ISSN 1870-4069

Application of Data Mining Techniques and Algorithms for the Detection of Breast Cancer

José E. Zagal Solano[1], Miguel A. Ruiz Jaimes[1], Juan J. Flores Sedano[1],
Mayra L. Bonilla Monzón[1], Gloria V. Urquiza Flores[1],
Marco A. Valerio Reyes[1], Jorge A. Ruiz Vanoye[2]

[1] Universidad Politécnica del Estado de Morelos,
Mexico

[2] Universidad Politécnica de Pachuca,
Mexico

{jzagal, mruiz, fsjo161286, bmmo16002, ufgo160163,
vrmo161129}@upemor.edu.mx, jorge@ruizvanoye.com

Abstract. Breast cancer is a disease that affects a large part of world society. It is the most diagnosed cancer in women and its prevention although it is not impossible, it seems really difficult since its cause is unknown, so early detection is based on the patient's forecast. The most common form for detection is self-exploration, however this is only detected in more advanced stages. That is why during the article some data mining techniques are presented along with three artificial intelligence algorithms, using data obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia and using the algorithms it is possible to identify patients likely to suffer from this disease.

Keywords: Classification algorithms, cancer of mama, k-nn, naïve Bayes, decision trees, machine learning.

1 Introduction

Breast cancer is the most common among women in the world, it represents 16% of all female cancers. It is estimated that in 2004, 519,000 women died from breast cancer.

Survival rates vary greatly, from 80% or more in the USA, Sweden and Japan, approximately 60% in middle-income countries, to figures below 40% in low-income countries (Coleman et al., 2008). Low survival rates can be mainly explained by the lack of early detection programs [2].

In Mexico also since 2006, breast cancer is the leading cause of cancer death in women. An occurrence of 20,444 cases in women is estimated annually, with an incidence of 35.4 cases per 100,000 women. The entities with the highest mortality from breast cancer are Coahuila (24.2), Sonora (22.6) and Nuevo León 22.4) [3].

During this article we will describe the algorithms k-NN, decision trees and Naive Bayes that were used for the detection of patients with the possibility of suffering cancer based on attributes such as tumor size, menopause, age, etc. Always looking for the highest accuracy range for each algorithm.

2 Related Works

There are currently several works focused on the detection of breast cancer, mainly those with a medical approach, which are based on the analysis of nutrition, lifestyle and interaction with the environment for the prevention of this cancer (Cuaya-Simbro et al., 2017).

On the other hand, in the area of computer science, studies have been carried out in which data mining is applied, which allow the evaluation of different models of prevention and diagnosis of breast cancer; These works make use of databases with information on patients with breast cancer and show the effectiveness of the models to detect or determine if a patient is at risk of suffering from this type of cancer, such as the following works.

1. Breast cancer detection using advanced data mining techniques with neural networks [3].
2. Design and implementation of fuzzy classification algorithms for breast cancer diagnosis [4].
3. Data mining as support in the diagnosis and treatment of breast cancer [6]. Also, there are other works that suggest and implement a solution such as the development of an application that supports the early detection of breast cancer, which is based on the capture of temporary data and its subsequent analysis.
4. Smart mobile application for breast cancer prevention [1].

3 Description of the Technique

3.1 K-NN Algorithm

The algorithm is based on the comparison of an unknown example with the training examples k that are the closest neighbors of the unknown example [7].

Classify a set of examples from others that we already know such as their class and we will call them training set. Calculate the distance of each example to classify with all the examples of the training set and classify the example according to the class to which those closest examples belong to. The variable k is used to determine how many of your closest examples of the training set have to be taken into account to classify it [9].

The main problem of the k-NN algorithm is to find the value of k with which a higher performance is obtained, a technique known as cross validation is generally used.

To calculate the distance, you can use different methods, such as Euclidean distance, Manhattan, Canberra or maximum distance [8].

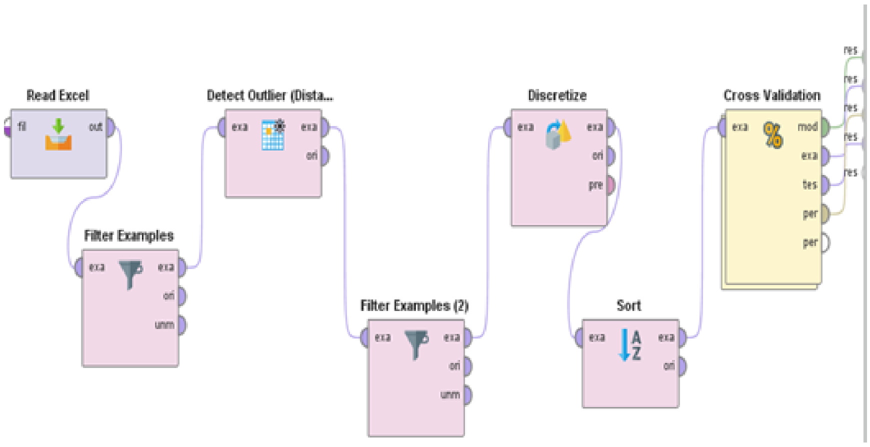


Fig. 1. Final configuration for the K-NN algorithm.

The advantages of this algorithm are the following:

1. **Not parametric**. It does not make explicit assumptions about the functional form of the data.
2. **Simple algorithm.** To explain, understand and interpret.
3. **High precision (relative).** High but not competitive against more supervised models.
4. **Insensitive to atypical values**. Accuracy may be affected by noise or irrelevant characteristics.

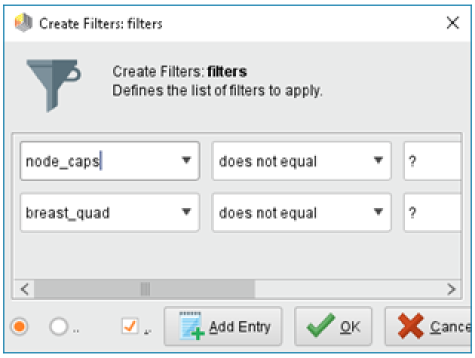
The disadvantages of this algorithm are:

1. **Instance based.** Do not explicitly learn a model. Memorize subsequent instances as knowledge for the prediction phase.
2. **Computationally expensive**. Store all training data.
3. **High memory requirement**. Store all (or almost all) training data[8].

The process that is implemented in Rapidminer to apply the algorithm is really very easy. The first step is to have a database, in this case there was a test database that was converted to Excel using the spoon tool. For this to be functional, an attribute such as "label" must be selected, which for this database is the "class" attribute. As shown in Figure 1, some operators were applied so that the accuracy range increased and better results were obtained.

When analyzing the data in the database, it was observed that, in certain attributes, there were missing data, so it was decided not to take them into account and eliminate them. For this, a filter was used as shown in Figure 2, where we can see that it is required that all data containing a “?” Be removed.

On the other hand, to avoid data that could affect the result by having extreme values, the operator that detects outliers was used. As shown in Figure 6, 10 neighbors and 2 outliers were chosen.

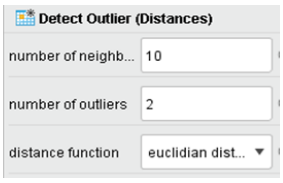


Create Filters: filters

Create Filters: **filters**
Defines the list of filters to apply.

node_caps | does not equal | ?
breast_quad | does not equal | ?

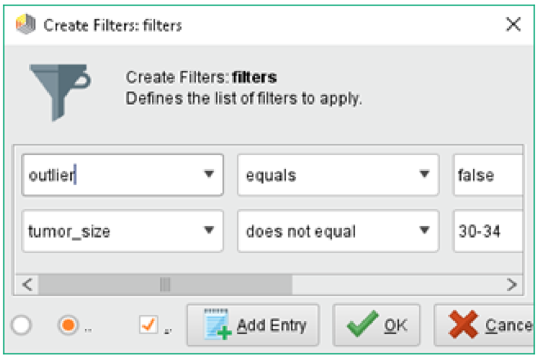
Fig. 2. Filter to eliminate missing data.



Detect Outlier (Distances)

number of neighb... 10
number of outliers 2
distance function euclidian dist...

Fig. 3. Filter for outliers.



Create Filters: filters

Create Filters: **filters**
Defines the list of filters to apply.

outlier | equals | false
tumor_size | does not equal | 30-34

Fig. 4. Filter for outliers.

The following filter, as shown in Figure 4, is used to filter outliers detected in the previous operation, and also filters data whose tumor size is not equal to 30-34.

The next step is discretization, and this is through two bins, the result of this is increasingly ordered by the deg_malig attribute.

Finally, in the cross validation the KNN algorithm is added, which is connected with an "Apply model", which in turn must be connected with a "Performance", as is shown in Figure 5.

3.2 Decision Tree

It is a type of supervised learning algorithm (with a predefined target variable), which is used in classification problems. It works for input and output variables both

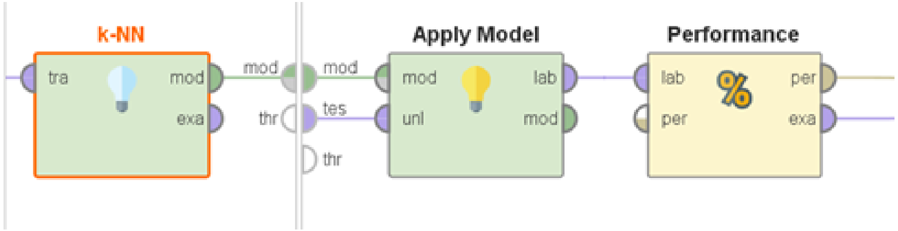


Fig. 5. Final configuration for the k-NN (Cross Validation).

categorical and continuous. They learn and train from given examples and predict for unseen circumstances [12].

Each node represents a division rule for a specific attribute. These can be expressed in a "Yes ... then ..." clause. Each data value or decision forms a clause, such that, for example, "if conditions 1, 2 and 3 are met, then result X will be the definitive result with certainty Y" [12].

The construction of new nodes is repeated until the detention criteria are met. A prediction for the class tag attribute is determined depending on most of the examples that reached this sheet during generation. Then, the tree model can be applied to the new examples using the Apply model operator.

Employing decision trees in machine learning has many advantages:

1. The cost of using the tree to predict the data decreases with each additional data point.
2. It works for numerical or categorical data.
3. The results are easy to explain.
4. Quantifiable reliability and can be tested.

But it also has some disadvantages:

1. When categorical data with multiple levels are presented, the results are in favor of attributes with more levels.
2. Calculations can become complex with numerous related results.
3. Conjunctions between nodes are limited to AND, while graphics support related nodes using OR.

For the application of this algorithm exactly the same database was used, and the final configuration can be seen in Figure 5. Because it is the same database, the missing data should be omitted, so a filter equal to the one shown in Figure 6 was added.

Similarly, it was decided to add a filter for data whose size of tumors were different from 30-34. The order was by this same attribute in decreasing form.

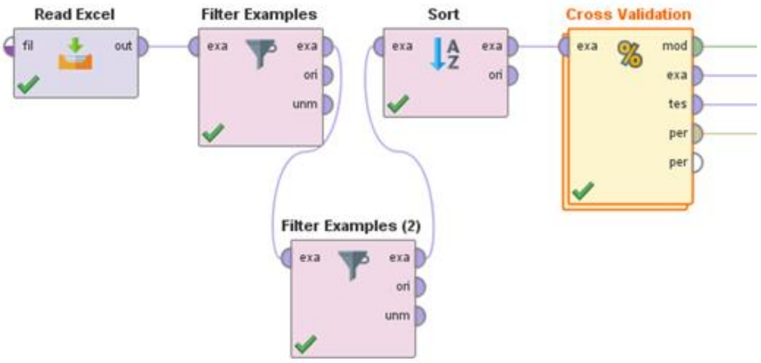


Fig. 6. Final configuration for the decision tree.

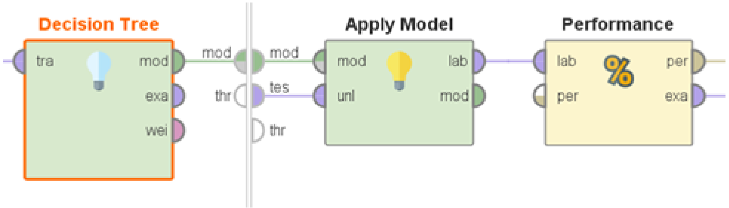


Fig. 7. Cross Validation for decision tree.

Finally, in the "Cross Validation" you must add the "Decision Tree", and for it to work you must apply a model and a performance, as is shown in Figure 7.

4 Problem Statement

4.1 Naive Bayes Classifier

Algorithm based on probabilities conditioned with known data. Its operation is based on calculating probabilities of known data and according to the results and a formula, it can calculate the probability that the entry is of one kind or another. It is based on Bayes' Theorem or conditional probability theorem. The probability that an event will occur having happened another that influences the previous one, is defined with the following formula:

$$P(C|W) = \frac{P(W|C)P(C)}{P(W)}\,. \qquad [10]$$

Finally, the Naive Bayes method classifies document D into one of all existing classes using the formula:

$$Mejor\ clase = \boldsymbol{argmax}_{\mathrm{cj\epsilon C}}\, P\,(Ci) \prod_{i=1}^{''} P(Wi\,|\,Ci)\,. \qquad [11]$$

Advantages of this method are:

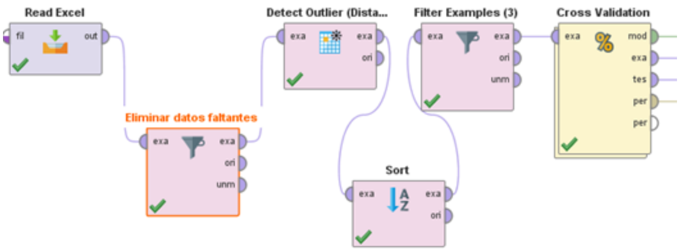
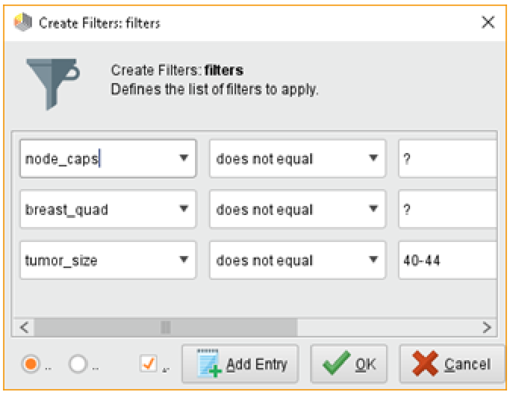


Fig. 8. Final configuration of the Naive Bayes algorithm.


Create Filters: filters

Create Filters: **filters**
Defines the list of filters to apply.

node_caps	does not equal	?
breast_quad	does not equal	?
tumor_size	does not equal	40-44

Fig. 9. Filter for missing data and tumor size.

1. Fast and simple to use.
2. Quick to train Quick to sort.
3. It is not sensitive to minor characteristics. Handles discrete and subjective information.
4. No problem handling continuous data streams.

Disadvantages are:

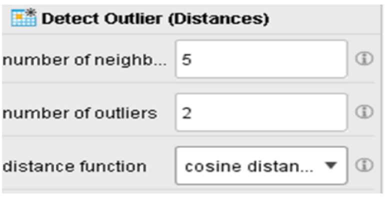
1. Not very accurate.
2. It assumes an independence of the characteristics [13].

The application of the Bayesian classifier is very similar to that of the other two. And the configuration is shown below in Figure 8.

The first step, as mentioned in the previous ones, is to filter the missing data, in addition there is added one more filter that is for the size of the tumor between 40-44 as seen in Figure 9.

On the other hand, it was decided to detect the outliers, and the number of neighbors chosen was 5 to have more proximity and the number of outliers was 2.

This time it was decided that the distance would be calculated with the "cosine distance" as shown in the following Figure 10.



Detect Outlier (Distances)

number of neighb... 5

number of outliers 2

distance function cosine distan...

Fig. 10. Outliers Detection.

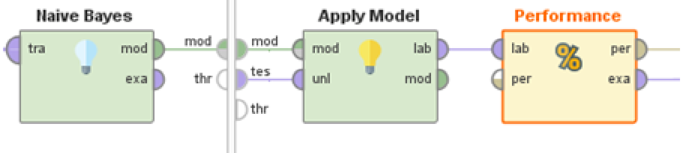


Fig. 11. Naive Bayes (Cross Validation).

Table 1. Data attributes in the database.

Data	
– Class	– Quadruple chest
– Age	– Node caps
– Menopause	– Deg – malig
– Tumor size	– Mama
– Nodes	– Irradiat

The sorting was given by the node_caps tribute in decreasing manner. The last filter was used for outliers.

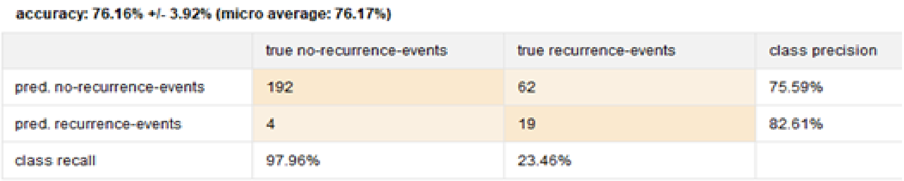
As in the previous algorithms, in the cross validation the algorithm must be put, which for this case is the “Naive Bayes”, applying a model and a performance for its correct operation, as is shown in Figure 11.

5 Analysis of Results

The data obtained as of July 11, 1988 have a number of 286 instances, each with 9 attributes and one attribute class, indicated below (Table 1).

Relevant information about the data set corresponds to 201 instances of the class: no recurrence and 85 instances of the class: recurrence, mentioning a range of precision of 4 systems tested with a result of 68% -73.5%. Subsequently, this data set was prepared to be processed and analyzed, submitting to the K-NN algorithms, decision tree and Bayesian classifier, using within the process operators to have the lowest number of data loss, with the main objective of obtaining the highest possible accuracy in the results of each.

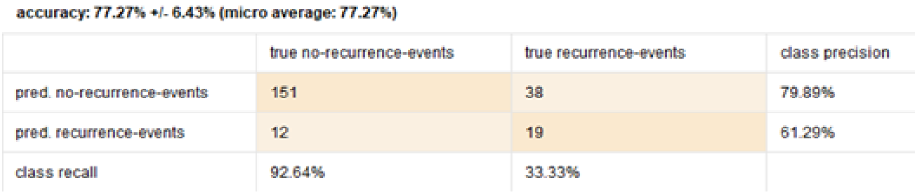
The results obtained are presented below, in each process of the set of training instances.



accuracy: 76.16% +/- 3.92% (micro average: 76.17%)

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	192	62	75.59%
pred. recurrence-events	4	19	82.61%
class recall	97.96%	23.46%	

Fig. 12. Precision of the K-NN algorithm.



accuracy: 77.27% +/- 6.43% (micro average: 77.27%)

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	151	38	79.89%
pred. recurrence-events	12	19	61.29%
class recall	92.64%	33.33%	

Fig. 13. Precision of the Decision tree.

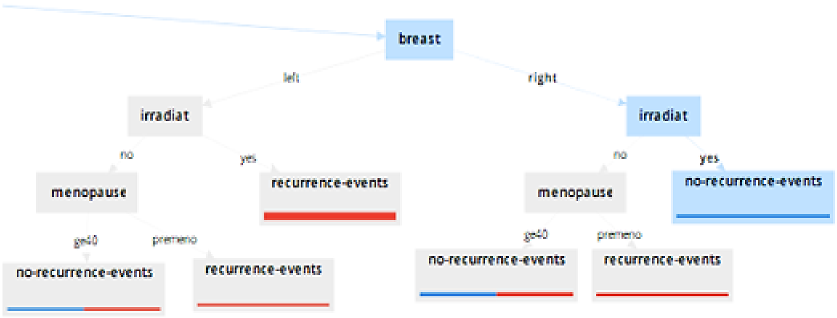


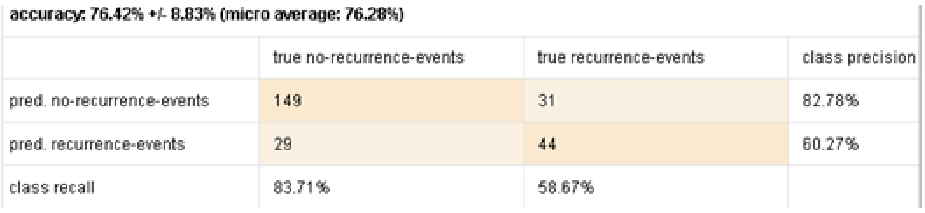
Fig. 14. Result of the Decision tree: Chest.

5.1 K-NN Algorithm

Accuracy result of 76.16%, with an amount of 192 data correctly classified in the "no-recurrence-events" class and 62 data not correctly classified. On the other hand, 19 correct data were classified in the "recurrence-events" class and only 4 were not classified well, as is shown in Figure 12. Exceeding the estimated accuracy of 2.66%, without the greatest loss of data.

5.2 Decision Tree

Accuracy result of 77.27%, obtaining 151 data classified correctly for the first class and 38 misplaced. And in the second class a total of 19 well-classified data and a total of 12 poorly classified, as is shown in Figure 13. Exceeding the estimated accuracy of 3.77%, without the greatest loss of data.



accuracy: 76.42% +/- 8.83% (micro average: 76.28%)

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	149	31	82.78%
pred. recurrence-events	29	44	60.27%
class recall	83.71%	58.67%	

Fig.15. Precision of the Naive Bayes classifier.

The result of this algorithm obtains a graph in which a tree is represented, detailing each of the attributes of the instances, due to being extensive a brief example is shown identifying the attributes, chest, tumor size and quadruple chest. as is shown in Figure 14.

5.3 Naive Bayes Classifier

Distribution model for the tag attribute class:

1. Class without recurring events (0.704).
2. 9 distributions.
3. Class with recurrence events (0.296).
4. 9 distributions.

Accuracy result of 76.42% with 149 data classified correctly for the first class and only 31 data classified in a wrong place. In the second class, 44 data classified in the correct class and 29 in the incorrect class were obtained. The percentage rose 0.02% when changing Euclidean distance to Cosine, so this solution was chosen as the best, as it is shown in Figure 15. Exceeding the estimated accuracy of 2.92%, without the greatest loss of data.

6 Conclusions and Future Research

The development of this research obtained favorable results, by using software and its operators to deal with issues pertaining to data mining, implementing the knowledge acquired throughout the four-month period.

In addition, to understanding and acquiring the knowledge of data analysis, the importance of lost values within the training set was understood, since the loss of data or values have a significant impact on the results obtained, as well as the operators used influence, you should look for the most favorable accuracy percentage, stating that the data is treated properly and implement a smaller number of data lost in the training process.

In the future, learning about tools for data mining, such as those implemented for the development of this work, will allow for the ability to process a large number of instances and submit the set of training instances to the different algorithms they offer, to analyze issues as important as breast cancer, as well as focusing this analysis on the possibility of analyzing relevant information of all kinds, within a wide number of examples, such as medicine, artificial intelligence, genetics, terrorism, as well as science and technology, etc.

Achieving such an in-depth analysis, will allows us in the future, to analyze patterns and obtain a high percentage of prediction in the events of such topics so relevant at present, not ruling out the possibility that, over the years, the algorithms and their set of training data can give solution to outstanding events in society within each of its areas.

References

1. Cuaya-Simbro, G., Ruíz-Hernández, E., Hernández-Hernández, L.Á., Lima-Luna, L.A.: Aplicación móvil inteligente para prevención de cáncer de mama. Revista de Ingeniería Tecnológica (2017)
2. WHO (World Health Organization). Obtenido de: Cáncer de mama: prevención y control (2016)
3. Ortiz-Murillo, J.A., Celaya-Padilla, J.M., López-Hernández, Y.: Detección de cáncer de mama usando técnicas avanzadas. In: Proccedings of the ISSSD (2016)
4. Secretaria de Salud, Gobierno de México. Obtenido de Programa de Acción Específico Prevención y Control del Cáncer de la Mujer 2013-2018 (2015)
5. De la Puente-Marrugo, Y.M., Milena-Rizzo D.R.: Diseño e implementación de algoritmos de clasificación borrosa para diagnóstico de cáncer de mama. Corporación Universitaria Rafael Nuñez (2006)
6. López-Portillo, C.C.: Minería de datos como soporte en el diagnóstico y tratamiento del cáncer de mama. Centro de Investigación Científica y de Educación Superior de Ensenada (2006)
7. Berástegui-Arbeloa, G.: Implementación del algoritmo de los k vecinos más cercanos y estimación del mejor valor local para su cálculo. Pamplona (2018)
8. González, L.: Obtenido de aprendizaje supervisado: K-Nearest Neighbors: http://ligdigonzalez.com/aprendizaje-supervisado-k-nearest-neighbors/ (2018)
9. Rapidminer: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/lazy/k_nn.html (2006)
10. Blogspot: Algoritmos de Mineria de datos: http://algoritmosmineriadatos.blogspot.com/2009/12/algoritmo-naive-bayes.html (2009)
11. Lucidchart: Que es un diagrama de árbol de decisión. https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision (2006)
12. Sloth's Lab: http://www.slothslab.com/python/2015/12/03/clasificador-bayesiano -ingenuo-python.html (2015)