



Predição da sobrevivência dos pacientes

Laura Damaceno de Almeida

Objetivo

Construir um modelo capaz de **prever a sobrevivência dos pacientes e ranquear as características clínicas** (riscos fatores) mais importantes incluídas nos prontuários médicos que podem indicar a insuficiência cardíaca.





Pipeline utilizada para o desafio



Variáveis disponíveis

Qualitativas Nominal:

- Gender (Gênero)
- Smoking (Fumante)
- Diabetes (Diabético)
- BP (se tem hipertensão)
- Anaemia (se tem anemia)
- Event (Sobreviveu)

Quantitativas Contínua:

- Sodium (Nível de sódio no sangue)
- Creatinine (Nível de creatinina no sangue)
- Pletelets (Plaquetas no sangue)

Quantitativas Discreta:

- Age (Idade)
- TIME (Tempo em observação)
- CPK (Nível da enzima CPK no sangue)
- Ejection.Fraction (Porcentagem de sangue que sai do coração em cada contração)



Qualidade dos dados

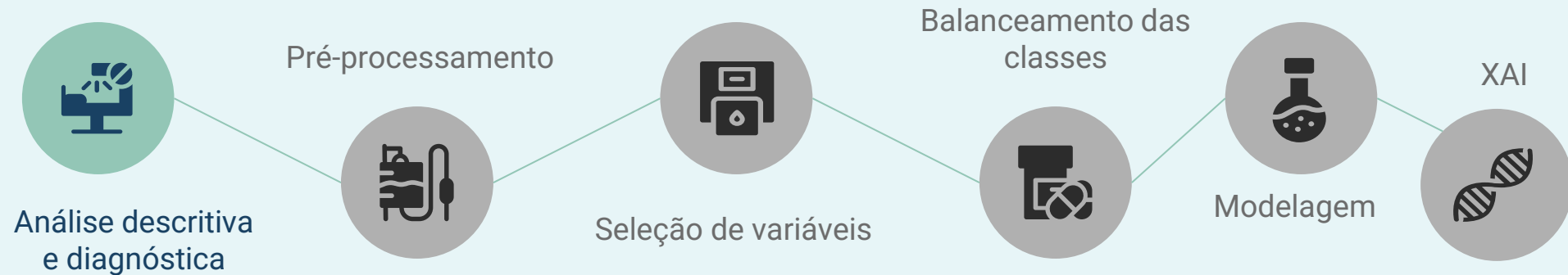
	index	colunas	tipo	Qtde valores NaN	% valores NaN	valores únicos por feature
0	TIME	TIME	int64	0	0.0	148
1	Event	Event	int64	0	0.0	2
2	Gender	Gender	int64	0	0.0	2
3	Smoking	Smoking	int64	0	0.0	2
4	Diabetes	Diabetes	int64	0	0.0	2
5	BP	BP	int64	0	0.0	2
6	Anaemia	Anaemia	int64	0	0.0	2
7	Age	Age	float64	0	0.0	47
8	Ejection.Fraction	Ejection.Fraction	int64	0	0.0	17
9	Sodium	Sodium	int64	0	0.0	27
10	Creatinine	Creatinine	float64	0	0.0	40
11	Pletelets	Pletelets	float64	0	0.0	176
12	CPK	CPK	int64	0	0.0	208

Pode-se perceber que para o conjunto de dados disponibilizado **não há valores faltantes nas variáveis**.

Entretanto em relação a variável idade (Age) o tipo de dado está errado, pois o deveria ser inteiro (int) e não float, com isso foi realizado um tratamento para transformar o tipo dessa variável.



Pipeline utilizada para o desafio



01

Entendimento do perfil dos pacientes

Análise descritiva das variáveis



Análise estatística descritiva*



01

Idade

75% dos pacientes tem idade abaixo de 70 anos

03

Tempo em observação

75% dos pacientes ficaram até 203 dias em observação

02

CPK no sangue

75% dos pacientes durante o acompanhamento tiveram o nível de CPK em até 582

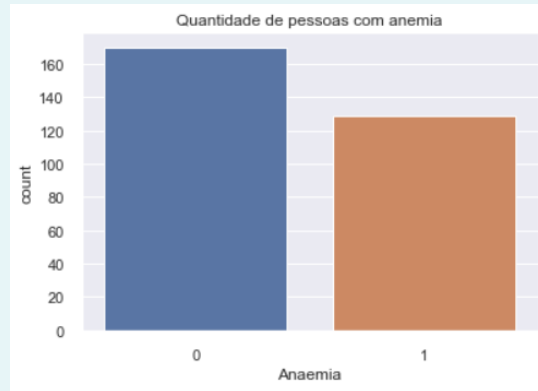
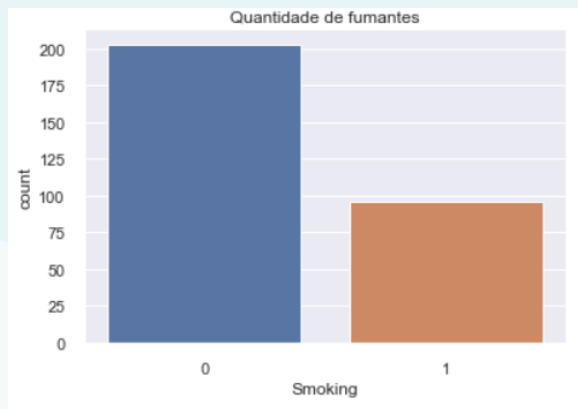
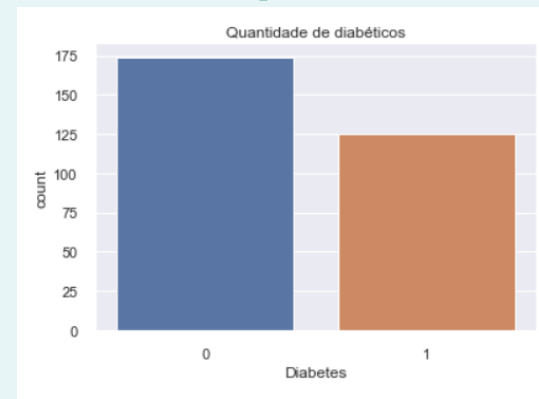
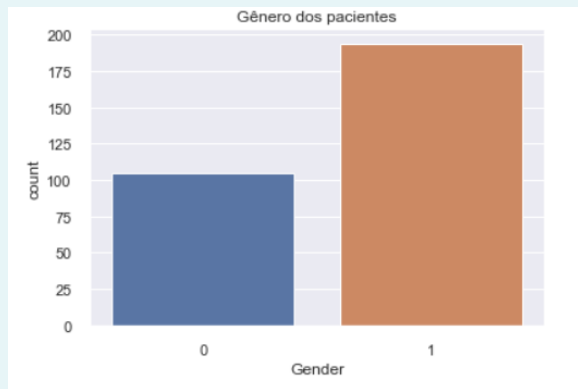
04

Saída do sangue

75% dos pacientes durante o acompanhamento tiveram até 45% da saída de sangue o coração a cada contração

* Tirado com o método describe() do pandas

Variáveis categóricas





Insights extraídos

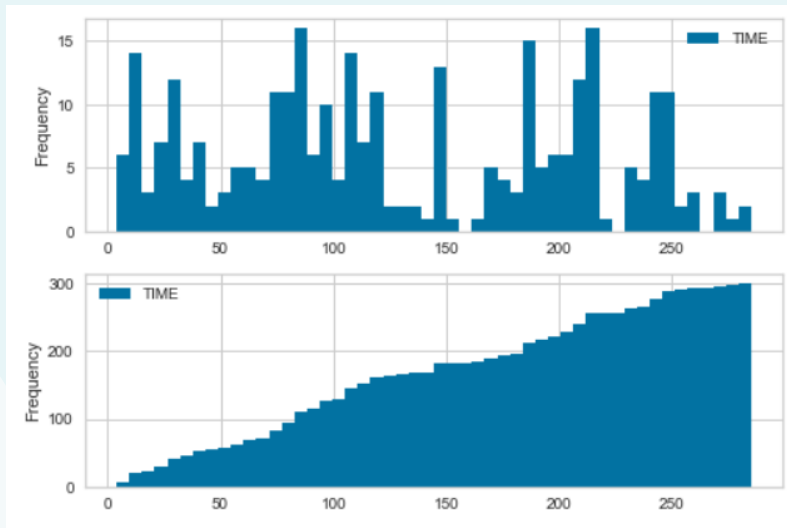
- 35.1% dos pacientes são do gênero feminino;
- 67.9% dos pacientes não fumam;
- 58.2% dos pacientes não são diabéticos;
- 64.9% dos pacientes não tem hipertensão;
- 56.9% dos pacientes não tem anemia;
- 67.9% dos pacientes sobreviveram;

Ponto de atenção: Há poucas amostras femininas no conjunto de dados e dependendo da forma como for passar como entrada para o modelo pode causar um viés de gênero nos resultados;

Além disso com a variável alvo é a Event (sobrevivência), entretanto está desbalanceada, necessitando de um tratamento antes de passar para o modelo, para evitar eviesamento do mesmo para a classe marjoritária.

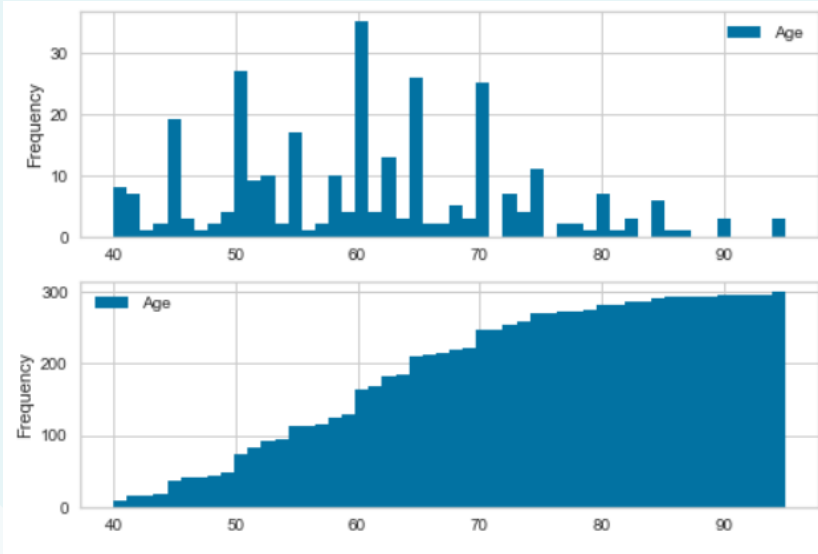
Variáveis quantitativas - Tempo

Para as variáveis numéricas, além de visualizar a frequência dos valores utilizando histograma e histograma acumulado calculei a assimetria das variáveis, pois ajuda a entender melhor a tendência dos valores de cada uma delas.



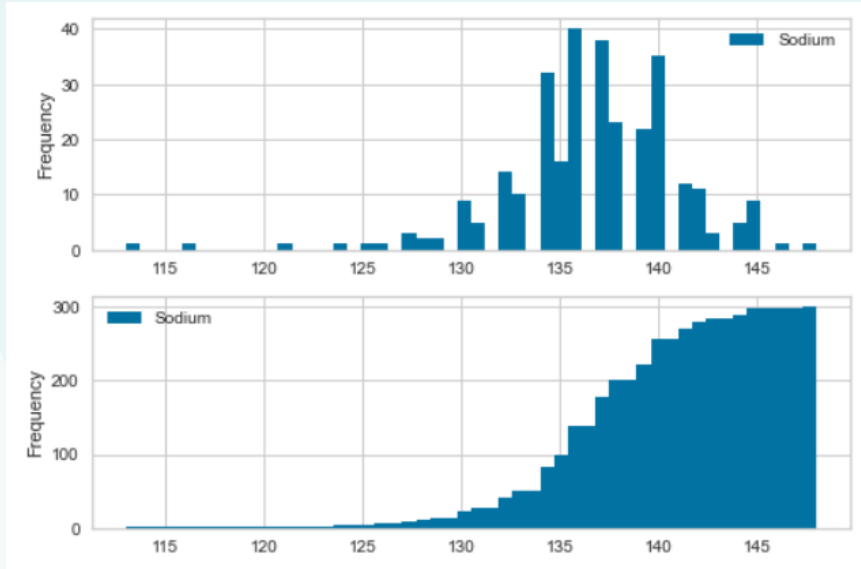
- Grande parte dos paciente ficaram de 50 a 100 dias em observação, e o valor de assimetria foi 0.12, sendo bem próximo de 0 o que indica que tem uma distriuição simétrica em relação ao valor central.
- No gráfico acumulativo pode-se perceber uma curva um pouco linear de 150 a 250 dias, o que indica que a quantidade de dias adicionados é aproximadamente constante;

Variáveis quantitativas - Idade



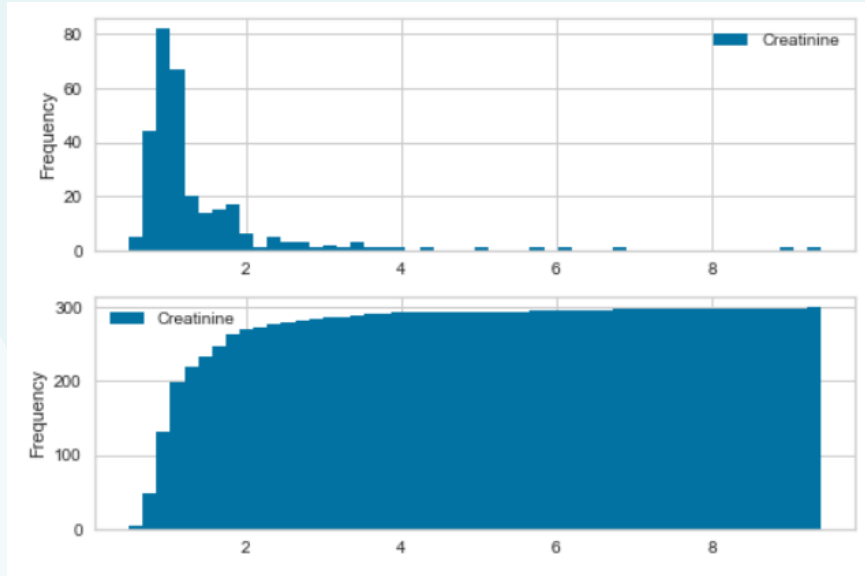
- A distribuição da idade dos pacientes parece ser normal, tendo uma concentração maior próximo de 60 anos.
- O valor retornado de assimetria foi 0.42, sendo também próximo de 0, justificando a hipótese acima, fazendo com que o valor de média, moda e mediana sejam parecidos.
- No gráfico acumulativo pode-se perceber uma curva um pouco linear, o que indica que a idade dos pacientes é aproximadamente constante;

Variáveis quantitativas - Sódio



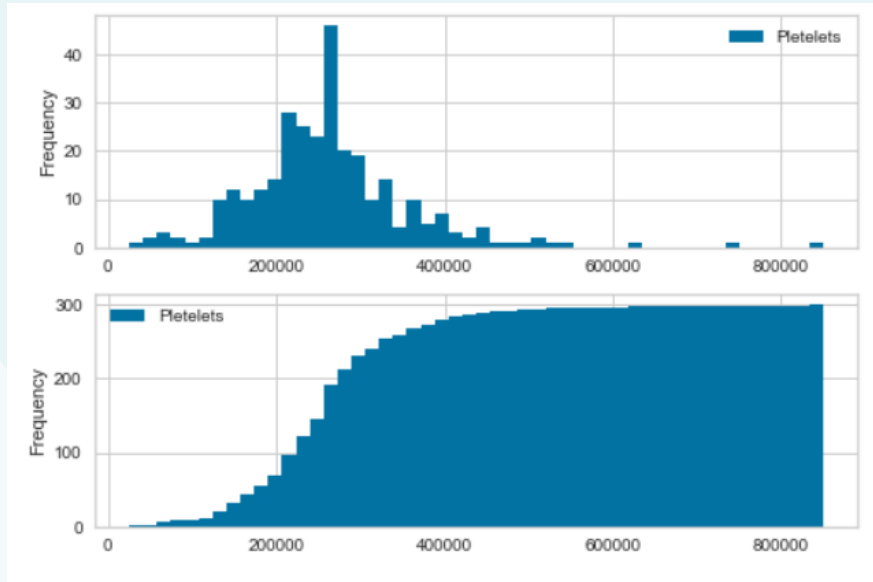
- A distribuição de sódio no sangue tem uma concentração maior entre 135 e 140.
- Valores estão mais concentrados a direita (justificado pelo valor negativo de assimetria: -1.04), com uma cauda mais longa indicando que há alguns valores discrepantes baixos que "puxam" a cauda. Fazendo com que o valor da média seja muito mais baixo que a mediana e a moda.
- No gráfico acumulativo pode-se perceber que a partir de 125 há um aumento exponencial na quantidade de sódio no sangue, ou seja quando a quantidade de sódio no sangue aumenta a curva acumulativa é "côncava para cima".

Variáveis quantitativas - Creatinina



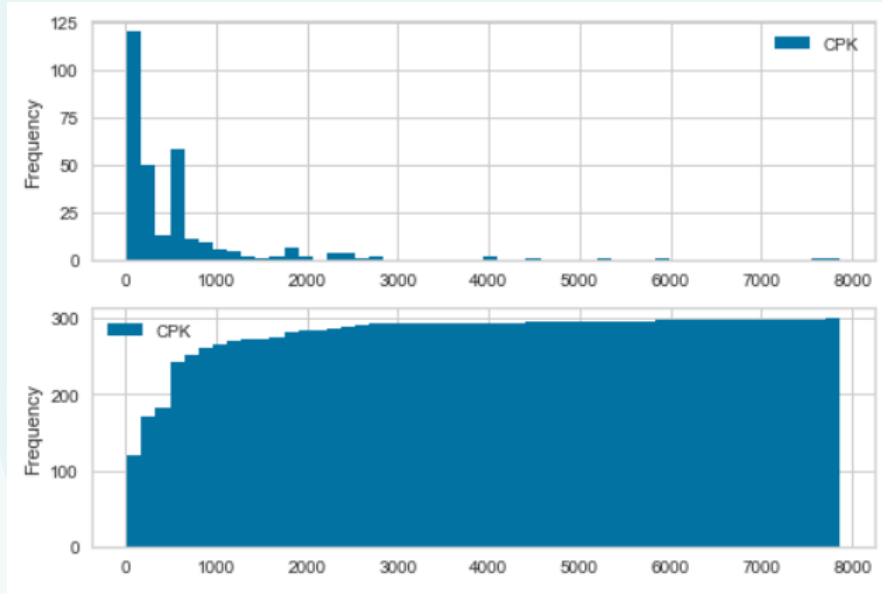
- Os níveis de creatinina estão mais concentrados entre 1 e 2.
- Os valores estão mais concentrados a esquerda (justificado pelo valor positivo de assimetria: 4.45), com uma cauda mais longa indicando que há alguns valores discrepantes (outliers) que "puxam" a cauda. Portanto o valor da média é muito maior que o valor da mediana e da moda.
- No gráfico acumulativo pode-se perceber que a partir de 1 há um aumento exponencial na quantidade de creatinina no sangue, ou seja quando a quantidade de creatinina no sangue aumenta a curva acumulativa é "côncava para cima".

Variáveis quantitativas - Plaquetas



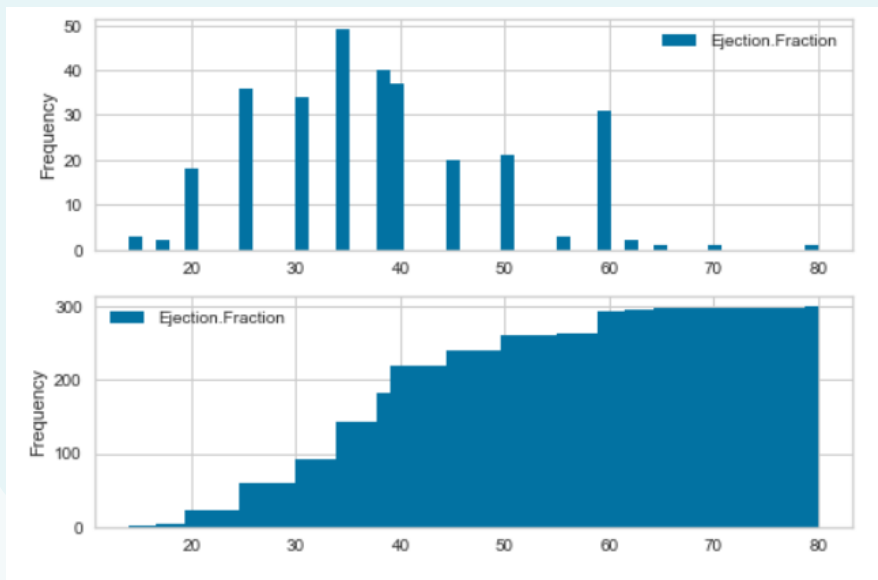
- Pode-se perceber que os níveis de plaqueta no sangue se concentram entre 200000 e mais ou menos 300000, e apesar de visualmente parecer uma distribuição normal, ele está levemente assimétrico à esquerda (justificado pelo valor de assimetria: 1.46).
- E com o gráfico acumulativo pode-se perceber que a partir de 20000 há um aumento exponencial na quantidade de plaquetas no sangue, ou seja quando a quantidade de plaquetas no sangue aumenta a curva acumulativa é "côncava para cima".

Variáveis quantitativas - CPK



- O níveis de CPK no sangue estão mais concentrados entre 0 e 1000, estando bem assimétrico à esquerda (justificado pelo valor de assimetria: 4.46), ou seja o valor da média é muito maior do que a mediana e a moda, devido aos outliers que "puxam" a cauda.
- E com o gráfico acumulativo pode-se perceber que a partir de 0 há um aumento exponencial na quantidade de CPK no sangue, ou seja quando a quantidade de CPK no sangue aumenta a curva acumulativa é "côncava para cima".

Variáveis quantitativas – Saída de sangue



- A maior parte da porcentagem de sangue que sai do coração dos pacientes estão mais concentrados entre 30 e 40%. +
- O valor de assimetria foi 0.55, indicando que a distribuição é simétrica, onde os valores de média, mediana e moda são bem próximos. +
- E com o gráfico acumulativo pode-se perceber que a partir de 20 há um aumento exponencial na porcentagem de sangue que sai do coração, ou seja quando a porcentagem de sangue que sai do coração aumenta a curva acumulativa é "côncava para cima".

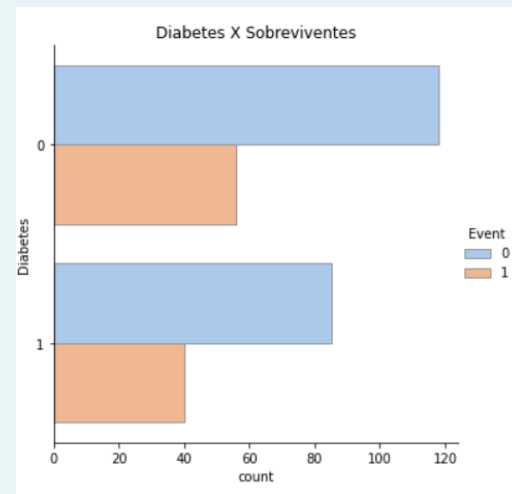
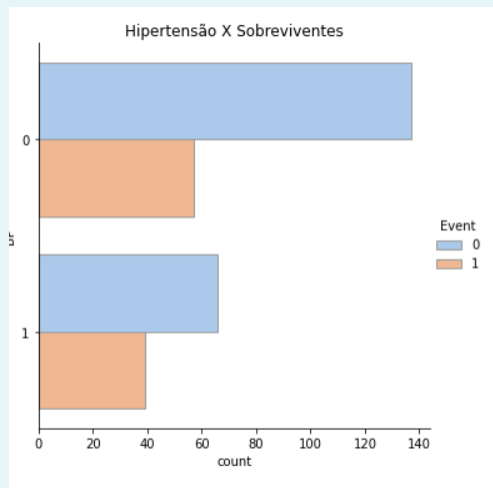
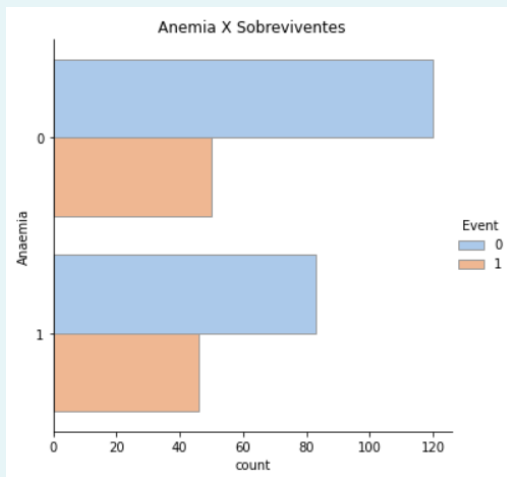
02

Quais variáveis impactam na sobrevivência?

Análise diagnóstica

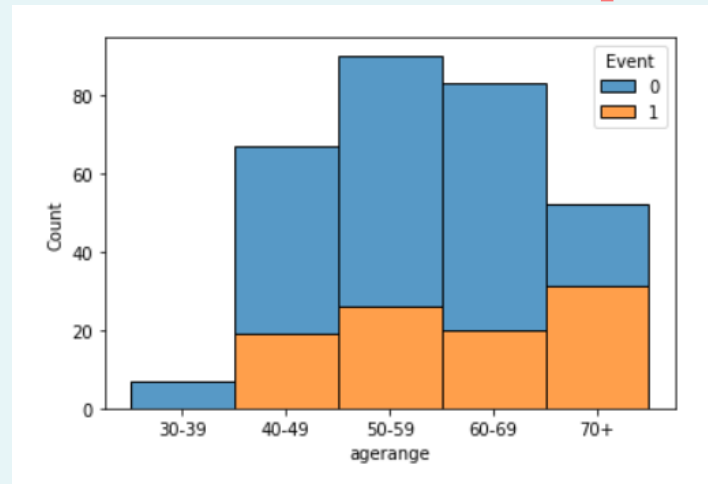
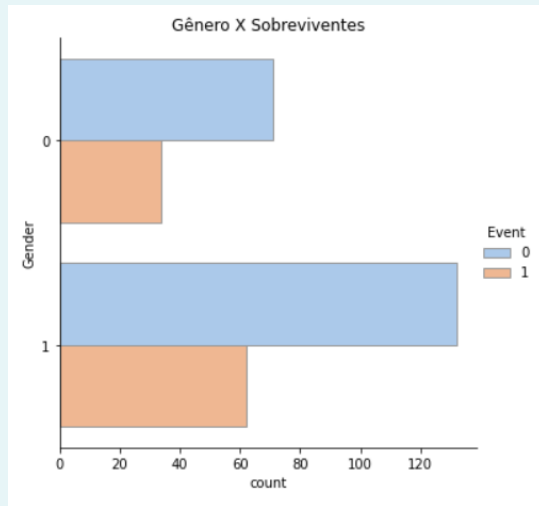
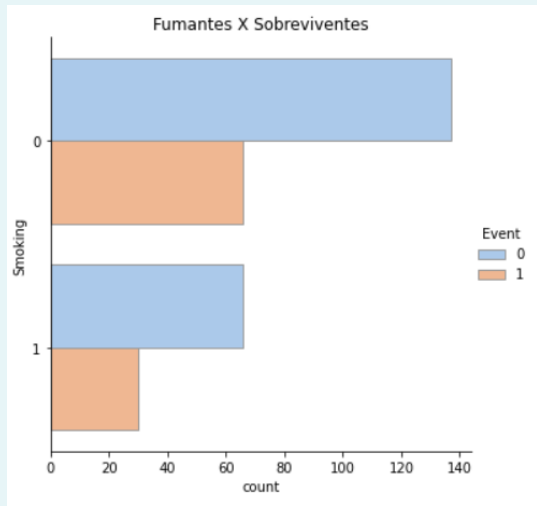


Variáveis categóricas



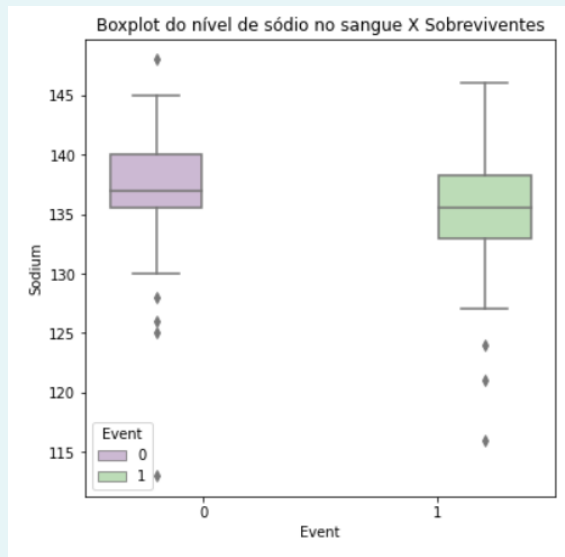
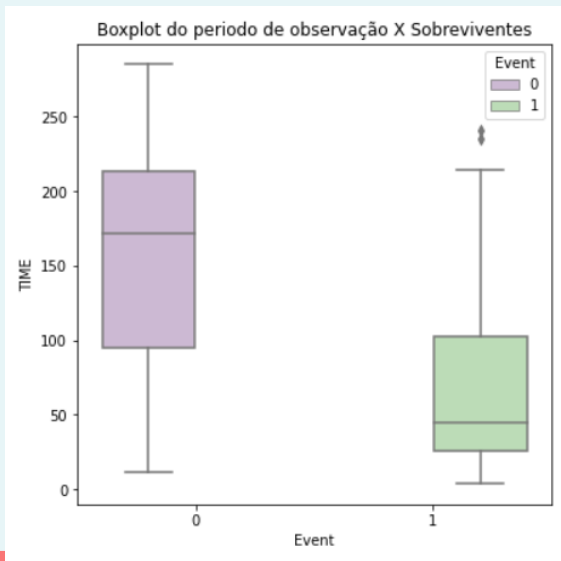
A maioria dos sobreviventes não tem anemia, hipertensão e diabetes. Entretanto a maioria dos pacientes que morreram também tem essas mesmas características

Variáveis categóricas



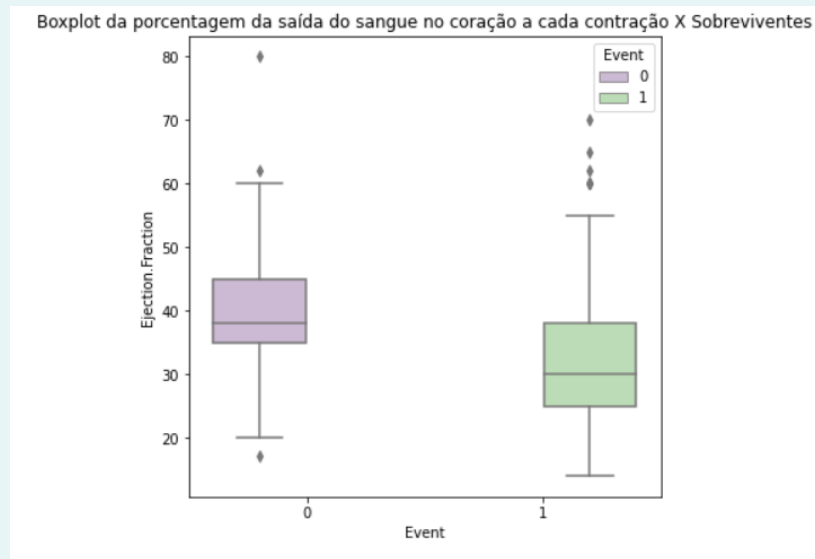
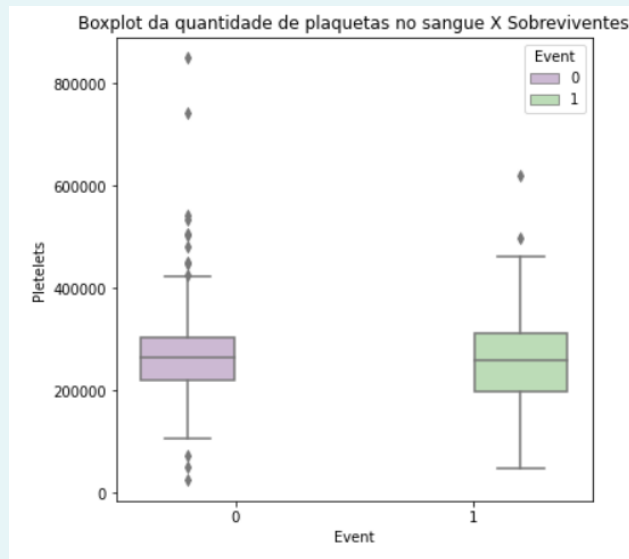
Tivemos mais sobreviventes do gênero masculino, entretanto isso pode-se dar ao fato que temos mais amostras do gênero masculino. E a maioria dos sobreviventes e pacientes que morreram não fumam. A taxa de sobreviventes é maior quando a pessoa tem entre 30-49 anos.

Variáveis quantitativas



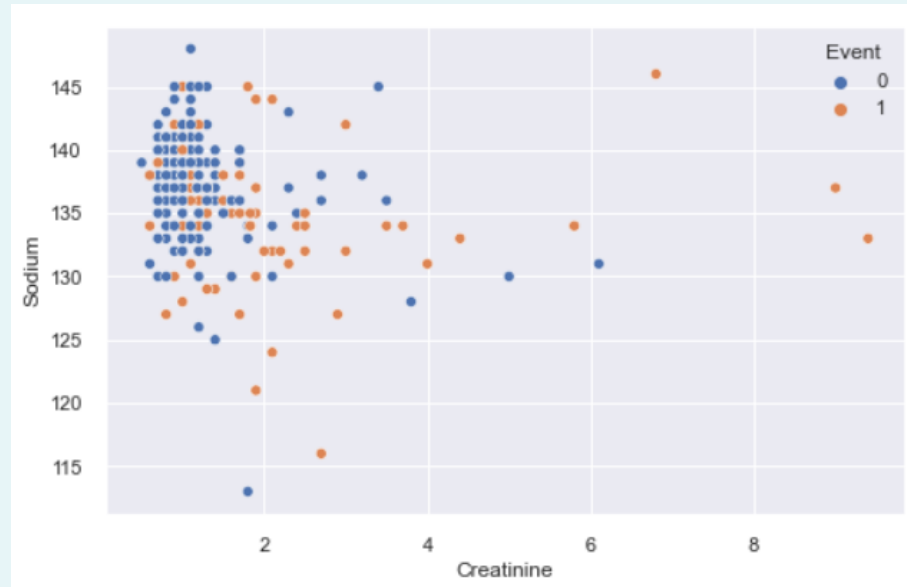
- Pessoas que sobreviveram ficaram mais tempo em observação do que as pessoas que morreram. A mediana do nível de sódio no sangue das pessoas que sobreviveram é maior que as que não sobreviveram. A mediana do nível de creatinina das pessoas que morreram é muito mais alta que os que não morreram, além disso pode-se perceber que há alguns outliers presentes, tanto quanto a classe é 0 quanto 1.

Variáveis quantitativas



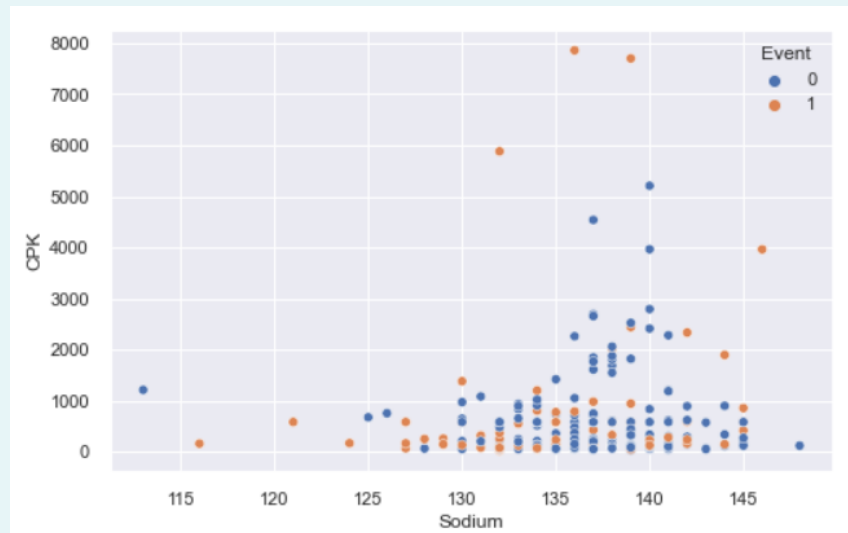
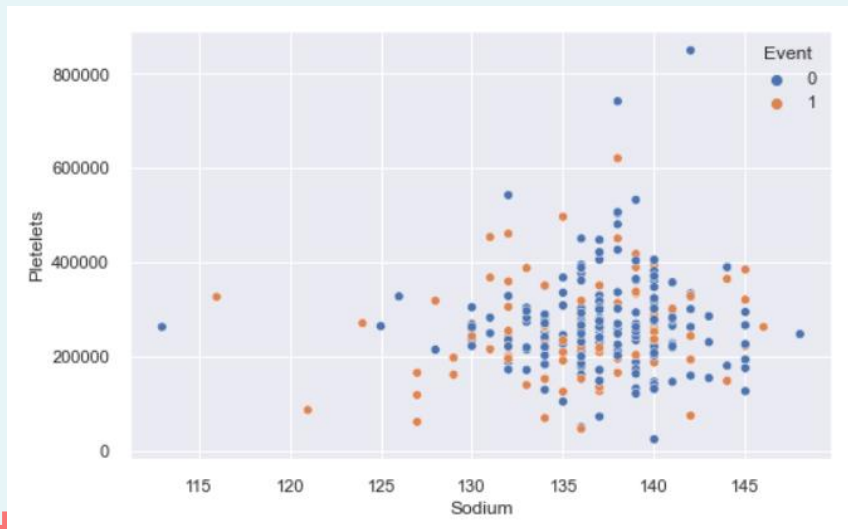
- Em relação as plaquetas no sangue podemos perceber que a distribuição de valores entre ambos os boxplots são bem semelhantes o que pode causar alguns problemas na hora do modelo diferenciar entre as classes 0 e 1 (causando alguns Falsos Positivos e Falsos Negativos), portanto não seria ideal utilizá-lo. Em relação a quem morreu a porcentagem da saída no sangue do coração de pessoas que morreram estava bem abaixo dos que sobreviveram

Variáveis quantitativas



Não existe uma correlação forte entre o nível de Creatinina e de sódio no sangue, mas podemos ver que quando o nível de creatinina está baixo o nível de sódio está mais alto. Já nesse gráfico conseguimos identificar alguns outliers no valor da creatinina (pontos bem isolados nos gráfico quando está acima de 5).

Variáveis quantitativas



Nos gráficos podemos perceber que quando o nível de sódio no sangue está alto, as plaquetas no sangue e a enzima CPK estão baixas. Entretanto olhando para essas variáveis e os pontos de sobreviventes (azul) e não sobreviventes (laranja), não tem como encontrar uma relação ou padrão.



Pipeline utilizada para o desafio



03

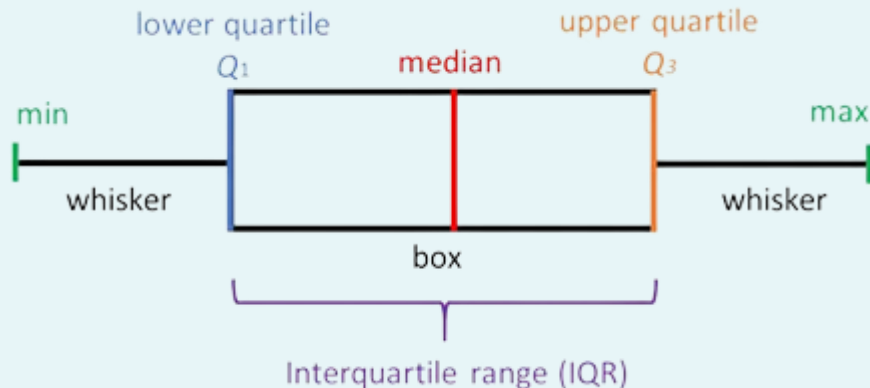
Pré-processamento

Tratamentos realizados nas variáveis



Outliers

Além das visualizações univariadas que identificamos a presença de outliers na variável Creatinina. Como essa variável é **assimétrica**, decidi fazer o tratamento e identificação da quantidade de outliers usando o método **Interquartil**, que leva em consideração os quartis (limite superior e inferior) para fazer o corte.



Informação dos outliers

Após executar o método foi retornado as seguintes informações:

Limite superior	0.150000000000000024
Limite inferior	2.1499999999999995
Quantidade de outliers	29
Porcentagem relativa no dataset	9.698996655518394%

Levando isso em consideração **removi esses outliers** pois correspondem a uma pequena parcela do dataset (9%). Entretanto seria ideal **validar com um especialista** da área se esse range de corte está certo, se os outliers são verdadeiros.



Pipeline utilizada para o desafio



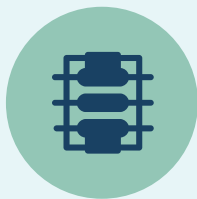
04

Seleção de variáveis

Utilização de métodos estatísticos



Métodos para seleção de variáveis



Teste-t

Usado para comparar a média de duas amostras dadas. Nesse teste irei comparar a variável considerando a sobrevivência do paciente e a variável considerando o óbito.



Spearman

Verificar a correlação e o impacto entre a variável categórica (que no caso é o nosso target) vs contínua




Chi-square

Para calcular a correlação entre as variáveis categóricas e o target.




Teste T

Pode-se usar o Teste T para seleção de variáveis pois nele conseguimos **identificar se com aquela variável há uma diferença significativa entre os casos de morte e sobrevivência**, através das hipóteses:




Por exemplo:

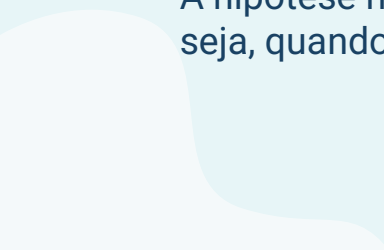
Hipótese nula: Não há diferença significativa entre o nível de creatinina e a sobrevivência do paciente.



Hipótese alternativa: Há uma diferença significativa entre o nível de creatinina e a sobrevivência do paciente.



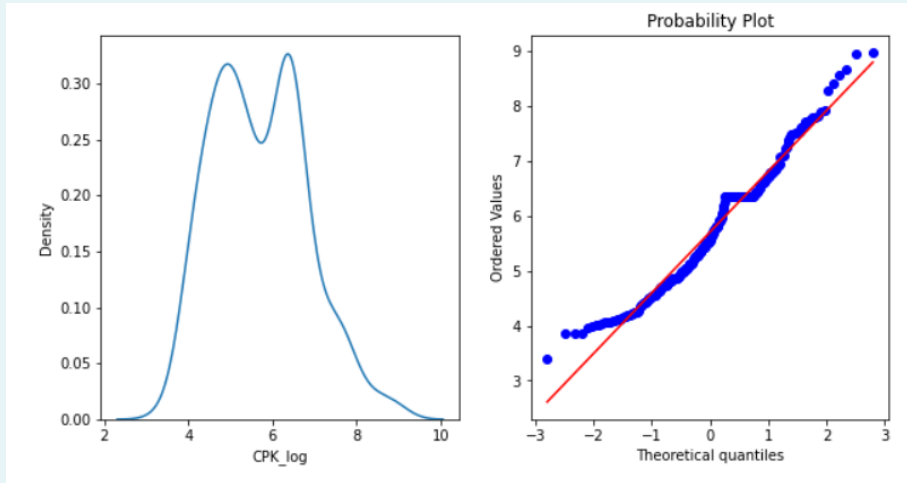
A hipótese nula é rejeitada se o valor p retornado pelo método for menor ou igual a 0.05, ou seja, quando o método tem 95% de certeza.



Teste T

Entretanto esse método espera que as amostras sigam uma distribuição normal e há algumas variáveis no dataframe que são assimétricas: CPK, plaquetas, creatinina e sódio. Então tive que criar uma nova variável com o valor log delas para ser ter uma distribuição normal.

Por exemplo a variável CPK, abaixo tem a visualização do gráfico de densidade e o QQ-plot:



Podemos ver que a distribuição está normal depois da transformação, pois os pontos dos dados estão em torno da linha vermelha no gráfico do QQ plot.

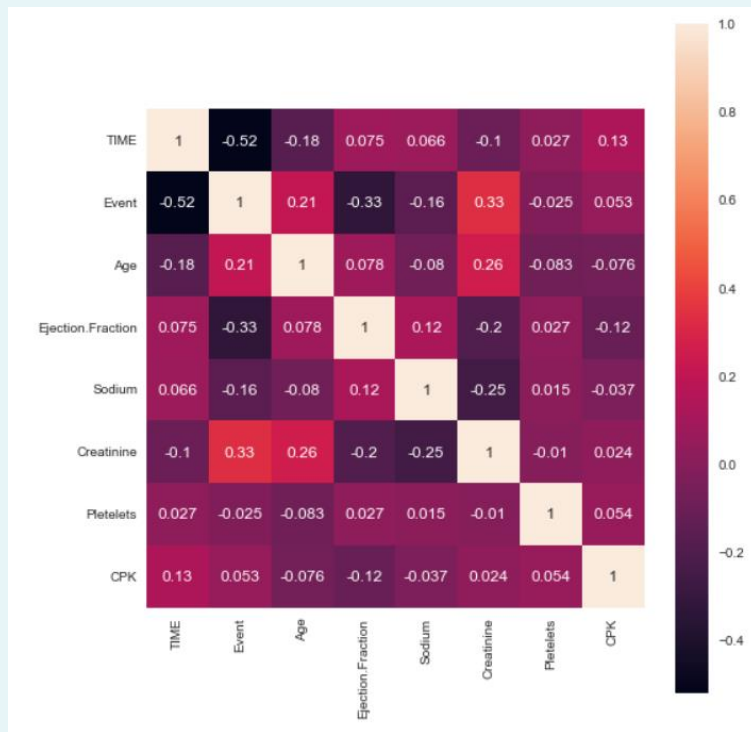
+ Teste – T: Resultados

	Passou?	Média sobreviventes	Média não sobreviventes
Age	✓	58	65
Time	✓	159	72
Ejection.Fraction	✓	40.48	32.64
Platelets*	✗	12.43	12.39
CPK*	✗	5.66	5.81
Sodium*	✓	4.92	4.91
Creatinine*	✓	0.0241	0.2406

* Valor do log

Spearman

E podemos ver que as variáveis apontadas pelo Teste T como importante, também foi refletido aqui na correlação de Spearman.

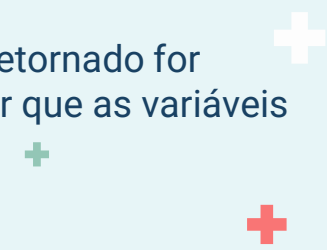




Chi-Square

Esse teste estatístico é um dos mais conhecidos quando queremos verificar se 2 variáveis categóricas são correlacionadas.

Nele a hipótese nula é que as variáveis são independentes e se o valor p retornado for menor ou igual a 0.05, então rejeitamos essa hipótese e podemos assumir que as variáveis na verdade são dependentes uma das outras.



Chi-Square: Resultados

	Passou?
Gender	✗
Diabetes	✗
Smoking	✗
BP	✗
Anaemia	✗

Pelo teste estatístico escolhido **nenhuma das variáveis categóricas tem uma relação com a variável alvo** (sobrevivência), levantando a hipótese que o estado do paciente durante os dias de observação (as alterações corporais: CPK, plaquetas, etc..) tem uma relação muito maior. Portanto vou **criar uma nova variável combinando as mesmas** para ver se muda a estatística de dependência.

Variável: `smoke_diabetes_anaemia_hip`

Criei essa variável com o intuito de falar de forma geral a condição do paciente, (se ele fuma, se tem diabetes, anemia e hipertensão). E juntei os valores binários dessa variável.

Por exemplo: se o paciente fuma e tem hipertensão, o valor da variável é:

`1_0_0_1`








Logo após isso eu transformo elas em números, usando o método One Hot Enconding.





Chi-Square: Resultados novas variáveis

	Passou?
smoke_diabetes_anaemia_hip_0_0_0_0	✗
smoke_diabetes_anaemia_hip_0_0_0_1	✗
smoke_diabetes_anaemia_hip_0_0_1_0	✗
smoke_diabetes_anaemia_hip_0_0_1_1	✗
smoke_diabetes_anaemia_hip_0_1_0_0	✓
smoke_diabetes_anaemia_hip_0_1_0_1	✗
smoke_diabetes_anaemia_hip_0_1_1_0	✗

Chi-Square: Resultados novas variáveis

	Passou?
smoke_diabetes_anaemia_hip_0_1_1_1	
smoke_diabetes_anaemia_hip_1_0_0_0	
smoke_diabetes_anaemia_hip_1_0_0_1	
smoke_diabetes_anaemia_hip_1_0_1_0	
smoke_diabetes_anaemia_hip_1_0_1_1	
smoke_diabetes_anaemia_hip_1_1_0_0	
smoke_diabetes_anaemia_hip_1_1_0_1	

Chi-Square: Resultados novas variáveis

	Passou?
smoke_diabetes_anaemia_hip_1_1_1_0	
smoke_diabetes_anaemia_hip_1_1_1_1	

Com as variáveis apresentadas a que teve dependência com a sobrevivência ou não do paciente é "smoke_diabetes_anaemia_hip_0_1_0_0", ou seja se o paciente não fuma, tem diabetes, não tem anemia e não tem hipertensão. Portanto no lugar das outras variáveis categóricas, irei utilizar ela como entrada pro modelo



Pipeline utilizada para o desafio



05

Balanceamento de classes

Tratamento na variável alvo



Tipo de aprendizado de máquina

Como o principal objetivo do desafio é identificar se o paciente tem chance de sobreviver ou não, o problema se trata de uma **classificação**.

E as principais métricas que considerei foram as que levam em consideração os **Falsos Negativos e Falsos Positivos**, dando ênfase no Falso Negativo pois o cenário mais crítico é se o modelo informar que o paciente tem irá sobreviver sendo que na verdade não. Portanto as métricas de classificação que utilizarei para selecionar o melhor modelo é:

- Recall
- F1 Score
- AUC

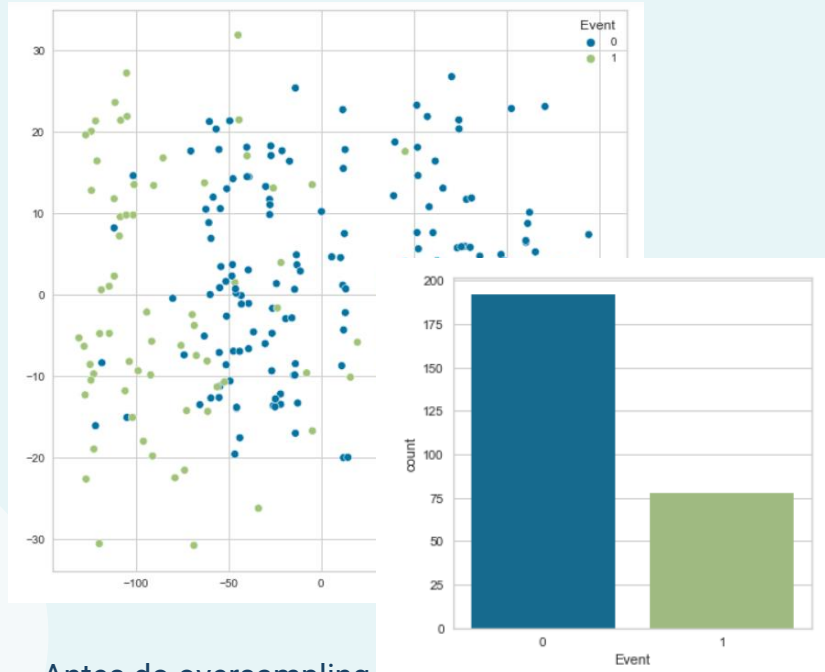


Desbalanceamento da classes

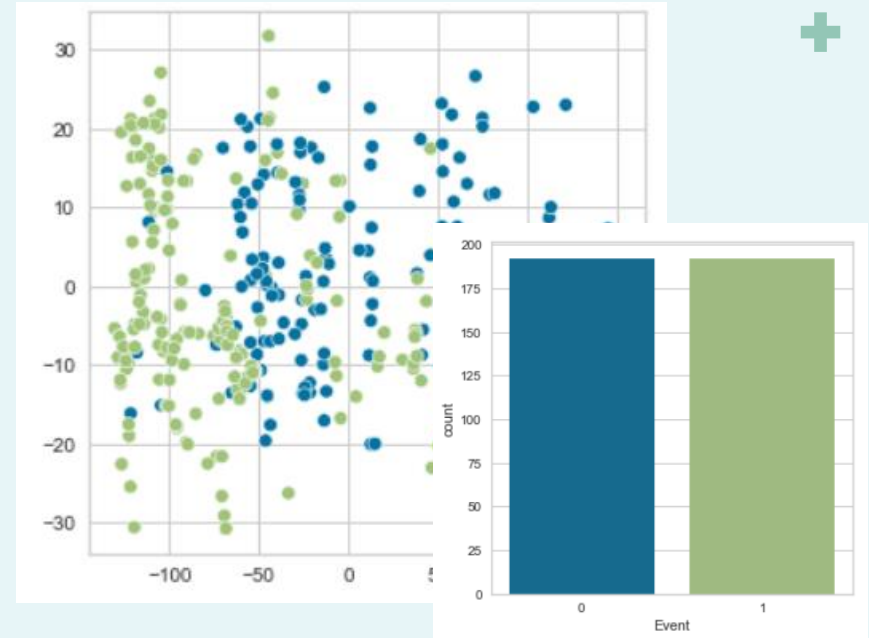


- Já no gráfico analisado na Análise descritiva podemos identificar que há um **desbalanceamento nas classes 0 e 1**;
- Temos mais amostras com pessoas que sobreviveram dos que morreram;
- Isso pode ser um problema na fase de modelagem, pois o modelo de aprendizado de máquina irá ficar **tendencioso/ especialista em uma classe** apenas (classe majoritária, que no nosso caso é o 0);
- Para evitar isso, irei realizar o método de **reamostragem oversampling**, que seleciona aleatoriamente exemplos da classe minoritária (no nosso caso é a classe 1) e vai criar pontos sintéticos na nossa amostra;

Desbalanceamento da classes

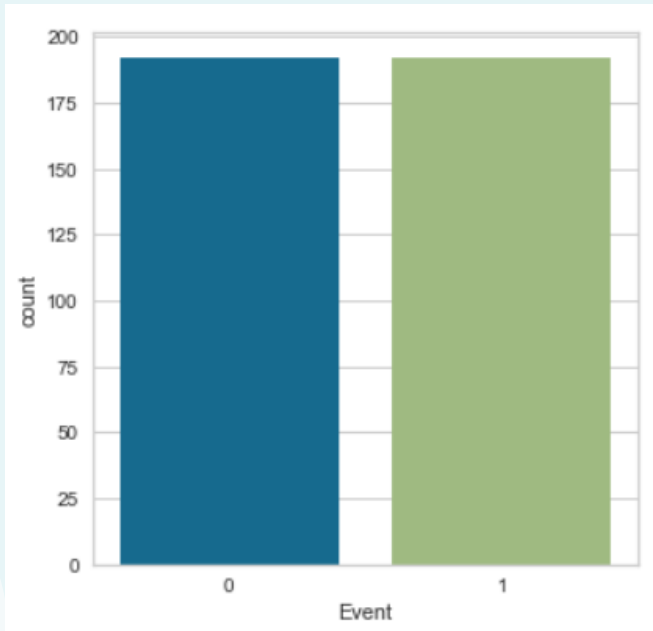


Antes do oversampling



Depois do oversampling

Desbalanceamento da classes



- Decidi utilizar essa técnica pois a classe 1⁺ corresponde apenas a 32.1% das amostras, então se usássemos o método de redução de amostras majoritárias (undersampling), ou seja, reduzir a classe 0 para a mesma quantidade da classe 1, teríamos poucas amostras para treinar o modelo;
- E depois do processo de oversampling podemos ver que as classes estão balanceadas;



Pipeline utilizada para o desafio



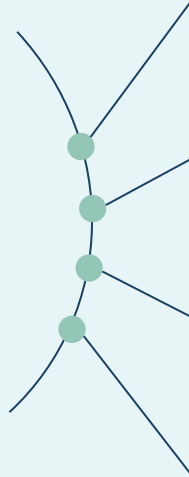
06

Modelagem

Criação do modelo e importância das variáveis



Etapas utilizadas



Step 1



Divisão da base em 30% pra teste e 70% pra treino;

Step 2



Pycaret (AutoML) para automatizar a etapa de seleção dos melhores modelos;

Step 3



Seleção dos 3 melhores modelos;

Step 4



Cross-validation para selecionar o melhor modelo generalista;

Pycaret - Resultados

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
nb	Naive Bayes	0.5984	0.9088	1.0000	0.5416	0.7020	0.2304	0.3543	0.0120
lightgbm	Light Gradient Boosting Machine	0.8612	0.9259	0.8876	0.8324	0.8566	0.7221	0.7281	0.1200
rf	Random Forest Classifier	0.8615	0.9283	0.8869	0.8349	0.8581	0.7228	0.7277	0.1300
et	Extra Trees Classifier	0.8774	0.9343	0.8765	0.8773	0.8699	0.7539	0.7641	0.1040
gbc	Gradient Boosting Classifier	0.8397	0.9160	0.8641	0.8087	0.8346	0.6792	0.6821	0.0320
xgboost	Extreme Gradient Boosting	0.8451	0.9275	0.8418	0.8333	0.8354	0.6889	0.6923	0.1180
knn	K Neighbors Classifier	0.8558	0.9318	0.8314	0.8620	0.8438	0.7102	0.7144	1.4040
ada	Ada Boost Classifier	0.8078	0.8636	0.8301	0.7778	0.8016	0.6152	0.6191	0.0400
lr	Logistic Regression	0.8185	0.9084	0.8196	0.8040	0.8102	0.6360	0.6382	2.7380
dt	Decision Tree Classifier	0.8238	0.8234	0.8190	0.8102	0.8144	0.6465	0.6467	0.0100
svm	SVM - Linear Kernel	0.7862	0.0000	0.7144	0.8282	0.7557	0.5676	0.5837	0.0120

- Variáveis de entrada: 'smoke_diabetes_anaemia_hip_0_1_0_0', 'Creatinine', 'Sodium', 'TIME', 'Age', 'Ejection.Fraction'
- Apesar do Naive Bayes ter tido melhor valor para o Recall, não vou considerá-lo na escolha pois ele não foi muito bem nas métricas F1 e AUC, o que indica uma dificuldade do modelo em diferenciar as classes 0 e 1;
- **Melhores modelos escolhidos:** Random Forest, ExtraTrees, LGBM;

Cross- Validation

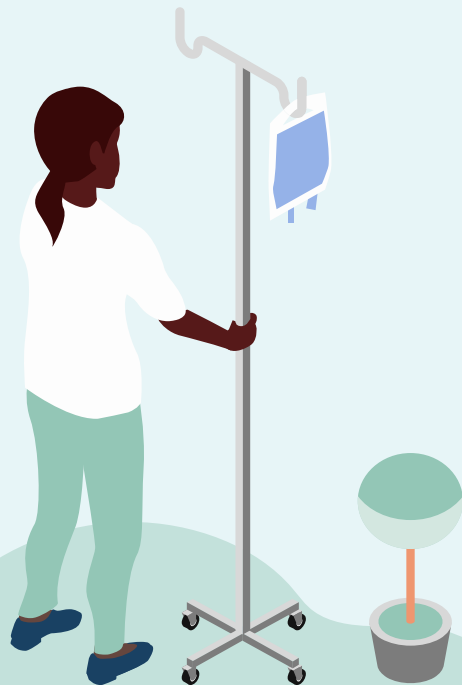
Avaliar a capacidade de generalização dos modelos, em outras palavras, verificar o quão pronto o modelo está para receber novos dados.

Modelo	AUC	Recall	F1
LGBMClassifier	0.9263363180029847	0.9	0.8771742543171115
RandomForestClassifier	0.935296356592653	0.8925925925925924	0.8760563823768173
ExtraTreesClassifier	0.9438092600129637	0.8854700854700855	0.8831059009298727

O modelo ExtraTrees foi o que apresentou melhor desempenho para as métricas AUC e F1 com um desvio padrão bem baixo. E apesar de ter sido o terceiro melhor no recall, comparado com o LGBM ele teve um desvio padrão mais baixo. Portanto escolherei ele como o melhor modelo.

Otimização de parâmetros

Antes de chegar no modelo final, otimizei os parâmetros do modelo ExtraTrees, com o método GridSearch que constroe um modelo para cada combinação possível de todos os valores de hiperparâmetro fornecidos, avaliando cada modelo e selecionando os parâmetros que tiverem melhor valor para o Recall.



Métricas finais - ExtraTrees



Recall

0.9



F1 Score

0.9



AUC

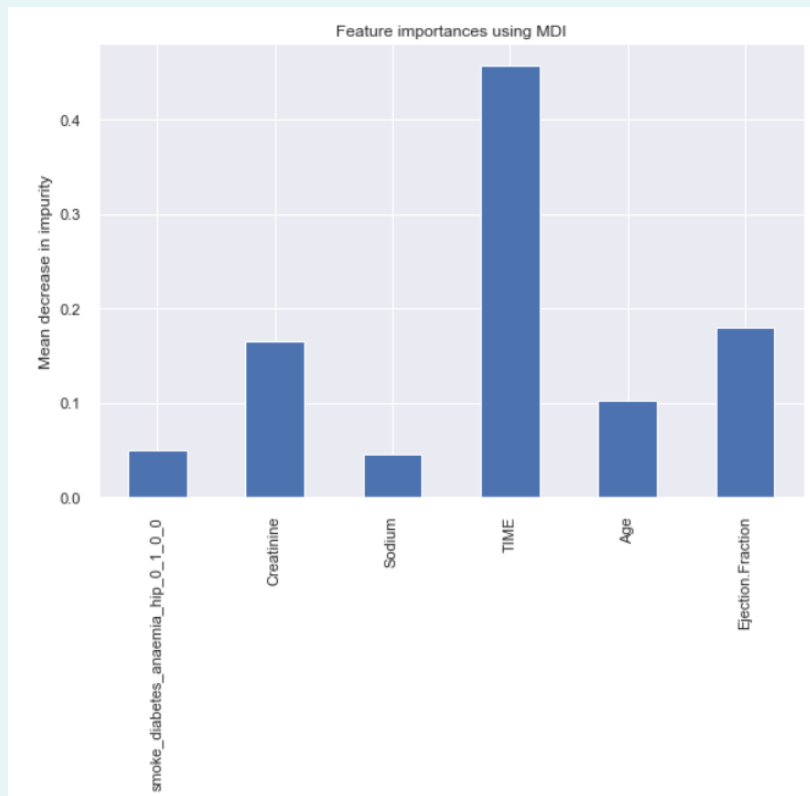
0.9

O modelo prevê 90% da classe 0 como corretas e 10% incorretamente. Modelos que tem valores de AUC e F1 próximos de 1 são considerados modelos bons, pois indica que ele consegue diferenciar bem as classes 0 e 1.

**Quais variáveis tem maior
impacto na sobrevivência
dos pacientes?**



Visualização gráfica da importância



+ Variáveis mais importantes para a sobrevivência do paciente



Tempo

Tempo em que o paciente está em observação



Creatinina

Nível de creatinina no sangue do paciente



Saída do sangue

Porcentagem da saída do sangue no coração do paciente a cada contração



Idade

Idade do paciente



A photograph of three medical professionals in a clinical setting. A male doctor with a beard and glasses, wearing a white lab coat, is seated at a desk. He is looking at a document held by a male colleague on the left. A female colleague on the right is also looking at the document. The background shows a window with a view of a city. The text is overlaid on the image in a dark teal color.

Qual o comportamento geral do modelo para cada variável?

Qual o impacto de cada variável na predição do modelo?



Pipeline utilizada para o desafio



07

XAI

Tornando transparente as decisões do modelo



Explainable AI (XAI)

Explainable AI ou Inteligência Artificial explicável, é uma área de estudo que visa tornar a saída dos modelos de IA de uma forma que os seres humanos possam entender.

Fazendo com que as pessoas entendam o “porque” da saída do modelo.





Vantagens do XAI

- Aumentar a confiança dos usuários (médicos) na inteligência artificial;
- Encontrar pontos de melhoria dos modelos de I.A e assim facilitar as correções;
- Levar maior transparência das decisões do modelo para as pessoas;

Perguntas que o XAI ajuda a responder

- Porque o modelo gerou essa saída e não outra?
- Quando e porquê eu posso confiar no modelo?
- Como eu posso ajustar o modelo?



Tipos de XAI



Global

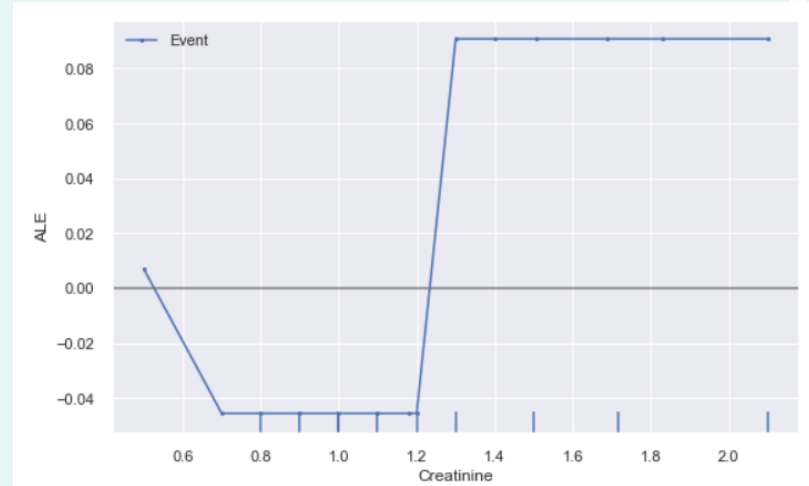
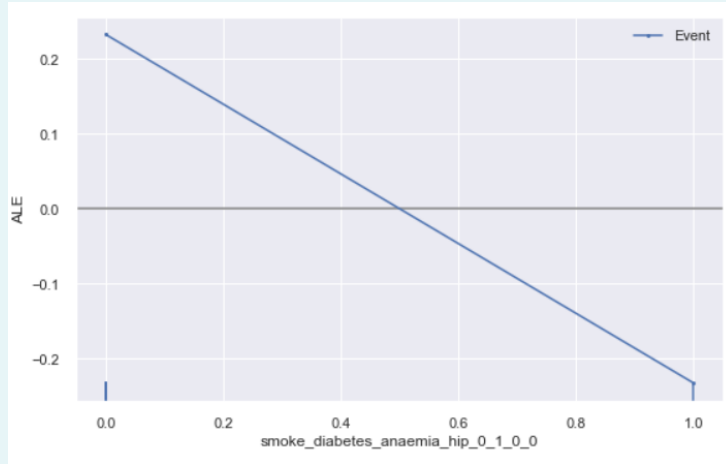
Descrevem o comportamento médio esperado do modelo de predição com base no efeito de uma variável. Nesse caso vou usar a biblioteca ALE



Local

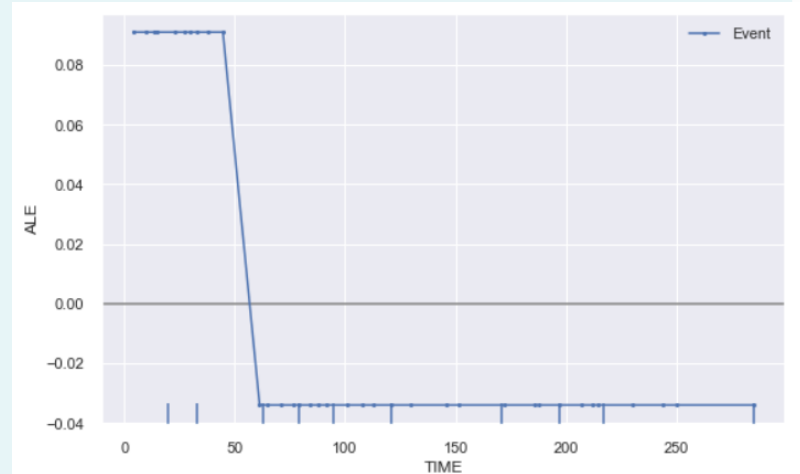
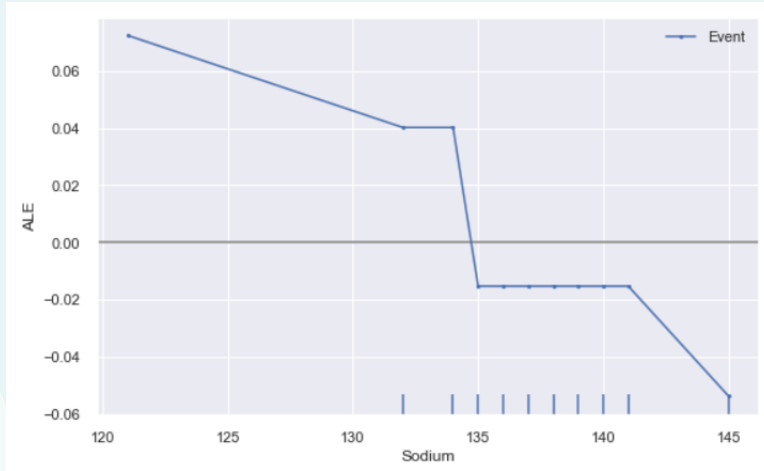
Para um conjunto específico de amostras, entender o impacto de cada variável na saída do modelo. Nesse caso vou usar a biblioteca Shap

XAI Global



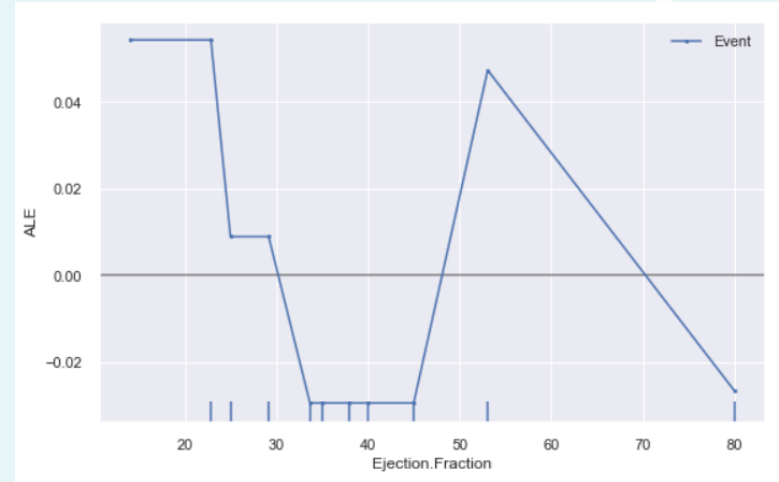
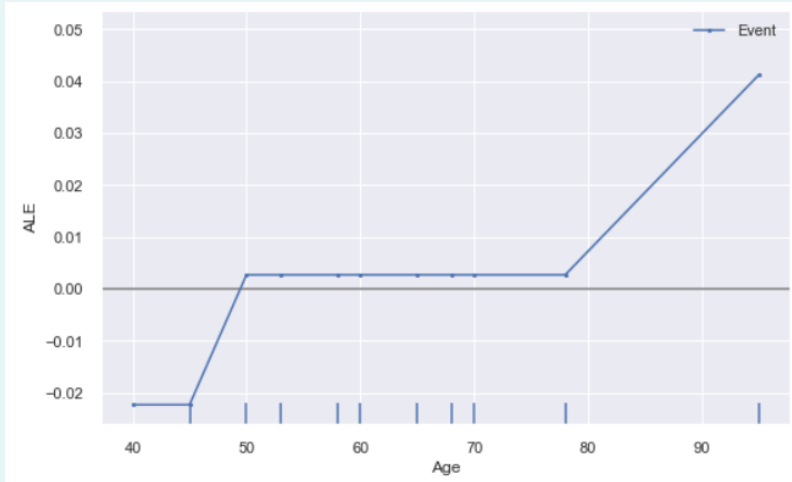
- Quando a variável da condição do paciente é 0 há uma influência de 0.2 na probabilidade do classificador;
- Quando o nível de Creatinina no sangue entre 0.7 e 1.2 tem uma influência de -0.04 na probabilidade do classificador, já de 1.3 a 2 há uma influência de 0.09 no classificador;

XAI Global



- Quando a porcentagem de sódio no sangue está entre 135 e 142 o efeito médio na probabilidade do classificador é de -0.01. Quando está próximo de 121 o efeito é de 0.06 no classificador;
- Quando o tempo de observação do paciente está entre 55 e 260 o valor médio de influência na probabilidade do classificador é de -0.03. E quando o valor está entre 1 e próximo de 50 dias o valor médio de influência é de 0.08;

XAI Global



- De 50 a aproximadamente 80 anos, não há impacto/influência dessa variável no valor de predição do classificador. De 80 a 90 anos há um impacto positivo no valor de probabilidade do classificador;
- De 20% a 45% de sangue que sai do coração a cada contração tem um impacto negativo na probabilidade do paciente morrer, entretanto essa influência se torna positiva (tem uma chance maior do paciente morrer) quando a porcentagem de sangue está entre 40% à 55%;

XAI Local



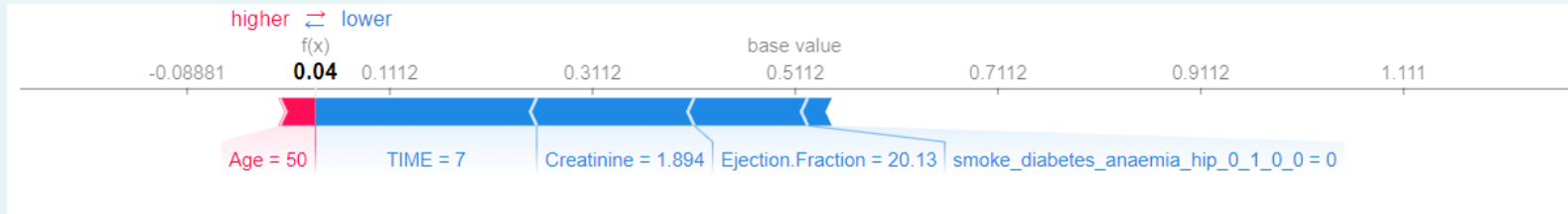
Pegar amostras

Para gerar essa explicabilidade peguei 3 amostras aleatórias no dataset de treino;

	smoke_diabetes_anaemia_hip_0_1_0_0	Creatinine	Sodium	TIME	Age	Ejection.Fraction
349	0	1.894091	137.0	7	50	20.132955
204	0	0.700000	139.0	73	60	20.000000
98	0	1.100000	145.0	200	65	50.000000

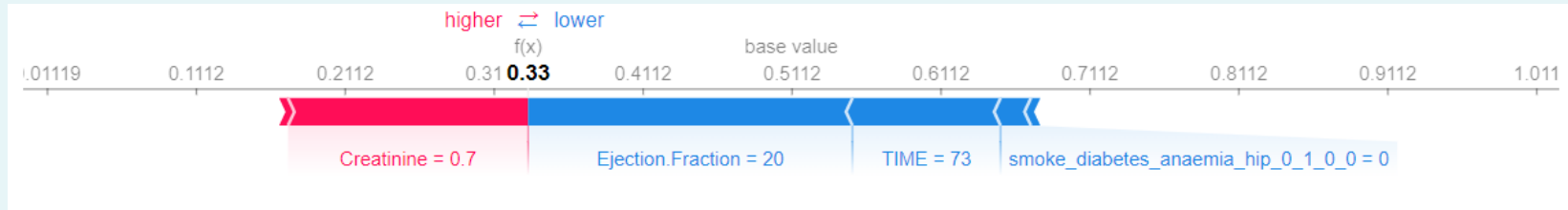
A visualização que estarei mostrando é o gráfico de força que mostra a força de cada variável na saída do modelo e a probabilidade da saída do modelo para aquela amostra ser da classe 0 (representado por $f(x)$)

XAI Local – Amostra 1



Esse paciente tem uma probabilidade de 4% de sobreviver (ser da classe 0), devido ao tempo em que a pessoa está internada, o nível de creatinina no sangue, a porcentagem de sangue que sai do coração a cada contração.

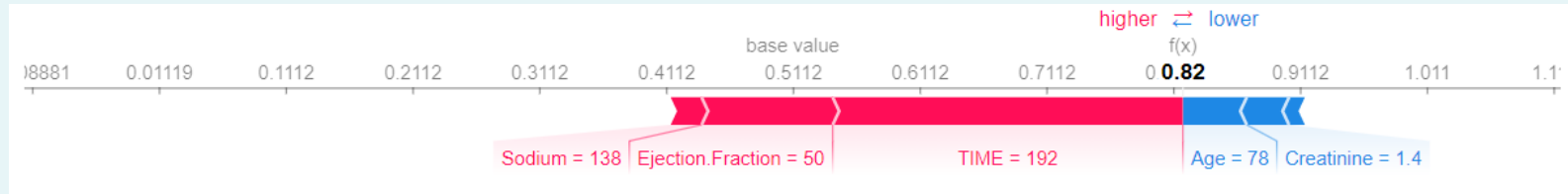
XAI Local – Amostra 2



Esse paciente tem 33% de chance de sobreviver, pois tivemos bastantes variáveis que contribuíram de forma negativa para a probabilidade como por exemplo: a porcentagem de sangue que sai do paciente a cada contração e o tempo em que esse paciente está em observação.

Entretanto pelo paciente ter um nível de creatinina no sangue de 0.7 isso contribuiu para o aumento da probabilidade.

XAI Local – Amostra 3



Esse paciente tem 82% de chance de sobreviver, pois tivemos muitas variáveis que contribuíram para o aumento dessa probabilidade como: o tempo que a pessoa está em observação, a quantidade de sódio no corpo e a porcentagem da saída do sangue no coração.

Já a idade da pessoa e o nível de creatinina no sangue contribuíram para a diminuição dessa probabilidade

08

Conclusão

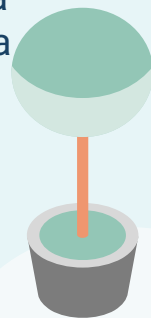
Análise dos impactos da IA na saúde pública





Conclusão e análise

- Problemas cardíacos têm se mostrado o maior causador de mortes do mundo, e estudos recentes aplicaram aprendizado de máquina como forma de predição e assim ajudar a tomada de decisão dos profissionais de saúde.
- Portanto é necessário ter modelos de aprendizado de máquina com o mínimo de Erros Tipo I (Falso Positivos) e Tipo II (Falsos Negativos) e o modelo apresentado em questão acerta 90% dos casos de sobrevivência.
- Além disso, modelos transparentes (que usam métodos de XAI) ajudam os médicos a ficarem mais confiantes em utilizarem esses sistemas e entenderem o impacto de cada variável na predição.



Obrigada pela oportunidade

Laura Damaceno de Almeida



<https://www.linkedin.com/in/laura-damaceno>



https://beacons.ai/laura_data



https://www.instagram.com/laura_data_talks/



<https://github.com/lauraDamacenoAlmeida>