Prova Prática para Cientista de Dados

Cargo Pretendido: Cientista de Dados
Candidato: Maicon Vinícius Ribeiro
(33) 9 8436 3870
maicon.ae@gmail.com

Pergunta

Como poderíamos prever e avaliar os impactos socioeconômicos do desmatamento no estado do Pará?

De forma geral, a previsão dos impactos socioeconômicos do desmatamento pode ser efetuada a partir da análise dos impactos já percebidos no decorrer dos últimos anos e projetada para o futuro.

O processo de previsão dos impactos depende diretamente da percepção de quais variáveis são mais ou menos relevantes para gerar determinados impactos. Isso é, faz-se necessário identificar adequadamente relações de causa e consequência entre os eventos, bem como identificar a importância das várias causas de um impacto socioeconômico.
Por exemplo, sabe-se que o desmatamento tem relação por exemplo com a frequência e intensidade de chuvas. Sabe-se também que as chuvas têm relação com a produção agrícola e geração de energia, o que por fim, gerará impactos econômicos. Então é possível inferir que o desmatamento gerará impactos dessa natureza. Cabe identificar com exatidão de quais formas (além dessa descrita) essa consequência pode se dar.

A partir da identificação das causas de determinados impactos socioeconômicos e percebida sua relação com o desmatamento, torna-se possível criar políticas de dirimição desses efeitos.

Sobre os dados utilizados nesta prova

Foram utilizados dados das seguintes origens:

- IBGE: Dados econômicos dos municípios do Brasil entre 2010 e 2019
- Mapbiomas: Dados de desmatamento e vegetação secundária entre 1986 e 2022
- INPE Prodes: Dados geográficos de regiões desmatadas na América do Sul
- IPEA: Dados socioeconômicos do estado do Pará: Impostos e receitas dos municípios

Essas bases de dados estão armazenadas dentro da pasta **dados** do projeto enviado junto a este documento. Exceto os dados do IPEA que, conforme explicado, foram obtidos via API e não possuem um arquivo armazenado localmente.

Sobre o desenvolvimento e organização do projeto

O projeto foi desenvolvido em python conforme solicitado e foram utilizadas bibliotecas específicas. Entre elas, destaco a biblioteca **Pandas**, utilizada nesse projeto para leitura de arquivos, exploração e análise de dados, a biblioteca **Numpy** utilizada para organização e preparação dos dados, a biblioteca **Sklearn** para desenvolvimento dos modelos de *machine learning*, a biblioteca **Matplotlib** para criação e visualização de dados de formas gráficas

Em relação ao aprendizado de máquina, foram utilizadas funções de regressão linear da biblioteca sklearn**.** Neste projeto, ela teve a intenção de identificar pontos comuns entre vários conjuntos de dados – correlacionando-os e percebendo valores não facilmente observáveis – e por fim, possibilitando a previsão de resultados futuros.

Tomei a liberdade de gravar uma breve demonstração sobre o projeto buscando explicar seu funcionamento de forma simplificada. É possível visualizar a estrutura e o funcionamento do projeto no vídeo a seguir:
https://youtu.be/WXUuqqPNFcE

Recomendações

Embora tenha sido possível analisar vários conjuntos de dados e relacioná-los uns aos outros, percebeu-se a necessidade de dados em um intervalo de tempo maior para obter resultados mais confiáveis.

Percebeu-se uma relação entre desmatamento e o produto interno bruto, na qual a medida em que o desmatamento diminui, o PIB apresenta um crescimento quase linear. Porém, por se tratarem de apenas 10 anos de dados, essa não é uma afirmação suficientemente segura.

Principais desafios observados

Tive algumas dificuldades específicas durante o processo de solução do exercício proposto. Tento a seguir descrevê-los seguidos de uma possível solução a ser implementada:

- Dificuldade em obter as bases de dados: Falta de experiência em buscar dados dessa natureza impediram a localização de dados organizados de forma útil em tempo satisfatório. Apenas a prática é suficiente para criar habilidades necessárias
- Dificuldade em realizar o treinamento de uma rede neural artificial: Dada a quantidade de dados utilizada, os resultados de treinamento não foram tão satisfatórios quanto o desejado, embora “funcionem”. Infelizmente não pude levantar mais dados para um treinamento apropriado
- Dificuldade em usar o Jupyter: A decisão de utilizar bancos de dados invés de ler os arquivos diretamente pela rede neural, impossibilitou o uso do jupyter. Optei pelo uso do banco de dados e não pude fazer uso dos notebooks pela dificuldade de transportar o banco de dados. Essa decisão se deu principalmente pela facilidade de tratar os dados com python.

Se houverem dúvidas, em relação aos procedimentos que foram adotados, me coloco a disposição para saná-las.