NATURENGS (MTU Journal of Engineering and Natural Sciences) Volume 4, Issue 1 (2023) 16-23

Malatya Turgut Özal University
Journal of Engineering and Natural Sciences

https://dergipark.org.tr/tr/pub/naturengs

Research Article

Prediction of Maternal Health Risk with Traditional Machine Learning Methods

Hursit Burak MUTLU[1], Nadide YÜCEL[1], Fatih DURMAZ[1]*, Emine CENGİL[2], Muhammed YILDIRIM[1]

[1]Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Malatya Turgut Ozal University, Malatya, Turkey.

[2]Department of Computer Engineering, Faculty of Engineering and Architecture, Bitlis Eren University, Bitlis, Turkey.

ARTICLE INFO

Received: 2023-05-05

Accepted: 2023-06-01

DOI: 10.46572/naturengs.1293185

ABSTRACT

In risky pregnancy, various diseases such as heart, lung, kidney, high blood pressure, diabetes and liver that pregnant women have before may aggravate the expectant mother's condition during pregnancy. By analyzing medical parameters such as maternal age, heart rate, blood oxygen level, blood pressure, body temperature, and the values corresponding to these parameters, information on risk intensity can be estimated for some patients. It is possible to reduce such pregnancy-related complications by classifying risk factors early in symptoms. It is possible to benefit from machine learning methods in determining maternal risk health. Therefore, in this study, six different machine learning methods were used to determine maternal risk health. The results obtained in these methods were compared with each other and it was observed that the most successful method in estimating maternal risk health was Decision Tree. The accuracy value obtained in the Decision Tree method was 89.16%. The lowest accuracy rate among the methods used in the paper was obtained in the k-nearest neighbors (KNN) method with 68.47%.

Keywords: *Artificial Intelligence, Classifiers, KNN, Machine Learning, Maternal Health*

Introduction

Maternal health is the physical, mental, and emotional well-being of the mother during pregnancy, childbirth, and all postpartum periods. Maternal morbidity and mortality rates during pregnancy are important health data, as they provide information on accessibility to maternal and other medical resources. Pregnancy complications such as hypertension, diabetes, bleeding, and premature birth are among the leading causes of maternal death. It is important to detect pregnancy-related risks before they cause premature birth or death and to provide treatment for them. Machine learning methods have an important place in determining maternal health risks [1, 2]. By analyzing the health data and risk factors of a pregnant woman with machine learning methods, the risk level can be monitored and estimated. In this way, the use of models based on machine learning is thought to be effective in reducing maternal mortality rates as a result of complications arising from changes in risk factors [3]. This study, it is aimed to use machine learning methods for the estimation of maternal health risk intensity level with the classification approach in the analysis of risk factors. During pregnancy, age, systolic and diastolic blood pressure (BP), body temperature (BodyTemp), pulse (heart rate), blood oxygen (BO), or breathing speed (BS) are among the risk factors that should be measured. In order to protect the health of the pregnant woman, taking these factors into account, timely identification of risks with machine learning algorithms can help reduce both maternal and infant mortality rates [4].

Pawar et al. [1] used machine learning methods in their study to determine the risk to maternal health. In this study, the researchers stated that machine learning methods can be used to determine maternal health risks. In addition, the researchers stated that the model they developed was more successful than traditional machine learning models. The accuracy value reached by the researchers in the developed model was 70.21%.

Ahmed et al. [4] developed an IoT-based maternal health system to detect maternal health risk factors. As a result of comparing the machine learning algorithm among some groups in the classification of the risk level in the analysis of maternal health risk factors, they obtained 97% accuracy by using the modified decision tree algorithm. They implemented machine learning algorithms in both Weka and using Python. They collected maternal health risk factor data from hospitals and maternity clinics in Bangladesh and classified them into three categories as low, medium, and high-risk levels and classified a total of 1014 data.

Ahmed et al. [5] used machine learning algorithms to discover the risk level in pregnancy based on risk factors in their research. They used the Pima-Indian diabetes dataset for risk factor analysis and comparison of some machine learning algorithms. They showed that the Logistic Model Tree (LMT) gives the highest accuracy when classifying and estimating the risk level. In addition, the data of a few selected pregnant women were collected via IoT-enabled devices, and the same process was applied to this dataset, as well as LMT. In the comparison results, they showed that the risk estimation

*Corresponding author : fth.drmz23@gmail.com
ORCID : 0009-0004-8363-1517

was the same for the current Pima-Indian diabetes dataset and the actual dataset.

Umoren et al. [3] examined risk estimation models for maternal mortality and made risk estimations applicable by using the model based on the decision tree classification approach. With the decision tree approach, 89.2% of 100% of clinical data samples were correctly classified and compared with the Support Vector Machine (SVM) model. In this analysis, they achieved an accuracy of 89.2% and 69.5%.

Rai et al. [6] suggested a method of evaluating the parameters of increased maternal and infant mortality risk levels as a result of pregnancy complications, according to a questionnaire conducted with experts in the field. They selected 14 features from the data taken from 117 pregnant women and used the feed-forward feature of artificial neural network (ANN) and Naïve Bayes (NB) algorithms to estimate the risk level. The number of data in the training of each class was determined as 80% and 20% for the test. While they achieved 80% accuracy with ANN, they reached an accuracy rate of 70% with NB. In addition, the researchers increased the level of accuracy by using a hybrid algorithm.

The rest of the article includes Materials and Method, Application Results, and Conclusion sections.

Materials and Methods

In this section, the machine learning methods and dataset used in the study to predict maternal risk health are examined.

Classifiers

In this study, which was conducted to determine maternal risk health, 6 different machine learning methods were used. In this study, the first method used for dataset analysis is the decision trees algorithm. Decision trees are a popular machine learning technique that has the ability to analyze datasets by generating a set of decision rules. These rules are organized in a tree-like model where each feature creates a decision node. The data travels through the tree and eventually ends up at a decision node. The decision tree model was applied to the training dataset and the parameters of the model were optimized by cross-validation method. The decision tree algorithm has advantages such as being able to work with categorical and continuous data, being easy to understand, and having fast computation times [7].

In this study, LightGBM (Light Gradient Boosting Machine) is another method used in the analysis of the dataset. LightGBM is an implementation of Gradient Boosting algorithms and has the advantage of being able to process large-scale datasets and offer faster training times. The LightGBM model was implemented on the training dataset and the hyperparameters of the model were optimized. The LightGBM algorithm provides low memory usage, high speed, and better accuracy, providing an effective classification method, especially for high-dimensional and large datasets [8].

Another method used in the paper is the CatBoost. CatBoost is one of the Gradient Boosting algorithms that can work with categorical and numerical features. CatBoost has advantages such as the automatic processing of categorical data and better accuracy. The CatBoost algorithm has the potential to offer faster training times and higher accuracy. Moreover, its ability to work directly with categorical data makes this algorithm an effective and suitable classification method in various application areas [9].

A stronger and more stable model is produced using the ensemble learning technique known as Random Forest, which integrates different decision trees. The Random Forest algorithm helps to improve the generalization ability by increasing the stability and accuracy of the model. The success of the algorithm can be measured by the accuracy provided in the classification process on the dataset. This makes Random Forest an effective and convenient classification method in various application areas [10].

Gradient Boosting Machines (GBM) is an ensemble learning technique based on decision trees. This technique involves combining a set of weak trees to form a stronger model. The GBM algorithm has the potential to offer high accuracy by increasing its generalization ability. The success of the algorithm can be measured by the accuracy obtained in the classification process on the dataset [11].

Another method used in the analysis of the dataset for the determination of maternal risk health is k-Nearest Neighbors (KNN). KNN is a simple and effective machine learning algorithm that works based on the distance between samples and determines the class of new samples by a majority vote of their nearest neighbors. The KNN algorithm draws attention, especially with its low computational cost and complexity, which provides an advantage in terms of understanding and implementation. The success of the algorithm can be measured by the accuracy it provides in the classification process of the dataset [12].

Dataset

The dataset used in the study to determine maternal risk health is a public dataset published on the internet [4, 13]. Researchers collected data from different places using IoT-based risk monitoring systems. The dataset consists of Age, SystolicBP, DiastolicBP, BS, and heart rate features. The last feature in the dataset is the label of the data. The dataset consists of three classes. These classes are "high risk", "middle risk" and "low risk". 10 examples of the dataset are in Table 1.

Table 1. Examples from the dataset

Age	SystolicBP	DiastolicBP	BS	Body Temp	Heart Rate	Risk Level
25	130	80	15.00	98.0	86	high risk
35	140	90	13.00	98.0	70	high risk
29	90	70	8.00	100.0	80	high risk
30	140	85	7.00	98.0	70	high risk
35	120	60	6.10	98.0	76	low risk
23	140	80	7.01	98.0	70	high risk
23	130	70	7.01	98.0	78	mid risk
35	85	60	11.00	102.0	86	high risk
32	120	90	6.90	98.0	70	mid risk
42	130	80	18.00	98.0	70	high risk

Age: Age of the pregnant woman,

SystolicBP: The upper value of blood pressure,

DiastolicBP: Low blood pressure value,

BS: Breathing speed value,

BodyTemp: Pregnant woman's body temperature,

HeartRate: Heart rate,

Risk Level (Classes): “high risk”, “middle risk” and “low risk”.

Data preprocessing steps were applied before the data in the dataset was classified. In this way, more successful results were obtained in machine learning methods. The most basic data preprocessing step applied at this stage is the data standardization step.

Application Results

Decision Tree, LightGBMClassifier, CatBoost, Random Forest, Gradient Boosting Machines, and KNN classifiers were used to classify the data in the dataset consisting of six different features obtained to assess maternal health risk. 80% of the maternal health risk dataset was reserved for train and 20% for testing. The confusion matrix and accuracy ratio of the dataset classified in six different classifiers were extracted. In the paper, various parameters were employed to evaluate how well the models performed [14,15].

The first model used in the study to predict maternal risk health is the Decision Tree. The confusion matrix of the Decision Tree method is given in Figure 1.

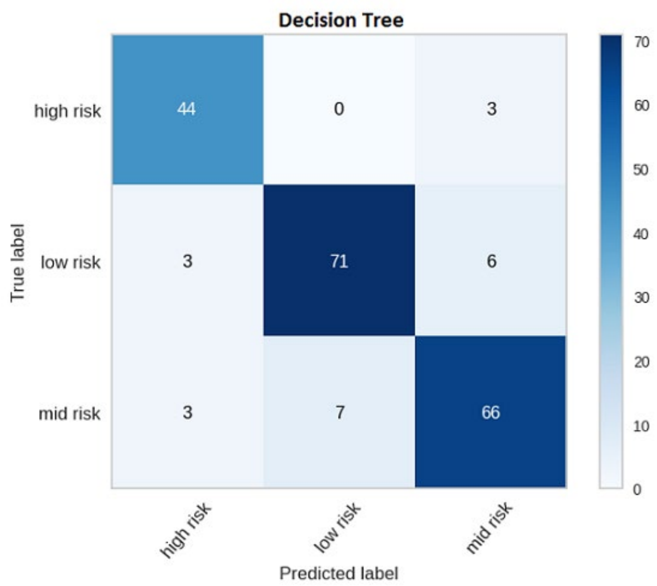

Figure 1. Confusion Matrix of Decision Tree

Examining the Decision Tree confusion matrix shown in Figure1, an accuracy rate of 89.16% was obtained in classifying the test data. The Decision Tree classifier predicted 181 correctly and 22 incorrectly out of 203 data allocated for testing. The class with the lowest performance of the Decision Tree classifier is the "mid risk" class.

The performance measurement metrics of the Decision Tree method are presented in Table 2.

Table 2. Decision Tree performance metrics (%).

	Accuracy	Sensitivity	Specificity	FPR	FDR	FNR	F1
High Risk	93.61	88.00	98.03	1.96	6.38	12.00	90.72
Low Risk	88.75	91.02	92.80	7.20	11.25	8.97	90.00
Mid Risk	86.84	88.00	92.18	7.81	13.15	12.00	87.00

When the performance metrics of the Decision Tree classifier are examined, the "high risk" class with the highest accuracy rate of 93.62%, and the "mid risk" class with the lowest accuracy rate of 86.84%.

The confusion matrix of the LightGBM Classifier is shown in Figure 2.

Examining the Light GBM Classifier confusion matrix shown in Figure 2, an accuracy rate of 84.24% was obtained in classifying test data. The Light GBM predicted 171 correctly and 32 incorrectly out of 203 data reserved for testing. The class with the lowest performance of the Light GBM is the “low risk” class.

The performance measurement metrics obtained in the Light GBM are presented in Table 3.

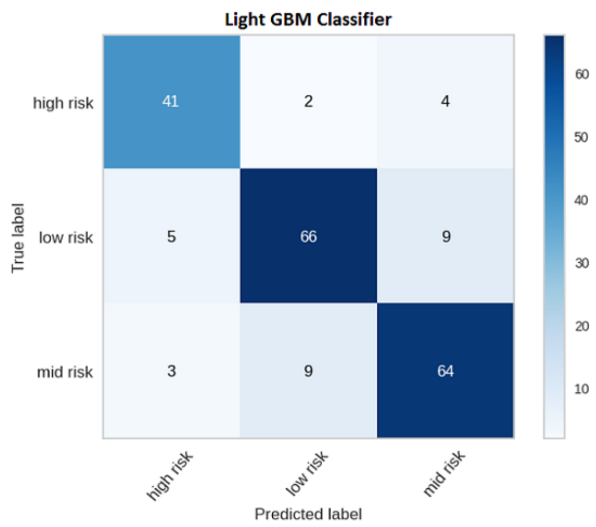

Figure 2. Confusion Matrix of Light GBM

Table 3. Light GBM Classifier performance metrics (%)

	Accuracy	Sensitivity	Specificity	FPR	FDR	FNR	F1
High Risk	87.27	83.67	96.10	3.89	12.76	16.32	85.41
Low Risk	82.50	85.71	88.88	11.11	17.50	14.28	84.07
Mid Risk	84.21	83.11	90.47	9.52	15.78	16.88	83.66

When the performance metrics of the Light GBM classifier are examined, the "High Risk" class with the highest accuracy rate of 87.23%, and the "Low Risk" class with the lowest accuracy rate of 82.50%.

The confusion matrix obtained in the CatBoost classifier is shown in Figure 3.

Examining the CatBoost confusion matrix shown in Figure 3, an accuracy rate of 83.74% was obtained in classifying test data. The CatBoost classifier predicted 170 correct and 33 incorrectly out of 203 test data. The class with the lowest performance of the CatBoost classifier is the “low risk” class.

Table 4 shows the performance measurement parameters attained by the CatBoost classifier.

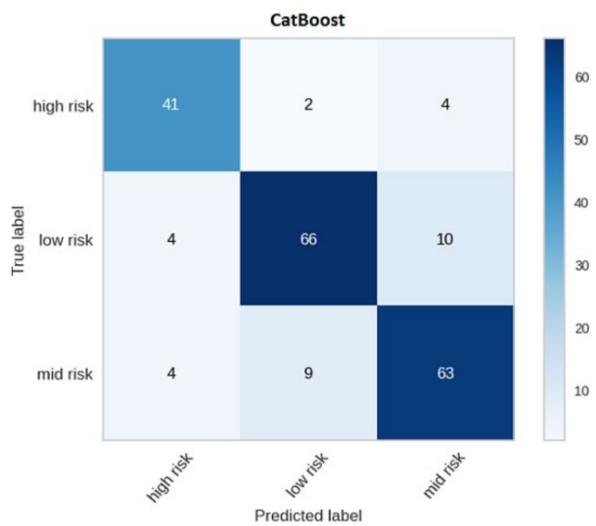

Figure 3. Confusion Matrix of CatBoost

Table 4. CatBoost performance metrics (%)

	Accuracy	Sensitivity	Specificity	FPR	FDR	FNR	F1
High Risk	87.23	83.67	96.10	3.89	12.76	16.32	85.41
Low Risk	82.50	85.71	88.88	11.11	17.50	14.28	84.07
Mid Risk	82.89	89.68	81.81	82.89	17.10	18.18	82.35

When the performance metrics of the CatBoost classifier are examined, the "High Risk" class with the highest accuracy rate of 87.23%, and the "Low Risk" class with the lowest accuracy rate of .82.50%.

The confusion matrix obtained in the Random Forest Classifier is shown in Figure 4.

Examining the Random Forest confusion matrix shown in Figure 4, an accuracy rate of 81.28% was obtained in classifying the test data. The Random Forest classifier predicted 165 correctly and 38 incorrectly out of 203 test data. The class with the lowest performance of the Random Forest classifier is the "low risk" class.

The performance measurement metrics obtained in the Random Forest classifier are presented in Table 5.

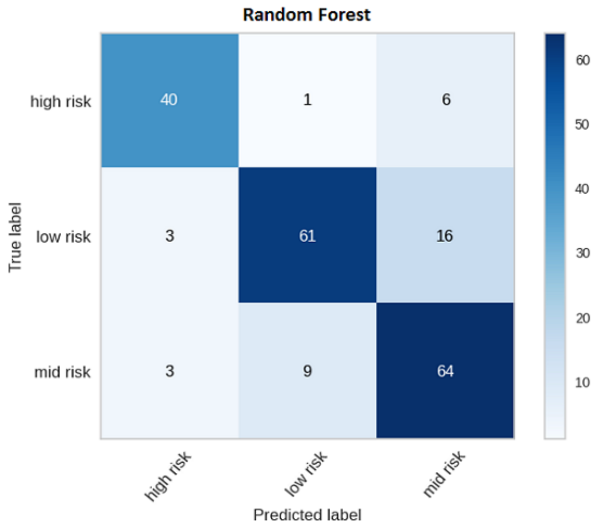

Figure 4. Confusion Matrix of Random Forest

Table 5. Random Forest performance metrics (%).

	Accuracy	Sensitivity	Specificity	FPR	FDR	FNR	F1
High Risk	85.10	86.95	95.54	4.45	14.89	13.04	86.02
Low Risk	76.25	85.91	85.60	14.39	23.75	14.08	80.79
Mid Risk	84.21	74.41	89.74	10.25	15.78	25.58	79.01

When the performance metrics of the Random Forest classifier are examined, the "high risk" class with the highest accuracy rate of 85.11%, and the "low risk" class with the lowest accuracy rate of .76.25%.

The confusion matrix obtained in the Gradient Boosting Machines classifier is shown in Figure 5.

Examining the Gradient Boosting Machines confusion matrix shown in Figure 5, an accuracy rate of 73.89% was obtained in classifying test data. The Gradient Boosting Machines classifier predicted 150 correctly and 53 incorrectly out of 203 test data. The class with the lowest performance of the Gradient Boosting Machines classifier is the "mid risk" class.

Table 6 shows the Gradient Boosting Machines classifier's performance measuring data.

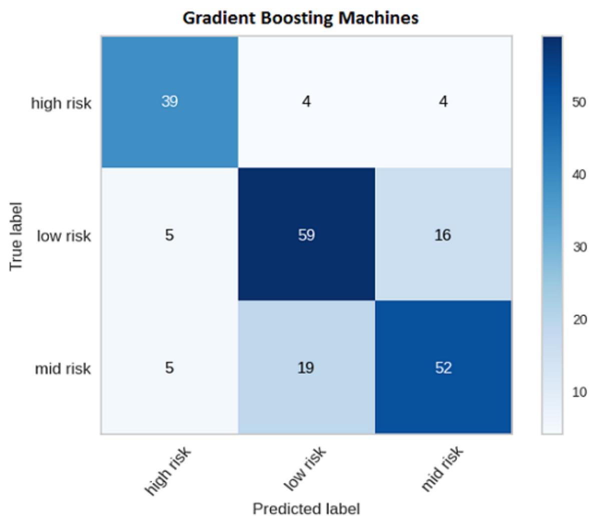

Figure 5. Confusion Matrix of Gradient Boosting Machines

Table 6. Gradient Boosting Machines performance metrics (%).

	Accuracy	Sensitivity	Specificity	FPR	FDR	FNR	F1
High Risk	82.97	74.41	89.74	10.25	15.78	25.58	79.01
Low Risk	73.75	71.95	82.64	17.35	26.25	28.04	72.83
Mid Risk	68.42	72.22	81.67	18.32	31.57	27.77	70.27

When the performance metrics of the Gradient Boosting Machines classifier are examined, the "high risk" class with the highest accuracy rate of 82.98%, and the "mid risk" class with the lowest accuracy rate of .68.42%.

The confusion matrix obtained in the KNN classifier is shown in Figure 6.

Examining the KNN confusion matrix shown in Figure 6, an accuracy rate of 68.47% was obtained in classifying the test data. The KNN classifier predicted 139 correctly and 64 incorrectly out of 203 test data. The class with the lowest performance of the KNN classifier is the "mid risk" class.

The performance measurement metrics obtained in the KNN classifier are presented in Table 7.

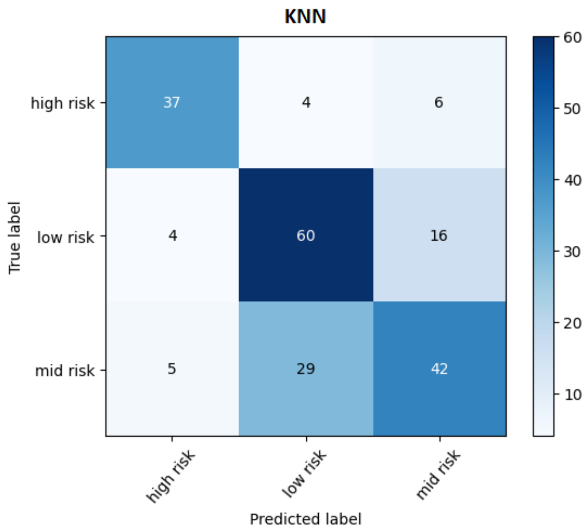

Figure 6. Confusion Matrix of KNN

Table 7. KNN performance metrics (%).

	Accuracy	Sensitivity	Specificity	FPR	FDR	FNR	F1
High Risk	78.72	80.43	93.63	6.36	21.27	19.56	79.56
Low Risk	75.00	64.51	81.81	18.18	25.00	35.48	69.36
Mid Risk	55.26	65.62	75.53	24.46	44.73	34.37	60.00

When the performance metrics of the KNN classifier are examined, the "high risk" class with the highest accuracy rate of 78.72%, and the "mid risk" class with the lowest accuracy rate of 55.26%.

Table 8 presents the accuracy results from the six classifiers used in the study.

Table 8. Accuracy rates of classifiers (%)

Decision Tree	Light GBM	CatBoost	Random Forest	Gradient Boosting Machines	KNN
89.16	84.24	83.74	81.28	73.89	68.47

The Decision Tree classifier, which provided the highest accuracy, predicted 181 correctly and 22 incorrectly out of 203 data reserved for testing. When the performance metrics of the Decision Tree classifier are examined, the "high risk" class with the highest accuracy rate of 93.62%, the "low risk" class with an accuracy rate of 88.75%, and the "mid risk" class with the lowest accuracy rate of .86.84%.

The classifier that provides the lowest accuracy is KNN. KNN predicted 139 correctly and 64 incorrectly out of 203 test data. When the performance metrics of the KNN classifier are examined, the "high risk" class with the highest accuracy rate of 78.72%, the "low risk" class with 75% accuracy, and the "mid risk" class with the lowest accuracy rate of .55.26%.

Diagnosis and diagnosis of maternal health risk is considered very important by experts. Biomedical data are essential for experts to identify the disease, ascertain at what stage it is, and ascertain the course of treatment. The results obtained in the Decision Tree classifier and other classifiers are compared in Figure 7.

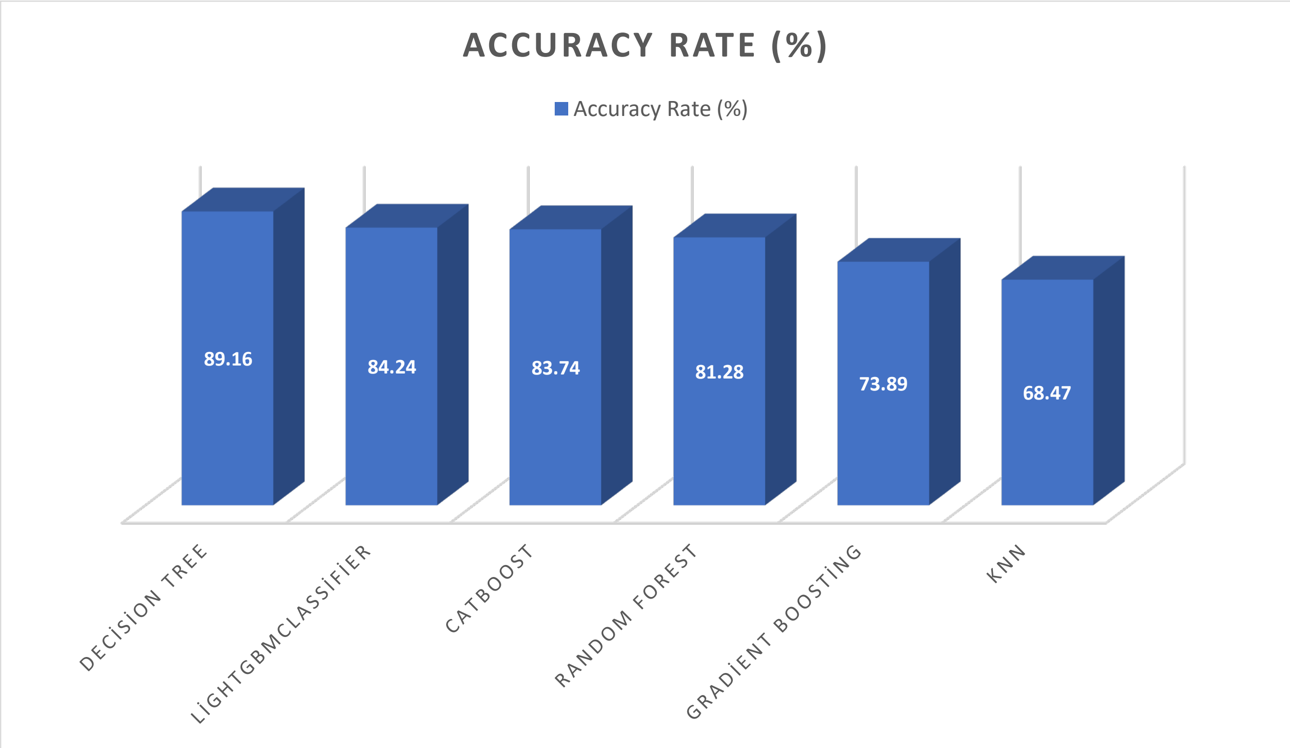

Figure 7. Accuracy rates of classifiers

In the classification process of three different conditions obtained from the maternal health risk dataset, the highest accuracy value was obtained in the Decision Tree classifier with a value of 89.16%. This accuracy rate was followed by Light GBM classifier at 84.24%, CatBoost at 83.74%, Random Forest at 81.28%, Gradient Boosting Machines at 73.89% and KNN at 68.47%.

Conclusions

There are some risks for the pregnant woman in the early stages of pregnancy and the current pregnancy process, depending on parameters such as age, number of births, birth frequency, socio-economic level, and alcohol and tobacco use. Risk factors that constitute various diseases such as hypertension, various heart diseases, and lung and kidney diseases that the expectant mother faces are determined and categorized with the help of medical experts. In order to categorize and forecast the amount of risk in an existing dataset for early identification of symptoms associated with these risk factors, various machine learning techniques are applied. Therefore, this study, it was aimed to predict maternal health risks by using computer-aided classifiers. In the proposed model, the Decision Tree classifier has shown great success compared to other classifiers. This classifier has been compared with other classifiers in the literature. 89.16% accuracy value was obtained in the proposed Decision Tree classifier.

Acknowledgments

The authors thank the owners of the database for sharing their data.

Declaration of Competing Interest

The authors declare that there is no conflict of interest in the study.

Author Contribution

The authors contributed equally to the article.

References

[1] Pawar, L. et al. (2022) A Robust Machine Learning Predictive Model for Maternal Health Risk. in 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE. 882-888.

[2] Varshavsky, J., et al. (2020) Heightened susceptibility: A review of how pregnancy and chemical exposures influence maternal health. Reproductive toxicology, 92: 14-56.

[3] Umoren, I., et al., Modeling and Prediction of Pregnancy Risk for Efficient Birth Outcomes Using Decision Tree Classification and Regression model.

[4] Ahmed, M. and M.A. Kashem. IoT based risk level prediction model for maternal health care in the context of Bangladesh. in 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI). 2020. IEEE.

[5] Ahmed, M., et al. Review and analysis of risk factor of maternal health in remote area using the internet of things (IoT). in InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019. 2020. Springer.

[6] Rai, S.K. and K. Sowmya (2018) A review on use of machine learning techniques in diagnostic health-care. Artificial Intelligent Systems and Machine Learning, 10(4): 102-107.

[7] Quinlan, J.R. (1987) Simplifying decision trees. International journal of man-machine studies, 27(3): 221-234.

[8] Ke, G., et al. (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 3149–3157.

[9] Hancock, J.T. and T.M. Khoshgoftaar (2020) CatBoost for big data: an interdisciplinary review. Journal of big data, 7(1): 1-45.

[10] Pal, M. (2005) Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1): 217-222.

[11] Natekin, A. and A. Knoll (2013) Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7: 21.

[12] Keller, J.M., M.R. Gray, and J.A. Givens (1985) A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics, SMC-15(4): 580-585.

[13] https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data?resource=download.

[14] Yildirim, M., et al. (2023) Automatic Classification of Particles in the Urine Sediment Test with the Developed Artificial Intelligence-Based Hybrid Model. Diagnostics, 13(7): 1299.

[15] Özbay, F. A., & Özbay, E. (2023). An NCA-based Hybrid CNN Model for Classification of Alzheimer's Disease on Grad-CAM-enhanced Brain MRI Images. Turkish Journal of Science and Technology, 18(1), 139-155.