# ANTES DE COMEÇARMOS,
# AS REGRAS DO JOGO:

1. O seu trabalho deve ser enviado em um arquivo de extensão .py ou .ipynb. Você deve enviar um único arquivo para o todo o seu trabalho, por meio do LMS e também uma cópia para o email ttavaresm@gmail.com, com o título “TRABALHO ADA”. A cópia é solicitada para prevenirmos o caso de alguma falha no envio ao LMS.

2. Na primeira linha do seu trabalho deve constar o seu nome completo.

3. Como o trabalho ficou um pouco grande, aceitarei trabalhos individuais e também em duplas. Se você estiver fazendo o trabalho em dupla: na primeira linha coloque os nomes completos das duas pessoas que compõem a dupla (para eu não precisar corrigir duas vezes o mesmo trabalho), mas peço que cada membro da dupla envie individualmente uma cópia do arquivo de resolução pela plataforma (para que eu possa atribuir notas para todos lá no LMS).

4. Leia com atenção todas as instruções. Não se esqueça de comentar os principais passos das suas soluções.

5. Você está estudando com professores que tem mais de 10 anos de experiência em Machine Learning. Acredite: se existe alguém que saberá caso você utilize o ChatGPT ou quaisquer IAs generativas para resolver estas questões, este alguém serei eu. Então, por favor, tenha em mente que este trabalho deve ser resolvido sem o uso de IAs generativas, e todas as respostas aqui fornecidas serão analisadas por arsenal técnico especializado, para garantir que não há plágios de quaisquer espécies (de fontes humanas ou não humanas). A entrega das resoluções solicitadas neste documento pressupõe o seu entendimento e a sua concordância com estes termos, e eventuais desvios serão acompanhados de consequências.

# CAPÍTULO 1: APRENDIZADO SUPERVISIONADO

Chegou o grande dia! Você finalmente foi admitido para trabalhar em um dos maiores bancos do país! É claro que você está de parabéns. E é claro também que nós vamos começar essa jornada através de um problemão!

Você precisa desenvolver um modelo de Machine Learning de crédito tendo em vista o histórico dos últimos clientes do banco, disponível aqui (arquivo train.py). Vamos chegar causando aquela boa impressão? Quem sabe não ganhamos um cartão de crédito sem anuidade para a vida toda?

Usando o dataset acima, desenvolva um modelo de crédito, passando por todas as principais etapas: pré-processamento de dados, escolha de hiperparâmetros, treinamento de modelo e avaliação de desempenho de modelo. Em suma, mostre que você é capaz de treinar um bom modelo de Machine Learning, que seja hábil para classificar um cliente em poor, standard ou good.

.

**EXTREMAMENTE IMPORTANTE**

- Nesta etapa você será avaliado quanto aos seus conhecimentos em (1) SVMs e (2) Boosting. Por isso, sua análise deve conter necessariamente ambas as abordagens (para boosting,

escolha um dos algoritmos demonstrados em aula). Uma boa ideia é treinar um modelo de cada tipo e depois comparar criticamente os resultados obtidos.
- Para cada abordagem acima, você deve necessariamente fornecer códigos em Python.
- Para cada abordagem acima, você deve fazer o possível para demonstrar ao máximo os seus conhecimentos. Testar explicitamente diferentes parâmetros, por exemplo, é uma boa ideia para isso. Realizar procedimentos de pré-processamento **quando necessário ou vantajoso**, é outra boa iniciativa.
- Para cada operação relevante que você fizer, eu espero um comentário que demonstre que você sabe o motivo de estar fazendo. Por exemplo, aplicar uma técnica de pré-processamento, ou aplicar um algoritmo de grid search para melhorar os hiperparâmetros de um modelo, são operações relevantes e merecem um comentário dizendo: “estou fazendo isto porque…”.
- Fique à vontade para manter no seu notebook todos os testes que fizer. De fato, é algo bastante positivo poder identificar que você executou vários testes antes de escolher qual caminho seguir. Mas eu espero que todas as células do notebook sejam executáveis com sucesso, isto é, sem gerar erros de código.

# CAPÍTULO 2: APRENDIZADO NÃO-SUPERVISIONADO

Você é o principal cientista de dados de uma grande cadeia de varejo no ramo de vestuário. Já houve situações em que a companhia tentou implementar planos ambiciosos de crescimento através da abertura das mais diversas lojas nas mais ousadas regiões possíveis. Mas todas as iniciativas foram frustradas de forma decepcionante. Em verdade, o que poderíamos esperar de uma sucessão de ideias tão mirabolantes quanto carentes de fundamento, não é mesmo?

No entanto, temos motivos para acreditar que tudo mudará! O novo presidente da companhia parece estar convencido de que guiar uma empresa que vale alguns milhões de dólares com base na aleatoriedade e em um punhado de devaneios provenientes das noites de domingo não parece uma estratégia lá muito inteligente. De fato, o sujeito está determinado a guiar os próximos planos a partir de dados concretos. Cá entre nós, ele ouviu falar sobre o termo “data driven” em algum congresso em New York e desde então repete isto como um mantra.

Por tudo isso, o ponto é que agora estamos diante de uma grande oportunidade: o próximo ciclo de expansão será guiado inteiramente por você, cientista de dados, e pelas suas análises. Se fizer um trabalho que chame atenção e que gere resultados, quem sabe você não pode se tornar um Chief Of Growth, ou algo parecido?

Considere, portanto, o dataset disponível aqui como o dataset oficial da companhia (arquivo shopping_behavior_updated.csv). Extraia comportamentos e padrões destes dados, entenda suas estruturas e faça recomendações. Por exemplo, onde temos mais gente interessada nos nossos produtos? Que tipo de produtos vendemos mais? Certas regiões podem combinar com lojas focadas em determinados produtos ou gêneros? Como devemos orientar o nosso marketing pensando nas faixas etárias dos nossos clientes?

Em suma, a ideia é: como podemos seccionar os clientes que compõem esta base de dados para tomar decisões com base nos padrões observados? O que os diferentes grupamentos possíveis nos dizem?

**EXTREMAMENTE IMPORTANTE**

- Nesta etapa você será avaliado quanto aos seus conhecimentos em (1) K-Means, (2) DBSCAN e (3) Agglomerative Clustering. Por isso, sua análise deve conter necessariamente todas estas três abordagens. Uma boa ideia é produzir uma análise individual para cada uma destas abordagens e, por fim, fazer comparações ou até mesmo oferecer recomendações baseadas nos pontos comuns que as três abordagens parecerem realçar. Uma restrição importante é que os insights que você vai fornecer devem ser obtidos exclusivamente por meio destas abordagens.
- Para cada abordagem acima, você deve necessariamente fornecer códigos em Python.
- Para cada abordagem acima, você deve fazer o possível para demonstrar ao máximo os seus conhecimentos. Testar explicitamente diferentes parâmetros, por exemplo, é uma boa ideia para isso. Realizar procedimentos de pré-processamento **quando necessário ou vantajoso**, é outra boa iniciativa.
- Para cada operação relevante que você fizer, eu espero um comentário que demonstre que você sabe o motivo de estar fazendo. Por exemplo, aplicar uma técnica de pré-processamento, ou aplicar um algoritmo de grid search para melhorar os hiperparâmetros de um modelo, são operações relevantes e merecem um comentário dizendo: “estou fazendo isto porque…”.
- Fique à vontade para manter no seu notebook todos os testes que fizer. De fato, é algo bastante positivo poder identificar que você executou vários testes antes de escolher qual caminho seguir. Mas eu espero que todas as células do notebook sejam executáveis com sucesso, isto é, sem gerar erros de código.