Universidade Federal do Rio Grande do Norte
Unidade Acadêmica Especializada em Ciências Agrárias
Escola Agrícola de Jundiaí
Curso de Análise e Desenvolvimento de Sistemas
TAD0204 - Fundamentos e Técnicas em Ciência de Dados - Turma 01
Projeto da 3ª Unidade

Objetivo

Consolidação das técnicas e métodos estudados ao longo da disciplina por meio da análise de um conjunto de dados e pelo compartilhamento de análises em diferentes contextos.

Especificação

O trabalho será desenvolvido em duplas com entregas semanais. Consiste na aplicação de diferentes técnicas de ciência de dados em um problema real. O desenvolvimento consiste na análise de um problema conforme as técnicas de ciências de dados.

Etapa 1: Escolha da base de dados e análise inicial (07/11/2025 à 14/11/2025)

A dupla deve buscar uma base de dados na plataforma Kaggle em uma área que possua conhecimento. O link da base de dados deve ser submetido à avaliação no fórum da disciplina até o dia 12 de novembro de 2025. O professor avaliará e enviará uma versão modificada para análise.

A base de dados deve satisfazer os seguintes critérios:

- Ao menos 2 variáveis quantitativas discretas;
- Ao menos 2 variáveis quantitativas contínuas;
- Ao menos 2 variáveis qualitativas nominais;
- Ao menos 2 variáveis qualitativas ordinais;
- Ao menos 20 variáveis totais;
- Ao menos 300 instâncias;
- As bases de dados de cada dupla devem ser diferentes.

Exceções dos requisitos podem ser consideradas, no entanto, o professor fará alterações na base de dados.

Para essa etapa, é esperado que a dupla elabore um relatório com uma descrição da base de dados (informações, autor, objetivo, suposições, etc.); uma descrição das variáveis, isto é, tipos, formatos, limites esperados e suposições; e ao menos cinco hipóteses ou informações não triviais que se acredita que podem ser obtidas dos dados. As hipóteses devem abranger ao menos as seguintes análises:

- Análise explicativa;
- Análise preditiva;
- Análise exploratória.

Considere o seguinte exemplo como guia: em uma base de dados de vendas de corretores de imóveis, desejamos verificar os fatores mais importantes que influenciam na venda do imóvel; ou então, podemos verificar se existe algum perfil entre os corretores, isto é, algum grupo que se destaca dos demais; ou então efetuar uma previsão da quantidade de vendas de um corretor no próximo ano.

Etapa 2: Qualidade dos Dados (14/11/2025 à 21/11/2025)

Nesta etapa a dupla deve efetuar:

1. Descrição das estatísticas básicas de cada variável: medidas de tendência central (média, moda, mediana), medidas de dispersão (variância, amplitude), estatísticas de ordem (mínimo e máximo) e quantidade de valores distintos, quando aplicável.
2. Análise das distribuições indicando se cada variável numérica segue uma distribuição normal ou não. São esperados histogramas, box-plots, gráficos de densidade, tabelas de valores, gráficos q-q, quando aplicável.
3. Identificação e tratamento de outliers. É esperada uma análise por meio do Isolation Forest.
4. Identificação de variáveis dependentes e independentes com justificação adequada conforme as hipóteses levantadas.
5. Identificação dos valores faltosos e análise se eles estão associados às classes do problema ou não.
6. Identificação e aplicação de etapas de limpeza de dados adicionais, conforme a natureza dos dados.

Etapa 3: Transformações e Análise Exploratória (21/11/2025 à 28/11/2025)

Nesta etapa é esperado:

1. Tratamento dos valores faltosos por meio de indutor, ou seja, deve ser modelado um indutor (regressão linear ou regressão logística) diferente para cada variável com valores faltosos, tomando-o como variável alvo e usando as demais variáveis como preditoras. As classes do problema podem ser consideradas ou não, dependendo da análise dos valores faltosos. Para todos os modelos deve ser apresentada alguma métrica de performance de generalização.
2. Transformações adequadas aos dados, conforme a necessidade das análises.
3. Análise de Agrupamentos:
    a. Análise das correlações entre as variáveis quantitativas.
    b. Análise de redução de dimensionalidade usando PCA ou filtros, conforme a natureza dos dados.
    c. Análise dos grupos identificados por métodos de visualização em baixa dimensionalidade com o auxílio do PCA.
    d. Análise descritiva identificando a melhor quantidade de grupos para o algoritmo k-Means por meio do método do cotovelo.
    e. Interpretação dos conjuntos identificados considerando o contexto do problema.

Etapa 4: Análise Preditiva e Explicativa (28/11/2025 à 05/12/2025)

Nesta etapa é esperado:

1. Escolha de uma variável alvo para a predição.
2. Análise da relação entre as variáveis preditivas e a variável alvo por meio de um modelo estatístico de regressão.
3. Análise da generalização dos indutores por holdout ou cross validation incluindo uma matriz de confusão (problemas de classificação) ou gráfico de dispersão da variável alvo pelos valores preditos (problemas de regressão).
4. Interpretação dos modelos considerando os contexto do problema.

Etapa 5: Apresentação dos Resultados (05/12/2025)

Cada dupla efetuará uma apresentação de até 15 minutos para a turma, destacando os principais pontos, problemas e soluções.

Considerações:

- Devem ser priorizados gráficos ou visualizações apropriados para representar as informações e contextualizar as explicações.
- O trabalho deve ser desenvolvido e entregue no formato de notebook Jupyter e submetido via tarefa no SIGAA.
- Toda análise deve gerar um conhecimento, mesmo que seja inconclusivo. Assim, devem ser discutidos os resultados de cada análise e isso deve estar presente no notebook.
- A base de dados **estruturada** precisa ser enviada ao professor para validação. Além disso, o professor fará modificações na base de dados, visando adequá-la aos objetivos didáticos da disciplina. As bases de dados devem ser enviadas até o dia 12 de novembro para a validação por meio do fórum “Escolhas das Bases de Dados” cadastrado no SIGAA, **apresentando o tema, o nome da dupla de trabalho e o link para a base de dados**. O professor responderá à submissão com a base validada ou solicitará outra base, se necessário. A resposta pode levar até três dias, porém, iniciem o desenvolvimento com a base escolhida, adaptando para a base validada pelo professor.
- Não serão aceitos trabalhos sem a validação da base de dados pelo professor.
- Se tiverem dúvidas em alguma parte, entrem em contato com o professor.
- O link a seguir direciona o aluno para a página do GitHub classroom. O uso do repositório é obrigatório: https://classroom.github.com/a/BeKg4DmJ