Machine Learning for the Evolutionary Analysis of Breast Cancer

Aprendizaje automático para el análisis evolutivo de Cáncer de mama

Alexander Mackenzie Rivero[1,*], Alberto Rodríguez Rodríguez[1,†], Edwin Joao Merchán Carreño[1,‡], and Rodrigo Martínez Béjar[2]
[1]Universidad Estatal del Sur de Manabí (UNESUM), Ecuador.
[2]Universidad de Murcia, Spain.
mackenzie.alexander@unesum.edu.ec

Received: August 15, 2017 — **Accepted:** September 15, 2017

How to cite: Mackenzie Rivero, A., Rodríguez Rodríguez, A., Merchán Carreño, E. J., & Martínez Béjar, R. (2018). Machine Learning for the Evolutionary Analysis of Breast Cancer. *Journal of Science and Research: Revista Ciencia e Investigación, 3(CITT2017), 44-49.* https://doi.org/10.26910/issn.2528-8083vol3issCITT2017.2018pp44-49

***Abstract*—The use of machine learning allows the creation of a predictive data model, as a result of the analysis in a data set with 286 instances and nine attributes belonging to the Institute of Oncology of the University Medical Center. Ljubljana. Based on this situation, the data are preprocessed by applying intelligent data analysis techniques to eliminate missing values as well as the evaluation of each attribute that allows the optimization of results. We used several classification algorithms including J48 trees, random forest, bayes net, naive bayes, decision table, in order to obtain one that given the characteristics of the data, would allow the best classification percentage and therefore a better matrix of confusion, Using 66 % of the data for learning and 33 % for validating the model. Using this model, a predictor with a 71,134 % effectiveness is obtained to estimate or not the recurrence of breast cancer.**

***Keywords*—Intelligent analysis, Breast cancer, Machine learning.**

***Resumen*—El uso del aprendizaje automático permite la creación de un modelo predictivo de datos, como resultado del análisis en un conjunto de datos con 286 instancias y nueve atributos pertenecientes al Instituto de Oncología del Centro Médico Universitario. Ljubljana. En función de esta situación, los datos se preprocesan aplicando técnicas inteligentes de análisis de datos para eliminar los valores perdidos, así como la evaluación de cada atributo que permite la optimización de resultados. Utilizamos varios algoritmos de clasificación incluyendo árboles J48, bosque aleatorio, bayes net, bayes naive, tabla de decisiones, para obtener uno que, dadas las características de los datos, permita el mejor porcentaje de clasificación y por lo tanto una mejor matriz de confusión, utilizando 66 % de los datos para aprendizaje y 33 % para validar el modelo. Con este modelo, se obtiene un predictor con una eficacia del 71,134 % para estimar o no la recurrencia del cáncer de mama.**

***Palabras Clave*—Análisis inteligente, Cáncer de mama, Aprendizaje automático.**

INTRODUCTION

Cancer represents the fifth leading cause of death worldwide. In 2015, 1.7 million people died as revealed by the World Health Organization's according to this report, breast cancer is the most common among women in both developed and developing countries, and reprecents 16 % of female cancer.

Breast cancer survival rates vary widely across the world, from 80 % or more in North America, Sweden and Japan, to around 60 % in middle-income countries, to below 40 % in low-income countries Coleman (Coleman et al., 2008).

In the last two decades, machine learning has become one of the pillars of information technology and a high potential of applicability. With the increasing amount of data available, there are sufficient reason to believe that intelligent data analysis is a technique that allows the extraction of knowledge as a necessary ingredient for technological progress (Quadrianto et al., 2010).

In a society where data have become the unit of measurement that records our behavior, every activity we do is more and more common to be stored in an electronic medium.

Having tools that allow the intelligent analysis of data makes it possible to extract knowledge, classify and detect trends in order to contribute to the advancement of science and technology.

It is useful to characterize learning problems according to the type of data, they use since problems with similar types of data can be solved with very ampls yous techniques. For example, natural language processing and bioinformatics. Vectors are the most basic entity that we could find in our work. In a life insurance company it may be interesting to obtain a vector of variables (blood pressure, heart rate, height, weight, cholesterol level, if you are a smoker, gender, among others) to infer the life expectancy of a potential customer . A farmer might be interested in determining a fruit's maturity based on (size, weight, spectral data). An engineer may want dependencies (voltage pairs, current). To end documents

[*]Master en tecnologías de la información y telemática avanzada.
[†]Doctor en ciencias pedagógicas.
[‡]Magister en docencia universitaria e investigación educativa.

may be represented by a summary vector that describe the occurrence of words Smola and Vishwanathan.

The objective of this research is to describe the technical aspects of the design of a model that uses machine learning to predict or not the recurrence of breast, based on the learning obtained from the analysis of a database of breast cancer, conformed by 286 instances and nine attributes pertaining to the Institute of Oncology of the University Medical Center. Ljubljana. Also the procedures performed in the development of the research are determined.

The research consisted of the development of 3 stages:

- Data pre-processing by applying intelligent data analysis techniques to eliminate missing values as well as the evaluation of each attribute to optimize results. As part of this process, the Feature selection algorithm (FS) is performed, the process of eliminating features from the data base that are irrelevant with respect to the task to be performed Liu.
- Data classification, with the purpose we used J48 tree algorithms, random forest, Bayesian algorithms like bayes net, naive bayes, and decision tables.
- Evaluation of models after the application of the algorithms using percentage split using 66 % of the data for learning and 33 % for validating the model to select the better classification algorithm in according to the type of data analyzed, so as to obtain one that given the characteristics of the data, allows to have the best classification and hence a better matrix of confusion.

MATERIALS AND METHODS

Data Sets

The data sets analysis has been performed using the WEKA 3.6 software, developed by the University of Waikato. The oncology data set was composed by 286 instances and 9 attributes of the Institute of Oncology of the University Medical Center, Ljubljana. The data set has been pre-processed as follows.

Confusion Matrix

A confusion matrix illustrates the accuracy of the solution to a classification problem. Given n classes a confusion matrix is a m x n matrix, where C, indicates the number of tuples were assigned to class C, but where the correct class is C. Obviously the best solution will have only zero values outside the diagonal (Dunham, 2006).

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study COE (2012).

1. a is the number of correct predictions that an instance is negative
2. b is the number of incorrect predictions that an instance is positive.

Data pre-processing

This first step was to apply intelligent data analysis techniques attributes at eliminating missing values using ReplaceMissing-Values: Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data. RemoveUseless: This filter removes attributes that do not vary at all or that vary too much. All constant attributes are deleted automatically, along with any that exceed the maximum percentage of variance parameter. The maximum variance test is only applied to nominal attributes.

MaximumVariancePercentageAllowed set the threshold for the highest variance allowed before a nominal attribute will be deleted.

Specifically, if (number of distinct values / total number of values x 100) is greater than this value then the attribute will be removed. Feature Selecction was also used feature selection is the process of eliminating features from the data set that are irrelevant with respect to the task to be performed. Its main aim is to determine a minimal subset of features from a problem domain while retaining a suitably high accuracy in representing the original features. Feature selection finds useful features to represent the data and remove non-relevant ones, and simplifies the implementation of the classifier itself by determining what features should be made available to it (Liu and Motoda, 1998).

The information shown in Table 1 corresponds to the first step of the research, where the structure of breast cancer data set used is described to identify each attribute.

Table 1. Attribute information for breast cancer data set.

Name	Description	Type	Limits
Age	Age range	Real	10-19;20-29;30-39; 40-49;50-59;60-69; 70-79;80-89;90-99
Menopause	Menopause momento	Discrete	lt40;ge40;premeno
Tumor-size	Tumor size excised in mm	Real	0-4;5-9;10-14;15-19; 20-24;25-29;30-34; 35-39;40-44;45-49; 50-54;55-59
Inv-nodes	A metric of presence	Real	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
Node-caps	Evidence that cancer cells	Discrete	Yes;no
Deg-malig	Tumor Histological grade	Real	1;2;3
Breast	Breast affected	Discrete	left; rigth
Breastquad	Breast quadrant	Discrete	left-up, left-low, right-up,right-low, central
Irradiat	Radiotherapy	Discrete	yes;no

Source: Prepared by the authors.

Data Classification

Five algorithms were selected and evaluated. The classifier J48 is the Weka implementation of the decision tree C4.5 Witten and Frank. It is known to be computationally very efficient and to guarantee the interpretability of the results. Briefly, C4.5 builds decision trees from a set of training data by using the information entropy gain criterion. At each node of the tree, C4.5 Witten et al. (2016) chooses the attribute of the data that

most effectively splits its set of samples into subsets, each one belonging to one of the predefined classes. The splitting criterion is the normalized information gain: the feature with the highest normalized information gain is chosen to make the decision (Quinlan, 1993).

Random Forest: The random forest machine learner is a meta-learner this meaning consisting of many individual learners (trees). The random forest uses multiple random trees classifications to vote on an overall classification for the given set of inputs. In general in each individual machine learner vote is given equal weight. The forest chooses the individual classification that contains most of the votes (Livingston, 2005).

Bayes net: Nodes in a Bayes net represent random variables with (usually) a discrete set of values (e.g. a utensil node could have values (knife, fork, spoon)). Linksin the net represent (via tables) conditional probabilities that a node has a particular value given that an adjacent node has a particular value. Belief in the values for node X is calculated as

$$BEL(x) = P(x/e) \tag{1}$$

where 'e' is the combination of all evidence present in the net. Evidence, produced by running a visual action, directly supports the possible values of a particular node (i.e. variable) in the net. There exist a number of evidence propagation algorithms, which recompute belief values for all nodes given one new piece of evidence (Rimey and Brown, 1992).

Naive Bayes: The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent of one another given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems (Dimitoglou et al., 2012).

Decision table: are one of the simplest hypothesis spaces possible, and usually they are easy to understand. Experimental results show that on artificial and real-world domains containing only discrete features, a decision table classier searches for exact matches in the decision table using only the features in the schema (Kohavi, 1995).

RESULTS AND DISCUSSIONS

We have performed classification using the Random Forest algorithm, J48 decision tree algorithm, Bayes Net algorithm, Naive Bayes algorithm and decision table on breast-cancer.arff in Weka software.

Results for Classification Using the Random Forest Algorithm

The confusion matrix is generated for class no-recurrence-events with two possible values no-recurrence-events or recurrence-events.

a b ¡– classified as
54 10 — a = no-recurrence-events
23 10 — b = recurrence-events

For the above confusion matrix, it was obtained that 10 patients with no recurrence- events were classified as recurrence-events and 23 patients with recurrence- events were classified as no-recurrence-events. Correctly classified instances were 64 = 65.9794 % and incorrectly classified instances were 33 = 34.0206 %.

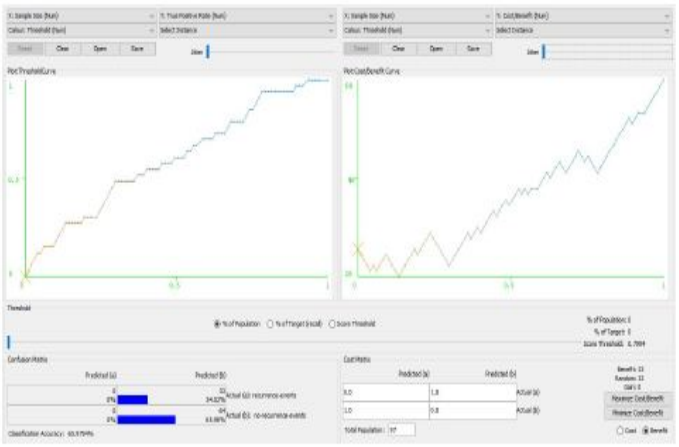

Figure 1. Weka cost / benefit analysis of recurrence events.
Source: Prepared by the authors.

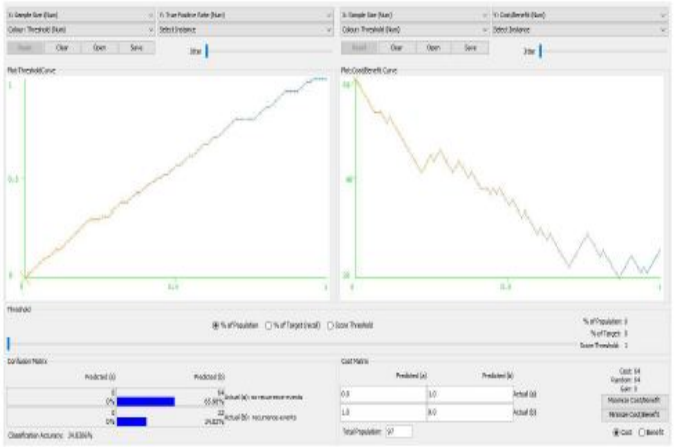

Figure 2. Weka cost / benefit analysis of no-recurrence events.
Source: Prepared by the authors.

Results for classification using the J48 decision tree algorithm

Confusion matrix:

a b ¡– classified as
56 8 — a = no-recurrence-events
23 10 — b = recurrence-events

For the above confusion matrix, it was obtained that 8 patients with no-recurrenceevents were classified as recurrence-events and 23 patients with recurrence-events were classified as no-recurrence-events. Correctly classified instances were 66 = 68.0412 % and the number of incorrectly classified instances were 31 = 31.9588 %.

J48 Tree, see Figure 2

node-caps = yes
— deg-malig = 1: recurrence-events (0.0)
— deg-malig = 2: no-recurrence-events (26.0/8.0)
— deg-malig = 3: recurrence-events (30.0/7.0)
node-caps = no: no-recurrence-events (230.0/54.0)
Number of leaves: 4
Size of the tree: 6

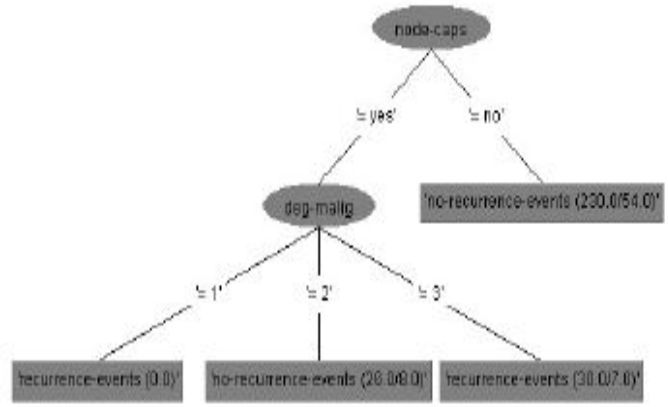

Figure 3. Tree of the J48 algorithm used.
Source: Prepared by the authors.

Results for classification using the Bayes Net algorithm

Confusion matrix:

a b ¡– classified as
52 12 — a = no-recurrence-events
17 16 — b = recurrence-events

For the above confusion matrix, it was obtained that 12 patients with norecurrence-events were classified as recurrence-events and 23 patients with recurrence-events were classified as no-recurrence-events. Correctly classified instances were 68 = 70.1031 % and incorrectly classified instances were 29 = 29.8969 %.

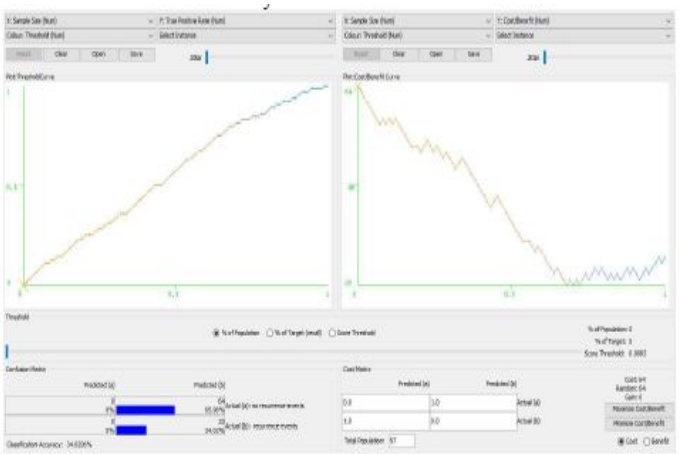

Figure 4. Weka cost / benefit analysis of no-recurrence events.
Source: Prepared by the authors.

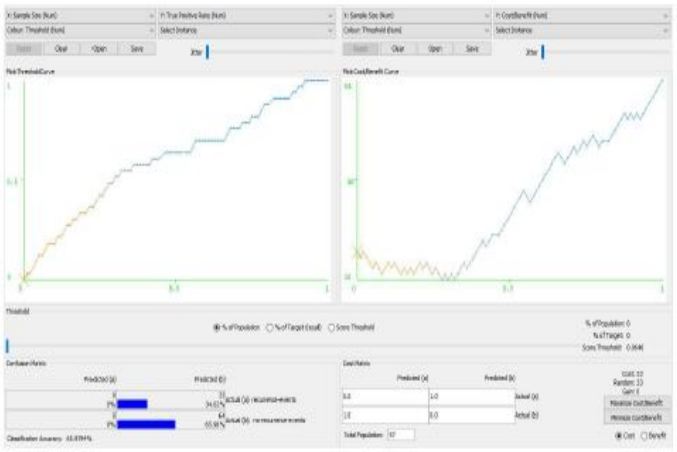

Figure 5. Weka cost / benefit analysis of recurrence events.
Source: Prepared by the authors.

Results for classification using the Naive Bayes algorithm

Confusion matrix:

a b ¡– classified as

53 11 — a = no-recurrence-events

17 16 — b = recurrence-events

For the above confusion matrix, was it obtained that 11 patients with norecurrence-events were classified as recurrence-events and 17 patients with recurrence-events were classified as no-recurrence-events. Correctly classified instances were 69 = 71.134 % and incorrectly classified instances were 28 = 28.866 %.

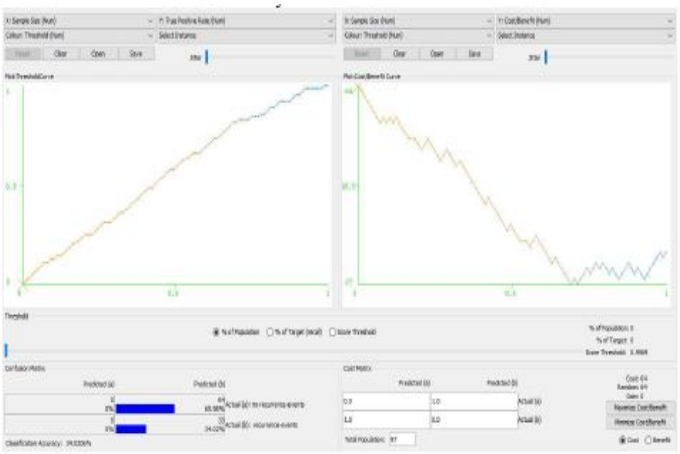

Figure 6. Weka cost / benefit analysis of no-recurrence events.
Source: Prepared by the authors.

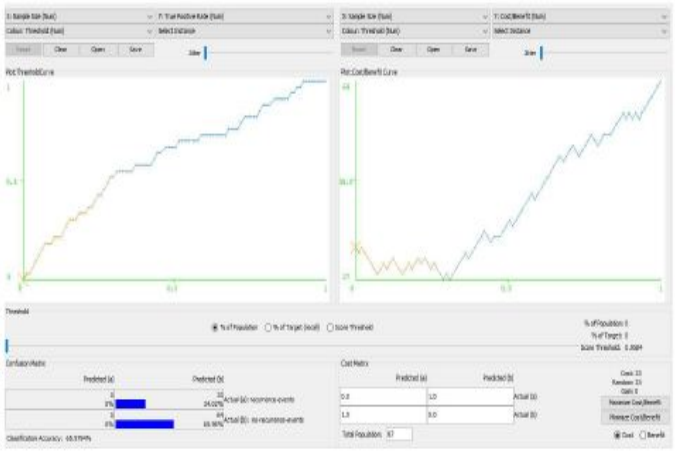

Figure 7. Weka cost / benefit analysis of recurrence events.
Source: Prepared by the authors.

Results for classification using the decision table algorithm

Confusion matrix:

a b ¡– classified as

60 4 — a = no-recurrence-events

27 6 — b = recurrence-events

For the above confusion matrix, it was obtained that 4 patients with no-recurrence-events were classified as recurrence-events and 27 patients with recurrence-events were classified as no-recurrence-events. Correctly classified instances were 66 = 68.0412 % and incorrectly classified instances were 31 = 31.9588 %.

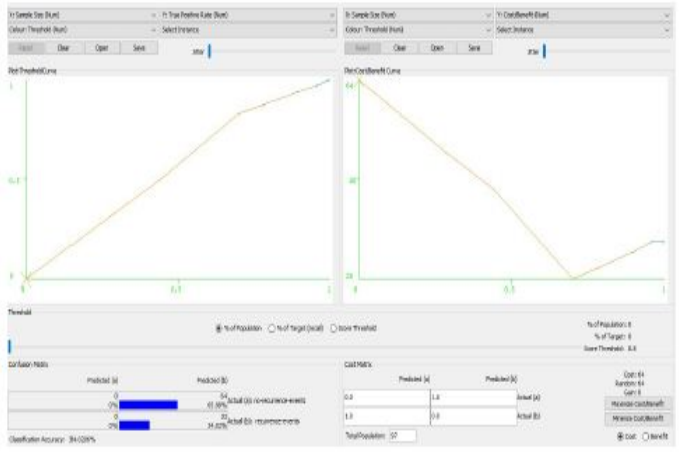

Figure 8. Weka cost / benefit analysis of no-recurrence events.
Source: Prepared by the authors.

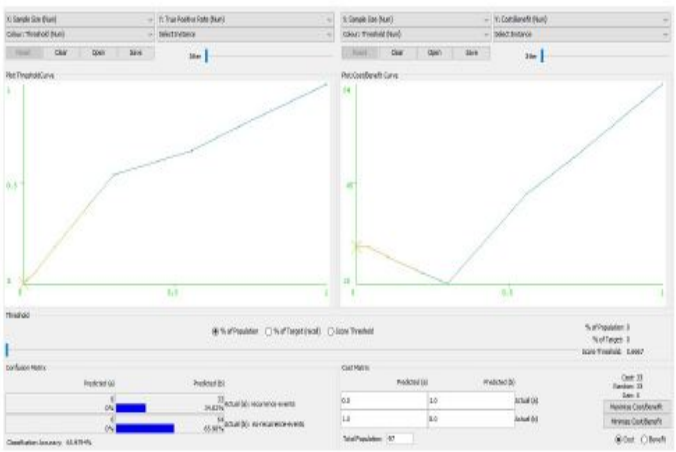

Figure 9. Weka cost / benefit analysis of recurrence events.
Source: Prepared by the authors.

Table 2. Results of the algorithms used.

Algorithm	Random Forest	J48	BayesNet	Naive Bayes	Decision table
Correctly classified	65,97 %	68,04 %	70,10 %	71,13 %	68,04 %
incorrectly classified	34,02 %	31,95 %	29,89 %	28,86 %	31,95 %
Precision no-recurrence-events	0,70	0,70	0,75	0,75	0,69
Precision recurrence-events	0,5	0,556	0,571	0,593	0,6

Source: Prepared by the authors.

CONCLUSIONS

This study examines the ability of a set of basic machine learning methods to accurately predict the recurrence or not of breast cancer, achieving an accuracy of 71,134 % using the Naïve bayes algorithm. Based on this model a physician can predict recurrence of breast cancer, having to enter the following patient data: age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat.

In order to continue this research, larger datasets will be used to increase the accuracy of this model.

BIBLIOGRAPHIC REFERENCES

COE, J. (2012). Performance comparison of naïve bayes and j48 classification algorithms. *International Journal of Applied Engineering Research*, 7(11):2012.

Coleman, M. P., Quaresma, M., Berrino, F., Lutz, J.-M., De Angelis, R., Capocaccia, R., Baili, P., Rachet, B., Gatta, G., Hakulinen, T., et al. (2008). Cancer survival in five continents: a worldwide population-based study (concord). *The lancet oncology*, 9(8):730–756.

Dimitoglou, G., Adams, J. A., and Jim, C. M. (2012). Comparison of the c4. 5 and a naïve bayes classifier for the prediction of lung cancer survivability. *arXiv preprint arXiv:1206.1121*.

Dunham, M. H. (2006). *Data mining: Introductory and advanced topics*. Pearson Education India.

Kohavi, R. (1995). The power of decision tables. In *European conference on machine learning*, pages 174–189. Springer.

Liu, H. and Motoda, H. (1998). *Feature extraction, construction and selection: A data mining perspective*, volume 453. Springer Science & Business Media.

Livingston, F. (2005). Implementation of breiman's random forest machine learning algorithm, in ece591q machine learning conference.

Quadrianto, N., Petterson, J., Caetano, T. S., Smola, A. J., and Vishwanathan, S. (2010). Multitask learning without label correspondences. In *Advances in Neural Information Processing Systems*, pages 1957–1965.

Quinlan, J. (1993). C4. 5: Programs for machine learning morgan kaufmann publishers san francisco. *CA Google Scholar*.

Rimey, R. D. and Brown, C. M. (1992). Where to look next using a bayes net: Incorporating geometric relations. In *European Conference on Computer Vision*, pages 542–550. Springer.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.