Análise do The Framingham Heart Study dataset

A análise a seguir se refere ao *The Framingham Heart Study* um *dataset* público que contém informações coletadas pela Universidade de Boston e o *National Heart, Lung and Blood Institute* nos EUA no final da década de 1940 e sua finalidade é avaliar quais fatores são de risco para desenvolver alguma doença arterial coronariana (CHD em inglês). Download feito no site https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset. Utilizou-se o método de *machine learning – Decision Tree* feito na linguagem de programação *python* para a análise do *dataset*.

O início do estudo teve pouco mais de 5000 pacientes, as pessoas escolhidas eram saudáveis, entre 30 e 59, da cidade de *Framingham*, *Massachusetts*. Eles participariam por 20 anos do estudo e, a cada dois anos, iriam a um centro médico para realizar exames e preencher questionários sobre seus hábitos de vida, por exemplo, se fumavam ou exercitavam.

age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
50	1.0	1	1.0	0.0	0	1	0	313.0	179.0	92.0	25.97	66.0	86.0	1
51	3.0	1	43.0	0.0	0	0	0	207.0	126.5	80.0	19.71	65.0	68.0	0
52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0	21.47	80.0	107.0	0
40	3.0	0	0.0	0.0	0	1	0	185.0	141.0	98.0	25.60	67.0	72.0	0
39	3.0	1	30.0	0.0	0	0	0	196.0	133.0	86.0	20.91	85.0	80.0	0

Tabela 1 – *dataframe*

Na tabela 1 temos os dados carregados. São 16 variáveis, 15 *features* e 1 *target*. As 15 *features* são os fatores de risco que foram estudados e o *target* é se a pessoa desenvolveu uma CHD ou não durante o estudo. Os fatores de risco são:

- *Male*: 1 indica se é homem e 0 mulher
- *Age*: A idade do participante
- *Education Level*: 1-Ensino médio incompleto, 2- Ensino médio completo, 3-Ensino superior incompleto, 4-Ensino superior completo
- *currentSmoker*: 1- Fumante, 0- Não fumante
- *cigsPerDay*: Número de cigarros por dia
- *BPMeds*: Quantos medicamentos para pressão está tomando
- *prevalentStroke*: 0- Nunca teve derrame, 1-Tem histórico de derrame
- *prevalentHyp*: 0-Não tem hipertensão, 1-Possui hipertensão
- diabetes: 0-Não tem diabetes, 1-Possui diabetes
- *totChol*: Colesterol total
- *sysBP*: Pressão sistólica
- *diaBP*: Pressão diastólica
- *BMI*: Índice de massa corpórea
- *Heart Rate*: Frequência cardíaca em batimentos por minuto
- *Glucose*: Glucose no sangue (mg/dL)

E o target, *TenYearCHD*: 0 significa que a pessoa não desenvolveu nenhuma doença e 1 se desenvolveu.

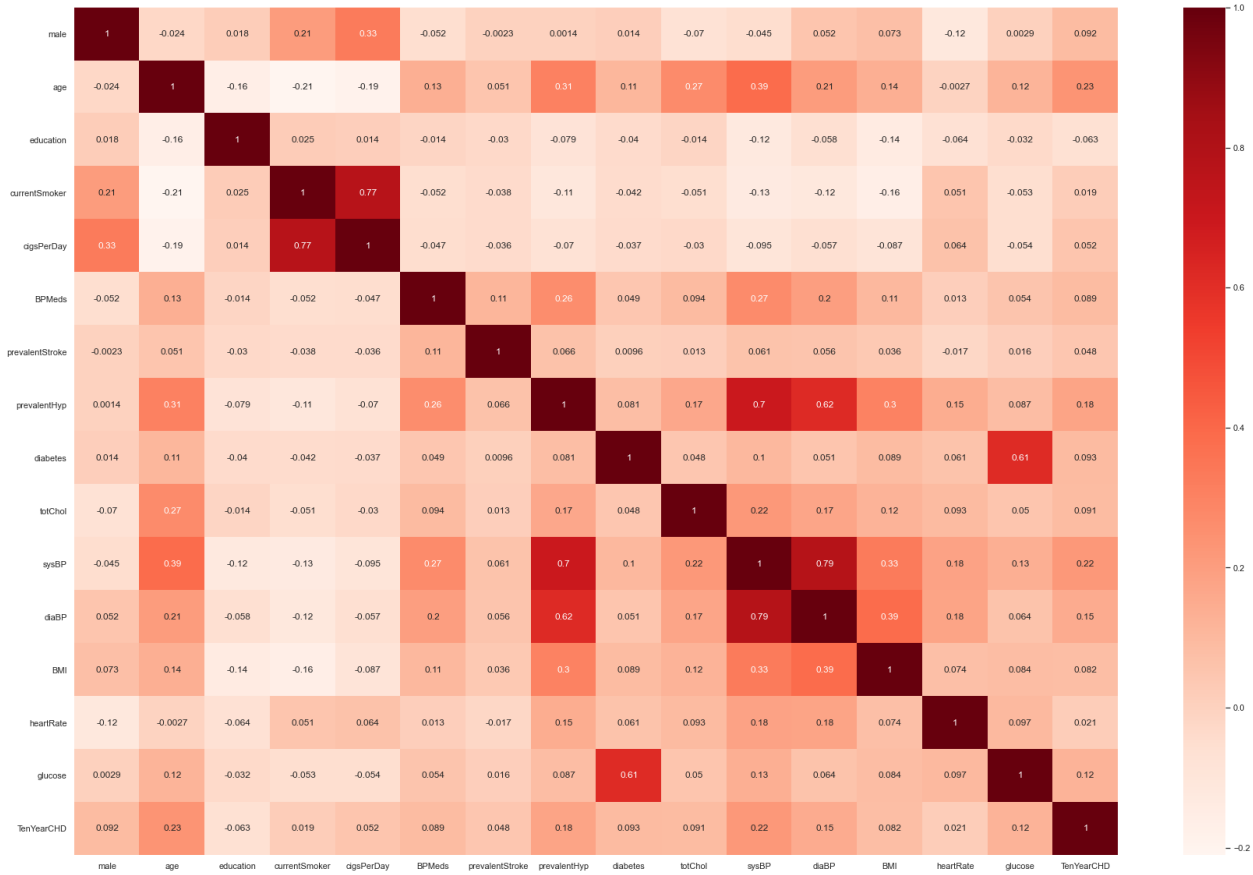

Figura 1 - Heatmap do dataset

Na figura 1 temos o *heatmap* do *dataset* para verificar a relação entre as variáveis e encontrar possíveis relações para explorar. 1 representa que as variáveis são muito relacionadas e -1 que elas são inversamente relacionadas. Vemos, na figura 1, que para *TenYearCHD* e o restante dos fatores de risco, não há valores próximos de 1 ou -1, então foi concluído que não é um único fator de risco que contribui para o desenvolvimento de uma CHD, mas sim o conjunto deles.

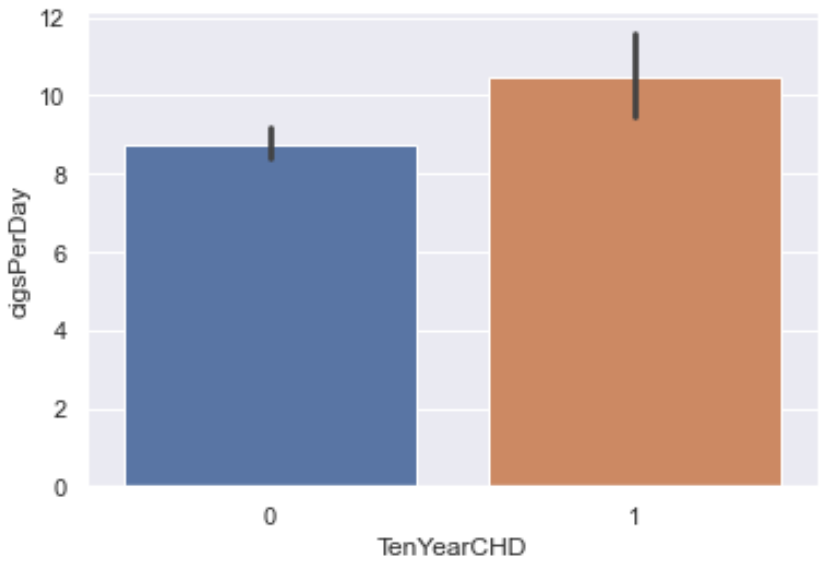

Figura 2 - Gráfico de barras - cigarros por dia e TenYearCHD

Na figura 2, foi observado que, as pessoas que fumam mais cigarros por dia, tem mais chances de desenvolver uma CHD, portanto, fumar mais é um fator de risco.

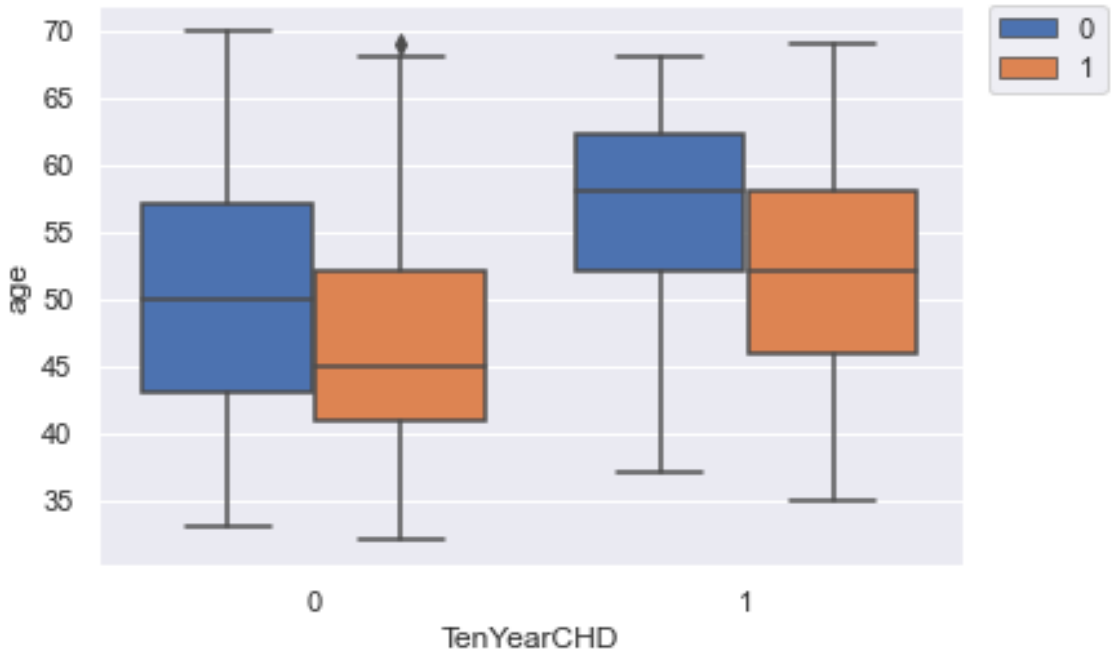

Figura 3 - Boxplot Age e TenYearCHD com hue=currentSmoker

Pessoas mais velhas são mais propensas a desenvolver CHD, como visto na figura 3. Podemos observar que 75 % das pessoas que fumam e desenvolveram CHD tem idade até 57-58 anos, a qual é a mediana para 50 % do grupo que não é fumante, portanto, fumantes têm uma chance maior de desenvolver CHD mais cedo, se comparado com o outro grupo não fumante.

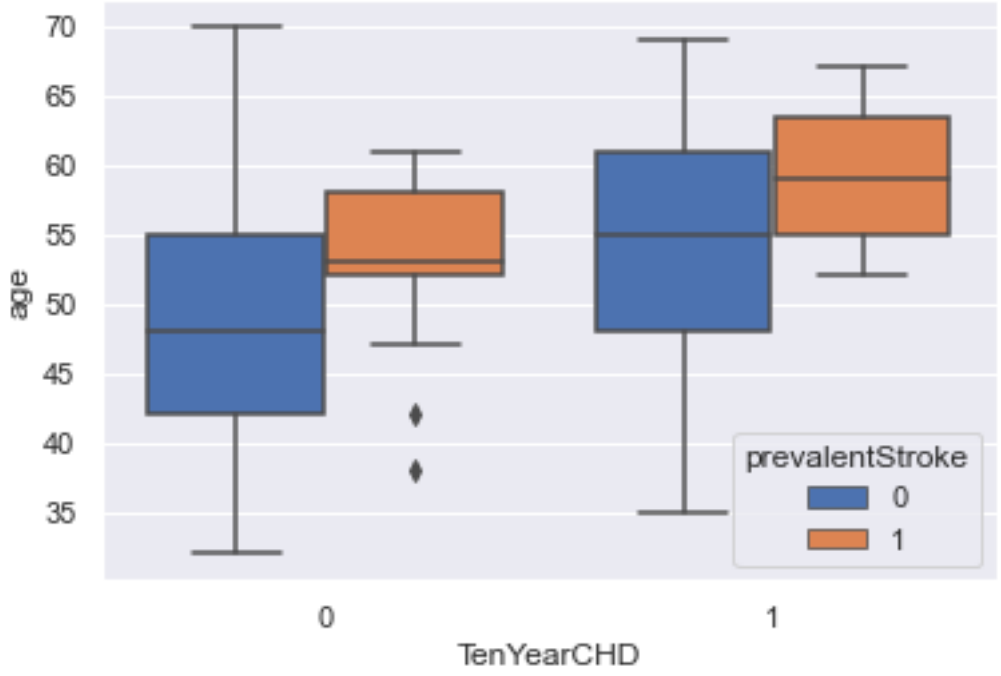

Figura 4 - Boxplot Age e TenYearCHD com hue=prevalentStroke

Nota-se que derrames são mais comuns em idades mais avançadas, visto na figura 4. Temos que as idades mínimas observadas para derrame são 47 e 53, aproximadamente, excluindo os 2 outliers que foram detectados.

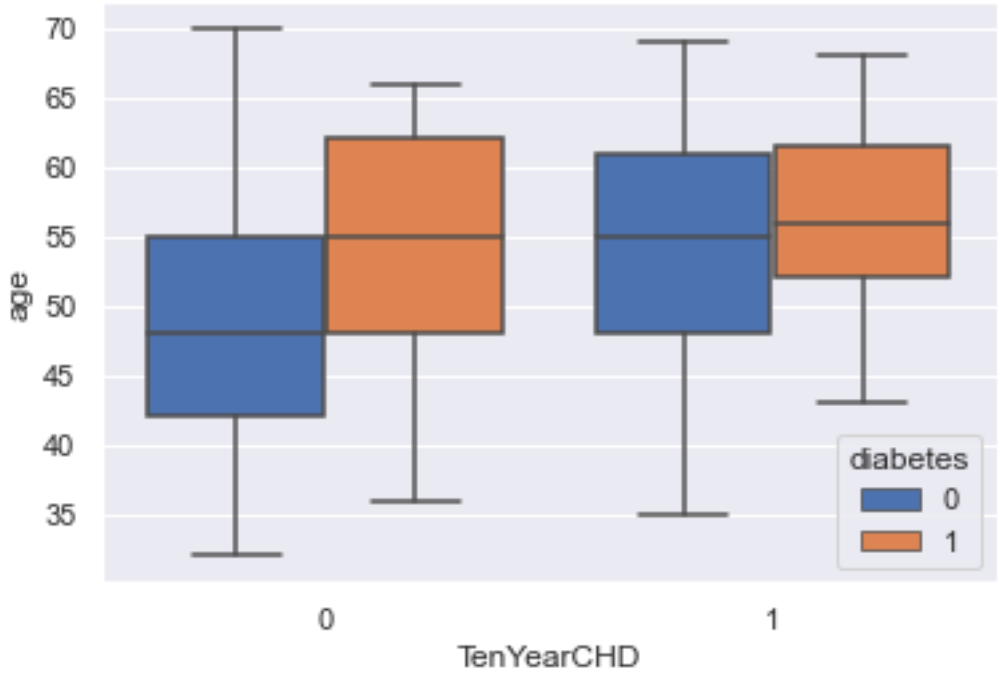

Figura 5 - Boxplot Age e TenYearCHD com hue=diabetes

Na figura 5, observamos o mesmo comportamento da figura passada, a presença de diabetes é mais comum em pessoas mais velhas.

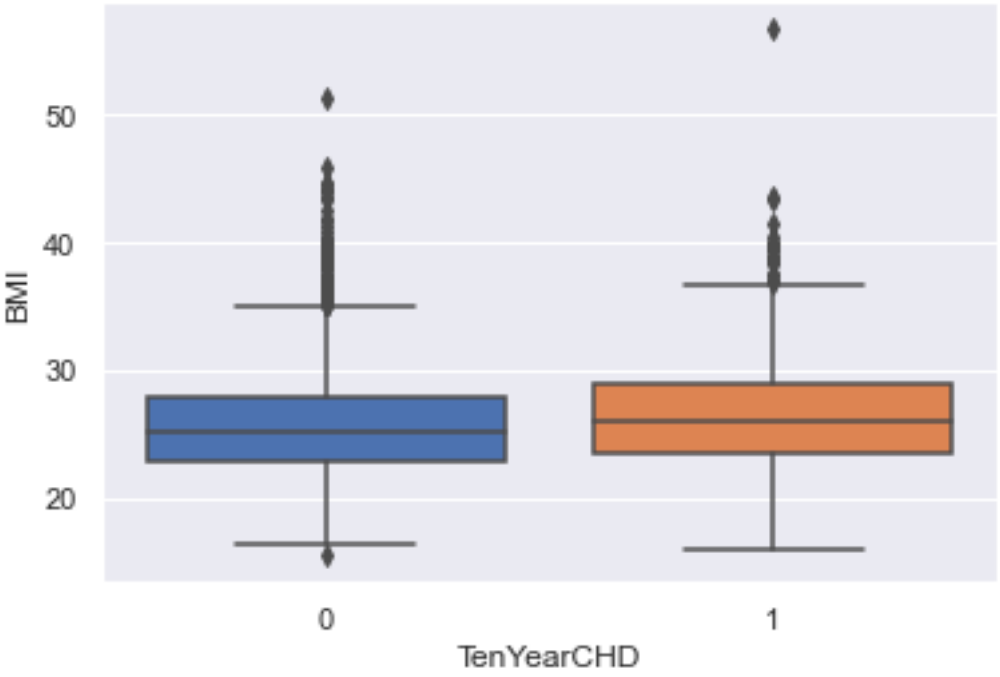

Figura 6 - Boxplot BMI e TenYearCHD

Já na figura 6, foi observado que BMI não é um fator de risco para o desenvolvimento de BMI, pois a diferença observada é baixa.

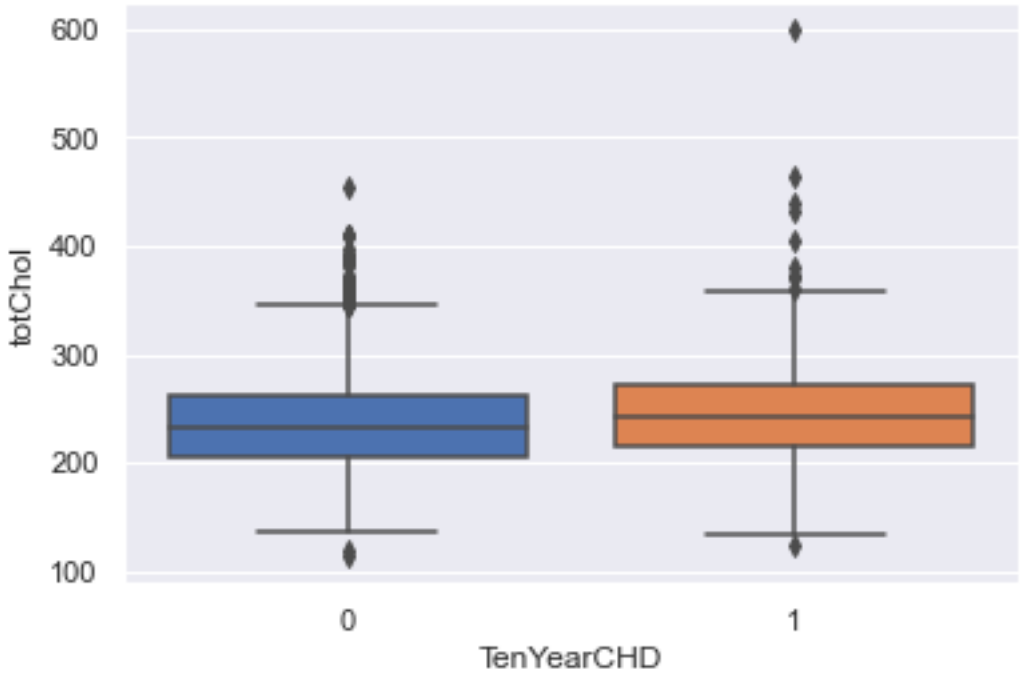

Figura 7 - Boxplot totChol e TenYearCHD

Já é conhecido que colesterol é um fator de risco para doenças do coração, mas é o colesterol LDL que é o fator de risco. O colesterol HDL é benéfico para saúde. Na figura 7, aparentemente colesterol não seria um fator de risco, mas, isso se deve que *totChol* é a combinação dos 2 tipos de colesterol. Caso essa variável fosse divindade em 2, LDL e HDL, teríamos, provavelmente, resultados mais claros, indicando que o LDL é um fator de risco.

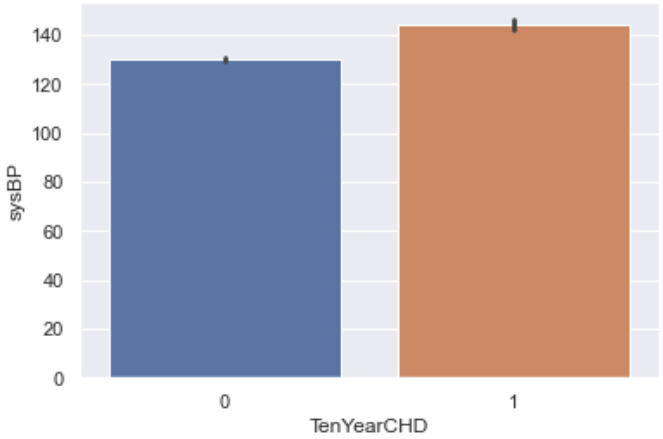

Figura 8 - Gráfico de barras - sysBP e TenYearCHD

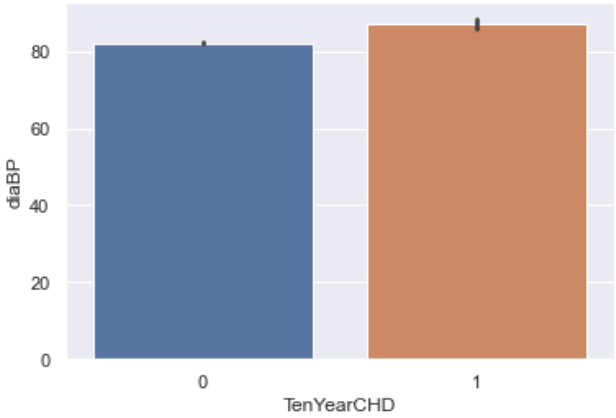

Figura 9 - Gráfico de barras - diasBP e TenYearCHD

Já nas figuras 8 e 9, observa-se que as pessoas com uma pressão arterial mais alta, possuem uma probabilidade maior de desenvolver CHD, ou seja, é um fator de risco.

No *dataset*, são mais de 3600 observações e, apenas 15,23 %, chegou a desenvolver CHD, ou seja, apenas 557 observações positivas. Isso representa um *dataset* desbalanceado que pode gerar problemas para o modelo, por exemplo, *overfitting*. Isso será resolvido aplicando a técnica de SMOTE e Pipeline em conjunto com *undersampling* para obter o melhor modelo possível.

Accuracy	
SMOTE	Pipeline (SMOTE + RandomUnderSampling)
69,67%	80,38%

Tabela 2 – Accuracy do modelo

O modelo proposto para esse estudo foi o *Decision Tree* para prever se uma pessoa irá ou não desenvolver CHD, levando em conta todos os fatores de risco apresentados anteriores. Foram retirados do modelo os fatores: *education*, *totChol*, *BMI* e *heartRat*, a fim de reduzir a complexidade da *Decision Tree* e tendo um ganho em *accuracy*, *precison* e *recall*. Pela tabela 2, temos que com apenas a técnica de SMOTE a *accuracy* foi menor do que com o conjunto de técnicas acoplados com a pipeline. A *accuracy* mede quantos casos foram corretamente identificados, porém não identifica falsos negativos e falsos positivos. Nesse caso do estudo de Framingham, as melhores medidas são *precision* e *recall*, pois identificam falsos positivos e falsos negativos, o que são importantes para um estudo sobre desenvolvimento de doenças. Não queremos ter uma taxa grande de falsos negativos quando se trata de doenças.

Técnica de balanceamento do dataset		
	SMOTE	Pipeline (SMOTE + RandomUnderSampling)
Precision	19,0%	79,0%
Recall	29,0%	82,0%

Tabela 3 – Precision e Recall para casos positivos

Observando as diferenças dos resultados apresentados na tabela 3, é importante notar as suas diferenças. Um modelo com 69,67 % de *accuracy* pode ser aceitável, mas com resultados baixos em *precision* e *recall*, o número grande de falsos negativos e positivos impactam de forma muito negativa a performance geral do modelo, então não é recomendado usá-lo. Já o outro modelo, obteve resultados melhores nas 3 medidas apresentadas, *accuracy*, *precision* e *recall*, então, esse é o modelo mais indicado para esse estudo, já que estamos tentando prever o desenvolvimento de uma CHD.