Guia do Projeto Prático
CSI 710 - Projeto e Análise de Experimentos

Prof. Carlos Henrique Gomes Ferreira

Objetivo do Projeto

O projeto prático tem como objetivo consolidar os conceitos abordados ao longo da disciplina por meio da aplicação real de técnicas experimentação e análise de dados. De forma obrigatória, os projetos devem contemplar:

- Análise exploratória dos dados, utilizando técnicas estatísticas e de probabilidade, como testes estatísticos, estudo de distribuições e visualizações descritivas;
- Aplicação de **uma** das técnicas principais discutidas na disciplina: **Regressão Linear Múltipla** ou **Projeto Fatorial 2^k com r replicações**.

Esses elementos devem estar presentes tanto na fase exploratória quanto na análise final dos resultados experimentais. Espera-se que os alunos demonstrem domínio das etapas do ciclo experimental completo, desde a definição do problema até a apresentação dos achados.

Formação dos Grupos

Os grupos deverão ser compostos **por até 4 discentes**. Recomenda-se que grupos maiores apresentem um projeto mais completo, considerando a divisão natural de responsabilidades e o ganho em capacidade de execução.

A composição dos grupos deve ser informada via Moodle até **sexta-feira, 23 de maio de 2025**. Após essa data, não serão permitidas alterações sem justificativa formal.

Entregas

O projeto será desenvolvido ao longo do semestre em quatro etapas principais, com entregas intermediárias e apresentações. As datas de apresentação serão definidas por sorteio e divulgadas com antecedência mínima de uma semana. **A participação na apresentação é obrigatória para todos os membros do grupo**.

Entrega e Apresentação 1: Proposta do Projeto

Data de entrega: 05/06
Formato: Apresentação em PDF enviada via Moodle
Tempo de apresentação: até 10 minutos
Pontuação: 20%

Conteúdo esperado:

- **Contextualização e definição clara do problema:** Apresente a motivação para o projeto, explicando o contexto prático ou teórico que justifica a investigação. Indique a relevância do problema a ser estudado.
- **Objetivos do estudo e hipótese:** Defina com clareza o que se pretende alcançar com o projeto. Esboce uma hipótese a ser testada, mesmo que ainda em caráter preliminar, especialmente se o projeto pretende usar o modelo fatorial ou regressão. A hipótese deve se conectar diretamente ao problema e orientar a escolha das variáveis e da técnica principal.

- **Técnica principal a ser utilizada:** Indique se será usado **Projeto Fatorial** $2^k.r$ ou **Regressão Linear Múltipla**, e justifique a escolha em função da hipótese e dos objetivos.
- **Esboço das variáveis experimentais:** Indique as variáveis candidatas a:
  - **Fatores** e seus **níveis** (no caso de projeto fatorial), ou
  - **Variáveis explicativas** (no caso de regressão),
  - **Variáveis resposta** associadas ao problema de interesse.

  Este item serve como uma primeira delimitação da estrutura experimental, ainda que sujeita a ajustes.
- **Fonte dos dados:** Especifique se os dados já estão disponíveis, se serão coletados de bases públicas (ex: Kaggle, GitHub, UCI, etc.), ou se haverá coleta direta (via API, formulários, logs, etc.). Indique também, se possível, o tamanho e a estrutura inicial do conjunto de dados.
- **Cenário e ambiente de experimentação:** Descreva o ambiente computacional onde o projeto será desenvolvido: ferramentas, bibliotecas, linguagens, plataformas (Google Colab, notebook pessoal,, etc.).
- **Referências iniciais:** Inclua links, artigos, repositórios ou outros materiais utilizados como base para o projeto. Isso pode incluir tanto as fontes de dados quanto abordagens similares que inspiraram o trabalho.

Entrega e Apresentação 2: Análise Exploratória dos Dados

Data de entrega: 08/07
Formato: Apresentação em PDF enviada via Moodle
Tempo de apresentação: até 10 minutos
Pontuação: 15%

Conteúdo esperado:

- Releitura dos objetivos do projeto, com eventuais atualizações e maior clareza quanto à motivação e ao que se pretende investigar;
- Apresentação clara da **hipótese do projeto**, explicitando o que se deseja testar ou entender com o uso da técnica principal (Projeto Fatorial $2^k.r$ ou Regressão Linear Múltipla). A hipótese deve estar diretamente conectada ao problema proposto e justificar a aplicação da técnica escolhida;
- Os seguintes elementos devem estar bem definidos e claramente conectados entre si: **Objetivo do estudo, Hipótese, Fatores, Níveis, Variáveis Respostas e Replicações a serem realizadas**. Não basta mencionar esses elementos, é essencial mostrar como eles se articulam no projeto experimental;
- Discussão sobre eventuais mudanças na estratégia de análise ou conjunto de dados inicialmente propostos;
- Comprovação de que os dados estão **prontos ou parcialmente prontos** para a fase de experimentação. Isso inclui conjuntos em fase de coleta, rotulagem, limpeza ou estruturação, desde que o andamento esteja claramente apresentado;
- Apresentação completa da **análise exploratória dos dados já disponíveis**, utilizando estatísticas descritivas, análise de distribuições e diferentes tipos de visualizações gráficas abordadas em sala de aula. A análise deve demonstrar entendimento do comportamento dos dados e como eles se relacionam com a hipótese experimental.

Entrega e Apresentação 3: Resultados Preliminares da Técnica Experimental

Data de entrega: 29/07
Formato: Apresentação em PDF enviada via Moodle
Tempo de apresentação: até 15 minutos
Pontuação: 15%

Conteúdo esperado:

- Apresentação dos primeiros resultados da técnica de experimentação escolhida (Regressão Linear ou Projeto Fatorial $2^k.r$);
- Evidências de adequação da técnica ao problema;
- Resultados preliminares com análise inicial.

Entrega Final: Relatório e Apresentação Final

Data de entrega: 19/08
Formato: Relatório em PDF enviado via Moodle + apresentação final em PDF
Tempo de apresentação: até 15 minutos
Pontuação: 50%

Relatório final: Deve ser redigido no **template da SBC** (Link aqui) contendo as seguintes seções:

- **Cabeçalho:** Título, resumo, autores, orientadores (se houver);
- **Introdução:** Contextualização, motivação, problema/lacuna (opcional) e objetivos;
- **Trabalhos relacionados (opcional):** Discussão sobre abordagens similares e suas limitações;
- **Metodologia:** Descrição da base de dados, variáveis, técnica aplicada;
- **Resultados:**
  - **Caracterização dos dados:** Análise exploratória final, descrevendo e caracterizando bem os dados;
  - **Resultados experimentais:** Saídas da técnica principal (regressão ou fatorial);
  - **Discussão:** Interpretação dos achados, limitações e implicações.
- **Conclusão e Trabalhos Futuros.**

A estrutura pode ser adaptada conforme a natureza do projeto, desde que contenha as seções centrais exigidas.

Nota: As datas das apresentações de cada etapa serão definidas por **sorteio**, com cronograma divulgado no Moodle e via e-mail. Todos os membros devem estar presentes na apresentação.

Sugestões de Projetos

Nesta seção, apresento algumas ideias de projetos organizadas por nível de complexidade e tipo de recurso necessário. Os projetos podem assumir diferentes naturezas:

- **Projetos educacionais:** Utilizam bases de dados públicas e tarefas bem definidas como classificação, análise textual ou exploração de dados.
- **Projetos integrados:** Podem estar associados a TCCs, ICs ou temas de interesse individual, permitindo aprofundamento técnico ou interdisciplinar.

As bases de dados podem ser obtidas a partir de repositórios consolidados como GitHub, Kaggle, Zenodo, UCI Machine Learning Repository ou mesmo por meio de coleta direta via APIs. Os projetos variam em termos de complexidade computacional. Por exemplo, alguns executam tranquilamente em notebooks pessoais, outros podem demandar uso de GPU, maior memória RAM ou execução em plataformas como Google Colab, servidores da UFOP ou máquinas de colegas.

Projetos com Baixa Complexidade Computacional (Notebook pessoal)

- **Análise de sentimentos ou toxicidade no Reddit:** Dados disponíveis neste repositório. Permite tarefas de NLP como extração de tópicos, detecção de discurso de ódio, análise de tendências temporais, entre outras.
- **Comparação de algoritmos de classificação:** Avaliar o impacto de técnicas de oversampling/undersampling em datasets desbalanceados para tarefas de classificação. Você pode varias algoritmos, estratégias diversas de tratamentos dos dados, balanceamento, etc. Para isso, você pode usar bases específicas da literatura. Por exemplo, a base `Olist` (link) disponível no Kaggle permite análises envolvendo regressão, classificação e segmentação de clientes com múltiplas tabelas.
- **Estudos com dados do IMDb ou Spotify:** Explorar correlações entre gênero, avaliação, duração de filmes ou músicas. Análise de tendências por país, ano ou gênero artístico.
- **Detecção de toxicidade textual com modelos prontos:** Aplicar o Google Perspective API ou a biblioteca Detoxify para avaliar toxicidade em conjuntos curtos de mensagens ou comentários coletados de fóruns online.
- **Exploração de bases do Telegram ou WhatsApp:** Utilização de conjuntos de dados anonimizados já existentes (disponíveis em papers e repositórios acadêmicos) para explorar volume de mensagens, horários de maior tráfego, palavras-chave e clusters de interação.

Projetos com Média Complexidade Computacional (Colab ou desktop com boa RAM)

- **Análise de redes de colaboração no GitHub:** Coleta via API ou uso de repositórios prontos. Investigação de métricas como centralidade, cliques, componentes e frequência de commits em projetos como Linux, TensorFlow ou Jupyter.
- **Avaliação de bibliotecas de grafos:** Comparar performance de bibliotecas como NetworkX, iGraph e SNAP em tarefas como detecção de comunidades, cálculo de centralidade e clustering.
- **Detecção de anomalias em redes sociais ou logs:** Análise de outliers em interações de usuários, falhas de sistema ou variações suspeitas em métricas operacionais (por exemplo, em logs de rede).
- **Análise de comentários no Twitch:** Coleta via API Helix. Aplicar classificadores ou análise léxica em comentários com base no tipo de conteúdo (games, esportes, política). Possível cruzamento com o uso de emojis, frequência de mensagens e tempo de sessão.
- **Estudo de tópicos em mensagens de grupos abertos do Telegram:** Aplicações análise de tópicos para inferência de tópicos em canais abertos com temáticas variadas (tecnologia, política, esportes, etc).

Projetos com Alta Complexidade Computacional (GPU ou execução externa)

- **Análise de vídeos curtos no TikTok:** Explorar dados coletados de perfis públicos ou APIs terceiras para análise de engajamento, viralidade de hashtags, tipos de filtros usados e tempo de retenção por categoria.
- **NLP com redes neurais profundas:** Aplicações como análise de sentimentos, classificação de tópicos, ou identificação de entidades nomeadas (NER) usando modelos como BERT, RoBERTa ou DistilBERT. Bases podem ser extraídas de tweets, Reddit ou artigos jornalísticos.
- **Modelagem de tendências em redes sociais:** Usar dados temporais para prever picos de engajamento, variações de humor da comunidade ou difusão informação. Envolve uso de séries temporais e estruturas recursivas (RNN/LSTM).
- **Experimentos com arquiteturas de deep learning:** Comparar diferentes configurações de modelos (número de camadas, função de ativação, otimizadores) para tarefas como classificação de imagens, texto ou sinais biométricos.

Antes de escolher seu projeto, reflita:

- Você já tem os dados ou sabe exatamente como coletá-los? Se você for fazer um coletor, tem certeza de que ele funciona?
- A tarefa é clara? É uma análise, predição ou classificação?
- O tamanho da base de dados é viável para processar com os recursos que você tem?
- Seu projeto roda no seu notebook? Precisa do Google Colab, de uma máquina da UFOP ou de ajuda externa?
- Você consegue concluir esse projeto dentro do semestre?

Lembre-se: um bom projeto não precisa ser o mais complexo. Precisa ser viável, bem planejado e bem executado.