## CIÊNCIA DE DADOS

E INTELIGÊNCIA ARTIFICIAL

**SEMANA DE IMERSÃO 02** 

# PROJETO DE MACHINE LEARNING

Pré-processamento e regressão















### PROJETO DE MACHINE LEARNING



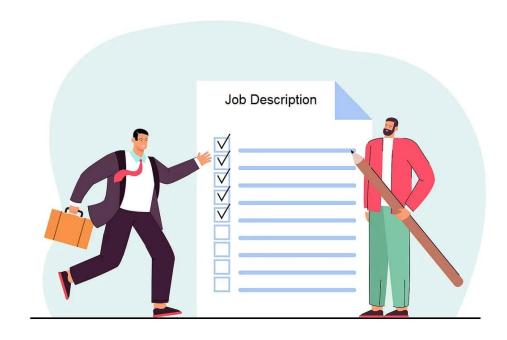
Nesta atividade, os participantes deverão utilizar bases de dados de seus próprios órgãos, setores ou de outra necessidade identificada nas atividades de contraturno para realizar **pré-processamento de dados e modelagem preditiva com regressão**.

O objetivo é **identificar padrões ou questões relevantes nos dados**, aplicando técnicas de tratamento e transformação para garantir sua qualidade. Os resultados deverão ser apresentados de forma clara e informativa, destacando insights obtidos por meio da análise e modelagem.

#### MONTE O SEU DREAM TEAM



Formação dos grupos: 3 a 4 participantes Converse com os integrantes: Defina tarefas claras para cada membro e o papel de cada um durante o desenvolvimento desse projeto





### DEFINIÇÃO E RESOLUÇÃO DO PROBLEMA



Conduzir uma análise exploratória e descritiva, com foco em pré-processamento de dados e modelagem preditiva via regressão. O objetivo é identificar padrões e insights relevantes, além de avaliar a capacidade do modelo em prever variáveis de interesse, apresentando os resultados por meio de visualizações claras e informativas.



### PRÉ-PROCESSAMENTO E PREPARAÇÃO DOS DADOS



Os grupos devem iniciar com a **limpeza e transformação dos dados**, incluindo:

- Tratamento de valores ausentes, duplicatas e outliers.
- Padronização e normalização das variáveis, quando necessário.
- Criação de novas variáveis (feature engineering) para melhorar a qualidade do modelo de regressão.



## ANÁLISE ESTATÍSTICA E SELEÇÃO DE VARIÁVEIS



- Realizem estatísticas descritivas, como médias, medianas, distribuições e correlações, para entender a estrutura dos dados.
- Identifiquem as variáveis mais relevantes para o modelo de regressão por meio de análise de correlação e outras técnicas de seleção de variáveis.



#### **MODELAGEM COM REGRESSÃO**



- Escolham e treinem um modelo de **regressão linear ou não linear**, dependendo das características dos dados.
- Avaliem o desempenho do modelo utilizando métricas como erro quadrático médio (MSE), R² e erro absoluto médio (MAE).



### CRITÉRIOS DE AVALIAÇÃO



- Organização dos Dados: Como os dados foram preparados e tratados.
- Análise Estatística: Qualidade das análises realizadas e sua capacidade de identificar padrões relevantes.
- **Apresentação**: Clareza, objetividade e coerência na apresentação dos resultados e conclusões.



#### **FORMATO E PRAZO PARA ENTREGA**



**TODOS** os participantes devem enviar a entrega no formato PDF e Notebook Jupyter (.ipynb) no portal do aluno.

- No **formato notebook**, certifique-se de anexar a base de dados utilizada.
- Se os dados forem sensíveis ou protegidos por lei, não inclua a base, mas garanta que os resultados estejam documentados e reproduzíveis no notebook.

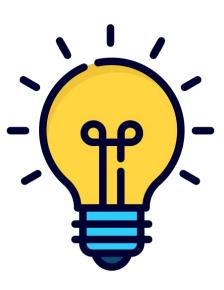
A entrega deve conter todas as etapas do pré-processamento, análise e modelagem de regressão, com visualizações e interpretações dos resultados de acordo com o template de entrega.

- A entrega será feita em duas etapas.
  - o 1ª entrega: 09/02/2025 até 23h59
  - o 2ª entrega: 16/02/2025 até 23h59

#### CONTEÚDO DAS ENTREGAS



- 1ª Entrega: 09/02/2025 até 23h59
  - Objetivos e Justificativas
  - Análises iniciais e pré-processamento
  - Modelagem, resultados e discussões preliminares
  - Apresentação dos achados, scripts e bases de dados (se aplicável)
- **2**<sup>a</sup> **Entrega:** 16/02/2025 até 23h59
  - o Todos os itens da 1ª entrega, revisados e aprimorados
  - Melhorias e ajustes no modelo, baseados na análise dos primeiros resultados.
  - Documento final, contendo:
    - Apresentação consolidada dos resultados, com visualizações claras.
    - Scripts comentados, garantindo reprodutibilidade.
    - Bases de dados (se aplicável e permitido).

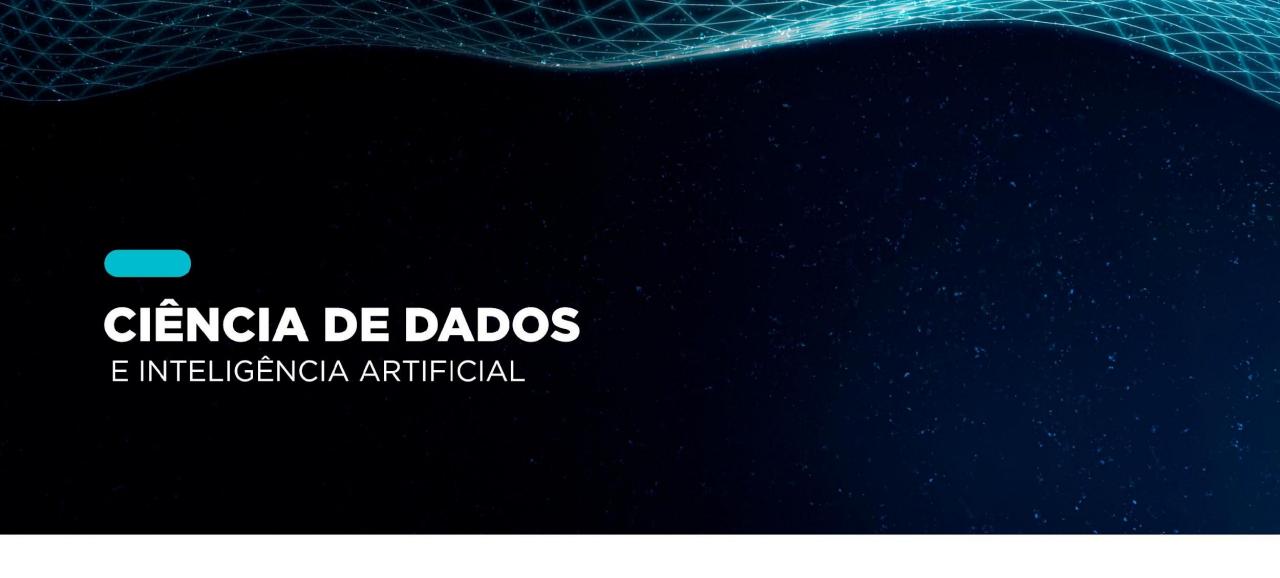


### **AVALIAÇÃO**



- A 1º entrega contribuirá para a nota da atividade de imersão, sendo parte da composição da nota de atividades. Nessa fase, serão avaliados a estruturação do problema, a qualidade do pré-processamento e a coerência das análises iniciais.
- A **2ª entrega** terá um peso maior e integrará a nota final da disciplina de Aprendizado de Máquinas. Além da revisão dos itens da primeira entrega, serão analisadas as melhorias implementadas, a robustez do modelo, a clareza na apresentação dos resultados e a organização dos scripts e documentação.

Se não houver melhorias entre as duas entregas, o participante deverá **reenviar a 1ª entrega**, garantindo que todos os requisitos mínimos da atividade sejam atendidos. No entanto, recomenda-se que os ajustes e aprimoramentos sejam realizados, pois a **2ª entrega terá um peso maior na composição da nota final**, e a qualidade das melhorias será um critério de avaliação.













MINISTÉRIO DA SAÚDE

