AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 54(2) (2022) 131-140
DOI: 10.22060/miscj.2022.21360.5283

Data mining approach for prediction umbilical cord wrapping around the fetus and investigating effective factors

N. Abedini[1], F. Moayedi[2]*, S. E. Dashti[3]

[1] Zand higher education, Shiraz, Iran
[2] Department of Computer Engineering, Larestan Higher Education Complex, Lar, Iran
[3] Department of Computer Engineering, Jahrom Branch, Islamic Azad University, Jahrom, Iran

ABSTRACT: Today, in medical knowledge, data collection on various diseases is very important. One of the important issues in the medical world is the baby's birth and its related issues. The relationship between mother and fetus is by the umbilical cord which is responsible for the development of the fetus. In this article, using data mining methods, the occurrence of umbilical cord torsion around the fetus is predicted, we also investigated some factors that can affect this event. Based on the studying articles on fetus birth and its factors, and consultation with gynecologists, the new and comprehensive questionnaire was designed on factors affecting the wrapping of the umbilical cord around the fetus, including 31 features that were completed by 140 samples of pregnant mothers. Then, the questionnaire was evaluated by Cronbach's Alpha. Since the obtained dataset was imbalanced it was balanced with SMOTE technique. We compared different classification methods, including SVM, Random Forest, KNN, and Naïve Base for prediction, which KNN had the best result accuracy of 81%. Finally, to extract effective factors some association rule mining methods such as Predictive Apriori, and FP-growth were applied. the results show nutrition, blood pressure, diabetes, fetus number, and Internet usage can have more impact on wrapping the umbilical cord around the fetus.

Review History:

Received: May, 02, 2022
Revised: Sep. 10, 2022
Accepted: Nov. 17, 2022
Available Online: Feb. 28, 2023

Keywords:

Association rules mining
SMOTE
KNN
fetus
umbilical cord prediction

1- Introduction

As technology progresses, a huge amount of data is expanding, and one of the important factors in the success of the use of science in various fields is the ability to take advantage of this information and data. This technology is widely used today in various fields, including banks, industrial centers, large factories, medical centers and hospitals, research centers, smart marketing, and many more. Data mining is the extraction of information and knowledge and the discovery of hidden patterns from very large databases, a bridge between statistical science, computer science, artificial intelligence, modeling, machine learning, and visual representation of data. Data mining is a complex process to identify correct, new, and potentially useful patterns and models in large volumes of data, in a way that these patterns and models are understandable to humans. Data mining techniques are popular among health researchers for an array of purposes, such as improving diagnostic accuracy, identifying high-risk patients, and extracting concepts from unstructured data [1]. Considering that the issue of wrapping the umbilical cord around the fetus is one of the most important factors during childbirth and the birth of a baby and raises concerns for mothers and doctors at birth, it is important to examine the factors influencing the occurrence of this issue. The umbilical cord is the source of blood and also provides oxygen to the baby, the umbilical cord supplies all the nutrients, calories, protein, fat, and vitamins to the baby. Umbilical cord wrapping is when the umbilical cord is twisted 360 degrees around the fetus. Wrapping the umbilical cord around the fetal neck can occur in two ways. The first type of umbilical cord around the neck occurs when the umbilical cord twists 360 degrees around the fetal neck. The second type of umbilical cord around the neck is when the umbilical cord is not able to return around the neck and is therefore like a knot and is dangerous for the fetus. One way to diagnose this phenomenon is ultrasound. When the wrapping is so tight that childbirth harms the baby, the baby is diagnosed by cesarean section. But sometimes this phenomenon due to reduced umbilical cord blood flow can lead to asphyxia, abnormal fetal heart rate pattern, fetal acidosis, severe cerebral palsy, and even decrease IQ during the postpartum period[2]. The umbilical cord is a very narrow strip that contains two arteries and a vein that connects the fetus to the placenta. The length of the umbilical cord in about 9 months of pregnancy reaches about 50 cm. Its thickness also varies in different fetuses. The larger the fetus, the larger the umbilical cord, and the diameter of the 9-month-old fetus varies somewhat from fetus to fetus. The umbilical cord carries nutrient-rich blood from the placenta to the fetus and carries fetal waste to the placenta for excretion. The umbilical cord floats in the amniotic fluid and is located between the organs and the body of the fetus and the wall of the uterus. Because asphyxia in many cases

*Corresponding author's email: fmoayyedi@gmail.com

leads to fetal death and in other cases due to hypoxia and brain damage leading to cerebral palsy, seizures, and learning disabilities. Since the umbilical cord may twist around parts of the fetus' body, especially the neck, and as the fetus progresses and the fetus descends into the birth canal, uterine contractions compress the umbilical cord and reduce the fetal heart rate. The prevalence of an umbilical cord torsion around the fetal neck is 33-23%,[2-4]], so umbilical cord torsion can be one of the causes of neonatal asphyxia[5]. The most common cause of abnormal heart rate changes during labor is umbilical cord complications, including twisting around the neck or trunk and limbs of the fetus[6]. Factors affecting the relationship between the umbilical cord around the neck and low birth weight[7] have been investigated: age, number of pregnancies, number of deliveries, history of umbilical cord around the neck, birth weight, being overweight, birth weight, diabetes, and blood pressure. Fayyad, U. M. et al.[8], defined the definitions and concepts of data mining. Their effort is the first step towards integrating a knowledge discovery framework into databases. Describe the relationship between data mining, knowledge discovery, and other related fields. In another study, Sónia Pereira et al. [9] used data mining models and real data to examine the factors influencing pregnancy and childbirth to predict the appropriate method of delivery. Chamidah Nurul[10] presented the embryo status classification by Cardiotocography based on feature extraction using a combination of K-Means and support vector machine. Based on 10 cross-validations, using the Cardiotocography of data set obtained from the UCI machine learning repository. Chen et al.[11] also examined the factors affecting preterm birth by observing and studying a total of 910 pregnant mothers using neural network data mining and decision tree C5.0. Bendon Robert W et al.[12] In the laboratory umbilical cord twisting and twisting to investigate the possible cause of Umbilical cord blood obstruction was studied. Fetal heart rate drops and reduced heart rate are often attributed to umbilical cord obstruction without knowing the anatomical basis of the obstruction. In another study, Tahereh Ashraf Ganjavi[13] evaluated the outcome of pregnancies that were complicated by the twisting of the umbilical cord around the neck or body of the fetus. The comparison was performed using the chi-square test and the means were compared by t-student analysis. Also Kaveh, Mahbod et al.[7] have reviewed the relationship between the umbilical cord around the neck and the number of its rings with the weight of the fetus Cross-sectional. Factors studied included age, number of pregnancies, number of deliveries, birth weight, sex of the baby etc. Mallick et al.[14] have reviewed related to thermal care and umbilical cord care practices in South Asia. They reviewed key coverage trends and used multivariate logistic regression. They also found that hygienic umbilical cord care was significantly associated with infant survival. Arkadiusz Krzyżanowski et al.[15] have studied the diagnosis of umbilical cord abnormalities and assessment of fetal wellbeing using modern ultrasonography. Early diagnosis and surveillance can minimize fetal mortality and help make decisions. Gede Angga Pradipta et al.[16] improved classification performance of the fetal umbilical cord using a combination of smote method and multi-classifier voting in the imbalanced dataset by image preprocessing. They compared different methods. Their work on the issue is similar to ours, but they use image processing. Also, they use smaller datasets with fewer features. Mirza Shuja et al.[17] used data mining classifiers and SMOTE to predict type II diabetes mellitus. For preprocessing they used SMOTE to balance their dataset, then they used Bagging, SVM, MLP, simple logistic, and Decision Tree with the preprocessed data to select the best classifier for a balanced dataset to predict diabetes.

After reviewing the articles, in similar areas related to umbilical cord issues, some factors in these articles helped to determine the features we wanted, and we selected a few features from them, while we created many more features. So far, less research has been done using data mining algorithms to investigate the factors affecting the umbilical cord wrapping around the fetus with the number of features that we have considered in this article. The new and comprehensive questionnaire was designed on factors affecting the wrapping of the umbilical cord around the fetus, for the first time, a native dataset of various features including 31 features such as maternal age, birthday week, nutrition, internet usage, etc. related to the wrapped umbilical cord was created. Also, a variety of physical factors, technology, nutrition, etc. were considered. In the next step, the questionnaires were completed by 140 pregnant mothers in a gynecological hospital. Since wrapped cords were lower than normal cases, the obtained dataset was imbalanced and this can have a negative effect on the validity of the results. Therefore, the dataset was balanced with SMOTE technique. Although less work has been done using data mining techniques in this area, we have reviewed some of them. We chose among the data mining methods classification such as SVM, Random Forest, KNN, and Naïve Base for prediction, and compared the results. We also chose the association rule mining algorithms to select the appropriate features.

The rest of the paper is structured as follows. Section 2 discusses the methodology. This section includes the design of the questionnaire, how to collect information and also the evaluation of the questionnaire, dataset balancing, and the definition of the classification algorithm that had the best result. In addition, some evaluation parameters are defined. In section 3 we discuss the result of the accuracy of different classification algorithms. finally, in section 4 the conclusion was provided.

2- Methodology

The research steps are shown in the figure below, which we will describe separately below(Fig. 1).

2- 1- Questionnaire design

Factors were selected through studies and reviews of similar articles in the field of newborn birth and consultation with obstetricians and gynecologists. Factors such as age, number of pregnancies, number of deliveries, history of

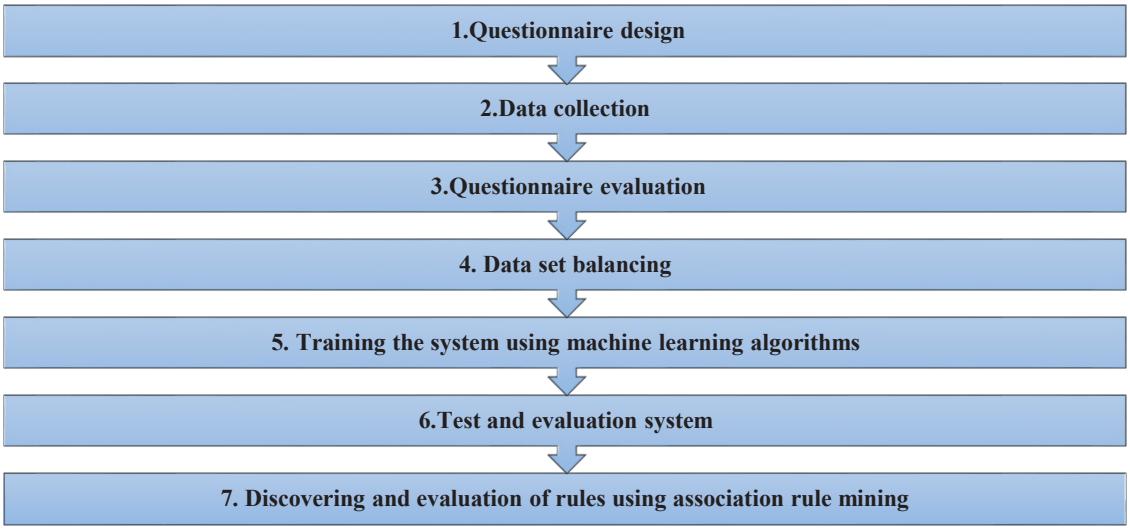

Fig. 1. Flowchart of the proposed method

umbilical cord around the neck, etc. about the umbilical cord and low birth weight had been previously studied [7]. In this article, in addition to such factors, the selected features are selected to be more biological factors related to mothers and fetuses. Also, other factors such as the impact of mobile phone technology and the Internet, nutrition, sleep, etc. have been considered. The features of our data set with their values are shown in Table 1.

2- 2- Data collection

The questionnaire was distributed among pregnant women. This information was collected by the mothers of the clinic and the maternity hospital. The total number of samples is 140. Of all these samples, 97 are without a cord wrapped and 43 with a cord wrapped around the fetus. The information was divided into two classes A and B. Class A includes information on mothers without wrapped cords and class B includes information on mothers with wrapped cords. To investigate the factors, for example, pregnant mothers of different ages and living conditions were examined. Finally, the desired dataset was collected with 31 features. Among these features, some were considered to be physiological factors, some to the use of technologies such as mobile phones, the Internet, and computer, as well as to the factors of nutrition and sleep, and electrical appliances. It can be said that until now, a dataset with these features and with this number of features had not been created and selected to investigate the factors affecting the umbilical cord wrapping around the fetus.

2- 3- Evaluation of the questionnaire

Reliability and validity are used to evaluate the questionnaire. Validity is defined as the amount that is measured in a qualitative study. The second measure of quality in a quantitative study is the reliability or accuracy of an instrument. In other words, the rate at which a research tool is used is the same as the results used. Reliability is about the integrity of a size. A participant completing a device must have the same answer each time the test is completed to measure motivation. Although it is not possible to accurately calculate reliability, reliability estimation can be obtained through various measures[18]. One of the most common methods for assessing the reliability of questionnaires is to determine the internal correlation of the questionnaire questions, which is done through Cronbach's alpha coefficient. Cronbach's alpha coefficient is closely related to the internal coordination of the questions and its value is theoretically between zero and one. Cronbach's alpha formula is as follows. Where N, $\bar{c}$, $\bar{v}$ are the number of items, average covariance between item-pairs, and average variance respectively.

$$\alpha = \frac{N.\bar{c}}{\bar{v} + (N-1).\bar{c}} \quad (1)$$

Here, after designing the questionnaire and reviewing the questions, to evaluate the questions of the questionnaire, its reliability was evaluated by Cronbach's alpha method, and also the degree of correlation and its relationship were evaluated using Reliability and Correlation in IBM SPSS statistics software.

2- 4- Dataset balancing

The problem of imbalanced data distribution in classes arises when there are more instances in some classes than in

Table 1. Data set features

Row	Feature	Row	Feature
1	Maternal age (18 to 24, 23 to 29, 30 to 35, 36 and up)	17	Decreased amniotic fluid (yes, no)
2	Maternal weight (55 to 65 kg, 65 to 75 kg, 75 to 85 kg, 85 kg and up)	18	Cord length (short, normal, long, I do not know)
3	Number of pregnancies (0, 1, 2, 3)	19	Mobile phone type (not used, simple, smart)
4	Fetal sex (girl, boy, twin girl, twin boy)	20	Mobile usage (I do not use, 1 to 2 hours, 2 to 4 hours, 4 to 6 hours, more than 6 hours)
5	Fetal weight (less than 2 kg 500 g, between 2 kg 600 g to 3 kg and 700 g, 3 kg, and 800 g to 4 kg and 200 g, more than 4 kg and 300 g)	21	Internet (yes, no)
6	Number of fetuses (1, 2, 3, 4, and more)	22	Internet usage (I do not use, 1 to 2 hours, 2 to 4 hours, 4 to 6 hours, more than 6 hours)
7	Birth week (28 to 32, 33 to 37, 38 to 40, 41, and more)	23	Computer (yes, no)
8	Placenta type (anterior, posterior, fundal, lateral, lower, unknown).	24	Computer usage (I do not use, 1 to 2 hours, 2 to 4 hours, 4 to 6 hours, more than 6 hours)
9	Fibroids (yes, no)	25	Hot foods (low, medium, high, very high)
10	Diabetes (yes, no)	26	Liquids (1 to 1.5 liters, 1.5 to 2 liters, 2 to 2.5 liters, more than 2.5 liters)
11	Hyper Blood pressure (yes, no)	27	Sleep (12 pm to 8 am, 12 pm to 8 am plus 1 hour of sleep per day, 12 pm to 8 am plus 2 hours of sleep per day, 12 pm to 8 am plus 3 hours of sleep per day, and more)
12	Smoking (yes, no)	28	Income (low, medium, high, very high)
13	Alcohol (yes, no)	29	Elemental electric kitchen appliances (yes, no)
14	History of abortion (yes, no)	30	Nutrition (vegetarian, meat, both)
15	Maternal activity during pregnancy (low, medium, high, very high)	31	Stress (very low, sometimes, most of the time, very high)
16	Increased amniotic fluid (yes, no)		

other classes, especially in two-class applications where one class has more instances than the other class. This situation becomes problematic in classification when a class, which is usually an absolute or minority class, is not shown in the data set, in other words, the number of incorrect observations exceeds the correct observations in a class. To deal with imbalanced data set issues, some techniques have been introduced, here we used SMOTE. The Synthetic Minority Oversampling Technique (SMOTE) is an oversampling approach to deal with imbalanced datasets. SMOTE creates new samples from the small class "synthetically", unlike duplicate samples used in traditional over-sampling methods [17].

In this method, for each case in a minority class (for example, a diabetic), the algorithm finds k samples (usually 5) at the closest distance to the minority sample (k nearest neighbors). This distance is obtained from the standard Euclidean distance. In the next step, new hybrid samples (xnew) are produced by the following method. The distance between the variables of a minority sample (xi) and its nearest neighbor (xj) is calculated. This distance is then multiplied by a random number between 0 and 1 (δ) and added to the value of the variables of the minority samples[19].

$$X_{new} = x_i + \left(x_j - x_i\right) \times \delta \quad (2)$$

Implementing this method on the primary data set led to better results, hence, the SMOTE technique was chosen to balance the primary data set.

2- 5- Train the system

Classification is one of the supervised learning methods for predicting data classes which identifies new data classes based on default and predefined classes. Some of the data classification algorithms are SVM, Decision Tree, Naïve Base, and K-Nearest Neighbor. In this article, we have examined different algorithms. Since k-nearest neighbors (KNN) had a better result, we will define it below.

k-nearest neighbors (KNN) algorithm

The k-nearest neighbor (KNN) algorithm is an easy and simple supervised machine learning algorithm that can be used to solve both classification and regression problems. This algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. A sample

is categorized by a majority vote of its neighbors, and this sample is determined in the most public class among k close neighbors. K is a positive integer value and is generally small. The best choice of k depends on the data. In general, a large amount of k reduces the effect of noise on the classification, but the boundary between classes becomes less distinct. Its decision-making idea is very simple, that is, the sample to be tested is the same as the sample category closest to it. It is different compared to the nearest neighbor by extending the nearest neighbor to k in the decision-making stage. The k extension allows this algorithm to achieve and use more information. It eliminates the learning processing process compared to other classification algorithms with distinct training steps[20] In the KNN algorithm, we need k nearest points, for this reason, we should calculate the distance between the input data point and other points in our training data. If we assume x is a point with the coordinates of (x_1, x_2, x_3,…,x_p) and y is a point with coordinates of (y_1,y_2, y_3,…,y_p) then we can achieve the distance between the two points from this formula:

$$d\left(x.y\right)=\sqrt{\sum_{i=1}^{p}\left(x_i-y_j\right)^2} \qquad (3)$$

SVM algorithm

In machine learning, support vector machines also called support vector networks, are supervised learning models that work with learning algorithms that analyze data and identify patterns and are used to classify and analyze regression. SVM has a high degree of generalization accuracy. In SVM, only the data contained in the support vectors form the basis of machine learning and modeling and this algorithm is not sensitive to other parts of the data and its goal is to find the best boundary between the data so that it has the greatest possible distance from all categories (their supporting vectors). In such a way that the two pages are far enough apart to collide with the data. As shown in Figure 2, the goal is to find the two pages that have the most distance, and as a result, the page between the two pages will be the best separator.

If the training points in the form $\left[x_{i\text{æ}}y_i\right]$ and input vector $x_i\in R^n$ And class value $\left(i=1.2.\ldots.N\right)$ and $y_i\in\left\{-1.+1\right\}$ Define, then in the case where the data are linearly separable, the decision rules that are defined and by an optimal plane that separates the binary decision classes are as follows:

$$y=sign\left(\sum_{i=1}^{N} y_i\propto_i\left(x.x_i\right)+b\right) \qquad (4)$$

Where Y is the output of the equation, y_i is the value of the instance class and represents the internal multiplication. The vector x = (x_1, x_2, …, x_n) represents an input data, and the vectors x_i,(i = 1,…, N) represent the backup vectors. And

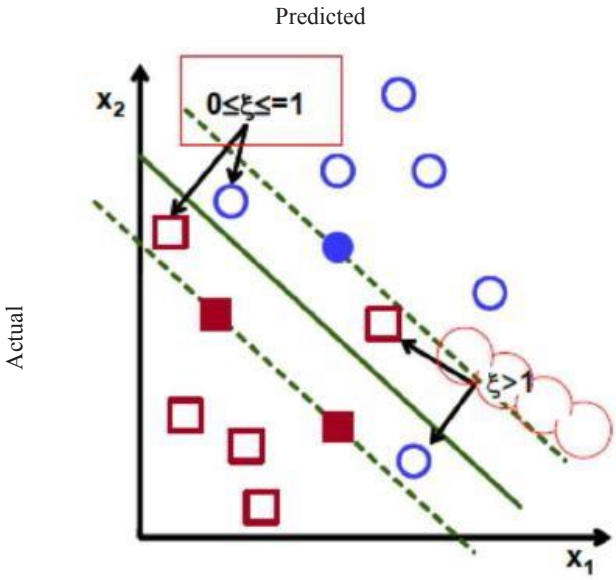

Fig. 2. View of the paging concept in the SVM algorithm

the $\propto_i$ b parameters determine the hyper page. If the data is not linearly separable, the above equation changes to the following equation:

$$y=sign\left(\sum_{i=1}^{N} y_i\propto_i k\left(x.x_i\right)+b\right) \qquad (5)$$

The k (x, x_i) function is a kernel function that generates internal multiplications to create machines with different types of nonlinear decision levels in the data space [21].

2- 6- Discovering and evaluation of rules using association rule mining

Next, we used association rules mining algorithms such as Apriori, Predictive Apriori, and FP-Growth to discover related rules with the wrapped umbilical cord.

Apriori

In data mining, the apriori algorithm is a sort of association rule. Data mining association approach Analysis, also known as association rule extraction, is used to discover the rules of a group of elements[22]. Apriori is one of the best algorithms for exploring dependency rules in this algorithm all subsets of each set are repeated. The basic idea of the Apriori algorithm is finding the maximum set of items: the first step is to simply count the number of items containing an element and find the project set that is greater than or equal to the minimum degree of support. From the second step, start the loop process until the failure to generate a higher dimension of the frequent item sets. The process is as follows: in step k - 1, the set of

Table 2. Confusion matrix and its various modes

	Positive	Negative
Positive	True positive(TP)	False negative(FN)
Negative	False positive(FP)	True negative(TN)

recurring items of dimension K produced by k - 1 is generated by the next set of candidate items of dimension K, and we can count the number of elements in k - 1 dimension[23]. The goal of this algorithm is to find the largest set of items that meet the minimum Support and Confidence. Apriori uses the breadth-first search and hash structure to be able to count candidate items effectively[24].

Predictive apriori

Another approach predictive Apriori can also generate rules; however, it receives unexpected results as it combines both the support and confidence. Apriori uses the property "all subsets of a frequent item set must be frequent; and if an item set is infrequent, then all its supersets must also be infrequent" [25]. Another approach predictive Apriori can also generate rules; however, it receives unexpected results as it combines both the support and confidence [26]. In Predictive Apriori, the confidence evaluation of rules depends on their support. The preferred criterion of this algorithm is how to improve the expected accuracy of unseen data. Once the law has been completed, the preferred measure of Predictive accuracy is the rules by which it is developed how well they are able to generalize and predict test sample labels using training data [24].

FP-growth

This algorithm, which uses a tree structure, first builds this tree and then uses the generated tree to look for a set of repetitive items. Using a compact structure called fp-tree, it maps the entire database space to a small amount of space in the main memory and uses fp-tree when counting the coverage of a set of items in the database. This reduces the amount of time it takes to read the database [27].

Comparison of the results shows that different features can have different effects, and factors such as nutrition, blood pressure, diabetes, smoking, birth week, and the amount of Internet are likely to have more impact on wrapping the umbilical cord around the fetus.

3- Results

The knowledge that is generated in the model learning phase must be analyzed in the evaluation phase to determine its value and then determine the efficiency of the model learning algorithm. To evaluate the results, criteria such as confusion matrix, accuracy, Recall, F-measure, support, and reliability are used, which we review.

Confusion matrix

A confusion matrix contains information about the actual classification predicted by a classification system. The performance of these systems is usually evaluated using data in the matrix. Confusion Matrix is a visual evaluation tool used in machine learning. The columns of a Confusion Matrix represent the prediction class results, and the rows represent the actual class results[28].

Each element of the matrix is as follows:

TN: indicates the number of records whose actual category is negative and the classification algorithm has correctly identified the category as negative.

TP: indicates the number of records whose actual category is positive and the classification algorithm has correctly identified the category as positive.

FP: indicates the number of records whose actual category is negative and the category classification algorithm has incorrectly detected them as positive.

FN: indicates the number of records whose actual category is positive and the category classification algorithm has incorrectly detected negative.

Accuracy

The most important criterion for determining the performance of a classification algorithm is the accuracy or classification rate, which calculates the total accuracy of classification. In fact, this criterion is the most famous and general criterion for calculating the efficiency of classification algorithms, which shows, the designed category what percentage of the total set of test records is properly categorized. Accuracy is one of the most widely used criteria in the field of machine learning. Accuracy is a useful measure of predictive success when classes are highly imbalanced. This criterion is often used to measure performance. Measures the number of objects classified as precision relative to all positive objects. This measurement indicates the algorithm's ability to make truly accurate predictions without erroneously accepted objects [18]. Accuracy (P) is defined as the real positive number (Tp) over the real positive number and the false positive number (Fp).

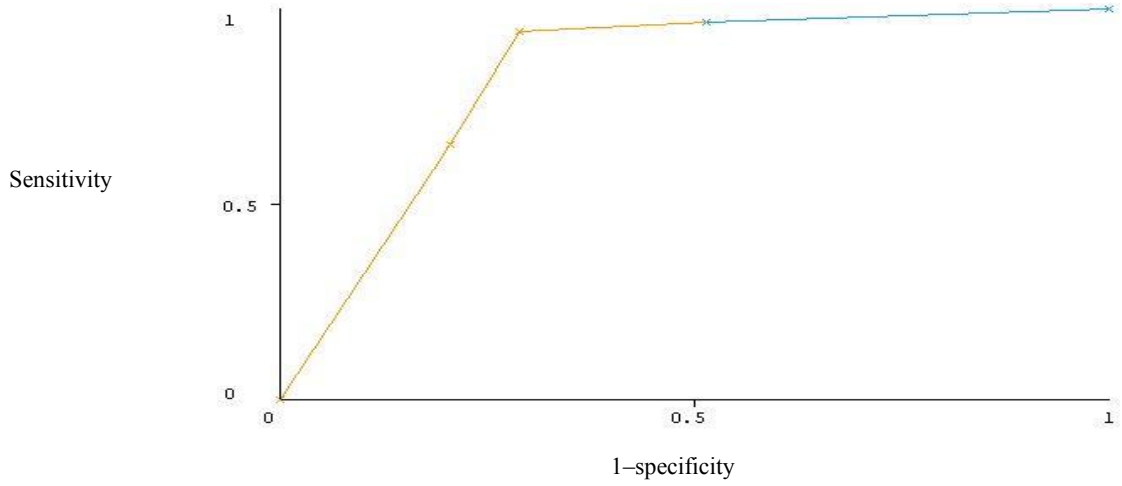

Fig. 3. ROC plot for knn classifier

$$P = \frac{T_P}{T_P + F_P} \tag{6}$$

Recall

Reading (R) is defined as a real positive number (Tp) to a real positive number and a false negative number (Fn). The formula of recall is as follows.

$$R = \frac{T_P}{T_P + F_N} \tag{7}$$

In addition to accuracy, it is important to consider Recall and accuracy. A data set can potentially rank with high accuracy, but show poor values in Recalling and accuracy. This is the result of an algorithm that rejects or accepts data points too much. In general, when t increases, the Recall r (t) increases with t, while the accuracy of p (t) decreases[29].

F-Measure

It is a good parameter for evaluating the quality of classification and also describes the weighted average between the two quantities Precision and Recall. For a classification algorithm, under ideal conditions, the value of this quantity is equal to 1, and in the worst case, it is equal to zero. This parameter is calculated according to the following equation:

$$f = 2 . \frac{percision.recall}{percision + recall} \tag{8}$$

ROC

One way to evaluate and check the performance of binary classification is the "Receiver Operating Characteristic" or ROC. The performance of "Binary Classifier" algorithms is usually measured by indicators called "Sensitivity" or "Recall". But in the ROC diagram, both of these indicators are combined and displayed as a curve. ROC curves are often used to evaluate the performance of classification algorithms or to generate string data. It is a diagram that displays the ability to evaluate a binary classification system with a variable detection threshold. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity or recall. The false-positive rate can be calculated as (1 − specificity). The ROC curve on our balanced data set for the knn classifier is shown in Figure 3.

A comparison of diagrams of several algorithms on our balanced data set is shown in Figure 4. As can be seen, KNN has the best result compared to Naïve base, Bagging, and Random forest algorithms.

The primary data set was created with a total number of 140, of which 97 samples are without wrapped umbilical cords and 43 samples are wrapped umbilical. After pre-processing it, and applying SMOTE, the total number of data set samples is183 , of which 97 samples have no wrapped umbilical cord and 86 samples have wrapped umbilical cord. To evaluate the accuracy and correctness of the classification, it was implemented with the SVM, Naïve Base, Random Forest, KNN, and Bagging algorithm in Weka software.

Since fewer works have been done to predict the umbilical cord wrapping around the fetus using data mining and to estimate its probability of occurrence, and we used our original native data set, it is not possible to compare the results with other works. Therefore, we used different classification methods to be able to evaluate different parameters of our

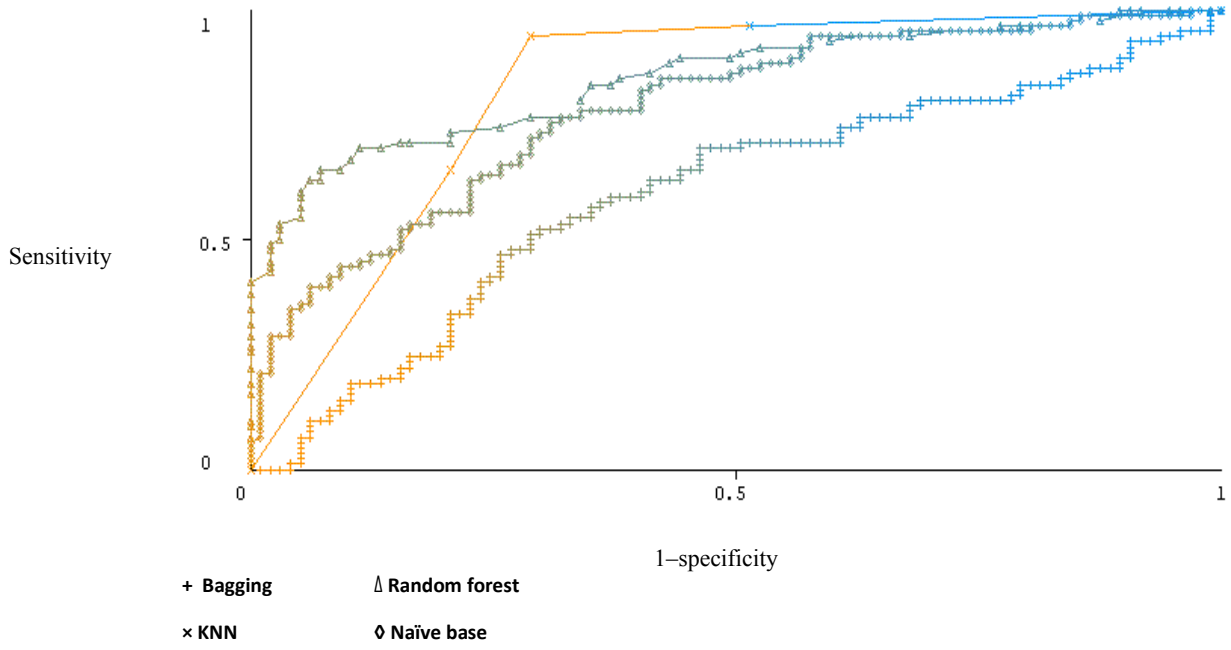

Fig. 4. Roc plot comparison for KNN, Naïve base, Bagging, Random forest

Table 3. The results of algorithms on the primary data set

Algorithms	Average Accuracy(%)	ROC Area	Average F-Measure	Average Recall	Average Precision	Average FP Rate	Average TP Rate
SVM	70	51.8	59.5	70.0	69.0	66.4	0.70
Naïve Base	62	54.4	60.9	62.9	59.8	57.9	62.9
Bagging	65	49.1	56.6	65.0	53.6	68.6	65.0
Random Forest	67	57.9	56.0	67.9	47.7	69.9	67.9
KNN	66	56.7	65.2	66.4	64.5	51.1	66.4

work apply different methods, and compare them with each other. The results of the primary data set are shown in Table 3.

As shown in Table 3, the accuracy, Recall, and ROC values, results showed. The SVM algorithm had the best result with 70% accuracy.

The results of the balanced data set by SMOTE are shown in Table 4.

As shown in Table 4, the accuracy of KNN is the best. The results are improved compared to the primary data set. The accuracy was improved by the SMOTE balancing method. This amount increased from 66% to 81%.

The results of the confusion matrix on the balanced data set by SMOTE are shown in Table 5.

Since predicting wrapped classes is more important a classifier is more valuable to us, which can classify wrapped class data better than unwrapped ones. The weight of the wrapped class is higher for us, so because in terms of the confusion matrix KNN is better at responding, this is a more acceptable result for us.

4- Conclusion

In this study, using the opinion of obstetricians and gynecologists, first, a questionnaire was designed to assess the different conditions and features of pregnant mothers in the twisting umbilical cord around the fetus. After evaluating the questionnaire with statistical methods and reviewing various methods of data balancing, the SMOTE technique has been selected to balance the initial data. We used the

Table 4. Results of algorithms on data set balanced by SMOTE technique

Algorithms	Average Accuracy(%)	ROC Area	Average F-Measure	Average Recall	Average Precision	Average FP Rate	Average TP Rate
KNN	81	85.0	79.5	79.8	80.4	16.6	82.0
Random Forest	79	83.4	73.2	73.2	73.3	21.4	79.8
Bagging	70	74.5	70.4	70.5	70.5	30.2	70.5
SVM	71	71.4	71.6	71.6	71.6	28.8	71.6
Naïve Base	67	72.8	67.7	67.8	67.7	32.5	67.8

Table 5. confusion matrix of algorithms on data set balanced by SMOTE technique

Algorithms	TP	FP	FN	TN
KNN	69	5	28	81
Naïve Base	68	30	29	56
Bagging	74	31	23	55
Random Forest	86	26	11	60
SVM	72	27	25	59

KNN algorithm in Weka to determine the accuracy of the classification. The accuracy of the initial data set with the KNN algorithm was about 66% and with the balanced data set by SMOTE method, it had an accuracy of 81%. In the next step, by applying algorithms to detect association rules such as Apriori, Predictive Apriori, and FP-growth, the effect of the mentioned factors on the wrapping of the umbilical cord around the fetus was investigated and rules were extracted. Although medically it is not possible to exactly predict the occurrence of the umbilical cord around the fetus, we tried to create a data set with various features that affect the condition of the mother and fetus to investigate it and some factors that can help the mother and fetus to improve their condition during pregnancy and with proper information and education, the impact of these factors can be reduced.

References

[1] A. Linden, P.R. Yarnold, Using data mining techniques to characterize participation in observational studies, Journal of Evaluation in Clinical Practice, 22(6) (2016) 839-847.

[2] J.F. Clapp, B. Lopez, S. Simonean, Nuchal cord and neurodevelopmental performance at 1 year, The Journal of the Society for Gynecologic Investigation: JSGI, 6(5) (1999) 268-272.

[3] G. Hankins, R.R. Snyder, J.C. Hauth, L. Gilstrap 3rd, T. Hammond, Nuchal cords and neonatal outcome, Obstetrics and gynecology, 70(5) (1987) 687-691.

[4] D. Shere, A. Anyaegbuam, Prenatal ultrasonographic morphologic assessment of the umblical cord: a review, Obstet Gynecol Surv (USA), 52(8) (1997) 506-523.

[5] G.R. Gutiérrez, S.E. Razo, A.C. Curiel, A.P.P. de León, Color Doppler flowmetry values in fetuses with nuchal cord encirclement, Ginecologia y obstetricia de Mexico, 68 (2000) 401-407.

[6] A. Funk, W. Heyl, R. Rother, M. Winkler, W. Rath, Subpartal diagnosis of umbilical cord encirclement using color-coded Doppler ultrasonography and correlation with cardiotocographic changes during labor, Geburtshilfe und Frauenheilkunde, 55(11) (1995) 623-627.

[7] Tehran University Medical Journal, 63(12) (2005) 991-997.

[8] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, in: KDD, 1996, pp. 82-88.

[9] S. Pereira, F. Portela, M.F. Santos, J. Machado, A. Abelha, Predicting type of delivery by identification of obstetric risk factors through data mining, Procedia Computer Science, 64 (2015) 601-609.

[10] N. Chamidah, I. Wasito, Fetal state classification from cardiotocography based on feature extraction using hybrid K-Means and support vector machine, in: 2015 international conference on advanced computer science and information systems (ICACSIS), IEEE, 2015, pp. 37-41.

[11] H.-Y. Chen, C.-H. Chuang, Y.-J. Yang, T.-P. Wu, Exploring the risk factors of preterm birth using data mining, Expert systems with applications, 38(5) (2011) 5384-5387.

[12] R.W. Bendon, S.P. Brown, M.G. Ross, In vitro umbilical cord wrapping and torsion: possible cause of umbilical blood flow occlusion, The Journal of Maternal-Fetal & Neonatal Medicine, 27(14) (2014) 1462-1464.

[13] t. ashraf, Umbilical cord entanglement and intrapartum complications, Journal of Shahrekord Uuniversity of Medical Sciences, 6(2) (2004) 44-49.

[14] L. Mallick, J. Yourkavitch, C. Allen, Trends, determinants, and newborn mortality related to thermal care and umbilical cord care practices in South Asia, BMC pediatrics, 19(1) (2019) 1-16.

[15] A. Krzyżanowski, M. Kwiatek, T. Gęca, A. Stupak, A. Kwaśniewska, Modern ultrasonography of the umbilical cord: prenatal diagnosis of umbilical cord abnormalities and assessement of fetal wellbeing, Medical science monitor: international medical journal of experimental and clinical research, 25 (2019) 3170.

[16] G.A. Pradipta, R. Wardoyo, A. Musdholifah, I.N.H. Sanjaya, Improving classifiaction performance of fetal umbilical cord using combination of SMOTE method and multiclassifier voting in imbalanced data and small dataset, Int. J. Intell. Eng. Syst., 13(5) (2020) 441-454.

[17] M. Shuja, S. Mittal, M. Zaman, Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE, in: Advances in computing and intelligent systems, Springer, 2020, pp. 195-211.

[18] R. Heale, A. Twycross, Validity and reliability in quantitative studies, Evidence-based nursing, 18(3) (2015) 66-67.

[19] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, D. Khalili, The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes, Medical decision making, 36(1) (2016) 137-144.

[20] W. Xing, Y. Bei, Medical health big data classification based on KNN classification algorithm, IEEE Access, 8 (2019) 28808-28819.

[21] S. Vijayarani, S. Dhayanand, Data mining classification algorithms for kidney disease prediction, Int J Cybernetics Inform, 4(4) (2015) 13-25.

[22] P. Edastama, A.S. Bist, A. Prambudi, Implementation Of Data Mining On Glasses Sales Using The Apriori Algorithm, International Journal of Cyber and IT Service Management, 1(2) (2021) 159-172.

[23] C. Wang, X. Zheng, Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint, Evolutionary Intelligence, 13(1) (2020) 39-49.

[24] B.M. Patil, R.C. Joshi, D. Toshniwal, Classification of type-2 diabetic patients by using Apriori and predictive Apriori, International Journal of Computational Vision and Robotics, 2(3) (2011) 254-265.

[25] I.H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science, 2(3) (2021) 160.

[26] T. Scheffer, Finding association rules that trade support optimally against confidence, Intelligent Data Analysis, 9(4) (2005) 381-395.

[27] M. Kavitha, S. Selvi, Comparative study on Apriori algorithm and Fp growth algorithm with pros and cons, International Journal of Computer Science Trends and Technology (I JCS T)–Volume, 4 (2016).

[28] J. Xu, Y. Zhang, D. Miao, Three-way confusion matrix for classification: A measure driven view, Information sciences, 507 (2020) 772-794.

[29] S. Stiernborg, S. Ervik, Evaluation of Machine Learning Classification Methods: Support Vector Machines, Nearest Neighbour and Decision Tree, in, 2017.

HOW TO CITE THIS ARTICLE

N. Abedini, F. Moayedi, S. E. Dashti, Data mining approach for prediction umbilical cord wrapping around the fetus and investigating effective factors, AUT J. Model. Simul., 54(2) (2022) 131-140.

DOI: 10.22060/miscj.2022.21360.5283