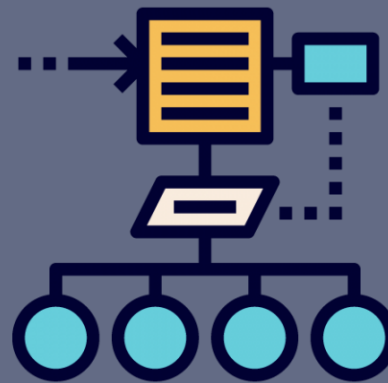


# Polytechnic University of the State of Morelos



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



October, 2019

Eng. Zagal Solano José Enrique

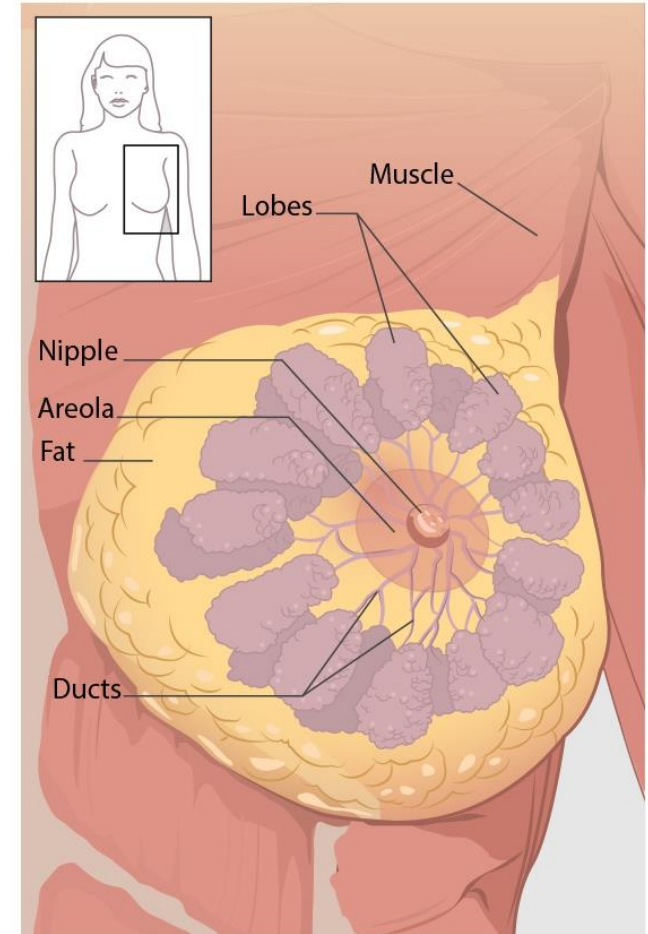
# Contents

- ❑ Introduction
- ❑ Related Works
- ❑ Description of the technique
- ❑ Problem Statement & Analysis of results
- ❑ Conclusions



# What Is Breast Cancer?

- Breast cancer is a disease that affects a large part of world society.
- Is disease that mainly affects women.
- It is estimated per year 552,000 women died.
- The most common form for detection is self-exploration, however this is only detected in more advanced stages.



# 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

Octubre, mes contra el

## CÁNCER DE MAMA

Es una de las principales causas de muerte a nivel mundial y tiene mayor incidencia en los países en desarrollo. Un diagnóstico a tiempo podría salvar la vida de hasta el 95 por ciento de las afectadas, sin embargo, en México, sólo el 15 por ciento de los casos se diagnostican en fases tempranas. Aquí los datos.

### SITUACIÓN MUNDIAL



### SITUACIÓN EN MÉXICO



### FACTORES DE RIESGO

- Vida sedentaria
- Mala alimentación
- Postergación de la edad de procreación
- Factores hereditarios\*

\* En 30% de los casos se encontraron antecedentes genéticos

### PREVENCIÓN

#### Autoexploración mamaria

Cada mes a partir  
de los 20 años

#### Mastografías

Mujeres mayores  
de 40 años

#### Poner atención en:

- Coloración anormal
- Hundimientos
- Bolitas o bultitos  
que se muevan
- Salida de líquido  
de las mamas

# Tools Used



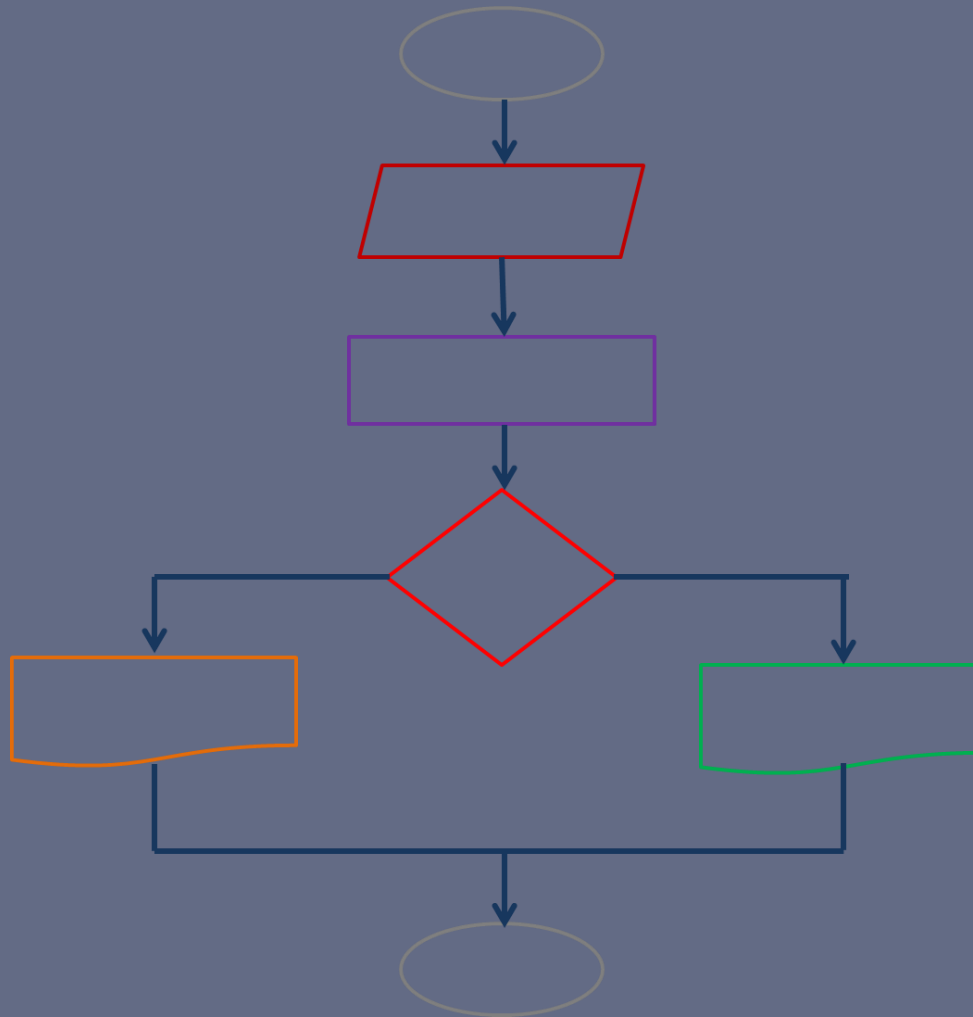
# Data mining for the detection of breast cancer

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 disease. (Institute of Oncology University Medical Center, 11 July 1988)

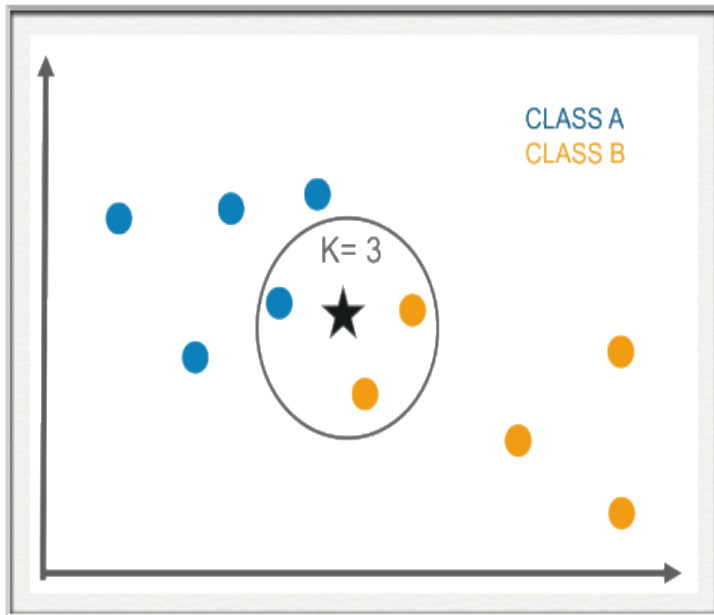


# Algorithms Used





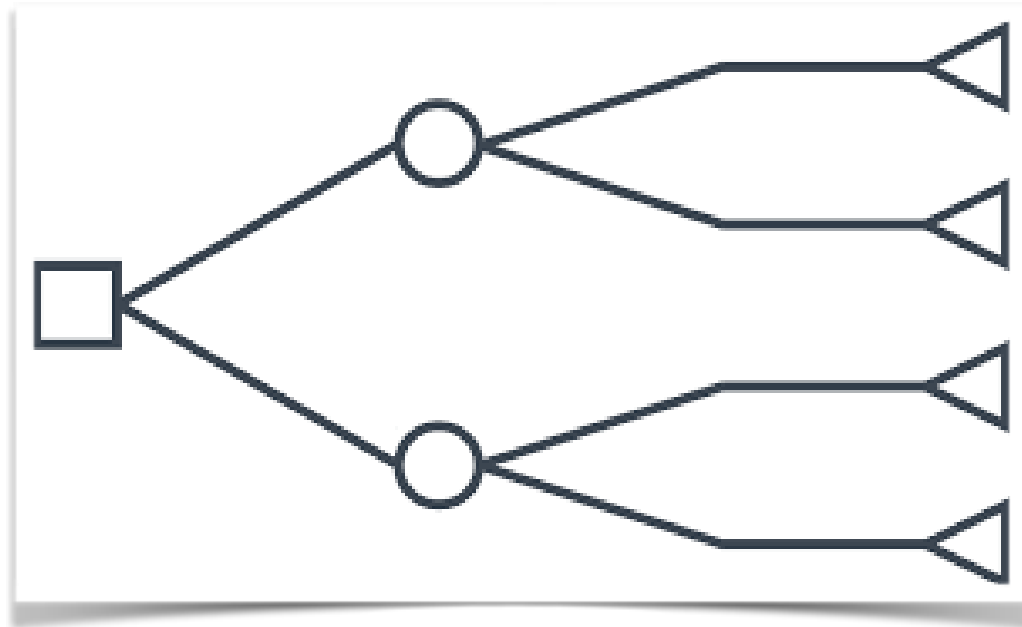
# K-NN Algorithm



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.

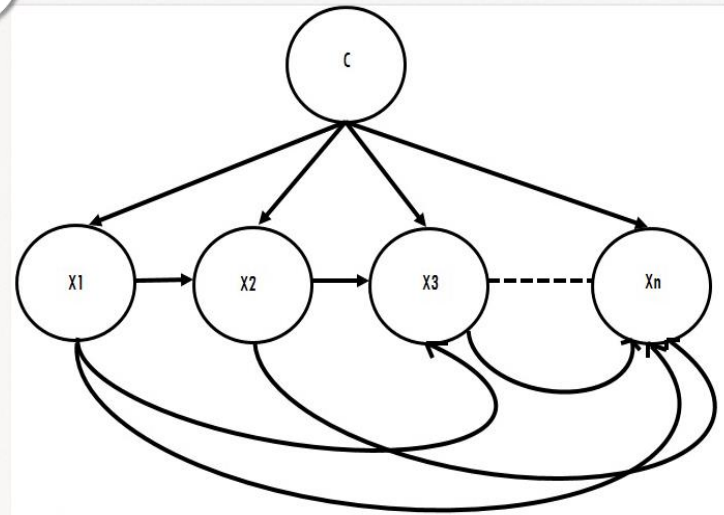
# Decision Tree

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 circumstances.



# 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.



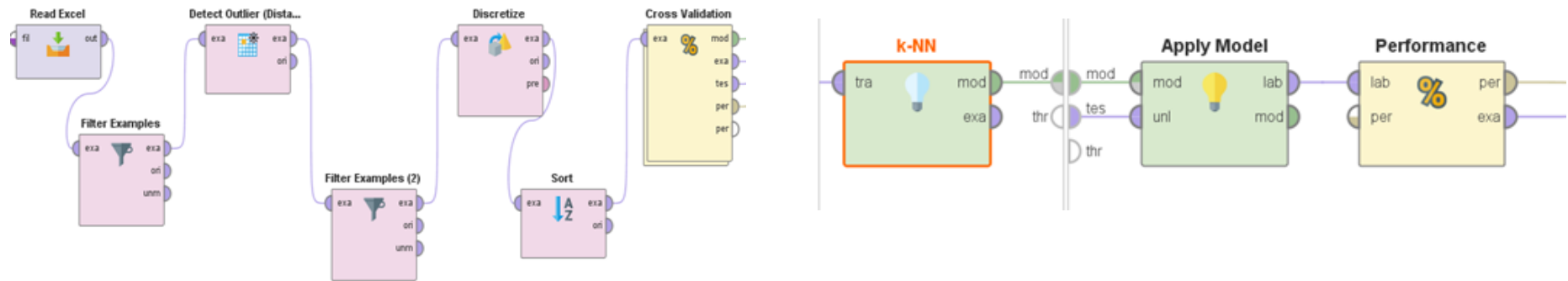
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# 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

| Attribute          |
|--------------------|
| 1. Class           |
| 2. Age             |
| 3. Menopause       |
| 4. Tumor size      |
| 5. Nodes           |
| 6. Node caps       |
| 7. Deg - malig     |
| 8. Mama            |
| 9. Quadruple chest |
| 10. Irradiat       |

# K-NN: Obtained results



Create Filters: filters

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

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

< Add Entry OK Cancel

Detect Outlier (Distances)

number of neighb... 10

number of outliers 2

distance function euclidian dist...

Create Filters: filters

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

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

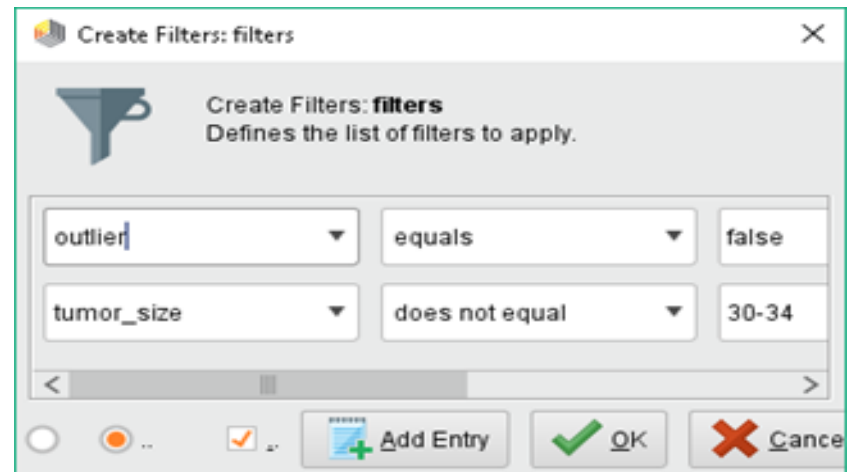
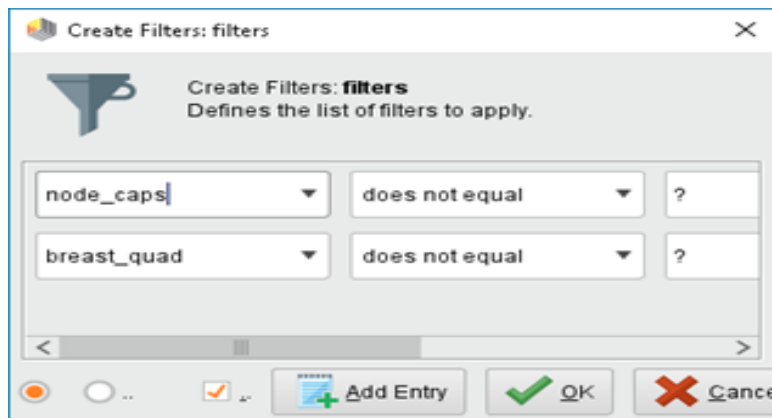
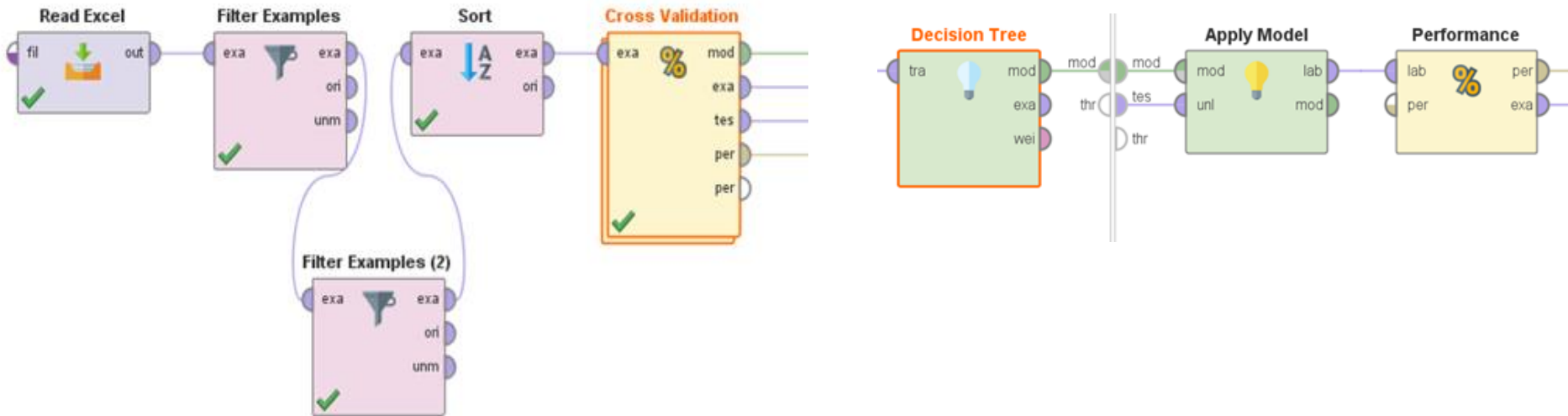
< Add Entry OK Cancel

# 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



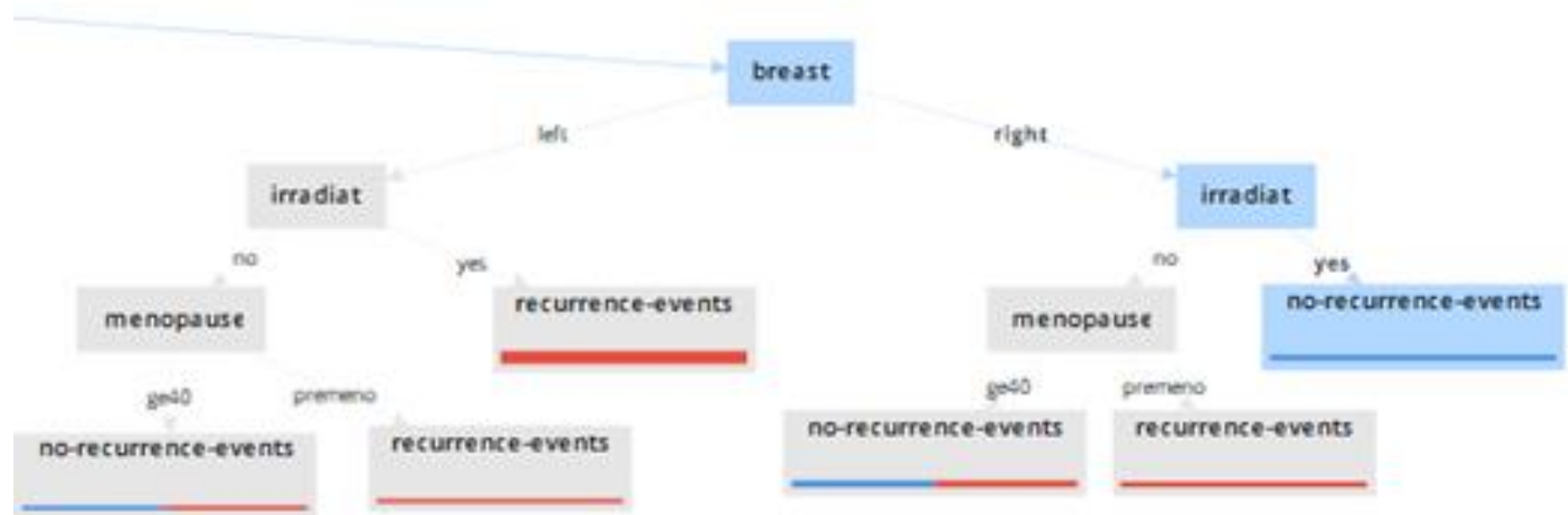
# Decision Tree: Obtained results

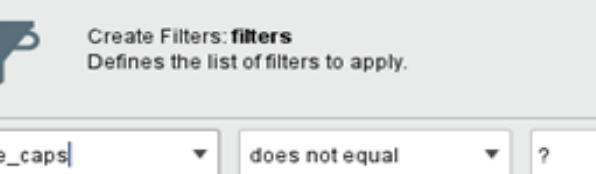
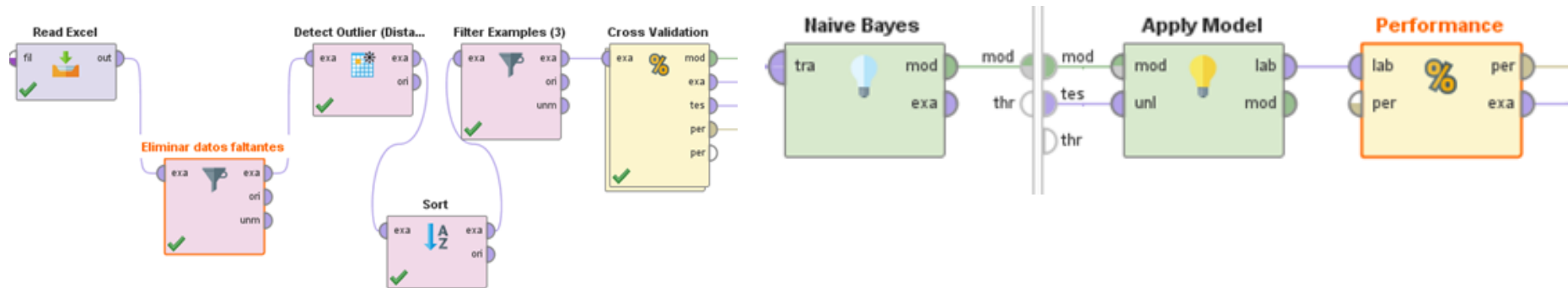
**accuracy: 77.27% +/- 6.43% (micro average: 77.27%)**

|                            | 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%                 |                 |



# Decision Tree: Obtained results









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 |

☐ ... ☐ ... ☒  Add Entry  OK  Cancel

 **Detect Outlier (Distances)**

number of neighbors  ⓘ

number of outliers  ⓘ

distance function  ⓘ

# 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.
- ❑ Stating that the data is treated properly and implement a smaller number of data lost in the training process.
- ❑ 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

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 minería 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

