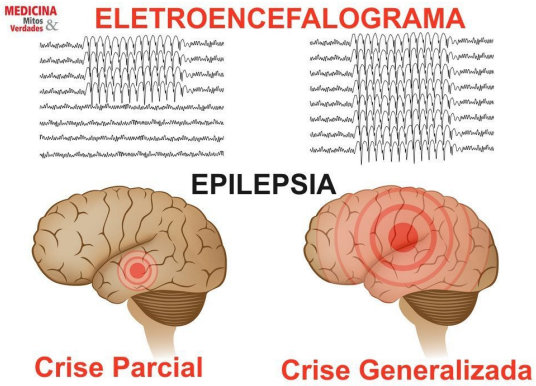
Aprendizado de máquina para análise da frequência cardíaca e detecção de epilepsia

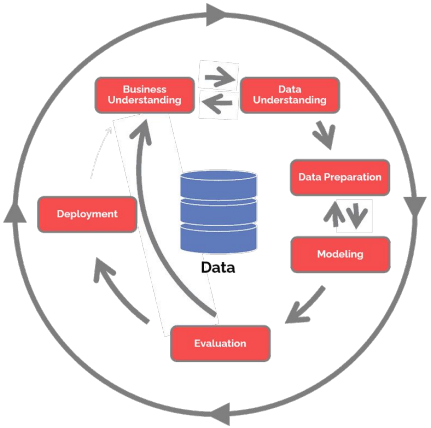
Ana Paula da Rocha, Alexandre Soli Soares , Vinicius Cin

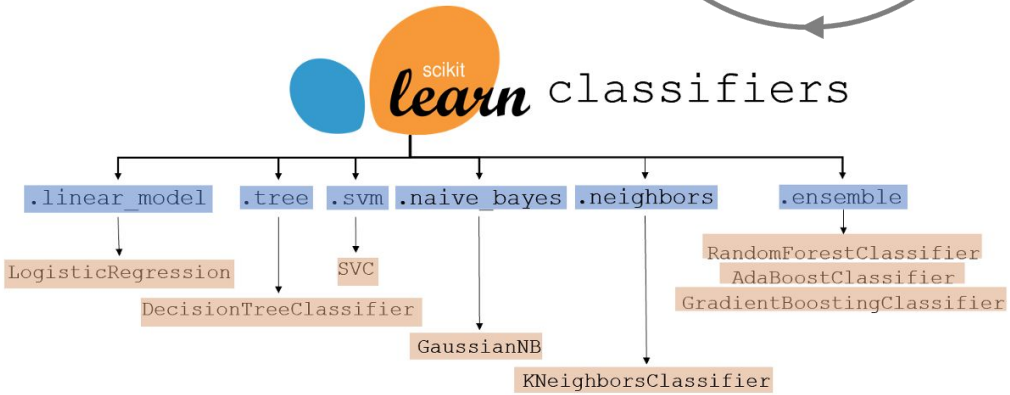
Introdução

Epilepsia



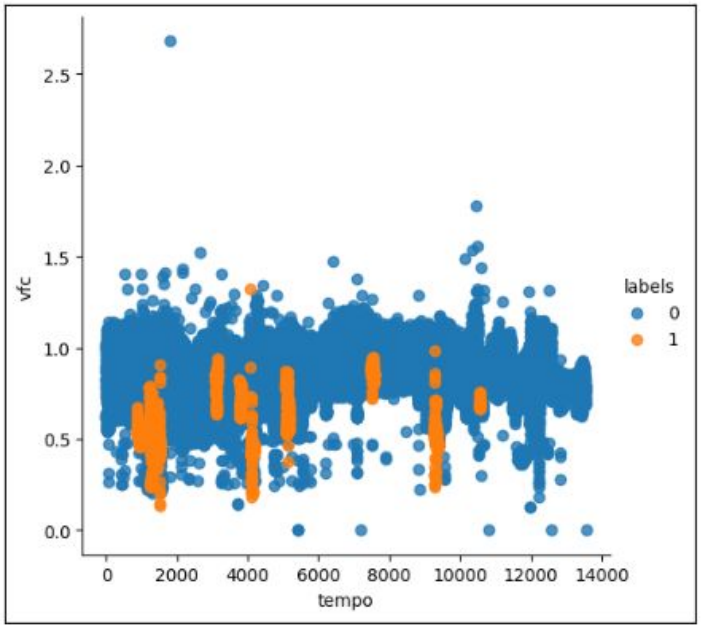
Mineração dos dados

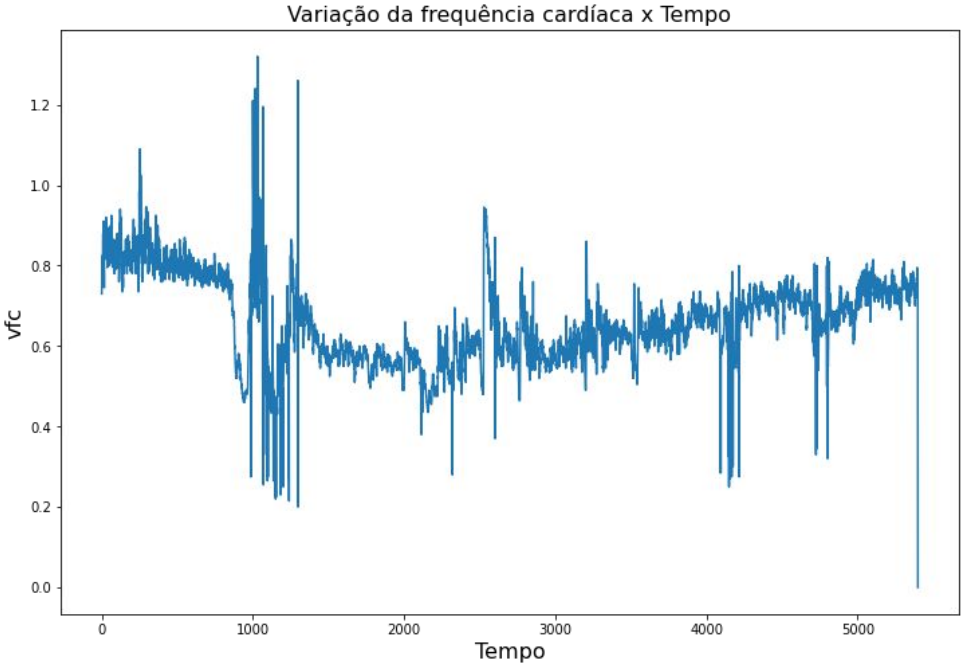




Análise Exploratória

Figura 1: Quantidade de amostras normais e com epilepsia no conjunto de dados.





Tratamento dos dados

- Classificação das amostras

```
Número de amostras= 1079998 , taxa de amostragem= 200
indexes "C" (1):  1088 1273
dif: 185
tamanho amostras 8385
--------------------------
Número de amostras= 2519998 , taxa de amostragem= 200
indexes "C" (1):  3875 3959
dif: 84
indexes "C" (1):  10842 10877
dif: 35
tamanho amostras 13195
--------------------------
Número de amostras= 2711998 , taxa de amostragem= 200
indexes "C" (1):  5790 5957
dif: 167
indexes "C" (1):  10954 11172
dif: 218
tamanho amostras 16384
--------------------------
Número de amostras= 1079998 , taxa de amostragem= 200
indexes "C" (1):  1226 1408
dif: 182
tamanho amostras 6229
--------------------------
```

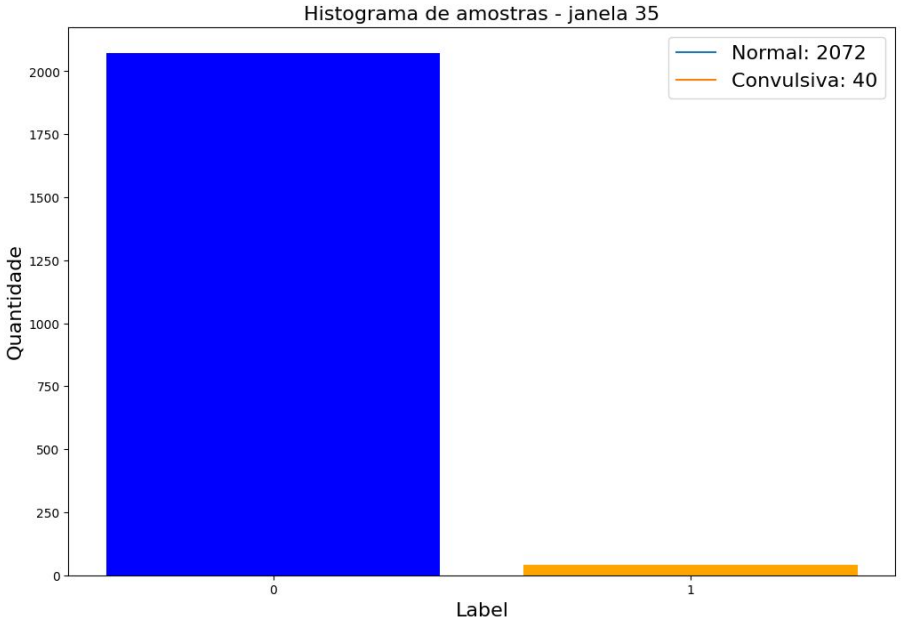
```
--------------------------
Número de amostras= 1080006 , taxa de amostragem= 200
indexes "C" (1):  2164 2324
dif: 160
tamanho amostras 8076
--------------------------
Número de amostras= 2159998 , taxa de amostragem= 200
indexes "C" (1):  3729 3796
dif: 67
indexes "C" (1):  8894 8991
dif: 97
tamanho amostras 12758
--------------------------
Número de amostras= 1439998 , taxa de amostragem= 200
indexes "C" (1):  4915 5122
dif: 207
tamanho amostras 8892
--------------------------
```

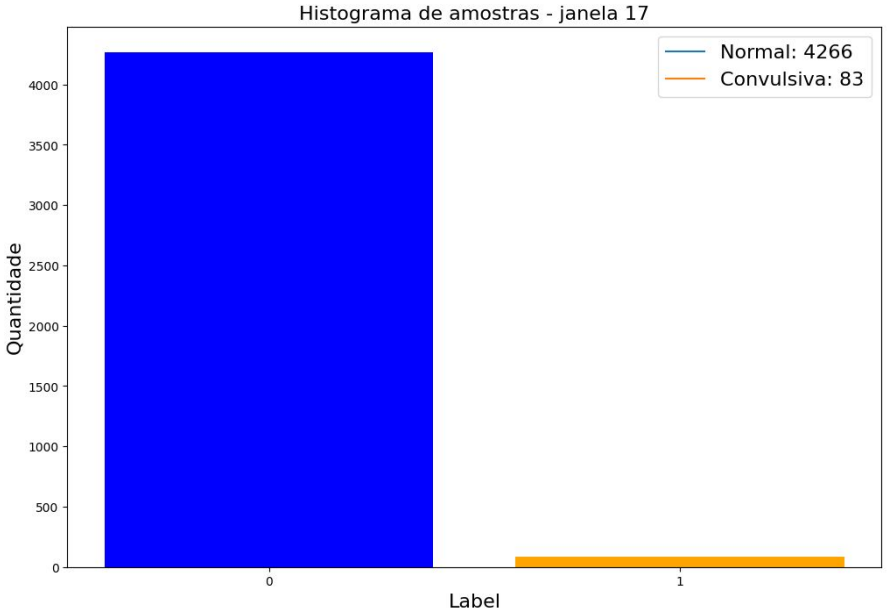
Tratamento dos dados

- Diferenciar labels com metadados
  - 72507 normais e 1412 com epilepsia

- Extração das features
  - A quantidade N de amostras ser o mínimo disponível entre os arquivos, que no caso é 35 amostras ou
  - utilizar 17 amostras, pois teria uma maior garantia de que todas as targets com epilepsia seriam usadas
    - N = 17: negativas: 4266, positivas: 83
    - N = 35: negativas: 2072, positivas: 40

Tratamento dos dados





Tratamento dos dados

- Dados nulos

Features Qtd. Amostras	'hfnu'	'lfnu'	'mean_hr'	'max_hr'	'std_hr'	'lf_hf_ratio'
17 amostras	1315	1315	7	7	7	1315
35 amostras	3	3	7	7	7	3

Tabela 1: Quantidade de valores nulos e infinitos para as *features*

- Normalização
  - Os modelos são afetados pelos valores

Resultados - Influência da Janela

Figura 2: Features para janelas de 17 amostras x score.

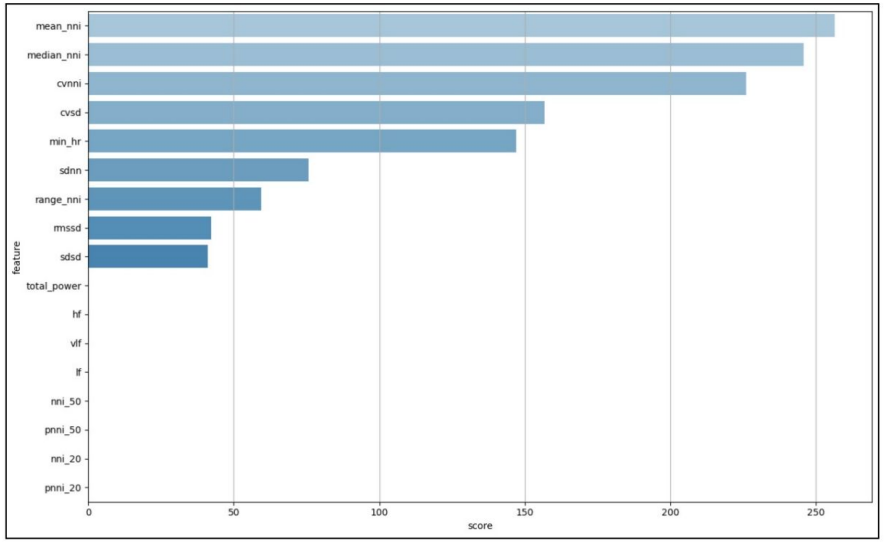
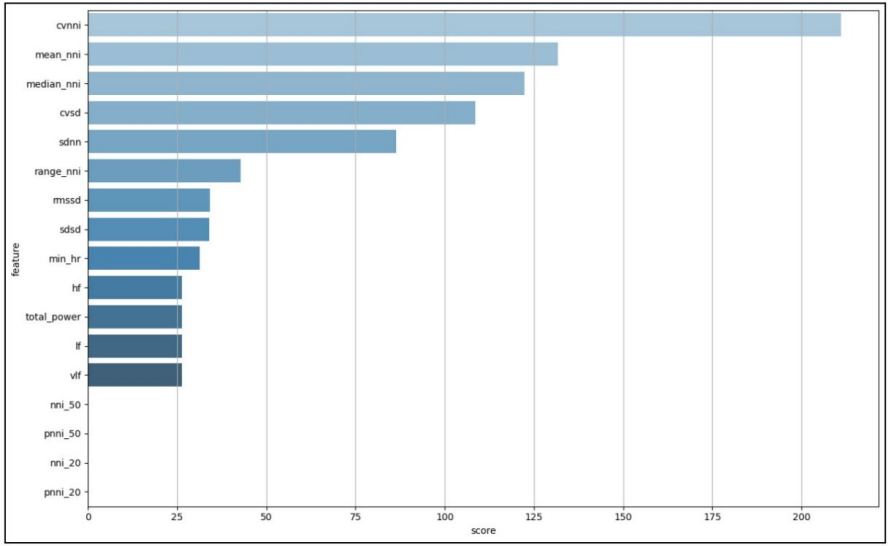


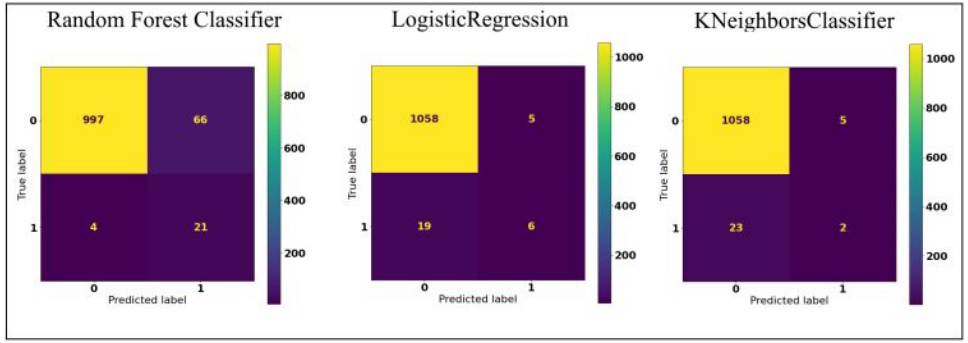
Figura 3: Features para janelas de 35 amostras x score.



Resultados - Métricas

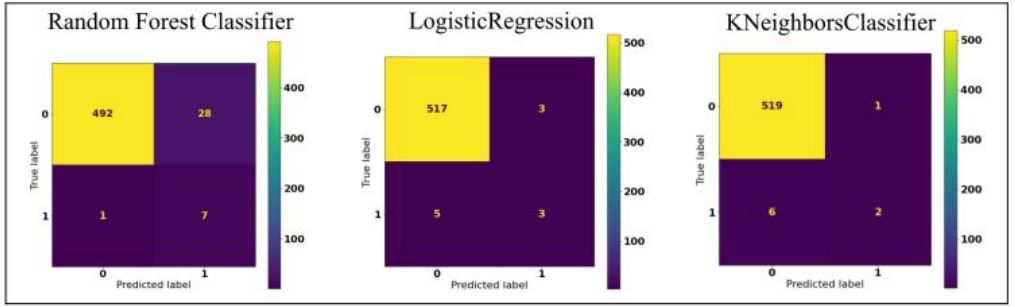
Modelo	Classe	Precisão	Recall	Score F1
Random Forest	0	0.9960	0.9379	0.9660
	1	0.2413	**0.8400**	0.375
Regressão Logística	0	0.9823	0.9952	0.9887
	1	**0.5454**	0.2400	0.3333
KNeighbors	0	0.9787	0.9952	0.9869
	1	0.2857	0.080	0.1250

Tabela 2: Métricas obtidas para os modelos com janela de 17 amostras



Modelo	Classe	Precisão	Recall	Score F1
Random Forest	0	0.9979	0.9461	0.9713
	1	0.2000	**0.8750**	0.3255
Regressão Logística	0	0.9904	0.9942	0.9923
	1	0.5	0.375	0.4285
KNeighbors	0	0.9885	0.9980	0.9933
	1	**0.6666**	0.25	0.3636

Tabela 3: Métricas obtidas para os modelos com janela de 35 amostras

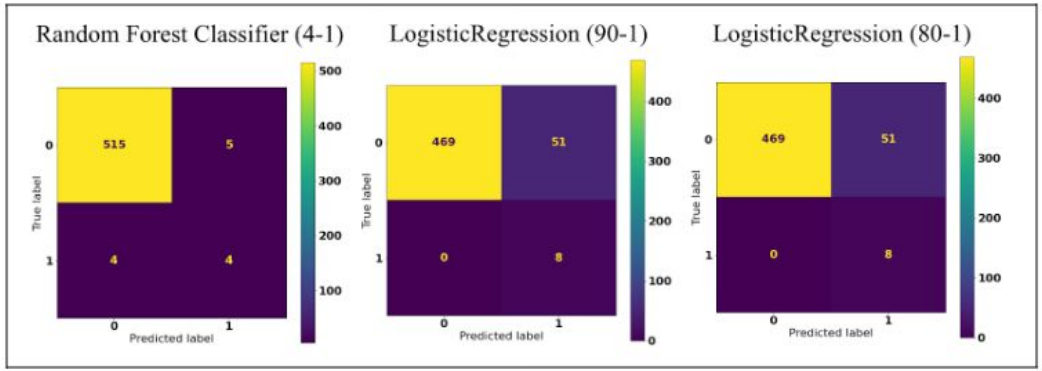


Resultados - Métricas - com peso

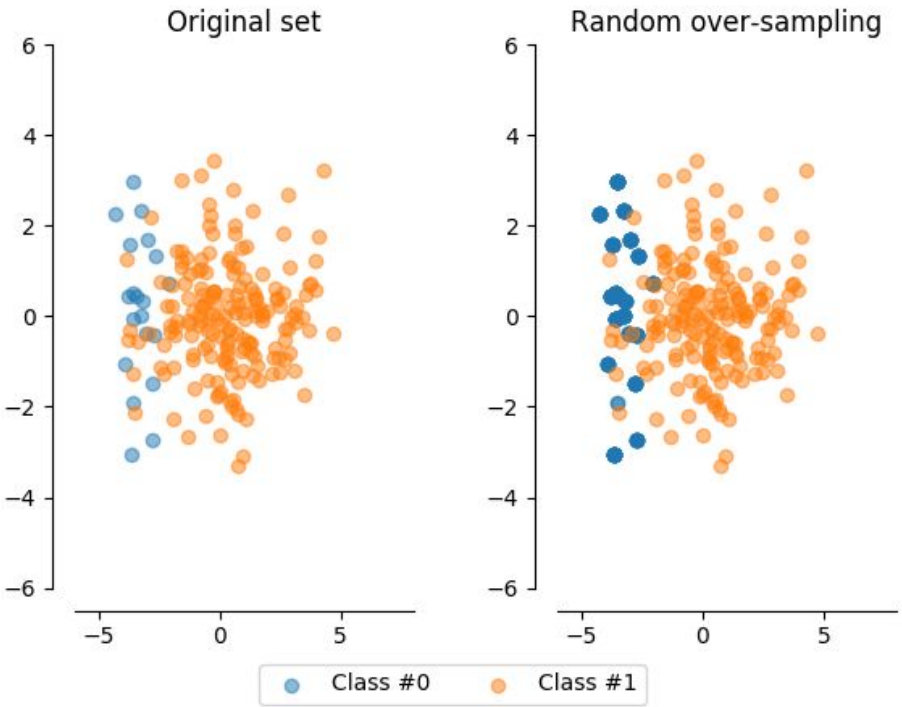
Modelo	Classe	Precisão	Recall	Score F1
Random Forest (pesos 4 para 1)	0	0.9922	0.9903	0.9913
	1	**0.4444**	0.5	0.4705
Regressão Logística (pesos 90 para 1)	0	1.0	0.9019	0.9484
	1	0.1355	1.0	0.2388
Regressão Logística (pesos 80 para 1)	0	0.9978	0.9076	0.9506
	1	0.1272	**0.875**	0.2222

Tabela 4: Métricas obtidas para os modelos com janela de 35 amostras e balanceamento através de pesos

Obs: O Sklearn não proporciona adição de pesos para o método *KNeighbors*

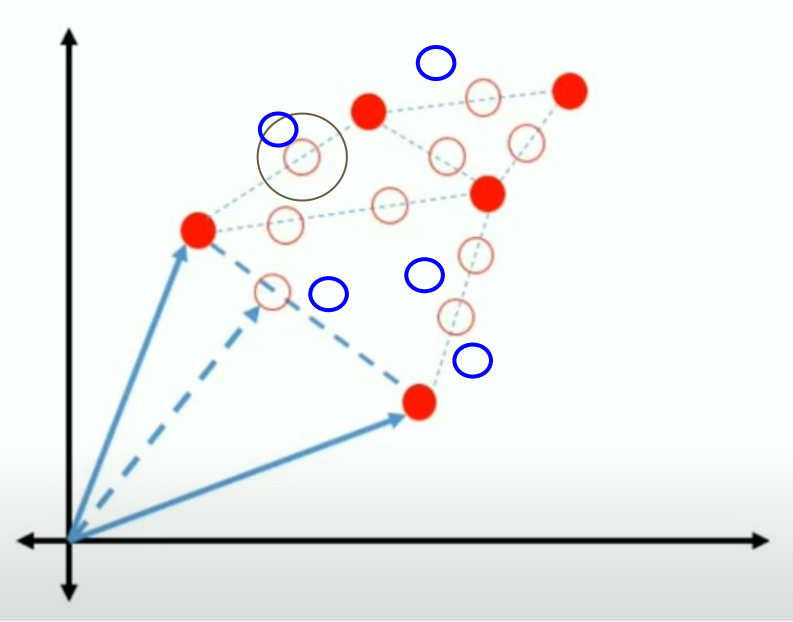


Random Over Sampler



SMOTE e ADASYN

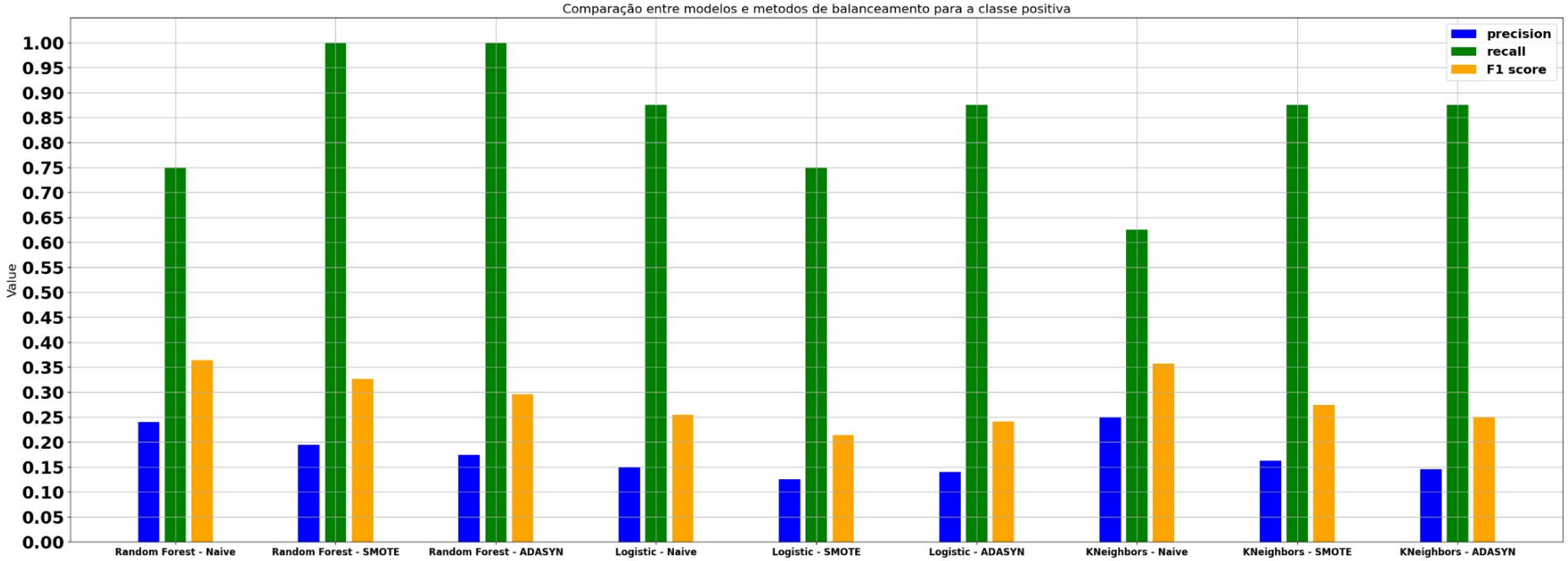
Synthetic Minority Over-sampling Technique
Adaptive Synthetic



SMOTE
Identifica cada feature através de um vetor
Obtém os K vizinhos mais próximos
Calcula a distância entre eles através do vetores
Multiplica esta distância por um número aleatório [0,1]
Adiciona um novo valor

ADASYN
Muito similar ao SMOTE, porém elimina a correlação linear entre as novas amostras utilizando uma distribuição normal

Resultados - Dados balanceados



Conclusão

Random Forest + SMOTE
Precisão: ~20%
Recall: 100%