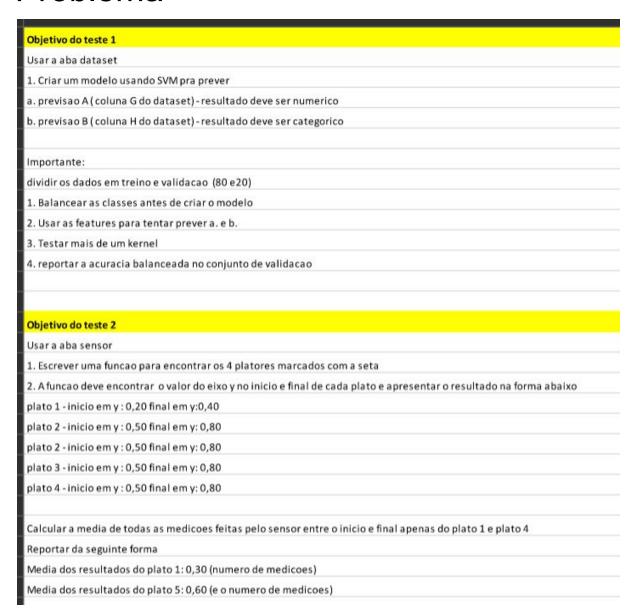
## **Problema**



## Observações sobre o teste 1

Se o objetivo é classificar ou predizer um modelo, ficar restrito a apenas um algoritmo não é uma boa estratégia. Sendo assim, eu sempre utilizo vários algoritmos para avaliar qual deles é mais adequado ao problema em questão.

Inicialmente, utilizei os seguintes algoritmos de classificação:

O que apresentou melhor resultado inicial para o problema foi o "KNN": KNeighborsClassifier().

Dividir os dados em 80/20 é uma boa estratégia para dados balanceados mas quando há restrição do número de exemplos da classe minoritária pode haver problemas. Utilizamos a técnica de Cross Validation para validar o modelo e separação dos dados de forma estratificada, ou seja, garantimos que haverá, nos subconjuntos dos dados, exemplos das duas classes (minoritária ou positiva e majoritária ou negativa).

O balanceamento, neste caso, é obrigatório pois há apenas 4 exemplos da classe minoritária, o que no meu ponto de vista, é irrisório para criarmos um modelo satisfatório. Mas, analisando o problema, minha proposta de solução é:

- codificar as features categóricas com técnicas mais sofisticadas para garantir um melhor aprendizado. Utilizei LabelEncoder para as classes e TargetEncoder para as demais features categóricas.
- 2. Eu testei a seguinte estratégia: como não há muitos exemplos disponíveis da classe positiva, peguei uma pequena amostra contendo todos os dados da classe positiva e alguns dados da classe negativa e apliquei o RandonOversampling para gerar mais dados da classe positiva. Em seguida, juntei estes dados novos com os dados originais (retirando os exemplos repetidos) e apliquei outras técnicas de oversampling como SMOTE, SVMSMOTE, DTOSMOTE, Geometric-SMOTE.
- 3. Até que tive bons resultados mas não consegui validar o modelo de forma aceitável, uma vez que não há exemplos extras disponíveis para validação.
- 4. O que percebi das classes é que há um dataset que diz "Bom" ou "Ruim" e outra numérica. Pude identificar que a Previsão A também é categórica pois 0.955 ocorre na maior parte das vezes (classe dominante) e os outros números são da classe minoritária. Portanto, pode parecer um resultado numérico (Regressão) mas não é. Trata-se de uma variável categórica, uma string e foi codificada também.
- 5. Percebi que na Feature 4 todos os exemplos são únicos. Data e hora pode ter relação com o clima, temperatura e até mesmo quem está operando as máquinas da produção. Foram codificados também. Resumindo, a única feature numérica é a Feature 3.
- 6. Não fiz a parte 2 dos sensores.