Ciência de Dados:
Análise de Dados Aplicada

Prof. Dr. Eduardo Pena

Projeto Final

Objetivo Geral

Nesta atividade, você será um(a) cientista de dados e deverá conduzir uma análise completa de dados, desde a definição do problema até a comunicação dos resultados. O objetivo é demonstrar competências em todas as etapas do processo analítico: integração de dados, análise exploratória, modelagem preditiva e comunicação (escrita de relatório e apresentação).

Tecnologias Obrigatórias:

- **SQL:** PostgreSQL ou DuckDB para consultas
- **Python/R:** Pandas, scikit-learn, matplotlib/seaborn/ggplot2
- **Versionamento:** Git com commits bem documentados
- **Outros:** Demais bibliotecas Python ou outras ferramentas que julgar necessárias

Grupos: Individual ou duplas.

1 Escolha do Dataset e Definição do Problema

Nesta etapa, você deve selecionar um conjunto de dados relevante e formular um problema analítico bem estruturado que será investigado ao longo do projeto. **Os dados devem obrigatoriamente ser provenientes de:**

Tabela 1: Fontes de dados permitidas

Categoria	Fontes
Brasil	Dados.gov.br
Internacional	NYC Open Data, Data.gov, EU Open Data
Rep.s Acadêmicos	UCI ML Repository, World Bank Open Data
Kaggle	Apenas datasets sem notebooks públicos existentes
Datasets de papers	Veja as instruções abaixo

Tabela 2: Venues recomendadas (Podem usar os papers de outros anos)

Conferência	Track/Seção
KDD	Applied Data Science Track
CIKM	Accepted Papers
ECML-PKDD	Applied Data Science Track
AAAI	Applications Track

1.1 Datasets de artigos científicos

Inspiração em Pesquisas Científicas (*Opcional – Desafiador e Altamente Recomendado*)

Os alunos podem basear seus projetos em estudos de **Applied Data Science** publicados em venues de alta qualidade. Essa abordagem permite trabalhar com problemas reais e bem fundamentados, além de comparar seus resultados com pesquisas já estabelecidas. É uma excelente forma de se familiarizar com a literatura da área. Preferência para artigos de 2020 em diante.

Caso não encontre opções adequadas nas fontes listadas e identifique um dataset em outra venue, certifique-se de que possui alto fator de impacto ou classificação Qualis A1 e consulte o professor previamente para aprovação.

Observações importantes: Utilize os mesmos dados originais do artigo, mas desenvolva perguntas de pesquisa diferentes, adote uma abordagem analítica distinta e explore aspectos não investigados no trabalho original. O objetivo não é replicar o estudo, mas sim complementá-lo com metodologias e perspectivas próprias. Espera-se que pelo menos 30% dos resultados sejam inéditos, provenientes das novas análises propostas.

1.2 Requisitos mínimos do dataset

O dataset selecionado deve atender aos seguintes critérios técnicos para garantir a viabilidade e complexidade adequada do projeto:

Tabela 3: Requisitos do dataset final

Critério	Especificação
Registros	Mínimo de 10.000
Preditores	Mínimo de 15 (após integração e feature engineering)
Tipos de variáveis	Combinação de variáveis numéricas e categóricas
Integração	Possibilidade de integração com fontes adicionais
Qualidade dos dados	Presença de erros de dados que demandem data cleaning
Potencial analítico	Adequado para análise preditiva (classificação ou regressão)

1.3 Formulação do Problema

Você deve desenvolver os seguintes temas sobre seu dataset e análise pretendida:

Perguntas de Pesquisa (*1-2 perguntas específicas*)

As perguntas devem ser específicas e mensuráveis, orientadas por dados e analiticamente interessantes.

Exemplos:

- *Como características socioeconômicas de um município brasileiro influenciam seu IDH?*
- *É possível predizer o nível de desenvolvimento municipal baseado em indicadores disponíveis?*

- *Quais fatores têm maior peso na classificação de municípios como de alta qualidade de vida?*

Suas perguntas devem ser investigáveis através dos dados disponíveis, ter relevância prática e permitir análises que ajudem na tomada de decisão.

Contexto e justificativa (*0,5-1 página*)
Após definir suas perguntas, desenvolva uma fundamentação sólida que responda:

- Por que este problema é relevante?
- Qual o impacto prático da análise proposta?
- Quem se beneficiaria dos resultados obtidos?
- Se inspirado em artigo: como sua abordagem difere do estudo original?

1.4 Hipóteses Testáveis

Desenvolva 2-3 hipóteses que possam ser testadas com o dataset escolhido, **priorizando aquelas relacionadas à modelagem preditiva** (classificação ou regressão).
Exemplos de hipóteses:

- *H1: Municípios com maior investimento em educação per capita apresentam IDH superior*
- *H2: A renda média é o preditor mais forte do IDH municipal*
- *H3: Modelos ensemble superam modelos lineares na predição de desenvolvimento municipal*

1.5 Entregável da Etapa 1

Tabela 4: Documento PDF (1-2 páginas)

Item
Identificação do grupo e link do **repositório Git** para acompanhamento
Descrição do dataset escolhido (fonte, estrutura, tamanho estimado)
Referência e diferenciação do trabalho original (se o tema for baseado em artigo)
Contexto e justificativa do problema
1-2 perguntas de pesquisa bem formuladas
2-3 hipóteses testáveis

Tabela 5: Critérios de avaliação da etapa 1

Critério
Relevância e originalidade do problema
Clareza e especificidade das perguntas
Viabilidade das hipóteses para modelagem preditiva
Qualidade da escrita e estruturação

2 Integração e Limpeza de Dados

Integrar dados de múltiplas fontes e preparar um dataset limpo para análise, resolvendo problemas de qualidade comuns em projetos reais.

2.1 Enriquecimento com Fontes Externas (*Opcional, mas encorajado*)

Esta etapa visa identificar e integrar dados externos que possam melhorar significativamente o poder preditivo do modelo. Utilize técnicas de **schema matching** e **data discovery** para encontrar datasets complementares que agreguem valor analítico ao problema proposto.

Esse enriquecimento deve ser aplicado quando o dataset principal possui limitações evidentes ou quando variáveis externas podem explicar melhor o fenômeno estudado.

Por exemplo, ao classificar casos de dengue baseado apenas em contagens históricas, a incorporação de dados meteorológicos (chuva, temperatura) pode revelar padrões sazonais cruciais para a predição.

Estratégias de integração recomendadas:

- **Matching por identificadores únicos**: CPF, CNPJ, códigos municipais IBGE
- **Integração temporal**: séries históricas, dados mensais/anuais compatíveis
- **Matching geográfico**: coordenadas, CEP, códigos de região
- **Atributos categóricos**: setores econômicos, faixas etárias, classificações padronizadas

Fontes típicas para enriquecimento:

- Dados demográficos e socioeconômicos (IBGE, censos)
- Indicadores econômicos e financeiros (Banco Central, IPEADATA)
- Informações climáticas e ambientais (INMET, INPE)
- APIs públicas de contexto (geolocalização, eventos, mercado)

Quando utilizar múltiplas fontes, siga estas etapas:

1. Padronizar nomenclaturas e tipos de dados entre fontes
2. Resolver conflitos quando fontes divergem para o mesmo registro
3. Detectar e tratar registros duplicados
4. Criar identificadores únicos consistentes

Importante: Proceda com o enriquecimento se identificar lacunas claras no dataset original que possam ser preenchidas com dados externos relevantes e de qualidade.

2.2 Limpeza de Dados (Obrigatório)

Implemente soluções para os principais problemas de qualidade:

- **Missing values**: escolher estratégias de imputação adequadas
- **Outliers**: detectar valores extremos e decidir sobre tratamento
- **Inconsistências**: verificar dependências entre variáveis (dependências funcionais, denial constraints – se existirem) – verificar e corrigir possíveis erros.
- **Padronização**: uniformizar formatos de datas, textos e códigos

Documente todas as transformações aplicadas e mantenha registro das decisões tomadas durante o processo de limpeza.

2.3 Consultas SQL e Visualizações para Diagnóstico de Qualidade

Desenvolva consultas SQL e visualizações que demonstrem os problemas encontrados nos dados brutos e validem a eficácia das correções aplicadas.

Crie gráficos que ilustrem visualmente a transformação: box plots para outliers, heatmaps para dados faltantes, e distribuições antes/depois da limpeza. As consultas devem ser acompanhadas de interpretações que expliquem os problemas identificados e como foram resolvidos (mesmo que parcialmente).

2.4 Entregável da etapa 2

2.5 Notebook Documentado Contendo

Item
Scripts/Processo de integração
Pipeline de limpeza reprodutível
Pelo menos 4 itens entre consultas SQL e visualizações com interpretação
Dataset final limpo

Tabela 6: Critérios de Avaliação da Etapa 2

Critério
Qualidade da integração e limpeza
Consultas SQL e visualizações de validação
Documentação e reprodutibilidade

3 Análise Exploratória e Consultas SQL

3.1 Objetivo

Compreender os dados através de análises estatísticas e consultas SQL que respondam às perguntas de pesquisa, garantindo que os dados estejam em formato adequado para análise.

3.2 Preparação dos Dados em Formato Tidy

3.2.1 Organização dos Dados

Transformar o dataset limpo para formato **tidy data** seguindo os princípios fundamentais: cada variável forma uma coluna, cada observação forma uma linha e cada tipo de unidade observacional forma uma tabela.

3.2.2 Reestruturação Necessária

- **Transformações estruturais**: Converter dados wide em long format quando necessário
- **Normalização**: Separar variáveis compostas em colunas distintas
- **Padronização de tipos**: Garantir tipos de dados apropriados
- Exportar dados preparados em formato **Parquet**

3.3 Consultas SQL Analíticas (5+ obrigatórias)

Desenvolver consultas que revelem insights interessantes a partir dos dados em formato tidy, explorando:

- **Agregações complexas**: Análises por grupos e períodos (ou equivalentes)
- **Análises temporais**: Funções de janela para tendências e rankings
- **Consultas hierárquicas**: CTEs para análises em múltiplos níveis

Exemplos de tipos de consultas:

1. Tendências temporais e sazonalidade
2. Comparações entre grupos e categorias
3. Análises de concentração e distribuição
4. Detecção de padrões e anomalias
5. Correlações e dependências entre variáveis

3.4 Análise Exploratória e Teste de Hipóteses

Compreender a estrutura dos dados através de análises estatísticas progressivas e verificação preliminar das hipóteses de pesquisa, incluindo:

- **Análise univariada**: Distribuições, medidas de tendência central e dispersão, identificação de padrões e valores atípicos
- **Análise bivariada**: Correlações entre variáveis, visualizações de relacionamentos, comparações entre grupos
- **Análise multivariada** (*quando relevante*): Matriz de correlações, análises de agrupamento, técnicas de redução dimensional
- **Teste de hipóteses**: Formulação e verificação preliminar das hipóteses através de testes estatísticos apropriados e validação visual

Todas as análises devem ser acompanhadas de visualizações e interpretações que conectem os achados às perguntas de pesquisa.

3.5 Entregável da Etapa 3

Notebook e demais recursos:

Item
Pipeline de transformação para formato tidy
Produção e exportação do Dataset final em formato Parquet
5+ consultas SQL documentadas com interpretações
Análises univariadas e bivariadas
Visualizações interpretadas
Testes de hipóteses
Insights que orientem a etapa de modelagem

Tabela 7: Critérios de Avaliação da Etapa 3

Critério
Preparação tidy e geração do Parquet
Consultas SQL analíticas e insights
Qualidade das análises e visualizações
Conexão com perguntas de pesquisa

4 Modelagem com Machine Learning

4.1 Objetivo

Desenvolver modelos preditivos robustos que respondam às perguntas de pesquisa, demonstrando evolução e melhoria ao longo do processo.

4.2 Implementação de Modelos

4.2.1 Diversidade de Algoritmos (2+ obrigatórios)

Implementar modelos com diferentes níveis de complexidade para comparação:

- **Modelo simples**: Regressão Linear ou Logística (baseline interpretável)
- **Modelo mais complexos**: Random Forest, Neural Networks, SVM, etc (procure mais opções)

4.2.2 Feature Engineering

Aplicar transformações necessárias nos dados:

- **Transformações básicas**: Encodings categóricos, normalização, tratamento de outliers
- **Criação de features** (se necessário): Variáveis derivadas quando relevantes para o problema
- **Seleção de features**: Identificar variáveis mais importantes para performance e interpretabilidade

4.3 Processo de Melhoria Iterativa

4.3.1 Desenvolvimento Incremental

- **Modelo baseline**: Implementação inicial simples para estabelecer referência
- **Iterações sequenciais**: Melhorias incrementais documentadas
- **Registro de experimentos**: O que funcionou, o que não funcionou e por quê

4.3.2 Otimização de Hiperparâmetros

- Tuning sistemático dos principais parâmetros
- Validação cruzada para generalização
- Comparação quantitativa antes/depois das otimizações

4.4 Avaliação e Interpretabilidade

4.4.1 Métricas de Performance

Utilizar métricas apropriadas ao tipo de problema (classificação, regressão) e contexto de negócio, incluindo análise de erro e validação em dados não vistos.

4.4.2 Análise de Explicabilidade

- **Importância das features**: Identificar variáveis mais influentes
- **Trade-off complexidade vs interpretabilidade**: Análise comparativa entre modelos
- **Interpretação contextual**: Conectar resultados ao problema
- **Limitações identificadas**: Reconhecer quando os modelos podem falhar

4.5 Entregável da Etapa 4

Notebook e demais recursos:

Item
Histórico de modelagem (baseline → final)
3+ algoritmos implementados e comparados
Processo de tuning com resultados antes/depois
Análise de interpretabilidade vs performance
Recomendação do melhor modelo com justificativa
Validação robusta dos modelos finais

Tabela 8: Critérios de Avaliação da Etapa 4

Critério
Diversidade e adequação dos algoritmos
Documentação do processo
Qualidade do tuning e validação
Interpretação no contexto de negócio

5 Relatório e Comunicação

Vocês devem comunicar de forma clara e profissional os resultados obtidos, demonstrando capacidade de traduzir análises técnicas em conclusões que podem gerar ações e basear a tomada de decisão.

5.1 Relatório Técnico (6-8 páginas sem contar referências)

Documento final que expande e consolida o relatório inicial, contendo:

- Resumo com principais descobertas
- Definição do problema e perguntas de pesquisa (expandido da Etapa 1)
- Metodologia completa e limitações

- Resultados das análises
- Discussão dos resultados
- Recomendações práticas
- Trabalhos futuros

Importante: O relatório deve ser escrito integralmente pelos alunos. O uso de IA é permitido apenas para correção do texto já produzido pelos estudantes.

5.2 Materiais de Apresentação

5.2.1 Apresentação Principal (23-30 minutos)

Apresentação gravada em vídeo com os integrantes visíveis nas câmeras. Deve ser interativa e dinâmica.

Estrutura da apresentação:

- **Contexto e problema** (3-4 min): Motivação e perguntas de pesquisa
- **Demonstração dos dados** (4-5 min): Dataset, integração e limpeza
- **Metodologia** (5-6 min): Pipeline analítico e modelos implementados
- **Resultados principais** (6-8 min): Descobertas e performance dos modelos
- **Recomendações** (3-4 min): Ações práticas baseadas nos resultados
- **Lições aprendidas** (2-3 min): O que funcionou, o que não funcionou, próximos passos

A apresentação deve incluir demonstração prática do pipeline quando possível.

5.2.2 Pitch de 2 minutos

Apresentação oral rápida e persuasiva do projeto, similar ao formato usado em competições de startups e apresentações executivas. Será apresentado em sala com slide/imagem de uma página como fundo que sintetize visualmente o projeto.

- Apresentação concisa: problema $\rightarrow$ solução $\rightarrow$ resultado $\rightarrow$ impacto
- Objetivo: despertar interesse dos colegas para a apresentação completa
- Criar slide caprichado e visualmente atrativo que resuma o projeto completo em uma página

5.3 Entregável da Etapa 5

Item
Relatório técnico em PDF
Slides da apresentação
Vídeo da apresentação
Demonstração do pipeline funcionando
Repositório Git organizado e documentado

Tabela 9: Critérios de Avaliação da Etapa 5

Critério
Clareza e qualidade da escrita
Estrutura e fluidez da apresentação
Demonstração técnica efetiva
Qualidade das recomendações práticas
Reflexão crítica (lições aprendidas)

Resumo das Entregas

Etapa	Entregável	Foco Principal	Critérios de Avaliação	Peso
1. Definição do Problema	**Documento PDF (1-2 páginas):** • Identificação do grupo + link Git • Dataset escolhido (fonte, estrutura, tamanho) • Contexto e justificativa • 1-2 perguntas de pesquisa específicas • 2-3 hipóteses testáveis • Diferenciação se baseado em paper	Formulação clara do problema analítico com dataset adequado	Relevância e originalidade do problema, clareza das perguntas, viabilidade das hipóteses, qualidade da escrita	15%
2. Integração e Limpeza	**Notebook documentado:** • Scripts de integração (fontes externas opcionais) • Pipeline de limpeza • Consultas SQL exploratórias e visualizações • Dataset limpo	Preparação robusta dos dados com validação de qualidade	Qualidade da integração e limpeza, consultas SQL e visualizações de validação, documentação e reprodutibilidade	20%
3. EDA + SQL	**Notebook + Dataset Parquet:** • Pipeline para formato tidy • Dataset final em Parquet • Consultas SQL analíticas • Análises univariadas e bivariadas • Visualizações • Testes de hipóteses	Compreensão dos dados através de análises estatísticas e consultas estruturadas	Preparação tidy e geração do Parquet, consultas SQL analíticas, qualidade das análises e visualizações, conexão com perguntas de pesquisa	20%
4. Modelagem ML	**Modelos implementados:** • Histórico baseline → modelo final • 2+ algoritmos implementados • Feature engineering quando necessário • Processo de tuning • Análise interpretabilidade vs performance • Recomendação do melhor modelo • Validação	Desenvolvimento iterativo de modelos preditivos com melhoria documentada	Adequação dos algoritmos, documentação do processo, qualidade do tuning e validação, e interpretação	30%
5. Comunicação	**Produtos finais:** • Relatório técnico (6-8 páginas) • Slides da apresentação • Vídeo apresentação (23-30 min) • Pitch 2 minutos com slide • Demonstração do pipeline • Repositório Git organizado	Comunicação dos resultados com capacidade de traduzir análises em ações práticas	Qualidade da escrita, estrutura da apresentação, demonstração técnica, qualidade das recomendações, reflexão crítica	15%

Tabela 10: Estrutura e Entregas do Projeto Final de Ciência de Dados