Mestrado Acadêmico em Bioestatística

Qualificação

Hevans Vinicius Pereira

DEPARTAMENTO DE ESTATÍSTICA

UNIVERSIDADE ESTADUAL DE MARINGÁ – UEM

31/08/2023

Descrição do Problema

Descrição dos Dados

Análise Bidimensional

Modelos

Descrição do Problema

Desde o seu surgimento em dezembro de 2019 (HUANG C.; WANG, 2020), o novo coronavírus se espalhou rapidamente. Em março de 2020, o governo federal decretou estado de emergência em saúde pública em razão da pandemia de COVID-19.

Neste trabalho iremos utilizar dados públicos, obtidos no Open Data SUS, sobre Síndrome Respiratória Aguda Grave e COVID para criar modelos de classificação que nos permitam estimar quais pacientes tem mais chance de vir a óbito com base em diversas características coletadas.

Para isso, consideramos apenas os casos de pacientes adultos, isto é, com 18 anos ou mais, hospitalizados por Covid-19 e que foram notificados no ano de 2022. O questionário que monitora a Síndrome Respiratória Aguda Grave já existia antes da COVID, mas foi alterado em virtude desta.

Descrição do Problema

Os modelos serão criados usandos técnicas de aprendizado supervisionado com o objetivo de predizer o desfecho de um paciente internado com COVID-19.

Além do modelo preditivo em si, haverá um ganho de informação se conseguirmos extrair do modelo as informações referentes à importância das variáveis, a fim de identificar os fatores que mais influenciam no desfecho dos pacientes.

Descrição dos Dados

- O dataset contém 556445 linhas e 173 colunas;
- Vamos usar apenas dados referentes ao ano de 2022;
- Vamos usar apenas as colunas que julgamos relevantes: sexo, idade, raça, escolaridade, região, sintomas e fatores associados;
- Ficamos com 40 colunas.

Tratamento de Dados

- Vamos trabalhar apenas com pacientes hospitalizados;
- Vamos trabalhar apenas com pacientes com idade maior ou igual a 18 anos;
- Vamos trabalhar apenas com pacientes que tiveram COVID-19;
- Vamos substituir dados faltantes e as observações não registradas (9 - Ignorado) para todas as variáveis de fatores associados pelo valor 0 (zero), considerando que não foi registrado porque não havia fator presente;
- No conjunto de dados, nas variáveis binárias o valor 2 indica 'Não', trocaremos para 0 (zero).

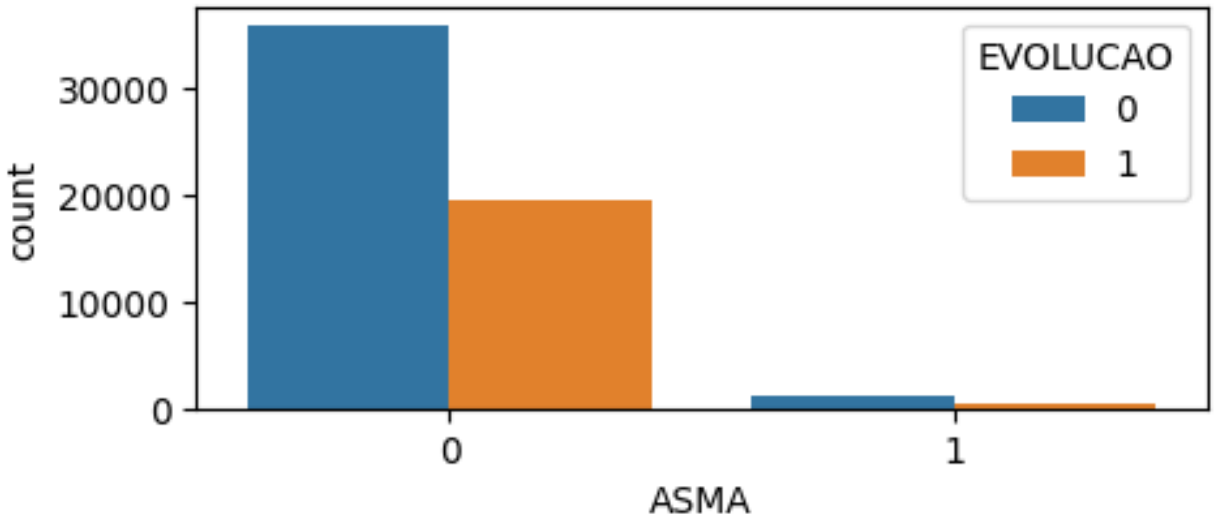
Tratamento de Dados

- Não temos como imputar valores para as variáveis de escolaridade, raça, zona urbana/rural e vacina então vamos trocar ‘Ignorado‘ por valor faltante;
- A variável alvo é a ‘EVOLUCAO‘, então vamos excluir dados faltante ou preenchidos com ‘9 - Ignorado‘;
- Na variável ‘EVOLUCAO‘ as observações ‘2-Óbito‘ e ‘3-Óbito por outras causas‘ serão juntadas na mesma categorias;
- Vamos juntar as observações ‘1-Sim, invasivo‘ e ‘2-Sim, não invasivo‘ na mesma categoria;
- Iremos trocar a unidade federativa pela região;

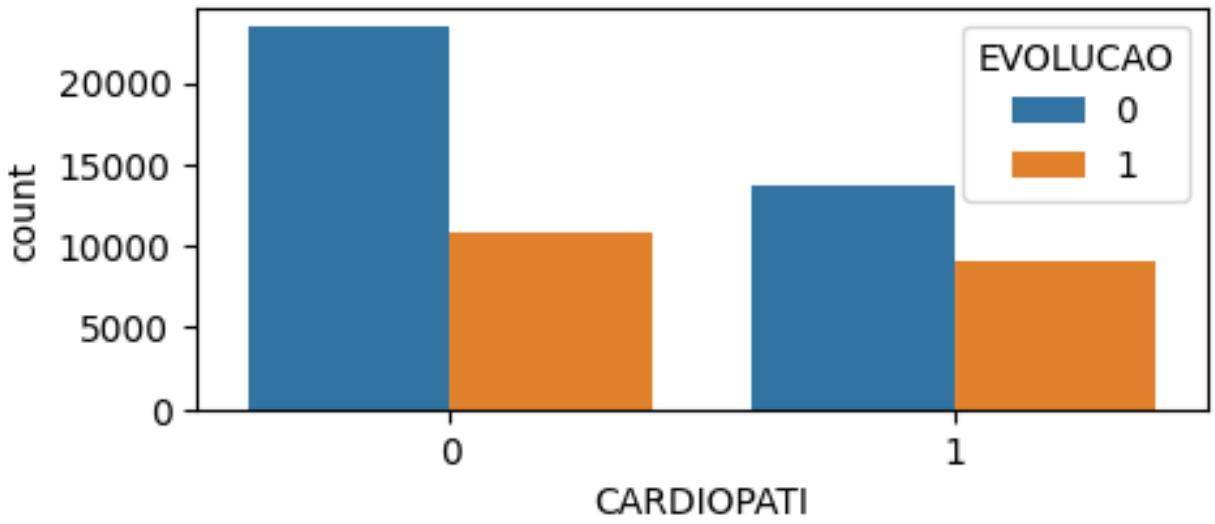
Tratamento de Dados

- Com os tratamentos aplicados até o momento não sobrou nenhuma observação na classe ‘2-Rural‘ para a variável ‘CS_ZONA‘, portanto vamos descartá-la.
- Na variável raça, vamos juntar as classes ‘Preta‘ e ‘Parda‘;
- Vamos descartar todas as observações que contenham dados faltantes;
- Isso nos deixa com 57016 linhas e 36 colunas;
- As variáveis restantes não apresentam correlação (de Spearman) relevante;

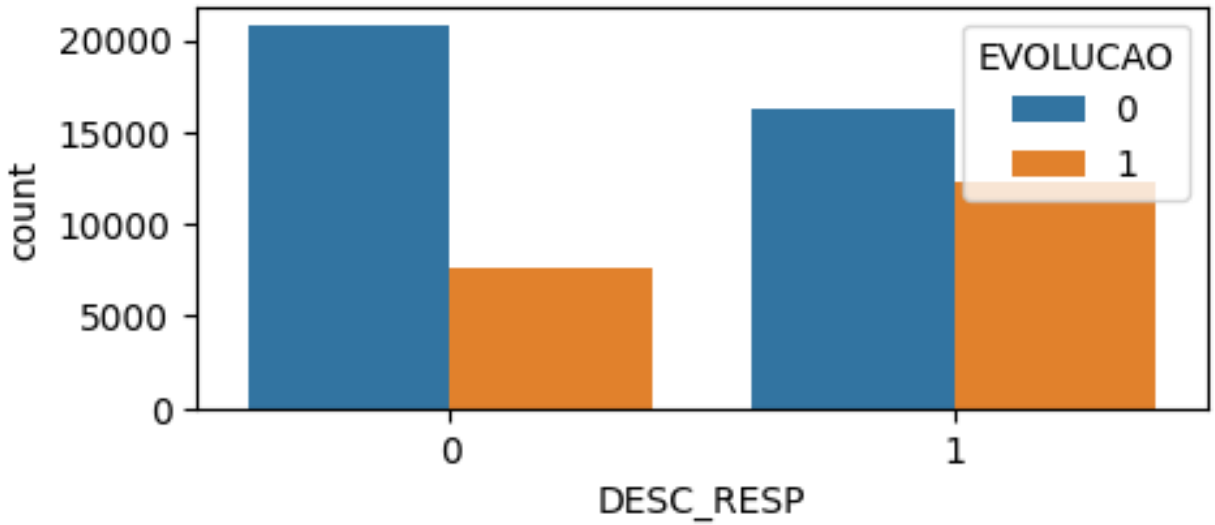
Asma



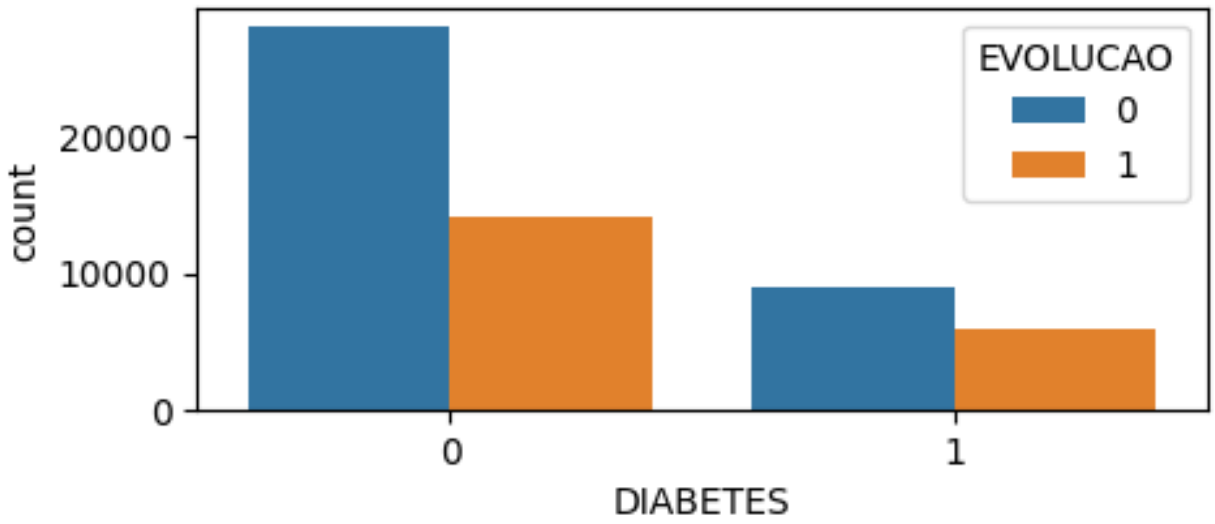
Cardiopatia



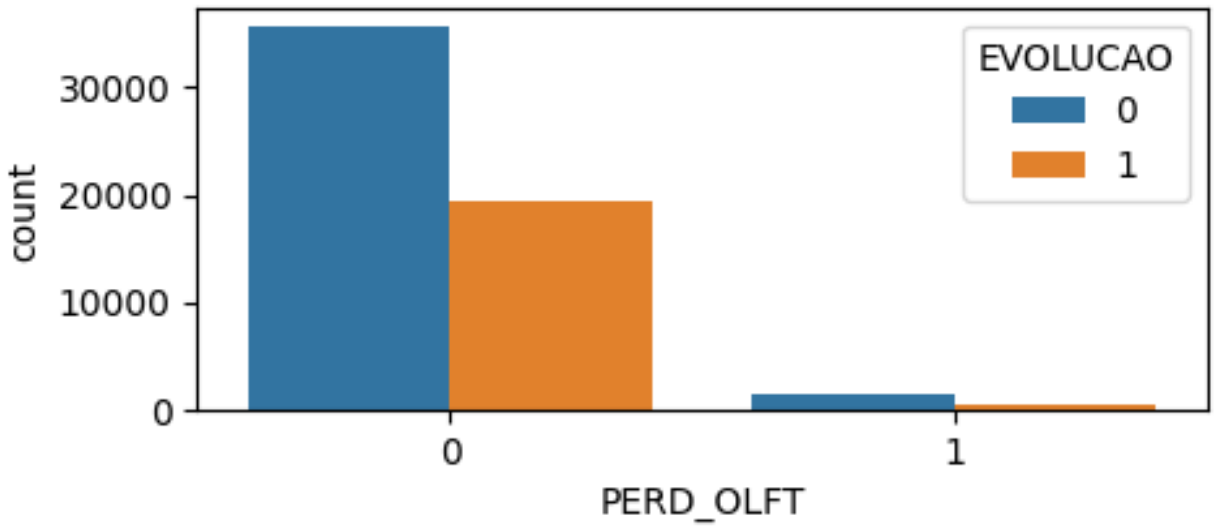
Desconforto Respiratório



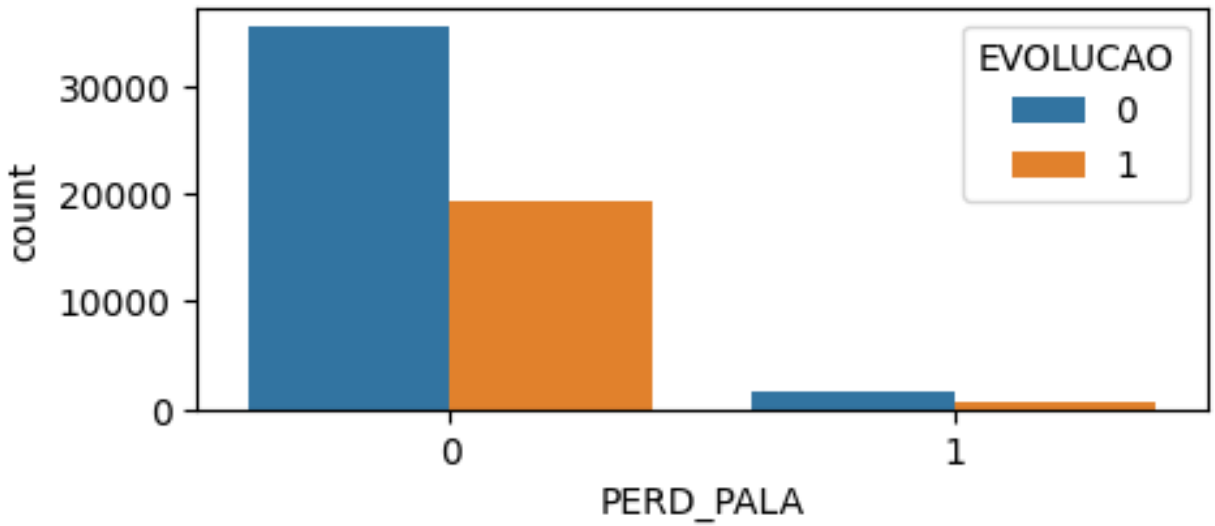
Diabetes



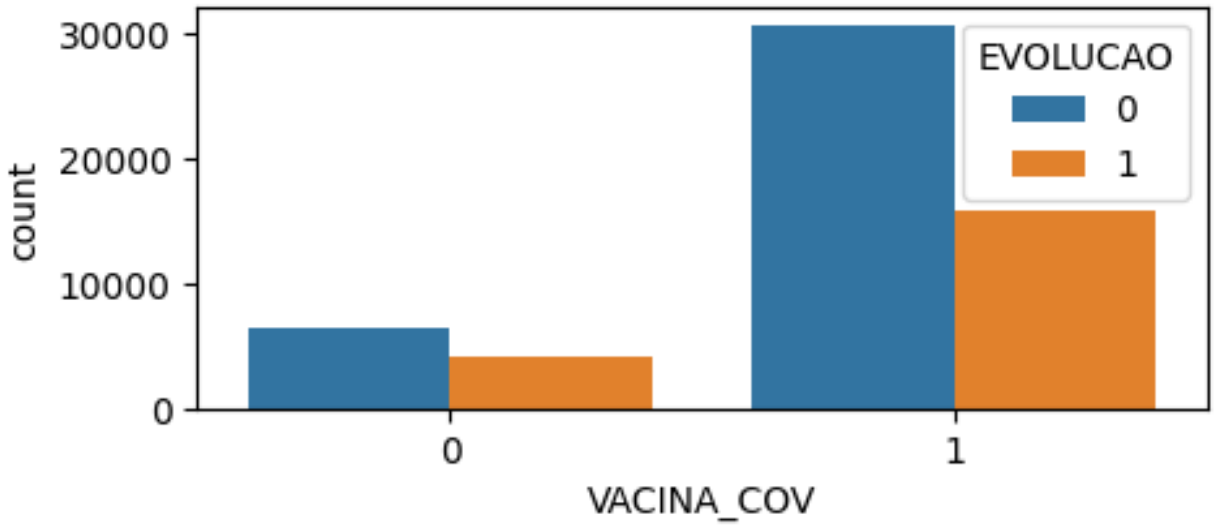
Perda de Olfato



Perda de Paladar



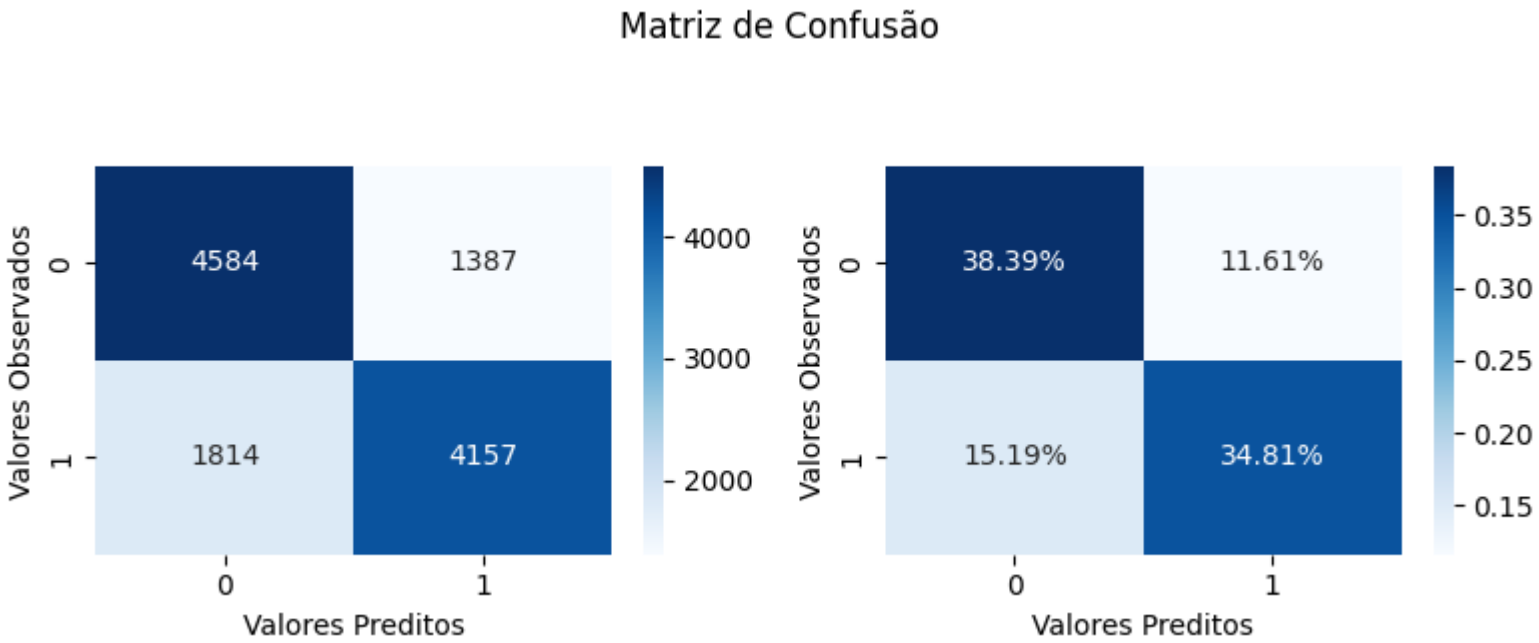
Vacina



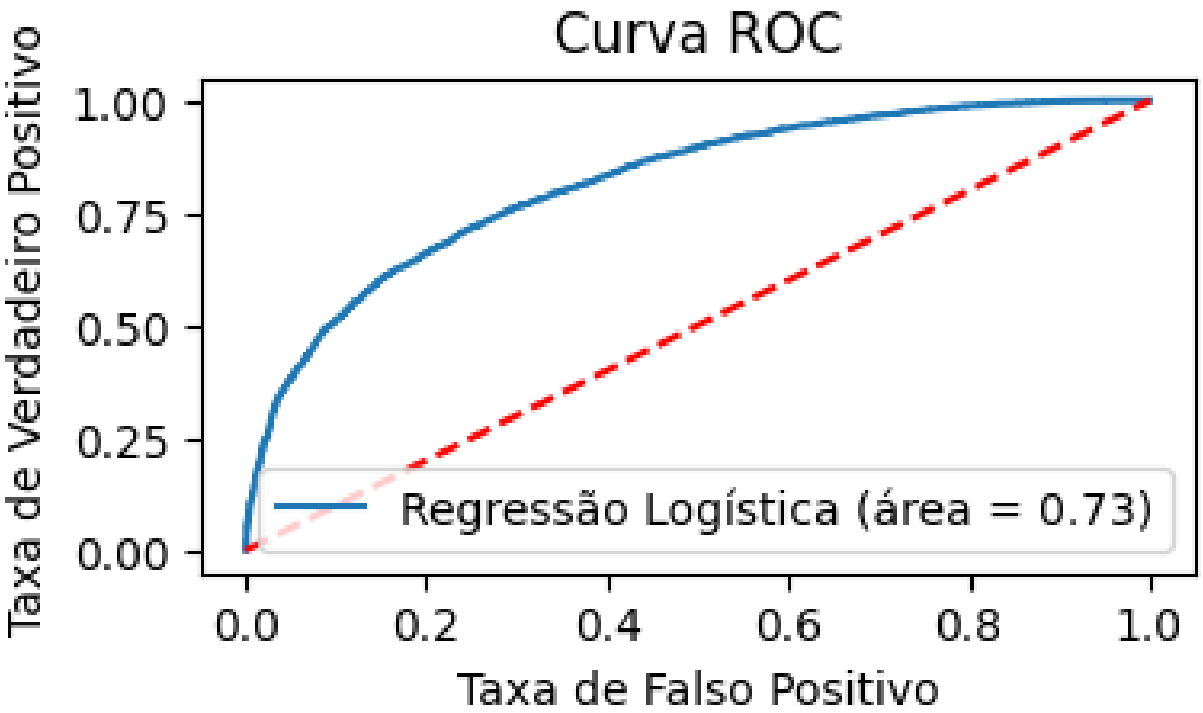
Modelos

- Como a variável EVOLUCAO está desbalanceada, vamos usar a técnica chamada de under sampling para fazer o balanceamento;
- Vamos separar 30% dos dados para teste;
- Vamos usar Regressão Logística, Floresta Aleatória e Rede Neural;
- Foi usado validação cruzada com 5 folds;
- Foi feita otimização de hiperparâmetros com RandomizedSearchCV ou Optuna;
- As variáveis categóricas foram transformadas em variáveis dummy.

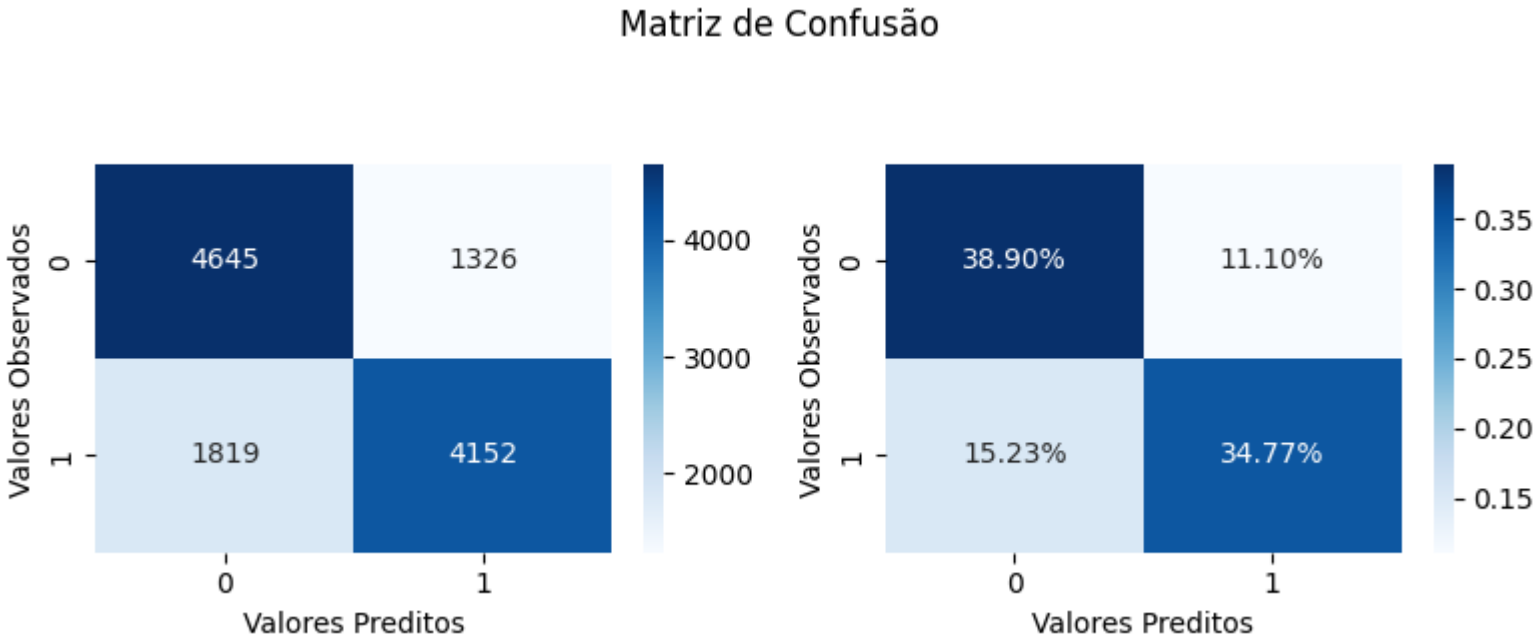
Regressão Logística



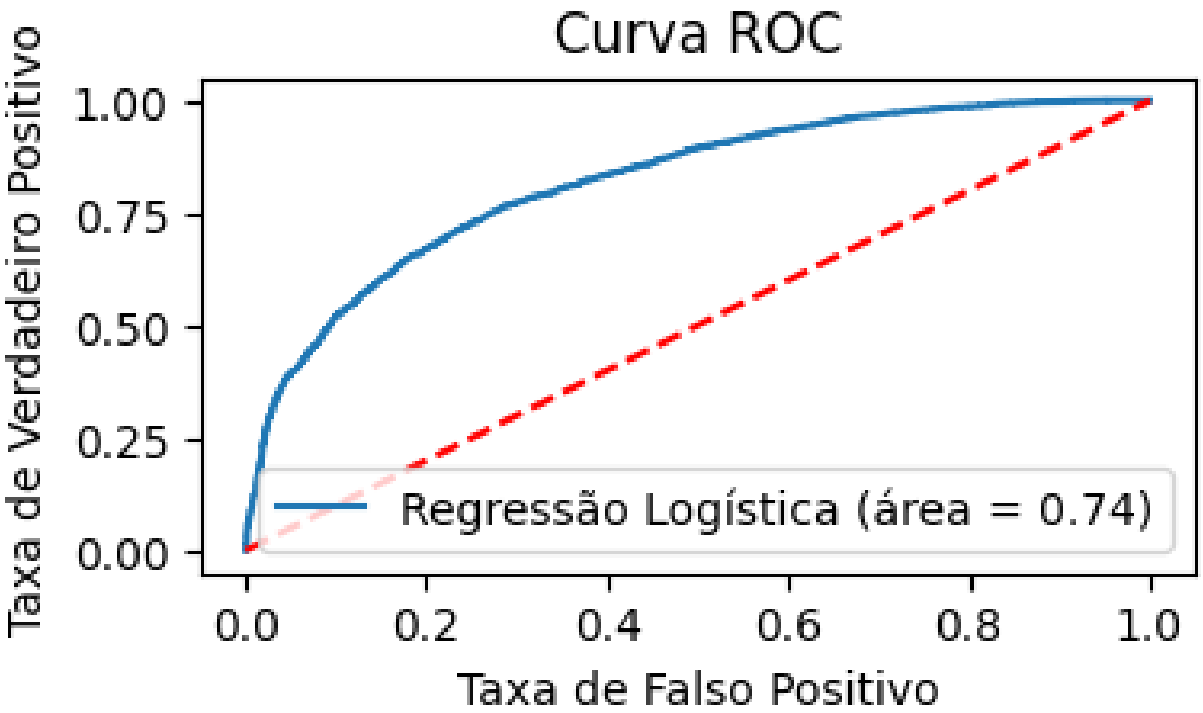
Regressão Logística



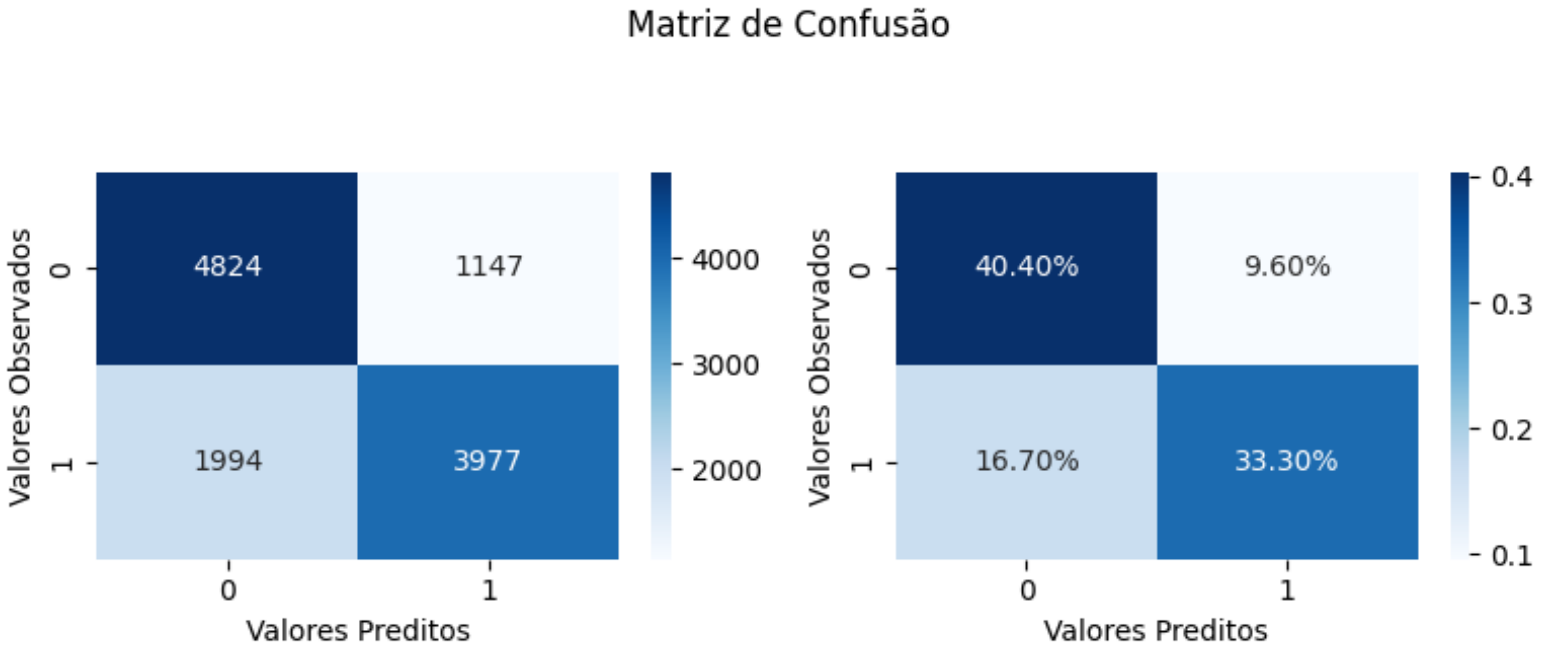
Floresta Aleatória



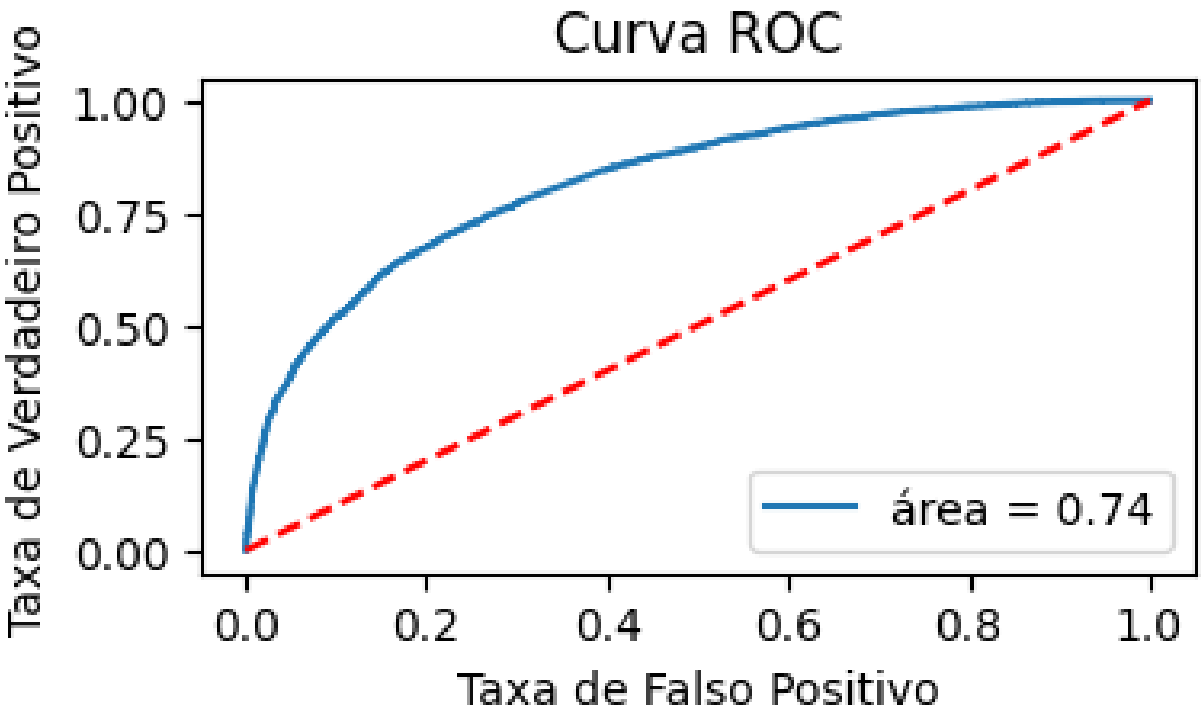
Floresta Aleatória



Rede Neural


Matriz de Confusão
Valores Observados
0
1
4824
1147
1994
3977
4000
3000
2000
0
1
Valores Preditos
Valores Observados
0
1
40.40%
9.60%
16.70%
33.30%
0.4
0.3
0.2
0.1
0
1
Valores Preditos

Rede Neural


Curva ROC
Taxa de Verdadeiro Positivo
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0
Taxa de Falso Positivo
área = 0.74

Comparando Modelos

Tabela: Métricas de Avaliação dos Modelos

Modelo	Acurácia	Recall	Precision	f1 score	ROC-AUC
Regressão Logística	0,73	0,70	0,75	0,72	0,73
Floresta Aleatória	0,74	0,69	0,76	0,72	0,74
Rede Neural	0,74	0,73	0,74	0,73	0,74

Para a sequência deste trabalho, consideraremos os seguintes pontos:
- mudar a categorização de algumas covariáveis (idade, uso de suporte ventilatório, entre outros) a fim de obter um maior poder preditivo;
- avaliar o efeito dos fatores associados na probabilidade de óbito dos pacientes;
- utilizar outras técnicas de aprendizado supervisionado.

Obrigado!