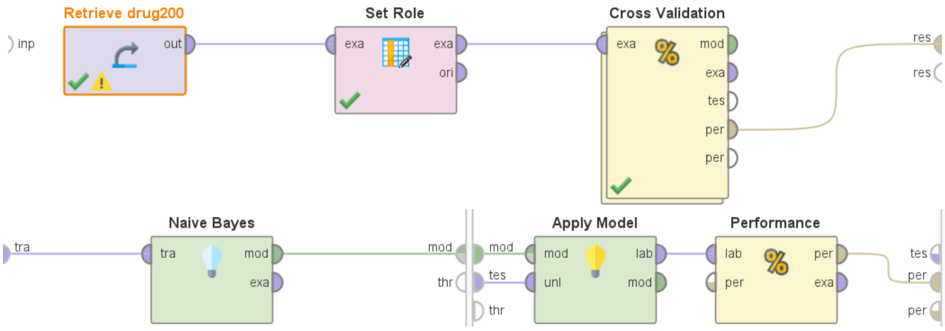
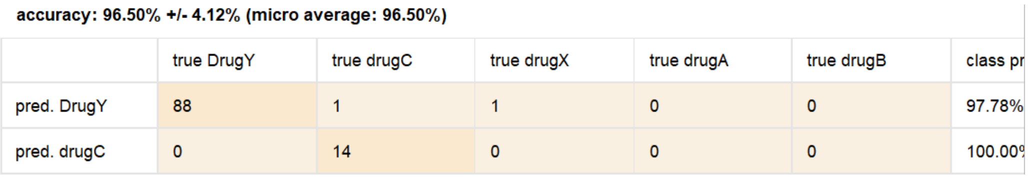


Figure 5 Naïve Bayes Algorithm (Cross Validation) Using Rapidminer



accuracy: 96.50% +/- 4.12% (micro average: 96.50%)

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	88	1	1	0	0	97.78%
pred. drugC	0	14	0	0	0	100.00

Figure 6 Results of Naïve Bayes Algorithm (Cross Validation) Using Rapidminer

Figure 5 is the data model using Naïve Bayes and Figure 6 is the result of the method using the Naïve Bayes algorithm in RapidMiner. After ensuring that the data is clean, the data is cross-validated in which classification has been carried out using the first stage of the naïve Bayes algorithm. Then the application of the model is carried out to get its performance. Accuracy obtained: is 96.50%

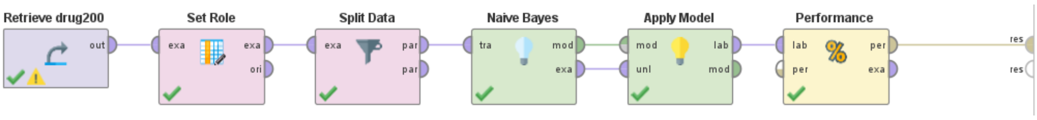
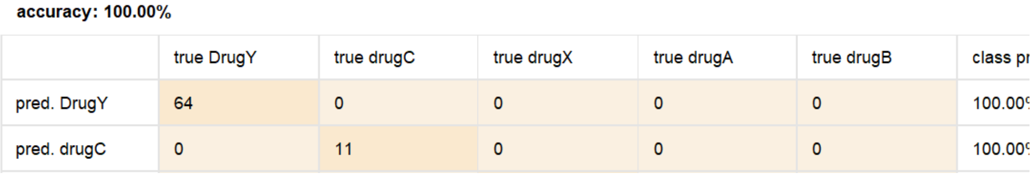


Figure 7 Naïve Bayes (Hold Out) Algorithm Using Rapidminer



accuracy: 100.00%

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	64	0	0	0	0	100.00
pred. drugC	0	11	0	0	0	100.00

Figure 8 Results of Naïve Bayes Algorithm (Hold Out) Using Rapidminer

Figure 7 is the data model using Naïve Bayes and Figure 8 is the result of the method using the Naïve Bayes algorithm in RapidMiner. After ensuring the data is clean, the data is divided into 70% for training and 30% for testing. Then the application of the model is carried out to get its performance. Accuracy result: is 100.00%

3.3.2. K-Nearest Neighbor (k-NN)

Cross Validation Method

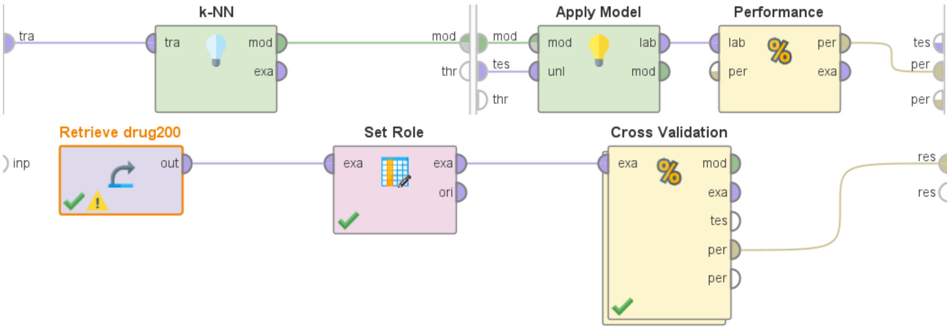
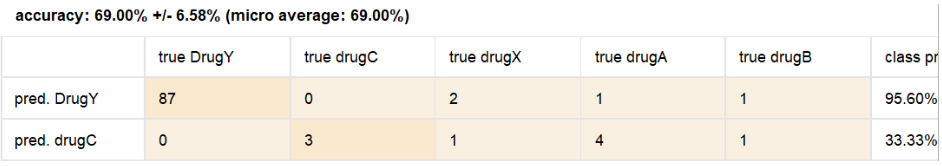


Figure 9 k-NN (Cross Validation) Algorithm Using Rapidminer



accuracy: 69.00% +/- 6.58% (micro average: 69.00%)

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	87	0	2	1	1	95.60%
pred. drugC	0	3	1	4	1	33.33%

Figure 10 Results of k-NN (Cross Validation) Algorithm Using Rapidminer

Figure 9 is the data model using K-Nearest Neighbor (k-NN) and Figure 10 is the result of the method using the K-Nearest Neighbor (k-NN) algorithm in RapidMiner. After ensuring that the data is clean, cross-validation is carried out in which classification has been carried out using the first stage of the K-Nearest Neighbor (k-NN) algorithm. Then the application of the model is carried out to get its performance. Accuracy obtained: is 69.00%

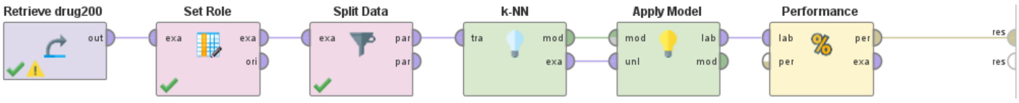
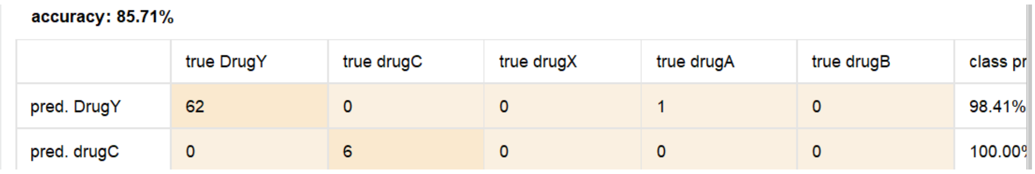


Figure 11 k-NN (Hold Out) Algorithm Using Rapidminer



accuracy: 85.71%

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	62	0	0	1	0	98.41%
pred. drugC	0	6	0	0	0	100.00

Figure 12 Results of k-NN (Hold Out) Algorithm Using Rapidminer

Figure 11 is the data model using K-Nearest Neighbor (k-NN) and Figure 12 is the result of the method using the K-Nearest Neighbor (k-NN) algorithm in RapidMiner. After ensuring the data is clean, the data is divided into 70% for training and 30% for testing. Then the application of the model is carried out to get its performance. Accuracy obtained: is 85.71%

3.3.3. Random Forest

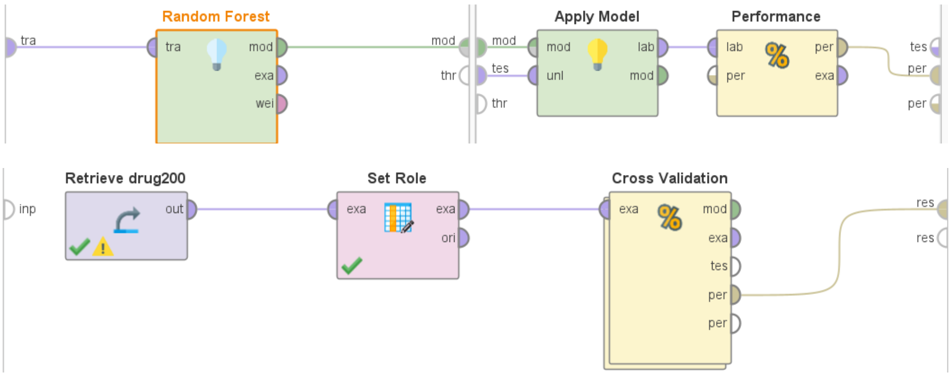
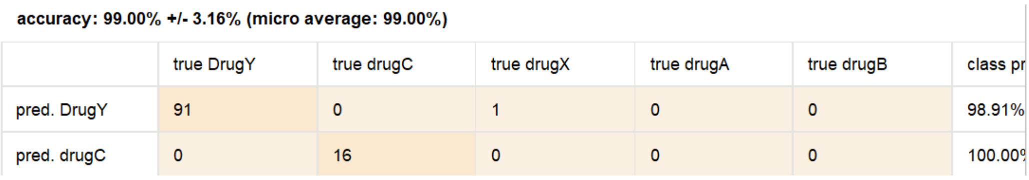


Figure 13 Random Forest Algorithm (Cross Validation) Using Rapidminer



accuracy: 99.00% +/- 3.16% (micro average: 99.00%)

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	91	0	1	0	0	98.91%
pred. drugC	0	16	0	0	0	100.00

Figure 14 Results of Random Forest Algorithm (Cross Validation) Using Rapidminer

Figure 13 is the data model using Random Forest and Figure 14 is the result of the method using the Random Forest algorithm in RapidMiner. After ensuring that the data is clean, the data is cross-validated in which classification has been carried out using the first stage of the Random Forest algorithm. Then the application of the model is carried out to get its performance. Accuracy obtained: is 99.00%

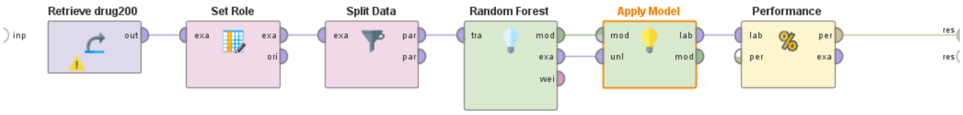
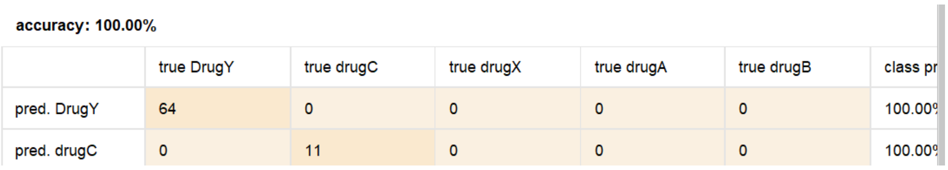


Figure 15 Random Forest (Hold Out) Algorithm Using Rapidminer



accuracy: 100.00%

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	64	0	0	0	0	100.00
pred. drugC	0	11	0	0	0	100.00

Figure 16 Results of Random Forest (Hold Out) Algorithm Using Rapidminer

Figure 15 is a data model using Random Forest and Figure 16 is the result of a method using the Random Forest algorithm in RapidMiner. After ensuring the data is clean, the data is divided into 70% for training and 30% for testing. Then the application of the model is carried out to get its performance. Accuracy result: is 100.00%

3.3.4. Random Tree Algorithm

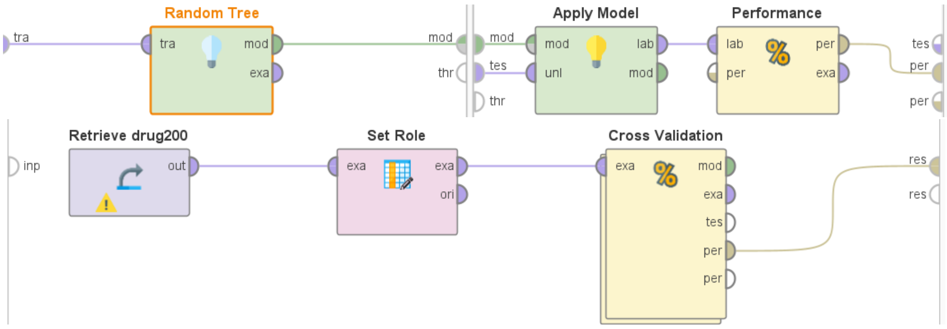
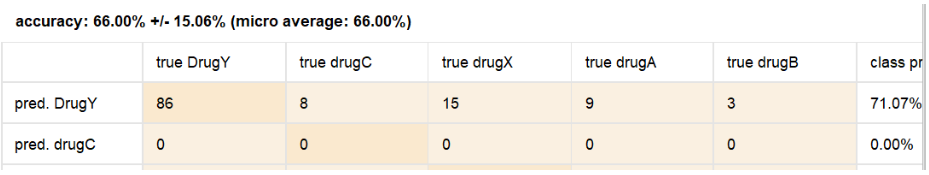


Figure 17 Random Tree Algorithm (Cross Validation) Using Rapidminer



accuracy: 66.00% +/- 15.06% (micro average: 66.00%)

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	86	8	15	9	3	71.07%
pred. drugC	0	0	0	0	0	0.00%

Figure 18 Random Tree Algorithm (Cross Validation) Using Rapidminer

Figure 17 is the data model using Random Tree and Figure 18 is the result of the method using the Random Tree algorithm in RapidMiner. After ensuring that the data is clean, the data is cross-validated in which classification has been carried out using the first stage of the Random Tree algorithm. Then the application of the model is carried out to get its performance. Accuracy obtained: is 66.00%

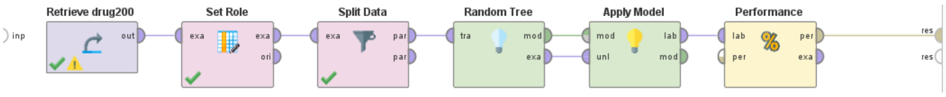
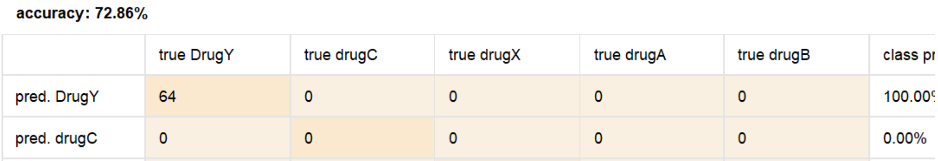


Figure 19 Random Tree (Hold Out) Algorithm Using Rapidminer



accuracy: 72.86%

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	64	0	0	0	0	100.00
pred. drugC	0	0	0	0	0	0.00%

Figure 20 Results of Random Tree (Hold Out) Algorithm Using Rapidminer

Figure 19 is a data model using a Random Tree and Figure 20 is a result method using the Random Tree algorithm in RapidMiner. After ensuring the data is clean, the data is divided into 70% for training and 30% for testing. Then the application of the model is carried out to get its performance. The accuracy obtained: is 72.86%

3.3.5. Deep Learning

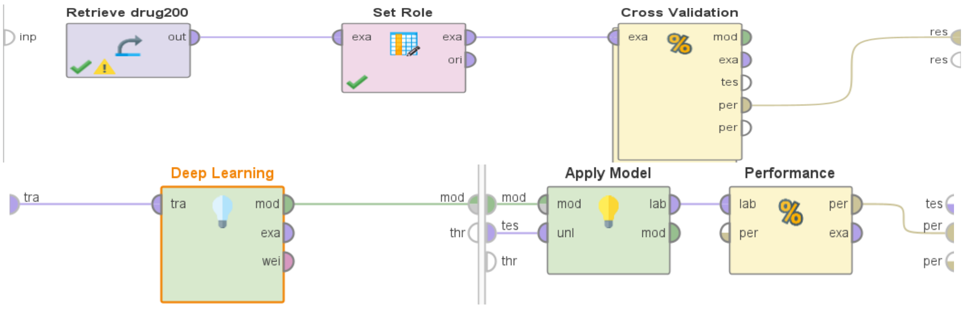
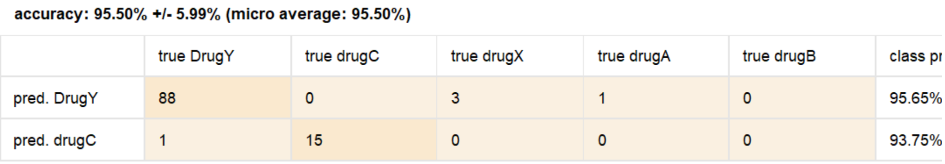


Figure 21 Deep Learning Algorithm (Cross Validation) Using Rapidminer



accuracy: 95.50% +/- 5.99% (micro average: 95.50%)

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	88	0	3	1	0	95.65%
pred. drugC	1	15	0	0	0	93.75%

Figure 22 Results of Deep Learning Algorithm (Cross Validation) Using Rapidminer

Figure 21 is a data model using Deep Learning and Figure 22 is the result of a method using the Deep Learning algorithm in RapidMiner. After ensuring that the data is clean, the data is cross-validated in which classification has been carried out using the first stage of Deep Learning algorithms. Then the application of the model is carried out to get its performance. Accuracy obtained: is 95.50%

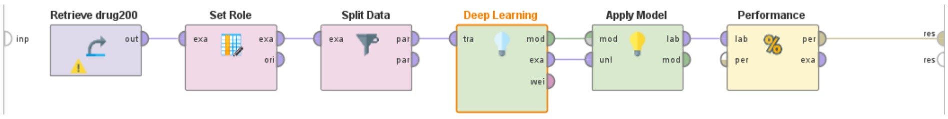
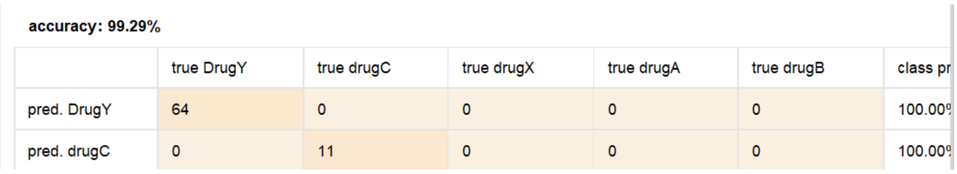


Figure 23 Deep Learning (Hold Out) Algorithm Using Rapidminer



accuracy: 99.29%

	true DrugY	true drugC	true drugX	true drugA	true drugB	class pr
pred. DrugY	64	0	0	0	0	100.00
pred. drugC	0	11	0	0	0	100.00

Figure 24 Results of Deep Learning Algorithm (Hold Out) Using Rapidminer

Figure 23 is the data model using Deep Learning and Figure 24 is the result of the method using the Deep Learning algorithm in RapidMiner. . After ensuring the data is clean, the data is divided into 70% for training and 30% for testing. Then the application of the model is carried out to get its performance. The accuracy obtained: is 99.29%

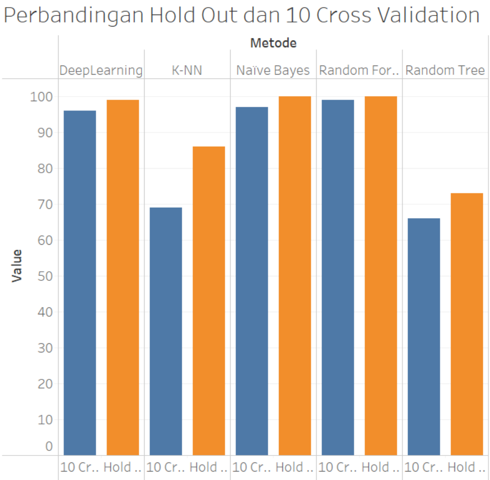


Figure 25 Comparison Results of Algorithms Using Tablau

After testing the five algorithms, the results show that the Random Forest algorithm is the most suitable algorithm for the purpose of this research. Based on Figure 13 and Figure 15, the Random Forest algorithm recorded the highest level of accuracy compared to the other four algorithms tested in this study, with an accuracy of 99.00% and 100.00%, respectively.

4. Conclusion

Fierce competition in the health sector drives the need for data-driven strategies to improve the effectiveness of drug management. This study aims to build a prediction system using a drug management dataset by testing five algorithms, namely deep learning, k-nearest neighbors (k-NN), naïve bayes, random tree, and random forest. The sample data was divided into 70% for training and 30% for testing, and cross-validation was applied to improve the reliability of the results. From the test results, the Random Forest algorithm shows the best performance with the highest level of accuracy compared to other algorithms. Therefore, random forest was chosen as a method implemented to support more effective drug data management in improving global health strategies.

References

[1] N. Mutmainah, s. Ernawati, and E. M. Sutrisna, "Identification of Potential Drug Related Problems (DRPS) Category of Inaccurate Drug Selection in Hypertensive Patients with Diabetes Mellitus in the Inpatient Installation of Hospital x Jepara in 2007" 2008.

[2] A. Halawa, s. Setiawan, and b. Syam, "nurses' perception of the role in improving patient safety," *Journal of Telenursing (Joting)*, Vol. 3, No. 1, pp. 73–84, Apr 2021, doi: 10.31539/joting.v3i1.2096.

[3] W. Hastomo, n. Aini, a. Satyo, b. Karno, and l. M. R. Rere, "Machine Learning Methods for Predicting Emissions of Manure Management," 2022.

[4] S. A. Rajagukguk, "Systematic Literature Review: Prediction of Students' Learning Achievement Using Machine Learning Algorithms," *Journal of Science, Reason, and Applications of Information Technology*, Vol. 1, No. 1, Aug 2021, doi: 10.20885/snati.v1i1.4.

[5] A. Sulistiyawati and e. Supriyanto, "Implementation of the k-means clustring algorithm in determining superior class students," vol. 15, no. 2.

[6] G. I. Webb, "naïve bayes," in *Encyclopedia of Machine Learning and Data Mining*, Springer US, 2016, pp. 1–2. Doi: 10.1007/978-1-4899-7502-7_581-1.

[7] A. R. Isnain, j. Supriyanto, and m. P. Kharisma, "implementation of k-nearest neighbor (k-nn) algorithm for public sentiment analysis of online learning," vol. 15, no. 2, p. 121, apr 2021, doi: 10.22146/ijccs.65176.

[8] R. Abraham and j.-f. Delmas, "record process on the continuum random tree," jul 2011, [online]. Available on: http://arxiv.org/abs/1107.3657

[9] I. J. Jacob and P. E. Darney, "design of deep learning algorithm for iot application by image based recognition," *Journal of ISMAC*, vol. 3, no. 3, pp. 276–290, Aug 2021, doi: 10.36548/jismac.2021.3.008.

[10] S. J. Rigatti, "random forest," 2017. [online]. Available on: http://meridian.allenpress.com/jim/article-pdf/47/1/31/1736157/insm-47-01-31-39_1.pdf

[11] A. Likas, n. Vlassis, and j. Verbeek, "the global k-means clustering algorithm the global k-means clustering algorithm.

[12] D. Xu Dan Yi. Tian, "a comprehensive survey of clustering algorithms," doi: 10.1007/s40745-015-0040-1.

[13] I. Z. Muflihah, "Analysis of Financial Distress of Manufacturing Companies in Indonesia with Logistic Regression".

[14] K. P. Fourkiotis and a. Tsadiras, "applying machine learning and statistical forecasting methods for enhancing pharmaceutical sales predictions," *Forecasting*, vol. 6, no. 1, pp. 170–186, Mar 2024, doi: 10.3390/forecast6010010.

[15] F. Provost and t. Fawcett, "data science and its relationship to big data and data-driven decision making," *Big Data*, Vol. 1, No. 1, pp. 51–59, Mar 2013, doi: 10.1089/big.2013.1508.