Journal of Advances in Medical and Biomedical Research | ISSN:2676-6264

Comparison of Four Data Mining Algorithms for Predicting Colorectal Cancer Risk

Mostafa Shanbehzadeh [1], Raoof Nopour [2], Hadi Kazemi-Arpanahi [3,4*]

1. *Dept. of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran*
2. *Dept. of Health Information Technology, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran*
3. *Dept. of Health Information Technology, Abadan Faculty of Medical Sciences, Abadan, Iran*
4. *Dept. of Student Research Committee, Abadan Faculty of Medical Sciences, Abadan, Iran*

Article Info

10.30699/jambs.29.133.100

Received: 2020/06/23;
Accepted: 2020/08/12;
Published Online: 04 Dec 2020

Use your device to scan and read the article online

Corresponding Information:
Hadi Kazemi-Arpanahi,
Dept. of Health Information Technology, Abadan Faculty of Medical Sciences, Abadan, Iran.
Dept. of Student Research Committee, Abadan Faculty of Medical Sciences, Abadan, Iran.
E-Mail:
H.kazemi@abadanums.ac.ir

ABSTRACT

Background & Objective: Colorectal cancer (CRC) is one of the most prevalent malignancies in the world. The early detection of CRC is not only a simple process but also is the key to treatment. Data mining algorithms could be potentially useful in cancer prognosis, diagnosis, and treatment. Therefore, the main focus of this study is to measure the performance of some data mining classifier algorithms in predicting CRC and providing an early warning to the high-risk groups.

Materials & Methods: This study was performed on 468 subjects, including 194 CRC patients and 274 non-CRC cases. We used the CRC dataset from Imam Hospital, Sari, Iran. The Chi-square feature selection method was utilized to analyze the risk factors. Next, four popular data mining algorithms were compared in terms of their performance in predicting CRC, and, finally, the best algorithm was identified.

Results: The best outcome was obtained by J-48 with F-measure=0.826, receiver operating characteristic (ROC)=0.881, precision=0.826, and sensitivity =0.827. Bayesian net was the second-best performer (F-Measure=0.718, ROC=0.784, precision=0.719, and sensitivity=0.722) followed by random forest (F-Measure=0.705, ROC=0.758, precision=0.719, and sensitivity=0.712). The multilayer perceptron technique had the worst performance (F-Measure=0.702, ROC=0.76, precision=0.701, and sensitivity=0.703).

Conclusion: According to the results of this study, J-48 could provide better insights than other proposed prediction models for clinical applications.

Keywords: Classification models, Colorectal cancer, Data mining, Prediction

Introduction

Colorectal cancer (CRC) is the most common gastrointestinal malignancy and the third leading cause of mortality in the world (1, 2). The CRC remains a critical challenge for communities' health with the estimated annual new case and mortality of one million and a half million, respectively (3, 4). The incidence of CRC has risen in low-income countries constantly over the past few decades (5, 6). This disease is becoming the first cause of cancer-related death in Asian developing countries (7). Iran has the third- and fourth-highest incidence rates among females and males, respectively (8). The CRC growth rate in Iran is expected to double over the next two decades and is considered a critical health challenge (9).

Despite advancements in diagnostic approaches, more than 90% of CRC cases had either progression or metastasis after diagnosis. Early detection can significantly improve the overall survival possibility in CRC patients (3, 10). Currently, colonoscopy and sigmoidoscopy are the most common CRC screening approaches. However, these techniques have some drawbacks, including being invasive, inconvenient, high costs, and suboptimal sensitivity and specificity. As a result, many people are reluctant to undergo these procedures.

Fecal occult blood test (FOBT) is another diagnostic method for detecting CRC. This technique has the benefits of being non-invasive and cost-effective. However, FOBT has not been broadly accepted due to its relatively low accuracy (3, 11). Therefore, establishing a CRC surveillance system for the regular screening of individuals using their risk factors is a priority for early and accurate CRC prediction. To meet this goal, accurate forecasting methods, such as data

mining with high-quality data and minimum error rate are needed (12).

Data mining is the process of selecting, discovering, and modeling huge volumes of data for extracting concealed patterns or potential relationships that provide valuable information. Mining healthcare data could improve medical evaluation screening, prognosis, diagnosis, treatment, and survival leading to enhanced clinical decision-making (13, 14).

According to the literature, little research has been conducted on using data mining to generate predictive models for CRC prognosis. Therefore, this study developed four widely-used data mining techniques, namely J-48, Bayesian net, random forest, and multilayer perceptron (MLP). Moreover, their CRC risk prediction performances were analyzed and compared.

Materials and Methods

This applied descriptive research was conducted in 2019. We applied data mining classification algorithms using a dataset of CRC risk factors. Our main goal was to compare the performance of different data mining methods for CRC risk prediction.

Dataset Description and Preprocessing

The dataset used in the present study was obtained from Imam Hospital, Sari, Iran. The medical records of the patients with CRC were reviewed by health information management experts. The inclusion criteria encompassed referring to the hospital for the screening, diagnosis, and treatment of CRC and signing informed consent.

The information content of 760 cases out of 800 records was complete. Forty incomplete case records were excluded due to the missing of more than 70% of the data. Moreover, to investigate on people without high-risk factors of CRC, the patient's records containing CRC high-risk factors according to the CRC screening guidelines (such as American Cancer Society and CRC Consortium) including the personal and family history of adenoma polyposis, Inflammatory Bowel Disease (IBD), CRC relative history, patients under 60 age years old with familial history of CRC and Hereditary cancer syndromes, such as Lynch syndrome, were excluded from the study (292 cases). This limitation was done to analyze the effects of other risk factors in developing CRC. Some dataset samples from the Imam Hospital data repository are depicted in Figure 1.

Age	Sex	High/ Fat	Red-meat	Fruit& Veg.	Exercise	Smoking-DAY	Smoking-Y	Fe(tablet)/ Y	Fe(tablet)/DAY2	Ca&D-DAY2	Ca&D-Y	Aspirin-Y	BMI	Contracep-DAY2	Contracep-Y	Alcohol- DAY2	Alcohol –Y	Endoscopy	Metabolic-syn	Fat.liv	Hormtherapy	Genetic	Class
25-44	Male	<1	<1	3-4	<1	<1	5-10	0	0	0	0	0	<=185	Non	Non	0	0	0	0	0	Non	0	0
25-44	Male	<1	2-3	3-4	<1	0	0	0	0	0	0	0	185-249	Non	Non	0	0	0	0	0	Non	0	0
45-64	Male	99	1-2	2-3	0	0	0	0	0	0	0	0	25-29.5	Non	Non	0	0	0	1	0	Non	0	1
45-64	Male	<1	1-2	2-3	0	1-2	>10	0	0	0	0	0	185-249	Non	Non	<50	>10	0	1	1	Non	99	1
45-64	Female	1-2	99	2-3	0	0	0	1-3	>5	99	99	>5	185-249	0	0	0	0	99	1	1	0	0	0
25-44	Female	2-3	2-3	2-3	<1	0	0	0	0	0	0	0	<=185	0	0	0	0	0	0	0	0	3	1
45-64	Male	2-3	2-3	<2	0	<1	1-5	1-3	0	0	0	0	<=185	Non	Non	0	0	1	0	0	Non	0	1
45-64	Female	<1	1-2	2-3	0	0	0	0	3-5	1-3	0	0	185-249	0	0	0	0	1	0	0	0	0	1
45-64	Male	<1	1-2	<2	0	1-2	>10	0	0	0	0	0	25-29.5	Non	Non	<50	5-10	1	1	0	Non	0	1
25-44	Female	<1	<1	3-4	0	0	0	0	0	0	3-5	>5	185-249	3-5	>5	0	0	99	1	99	0	0	0
45-64	Male	2-3	2-3	2-3	0	<1	>10	0	0	0	0	>5	30-34.9	Non	Non	0	0	1	1	0	Non	0	1
45-64	Male	99	2-3	<2	0	1-2	>10	0	0	0	0	0	25-29.9	Non	Non	0	0	1	1	0	Non	0	1
25-44	Male	2-3	2-3	2-3	<1	1-2	>10	0	0	0	0	0	30-34.9	Non	Non	0	0	0	1	0	Non	0	1
45-64	Female	1-2	1-2	99	<1	0	0	0	0	0	0	99	30-34.9	0	0	0	0	99	0	0	0	0	0
45-64	Female	1-2	<1	<2	0	<1	99	0	0	1-3	0	0	30-34.9	2-3	>5	0	0	1	1	0	0	0	1
45-64	Male	<1	<1	<2	0	<1	>10	0	0	0	>5	0	30-34.9	Non	Non	<50	5-10	1	1	0	Non	0	1
>65	Male	<1	99	<2	<1	1-2	>10	0	0	0	0	0	25-29.5	Non	Non	0	0	1	0	0	Non	99	1
>65	Male	2-3	2-3	<2	0	1-2	>10	0	0	0	0	0	25-29.5	Non	Non	0	0	0	0	0	Non	0	1
45-64	Male	2-3	2-3	<2	0	1-2	>10	0	0	0	0	>5	30-34.9	Non	Non	0	0	1	1	99	Non	0	1

Figure 1. Dataset sample associated with CRC risk factors

Feature Selection

To reduce the dimensions of the dataset and improve the efficiency of data mining algorithms, the possible factors for CRC were scored using the Chi-square correlation coefficient technique. Feature selection automatically selects the most important input features known as independent variables from the dataset contributing to the classification and assignment of cases based on the target output known as dependent variables.

Appropriate feature selection methods, including univariate selection, recursive feature elimination,

principal component analysis, and feature importance must be used to elevate the performance of data mining algorithms. In this research, we applied the weight statistical Chi-square test in Rapid miner software to identify the most important attributes in the CRC dataset. This test is based on the difference between the observed and expected values (Equation 1) and is used to determine the significance of the relationship between independent and dependent variables. The feature importance is calculated based on Equations 2 and 3.

$$Fe_i = \frac{(n_i * n_j)}{n} \qquad (1)$$

$$X^2 = \sum_{n=1}^{k} \frac{(\mathrm{Fo_i} - \mathrm{Fe_i})^2}{\mathrm{Fe_i}} \qquad (2)$$

$$\text{Number of samples} \geq (\text{Number of variables}) * 30 \qquad (3)$$

Predictive Models

Four selected data mining algorithms were utilized due to their common usage in recently published studies with the high performance of classification. We present a brief description of these four algorithms is here.

J-48

The J-48 is an important decision tree algorithm. Its capabilities include missing values accounting, decision tree pruning by determining confidence factors, extracting rules, and considering continuous attribute value ranges. These features make the J-48 algorithm a better choice than the other tree algorithms. This algorithm uses divide and conquers strategies for decision tree making based on independent and dependent variables.

In each node of the tree, the splitting function is completed by an attribute that can predict samples in each class more precisely. Initially, the J-48 rule sets are made by an unpruned tree, and each path from the root node to leaf is transformed into a prototype rule associated with the leaf node label. In the current study, the decision tree was made with the confidence of 0.2 to include all independent variables with maximum performance **(15).**

Random Forest

This algorithm is applied in datasets with a large dimension. It applied additional layers of randomness than other decision tree algorithms. Node splitting processes in the RF algorithm is performed by a random subset of predictors. The latter process in RF is different from other algorithms in which it is completed by the best all variable splitter. The diversity of trees is important in random forest performance **(16)**.

Bayesian Network

In this method, the degree of dependency between the independent variables and the output class can be shown by directed acyclic graph conditional probability methods. This graph shows the variables, each of which occurs independently **(17, 18)**. In the present study, the Bayesian network determined the probability of CRC occurrence based on the factors occurrence and frequency independently.

Multilayer Perceptron

An MLP is a feed-forward artificial neural network (ANN) model for predicting the class label of tuples. An MLP is composed of multiple layers of nodes in a directed graph, every layer of which is fully connected to the next one. Except for the input nodes, every node is a neuron (or processing element) with a nonlinear activation. An ANN consists of input, output, and processing (hidden) layers. Each layer contains a group of neurons that are generally associated with all the neurons of the other layers **(19)**. In this study, an MLP with 45 sigmoid nodes was used to develop a CRC risk prediction model.

Performance Evaluation Measures

In order to evaluate the predictive performance of models, we applied some evaluation measures, including precision, sensitivity, and F-measure. The first measure is precision as shown in Equation 4, which measures the probability of a positive prediction being correct. The second measure is sensitivity as shown in Equation 5 and is referred to as the proportion of positive cases that are classified as positive. Specificity refers to the proportion of negative cases classified as negative. The last measure is the F-measure as shown in Equation 6, which measures the probability of a positive prediction being correct.

The confusion matrix (Table 1) helps implement an evaluation step in classifying for prediction. For the prediction process, each sample can be classified into two classes of CRC and non-CRC. This matrix consists of four elements, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The TP indicates that the prediction result of the representation is CRC and is consistent with its real

class. The TN means that the prediction result of the sample is non-CRC and consistent with its actual class. The FP is the non-CRC sample predicted as CRC, and FN refers to a result expected as non-CRC with the actual result being CRC **(20, 21)**.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

$$\text{F – Measure} = 2.\frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (6)$$

Table 1. Confusion matrix

Real	Predicted	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Results

The present study was conducted on 468 participants, including 194 colorectal cancer patients (42.5%) and 274 (58.5%) control subjects. After feature selection, we obtained 15 clinical features as the most important risk factors of CRC prediction according to Equation 2 and 3. The results indicated that cigarette smoking (a packet per day) and a history of metabolic syndromes with the values of 63.046 and 0.01 were the most and least important risk factors of CRC, respectively (Table 2).

Table 2. Important variables selected for CRC risk prediction

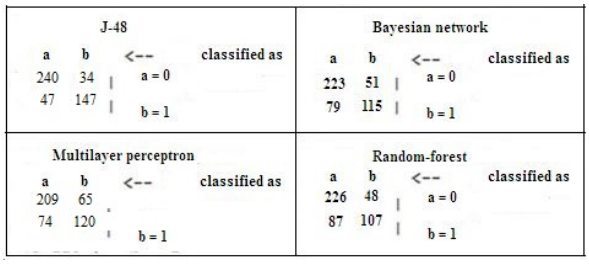
Variable	Variable values	Chi-square value
Smoking (In day)	No smoking, 1–4 cigarettes, 4-8 cigarettes, 8-12 cigarettes, 12 > cigarettes, unknown	63.046
Smoking (In years)	Day variable ranges for 1 year (365 day)	36.225
Exercise (In day)	No exercise, <15 minute, 15 - 30 minute, 30 - 45 minute, >30 minute. , unknown	43.328
Animal Fat (In day)	Not consumption, < 50 gram, 50-100 gram, 100-150 gram, 150-200 gram, 200> gram, Unknown	42.038
Red meat (In day)		41.727
Fruits and vegetables (In day)	Not consumption, < 100 gram., 100-200-gram, 200-300 gram, 300-400 gram, 400> gram, Unknown	27.080
Aspirin pill (In day/2)	Not taking, < 50 mili gram, 50-100 mili gram, 100-200 mili ram, 200-300 mili gram, 300> mili gram, Unknown	27.069
Contraceptive pill (In day)		27.080
Iron supplement (In day)		8.464
Contraceptive pill (In year)	Day variable ranges for 1 year (365 day)	19.416

Variable	Variable values	Chi-square value
Aspirin pill (In year)	Day variable ranges for 1 year (365 day)	11.263
Iron supplement (In year)	Day variable ranges for 1 year (365 day)	7.886
Body Mass Index (BMI)	<18.5 kg/m^2 , 18.5-24.9 kg/m^2, 25-29.9 kg/m^2, 30 > kg/m^2 Unknown	9.235
Alcohol (In day)	No alcohol drinking , <20 gram, 20–59 gram, 60–139 gram, 140–179 gram, ≥180 gram, unknown	18.172
Alcohol (In year)	Day variable ranges for 1 year (365 day)	12.389

A 10% cross-validation was considered for bias embedded in the performance of data mining algorithms. The result of comparing the four data mining algorithms based on the evaluation criteria showed the CRC prediction precision of 0.826, 0.709, 0.719, and 0.701 for J-48, random forest, Bayesian net, and MLP, respectively. The F-measure values for CRC prediction were found as 0.826 in the J-48 model, 0.705 in a random forest, 0.718 in the Bayesian net, and 0.702 in MLP.

Moreover, the sensitivity of CRC prediction was 0.827, 0.712, 0.722, and 0.703 for the J-48 model, random forest, Bayesian net, and MLP, respectively. The AUC values in CRC prediction were 0.881 in the J-48 model, 0.758 in a random forest, 0.784 in the Bayesian net, and 0.765 in MLP (Table 3). The findings of comparing the receiver operating characteristic curves for selected data mining algorithms are shown in Figure 2.

Table 3. DM algorithm confusion matrix



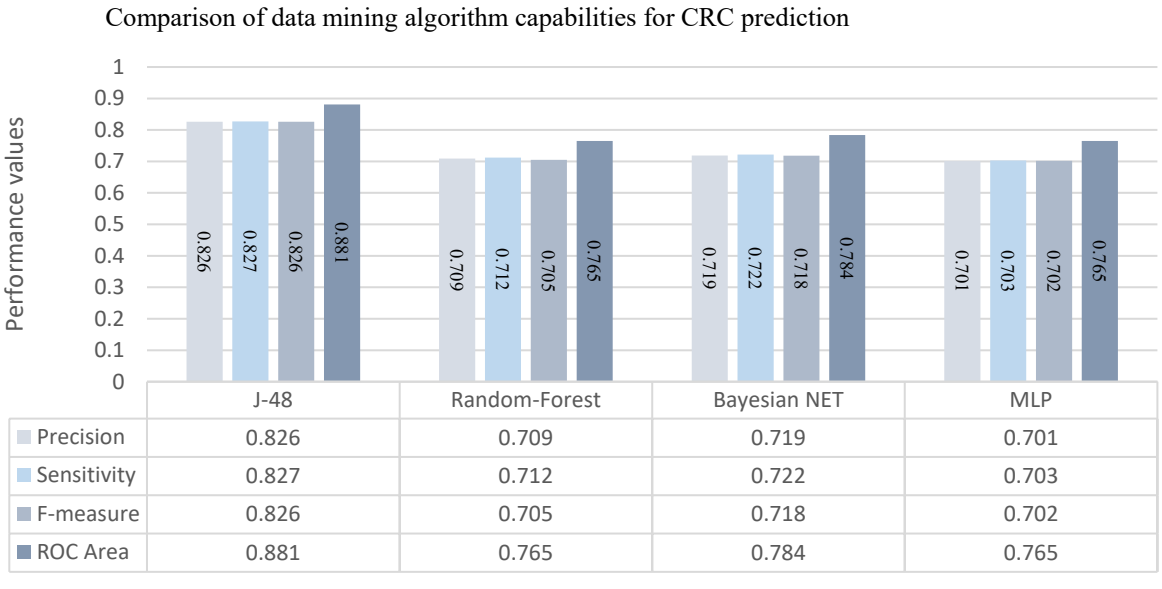


Figure 2. Comparison of data mining algorithms capabilities for CRC risk prediction

The comparison of these four data mining algorithms revealed that the J-48 decision tree algorithm (Figure 3) is better than other algorithms in all the investigated performance aspects. Therefore, we extracted some rules with the structure of IF-THEN from this algorithm for interpreting the clinical findings acquired in this algorithm.

1. IF (Smoking-day==0 && Aspirin-(Days/2) ==0 && Alcohol (In years) ==0 && Fruits and vegetables <=2 && Fat-meal <=1) THEN Class= 0 (Low risk).
2. IF (Smoking-day==0 && Aspirin-(Days/2) ==0 && Alcohol (In years) ==0 && Fruits and vegetables <=2 && Fat-meal == 1-2) THEN Class= 1 (High risk).

The interpretation of Rule 1 demonstrated that, for example, if a person did not consume any cigarettes, aspirin tablets, or alcohol in a day, and consumed < 200 g of fruits and vegetables and < 50 g of animal fats in a day, this algorithm classified him as low-risk similar to the pattern of non-CRC people investigated in this research. With the characteristics of Rule 1 (except animal fat consumption of 50-100 g in Rule 2), this algorithm classified this person into the high-risk group, which is similar to the pattern of CRC people.

```
J48 pruned tree
------------------

smoking-(day) = 0
|   Aspirin-(P.days/2) = 0
|   |   Alcohol(years) = 0
|   |   |   fruits-and-vegetables = <2
|   |   |   |   fat-meal = <1: 0 (14.0/4.0)
|   |   |   |   fat-meal = 1-2: 1 (5.0/2.0)
|   |   |   |   fat-meal = 2-3: 1 (5.0)
|   |   |   |   fat-meal = >3: 1 (0.0)
|   |   |   fruits-and-vegetables = 2-3
|   |   |   |   exercise = 0
|   |   |   |   |   red-meat = <1
|   |   |   |   |   |   contraceptive-tablets-(P.days/2) = 0: 1 (4.0/1.0)
|   |   |   |   |   |   contraceptive-tablets-(P.days/2) = <1: 0 (0.0)
|   |   |   |   |   |   contraceptive-tablets-(P.days/2) = 1-3: 0 (2.0)
|   |   |   |   |   |   contraceptive-tablets-(P.days/2) = 3-5: 1 (1.0)
|   |   |   |   |   |   contraceptive-tablets-(P.days/2) = >5: 0 (0.0)
|   |   |   |   |   red-meat = 1-2: 1 (19.0/7.0)
```

Figure 3. A part of the J-48 decision tree algorithm

Discussion

The main goal of this study was to compare the performance of four classification algorithms, namely J-48, Bayesian net, random forest, and MLP for predicting CRC. Results showed that all classification algorithms were acceptable and could give reasonable responses. However, J-48 had the best performance for all evaluation measures.

Much research has compared the performance of different data mining techniques in medicine (22-24). Some have focused on early detection, risk assessment, diagnosis, treatment, and survivability estimation of CRC (25). Studies conducted by Nartowt *et al.* (26), Sha *et al.* (27), Chau *et al.* (28), and Wang and Lualdi (29, 30) showed that using ANN for CRC prediction, early diagnosis, and screening had high classification performance.

Pourhoseingholi *et al.* (2017) demonstrated that among the multiple data mining models, the random forest had the best capability for estimating CRC five-year survival (31). Zhang *et al.* (2016) evaluated the application of three machine learning algorithms, including logistic regression, SVM, and ANN for CRC diagnosis based on a serum tumor marker. Finally, the results indicated better performance for logistic regression in terms of early CRC diagnosis (32).

Pourahmad *et al.* (2016) presented that the hierarchal clustering method had higher sensitivity and fuzzy c-means with maximum specificity. As a result, these authors introduced it as a non-invasive, efficient, and effective model for CRC staging (33). We reviewed four data mining techniques demonstrating that the J-48 algorithm had the best performance. In conclusion, the J-48 algorithm is recommended for predicting CRC

cases as a common model. Furthermore, it might be applied clinically in the future.

The CRC can be caused by numerous clinical and non-clinical factors (34). Given the multi-causal nature of CRC, predictive models can be useful for recognizing high-risk groups leading to early detection and the adoption of effective treatment plans (35). The CRC early diagnoses through scientific screening methods have been shown to increase survival chances (36). Regarding the timely and accurate prediction of CRC, a neoplasm with a high incidence and mortality rate provides a better plan for health policy to decrease complications and improve the patient's survival probability (37).

The true prediction may enhance CRC treatment and elevate the survival rate of patients. The predictive models in our study can discriminate the high- and low-risk individuals for CRC. Individuals with the prediction values of 1 or 0 were judged as high- and low-risk for CRC, respectively.

The current study had some limitations. Firstly, the research database lacked enough quantitative data, which may diminish the precision in data mining. Secondly, this investigation was a retrospective single-center experience. Thirdly, the absence of integrated EMR with machine learning tools and manual data entry had a negative effect on data mining quality. Finally, the research dataset lacked some prognostic factors, such as the history of patients and their families, which might have negatively impacted CRC predictions. Evaluation of more data mining techniques, larger databases in different organizations, and prospective data collection approaches are recommended for improving the data quality criteria.

Conclusion

The present study evaluated and compared the efficiency of some data mining classifier algorithms in terms of the early prediction of CRC risk. We compared four prediction models for CRC incidence considering the most important factors. The results indicated that the J-48 algorithm had the best classification performance. This study may assist future researchers in choosing the optimal predictive models for implementing community lifestyle interventions to reduce the incidence of CRC. The results of comparing the performance of prediction models in this study were satisfactory and we believe that this process would be improved in case we could use more data samples in the study dataset.

Acknowledgements

This article is extracted from a research project which supported by Tehran University of Medical Sciences with the contract no IR.TUMS.SPH.REC.1398.191. We appreciate Research Deputy of Tehran University of Medical Sciences who sponsored this project financially.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Funding and support

This study was part of an approved project funded by the Tehran Health Services Management Research Centre affiliated with the Tehran University of Medical Sciences particularly focusing at patient's rights.

Conflict of Interest

Authors declared no conflict of interest.

References

1. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. CA: CancerJ Clin. 2020;70(3):145-64. [DOI:10.3322/caac.21601]
2. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nature Rev Gastroenterol Hepatol. 2019;16(12):713-32. [DOI:10.1038/s41575-019-0189-8]
3. Kinar Y, Akiva P, Choman E, et al. Performance analysis of a machine learning flagging system used to identify a group of individuals at a high risk for colorectal cancer. PloS one. 2017;12(2): e0171759. [DOI:10.1371/journal.pone.0171759]
4. Ge H, Yan Y, Di Wu YH, Tian F. Potential role of LINC00996 in colorectal cancer: a study based on data mining and bioinformatics. OncoTarget Ther. 2018;11:4845. [DOI:10.2147/OTT.S173225]
5. Roberts PO, de Souza TG, Grant BM, et al. Pathologic factors affecting colorectal cancer survival in a Jamaican population-the UHWI experience. J Racial Ethnic Health Disparities. 2020;7(3): 413-20. [DOI:10.1007/s40615-019-00669-7]
6. Tsoi KK, Hirai HW, Chan FC, Griffiths S, Sung JJ. Predicted increases in incidence of colorectal cancer in developed and developing regions, in association with ageing populations. Clin Gastroenterol Hepatol. 2017;15(6):892-900. e4. [DOI:10.1016/j.cgh.2016.09.155]
7. Rieger AK, Mansmann UR. A Bayesian scoring rule on clustered event data for familial risk assessment-

An example from colorectal cancer screening. Biometric J. 2018;60(1):115-27. [DOI:10.1002/bimj.201600264]

8. Goshayeshi L, Pourahmadi A, Ghayour-Mobarhan M, et al. Colorectal cancer risk factors in north-eastern Iran: A retrospective cross-sectional study based on geographical information systems, spatial autocorrelation and regression analysis. Geospatial Health. 2019;14(2). [DOI:10.4081/gh.2019.793]
9. Taheri M, Tavakol M, Akbari ME, Almasi-Hashiani A, Abbasi M. Associations of demographic, socioeconomic, self-rated health, and metastasis in colorectal cancer in Iran. Med J Iran. 2019;33:17.
10. Chen H, Lin Z, Wu H, Wang L, Wu T, Tan C. Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest. Spectrochimica Acta Part A: Molec Biomolec Spectrosc. 2015;135:185-91. [DOI:10.1016/j.saa.2014.07.005]
11. Kop R, Hoogendoorn M, Ten Teije A, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med. 2016;76:30-8. [DOI:10.1016/j.compbiomed.2016.06.019]
12. Gage MM, Hueman MT. Colorectal cancer surveillance: What is the optimal frequency of follow-up and which tools best predict recurrence? Curr Colorect Cancer Reports. 2017;13(4):316-24. [DOI:10.1007/s11888-017-0382-5]
13. Nartowt B, Hart G, Muhammad W, Liang Y, Deng J. A model of risk of colorectal cancer tested between studies: building robust machine learning models for colorectal cancer risk prediction. Int J Radiation Oncol. 2019;105(1):E132. [DOI:10.1016/j.ijrobp.2019.06.2265]
14. Safdari R, Arpanahi HK, Langarizadeh M, Ghazisaiedi M, Dargahi H, Zendehdel K. Design a fuzzy rule-based expert system to aid earlier diagnosis of gastric cancer. Acta Informatica Medica. 2018;26(1):19. [DOI:10.5455/aim.2018.26.19-23]
15. Wu X, Kumar V, Quinlan JR, et al. Top 10 algorithms in data mining. Knowledge Inform Sys. 2008;14(1):1-37. [DOI:10.1007/s10115-007-0114-2]
16. Liaw A, Wiener M. Classification and regression by random forest. R news. 2002;2(3):18-22.
17. Amirkhani H, Rahmati M, Lucas PJ, Hommersom A. Exploiting experts' knowledge for structure learning of bayesian networks. IEEE transactions on pattern analysis and machine intelligence. 2016;39(11):2154-70. [DOI:10.1109/TPAMI.2016.2636828]
18. Zhang S, Tjortjis C, Zeng X, Qiao H, Buchan I, Keane J. Comparing data mining methods with logistic regression in childhood obesity prediction. Inform Sys Front. 2009;11(4):449-60. [DOI:10.1007/s10796-009-9157-0]
19. Baitharu TR, Pani SK. Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. Proc Comput Sci. 2016;85:862-70. [DOI:10.1016/j.procs.2016.05.276]
20. Liu RS, Li HJ, Liang FX, et al. Diagnostic accuracy of different computer-aided diagnostic systems for malignant and benign thyroid nodules classification in ultrasound images A systematic review and meta-analysis protocol. Medicine. 2019;98(29):4. [DOI:10.1097/MD.0000000000016227]
21. Pillai L, Chouhan U. Comparative Analysis of machine learning algorithms for Mycobacterium Tuberculosis protein sequences on the basis of physicochemical parameters. J Medical Imag Health Inform. 2014;4(2):212-9. [DOI:10.1166/jmihi.2014.1241]
22. Vijayarani S, Dhayanand S. Data mining classification algorithms for kidney disease prediction. Int J Cybernetics Inform. 2015;4(4):13-25. [DOI:10.5121/ijci.2015.4402]
23. Shah C, Jivani AG. Comparison of data mining classification algorithms for breast cancer prediction. 2013 Fourth international conference on computing, communications and networking technologies (ICCCNT); 2013: IEEE. [DOI:10.1109/ICCCNT.2013.6726477]
24. Abdar M, Kalhori SRN, Sutikno T, Subroto IMI, Arji G. Comparing performance of data mining algorithms in prediction heart diseases. Int J Electric Computer Engin. 2015;5(6):1569-76. [DOI:10.11591/ijece.v5i6.pp1569-1576]
25. Sabouri S, Esmaily H, Shahidsales S, Emadi M. Survival prediction in patients with colorectal cancer using artificial neural network and cox regression. Int J Cancer Manag. 2020;13(1):6. [DOI:10.5812/ijcm.81161]
26. Nartowt BJ, Hart GR, Roffman DA, et al. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. PloS one. 2019;14(8). [DOI:10.1371/journal.pone.0221421]
27. Sha S, Du W, Parkinson A, Glasgow N. Relative importance of clinical and socio demographic factors in association with post-operative in-hospital deaths in colorectal cancer patients in New South Wales: An artificial neural network approach. J Eval Clin Pract. 2020; 26(5):1389-98. [DOI:10.1111/jep.13318]
28. Chau R, Jenkins MA, Buchanan DD, et al. Determining the familial risk distribution of colorectal cancer: a data mining approach. Familial Cancer. 2016;15(2):241-51. [DOI:10.1007/s10689-015-9860-6]

29. Wang Q, Wei J, Chen Z, et al. Establishment of multiple diagnosis models for colorectal cancer with artificial neural networks. Oncol Lett. 2019;17(3):3314-22. [DOI:10.3892/ol.2019.10010]

30. Lualdi M, Cavalleri A, Battaglia L, et al. Early detection of colorectal adenocarcinoma: a clinical decision support tool based on plasma porphyrin accumulation and risk factors. BMC Cancer. 2018;18(1):841. [DOI:10.1186/s12885-018-4754-2]

31. Pourhoseingholi MA, Kheirian S, Zali MR. Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. Acta Informatica Medica. 2017;25(4):254. [DOI:10.5455/aim.2017.25.254-258]

32. Zhang B, Liang X, Gao H, Ye L, Wang Y. Models of logistic regression analysis, support vector machine, and back-propagation neural network based on serum tumor markers in colorectal cancer diagnosis. Genet Mol Res. 2016;15(2):10.4238. [DOI:10.4238/gmr.15028643]

33. Pourahmad S, Pourhashemi S, Mohammadianpanah M. Colorectal cancer staging using three clustering methods based on preoperative clinical findings. Asian Pacific J Cancer Prevent. 2016;17(2):823-7. [DOI:10.7314/APJCP.2016.17.2.823]

34. Myte R, Gylling B, Häggström J, et al. One-carbon metabolism and colorectal cancer risk according to molecular subtypes: a Bayesian network learning approach. Cancer Res.2016. [DOI:10.1158/1538-7445.AM2016-4294]

35. Lu W, Fu DL, Kong XX, et al. FOLFOX treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms. Cancer Medicine.2020;9(4):1419-29 [DOI:10.1002/cam4.2786]

36. Ai D, Pan H, Han R, Li X, Liu G, Xia LC. Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. Genes. 2019;10(2):112. [DOI:10.3390/genes10020112]

37. Afshar S, Warden E, Manochehri H, Saidijam M. Application of artificial neural network in miRNA biomarker selection and precise diagnosis of colorectal cancer. Iran Biomed J. 2019;23(3):175-83. [DOI:10.29252/ibj.23.3.175]

How to Cite This Article:

Shanbehzadeh M, Nopour R, Kazemi-Arpanahi H. Comparison of Four Data Mining Algorithms for Predicting Colorectal Cancer Risk. J Adv Med Biomed Res. 2021; 29 (133) :100-108

Download citation:
BibTeX | RIS | EndNote | Medlars | ProCite | Reference Manager | RefWorks

Send citation to:
Mendeley Zotero RefWorks