Um Estudo sobre a Influência da Telecomunicação nos Estados Brasileiros

Gabriel N. Santos[1], Victor H. L. Bauer[2]

FACOM – Universidade Federal do Mato Grosso do Sul (UFMS)
Campo Grande – MS – Brazil

g.neris@ufms.br, victor.bauer@ufms.br

1. Introdução

O acesso à internet é imprescindível, sendo o maior meio de comunicação da atualidade, porém requer infraestrutura dedicada para que seus inúmeros benefícios sejam alcançados pela população. Segundo o Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic), em 2021 haviam 35,5 Milhões de pessoas sem conexão com a internet, um número muito expressivo visto o valor que a esta agrega na vida das pessoas.

É necessário a realização de investimentos na área de telecomunicações para que a tecnologia alcance as pessoas que ainda não possuem acesso. O processo deve ser feito de forma estratégica, visando atender principalmente as áreas mais necessitadas para que estas recebam investimento de forma adequada para evoluírem tecnologicamente. Portanto seria essencial apurar quais são as áreas com maior déficit na área de telecomunicações, além dos fatores que possam influenciar no crescimento do setor na localidade, compreendendo adequadamente o cenário.

O objetivo deste estudo é responder os questionamentos contidos na Tabela 1, utilizando os conceitos de sistemas de apoio a decisão, a fim de compreender o entendimento do cenário atual da internet do Brasil, focando em possíveis melhorias que podem ser realizadas para que mais pessoas tenham acesso a essa tecnologia,

Tabela 1. Questionamentos a serem respondidos

N°	Questionamento
1	Quais as regiões brasileiras mais necessitadas de investimento na área de telecomunicações?
2	Qual a relação entre o crescimento na área de telecomunicações e o PIB por estado?

3	Quanto anos, de acordo com a taxa de aumento do número de residências com conexão à internet nos últimos anos, serão necessários para que os 3 estados com menor número acompanhem a média dos 3 estados com maior número da pesquisa mais recente(2018)?
4	Qual será o PIB do Brasil caso pelo menos 90% dos domicílios e escolas de cada estado tenham acesso à internet?

2. Materiais e Métodos

Para responder os questionamentos já apresentados, selecionamos três bases de dados públicas, sendo elas a Pesquisa Nacional por Amostra de Domicílios, os indicadores do Produto Interno Bruto dos Municípios e os Microdados da Educação Básica, todos sendo compreendidos dos anos de 2011 a 2018. As bases de dados e seus respectivos endereços seguem na tabela abaixo.

Tabela 2. Bases de dados

Base de dados (2011-2018)	Endereço eletrônico
Microdados Educação Básica	https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar
Pesquisa Nacional por amostra de domicílios	https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=downloads
Produto Interno Bruto dos Municípios	https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html

Após realizado o download dos arquivos correspondentes aos anos selecionados, foi feita uma análise a respeito da estrutura dos mesmos e definido quais seriam os dados a serem utilizados em nossa análise.

2.1. Data Warehouse

A agilidade e eficiência nas tomadas de decisões empresariais e o aumento na quantidade de dados armazenados em um banco de dados criaram uma lacuna nas tecnologias usadas para análises de dados(Inmon, 1996).

Focando no âmbito do armazenamento de dados, os chamados bancos de dados transacionais não possuem um bom desempenho para demandas onde precisa-se fazer um processamento analítico dos dados. Para isso, foi criado o conceito de *Data Warehouse*, sendo este segundo Date (2004) “um depósito de dados orientado por assunto,

integrado, não volátil, variável com o tempo, para apoiar as decisões gerenciais". Tal sistema oferece grande otimização para consultas, possibilitando análises avançadas com uma grande volumetria de dados.

2.2. ETL - Extração, Transformação e Carga dos Dados

O processo de ETL (Extract, Transform, Load), que segundo BARBOSA e CORREIA (2010), é o processo de extração de dados, que podem ser oriundo de diversas fontes, transformação desses dados para manter uma organização necessária, e carregamento do produto gerado para uma nova base, assim tornando-o mais fácil para análise.

Dessa forma, foi desenvolvido um programa em Java que fizesse o *download* e a descompactação das bases da PNAD, Microdados de educação básica e PIB dos municípios referentes aos anos de 2011 a 2018. foi necessário realizar uma série de varreduras pelas bases de dados da PNAD, Microdados da educação básica e PIB, a fim de obtermos as tabelas-fato e tabelas-dimensão para montar o nosso DW.

Com base em uma análise feita para podermos responder às perguntas propostas, foi definido que os dados necessários para respondê-las eram ano, estado, quantidade total de domicílios, quantidade de domicílios com acesso à internet, quantidade total de escolas, quantidade de escolas com acesso à internet e o valor do PIB do estado naquele ano. Assim foi criado uma classe EtlData contendo os campos "uf", "quantDomicilios" "quantEscolas", "quantDomiciliosAcesso", "quantEscolasAcesso", "nuAnoCenso", "pib".

2.2.1. Microdados Educação Básica

É disponibilizado pelo INEP anualmente uma planilha contendo diversos dados relevantes sobre as escolas brasileiras. Para nossa análise foi preciso analisarmos as colunas "SG_UF", "NU_ANO_CENSO" e "IN_INTERNET", a partir disso foi criado um HashMap<String, EtlData> etlDataMap, com chave "UF-nuAnoCenso" para guardar os somatórios de quantEscolas e quantEscolasAcesso para cada estado e ano, foi feita uma iteração pelas linhas da planilha buscando os dados

```
while ((line = br.readLine()) != null) {
    if (line.contains("NU_ANO_CENSO"))
        continue;
    String[] data = line.split(csvSeparator);

    // Extract the required column values
    String nuAnoCenso = data[0];
    String sgUf = data[4];
    String inInternet = data[187];
    String key = sgUf + "-" + nuAnoCenso;
    EtlData etlData = etlDataMap.get(key);
    // Create an instance of EtlData and populate it with the extracted values
    if (etlData == null) {
        etlData = new EtlData();
        etlData.setNuAnoCenso(nuAnoCenso);
        etlData.setUf(sgUf);
        etlData.setQuantEscolas(1.0);
        etlData.setQuantEscolasAcesso(equals(inInternet, "1") ? 1.0 : 0.0); //Sum schools with internet access
        etlData.setInInternet(inInternet);
    }
    else {
        etlData.setQuantEscolas(etlData.getQuantEscolas() + 1.0);
        etlData.setQuantEscolasAcesso(equals(inInternet, "1") ? etlData.getQuantEscolasAcesso() + 1.0 : etlData.getQuantEscolasAcesso());
    }
    // Add the EtlData object to the list
    etlDataMap.put(key, etlData);
}
```

Figura 1: Processo de ETL Microdados de educação.

Dessa forma, conseguimos fazer um incremento na nossa tabela AcessoInternet, com os campos quantEscolas e quantEscolasAcesso, que representam respectivamente a quantidade total de escolas e a quantidade de escolas com acesso à internet no estado.

2.2.2 PIB dos Municípios

Como as informações de PIB estavam separadas por município, foi preciso realizar o processo de ETL, com objetivo de obter uma relação dos valores de PIB por estado e ano. Para conseguir tal agrupamento, precisamos varrer os dados agrupando o campo Produto Interno Bruto, a preços correntes (R$ 1.000) pelos campos Ano e Sigla da Unidade da Federação. Com uma tabela contendo essas relações, pudemos fazer o incremento do campo pib na tabela AcessoInternet, já que os campos referentes à ano e UF já estão sendo cobertos pelas relações com as tabelas Tempo e Estado.

2.2.3. PNAD - Pesquisa Nacional por amostra de domicílios

Anualmente é disponibilizada pelo IBGE uma planilha contendo o levantamento de informações relevantes de cada estado, como quantidade total de domicílios, por tipo de conexão e com acesso à internet. Como a disposição dos dados são diferentes dos levantamentos de 2011 a 2015 dos anos subsequentes, foi preciso analisarmos as colunas “total” e “com acesso a internet” de cada estado e ano, já para os anos de 2016 a 2018 as colunas analisadas foram “total” e “havia” e “grandes regiões e unidades da federação” de cada ano. Assim, foram geradas novas planilhas para cada ano, contendo as colunas “

o processo de ETL em duas etapas. Assim, foi feito um processo de análise em todos os anos, a fim de gerarmos uma planilha para os anos de 2011 a 2015 com objetivo de termos amostras semelhantes para todos os anos da pesquisa, pois assim, se tornaria mais palpável o ETL para obtermos nossa base de dados final.

2.2.4 Taxa de crescimento dos domicílios com acesso nos últimos anos

Para responder um dos questionamentos propostos, vimos a necessidade de efetuar o cálculo da taxa de crescimento dos domicílios com acesso à internet nos últimos anos. No entanto, para obter tal informação, seria preciso fazer uma série de cálculos utilizando a colunas quantDomiciliosAcesso ao longo dos anos por estado, isso poderia acabar impactando no desempenho dessa consulta, por isso decidimos realizá-la apenas uma vez e transformá-la em uma nota tabela Crescimento. Essa tabela irá possuir os campos codTempo, que é a relação com a tabela Tempo, codEstado, que faz referência à tabela Estado e taxa, que será determinado pela fórmula (quantDomiciliosAcesso do ano atual/ quantDomiciliosAcesso do ano anterior)*100.

A partir das bases de dados já apresentadas, foi desenvolvida a modelagem do Data Warehouse, tendo em vista a resolução das perguntas propostas.

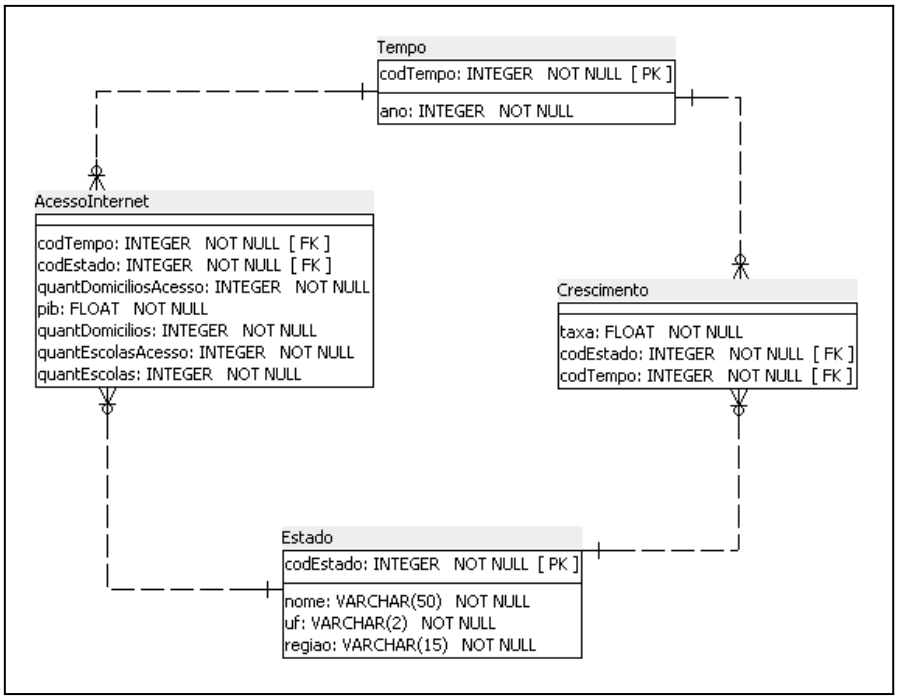


Figura 2: Modelagem do Data Warehouse.

2.3. Análise e Predição dos dados

Após o Data Warehouse estar pronto para consultas como planejado, iniciamos as consultas e manipulação dos dados a fim de responder às perguntas propostas. Para tais ações, foi adotada a linguagem de programação Python. Tal linguagem foi escolhida devido sua facilidade para manipular dados através de suas bibliotecas, além da sua versatilidade.

Visto que estamos utilizando o PostgreSQL para armazenar nosso Data Warehouse, utilizamos o módulo psycopg, que possibilita o trabalho com este SGBD. Através dele, foi desenvolvida uma classe de conexão, para estabelecer conexão com o Banco de Dados e realizar consultas.

```python
class Connection:
    def __init__(self, host, port, database, username, password):
        self.host = host
        self.port = port
        self.database = database
        self.username = username
        self.password = password
        self.conn = None

    def connect(self):
        try:
            self.conn = psycopg2.connect(
                host=self.host,
                port=self.port,
                database=self.database,
                user=self.username,
                password=self.password
            )
            print("Conexão estabelecida com sucesso!")
        except (psycopg2.Error) as e:
            print(f"Erro ao conectar-se ao banco de dados: {e}")
```

Figure 3. Trecho da classe de conexão.

Em todos os scripts foi utilizado o módulo pandas, pois este fornece estruturas e ferramentas para análise de dados de forma eficiente. A partir dele, foi possível trabalhar com DataFrame e realizar as análises. Somado ao uso dos módulos já mencionados, também foi empregado o uso do matplotlib para plotar gráficos a partir dos dados, facilitando a análise, o numpy para operações numéricas manipulações de dados, além do scikit-learn, sendo principal responsável pelas análises preditivas através de machine learning.

2.3.1 Questão 1

Foi realizada uma consulta SQL utilizando a classe de conexão, selecionando os dados das quantidades de domicílios e escolas de cada região, além de quantos destes possuem acesso à internet. Tais dados foram armazenados em um DataFrame e a partir deles, foram calculadas a porcentagem de escolas e domicílios com acesso. Foi utilizado então realizado um plot de gráfico com as valores para que seja facilitada a análise.

```python
#Consulta para resposta da questão 1

dfRegioes = connection.execute_query("SELECT e.regiao ,SUM(ai.quantdomiciliosacesso) as dom_acesso, SUM(ai.

dfRegioes['taxa_dom'] = round((dfRegioes['dom_acesso'] / dfRegioes['quant_dom']) * 100, 2)

dfRegioes['taxa_escolas'] = round((dfRegioes['escolas_acesso'] / dfRegioes['quant_escolas']) * 100, 2)

print(dfRegioes)

dfRegioes.plot(x='regiao', y=['taxa_dom', 'taxa_escolas'], kind='line')
plt.xlabel('Regiões Brasileiras')
plt.ylabel('Porcentagem de domicílios e escolas com Acesso à Internes')
plt.legend()
plt.show()
plt.show()
```

Figura 4. Código Python - Questão 1

2.3.2 Questão 2

Para responder a pergunta 2, realizamos uma consulta que nos traz como resultado os dados referente a escolas, domicílios e ao pib para todos os anos e estados presentes. A partir deles foi possível calcular a taxa de crescimento anual do PIB, dos domicílios com acesso a internet e de escolas com acesso à internet através de funções específicas para cada um.

```
#Consulta para a resposta da questão 2 - Todos os estados
dfPibInternet = connection.execute_query("SELECT e.nome, t.ano, ai.pib, c.taxa, ai.quantdomiciliosacesso as dom_acesso
dfPibInternet['taxa_dom'] = round((dfPibInternet['dom_acesso'] / dfPibInternet['quant_dom']) * 100, 2)
dfPibInternet['taxa_escolas'] = round((dfPibInternet['escolas_acesso'] / dfPibInternet['quant_escolas']) * 100, 2)

dfPibInternet = dfPibInternet.groupby('nome').apply(calcular_taxa_crescimentoDom)
dfPibInternet = dfPibInternet.reset_index(drop=True)
dfPibInternet = dfPibInternet.groupby('nome').apply(calcular_taxa_crescimentoEscolas)
dfPibInternet = dfPibInternet.reset_index(drop=True)
dfPibInternet = dfPibInternet.groupby('nome').apply(calcular_taxa_crescimentoPIB)

dfPibInternet.plot(x='ano', y=['Tx_Cresc_pib', 'Tx_Cresc_Dom', 'Tx_Cresc_Escolas'], kind='line')
plt.show()
```

Figura 5. Código Python - Questão 2

```
# Função para calcular a taxa de crescimento dentro de cada grupo
def calcular_taxa_crescimentoDom(group):
    group['Tx_Cresc_Dom'] = group['dom_acesso'].pct_change() * 100
    return group

# Função para calcular a taxa de crescimento de escolas com acesso dentro de cada grupo
def calcular_taxa_crescimentoEscolas(group):
    group['Tx_Cresc_Escolas'] = group['escolas_acesso'].pct_change() * 100
    return group

# Função para calcular a taxa de crescimento do PIB com acesso dentro de cada grupo
def calcular_taxa_crescimentoPIB(group):
    group['Tx_Cresc_pib'] = group['pib'].pct_change() * 100
    return group
```

Figura 6. Funções de cálculo de taxa de crescimento - Questão 2

2.3.3 Questão 3

Sendo necessária a utilização de predição temporal para a responder a questão 3, adotamos o modelo de Regressão Linear através do módulo scikit-learn, para então realizar a previsão dos valores futuros com base nos anos anteriores.

Porém, foi necessário estruturar os dados antes de executar um algoritmo de regressão linear, dessa forma realizamos as consultas responsáveis por trazer os dados dos 3 estados com as menores porcentagem de domicílios com acesso a internet, assim também os 3 estados com as maiores porcentagens. Foi calculada a média das porcentagens dos 3 estados com maior taxa, estabelecendo assim o valor no qual servirá de parâmetro para a predição.

Foi desenvolvida então a função prever(), responsável por executar o algoritmo de regressão linear, que recebe o Data Frame com os dados de anos anteriores dos três

estados com as menores porcentagens e a porcentagem que será alcançada. O código gera novos anos no Data Frame, baseado na sequência de valores anteriores, até que a porcentagem dos três estados seja igual ou maior a média das porcentagens dos três estados com as maiores porcentagens.

```
def prever(df, valor_desejado):
    # Lista para armazenar os anos preditos
    anos_preditos = []

    for estado, dados_estado in df.groupby('nome'):
        anos = dados_estado['ano'].tolist()
        porcentagemdomiciliosacesso = dados_estado['porcentagemdomiciliosacesso'].tolist()

        # Criação de matriz de características e vetor de alvo para treinamento
        X = [[ano] for ano in anos]
        y = porcentagemdomiciliosacesso

        # Criar e ajustar o modelo de regressão linear
        model = LinearRegression()
        model.fit(X, y)

        while porcentagemdomiciliosacesso[-1] < valor_desejado:
            prox_ano = anos[-1] + 1
            predicao = model.predict([[prox_ano]])

            anos.append(prox_ano)
            porcentagemdomiciliosacesso.append(predicao[0])

        if porcentagemdomiciliosacesso[-1] >= valor_desejado:
            for i in range(len(anos)):
                anos_preditos.append((estado, anos[i], porcentagemdomiciliosacesso[i]))

    return pd.DataFrame(anos_preditos, columns=['nome', 'ano', 'porcentagemdomiciliosacesso'])
```

Figura 7. Função responsável pela Regressão Linear - Questão 3

A mesma retorna então um Data Frame que compreende tanto os valores que foram utilizados na predição, quanto aqueles que foram gerados a partir dela. O Data Frame então é impresso e plotado para análise

```
#Chamada da função de previsão
df_previsao = prever(df_menores_estados, mediaPorcentagemMaior)
print(df_previsao)

cores = ['blue', 'red', 'green']

for estado, dados_estado in df_previsao.groupby('nome'):
    plt.plot(dados_estado['ano'], dados_estado['porcentagemdomiciliosacesso'], label=estado, color=cores.pop(0))
plt.xlabel('Ano')
plt.ylabel('Porcentagem de Domicílios com Acesso à Internet')
plt.axhline(y=mediaPorcentagemMaior, color='y', linestyle='--')
plt.legend()
plt.show()
```

Figura 8. Chamada da função prever() e plot do Data Frame Final - Questão 3

2.3.4 Questão 4

Para a predição do PIB utilizamos aprendizado de máquina através do módulo scikit-learn. Para o treinamento do modelo, utilizamos como entrada os dados de domicílios e escolas, além do próprio PIB dos anos anteriores, agrupando-os tanto por estado, região e país, para um maior número e variação de entradas.

O método desenvolvido prever_pib() é responsável por treinar e realizar a predição, recebendo o Data Frame e as colunas de entrada que servirão base para predizer o PIB.

```python
#Funcao realiza treinamento e retorna datagrama novo
def prever_pib(df, quantdomicilios, quantdomiciliosacesso, quantescolas, quantescolasacesso,
               porcDomAcesso, porcEscAcesso):

    # Definir as features (colunas de entrada) e o target (coluna a ser prevista)
    features = ['quantdomicilios', 'quantdomiciliosacesso', 'quantescolas', 'quantescolasacesso',
                'porcDomAcesso', 'porcEscAcesso']
    target = 'pib'

    # Dividir o dataframe em conjuntos de treinamento e teste
    X_train, X_test, y_train, y_test = train_test_split(df[features], df[target], test_size=0.2, random_state=42)

    # Criar o modelo de regressão linear
    model = LinearRegression()

    # Treinar o modelo com os dados de treinamento
    model.fit(X_train, y_train)

    # Fazer a previsão para os novos dados
    novos_dados = pd.DataFrame([[quantdomicilios, quantdomiciliosacesso, quantescolas, quantescolasacesso,
                                 porcDomAcesso, porcEscAcesso]], columns=features)
    predicao_pib = model.predict(novos_dados)
    print("---------------------------")
    print('A previsão do PIB é:', predicao_pib)
    print("---------------------------")

    # Criar um novo dataframe com as colunas utilizadas na aprendizagem
    df_novos_dados = pd.DataFrame([[quantdomicilios, quantdomiciliosacesso, quantescolas, quantescolasacesso,
                                    porcDomAcesso, porcEscAcesso]], columns=features)

    # Adicionar as colunas restantes do dataframe original ao novo dataframe
    for coluna in df.columns:
        if coluna not in features and coluna != target:
            df_novos_dados[coluna] = df[coluna].iloc[0]

    # Adicionar a coluna de previsão do PIB ao novo dataframe
    df_novos_dados[target] = predicao_pib[0]

    # Concatenar o novo dataframe com o dataframe original
    df_resultado = pd.concat([df, df_novos_dados], ignore_index=True)

    return df_resultado
```

Figura 9. Função prever_pib(). - Questão 3

A função train_test_split() proveniente do módulo scikit-learn é responsável por separar o Data Frame em um conjunto que será utilizado para o treinamento e outro que será utilizado na predição. Após a predição o novo valor é adicionado ao Data Frame.

3. Resultado e Discussão

Através dos processos mencionados, foram obtidas informações para as respostas das perguntas propostas. Dessa forma podemos analisar cada resultado.

3.1. Questão 1

A partir do código desenvolvido em Python, foi realizado o plot do gráfico a respeito de cada região e suas respectivas porcentagens de acesso a internet.

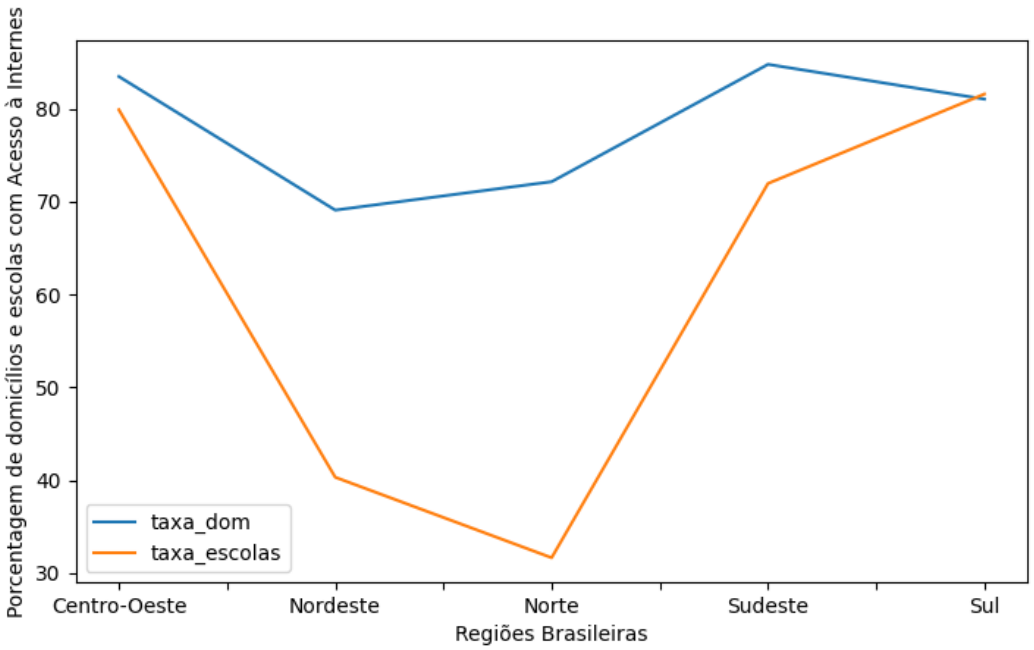


Figura 10.Gráfico Questão 3

Através dele é possível notar que Nordeste e Norte são os estados com as menores porcentagens de acesso a internet. Visto que os dados foram apenas consultados e estruturados para análise, o gráfico representa uma situação real.

3.1. Questão 2

Utilizando o algoritmo desenvolvido, conseguimos verificar a variação da taxa de crescimento do PIB juntamente com a variação da taxa de crescimento do número de escolas e domicílios com acesso a internet ao longo dos anos, por estado. Porém ao plotar o gráfico, devido ao grande número de estados, o resultado foi apresentado de forma poluída. Dessa forma, decidimos escolher aleatoriamente um estado de cada região e plotar o gráfico individual dele. Abaixo, podemos observar o estado de Mato Grosso do Sul e de São Paulo.

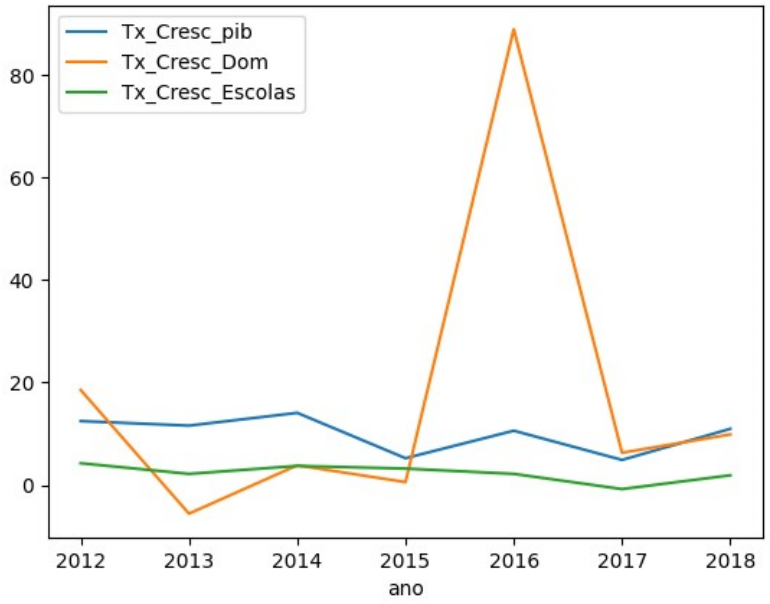


Figura 11.Gráfico Mato Grosso do Sul

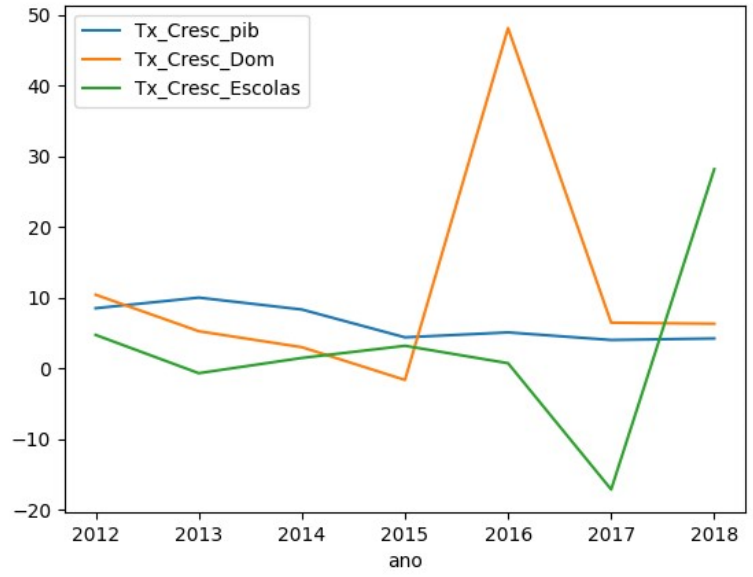
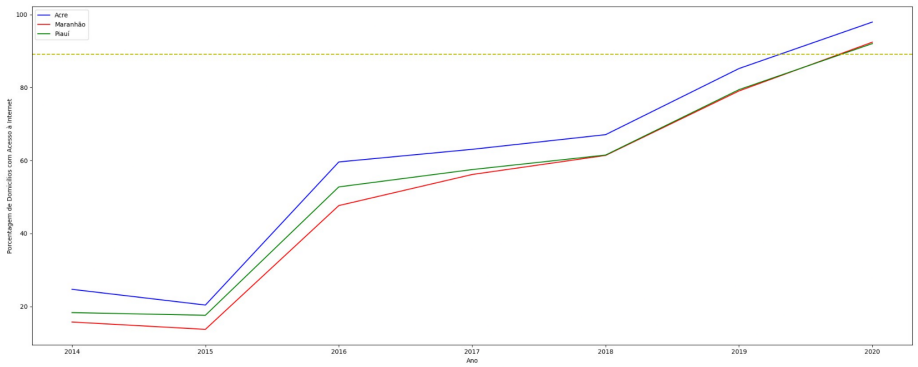


Figura 12.Gráfico São Paulo

Podemos notar que o PIB e as porcentagens de acesso a internet não possuem tanta relação assim, visto que nos casos apresentados acima, tivemos um caso onde houve acompanhamento e outro.

3.1. Questão 3

Através do código desenvolvido foi possível predizer em quantos anos ocorrerá o perguntado conforme é visto no gráfico.



7. References

INMON, W. H. (1996a) Como construir o data warehouse. Rio de Janeiro, RJ: Campus, 1996a.

DATE, C. J. (2004) - Introdução a Sistemas de Bancos de Dados. 8ª Ed., Rio de Janeiro: Campus, 2004.

BARBOSA, Luís ; CORREIA, Miguel P., eds.(2010) – "INForum 2010 : actas do II Simpósio de Informática, Braga, 2010." [Braga] : Universidade do Minho, 2010. ISBN 978-989-96863-0-4.

Knuth, D. E. (1984), The TeXbook, Addison Wesley, 15th edition.

CETIC.BR (2022)- Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros. Núcleo de Informação e Coordenação do Ponto BR. --