DOI: 10.15514/ISPRAS-2024-36(1)-14

Determining Relevant Risk Factors for Breast Cancer

Z.J. Ibarra-Cuevas, ORCID 0000-0002-0084-2393 <*zaziil.97@gmail.com*>
J.I. Nunez-Varela, ORCID 0000-0002-9633-3453 <*jose.nunez@uaslp.mx*>
A. Nunez-Varela, ORCID 0000-0003-4813-8992 <*alberto_snv@hotmail.com*>
F.E. Martinez-Perez, ORCID 0000-0002-3133-9045 <*eduardo.perez@uaslp.mx*>
S.E. Nava-Muñoz, ORCID 0000-0001-9345-4391 <*senavam@uaslp.mx*>
C.A. Ramirez-Gamez, ORCID 0000-0002-1509-0980 <*crgamez@uaslp.mx*>
H.G. Perez-Gonzalez, ORCID 0000-0003-3331-2230 <*hectorgerardo@uaslp.mx*>
School of Engineering, Universidad Autónoma de San Luis Potosí, San Luis Potosí, México.

Abstract. Breast cancer is a serious threat to women's health worldwide. Although the exact causes of this disease are still unknown, it is known that the incidence of breast cancer is associated with risk factors. Risk factors in cancer are any genetic, reproductive, hormonal, physical, biological, or lifestyle-related conditions that increase the likelihood of developing breast cancer. This research aims to identify the most relevant risk factors in patients with breast cancer in a dataset by following the *Knowledge Discovery in Databases* process. To determine the relevance of risk factors, this research implements two feature selection methods: the *Chi-Squared test* and *Mutual Information*; and seven classifiers are used to validate the results obtained. Our results show that the risk factors identified as the most relevant are related to the age of the patient, her menopausal status, whether she had undergone hormonal therapy, and her type of menopause.

Keywords: data mining; breast cancer; risk factors.

For citation: Ibarra-Cuevas Z.J., Nunez-Varela J.I., Nunez-Varela A., Martinez-Perez F.E., Nava-Muñoz S.E., Ramirez-Gamez C.A., Perez-Gonzalez H.G. Determining Relevant Risk Factors for Breast Cancer. *Trudy ISP RAN/Proc. ISP RAS*, vol. 36, issue 1, 2024. pp. 225-238. DOI: 10.15514/ISPRAS-2024-36(1)-14.

Full text: Ibarra-Cuevas Z.J., Nunez-Varela J.I., Nunez-Varela A., Martinez-Perez F.E., Nava-Muñoz S.E., Ramirez-Gamez C.A., Perez-Gonzalez H.G. Determination of Relevant Risk Factors for Breast Cancer Using Feature Selection. *Programming and Computer Software*, 2023, Vol. 49, No. 8, pp. 671–681. DOI: 10.1134/S0361768823080091.

Определение релевантных факторов риска для рака молочной железы

С.Х. Ибарра-Куэвас, ORCID 0000-0002-0084-2393 <*zaziil.97@gmail.com*>
Х.И. Нунес-Варела, ORCID 0000-0002-9633-3453 <*jose.nunez@uaslp.mx*>
А. Нунес-Варела, ORCID 0000-0003-4813-8992 <*alberto_snv@hotmail.com*>
Ф.Э. Мартинес-Перес, ORCID 0000-0002-3133-9045 <*eduardo.perez@uaslp.mx*>
С.Э. Нава-Муньос, ORCID 0000-0001-9345-4391 <*senavam@uaslp.mx*>
С.А. Рамирес-Гамес, ORCID 0000-0002-1509-0980 <*crgamez@uaslp.mx*>
Э.Х. Перес-Гонсалес, ORCID 0000-0003-3331-2230 <*hectorgerardo@uaslp.mx*>

Инженерная школа Автономного университета Сан-Луис-Потоси, Сан-Луис-Потоси, Мексика.

Аннотация. Рак молочной железы представляет собой серьезную угрозу для здоровья женщин во всем мире. Хотя точные причины этого заболевания до сих пор неизвестны, известно, что заболеваемость раком молочной железы связана с некоторыми факторами. Факторы риска при раке – это любые генетические, репродуктивные, гормональные, физические, биологические или связанные с образом жизни состояния, которые увеличивают вероятность развития рака молочной железы. Настоящее исследование направлено на выявление наиболее значимых факторов риска у пациентов с раком молочной железы по набору данных, следуя процессу «Обнаружение знаний в базах данных». Чтобы определить актуальность факторов риска, реализованы два метода отбора признаков: критерий Хи-квадрат и взаимная информация; для проверки полученных результатов используются семь классификаторов. Результаты показывают, что наиболее важные факторы риска связаны с возрастом пациентки, ее менопаузальным статусом, прохождением гормональной терапии и типом менопаузы.

Ключевые слова: добыча данных; рак молочной железы; факторы риска.

Для цитирования: Ибарра-Куэвас С.Х., Нунес-Варела Х.И, Нунес-Варела А., Мартинес-Перес Ф. Э., Нава-Муньос С.Э., Рамирес-Гамес С.А., Перес-Гонсалес Э.Х. Определение релевантных факторов риска для рака молочной железы. Труды ИСП РАН, том. 36, вып. 1, 2024. стр. 225-238 (на английском языке). DOI: 10.15514/ISPRAS-2024-36(1)-14.

Полный текст: Ибарра-Куэвас С.Х., Нунес-Варела Х.И, Нунес-Варела А., Мартине-Перес Ф.Э., Нава-Муньос С.Э., Рамирес-Гамес С.А., Перес-Гонсалес Э.Х. Определение релевантных факторов риска для рака молочной железы на основе отбора признаков. Programming and Computer Software, 2023, т. 49, № 8, стр. 671–681 (на английском языке). DOI: 10.1134/S0361768823080091.

1. Introduction

Globally, breast cancer is the most common and widespread type of cancer among women with more than 2.2 million new cases and about 680,000 deaths in 2020, according to the Global Cancer Observatory [1]. The early detection of breast cancer is key to increase the chance of treatment and recovery; this is normally done by screening tests, such as a mammography. Studies have also identified what are known as risk factors, that are associated with the likelihood of developing breast cancer. There are a wide variety of risk factors that include genetic, reproductive, hormonal, physical, biological, lifestyle-related, among others [2]. It is important to analyze and understand the possible impact each factor could have in the development of breast cancer so that physicians could suggest preventive strategies to women who are known to have some of these risk factors.

A common trend in recent years is the analysis of data obtained from clinical records [3, 13, 14]. This has been achieved by using methodologies that extract potentially valuable information. Knowledge Discovery in Databases (KDD) [4] is a process that follows different phases or stages (Figure 1), such as selection, preprocessing and transformation of data, so that machine learning

methods could be applied with the aim of classifying information (prediction) or identifying new knowledge (discovery).

In this research we follow the KDD process, and our main contribution is the integration of feature selection methods and ensemble learning algorithms to determine and validate relevant risk factors from a breast cancer dataset. The most relevant factors identified are related to the patient's age, whether she had undergone hormone therapy, her type of menopause, and her menopausal status.

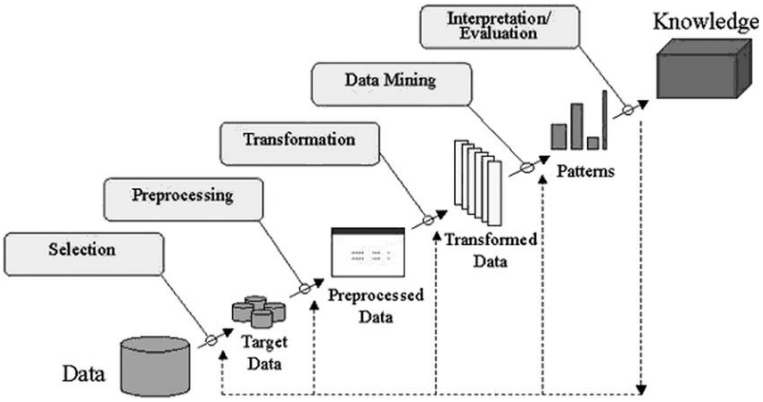

Fig. 1. Knowledge Discovery in Databases Process (taken from [4])

Being able to determine whether there is a risk of breast cancer or not solely from information readily known to most people is an important tool that would be widely available without the need to have specialized equipment. Of course, this is not meant to substitute screening tests and the knowledge of medical personnel. However, these tools could provide useful information and be part of the strategies for breast cancer risk control.

The rest of this paper is organized as follows. Section 2 reviews the related works for determining breast cancer risk factors. Section 3 explains the dataset used in this research. Section 4 describes the data pre-processing stage. Section 5 explains how the relevant risk factors are selected. Section 6 shows the results of classification methods on the dataset. Section 7 presents the validation of those selected risk factors. Section 8 provides our final conclusions.

2. Related work

Li et al. [5] present a prevention and control system for breast cancer by means of item rule association algorithms applied on a private dataset with 2,966 records and 83 attributes. An important characteristic of their work is the creation of their own dataset by interviewing patients from 22 hospitals over a one-year period and storing clinical, personal, and socio-economical information. Three types of rules defining the more relevant risk factors were identified; 35 rules were obtained using a single factor, 19 rules were obtained combining two factors, and 9 rules were obtained combining three factors. The main difference with our work is the creation of their own dataset, that provides more information and control. Kabir et al. [6] also generated risk factor rules by means of association rule mining, using the Breast Cancer Surveillance Consortium's (BCSC)[1] Risk Factors dataset. This public dataset contains 6,318,638 cases and 13 attributes, although all records containing at least one missing value were discarded. The logit model was used to select those factors that may affect the likelihood of breast cancer. A set of 5 rules was obtained for breast cancer cases and 4 rules for non-cancer cases. However, because of the class imbalance problem, they had to adjust the algorithm for the breast cancer cases.

[1] Breast Cancer Surveillance Consortium page: https://www.bcsc-research.org

The class imbalance is a problem that is commonly found in cancer-related datasets, since there are fewer positive cases compared to the number of negative cases. Kabir and Ludwig [7] focused on this issue by implementing six data-level resampling approaches. These techniques were applied on the BCSC's Risk Factors dataset, after discarding all records containing at least one missing value. The authors used three different classification algorithms: Decision Tree, Random Forest, and XGBoost. Their results showed that performance improves when resampling techniques are used compared to when no techniques are applied. The difference with our work is that we use a resampling approach at the algorithm level.

In summary, the main distinction between the described works and ours is that we make use of feature selection methods, a resampling technique, and use classification to validate the relevance of the selected risk factors.

3. Data selection

The breast cancer dataset used in this research was obtained from the Breast Cancer Surveillance Consortium (BCSC)[2]. The BCSC provides four datasets related to risk factors of breast cancer. For our analysis, the Risk Estimation (v.2) dataset [8] was selected (with information ranging from 1996-2002) for three reasons: i) it provides an attribute indicating the presence of breast cancer, that is used to classify each case, ii) it contains information about 11 risk factors, and iii) patients had no previous diagnosis of breast cancer up until the screening test recorded in the dataset. This last point is important because we are interested in determining relevant risk factors when no cancer has been diagnosed before. For instance, the Risk Factors (v.2) dataset also includes information of patients that have had cancer at some point in their life. This dataset could be useful to analyze the relationship between risk factors in women that have had cancer and those that have not.

Table 1 contains the description of the 16 attributes within the Risk Estimation (v.2) dataset and the values that can be assigned to each attribute, as well as their meaning. Table 2 shows the number of breast cancer cases, and their corresponding percentage, within the Risk Estimate (v.2) dataset. In total the dataset contains 1,007,660 cases. However, notice the difference between positive cancer (0.73%) and non-cancer (99.27%) cases. This imbalance in the data is an issue commonly present in this type of problems and will be further discussed in Section 6.

4. Data preprocessing

The preprocessing phase for our research consisted in taking the original dataset and apply four different operations.

4.1 Simple conversion operations

First, we converted all data types from numerical to categorical, except the count attribute which remained as a numeric attribute. Second, we converted all 9 values to the categorical value of unknown in all attributes that contain this value (i.e., attributes 1 and 3 to 12).

4.2 Attribute transformation

After analyzing the values of three attributes, specifically, value 1 of the menopause attribute, value *9* of the *surgmeno* attribute, and value *9* of the *hrt* attribute (attributes 1, 11 and 12 in Table 1 respectively); we decided to transform these three attributes to clarify the information given by those values. For the *menopause* attribute, value *1* refers to postmenopausal women *or* women of more than 55 years old. It is possible to identify true postmenopausal cases by means of the *surgmeno*

[2] Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). http://www.bcsc-research.org/

attribute. If the *surgemeno* attribute contains a *0* or *1*, it means that the record refers to a postmenopausal woman, and these records are assigned a value of *1* in the *menopause* attribute. A new value *2* was created and assigned to those cases where it is not possible to define whether a woman is postmenopausal or is older than 55 years. The attribute was renamed as *menopause_new* to differentiate from the original (see Table 3). Originally, value *1* was assigned to 140,843 records; after the transformation 107,810 records were detected as true postmenopausal cases (that were left with a value of *1*), and the rest were assigned the new value of *2*.

Table 1. Description of attributes of the Risk Estimation (v.2) dataset

No.	Attribute	Description	Values
1	*menopause*	Menopausal status	0 = premenopausal 1 = postmenopausal or age>=55 9 = unknown
2	*agegrp*	Age (years) in 5-year groups	1 = 35-39 2 = 40-44 3 = 45-49 4 = 50-54 5 = 55-59 6 = 60-64 7 = 65-69 8 = 70-74 9 = 75-79 10 = 80-84
3	*density*	BI-RADS breast density codes	1 = Almost entirely fat 2 = Scattered fibro glandular densities 3 = Heterogeneously dense 4 = Extremely dense 9 = Unknown or different measurement system
4	*race*	Race	1 = white 2 = Asian/Pacific Islander 3 = black 4 = Native American 5 = other/mixed 9 = unknown
5	*hispanic*	Patient is Hispanic	0 = no 1 = yes 9 = unknown
6	*bmi*	Body mass index	1 = 10-24.99 2 = 25-29.99 3 = 30-34.99 4 = 35 or more 9 = unknown
7	*agefirst*	Age at first birth	0 = Age < 30 1 = Age 30 or greater 2 = Nulliparous 9 = unknown
8	*nrelbc*	Number of first-degree relatives with breast cancer	0 = zero 1= one 2 = 2 or more 9 = unknown
9	*brstproc*	Previous breast procedure	0 = no 1 = yes 9 = unknown
10	*lastmamm*	Result of last mammogram before the index mammogram	0 = negative 1 = false positive 9 = unknown
11	*surgmeno*	Type of menopause	0 = natural 1 = surgical 9 = unknown or not menopausal (menopause=0 or menopause=9)
12	*hrt*	Current hormone therapy	0 = no 1 = yes 9 = unknown or not menopausal (menopause=0 or menopause=9)
13	*invasive*	Diagnosis of invasive breast cancer within one year of the index screening mammogram	0 = no 1 = yes

14	*cancer*	Diagnosis of invasive or ductal carcinoma in situ breast cancer within one year of the index screening mammogram	0 = no 1 = yes
15	*training*	Training data	0 = no (validation) 1 = yes (training)
16	*count*	Frequency count of this combination of covariates and outcomes (all variables 1 to 15)	

Table 2. Distribution of positive and non-cancer cases

Breast Cancer Diagnosis	Cases	%
Yes	7,319	0.73
No	1,000,341	99.27
Total	1,007,660	100

For the *surgmeno* attribute, value *9* is given to women that have not undergone menopause yet *or* the status of menopause is unknown. A new value *2* was created to refer to cases that are still not menopausal by checking if the *menopause* attribute is *0*. The attribute was renamed as *surgmeno_new* to differentiate from the original (see Table 3). Originally, value *9* was assigned to 83,545 records; after this operation 29,542 records were given the value of *2*, and 54,003 remained as *unknown*.

Similarly, for the *hrt* attribute, the same value *9* is assigned to cases that have not presented menopause *or* to cases where the use of hormone restitution therapy is unknown. A new value *2* was created to refer to cases that are still not menopausal by checking if the *menopause* attribute is *0*. The attribute was renamed as *hrt_new* to differentiate from the original (see Table 3). Originally, value *9* was assigned to 64,489 records; after this operation 29,542 records were given the value of *2*, and 34,947 remained as *unknown*.

Table 3. New attributes after being transformed

Attribute	Values
menopause_new	0 = premenopausal 1 = postmenopausal 2 = postmenopausal or age>=55 9 = unknown
surgmeno_new	0 = natural 1 = surgical 2 = not menopausal 9 = unknown or unknown menopausal (menopause=9)
hrt_new	0 = no 1 = yes 2 = not menopausal 9 = unknown or unknown menopausal (menopause=9)

4.3 Attribute removal

Three attributes were removed from the dataset. The *invasive* attribute, that refers to the diagnosis of invasive or ductal carcinoma, was not considered due to the causality of correlation with the *cancer* attribute of interest. The *training* attribute suggests whether that record in the dataset is to be considered for training or validation. However, because of the next transformations to be described we cannot use this division of records, thus the attribute is removed. Finally, the *last_mammogram* attribute indicates the result of the last mammogram taken before the index mammogram that relates to the *cancer* attribute. Since it only contains information about negative and false positive results, then, it can be removed without affecting our analysis.

4.4 Elimination of records with unknown values

Most of the attributes, as shown in Table 1, contain the *unknown* value. After careful analysis we decided to remove all records containing one or more *unknown* values and work only with records containing true values. After this operation, out of the 1,007,660 cases in the dataset (see Table 2), we are left with 160,390 cases.

5. Risk factors selection

To determine the ranking of attributes, this research makes use of two feature selection methods: *Chi-squared test* and *Mutual Information*.

5.1 Chi-squared test

The *Chi-squared test* is a nonparametric statistical technique used to determine if a distribution of observed frequencies differs from the theoretical expected frequencies [9]. Table 4 presents the *Chi-squared* values obtained for each of the 11 risk factors within the dataset. The values are sorted in descending order. The higher the value of an attribute the more relevant it is considered. We also verified the resulting values with a confidence of 95% (p-value of 0.05). Attributes from 1 to 9 are statistically significant at the 0.05 level. Only attributes 10 and 11 are not statistically significant.

According to the obtained values the first four attributes could be considered as more relevant, i.e., the patient's age (*agegrp*), whether she had undergone hormone therapy (*hrt_new*), her type of menopause (*surgmeno_new*), and her menopausal status (*menopause_new*). The next two attributes are also interesting, whether the patient have had a breast procedure (*brstproc*) and the patient's breast density (*density*). The rest of the attributes could be considered less relevant for this specific dataset.

Table 4. Chi-squared results for all risk factors

No.	Attribute	Chi-squared
1	*agegrp*	170.285
2	*hrt_new*	84.667
3	*surgmeno_new*	82.352
4	*menopaus_new*	82.306
5	*brstproc*	49.163
6	*density*	40.555
7	*nrelbc*	21.018
8	*Hispanic*	16.404
9	*agefirst*	6.721
10	*race*	4.456
11	*bmi*	1.374

5.2 Mutual Information

Mutual Information [10] is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. Table 5 presents the values obtained from the *Mutual Information* with normalization. Again, the values are sorted in descending order. The higher the value of an attribute the more relevant it is considered. Here, a threshold (cutoff) value was calculated in order to determine which attributes should be selected. Our threshold value was calculated by means of the standard deviation (*S*). For an attribute to be selected, its *Mutual Information* value must be greater than the threshold value *S*. In this case, only the first four attributes are greater than our calculated $S = 0.00022$. Notice that these four selected attributes are the same most relevant calculated by the *Chi-squared test*. The rest of the attributes have a similar ranking as given by the *Chi-squared test*.

Table 5. Mutual Information results for all risk factors

No.	Attribute	Mutual Information
1	*agegrp*	0.000740
2	*hrt_new*	0.000398
3	*surgmeno_new*	0.000390
4	*menopaus_new*	0.000390
5	*brstproc*	0.000202
6	*density*	0.000196
7	*Hispanic*	0.000092
8	*nrelbc*	0.000085
9	*agefirst*	0.000032
10	*race*	0.000021
11	*bmi*	0.000006

5.3 Definition of subsets of relevant attributes

To synthesize and validate the results obtained by the *Chi-Squared test* and *Mutual Information*, three subsets are defined based on the values given in the rankings of both methods as seen in Table 6.

Table 6. Attributes of the defined subsets

Subset ID	Attributes
Subset(4)	*{agegrp, hrt_new, surgmeno_new, menopause_new}*
Subset(7)	*{Subset(4), brstproc, density, nrelbc}*
Subset(11)	*{Subset(7), Hispanic, agefirst, race, bmi}*

6. Imbalance classification problem

This type of problem occurs when the number of records of some class label is much larger than the other class (as shown in Table 2). This problem remains after the preprocessing phase described in Section 4, where all records with an unknown value were eliminated. The resulting dataset ended up with 95.83% of non-cancer records versus 4.17% of positive cancer records. The problem of class imbalance has been actively addressed and several techniques to deal with this problem have been proposed, both at the data-level and algorithm-level [11]. Because it is important to maintain the integrity of our dataset, we follow an algorithm-level approach by implementing an ensemble learning method, particularly the *Bagging* method [12].

The *Bagging* method creates independent and parallel sub-classifiers with a single machine learning algorithm. First, from the initial data, several subsets of the same size are generated, thus ensuring diversity and independence. Then, for each sample, a sub-classifier is constructed and, finally, using a majority vote the final classification is obtained (Fig. 2).

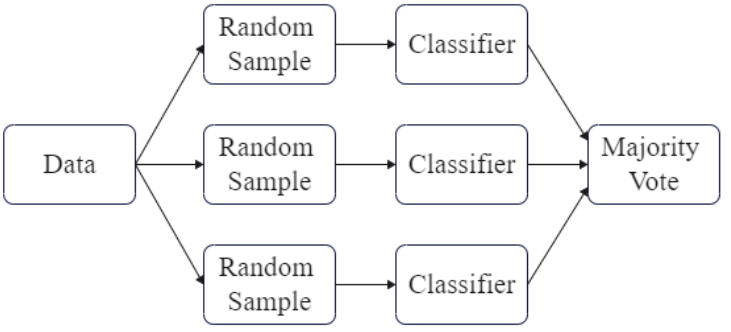

Fig. 2. Bagging diagram

Following this method, it was necessary to create a resampling of the data according to the *cancer* attribute. Fig. 3 shows the process used to perform such resampling. From the dataset, after being preprocessed, twenty-three sample groups were randomly generated, combining all the positive cancer records with a subset of the same number of randomly selected non-cancer records. Since the dataset ended up with 1,053 positive cancer records after the preprocessing phase, each sample group contains that number of records plus a random selection of 1,053 non-cancer records (2,106 records per sample group).

7. Risk factors validation

Section 5 defined two similar rankings for the risk factors within the dataset. The aim of identifying which risk factors are more relevant than others, is to use those relevant attributes to determine breast cancer cases, or at least, to pay more attention to those specific factors. In this section, experiments will be performed to determine the predictive performance of the attribute subsets as defined in Table 6, i.e., *Subset(4)*, *Subset(7)*, and *Subset(11)*, where the latter will be used as baseline for the previous two subsets. For our experiments, the RapidMiner software platform[3] was used, as it provides preprocessing procedures and the implementation of machine learning algorithms, among other features. Seven different algorithms were selected to cover multiple machine learning techniques: Decision Tree, Decision Stump, Random Tree, Deep Learning, Generalized Linear Model, Naïve Bayes, and k-NN (k-Nearest Neighbors). All algorithms were executed considering the default settings given by the software platform. To validate each subset of attributes, the seven classification algorithms were trained only with the attributes that belong to the subset being evaluated. Also, a 10-fold cross validation was used to obtain the performance metrics of accuracy, precision, and recall.

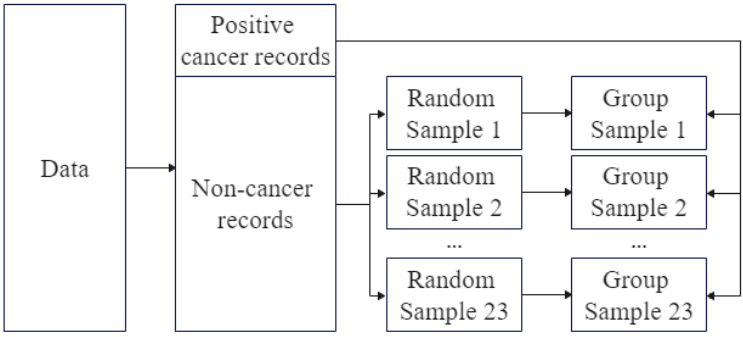

Fig. 3. Resampling process for the class imbalance problem

Table 7 presents the results for the three subsets of attributes as defined in Table 6. The first thing to note is the column that refers to *Subset(11)*; this is our baseline, as it considers all attributes. The classifiers with the highest accuracy (Acc.) are Decision Tree and Deep Learning with 97.45% and 97.21% respectively, while the least accurate is Random Tree with 78.38%.

Table 7. Performance metrics of the subsets of relevant attributes.

Algorithm	Metric	Subset(4)	Subset(7)	Subset(11)
Decision Stump	Acc.	86.32%	86.32%	86.32%
	Prec.	99.79%	99.79%	99.79%
	Rec.	72.83%	72.83%	72.83%
Decision Tree	Acc.	86.32%	96.18%	97.45%
	Prec.	99.79%	99.82%	99.77%

[3] Rapid Miner page: https://rapidminer.com

	Rec.	72.83%	92.53%	95.12%
Random Tree	Acc.	85.24%	80.67%	78.38%
	Prec.	98.29%	94.39%	87.79%
	Rec.	72.15%	67.16%	70.08%
Deep Learning	Acc.	93.32%	96.16%	97.21%
	Prec.	99.41%	99.65%	99.52%
	Rec.	87.19%	92.65%	94.88%
Generalized Linear Model	Acc.	92.87%	95.56%	96.62%
	Prec.	99.62%	99.68%	99.71%
	Rec.	86.09%	91.44%	93.51%
Naïve Bayes	Acc.	92.51%	93.87%	93.93%
	Prec.	98.77%	98.64%	98.70%
	Rec.	86.31%	89.11%	89.16%
k-NN	Acc.	93.10%	87.27%	81.30%
	Prec.	100.00%	99.94%	99.91%
	Rec.	86.20%	74.59%	62.67%

It is important to also consider the metrics of precision (Prec.) and recall (Rec.), that provide more information with regard of the classification of positive cancer cases. The higher the precision value the fewer false positives being classified. On the other hand, the higher the recall value the more positive records are classified correctly. In our experiments for *Subset(11)*, the precision values for all algorithms are high. However, the recall value for k-NN is low, which means that only 62.67% of the positive cancer cases were correctly classified. In terms of the three metrics, Decision Tree, Deep Learning, and Generalized Linear Model obtained the best results for all attributes.

In order to validate whether the selected attributes could be truly relevant in our study, we need to compare the results against those obtained by the baseline (*Subset(11)*). First, notice that Decision Stump reported the same results for the three subsets. This is because the algorithm generates a decision tree with only one division obtained from the evaluation of one of the most significant attributes. In our case, the algorithm chose the attributes of *agegrp* and *menopause_new* as a single node, and since both attributes are part of the three subsets then the results are the same. Although these results do not provide new information, as they are the same, the algorithm does support the relevance of these two attributes as stated in Section 5.

After analyzing these results, it is possible to conclude that the four selected risk factors: the patient's age (*agegrp*), whether she had undergone hormone therapy (*hrt_new*), her type of menopause (*surgmeno_new*), and her menopausal status (*menopause_new*); are relevant for the classification of positive cancer cases.

8. Conclusions

Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has a direct influence on their daily practice and clinical service. The study of risk factors for breast cancer is an option that has been investigated to create control and risk assessment strategies in women. The main objective of this research is to identify relevant risk factors that could accurately predict whether a woman can develop breast cancer or not. Our research explores two feature selection techniques, *Chi-squared test* and *Mutual Information*, combined with an ensemble method (*Bagging*) to detect breast cancer cases with information on risk factors. We found that the most relevant risk factors in breast cancer cases, according to the dataset analyzed, are the patient's age (*agegrp*), whether she had undergone hormone therapy (*hrt_new*), her type of menopause (*surgmeno_new*), and her menopausal status (*menopause_new*). These four risk factors were validated by means of seven classification algorithms. It is possible to obtain a predictive performance similar to that obtained using all 11 attributes of the dataset. These are significant results that should also be validated by physicians. It is difficult to directly compare our results with

other similar works because of the different datasets and methods being used. Datasets may contain clinical, personal, demographical, therapeutical, or pathological information, and the availability of this information and the number of attributes of each type will affect the results obtained. As future work, one of the most important issues is to have as much data as possible. We are looking at the possibility of creating our dataset in collaboration with local hospitals. Also, we are interested in exploring other feature selection methods and resampling techniques, along with other classification algorithms. We expect that this work could further advance our understanding in topics as relevant such as this.

References

[1]. Global Cancer Observatory, “Cancer Today”, https://gco.iarc.fr/today/online-analysis-pie (accessed Apr. 25, 2023).
[2]. Cancer.Net, “Breast Cancer: Risk Factors and Prevention”, https://www.cancer.net/cancer-types/breast-cancer/risk-factors-and-prevention (accessed Apr. 25, 2023).
[3]. P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, “Predicting Breast Cancer Recurrence Using Machine Learning Techniques”, ACM Comput. Surv., vol. 49, no. 3, pp. 1–40, Dec. 2016, doi: 10.1145/2988544.
[4]. H. Kawano, “Knowledge Discovery and Data Mining”, J. Japan Soc. Fuzzy Theory Syst., vol. 9, no. 6, pp. 851–860, 1997, doi: 10.3156/jfuzzy.9.6_851.
[5]. A. Li et al., “Association Rule-Based Breast Cancer Prevention and Control System”, IEEE Trans. Comput. Soc. Syst., vol. 6, no. 5, pp. 1106–1114, Oct. 2019, doi: 10.1109/TCSS.2019.2912629.
[6]. M. F. Kabir, S. A. Ludwig, and A. S. Abdullah, “Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining”, in 2018 IEEE International Conference on Big Data (Big Data), Dec. 2018, pp. 2433–2441, doi: 10.1109/BigData.2018.8622028.
[7]. M. F. Kabir and S. Ludwig, “Classification of Breast Cancer Risk Factors Using Several Resampling Approaches”, in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Dec. 2018, pp. 1243–1248, doi: 10.1109/ICMLA.2018.00202.
[8]. W. E. Barlow et al., “Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography”, JNCI J. Natl. Cancer Inst., vol. 98, no. 17, pp. 1204–1214, Sep. 2006, doi: 10.1093/jnci/djj331.
[9]. K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”, London, Edinburgh, Dublin Philos. Mag. J. Sci., vol. 50, no. 302, pp. 157–175, Jul. 1900, doi: 10.1080/14786440009463897.
[10]. D. J. C. MacKay, "Information Theory, Inference & Learning Algorithms". USA: Cambridge University Press, 2002.
[11]. H. Kaur, H. S. Pannu and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions", ACM Computing Surveys, vol. 52, no. 4, pp. 1-36, 2019, doi: 10.1145/3343440.
[12]. L. Breiman, “Bagging Predictors”, Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1023/A:1018054314350.
[13]. I. Volkov, G. Radchenko, and A. Tchernykh, "Digital Twins, Internet of Things and Mobile Medicine: A Review of Current Platforms to Support Smart Healthcare". Programming and Computer Software, vol. 47, pp. 578–590, 2021, doi: 10.1134/S0361768821080284.
[14]. I. Vasilev, M. Petrovskiy, I. Mashechkin, et al. "Predicting COVID-19-Induced Lung Damage Based on Machine Learning Methods". Programming and Computer Software, vol. 48, pp. 243–255, 2022, doi: 10.1134/S0361768822040065.

Информация об авторах / Information about authors

Сасиль Хосефина ИБАРРА-КУЭВАС – магистр компьютерных наук и разработчик программного обеспечения. С 2022 года работает в коммерческой компании, где ведет разработку программного обеспечения. Научные интересы: интеллектуальный анализ данных, базы данных и разработка программного обеспечения.

Zazil Josefina IBARRA-CUEVAS – Master of Science in Computer Science and software developer working at a private company since 2022. Research interests: Data mining, data bases and software engineering.

Хосе Игнасио НУНЕС-ВАРЕЛА – доктор компьютерных наук, профессор, координатор бакалаврской программы по инженерии интеллектуальных систем в Автономном университете Сан-Луис-Потоси с 2017 года. Научные интересы: машинное обучение, наука о данных, интеллектуальная робототехника.

Jose Ignacio NUNEZ-VARELA – Doctor of Computer Science, professor, coordinator of the Intelligent Systems Engineering undergraduate program at the Autonomous University of San Luis Potosi since 2017. Research interests: Machine learning, data science, intelligent robotics.

Альберто НУНЕС-ВАРЕЛА – доктор компьютерных наук и доцент Автономного университета Сан-Луис-Потоси с 2014 года. Область научных интересов: разработка программного обеспечения, вывод на основе формальных грамматик, обработка естественного языка и машинное обучение.

Alberto NUNEZ-VARELA – Doctor of Computer Science and associate professor at the Autonomous University of San Luis Potosi since 2014. Research interests: Software engineering, grammatical inference, natural language processing, and machine learning.

Франсиско Эдуардо МАРТИНЕС-ПЕРЕС – доктор компьютерных наук, профессор, координатор бакалаврской программы по программированию в Автономном университете Сан-Луис-Потоси с 2023 года. Научные интересы: обработка изображений, окружающий интеллект, повсеместные вычисления, человеко-машинное взаимодействие и медицинская информатика.

Francisco Eduardo MARTINEZ-PEREZ – Doctor of Computer Science, professor, coordinator of the Computer Engineering undergraduate program at the Autonomous University of San Luis Potosi since 2023. Research interests: Image processing, ambient intelligence (AmI), ubiquitous computing, human–computer interaction, and medical informatics.

Сандра Э. НАВА-МУНЬОС – доктор компьютерных наук, профессор, координатор аспирантской программы по информатике в Автономном университете Сан-Луис-Потоси с 2023 года. Научные интересы: разработка программного обеспечения, человеко-машинное взаимодействие, контекстно-зависимые вычисления и медицинская информатика.

Sandra E. NAVA-MUÑOZ – Doctor of Computer Science, professor, coordinator of the Computer Science postgraduate program at the Autonomous University of San Luis Potosi since 2023. Research interests: Software engineering, human-computer interaction, context aware computing, and medical informatics.

Сесар Аугусто РАМИРЕС-ГАМЕС имеет степень магистра компьютерных наук, соискатель степени доктора философии. С 2023 года работает в коммерческой компании, где ведет разработку программного обеспечения. Научные интересы: компьютерное зрение, обработка изображений и машинное обучение.

César Augusto RAMÍREZ-GÁMEZ – Master of Science in Computer Science, his Ph.D. degree, and software developer working at a private company since 2023. Research interests: Computer vision, image processing, and machine learning.

Эктор Херардо ПЕРЕС-ГОНСАЛЕС – штатный профессор-исследователь Автономного университета Сан-Луис-Потоси (Мексика), имеет ученую степень доктора компьютерных

наук. Автор научных статей и глав в книгах по автоматизации проектирования программного обеспечения и человеко-машинного взаимодействия, выступал с научными докладами на международных конференциях в США, Канаде, Великобритании, Португалии и в Сингапуре. Область научных интересов: проектирование программного обеспечения, преподавание методов разработки программного обеспечения, обработка цифровых изображений, разработка программного обеспечения для квантовых компьютеров.

Hector Gerardo PEREZ-GONZALEZ – Full-time research professor at Universidad Autónoma de San Luis Potosi, Mexico. PhD in Computer Science from the University of Colorado in 2003. Author of research articles and book chapters on Automatic Software Design and Human-Computer Interaction. He has been a speaker at international conferences in the USA, Canada, UK, Portugal, and Singapore. His research areas are software design, computer science education, and quantum software engineering. He is a member of the National Researchers System in Mexico.