Projeto 1 – Classificação
Disciplina: SIN 460 – Mineração de Dados
Professores: Felipe Provezano Coutinho e Joelson Antônio dos Santos

Base de Dados

Nome da base de dados: Heart Disease UCI

A base de dados pode ser encontrado no seguinte link:
https://www.kaggle.com/ronitf/heart-disease-uci

Criadores:
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Informações da Base de Dados

A base de dados contém informações de 303 indivíduos e tem como objetivo oferecer informações relevantes para classificar se dado indivíduo possui ou não problemas cardíacos. Na composição original desta base de dados estão presentes 76 atributos mas, para fins de estudo e seguindo outros experimentos publicados, foram utilizados 14 atributos. É importante frisar que esse pré processamento nos atributos já foi realizado por outros estudiosos. O link da base de dados desse relatório é referente a base que possui os atributos reduzidos para 14, sendo eles:

1. **age:** Representa a idade dos indivíduos (29 até 77 anos).
2. **sex:** Representa o sexo dos indivíduos (1 = masculino 0 = feminino).
3. **cp:** Representa o tipo de dor no peito, os valores variam de 0 até 3 porém, o objetivo é apenas saber se há ou não dor no peito: (0 = não há dor no peito, [1,2,3] = há dor no peito).
4. **trestbps:** Representa a pressão arterial em repouso: (94 até 200,valores em mmHG).
5. **chol:** Níveis séricos de colesterol (126 até 564 valores em mg/dl).
6. **fbs:** Níveis de açúcar no sangue em jejum (1 > 120 mg/dl e 0 < 120 mg/dl).
7. **restecg:** Resultados eletrocardiográficos em repouso, três tipos de resultados diferentes (valores entre 0 e 2).
8. **thalach:** frequência cardíaca máxima alcançada (71 até 202 bpm).
9. **exang:** angina(um tipo de dor no peito) induzida por exercício(1 = True; 0 = False).
10. **oldpeak:** Depressão do segmento ST induzida pelo exercício em relação ao repouso (números racionais, entre 0 e 6.2 mm).
11. **slope:** Representa o tipo de inclinação do segmento ST durante o pico de exercício, três resultados diferentes: (valores entre 0 e 2).
12. **ca:** Representa o número de grandes vasos(valores entre 0 e 3).

13. **thal:** Resultados do teste de estresse do tálio, quatro resultados possíveis (valores entre 0 e 3).
14. **target:** (1 = doente e 0 = Não doente).

Todos os nomes dos atributos citados seguem os nomes da base de dados original, contudo, para as técnicas de mineração no jupyther todos os atributos foram renomeados para o português.

Normalização da Base de dados

Não foi necessária nenhuma grande mudança dos atributos desta base de dados visto que todos são numéricos e não há nenhum campo ausente. Uma das modificações necessária foi transformar todos os atributos numéricos para o tipo float e, por fim, a normalização dos dados, onde as características (variável X no código) foram normalizadas seguindo o conceito de manter a média = 0 e desvio padrão = 1.

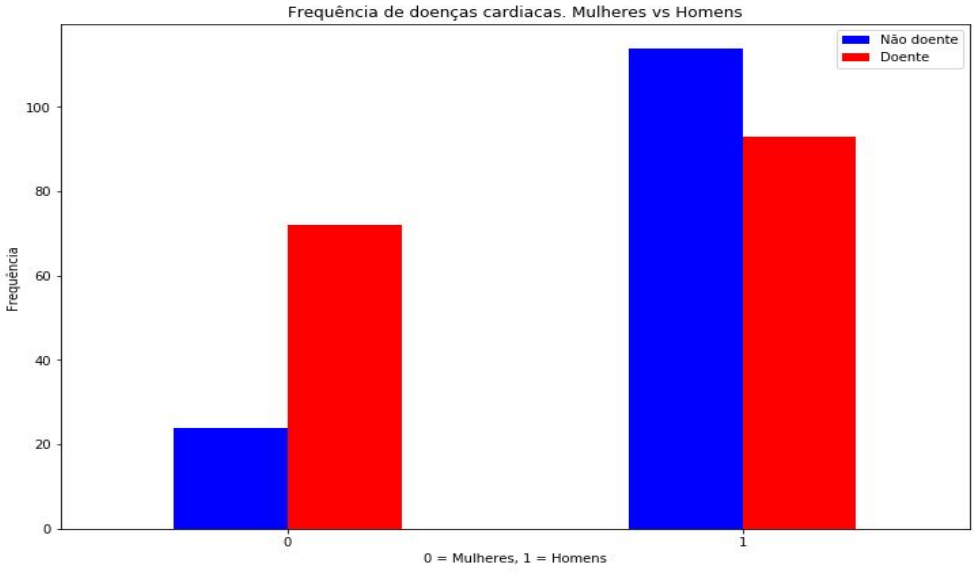
Resultados Esperados

Por mais que a base de dados utilizada em nosso estudo não contenha todos os 76 atributos disponíveis, é esperado que com o uso das técnicas de mineração seja possível classificar com uma alta taxa de acerto se dado indivíduo possui ou não doença cardíaca, além de encontrar qual ou quais atributo tem maior impacto no processo de classificação.

Extração de informações da base de dados

Pacientes com doença no coração: 54.46%
Pacientes sem doença no coração: 45.54%

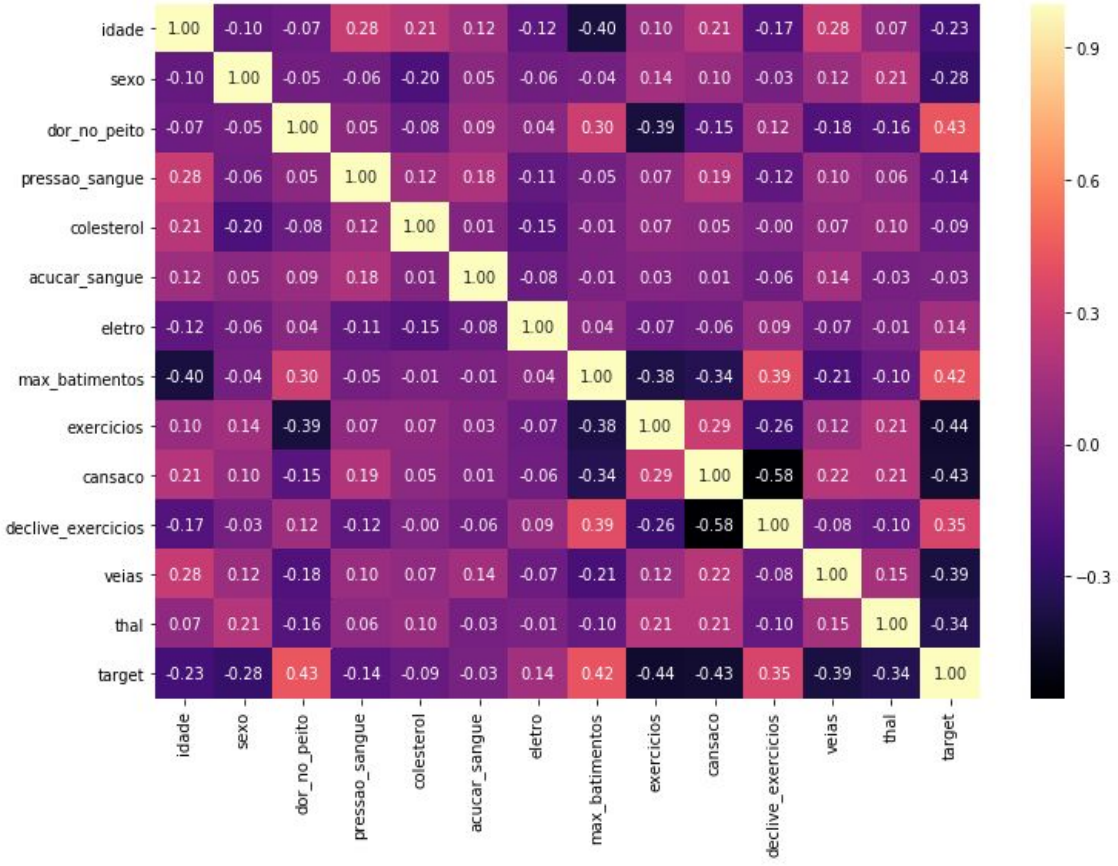
Homens no dataset: 68.32%
Mulheres no dataset: 31.68%



Processo de Mineração de dados

A correlação é importante na ciência de dados pois é capaz de mostrar o grau de dependência entre duas variáveis, ou seja, quantificar a influência de um atributo sobre outro. Dessa forma, foi gerada uma matriz de correlação entre todos os atributos da base de dados utilizada, como mostra a figura a seguir.

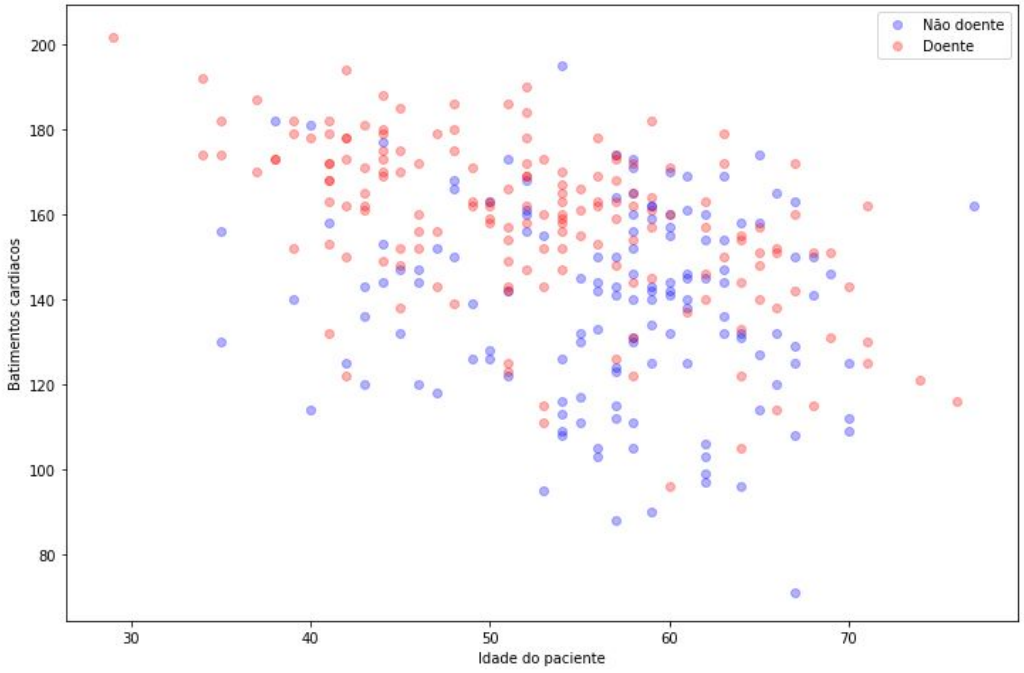
Matriz de Correlação entre os atributos



Observando a matriz é possível perceber que os atributos max_batimentos e dor_no_peito possuem os maiores valores de correlação em relação ao atributo target, o qual é responsável por classificar o paciente como doente ou não doente.

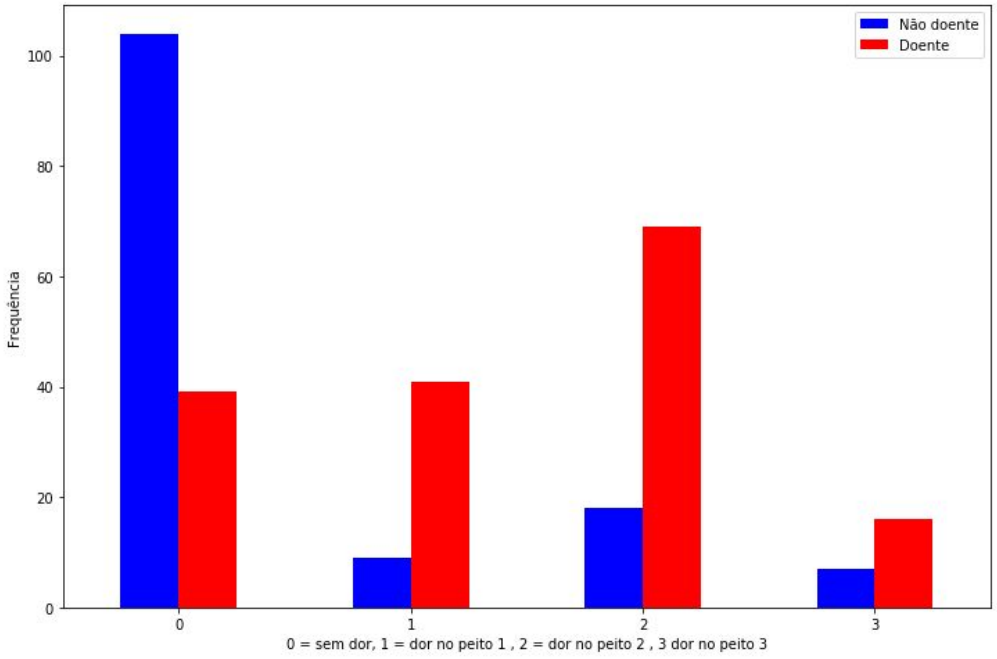
Constatado esse fato, é esperado que ao se realizar uma representação gráfica dos dois atributos(max_batimentos, dor_no_peito), com o atributo target, seja possível encontrar um resultado relativamente homogêneo, onde seria possível através de uma simples visualização encontrar os intervalos onde o paciente é doente ou não. Para comprovar essa hipótese foram traçados três gráficos utilizando o jupyther notebook.

Gráfico: Idade do Paciente vs Batimentos Cardíacos



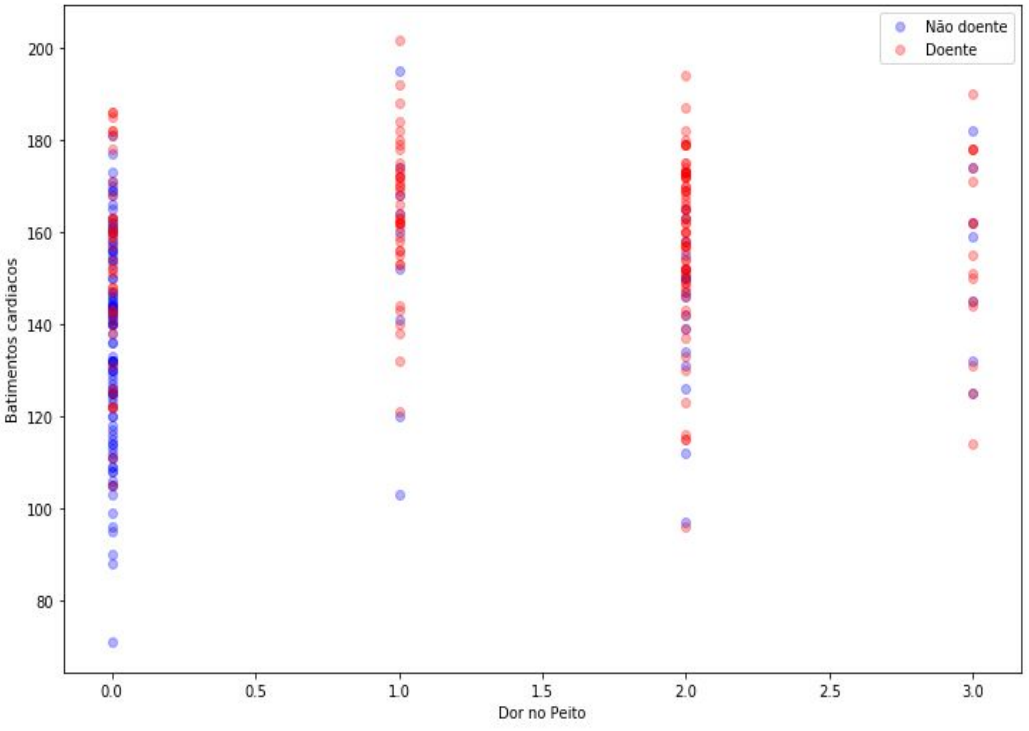
Ao analisar o gráfico, pode-se perceber que quando a faixa de batimentos cardíacos é de até 140 bpm, o número de indivíduos não doentes é alto. Porém, quando os batimentos cardíacos ultrapassam a faixa dos 140 bpm, o número de indivíduos com doença cardíaca é consideravelmente maior do que os que não apresentam doença.

Gráfico: Dor no peito x Frequência de Indivíduos



Ao analisar o gráfico fica claro que o número de indivíduos não doentes é muito maior quando não há a presença de dor no peito, porém, quando o indivíduo apresenta algum tipo de dor no peito, principalmente do tipo 1 e 2, o número de indivíduos doentes é cerca de 3 vezes maior do que os não doentes.

Gráfico: Dor no peito vs Batimentos Cardíacos



Por fim, ao esboçar um gráfico dos dois atributos de maior valor de correlação é possível observar um nível ainda maior de homogeneidade. Quando não há a presença de dor no peito e batimentos cardíacos até a faixa de 140 bpm, a concentração de indivíduos não doente é consideravelmente alta. Em contrapartida, se os batimentos cardíacos forem maiores que 140 bpm e o indivíduo apresentar algum tipo de dor no peito, pode-se inferir que grande parcela dos que se encontram nessa faixa possuem doença cardíaca.

Utilização de Redes Neurais para a classificação

A técnica escolhida para a classificação foi a de redes neurais artificiais devido a facilidade em computar e encontrar padrões. Essa técnica é utilizada para problemas de classificação pela capacidade de aprender de seu ambiente e com isso melhorar gradativamente seu desempenho.

A estrutura da rede neural é formada por neurônios e camadas. Na camada de entrada os padrões são apresentados à rede, na camada de saída o resultado é apresentado e por último a camada intermediária ou escondida, é responsável por todo o processamento da rede, extraindo assim as características da mesma. O número de

camadas intermediárias varia entre 1 ou mais, não havendo um número específico para melhores resultados.

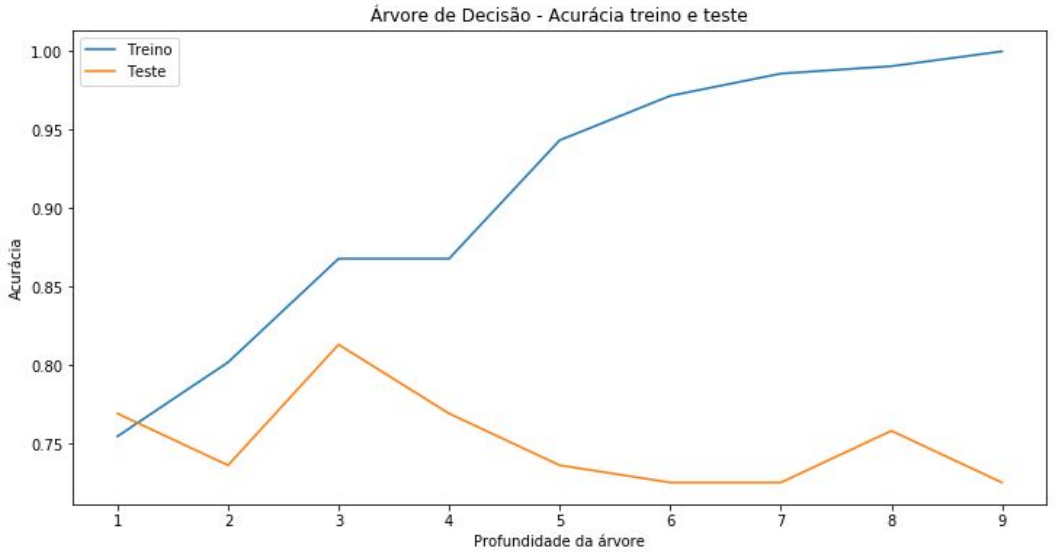
O algoritmo utilizado é do tipo aprendizado supervisionado e segue os conceitos de feedforward, onde cada camada é conectada com uma ou mais camadas, porém, sempre seguindo o mesmo sentido, sem caminho de volta.

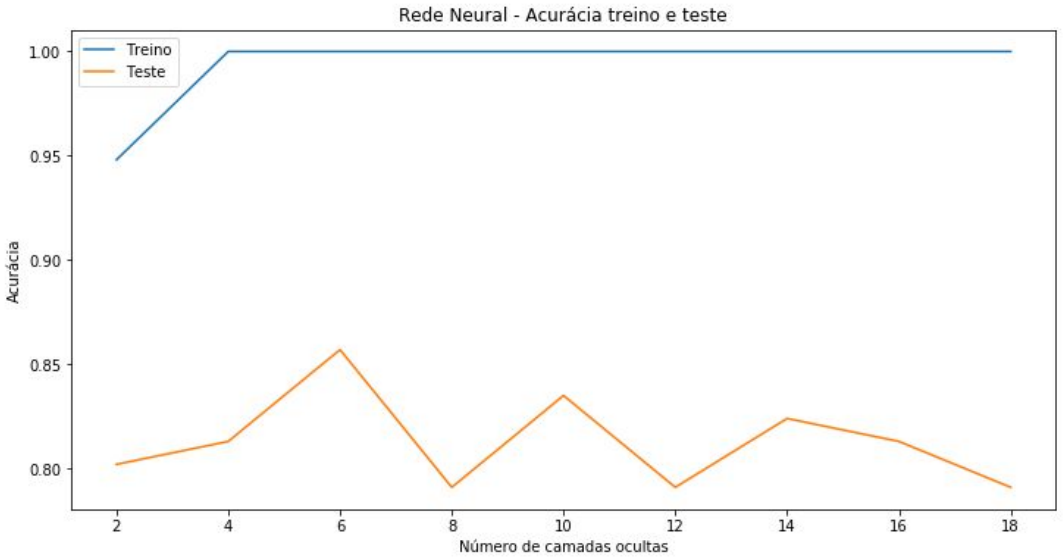
O treinamento da rede neural se deu pela utilização de backpropagation, que é um treinamento supervisionado responsável por calcular um medida de erro e em seguida corrigir os pesos nas camadas a fim de reduzir o número de erros.

Os dados foram divididos em 70% para treinamento e 30% para teste, e o número de camadas ocultas e de neurônios da rede foi escolhido de forma empírica.

Foi realizada uma função denominada Tree(), responsável por definir qual a melhor profundidade da rede para resolver o problema. Essa definição é feita pelo valor que representa a menor diferença entre a acurácia no treino e no teste. Aliado a função Tree(), há uma função responsável por evidenciar qual o melhor número de camadas ocultas que diminuem a diferença entre a acurácia do treino e teste.

Após o treinamento da rede neural os resultados foram representados pelos seguintes gráficos:





Concluímos que uma árvore com profundidade de tamanho 3 e a utilização de 6 camadas, entregam uma acurácia de teste na faixa de 80% a 85%, enquanto mantém um valor próximo da acurácia nos treinos.

Experiência

Em um contexto geral o projeto foi de grande importância para o grupo visto que ficou evidenciado de forma prática as dificuldade que cercam o estudo da mineração de dados e, junto a isso, como superá-las. Logo no início do projeto foi possível perceber que encontrar uma base de dados relevante e sem dados ausentes não é uma tarefa fácil mas, após um bom tempo de pesquisa, o grupo encontrou uma base do interesse de todos os membros. A base foi relativamente fácil de manipular, já que todos os atributos se encontravam na forma numérica, o que facilitou muito o processo de normalização dos dados.

Na sequência houve uma pequena dificuldade com a ferramenta jupyther, justamente por ser algo novo e utilizar a linguagem de programação python, que ainda não era muito conhecida pelo grupo. A curva de aprendizado foi muita rápida devido a simplicidade da linguagem python, o que foi de extrema importância para o grupo, pois mostrou a todos que as dificuldades em relação à linguagem poderiam ser superadas sem grande esforço, encorajando todos a darem continuidade no projeto e futuramente se envolverem mais com a linguagem e sua infinidade de recursos.

A dúvida que veio logo em seguida foi na escolha da técnica de mineração que seria utilizada para a classificação dos indivíduos. Após alguns debates, foi escolhido em consenso o uso de redes neurais por dois motivos:

1- Ao se realizar uma pesquisa pela literatura, o grupo percebeu uma grande eficiência na utilização das redes neurais para problemas de classificação.

2- Um dos membros do grupo já estava familiarizado com o estudo de redes neurais, o que foi muito importante, pois tornou o projeto mais simples além de ter introduzido os outros membros do grupo em uma área de estudo que ainda não era conhecido por todos.

A experiência ganha foi imensa, pois os membros que ainda não eram familiarizados com o uso de redes neurais artificiais puderam começar a aprender sobre o assunto, vendo os resultados na prática. Além disso, o membro que já possuía um conhecimento prévio pode se aprofundar ainda mais na área de seu interesse.

Por fim, os gráficos gerados e os resultados da utilização das redes neurais possibilitou ao grupo realizar uma análise e reflexão, o que permitiu tirar conclusões concretas e por fim evidenciar que muitas vezes dados que não entendemos inicialmente podem na verdade ter um significado muito grande, e assim serem utilizados para resolver diversos problemas presentes no cotidiano da vida humana.

Discussão dos resultados

Ao fim do projeto os resultados encontrados ficaram dentro do esperado. A acurácia encontrada pela rede neural artificial usando o melhor valor para o número de camadas ocultas e profundidade da árvore foi de 80% a 85%. Pode ser considerado um valor consideravelmente alto, visto que seria muito difícil para um médico conseguir um diagnóstico com essa precisão apenas com os atributos da base de dados.

Infelizmente não é possível encontrar com exatidão um número que representa a acurácia dos médicos para detecção da presença de doenças cardíacas, o que dificuldade a capacidade da comparação do resultado gerado com o resultados de médicos na vida real. Porém um dos métodos usados na medicina e que nos permite fazer um “grosso comparativo” é o Score. O método é utilizado para detecção de risco cardiovascular e necessita do resultado de exames de sangue possui uma taxa de acerto de 72 por cento.

Por fim, foi possível perceber que os atributos max_batimentos e dor no peito são fatores de grande peso para a classificação de um indivíduo em doente ou não, mas também fica claro que somente os dois atributos não seriam capazes de oferecer a taxa de acurácia encontrada no fim do projeto.

Referências:

http://conteudo.icmc.usp.br/pessoas/andre/research/neural/

https://itmidia.com/cientistas-da-alphabet-usam-ia-para-detectar-riscos-de-doenca-cardiovascular/

HAYKIN, Simon. **Redes neurais: princípios e prática**. Bookman Editora, 2007.