# Simulator Télos - Nível#6

# (Introdução ao Machine Learning)

Sejam muito bem-vindos ao **Simulator - Nível #6**! Neste nível, você terá a oportunidade de aplicar os conhecimentos adquiridos sobre análise de dados, desde a exploração inicial até a construção e avaliação de modelos preditivos. Você realizará: a análise exploratória de dados (EDA); o pré-processamento; o treino e avaliação dos modelos de machine learning. Que tal agora colocar todo esse conhecimento em prática em um projeto real e desafiador? 🙂

## Contextualização

Agora vamos atuar com um mini-projeto real (com o que há de melhor no mercado hehe), onde você terá a possibilidade de construir um projeto desde a sua idealização, organização e atuação.
Usem e abusem da criatividade 😉

## Descrição

Como data scientist em uma empresa vinícola, seu papel é crucial para otimizar diversos aspectos da produção e da qualidade dos vinhos, utilizando dados para guiar decisões estratégicas. A indústria do vinho, embora tradicional, está cada vez mais voltada para o uso de tecnologias avançadas, como machine learning, para melhorar seus processos. Ao trabalhar com o Wine Quality Dataset, você terá a oportunidade de aplicar técnicas analíticas para prever a qualidade do vinho com base em variáveis físico-químicas, como acidez, teor alcoólico, e pH. Esses insights podem ser usados para aprimorar a seleção de uvas, ajustar processos de fermentação, melhorar a consistência entre safras e até auxiliar na definição de preços e marketing. A capacidade de prever a qualidade final de um vinho antes que ele chegue ao consumidor não apenas agrega valor ao produto, mas também ajuda a empresa a reduzir custos, minimizar desperdícios e manter um padrão de excelência, diferenciando-se no mercado competitivo de vinhos.

O mínimo esperado para o projeto é:

1- Análise Exploratória de Dados (EDA): Visualização básica das variáveis e correlações principais.

2- Pré-processamento de Dados: Tratamento de valores ausentes e normalização dos dados (se necessário).

3- Treinamento de Modelo: Treinamento de pelo menos um algoritmo de machine learning adequado para o problema.

4- Avaliação de Desempenho: Cálculo de métricas básicas de desempenho ($R^2$, RMSE, MAE) para regressão.

5- Código Funcional: Entrega de um código que execute todo o pipeline de forma reprodutível.

**OBS**: A entrega deverá ser um notebook e deverá ser disponibilizado no Github do aluno.

Segue a lista de requisitos:

## 1. Análise Exploratória de Dados (EDA)

- **Descrição**: A EDA é o processo inicial de análise dos dados, onde se busca entender as distribuições das variáveis, suas relações e a presença de outliers ou valores ausentes. É fundamental para orientar as próximas etapas do pipeline de machine learning.
- **Critérios de Aceitação**:
  - Visualizações básicas (gráficos de dispersão, histogramas) são apresentadas para as principais variáveis.
  - Relações entre variáveis (como correlação) são exploradas.
- **#DicaTelos** 💡: Use bibliotecas como **matplotlib** ou **seaborn** para criar gráficos simples e mostrar a distribuição das variáveis. Focar nas variáveis mais correlacionadas com a qualidade do vinho pode direcionar melhor seu modelo.

## 2. Pré-processamento de Dados

- **Descrição**: O pré-processamento inclui a preparação dos dados para serem utilizados pelos algoritmos, como tratamento de valores ausentes, normalização, ou padronização das variáveis numéricas.
- **Critérios de Aceitação**:
  - Qualquer valor ausente é devidamente tratado (remoção ou imputação).
  - Os dados foram normalizados ou padronizados se necessário para o algoritmo escolhido.
- **#DicaTelos** 💡: Verifique a necessidade de normalizar os dados para algoritmos que sejam sensíveis à escala, como **SVM** ou **kNN**. A função **StandardScaler** do Scikit-learn pode ajudar.

### 3. Treinamento de Modelo

- **Descrição**: O treinamento do modelo envolve aplicar um algoritmo de machine learning aos dados processados para construir um modelo preditivo que possa ser avaliado e ajustado.
- **Critérios de Aceitação**:
    - Pelo menos um modelo de machine learning é treinado nos dados.
    - O modelo escolhido é apropriado para o problema (regressão ou classificação).
- **#DicaTelos** 💡 : Comece com modelos simples como **regressão linear** ou **árvores de decisão**. Você pode compará-los com modelos mais complexos depois, mas um modelo simples e bem treinado pode trazer bons resultados rapidamente.

### 4. Avaliação de Desempenho

- **Descrição**: A avaliação de desempenho é essencial para verificar o quão bem o modelo está funcionando. Métricas de regressão como $R^2$, RMSE, ou MAE são usadas para quantificar o erro do modelo em prever a qualidade do vinho.
- **Critérios de Aceitação**:
    - Pelo menos uma métrica de desempenho apropriada é calculada ($R^2$, RMSE, MAE).
    - A métrica é interpretada e discutida no relatório.
- **#DicaTelos** 💡 : Use o método **train_test_split** do Scikit-learn para separar seus dados em treino e teste, garantindo uma avaliação mais realista do modelo. Depois, calcule as métricas usando a biblioteca **Scikit-learn**.

### 6. Código Funcional

- **Descrição**: O código entregue deve ser claro e reprodutível, permitindo que outros possam rodá-lo facilmente e alcançar os mesmos resultados.
- **Critérios de Aceitação**:
    - O código executa todas as etapas (EDA, pré-processamento, treinamento e avaliação) corretamente sem erros.
    - O código está bem documentado, com comentários explicativos nas partes mais complexas.
- **#DicaTelos** 💡 : Use **Notebooks** para organizar seu código de forma clara, com cada etapa do processo em células separadas, facilitando a execução e interpretação.

2. **Desafio Bônus! (OPCIONAL)**: Organizar os resultados em uma apresentação no PowerPoint

- **Descrição**: A apresentação deve explicar as etapas realizadas, o raciocínio por trás das escolhas e uma interpretação dos resultados. Ele demonstra a capacidade de comunicação do processo técnico.
- **Critérios de Aceitação**:
    - O relatório ou apresentação contém uma explicação breve e clara das etapas realizadas e do desempenho do modelo.
    - Os resultados são discutidos e apresentados de forma compreensível.
- **#DicaTelos** 💡: Estruture o relatório em seções claras (introdução, metodologia, resultados, conclusão). Use gráficos e tabelas para ilustrar suas descobertas.

## Tempo de desenvolvimento

Espera-se que o aluno seja capaz de desenvolver o projeto em um prazo máximo de **6 horas**. Sugere-se o seguinte cronograma de desenvolvimento (que pode ser modificado livremente pelo aluno):

- **Dia 1**:
    - **2h** - Realizar a Análise Exploratória de Dados (EDA), identificando as principais variáveis e correlações.
- **Dia 2**:
    - **2h** - Pré-processamento dos dados, incluindo tratamento de valores ausentes e normalização/padronização.
- **Dia 3**:
    - **1h** - Treinar um modelo de machine learning .
- **Dia 4**:
    - **1h** - Avaliar com uma métrica de desempenho.

## Avaliação

A avaliação desta atividade levará em consideração os critérios a seguir:

1. Os alunos conseguiram realizar a Análise Exploratória de Dados (EDA) de forma clara e completa?

2. Os alunos foram capazes de pré-processar os dados corretamente?

3. Os alunos treinaram e selecionaram um modelo apropriado para o problema?

4. Os alunos conseguiram avaliar o modelo utilizando métricas de desempenho adequadas?

## Competências avaliadas

Para o desenvolvimento deste projeto utilizando o **Wine Quality Dataset**, são necessárias aos alunos as seguintes competências:

- Entendimento básico de análise exploratória de dados (EDA)
- Conhecimento sobre pré-processamento de dados (normalização, tratamento de valores ausentes)
- Familiaridade com algoritmos de machine learning supervisionado (regressão linear, random forest, etc.)
- Proficiência em métricas de avaliação de modelos de regressão ($R^2$, RMSE, MAE)
- Conhecimento básico em bibliotecas de machine learning (Scikit-learn)
- Habilidade em programação com Python, incluindo manipulação de dados com pandas e visualização com matplotlib/seaborn
- Capacidade de organizar o código de forma clara e reprodutível
- Aplicação de boas práticas na documentação e explicação dos resultados