Desmatamento no Brasil - Análise de Dados

1 - Descrição do problema

A análise de dados, de forma simples e direta, visa extrair informações e percepções valiosas de dados brutos. O trabalho do analista de dados consiste em executar uma sequência de tarefas para alcançar essas informações e embasar a tomada de decisões. Muitas vezes os dados brutos não apresentam em um “primeiro olhar” essas informações e percepções que o ciclo de análise de dados apresenta ao final do processo, provando a importância deste procedimento.

Dessa forma, é possível aplicar este conceito para muitos segmentos, como saúde, educação, economia, transporte, meio ambiente, e etc., de forma a obter novas e relevantes informações para entender a evolução de cada segmento.

Especificamente sobre o meio ambiente, para fazer o monitoramento de queimadas e desmatamento no passado, era necessário deslocar funcionários até o local para fazer este acompanhamento de forma presencial.

Com a evolução da tecnologia, não é obrigatório manter um controle da situação de forma presencial. Atualmente existem sensores IoT que fornecem dados praticamente em tempo real sobre as condições do ambiente monitorado, e dessa forma, é possível manter um acompanhamento bastante apurado sobre tais condições de forma remota, apenas analisando os dados capturados e enviados pelos sensores IoT.

Os dados capturados e analisados permitirão uma tomada de decisão sobre quais medidas devem ser tomadas e em quais locais específicos, tornando o entendimento e visualização da situação mais fáceis.

Este projeto visa analisar o desmatamento por município e bioma no território brasileiro entre os anos 2000 e 2023, apresentando os resultados de forma visual em um relatório no Google Looker Studio.

2 - Sobre a fonte de dados

A fonte de dados utilizada para elaboração deste projeto foi adquirida no portal basedosdados.org, via *download* direto de arquivo em formato .csv. O arquivo já apresenta dados estruturados, ou seja, em formato tabular, contendo os seguintes campos e seus respectivos tipos:

- ano: INT;
- id_municipio: STRING;
- bioma: STRING;
- area_total: INT - (área total do bioma);
- desmatado: FLOAT - (área total desmatada);
- vegetacao_natural: FLOAT - (área de vegetação natural);
- nao_vegetacao_natural: FLOAT - (área sem vegetação natural);
- hidrografia: FLOAT - (área de hidrografia);

Todos os campos que trabalham com áreas possuem km² como unidade de medida.

3 - Análise Exploratória dos Dados(EDA)

Inicialmente foi utilizado o *StringIndexer* para criar uma codificação numérica de cada bioma para facilitar na aplicação de modelos de *Machine Learning*; em seguida foi criada uma coluna de “caracteristicas” que corresponde a um vetor que reúne as principais

características de cada registro, como:
“area_total”,”codigo_bioma”,”vegetacao_natural”,”nao_vegetacao_natural”, e ”hidrografia”. Uma nova coluna foi criada para armazenar um novo vetor contendo os valores normalizados da coluna “caracteristicas”, com base na classe *MinMaxScaler*.

Foram calculadas as correlações entre algumas variáveis, com destaque para a correlação entre “area_total” e “vegetacao_natural”, que apresentou o valor de aproximadamente 0.99. Ou seja, com uma alta correlação é possível estimar o valor de uma variável com base nas alterações desta outra variável, já que um movimento em uma delas irá refletir na outra variável.

Em seguida foi implementado o modelo de regressão linear, que utilizou-se da coluna “caracteristicas_normalizadas” para estimar os valores da coluna “desmatado”. Ao alimentar o modelo com os dados, separamos em dados de treinamento (70%) e dados de teste (30%). Primeiramente treinamos o modelo de regressão linear, e em seguida usamos a parcela de teste dos dados para estimar os valores da coluna “desmatado”.

O modelo de regressão linear em questão apresentou um coeficiente de determinação (R2) igual a 0.99. Ou seja, como o valor máximo deste coeficiente é igual a 1, podemos inferir que o modelo de regressão linear implementado tem um funcionamento positivo na estimativa de valores. Na prática, movimentos ou alterações em “caracteristicas_normalizadas” terão impactos nos valores de “desmatado”. Com isso, é possível estimar valores para “desmatado” com base em valores de “caracteristicas_normalizadas”, já que possuem uma alta correlação.

Já no erro quadrático médio (RMSE), que corresponde a outra métrica de avaliação do modelo de regressão linear, a ideia é que o valor seja o menor possível, pois corresponde a percepção de possíveis erros. No caso do modelo implementado foi encontrado o valor de 0.02 para o erro quadrático médio (RMSE).

Por fim foi implementado o modelo de agrupamento *KMeans* de modo a criar 3 grupos para armazenar os registros do *dataframe*.

4 - Relatório de *Insights*

No relatório são apresentadas algumas abordagens contendo agregações diferentes, para possibilitar a extração de diversas informações e *insights*, como: total desmatado por ano, maior área desmatada por ano, maior desmatamento por bioma, entre outros. Com o uso de controles interativos de “bioma” e “ano” o usuário consegue personalizar e refinar ainda mais a visualização dos dados que realmente são necessários para extrair *insights* relevantes.

Os dados coletados foram registrados entre 2000 e 2023, e com base nas agregações realizadas é possível perceber uma tendência clara de crescimento em alguns aspectos.

O primeiro ponto a ser abordado é o maior desmatamento registrado por ano, independente do bioma atingido, e neste caso, há um crescimento progressivo de 2000 até 2023, sendo 2023 o ano que registrou o maior desmatamento até o momento.

Em seguida, ao considerar a média de desmatamentos por ano, temos um cenário bem parecido, em que há uma progressão nos valores numéricos de 2000 a 2023, sendo 2023 o ano com a maior média de desmatamentos, independente dos biomas mais afetados.

É possível perceber, com base no relatório, que o maior desmatamento realizado por bioma foi na Amazônia, independente do ano em que ocorreu. Assim como a média da área desmatada por bioma também aponta a Amazônia como maior valor. Por outro lado, a soma

de todas as áreas desmatadas indica que o Cerrado foi o bioma mais atingido neste intervalo de tempo. Outra informação relevante é que a Mata Atlântica possui o maior número de registros de desmatamentos.

5 - Conclusões

Dessa forma, é possível notar que, apesar de todos os biomas serem atingidos de forma sequencial ao longo dos anos, é muito forte o impacto de tais ações sobre a Amazônia, Mata Atlântica e Cerrado. De modo que é possível concluir que esse problema tende a piorar, ao olharmos para os números de cada ano e percebermos como crescem de forma linear.

O desmatamento se mostra diferente entre os biomas. Por exemplo, a Mata Atlântica não apresenta registro de uma grande área desmatada de uma só vez, como é o caso da Amazônia. Porém, a Mata Atlântica possui o maior registro de desmatamentos, ou seja, são pequenos danos que quando são somados mostram um cenário destrutivo para este bioma.

O ideal seria o desenvolvimento de políticas e programas de proteção a essas regiões de forma bem direcionada, analisando o “id_municipio” e verificando quais municípios possuem mais registros de desmatamento.

É necessário olhar para os municípios para conseguir aplicar uma mudança no bioma como um todo; uma vez que temos esses dados, fica mais fácil saber quais municípios são mais visados para esta prática.