IRAQI
Academic Scientific Journals

University of Kirkuk Journal For Administrative and Economic Science (2024) 14 (3): 282-291

UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN 2222-2995
University of Kirkuk Journal For Administrative and Economic Science

Faraj Shahen Mohammed. Performance Evaluation of Machine Learning Algorithms for Predictive Classification of Anemia Data. *University of Kirkuk Journal For Administrative and Economic Science* (2024) 14 (3):282-291.

Performance Evaluation of Machine Learning Algorithms for Predictive Classification of Anemia Data

Shahen Mohammed Faraj [1]

[1] *Department of International Trade, College of Law and Administrative, University of Halabja, Halabja, Iraq*

shahen.faraj@uoh.edu.iq [1]

Abstract: Anemia is a prevalent condition in women, often due to iron deficiency, pregnancy, or menstrual blood loss. Early detection is crucial for effective treatment and management. The objective is to evaluate and compare various machine-learning algorithms to determine which ones are most effective for predictive classification tasks in the context of medical data. The dataset consists of records from 180 women, randomly selected from several hospitals and clinics in Sulaimani City, to investigate anemia. The algorithms used for comparison of Bayes Net (BN), Random Forest (RF), Random Tree (RT), Logistic Model Tree (LMT), trees J48 and Reduced Error Pruning (REP) Tree are supervised learning algorithms used extensively for classification purposes and are chosen for their variety. The performance of these algorithms is compared using various performance criteria based on the confusion matrix, such as Accuracy, Recall/Sensitivity, Specificity, True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, F-Score, Matthews Correlation Coefficient (MCC), and the Receiver Operating Characteristic (ROC). Based on the analysis of this data, the Logistic Model Tree (LMT) demonstrates superior classification performance, achieving 97.78% in Accuracy (Correctly Classified), 94.3% in MCC, and 99.6% in ROC compared to other classifiers.

Keywords: Bayes Net, Random Forest, Reduced Error Pruning, Logistic Model Tree, Reduced Error Pruning.

تقييم أداء خوارزميات تعلم الآلة للتصنيف التنبؤي لبيانات فقر الدم

م. شاهين محمد فرج [1]

[1] *قسم التجارة الدولية، كلية القانون والإدارية، جامعة حلبجة، حلبجة، العراق*

المستخلص: فقر الدم هو حالة شائعة بين النساء، وغالبًا ما يكون بسبب نقص الحديد أو الحمل أو فقدان الدم أثناء الدورة الشهرية. الكشف المبكر أمر بالغ الأهمية للعلاج والإدارة الفعالة. والهدف هو تقييم ومقارنة خوارزميات التعلم الآلي المختلفة لتحديد أيها الأكثر فعالية لمهام التصنيف التنبؤي في سياق البيانات الطبية. تتكون مجموعة البيانات من سجلات من ١٨٠ امرأة، تم اختيارهن عشوائيًا من مدينة السليمانية، للتحقيق في فقر الدم. الخوارزميات المستخدمة لمقارنة Bayes Net (BN) وRandom Forest (RF) وRandom Tree (RT) و Logistic Model Tree (LMT) والأشجار J48 وReduced Error Pruning (REP) Tree هي خوارزميات تعلم

خاضعة للإشراف تستخدم على نطاق واسع لأغراض التصنيف ويتم اختيارها لتنوعها. تتم مقارنة أداء هذه الخوارزميات باستخدام معايير أداء مختلفة تستند إلى مصفوفة الارتباك، مثل الدقة، والتذكير/الحساسية، والخصوصية، ومعدل الإيجابية الحقيقية (معدل TP)، ومعدل الإيجابية الكاذبة (معدل FP)، والدقة، ودرجة F، ومعامل ارتباط ماثيوز (MCC)، وخاصية تشغيل المستقبل (ROC). بناءً على تحليل هذه البيانات، تُظهر شجرة النموذج اللوجستي (LMT) أداء تصنيف متفوقًا، حيث حققت ٩٧,٧٨٪ في الدقة (التصنيف الصحيح)، و٩٤,٣٪ في MCC، و٩٩,٦٪ في ROC مقارنة بالمصنفات الأخرى.

الكلمات المفتاحية: شبكة بايز، الغابة العشوائية، تقليم الأخطاء المخفضة، شجرة النموذج اللوجستي، تقليم الأخطاء المخفضة.

Corresponding Author: E-mail: shahen.faraj@uoh.edu.iq

Introduction

Anemia is a widespread health issue that disproportionately affects women due to various physiological and nutritional factors such as menstruation, pregnancy, and iron deficiency. This condition can lead to severe fatigue, weakness, and other health complications, making early and accurate diagnosis essential for effective treatment and management. With the advancement of technology, machine-learning algorithms have emerged as powerful tools for predictive analytics in healthcare. Moreover, Machine-learning algorithms are quintessential aspects of synthetic Genius that enable systems to study from data, discover patterns, and make selections autonomously. By processing giant datasets, these algorithms improve their overall performance over time without explicit programming instructions. This capability has revolutionized industries ranging from finance and healthcare to entertainment and transportation. Additionally, Machine-learning algorithms are widely classified into supervised, unsupervised, and semi-supervised learning. Supervised getting to know uses labeled data to train fashions for duties like regression and classification. Unsupervised getting to know identifies patterns in unlabeled facts thru clustering or affiliation techniques. Semi-supervised getting to know combines labeled and unlabeled data to beautify mannequin performance. Then, Machine-learning algorithms are extensively applied throughout more than a few industries. In healthcare, they are used for diagnosing diseases and predicting patient outcomes. In finance, these algorithms detect fraud, predict stock market trends, and assess risks. Marketing utilizes them for customer segmentation and personalized recommendations. Technological applications include speech recognition, image classification, and the development of autonomous vehicles. These algorithms continue to evolve alongside advancements in data collection, computational power, and algorithmic techniques, fostering innovation and catalyzing transformations across global industries.

1. Objective of the Research

This study aims to conduct a comparative analysis of various machine-learning algorithms to determine their effectiveness in the predictive classification of anemia in women. By leveraging medical data, the dataset comprises records from 180 women, randomly selected from Sulaimani City, to study anemia. In this study, six different machine-learning techniques were used such as Bayes Net (BN), Random Forest (RF), Random Tree (RT), Logistic Model Tree (LMT), trees J48 and Reduced Error Pruning (REP) Tree.

2. Materials and Methods

There are a number of classification algorithms that has (Bayes Net (BN), Random Forest (RF), Random Tree (RT), Logistic Model Tree (LMT), bushes J48 and Reduced Error Pruning (REP) Tree) classification purposes and investigated in anemia data. The working machine is proven in Figure 1. Classification algorithm used to be as soon as chosen to stumble on the most suitable one for predicting anemia.

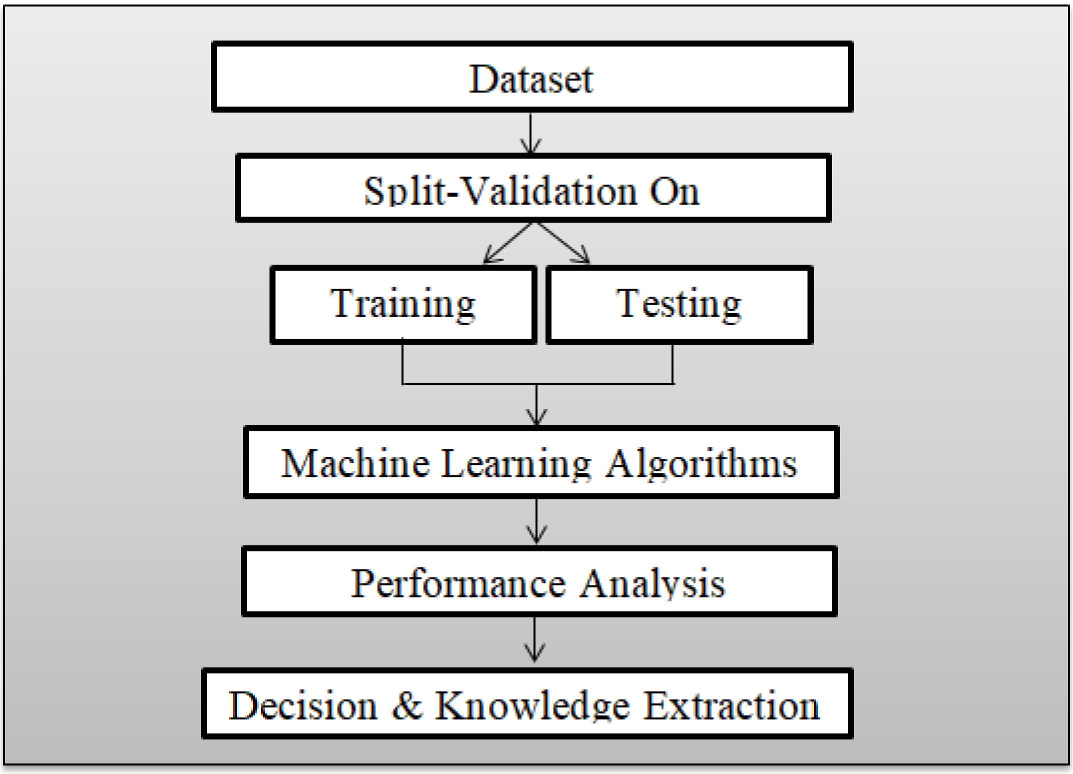

Figure (1): working process

In this lookup work, the anemia facts of female is to be classified using often used forty-machine mastering algorithms are given below:

A. Bayes Net (BN)

Bayesian networks, or Bayes Nets, are graphical models that depict the probabilistic relationships among a collection of random variables. These networks are represented as annotated Directed Acyclic Graphs (DAGs) that encapsulate a joint probability distribution. Each node in the graph corresponds to a random variable, and the directed edges between nodes represent conditional dependencies. One key interpretation of Bayesian networks is that they provide a compact, intuitive way to model complex systems and reason about uncertainty. By breaking down the joint distribution into a series of local distributions, Bayesian networks simplify the computation of probabilities and allow for efficient inference, making them powerful tools in fields such as machine learning, artificial intelligence, and statistics.

B. Random Forest (RF)

Random Forest (RF) is an ensemble learning method composed of multiple decision trees that collectively determine the output class, which is the mode of the classes predicted by individual trees. Unlike single decision trees, random forests generate numerous classification trees without pruning, enhancing their robustness and accuracy. This technique leverages the diversity of the trees to mitigate over fitting and improve predictive performance. Random forests are highly effective in handling large datasets with high dimensionality and can manage missing values and maintain accuracy for a significant portion of the data, making them valuable for a wide range of classification tasks.

C. Random Tree (RT)

Random Tree (RT) is a versatile and powerful machine-learning algorithm used for both classification and regression tasks. Unlike traditional decision trees, which are prone to over fitting, random trees introduce randomness in the selection of features and splits during the tree-building process. This results in a more robust and generalized model. Each random tree is constructed by selecting a random subset of features at each node and using these features to determine the best split. This randomness helps in capturing a diverse range of patterns within the data, leading to improved performance, especially on complex datasets. Random trees are the building blocks of more sophisticated ensemble methods, such as Random Forests, where multiple random trees are aggregated to enhance predictive accuracy. Due to their simplicity, interpretability, and ability to handle large datasets with high dimensionality, random trees are widely used in various fields, including finance, healthcare, and bioinformatics

D. Logistic Model Tree (LMT)

Logistic Model Tree (LMT) is a hybrid machine-learning algorithm that combines the interpretability of decision trees with the predictive power of logistic regression. This model builds a decision tree where each leaf node contains a logistic regression model, enabling it to handle both categorical and continuous variables effectively. The structure of the LMT allows for the partitioning of the data into smaller, more manageable subsets, while the logistic regression at the leaves ensures robust prediction within each subset. This approach enhances the model's flexibility and accuracy, making it suitable for complex classification tasks. LMTs are particularly useful in scenarios requiring both high interpretability and strong predictive performance, such as in medical diagnosis and financial forecasting.

E. Trees J48

J48 is a famous machine-learning algorithm used for classification tasks, derived from the C4.5 algorithm. It constructs decision trees with the aid of recursively splitting the data based totally on the attribute that affords the perfect facts gain. Each interior node in the tree represents a choice primarily based on an attribute, and every leaf node represents a classification label. J48 is valued for its simplicity, interpretability, and effectivity in managing both continuous and specific data. Additionally, it includes mechanisms for pruning, which assist in reducing over fitting and improving the model's generalization. Due to its robustness and ease of use, J48 is widely applied in a number of fields; such as statistics mining, pattern recognition, and bioinformatics, making it a versatile device for predictive modelling.

F. Reduced Error Pruning (REP) Tree

Reduced Error Pruning (REP) Tree is a decision tree algorithm that incorporates a pruning technique to enhance model accuracy and prevent over fitting. This method involves initially creating a fully-grown tree and then systematically removing branches that do not improve predictive performance based on a validation set. By evaluating the tree's performance on unseen data, REP ensures that only the most relevant branches are retained, thus simplifying the model while maintaining its predictive power. This pruning process helps in producing a more generalizable and efficient model, making REP Trees particularly useful in scenarios where interpretability and accuracy are crucial. REP Trees are widely applied in fields such as finance, healthcare, and marketing for reliable and robust classification tasks.

3. Performance Evaluation

A confusion matrix, illustrated in Table 1, evaluates a classifier's performance. True positives (TP) denote accurately identified positive cases, whereas true negatives (TN) indicate correctly identified negative cases. False positives (FP) refer to negative instances incorrectly classified as positive, and false negatives (FN) are positive instances wrongly classified as negative. This matrix is essential for evaluating the classifier's accuracy and error distribution, which helps in identifying areas for improvement and refining model performance. By analysing these metrics, one gains insights into the classifier's strengths and weaknesses, guiding adjustments to enhance its predictive reliability.

Table (1): Confusion Matrix

Actual	**Predicted**	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

Table 2 summarizes the classifier evaluation metrics, including Classification Accuracy, Sensitivity, Specificity, TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, and ROC Area. These metrics provide a comprehensive assessment of the classifier's performance based on the contingency table. Among these, Classification Accuracy is a widely used measure that quantifies the proportion of instances correctly classified by the model. Sensitivity and Specificity evaluate the classifier's ability to identify positive and negative cases, respectively, while Precision, Recall, and F-Measure offer insights into the classifier's effectiveness in managing false positives and false negatives. The ROC Area and MCC further refine performance evaluation by measuring the model's overall ability to discriminate between classes and its correlation with actual outcomes.

Table (2): Detailed Accuracy by Classes

Tools	Statistic
Accuracy	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \times 100$
Sensitivity (Recall or True Positive Rate)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100$
Specificity (True Negative Rate)	$\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100$
TP Rate (True Positive Rate)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100$
FP Rate (False Positive Rate)	$\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \times 100$
Precision (Positive Predictive Value)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100$
Recall (Sensitivity or True Positive Rate)	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100$
F-Measure (F1 Score)	$\frac{2\,.\,(\mathrm{Precision}\,.\,\mathrm{Recall})}{\mathrm{Precision}+\mathrm{Recall}} \times 100$
Matthews Correlation Coefficient (MCC)	$\frac{\mathrm{TP}\,.\,\mathrm{TN}-\mathrm{FP}\,.\,\mathrm{FN}}{\sqrt{(TP+FP)(TP+FN)(TN+Fp)+(TN+FN)}} \times 100$
Receiver Operating Characteristic (ROC)	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}} \times 100$

4. Data Description and Analysis

The objective of this study is to conduct a parametric analysis of the obtained dataset using different machine learning algorithms to classify the data. The dataset consists of records from 180 women, randomly selected from several hospitals and clinics in Sulaimani City, to investigate anemia. The data set included the following variables: The first group represents explanatory variables: Age, Blood Pressure, Body Mass Index, Presence of Chronic Disease, MCH, MCHC, RBC, MCV, HCT, RDW%, S. Ferritin, and V. B12. The second group represents the response variable, which is Hemoglobin (HGB), with the code for HGB being: yes - anemia, and no - anemia. Machine learning algorithms were used in Weka.

A. Confusion Matrix for Machine learning algorithms

Table (3): Confusion matrix for machine learning algorithms

Bayes Net (BN)				Random Forest (RF)				Random Tree (RT)			
Predicted Class				Predicted Class				Predicted Class			
Actual Class	Yes	No	SUM	Actual Class	Yes	No	SUM	Actual Class	Yes	No	SUM
Yes	41	7	48	Yes	42	6	48	Yes	40	8	48
No	4	128	132	No	0	132	132	No	1	131	132
SUM	45	135	180	SUM	42	138	180	SUM	41	139	180
Logistic Model Tree (LMT)				Trees J48				Reduced Error Pruning (REP) Tree			
Predicted Class				Predicted Class				Predicted Class			
Actual Class	Yes	No	SUM	Actual Class	Yes	No	SUM	Actual Class	Yes	No	SUM
Yes	45	3	48	Yes	40	8	48	Yes	39	9	48
No	1	131	132	No	3	129	132	No	6	126	132
SUM	46	134	180	SUM	43	137	180	SUM	45	135	180

The confusion matrix presented in Table 3 provides a detailed comparison of the performance of various machine learning algorithms, including Bayes Net (BN), Random Forest (RF), Random Tree (RT), Logistic Model Tree (LMT), J48 trees, and Reduced Error Pruning (REP) Tree.
For Bayes Net (BN), the algorithm correctly classified 41 out of 48 positive cases (Yes) and 128 out of 132 negative cases (No), resulting in a total of 45 correctly classified positives and 135 negatives. This indicates an overall accuracy in predicting the outcomes.
Random Forest (RF) showed slightly better performance with 42 correct classifications for positives and 132 for negatives, achieving a similar overall accuracy but with a slightly different distribution of errors.
Random Tree (RT) also demonstrated competitive results with 40 correct positives and 131 negatives, suggesting it performs comparably to BN and RF.
The Logistic Model Tree (LMT) exhibited the highest accuracy, correctly classifying 45 positives and only 1 negative incorrectly, leading to the highest total correct classifications among all algorithms.
The J48 trees showed reasonable performance with 40 correct positives and 129 negatives, while the Reduced Error Pruning (REP) Tree algorithm classified 39 positives and 126 negatives correctly.
Overall, the Logistic Model Tree (LMT) demonstrates the highest accuracy at 97.78% in classifying both positive and negative cases, setting it apart from the other algorithms. Although other algorithms also perform competitively, LMT's superior accuracy highlights its effectiveness in predicting anemia status more reliably

B. Classification Accuracy, Sensitivity and Specificity of Proposed

Table (4): The Classification Accuracy, Sensitivity and Specificity of Classifier

Classifier	Sensitivity %	Specificity %	Accuracy %	Correctly Classified %		Incorrectly Classified %	
Bayes Net (BN)	85.42	96.97	93.89	169	93.89	11	6.11
Random Forest (RF)	87.50	100.00	96.67	174	96.67	6	3.33
Random Tree (RT)	83.33	99.24	95.00	171	95.00	9	5.00
Logistic Model Tree (LMT)	93.75	99.24	97.78	176	97.78	4	2.22
trees J48	83.33	97.73	93.89	169	93.89	11	6.11
Reduced Error Pruning (REP) Tree	81.25	95.45	91.67	165	91.67	15	8.33

Table 4 provides a comprehensive summary of the performance metrics for various classifiers, including Sensitivity, Specificity, Accuracy, Correctly Classified %, and Incorrectly Classified %.
The Logistic Model Tree (LMT) exhibits the highest accuracy at 97.78%, reflecting its exceptional ability to correctly classify both positive and negative cases. It also shows strong Sensitivity at 93.75% and high Specificity at 99.24%, indicating its effectiveness in identifying true positives while minimizing false positives and negatives. The LMT classifier correctly classified 176 out of 180 cases, with only 4 incorrectly classified, yielding an incorrect classification rate of 2.22%.
Random Forest (RF) follows closely with an accuracy of 96.67%, demonstrating excellent performance with a perfect Specificity of 100% and a Sensitivity of 87.50%. This classifier correctly classified 174 cases and had an incorrect classification rate of 3.33%.
The Random Tree (RT) achieved an accuracy of 95.00%, with a Sensitivity of 83.33% and a Specificity of 99.24%. It correctly classified 171 cases, resulting in a slightly higher incorrect classification rate of 5.00%.
Bayes Net (BN) and J48 trees both have accuracies of 93.89%, with BN showing a Sensitivity of 85.42% and a Specificity of 96.97%, while J48 trees have a Sensitivity of 83.33% and a Specificity of 97.73%. Both classifiers have relatively high correct classification rates (169 cases), but also higher incorrect classification rates of 6.11%.
The Reduced Error Pruning (REP) Tree has the lowest accuracy at 91.67%, with Sensitivity at 81.25% and Specificity at 95.45%. It correctly classified 165 cases, with an incorrect classification rate of 8.33%.
Overall, the Logistic Model Tree (LMT) stands out as the most accurate and reliable classifier, while other algorithms, though competitive, show slight variations in performance.

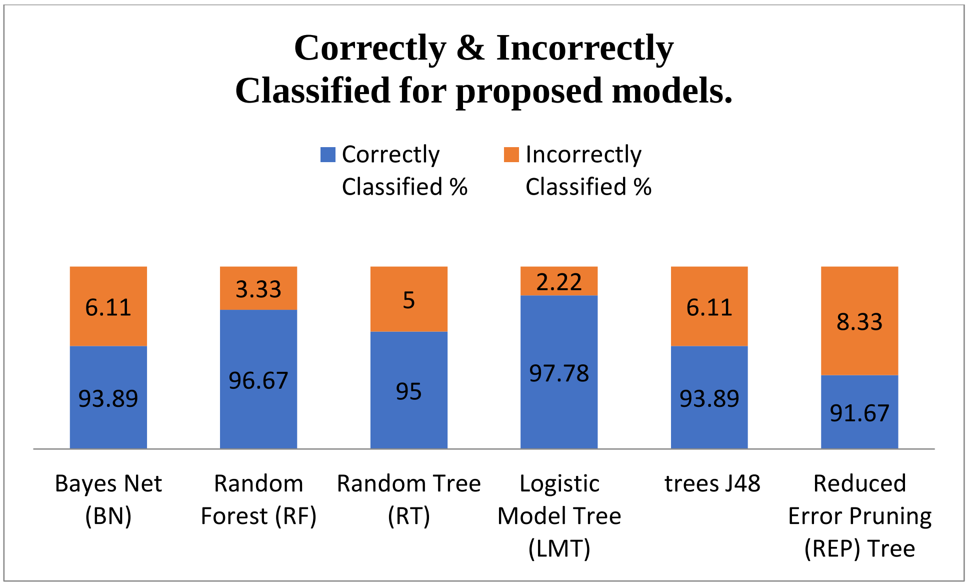

Figure (2): illustrate the difference in Accuracy, Sensitivity and Specificity of the proposed models.

C. Calculation Detailed Performance Metrics

Table (5): Detailed Performance Metrics by Class for Classifiers

Detailed Accuracy By Classes		TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
Bayes Net (BN)	Yes	0.854	0.03	0.911	0.854	0.882	0.841	0.963
	No	0.97	0.146	0.948	0.97	0.959	0.841	0.963
	Weighted Avg.	0.939	0.115	0.938	0.939	0.938	0.841	0.963
Random Forest (RF)	Yes	0.875	0	1	0.875	0.933	0.915	0.997
	No	1	0.125	0.957	1	0.978	0.915	0.997
	Weighted Avg.	0.967	0.092	0.968	0.967	0.966	0.915	0.997
Random Tree (RT)	Yes	0.833	0.008	0.976	0.833	0.899	0.871	0.913
	No	0.992	0.167	0.942	0.992	0.967	0.871	0.913
	Weighted Avg.	0.95	0.124	0.951	0.95	0.949	0.871	0.913
Logistic Model Tree (LMT)	Yes	0.938	0.008	0.978	0.938	0.957	0.943	0.996
	No	0.992	0.063	0.978	0.992	0.985	0.943	0.996
	Weighted Avg.	0.978	0.048	0.978	0.978	0.978	0.943	0.996
trees J48	Yes	0.833	0.023	0.93	0.833	0.879	0.841	0.947
	No	0.977	0.167	0.942	0.977	0.959	0.841	0.947
	Weighted Avg.	0.939	0.128	0.939	0.939	0.938	0.841	0.947
Reduced Error Pruning (REP) Tree	Yes	0.813	0.045	0.867	0.813	0.839	0.783	0.922
	No	0.955	0.188	0.933	0.955	0.944	0.783	0.922
	Weighted Avg.	0.917	0.15	0.916	0.917	0.916	0.783	0.922

Table 4 provides a detailed analysis of performance metrics for various classifiers, including Bayes Net (BN), Random Forest (RF), Random Tree (RT), Logistic Model Tree (LMT), J48 trees, and Reduced Error Pruning (REP) Tree. The metrics include True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, Recall, F-Measure, Matthews Correlation Coefficient (MCC), and ROC Area.

The Logistic Model Tree (LMT) exhibits the highest overall performance with a TP Rate of 93.8%, an FP Rate of 0.048, and an impressive ROC Area of 0.996. It achieves a Precision and Recall of 97.8% for both classes, resulting in a high F-Measure of 97.8% and a high MCC of 0.943. This indicates that LMT is highly effective in both identifying positive cases and avoiding false positives, leading to robust overall classification performance.
Random Forest (RF) follows closely, with a TP Rate of 96.7% and an FP Rate of 0.092. Its Precision and Recall are both 96.8%, and its F-Measure is 96.6%, demonstrating strong predictive capabilities. The ROC Area of 0.997 and MCC of 0.915 further highlight its high performance in distinguishing between classes.
Random Tree (RT) has a TP Rate of 95.0% and an FP Rate of 0.124. It achieves a Precision of 95.1% and Recall of 95.0%, with an F-Measure of 94.9% and an ROC Area of 0.913. While it is slightly less accurate than RF and LMT, it remains highly effective.
Bayes Net (BN) shows a TP Rate of 93.9% and an FP Rate of 0.115. Its Precision and Recall are 93.8% and 93.9%, respectively, with an F-Measure of 93.8% and an ROC Area of 0.963. Although its performance is slightly lower than RF and LMT, it still provides reliable classification.
J48 trees achieve a TP Rate of 93.9% with an FP Rate of 0.128. Its Precision is 93.9%, and its Recall is 93.9%, resulting in an F-Measure of 93.8% and an ROC Area of 0.947. The performance is competitive but not as high as LMT or RF.
The Reduced Error Pruning (REP) Tree has the lowest metrics, with a TP Rate of 91.7% and an FP Rate of 0.150. Its Precision and Recall are 91.6% and 91.7%, respectively, with an F-Measure of 91.6% and an ROC Area of 0.922. While it performs well, it lags behind the other classifiers.
Overall, the Logistic Model Tree (LMT) stands out with superior performance across most metrics, followed closely by Random Forest and Random Tree, while Bayes Net, J48 trees, and REP Tree offer competitive but slightly lower results.

5. Discussion

The comparative analysis of machine learning algorithms for classifying anemia data reveals notable differences in performance across various classifiers. The Logistic Model Tree (LMT) demonstrates the highest overall accuracy and robustness, with superior performance in True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, Recall, and ROC Area. Its ability to accurately classify both positive and negative cases with minimal false positives and negatives underscores its effectiveness in anemia classification. Random Forest (RF) also performs exceptionally well, showing high accuracy and strong class discrimination, although it slightly trails behind LMT in terms of TP Rate and Precision. Random Tree (RT) provides competitive results, demonstrating effective classification with slightly lower accuracy compared to RF and LMT. Bayes Net (BN) and J48 Trees offer reliable performance but does not outperformed by LMT and RF. The Reduced Error Pruning (REP) Tree, while still effective, has the lowest accuracy and highest rate of incorrect classifications among the classifiers evaluated.

Overall, these findings indicate that while LMT is the most accurate and reliable classifier, other algorithms perform well, with each offering specific strengths in different aspects of classification performance.

6. Conclusions and recommendations

A. Conclusion

The researcher has the following conclusions and recommendations based on the analysis:

1. The comparative analysis of machine learning classifiers for anemia prediction reveals distinct differences in performance metrics among various algorithms. The Logistic Model Tree (LMT) emerges as the top performer, highlighting the highest overall accuracy at 97.78%. It demonstrates superior classification abilities, with an impressive True Positive Rate (TP Rate) of 93.8%, a low False Positive Rate (FP Rate) of 0.048, and a high ROC Area of 0.996. This indicates LMT's exceptional effectiveness in correctly identifying positive and negative cases while minimizing errors.
2. Random Forest (RF) follows closely, achieving a high accuracy of 96.67% and showing strong performance metrics with a TP Rate of 96.7% and a perfect Specificity of 100%. Although slightly behind LMT, RF's strong Sensitivity and ROC Area highlight its robust predictive capabilities. Random Tree (RT) also performs well, with an accuracy of 95.00%, showing effective classification but slightly lower compared to LMT and RF.
3. Bayes Net (BN) and J48 Trees provide reliable results with accuracies of 93.89%, but they show higher incorrect classification rates compared to the top performers. The Reduced Error Pruning (REP) Tree, while still effective, has the lowest accuracy at 91.67% and the highest rate of incorrect classifications among the evaluated classifiers.

Over all, The Logistic Model Tree (LMT) excels with the highest accuracy of 97.78%, demonstrating superior classification performance. Random Forest (RF) and Random Tree (RT) follow, showing strong predictive capabilities. Bayes Net (BN) and J48 Trees are reliable but less accurate, while Reduced Error Pruning (REP) Tree has the lowest accuracy and highest error rate.

B. Recommendations

Based on the conclusions, the following are some recommendations for this study.

- Adopt LMT for anemia classification where high accuracy and reliable performance are crucial, as it demonstrates superior overall metrics.
- Consider RF as a strong alternative to LMT, particularly when a high Specificity and a good balance of Precision and Recall are needed.
- Utilize BN and J48 Trees in cases where their specific performance characteristics, such as lower computational requirements, may offer practical advantages.

Reconsider the use of REP Tree for anemia classification, as its lower accuracy and higher incorrect classification rate suggest it may not be the best choice compared to other evaluated classifier.

References

1. Ahmed, N. M., Hamdin, A. O., & Mohamad, S. H. (2023). Copula Probability Model for Determining Dependency and Generations. University of Kirkuk Journal for Administrative and Economic Science, 13(2).
2. Ali, F. K. H., Abdullah, S. A., & Mohamad, S. H. (2023). Relationship Between Socio-Demographic Characteristics and Food Labeling Consumption in Sulaimani City by Using Chi-Square Test. Al-Qadisiyah Journal For Agriculture Sciences, 13(1).
3. Doulah, M. (2019). Application of machine learning algorithms in bioinformatics. J Proteom Bioinform, 3(1), 1-11.
4. Elbeltagi, A., Pande, C. B., Kumar, M., Tolche, A. D., Singh, S. K., Kumar, A., & Vishwakarma, D. K. (2023). Prediction of meteorological drought and standardized precipitation index based on the random forest (RF), random tree (RT), and Gaussian process regression (GPR) models. Environmental Science and Pollution Research, 30(15), 43183-43202.
5. Fayaz, S. A., Zaman, M., & Butt, M. A. (2021). An application of logistic model tree (LMT) algorithm to ameliorate Prediction accuracy of meteorological data. International Journal of Advanced Technology and Engineering Exploration, 8(84), 1424-1440.
6. Friedman, A. J., Chen, Z., Ford, P., Johnson, C. A., Lopez, A. M., Shander, A., ... & Van Wyck, D. (2012). Iron deficiency anemia in women across the life span. Journal of women's health, 21(12), 1282-1289.
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.
8. Jain, R. (2023). Comparative Study of Machine Learning Algorithms for Document Classification. International Journal of Computer Sciences and Engineering IJCSE.
9. Mitchell, T. M., & Mitchell, T. M. (1997). Machine learning (Vol. 1, No. 9). New York: McGraw-hill.
10. Mohamed, W. N. H. W., Salleh, M. N. M., & Omar, A. H. (2012, November). A comparative study of reduced error pruning method in decision tree algorithms. In 2012 IEEE International conference on control system, computing and engineering (pp. 392-397). IEEE.
11. Munro, M. G. (2023). Heavy menstrual bleeding, iron deficiency, and iron deficiency anemia: Framing the issue. International Journal of Gynecology & Obstetrics, 162, 7-13.
12. Omer, A., Faraj, S. M., & Mohamad, S. H. (2023). An application of two classification methods: hierarchical clustering and factor analysis to the plays PUBG. Iraqi Journal of Statistical Sciences, 20(1), 25-42.
13. Posonia, A. M., Vigneshwari, S., & Rani, D. J. (2020, December). Machine learning based diabetes prediction using decision tree J48. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 498-502). IEEE.
14. Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., & Mewes, H. W. (2005). Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry, 29(1), 37-46.
15. Yousefi, S., & Poornajaf, M. (2023). Comparison of result of machine learning algorithms in predicting heart disease. Frontiers in Health Informatics, 12.