INF-0615 – Aprendizado de Máquina Supervisionado I

# Trabalho 3 - Árvores de Decisão e Florestas Aleatórias

COVID-19 World Outbreak

**Data de Entrega: 06/05/2020**

## 1 Descrição do Problema

A pandemia do vírus COVID-19 tem afetado diretamente e indiretamente todas as sociedades do mundo. Muitas pessoas de diferentes nações, idades e classes sociais tem sido contaminadas, apresentando diferentes quadros clínicos de reação ao vírus. A doença, que iniciou-se na cidade Wuhan na China, em três meses, já havia se alastrado para grande parte dos países do globo, levando a Organização Mundial da Saúde a decretá-la como uma pandemia em 12/03/2020 [1].

Nesse trabalho, **você irá inferir o possível estado do paciente diagnosticado com o vírus COVID-19 dentre as três possíveis classes: em tratamento, falecido ou recuperado.** A base de dados contém os seguintes atributos:

- **Case in Country**: Qual é o número desse caso no país referenciado;
- **Reporting Date**: Data do relatório feito sobre o estado clínico do paciente;
- **Country**: País referente ao caso reportado;
- **Gender**: Gênero do paciente;
- **Age**: Idade do paciente;
- **Symptom Onset**: Data em que o paciente começou a sentir os sintomas;
- **If Onset Approximated**: Valor binário. Informa se o valor do atributo *Symptom Onset* é aproximado ou exato;
- **Hosp Visit Date**: Data em que o paciente foi ao hospital;
- **International traveler**: Informa se o paciente fez viagens internacionais;
- **National traveler**: Informa se o paciente fez viagens nacionais;
- **Exposure Start**: Data em que iniciou-se a exposição ao vírus;
- **Exposure End**: Data em que terminou a exposição ao vírus;
- **Traveler**: Informa se o paciente realizou alguma viagem;
- **Visiting Wuhan**: Valor binário. Informa se o paciente visitou a cidade de Wuhan;
- **From Wuhan**: Valor binário. Informa se o paciente mora em Wuhan;
- **Label (target): Indica o estado do paciente no dia que foi feito o report sobre seu estado clínico. Os possíveis valores são: "OnTreatment", "dead"ou "recovered"**.

Todos os atributos relacionados às datas foram divididos em 15 categorias, de 0 a 14, representando períodos a partir do dia 01/01/2020. Assim, um atributo na categoria 0, refere-se a uma data anterior ao dia 01/01/2020 (dezembro de 2019). Um atributo na categoria 1, refere-se a uma data entre o dia 01/01/2020 e o dia 05/01/2020. Atributo com valor na categoria 2, refere-se do dia 06/01/2020 ao dia 10/01/2020, e assim sucessivamente com intervalos de 5 em 5 dias.

O atributo relacionado à idade também foi categorizado em 8 faixas etárias: categoria 1: de 0 a 10 anos; categoria 2: de 11 a 20 anos; categoria 3: de 21 a 30 anos; categoria 4: de 31 a 40 anos e assim sucessivamente de 10 em 10.

Por fim, para facilitar a modelagem, criamos 10 categorias referente ao atributo *Case in Country*, começando em 0 até 200, em intervalos de 20 em 20. Ou seja, a categoria 1 refere-se ao intervalo de 0 a 20, a categoria 2 ao intervalo de 21 a 40 e assim sucessivamente.

Repare que haverá valores negativos $(-1)$ na base de dados. Eles se referem a Not Assigned (NA) na base original antes do processamento. Esse valor apenas transforma o valor NA em uma nova categoria de forma que vocês não precisam se preocupar com atributos NA. Além disso outras features, que tinham valor NA originalmente, tiveram esse valores substituídos pela categoria mais comum em um mesmo país (atributo *country*).

Este dataset provém originalmente de um conjunto de dados reunidos por diversos países no mundo e colocados à disposição para fins acadêmicos e de pesquisa.

## 2 Tarefas

Neste Trabalho, pedimos que você:

1. Inspecionar os dados de treinamento. Quantos exemplos há de cada classe? O dataset está desbalanceado ?
2. Treinar uma árvore de decisão como baseline.
3. Treine outras árvores de decisão variando o tamanho das árvores geradas. Plote o erro no conjunto de treinamento e validação pela profundidade da árvore de decisão.
4. Explore pelo menos 2 possíveis subconjuntos de features para treinar uma árvore de decisão. Reporte o erro no treino, validação e teste.
5. Treine várias florestas aleatórias variando o número de árvores. Plote o erro no conjunto e treinamento e validação variando o número de árvores geradas.
6. Calcule a matriz de confusão, os verdadeiros positivos para cada classe e a acurácia normalizada no teste para os melhores modelos (árvore com melhor profundidade, floresta com melhor número de árvores e árvore com melhor subconjunto de features).
7. Escreva um relatório de no máximo 5 páginas reportando:
   (a) A diferença de desempenho entre o *baseline* e os outros modelos mais complexos gerados.
   (b) Houve overfitting ? Houve underfitting ? Como você lidou com o desbalanceamento ?
   (c) 2 páginas com conclusões explicando a diferença entre os modelos e o porquê que estas diferenças levaram a resultados piores ou melhores.

Repare que agora você define o conjunto de validação. Tome a base *COVID19_training_validation_set_cleaned.csv* e faça o split em 80% para treinamento e 20% para validação. Lembre-se de manter o mesmo conjunto de validação para todos os modelos.

## 3 Opcionais

Como mencionado, a base original apresentava valores NA e possíveis dados espúrios. Neste exercício opcional, disponibilizaremos também esta base (treinamento e teste) e pedimos que você:

1. Aplique algum processamento para substituir o valor NA por algum outro valor. Neste link [2] há a explicação dos métodos mais comuns.
2. Treine uma árvore de decisão.

3. Treine uma floresta aleatória.

4. Reporte o erro no conjunto de teste e escreva suas conclusões. Esse resultado foi melhor que os modelos treinados com as bases processadas (cleaned) ?

Se você optar fazer esta parte, seu relatório pode conter até 7 páginas.

# 4 Arquivos

Os arquivos disponíveis no Moodle são:

- *COVID19_training_validation_set_cleaned.csv*: conjunto de dados processados sem atributos NA (cleaned) para serem utilizados como treinamento e validação;
- *COVID19_test_set_cleaned.csv*: dados de teste processados sem atributos NA que serão **disponibilizados 2 dias antes do prazo final de submissão**;
- *COVID19_training_validation_set_original.csv*: conjunto de dados sem processamento com atributos NA para a parte opcional;
- *COVID19_test_set_original.csv*: dados de teste sem processamento com atributos NA para a parte opcional. Serão também **disponibilizados 2 dias antes do prazo final de submissão**;

# 5 Referências


1. *WHO announces COVID-19 outbreak a pandemic.* World Health Organization. `http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic`

2. *All About Missing Data Handling.* `https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184`