Conteúdo

1. Introdução

Neste projeto foi aplicado o processo de classificação de machine learning na qual permitiu criar um classificador através de um dataset trabalhando de maneira supervisionada, desta forma a partir de uma série de dados gerar uma classificação de exemplares futuros. O conjunto de dados trabalhado foi obtido do dataset público *cardiovascular disease* disponível no site kaggle, que se trata de um dataset sobre doença coronária (infarto), contemplando dados dos pacientes que tiveram a doença e que não tiveram, bem como comorbidades, idade, histórico familiar de doença coronariana entre outras características.

O dataset doença cardiovascular é uma amostra retrospectiva de homens em uma região de alto risco para doenças cardíacas do Cabo Ocidental, África do Sul, desta forma contém 462 linhas e 11 colunas, sendo uma coluna identificador. Esses dados são retirados de um conjunto de dados maior, descrito em Rousseauw et al, 1983, South African Jornal Médico.

Os tratamentos mais utilizados para o combate dessa doença é a redução da pressão arterial e a redução de outros fatores vistos como potencial após um evento de doença cardíaca coronária no paciente.

Um dos grandes campos de atuação despertados em aulas sobre a inteligência artificial foi o grande nível potencial que machine learning pode contribuir para população em áreas relacionada à saúde no combate e entendimento de doenças, sendo um dos principais motivadores para a escolha desse dataset, que segundo o site http://www.cardiometro.com.br/ é uma das principais coenças causadoras de mortes no Brasil, com 377997 mortes em novembro de 2020. Desta forma o foco do trabalho foi identificar as principais características que levam a desenvolver a doença, assim como outros insights a partir de uma classificação, podendo ajudar um médico a visualizar um melhor tratamento se um paciente tem ou não a doença cardíaca coronariana visto que são mais de uma característica que leva a tal doença.

As características oferecidas pelo dataset com seus detalhes são:

- **sbp -> systolic blood pressure (pressão arterial sistólica):**
  Exemplo: 120x80 significa:
  (PAS) é o maior valor verificado durante a aferição de pressão arterial
  120 refere-se à pressão arterial sistólica e 80 refere-se à pressão arterial diastólica, ambas medidas em milímetros de mercúrio (mmHg).

- **tobacco -> cumulative tobacco (tabaco cumulativo (kg))**

- **ldl -> low density lipoprotein cholesterol** (colesterol de lipoproteína de baixa densidade): É o colesterol ruim porque em altas taxas ela está relacionada com a aterosclerose e, portanto, está também indiretamente relacionada ao infarto e AVC.

- **adiposity -> Adiposidade localizada**, popularmente conhecida como gordura localizada, se refere a um excessivo acúmulo de tecido adiposo em uma determinada parte do corpo.

- **famhist -> histórico familiar de doença cardíaca**, um fator com níveis "ausente" e "presente"

- **typea -> type - A behavior (comportamento tipo A):** São fatores psicológicos que podem influenciar este tipo de doença cardíaca, ou seja, não é apenas desenvolvido com fatores físicos como consumo de tabagismos e álcool entre outros que são comportamentos classificados como B.

  Indivíduos com comportamento do tipo A tem certas desordens de personalidade, particularmente para aqueles que não admitem não serem bem-sucedidos na sua vida. Nesta perspetiva, muitos homens "escolhem" inconscientemente a crise cardíaca, evitando mostrar as suas fraquezas e vulnerabilidades.

  Pacientes do grupo A apresentam excesso de competitividade, perfeccionismo e gostam de realizar várias tarefas no menor tempo possível. Podem ser explosivos. Estabelecem metas e objetivos e quando os

alcançam, continuam a sentir-se insatisfeitos. Ou seja, atitudes geradoras de tensões internas e conflitos internos.

- **obesity -> obesidade:** vetor numérico representando o nível de obesidade do paciente.

- **alcohol -> current alcohol consumption (consumo atual de álcool):**

  Considera-se que o consumo moderado de álcool está associado com um risco reduzido de desenvolver doenças cardiovasculares em comparação com a abstinência ou o consumo em excesso.

  Embora haja evidências de efeitos benéficos do uso moderado das bebidas alcoólicas, existem estudos mostrando um maior risco de doença coronariana associado ao padrão de uso excessivo do álcool.

- **age -> age at onset (idade no início)**

- **chd** -> **response, coronary heart disease**
  resposta, doença cardíaca coronária

Assim foi escolhido o método de classificação com a feature chd como classificador, para classificar futuras observações coletadas de pacientes com as demais características permitindo predizer se tem ou não resposta a doença cardíaca coronária e indicando quais características são mais predominantes e com maior influência para diferenciar se tem maior tendência ou não desenvolver tal doença.

2. DESENVOLVIMENTO

2.1 Preparação dos dados

Importamos o dataset através do arquivo cardiovascular.txt disponível em https://www.kaggle.com/yassinehamdaoui1/cardiovascular-disease, na qual as divisões dos dados foram feitas por ponto-virgula, e afim de uma padronização e melhor entendimento na importação foi feito um tratamento para todos os possiveis valores faltantes ser importado como NA, além de habilitarmos o header do dataset para ter referência do que se trata cada coluna.

Verificamos a quantidade de dados de dados vazios por feature e nenhuma coluna apresentou dados perdidos, como podemos ver na tabela abaixo.

	V1
ind	0
sbp	0
tobacco	0
ldl	0
adiposity	0
famhist	0
typea	0
obesity	0
alcohol	0
age	0
chd	0

Realizando uma visualização geral com as 5 primeiras linhas do dataset, podemos observar que a primeira coluna se refere a uma feature de identificação, e não é relevante para exploração, desta forma optamos por removê-la. Outra observação é que existem dados numéricos inteiros exemplo idade e dados em números com ponto flutuante exemplo tabaco,

assim analisando o tipo de cada feature, vimos diferença com tipos num e int, além da feature famhist que é do tipo caracteres.

ind	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
1	160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
2	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
3	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
4	170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
5	134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1

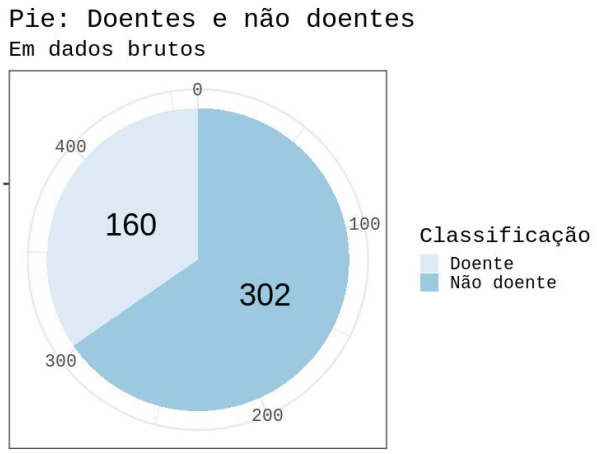
```
$ ind      : int  1 2 3 4 5 6 7 8 9 10 ...
$ sbp      : int  160 144 118 170 134 132 142 114 114 132 ...
$ tobacco  : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
$ ldl      : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
$ adiposity: num  23.1 28.6 32.3 38 27.8 ...
$ famhist  : chr  "Present" "Absent" "Present" "Present" ...
$ typea    : int  49 55 52 51 60 62 59 62 49 69 ...
$ obesity  : num  25.3 28.9 29.1 32 26 ...
$ alcohol  : num  97.2 2.06 3.81 24.26 57.34 ...
$ age      : int  52 63 46 58 49 45 38 58 29 53 ...
$ chd      : int  1 1 0 1 1 0 0 1 0 1 ...
```

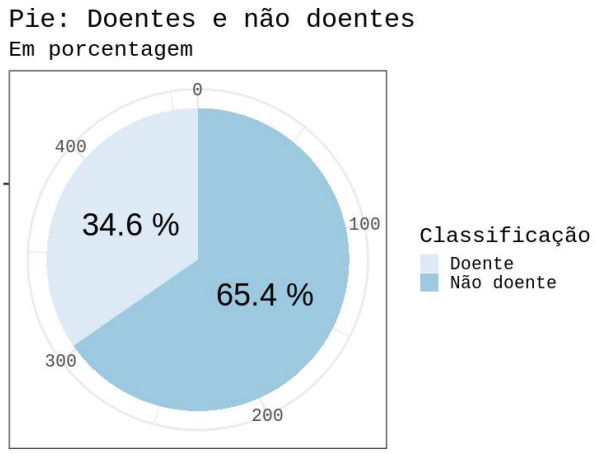
Assim com limpeza de dados inicial padronizamos a features como tipo number, substituindo em famhist “Present” como 1 e “Absent” como 2.

```
$ sbp      : num  160 144 118 170 134 132 142 114 114 132 ...
$ tobacco  : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
$ ldl      : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
$ adiposity: num  23.1 28.6 32.3 38 27.8 ...
$ famhist  : num  1 2 1 1 1 1 2 1 1 1 ...
$ typea    : num  49 55 52 51 60 62 59 62 49 69 ...
$ obesity  : num  25.3 28.9 29.1 32 26 ...
$ alcohol  : num  97.2 2.06 3.81 24.26 57.34 ...
$ age      : num  52 63 46 58 49 45 38 58 29 53 ...
$ chd      : num  1 1 0 1 1 0 0 1 0 1 ...
```

2.2 Visualização dos dados

A primeira exploração estatística feita foi a respeito do balanceamento dos exemplares em relação a feature chd, ou seja, quantos observações contém resposta positiva para doença cardíaca coronariana ou não. Como resultado percebemos um desbalanceamento sendo 160 exemplares (34,6%) com presença da doença e 302 sem a presença da doença (65, 4%).

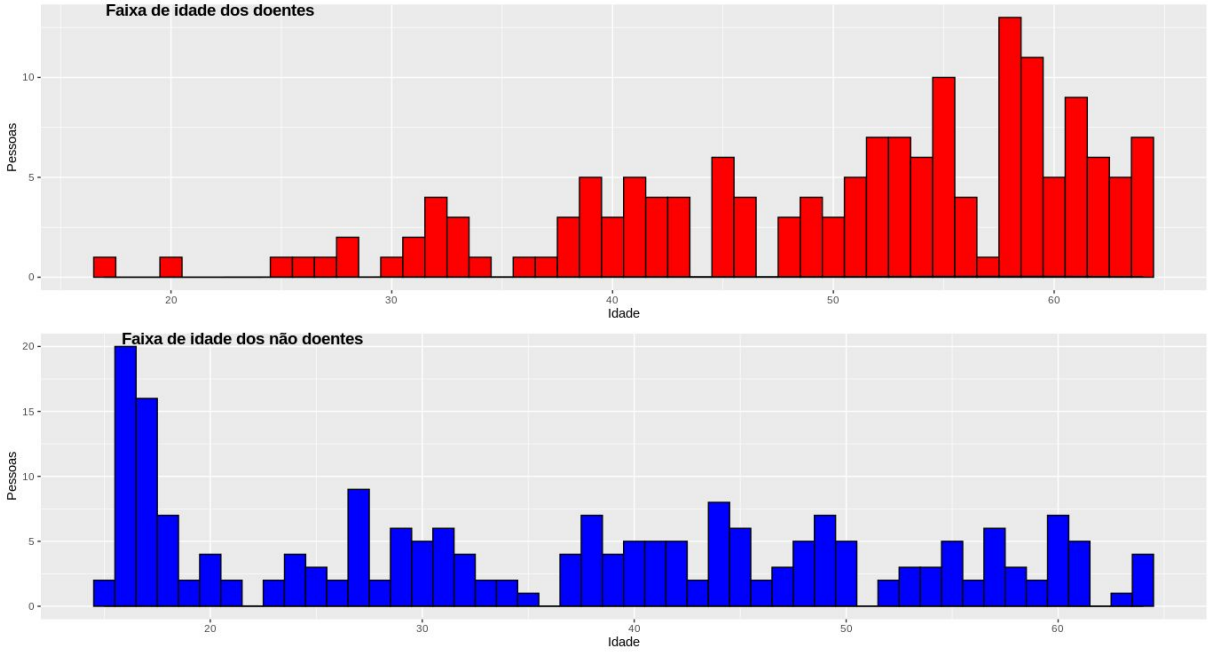





```
# Gráfico de pizza: Doentes e não doentes
pie_dataframe <- data.frame(
  group = c("Doente", "Não doente"),
  value = c(sum(data$chd == 1), sum(data$chd == 0)) # 160
tiveram doença na coronária e 302 Não tiveram doença na
coronária
)

pie_dataframe_rate <- ggplot(pie_dataframe, aes(x = "", y =
value, fill = group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  scale_fill_brewer(palette="Blues") + theme_minimal() +
  geom_text(aes(x = 1,label = paste(round(value / sum(value)
*100,1),"%"), family = "sans"),
            position = position_stack(vjust = 0.5), size = 10)+
  labs(fill = "Classificação",
       x = NULL,
       y = NULL,
       title = "Pie: Doentes e não doentes",
       subtitle = "Em porcentagem") +
  theme_bw(base_size = 20, base_family = "mono")

pie_dataframe_no_rate <- ggplot(pie_dataframe, aes(x = "", y =
value, fill = group)) +
  geom_bar(width = 1, stat = "identity") +
```

```
  coord_polar("y", start = 0) +
  scale_fill_brewer(palette="Blues") + theme_minimal() +
  geom_text(aes(x = 1,label = value, family = "sans"),
            position = position_stack(vjust = 0.5), size = 10)+
  labs(fill = "Classificação",
       x = NULL,
       y = NULL,
       title = "Pie: Doentes e não doentes",
       subtitle = "Em dados brutos") +
  theme_bw(base_size = 20, base_family = "mono")

ggarrange(pie_dataframe_no_rate, pie_dataframe_rate, ncol = 2)
```

Por ser um dataset desbalanceado nos algoritmos de classificação de machine learning pode ocorrer alguma influência da classificação para o conjunto de dados mais presente, podendo ocorrer maior número de falso negativo do que falso positivo na matriz de confusão.

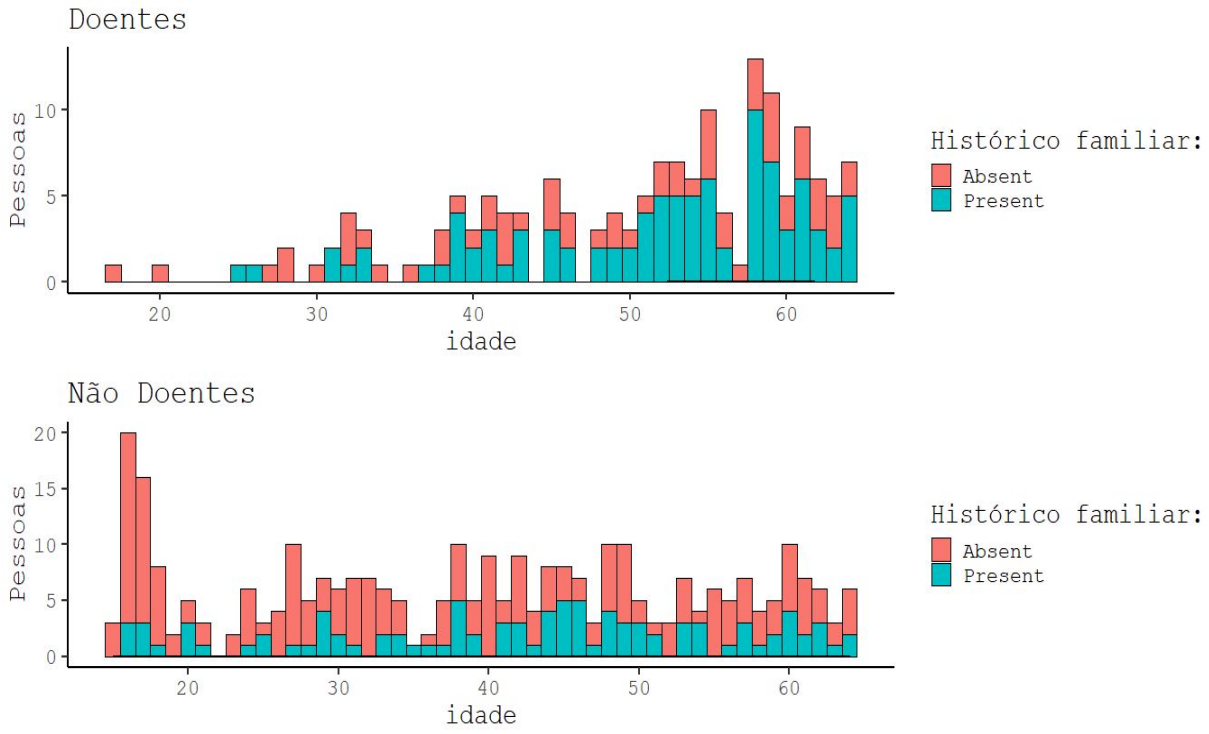



```
# Filtro exemplares que tem e não tem resposta a doença
cardiaca
coronary = filter(data, chd == 1)
```

```
nonCoronary = filter(data, chd == 0)

# Faixa de idade do dataset
ageCoronary <- ggplot(coronary, aes(x=age)) +
  geom_histogram(color="black", fill="red", bins = 30, binwidth
= 1)+ geom_density(fill="#FF6666") +
  scale_x_continuous(name = "Idade") +
  scale_y_continuous(name = "Pessoas")

ageNoncoranary <- ggplot(nonCoronary, aes(x=age)) +
  geom_histogram(color="black", fill="blue", bins = 30,
binwidth = 1)+ geom_density(fill="#FF6666") +
  scale_x_continuous(name = "Idade") +
  scale_y_continuous(name = "Pessoas")

ggarrange(ageCoronary, ageNoncoranary, labels = c("Faixa de
idade dos doentes", "Faixa de idade dos não doentes"), nrow =
2)
```

Analisamos a faixa de idade do dataset em relação a feature chd, e podemos observar que pessoas com idades mais avançadas tende a ter uma maior probabilidade de ter a doença, tendo uma crescente considerável a partir dos 50 anos, com pico de ocorrências na faixa de idade próxima à 58, 59, 60 anos.

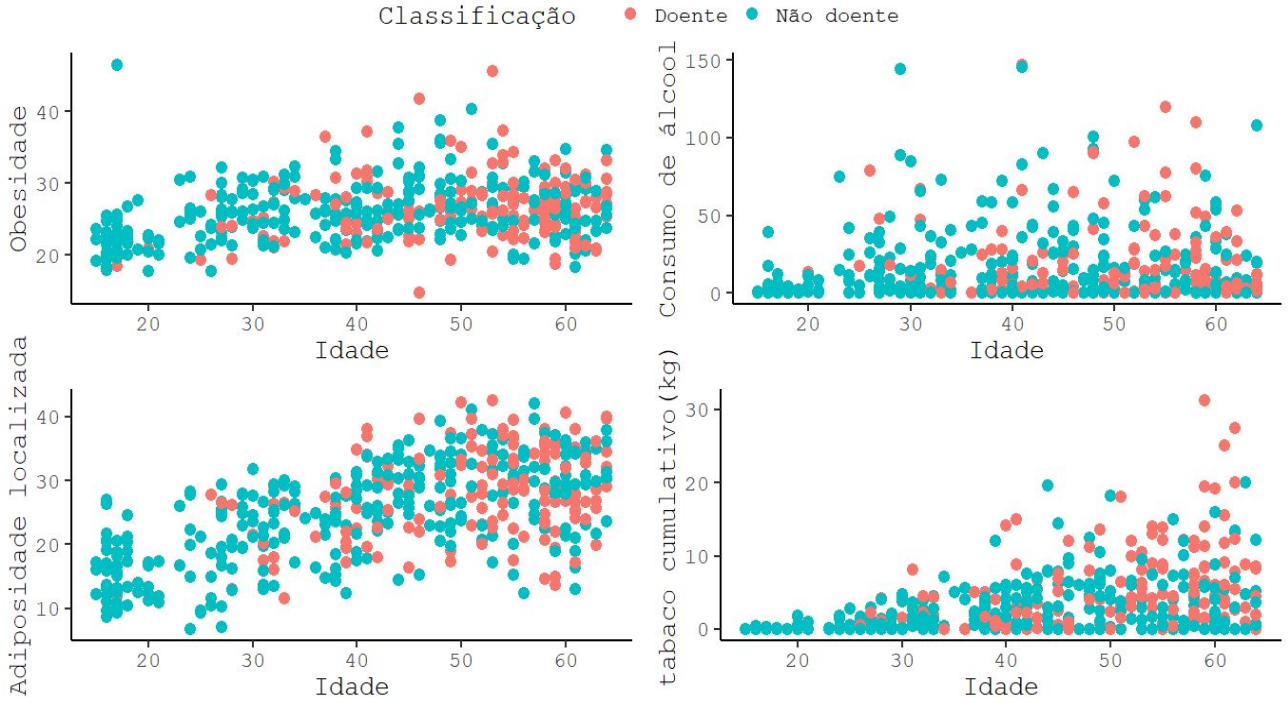
E o pico de pessoas que não desenvolveram a doença está em regiões da adolescência próxima a 16 e 17 anos, sendo possível visualizar que mesmo pessoas com idades avançadas existem pacientes que não tem o desenvolvimento da doença, portanto, por mais que seja um fator que contribui, o avançado de idade não é uma verdade absoluta que a doença cardíaca coronariana vai estar presente, desta forma, realizamos explorações de outras features que possam contribuir nessa relação com idade e presença da doença.




```r
# Quantos dos exemplares que desenvolveram a doença possuiam historico familiar
ageCoronary <- ggplot(coronary, aes(x=age, fill=famhist)) +
  geom_histogram(color="black", bins = 30, binwidth = 1)+
geom_density(fill="#FF6666") +
  labs(fill = "Histórico familiar:",
       x = "idade",
       y = "Pessoas",
       title = "Doentes") +
  theme_classic(base_size = 20, base_family = "mono")

ageNoncoranary <- ggplot(nonCoronary, aes(x=age, group = famhist, fill=famhist)) +
  geom_histogram(color="black", bins = 30, binwidth = 1)+
geom_density(fill="#FF6666") +
  labs(fill = "Histórico familiar:",
       x = "idade",
       y = "Pessoas",
       title = "Não Doentes") +
  theme_classic(base_size = 20, base_family = "mono")
```

```
ggarrange(ageCoronary, ageNoncoranary, nrow = 2)
```

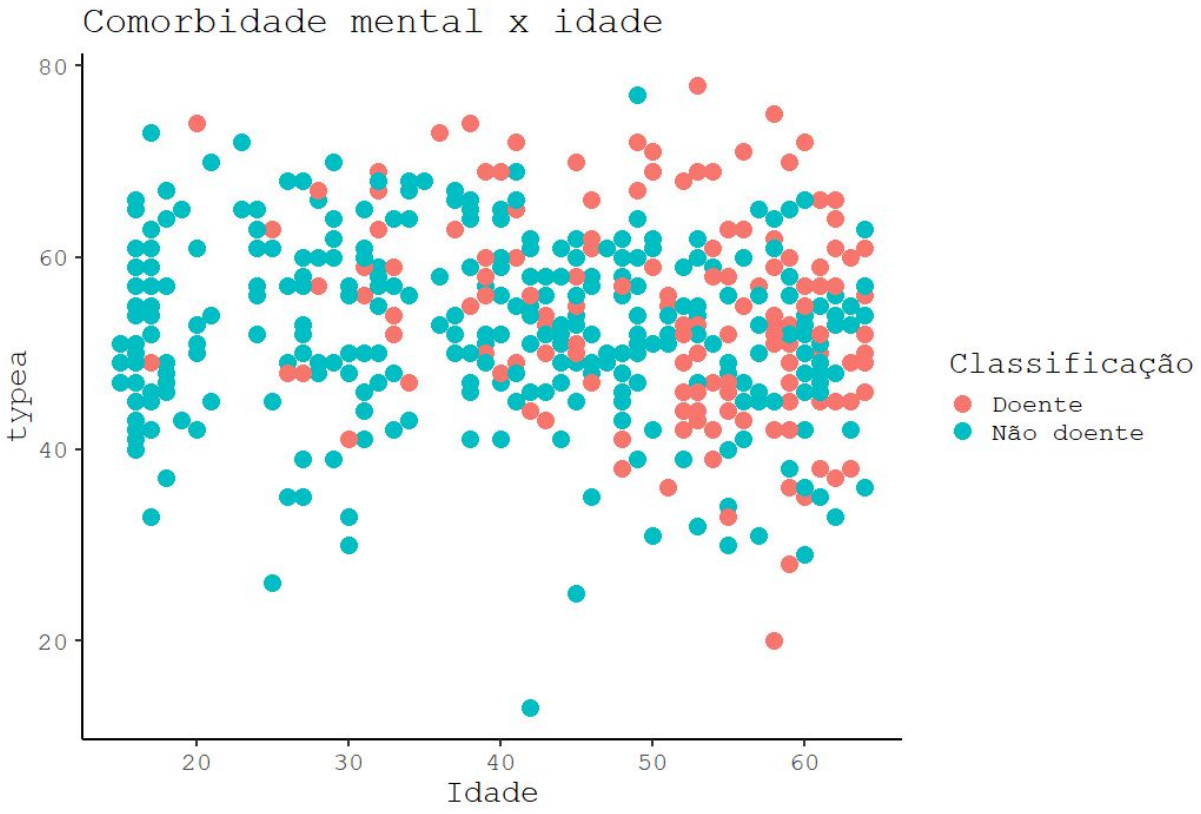
Foi contabilizado quanto exemplares tem em seu histórico familiar a presença da doença sendo 158 observações que tiveram e 214 que não tiveram a doença cardíaca coronariana. Gerando gráfico de quantidade de pessoas a presença da doença cardíaca ao longo da idade filtrando pelo histórico familiar sendo número *"Present"* representando a presença de doença na família e *"Absent"* a ausência dessa doença em seu histórico familiar, podemos observar que a maior parte das pessoas que desenvolvem a doença quando chegam em idades avançadas tem o histórico familiar positivo para presença da doença, portanto, a doença coronariana pode ter fatores hereditários que contribuem o seu aparecimento, e junto a outras característica no caso a idade contribui ainda mais para o seu surgimento. Da mesma forma o gráfico nos indica que pessoas jovens próxima à faixa de idade entre 15 e 20 anos, se apresentar um histórico familiar ausente, existe uma probabilidade de ter resposta a doença cardíaca muito baixa, desta maneira, essa informação poderia auxiliar médicos em diagnósticos onde se tem o conhecimento da idade e dos histórico familiar de seus pacientes.



```r
# Comparando as principais comorbidades x idade: tobacco,
adiposity and alcohol
comorbity <- data
comorbity$chd <- replace(comorbity$chd, comorbity$chd == 1,
"Doente")
comorbity$chd <- replace(comorbity$chd, comorbity$chd == 0,
"Não doente")

cmb_adiposity <- ggplot(comorbity, aes(x=age, y=adiposity,
color = chd)) +
  geom_point(size = 4, show.legend = FALSE) +
  theme_classic2(base_size = 20, base_family = "mono")+
  labs(x = "Idade", y = "Adiposidade localizada")

cmb_tobacco <- ggplot(comorbity, aes(x=age, y=tobacco, color =
chd)) +
  geom_point(size = 4) +
  theme_classic2(base_size = 20, base_family = "mono")+
  labs(color = "Classificação", x = "Idade", y = "tabaco
cumulativo(kg)")

cmb_alcohol <- ggplot(comorbity, aes(x=age, y=alcohol, color =
chd)) +
  geom_point(size = 4) +
  theme_classic2(base_size = 20, base_family = "mono") +
  labs(color = "Classificação", x = "Idade", y = "Consumo de
álcool")

cmb_obesity <- ggplot(comorbity, aes(x=age, y=obesity, color =
chd)) +
  geom_point(size = 4, show.legend = FALSE) +
  theme_classic2(base_size = 20, base_family = "mono") +
  labs(x = "Idade", y = "Obesidade")

# Agrupando comparações
ggarrange(cmb_obesity, cmb_alcohol, cmb_adiposity, cmb_tobacco,
nrow = 2, ncol = 2, common.legend = TRUE, legend = "top")
```

Decidimos analisar a *comorbidade física* que pode desencadear no infarto, sendo elas: Adiposidade, tabagismo, consumo de álcool e nível de obesidade do indivíduo. Comparamos essa análise na faixa de idade dos exemplares, mostrando que a obesidade e adiposidade (gordura encontrada no sangue), aliadas uma faixa de idade entre 45 - 65 anos, possuem alto risco em desenvolver a doença.

Concluímos que comorbidades físicas possuem ampla relação com a idade do indivíduo, aumentando ou não suas chances de desenvolver a doença coronária.



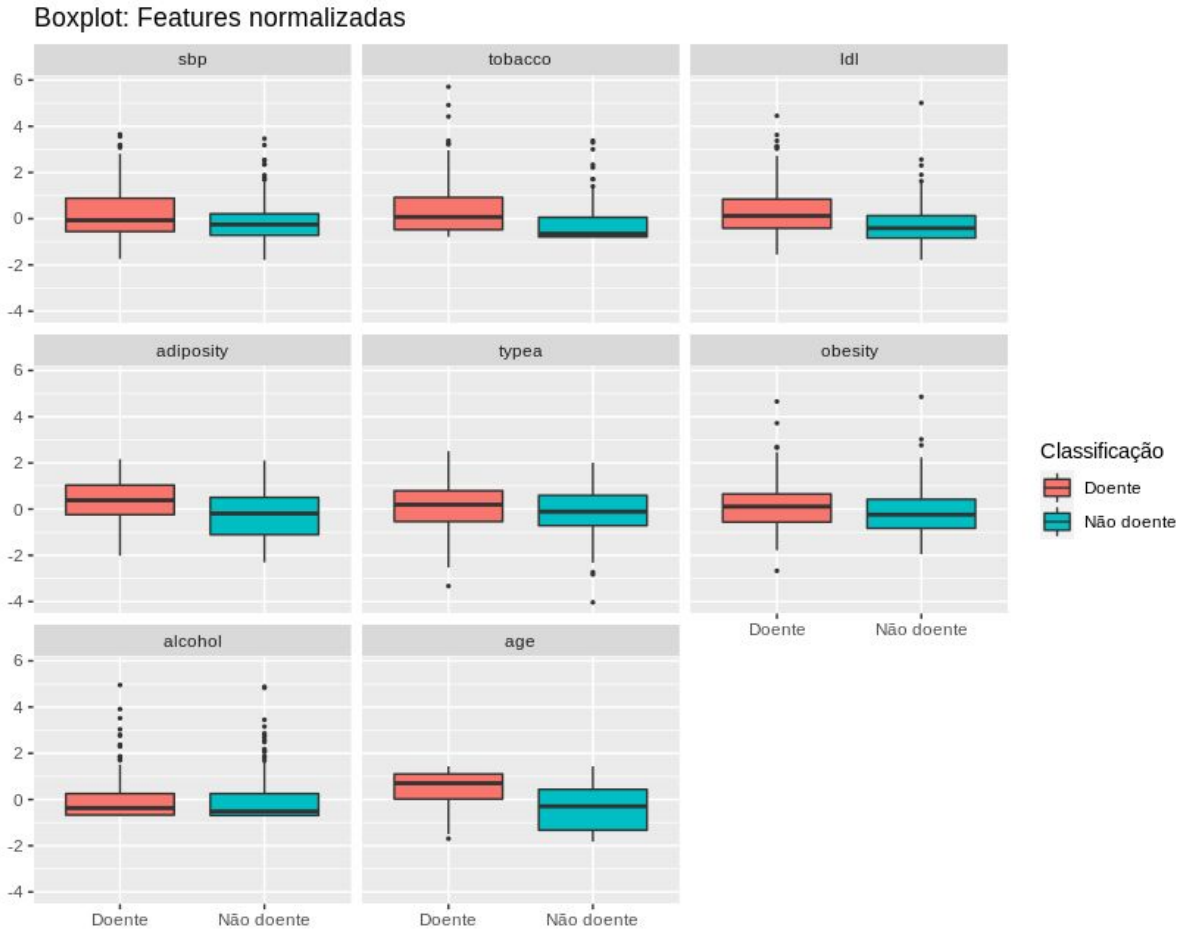

```r
# Analisando a comorbidade mental
ggplot(comorbity, aes(x=age, y=typea, color = chd)) +
  geom_point(size = 5) +
  theme_classic2(base_size = 20, base_family = "mono") +
  labs(x = "Idade", y = "typea", title = "Comorbidade mental x
idade", color = "Classificação")
```

Analisando também um dado bem relevante ao nosso dataset, o **typea** (tipo de personalidade A)**,** que mapeia o quão o indivíduo do dataset se encaixa no tipo de personalidade comportamental A.

Características relevantes desse tipo de personalidade:

- Rigidamente organizado;
- Impaciente;
- Distúrbio de ansiedade;
- Workaholic;
- Preocupado com o todo de quaisquer situações.

Dados essas informações, essa feature é classificada como uma *comorbidade mental* que pode também desencadear a doença coronária. Olhando para o gráfico conseguimos verificar que quem se encaixa em 60% dessas características, possuem uma faixa considerável de desenvolvimento da doença a partir dos 35 anos de idade.



```
ggplot(data_mt, aes(x=chd,y=value,fill=chd)) +
  geom_boxplot(outlier.size = 0.5) +
  facet_wrap(~variable) +
  labs(title = "Boxplot: Features normalizadas",fill =
"Classificação", x= element_blank(), y = element_blank())
```

Fazendo uma análise de todas as features disponíveis conseguimos observar as tendências entre quem teve a doença e quem não desenvolveu. Em média as features de **sbp** (pressão arterial) costuma ser alterada em pacientes doentes, assim como **adiposity** (níveis de gordura no sangue), **ldl** (colesterol ruim) e **tabacco** (níveis de tabagismo). O que concluímos a partir dessa visualização é a confirmação de que em média as comorbidades físicas são os principais propulsores da doença.

Com essa separação em boxplots das features conseguimos visualizar uma prévia das features que serão relevantes a atividade de classificação em machine learning.

3. APLICAÇÃO DOS ALGORITMOS

Foi decidido trabalhar com dois algoritmos de classificação, o KNN e Árvore de Decisão, desta forma comparar suas taxas de acerto e qual algoritmo se adequaria melhor para uma classificação real com um conjunto de dados maior disponível. Assim foi separado o dataset em conjunto de treino e teste, utilizando 80% dos exemplares para treino e 20% para teste.

O conjunto de treino foi separado no intuito de ser comparado com o conjunto de teste no algoritmo KNN, pois esse algoritmo memorizar um conjunto de dados para assim ser comparado com outro conjunto, diferentemente da árvore de decisão na qual não memorizam um conjunto, mas faz as comparações nos input oferecidos a árvore, desta forma, uma vez criada a árvore, pode ser inserido outros conjuntos de dados referente ao dataset trabalhado.

```
# Dataset de treino 80%  da base original e teste 20% base
original.
set.seed(123)
smp_size <- floor(0.80 * nrow(data))
train_ind <- sample(seq_len(nrow(data)), size = smp_size)

train <- data[train_ind, ]
test <- data[-train_ind, ]

dim(train)
dim(test)
```

3.1 KNN

A escolha do índice K que indica o número de vizinhos mais próximos na iterações do algoritmo não é trivial, optamos por valores pequenos e ímpares, sendo eles 1, 3, 5 e 11.

Implementamos uma função geradora do KNN, executamos os testes com K vizinhos coletando os resultados. A coluna chd que é a classificadora escolhida após a remoção do identificador se encontrou na 10º coluna do conjunto, pensando nisso, na função que implementa o KNN, além, do conjunto de treino, teste e o índice K, passamos um quarto parâmetro referente a essa posição do classificador no dataset, assim, essa mesma função poderia ser reaproveitada para outras implementações desse algoritmo a outros datasets, realizando assim uma boa prática de programação em R.

Como saída, a função permite visualizar a árvore de decisão e a taxa de acerto.

```
# train[,10] representa coluna chd
verificaknn <- function(datasetTrain, datasetTest, vetorK,
posicaoClassificador){
  classesTrain <- datasetTrain[ ,posicaoClassificador]
  datasetTrain <- datasetTrain[ , -posicaoClassificador]

  classesTest <- datasetTest[ , posicaoClassificador]
  datasetTest <- datasetTest[ , -posicaoClassificador]

  result <- knn(datasetTrain, datasetTest, classesTrain, vetorK)

  # Matriz de confusã
  print("Matriz confusão:")
```

3.1.1 RESULTADOS KNN

Para K sendo 1 o índice de acerto foi de 59% com a seguinte matriz de confusão:

x	0	1
0	42	16
1	22	13

De 58 exemplares sem a presença da doença, 42 foram classificados de maneira correta e 16 de maneira incorreta. Da mesma forma de 35 exemplares com a presença da doença. 22 foram classificados de maneira incorreta e 13 obtiveram acertos.

Para K sendo 3 o índice de acerto foi menor com 54% e a seguinte matriz de confusão:

x	0	1
0	42	16
1	26	9

Nos 58 exemplares sem a presença da doença, foi obtido a mesma classificação anterior A diferença foi nos 35 exemplares com a presença da doença onde aumentou mais 4 observações classificadas de maneira incorreta, ou seja 26 com 9 exemplares em acerto.

Para K sendo 5 o índice de acerto foi de 58 % com a seguinte matriz de confusão:

x	0	1
0	46	12

1	27	8

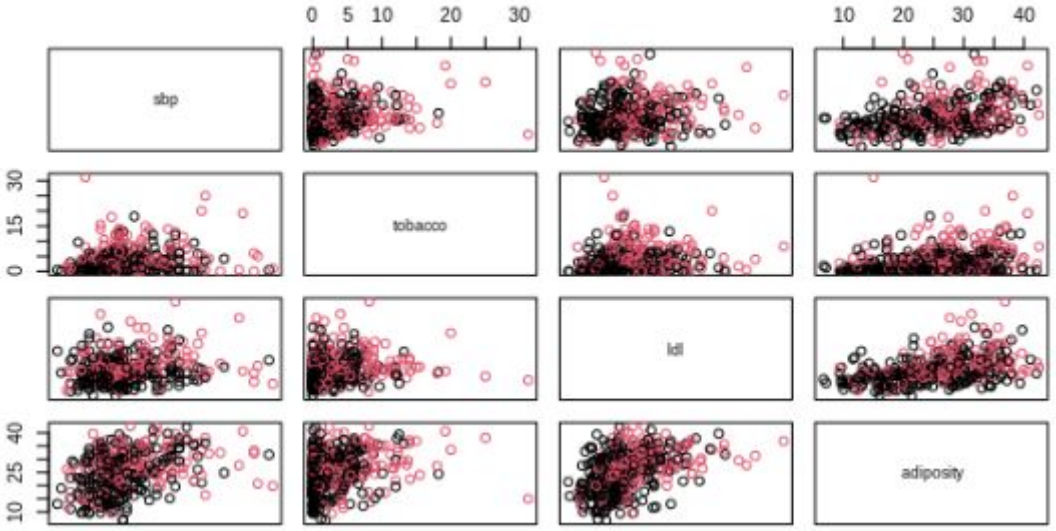
Dos 58 exemplares sem a presença da doença, 46 de maneira correta e 12 de maneira incorreta.

Para K sendo 11, obtivemos a melhor taxa de acerto na classificação com 63%, a partir se aumentar o índice K, a acurácia decai.

x	**0**	**1**
0	51	7
1	27	8

Os resultados apresentaram um alto índice de acerto na predição, mostra que a classificação de benigno e maligno foi aplicada com sucesso pelo algoritmo de classificação KNN.

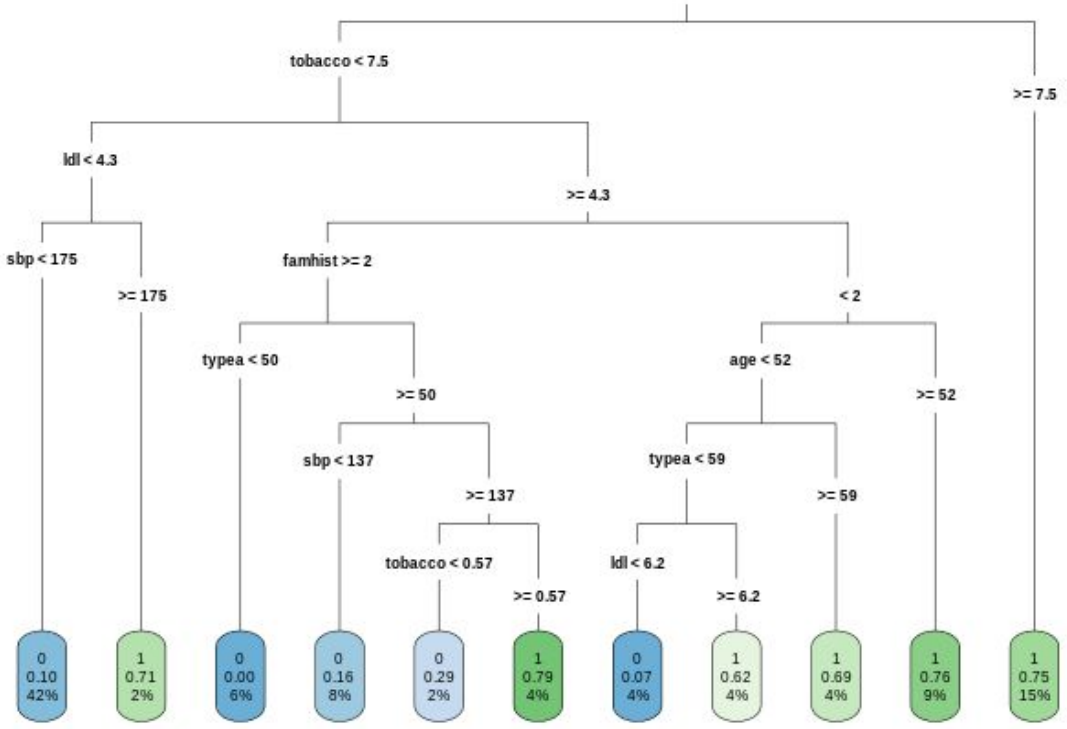
Com isso concluímos que os dados têm uma distância euclidiana muito próxima, e possivelmente os dados se misturam em relação a chd. Assim verificamos a densidade dos dados e proximidade de valores na exploração estatística a seguir sendo possível a visualização da sobreposição já nas primeiras features.



3.2 ÁRVORE DE DECISÃO

Aplicamos a árvore de decisão gerada a partir do método rpart.

```
modelo <- rpart(diagnosis~., train, method = "class", control =
rpart.control(minisplit = 1))
plot <- rpart.plot(modelo, type = 3)
```



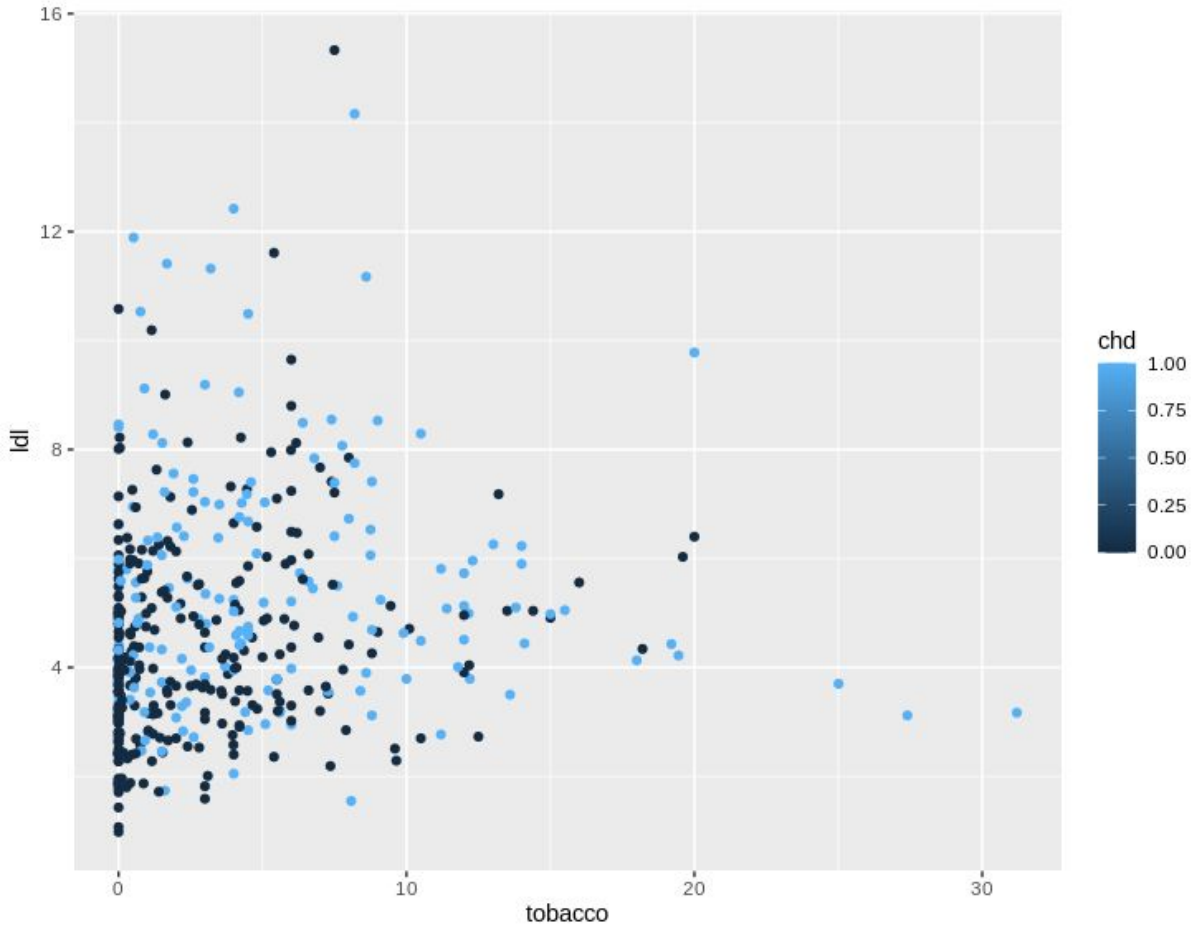
Foram obtidos os seguintes resultados para os conjuntos aplicados ao algoritmo de machine learning árvore de decisão.

- Dataset de teste: 0.5913978 ou 59% índice de acerto
- Dataset de treino: 0.8346883 ou 83% índice de acerto
- Dataset de completo: 0.7857143 ou 78% índice de acerto

3.2.2 RESULTADOS ÁRVORE DE DECISÃO

Para o conjunto de teste o resultado de acerto foi semelhante ao obtidos por KNN, poŕem ao aplicar conjunto de dados maiores há uma melhora significativa para o modelo de machine learning gerado, atingindo 83 de acurácia para dataset de treino. Assim acreditamos que por ser uma mostra retrospectiva de um conjunto de dados que não está público segundo a descrição do dataset no site kaggle, os resultados poderiam ser melhores se tivéssemos um maior número de exemplares para o treinamento. Uma hipótese para o não atingimento de acurácias acima de 90% é os inúmeros fatores que podem contribuir para o surgimento da doença, como podemos notar que a árvore gerada tem muitos ramos, ou seja, diferentes medidas de features podem ter inúmeras interpretações, bem próximas, assim entendemos que um maior conjunto de exemplares poderia contribuir para uma melhor classificação.

As duas melhores features que contribuíram para classificação entre um diagnóstico com resposta à doença cardíaca coronária segundo o modelo de árvore gerada foram o consumo de tabaco e ldl (colesterol de lipoproteína de baixa densidade), ao analisar o plot entre essas duas features reforça a ideia que muitos fatores podem influenciar o surgimento dessa doença cardiovascular, pois muitos pontos nos indicando que esses pacientes que estão mais misturado muito provável tem outras característica presentes com por exemplo o fator idade avançada, fatores psicológicos e físicos como o alto consumo de álcool entre outros.



Desta forma, a função de custos dessas features presentes na árvore teve um valor próximo, considerada assim como relevante para diferenciar um diagnóstico com resposta a doença ou não, visto que o dataset é complexo considerando o número de informações presentes. Portanto, um médico deverá olhar mais de uma ampla todas as características presentes para indicar um tratamento adequado no sentido de querer diagnosticar o paciente se tem ou não a doença cardíaca coronariana.

4. CONCLUSÃO E ANÁLISE DE RESULTADOS

Para a visualização de dados conseguimos notar tendências relevantes de comorbidades físicas e psicológicas, que somadas a idade levam a um infarto. Um insight também relevante é o histórico familiar, que se mostrou um fator pesado para pacientes que desenvolveram a doença. Comorbidades físicas como consumo de álcool e tabagismo também podem desencadear a doença, porém não é uma unanimidade, variando de caso para caso.

Conseguimos notar que a melhor idade, considerando-a a partir dos 50 anos requer um cuidado e medidas de mitigação contra a doença mais rígidas para que a mesma não se desenvolva, fatores como estilo de vida (personalidade tipo a) também influenciam, portanto com a visualização, conseguimos notar que uma vida desequilibrada e com tendências familiares acabam acarretando ao infarto.

Entrando em Machine learning, a classificação por KNN o índice de acerto foi relativamente baixo com 63% utilizando o melhor K, isso devido a muitos dados de diferentes features com distâncias muito próximas, nos revelando uma possível tendência a ter diversas features que contribuem para o surgimento de resposta à doença cardíaca. Pensando assim, o aprendizado de máquina por KNN, teve uma acurácia em mais da metade dos dados comparados com o dataset de teste, ou seja, esse processo de machine learning poderia contribuir com mais da metade de chance de acerto na predição de um diagnóstico de pacientes ter ou não a doença, e se alinhar as outras porcentagem conhecimento próprio do médico, poderia ter bons resultado em predição desse tipo de diagnóstico.

Na classificação utilizando a árvore de decisão aplicando o conjunto de teste para ser classificado, o resultado obtido foi um pouco menor que ao KNN com 59% de taxa de acerto, e como nesse algoritmo não há uma memorização para ocorrer a comparação, podemos aplicar outros conjuntos de dados, assim utilizamos o conjunto de treino e obtivemos uma acurácia de 83% poŕém são dados já conhecido pelo modelo, nos evidenciando a tendência do algoritmo errar a classificação partir de dados não conhecidos como o conjunto de teste.

Com a árvore de decisão foi possível entender que as features que mais contribuem para diferenciar se há resposta para doença cardíaca são respectivamente são consumo de tabaco, colesterol de lipoproteína de baixa

densidade, pressão arterial sistólica, histórico familiar, comportamento do tipo A e idade.

O principal objetivo do trabalho era analisar os dados de um paciente e predizer se tem ou não doença cardíaca, assim como o grau de tendência a ser desenvolvido. Por isso a ideia da aplicação de métodos de classificação de machine learning para predizer esse tipo de diagnóstico. A expectativa inicial era atingir alto índice de acurácia acima de 90%, por entender que o dataset apresenta um bom classificador para a questão que estávamos procurando responder, o classificador chd, no entanto, o nível de acurácia foi mais baixo. Procuramos entender o motivo para o baixo índice de acerto, adotamos primeiramente o balanceamento do dataset em relação a chd, onde foi mantido o nível de acerto. Ao analisar a árvore gerada entendemos que contém muitos ramos, ou seja, muitas features com alta função de custo. Desta forma, seria errado dizer que se um paciente tem alto consumo de tabaco e está em idade avançada que ele vai ter a doença cardíaca coronariana, pois isso depende de mais fatores como qual o seu nível de gordura localizada entre outros.

Portanto entendemos que a classificação oferecida pelo machine learning tem grande relevância para suporte na área da saúde uma vez que são mais de um tipo de característica que pode levar um pessoa ter uma doença cardíaca coronariana, assim juntamente com o conhecimento de um médico esse trabalho de classificação sobre tais dados podem contribuir em situações reais ligadas a area da saude.

5. CÓDIGO

```
# install.packages("RColorBrewer")
# install.packages("tidyverse")
# install.packages("RColorBrewer")

library(tidyr)
library(dplyr)
library(dslabs)
library(rpart.plot)
library(tidyverse)
library(cluster)
library(factoextra)
library(FactoMineR)
library(ggplot2)
library(RColorBrewer)
library(ggpubr)
library(class)
library(rpart)
library(rpart.plot)
library(reshape2)
myPalette <- brewer.pal(5, "Set2")

data <- read.csv("./dataset.txt", sep =";", na.strings =
c('','NA','na','N/A','n/a','NaN','nan'), header = TRUE)

#Parte 1 - Entendimento prévio e limpeza do dataset
--------------------------------------------------------------------
View(head(data, n=5))
dim(data) # 462 linhas e 11 colunas (sendo uma coluna identificador)
str(data)

# Verificar dados vazios em cada coluna e contabilizar
View(sapply(data, function(x) sum(is.na(x))))
sum(is.na(data))

# retirar coluna de identificação
data <- data[-1]

# tranformando "Present" e "Absent" em dados númericos
```

```r
data$famhist <- replace(data$famhist, data$famhist == "Present", "1") #
Atribuindo Present com 1
data$famhist <- replace(data$famhist, data$famhist == "Absent", "2") #
Atribuindo Absent com 2
data$famhist <- as.numeric(data$famhist)

# Padronizando o dataset com o tipo número para todas as colunas
data[1:462,c(1,6,9,10)] <- sapply(data[1:462,c(1,6,9,10)], as.numeric)
str(data)

comDoenca = sum(data$chd == 1) # 160 tiveram doença na coronaria
semDoenca = sum(data$chd == 0) # 302 Não tiveram doença na coronaria

# Gráfico de pizza: Doentes e não doentes
pie_dataframe <- data.frame(
 group = c("Doente", "Não doente"),
 value = c(sum(data$chd == 1), sum(data$chd == 0)) # 160 tiveram doença
na coronária e 302 Não tiveram doença na coronária
)

pie_dataframe_rate <- ggplot(pie_dataframe, aes(x = "", y = value, fill
= group)) +
 geom_bar(width = 1, stat = "identity") +
 coord_polar("y", start = 0) +
 scale_fill_brewer(palette="Blues") + theme_minimal() +
   geom_text(aes(x   =   1,label   =   paste(round(value   /   sum(value)
*100,1),"%"), family = "sans"),
           position = position_stack(vjust = 0.5), size = 10)+
 labs(fill = "Classificação",
      x = NULL,
      y = NULL,
      title = "Pie: Doentes e não doentes",
      subtitle = "Em porcentagem") +
 theme_bw(base_size = 20, base_family = "mono")

pie_dataframe_no_rate <- ggplot(pie_dataframe, aes(x = "", y = value,
fill = group)) +
 geom_bar(width = 1, stat = "identity") +
 coord_polar("y", start = 0) +
 scale_fill_brewer(palette="Blues") + theme_minimal() +
 geom_text(aes(x = 1,label = value, family = "sans"),
           position = position_stack(vjust = 0.5), size = 10)+
 labs(fill = "Classificação",
```

```
      x = NULL,
      y = NULL,
      title = "Pie: Doentes e não doentes",
      subtitle = "Em dados brutos") +
 theme_bw(base_size = 20, base_family = "mono")

ggarrange(pie_dataframe_no_rate, pie_dataframe_rate, ncol = 2)

# conferir balanceamentos das observações em prol do histórico familiar
famhist_present = sum(data$famhist == 1) # 158 tiveram doença na coronariana
famhist_ausent = sum(data$famhist == 2) # 214 Não tiveram doença na coronariana

# tranformando "Present" e "Absent" em dados númericos
data$famhist <- as.character(data$famhist)
data$famhist <- replace(data$famhist, data$famhist == "1", "Present") # Atribuindo Present com 1
data$famhist <- replace(data$famhist, data$famhist == "2", "Absent") # Atribuindo Absent com 2

# Filtro exemplares que tem e não tem resposta a doença cardiaca
coronary = filter(data, chd == 1)
nonCoronary = filter(data, chd == 0)

# Faixa de idade do dataset
ageCoronary <- ggplot(coronary, aes(x=age)) +
  geom_histogram(color="black", fill="red", bins = 30, binwidth = 1)+
geom_density(fill="#FF6666") +
 scale_x_continuous(name = "Idade") +
 scale_y_continuous(name = "Pessoas")

ageNoncoranary <- ggplot(nonCoronary, aes(x=age)) +
  geom_histogram(color="black", fill="blue", bins = 30, binwidth = 1)+
geom_density(fill="#FF6666") +
 scale_x_continuous(name = "Idade") +
 scale_y_continuous(name = "Pessoas")

ggarrange(ageCoronary, ageNoncoranary, labels = c("Faixa de idade dos doentes", "Faixa de idade dos não doentes"), nrow = 2)

# Quantos dos exemplares que desenvolveram a doença possuiam historico familiar
```

```r
ageCoronary <- ggplot(coronary, aes(x=age, fill=famhist)) +
   geom_histogram(color="black", bins = 30, binwidth = 1)+
geom_density(fill="#FF6666") +
 labs(fill = "Histórico familiar:",
      x = "idade",
      y = "Pessoas",
      title = "Doentes") +
 theme_classic(base_size = 20, base_family = "mono")

ageNoncoranary <- ggplot(nonCoronary, aes(x=age, group = famhist,
fill=famhist)) +
   geom_histogram(color="black", bins = 30, binwidth = 1)+
geom_density(fill="#FF6666") +
 labs(fill = "Histórico familiar:",
      x = "idade",
      y = "Pessoas",
      title = "Não Doentes") +
 theme_classic(base_size = 20, base_family = "mono")

ggarrange(ageCoronary, ageNoncoranary, nrow = 2)

# Comparando as principais comorbidades x idade: tobacco, adiposity and
alcohol
comorbity <- data
comorbity$chd <- replace(comorbity$chd, comorbity$chd == 1, "Doente")
comorbity$chd <- replace(comorbity$chd, comorbity$chd == 0, "Não
doente")

cmb_adiposity <- ggplot(comorbity, aes(x=age, y=adiposity, color =
chd)) +
 geom_point(size = 4, show.legend = FALSE) +
 theme_classic2(base_size = 20, base_family = "mono")+
 labs(x = "Idade", y = "Adiposidade localizada")

cmb_tobacco <- ggplot(comorbity, aes(x=age, y=tobacco, color = chd)) +
 geom_point(size = 4) +
 theme_classic2(base_size = 20, base_family = "mono")+
   labs(color = "Classificação", x = "Idade", y = "tabaco
cumulativo(kg)")

cmb_alcohol <- ggplot(comorbity, aes(x=age, y=alcohol, color = chd)) +
 geom_point(size = 4) +
 theme_classic2(base_size = 20, base_family = "mono") +
```

```r
 labs(color = "Classificação", x = "Idade", y = "Consumo de álcool")

cmb_obesity <- ggplot(comorbity, aes(x=age, y=obesity, color = chd)) +
 geom_point(size = 4, show.legend = FALSE) +
 theme_classic2(base_size = 20, base_family = "mono") +
 labs(x = "Idade", y = "Obesidade")

# Agrupando comparações
ggarrange(cmb_obesity, cmb_alcohol, cmb_adiposity, cmb_tobacco, nrow =
2, ncol = 2, common.legend = TRUE, legend = "top")

# Analisando a comorbidade mental
ggplot(comorbity, aes(x=age, y=typea, color = chd)) +
 geom_point(size = 5) +
 theme_classic2(base_size = 20, base_family = "mono") +
 labs(x = "Idade", y = "typea", title = "Comorbidade mental x idade",
color = "Classificação")

data$famhist <- replace(data$famhist, data$famhist == "Present", "1") #
Atribuindo Present com 1
data$famhist <- replace(data$famhist, data$famhist == "Absent", "2") #
Atribuindo Absent com 2
data$famhist <- as.numeric(data$famhist)

# separação e armazenamento das features chd e famhist
chd_feature <- data[10]
famhist_feature <- data[5]

# Novo data frame sem as features chd e famhist, features com potencial
para classificador
df <- data[-10]
df <- df[-5]

# dataset comtém com diferentes precisões decimais entre as features,
aplicar escala
data_scale <- as.data.frame(scale(df))

# new_df é um data frame padronizado em escala mais as features chd e
famhist
new_df_scale <- cbind(data_scale, famhist_feature)
new_df_scale <- cbind(new_df_scale, chd_feature)
View(new_df_scale)
```

```r
# Comparando features com boxplot entre quem teve ou não a doença
data_mt <- new_df_scale[-9]
data_mt <- melt(new_df_scale[-9], id.vars = c("chd"))
data_mt$chd <- replace(data_mt$chd, data_mt$chd == 1, "Doente")
data_mt$chd <- replace(data_mt$chd, data_mt$chd == 0, "Não doente")

ggplot(data_mt, aes(x=chd,y=value,fill=chd)) +
 geom_boxplot(outlier.size = 0.5) +
 facet_wrap(~variable) +
 labs(title = "Boxplot: Features normalizadas",fill = "Classificação",
x= element_blank(), y = element_blank())

#---------- 2 parte - IA ----------------------------------------

# Conjunto de treino e teste (80% 20%) - Dados original
set.seed(123)
smp_size <- floor(0.80 * nrow(data))
train_ind <- sample(seq_len(nrow(data)), size = smp_size)

train <- data[train_ind, ]
test <- data[-train_ind, ]

dim(train)
dim(test)

# Verificando se o dataset de treino é classificavel e possibilita
exploração.

# preparando chd para colorir plot
chd <- as.factor(train$chd)

# Comparativo de todas
plot(train, col = chd)

# É possivél ver um excesso de gordura no organismo ao longo da idade e
também o aumento de numero
# de paciantes com resultado de doença coronária.
plot(x = train$age, y = train$adiposity, col = chd, pch =19,
main="Adiposidade x Idade",
    ylab="Adiposidade", xlab="Idade")

#--Algoritmo KNN com k = 1,3,5 e 11 vizinhos-------------------------
# train[,10] representa coluna chd
```

```r
verificaknn <- function(datasetTrain, datasetTest, vetorK, posicaoClassificador){
 classesTrain <- datasetTrain[ ,posicaoClassificador]
 datasetTrain <- datasetTrain[ , -posicaoClassificador]
  classesTest <- datasetTest[ , posicaoClassificador]
 datasetTest <- datasetTest[ , -posicaoClassificador]
  result <- knn(datasetTrain, datasetTest, classesTrain, vetorK)
  # Matriz de confusão
 print("Matriz confusão:")
 print(as.matrix(table(classesTest, result)))
 matriz <- as.matrix(table(classesTest, result))
  # Indíce de acerto
 acc <- sum(diag(matriz))/nrow(datasetTest)
 print("Indice de acerto:")
 print(acc)
}

# KNN com dataset sem redução
# k = 1
verificaknn(train, test, 1, 10)
# k = 3
verificaknn(train, test, 3, 10)
# k = 5
verificaknn(train, test, 5, 10)
# k = 11
verificaknn(train, test, 11, 10)

#  -------------------- Algoritmo de árvore de decisão --------------------------
modelo <- rpart(chd~., train, method = "class", control = rpart.control(minisplit = 1))
plot <- rpart.plot(modelo, type = 3)

verificaDesicionTree <- function(modelo, datasetTest, posicaoClassificador){
 classesTest <- datasetTest[ , posicaoClassificador]
 datasetTest <- datasetTest[ , -posicaoClassificador]
  pred <- predict(modelo, datasetTest, type = "class")
  # Matriz de confusão
 print("Matriz confusão:")
 matriz <- as.matrix(table(classesTest, pred))
 print(matriz)
  # Indíce de acerto
```

```
 acc <- sum(diag(matriz))/nrow(datasetTest)
 print("Indice de acerto:")
 print(acc)
}

verificaDesicionTree(modelo, test, 10)
verificaDesicionTree(modelo, train, 10)
verificaDesicionTree(modelo,  data,  10)  #  Utilizando  o  modelo  para
predizer todo o data set

p <- ggplot(data, aes(tobacco, ldl, group=chd, colour = chd))
p + geom_point()
```

6. REFERÊNCIAS

- https://www.kaggle.com/yassinehamdaoui1/cardiovascular-disease - Visitado para escolha do dataset.

- https://www.infopedia.pt/$comportamentos-de-tipo-a-e-b
  Visitado para entendimento da feature “type a”.

- https://pebmed.com.br/consumo-de-alcool-e-doencas-cardiovasculares-o-que-dizem-os-estudos/ - Visitado para entender mais sobre as comorbidades expressas nos dataset e sua influência em doença coronaria.

- https://rdrr.io/cran/unbalanced/man/ubNCL.html - Visitado para aplicação de um balanceamento do dataset ao aplicar algoritmos de Machine learning para buscar melhores resultados.