Universidade Federal de Santa Maria
Engenharia de Computação
UFSM000283 – Ciência de Dados II
Prof. Dr. Gabriel Machado Lunardi

Desafio DataDriver:
O futuro Elétrico dos Municípios Brasileiros

Este trabalho **possui caráter avaliativo** e contribuirá com **50% da nota final** da disciplina, conforme critérios acordados no plano de ensino no início do semestre. Portanto, **atente-se a todas as instruções deste documento**.

A transição para veículos elétricos (VEs) é uma realidade global e o Brasil está começando a trilhar esse caminho. Entender como essa tecnologia se difunde pelo território nacional, identificando os fatores que impulsionam ou dificultam sua adoção em nível municipal, é muito importante para o planejamento de infraestrutura, o desenvolvimento de políticas de incentivo eficazes e a identificação de oportunidades de mercado.

Para explorar essa questão de forma engajadora, este trabalho está estruturado no estilo de uma competição *Kaggle*. Se você nunca ouviu falar, o *Kaggle.com* é uma plataforma *online* que hospeda desafios de ciência de dados e *machine learning*. Nesses desafios (ou competições), os participantes recebem um conjunto de dados (um para treinar seus modelos e outro para teste, sem as respostas) e o objetivo é construir o modelo preditivo mais preciso. As submissões são avaliadas com base em quão bem o modelo acerta as previsões para o conjunto de teste (cujas respostas reais são mantidas em segredo pelo organizador), e um *ranking* (*leaderboard*) é gerado. Nosso desafio seguirá essa mesma dinâmica!

1. Objetivo

Desenvolver o melhor **modelo preditivo** para **estimar o número de veículos elétricos em cada município brasileiro**. Vocês aplicarão os conhecimentos e algoritmos de *Machine Learning* estudados na disciplina para construir, treinar, avaliar e otimizar seus modelos, competindo pela solução mais precisa em um *ranking* direto.

2. Organização

A execução deste trabalho se organizará da seguinte maneira:

- **Equipes:** 4 integrantes.
- **Duração e entrega:** 3 aulas dedicadas nas quais o professor atuará como orientador. Além delas, o grupo deve considerar se reunir fora da sala de aula, afinal, alguns modelos podem demorar para "rodar". Data de entrega: 26/06/2025 até às 15h 30min.
- **Entregáveis:** deverão ser entregues dois arquivos
    - **Um notebook Jupyter por equipe,** bem-organizado, contendo todo o processo de desenvolvimento do modelo: desde a importação e exploração dos dados até a avaliação final e conclusões. Não haverá apresentação formal.
    - **Um arquivo de submissão de previsões** (formato .csv) para o conjunto de teste da competição, conforme especificações detalhadas abaixo.

4. Conjuntos de dados

Na primeira aula do projeto, serão disponibilizados dois arquivos:

1. `dados_treino_alunos.csv:` Contendo as features (variáveis preditoras) e a variável alvo (*target*) para o treinamento e validação interna dos seus modelos.
2. `dados_teste_competicao_features.csv:` Contendo as mesmas features do conjunto de treino, mas **sem a variável alvo**. É para este conjunto que vocês deverão gerar as previsões que serão submetidas para o *ranking* da competição. O gabarito (respostas corretas) deste arquivo é privado e será usado pelo professor para avaliar as submissões.

5. Cronograma de atividades sugerido

Sugere-se a seguinte distribuição de atividades para cada aula prevista para a realização e orientação do trabalho. Lembrando que mais tempo pode ser necessário fora de sala de aula.

5.1 Aula 1: Imersão nos Dados

- **Carregamento e Compreensão Inicial:** Importar dados_treino_alunos.csv, verificar tipos de dados, dimensões.
- **Análise Exploratória de Dados (AED).**
- **Limpeza e Pré-processamento.**
- **Engenharia de Features (Inicial):** Criar novas features a partir das existentes que possam ter poder preditivo. Lembre-se que qualquer feature criada aqui precisará ser replicada de forma idêntica no conjunto de teste da competição.
- **Divisão dos Dados de Treino:** Separar dados_treino_alunos.csv em subconjuntos de treino e validação interna (para avaliar seus modelos antes da submissão final). É crucial que vocês usem boas práticas de validação (ex: validação cruzada) para evitar overfitting e ter uma estimativa realista do desempenho do modelo.

5.2 Aula 2: Modelagem e Otimização

- **Seleção e Treinamento de Modelos:**
    - Treinar **pelo menos 3 algoritmos diferentes** estudados em aula.
    - Para cada modelo, justificar brevemente por que ele pode ser adequado para este problema.
- **Métricas de Avaliação Interna:** Definir e utilizar métricas apropriadas para problemas de regressão.
- **Avaliação Inicial e Comparação (nos seus dados de validação interna).**
- **Otimização de Hiperparâmetros (Tuning):** Selecionar o(s) modelo(s) mais promissor(es) e tentar otimizar seus hiperparâmetros.

5.3 Aula 3: Seleção Final, Geração de Previsões e Documentação

- **Seleção do Melhor Modelo:** Escolher o modelo final com base no desempenho na sua validação interna e justificá-lo.
- **Treinamento Final do Modelo:** Re-treinar o modelo escolhido usando todos os dados_treino_alunos.csv (ou a melhor estratégia de treino que definiram).
- **Geração de Previsões para Competição:**
    - Carregar dados_teste_competicao_features.csv.
    - **Aplicar rigorosamente os mesmos passos de pré-processamento e engenharia de features que foram aplicados aos dados de treino.** Isso inclui a criação de quaisquer novas features. **A consistência aqui é fundamental!**
    - Usar o modelo final treinado para gerar as previsões para cada observação em dados_teste_competicao_features.csv.
    - Salvar essas previsões em um arquivo .csv conforme o formato especificado na seção “Competição”.
- **Análise do Modelo Escolhido no Notebook:** Discutir a importância das features (se o modelo permitir), limitações, etc.
- **Registro de Prompts de IA no Notebook:** Em uma seção dedicada, reportar os principais prompts utilizados com ferramentas de IA que auxiliaram no desenvolvimento.
- **Conclusões Finais no Notebook.**

6. Competição 🏆

A competição será baseada no desempenho do seu melhor modelo no conjunto dados_teste_competicao_features.csv, avaliado pelo professor com um gabarito privado. **O desempenho nesta competição contribuirá com até 2,0 pontos para a nota final do trabalho**.

6.1 Submissão de Previsões

- Cada equipe deverá submeter um único arquivo no formato .csv contendo as previsões para o dados_teste_competicao_features.csv.
- **Formato do arquivo de submissão:** O arquivo deve ter exatamente duas colunas:
    - A primeira coluna deve se chamar Cod_municipio no arquivo dados_teste_competicao_features.csv.
    - A segunda coluna deve se chamar Qtd. Veículos
- Certifique-se de que a ordem dos Cod_municipio no seu arquivo de submissão corresponda à ordem do arquivo dados_teste_competicao_features.csv ou que os IDs estejam corretos. Não deve haver valores ausentes (NaN) nas previsões.

6.2 Métrica de Competição Oficial

- O desempenho no ranking será medido exclusivamente pela métrica **RMSE (Raiz do Erro Quadrático Médio)**, calculada comparando as previsões submetidas com o gabarito privado. Quanto menor o RMSE, melhor o desempenho.

6.3 *Leaderboard* e Atribuição de Pontos da Competição

- Após o prazo de submissão, o professor calculará o RMSE para todas as equipes e divulgará um *leaderboard*.
- A pontuação será atribuída da seguinte forma:
    - **1º lugar Ouro:** 2,0 pontos
    - **2º lugar Prata:** 1,6 pontos
    - **3º lugar Bronze:** 1,2 pontos
    - **Demais equipes com submissão válida e funcional:** 0,8 ponto (como incentivo pela participação e esforço na submissão, desde que o modelo demonstre um esforço genuíno e não seja trivial. A critério do professor, resultados excessivamente ruins podem não receber esta pontuação mínima).

O bom desempenho na competição é valorizado, mas lembrem-se que a qualidade do processo, documentada no notebook, também é fundamental para a avaliação geral.

7. Critérios avaliativos

A nota final do trabalho (total de 10,0 pontos) será composta da seguinte forma:

- **A. Desempenho e Empenho em Aula (até 2,0 pontos)**: Nota individual atribuída pelo professor, considerando a participação ativa, proatividade, colaboração com a equipe e o progresso demonstrado durante as aulas dedicadas ao projeto.
- **B. Avaliação por Pares (até 1,0 ponto)**: Nota individual. Ao final do projeto, cada aluno avaliará anonimamente o empenho, a contribuição e a colaboração de seus colegas de equipe. Um formulário específico será disponibilizado.
- **C. Desempenho na Competição (até 2,0 pontos)**: Nota da equipe, atribuída com base na classificação no *leaderboard*, conforme descrito na seção “Competição”.
- **D. Notebook Jupyter (até 5,0 pontos):** Nota da equipe, avaliada conforme os critérios detalhados na tabela abaixo.

Critério		Nota
1. Organização, Clareza e Reprodutibilidade	Notebook bem estruturado com seções claras (uso de *Markdown* para títulos e explicações). Código limpo, comentado e que pode ser executado do início ao fim sem erros, gerando os resultados apresentados e o arquivo de submissão.	1,0
2. Análise Exploratória de Dados (AED)	Profundidade e qualidade da exploração dos dados de treino. Uso eficaz de visualizações para extrair padrões relevantes. Identificação e discussão de padrões, *outliers* e correlações.	1,5
3. Pré-processamento e Engenharia de *Features*	Justificativa e aplicação correta e consistente das técnicas de limpeza e pré-processamento tanto nos dados de treino/validação quanto nos dados de teste da competição. Criatividade, relevância e implementação correta de novas *features*.	1,5
4. Modelagem, Treinamento e Validação de Algoritmos	Aplicação correta e justificada de pelo menos 3 algoritmos de regressão diferentes. Qualidade do treinamento, da validação interna (ex: uso de validação cruzada) e, se aplicável, da otimização de hiperparâmetros.	1,0

8. Comentário adicionais

O uso de ferramentas de IA como assistentes é permitido e incentivado, **desde que devidamente reportado.** O objetivo é que vocês aprendam a usar essas ferramentas de forma ética e eficaz para potencializar seu trabalho, e não para substituir o pensamento crítico e o desenvolvimento das soluções. Sejam criativos na abordagem do problema!

Atenção: A consistência no pré-processamento e na engenharia de *features* entre os dados de treino e os `dados_teste_competicao_features.csv` é fundamental para um bom desempenho no *ranking* da competição. Qualquer inconsistência aqui provavelmente levará a previsões ruins.

Suspeitas de **plágio** estão sujeitas a nota **zero.**