International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Cancer Diseases

Vibha Sahu [1], Prof. Ritu Chaturvedi [2], Dr. Kaptan Singh [3]

PG Scholar[1], Professor[2,3]

Department of Computer Science and Engineering, TRUBA College, Bhopal, Madhya Pradesh, India

Article Info

Publication Issue :
Volume 9, Issue 1
January-February-2023

Page Number : 32-37

Article History
Accepted: 03 Jan 2023
Published: 18 Jan 2023

ABSTRACT

This paper talks about a healthcare operational decision-making system that uses machine learning classifiers to predict decisions based on the actual decisions made by the doctor during healthcare operations. In this type of system for making decisions, most of the supervised machine learning classification and optimization techniques are used. This system can help the doctor decide what to do in the best way. We testify to this system on the caesarian section, which is the most common obstetric operation in the world to help save both mother and baby. This system helps us figure out when surgery is a good idea. This study shows how machine learning algorithms can be used to figure out how to do medical procedures. For this case study, the results show that both k nearest neighbours and Random Forest have an accuracy of 95.00%.

Keywords : Machine Learning Classifiers, Healthcare Decision Making, K Nearest Neighbor, Random Forest

I. INTRODUCTION

Today's healthcare needs effective methods and research methods to save lives, lower the cost of healthcare, and find contagious diseases early on. Using machine learning, healthcare organisations can predict trends in how patients are feeling and what they do. Recent discoveries in the health care field have led to a large amount of rich data being collected. McKinsey thinks that big data and machine learning could be worth $100 billion a year because they could help people make better decisions, improve innovation, and make clinical trials more efficient. Getting useful information and patterns from datasets can be a big chance to improve healthcare in the real world. This kind of information can be used to predict changes in a patient's condition as quickly as possible and lower the cost of health care. Information technologies are being used more and more in healthcare organisations to meet the needs of doctors as they make decisions about how to run their businesses. Machine Learning can not only help doctors decide what to do in an emergency, but it can also help with primary care in general. Also, machine learning techniques can be used to help doctors diagnose patients, especially when it's hard to predict what will happen, and choose the best way to operate [1]. The term machine learning was introduced in 1959 by Arthur Samuel. The above points indicate that there is a great need of new computational theories and tools to extract information from large volume of datasets [2]. The role of machine learning within the field of data mining and processing of large datasets increased with the discoveries of several algorithms such as support vector machine methods in 1990s [3]. Since the 1970s, machine learning and data analytics have been used in health care information systems [4]. Machine learning can handle things that people can't do, like being subjective because they're tired, and it can give people clues to help them make decisions [5].Here's how this paper is put together. In Section 2, machine learning in healthcare is explained. In Section 3, you can read about supervised machine learning classifiers like k-nearest neighbours, random forest, logistic regression, naive bayes, and support vector machine. Section 4 talks about the caesarean section, and then it talks about how it was done. Section 5 talks about the section on the results and what they mean.

MACHINE LEARNING IN HEALTHCARE

Machine learning is a necessary part of healthcare today. Optimists think that machine learning and artificial intelligence will help doctors find diseases earlier and better, treat them more accurately, and work better with patients in the future. Recent improvements in machine learning have shown that it can be used to make algorithms that are just as good as doctors.In the past few years, medical image processing and analysis, predicting healthcare operational decisions, dosage trials for intravenous tumour treatment, and detecting and treating prostate cancer are just some of the things that have been done in healthcare.

1. SUPERVISED MACHINE LEARNING CLASSIFIERS

Machine learning is all about making computer programmes that can look at data to find patterns and make better predictions about future cases based on the cases we give them without being explicitly programmed. In the training phase, the machine learning is called "supervised" if the examples have known labels. In the training phase of "unsupervised" machine learning, the examples do not have labels. This section is all about machine learning techniques like support vector machine (SVM), naive Bayes classifier (NB), random forest (RF), k-nearest neighbour (KNN), and logistic regression.

1.1RANDOM FOREST (RF)

Random forests, also called random decision forests, are a type of ensemble learning method that can be used for classification, regression, and other tasks. They work by building a lot of decision trees at training time and outputting the class that is the average of the classes (classification) or the mean prediction (regression) of the individual trees. Random decision forests fix the fact that decision trees tend to fit their training set too well.

1.2 LOGISTIC REGRESSION (LR)

It's a way of applying statistics to datasets in which the results are determined by a number of factors that may be controlled independently of one another. A binary measure of success is used to assess the final result (in which there are only two possible outcomes). Logistic regression is used to determine which model best describes the association between a collection of independent (predictor or explanatory) factors and the dichotomous feature of interest (dependent variable = response or outcome variable).

1.3 NAIVE BAYES (NB)

It is a Bayesian network-based classification method that assumes predictors are unrelated to one another. Naive Bayes classifiers, in their simplest form, work on the assumption that the existence of one feature in a class has no bearing on the presence of any other feature. All of these characteristics contribute to the likelihood, regardless of whether they are reliant on one another or on the existence of the other characteristics. Probability calculations, or what statisticians term a "posterior probability," are the foundation of NB.

1.4 K-NEAREST NEIGHBOURS (kNN)

As a supervised classification technique, the k-nearest-neighbors method learns to classify new data by examining an existing labelled dataset. To assign a label to a new point, it polls its nearest neighbours, which are also given labels, to see which label is preferred by the majority of its neighbours (k is the number of neighbours it checks).

1.5 SUPPORT VECTOR MACHINE (SVM)

You may utilise a Support Vector Machine (SVM) to solve both classification and regression issues, as it is a discriminative classifier. The purpose of support vector machines (SVMs) is to locate a separating hyperplane that maximises the gap between the various classes in the training data. In other words, the method creates an ideal hyperplane that classes fresh samples in order to generate the greatest distance feasible, hence lowering an upper bound, using a set of labelled training data (supervised learning). To help SVM classify data, we employ what are called "Supports Vectors," which are just the coordinates of data points that are closest to the ideal separation hyperplane. In addition, the data is transformed into a high-dimensional space suitable for use with linear discriminate functions using an appropriate kernel function.

II. CAESARIAN SECTION

It is common practice in the field of obstetrics to deliver babies through caesarean section. Since 1996, caesarean section rates have increased dramatically in the United States [6]. From 1996 to 2009, the rate of this surgery rose by 60%. In 2017, 32% of births were C-sections [7]. According to the most recent birth statistics from the Centers for Disease Control and Prevention, about one-third of pregnant women in the United States gave birth by Caesarean section in 2015. (CDC). C-sections are lifesaving in an emergency and can help high-risk mothers avoid harm during childbirth. In June of 2010, the World Health Organization formally reversed its earlier guideline of a C-section rate of 15%. There is no empirical evidence for an optimal proportion," their official statement said. The most important thing is to ensure that all women who require caesareans actually get them" [8].The UK Royal College of Obstetricians and Gynecologists (RCOG) and the UK Royal College of Anesthetists (RCA) have both pledged support for the UK National Confidential Enquiry into Patient Outcome and Death (NCEPOD) [9].

Category	*Criterion*
1	Immediate threat to the life of the woman or fetus
2	Maternal or fetal compromise that is not immediately life threatening
3	No maternal or fetal compromise but early delivery required
4	Delivery timed to suit woman and staff (elective)

Table 1. Categories of emergency for caesarian section

III. METHODOLOGY

Fig. illustrates the methodology of decision-making system based on supervised machine learning classifiers.

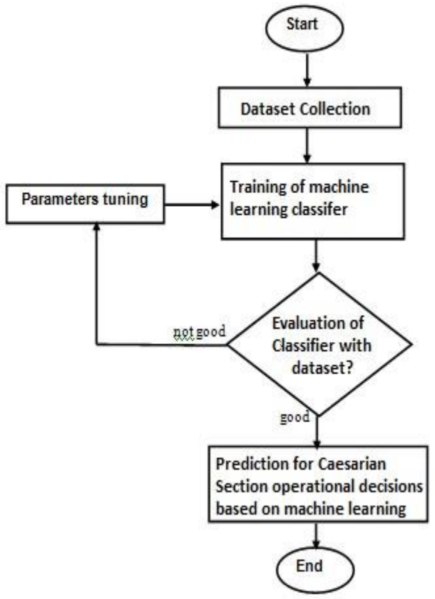
Start
Dataset Collection
Parameters tuning
Training of machine learning classifer
Evaluation of Classifier with dataset?
not good
good
Prediction for Caesarian Section operational decisions based on machine learning
End

Fig.1. A methodology of decision making system based on machine learning classifiers

As shown in Fig 1, three steps are required to be accomplished in order to predict decisions: dataset collection, training of machine learning classifiers and evaluation of machine learning classifiers.

DATASET COLLECTION

"Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study" [10] provided the caesarian section dataset used in this analysis. The 80 cases in the dataset are comprised of the five attributes—age, number of pregnancies, delivery duration, blood pressure, and heart status—that best capture the most salient features of common delivery complications. Table 2 categorises these features based on recommendations from the UK's Royal College of Anesthetists (RCA) [9].

Data set	No. of	No. of Positive	No. of Negative	Imbalance Ratio
Cancer-I	08	248	452	1.8
Cancer-II	12	249	452	1.8
Cancer-	22	272	508	1.6

TRAINING OF MACHINE LEARNING CLASSIFIERS

Classification refers to the process of teaching supervised machine learning classifiers. This caesarian section dataset is labelled, and that's what the training is based on. Caesarean section decision-making is successfully predicted by a classifier generated by a supervised machine learning algorithm that studies a training dataset of previous cases.

EVALUATION OF CLASSIFIERS:

The machine learning algorithm creates the categorization model during data training by analysing the data. The machine learning classifiers employed in this work to foretell the need for a C-section surgery were evaluated using a training dataset.

Table 3. CONFUSION MATRIX

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

The confusion matrix was mined for useful metrics that could be used to assess the performance of the machine learning classifiers. The True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1 score, and ROC area were used to evaluate the machine learning classifiers alongside the accurate classification rate or accuracy.

Precision and recall

Precision and recall, which are key components, provide insight into performance levels. If the precision number is 1.000, it represents 100% accuracy, according to the set of rules that classify the operation as a response.

$$Sensitivity\,(\%) = \frac{TP}{TP+FN} \times 100$$

$$Specificity\,(\%) = \frac{TN}{FP+TN} \times 100$$

The percentage of class-related information that is correctly classified, however, is what is referred to as recall value.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

Accuracy

The ratio of correctly categorized examples to all instances in the dataset serves as the definition of accuracy

$$Accuracy\,(\%) = [\frac{TP+TN}{TP+FP+TN+FN}] \times 100$$

Given that TP stands for True Positive, FP for False Positive, and TN for True Negative and FN for False Negative.

Truncated Positive Ratio

The ability is what is used to identify the high true positive rate. The term "sensitivity also refers to the true positive rate.

$$TPR = [\frac{TP}{TP+FN}]$$

Error Ratio

The error ratio is a measure of how many test cases were unsuccessful.

$$\mathbf{Re = 1- Acc\ (M)}$$

Rerepresents the error ratio, and Acc (M) represents the certainty of the chosen variables

IV. RESULTS AND DISCUSSION

The Caesarean Section Dataset used in this study was taken from "Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study" [10]. Eighty cases involving caesarean sections were collected for the purpose of testing and developing supervised machine learning classifiers used for predicting healthcare operational choices. In our training data, we labelled occurrences in which caesarean section surgery was decided upon as "yes" with a 1 and instances in which it was decided upon as "no" with a 0.The CANCER-I dataset's selection features were compared using a variety of classifiers and a combined classifier, and the findings were summarised in Table 5.2 in terms of accuracy, mean absolute error, mean relative error, and elapsed time. In this table showed or describe the outcomes of several classifier like decision tree, k-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest, and combination of classifier SVM-RBF and Random Forest. According to the table below, the Ensemble method of classification (SVM-RF) gave the greatest accuracy of any method or data mining strategy tested. In this table we explain the accuracy, mean absolute error, mean relative error and elapse time utilising different data mining algorithms.

Parameter Methods	ACC.	MAE	MRE	TIME
DT	**86.09**	**22.22**	**32.66**	**14.98**
KNN	**85.19**	**23.01**	**36.58**	**16.97**
SVM	**96.19**	**21.52**	**22.46**	**13.88**
RF	**97.10**	**21.01**	**29.04**	**13.88**
SVM-RF	**98.10**	**18.88**	**24.78**	**17.35**

Table 5.:Comparison of Different Classifier Results of the Selective Attributes (CANCER-

We also determined how well each classifier performed by computing its kappa static, mean absolute error, root mean squared error, relative absolute error, and root relative squared error.As can be seen, both k-nearest neighbours and Random Forest correctly predicted results in 95% of the tests. Consequently, the data gleaned through machine learning algorithms may be applied to improve.

V. CONCLUSION

In this research, we used a variety of data mining classification techniques, including decision tree, KNN, random forest, and support vector machine analysis. This dissertation combines the support vector machine (SVM) and random forest (RF) classifiers and proposes the majority vote ensemble classification technique for healthcare or cancer datasets, comparing their performance to that of other classification mechanisms such as the decision tree, the RF, the SVM, and the KNN, and obtaining results in terms of accuracy, mean absolute error, mean relative error, and elapsed time. The accuracy of our suggested (SVM-RF) system was 99.96% for cancer-I (Reduced data set), 96.35% for cancer-II (Original or big data set), and 96.26% for cancer-III (Reduced data set). Our suggested majority vote ensemble classifier outperforms the twin SVM method, with an improvement in accuracy of 1.29 percentage points for a smaller dataset and 0.60 percentage points for a larger one.

VI. REFERENCES

[1]. Boris Milovic, Milan Milovic. Prediction and Decision Making in Health Care using Data Mining,

[2]. nternational Journal of Public Health Science (IJPHS), Vol. 1, No. 2, December 2012, pp. 69~78[2] Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 37-54.

[3]. Kantardzic, Mehmed. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2003.

[4]. H. Jiawei and K. Micheline, Data Mining: Concepts and Techniques, vol. 2, Morgan Kaufmann Publishers, 2006.

[5]. Candelieri, A., Dolce, G., Riganello, F., &Sannita, W. G. (2011). Data Mining in Neurology. In KnowledgeOriented Applications in Data Mining (pp. 261-276). InTech.

[6]. Brady E. Hamilton, Ph.D.; Joyce A. Martin, M.P.H.; and Stephanie J. Ventura, M.A., Division of Vital Statistics,Births: Preliminary Data for 2007, National Vital Statistics Report.

[7]. Births: Provisional Data for 2017 USA. CDC. May 2018. Retrieved 18 May 2018.

[8]. World Health Organization (WHO) statement “Should there be a limit on Caesareans?". BBC News. 30 June 2010.

[9]. Andrew Simm, Darly Mathew, Caesarian section: techniques and complications, Obstetrics, Gynaecology & Reproductive Medicine, Volume 18, Issue 4, April 2008, Pages 93-98.

[10].FarhadSoleimanianGharehchopogh, Peyman Mohammadi, Parvin Hakimi, Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study, International Journal of Computer Applications (0975 – 8887) Volume 52 – No. 6, August 2012,Pages 21-26.

Cite this article as :

Vibha Sahu, Prof. Ritu Chaturvedi, Dr. Kaptan Singh, "Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Cancer Diseases ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9 Issue 1, pp. 32-37, January-February 2023.

Journal URL : https://ijsrcseit.com/CSEIT2390113