VOL.8 NO.2 September 2019 - ISSN 2233 – 1859

UOIuBIH
ORSinBIH
Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Comparison of Different Machine Learning Algorithms for Breast Cancer Recurrence Classification

M. Haskul

E.Yaman

Faculty of Engineering and Natural Sciences,
International University of Sarajevo International University of Sarajevo,
Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
mhaskul@ius.edu.ba

Article Info

Article history:
Article received on August 17, 2019
Received in revised form: September 1, 2019

Keywords:
data mining, breast cancer, classification, machine learning, Weka, C4.5, random forest, KNN

ABSTRACT: In this paper we compared some machine learning algorithms to predict recurrence of breast cancer and see which model used gives best accuracy for the prediction. In this study we used database donated by University Medical Centre, Institute of Oncology, Ljubljana, Slovenia. The preprocessed dataset includes 286 instances, 9 attributes and 1 class attribute. Firstly, we used attribute evaluation to see which attribute is more effective on class attribute. Secondly we have explored three different algorithms: C4.5, Random Forest and K Nearest Neighbor. Several data mining tools have been applied with these 3 algorithms to explore which model is better on accuracy. Finally we have found that C4.5 algorithm is the best for our dataset: breast cancer recurrence.

1. INTRODUCTION

A cancer is a broad term for a type of diseases characterized by abnormal cells which grow and invade healthy cells in the body. As one of cancer types breast cancer is the most common one especially diagnosed with the women all around the world, comprising 23% of all females around the globe [Ozmen, V. 2001, p.8]. The breast cancer starts in the cells of the breast as a group of cancer cells that can then invade surrounding tissues or spread to other areas of the body. In 2012, a diagnosis of breast cancer was received in the United States and almost 226.870 women had breast cancer, however, 39.510 of them died of breast cancer in the same year. It is really crucial to have earlier detection, a new personalized approach to treatment and better understanding of this cancer in order to increase survival rates and reduce the number of deaths associated with this disease [Tuncer, 2007]. World Health Organization, International Agency for Research on Cancer (WHO-IARC) predictions shows that the annual global burden of new breast cancer cases will reach 1.5 million and the majority of these will be seen in low-income countries [Yildirim, A., D. & Özaydin A., N. 2014, p. 47].

This study aims to compare three different algorithms which are C4.5, Random Forest and K-Nearest Neighbors (KNN) with using WEKA machine learning software to see which model gives us better accuracy for this case. After this comparison, also we can decide that which classification method should be used with the new data. In this study we used database donated by University Medical Centre, Institute of Oncology, Ljubljana, Slovenia. Thanks go to M. Zwitter and M. Soklic for providing the data.

We firstly have started with attribute evaluation to see which attribute is more effective on class attribute which is recurrence status for this database. We have applied chosen algorithms for two different cases. For the first one, we have used all attributes to see how accuracy will be obtained if all attributes are used. For the second one, attributes which are less effective on class attribute have been eliminated. And same algorithms have been used also for this case. C4.5, Random Forest and K-Nearest Neighbors (KNN) were used as algorithms to generate decision trees and classify dataset and finally to compare these three algorithm's results for two cases explained before.

2. LITERATURE REVIEW AND MATERIAL

A number of studies have been undertaken using data mining techniques applied in breast cancer dataset. Zand, H., K., K. (2015) and Gupta, S. & Kumar, D. & Sharma, A. (2011) for example, investigated the data mining techniques for breast cancer diagnosis and prediction. Testard P.Vaillant (2010) built a risk prediction model and Idowu, P.,A. & Williams, K., O. & Balogun, J., A. & Oluwaranti, A., I. (2015) focused on data mining techniques to predict breast cancer risks in Nigeria as well. Moreover, In A. Endo, T. Shibata, and H. Tanaka,(2008) implemented common machine learning algorithms to predict survival rate of breast cancer patient.

According to our research, there have not been many famous studies conducted regarding the case of recurrence of the cancer after pulling through. Beside this, using and comparing multiple data mining techniques which are C4.5, Random Forest and KNN makes this study distinguished from others.

Data title is breast cancer data and this database generated on 11 July 1988.

Here are some past usage of breast cancer database which we used for our work and their accuracy results:

Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986). The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann. – accuracy range: 66%-72% – Clark,P. & Niblett,T. (1987). Induction in Noisy Domains. In Progress in Machine Learning (from the Proceedings of the 2nd European Working Session on Learning), 11-30, Bled, Yugoslavia: Sigma Press. – 8 test results given: 65%-72% accuracy range – Tan, M., & Eshelman, L. (1988). Using weighted networks to represent classification knowledge in noisy domains. Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI. – 4 systems tested: accuracy range was 68%-73.5% – Cestnik,G., Konenenko,I, & Bratko,I. (1987). Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press. – Assistant-86: 78% accuracy.

Number of Instances: 286

This data set includes 201 instances of one class which is no-recurrence-events and 85 instances of another class that include recurrence-events. In this data there are 9 attributes which some are linear and some are nominal.

Number of Attributes: 9 plus the class attribute.

Attribute Information:

1- age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

2- menopause: lt40, ge40, premeno.

3- tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.

4- inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.

5- node-caps: yes, no.

6- deg-malig: 1, 2, 3.

7- breast: left, right.

8- breast-quad: left-up, left-low, right-up, right-low, central.

9- irradiat: yes, no.

Class Distribution:

1- no-recurrence-events: 201 instances

2- recurrence-events: 85 instances

[datahub.io, 2019]

3. ATTRIBUTE EVALUATION

One of the most important preliminary steps of data mining and machine learning solutions is to determine an appropriate subset of the attributes of the data which will be used in the analysis. For classification methods, this is done by looking at the ratio of an attribute to the class attribute. The attribute selection is the general name of the methods that determine the subset of attributes that will provide the highest benefit for analysis within a particular attribute space. For a set of n attributes, the size of this space will be 2n. Using all the attributes in the data set can be unnecessary or even harmful. The unnecessary attribute concept for the classification process refers to attributes that have no relation to the class value. For example, it is expected that there will be no relationship between the fuel consumption and color of a car. Keeping these attributes in the data set can lead complexity of the model that will be developed. The combination of highly correlated attributes (e.g., birth date and age) refers to overweight of the relevant pair of attributes and misleads the method of classification. For these reasons, attribute selection is vital for many machine learning and data mining methods. [Var, E.,2018] Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. [Kim,Y.,2003]

Table 1. The most efficient attributes of the dataset for recurrence prediction

NAME	SCORE	DEFINITION
deg-malig	0.07701	degree of malignancy
inv-nodes	0.069	the number (range 0 - 39) of axillary lymph nodes
tumor-size	0.05717	size of tumor
node-caps	0.05126	infection status
irradiat	0.02582	irradiation
age	0.01061	
breast-quad	0.00885	breast quadrant
breast	0.00249	left or right
menopause	0.002	

Attribute Evaluator : InfoGainAttributeEval

This evaluator explores how much important an attribute for class attribute [Weka, v3.8.3].

binarizeNumericAttributes is false

doNotCheckCapabilities is false

missingMerge is true

Search Method : Ranker

This searching method ranks attributes according to their own evaluations [Weka, v3.8.3].

Generate rating is true

numToSelect : -1

Attribute Selection Mode : Use full training set

We also used cross-validation mode to see ranks of attributes with 10 folds and 1 seed. Rank order of attributes is the same with full training set mode. Only difference is on ranks and they are changing between 0.001 and 0.011. Because of this high similarity, there is no need to show again these results after Cross-Validation Mode used.

According to rank results we have gotten from Attribute Evaluation, it can be said that malignancy degree is the most effective attribute on the class attribute which is recurrence status. And the number of axillary lymph nodes, tumor size, infection status, irradiation treatment and age are other attributes that have considerable amount of effect on the class attribute prediction respectively. Firstly we have applied various algorithms to the database with all attributes. Also according to rank list last three attributes have been eliminated and we have applied same algorithms to the database to see how much important attribute selection is. [Greenough, B.,1925]

4. CLASSIFICATION METHODS

The concept of classification is distributing data between the various classes defined on a data set simply. The classification algorithms learn this distribution from the given training set and then try to classify data properly when the new data comes. In this work, three different algorithms have been examined and compared which are C4.5, Random Forest and K-Nearest Neighbors (KNN) for the breast cancer dataset.

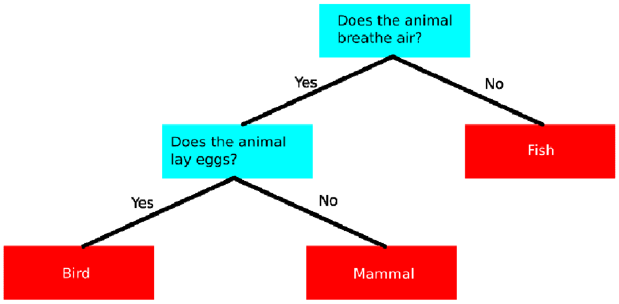

Figure 1. A simple decision tree model for classification. [Bickerton, C., 2018]

4.1 C4.5

In the late 1970s and early 1980s, J. Ross Quinlan who is a machine learning researcher developed a decision tree algorithm known as ID3. When Quinlan was working on developing ID3, a group of statisticians (L. Breiman, J. Friedman, R. Olshen and C. Stone) published a book called “Classification and Regression Trees” (CART) describing the formation of binary decision trees. ID3 and CART were developed independently of each other. However, they had a similar approach for decision trees by using training variables. These two essential algorithms were the source of many studies on decision tree. Quinlan continued his study on data mining and developed C4.5 and ID3 became an ancestor of C4.5 algorithm. [Alan, M.,2014]

C4.5 is the most used decision tree algorithm nowadays. C4.5 algorithm is an example of decision tree algorithms. This algorithm checks all attributes in every step and it calculates information gain after normalized them. Actually this normalization is the main difference between C4.5 and ID3 decision trees. If best information gain is given by which attribute, that attribute becomes a new decision on the tree. And for the below decisions, same method continuous until when class attribute is reached. [Seker, S., 2012]

Due to copyright situation, Weka presents C4.5 algorithm as J48 to the users.

4.2 Random Forest

Random Forest is a controllable machine learning algorithm. As the name suggests, it creates a forest completely randomly. The forest it established is a collection of decision trees trained by begging method. Begging method works as an ensemble of learning models makes more trustable the overall result. This algorithm provides us stably and fast prediction. The best important advantage of Random Forest is that it can be used for both classification and regression which these are most used techniques in machine learning. Random Forest, while growing trees, add additional randomness to the model. Instead of looking for the most important feature when dividing a node into pieces, it searches for the best feature among a random feature subset. This usually results in a wide variety that results in a better model. [Donges, N., 2018]

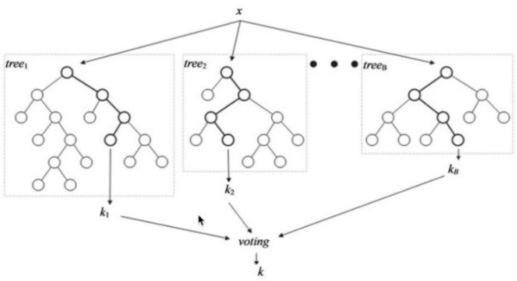

Figure 2. An example model for random forest algorithm. [Seker, S., Erdogan, D.]

It creates different decision trees and branches and more than one tree is produced to make a decision. In Figure 2, more than one decision tree was created from the same data set. The decision which is generated from each of them may be different. For example, k1 from the first, k2 from the second and k3 from the third became decisions separately. Finally, they will be voted among themselves. As a result, it acts as a single algorithm. [Seker, S., Erdogan, D.]

4.3 K-Nearest Neighbors (KNN)

The KNN algorithm or the K-Nearest neighbor algorithm is one of the most known and used algorithms in machine learning algorithms. Classification is applied by using the closeness of a selected property between it and its nearest neighbor. The value of K in this case is denoted by a number, for example 4 or 6. For determining the distance between objects, the formula in Equation-1 is used.

$$d_{(ij)} = \sqrt{\Sigma} \quad (1)$$

According to the data defined, firstly K value is checked when a new object that needs to be defined comes. The K number is usually selected as the odd number since there should not be equality [Kilinc, D.,2016].

Due to copyright situation, Weka presents KNN algorithm as IBk to the users.

5. RESULTS AND DISCUSSION

This part of our work shows us the results which are obtained from three different algorithms explained previous part. C4.5, Random Forest and KNN algorithms have been used for breast cancer recurrence.

In data mining studies, the data set is separated into two groups as training and test sets to examine the success of the applied method. In the k fold cross-validation method, firstly k value is selected. The dataset is divided into number of k parts. Firstly, one of the pieces is selected for testing and the rest are used for training. It doesn't matter where you start from. Also for the other folds this process is repeated. As a result, we run the same method k times in k different training and test sets. [Seker, S.,2013]

5.1 C4.5

C4.5 algorithm has been applied with k fold cross validation. We chose 10 folds for the first time and we also applied C4.5 with K=5, 15, 20, 30, 40, 50, 60, 70, 80, 100 folds. We wanted to see how much important are number of folds for this case. But each time when we change number of folds, accuracy did not change significantly with database used.

With using C4.5 algorithm, it can be seen from the matrix below that 7 patient data are classified wrong for the class “no-recurrence-events”. And also it can be seen that 62 data are misclassified as no-recurrence-event which this number is bigger than the before one relatively. As is seen, this type of decision tree model classified "no recurrence event" predictions almost correctly but it misclassified "recurrence event" predictions mostly.

Table 2. Confusion matrix of C4.5 (All attributes are used.)

Classified as:	a	b	Total
a= no-recurrence-events	194	7	201
b= recurrence-events	62	23	85

According to attributes, which are infection status and malignancy degree, C4.5 algorithm has generated the decision tree as in figure 3.

217 instances classified correctly with this algorithm and it can be seen that accuracy of this model is 75.8741%.

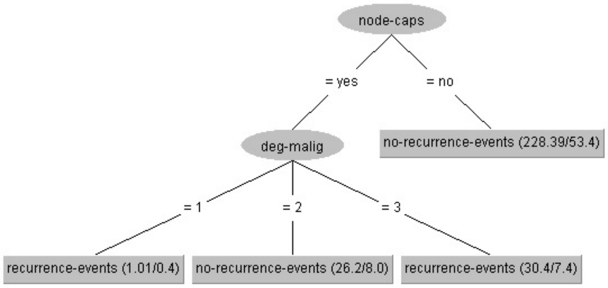
node-caps
= yes
= no
deg-malig
no-recurrence-events (228.39/53.4)
= 1
= 2
= 3
recurrence-events (1.01/0.4)
no-recurrence-events (26.2/8.0)
recurrence-events (30.4/7.4)

Figure 3. Decision tree from breast cancer data based on recurrence situation.

C4.5 algorithm has been used with k=10 cross validation for two situation. In the table 2, we reported results with using all attributes. In the table 3, we also reported results with selected attributes in the attribute evaluation part of this study.

Table 3. Confusion matrix of C4.5 (3 attributes are eliminated)

Classified as:	a	b	Total
a= no-recurrence-events	193	8	201
b= recurrence-events	63	22	85

215 instances classified correctly with this algorithm and it can be seen that accuracy of this model is 75.1748 %. Menopause, breast and breast-quad are eliminated because their ranks are low relatively to others and it means these attributes are not effective too much on class attribute. But we did not see considerable difference between those two accuracy.

5.2 Random Forest

We have used Random Forest algorithm for two different model to see the difference between their accuracy when k value is changed. Firstly we applied random forest with using k fold cross validation with k=10. After that we applied it with k=40. We have applied this algorithm with different k values because we wanted to see how can we increase accuracy by changing number of k. For the second one, k=40 has been chosen because of our tries we got after changing k value a lot of times. We have gotten most accurate result at k=40. We entered 0 to the max depth option because we already have a small size of database and small number of attributes and 0 means there is no limit of number of branch. Bagging is applied with 100 iteration for all works with Random Forest.

Also we have applied these two different models two times by eliminating some attributes to see the difference.

Table 4. Confusion matrix of Random Forest algorithm. (10 folds cross validation, all attributes are used)

Classified as:	a	b	Total
a= no-recurrence-events	175	26	201
b= recurrence-events	61	24	85

In table 4, 199 instances classified correctly with Random Forest algorithm applied with 10 folds cross validation with all attributes. And its accuracy is 69.5804%.

Table 5. Confusion matrix of Random Forest algorithm. (10 folds cross validation, 3 attributes are eliminated)

Classified as:	a	b	Total
a= no-recurrence-events	169	32	201
b= recurrence-events	62	23	85

In table 5, 192 instances classified correctly with Random Forest algorithm applied with 10 folds cross validation with selected attributes. Menopause, breast and breast-quad are eliminated like for C4.5 as well. And its accuracy is 67.1329% now.

It can be seen that, when we eliminate some attributes accuracy of Random Forest algorithm is reduced a bit. So we can say that using all attributes can give better accuracy for this case. Also we can see that when we eliminate some attributes, big difference is occurred on no-recurrence-events.

Table 6. Confusion matrix of Random Forest algorithm. (40 folds cross validation, all attributes are used.)

Classified as:	a	b	Total
a= no-recurrence-events	180	21	201
b= recurrence-events	57	28	85

208 instances have been classified correctly by applying Random Forest algorithm with 40 folds cross validation with all attributes in table 6. And its accuracy is 72.7273%. According to our set of trying, applying 40 folds cross validation to the Random Forest algorithm has been given us best accuracy with using all attributes.

Table 7. Confusion matrix of Random Forest algorithm. (3 attributes are eliminated)

Classified as:	a	b	Total
a= no-recurrence-events	167	34	201
b= recurrence-events	56	29	85

In table 7, 196 instances have been classified correctly by applying Random Forest algorithm with 40 folds cross validation with selected attributes. And its accuracy is 68.5315%. Reducing attribute numbers does not give us better option for Random Forest models as is seen.

5.3 K-Nearest Neighbors (KNN)

KNN algorithm has been applied for two cases. Firstly it has been applied with all attributes and secondly we have applied KNN with selected attributes. For this algorithm, KNN -nearest neighbor number- should have been selected in the beginning. It shows us how many neighbor of an object selected to calculate distance between an object and its nearest neighbor. According to our tries on it, the number of neighbor used should be 4 for this model since we have gotten best accuracy with 4 neighbors.

Table 8. Confusion matrix of KNN algorithm. (All attributes are used)

Classified as:	a	b	Total
a= no-recurrence-events	194	7	201
b= recurrence-events	66	19	85

210 instances have been classified correctly by applying KNN algorithm with number of KNN= 4 with all attributes. And its accuracy is 74.4755%. As it is seen in table 8, KNN has misclassified 66 instances for the class attribute, recurrence-events whereas 7 instances have been misclassified for no-recurrence-events class attribute.

Table 9. Confusion matrix of KNN algorithm. (3 attributes are eliminated)

Classified as:	a	b	Total
a= no-recurrence-events	190	11	201
b= recurrence-events	61	24	85

214 instances have been classified correctly by applying KNN algorithm with same number of KNN with previous one. But in this try 3 attributes are eliminated. Its accuracy is 74.8252%. If we compare table 8 and table 9, it can be seen that accuracy of class attributes changed. When we eliminate some attributes, accuracy of no-recurrence-events classification reduced a bit whereas accuracy of recurrence-events increased a bit. When we look at total accuracy difference, it can be said that accuracy is increased by 0.4% when we reduce number of attributes.

6. CONCLUSION

Nowadays data mining tools have very important area in the variety of sectors. Data mining, which is used to reveal confidential, valuable, usable information and provide strategic decision support from a large amount of data, has created a new perspective in the use of health data as well as responding to problem areas related to large amounts of data.

In this study we applied mainly three different machine learning algorithms to the breast cancer database. We tried to explain the basic definitions of these algorithms. C4.5, Random Forest decision tree models and K-Nearest Neighbor are used widely today and before. They have significant importance in the related sectors. In this work, we compared their results after applied these methods to the our database. Also we compared two different test options on Random Forest algorithms and we got accuracy results of them. In the following tables, comparisons are shown.

Table 10. Comparison of accuracy between models used.

Algorithms	Models	Folds	Accuracy
C4.5	All attributes	10	75.87%
C4.5	Selected attributes	10	75.17%
Random Forest	All attributes	10	69.58%
Random Forest	Selected attributes	10	67.13%
Random Forest	All attributes	40	72.73%
Random Forest	Selected attributes	40	68.53%
KNN	All attributes	10	74.48%
KNN	Selected attributes	10	74.83%

It can be seen that for our database and attributes, applying C4.5 algorithm with all attributes given gives us best accuracy comparing to other models used. But also the other algorithms in compatible models gave us good result close to C4.5.

Except KNN algorithm, attribute selection did not give us better accuracy at all. We could increase accuracy a bit with using it on KNN only. According to our dataset, increasing number of folds from 10 to 40 also is useful to increase accuracy in Random Forest algorithm.

Finally, it can be said that this work ended up with accuracy range 67%-75%.

This study shows us some parameters are very important to predict for the new cases related to the breast cancer. Data mining tools gives us an opinion which we can use this opinion to predict some important cases. But of course this is too risky for health area. Even if we have 100% accuracy this means all predictions done are correct but we cannot talk 100% sure about the new cases. It gives us an opinion.

REFERENCES

Alan, M., “Karar Ağaçlarıyla Öğrenci Verilerinin Sınıflandırılması”,Ataturk Universitesi Iktisadi ve Idari Bilimler Dergisi, Cilt: 28, Sayı: 4, 2014

Bickerton, C., “A beginner’s guide to decision tree classification”,[https://towardsdatascience.com/a-beginners-guide-to-decision-tree-classification-6d3209353ea], 2018

Dias, J., "Breast cancer diagnostic typologies by grade-of-membership fuzzy modeling", Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering (P-131)

Endo A., Shibata T., and H. Tanaka,. “Comparison of seven algorithms to predict breast cancer survival,‖ Biomedical Soft Computing and Human Sciences”, vol. 13 2, pp. 11–16, 2008

Greenough, R., “Varying Degrees of Malignancy in Cancer of the Breast”, 10.1158/jcr.1925.453 Published,December,1925

Gupta, S.; Kumar, D., Sharma, A. 2011. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE). Vol. 2 No. 2 pg 198-195, April, 2011. ISSN: 0976-5166. Accessed on June 24, 2014.

Idowu, PA, Williams KO, Balogun JA, Oluwaranti AI. 2015. Breast Cancer Risk Prediction Using Data Mining Classification Techniques. Transactions on Networks and Communications. 3(2):1-11.

Kilinc, D., Borandag, E., Yucalar, F., Tunali, V., Simsek, M., Ozcift, A., “KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi”, Marmara Fen Bilimleri Dergisi, 3: 89-94, 2016

Kim,Y., Street,N.,Menczer,F., “Feature selection in data mining”, Conference Paper,USA,2003

Koyuncugil, A., Ozgulbas, N., “Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları”, Bilisim Teknolojileri Dergisi,May,2009

Ozmen, V., “Breast cancer in the world and Turkey”. J Breast Health; 4:7-11, 2008.

Seker,S.,“C4.5Ağacı”,[http://bilgisayarkavramlari.sadievrenseker.com/2012/11/13/c4-5-agaci-c4-5-tree/], 2012

Seker,S.,“K Fold Cross Validation”, [http://bilgisayarkavramlari.sadievrenseker.com/2013/03/31/k-fold-cross-validation-k-katlamali-carpraz-dogrulama/],2013

Testard, P.-Vaillant, “The war on cancer” CNRS international magazine, vol. 17, pp. 18– 21, 2010.

Tuncer,M., “Turkiye’de kanser kontrolu”, Saglik Bakanligi, Kanserle Savas Daire Baskanligi, Ankara: Onur Matbaacılık Ltd. Şti, 2007.

Var, E., Inan, A., “Differentially private attribute selection for classification”, Journal of the Faculty of Engineering and Architecture of Gazi University 33:1, 2018

Yildirim, A., D. & Ozaydin, N., A. “Sources of Breast Cancer Knowledge of Women Living in Moda / İstanbul and Their Attendance to Breast Cancer Screening” J Breast Health; 10: 47-56, 2014.

Zand, H., “A Comparative Survey On Data Mining Techniques For Breast Cancer Diagnosis And Prediction.” Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online) An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/01/jls.htm 2015 Vol.5 (S1), pp. 4330-4339/Karim,2015