📘 Organização da Semana 2 – EDA e Estruturação dos Dados

Após compreendermos o comportamento dos dados na primeira etapa, a Semana 2 será dedicada a transformar esses aprendizados em um pipeline funcional que permita iniciar as primeiras previsões. Nesta fase, a equipe de Data Science trabalhará para:

- unificar o pré-processamento, garantindo que todas as transformações dos dados sejam consistentes e reutilizáveis;
- criar as primeiras features derivadas, baseadas nos padrões identificados na EDA;
- treinar os modelos baseline, que servirão como referência inicial de performance;
- desenvolver uma API em Python, capaz de receber um JSON de entrada e retornar uma previsão simples via endpoint /predict; e
- validar a integração com o time de Dev.

Assim como na semana anterior, o trabalho será distribuído entre dimensões independentes, permitindo que todos os membros contribuam simultaneamente. Ao final desta etapa, teremos o primeiro pipeline operacional da equipe DS.

🛠️ DS1 – Pré-Processamento Unificado (Limpeza + Encoding + Normalização) | Ana

Tarefas

- Reaplicar as correções levantadas no Data Quality da Semana 1
- Criar funções reutilizáveis para:
  - tratamento de nulos
  - padronização de tipos
  - normalização de variáveis numéricas (MinMax ou StandardScaler)
  - encoding de categóricas (OneHot ou Ordinal)

- Garantir que todas as transformações funcionem para toda a base
- Gerar um notebook e um script com as funções

Principal questão a ser respondida:

❓ “Quais transformações são indispensáveis para que os modelos recebam dados consistentes e previsíveis?”

🧩 DS2 – Feature Engineering Inicial (Criação de Novas Variáveis) | Amélia

Tarefas

- Criar features derivadas de tempo:
  - horário (manhã/tarde/noite)

    - atraso previsto — transformação log ou caps
- Criar features de companhia/aeroporto:
    - frequência de voos por companhia
    - rotas mais utilizadas
- Testar impacto de cada feature com análises rápidas (correlação, separabilidade)
- Documentar as features criadas

Principal questão a ser respondida:

❓ **"Quais novas variáveis parecem adicionar mais sinal para predizer atraso?"**

DS1 e DS2, nesta etapa o foco não é apenas preparar os dados, mas garantir que todo o pré-processamento possa ser reproduzido quando o modelo estiver em uso. Como o modelo será integrado a uma API, ele não vai receber um dataframe pronto, e sim dados novos. Por isso, todas as transformações precisam ser feitas por meio de um pipeline reutilizável, e não de forma manual no notebook.

Transformações feitas diretamente no dataframe, como mapeamentos ou cálculos "na mão", funcionam para análise, mas não são reproduzidas na predição e acabam gerando inconsistência entre treino e uso real do modelo.

Um exemplo é a criação da variável de média de atraso por companhia aérea. Esse tipo de feature não pode ser criada apenas com um groupby no dataframe. Essa lógica precisa estar dentro do pipeline, por meio de uma função ou transformador que aprenda essa média no treinamento e consiga aplicá-la novamente quando um novo dado chegar via API.

Na prática, a sigla da companhia não é usada diretamente pelo modelo. Ela serve apenas como entrada para o pipeline gerar uma informação numérica. O mesmo vale para tratamento de nulos, encoding de variáveis categóricas e normalização das numéricas: tudo precisa estar concentrado no pipeline.

Para esta semana, não é necessário nada avançado. Basta aplicar o uso básico de pipelines de transformação no scikit-learn para garantir um pré-processamento consistente e reutilizável. A regra geral é simples: se a transformação não estiver no pipeline, ela não deve ser usada pelo modelo.

⚖️ **DS3 – Balanceamento & Preparação do Dataset para o Modelo | Enoque + suporte Helena, se necessário**

Tarefas

- Avaliar o desbalanceamento da variável alvo
- Testar técnicas de undersampling ou oversampling
- Criar função para a separação dos dados em treino e teste
- Documentar qual estratégia funcionou melhor *para o baseline*

Principal questão a ser respondida:

❓ **"Qual estratégia de balanceamento preserva melhor os padrões reais dos dados sem gerar distorções e porquê?"**

DS3, nesta etapa o seu foco será preparar os dados para a etapa de modelagem, a partir do dataset já transformado pelo pipeline de pré-processamento e feature engineering. A ideia aqui não é criar novas regras de transformação, mas trabalhar sobre os dados que chegam prontos para o modelo.

O primeiro ponto é avaliar o comportamento da variável alvo, identificando se existe desbalanceamento entre as classes e qual o impacto disso nos modelos baseline. A partir disso, você pode testar estratégias simples de balanceamento, como undersampling ou oversampling, sempre com cuidado para não distorcer os padrões reais observados nos dados.

Além disso, será necessário definir uma estratégia clara de separação entre treino e teste, garantindo que essa divisão seja consistente e reproduzível. Essa etapa é importante para que os resultados dos modelos possam ser comparados de forma justa e interpretável.

Mesmo antes da entrega final do pipeline pelas DS1 e DS2, você já pode estudar métricas, pensar nos modelos baseline mais adequados e estruturar o código de treino. Quando o dataset final estiver disponível, a ideia é apenas aplicar essas decisões e validar os resultados.

O objetivo desta etapa não é maximizar performance, mas garantir que o modelo baseline funcione corretamente, com dados bem preparados e resultados confiáveis, servindo como base para a integração com a API.

🍼 **DS5 – Validação Técnica do Baseline + Testes do JSON | Helena**

Tarefas

- Validar se o modelo baseline funciona com o JSON definido
- Criar função enviando o predict do json
- Testar com exemplos reais e exemplos inválidos

Principal questão a ser respondida:

❓ **“O modelo baseline e o JSON de entrada são suficientemente robustos para evitar erros na API?”**

Assim como na primeira semana, organizei as atividades em cinco frentes diferentes, cada uma com tarefas definidas. Essas frentes de pesquisa não foram atribuídas individualmente, porque a ideia é que cada pessoa escolha aquela com a qual mais se identifica, seja por afinidade, curiosidade ou por achar que faz mais sentido começar por ali. Mas, para que tudo funcione bem e ninguém acabe fazendo a mesma coisa, é importante que cada integrante avise a equipe e a liderança qual dimensão pretende assumir antes de começar.

🗓️ Cronograma da Semana 2 – Datas Importantes

📌 Segunda-feira — 22/12
Reunião de planejamento semanal

- Alinhamento das responsabilidades de cada integrante
- Revisão do pipeline definido
- Dúvidas técnicas e checklist dos arquivos necessários
- Ajustes no cronograma, se necessário

📌 Quinta-feira — 25/12
Demonstração das Entregas

- Cada membro apresenta a parte do pipeline que desenvolveu
- Validação do pré-processamento
- Demonstração do baseline funcionando

📌 Sexta-feira — 26/12
Consolidação e documentação

- Consolidação do pipeline inicial
- Padronização dos scripts e da estrutura de pastas
- Ajustes finais da API e testes adicionais

🔴 **Observação**
Planejei apenas as reuniões obrigatórias da semana, mas poderemos marcar encontros adicionais conforme necessidade da equipe, especialmente para revisar passos técnicos mais complexos, como testes da API ou validação.