IRAQI Academic Scientific Journals

KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES (2024) (01) (02): P(52-64)

خزائن للعلوم الاقتصادية والإدارية
KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES
ISSN: 2960-1363 (Print)
ISSN: 3007-9020 (Online)

Machine Learning Approaches to Classify and Predict Congenital Jaundice

Dr.Mohammad Mahmood Faqe Hussein [1], Soran Husen Mohamad [2], Bakhan Hoshyar Qadir [3]

[1] -statistics and informatics department - College of Administration and Economics -University of Sulaimani- Sulaimani city – Iraq
[2]-statistics and informatics department - College of Administration and Economics -University of Sulaimani- Sulaimanicity – Iraq
[3] -statistics and informatics department - College of Administration and Economics -University of Sulaimani- Sulaimanicity – Iraq

Mohammad.faqe @univsul.edu.iq[1]
Soran.abdulrahman@univsul.edu.iq[2]
Bakhan.qadir@univsul.edu.iq[3]

Abstract. Neonatal Jaundice, a common condition in newborns, results from elevated bilirubin levels, leading to the characteristic yellowing of the skin and eyes. Timely and precise classification and prediction of neonatal jaundice are crucial for early medical intervention. Machine learning has emerged as a powerful tool in healthcare for creating predictive models. This abstract offers an overview of machine learning methods used to classify and predict neonatal jaundice, incorporating clinical, laboratory, and genetic data. The dataset used in this research has been collected from patients at the Sulaimani Children's Hospital. The sample was collected over 5 months, from 2022 to 2023. Number of rows: 130 (one for each patient's data) There are 8 attributes in each record with Y as the target attribute consisting of jaundice levels The remaining seven characteristics are what the algorithm uses to predict. All characteristics are discrete. Table 1 / Dataset used in this study Full size table. Two types of machine learning algorithms are used: KNN and Naive Bayes. The assessment of model performance relies on metrics such as accuracy, sensitivity, specificity, and AUC-ROC, revealing encouraging outcomes. The results show that the classification of congenital jaundice using the KNN (k = 15) algorithm gives us accurate classification results when compared with the Naive Bayesian algorithm, and its classification percentage is equal to 64.34%.

Keywords: Machine Learning, Categorization, Congenital jaundice, Model Evaluation, AUC-ROC Analysis

DOI: 10.69938/Keas.2401025

آلة التعلم لطرق التصنيف والتنبؤ اليرقان الخلقي

د.محمد محمود فقي حسين [١] ، سوران حسين محمد [٢]، بخان هوشيار قادر [3]

[1] قسم الاحصاء و المعلوماتية- كلية الإدارة والاقتصاد- جامعة السليمانية - المدينة السليمانية - العراق
[2] قسم الاحصاء و المعلوماتية- كلية الإدارة والاقتصاد- جامعة السليمانية - المدينة السليمانية - العراق
[3] قسم الاحصاء و المعلوماتية- كلية الإدارة والاقتصاد- جامعة السليمانية - المدينة السليمانية - العراق

Mohammad.faqe @univsul.edu.iq[1]
Soran.abdulrahman@univsul.edu.iq[2]
Bakhan.qadir@univsul.edu.iq[3]

المستخلص. اليرقان الخلقي، وهو حالة شائعة عند الأطفال حديثي الولادة، ينتج عن ارتفاع مستويات البيليروبين، مما يؤدي إلى الاصفرار المميز للجلد والعينين. يعتبر التصنيف الدقيق والتنبؤ المبكر باليرقان الخلقي أمرًا بالغ الأهمية للتدخل الطبي المبكر. ظهرت تقنيات التعلم الآلي كأداة قوية في الرعاية الصحية لإنشاء نماذج تنبؤية. يقدم هذا الملخص نظرة عامة على طرق التعلم الآلي المستخدمة لتصنيف وتوقع اليرقان الخلقي، ويشمل البيانات السريرية

والمختبرية والجينية. تم جمع البيانات المستخدمة في هذا البحث من المرضى في مستشفى الأطفال في السليمانية. تكونت العينة خلال 5 أشهر، من 2022 إلى 2023. يتكون هذا النموذج من 130 سجل مرضى. كل سجل يمثل مريضًا واحدًا. يحتوي السجل على 8 سمات، واحدة منها هي السمة القابلة للتنبؤ والتي تسمى Y، وقيمتها تشير إلى معدل اليرقان. السبعة سمات الأخرى تُستخدم في الجزء التنبؤي من الخوارزمية. جميع السمات الثمانية هي سمات فئوية. يوضح الجدول التالي البيانات المستخدمة في هذه الدراسة. تم استخدام نوعين من خوارزميات التعلم الآلي: KNN وNaïve Bayesian. يعتمد تقييم أداء النموذج على مقاييس مثل الدقة، الحساسية، النوعية، وAUC-ROC، مما يكشف عن نتائج مشجعة. تظهر النتائج أن تصنيف اليرقان الخلقي باستخدام خوارزمية KNN (k = 15) يعطينا نتائج تصنيف دقيقة مقارنةً مع خوارزمية Naïve Bayesian، ونسبة التصنيف تساوي 64.34%.

الكلمات المفتاحية: التعلم الآلي، التصنيف، فرط بيليروبين الدم الوليدي، تقييم النموذج، تحليل AUC-ROC.

Corresponding Author: E-mail: Mohammad.faqe @univsul.edu.iq1

1 Introduction

Zulkarnain, Z., et al. Naive bayes performed better in Netflix Rating classification with 72% accuracy whereas KNN is having only 61%. Netflix uses both these algorithms to classify movie and TV show ratings as seen by the user. Puteri, Q. A., et al. The model based on Naive Bayes shows higher efficiency rate in predicting Diabetes compares to a K-Nearest Neighbour of 71% and around 77%, respectively, for this study. Novianto, E., et al. (2023) K-Nearest Neighbor (KNN):96.67%Naive Bayes:77.33%Law undergraduate timely graduation, as per the study. Muzakir, A., et al. (2023) K-Nearest Neighbor (KNN) and Naive Bayes are utilized for prostate cancer classification, with KNN showing higher accuracy (90%) compared to Naive Bayes (80%). In the paper, we want to find a suitable classification for our data variable by using two steps, the first step we used K-Nearest Neighbor (KNN) and Naive Bayes Classifier techniques in the second steps we want to compare between the two roads and know which way is a better classification for our data set by using Sensitivity, Specificity, Accuracy and Correctly Classified to check the effectiveness of the evaluated techniques.

1.1: Machine learning:

The area of machine learning has developed from the wide field of artificial intelligence, which seeks to emulate machines with the intelligent skills of humans. It is concerned with designing and developing algorithms that enable data-based computers to learn. Fields such as statistics, probability theory, data mining, or pattern recognition are closely related to machine learning. Machine Learning is a branch of computer science that uses previous experience to make future decisions by learning from and using its knowledge. Computer science, engineering, and mathematics are at the intersection of machine learning. Machine learning uses statistical theory in order to build mathematical models. Finding algorithms to solve problems is important. The aim of machine learning is to generalize a pattern that can be observed or establish an unknown basis for such instances. Machine learning can take many different forms.

1.1.1: Supervised Learning: In this type, the model is trained on a labeled dataset where the input data is paired with the correct output labels. It learns to map inputs to outputs and can then make predictions on new, unseen data.

1.1.2: Unsupervised Learning: Here, the model is given data without explicit labels. It aims to find patterns, relationships, or structures within the data. Clustering and dimensionality reduction are common tasks in unsupervised learning.

1.1.3: Semi-Supervised Learning: This approach combines elements of both supervised and unsupervised learning. It leverages a small amount of labeled data along with a larger amount of unlabeled data.

1.1.4: Reinforcement Learning: In reinforcement learning, an agent learns to interact with an environment to achieve a specific goal. It receives feedback in the form of rewards or penalties based on its actions and uses this feedback to improve its decision-making.

Machine learning has found applications in a wide range of fields, including image and speech recognition, natural language processing, recommendation systems, autonomous vehicles, healthcare diagnostics, finance, and more[1][6].

1.2: Use Machine Learning

Machine learning should be considered when you have a problem or task that involves making predictions, classifications, or decisions based on data, especially when traditional rule-based programming or manual analysis becomes complex or infeasible.

It's important to note that using machine learning requires having sufficient and relevant data, as well as expertise in the field. It's not a one-size-fits-all solution and might not be appropriate for every problem. Additionally, the decision to use machine learning should be based on a cost-benefit analysis, as it can be resource-intensive in terms of data preparation, model training, and maintenance[3][14].

1.3: K -Nearest Neighbor (KNN)

K-NN, or k-nearest neighbors, is an instance-based, or lazy learning method, used for classification in data mining. Due to its reliance on learning from examples rather than constructing an explicit model during training, K-NN is often referred to as a lazy learner. It uses training examples to form its model. K-NN is a straightforward and effective classification technique. For our model training, we employ the K-NN classifier using the heart disease dataset. The concept of "closeness" is based on the Euclidean distance measure, which calculates the distance between two points, X and Y, as shown in the equation below (1.1)

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \ldots(1)$$

Where:

y_i = Data Samples : x_i = Testing Data

i = Data Variable

D = Distance

n = Data Dimension

We have repeatedly applied the K-NN classifier, running it through multiple iterations while adjusting the value of 'K' to achieve optimal accuracy. To measure precision, sensitivity, and specificity, we use the confusion matrix. The K-Nearest Neighbors (K-NN) algorithm, a non-parametric method used for both classification and regression, operates within the realm of pattern recognition. It is a supervised learning approach where the 'K' in the algorithm refers to the number of nearest neighbors considered [7].

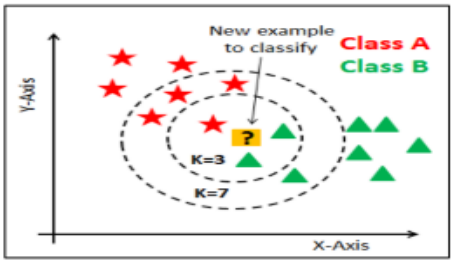

Figure (1): KNN classification example[17]

The figure above Shows how KNN classification can differ based on the K-value. If (K=3) the unknown data-point would be categorized as green [Class B] and if (K = 7) the unknown data-point would be categorized as red [Class A] [17]

When we input an unknown data point to the K-NN algorithm , if K= 3, a circle would be drawn around the data point with the nearest 3 known points inside, whatever the majority of the three data is, the unknown point would be classified as that type. For example, when K=3 in the above figure, two of the points are green, therefore the unknown point would be categorized as green. Same process

is done for the situation when K=7. In the above figure majority of pints is red when K=7, therefore it is classified as red.

To classify a dataset using the K-Nearest Neighbor (KNN) algorithm, we adhered to the following steps:

We started by determining the parameter K, representing the number of nearest neighbors. After testing several K-values, we found that K=7 provided the highest accuracy compared to K=3 and K=5. We then set the distance between data points to be equal to (1/d). The Weka program sorted these distances from the highest to the lowest value and identified the closest distance to the k-th order. Finally, we classified the test data based on the training datasets as outlined in the confusion matrix provided by the Weka program.

The K value is recommended to be an odd number greater than one. Increasing the K-value reduces the impact of classification noise. For evaluating distances, we used the Euclidean distance metric. The K-Nearest Neighbor algorithm is advantageous as it is resistant to noise in the training data and performs well when the training dataset is sufficiently large. Distance between neighbors can be measured using the Euclidean distance, as described in equation (1) [9].

1.4: Advantages and Disadvantages of K -Nearest Neighbor (KNN)

The following are some advantages of K-NN :

1– Since the process is clear, implementing and debugging is simple.

2 – If an analysis of the neighbors is useful as an explanation, K-NN can be very powerful in cases where an explanation of the performance of the classifier is useful.

3– There are several strategies for noise reduction that operate only for K-NN that can be useful in improving the classifier's accuracy.

4– Case-Retrieval Nets are an elaboration of the concept of the Memory-Based Classifier that can dramatically increase runtime efficiency on large case-bases[5].

Some of the main disadvantages of KNN method are as follows:

1– Because all the work is done at runtime, if the training set is big, K-NN may have poor run-time output.

2– Since all features contribute to the similarity and thus to the classification, K-NN is very sensitive to irrelevant or redundant features. This can be enhanced by careful selection of features or the weighting of features.

3– On very difficult classification tasks, more exotic techniques such as Support Vector Machines or Neural Networks can outperform K-NN[5].

1.5: Naive Bayes Classifier:

is a probabilistic classifier that applies the Bayes theorem with strong (naive) independence assumptions about the relationships between the features. The model of Naive Bayesian is easy to build, with no complicated iterative parameter estimation. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is popularly used because it often outperforms more complex classification methods. Bayes theorem provides a way of calculating the posterior probability, P(A|B), from P(A), P(B), and P(B|A). Naive Bayes makes the assumption that the effect of the value of a predictor (B) on a given class (A) is independent of the values of other predictors. The term "class conditional independence" refers to this presumption[4].

Bayes theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Mathematically, it's represented as:

$$P(y/X) = \frac{p(y,x)}{p(x)} \quad \rightarrow P(y/X) = \frac{P(X/y)P(y)}{P(X)} \qquad \dots (2)$$

where:

$P(y/X)$ = is the probability of event A occurring given that event B has occurred.

$P(X/y)$ = is the probability of event B occurring given that event A has occurred.

$P(y)$ and $P(X)$ are the probabilities of events A and B occurring independently [21][23].

Bayes theorem also can be rewritten as:

$$P(y/x_1, \dots, x_n) = \frac{P(x_1/y)P(x_2/y)\dots P(x_n/y)P(y)}{P(x_1)P(x_2)\dots P(x_n)} \quad \dots (3)$$

Or

$$P(y/x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i/y) \quad \dots (4)$$

The "naive" part of Naive Bayes comes from the assumption that all features (attributes or variables) are independent of each other given the class label. This is often not true in reality, but the simplification allows for a more tractable and computationally efficient model[16].

1.6: Types of Naive Bayes Algorithms:

1: Multinomial Naive Bayes: Used for text classification and dealing with discrete data, like word counts in text.

2: Gaussian Naive Bayes: Assumes that features follow a Gaussian distribution. Suitable for continuous numerical features.

3: Bernoulli Naive Bayes: Similar to multinomial but used for binary or Boolean features.

Naive Bayes is relatively simple, interpretable, and computationally efficient. It's often used as a baseline model for text classification tasks and can perform surprisingly well even when the independence assumption doesn't strictly hold. However, more advanced algorithms may outperform it in complex scenarios where feature interdependencies are significant[21].

1.7: Advantages and Disadvantages of Naïve Bayesian:

The following are some advantages of Naïve Bayesian:

1. Handling quantitative and discrete data
2. It only requires a small amount of training data to estimate the parameters (average and variance of variables) required for classification
3. Handle the lost value by ignoring the agency during the estimated opportunity calculation
4. Fast and space efficiency
5. Strong against irrelevant attributes[12].

Some of the main disadvantages of Naïve Bayesian method are as follows:

1. Not applicable if the conditional probability is zero, if zero then the predicted probability will be zero as well
2. Assume independent variables[12].

1.8: Measures for Performance Evaluation:

One of the key steps in developing any machine learning model is assessing its performance. Evaluation metrics are closely linked to specific machine learning tasks, with different metrics suited for classification and regression tasks. Metrics such as precision and recall are valuable across multiple tasks. Classification and regression are examples of supervised learning, which dominates most machine learning applications. The Confusion Matrix serves as a performance measurement tool for classification problems where the output involves two or more classes, as illustrated in Table 1.1. [2][15]

Table (1) :Confusion Matrix[2]

Actual Class	**Predicted Class**	
	Predictive Positive	Predictive Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

1.8.1: Accuracy

Accuracy is a common metric used to measure the performance of classification models, and it represents the ratio of correct predictions to the total number of predictions made by the model[1].

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} * 100\% \quad \ldots(5)$$

Where,

TP= True Positive : TN = True Negative : FP = False Positive : FN = False Negative

1.8.2: Sensitivity and Specificity

The metric known as Sensitivity measures a model's ability to correctly predict true positives for each category. In contrast, Specificity assesses a model's effectiveness in predicting true negatives for each category. These metrics are applicable to any categorical model. The equations used to calculate Sensitivity and Specificity are provided below [8].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} * 100 \quad \ldots(6)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} * 100 \quad \ldots(7)$$

1.8.3: Kappa Statistic

Kappa Statistics is a metric that compares observed accuracy with expected accuracy, accounting for random chance. It is used not only to assess the performance of a single classifier but also to compare different classifiers against each other. By considering the possibility of agreement occurring by random chance, Kappa Statistics often provides a more accurate assessment than using accuracy alone. For instance, an observed accuracy of 80% is less significant if the expected accuracy is 75% compared to an expected accuracy of 50% [17]

$$\text{Kappa Statistics} = \frac{\text{Observed Accuracy}-\text{Expected Accuracy}}{1-\text{ Expected Accuracy}} \quad \ldots(8)$$

1.8.4: Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a measurement of the average size of errors in a set of predictions, without taking into account their direction. It is used to assess the effectiveness of a regression model and is measured as the average absolute difference between the predicted values and the actual values[19].

$$MAE = \frac{\sum_{i=1}^{n}|y_i-\hat{y}|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} \quad \ldots(9)$$

Where,

y_i = Actual value

$\hat{y}$ = Predictive Value

n = Number of Testing Data

1.8.5: Receiver Operating Characteristic (ROC)

A curve of the Receiver Operator Characteristic (ROC) is a graphical plot used to illustrate binary classifiers' diagnostic ability. It was first used in the theory of signal detection, but is now used in many other fields, including medicine, radiology, natural hazards, and machine learning. The ROC is not based upon the distribution of groups. This makes it helpful for classifiers that predict rare events such as illnesses or disasters to be evaluated. In comparison, performance assessment using accuracy[20].

$(TP + TN)\big/(TP + TN + FN + FP)$ *would favor classifiers that always predict a* negative outcome for rare events.

3.1 Data Description and Analysis

The dataset utilized in this study was gathered from patients at Sulaimani Children's Hospital over a five-month period from 2022 to 2023. It comprises a total of 130 patient records, with each row representing an individual patient's data. The dataset includes eight attributes, one of which is the target variable labeled Y, indicating the Jaundice Rate. The remaining seven attributes are used for

the prediction aspect of the algorithm. All eight attributes are categorical. The dataset used in this study is illustrated in the following table.

Table (2): Variables of this Study

X_1: Gender	X_5: Blood Group (Mother)
X_2: Type of Birth	X_6: Pregnancy Period (week)
X_3: Age (Day)	X_7: Weight (Kg)
X_4: Blood Group (Father)	Y: Jaundice Rate $< 14 = a$ Jaundice Rate $\geq 14 = b$

The aim of this study is to perform a parametric analysis of the dataset by applying two distinct machine learning algorithms. To classify the data, we employed both the K-Nearest Neighbor (KNN) and Naive Bayes Classifier techniques. The Weka program was utilized to generate confusion matrices, which were instrumental in evaluating the effectiveness of these techniques

3.2 Analysis with K -Nearest Neighbor (KNN)

Table (3): Confusion Matrix KNN

When K=3				**When K=5**				**… When K=15**			
Predicted Class				**Predicted Class**				**Predicted Class**			
Actual Class	A	b	SUM	Actual Class	a	b	SUM	Actual Class	a	b	SUM
a	25 (TP)	34 (FN)	59	a	22 (TP)	37 (FN)	59	a	27 (TP)	32 (FN)	59
b	28 (FP)	42 (TN)	70	b	23 (FP)	47 (TN)	70	b	14 (FP)	56 (TN)	70
SUM	53	76	129	SUM	45	84	129	SUM	41	88	129

The table above indicates that the model correctly classified the conditions of 67 patients (true positives and true negatives) when K=3. Conversely, the model incorrectly classified the conditions for 62 patients (false positives and false negatives). In other words, 25 patients (out of 53) diagnosed with (a) (Jaundice Rate < 14) have been classified correctly and the remaining 28 patients have been predicted to have (b) (Jaundice Rate ≥ 14) instead of (a) (Jaundice Rate < 14). Similarly, 42 patients (out of 76) who were diagnosed with (b) (Jaundice Rate ≥ 14) have been classified correctly and the remaining 34 patients have been misclassified as having (a) (Jaundice Rate < 14).
The table above reveals that with K=5, the model correctly classified the conditions for 69 patients (true positives and true negatives). However, it inaccurately classified the conditions for 60 patients (false positives and false negatives). In other words, 22 patients (out of 45) diagnosed with (a) have been classified correctly and the remaining 23 patients have been predicted to have (b) instead of (a). Similarly, 47 patients (out of 84) who were diagnosed with (b) have been classified correctly and the remaining 37 patients have been misclassified as having (a).
The table above illustrates that when K=15, the model accurately classified the conditions for 83 patients (true positives and true negatives). However, it misclassified the conditions for 46 patients (false positives and false negatives). In other words, 27 patients (out of 41) diagnosed with (a) have been classified correctly and the remaining 14 patients have been predicted to have (b) instead of (a). Similarly, 56 patients (out of 88) who were diagnosed with (b) have been classified correctly and the remaining 32 patients have been misclassified as having (a).

3.3 Analysis with Naïve Bayes Classifier

Table (4): Confusion Matrix Naïve Bayesian

Predicted Class			
Actual Class	**a**	**b**	**SUM**

A	22 (TP)	37 (FN)	59
B	15 (FP)	55 (TN)	70
SUM	37	92	129

The table above demonstrates that the model correctly classified the conditions for 77 patients (true positives and true negatives). Conversely, it misclassified the conditions for 52 patients (false positives and false negatives). In other words, 22 patients (out of 37) diagnosed with (a) (Jaundice Rate < 14) have been classified correctly and the remaining 15 patients have been predicted to have (b) (Jaundice Rate ≥ 14) instead of (a) (Jaundice Rate < 14) . Similarly, 55 patients (out of 92) who were diagnosed with (b) (Jaundice Rate ≥ 14) have been classified correctly and the remaining 37 patients have been misclassified as having (a) (Jaundice Rate < 14) .

3.4 Calculating Results for K -Nearest Neighbor (KNN)

When K=3

$$\text{Sensitivity} = \frac{25}{25+34} * 100 = 42.37\%$$

$$\text{Specificity} = \frac{42}{42+28} * 100 = 60\%$$

$$\text{Accuracy} = \frac{(25+42)}{(25+28+34+42)} * 100\% = 51.93\%$$

$$\text{Error Rate } (a,b) = \frac{(FP+FN)}{(TP+FP+FN+TN)} * 100\%$$

$$\text{Error Rate } (a,b) = \frac{(28+34)}{(25+28+34+42)} * 100\% = 48.06\%$$

When K=5

$$\text{Sensitivity} = \frac{22}{22+37} * 100 = 37.28\%$$

$$\text{Specificity} = \frac{47}{47+23} * 100 = 67.14\%$$

$$\text{Accuracy} = \frac{(22+47)}{(22+23+37+47)} * 100\% = 53.48\%$$

$$\text{Error Rate } (a,b) = \frac{(23+37)}{(22+23+37+47)} * 100\% = 46.51\%$$

When K=15

$$Sensitivity = \frac{27}{27+32} * 100 = 45.76\%$$

$$Specificity = \frac{56}{56+14} * 100 = 80\%$$

$$Accuracy = \frac{(27+56)}{(27+14+32+56)} * 100\% = 64.34\%$$

$$Error\ Rate\ (a,b) = \frac{(14+32)}{(27+14+32+56)} * 100\% = 35.65\%$$

Table (5): The Classification Accuracy, Sensitivity and Specificity of Proposed model (KNN)

Classifier	Sensitivity	Specificity	Accuracy	Correctly Classified		Incorrectly Classified	
KNN = 3	42.37%	60%	51.93%	67	51.93%	62	48.06%
KNN = 5	37.28%	67.14%	53.48%	69	53.48%	60	46.51%
KNN = 15	45.76%	80%	64.34%	83	64.34%	46	35.65%

3.5 Calculating Results for Naïve Bayes Classifier

$$\text{Sensitivity} = \frac{22}{22+37} * 100 = 37.28\%$$

$$\text{Specificity} = \frac{55}{55+15} * 100 = 78.57\%$$

$$\text{Accuracy} = \frac{(22+55)}{(22+15+37+55)} * 100\% = 59.68\%$$

$$\text{Error Rate}\,(a,b) = \frac{(15+37)}{(22+15+37+55)} * 100\% = 40.31\%$$

Table (6): The Classification Accuracy, Sensitivity and Specificity of Proposed model Naïve Bayes and KNN (K=15)

Classifier	Sensitivity	Specificity	Accuracy	Correctly Classified		Incorrectly Classified	
Naïve Bayesian	37.28%	78.57%	59.68%	77	59.68%	52	40.31%
KNN (K= 15)	45.76%	80%	64.34%	83	64.34%	46	35.65%

The above table shows that the K -Nearest Neighbor (KNN) yields higher and better results compared to Naïve Bayes in terms of Sensitivity, Specificity and accuracy. Specificity for KNN method with three K values (3, 5, 15) the obtained values are (60,67.14,80) while for SVM is (78.57). Additionally, Sensitivity for KNN method with three K values (3, 5, 15) the obtained values are (42.37,37.28,45.76) while for SVM is (37.28). Finally, accuracy for KNN method with three K values (3, 5, 15) the obtained values are (51.93,53.48,64.34) while for SVM is (59.68).

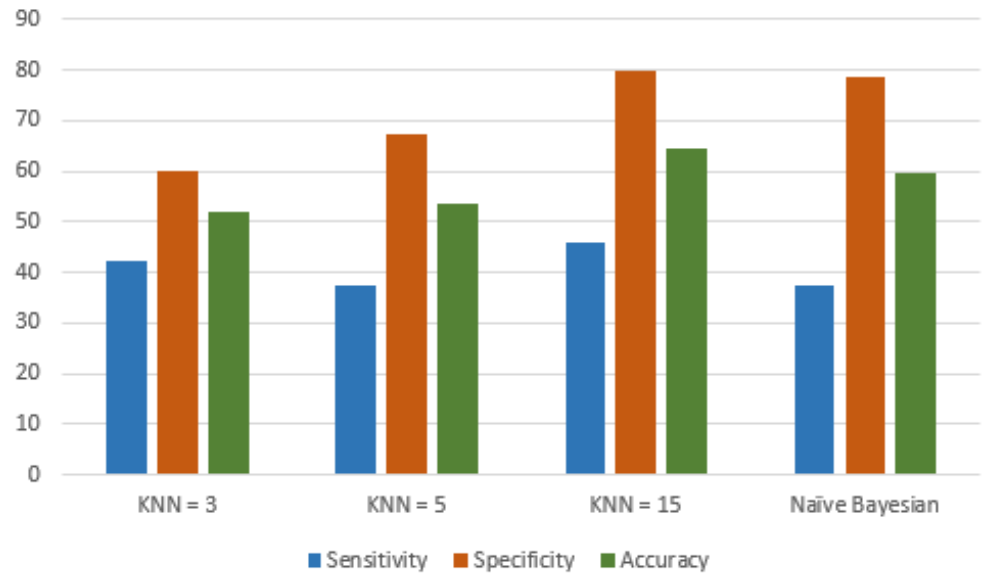

Figure (2) Illustrate the difference in Accuracy, Sensitivity and Specificity of the proposed models.

3.6: Calculation Receiver Operating Characteristic, Kappa Statistic and Mean Absolut Error for K -Nearest Neighbor (KNN)

When (K=15)

$$\text{Observed Avvuracy} = \frac{(\text{TP} + \text{TN})}{\text{Total}} = \frac{(27+56)}{129} = 0.64$$

$$\text{Expected Accuracy} = \frac{\frac{(\text{TP}+\text{FN})*(\text{TP}+\text{FP})}{\text{Total}} + \frac{(\text{FP}+\text{TN})*(\text{FN}+\text{TN})}{\text{Total}}}{\text{Total}}$$

$$\text{Expected Accuracy} = \frac{\frac{(27+32)*(27+14)}{129} + \frac{(14+56)*(32+56)}{129}}{129} = \frac{18.75+47.75}{129} = 0.515$$

$$\text{Kappa Statistics} = \frac{(\text{Observed Avvuracy} - \text{Expected Accuracy})}{1 - \text{Expected Accuracy}} = \frac{(0.64 - 0.515)}{1 - 0.515} = 0.257$$

$$\text{Mean Absolute Error(MAE)} = \frac{\sum_{i=1}^{n}|y_i - y^n|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n} = \frac{46}{129} = 0.356$$

$$\text{Weighted Avg(ROC)} = \frac{(\text{ROC(a)} * \text{P}) + (\text{ROC(b)} * \text{N})}{\text{P} + \text{N}} = \frac{(0.605 * 41) + (0.605 * 88)}{41 + 88}$$
$$= \frac{(24.805 + 53.24)}{129} = 0.605$$

Table (7): Detailed Accuracy by Class KNN(when K=3,5,15)

When K=3				**When K=5**				**When K=15**			
Class	A	B	Weighted Avg.	Class	A	b	Weighted Avg.	Class	a	b	Weighted Avg.
TP Rate	0.424	0.6	0.519	TP Rate	0.373	0.671	0.535	TP Rate	0.458	0.8	0.643
FP Rate	0.4	0.576	0.496	FP Rate	0.329	0.627	0.491	FP Rate	0.2	0.542	0.386
Precision	0.472	0.553	0.516	Precision	0.489	0.56	0.527	Precision	0.659	0.636	0.647
Recall	0.424	0.6	0.519	Recall	0.373	0.671	0.535	Recall	0.458	0.8	0.643
ROC Area	0.538	0.538	0.538	ROC Area	0.574	0.574	0.574	ROC Area	0.605	0.605	0.605

The data in Table (7) as shown that the weighted average of Roc in KNN method has the highest value (0.538, 0.574) for k-values of (3, 5) compared to the weighted average of precision, Recall, TP Rate, FP Rate. On the other hand, the weighted average of Precision in KNN method when K-value of (15) has the highest value (0.647) compared to the weighted average of ROC, Recall, TP Rate, FP Rate . Additionally, the weighted average of false positive yields the lowest results for the three k-values (3, 5, 15) which are (0.496, 0.491, 0.386). We also noticed that the lowest weighted average value for Naïve Byes and KNN is FP rate.

Table (8): Detailed Accuracy by Class Naïve Byes

Class	**a**	**b**	**Weighted Avg.**
TP Rate	0.373	0.786	0.597
FP Rate	0.214	0.627	0.438
Precision	0.595	0.598	0.596
Recall	0.373	0.786	0.597
ROC Area	0.579	0.579	0.579

The above Table (8) has been obtained from the Weka program; We learned that the weighted average of Recall has the highest value (0.597) when compared to weighted average of each of Roc area, TP Tate, FP Rate, Precision. Additionally, FP Rate has the least value (0.438).

4 Conclusions

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

1.Machine learning methods offer promising avenues for accurately categorizing and forecasting congenital jaundice based on diverse input parameters such as bilirubin levels, clinical

manifestations, and patient characteristics. Key considerations include the quality and quantity of data, the selection of pertinent features, model performance metrics, and the balance between complexity and interpretability.

2 In the KNN algorithm, the classification percentage for (K=3, 5, and 15) is (51.93%, 53.48%, 64.34%) respectively, we see that the best classification between those is (K=15), because the classification percentage is higher and better when compared with the other classifications.

3. by using two algorithms of classification (KNN) and (Naïve Bayesian), we see that classification using the (KNN)➔ k=15) algorithm gives us accurate classification results, and its classification percentage is equal to (64.34%), while classification percentage is equal to (59.68%) by using Naïve Bayesian algorithms.

4. In table (8): We learned that the weighted average of Recall has the highest value (0.597) when compared to weighted average of each of Roc area, TP Tate, FP Rate, Precision. Additionally, FP Rate has the least value (0.438).

References

1. Alpaydin, E. (2014). Introduction to machine learning. MIT press. Edition 3rd. United States of America.
2. بديعة رحمن خليل, أ. م. د. محمد محمود فقي, & أ. م. د. سوزان صابر حيدر. (2020). Classifying Patients with Myocardial Infarction and Heart Failure by Using SVM and KNN Learning Techniques .*Journal of Administration and Economics*.327-315 ,(126) ,
3. Amin, M., & Ali, A. (2018). Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decisions. *Wavy AI Research Foundation: Lahore, Pakistan*, *90*.
4. Anwar A. N. Abu Alhussein (2019) Comparison Between Support Vector Machines and Artificial Neural Networks for Time Series Forecasting.
5. Christopher M. Bishop . Pattern Recognition and Machine Learning.
6. Cunningham, P., & Delany, S. J. (2020). k-Nearest Neighbour Classifiers--. arXiv
7. Dangeti, P. (2017). Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. Packt Publishing.
8. Lutz H.Hamel (2009) Knowledge Discovery with Support Vector Machines (Wiley Series on Methods and Applications Data Mining).
9. Mitrani, A. (2019). Evaluating Categorical Models II: Sensitivity and Specificity. Towards Data Science, Medium, 6.
10. Muzakir, A., Desiani, A., & Amran, A. (2023). Klasifikasi Penyakit Kanker Prostat Menggunakan Algoritma Naïve Bayes dan K-Nearest Neighbor. *Komputika: Jurnal Sistem Komputer*, *12*(1), 73-79.
11. Novianto, E., Hermawan, A., & Avianto, D. (2023). Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, Decision Tree Untuk Prediksi Status Kelulusan Mahasiswa S1. Rabit: Jurnal Teknologi dan Sistem Informasi Univrab, 8(2), 146-154.
12. Olson, D. L., & Delen, D. (2008). Advanced data mining techniques. Springer Science & Business Media.
13. Puteri, Q. A., Sagirani, T., & Lemantara, J. (2023). Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor (KNN) untuk Mengetahui Keakuratan Diagnosa Penyakit Diabetes. *Jurnal Nasional Teknologi dan Sistem Informasi*, *9*(3), 247-254.
14. Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. (2017). Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. *Int. J. Eng. Emerg. Technol*, *2*(1), 53. preprint arXiv:2004.04523.
15. Russell.R.(2018). Machine Learning Step-by-Step Guide To Implement Machine Learning Algorithms with Python. CreateSpace Independent Publishing Platform, ISBN: 1719528403,9781719528405.
16. Tomar, D., & Agarwal, S. (2014). Feature selection based least square twin support vector machine for diagnosis of heart disease. International Journal of Bio-Science and Bio-Technology, 6(2), 69-82

17. Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. International Journal of Innovative Science, Engineering & Technology, 2(9), 441-444.
18. Zulkarnain, Z., Mutia, R., Ariani, J. A., Barik, Z. A., & Azmi, H. (2024). Performance Comparison K-Nearest Neighbor, Naive Bayes, and Decision Tree Algorithms for Netflix Rating Classification. *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, *1*(1), 16-22.
19. Hussein, M. M. F., Saeed, A. A., & Mohamad, S. H. (2023). Comparison Markov Chain and Neural Network Models for forecasting Population growth data in Iraq. University of Kirkuk Journal For Administrative and Economic Science, 13(4).
20. 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.
21. Ahmed, D. H., Mohamad, S. H., & Karim, R. H. R. (2023). Using Single Exponential Smoothing Model and Grey Model to Forecast Corn Production in Iraq during the period (2022-2030). University of Kirkuk Journal For Administrative and Economic Science, 13(3).
22. Hamad, A. P. D. A. S., Faqe, A. P. D. M. M., & Mohamad, A. L. S. H. (2023). Forecasting Life-Expectancy in Iraq During the Period (2022-2035) Using Fuzzy Markov Chain. University of Fallujah, Journal of Business Economics for Applied Research, 5(3), 347-372.
23. Karim Hama Ali, F., A Abdullah, S., & Husen Mohamad, S. (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), 139-146.
24. Hussain, M. M. F., & Hamad, A. S. (2012). Use K-Means Cluster Analysis to Study the Classification Of Some Water Factories, According To Some Specifications On The Cover Of The Bottle
25. Website: K-Nearest Neighbors (KNN) Classification with scikit-learn : https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn
26. Website: Mean Absolute Error | Deep checks : https://deepchecks.com/glossary/mean-absolute-error/
27. Website: ROC Curve - How to Interpret ROC Curves - Displayr https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/
28. Website: Towards Data Science article on Naive Bayes: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
29. Website: Towards Data Science article on Naive Bayes: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c
30. Website: Machine Learning Mastery's guide on Naive Bayes: https://machinelearningmastery.com/naive-bayes-for-machine-learning/
31. Website: K-Nearest Neighbors (KNN) Classification with scikit-learn | DataCamp https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn

Appendix

Table(9) summary of data

No.	gender	type of birth	age(Day)	bloodGroupe (father)	blood Groupe (mother)	period	Weight (Kg)	level
1	0	1	5	1	2	40	3.5	a
2	1	1	4	0	2	37	3.25	a
3	1	1	5	0	2	40	3.5	b
4	0	1	2	4	0	40	3	a
5	1	1	4	2	2	39	4.4	b
6	1	0	6	0	0	36	2.9	b

. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
129	0	1	6	2	3	38	3	a
130	0	1	7	2	2	37	2.58	b

Note: for gender variable ((0) for female, (1) for male). for type of birth variable ((0) for Natural birth (1)for C-section birth). For blood Groupe (father) and blood groupe (mother) variables ((0) for O^+ , (1) for AB^+, (2) for A^+, (3) for B^+, (4) for B^-, (5) for O^-, (6) for AB^-, (7) for A^-).