

Especificação dos Trabalhos Práticos de Mineração de Dados

Fase 1: Proposta

Na primeira fase, os alunos devem descrever o conjunto de dados e a tarefa de mineração de dados a ser executada, detalhando o contexto dos dados. É o momento de explicar o problema que se pretende resolver e justificar a relevância dos dados para esse problema. De forma intuitiva, o documento a ser submetido deve conter respostas para:

- qual a tarefa a ser executada?
- quais dados vão ser utilizados?
- por que a tarefa é relevante e significativa?
- Como a tarefa vai ser executada?

Essa fase serve como base para as fases posteriores, pois ajuda a estabelecer uma compreensão clara dos dados e do contexto de negócio. Aqui, os alunos começam a explorar e se familiarizar com os dados.

Fase 2: Utilização de LLM

Na segunda fase, os alunos farão uso de uma ou mais LLMs (Large Language Models) para auxiliar na execução da tarefa proposta. A interação com a LLM permitirá que os alunos obtenham sugestões e resolvam dúvidas, como problemas de limpeza de dados, seleção de variáveis ou aplicação de algoritmos. Toda a interação com a(s) LLM(s) deve ser registrada, assim como criticada sob a perspectiva da sua eficácia, precisão e outras dimensões pertinentes.

Fase 3: Trabalho Completo

Na terceira fase, os alunos deverão aplicar o que aprenderam nas fases anteriores para realizar o trabalho completo proposto, em particular corrigindo eventuais erros ou incapacidades da(s) LLM(s) usadas na Fase 2. Nesta fase final, os alunos integrarão todos os elementos das fases anteriores, realizando a modelagem completa e submetendo o relatório para avaliação.

Para essa entrega, não é necessário incluir os links para as conversas realizadas na Fase 2 — apenas uma análise comparativa entre as soluções sugeridas pelas LLMs e a solução final desenvolvida pelo grupo deverá ser apresentada.

Restrições do Dataset

Com o objetivo de manter um padrão no momento da avaliação dos trabalhos, é necessário que o dataset possua:

- **Mínimo de Linhas:** 1000
- **Mínimo de Colunas:** 4
- **Entrega:** Um Jupyter Notebook contendo as seções especificadas abaixo. Um template será fornecido aos alunos para facilitar a organização e estruturação do trabalho.

Estrutura do Notebook (seguindo formato CRISP)

1. Business Understanding

Nesta seção, os alunos deverão fornecer uma explicação detalhada do problema de negócio que está sendo resolvido com os dados fornecidos. O foco será entender o propósito do dataset, sua origem e o valor que ele pode agregar na solução de problemas.

- **Objetivo do Dataset:** Descreva o propósito do dataset e como ele pode ser usado para resolver problemas de negócio.
- **Origem dos Dados:** Explique a fonte do dataset (seu autor, instituição de origem ou como ele foi coletado).
- **Características do Dataset:** Descreva as colunas e as informações contidas em cada uma delas. Explique o que as linhas e colunas representam.
- **Relação com o Problema de Negócio:** Justifique a escolha do dataset e sua relevância para a aplicação que se pretende explorar.

Link para conversa com LLM: Nesta seção, inclua o link da conversa com a LLM que ajudou a desenvolver a compreensão do negócio, se aplicável.

2. Data Understanding & Data Preparation

Nesta segunda fase, o foco será explorar o dataset em mais detalhes e realizar a preparação dos dados para os próximos passos.

Data Understanding:

- **Exploração Inicial:** Análise inicial dos dados (distribuição de valores, estatísticas descritivas, valores nulos, etc.).
- **Análise Visual:** Geração de gráficos e visualizações que ajudem a compreender melhor o dataset.
- **Insights sobre os Dados:** Identifique padrões, outliers e outras características relevantes no dataset.

Data Preparation:

- **Limpeza de Dados:** Tratamento de dados nulos, duplicados ou incoerentes.
- **Transformação de Dados:** Normalização, discretização ou outras transformações necessárias para o correto funcionamento dos algoritmos.
- **Seleção de Features:** Escolha das colunas mais relevantes para a modelagem, se necessário.
- **Divisão dos Dados:** Separação entre dados de treino e teste (se aplicável).

Link para conversa com LLM: Inclua aqui o link para a conversa com a LLM que foi utilizada para solucionar eventuais problemas de preparação ou entendimento dos dados. (Exemplo será anexado ao final do documento)

3. Modeling

A fase de modelagem consiste na aplicação de algoritmos de mineração de dados sobre o dataset preparado, com o objetivo de extrair padrões e insights. Cada aluno deverá aplicar uma seleção de algoritmos e comparar seus resultados.

Algoritmos para mineração de padrões frequentes: (Tópico que será abordado no primeiro trabalho)

- **Apriori**
- **Eclat**
- **FP-Growth**
- **Association Rules**

Nesta fase, os alunos deverão:

- **Aplicar os Algoritmos:** Executar algum dos algoritmos no dataset.
- **Comparar Desempenho:** Comparar os resultados em termos de tempo de execução, qualidade dos padrões encontrados, etc. (Caso utilize mais de um algoritmo)
- **Justificar Escolhas:** Explicar por que determinado algoritmo apresentou melhor performance no dataset.

Link para conversa com LLM: Anote aqui as interações que ajudaram a desenvolver a modelagem ou a resolver desafios específicos relacionados aos algoritmos.

4. Evaluation

Na fase de avaliação, os alunos irão discutir os resultados obtidos e as lições aprendidas. É importante que seja feita uma reflexão crítica sobre o desempenho dos algoritmos e a qualidade dos padrões encontrados.

- **Análise de Resultados:** Discutir os padrões identificados e sua relevância para o problema de negócio.
- **Avaliação dos Algoritmos:** Criticar o desempenho do(s) algoritmo(s) em termos de eficiência e precisão.
- **Considerações Finais:** Reflexão sobre os desafios enfrentados, o que funcionou bem e o que poderia ser melhorado.

Link para conversa com LLM: Incluir link para interações finais com a LLM, onde ajustes e análises adicionais foram realizados.

Entrega Final

Na primeira fase, os alunos deverão entregar um PDF organizado com os tópicos propostos a serem cobertos no trabalho. A estrutura e a organização desse documento fazem parte da avaliação.

Na segunda fase, os alunos devem utilizar uma LLM (como o ChatGPT, por exemplo) para obter uma primeira versão da solução do trabalho. É necessário identificar e anotar os acertos e erros apresentados pela LLM. A entrega dessa fase deve ser feita em um Jupyter Notebook organizado, contendo o que foi proposto, com comentários e observações sobre o desempenho da LLM.

Na terceira fase, os alunos deverão aplicar o que aprenderam nas fases anteriores para realizar o trabalho completo, corrigindo eventuais erros ou limitações da(s) LLM(s) usadas na Fase 2. Nesta fase final, os alunos integrarão todos os elementos desenvolvidos anteriormente, realizando a modelagem completa e submetendo o trabalho final para avaliação. A entrega da Fase 3 também deve ser feita em Jupyter Notebook organizado.

Para essa entrega final, não é necessário incluir os links para as conversas realizadas na Fase 2 — apenas uma análise comparativa entre as soluções sugeridas pelas LLMs e a versão final desenvolvida pelo grupo deverá ser apresentada.

Atenção: Qualquer outro formato de entrega que não seja o especificado acima não será aceito. A organização das entregas faz parte da avaliação.

Exemplo de chat: <https://chatgpt.com/share/6800faa3-d170-8005-8f63-d9ec37ac5775>