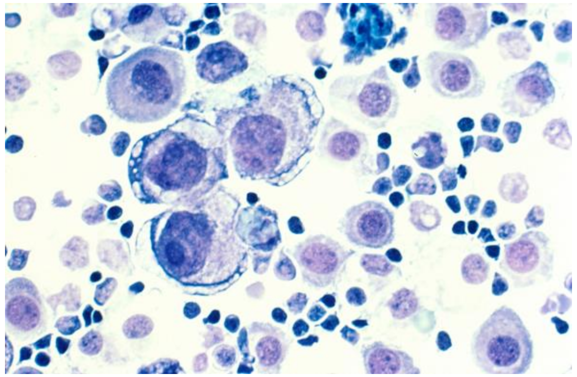
Individual Project: Breast Cancer



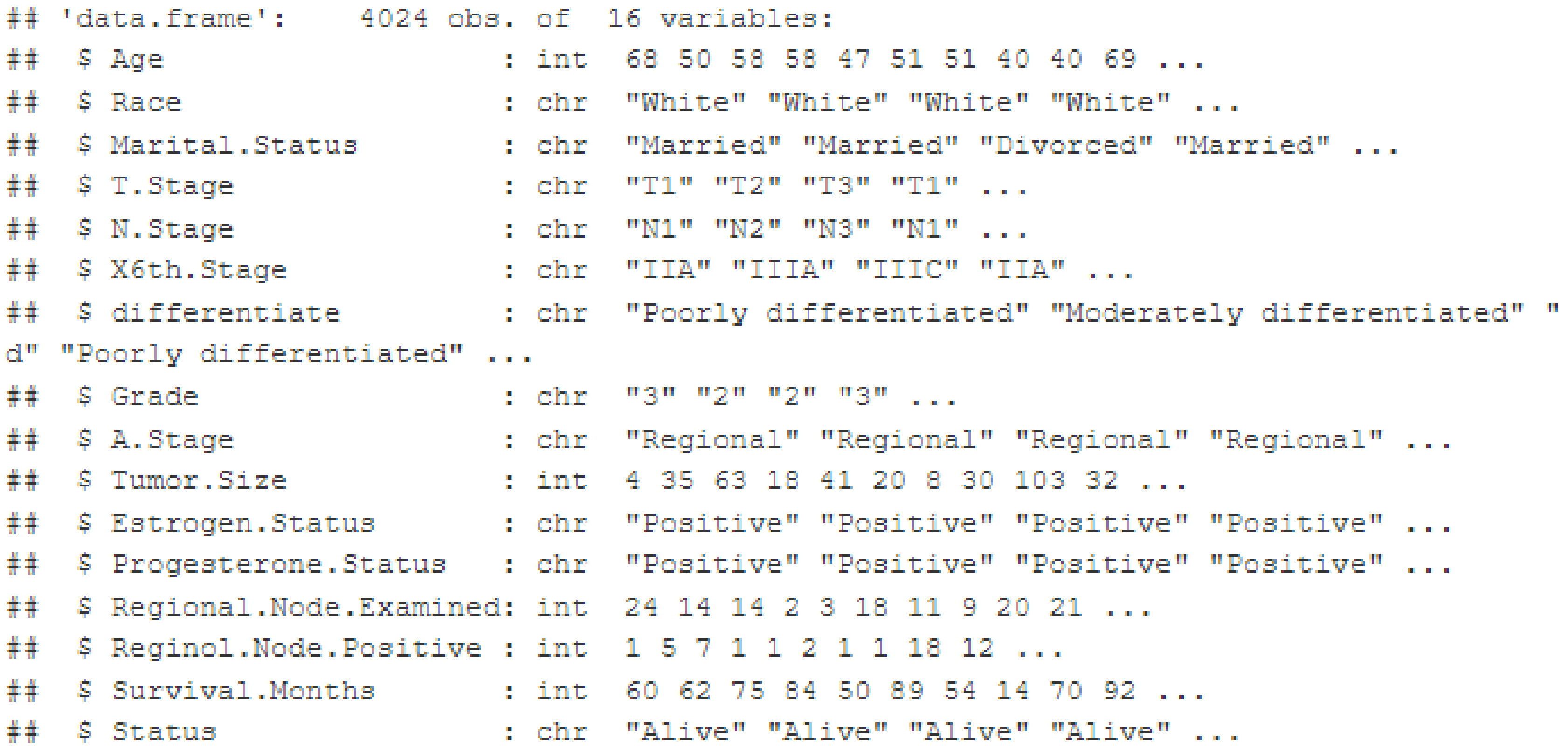
Sarai Ramirez

STA5206

Question:

- Could different measurements related to breast cancer and personal information predict the death of the patient?

Dataset:

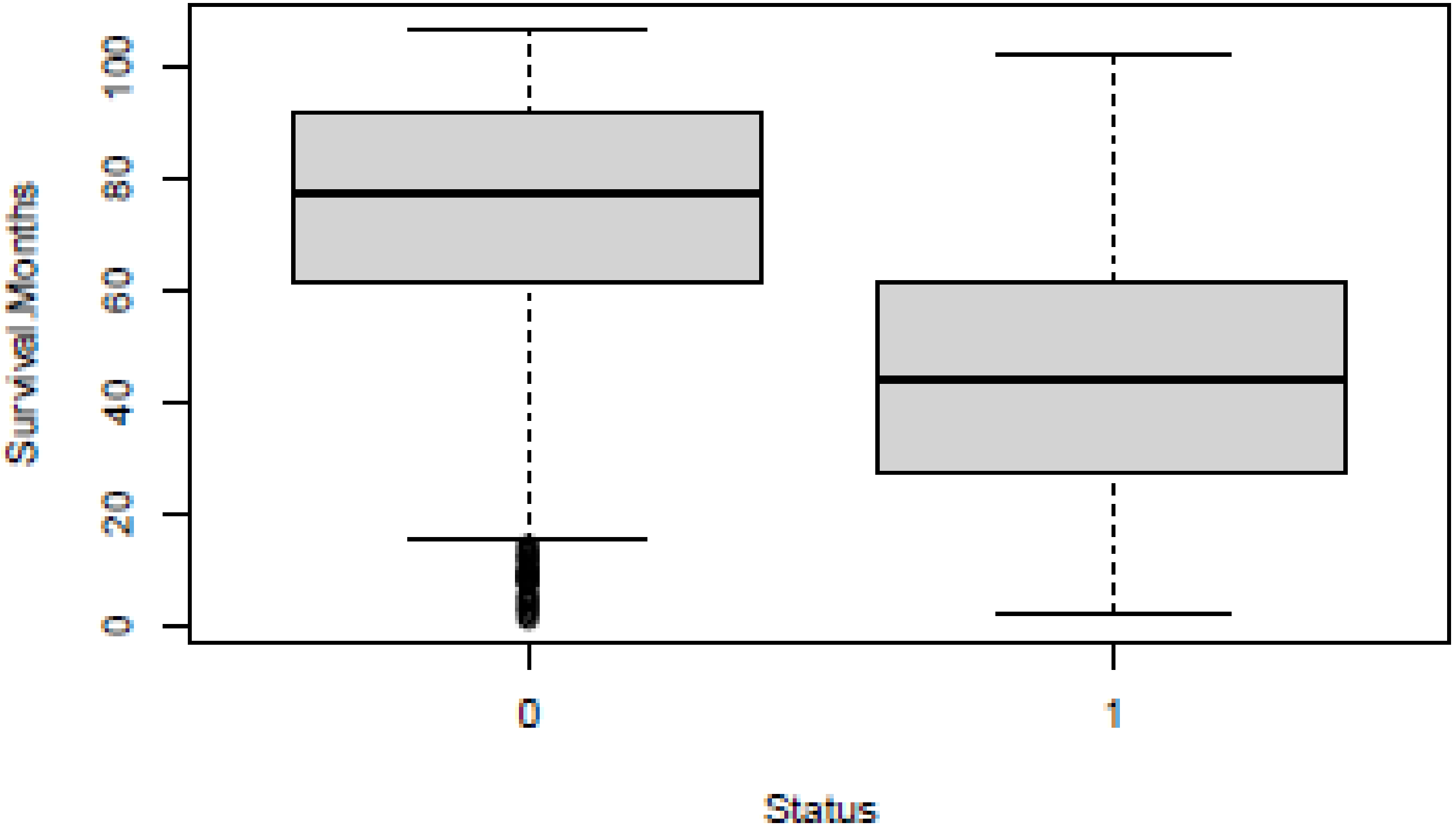


```
## 'data.frame':    4024 obs. of  16 variables:
##  $ Age                   : int  68 50 58 58 47 51 51 40 40 69 ...
##  $ Race                  : chr  "White" "White" "White" "White" ...
##  $ Marital.Status        : chr  "Married" "Married" "Divorced" "Married" ...
##  $ T.Stage               : chr  "T1" "T2" "T3" "T1" ...
##  $ N.Stage               : chr  "N1" "N2" "N3" "N1" ...
##  $ X6th.Stage            : chr  "IIA" "IIIA" "IIIC" "IIA" ...
##  $ differentiate         : chr  "Poorly differentiated" "Moderately differentiated" "
d" "Poorly differentiated" ...
##  $ Grade                 : chr  "3" "2" "2" "3" ...
##  $ A.Stage               : chr  "Regional" "Regional" "Regional" "Regional" ...
##  $ Tumor.Size            : int  4 35 63 18 41 20 8 30 103 32 ...
##  $ Estrogen.Status       : chr  "Positive" "Positive" "Positive" "Positive" ...
##  $ Progesterone.Status   : chr  "Positive" "Positive" "Positive" "Positive" ...
##  $ Regional.Node.Examined: int  24 14 14 2 3 18 11 9 20 21 ...
##  $ Reginol.Node.Positive : int  1 5 7 1 1 2 1 1 18 12 ...
##  $ Survival.Months       : int  60 62 75 84 50 89 54 14 70 92 ...
##  $ Status                : chr  "Alive" "Alive" "Alive" "Alive" ...
```

-This variables indicate tumor size, lymph nodes affected, characteristics of cancer cells, metastasis, personal information, Life expectancy and Status for each patient.

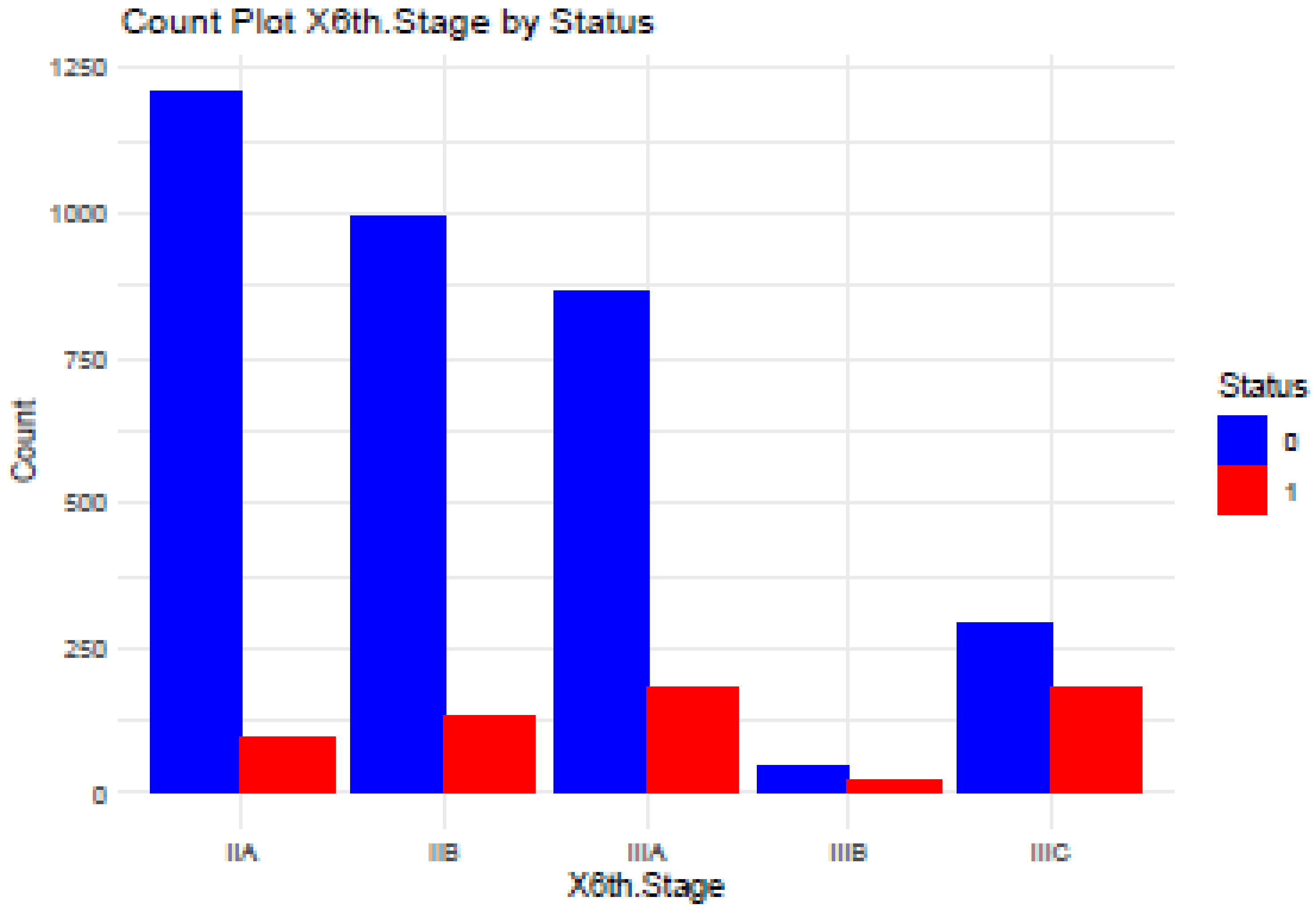
- T-stage, N-stage, X6th.Stage and Differentiated columns are ordinal categorical variables and were changed to numeric. Estrogen, Progesterone, A.Stage and Status were changed to binary.

-I decided to predict Status.



- Survival months variable was eliminated because if I want to predict the death event, is not possible to know that information.

- Grade variable was eliminated because it was the same as differentiated variable.



The X6th Stage discloses information about the variables: T-Stage, N-Stage and A-Grade at once

Methodology

Statistical models used:

- To predict the State variable I used Logistic Regression, Random Forest and XGBoost, because state is a binary variable, and I wanted to compare the accuracy of different models.

- PCA was used for dimensionality reduction.

Programing language: R

Libraries: webshot2,dplyr, data.table, caret, ggplot2, ROSE, randomForest, xgboost, pROC, rpart, rpart.plot, DiagrammeR, MASS, Rtsne and fastDummies

RESULTS:

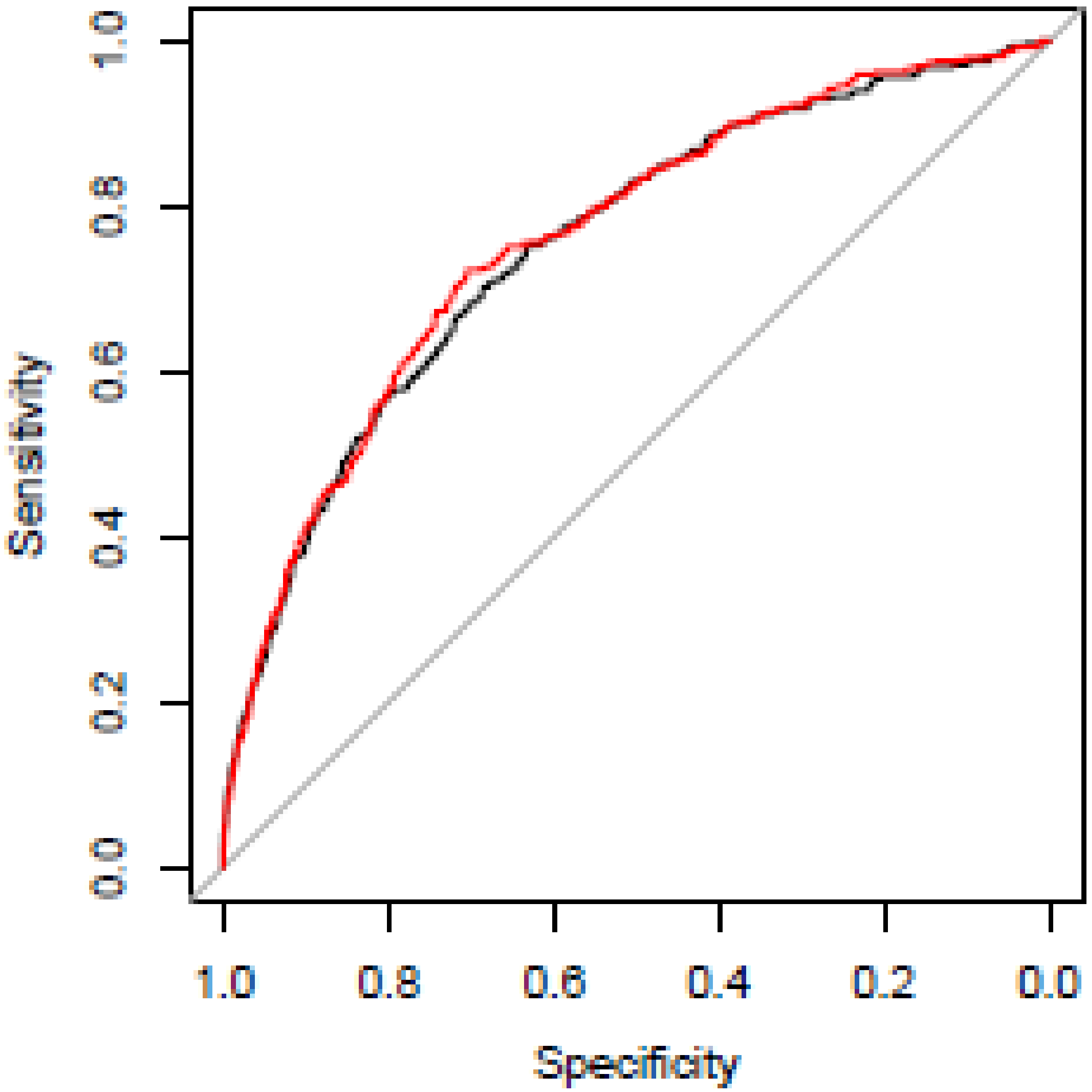
Model	Characteristics	Accuracy	Specificity	AIC
Logistic Regression	All the variables are used	*0.8897*	*0.4573*	1580.5
Logistic Regression	*all the variables, except survival months*	*0.851*	*0.1508*	*2086.3,*
Logistic Regression	*selected variables*	*0.8477*	*0.09548*	*2115.3*
Logistic Regression	Trained with 30% over sampling, and all variables.	*0.828*	*0.3719*	3518.8
Logistic Regression with PCA	Selected Variables	0.8447	0.1011	2090.2

AUC-ROC

Comparison between two models:

red – over sampling and all variables, AUC = 0.759

Black – all variables except survival months, AUC= 0.75.

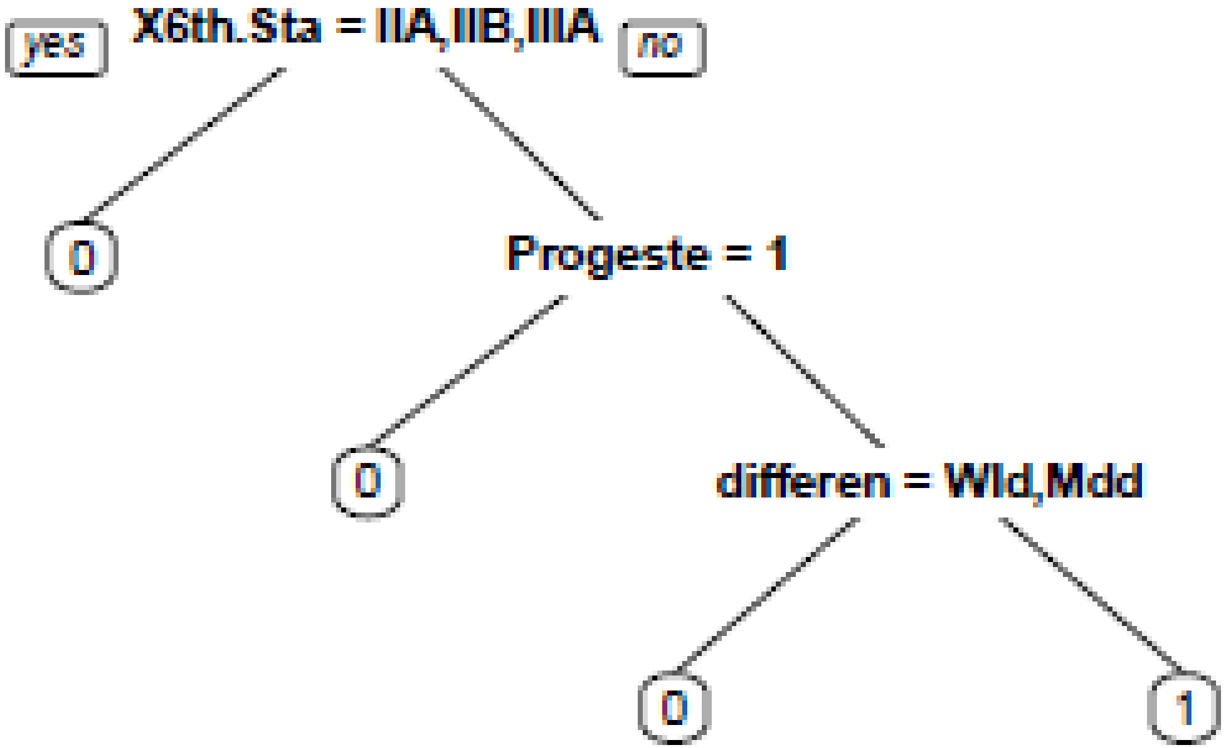


Model	Characteristics	Accuracy	Specificity
Decision Tree	All variables	0.8378601	0.07035176
Random Forest	Selected Variables	0.8444444	0.1005025
Random Forest	Selected Variables and 30% oversampling	*0.8222222*	*0.2763819*
Random Forest with PCA	Selected Variables	0.8354324	0.1755319

Model	Characteristics	Accuracy	Specificity
XGBoost	Scale pos weight and max_depth=10	*0.7835*	*0.2161*

Decision Tree with all the variables:

- Root node is the diagnosis, and contains information about tumor size, lymph nodes and metastasis.
-Progesterone is a hormone receptor in the cancer cell
- The differentiated variable indicates how mutated the cancer cell is.



CONCLUSIONS

- There is a clear trade-off between accuracy and specificity, when the model improves the detection of death events the false positive predictions increase also.

- Models that accurately predict death events could be used to detect high-risk patients, who could benefit from participating in medical trials. However because the accuracy of the model is not high enough and there are false positives people who could live with a normal treatment could risk their life in medical trials, thus all of these models should reach a higher accuracy.