Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Development of Hybrid Machine Learning in Patient Diagnosis Classification Using the XRP Model (Extraction, Reduction & Prediction)

Hendra Nusa Putra*[1], ***Sarjon Defit***[2], ***Gunadi Widi Nurcahyo***[2]

[1]Medical Record Department, STIKES Dharma Landbouw, Padang, Indonesia.
[2] Information Technology Doctoral Department, Faculty of Computer Science, UPI YPTK, Padang, Indonesia.
*Corresponding Author.

Received 11/06/2023, Revised 15/03/2024, Accepted 17/03/2024, Published Online First 20/07/2024

Abstract

This study, carried out over six months at an Indonesian hospital, explores the benefits of standardizing medical record data and integrating health information systems for healthcare delivery. Utilizing a quantitative research approach, it focuses on the impact of precise data mining extraction on data analysis and the advantages of an integrated system for accessing patient records. Advanced data mining methods were employed for feature extraction, selection, and dataset reduction to enhance data classification accuracy. Findings revealed a direct correlation between the accuracy of data extraction and the reliability of data classification, highlighting the significant role of dataset reduction in improving analysis precision. The introduction of the XRP Model, a new predictive tool for assessing disease likelihood, marked a notable advancement, demonstrating high accuracy rates in predicting diabetes and heart disease (96.8% and 88%, respectively). The model's consistent performance across various outcome scenarios underscores its potential in healthcare decision-making. This research evidences the value of advanced data mining and dataset reduction in refining data classification, thus facilitating better healthcare decisions. The XRP Model's success in disease prediction suggests considerable benefits for healthcare services, offering insights crucial for the development and optimization of health information systems. These findings have the potential to influence healthcare policy and practice, advocating for a new standard in healthcare data management.

Keywords: Disease Prediction, Feature Reduction, Feature Selection, Machine Learning, Medical Record.

Introduction

The development of electronic medical records has significantly facilitated the analysis and storage of medical data. Despite this, it is challenging to facilitate the sharing of medical information among various healthcare facilities because electronic medical records contain a sizable amount of personal privacy information[1].

The healthcare and medical sector are more in need of data mining today. When specific data mining methods are used correctly, valuable information can be extracted from large databases, which can help the medical practitioner make early decisions and improve health services. The accurate analysis of medical databases is helpful in early illness prediction, patient care, and community services. There are several applications where methodologies based on machine learning have been utilized successfully, including disease prediction. By

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

assisting clinicians in the early identification and prediction of diseases, developing a classifier system employing machine learning algorithms aims to contribute significantly to solving health-related problems.

Machine learning algorithms and electronic medical records (EMR) can help detect disease and determine its adverse effects. This is accomplished by utilizing diverse healthcare data types through algorithms and practical use cases, helping people to understand the basic ideas behind healthcare data analysis. As the volume of data grows, machine learning helps doctors speed up examinations and produce more accurate results. Large datasets that are too complex for human study can be mined for knowledge with astonishing ease thanks to machine learning. This research aims to help doctors specializing in specific fields get other perspectives and determine whether certain scenarios are feasible[2].

The weighting of electronic medical record features in the extraction assessment model will produce more accurate knowledge in producing predictive outcomes. It will go through stages of crawling, data preparation, and data reduction during the process, and it will be polished using several algorithms. An electronic medical record dataset will be used to test the outcomes. Prediction of patient disease is crucial to patient care and handling in the medical field[3]. Correct disease identification and prediction can impact clinical judgments made by doctors, allow for early intervention, and enhance the overall quality of patient care. Patients frequently experience symptoms in clinical settings that could point to a particular illness. Doctors may be unable to identify the underlying condition due to the numerous symptoms that must be assessed and studied. The proper features must be chosen to improve the disease prediction and evaluation system's performance because each symptom may have varying degrees of correlation with particular diseases[4, 5].

The quantity of usable examples (instances) directly influences the effectiveness of the data mining process. Although its use can significantly reduce the risk of making poor decisions, it does not guarantee optimal business outcomes[6]. The findings demonstrate that no single optimal algorithm can consistently outperform other algorithms[7].

This research's main goal is to improve the patient illness modeling system by strategically enhancing feature selection methods. In this intervention study, the most important indicators or traits influencing the prognosis of the disease are identified by means of active testing and use of sophisticated feature selection techniques. Our goal is to improve the quality of the system's analysis and its ability to anticipate diseases by carefully implementing these analytical techniques and selecting the most suitable attributes. It is anticipated that this integration will reduce computational load during data processing, speed up the model-building process, and ultimately increase the accuracy of disease detection rates and without requiring extra testing, the patterns found may aid in diagnosing problems and identifying diseases. This can facilitate better illness diagnosis in the future and assist physicians in making well-informed decisions[8, 9].

This research aims to pick features more effectively in order to improve the accuracy of patient disease prediction models. To do this, we use a variety of methodologies like variance computation, correlation analysis, and prediction models like Decision Trees and Logistic Regression[10]. We hope to simplify the model and increase the speed and precision of illness diagnosis by identifying the most pertinent elements. In order to guarantee the legitimacy and dependability of our research, patient data from other medical sources is being gathered for this project.

Materials and Methods

We implemented an experimental methodology for our work that was centered on developing and evaluating an ensemble predictive model. To guarantee excellent data quality, the procedure started with the careful collection of data and continued with rigorous pre-processing. The features were chosen according to their correlations, and then Principal Component Analysis (PCA) was used to decrease them in order to improve computing efficiency and minimize dimensionality. An ensemble method was used to build our prediction model, and its parameters were carefully adjusted to get the best possible results. A preset dataset was used for the model's evaluation, and established metrics were applied. This study's objective was to evaluate the predictive model's accuracy in estimating the given target variable while adjusting for the data's limitations and the methods used. By focusing on the most relevant information, our method reduces the number of components required

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

for a functional network traffic analysis, hence simplifying the model. This simplified feature selection improves detection performance by reducing the complexity of the model and the time required to build the classifier model. Our system is made to function with an effective classifier by making the most of its attributes[11]. The results of our research experiments are shown in Fig. 1 below:

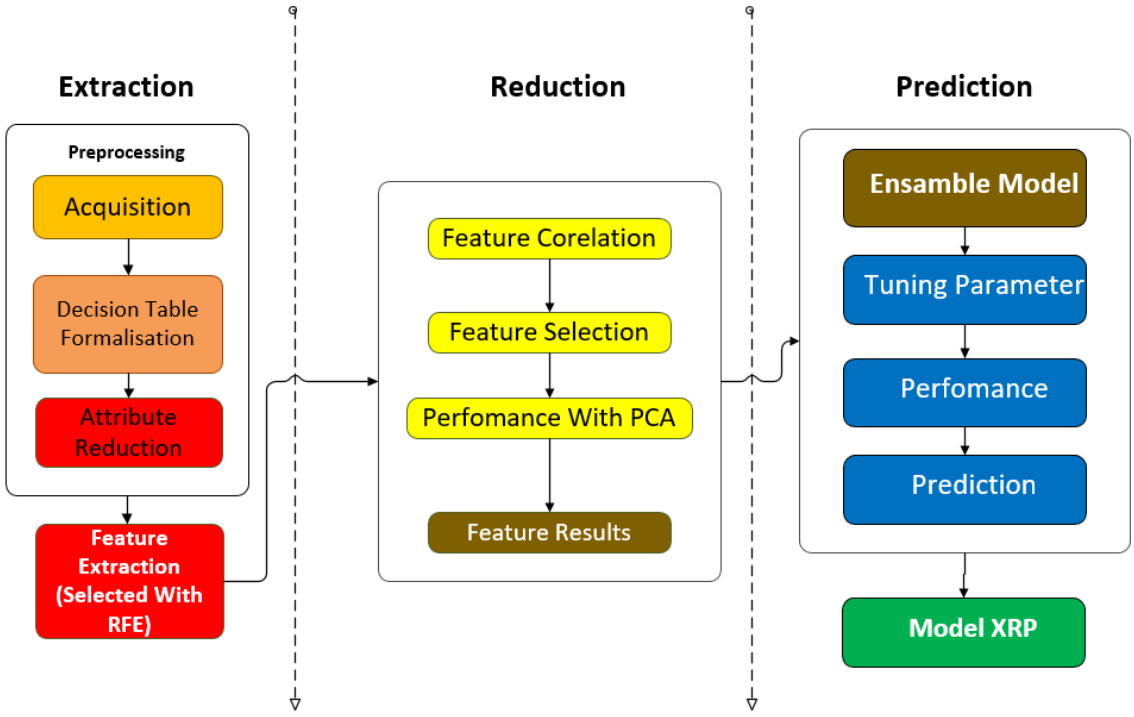


Figure 1. Developed Experiment Flow Model Diagram (XRP Model).

Based on Fig. 1 provided, the study design appears to be a structured approach to data analysis with the goal of disease prediction. The design is divided into three main stages: Extraction, Reduction, and Prediction. Here's a breakdown of each stage:

1. **Extraction**:

Pre-processing: Involves initial data acquisition and preparation.

- Acquisition: The first step where data is collected.
- Decision Table Formalisation: Likely involves structuring the collected data into a format suitable for analysis (e.g., a decision table).
- Attribute Reduction: Reducing the number of data attributes to those most relevant.
- Feature Extraction (Selected with RFE): Extracting features from the data using Recursive Feature Elimination (RFE), which is a method to select features by recursively considering smaller and smaller sets of features.

2. **Reduction**:

This stage seems focused on identifying the most significant features from the data.

- Feature Correlation: Assessing the interdependencies between different features.
- Feature Selection: Selecting the most important features to include in the predictive model.
- Performance with PCA: Possibly evaluating the performance of the feature selection by using Principal Component Analysis (PCA) to reduce dimensionality and highlight variation.
- Feature Results: The outcome of the feature selection and reduction process.

3. **Prediction**:

This is the stage where the actual prediction model is built and evaluated.

- Ensemble Model: Utilizing multiple models to make a prediction, likely to improve the robustness and accuracy of the results.
- Tuning Parameter: Adjusting the parameters of the model(s) for better performance.

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

- Performance: Assessing the performance of the model(s), probably through metrics like accuracy, precision, recall, etc.
- Prediction: The final step where the model is used to make predictions.
- Model XRP: specific model used or developed in research

Results and Discussion

Using advanced analytic methods to forecast disease outcomes has become more and more important in the field of medical research. In order to improve disease forecast accuracy, this study presents a sophisticated method that uses a stacking ensemble model that integrates various predictive algorithms. With a careful blend of DecisionTreeClassifier and Logistic Regression as base estimators and Logistic Regression as the meta-classifier, we hope to build a strong model that can provide accurate predictions. We aim to demonstrate the efficacy of the stacking method in enhancing disease prediction accuracy through a methodical process that involves data preparation, model training, and performance evaluation. This introduction provides the framework for a detailed analysis of our study procedures and findings, which are outlined in the ensuing sections.

1. Import essential libraries for the research experiment, including pandas for data manipulation, StackingClassifier for ensemble model construction, base estimators LogisticRegression and DecisionTreeClassifier for the foundational models, and train_test_split and accuracy_score for evaluating the performance of the models.
2. Load the dataset for the disease from a CSV file into a pandas DataFrame to facilitate data handling and manipulation.
3. Extract the feature set (independent variables) and the target variable (dependent variable, 'prognosis') from the dataset to prepare for model training and testing.
4. Utilize the train_test_split function to divide the dataset into a training set (80%) for model fitting and a testing set (20%) to evaluate the model's performance.
5. Define the base estimators: LogisticRegression and DecisionTreeClassifier that will serve as the foundational predictive models within the stacking framework.
6. Train the base estimators on the training data using their respective fitting methods to ensure they are ready to generate predictions.
7. Generate and store the predictions from the base estimators on the test data, ensuring each estimator's outputs are captured for further analysis.
8. Aggregate the predictions from the base estimators into a new feature set, which will serve as input for the meta-classifier, showcasing an innovative approach to model enhancement.
9. Establish a meta-classifier another instance of LogisticRegression—which will act as the final decision-maker in the stacking ensemble.
10. Train the meta-classifier on the aggregated predictions, thus integrating the base estimators' insights into a singular predictive model.
11. Utilize the meta-classifier to predict disease outcomes based on the test data's aggregated features, thereby demonstrating the efficacy of the stacking method.
12. Evaluate the accuracy of the ensemble model by comparing its predictions to the actual disease outcomes, using the accuracy score as a metric of success.
13. Present the model's accuracy on the screen, thereby concluding the experimental phase and highlighting the research's tangible outcomes.

A.Diabetes Disease Data Trial

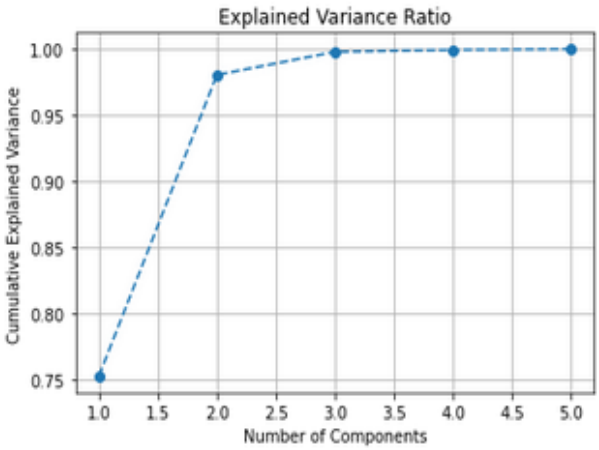
The data above are a feature preview; some are displayed with eight features and one label with 100,000 records. In this extraction step, we will perform data pre-processing because several features are of the object type. Then, the data is changed to a type integer for gender and smoking history features. The dataset that has the same type (integer) is then followed by extracting the features into the desired feature selection using Recursive Feature Elimination (RFE) by selecting through the five estimators, and the results are stored in a new file for each estimator with the specified features including 5 features, 6 features, and 7 features. Table 1. shows the best accuracy after feature selection, found in 6 features using the Random Forest technique.

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 1. Accuracy Results of Selected Diabetes Features.

No	Model / Estimator	Feature	Training Accuracy (%)	Testing Accuracy (%)
1	Logistic Regression	5	93,78	93,66
		6	94,22	94,08
		7	95,96	96,08
		8	**96,04**	**96,19**
2	Decission Tree	5	96,00	96,01
		6	95,99	96,12
		7	96,03	96,12
		8	**96,04**	**96,19**
3	Random Forest	5	96,00	96,10
		6	**96,04**	**96,15**
		7	96,04	96,14
		8	96,04	96,19
4	SVM	5	94,22	94,09
		6	94,22	94,08
		7	94,30	94,22
		8	**96,04**	**96,19**
5	Gradient Boosting	5	96,03	96,09
		6	96,04	94,10
		7	96,04	95,80
		8	**96,04**	**96,19**

The best accuracy data is 96% training data and 96% test data and using a PCA of 95%, this data was chosen because it has the highest accuracy among others and is close to the accuracy value with 8 features. This selected dataset is used as a result of the reduction. The Improvement observed in Fig. 2. indicates that each additional component significantly contributes to explaining the variation in the data.



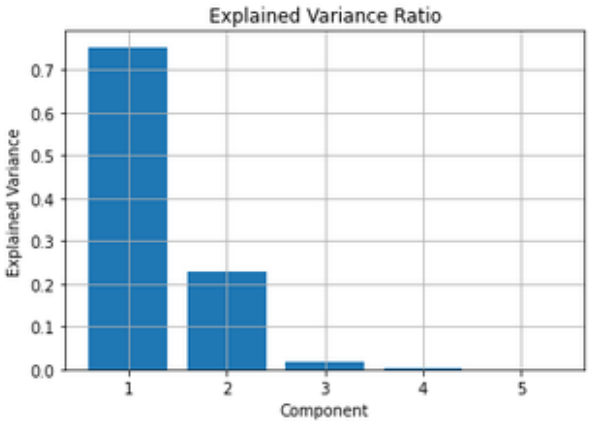


Figure 2. Plot Explained Variance, and Histogram Explained Variance Diabetes.

The ROC graph in Fig. 3. with an AUC of 0.86 shows that the classification model distinguishes between positive and negative classes well. The higher the AUC value, the better the mode performance.

In this step, training and testing of selected features are carried out using the ensemble model algorithm by trying 4 techniques: Random Forest, Adaboost, Stacking, and Bagging; the following accuracy is obtained: For the assessment of machine learning models, our findings demonstrate impressive performance measures. With an accuracy rate of 97.25%, Adaboost surpassed other models, with Random Forest coming in second with an accuracy rate of 97.24%. Notably, stacking performed competitively with an accuracy rate of 96.39%, while

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

bagging showed good predictive power with an accuracy score of 96.84%. Data from the training results and prediction testing above shows an increase in accuracy carried out with the ensemble technique above, where previously assessed accuracy obtained a final accuracy of 0.9724, so this model can be stored and continued and utilized for the implementation stage.

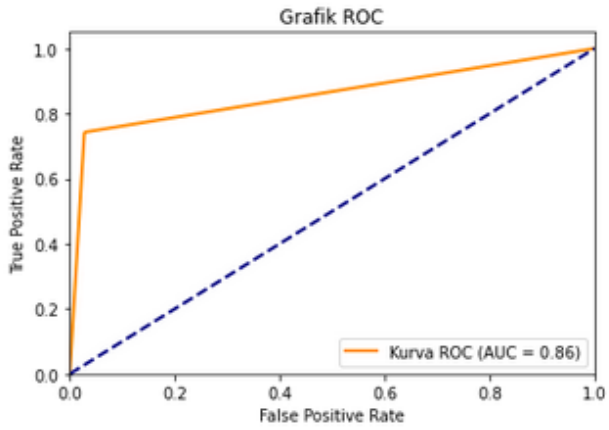


Figure 3. Graphic ROC.

The confusion matrix in Fig. 4. shows that the classification model succeeded in identifying 1,181 cases as positive which were truly positive (True Positives) and 18,212 cases as negative which were truly negative (True Negatives), while the model also made an error by classifying 80 negative cases as positive (False Positives) and did not recognize 527 positive cases as negative (False Negatives), which indicates that the model has room for improvement, especially in reducing the number of False Negatives to increase its sensitivity.

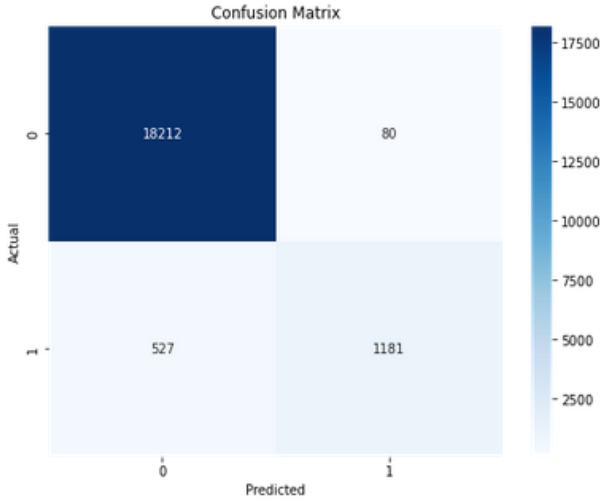


Figure 4. Confusion Matrix.

Fig. 5. shows that a Precision value of 0.954218 means that around 95.42% of the results classified as positive by the model are true positives, while the rest may be false positives. A recall value of 0.843539 means the model can find and classify around 84.35% of all positive samples in the dataset. The F1-Score value of 0.889581 indicates a good balance between Precision and Recall.

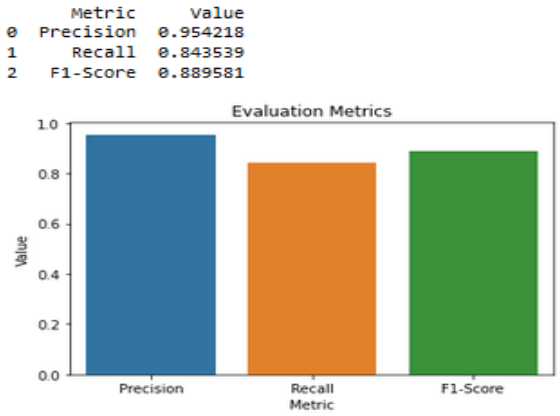


Figure 5. Evaluation Matrix.

Table 2. is comparative data on the accuracy of diabetes prediction, which was then carried out by applying the XRP model to 96.8%

Table 2. Model Comparison Dataset Diabetes.

Dataset	Accurate
https://www.kaggle.com/code/enochadjei/diabetes-analysis-and-predictions	95,8%
https://www.kaggle.com/code/gabrielfachet i/diabetes-prediction-eda-votingclassifier,	92%

B.Heart Disease Data Trial

This test phase attempts to predict heart disease, with a total of 13 features, which are then carried out using the XRP, and then proceed with extracting the features into the desired feature selection using Recursive Feature Elimination (RFE) by selecting through the five estimators. The results are stored in a new file for each estimator with the specified features, including 9 features, 10 features, and 11 features.

a) Reduction

Table 3. shows the best accuracy after feature selection and is carried out in 11 features using the SVM technique. The best accuracy data using PCA is 0.9259; this data was chosen because it has the highest accuracy among the others and has the

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

highest accuracy of the 11 features. This selected dataset is used as a result of reduction.

Table 3. Accuracy Results of Selected Heart Features.

No	Model / Estimator	Feature	Training Accuracy (%)	Testing Accuracy (%)
1	Logistic Regression	9	85,64	83,33
		10	86,11	83,33
		11	86,11	81,48
		12	**87,50**	**77,77**
2	Decission Tree	9	86,11	79,62
		10	87,03	83,33
		11	87,96	81,48
		12	**87,50**	**77,77**
3	Random Forest	9	85,64	85,18
		10	87,03	83,33
		11	87,96	81,48
		12	**87,50**	**77,77**
4	SVM	9	85,64	83,33
		10	86,11	83,33
		11	**87,96**	**81,48**
		12	**87,50**	**77,77**
5	Gradient Boosting	9	86,11	81,48
		10	87,03	83,33
		11	87,96	81,48
		12	**87,50**	**77,77**

The Improvement observed in Fig. 6. indicates that each additional component significantly contributes to explaining the variation in the data.

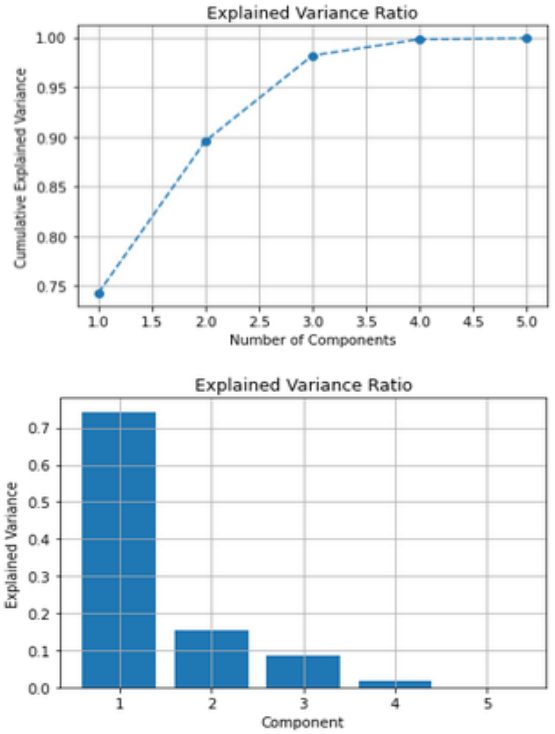


Figure 6. Explained Variance and Histogram Explained Variance Heart Disease.

b) Prediction

In this step, training and testing of selected features are carried out using the ensemble model algorithm by trying 4 techniques: Random Forest, Adaboost, Stacking, and Bagging; the following accuracy is obtained:

In the analysis of machine learning model performance, Random Forest achieved an accuracy rate of 87.03%, establishing itself as a strong contender. Stacking also demonstrated impressive results with an accuracy of 88.88%. The accuracy rates of 83.33% for both Adaboost and Bagging, however, showed similar performance.

Fig. 7, with a Precision value of 0.800987 indicates that approximately 80.1% of the results classified as positive by the model are true positives, while the rest may be false positives. A recall value of 0.764069 indicates that the model can find and classify around 76.4% of all positive samples in the dataset. The F1-Score value of 0.773887 indicates a good balance between Precision and Recall. These results indicate that the model accurately identifies true positives, can find most of the positive samples, and achieves a good balance between Precision and Recall. The higher the value of Precision, Recall, and F1-Score.

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

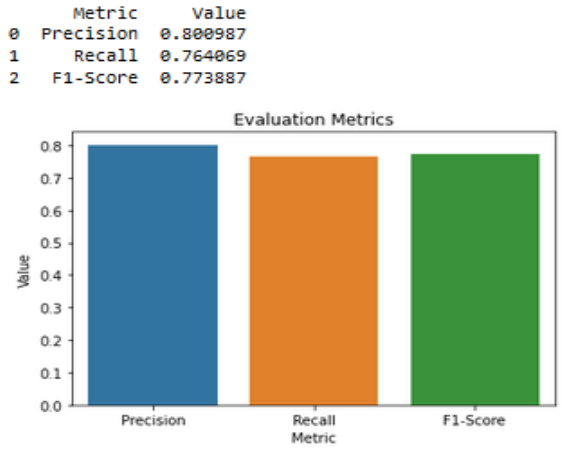


Figure 7. Evaluation Matrix.

Table 4. Comparative data on the accuracy of heart prediction, which was then carried out by applying the XRP model to 88%

Table 4. Model Comparison Dataset Heart.

Dataset	Accurate
https://www.kaggle.com/code/utkarshx27/heart-disease-prediction-rf-model-85-acc/notebook	85%
.https://www.kaggle.com/code/bulentsiyah/heart-disease-prediction-using-neural-networks	83%

The XRP model has the greatest accuracy rating, 88%, according to Table 5. which contains research data from earlier studies with many comparisons.

Table 5. Model Comparison Research.

Classification	Feature Selection	Dataset	Accuracy	Ref
Naïve Bayes	Logistic Regression	UCI	85.24 %	[12]
SVM	MRMR	UCI	84.85 %	[13]
Logistic Regression	Logistic Regression	Kaggle	78.66 %	[14]
Decision tree	Gain ratio decision tree	NM*	85%	[15]
Random Forest	Random Forest	Kaggle	77.25 %	[16]
cK-NN	-	UCI	79.9%	[17]
Neural Network	-	Cleveland	87.70 %	[18]
SVM with boosting	-	Cleveland	84.81 %	[19]
Ensemble based on distances for K-NN.	-	Cleveland	84.83 %	[20]
Deep Neural Network	-	Cleveland	83.67 %	[21]
Recursive Feature Elimination (RFE).	**SVM + Stacking for Prediction**	**Kaggle**	**88%**	**Model XRP**

Conclusion

The developed model, identified as the XRP Model, is based on an evaluation pattern that consists of prediction, reduction, and extraction. It has been shown to be highly accurate in predicting diseases, and the chosen traits are essential to obtaining precise label predictions. This model has undergone several rounds of development and has proven successful not just in early trials but also in increasing the accuracy of disease prediction when tested on other cases with different target data and labels. We recognize that healthcare providers in other countries may not be familiar with models like XRP. Therefore, we recommend introducing these models in stages, starting with workshops and training to increase understanding of the prediction models. This requires offering technical assistance for early implementation in addition to specific training on how the model can be implemented into currently operating healthcare systems. If used correctly, the XRP Model can improve the efficacy and efficiency of medical diagnosis and treatment in any country, empowering medical professionals to make more informed choices.

Acknowledgement

The cooperation of the RSUP M. Djamil Padang Hospital is appreciated

Authors' Declaration

- Conflicts of Interest: None.

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

- We hereby confirm that all the Ffigures and tables in the manuscript are ours. Besides, the figures and images, which are not ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at Putra Indonesia University "YPTK" Padang
- Ethics Approval: Health Research Ethics Committee RSUP Dr. M. Djamil Padang, West Sumatera Indonesia has approved this research with protocol number: LB.02.02/5.7/400/2023. Each participant was given written information about the purpose of the study before the study began, and written informed consent was obtained
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- No potentially identified images or data are present in the manuscript.

Authors' Contribution Statement

H.N.P conceived of the presented idea, developed the theory and performed the computations and then implemented the practical side. S.D supervised the findings of this work. G.W.N verified the analytical methods. H.N.P, S.D and G.W.N participated in drafting and revision of the manuscript. All authors read and approved the final manuscript.

References

1. Basil NN, Ambe S, Ekhator C, Fonkem E. Health records database and inherent security concerns: A review of the literature. Cureus. 2022; 14(10): e30168. https://doi.org/10.7759/cureus.30168
2. Farooqui ME, Ahmad DJ. A detailed review on disease prediction models that uses machine learning. Int J Innov Res Comput Sci Technol. 2020; 8(4): 326-330. https://doi.org/10.21276/ijircst.2020.8.4.14
3. Joseph N, Lindblad I, Zaker S, Elfversson S, Albinzon M, Hantler L, et al. Automated data extraction of electronic medical records: Validity of data mining to construct research databases for eligibility in gastroenterological clinical trials. Ups J Med Sci. 2022; 127(1). https://doi.org/10.48101/ujms.v127.8260
4. Alanazi R. Identification and prediction of chronic diseases using machine learning approach. J Healthc Eng. 2022; 2826127. https://doi.org/10.1155/2022/2826127
5. Ghaffar Nia N, Kaplanoglu E, Nasab A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. Discov Artif Intell. 2023; 3: 5. https://doi.org/10.1007/s44163-023-00049-5
6. Qu K, Wang L. Research on visual data mining technology. J Phys Conf Ser. 2021; 1748(3). https://doi.org/10.1088/1742-6596/1748/3/032056
7. Miao C, An TS. Application of data mining techniques on tourist expenses in Malaysia. Baghdad Sci J. 2021; 18: 737-745. https://doi.org/10.21123/bsj.2021.18.1(Suppl.).0737
8. Mahmood RAR, Abdi AH, Hussin M. Performance evaluation of intrusion detection system using selected features and machine learning classifiers. Baghdad Sci J. 2021; 18: 884-898. https://doi.org/10.21123/bsj.2021.18.2(Suppl.).0884
9. Sameer S, Behadili SF. Data mining techniques for Iraqi biochemical dataset analysis. Baghdad Sci J. 2022; 19(2): 385-398. https://doi.org/10.21123/bsj.2022.19.2.0385
10. Morales A, Villalobos FJ. Using machine learning for crop yield prediction in the past or the future. Front Plant Sci. 2023; 14: 1-13. https://doi.org/10.3389/fpls.2023.1128388
11. Hu F, Situo Z, Xubin L, Liu W, Niandong L, Yanqi S, et al. Network traffic classification model based on attention mechanism and spatiotemporal features. Eurasip J Inf Secur. 2023; 1: 6. https://doi.org/10.1186/s13635-023-00141-4
12. Julian A, Deepika R, Geetha B, Sweety VJ. Heart disease prediction using machine learning. In: Artificial Intelligence, Blockchain and Computing Security. 2023; 2: 248-253. https://doi.org/10.1201/9781032684994-38
13. Bashir S. Improving heart disease prediction using feature selection approaches. In: Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST). 2019; 619-623. https://doi.org/10.1109/IBCAST.2019.8667106
14. Hoque S, Khatun SS, Khurshid AB, Peal MD, Salam KMA. Prediction of heart disease using machine learning. In: 2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC). 2022; 471–476. https://doi.org/10.1109/ICMACC54824.2022.10093246
15. Zeniarja J, Ukhifahdhina A, Salam A. Diagnosis of

Published Online First: July, 2024
https://doi.org/10.21123/bsj.2024.9695
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

heart disease using K-nearest neighbor method based on forward selection. J Appl Intell Syst. 2020; 4(2): 39–47. https://doi.org/10.33633/jais.v4i2.2749
16. Pemmaraju AG, Asish A, Das S. Heart disease prediction using feature selection and machine learning techniques. In: 2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS). 2022; 28–33. https://doi.org/10.1109/MLCSS57186.2022.00014
17. Pious IK, Antony Kumar K, Soulwin YC, Reddy EN. Heart disease prediction using machine learning algorithms. In: 2022 International Conference on Innovative Computing. Intelligent Communication and Smart Electrical Systems (ICSES). 2022; 1–6. https://doi.org/10.1109/ICSES55317.2022.9914207
18. Modak S, Abdel-Raheem A, Rueda E. "Heart Disease Prediction Using Adaptive Infinite Feature Selection and Deep Neural Networks," *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC).* 2022; 235-240. https://doi.org/10.1109/ICAIIC54071.2022.9722652
19. Gupta A, Yadav S, Shahid S, Venkanna U. Heart Care: IoT based heart disease prediction system. In: International Conference on Information Technology (ICIT). 2019; 88-93 https://doi.org/10.1109/ICIT48102.2019.00022
20. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inform Med Unlocked. 2019; 16: 100203. https://doi.org/10.1016/j.imu.2019.100203
21. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019; 7: 81542-81554. https://doi.org/10.1109/ACCESS.2019.2923707

تطوير التعلم الآلي المختلط في تصنيف تشخيص المرضى باستخدام نموذج XRP (الاستخراج والاختزال والتنبؤ)

هيندرا نوسا بوترا[1]، سارجون ديفيت[2]، جونادي ويدي نوركاهيو[2]

[1] قسم السجلات الطبية، ستيكس دارما لاندبو، بادانج، إندونيسيا.
[2] قسم تكنولوجيا المعلومات ،كلية علوم الحاسب، UPI YPTK، بادانج، إندونيسيا.

الخلاصة

تستكشف هذه الدراسة، التي أجريت على مدار ستة أشهر في أحد المستشفيات الإندونيسية، فوائد توحيد بيانات السجلات الطبية وتكامل أنظمة المعلومات الصحية لتقديم الرعاية الصحية باستخدام نهج البحث الكمي، فإنه يركز على تأثير استخراج البيانات الدقيقة على تحليل البيانات ومزايا النظام المتكامل للوصول إلى سجلات المرضى. تم استخدام طرق متقدمة لاستخراج البيانات لاستخراج الميزات واختيارها وتقليل مجموعة البيانات لتعزيز دقة تصنيف البيانات. وكشفت النتائج عن وجود علاقة مباشرة بين دقة استخراج البيانات وموثوقية تصنيف البيانات، مما يسلط الضوء على الدور الهام للحد من مجموعة البيانات في تحسين دقة التحليل. كان إدخال نموذج XRP، وهو أداة تنبؤية جديدة لتقييم احتمالية الإصابة بالأمراض، بمثابة تقدم ملحوظ، حيث أظهر معدلات دقة عالية في التنبؤ بمرض السكري وأمراض القلب (96.8% و88% على التوالي). يؤكد الأداء المتسق للنموذج عبر سيناريوهات النتائج المختلفة على إمكاناته في اتخاذ القرارات في مجال الرعاية الصحية. يوضح هذا البحث قيمة التنقيب المتقدم في البيانات وتقليل مجموعة البيانات في تحسين تصنيف البيانات، وبالتالي تسهيل اتخاذ قرارات أفضل في مجال الرعاية الصحية. يشير نجاح نموذج XRP في التنبؤ بالأمراض إلى فوائد كبيرة لخدمات الرعاية الصحية، حيث يقدم رؤى حاسمة لتطوير وتحسين أنظمة المعلومات الصحية. هذه النتائج لديها القدرة على التأثير على سياسة وممارسات الرعاية الصحية، والدعوة إلى معيار جديد في إدارة بيانات الرعاية الصحية.

الكلمات المفتاحية: التنبؤ بالمرض، تقليل الميزات، اختيار الميزات، التعلم الآلي، السجل الطبي.