Relatório sobre Business Understanding (CRISP-DM - Fase 1)

Dataset: Gender Gap in Spanish WP

Grupo:
Nilo Bemfica Mineiro Campos Drumond *(nbmcd)*
Pedro Didier Maranhão *(pdm)*
Pedro Tenório Lemos *(ptl)*

A primeira fase do CRISP-DM é o ***"Business Understanding"***, que visa garantir que a equipe de ciência de dados e inteligência artificial tenha uma compreensão clara e precisa do problema de negócios em questão. A base de dados a qual nosso grupo ficou responsável se concentrou na análise das diferenças de gênero no Wikipedia em espanhol, especificamente na proporção de editores do gênero feminino e masculino e nas práticas de edição.

A Wikipedia, sendo uma das maiores fontes de informação disponíveis gratuitamente, reflete a diversidade da sociedade global, nesse caso dos falantes de espanhol. A representação desigual de gêneros entre os editores pode levar a um viés no conteúdo e na cobertura de tópicos. Entender essa disparidade é crucial para garantir que a Wikipedia seja um espaço **equitativo e representativo**.

Algumas questões críticas do negócio que pontuamos e podem guiar futuras análises:

- Qual é a proporção atual de editores de gênero feminino para masculino no Wikipedia em espanhol?
- Existem diferenças significativas nas práticas de edição entre os gêneros?
- Como podemos usar os dados para entender melhor e, possivelmente, abordar esse desequilíbrio?

1. Determinando Objetivos de Negócio:

Os principais objetivos que visamos para o projeto incluem:

- Estimar a porcentagem de editores do sexo feminino na Wikipedia em espanhol.
- Medir o engajamento e as práticas de edição em relação aos seus homólogos masculinos.

2. Avaliando a Situação:

O conjunto de dados fornece informações sobre o gênero dos editores, suas práticas de edição e outras métricas relacionadas principalmente à frequência e quantidade de postagem. Algumas suposições podem incluir a precisão da identificação de gênero e a representatividade dos dados em relação à população total de editores para termos um estudo efetivo.

3. Determinar Metas da Equipe:

O principal objetivo técnico seria classificar e analisar os editores com base em seu gênero e avaliar as práticas de edição associadas. Isso pode envolver:

- Análise exploratória para entender diferenças nas práticas de edição entre gêneros.
- Classificação de gênero a partir dos padrões de utilização do site.

4. Produzindo um Plano de Projeto:

O plano pode ser dividido em várias etapas:

1. Limpeza e preparação de dados.
2. Análise exploratória.

   As duas primeiras etapas funcionam juntas no **CRISP-DM.**

3. Modelagem (classificador de gênero conforme comportamento para indicar viés).
4. Avaliação do modelo e validação.
5. Interpretação e comunicação dos resultados.

A diferença dos gêneros no Wikipedia em espanhol é uma área de foco crucial, dada a importância da representação equiparada na maior enciclopédia do mundo. Com o conjunto de dados fornecido, estamos posicionados para entender e quantificar esse desequilíbrio, fornecendo insights valiosos para a comunidade Wikipedia falante dessa língua. Nossa equipe julga que a principal entrega desse estudo seria, de fato, uma análise descritiva dos dados, mas o modelo pode ter um grande valor de negócio, dada uma alta performance, para incentivar usuários do gênero feminino (sub representado) a comentar mais, até atingir um nível de postagem mais parecido com um usuário médio.