

# Analizando o comportamento de modelos de aprendizado de máquina interpretáveis com modelos explicadores no cálculo do plano de saúde

Laura Damaceno de Almeida<sup>1</sup>

<sup>1</sup>Centro de Matemática, Computação e Cognição, Universidade Federal do ABC Av.dos Estados, 5001 - Bairro S. Terezinha, Santo André, SP, Brasil - CEP: 09210-580

`l.damaceno@ufabc.edu.br`

**Abstract.** *Health insurance is a means of financing a person's healthcare expenses, ensuring his access to quality medical care, reducing inequalities and improving the life's quality of a country's population. This study demonstrates the application of machine learning systems to calculate an accurate and fair value of the health insurance charge and analyze the behavior of interpretable models and black-box models with explainable models.*

**Resumo.** *O plano de saúde é um meio de financiar as despesas de saúde de uma pessoa, garantindo seu o acesso à assistência médica de qualidade, reduzindo desigualdades e melhorando a qualidade de vida da população de um país. Este estudo demonstra a aplicação de sistemas de aprendizado de máquina para calcular um valor preciso e justo da cobrança do plano de saúde e analisar o comportamento dos modelos interpretáveis e modelos caixa-preta com auxílio de modelos explicadores.*

## 1. Introdução

Nos países em desenvolvimento, as pessoas de baixa-renda não podem arcar com o alto custo de serviços médicos que estão crescendo exponencialmente com o tempo. A prestação de cuidados de saúde é o principal desafio na indústria de assistência médica ou *healthcare*. [Leung and Van Merode 2019] em seu artigo, comenta que a tecnologia na área de *healthcare* é fundamental, devido ao aumento da carga de doenças e problemas de acessibilidade para cuidados de qualidade.

O plano de saúde é um meio de financiar as despesas de saúde de uma pessoa. Embora a maioria das pessoas tenha um plano de saúde privado, principalmente através de um empregador, muitos outros obtêm cobertura através de programas oferecidos pelo governo [Berchick et al. 2019]. O plano de saúde acaba sendo uma necessidade social da população de qualquer país [Carrin 2002], pois garante o acesso adequado a assistência médica de qualidade. E a tecnologia desempenha um papel importante na redução de desigualdades na prestação de assistência médica e melhoria da qualidade de vida da população, fortalecendo assim inclusão [Nayak et al. 2019].

O Aprendizado de Máquina, subdisciplina popular da Inteligência Artificial (IA), utiliza grandes conjuntos de dados e identifica os padrões entre as variáveis. Essas técnicas podem descobrir associações anteriormente desconhecidas, gerando novas hipóteses e direcionar pesquisadores e recursos para melhores tomadas de decisão

[Noorbakhsh-Sabet et al. 2019]. Há vários benefícios com a utilização desses sistemas em *healthcare*, por exemplo melhorar a segurança do paciente, melhorar a qualidade do atendimento e reduzir custos de saúde [Waring et al. 2020].

O estudo apresentado neste artigo irá demonstrar a aplicação de sistemas de aprendizado de máquina de regressão para calcular o valor preciso e justo da cobrança do plano de saúde e analisar o comportamento dos modelos interpretáveis e modelos caixa-preta com auxílio de modelos explicadores.

### 1.1. Base de dados

Para o desenvolvimento deste estudo foi utilizado a base de dados do plano de saúde coletada nos Estados Unidos, disponibilizada no repositório Kaggle. Esta base contém 1338 amostras de segurados, onde as cobranças do plano (*Charge*) são fornecidas em relação aos seguintes atributos do segurado: *Age*, *Sex*, *BMI*, *Children*, *Smoker* e *Region* (Tabela 1.).

**Tabela 1. Informações das variáveis disponíveis na base de dados**

<i>Nome da variável</i>	<i>Descrição</i>
<b>Age</b>	Idade do beneficiador primário
<b>Sex</b>	Sexo do contratante de seguros: feminino/masculino
<b>BMI</b>	Índice de massa corporal
<b>Children</b>	Número de filhos cobertos pelo seguro saúde / Número de dependentes
<b>Smoker</b>	Fumante / Não-fumante
<b>Region</b>	A área residencial do beneficiário nos EUA
<b>Charge</b>	Cobrança do plano de saúde.

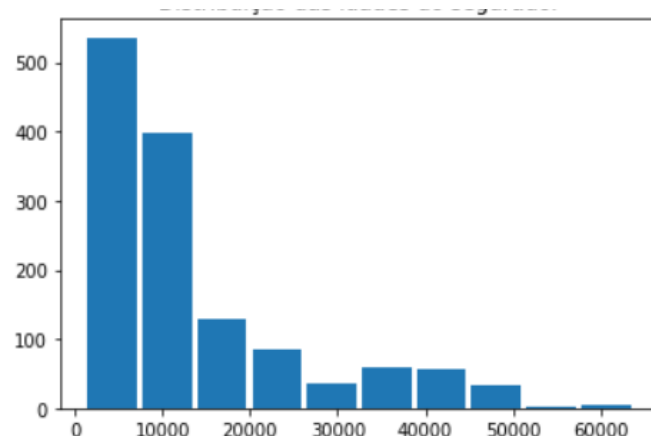
Portanto na base de dados há 4 variáveis numéricas (*Age*, *BMI*, *Children* e *Charge*) e 3 variáveis categóricas (*Sex*, *Smoker* e *Region*). Todas as variáveis categóricas serão transformadas em inteiros a partir da função *get dummies* da biblioteca Pandas e assim tornar mais acessível para os modelos de aprendizado de máquina.

## 2. Análise de dados

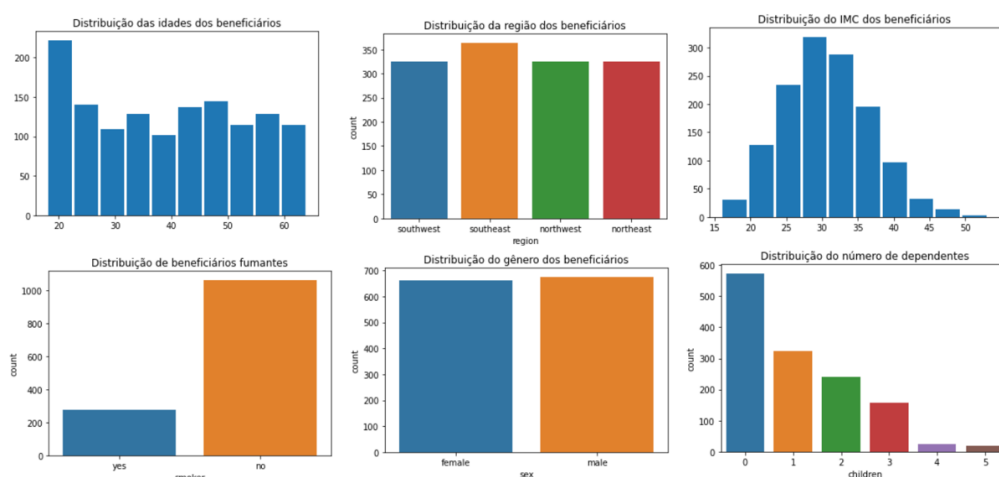
Durante a análise de dados foi identificado através do histograma (Figura 1), que a variável *target Charge* não apresenta distribuição normal, pois de forma visual ela segue uma tendência assimétrica. E essa hipótese foi posteriormente comprovada a partir do teste estatístico de normalidade com valor de significância de 0.05, pois o valor do *p-value* retornado pelo teste foi menor do que a significância definida. Levantando mais uma hipótese de que o modelo de regressão linear poderá não apresentar a melhor performance.

Além disso, com o objetivo de explorar, identificar padrões e compreender melhor a base de dados, foi realizado uma análise exploratória onde foi notado que grande parte dos beneficiários não tem dependentes registrados e a grande maioria deles não são fumantes. E a quantidade de homens e mulheres aparentam ser equivalentes (Figura 2).

Ao gerar o gráfico de dispersão entre a variável *Charges* e *Age* (Figura 3), nota-se que o valor do preço do plano de saúde apresenta 3 comportamentos diferentes: um preço



**Figura 1. Distribuição dos preços do plano de saúde**

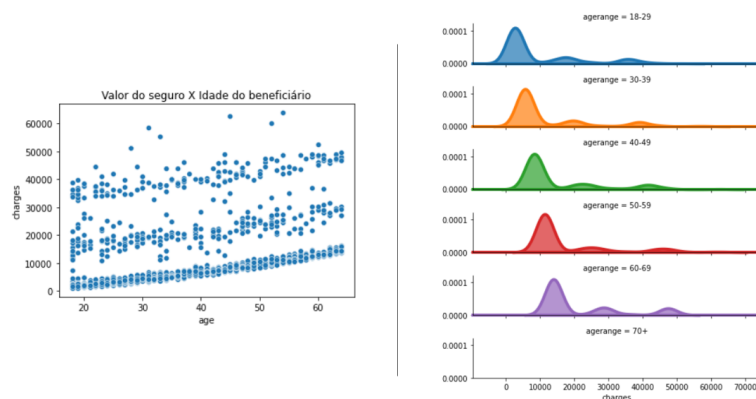


**Figura 2. Distribuição das variáveis**

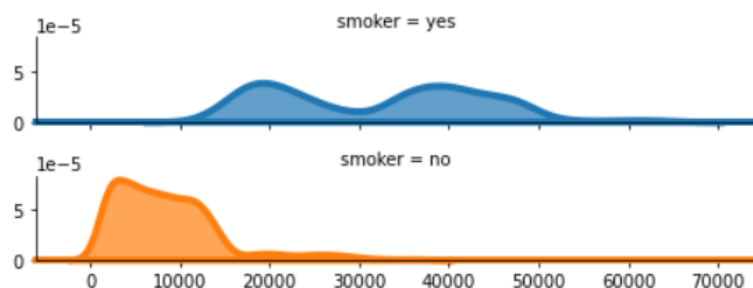
que varia até 10.000 dólares, outro preço que vai de 10.000 à 30.000 dólares e outro que é acima de 30.000, levantando a hipótese que há 3 tipos de seguros de vida que apresenta valores diferentes para cada beneficiário, além disso pode-se identificar que os valores aumentam conforme a idade do mesmo, apresentando uma correlação linear forte. E ao quebrar a idade em ranges diferentes 18-29, 30-39, 40-49, 50-59, 60-69, e mais de 70 anos pode-se perceber melhor a variação dos preços do seguro (Figura 3).

Ademais o beneficiário sendo fumante também contribui para um preço maior no seguro, isso porque no gráfico gerado pode-se perceber que a distribuição dos valores quando a pessoa é fumante acaba sendo bem maior que a de um não fumante (Figura 4). E pode-se perceber uma relação um pouco mais fraca entre a quantidade de dependentes e o valor do seguro, levando a hipótese de que alguma outra variável exerce uma terceira influência de forma indireta nessa relação (Figura 5).

Entretanto não foi identificado esse mesmo comportamento nas variáveis *Sex*, *Region* e *Charges* como mostrado na Figura 6 pois ambos não contribuem de forma significativa e direta na mudança do valor do seguro.



**Figura 3. Relação entre a idade e o preço do seguro**



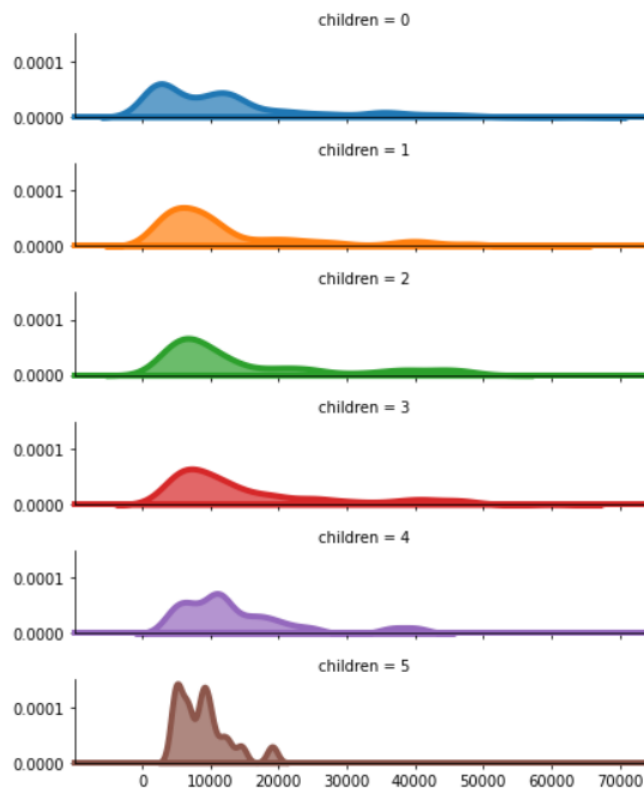
**Figura 4. Diferença de preços entre beneficiários fumantes e não fumantes**

Através das conclusões apresentadas acima, para fins de estudo será considerado 2 cenários na construção do modelo: 1- Todas as variáveis do conjunto de dados serão passadas para os modelos e 2- Somente as variáveis: *Children*, *BMI*, *Smoker* e range da idade serão passadas para o modelo. O conjunto de dados foi dividido em 70% para treino e 30% para teste, e as variáveis categóricas foram transformadas em valores numéricos. No segundo cenário foi desconsiderado o valor bruto da idade do beneficiário e foi criada a variáveis binária que categorizam a idade do beneficiário nas seguintes divisões: 18-29, 30-39, 40-49, 50-59, 60-69 e 70+, apresentadas na Figura 3.

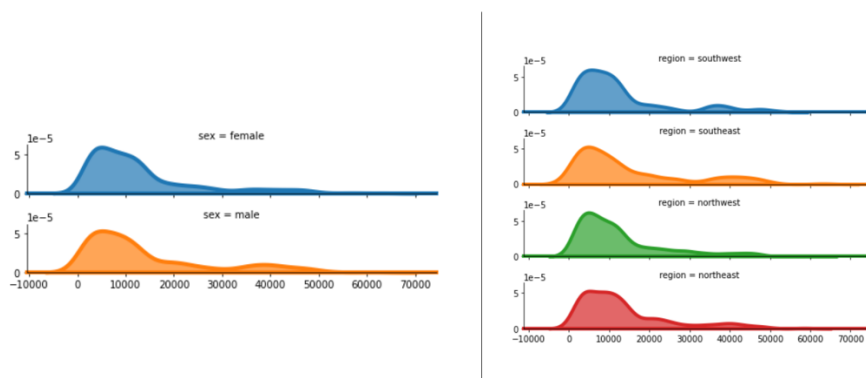
### 3. Modelagem e resultados iniciais

Neste trabalho foi considerado os seguintes algoritmos de aprendizado de máquina: Regressão Linear, Árvore de Decisão, *RuleFit* e *Gradient Boosting*. Os mesmos foram implementados utilizando a linguagem de programação Python e a biblioteca *Scikit-learn*.

Não foi utilizado os parâmetros padrões dos algoritmos, com exceção da Regressão Linear, foi utilizado o método do *GridSearch* para a seleção dos melhores parâmetros durante a etapa de treino. Os parâmetros e valores testados para o modelo de árvore de decisão, pode ser encontrado na Tabela 2 e a melhor configuração encontrada para o primeiro e segundo cenário foi: *max depth* = 6, *max leaf nodes* = 20, *min samples leaf* = 20 e *min samples split* = 10.



**Figura 5. Diferença de preços entre a quantidade de dependentes**



**Figura 6. Diferença de preços entre a quantidade de dependentes**

**Tabela 2. Parâmetros e valores testados para Árvore de Decisão**

<i>Parâmetros</i>	<i>Valores testados</i>
<b>min_samples_split</b>	10, 20, 40
<b>max_depth</b>	2, 6, 8
<b>min_samples_leaf</b>	20, 40, 100
<b>max_leaf_nodes</b>	5, 20, 100

Para o modelo *RuleFit*, os parâmetros e valores testados pode ser encontrado na Tabela 3 e a melhor configuração para o primeiro e segundo cenário foi: *exp rand tree*

*size = False*, *lin trim quantile = 0.05*, *max rules = 200*, *model type = rl*, *rfmode = regress* e *tree size = 5*. Os parâmetros e valores testados para *Gradient Boosting* está na Tabela 4, a configuração final para o primeiro cenário foi: *criterion = mse*, *max depth = 2*, *max features = auto*, *min samples leaf = 40*, *min samples split = 40*, *n estimators = 100*. E para o segundo cenário a configuração foi: *criterion = mse*, *max depth = 2*, *max features = auto*, *min samples leaf = 40*, *min samples split = 10*, *n estimators = 100*.

**Tabela 3. Parâmetros e valores testados para *RuleFit***

<i>Parâmetros</i>	<i>Valores testados</i>
<b>model_type</b>	"rl", "r", "l"
<b>max_rules</b>	100, 200, 500
<b>lin_trim_quantile</b>	0.05, 0.025, 0.075
<b>exp_rand_tree_size</b>	True, False
<b>tree_size</b>	5, 8, 10, 20

**Tabela 4. Parâmetros e valores testados para *Gradient Boosting***

<i>Parâmetros</i>	<i>Valores testados</i>
<b>max_depth</b>	2, 6, 8, 10
<b>max_features</b>	"auto", "sqrt", "log2"
<b>min_samples_leaf</b>	20, 40, 100
<b>min_samples_split</b>	10, 20, 40
<b>n_estimators</b>	100, 200, 300, 1000

### 3.1. Avaliação do desempenho

Em problemas de regressão, o erro da hipótese definida pelo modelo pode ser calculada pela distância entre o valor real e o predito pelo modelo [Lorena et al. 2000]. Para avaliar a performance e o desempenho dos modelos construídos em ambos os cenários, foi considerado as métricas R<sup>2</sup> (R-Quadrado ou Coeficiente de determinação), MSE (Erro Quadrático Médio) e MAPE (Erro Percentual Absoluto Médio).

Na Tabela 5 pode-se perceber que o modelo que teve o menor desempenho foi a Regressão Linear com 0.73 de R<sup>2</sup> e 37091849.82 de MSE, e os melhores foram: *Gradient Boosting* com 0.86 de R<sup>2</sup>, 19492281.88 de MSE e 0.32 de MAPE, e *RuleFit* com 0.85, MSE de 19313425.66 e 0.32 de MAPE. Com a diminuição das variáveis de entrada, cenário 2 (Tabela 6), não teve mudança no valor de R<sup>2</sup> para os modelos *Gradient Boosting* e Árvore de Decisão e os mesmos apresentaram melhora no valor de MSE. No cenário 2 o modelo *RuleFit* apresentou melhora em ambas as métricas, indo de 0.85 no R<sup>2</sup> para 0.86 e de 19313425.66 de MSE para 19228521.56.

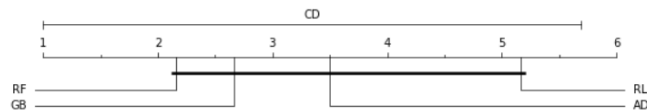
Portanto a redução de dimensionalidade, de 6 para 4 variáveis de entrada contribuiu de forma significativa para a melhora do desempenho dos modelos, todavia o modelo de Regressão Linear foi o mais afetado pela mudança, apresentando piora nos valores das métricas, podendo ser justificado pelos gráficos de diferença crítica (Figura 7) onde pode-se notar um deslocamento para a esquerda dos modelos *RuleFit* (RF) e *Gradient Boosting* (GB) no cenário 2 (Figura 8). A Regressão Linear (RL) manteve as mesmas posições em ambos os gráficos e a Árvore de Decisão (AD) apresentou um leve deslocamento para a direita no cenário 2 mostrando que o *ranking* caiu.

**Tabela 5. Resultados dos modelos no cenário 1**

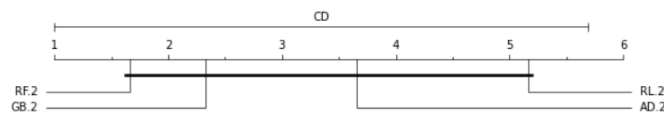
<i>Modelo</i>	<i>R2</i>	<i>MSE</i>	<i>MAPE</i>
<b>Regressão Linear</b>	0.73	37091849.82	0.44
<b>Árvore de Decisão</b>	0.84	21829685.85	0.31
<b>Gradient Boosting</b>	0.86	19492281.88	0.32
<b>RuleFit</b>	0.85	19313425.66	0.32

**Tabela 6. Resultados dos modelos no cenário 2**

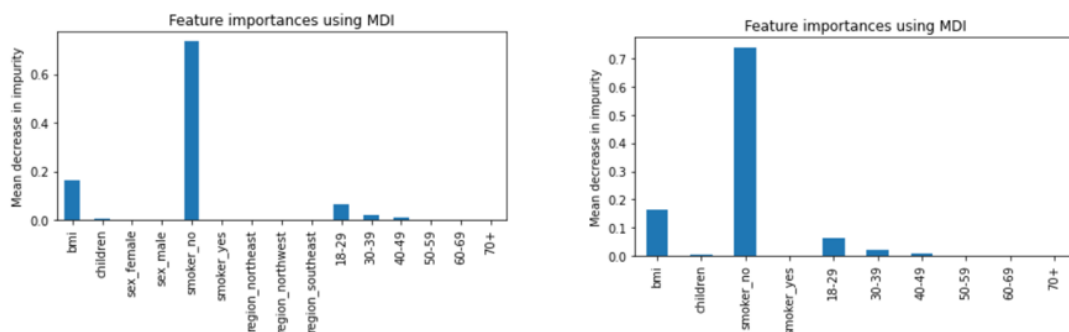
<i>Modelo</i>	<i>R2</i>	<i>MSE</i>	<i>MAPE</i>
<b>Regressão Linear .2</b>	0.72	37222809.17	0.44
<b>Árvore de Decisão .2</b>	0.84	21372143.69	0.32
<b>Gradient Boosting .2</b>	0.86	19382762.62	0.32
<b>RuleFit .2</b>	0.86	19228521.56	0.32



**Figura 7. Diferença crítica dos modelos no cenário 1**



**Figura 8. Diferença crítica dos modelos no cenário 2**

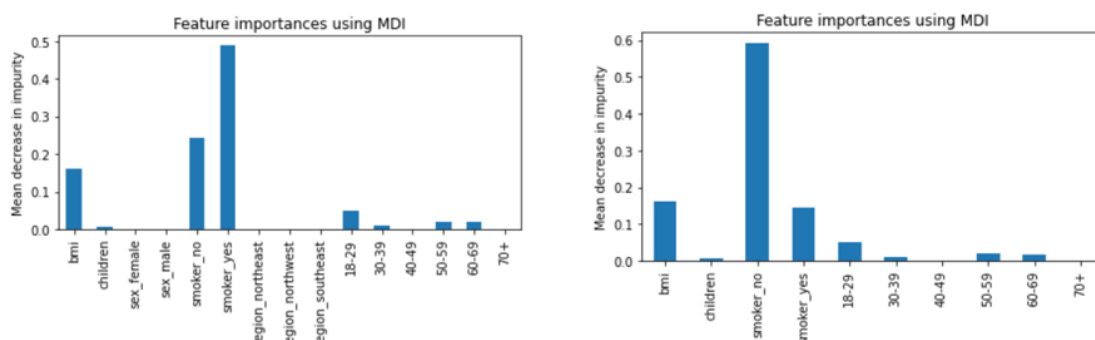


**Figura 9. Importância dos atributos para o modelo Árvore de Decisão- cenário 1 x cenário2**

### 3.2. Avaliação da explicação natural dos modelos

Para os modelos *Gradient Boosting*, *Árvore de decisão* e *RuleFit* foi avaliado a importância dos atributos, além disso para o *Regressão linear* e *RuleFit* foi analisado o valor

dos coeficientes gerado pelos modelos para os atributos em ambos os cenários.



**Figura 10. Importância dos atributos para o modelo *Gradient Boosting*- cenário 1 x cenário2**

Pode-se identificar que não houve mudança significativa na importância das variáveis nos modelos nos cenários 1 e 2, pois a remoção de algumas variáveis no cenário 2 correspondeu com as variáveis que apresentaram menor importância para os modelos no primeiro cenário. Por exemplo nos modelos *Gradient Boosting* (Figura 10) e *Árvore de Decisão* (Figura 9) a variável binária *Smoker*, se a pessoa é fumante ou não, continua sendo uma das mais importantes seguido pela variável de massa corporal (BMI).

	rule	type	coef	support	importance
15	bmi > 30.005000114440918 & smoker_no <= 0.5	rule	13056.070293	0.113074	4134.641829
4	smoker_no	linear	-9902.549075	1.000000	3998.602533
9	18-29	linear	-5735.185831	1.000000	2728.474265
65	bmi > 28.4399995803833 & smoker_yes > 0.5	rule	3877.061888	0.141343	1350.670550
26	60-69 <= 0.5 & smoker_no > 0.5 & 50-59 <= 0.5	rule	-2533.051858	0.586572	1247.396785
13	60-69	linear	4511.609769	1.000000	1163.062275
10	30-39	linear	-2854.137369	1.000000	1102.878818
1	children	linear	673.601423	1.000000	738.054732
27	smoker_no <= 0.5 & bmi > 36.85499954223633 & b...	rule	3323.511448	0.042403	669.708810
48	18-29 <= 0.5 & smoker_no > 0.5	rule	-985.822345	0.508834	492.834234

	rule	type	coef	support	importance
29	bmi > 30.00999927520752 & smoker_no <= 0.5	rule	13991.146621	0.106007	4307.127820
2	smoker_no	linear	-8510.417978	1.000000	3436.466573
4	18-29	linear	-5824.532855	1.000000	2770.980482
48	bmi > 28.4399995803833 & smoker_no <= 0.5	rule	4488.624160	0.141343	1563.723416
25	smoker_yes <= 0.5 & 60-69 <= 0.5 & 50-59 <= 0.5	rule	-2967.169008	0.586572	1313.442532
8	60-69	linear	4806.633549	1.000000	1239.117396
5	30-39	linear	-2867.334678	1.000000	1107.978444
1	children	linear	681.919308	1.000000	747.168510
32	bmi > 36.85499954223633 & smoker_no <= 0.5 & b...	rule	3441.001085	0.042403	693.383723
23	60-69 > 0.5 & smoker_no > 0.5 & children > 3.5	rule	9764.973983	0.003534	579.440983

**Figura 11. Importância dos atributos para o modelo *RuleFit*- cenário 1 x cenário2**

Considerando o modelo *RuleFit*, pode-se notar que as regras geradas com um valor de importância maior em ambos os cenários, foram muitos similares bem como os atributos originais (Figura 11). A regra com o maior coeficiente e maior nível de importância é:  $bmi \geq 30.005000114440918$  &  $smoker\_no \leq 0.5$ , ou seja, se a pessoa tiver o peso acima de 30 e for fumante, quando todas as outras variáveis tiver um valor constante então a variação média do valor do preço do plano de saúde aumenta 13056 (cenário 1) e 13991 dólares (cenário 2). E isso é bem interessante pois, de acordo com o Centro de controle e prevenção de doenças dos Estados Unidos, as pessoas norte americanas com BMI acima de 30 são consideradas obesas e acima de 35 são extremamente obesas, e o modelo conseguiu identificar essa relação de obesidade e ser fumante, implicando em um gasto maior no plano de saúde dessas pessoas por terem uma tendência maior à problemas de saúde.



Outro ponto interessante a ser ressaltado é que em todas as regras geradas pelo modelo *RuleFit*, em ambos os cenários, foi utilizado a variável *Smoker* onde pode-se perceber na Figura 11, que a pessoa não sendo fumante causa uma queda no valor do preço do plano de saúde de 9.902 dólares (cenário 1) e 8.510 dólares (cenário 2). Uma das diferenças identificadas foi a quantidade de regras geradas no cenário 1 foi 16 regras e no cenário 2, mesmo com uma diminuição na quantidade de atributos de entrada, teve 18 regras geradas.

Parametros		Parametros	
Intercept	11098.707519	Intercept	11964.224671
bmi	310.378925	bmi	297.650509
children	632.262709	children	637.461921
sex_female	70.891189	smoker_no	-12095.730960
sex_male	-70.891189	smoker_yes	12095.730960
smoker_no	-12101.831554	18-29	-4816.219431
smoker_yes	12101.831554	30-39	-3190.079832
region_northeast	921.452120	40-49	-127.891624
region_northwest	878.599965	50-59	3343.065573
region_southeast	190.979708	60-69	4791.125314
18-29	-4796.624893	70+	0.000000
30-39	-3196.636245		
40-49	-134.668139		
50-59	3349.871819		
60-69	4778.057459		
70+	0.000000		

**Figura 12. Importância dos atributos para o modelo *RuleFit*- cenário 1 x cenário2**

Os valores dos coeficientes gerados pela Regressão linear foram muito semelhantes em ambos os cenários (Figura 12), assim como os outros modelos analisados anteriormente. Um dos pontos interessantes levantados pelo modelo foi em relação a variável *children* que teve um valor de 632,26 e 637,46 aproximadamente nos coeficientes, o que indica que quando os valores das outras variáveis são constantes, com o aumento de 1 dependente a variação média do valor do preço do seguro aumenta em 632,26 dólares para o cenário 1 e 637,46 dólares para o cenário 2. Além disso para a variável *Smoker* pode-se notar que a variação média no valor do preço do seguro quando a pessoa é fumante é de 12.095,73 dólares para o cenário 2 e 12.101,83 dólares no cenário 2.

Entretanto um dos pontos que chamou mais atenção é em relação a variável de gênero do contratante no cenário 1, pois de acordo com o modelo de Regressão linear se o contratante for mulher tem um aumento de 70 dólares no preço do plano, e essa conclusão gerada pelo modelo pode levantar a hipótese da existência de viés de gênero na base de dados, entretanto os modelos anteriores (*RuleFit*, *Gradient Boosting* e *Árvore de Decisão*) não consideraram o atributo de gênero como relevante para a predição.

Através da avaliação da explicação natural dos modelos pode-se concluir que a variável *Smoker* foi a que apresentou mais importância para todos os modelos, onde se a

pessoa é fumante acaba tendo um valor muito maior no plano de saúde que uma pessoa não fumante. Seguido pela condição da massa corporal da pessoa, onde pessoas mais obesas tendem a ter um gasto maior no plano de saúde.

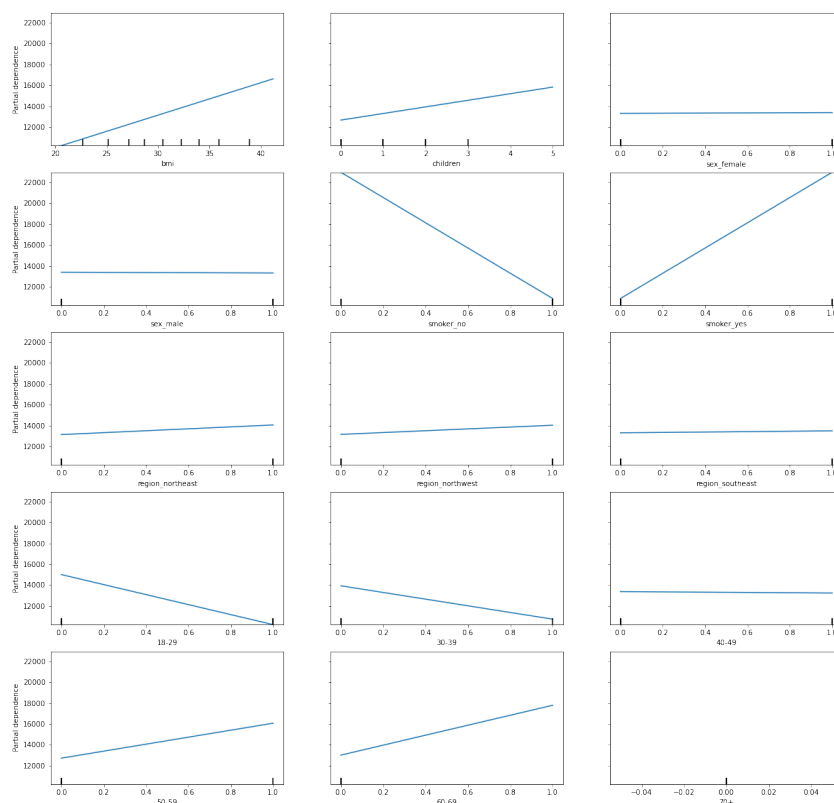


Figura 13. PDP da Regressão linear - cenário 1

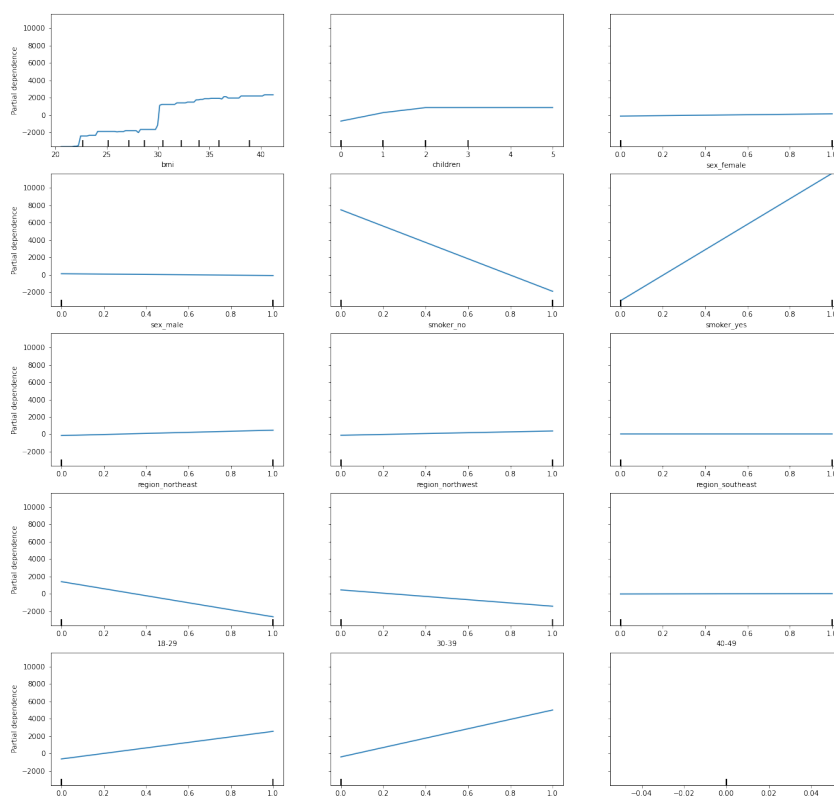
#### 4. Comportamento dos modelos com explicadores globais e locais

Para avaliar o comportamento dos modelos gerados, será considerado os seguintes explicadores globais: *Partial Dependence Plot (PDP)*, *Accumulated Local Effects (ALE)* e Protótipos. E para as explicações locais será considerado: SHAP (*SHapley Additive ex-Planations*), *Counterfactual*, e *Individual Conditional Expectation (ICE)*. As explicações globais foram geradas a partir de base de treino e para a geração das explicações locais foi criado 3 exemplos fictícios fora da base de treino e validação (Tabela 7).

Este capítulo é dividido em subseções com as análises nos resultados de cada um dos explicadores. Importante reforçar que o modelo *RuleFit* só foi avaliado nas sessões 4.2 e 4.3, pois durante a execução das bibliotecas utilizadas nos outros explicadores (PDP, SHAP, *Counterfactual* e ICE) apresentou erro na execução, por não ser suportado pelas bibliotecas utilizadas.

**Tabela 7. Exemplos para as explicações locais**

Age	Sex	BMI	Children	Smoker	Region
65	male	50.5	0	yes	northeast
30	female	30.3	5	yes	southeast
23	male	29.5	2	no	southwest

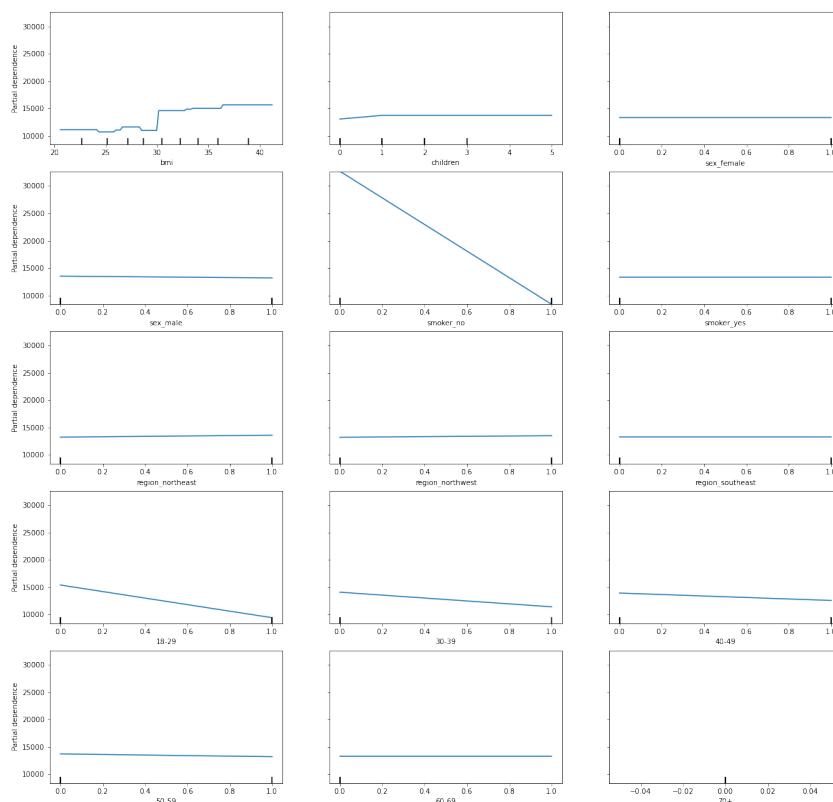


**Figura 14. PDP do *Gradient Boosting* - cenário 1**

#### 4.1. *Partial Dependence Plot (PDP)*

Durante as análises não foi identificada mudança expressiva nos explicadores ao comparar as explicações entre os cenários 1 e 2. Para gerar os gráficos de dependência parcial foi utilizado a biblioteca *Sklearn*.

Olhando para o PDP que mostra o efeito parcial que o atributo tem na predição pode-se perceber que no cenário 1 as variáveis tem efeitos bem parecidos nos modelos (Figura 13, Figura 14, Figura 15), um ponto interessante a ser ressaltado em todos os modelos é que conforme a variável BMI vai aumentando, o efeito local da variável também vai aumentando, ou seja o valor do plano de saúde também aumenta, e quando a massa corporal da pessoa chega em mais ou menos 35 há uma estabilidade no valor, informando

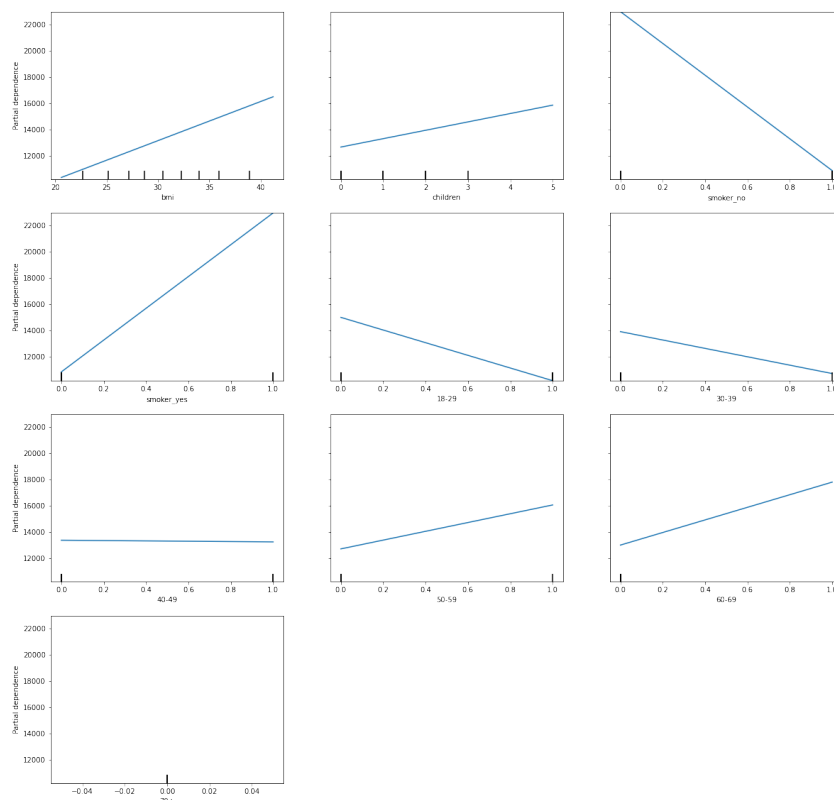


**Figura 15. PDP da Árvore de decisão - cenário 1**

que não há mais influência no valor de predição.

Além disso, na variável *children* pode-se observar que o modelo *Gradient Boosting* o efeito local da variável vai aumentando até chegar em 2 dependentes, depois disso o efeito local se mantém constante, ou seja, não há mais influência no valor de predição. Esse comportamento é bem parecido no modelo de Árvore de decisão, só que o valor do efeito local para de aumentar até chegar em 1 dependente. Entretanto para o modelo de Regressão Linear, pode-se observar um comportamento linear que não fica estável em nenhum valor específico.

Como o restante das outras variáveis são binárias, pode-se observar no caso da Regressão linear e *Gradient Boosting* (Figura 13 e Figura 14) um comportamento semelhante, mas invertido nas variáveis *smoker yes* e *smoker no*, resumidamente indica que quando a pessoa é fumante o efeito local da variável na predição aumenta. Entretanto esse efeito da variável *smoker yes* não é o mesmo para o modelo de Árvore de decisão (Figura 15), cuja variável não apresenta influência na mudança da predição do preço, isso pode ser justificado pelo fato que essa variável não apresenta importância na predição conforme mostrado na Figura 9.



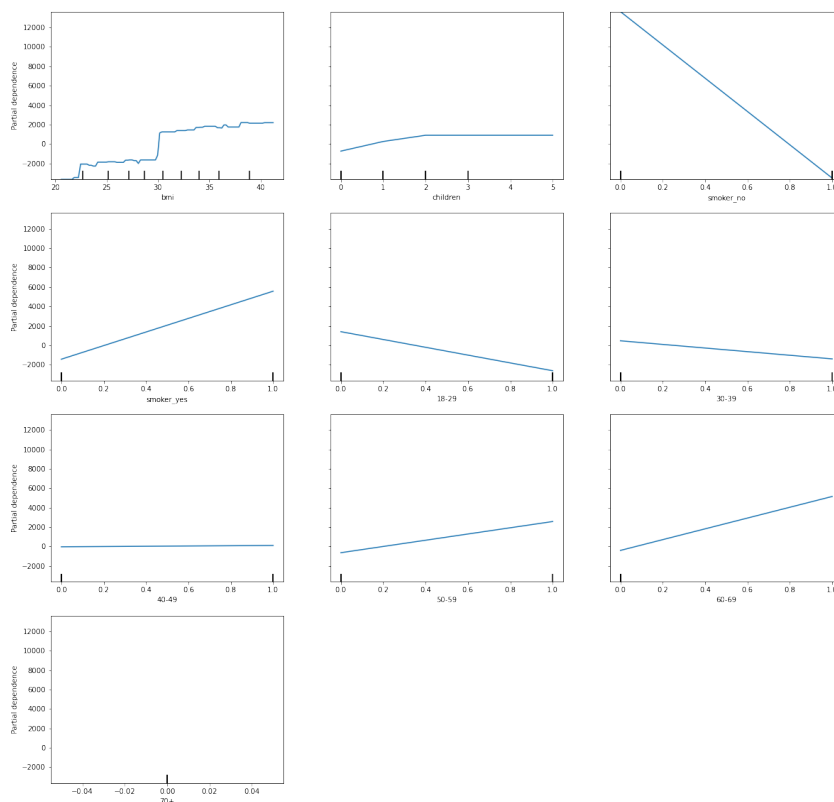
**Figura 16. PDP da Regressão linear - cenário 2**

Em relação as variáveis de idade, quando a pessoa tem idade entre 18-29 ou 30-39 o efeito local é negativo, ou seja, há uma diminuição no preço (Figura 13, Figura 14 e Figura 15). Ademais, observando o comportamento dos modelo *Gradient Boosting* e Regressão linear, pode-se observar que quando a pessoa tem entre 50-59 ou 60-69 anos há um aumento no efeito local na predição do preço do plano de saúde.

Pode-se observar que o mesmo comportamento das variáveis comentados acima para o cenário 1 é semelhante ao cenário 2 (Figura 16, Figura 17 e Figura 18). Entretanto, um ponto de atenção é que a biblioteca utilizada não gera uma distribuição de frequência junto com a visualização do PDP, podendo gerar conclusões equivocadas.

#### **4.2. Accumulated Local Effects (ALE)**

Para gerar os gráficos do ALE foi utilizado a biblioteca *Alibi*, onde descreve como as variáveis influenciam a previsão do modelo de aprendizado de máquina em média, sendo mais imparcial que os gráficos de dependência parcial (PDPs) quando a variável de interesse é altamente correlacionado com outras variáveis. Todavia, ao tirar o valor de correlação de Person (avalia pressupostos lineares) e Spearman (avalia pressupostos não

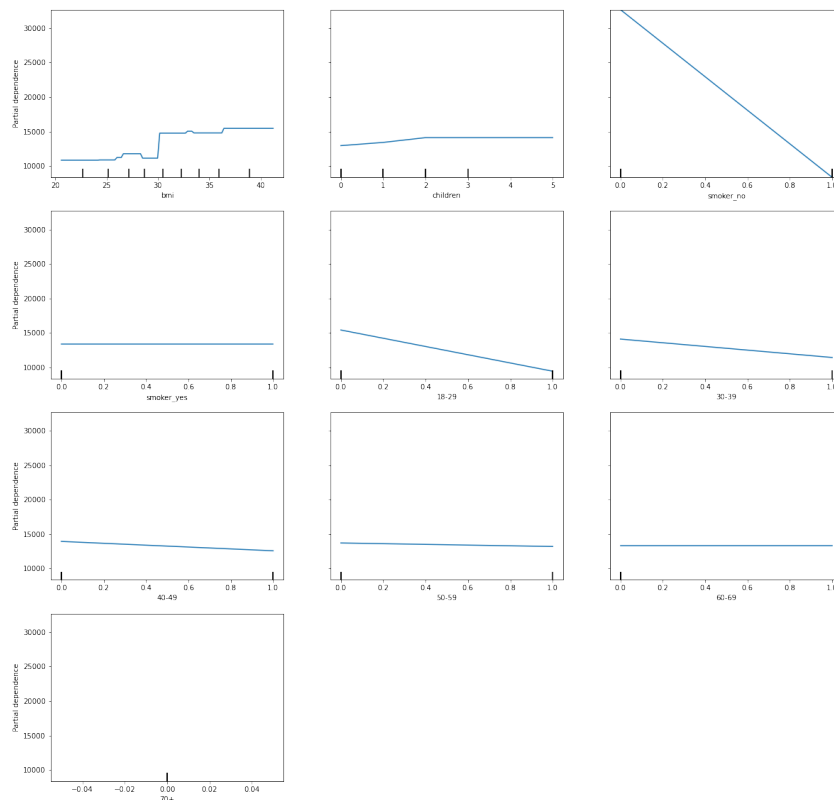


**Figura 17. PDP do *Gradient Boosting* - cenário 2**

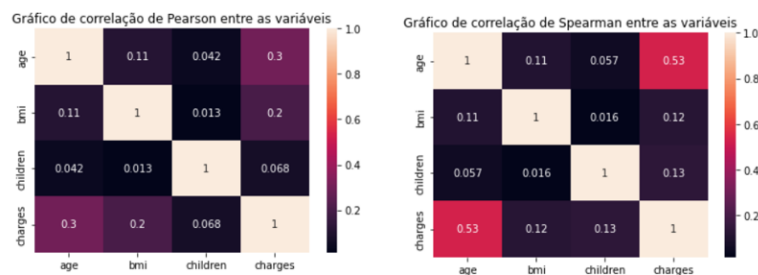
lineares) nas variáveis numéricas não há correlação forte entre as variáveis (Figura 19), levantando a hipótese que não teria uma diferença tão expressiva entre as influências locais das variáveis mostradas no ALE e PDP, mas pode ser que haja uma correlação implícita que esses métodos estatísticos não consideraram.

Nas Figuras 20 e 21 pode-se visualizar os gráficos do ALE gerados com todos os modelos para cada variável, na cor azul há o modelo de Regressão linear, a cor laranja representa a Árvore de decisão, verde para o modelo *RuleFit* e vermelho para o *Gradient Boosting*. Entre os cenários 1 e 2 foi constatado uma diferença bem acentuada na variável *children* para o modelo *Gradient Boosting*, onde no cenário 1 foi observado uma crescente na influência da variável no valor da predição até chegar no valor 4 onde a partir disso não apresenta mais influência local na predição. Já no cenário 2 o valor de influência sobe até chegar no valor 2 e depois disso a variável não apresenta influência local no valor da predição. E essa diferença não foi identificada nos gráficos gerados pelo PDP (Figura 14 e Figura 17).

O modelo *RuleFit* apresentou um comportamento bem semelhante ao do *Gradient Boosting*, só que no primeiro cenário quando o valor da quantidade de dependentes vai



**Figura 18. PDP da Regressão Linear - cenário 2**



**Figura 19. Gráfico de correlações entre as variáveis**

aumentando, também aumenta a predição até chegar no valor 2 e depois disso não há influência no valor de predição. Já no segundo cenário o efeito da variável *children* sobe até chegar no valor 4 que se mantém constante, não apresentando mais influência. Um ponto interessante mostrado nesse gráfico foi que quando a pessoa tem 0 dependentes registrados tem um efeito local de -500 no valor da predição do plano de saúde, ou seja há uma diminuição no valor predito.

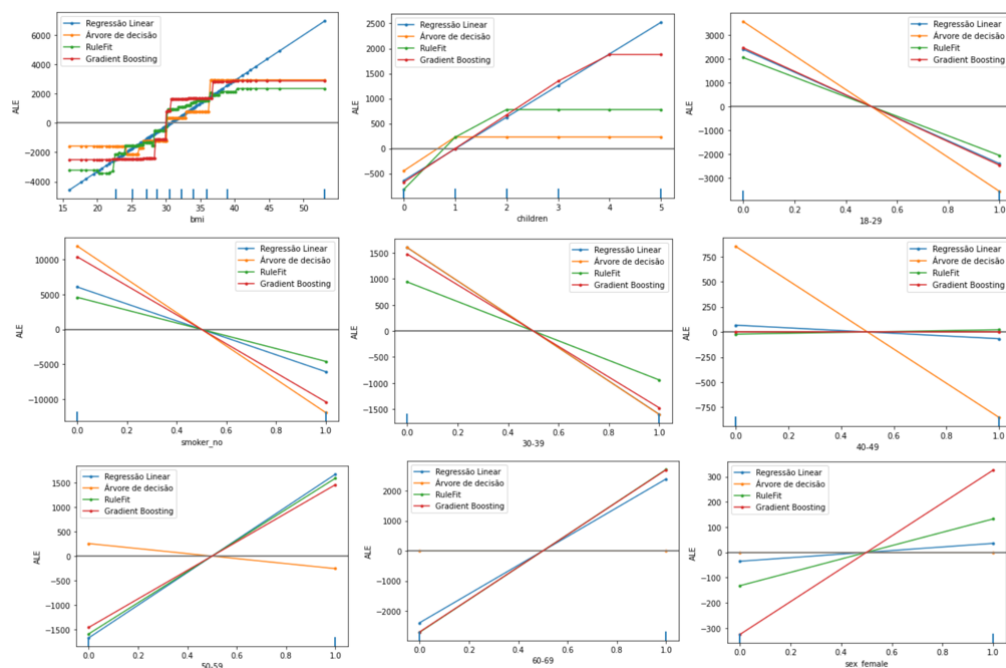


Figura 20. Gráfico ALE - cenário 1

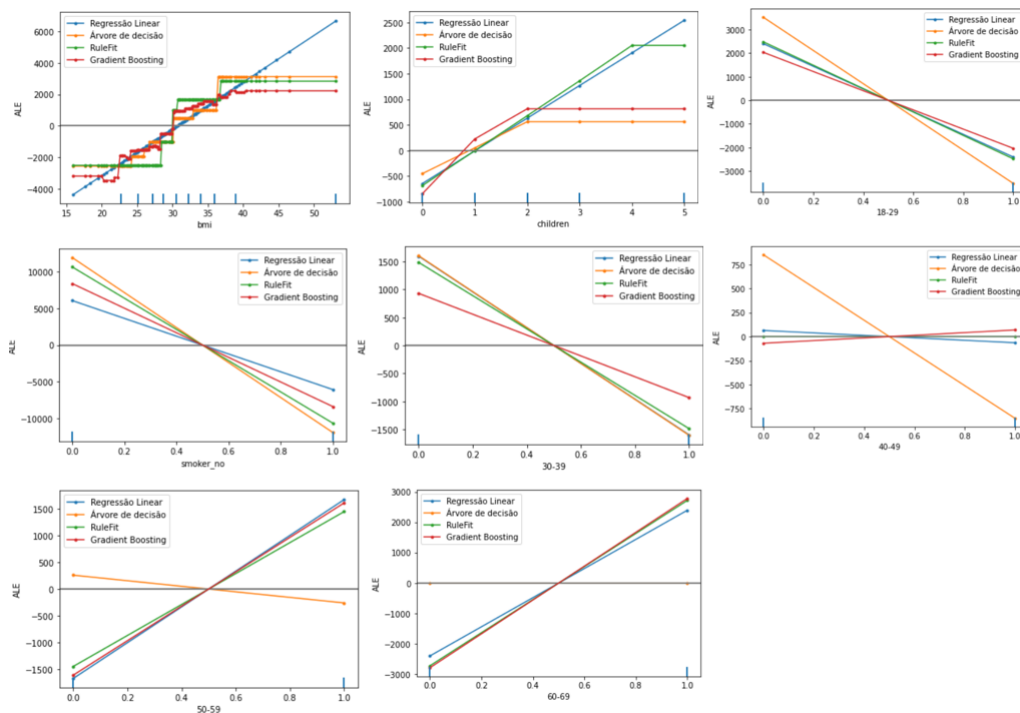


Figura 21. Gráfico ALE - cenário 2

Um ponto importante identificado, foi que no cenário 1 (Figura 20) pode-se perceber que todos os modelos, exceto Árvore de decisão, apresenta viés de gênero, pois quando a pessoa é do gênero masculino tem uma efeito local negativo no valor da predição, já quando é do gênero feminino tem um efeito oposto no preço do plano de



saúde e só com a visualização da importância das variáveis (sessão 3.2) não foi possível identificar esse viés nos modelos *RuleFit* e *Gradient Boosting*.

Outro ponto observado é que quando o contratante for entre 18 e 39 anos o efeito local na predição dos modelos do valor do plano diminui (Figura 20 e Figura 21), entretanto se o contratante tiver entre 50 e 69 anos o efeito local aumenta. Levantando a hipótese que pessoas mais velhas tendem a ser suscetíveis a doenças e problemas de saúde, levando a um gasto maior no plano de saúde.

Uma das diferenças mais acentuadas entre os gráficos ALE e o PDP foi em relação a variável BMI, onde houveram quebras mais constantes nos gráficos gerados com o ALE até a massa corporal chegar em 35, depois disso a variável não apresenta efeito no valor da predição.

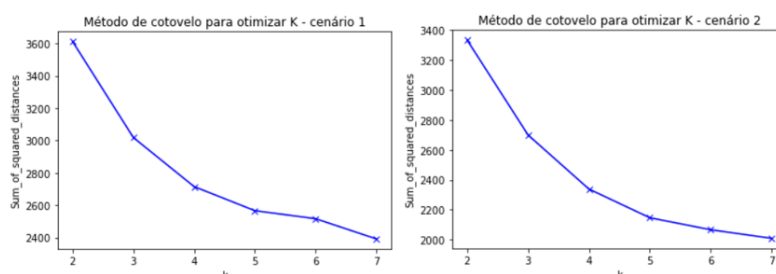


Figura 22. Método de Cotovelo cenário 1 e 2

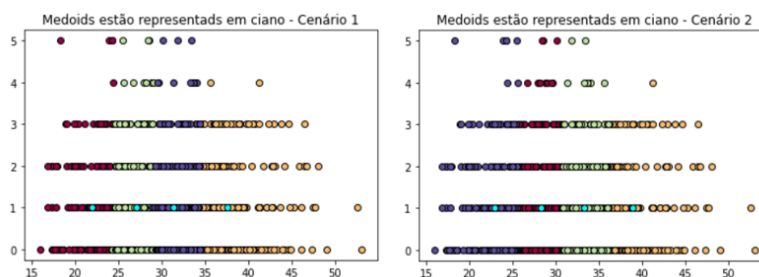


Figura 23. Visualização da localização dos protótipos

### 4.3. Protótipos

Para encontrar pontos representativos na base de dados, foi utilizado o método *K-Medoids*, implementado através da biblioteca do *Scikit-learn*. Para encontrar o melhor valor para K (número de protótipos), foi utilizado o método Cotovelo que busca uma quantidade de agrupamentos (k) em que a soma dos quadrados intra-clusters seja a menor possível, otimizando assim a busca pelo melhor valor de K. Conforme observado na Figura 22 tanto para o cenário 1 quanto para o cenário 2 a quantidade indicada de clusters foi 4, portanto foram gerados 4 protótipos para esses subconjuntos (Tabela 8 e Tabela 10).

Na Figura 23 pode-se visualizar a localização dos protótipos e a distribuição da divisão dos grupos, neste gráfico o eixo X é representado pela variável BMI e o Y *children*, e tanto nos cenários 1 e 2 a divisão dos grupos está bem parecida. Ao analisar os protótipos e as saídas geradas pelos modelos no cenário 1 (Tabela 8 e Tabela 9), o terceiro protótipo

**Tabela 8. Protótipos cenário 1**

BMI	Children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	18-29	30-39	40-49	50-59	60-69	70+
22.00	1	0	1	1	0	0	0	0	1	0	0	0	0	0
37.62	1	1	0	1	0	0	0	1	1	0	0	0	0	0
27.10	1	1	0	1	0	0	0	0	0	0	1	0	0	0
31.40	1	0	1	1	0	0	0	0	1	0	0	0	0	0

recebeu maior valor de predição em todos os modelos, e pode-se observar que ele é o único exemplo em que a pessoa tem entre 40-49 anos, levantando o ponto que o fator da idade contribui para o valor da predição. Ademais em ambos os protótipos que são do gênero feminino tiveram os maiores valores gerados, abordando novamente o cenário de viés de gênero identificado na sessão 4.2.

No cenário 2 não há como identificar a diferença de um grupo pro outro pois todos os protótipos tiveram valores de BMI bem próximos e os valores das outras variáveis binárias estão iguais. Além disso infelizmente ao utilizar esse explicador não há como analisar o impacto da variável *children*, pois todos os protótipos gerados foi com uma quantidade fixa de 1 dependente.

**Tabela 9. Resultados dos modelos nos protótipos - cenário 1**

RL	DT	GB	RF
1589.958933	3315.009129	3204.323699	3804.147595
6770.839820	5514.223668	5459.101028	4869.061158
7976.630580	13047.396528	9995.324520	9265.955518
4507.520824	3315.009129	4982.251584	3915.822706

**Tabela 10. Protótipos cenário 2**

BMI	Children	smoker_no	smoker_yes	18-29	30-39	40-49	50-59	60-69	70+
28.31	1	1	0	1	0	0	0	0	0
38.94	1	1	0	1	0	0	0	0	0
33.33	1	1	0	1	0	0	0	0	0
22.895	1	1	0	1	0	0	0	0	0

**Tabela 11. Resultados dos modelos nos protótipos - cenário 2**

RL	DT	GB	RF
4116.222099	5821.882793	5516.101751	4746.237877
7280.247005	5142.617195	5721.376043	4746.690356
5610.427652	5821.882793	5516.101751	4746.451559
2504.444595	3535.249633	5151.178565	4746.007380

#### 4.4. SHAP (*SHapley Additive exPlanations*)

Neste explicador local foi utilizado o gráfico de força, para avaliar a força que cada atributo teve no resultado da predição para os exemplos apresentados no capítulo 4 e para a execução desse explicador foi utilizado a biblioteca Shap.



Figura 24. Shap dos modelos para o exemplo 1 - cenário 1

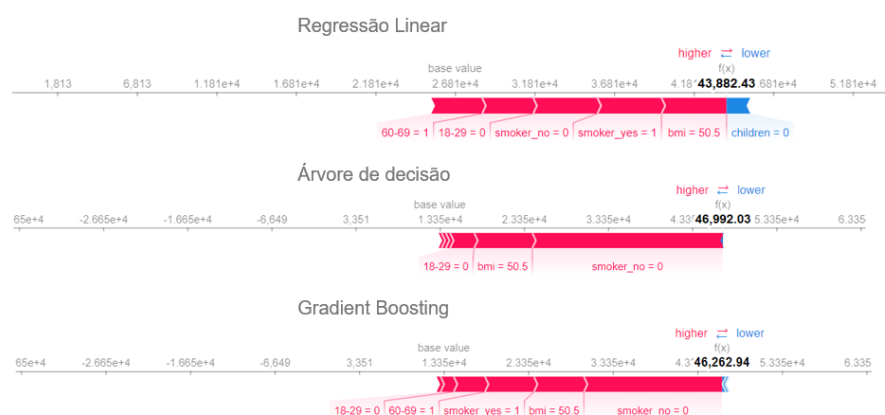


Figura 25. Shap dos modelos para o exemplo 1 - cenário 2

No primeiro exemplo, considerando como entrada uma pessoa de 65 anos, do sexo masculino, com BMI de 50.5, 0 dependentes cadastrados, fumante e mora na região nordeste dos Estados Unidos, pode-se observar no gráfico de força (Figura 24 e Figura 25) que em ambos os cenários as variáveis que apresentaram um aumento para o valor do preço do plano de saúde foram quase as mesmas (lembrando que no cenário 2 houve a redução de algumas variáveis como entrada para o modelo, Capítulo 2). O fato desse contratante ser fumante e acima do peso foi uma das variáveis que contribuiu para o aumento do preço em todos os modelos.

Considerando o segundo exemplo, uma pessoa de 30 anos, do gênero feminino, com BMI de 30.3, 5 dependentes registrados, fumante e mora na região sudeste dos Estados Unidos, novamente o fato da pessoa ser fumante contribuiu para o aumento no preço (Figura 26 e Figura 27), seguido pelo fato de ser acima do peso. Um ponto importante que foi possível identificar foi um problema no código, pois a idade da pessoa entrou no *range* de 18-29 anos, sendo assim necessário o ajuste no código do pré-processamento dos dados. E por ser considerado o *range* de 18-29 anos isso contribuiu para a diminuição do

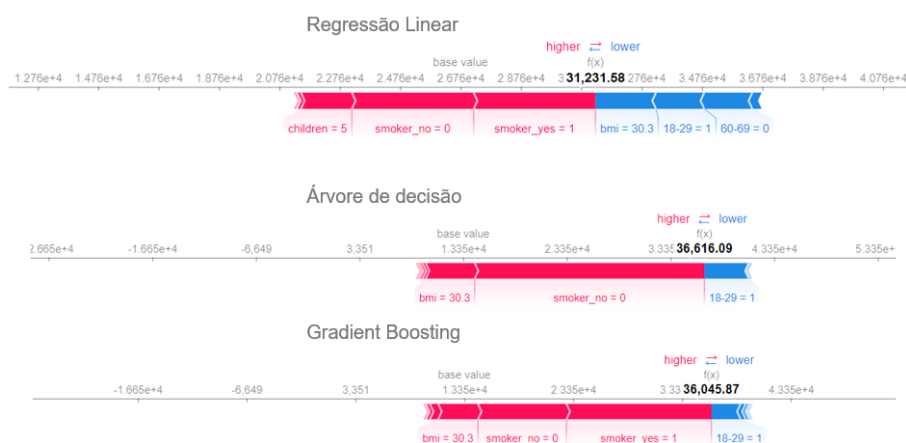


Figura 26. Shap dos modelos para o exemplo 2 - cenário 1

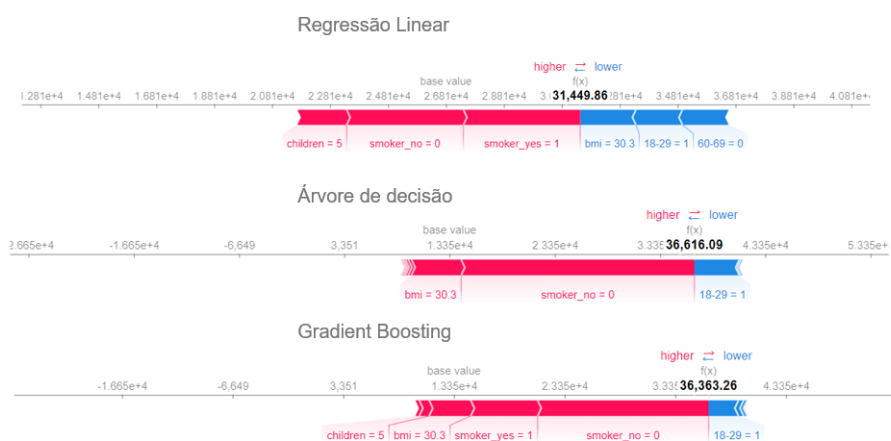


Figura 27. Shap dos modelos para o exemplo 2 - cenário 2

valor de predição. Apesar deste exemplo ter variáveis que contribuíram para a diminuição do preço, o valor de predição continuou alto, pois a força das variáveis que aumentaram era muito maior.

No último exemplo: homem de 23 anos, com BMI de 29.5, com 2 dependentes, não é fumante e mora na região sudoeste, pode-se perceber que o valor de predição do modelo foi bem baixo comparado com os outros exemplos acima (Figura 28 e Figura 29), pois houveram muitas variáveis que contribuíram para a redução do valor de predição, como por exemplo: o fato de não ser fumante, ter entre 18-29 anos, e não ser acima do peso. Entretanto ter 2 dependentes contribuiu para um acréscimo no valor de predição para os modelos de Árvore de decisão e *Gradient Boosting*. Outro ponto interessante é que no modelo de Árvore de decisão no primeiro cenário (Figura 28) demonstrou que o fato dessa pessoa estar morando na região sudoeste também contribuiu para a diminuição do valor.



Figura 28. Shap dos modelos para o exemplo 3 - cenário 1



Figura 29. Shap dos modelos para o exemplo 3 - cenário 2

#### 4.5. Counterfactual

Para a execução desse explicador local foi utilizado a biblioteca Dice, onde para cada exemplo em cada um dos cenários foram gerados 4 exemplos contrafactuais.

Nas Figuras 30 e 31 foi fornecido como entrada o exemplo 1 e pode-se perceber que na maioria dos exemplos contrafactuais gerados, se o contratante não fosse fumante e tivesse um BMI abaixo de 30, teria uma queda no valor do preço do seguro calculado pelos modelos. Considerando o segundo exemplo (Figura 32 e Figura 33), pode-se tirar conclusões bem semelhantes ao do primeiro entretanto essas conclusões levantaram um questionamento em relação a massa corporal de uma pessoa, na Figura 30 foi apresentado nos modelos de *Árvore de Decisão* e *Gradient Boosting* 16,6 para BMI, entretanto de acordo com o Centro de controle e prevenção de doenças dos Estados Unidos, se uma pessoa adulta tem peso menor que 18,5 é considerado abaixo do peso o que acarretaria em um custo maior no plano de saúde pois ela estaria mais suscetível a problemas de saúde.

No último exemplo fornecido, pode-se notar que com o aumento de peso há um aumento no valor da predição (Figura 34 e Figura 35), além disso levanta o fato da pessoa

### Regressão linear

bmi	children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	18-29	30-39	40-49	50-59	60-69	70+	charges
50.5	0	0	1	0	1	1	0	0	0	0	0	0	1	0	44503.0

Diverse Counterfactual set (new outcome: [500, 20000])

bmi	children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	18-29	30-39	40-49	50-59	60-69	70+	charges
19.9	-	-	-	-	0	-	-	0	-	1	-	-	-	-	19707.2
17.9	-	-	-	1	-	-	-	-	1	-	1	-	-	-	17351.8
32.5	-	-	-	-	0	-	-	-	1	-	-	-	0	-	17239.95
31.2	-	-	-	1	-	-	-	-	-	1	-	-	0	-	18436.45

### Árvore de decisão

bmi	children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	18-29	30-39	40-49	50-59	60-69	70+	charges
50.5	0	0	1	0	1	1	0	0	0	0	0	0	1	0	46992.0

Diverse Counterfactual set (new outcome: [500, 20000])

bmi	children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	18-29	30-39	40-49	50-59	60-69	70+	charges
-	-	-	-	1	-	-	-	-	0	-	-	-	-	-	15208.670
-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	15208.670
-	3.0	-	-	1	-	-	-	-	-	-	-	-	-	-	15208.670
-	-	-	-	1	0	-	-	-	-	-	-	-	-	-	15208.670

### Gradient Boosting

bmi	children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	18-29	30-39	40-49	50-59	60-69	70+	charges
50.5	0	0	1	0	1	1	0	0	0	0	0	0	1	0	46057.0

Diverse Counterfactual set (new outcome: [500, 20000])

bmi	children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	18-29	30-39	40-49	50-59	60-69	70+	charges
-	-	-	-	1	0	-	-	0	-	-	-	-	-	-	14897.6
-	-	-	-	1	0	-	-	-	-	-	-	0	-	-	14897.6
26.1	-	-	-	-	0	-	-	-	-	-	-	-	-	1	18636.116
18.9	-	-	-	-	0	-	-	-	0	-	-	-	-	-	16973.1

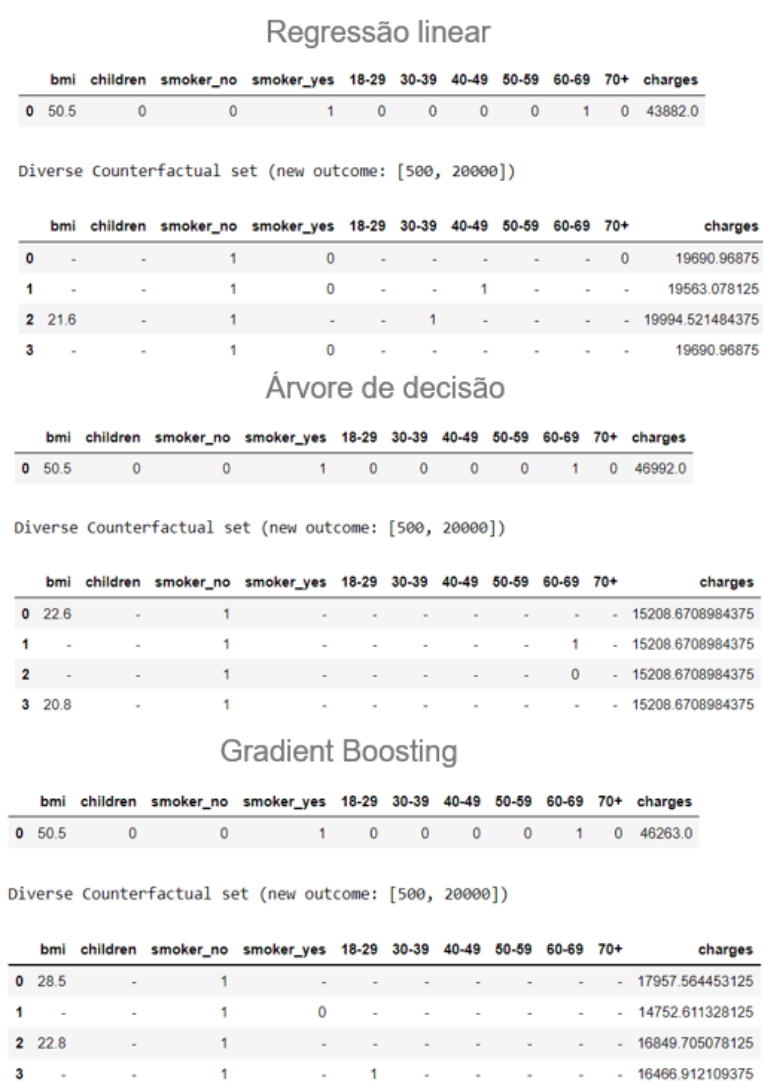
**Figura 30. Counterfactual dos modelos para o exemplo 1 - cenário 1**

não fumar já há uma tendência nos modelos em diminuir o valor predito, podendo ser justificado pelo fato que exemplos gerados onde *Smoker yes* é igual a 1, há um aumento no valor predito para o plano de saúde. Entretanto na Figura 34 foi trazido alguns exemplos um pouco irreais, como o fato da pessoa não ser nem gênero feminino e nem do gênero masculino.

#### 4.6. Individual Conditional Expectation (ICE)

Para gerar essa explicação foi utilizado a biblioteca Shap, e foi fornecido como entrada todos os 3 exemplos utilizados anteriormente.

Pode ser observado que no cenário 1 as amostras se comportam de forma igual para todas as variáveis (Figura 36, Figura 37, Figura 38), onde novamente com o aumento da massa muscular (BMI) indica um preço mais alto em todas as amostras. Um outro ponto interessante é que com o ICE pode-se ver com mais clareza que a quantidade de dependentes nas amostras não exerceram influência para a variação na predição do modelo de Árvore de decisão, e esse comportamento não é o mesmo para o *Gradient Boosting* e para o modelo de Regressão linear, onde demonstram que um aumento na quantidade de



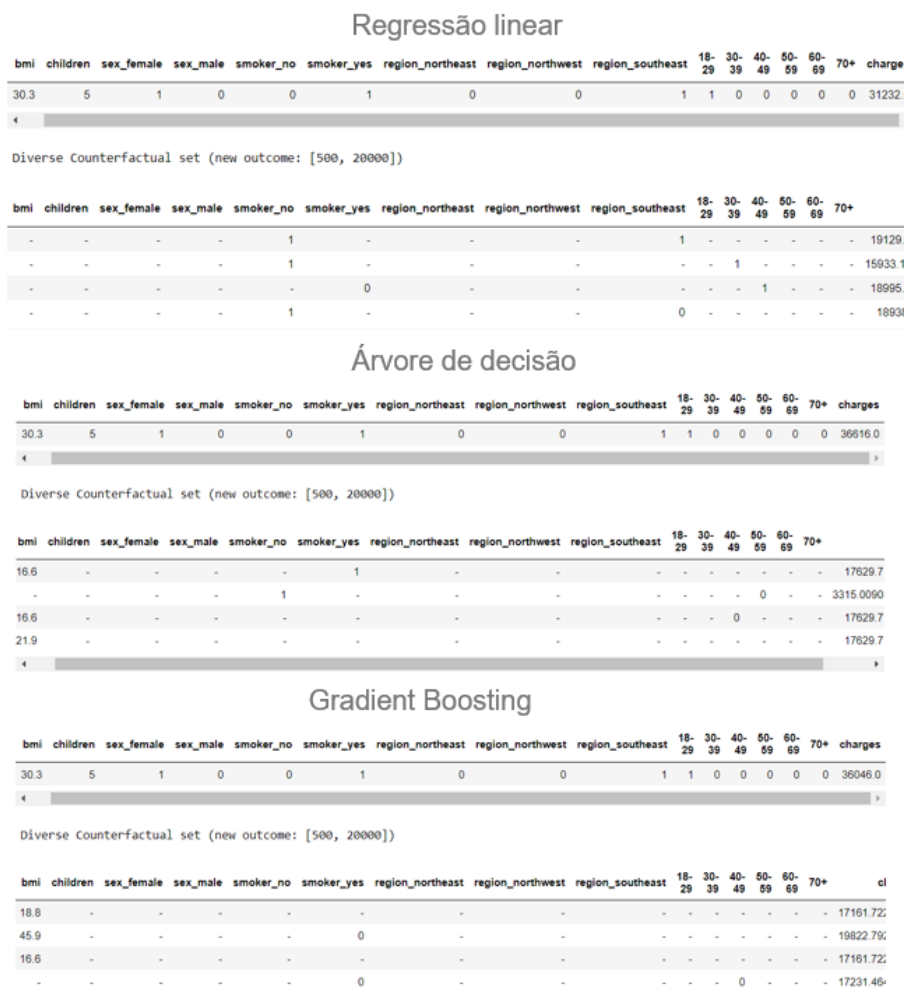
**Figura 31. Counterfactual dos modelos para o exemplo 1 - cenário 2**

dependente indica em uma variação mais alta no preço do plano de saúde.

No cenário 1 (Figura 36, Figura 37, Figura 38) e cenário 2 (Figura 39, Figura 40 e Figura 41) ficou mais evidente a queda no valor da predição nos modelos de Árvore de decisão e *Gradient Boosting* quando as variáveis *Smoker no* e 18-29 são iguais a 1. Uma das diferenças encontradas entre os cenários 1 e 2 foi em relação a Árvore de decisão, pois no cenário 1 não houve influência da quantidade de dependentes no valor da predição, já no cenário 2 apresenta uma leve alteração no valor do preço de forma crescente quando há um aumento na quantidade de dependentes.

## 5. Conclusão

Pode-se concluir que modelos que levaram em consideração as interações entre as variáveis foram os que apresentaram melhor performance para o cálculo do valor do plano de saúde, como por exemplo: Árvore de decisão, *RuleFit* e *Gradient Boosting*. A partir das explicações naturais geradas pelos modelos interpretáveis e entendimento das explicações locais e globais, as variáveis que apresentam maior contribuição e influência



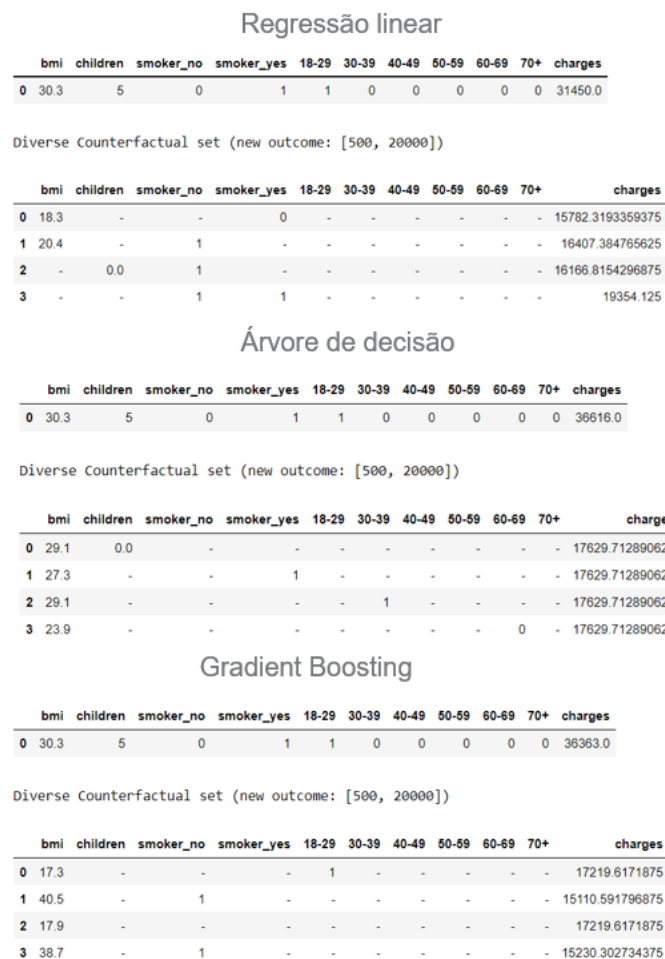
**Figura 32. Counterfactual dos modelos para o exemplo 2 - cenário 1**

na predição é *Smoker*, que informa se a pessoa contratante fuma ou não, e a variável BMI, massa corporal da pessoa. Com a utilização dos explicadores globais e locais (Capítulo 4) foi possível observar que se a pessoa contratante for obesa e fumante ocasiona em um gasto maior no plano de saúde. E pode-se notar que quando o contratante tem 0 dependentes cadastrados gera uma influência negativa no valor do plano de saúde predito, e que vai aumentando conforme o número de dependentes aumenta.

Entretanto foi identificado que os modelos de *Regressão linear*, *Gradient Boosting* e *RuleFit* apresentaram viés de gênero, e um trabalho futuro a ser considerado é investigar porque essa variável gerou o viés ou como alternativa não considerar mais ela como entrada para os modelos, pois como foi comprovado a partir das métricas de desempenho tiradas (sessão 3.1), com a remoção dessa variável não há uma perda na performance dos modelos.

Ademais foi observado no capítulo 4 que o fato de fornecer como entrada para os modelos variáveis que dizem a mesma coisa, como o caso de *Smoker yes* e *Smoker no*, pode-se estar dobrando o valor da importância delas no modelo podendo causar um enviesamento para essas variáveis. Por exemplo o modelo *RuleFit*, que em todas as regras geradas essa variável *Smoker* está presente. No caso do *Gradient Boosting* pode-se





**Figura 33. Counterfactual dos modelos para o exemplo 2 - cenário 2**

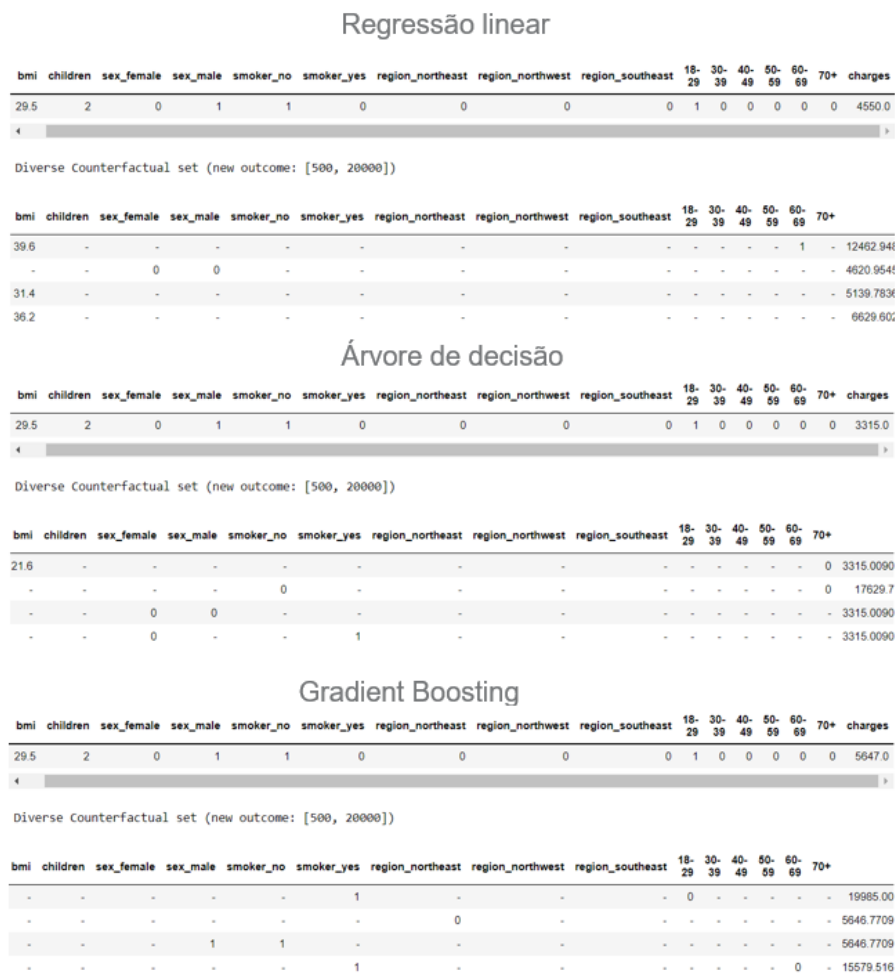
perceber esse viés na Figura 10, onde ambas as variáveis são importantes para a predição.

Todavia esse comportamento não foi observado na Árvore de decisão, que conseguiu balancear melhor o peso de importância e influência das variáveis na predição, por exemplo no gráfico de importância dos atributos (Figura 9), a variável *Smoker no* tem uma grande importância na predição justificado também pelos explicadores globais e locais, do que a variável *Smoker yes* que não exerceu nenhuma influência. E esse comportamento também pode ser observado para as variáveis de gênero e idade.

Durante a análise dos modelos com o explicador Shap na sessão 4.4 foi possível identificar um problema no pré-processamento dos dados, onde no exemplo fornecido como entrada o contratante tem 30 anos e foi considerado que a idade dele está entre 18-29 anos, causando impacto no valor da predição do plano de saúde. E isso ressalta uma das vantagens na utilização de explicadores, ficando para trabalhos futuros realizar o ajuste no pré-processamento.

## Referências

Berchick, E. R., Hood, E., and Barnett, J. C. (2019). *Health insurance coverage in the United States: 2018*. Washington, DC: US Department of Commerce.



**Figura 34. Counterfactual dos modelos para o exemplo 3 - cenário 1**

- Carrin, G. (2002). Social health insurance in developing countries: a continuing challenge. *International social security review*, 55(2):57–69.
- Leung, T. I. and Van Merode, G. (2019). Value-based health care supported by data science. *Fundamentals of clinical data science*. Cham: Springer International Publishing, pages 193–212.
- Lorena, A. C., Gama, J., and Faceli, K. (2000). *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.
- Nayak, B., Bhattacharyya, S. S., and Krishnamoorthy, B. (2019). Democratizing health insurance services; accelerating social inclusion through technology policy of health insurance firms. *Business Strategy & Development*, 2(3):242–252.
- Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., and Abedi, V. (2019). Artificial intelligence transforms the future of health care. *The American journal of medicine*, 132(7):795–801.
- Waring, J., Lindvall, C., and Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, 104:101822.

## Regressão linear

	bmi	children	smoker_no	smoker_yes	18-29	30-39	40-49	50-59	60-69	70+	charges
0	29.5	2	1	0	1	0	0	0	0	0	5108.0

Diverse Counterfactual set (new outcome: [500, 20000])

	bmi	children	smoker_no	smoker_yes	18-29	30-39	40-49	50-59	60-69	70+	charges
0	34.2	-	-	-	-	-	-	-	0	-	6506.845703125
1	46.6	-	1	-	-	-	-	-	-	-	10197.7119140625
2	-	4.0	1	-	-	-	-	-	-	-	6382.81201171875
3	37.7	-	-	-	-	-	-	-	-	-	7548.6220703125

## Árvore de decisão

	bmi	children	smoker_no	smoker_yes	18-29	30-39	40-49	50-59	60-69	70+	charges
0	29.5	2	1	0	1	0	0	0	0	0	8309.0

Diverse Counterfactual set (new outcome: [500, 20000])

	bmi	children	smoker_no	smoker_yes	18-29	30-39	40-49	50-59	60-69	70+	charges
0	44.6	-	-	-	-	-	-	-	-	-	5142.6171875
1	38.4	-	-	-	-	1	-	-	-	-	5142.6171875
2	-	1.0	-	-	-	-	1	-	-	-	5821.8828125
3	-	5.0	-	-	-	-	1	-	-	-	8308.5458984375

## Gradient Boosting

	bmi	children	smoker_no	smoker_yes	18-29	30-39	40-49	50-59	60-69	70+	charges
0	29.5	2	1	0	1	0	0	0	0	0	6433.0

Diverse Counterfactual set (new outcome: [500, 20000])

	bmi	children	smoker_no	smoker_yes	18-29	30-39	40-49	50-59	60-69	70+	charges
0	-	3.0	-	-	-	-	-	-	1	-	13092.5595703125
1	44.6	-	-	-	-	-	-	-	0	-	6072.57958984375
2	47.6	-	-	-	-	-	-	-	-	-	6072.57958984375
3	-	-	0	-	-	-	-	0	-	-	17783.1796875

Figura 35. *Counterfactual* dos modelos para o exemplo 3 - cenário 2

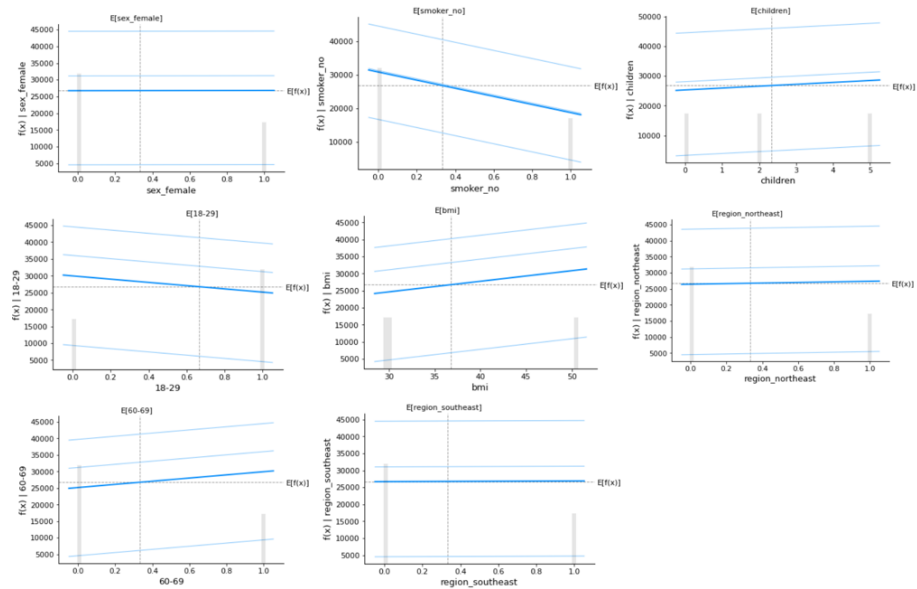


Figura 36. ICE da Regressão Linear para os exemplos no cenário 1

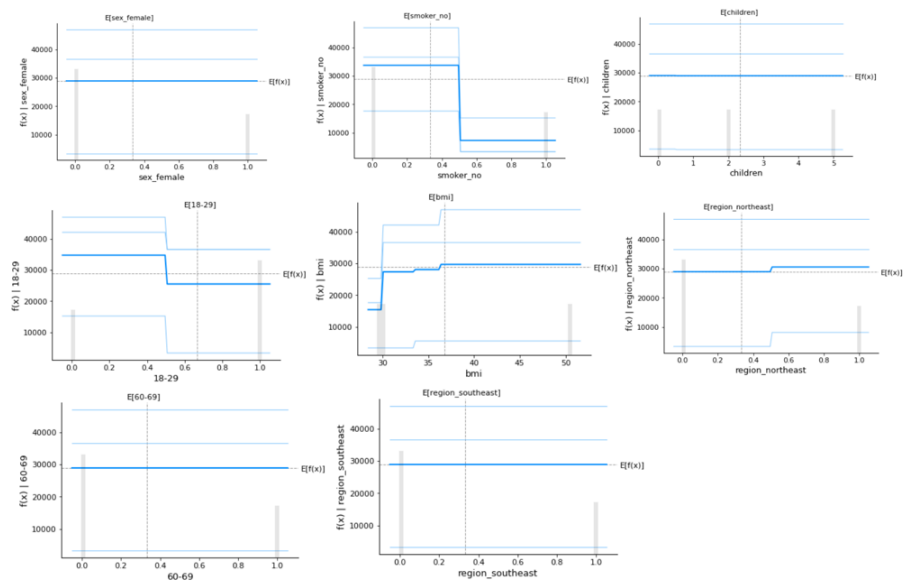


Figura 37. ICE da Árvore de decisão para os exemplos no cenário 1

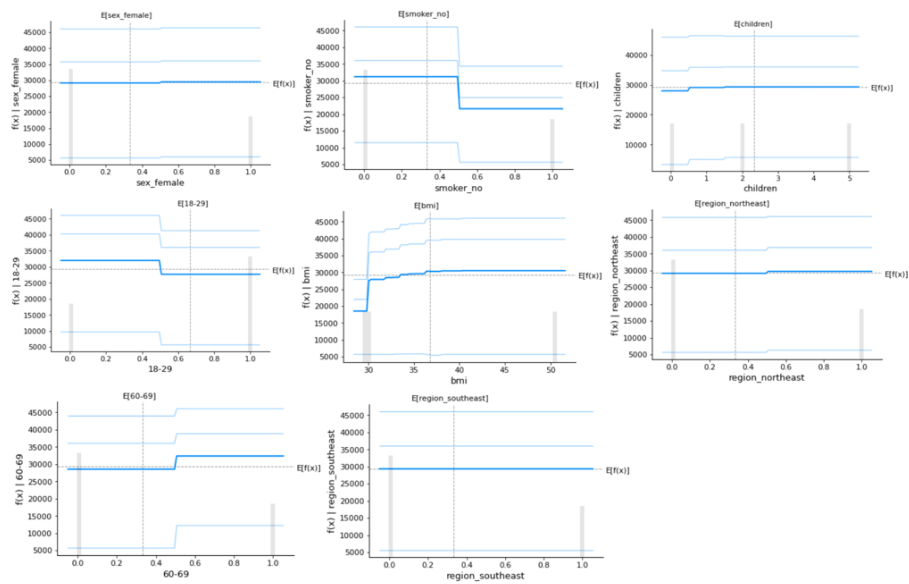


Figura 38. ICE do *Gradient Boosting* para os exemplos no cenário 1

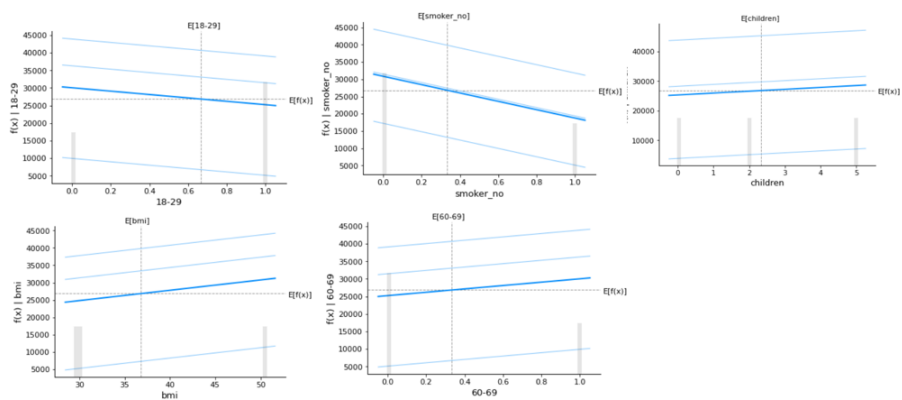


Figura 39. ICE da Regressão Linear para os exemplos no cenário 2

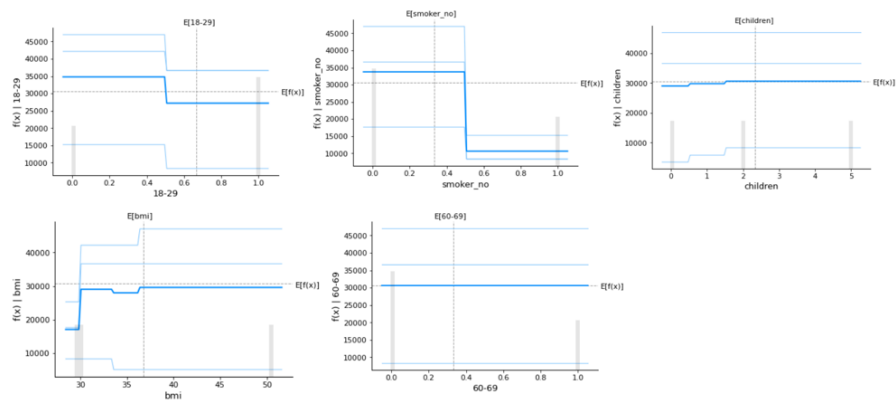
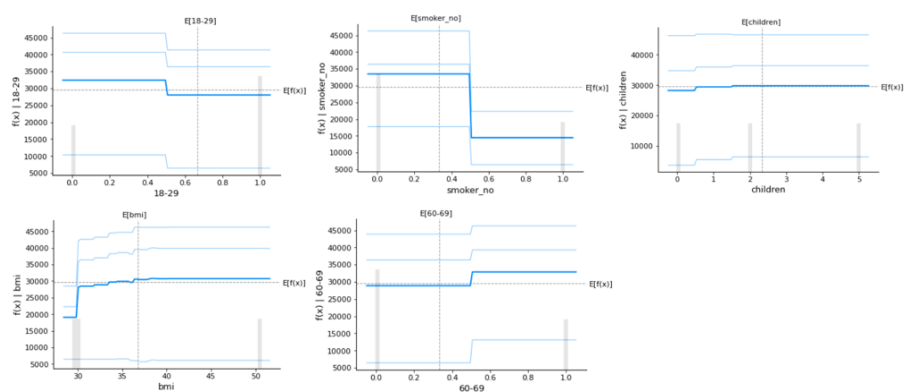


Figura 40. ICE da Árvore de decisão para os exemplos no cenário 2



**Figura 41. ICE do *Gradient Boosting* para os exemplos no cenário 2**