Análise de Sobrevivência

Trabalho 3

Modelos de classificação - Câncer de Mama

Grupo 1 - Deusdedith, Eduardo Henrique,
Felipe Ferraz, Lucas Ribeiro, Thiago Ferro, Inês Dantas

O que é o câncer de mama?

O câncer de mama é uma doença causada pela multiplicação desordenada de células anormais da mama, que forma um tumor com potencial de invadir outros órgãos.

De acordo com o Instituto Nacional de Câncer (INCA), para cada ano do triênio 2020-2022 a estimativa é de 66.280 novos casos de câncer de mama no país.

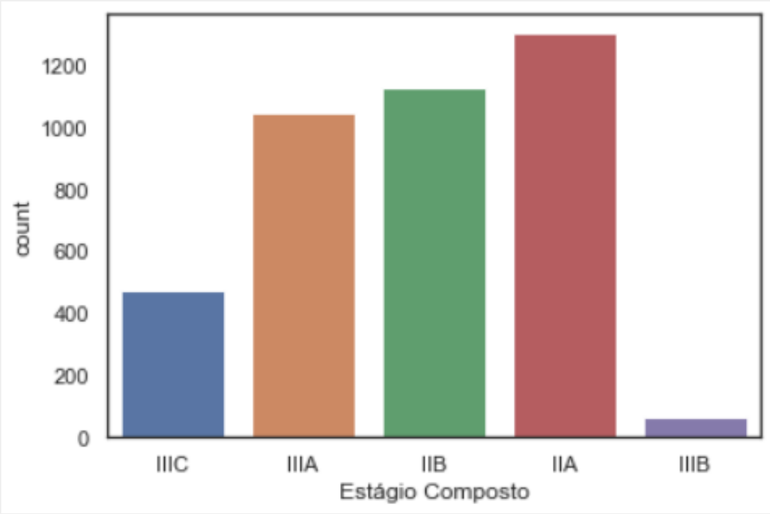
Dados

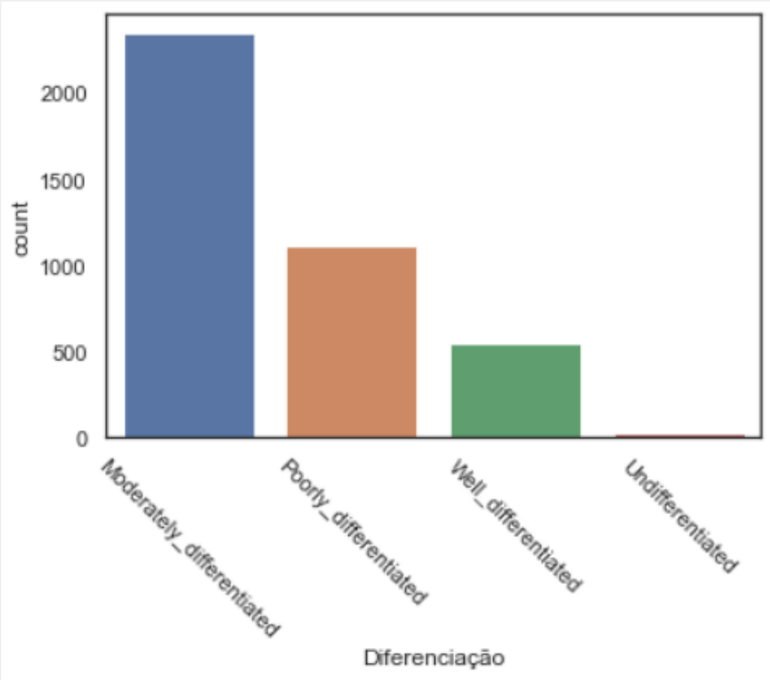
Os dados escolhidos foram obtidos da database do National Cancer Institute dos EUA, na plataforma para compartilhamento de pesquisas **zenodo**. Possui 4024 observações e 15 colunas, obtidas de 2006 à 2010 no Programa 'Surveillance, Epidemiology, and End Results' (SEER), que constam:

- Idade no momento do diagnóstico;
- Etnia do paciente;
- Estado Civil;
- T Stage - tamanho do tumor;
- N Stage - Disseminação da doença ;
- 6th Stage Grade - Estágio do câncer;
- A Stage - Localização do câncer;
- Tamanho do câncer (mm);
- Status do estrogênio;
- Status da progesterona;
- Quantidade de nódulos inspecionados;
- Quantidade de nódulos com metástase;
- Tempo de sobrevivência;
- **Status do paciente (vivo ou morto).**

EDA e Data Cleaning

- Idades concentrada acima de 45 anos (82%);
- Viés étnico branco (85%);
- Maioria de casados (~66%);
- Estágio;
- Diferenciação;
- Metástase;
- Estrogênio e Progesterona

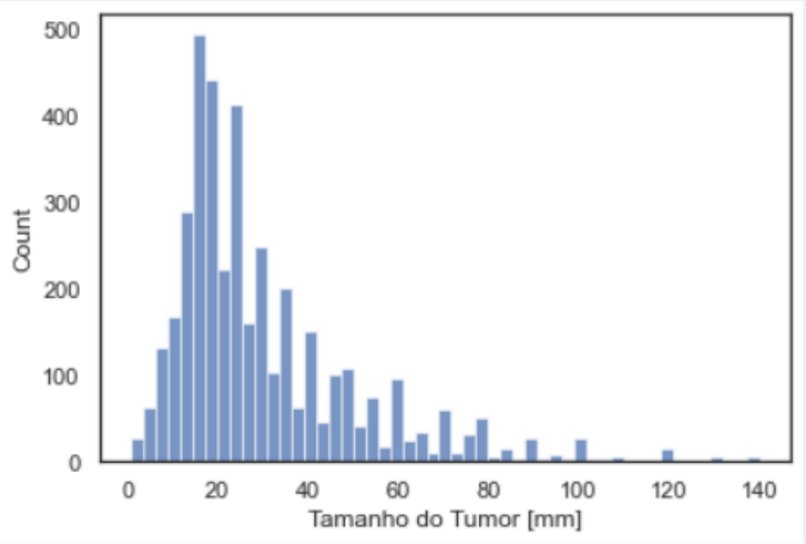


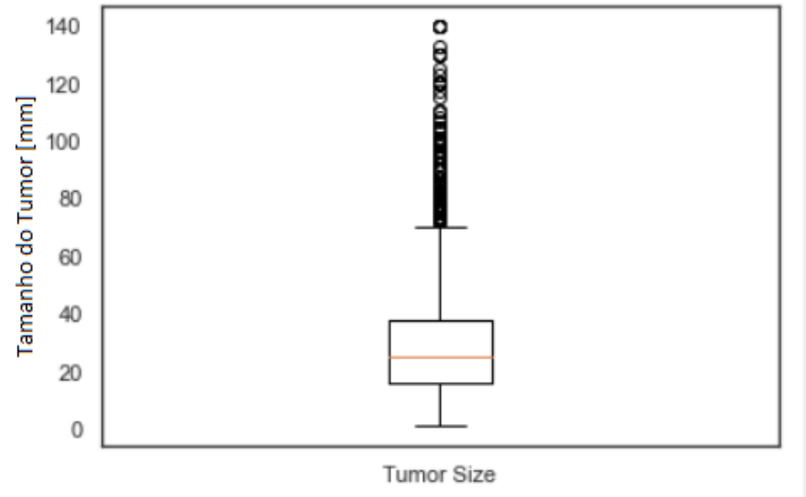


EDA e Data Cleaning

Tamanho do Tumor

- Feature numérica, boa preditora pois define o estágio e avanço da doença;
- Média 30.5mm, mediana 25mm;
- Limpeza: Fórmula Interquartil.





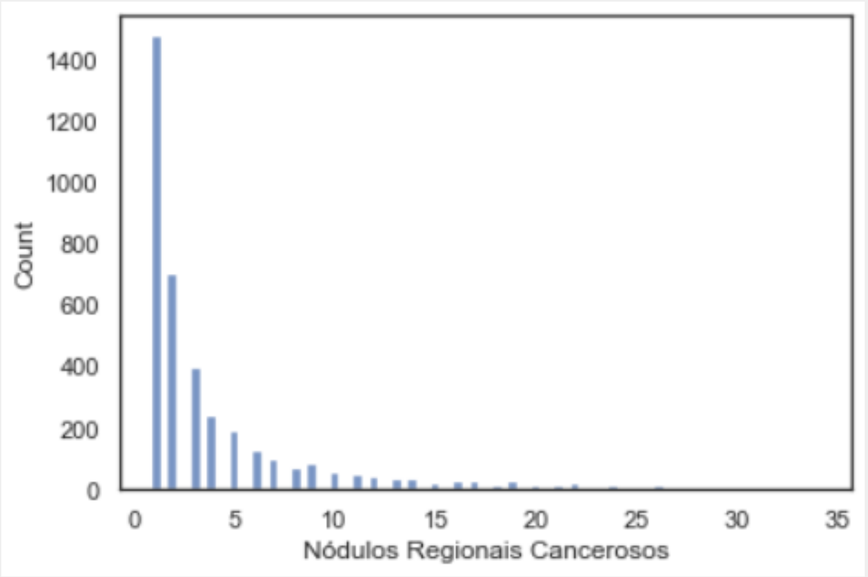
EDA e Data Cleaning

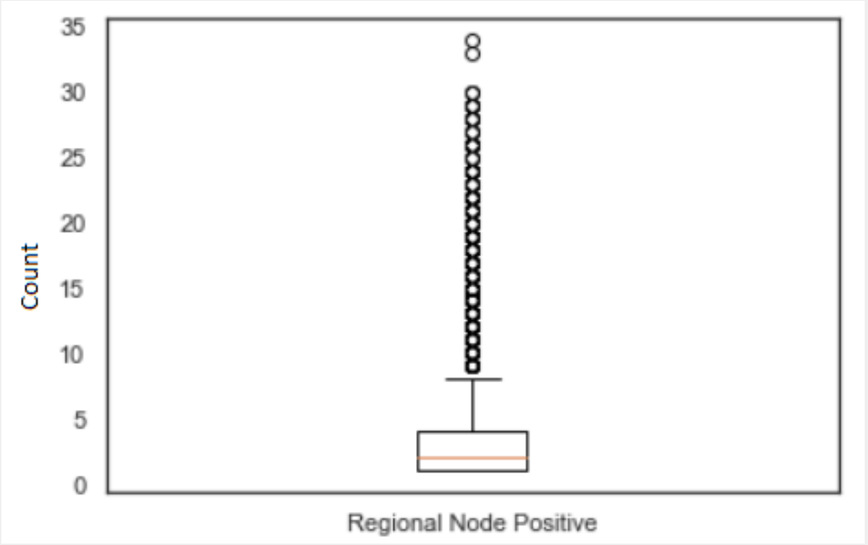
Nódulos Linfáticos Examinados e Nóds. Cancerosos

- Feature numérica, boa preditora pois também determina o avanço da doença;

- Examinados: média ~14 e medi. 14;
- Cancerosos: média ~4 e medi. 2;

- Mesmo método de limpeza (tanto exam. qto. positive).

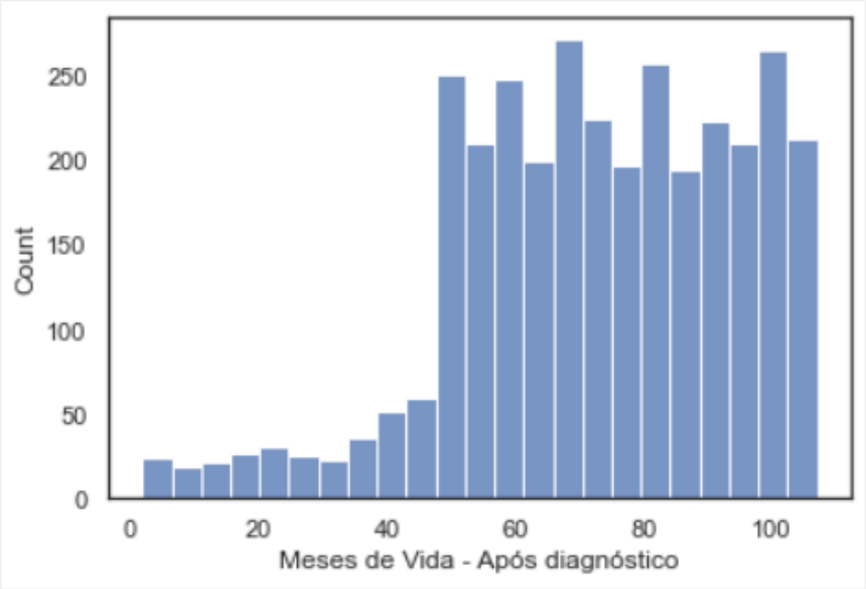




EDA e Data Cleaning

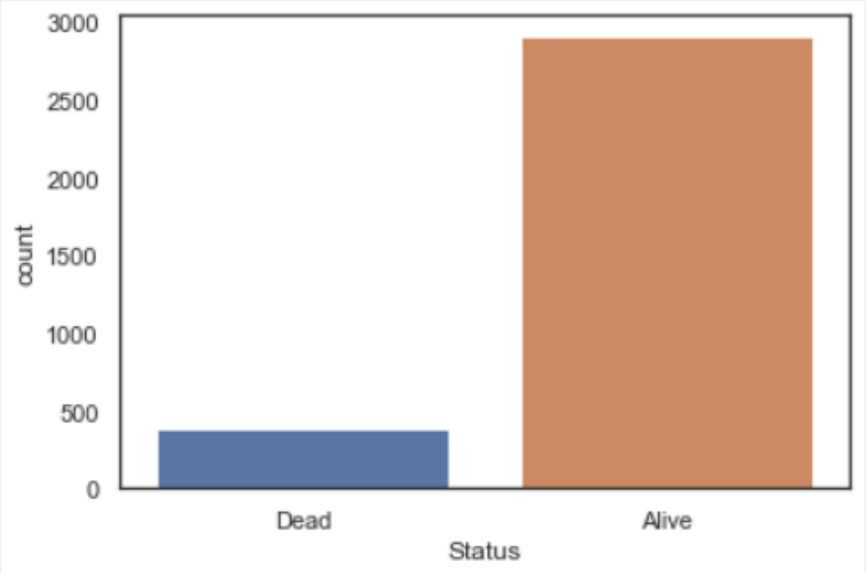
Tempo após Diagnóstico (Meses)

- Feature numérica, boa preditora pois quanto mais tempo com a doença, maior chance de morrer;
- Média ~71 e medi. 73;
- Sem Outliers.



Estado do Paciente

- Variável target
- Assimétrico (~88% vivo)



Feature Engineering

Encoding e Normalização

Categóricas não ordinais (Get Dummies)

- Etnia (Race), Estado Civil (Marital_Status), Metástase (A_Stage), Estrogênio e Progesterona.

Categóricas Ordinais (Ordinal Encoder, 0 1 2 3 4...)

- Tamanho categórico (T-Stage), Disseminação (N-Stage), Estágio Composto (6th-Stage) e Diferenciação (Grade).

Numéricas (Standard Scaler)

- Idade (Age), Tamanho do Tumor (Tumor_Size), Regional Nodes, Tempo após diagnóstico (Survival_Months).

Target binária → 0 = Morto, 1 = Vivo

Grade antes de Codificação Ordinal

```
Moderately differentiated; Grade II      2351
Poorly differentiated; Grade III         1111
Well differentiated; Grade I              543
Undifferentiated; anaplastic; Grade IV     19
Name: Grade, dtype: int64
```

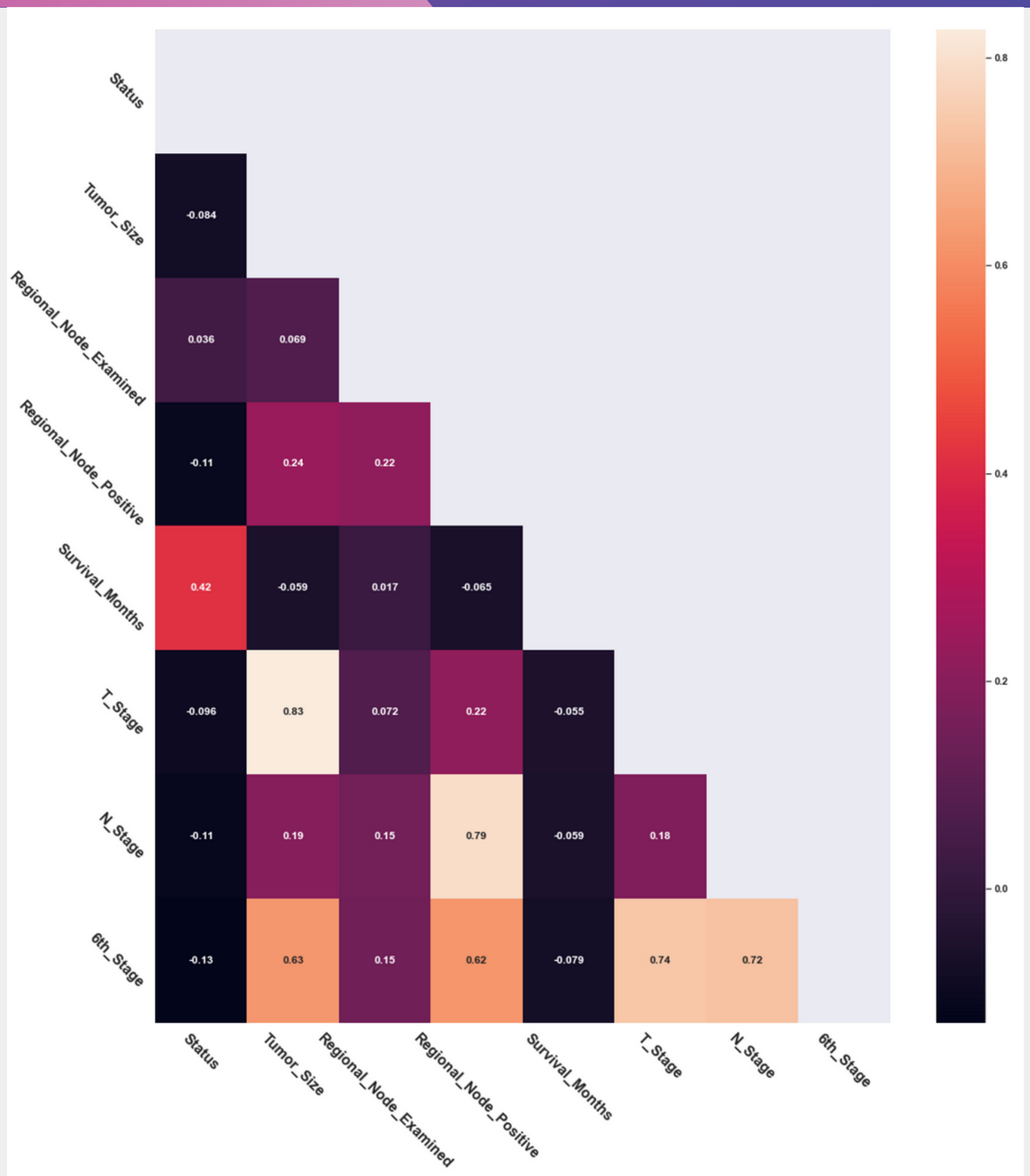
Grade Codificado

```
0    1973
1     830
3     472
2      10
Name: Grade, dtype: int64
```

Feature Selection

Correlações

- Alta correlação entre T_Stage, N_Stage, Tumor_Size com 6th_Stage (Estágio Composto)
- T_Stage com Tumor_Size (categórica de tamanho)
- N_Stage com Nódulos Cancerosos
- Correlação de dummies uma entre a outra
- Quantia de Nódulos Cancerosos, Meses após diagnóstico com a Target



Feature Selection

	coef	std err	z	P>|z|	[0.025	0.975]
T_Stage	-0.2715	0.258	-1.051	0.293	-0.778	0.235
N_Stage	-0.0265	0.393	-0.067	0.946	-0.798	0.745
6th_Stage	-0.0925	0.244	-0.379	0.705	-0.571	0.386
Grade	0.0733	0.078	0.942	0.346	-0.079	0.226
Regional_Node_Positive	-0.1411	0.061	-2.323	0.020	-0.260	-0.022
Race_Other	0.7246	0.378	1.917	0.055	-0.016	1.466
Race_White	0.5387	0.254	2.124	0.034	0.042	1.036
Marital_Status_Married	0.5648	0.213	2.646	0.008	0.147	0.983
Marital_Status_Separated	0.3035	0.737	0.412	0.680	-1.140	1.747
Marital_Status_Single	0.4049	0.268	1.509	0.131	-0.121	0.931
Marital_Status_Widowed	0.2567	0.335	0.767	0.443	-0.399	0.912
A_Stage_Regional	1.3127	0.352	3.731	0.000	0.623	2.002
Estrogen_Status_Positive	0.8149	0.291	2.796	0.005	0.244	1.386
Progesterone_Status_Positive	0.3853	0.208	1.852	0.064	-0.023	0.793
Age	-0.3008	0.080	-3.763	0.000	-0.457	-0.144
Tumor_Size	0.0346	0.113	0.307	0.759	-0.186	0.256
Regional_Node_Examined	0.2051	0.079	2.609	0.009	0.051	0.359
Survival_Months	1.3433	0.082	16.426	0.000	1.183	1.504

Summary - Todas as Features

Testes de Hipótese (Stats Models Summary)

	coef	std err	z	P>|z|	[0.025	0.975]
Regional_Node_Examined	0.0355	0.011	3.307	0.001	0.014	0.056
Regional_Node_Positive	-0.1603	0.039	-4.147	0.000	-0.236	-0.085
Estrogen_Status_Positive	1.1679	0.248	4.712	0.000	0.682	1.654
A_Stage_Regional	2.5241	0.409	6.168	0.000	1.722	3.326
Race_White	0.2411	0.197	1.225	0.221	-0.145	0.627
Age	-0.0214	0.007	-2.999	0.003	-0.035	-0.007
Tumor_Size	-0.1508	0.071	-2.115	0.034	-0.291	-0.011
Survival_Months	1.3252	0.080	16.546	0.000	1.168	1.482

Summary - Features Seletas

Feature Selection

Variance Inflation Factor (VIF)

- Forma de medir Multicolinearidade
- VIF > 5 indica variável com alta multicolinearidade
- Altera o cálculo em combinações diferentes de features

	Var	VIF
2	6th_Stage	22.70
0	T_Stage	14.65
1	N_Stage	8.88
4	Regional_Node_Positive	8.43
15	Tumor_Size	3.02
5	Race_Other	2.21
11	A_Stage_Regional	2.05
10	Marital_Status_Widowed	1.54
3	Grade	1.45
14	Age	1.09
8	Marital_Status_Separated	1.09
16	Regional_Node_Examined	1.06
17	Survival_Months	1.02
9	Marital_Status_Single	0.89
6	Race_White	0.35
7	Marital_Status_Married	0.33
12	Estrogen_Status_Positive	0.21
13	Progesterone_Status_Positive	0.09

VIF - Todas as Features

	Var	VIF
5	Age	32.30
0	Regional_Node_Examined	4.57
1	Regional_Node_Positive	3.10
3	A_Stage_Regional	2.58
6	Tumor_Size	1.07
7	Survival_Months	1.02
4	Race_White	0.17
2	Estrogen_Status_Positive	0.16

VIF - Features Seletas

Modelos

Regressão Logística

K-Nearest Neighbors

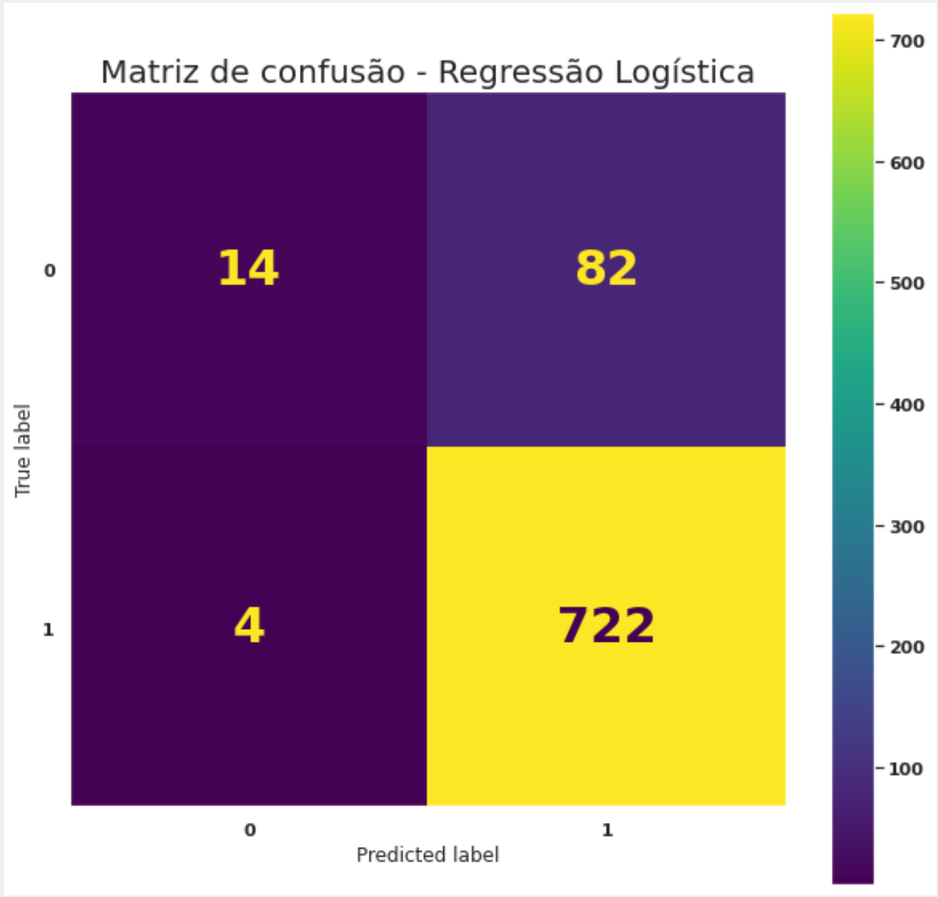
GaussianNB

Random Forest Classifier

XGBoost Classifier

*Melhor Threshold = 0.3 p/ todos modelos

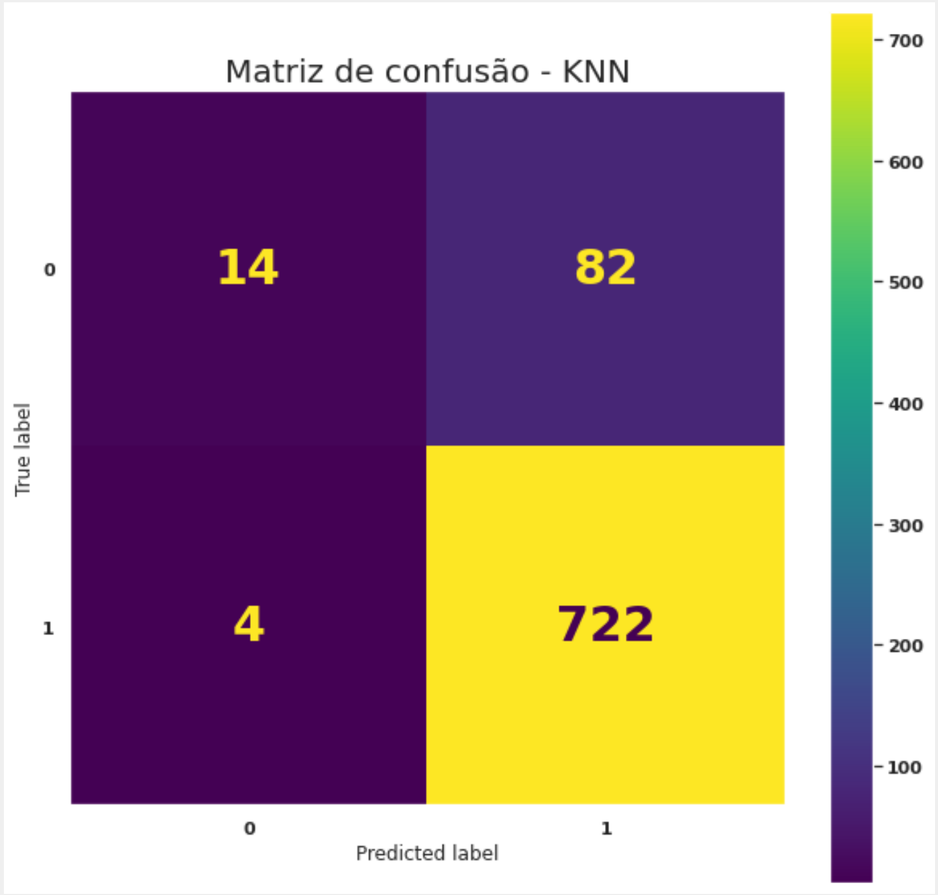
Regressão Logística



- Acurácia, Recall e Falso Negativo

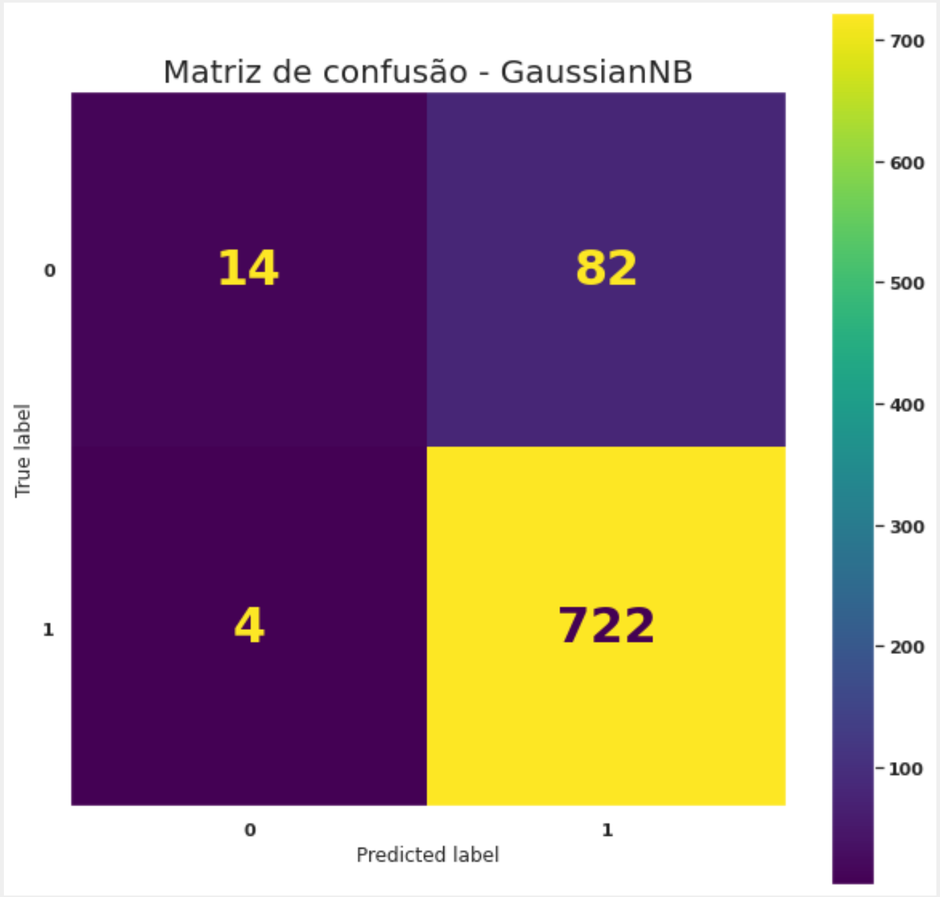
	precision	recall	f1-score	support
0	0.78	0.15	0.25	96
1	0.90	0.99	0.94	726
accuracy			0.90	822
macro avg	0.84	0.57	0.59	822
weighted avg	0.88	0.90	0.86	822

K-Nearest Neighbors



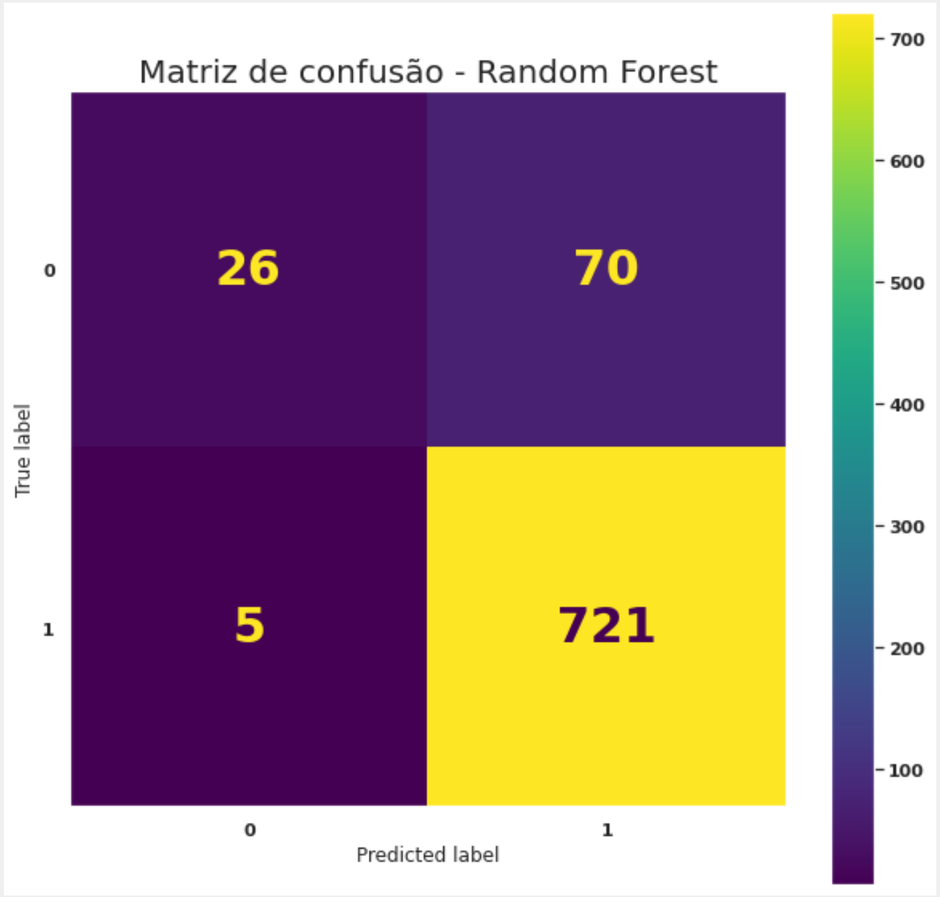
	precision	recall	f1-score	support
0	0.78	0.15	0.25	96
1	0.90	0.99	0.94	726
accuracy			0.90	822
macro avg	0.84	0.57	0.59	822
weighted avg	0.88	0.90	0.86	822

GaussianNB



	precision	recall	f1-score	support
0	0.78	0.15	0.25	96
1	0.90	0.99	0.94	726
accuracy			0.90	822
macro avg	0.84	0.57	0.59	822
weighted avg	0.88	0.90	0.86	822

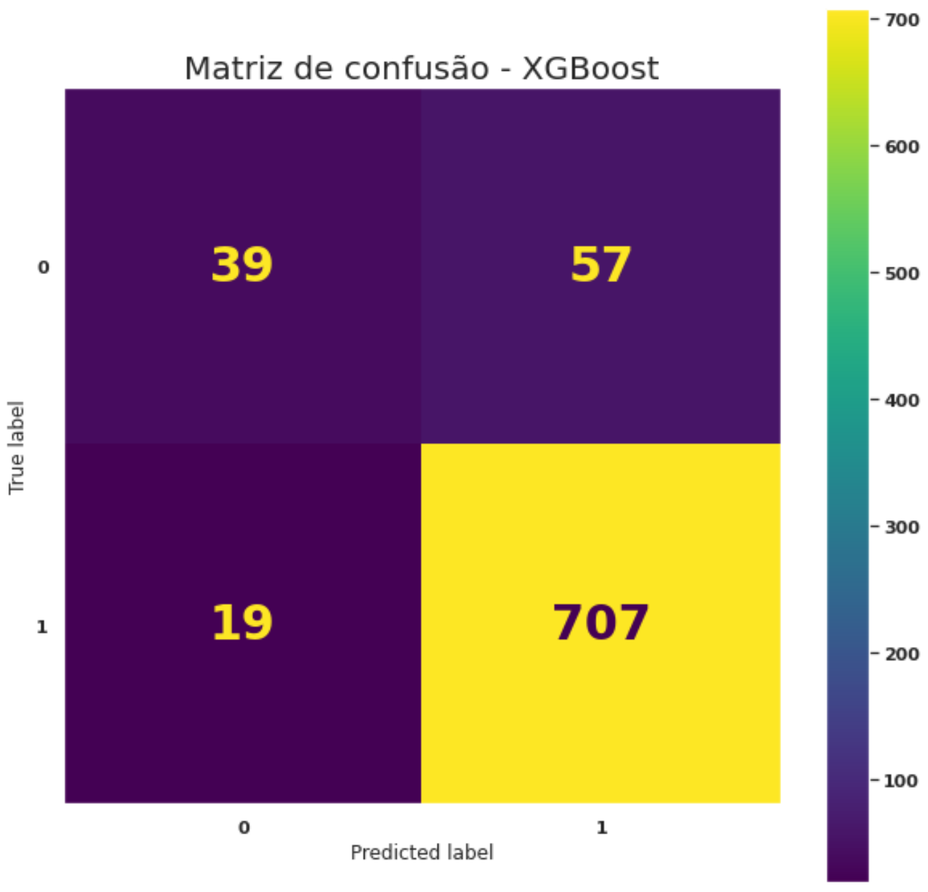
Random Forest



- Todas as Features

	precision	recall	f1-score	support
0	0.84	0.27	0.41	96
1	0.91	0.99	0.95	726
accuracy			0.91	822
macro avg	0.88	0.63	0.68	822
weighted avg	0.90	0.91	0.89	822

XGBoost



- Todas as Features (cru, sem pré processamento)

********** Melhor XGBoost Classifier **********

	precision	recall	f1-score	support
0	0.67	0.41	0.51	96
1	0.93	0.97	0.95	726
accuracy			0.91	822
macro avg	0.80	0.69	0.73	822
weighted avg	0.90	0.91	0.90	822

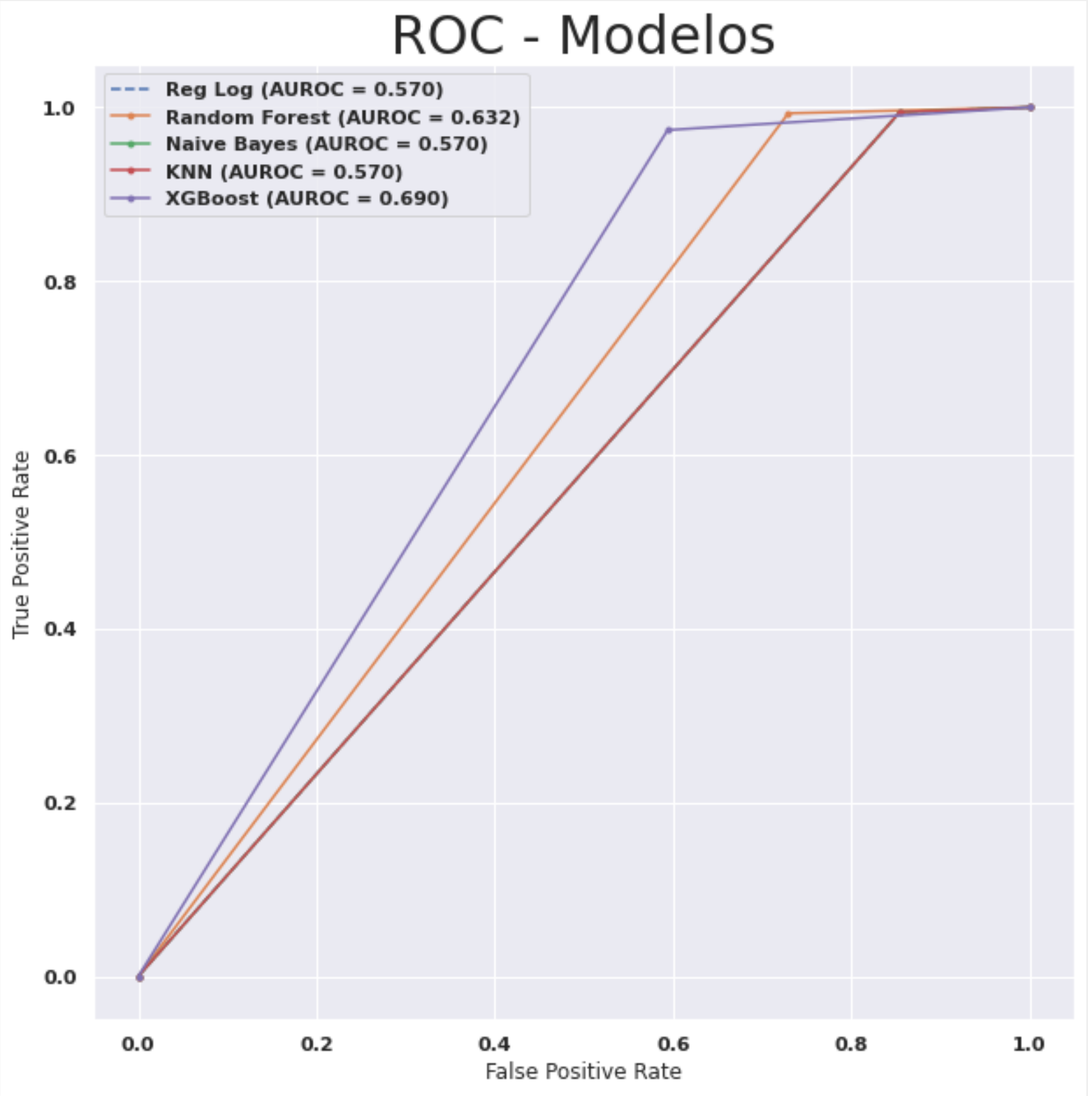
Acurácia: 0.9075

Recall: 0.9738

Performance

- XGBoost melhor modelo (>AUC), apesar de RF ter melhor Acurácia e Recall;
- Reg. Log., KNN e Gaussian NB performance idêntica, chegaram ao mesmo resultado;
- Dummy Classifier performa similar pois threshold de todos é 30% de probabilidade de Vivo.

	Acurácia	Recall	FN	F1-Score	AUROC
Random Forest	0.9075	0.9904	7.0	0.9498	0.6358
XGBoost - All Feats.	0.9075	0.9738	19.0	0.9490	0.6900
Reg. Logística	0.8954	0.9945	4.0	0.9438	0.5702
KNN	0.8954	0.9945	4.0	0.9438	0.5702
Gaussian Naive-Bayes	0.8954	0.9945	4.0	0.9438	0.5702
Dummy Classifier	0.8832	1.0000	0.0	0.9380	NaN



Fontes

O que é o câncer de mama?

https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama

O que é o câncer de mama?

https://fcmsantacasasp.edu.br/blog/estatisticas-sobre-o-cancer-de-mama-no-brasil/

Dados

https://zenodo.org/record/5120960#.YtbR7ITMKUI