Neste laboratório, você será encarregado de criar um modelo preditivo de Machine Learning que será então implantado como uma ferramenta de Triagem Virtual (TV). A TV é uma técnica computacional que avalia a atividade biológica desejada de milhares de moléculas. Ao utilizar essa tecnologia, podemos diminuir significativamente o tempo e os custos envolvidos no pipeline de desenvolvimento de medicamentos.

Neste laboratório, você será o Engenheiro de Machine Learning em uma empresa farmacêutica em ascensão. Sua equipe vem trabalhando no problema de identificar novos tratamentos para a doença de Chagas. Seu trabalho é treinar um modelo de machine learning que possa ser usado como ferramenta de triagem virtual para avaliar muitas moléculas e identificar possíveis candidatos para o tratamento. Após algumas pesquisas, você encontrou um conjunto de dados promissores de um ensaio in vivo detalhado em que milhares de moléculas foram avaliadas para determinar sua atividade biológica para suprimir o Trypanosoma Cruzi, o agente causal da doença de Chagas. Com sua equipe, você decide que este conjunto de dados de alta qualidade pode ser usado para treinar o modelo de machine learning.

Os Engenheiros de Características em sua equipe foram capazes de extrair diferentes tipos de características moleculares que descrevem essas moléculas e geraram o seguinte conjunto de dados: cruzi\_comp\_activity\_molecular\_features.csv. No entanto, eles também estão notificando você de que, para algumas moléculas, algumas características moleculares não puderam ser extraídas ("NumberOfAromaticRings" e "TPSA"). Eles deixaram esse problema para você, o Engenheiro de ML, decidir como lidar com ele.

O resultado do ensaio in vivo é registrado na coluna "EC50" do conjunto de dados. De acordo com o relatório, qualquer molécula com uma atividade de EC50 de menos ou igual a 1 𝜇M é considerada ativa. Em outras palavras, o relatório especifica que, para uma molécula ser considerada ativa, seu valor de EC50 não deve ser maior que 1. Essas moléculas ativas podem então ser priorizadas como possíveis candidatos para um novo tratamento contra a doença de Chagas.

Você precisará inspecionar os dados, pensar no problema e decidir se irá treinar um modelo de regressão ou classificação para a tarefa em questão. Você precisará relatar uma métrica de desempenho correta e relevante para seu(s) modelo(s).

Algumas observações:

Você pode usar qualquer um dos algoritmos que foram abordados na aula até agora, seja para regressão ou classificação.

Não importe nenhum outro classificador, como Random Forests, XgBoost, etc.

Round 2

Depois de conduzir alguns experimentos iniciais, sua equipe não está muito satisfeita com os resultados obtidos. Nesse ponto, há diferentes fatores que podem estar influenciando o desempenho do modelo. Um deles pode ser o tamanho insuficiente do conjunto de dados, ou seja, o número limitado de moléculas avaliadas. Outra possível razão pode ser a falta de descritividade das características utilizadas. Para investigar isso, você solicita aos Feature Engineers que extraiam outro tipo de características para as moléculas. Desta vez, eles geram um conjunto de dados chamado `cruzi\_comp\_activity\_structural\_features.csv`. Este arquivo contém as subestruturas químicas presentes em cada molécula que são conhecidas por serem informativas da atividade biológica.

Sua próxima tarefa é avaliar se essas novas características funcionam de fato melhor. Além disso, você pode experimentar uma combinação de ambos os tipos de características e ver o que funciona melhor.