Universidade Federal do Rio Grande do Norte
Centro de Ciências Exatas e da Terra
Departmento de Informática e Matemática Aplicada

Inteligência Artificial para diagnóstico de problemas cardíacos

Daniel Sehn Colao

Natal-RN

Dezembro, 2022

Inteligência Artificial para diagnóstico de problemas cardíacos

Autor: Daniel Sehn Colao
Professor: Ranniery da Silva Maia

Resumo

Este documento tem como objetivo descrever um sistema classificador de problemas cardíacos com base em indicadores pessoais. Para isso, foi utilizado um modelo baseado em redes neurais aplicado à base de dados https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

Palavras-chave: Inteligência Artificial, Deep learning, Problemas cardíacos.

Lista de figuras

Sumário

1 Introdução

Neste trabalho, foi implementado um classificador baseado em Redes Neurais Feed-Forward com o intuito de realizar previsão de problemas cardíacos com base em respostas dadas por cidadãoes estadunidenses em um questionário feito pela instituição Behavioral Risk Factor Surveillance System (BRFSS), em 2020.

O questionário promovido pela BRFSS consiste em perguntas acerca de indicadores e status da saúde do entrevistado, como o índice BMI e se o usuário fumou mais de 100 cigarros na vida. Ao final deste, o cidadão responde se, anteriormente, reportou doença cadíaca coronária ou infarto do miocárdio.

O modelo em questão foi escolhido para testar o comportamento de uma rede neural feed-forward em tarefas de classificação que não exigem conhecimento de estados e informações anteriores. Além disso, para a base de dados escolhida, poucos dos trabalhos divulgados fazem uso de redes neuráis, optando por algoritmos de aprendizado de máquina, como regressão logística, K vizinhos mais próximos e árvore de decisão, como também as técnicas de ensemble learning RandomForest e Adaboost.

2 Coleta de dados

Conduzida nos Estados Unidos em 2020, a pesquisa feita pelo órgão BRFSS foi realizada para coletar dados de cidadãos estadunidenses com o propósito de investigar o relacionamento entre doença cadíaca coronariana ou infarto do miocárdio e indicadores e status de saúde.

Foram feitas as seguintes 18 perguntas:

- Você já apresentou doença cardíaca coronária ou infarto do miocárdio?

  Variável (dataset): HeartDisease. Esta será a nossa variável target no treinamento do modelo.

- Qual é o seu BMI?

  Variável (dataset): BMI.

  Variável contínua calculada pela divisão do peso da pessoa, em kilogramas, pelo quadrado da altura, em metros. Este indicador é mensurado para analisar se o peso de uma pessoa está saudável.

- Você fumou, pelo menos, 100 cigarros na sua vida?

  Variável (dataset): Smoking.

- Você toma mais de 14 bebidas alcoólicas por semana (homens) ? Para mulheres, o número perguntado foi pelo menos 7 por semana.

  Variável (dataset): AlcoholDrinking.

- Você já teve algum derrame?

  Variável (dataset): Stroke.

- Agora, pensando em sua saúde física, que inclui doenças físicas e lesões, por quantos dias nos últimos 30 não foi boa?

  Variável (dataset): PhysicalHealth.

- Pensando em sua saúde mental, por quantos dias nos últimos 30 dias sua saúde mental não foi boa?

  Variável (dataset): MentalHealth.

- Você tem muita dificuldade para andar ou subir escadas?

  Variável (dataset): DiffWalking.

- Qual é o seu sexo?

  Variável (dataset): Sex.

- Qual a sua categoria de idade?

  Variável (dataset): AgeCategory.

- Qual a sua raça?

  Variável (dataset): Race. Este atributo será removido na fase de pré-processamento por questões éticas.

- Você tem diabetes?

  Variável (dataset): Diabetic.

- Você pratica atividades físicas?

  Variável (dataset): PhysicalActivity.

- Em média, quantas horas você dorme em um período de 24 horas?

  Variável (dataset): SleepTime.

- Você diria que, em geral, sua saúde é...?

  Respostas: Poor, fair, good, very good e excellent.

  Variável (dataset): GenHealth.

- Você tem asma?

  Variável (dataset): Asthma.

- Você tem câncer de pele?

  Variável (dataset): SkinCancer.

- Sem incluir cálculos renais, infecção da bexiga ou incontinência, alguma vez te disseram que tinha doença renal?

  Variável (dataset): KidneyDisease.

3 Pré-processamento

A tarefa de pré-processamento consiste em manipular e preparar o dataset obtido na fase de coleta de dados para garantir ou aprimorar o desempenho de um modelo de aprendizado de máquina. Essa atividade é crucial em atividades dessa área de Inteligência Artificial, pelo motivo dos modelos aprenderem a partir das amostras disponibilizadas, pois os parâmetros são estimados com base na experiência.

O foco é transformar dados crus em um formato útil e eficiente, para que os modelos não sejam afetados por erros, inconsistências e ausência de valores presentes no dataset.

3.1 Eliminação de Atributos

Por questões éticas e não interferência nos resultados, o atributo "Race", o qual indica a raça do usuário entrevistado, foi removido do dataset.

3.2 Limpeza de Dados

O dataset foi carregado na estrutura de dados dataframe, disponibilizada pela biblioteca pandas, que contém o método dropduplicates() para eliminar amostras duplicadas dessa estrutura de dados. Essa tarefa contribui para o treinamento pelo fato de reduzir o viés (bias) da base de dados, reduzindo, dessa maneira, a possibilidade de overfitting. Cerca de 18.078 amostras foram eliminadas com a aplicação do método em questão.

Ademais, o atributo "BMI" é a única variável quantitativa contínua. Logo, a atividade de verificação e exclusão de outliers foi empregada apenas sobre tal atributo. A figura 1 mostra a distribuição das amostras com relação à feature BMI.

Foi adotada a estratégia de eliminação de amostras cujo valor do atributo BMI estivesse antes do 1o percentil ou após o 99o percentil. Cerca de 6.053 amostras foram excluídas. A figura 2 mostra o resultado dessa aplicação.

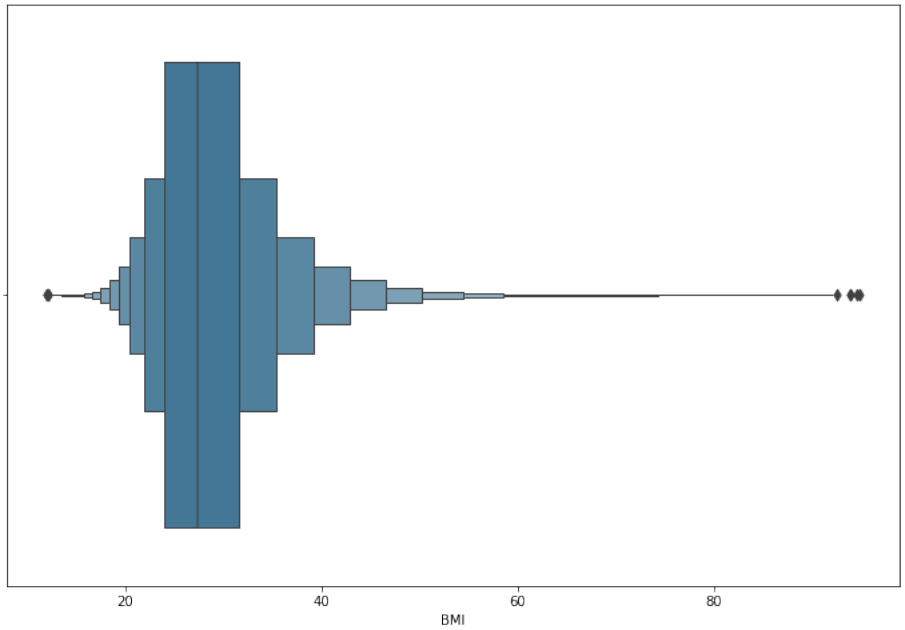

Figura 1: Outliers - atributo BMI

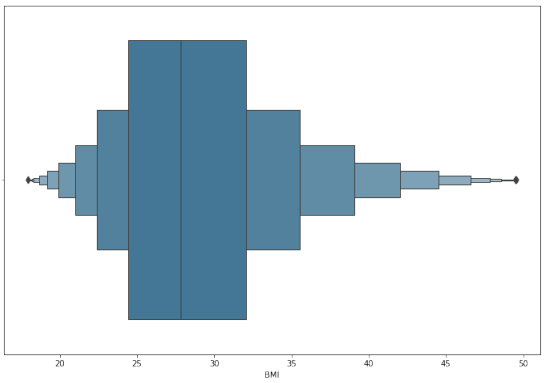

Figura 2: Resultado após remoção de outliers - atributo BMI

O dataset não apresentou valores ausentes, valores nulos nem incompletos.

3.3 Transformação de Dados

Redes Neurais trabalham somente com dados quantitativos. No entanto, o dataset utilizado neste trabalho contém, majoritariamente, variáveis categóricas. Com isso, foi necessário realizar uma transformação de variável qualitativa para quantitativa.

A estratégia adotada nessa tarefa foi o mapeamento de cada categoria de uma featura a um número inteiro. Por exemplo, as respostas “Yes” e “No” foram convertidas para 1 e 0, respectivamente.

Inicialmente, houve a tentativa de utilizar o OneHotEncoder para converter uma categoria em um número binário, visto que números inteiros são manuseados quando há uma certa ordem entre os possíveis resultados. No entanto, os modelos treinados apresentaram performances insatisfatíveis nesse teste.

3.4 Balanceamento de Dados

O dataset apresentou um desbalanceamento significativo no atributo "HeartDisease", classe que o modelo tentará realizar a previsão. A figura 3 retrata o desbalanceamento.

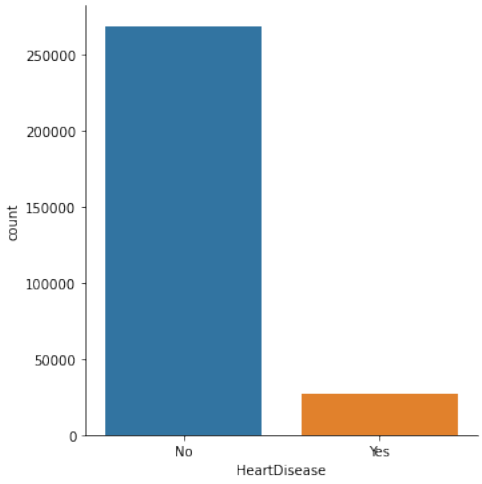

Figura 3: Desbalanceamento do atributo HeartDisease

Foi aplicada a técnica de Undersampling, caracterizada pela redução de amostras da classe predominante que, pela figura 3, trata-se da classe "No". Antes de realizar essa tarefa, foi particionado o dataset em conjuntos de treino e teste, pois o ideal é que se aplique apenas sobre o conjunto de treino, para não perder a distribuição das amostras reais.

Dessa maneira, o conjunto de treino ficou com um total de 42.674 amostras, sendo 21.337 amostras categorizadas com ´´Yes" (HeartDisease) e a outra metade com "No". O conjunto de teste, 59.133 amostras, representando 20% das amostras antes de realizar o balanceamento.

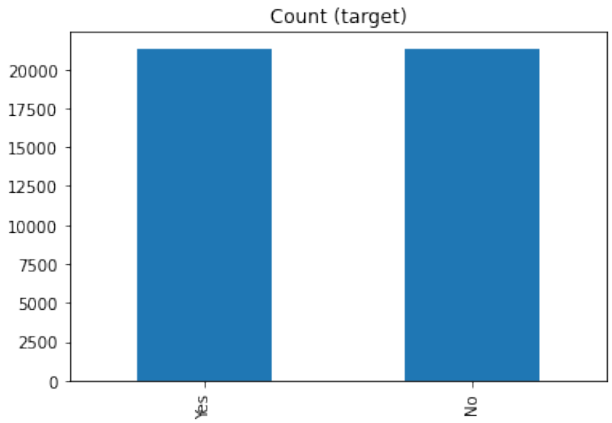

Figura 4: Resultado da aplicação de Undersampling

4 Fundamentação teórica

Em inteligência artificial, classificadores formam um conjunto de algoritmos de aprendizado de máquina que visam, a partir de experências, prever a categoria de uma determinada amostra. Esses algoritmos recebem dados para identificar relações entre os atributos e a saída, que é uma classe/categoria, para, posteriormente, receber dados desconhecidos e atribuir, cada um deles, a uma classe.

Várias técnicas de classificação encontram-se disponíveis atualmente, a Árvore de Decisão, o algoritmo SVM e a Regressão Logística são os principais algoritmos de Aprendizado de Máquina para a tarefa em questão. Além dessas, encontram-se disponíveis as técnicas de Ensemble Learning, como Random Forest e Bagging, que fazem uso de mais de um algoritmo de Aprendizado de Máquina com a finalidade de obter melhor performance preditiva em comparação com um algoritmo ML sozinho.

Deep Learning é uma subárea de Aprendizado de Máquina que visa reproduzir, por meio de redes neurais profundas, a maneira pela qual nós humanos aprendemos, ou seja, simular o comportamento do cérebro humano, que é regido por neurônios. Essa área evoluído bastante recentemente pela alta disponibilidade de dados e desenvolvimento de aparatos computacionais.

Para este trabalho, foi implementada uma rede neural e um modelo Random Forest com o objetivo de encontrar o relacionamento entre os atributos de saúde questionados na pesquisa e doença cardíaca coronariana ou infarto do miocárdio.

5 Modelos implementados

5.1 Rede Neural

O modelo consiste em uma rede neural profunda composta por 6 camadas: uma de entrada, quatro ocultas e outra de saída.

Por se tratar de redes neurais, foi necessário converter variáveis categóricas para variáveis numéricas durante a fase de pré-processamento. As variáveis qualitativas convertidas foram: Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Diabetic, Physical Activity, GenHealth, Asthma e SkinCancer.

A ferramenta manuseada para treinar este modelo foi a biblioteca Keras.

Os hiperparâmetros ajustados foram o número de épocas e o tamanho do batch, o qual se refere ao número de amostras processadas antes de atualizar os parâmetros do modelo.

A base de dados havia sido particionada, previamente, em conjunto de treino e teste na fase de pré-processamento. O conjunto de treino foi particionado novamente, sendo 20% dos dados inseridos em um novo conjunto, o conjunto de validação.

5.2 Random Forest

O modelo consiste em uma implementação da técnica Random Forest de Ensemble Learning, a qual emprega várias árvores de decisão com a finalidade de obter uma performance melhor em comparação com apenas o treinamento de uma única árvore de decisão.

Ao realizar o primeiro treinamento, foram ajustados duas vezes os hiperparâmetros visando melhorias na acurácia do modelo. O número de árvores de decisão, o número mínimo de amostras em uma folha, o número mínimo de amostras para ramificação e a profundidade máxima da árvore foram os hiperparâmetros ajustados nesse trabalho.

A ferramenta manuseada para treinar este modelo foi a biblioteca Scikit Learn.

No treinamento da Random Forest, foram utilizados apenas os conjuntos de treino e teste, construídos na fase de balanceamento (pré-processamento).

6 Resultados

6.1 Modelo Rede Neural

Hiperparâmetros:

- Número de épocas: 50
- Batch size: 64

6.1.1 Curvas de Aprendizado

A figura 5 mostra o gráfico da acurácia de treino e validação ao longo das épocas. Nota-se que, a partir da época 40, o modelo começa a apresentar comportamento de overfitting, por causa do aumento do erro de generalização. O mesmo acontece na figura 6, a qual demonstra a função Loss ao longo das épocas do treinamento.

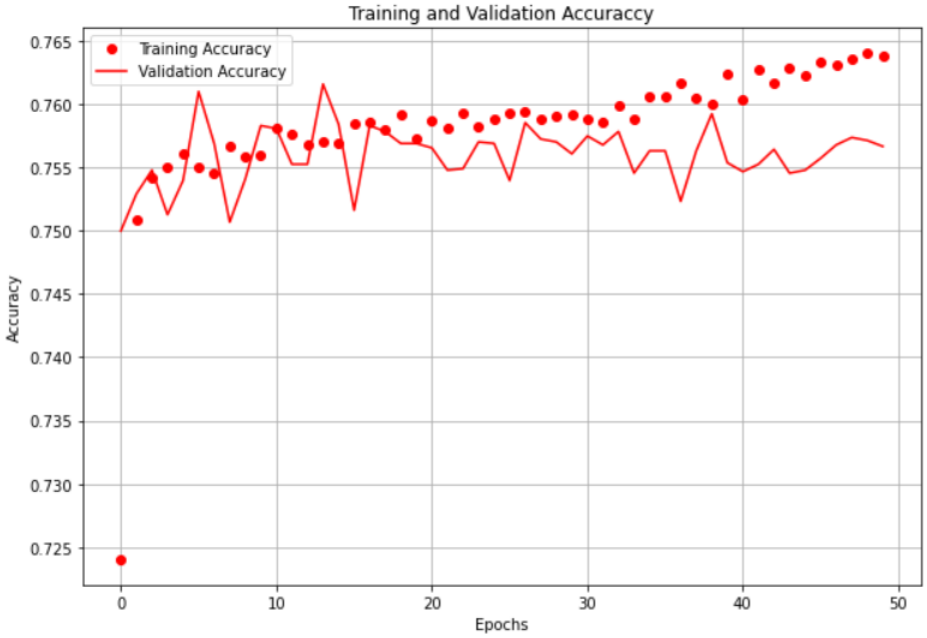

Figura 5: Curva de aprendizado - Acurácia

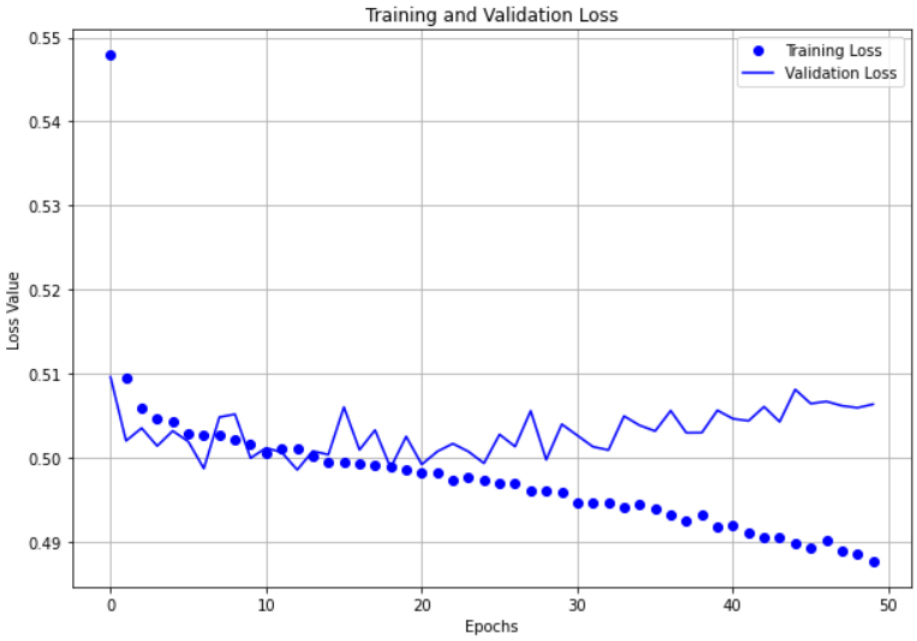

Figura 6: Curva de aprendizado - Loss

6.1.2 Matriz de Confusão

A matriz de confusão, descrita na figura 7, foi construída com base nos resultados obtidos no conjunto de teste, apresentando uma acurácia de, aproximadamente, 72.38%.

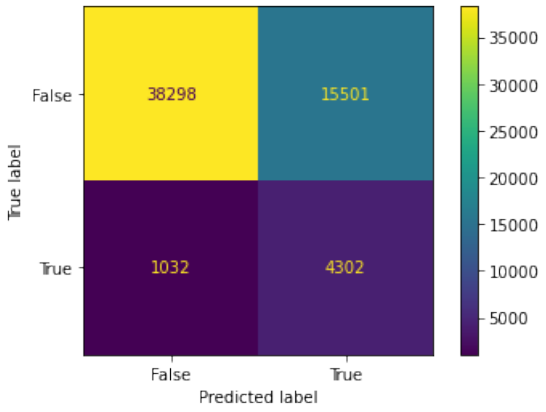

Figura 7: Matriz de Confusão

É importante salientar o erro no quadrante inferior esquerdo. 1.032 pessoas que tiveram problemas cardíacos foram diagnosticas pelo modelo como se não os tivessem.

6.1.3 Ajuste de Hiperparâmetros

Foram ajustados os hiperparâmetros número de épocas e batch-size com o intuito de melhorar a performance da rede neural.

Hiperparâmetros:

- Número de épocas: 140
- Batch size: 128

As figuras 8 e 9 mostram as curvas de aprendizado após o ajuste desses hiperparâmetros.

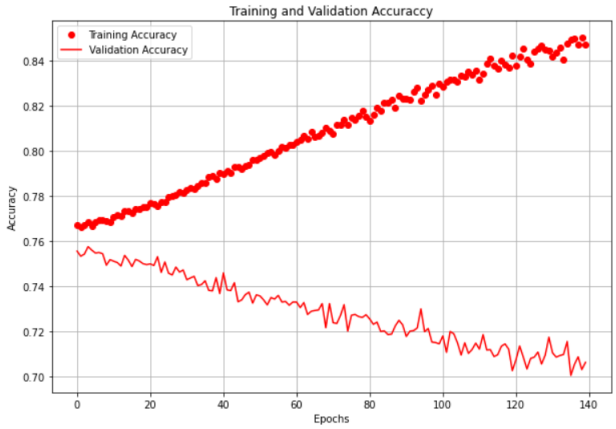

Figura 8: Curva de aprendizado - Acurácia

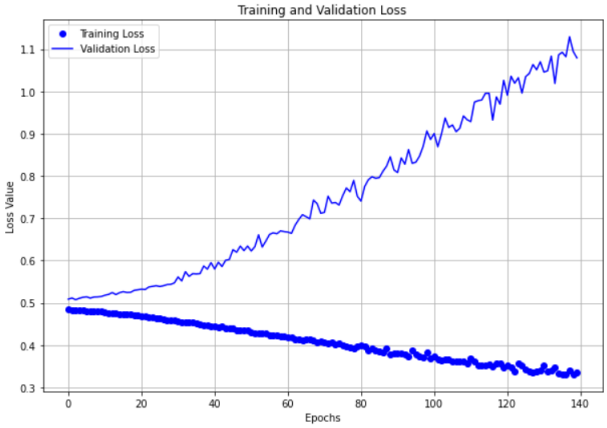

Figura 9: Curva de aprendizado - Loss

Claramente houve overfitting, pelo acentuado erro de generalização conforme o aumento da época.

6.2 Modelo Random Forest

6.2.1 Primeiro treino

Acurácia: 72.506%

Hiperparâmetros:

- max-depth: 20
- n-estimators: 100
- min-samples-leaf: 2
- min-samples-split: 5

A figura 10 mostra a matriz de confusão deste treinamento.

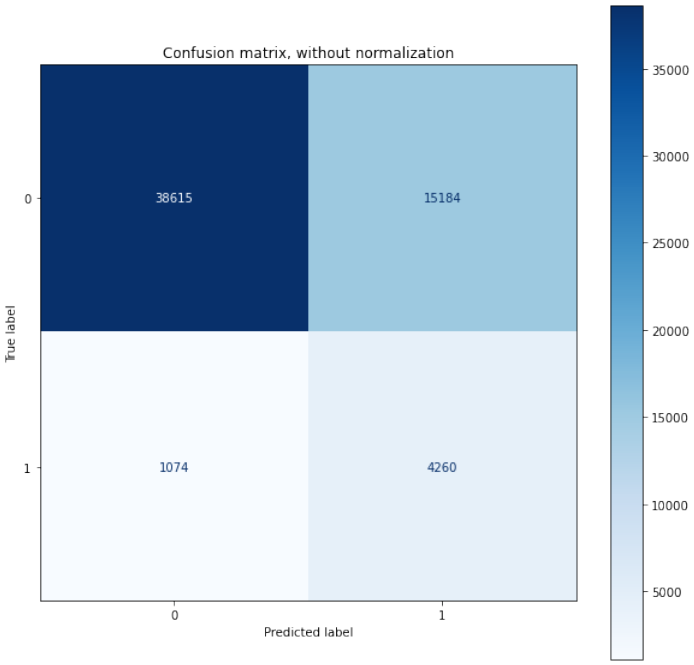

Figura 10: Matriz de Confusão - Treino 1

6.2.2 Segundo treino

Acurácia: 72.420%

Hiperparâmetros:

- max-depth: 50
- n-estimators: 200
- min-samples-leaf: 3
- min-samples-split: 8

A figura 11 mostra a matriz de confusão deste treinamento.

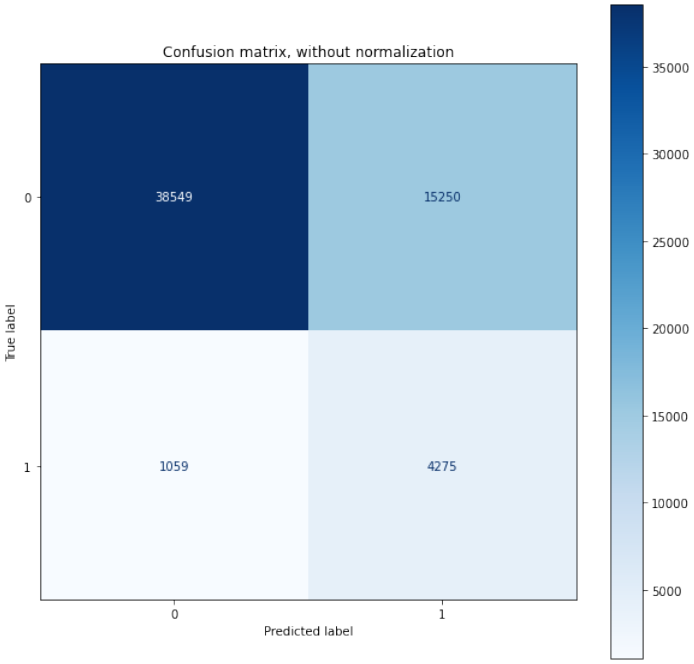

Figura 11: Matriz de Confusão - Ajuste de hiperparâmetros 1

6.2.3 Terceiro treino

Acurácia: 72.508%

Hiperparâmetros:

- max-depth: 80
- n-estimators: 400
- min-samples-leaf: 4
- min-samples-split: 10

A figura 10 mostra a matriz de confusão deste treinamento.

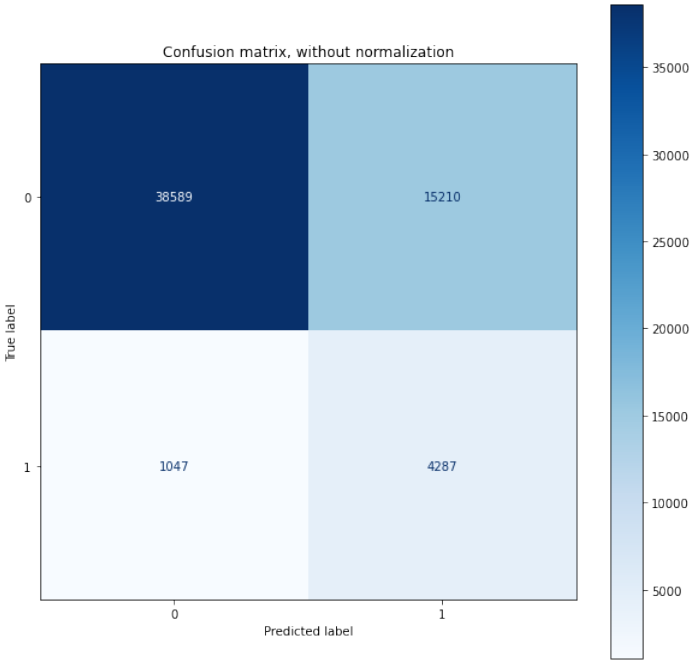

Figura 12: Matriz de Confusão - Ajuste de hiperparâmetros 2

7 Conclusão

Levando em consideração o que foi apresentado neste trabalho, foram implementados dois modelos classificadores com o objetivo de realizar o diagnóstico de problemas cardíacos. O primeiro consiste em uma rede neural feed-forward, enquanto o segundo, uma Random Forest. A base de dados utilizada pode ser encontrada no link https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

O classificador baseado em rede neural obteve uma acurácia de, aproximadamente, 75%, e o segundo classificador alcançou uma acurácia de 72%. Tais perfomances são boas, porém há margens para melhoria, que serão discutidas na próxima seção (7.1).

7.1 Trabalhos futuros

Como sugestões para melhorar a performance dos classificadores treinados, podemos destacar as seguintes: normalização de atributos quantitativos contínuos, verificação de inconsistências no dataset, como a verificação do atributo "SleepTime" para garantir que não haja amostras com valores superiores a 20 horas, além da opção de não utilizar undersampling para o balanceamento de classes, com o objetivo de não perder informações.

Referências