DHDS - TURMA 6
PROJETO INTEGRADOR: COVID-19

CÁLCULO DO RISCO DE ÓBITO POR COVID-19 (%) FRENTE A DIVERSOS PARÂMETROS PESSOAIS PRÉ-EXISTENTES.

Fernando Rodrigues
Paulo Assunção
João Pacher

Outubro/2020

OBJETIVO

- Risco individual de óbito (%) ao se contaminar com o vírus SARS2 da COVID19.

- Abordagem alternativa ao isolamento horizontal adotado no Brasil pelo vertical que implica manter apenas as pessoas dos grupos de riscos isolados.

ABORDAGEM

MODELAGEM DE CLASSIFICAÇÃO COM CÁLCULO DE PROBABILIDADE

- Dataset público SIVEP-Gripe
(Sistema de Informação de Vigilância Epidemiológica da Gripe) .

```
raw_df.shape
```

```
(291130, 134)
```

```
# número de casos covid
raw_df['PCR_SARS2'].value_counts(dropna=False)
```

```
NaN    193206
1.0     97924
Name: PCR_SARS2, dtype: int64
```

```
# Guardando apenas as features de interesse e disponíveis

df = df[['EVOLUCAO', 'DT_NOTIFIC', 'SG_UF_NOT', 'NU_IDADE_N', 'CS_SEXO', 'CS_RACA',
         'CS_ESCOL_N', 'CS_GESTANT', 'PUERPERA', 'CARDIOPATI', 'HEMATOLOGI', 'SIND_DOWN','HEPATICA','ASMA',
         'DIABETES','NEUROLOGIC','PNEUMOPATI', 'IMUNODEPRE','RENAL','OBESIDADE', 'OUT_MORBI']]

df.shape
```

```
(43146, 21)
```

291.130
Observações

97.924
Casos de Covid-19

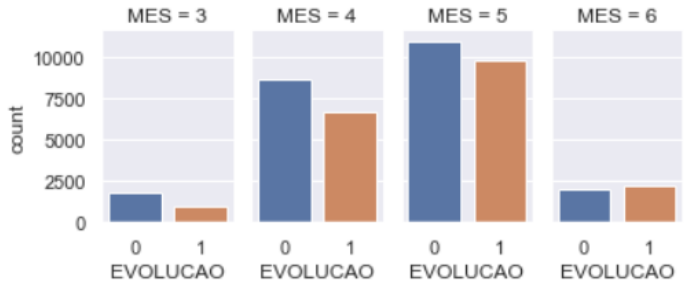
43.146
Casos com Dados Correlacionáveis

1
Target [‘EVOLUCAO’]

133
Atributos no dataset

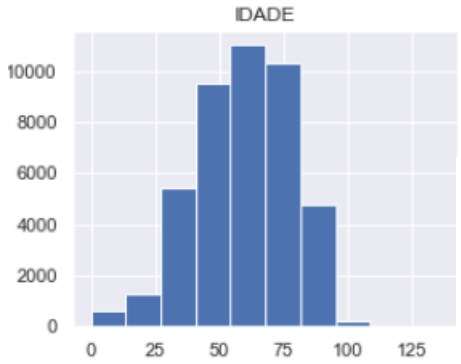
20
Atributos de interesse (“leakage”)

ETAPAS

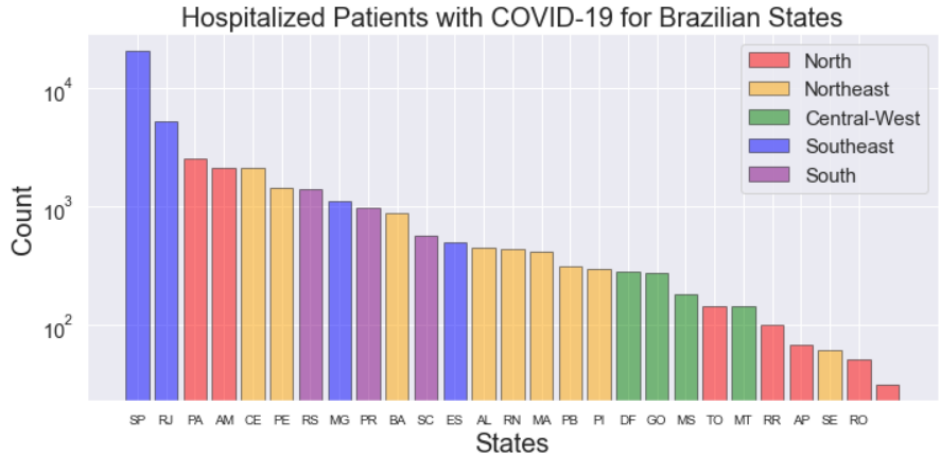


ATRIBUTOS:
Idade
UF
Sexo
Raça
Escolaridade
Comorbidades

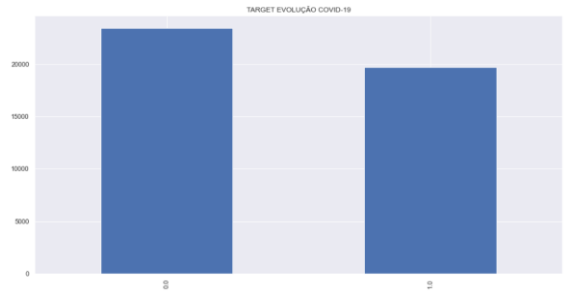
- Cleaning, Wrangling e Feature Engineering
  - Tratamento dos Nulos,
  - Agrupamentos,
  - Inputações,
  - Alteração de Data Types



- EDA
  - Histogramas,
  - Dados dos atributos em relação ao target,
  - Codificação de variáveis categóricas

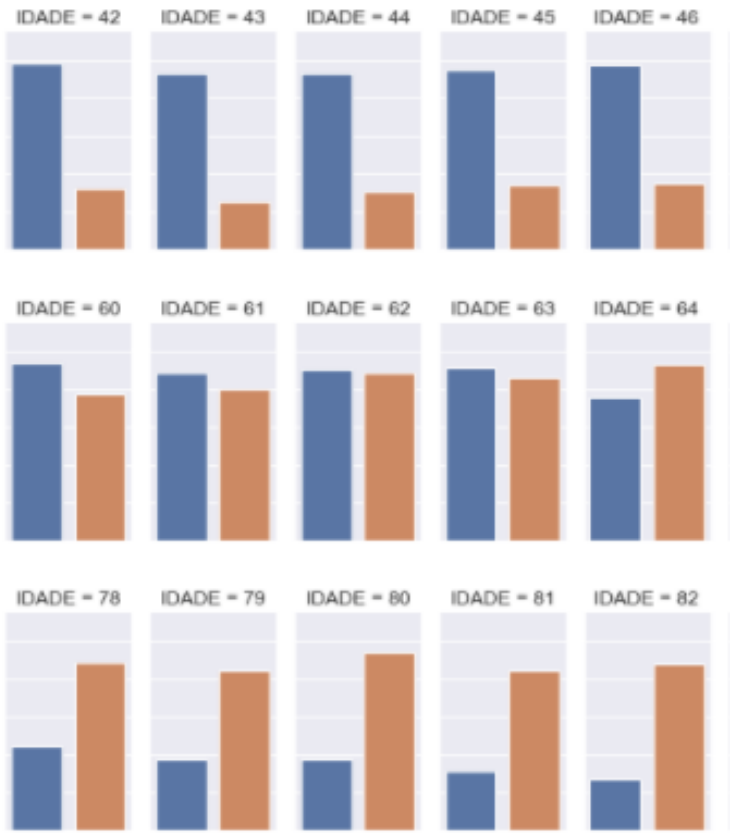


- Pré-processamento
  - Redução de dimensionalidade (PCA, Feature Selection com o XGBClassifier, ExtraTreesClassifier, PPS, Sumário estatístico LogReg)
  - Separação em subconjuntos Treino/Teste
  - Declaração das variáveis categóricas

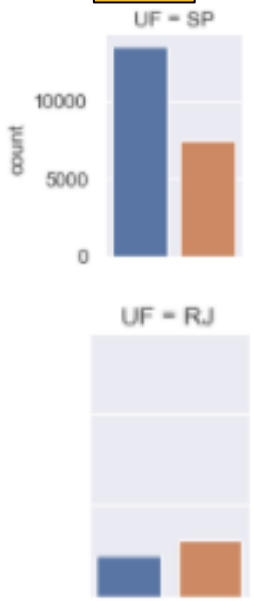


ATRIBUTOS

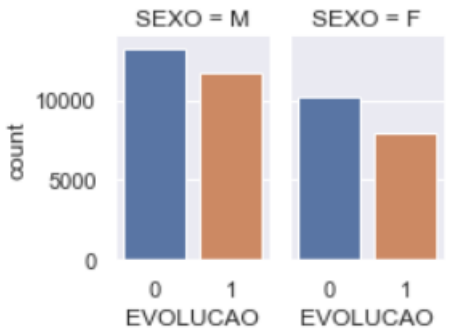
Idade



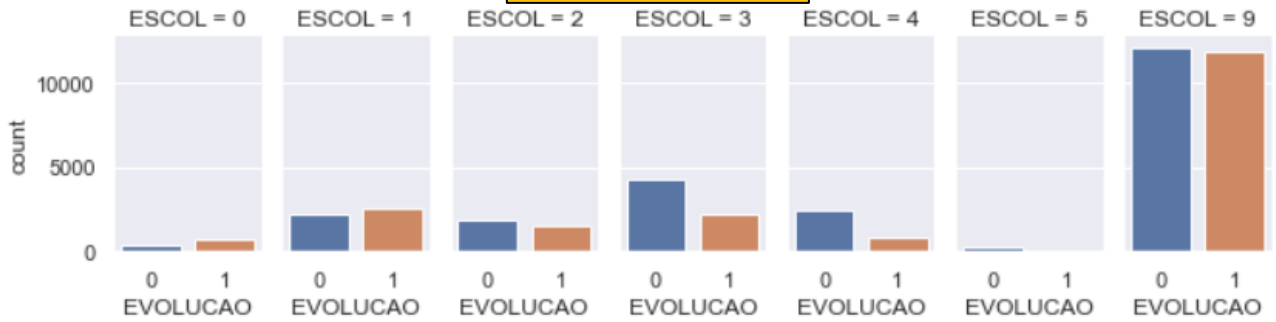
UF



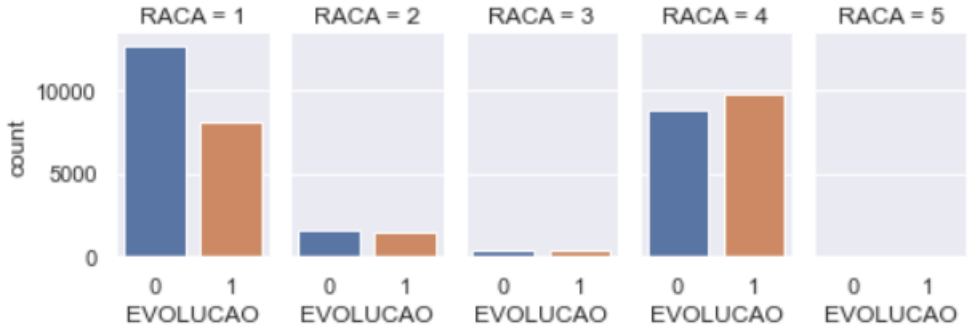
Sexo



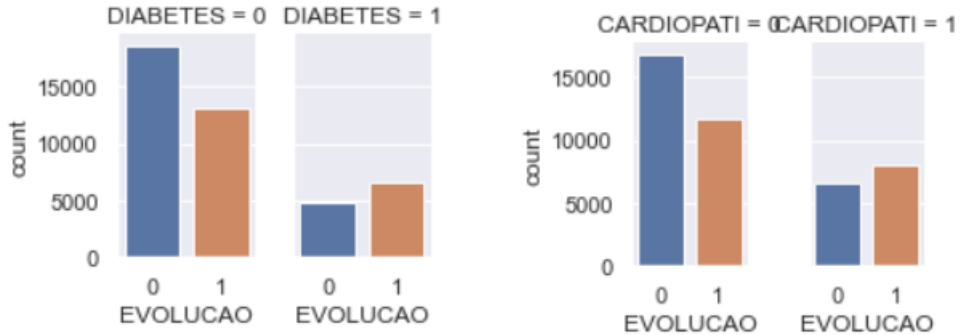
Escolaridade



Raça



Comorbidades



FORTE: IDADE
MÉDIA: UF, SEXO, ALGUMAS COMORBIDADES
FRACA: DEMAIS EM GERAL

RESULTADOS

- Modelos de Classificação
  - Sem otimização
  - Com otimização
  - Ensemble

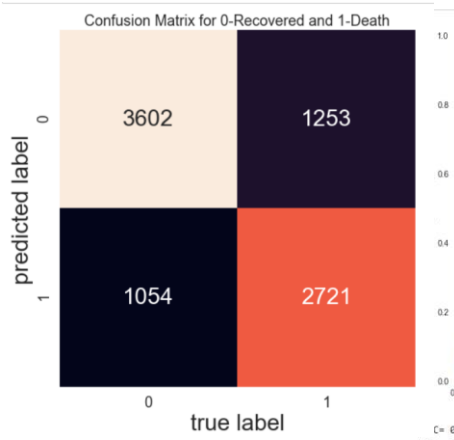
- Pickling

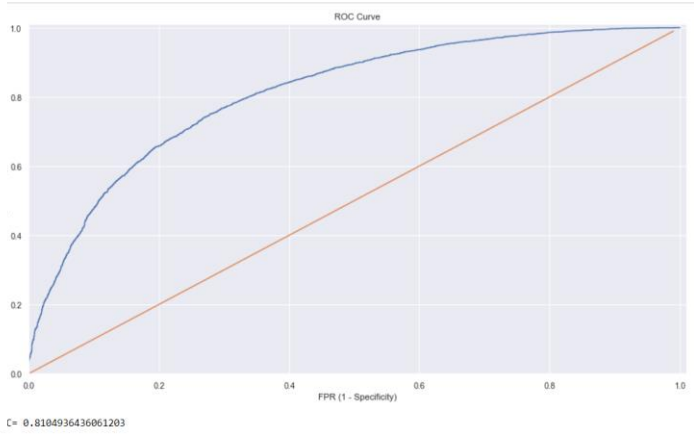
	Trainning_Accuracy	Validation_Accuracy
Logistic	0.725055	0.733372
BernoulliNB	0.669545	0.674739
RandomForest	0.914735	0.682387
Decision Tree	0.914735	0.649131
GradientBoost	0.729343	0.728042
LightGBM	0.741482	0.731981
XGBoost	0.755041	0.730359
AdaBoost	0.728387	0.729316
SVC	0.703790	0.706373
CatBoost	0.754288	0.732445

```
Accuracy: 0.73 (+/- 0.00) [Voting_Classifier_Hard]
Recall: 0.70 (+/- 0.01) [Voting_Classifier_Hard]
Accuracy: 0.73 (+/- 0.00) [Voting_Classifier_Soft]
Recall: 0.70 (+/- 0.01) [Voting_Classifier_Soft]
```

PROBA_THRESHOLD: 50%
ACCURACY: 73%
RECALL: 70%

PROBA_THRESHOLD: 20%
ACCURACY: 64%
RECALL: 95%





RISCO ALTO	RISCO MÉDIO	RISCO BAIXO
PROBA > 50%	PROBA 20 A 50%	PROBA < 20%

NEXT STEPS / MELHORIAS

- Usar outras bibliotecas para codificar as variáveis categóricas:
https://contrib.scikit-learn.org/category_encoders/ e/ou
https://feature-engine.readthedocs.io/en/latest/index.html ;

- Estudar redução de dimensionalidade com o Features Importance do CatBoost;

- Rodar o CatBoost com os atributos categóricos antes de serem dummiezados;

- Aumentar de 3 modelos para Fine tunning entre os 10 prospectados inicialmente;

- Aplicar Deep Learning;

- Buscar um dataset com mais atributos que possam explicar a previsão.