Enap

coding BOOTCAMP

Bootcamp Machine Learning – Turma 1/21

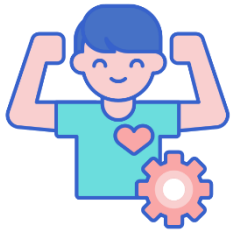
Predição de abandono de tratamento da Tuberculose

Daniel Castro, Albino Aveleda, Andre Aloise

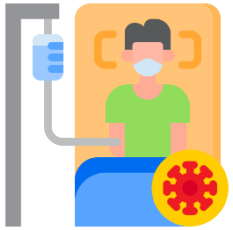
Descrição do Problema e da Solução

- O Brasil é um dos países prioritários para controle da Tuberculose
- O Amazonas apresenta uma das maiores incidências entre as UFs

Diagnosticado com TB



Cura
Conclusão do tratamento



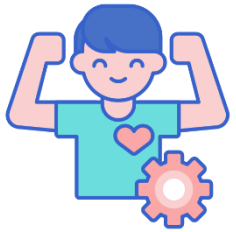
15 a 20%
Abandono do tratamento
Recidiva da doença
Formas resistentes
Óbito

https://creativemarket.com/eucalyp

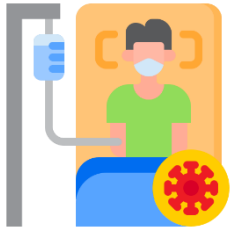
Descrição do Problema e da Solução

- O Brasil é um dos países prioritários para controle da Tuberculose
- O Amazonas apresenta uma das maiores incidências entre as UFs

Diagnosticado com TB

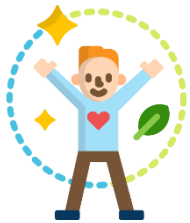


Cura
Conclusão do tratamento



15 a 20%
Abandono do tratamento
Recidiva da doença
Formas resistentes
Óbito

tratamento observado



Cura
Conclusão do tratamento

https://creativemarket.com/eucalyp

Fonte de Dados e Variáveis

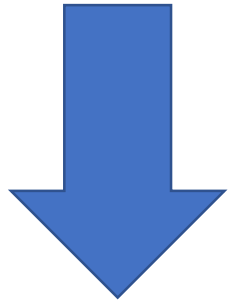
- Sistema de informação da Tuberculose no Amazonas, 2007 a 2020
- Dimensão do dataset original: (63271, 160)

Variáveis independentes:

- Gênero (Masculino, Feminino, Indefinido)
- Idade (anos)
- Gestante (sim, não)
- Raça/cor (Amarela, Branca, Parda, Preta, Indígena)
- Escolaridade (Sem estudo, fundamental, médio, superior)
- Portador HIV (sim, não)
- Aids (sim, não)
- Etilista (sim, não)
- Diabetes (sim, não)
- Neuropatologia (sim, não)
- Drogas (sim, não)
- Fumante (sim, não)
- Forma (pulmonar, extrapulmonar)
- Comorbidade (sim, não)
- Município
- Bairro
- CEP

Variável dependente:

- Abandono (Sim, Não)

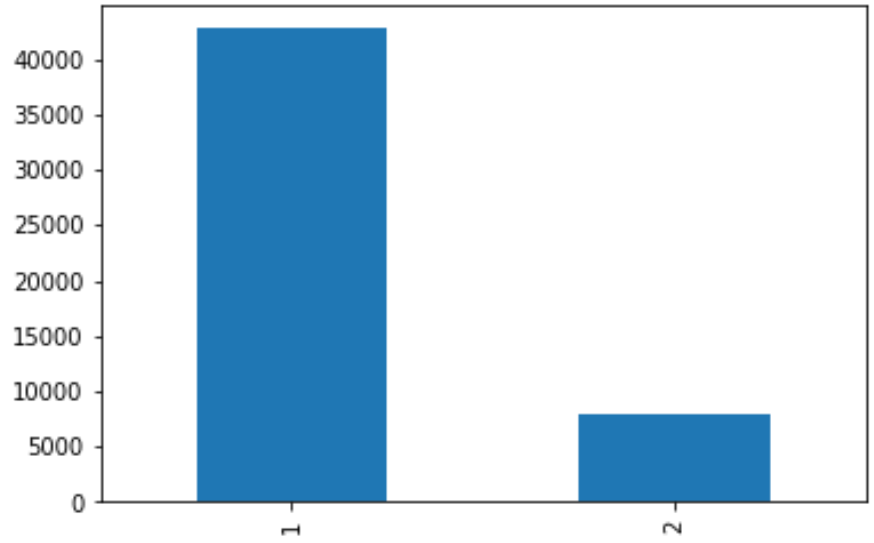


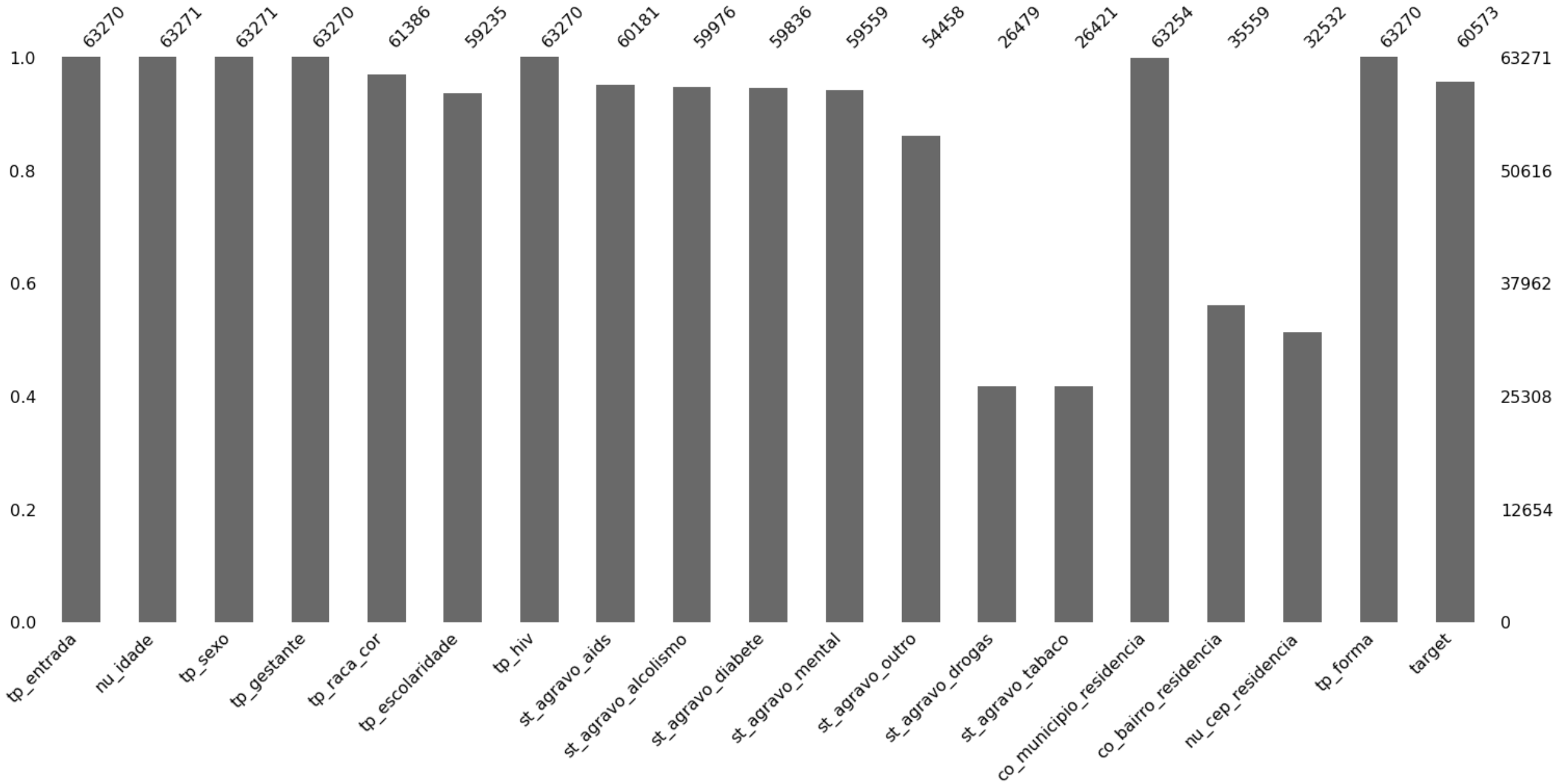
Modelo de Classificação

Resumo da Análise de Dados

- Redução das variáveis independentes
  - **160 -> 18**
- Campos não preenchidos
  - + 160.000
  - CEP 48.58%
  - Bairro 43.80%
  - Target 4.26%
- CEP
  - Considerado até o subsetor

- Desbalanceamento do target
  - Cura **84.29%**
  - Abandono **15.71%**



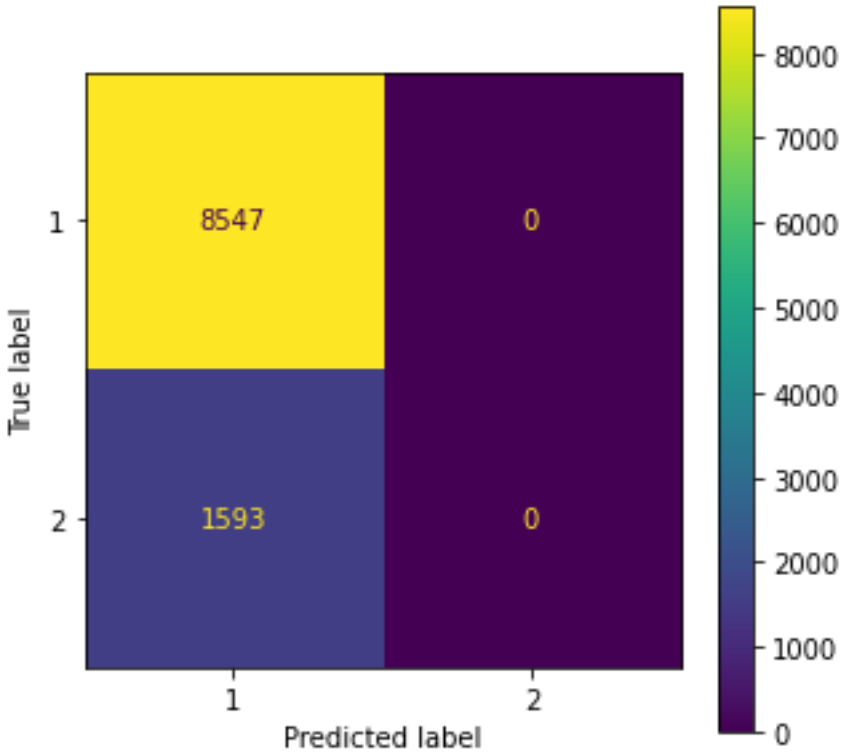

63270
63271
63271
63270
61386
59235
63270
60181
59976
59836
59559
54458
26479
26421
63254
35559
32532
63270
60573
1.0
0.8
0.6
0.4
0.2
0.0
63271
50616
37962
25308
12654
0
tp_entrada
nu_idade
tp_sexo
tp_gestante
tp_raca_cor
tp_escolaridade
tp_hiv
st_agravo_aids
st_agravo_alcolismo
st_agravo_diabete
st_agravo_mental
st_agravo_outro
st_agravo_drogas
st_agravo_tabaco
co_municipio_residencia
co_bairro_residencia
nu_cep_residencia
tp_forma
target

Modelos Utilizados

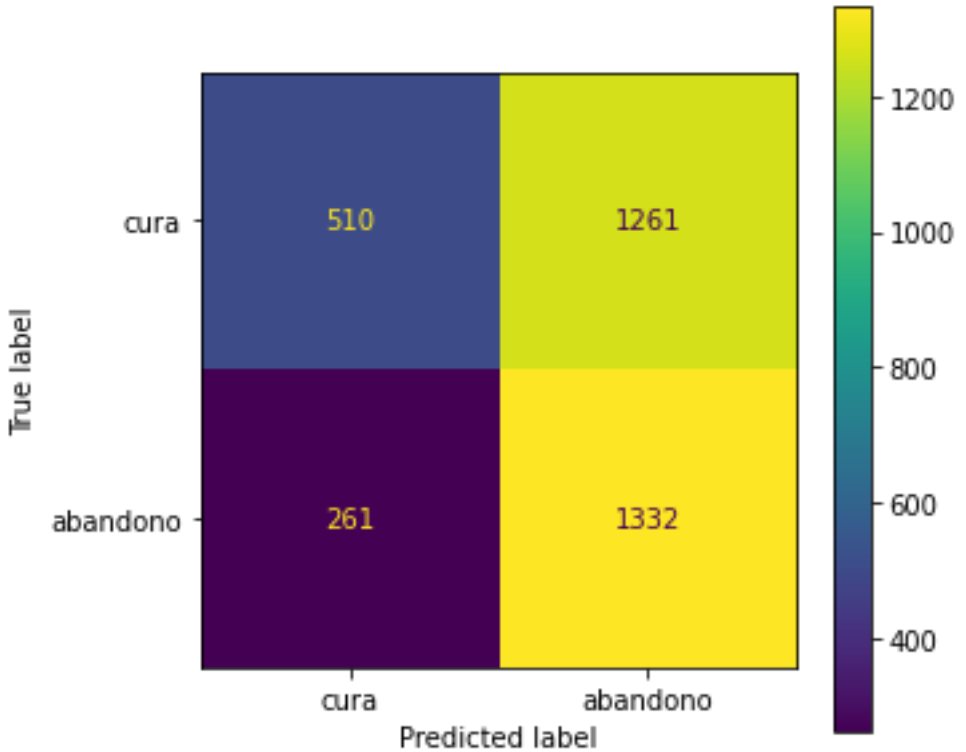
- Logistic Regression
- K Nearest Neighbor (Knn)
- Decision Tree Classifier
- Random Forest
- Support Vector Machine (SVM)
- Extreme Gradient Boosting (XGBoost)
- Gaussian Naive Bayes

Resultados intermediários

Regressão Logistica - Acurácia: 84.29%



SVM - Acurácia: 54.76%



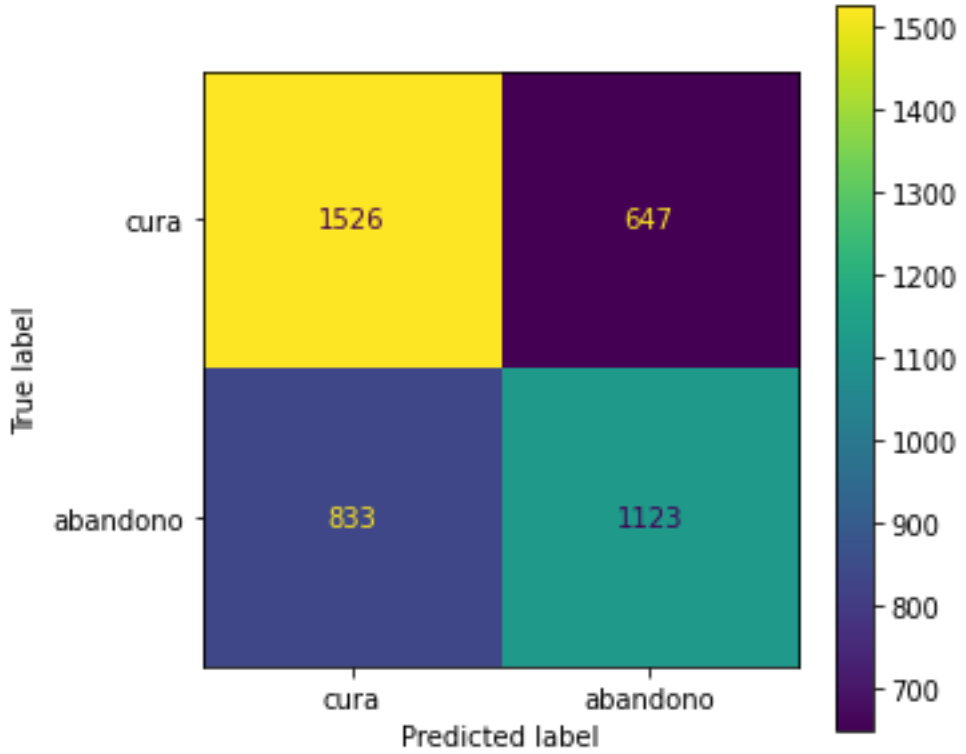
Técnicas Utilizadas

- Ensemble
- Grid Search
- Cross Validation
- Encoder
- Normalization
- Feature importances
- Metrics
  - Accuracy, Precision, Recall, F1, ROC AUC
- Imbalanced
  - Under, Over, SMOTE

Resultados alcançados

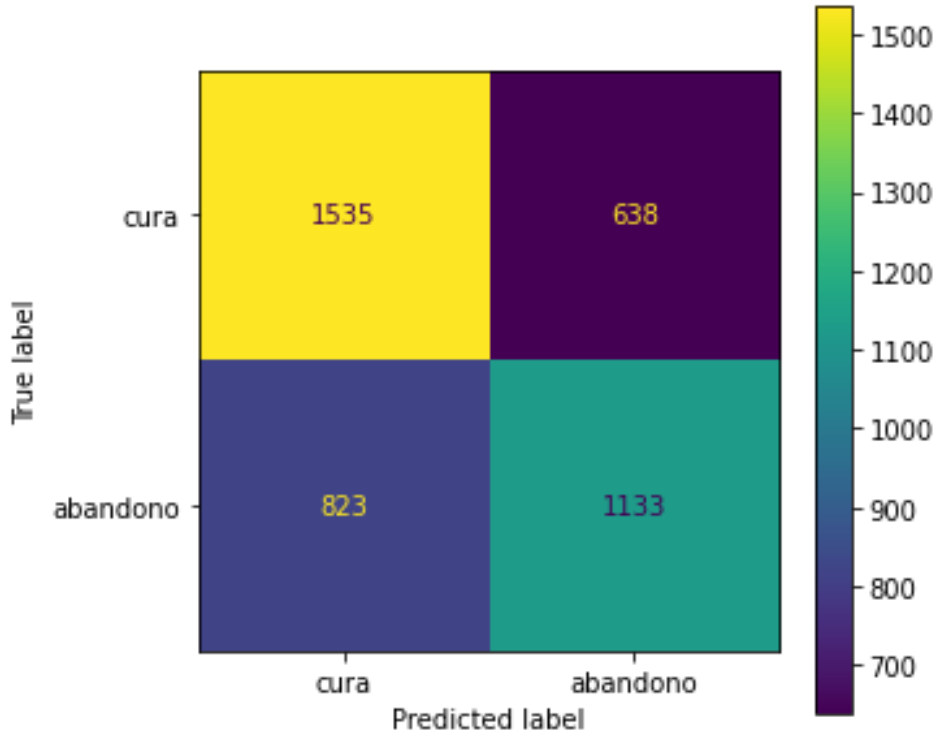
Random Forest

```
Accuracy Score : 0.6415596996851538
Precision Score: 0.6410001682523802
Recall Score   : 0.6415596996851538
F1 Score       : 0.6399685828492012
ROC AUC        : 0.6381929132116879
```

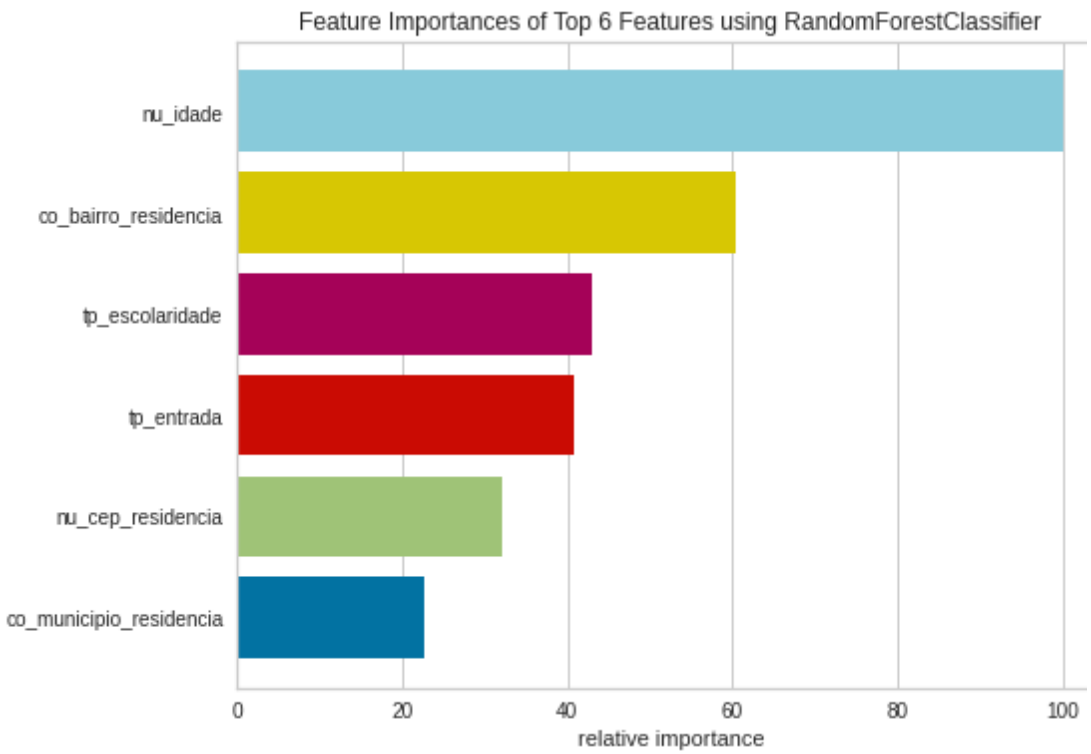


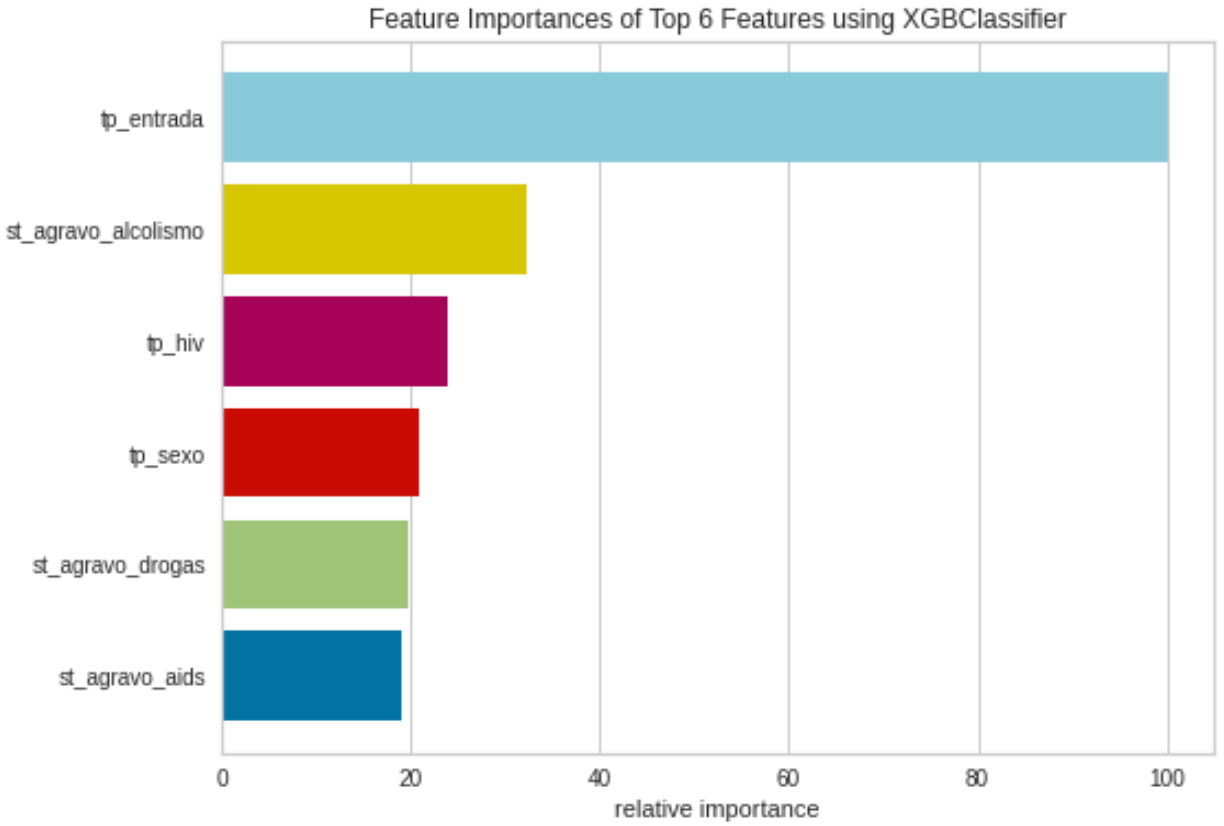
XGBoost

```
Accuracy Score : 0.6461612981351417
Precision Score: 0.6456584131105783
Recall Score   : 0.6461612981351417
F1 Score       : 0.644603003198162
ROC AUC        : 0.6428200201958033
```

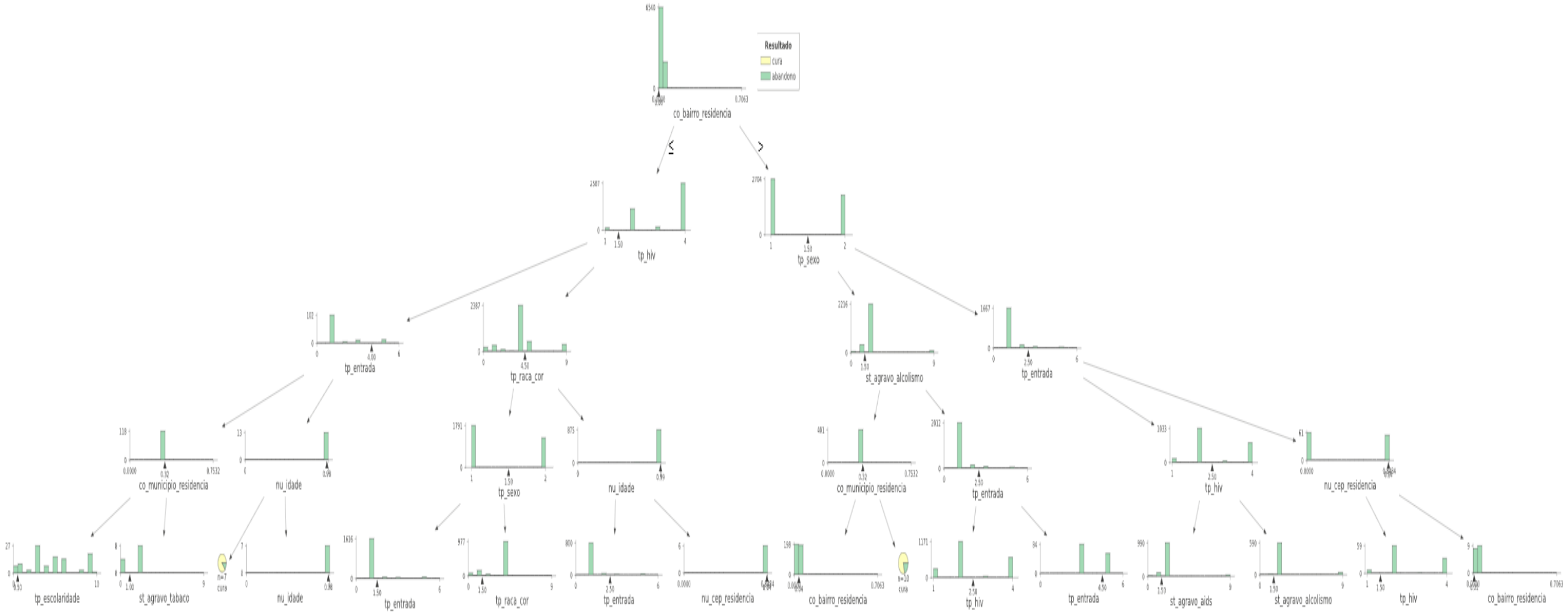


Feature importances (RandomForest e XGBoost)

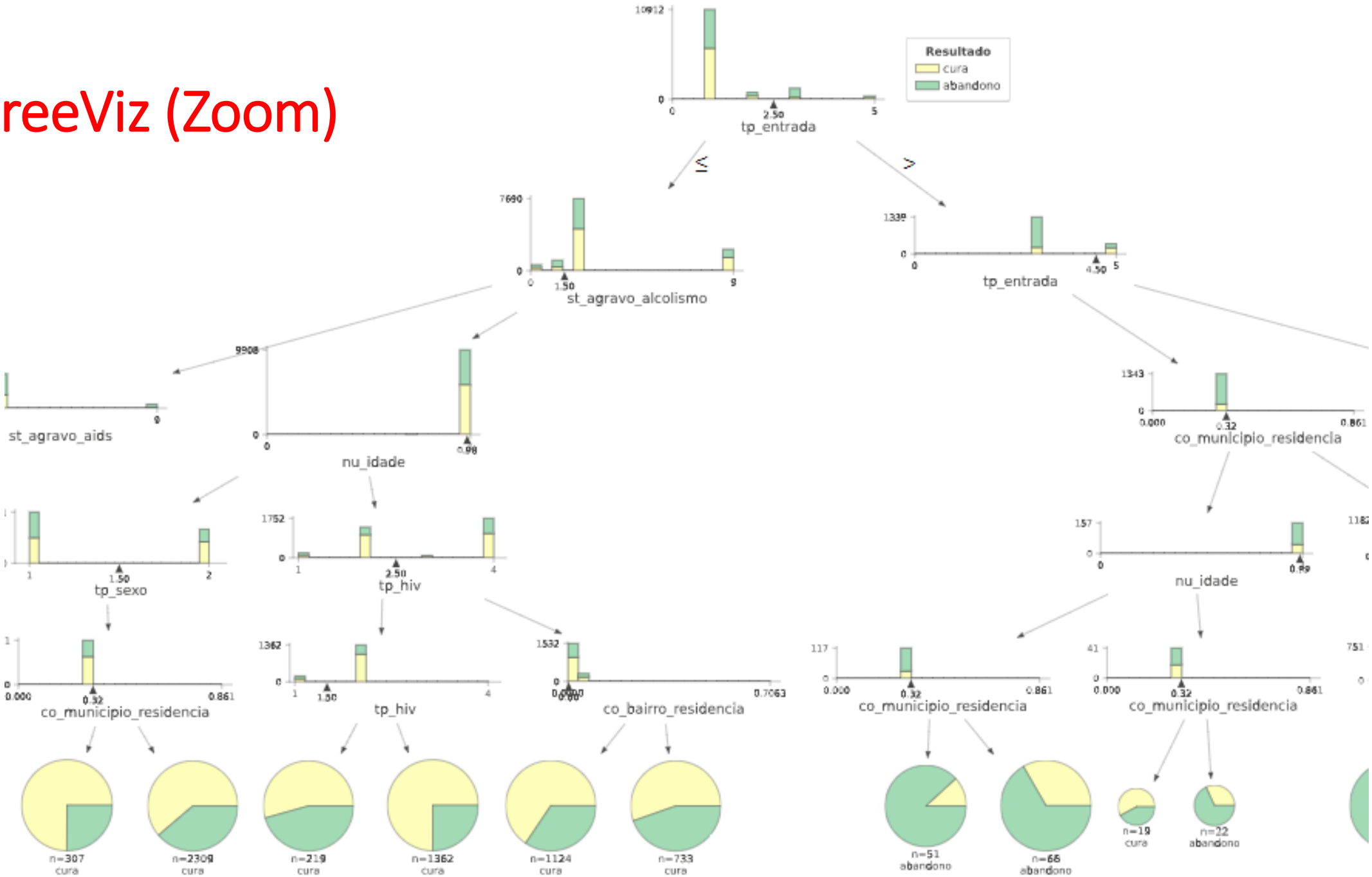




DtreeViz


Resultado
cura
abandono
co_bairro_residencia
≤
>
tp_hiv
tp_sexo
tp_entrada
tp_raca_cor
st_agravo_alcolismo
tp_entrada
co_municipio_residencia
nu_idade
tp_sexo
nu_idade
co_municipio_residencia
tp_entrada
tp_hiv
nu_cep_residencia
tp_escolaridade
st_agravo_tabaco
nu_idade
tp_entrada
tp_raca_cor
tp_entrada
nu_cep_residencia
co_bairro_residencia
tp_hiv
tp_entrada
st_agravo_aids
st_agravo_alcolismo
tp_hiv
co_bairro_residencia

DtreeViz (Zoom)



Conclusão – “Pipeline”

Pré-processamento

Undersampling

Todos os dados

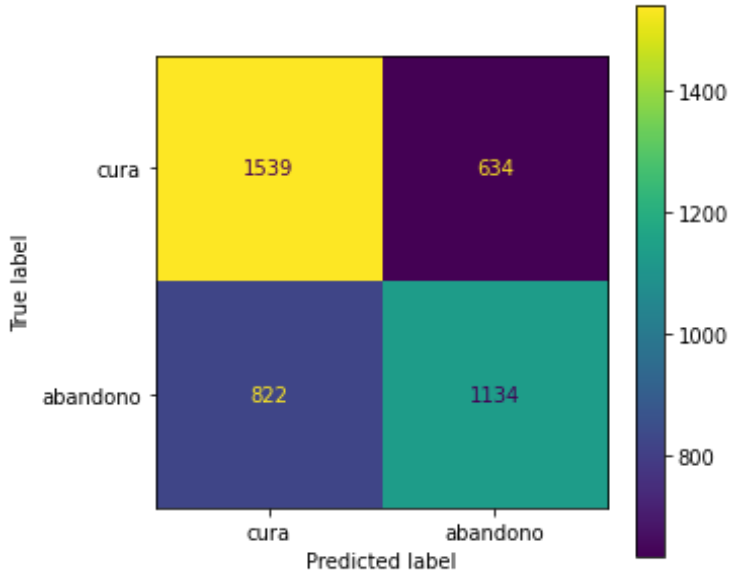
Treinamento

Grid Search

(Embaraçosamente Paralelo)

Ensemble

Voting Classifier



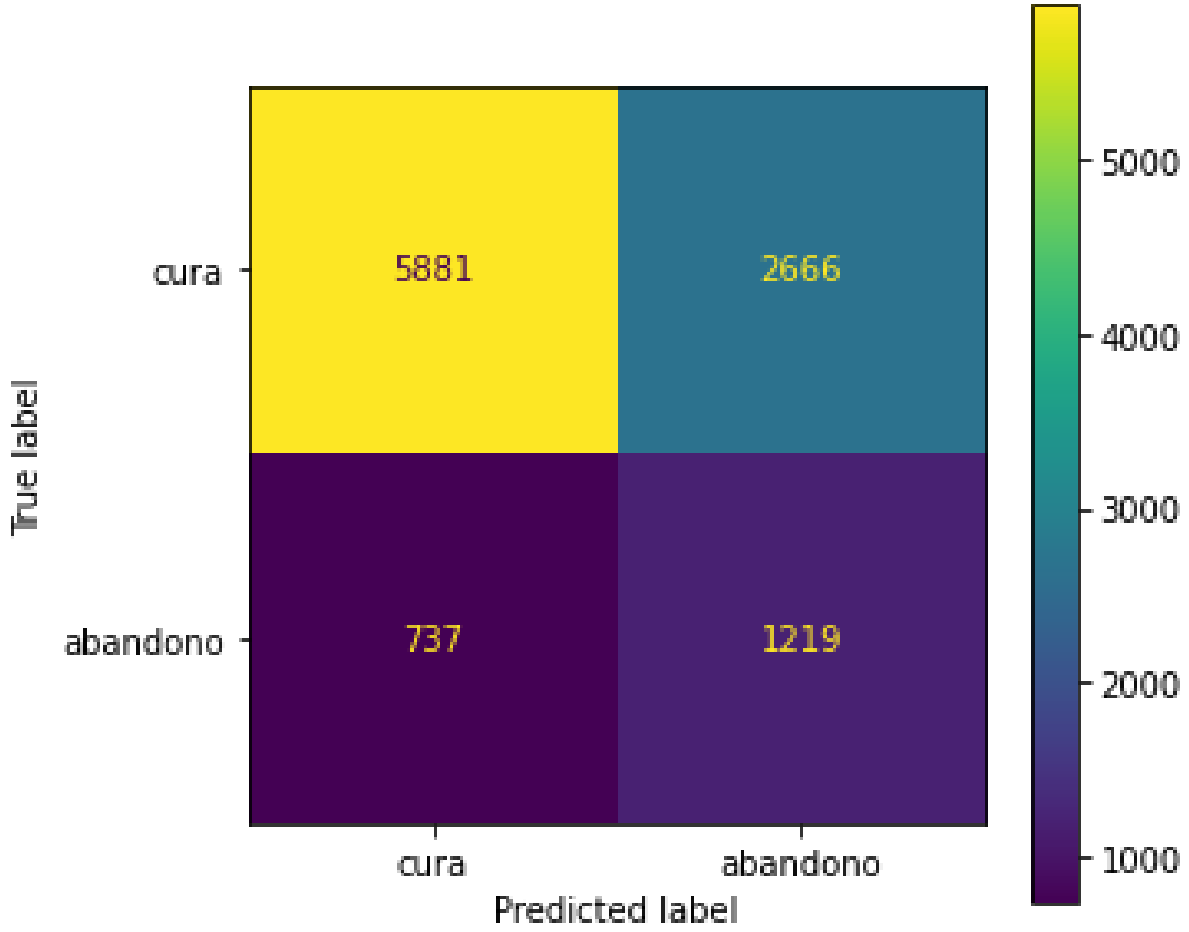

```
Accuracy Score : 0.6473722450956648
Precision Score: 0.6468969076662938
Recall Score   : 0.6473722450956648
F1 Score       : 0.645782094792074
ROC AUC        : 0.6439960304800408
```

Melhor resultado

Random Forest

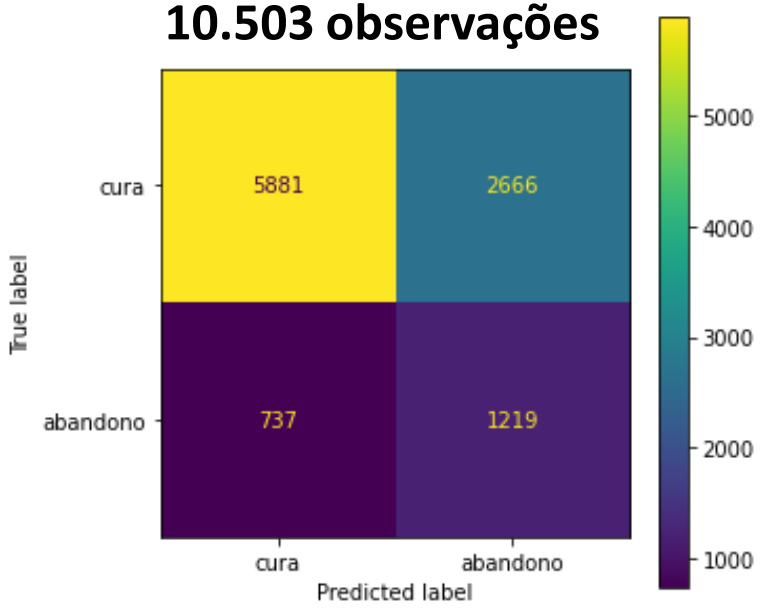
Accuracy Score	67.60%
Precision Score	78.16%
Recall Score	67.60%
F1 Score	70.89%
ROC AUC	65.56%

- class_weight
  - Bug no VotingClassifier



Considerações finais

10.503 observações



Necessidade de melhoria da qualidade dos registros (CEP, bairro, município)

Utilizando o modelo atual:

- 56% pacientes não precisarão de TDO e não abandonarão
- 62% de abandono não ocorrerá, pois estarão no TDO

Próximos passos:

- relacionamento com outras bases para completar os dados de localização da residência
- possibilidade de incluir novas variáveis independentes
- implementar como ferramenta que auxilia na seleção dos pacientes que receberão Tratamento Diretamente Observado