Relatório do Trabalho 3 - Introdução à Inteligência Artificial

Ian Nery Bandeira - 17/0144739
André Carvalho Marques - 15/0005491

Resumo: Esse trabalho consiste em desenvolver e aplicar um código de "Random Forest" em dados provenientes do Hospital Israelita Albert Einstein de 5644 casos de COVID-19, para predizer questões como a relevância de outros exames laboratoriais no diagnóstico da COVID, ou como a relevância de outros exames predizem a severidade de internação do paciente, seja com COVID ou não.

Palavras-chave: Inteligência Artificial; Random Forest; COVID-19; Kaggle;

1 - Introdução

O algoritmo de Random Forest (RF) é um método de aprendizado por agrupamento que pode ser utilizado para tarefas de classificação ou regressão, e se destaca pela sua simplicidade e capacidade de predição.

O conceito de aprendizado por agrupamento se baseia na idéia de combinar vários modelos de predição menores, treiná-los para uma mesma tarefa, e convergir seus resultados individuais em um modelo de predição que os agrega, de forma a ter um resultado de predição mais robusto. No caso de RF's, o modelo de predição menor utilizado para criar a floresta, é chamado de Árvore de Decisão, que consiste em um mapa dos possíveis resultados de uma série de escolhas relacionadas para tomar uma decisão. O método que dá nome a aleatoriedade de uma RF consiste em selecionar linhas do grupo de dados original aleatoriamente, tendo como inclusive uma única linha ser selecionada duas ou mais vezes, para compor uma árvore de decisão; e o método que dá nome à "Floresta", consiste na execução do primeiro método para várias árvores.

No problema apresentado, nos foi dada uma planilha com 5644 diagnósticos de pacientes não identificados quanto a nome e idade, e com exames normalizados para média 0 e variância 1; e duas tarefas foram requeridas: Primeiro, se é possível predizer o diagnóstico de COVID-19 de um paciente baseado nos outros exames presentes na planilha, e então se é possível a predição de "severidade" para com a internação do paciente, ou seja, se ele foi tratado em casa, em uma enfermaria, em uma unidade de tratamento semi-intensiva ou em uma UTI.

Esse relatório é organizado como segue. Na seção 2, são expostos os materiais de base utilizados para o funcionamento do programa, bem como os métodos empregados para gerar e sintetizar os dados apresentados na seção seguinte. Na seção 3 são apresentados os dados e gráficos provenientes da execução do programa para ambas as predições requeridas. Na seção

4, serão analisados os dados apresentados na seção anterior, e na seção 5 teremos as considerações finais acerca do projeto.

2 - Materiais e métodos

Como fundamento teórico, foram utilizados os slides e vídeo aulas do professor Díbio, bem como o link disponibilizado por ele:

An Implementation and Explanation of the Random Forest in Python

Como ferramentas empregadas no trabalho, foram utilizadas para a execução das Tarefas 1 e 2, o tratamento de planilhas .xlsx em um dataset manipulável pela biblioteca "*pandas*". Pela biblioteca "*sklearn*", utilizamos: o modelo de seleção de índices *KFold*, para criar os índices utilizados pelo training e test sets; o codificador *OrdinalEncoder*, que codifica valores categóricos (strings) em valores numéricos, para serem utilizados corretamente pela RF; o *RandomForestClassifier*, que gera o modelo de RF para depois ser executado utilizando os dados do training set; os medidores de predição *predict_proba* e *predict* para as Tarefas 1 e 2 respectivamente, com o objetivo de validar o modelo de RF já treinado com os test sets; e a métrica de predição *accuracy_score*, para gerar a acurácia desejável de predição para os valores da Tarefa 2. Para a Tarefa 2, especificamente, foi criado um índice de "severidade", que determina se a pessoa foi tratada em casa (com peso 0), em enfermarias (peso 1), em unidades semi intensivas de tratamento (peso 2), ou em UTI's (peso 3).

Para síntese dos dados sensíveis ao que foi pedido no trabalho, foram utilizadas as bibliotecas "*os",* para receber o caminho onde o projeto está, para salvar os arquivos e imagens nas pastas correspondentes; e a biblioteca "*pyplot*", proveniente da *"matploptlib",* utilizada para auxiliar as bibliotecas "*scikitplot*", responsável por gerar o gráfico da curva ROC e o índice AUC para a Tarefa 1, e "*shap*", responsável por gerar os gráficos que sumarizam as colunas mais importantes para a decisão da RF.

3 - Resultados: quadros, gráficos e figuras

Segue o gráfico da curva ROC para a Tarefa 1, junto de *summary plots* e *dependency plots* que indicam a importância de cada coluna para a predição dos dados das Tarefas 1 e 2:

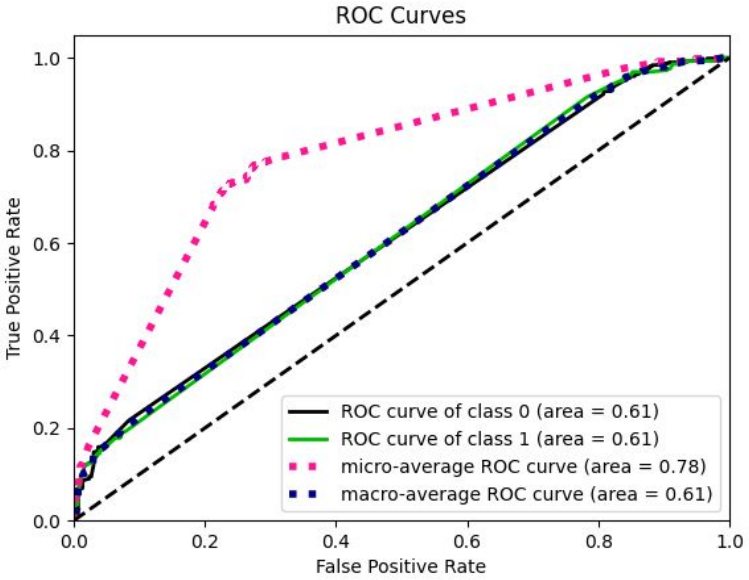

Gráfico 1 – Diagrama da curva ROC para a Tarefa 1, com AUC de 61%

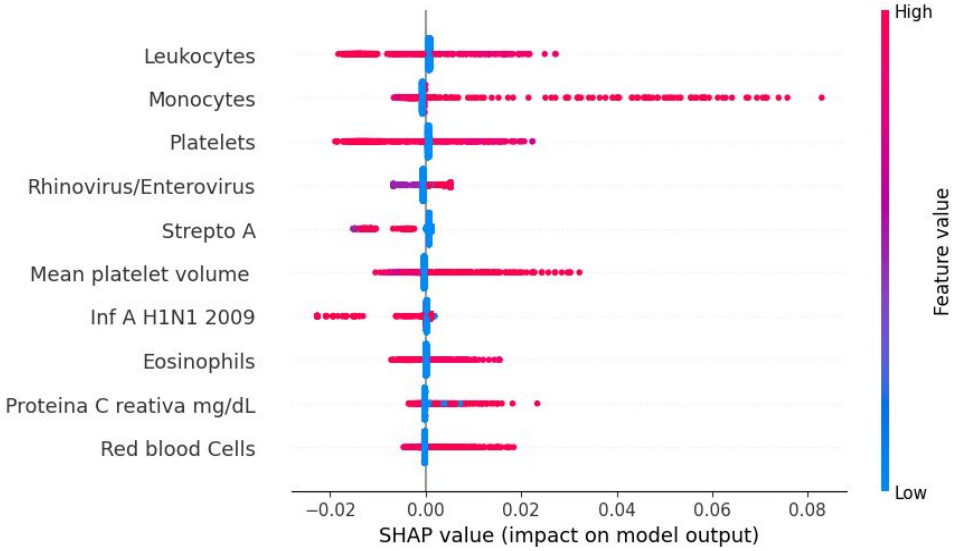

Gráfico 2 – *Summary plot* do modelo de RF para Tarefa 1, com as 10 colunas mais significantes.

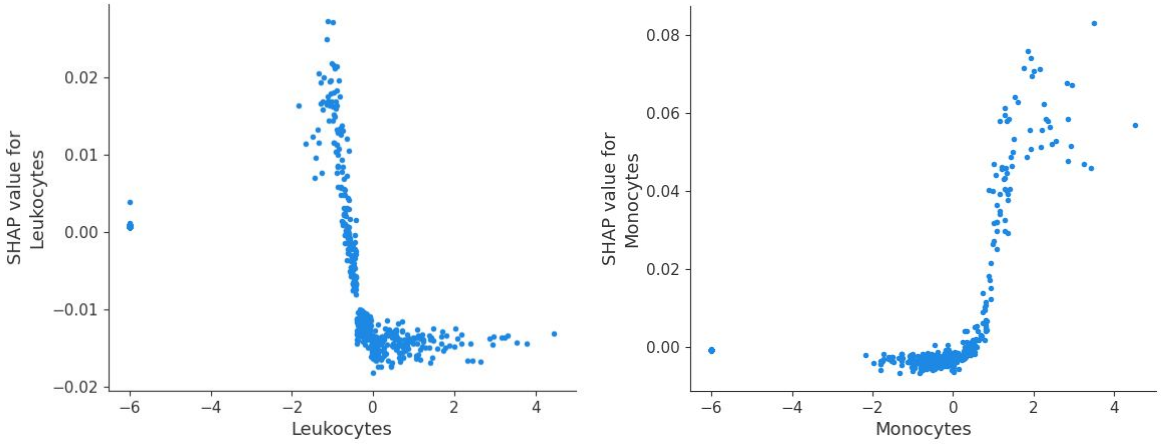

Gráfico 3 – *Dependency plots* das duas colunas mais significativas provenientes da análise do Gráfico 2.

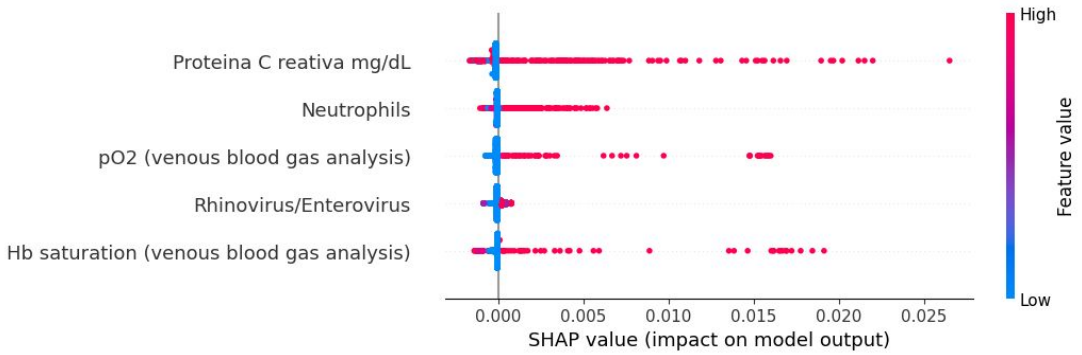

Gráfico 4 – *Summary plot* do modelo de RF para Tarefa 2, com as 5 colunas mais significantes.

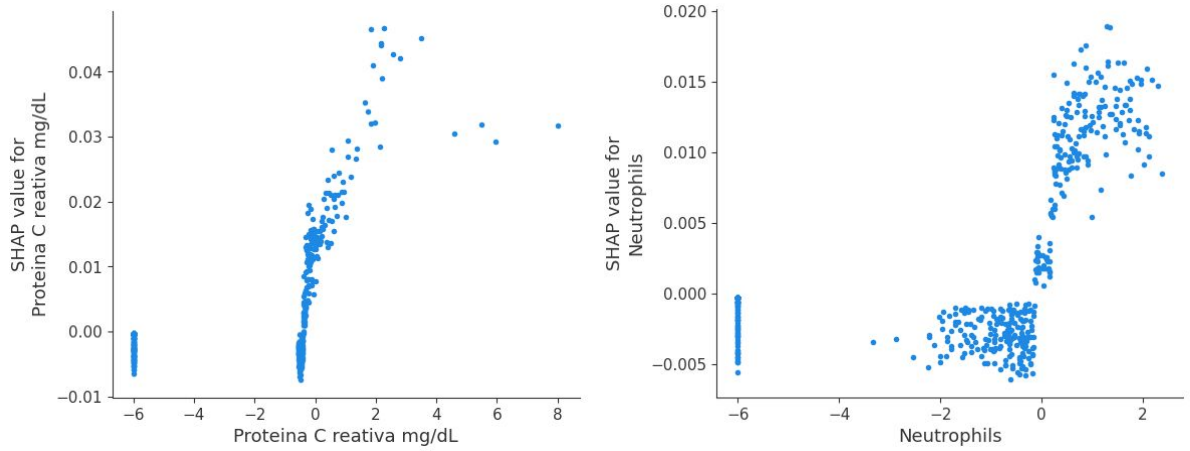

Gráfico 5 – *Dependency plots* das duas colunas mais significativas provenientes da análise do Gráfico 4.

4 - Análise de Resultados

Para a Tarefa 1, temos que a curva ROC gerou um AUC, ou seja, um índice de desempenho do modelo de **0.61,** e como pode ser observado, o *micro-average* teve índice relativamente alto, que mostra que as amostras não estão balanceadas entre si, apesar de normalizadas**.** O índice poderia alcançar até 0.64, mas como foi requerido que apenas exames laboratoriais fossem utilizados, retiramos o quantil de idade do dataset. Para a análise do Gráfico 2, criamos dois dependency plots dos dois fatores mais determinantes, a quantia de Leucócitos e Monócitos, mostrados no Gráfico 3. Pode-se observar, que apesar de muitos casos caírem no valor "-6", utilizado para substituir os valores nulos da tabela, eles não possuem importância nos valores do gráfico, enquanto os valores reais, para os altos "valores shap", que é um índice representativo da mudança dos valores em escala log, representam casos de leucopenia e monocitose, ambas sendo importantes para o diagnóstico de infecções virais, enquanto a segunda atrelada a idade do paciente, pode ser um índice de severidade da doença.

Para a Tarefa 2, que teve um *accuracy_score* de 90%, faremos uma análise similar ao que foi feito no parágrafo acima, para os Gráficos 4 e 5. Pode ser observado, pelo gráfico 4, que a importância do valor do PCR se dá principalmente nos valores mais elevados do exame,

que realmente adiciona à severidade da condição do paciente, já que o PCR sinaliza o grau de inflamação persistente e é um marcador de mau prognóstico na síndrome de angústia respiratória aguda (SARA). Já a análise dos neutrófilos, temos que a leucocitose neutrofílica é também é um indicativo infeccioso importante, já que é uma resposta natural do corpo à infecções. Outros indicativos importantes são os altos níveis de saturação de hemoglobina e pO2 no sangue venoso, que também é um indicativo de mal funcionamento do corpo em vista a problemas respiratórios. Nestes dados, o valor "-6" também foi utilizado para substituir os valores nulos da tabela.

5 Considerações Finais/Conclusões

Apesar dos dados apresentados estarem próximos do que seria a realidade, e isso acarretar na falta de diversos dados para uma comparação não enviesada, ou seja, é extremamente improvável que uma pessoa que tenha dado negativo no teste para a COVID-19 continue a fazer exames de diagnóstico, a análise feita pela RF foi satisfatória de modo que previu com um bom grau exatidão exames pertinentes com um diagnóstico da doença ou de uma maior severidade desta. Como não foram analisados os quantis de idade, não pudemos fazer a correlação deste com a severidade, pois como sabemos, idade é um dos fatores que leva um indivíduo a estar no "grupo de risco", e isso aumentaria bastante os índices de predição tanto da Tarefa 1 quanto da Tarefa 2.

Referências Bibliográficas

PABLO CASAS **A gentle introduction to SHAP values in R,** 2019, disponível Aqui.

Micro Average vs Macro average Performance in a Multiclass classification setting, disponível Aqui.

LÍVIA PESSÔA DE SANT'ANNA, **Alterações dos leucócitos na Covid-19 e valor prognóstico**, disponível Aqui.

Pence B. D. (2020). **Severe COVID-19 and aging: are monocytes the key?**. GeroScience, 42(4), 1051–1061. disponível Aqui.

MARI TERRITO, **Leucocitose neutrofílica**, 2020, disponível Aqui.

LÍVIA PESSÔA DE SANT'ANNA, **Parâmetros hematológicos em pacientes com infecção por coronavírus**, disponível Aqui.

Exames laboratoriais na Covid-19 – Quais solicitar e o que esperar?, disponível Aqui.

CARLOS HENRIQUE DE SOUSA, **Quais sinais ajudam a predizer má evolução na Covid-19?**, disponível Aqui.

WILL MCGUINIS, **Category Encoders Documentation**, disponível Aqui.

SCOTT LUNDBERG, **SHAP Documentation**, disponível Aqui.