Laboratório 2
Estimativa de Volume de Chuva

O arquivo "*Dados_Radar_Estacao_Completo_2018_2022.csv*" contém 2.856.380 registros capturados por radares entre os anos de 2018 e 2022. Os dados são capturados a cada 15 minutos em 27 estações meteorológicas espalhadas pelo estado do Paraná. A figura abaixo mostra algumas das variáveis disponíveis. O valor estimado de chuva (*Tp_es*t), capturado pelo pluviômetro é a variável alvo.

	time	UH	UV	DBZH	DBZV	KDP	ZDR	RHOHV	lat	lon	Est	Tp_est
0	2018-01-09 12:30:00	NaN	NaN	NaN	NaN	NaN	-0.72	0.40	-22.919829	-53.156094	Loanda	0.0
1	2018-01-09 12:30:00	-8.03	NaN	-9.92	NaN	-0.50	2.41	0.54	-24.229322	-53.341064	Porto_Formosa	0.0
2	2018-01-09 12:30:00	6.00	5.17	4.74	3.38	1.84	1.33	0.68	-23.083564	-52.455313	Paranavai	0.2
3	2018-01-09 12:30:00	NaN	NaN	NaN	NaN	NaN	5.56	0.47	-24.064595	-52.406170	Campo_Mourao	0.0
4	2018-01-09 12:30:00	NaN	-8.08	NaN	NaN	NaN	-4.25	0.49	-24.542079	-52.975897	Ubirata	0.0

A base de dados é fortemente desbalanceada com a maior parte dos valores para **Tp_est** igual a zero. Todos os registros com Tp_est = 0 podem ser descartados. Outro fato a ser observado nessa base é a ausência de valores em algumas características. Quando o valor do radar não pode ser capturado, NaN é colocado no valor da característica, como ilustrado na figura abaixo. Para esses casos, existem duas alternativas: eliminar as linhas que tem alguma variável NaN ou preencher esses valores de alguma forma, como por exemplo com a média ou vizinhos mais próximos dos valores daquela característica. Esse processo deve ser realizado com cuidado, pois a base de dados tem um aspecto temporal.

Seu trabalho consiste em treinar um regressor para estimar *Tp_est*. Você pode usar todos os dados entre 2018 e 2021 para treinar e validar o modelo. O ano de 2022 deve ser usado somente para testes.

Importante: Lembre-se que 2022 não deve ser usado para escolha dos parâmetros do classificador nem para o processo de limpeza e correção dos valores NaN. Isso configura vazamento de dados (*data leaking*).

Nos seus experimentos você deve reportar o erro médio quadrado (MSE) e o erro médio absoluto (MAE). Para cada experimento, apresente também um gráfico de dispersão no qual o eixo X tem o valor de **Tp_est** e o eixo Y tem o valor predito pelo seu algoritmo.

Você pode usar todos os algoritmos vistos em sala de aula, ou seja, não utilize técnicas de ensemble como Random Forest e Gradient Boosting.

Finalmente, você pode escolher treinar um modelo somente, ou vários modelos, um para cada estação, ou grupo de estações, tendo em vista a particularidade de cada estação meteorológica.

O que deve ser entregue:

Um arquivo python (jupyter notebook) documentado.