**Avaliação A1 (5,0)**

**Use este documento para guiar as análises para a avalição A1. Os trabalhos são em equipe, então dividam para conquistar, procurem delegar funções para os membros das equipes, responsabilizando cada membro por uma parte menor do trabalho**, **que será cobrada posteriormente na avaliação. A princípio, a nota da avaliação A1 iria ser a mesma para o grupo todo, mas devido a diversas reclamações dos integrantes dos grupos, a análise será individualizada, recompensando aqueles que fizeram sua parte no trabalho.**

**Para fazer esta análise, utilize como base o código fornecido para a avaliação A2 e também as explicações fornecidas nas últimas aulas de como utilizar as ferramentas CHAT GPT e Gemini para fazer análises de dados. O uso destas ferramentas é permitido e recomendado. Utilizando a base NPS fornecida, siga os seguintes passos:**

1. Utilizando a base inteira, calcule o target principal, formando uma nova coluna na base, onde a variável nota será analisada em cada linha:
   1. Notas 9 ou 10 : a coluna target irá conter a classe “promotor”. No total a coluna deve conter 18.251 promotores, sendo essa a classe majoritária.
   2. Notas 8 ou 7: a coluna target irá conter a classe “neutro”. No total a coluna deve conter 4.738 neutros.
   3. Notas < 7: a coluna target irá conter a classe “detrator”. No total a coluna deve conter 2.185 detratores, sendo essa a classe minoritária.
   4. Executar o comando df.groupby(“target”).count() e verificar se as contagens do target batem com a descrição dos itens a, b e c.
2. Localizar a variável “mercado” e filtrar para reter apenas instâncias que são do brasil, para isso você pode utilizar um comando como:

df = df.loc[df[”mercado”] ==”brasil”].

1. Assim como no item 2, executar um comando de filtragem para trabalhar com o grupo que a sua equipe ficou encarregada. O nome da variável a ser filtrada é “Grupo de Produto”, exemplo : df = df.loc[df[”Grupo de Produto”] == “Grupo x”], onde x é o número de seu grupo.
2. Fazer a volumetria de target, calculando para o seu grupo quantos promotores, neutros e detratores ficaram na base, calcular também o percentual de cada classe.
3. Criar uma coluna chamada região, que irá corresponder as 5 regiões do país, baseado na informação da coluna “estado”. Ex: se o estado for Paraná, Santa Catarina ou Rio grande do Sul, a coluna região deve conter a string “sul”.
4. Criar uma coluna chamada safra, baseada na variável “data\_resposta”, capturando apenas o ano da resposta.
5. Calcular a volumetria do target Safrada (pelo ano) para a base total (sempre filtrada para o seu grupo), fazer a mesma volumetria para cada região do país e para cada um dos quatro períodos de pesquisa baseado na coluna “Período de Pesquisa”. A volumetria deve ser calculada em valores absolutos e percentuais. Além do cálculo safrado, a volumetria deve somar todas as safras no final. Fazer verificações se o total é coerente com a contagem obtida diretamente da base.

# Tabela exemplo de Volumetria base Total por Safra

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Safra** | **Total** | **Promotores** | **Neutros** | **Detratores** | **%Promotores** | **% Neutros** | **% Detratores** |
| 2020 | 100 | 69 | 6 | 25 | 69.0 | 6.0 | 25.0 |
| 2021 | 150 | 91 | 41 | 18 | 60.6 | 27.3 | 12.0 |
| 2022 | 200 | 125 | 23 | 52 | 62.5 | 11.5 | 26.0 |
| 2023 | 180 | 122 | 29 | 29 | 67.7 | 16.1 | 16.1 |
| Total | 630 | 407 | 99 | 124 | 64.6 | 15.7 | 19.6 |

Após o cálculo das volumetrias, inferir se existe diferença de balanceamento entre as classes, fazer isso para as safras e para o total, indicando no relatório a classe majoritária e a classe minoritária. Também indicar no relatório se alguma safra apresenta uma volumetria muito diferente comparada com as demais safras. Repetir essa análise fazendo novas tabelas de volumetria para cada filtragem do seu grupo, sendo que no total deve ter: (i) uma tabela de volumetria para Base total, (ii) 5 tabelas de volumetrias para as regiões e (iv) 4 tabelas de volumetrias para os períodos de pesquisas. No total, o relatório deverá conter 10 tabelas de volumetria igual a este exemplo com suas respectivas análises.

1. Filtrar as perguntas pertencentes ao seu grupo. Descartar quaisquer colunas que não sejam necessárias para a continuidade da análise.
2. Em um primeiro momento fazer para o seu grupo a **correlação de spearman**, entre a **variável nota e as demais variáveis de perguntas (contendo a nota de 0 a 10)**. Ordenar a lista de correlações da maior correlação para a menor, grifando em vermelho as correlações fortes, grifando em azul as correlações médias e grifando em verde as correlações fracas. Fazer isso para o seu grupo inteiro, por região e por período de pesquisa. **Apresentar também a lista de correlações Safrada, uma lista por safra, replicando o trabalho para cada região e período de pesquisa.**
3. **Faça 2 modelos de classificação binária por análise (um para neutro e outro para detrator)**. Para isso, use apenas as variáveis numéricas (perguntas) como variáveis de entrada (X) e treine o modelo com um novo target, reduzindo o target de 3 classes para 2 classes, transformando o problema multi-classe e um problema de classificação binária (exemplo 1: modelo de detratores, positivar caso detrator e negativar caso neutro ou promotor) (y) (exemplo 2: modelo de neutros, positivar caso neutro e negativar caso detrator ou promotor) (y). Importante!! Ao criar o target binário, não utilize no espaço de características do modelo (X) o target de 3 classes como entrada, nem a variável nota, pois ambas são consideradas vazamento neste contexto (pois derivam o target). Sendo assim, para cada modelo, o X (espaço de características) deve conter todas as colunas de perguntas, menos a variável nota e o y (target) deve conter apenas o target binário adaptado.
4. No total devem ser feitos 20 modelos: um modelo detrator e um modelo neutro para a base inteira filtrada pelo seu grupo, repetindo a análise por: região e período de pesquisa. Não precisa fazer modelos safrados!
5. Colocar no relatório as top 10 variáveis de cada modelo, para isso use um modelo como RandomForest ou XGbooost. Tirar as conclusões sobre top variáveis, podendo comparar esta análise com a lista de correlações.
6. Tentar avaliar as 5 top-variáveis de cada modelo com uma das técnica de gráfico, podendo ser PDP, ALE ou SHAP. A explicação de como fazer estes gráficos será fornecida nas próximas aulas, antes da avaliação.

**O relatório final deve todas as análises enumeradas neste documento e obrigatoriamente deve responder as 4 perguntas formuladas pelo Jhonny, que estão nos slides (arquivo PFD) disponibilizado**. **Para cada pergunta, elabore uma resposta utilizando como base a correlação de spearman e as top variáveis de cada modelo, considerando em cada resposta a base toda, as regiões do Brasil e os períodos de pesquisa. Comente também sobre a análise safrada em cada pergunta.**

**Metodologia avaliação:**

1. **Análise quantitativa e qualitativa do relatório, ou seja, verificação se o grupo entregou todos os itens solicitados (1,0).**
2. **Responsabilidade individual e contribuição para o grupo, sendo inquerido sobre a análise que ficou delegada a seu encargo (4,0).**