**Validação Cruzada**

**Uma conversa com o aluno do terceiro semestre**

-Oi Lucas, tudo bem?

-Oi João, mais ou menos e você?

-Eita Lucas mais ou menos porquê?

-Ah, eu estou fazendo um projetinho individual aqui e estou passando um sufoco, é sobre a validação de um modelo, não sei direito como eu devia fazer, nem se foi a melhor escolha de modelo, já que eu tenho outras opções.

-Já pensou em fazer uma Validação Cruzada?

-Eita, não sei nem o que é isso.

-Bem, validação cruzada é uma forma de validação do nosso modelo. Bem como você já está na etapa de validação, eu imagino que você já tenha feito a análise exploratória e a separação de teste e treinamento ne?

-Sim.

-Bem, primeiramente, na Validação Cruzada, é feita uma partição em nosso banco de treinamento, digamos que temos 600 dados, nesse caso, poderia ser feita 3 partições, onde cada partição teria 200 dados. Com isso, seria aplicado nosso modelo em cada combinação de nossas sessões. Ou seja, seria gerado um novo agrupamento de treinamento com as sessões 1 e 2, depois 2 e 3 e por fim 3 e 1 e para cada agrupamento de treinamento a sessão restante (no 1 e 2, seria a 3 por exemplo) seria nosso teste. Então basicamente, a Validação Cruzada cria novos treinamentos e testes a partir da divisão dos dados de nosso treinamento.

E não para por ai, é possível que a gente especifique a forma com a qual a validação cruzada avalie o desempenho de nosso modelo, checando se o y previsto das sessões batem com a realidade. Por exemplo, em uma aula de Machine Learning, minha validação cruzada de 4 sessões gerou 4 RMSEs, já que a primeira era a combinação de treino 1,2,3 para ver o teste da 4 e assim por diante até ter uma faixa de valores.

Agora, você não só tem um valor que avalia se seu modelo é bom, que é o RMSE, que você sabe que tem que ser baixo (com o cuidado de não ser muito baixo para o caso de Overfitting), como tem 4 ou 5, a partir do seu treinamento. E se você quiser, você pode fazer isso com todos os modelos em mente colocar esses valores de RMSE em uma lista e plotar um histograma com cada modelo e ver quem tem o menor RMSE, para escolher o modelo ideal! Espero ter ajudado!