Polytechnic University of the State of Morelos

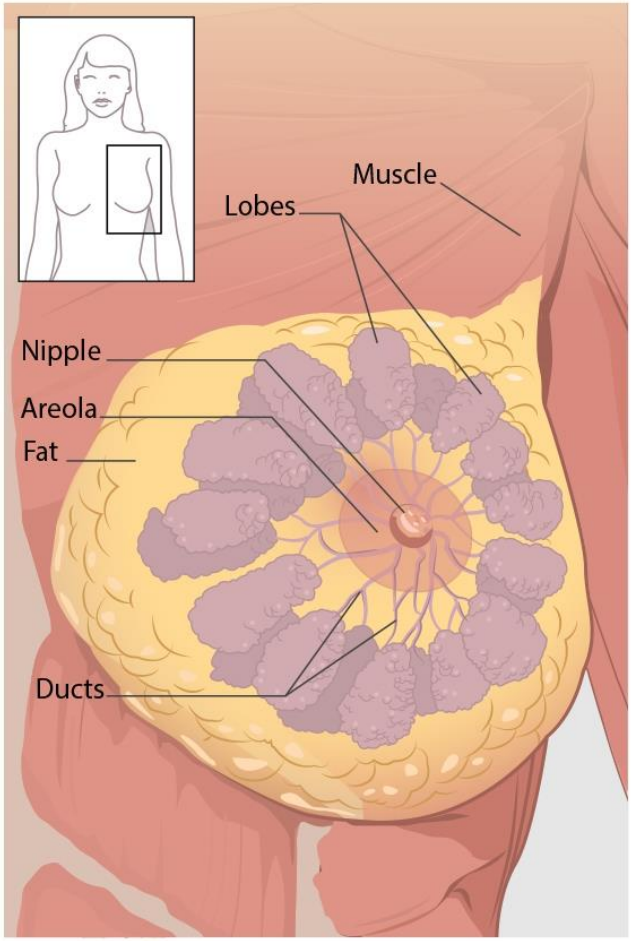
Upemor
Universidad Politécnica

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

October, 2019

What Is Breast Cancer?

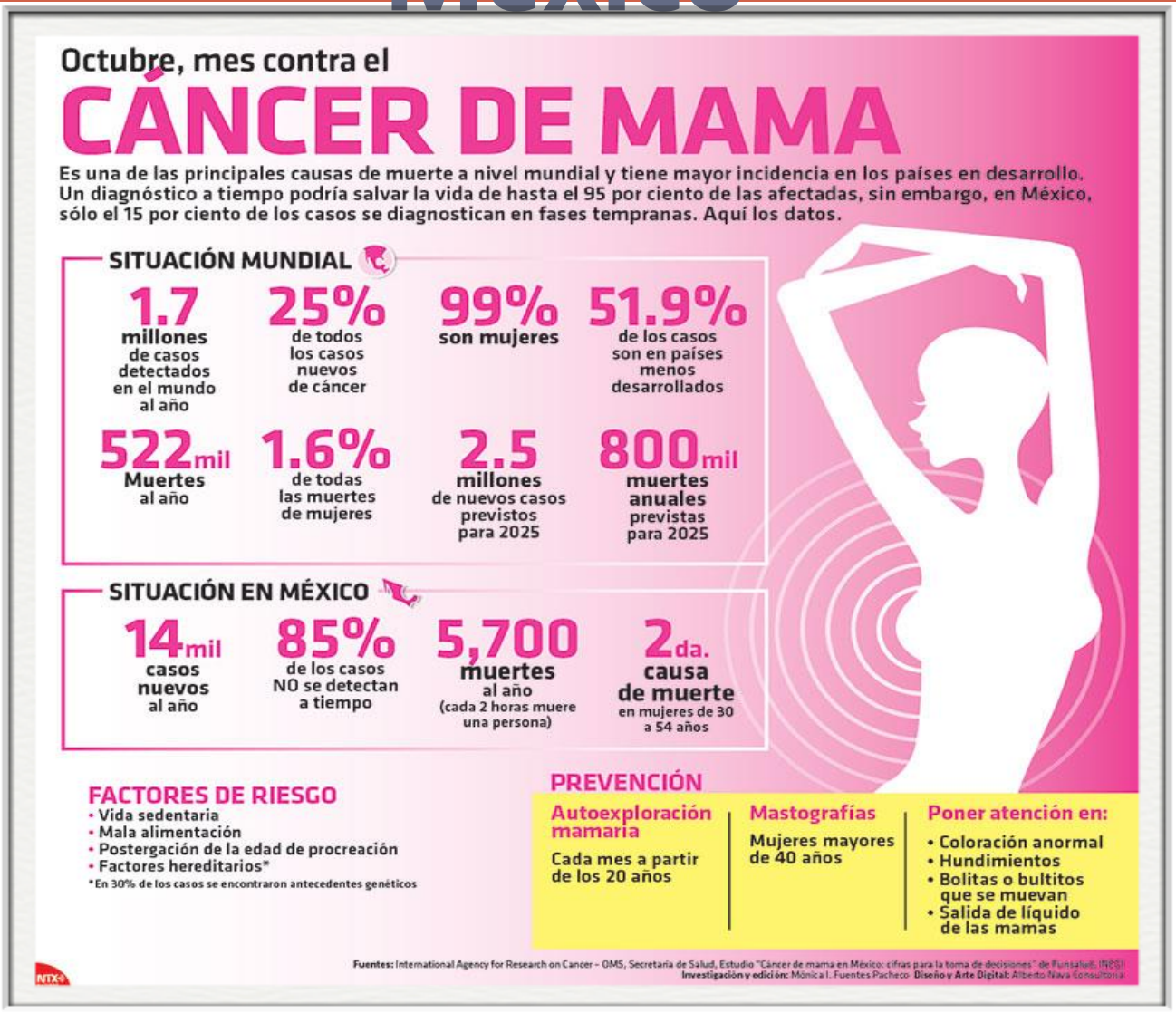
Breast cancer is a disease that affects a large part of world society. It is the most diagnosed cancer in women and its prevention although it is not impossible, it seems really difficult since its cause is unknown, so early detection is based on the patient's forecast. The most common form for detection is self-exploration, however this is only detected in more advanced



Cancer Trends in Mexico

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, Sonora and Nuevo León

Cancer Trends in Mexico



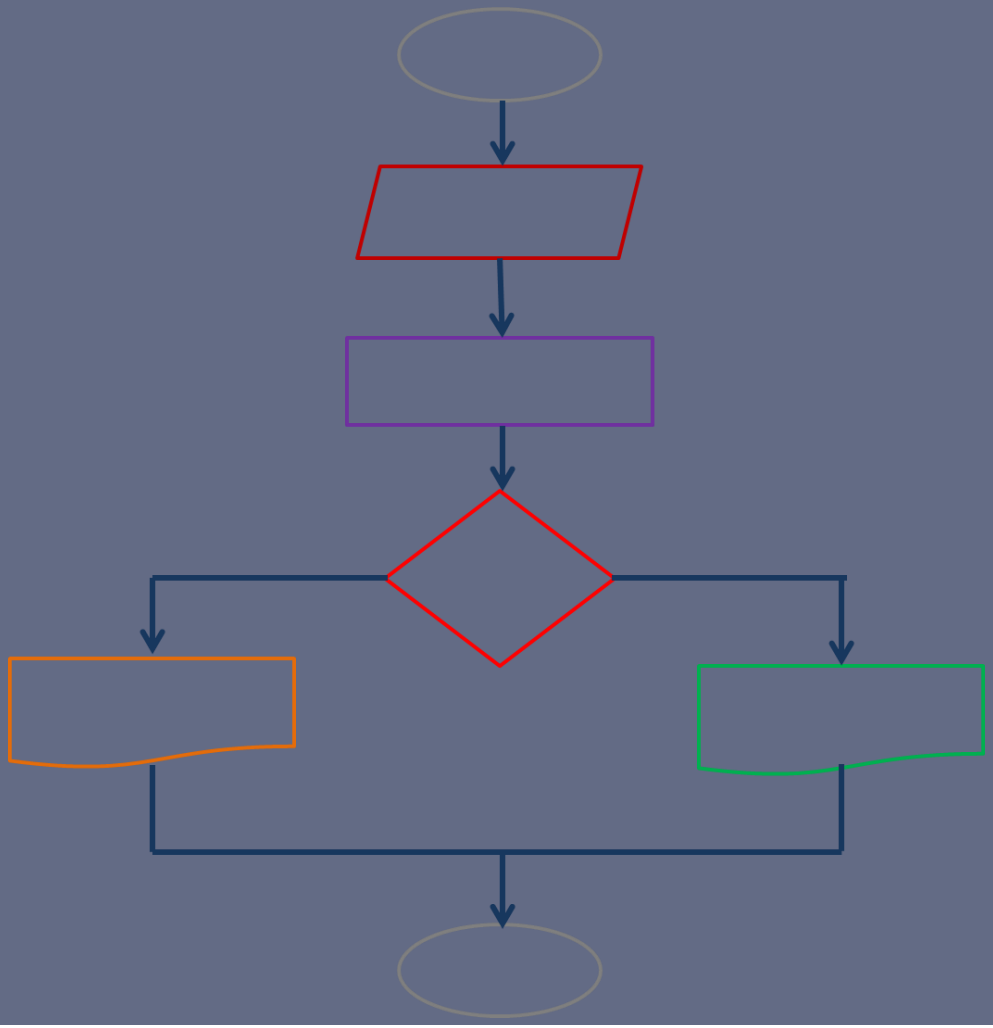
Tools Used

Data mining for the detection of breast cancer

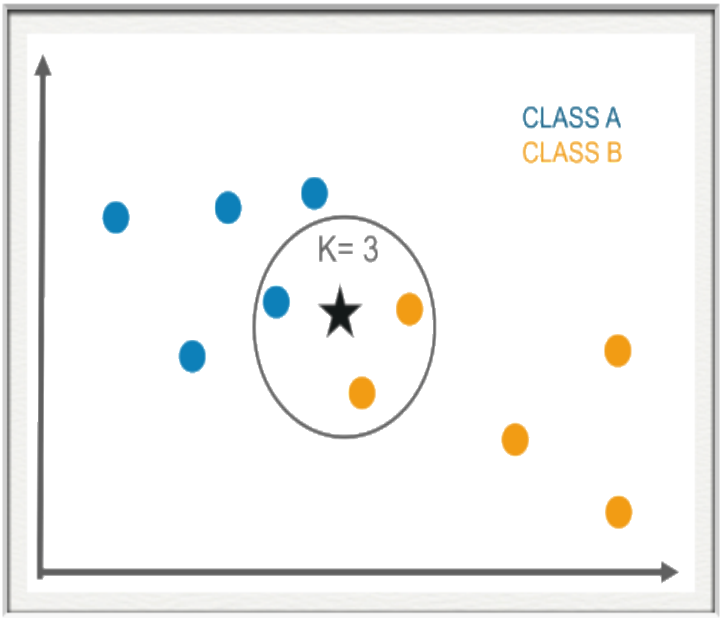
Data mining allow the evaluation of different models about prevention and diagnosis of breast cancer.

We use DB with patient information to prevent or detect di

Algorithms Used



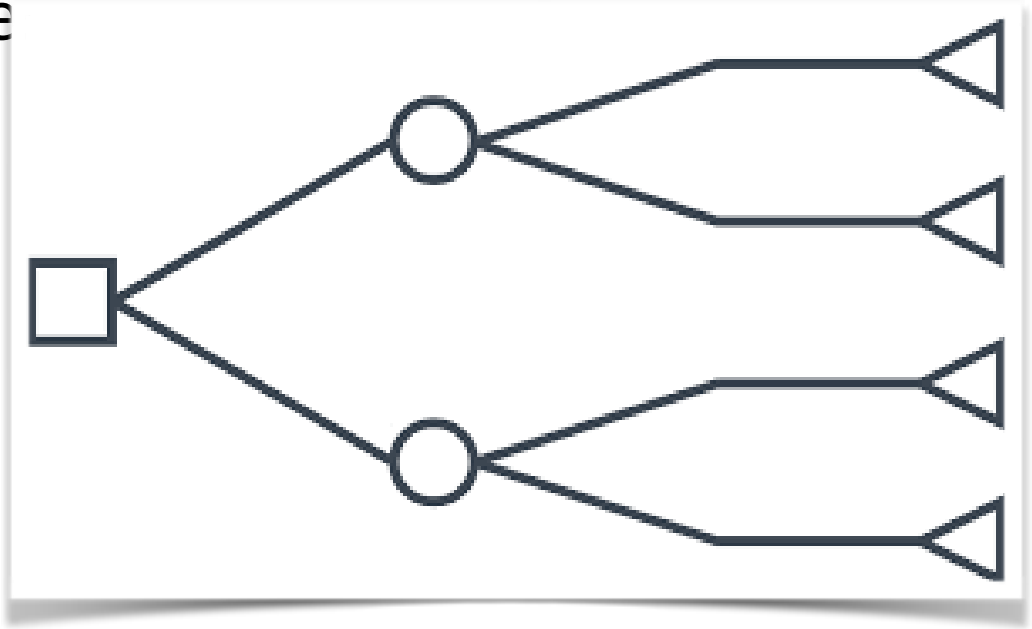
K-NN Algorithm



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.

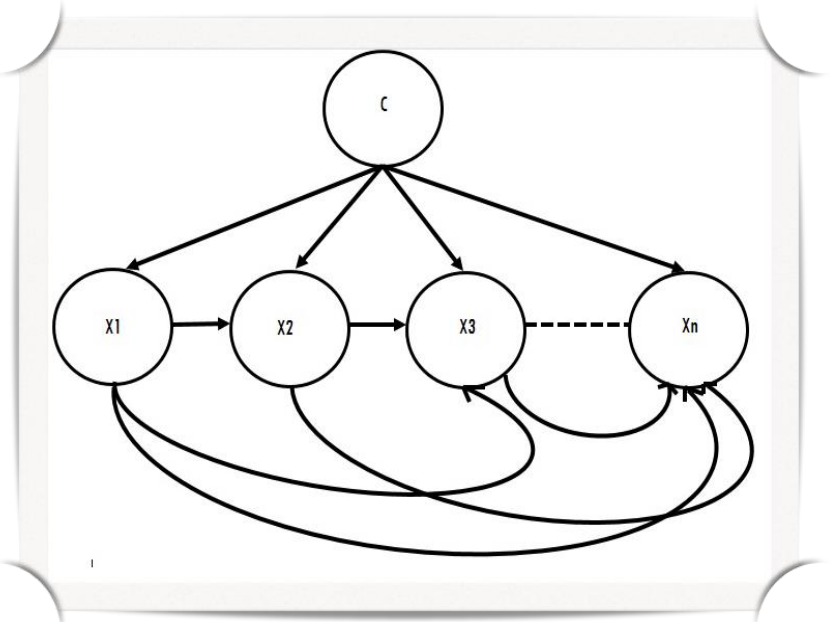
Decision Tree

It is a type of supervised learning algorithm (with a predefined target variable) that is used in classification problems and it works for input and output variables both categorical and continuous. They learn and train from given examples and predict for unseen circumstance



Naive Bayes Classifier

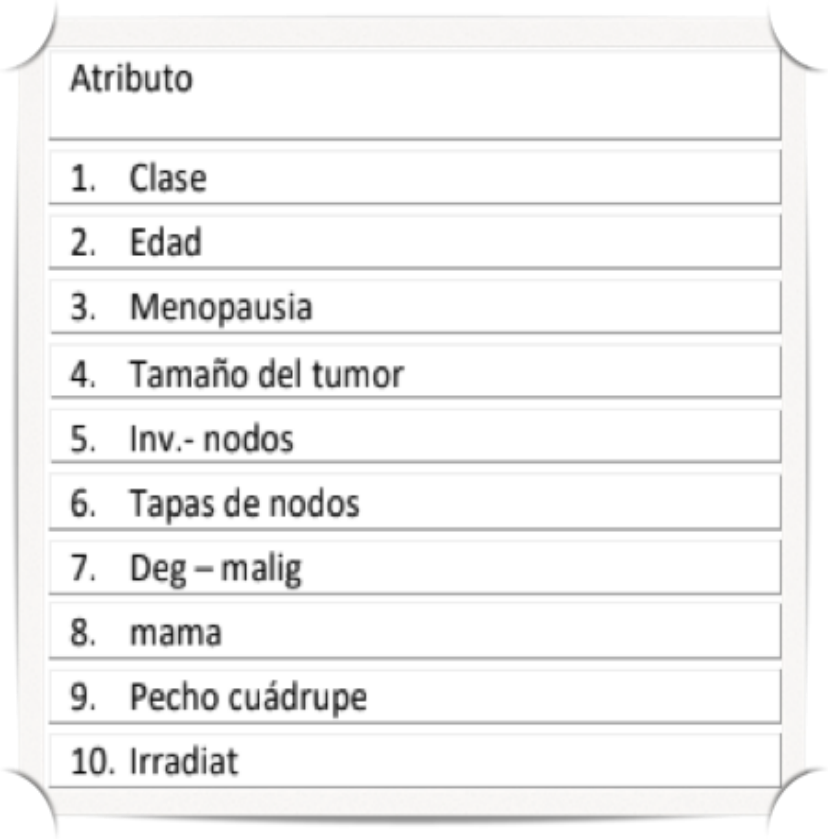
Algorithm based on probabilities conditioned with known data. Its operation is based on calculating probabilities of known data and according to the results and a formula, it can calculate the probability that the entry is of one kind or another. It is based on Bayes' Theorem or conditional probability



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Analysis of results

- 201 instances of the class: no recurrence
- 85 instances of the class: recurrence,
- Mentioning a range of precision of four systems tested with a result of 68% -73.5%.
- Number of Instances: 286



K-NN: Obtained results

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%	

Decision Tree: Obtained results

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%	

Naive Bayes: Obtained results

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%	

Conclusions

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.

References

Salud, S. d. (8 de Septiembre de 2015). Gobierno de México. Obtenido de Programa de Acción Específico Prevención y Control del Cáncer de la Mujer 2013 - 2018: https://www.gob.mx/salud/acciones-y-programas/informacion- estadística

Berástegui Arbeloa, G. (2018). Implementación del algoritmo de los k vecinos más cercanos y estimación del mejor valor local para su cálculo. Pamplona

Gabits. (2 de Diciembre de 2009). Blogspot. Obtenido de Algoritmos de mineria de datos: http://algoritmosmineriadatos.blogspot.com/2009/12/algoritmo-naive-bayes.html

Sloth's Lab. (3 de Diciembre de 2015). Obtenido de http://www.slothslab.com/python/2015/12/03/clasificador- bayesiano-ingenuo-python.html

THANK YOU