**Análise dos Rendimento Escolar em 2010**

Ramon de Saboya Gomes, Maria Luiza Menezes Vieira, Ullayne Fernandes Farias de Lima

Centro de Informática, UFPE

**INTRODUÇÃO**

Segundo o IBGE (2016), 51% dos brasileiros com mais de 25 anos possuem apenas o ensino fundamental completo, sendo uma taxa preocupante para a qualidade dos profissionais do país, bem como a própria qualidade de vida. Por isso, o grupo achou importante avaliar qual o rendimento dos estudantes brasileiros, bem como analisar a situação em que se encontram para tal rendimento e como melhorá-lo.

1. **OBJETIVOS**

Este projeto tem com objetivo fazer a análise da base de dados “Taxas de Rendimento Escolar na Educação Básica: Municípios 2010” a fim de observar as seguintes hipóteses:

1. O ensino médio têm o maior índice de abandono e/ou reprovação
2. O abandono/reprovação é maior em escolas do ensino público e em zonas rurais.
3. O índice de abandono está relacionado com o pib per capita do estado.
4. As regiões norte e nordeste possuem a taxa de aprovação média menor que nas demais regiões.

Para observar se as hipóteses levantadas estão corretas, será analisada, além da base citada, a que relaciona cada estado com o seu respectivo Produto Interno Bruto (PIB).

1. **JUSTIFICATIVA**

Para este projeto, o grupo escolheu o tema educação a fim de observar alguns dos fatores que possam influenciar no rendimento acadêmico de um estudante como condições do local de estudo e índice de desenvolvimento do estado que vive, pois acreditamos que observar esses fatores, pode ser benéfico para entender os motivos da desmotivação dos alunos em relação aos estudos.

1. **METODOLOGIA**

Os dados que serão analisados neste projeto estão divididos em 2 base de dados:

**Taxas de Rendimento Escolar na Educação Básica**

A base de dados do ministerio da educação está disponível no site do governo federal no portal de dados do Brasil atraves link <http://dados.gov.br/dataset/taxas-de-rendimento-escolar-na-educacao-basica?fbclid=IwAR1k-zIfTDd7K6rWWhIYY9JPyfkjYI8Pm83tY7_K29ZWntVCx5vMGoPCpMM>. Esta base possui dados de 2007 a 2010 e, por simplicidade, a análise realizada pelo grupo tem foco apenas no ano de 2010. Ela possui dados sobre: taxas de aprovação, reprovação e abandono etc. Os principais atributos são: ano, região, UF, código e nome do município, localizador (rural e urbana), rede (particular/estadual/municipal/federal) e taxa de aprovação, reprovação e abandono dos alunos por série.

**Contas Regionais do Brasil - Ano base 2010**

A base de dados esta disponivel no site do governo federal, no portal de dados do brasil atraves do link: <http://dados.gov.br/dataset/contas-regionais-do-brasil>. Esta base de dados conta com várias informações sobre o PIB, PIB per capita, população, variação do volume do PIB etc. Neste projeto, o grupo tomará proveito apenas do PIB per capita. O arquivo utilizado pelo grupo tem como principais atributos o ano e o estado.

* 1. **Desenvolvimento do Projeto**

A seguir, será listada as etapas de desenvolvimento do projeto como o processo de extração e pré processamento de dados, análise dos resultados, etc.

* + 1. **Coleta e Pré-Processamento**

Os dados foram obtidos no site do governo federal, portanto não se faz necessário coletar novos pois as bases de dados descritas são suficientes para analisar as hipóteses. Contudo, é necessário fazer pré-processamento dos dados. Será aplicado o ajuste dos tipos de atributos com pandas e o tratamento de dados que estão faltando. Além disso, é possível discretizar alguns dos atributos como o localizador para um atributo binário.

* + 1. **Análise das Hipóteses**

Para analisar as hipóteses, será aplicado algoritmos de regressão a fim de observar o comportamento dos dados.

Em primeira instância, o grupo pretende aplicar algoritmos de análise de comunidade baseado em três atributos: taxa de reprovação e abandono, taxa de aprovação e taxa de abandono isolada. A comunidade baseada em taxa de reprovação ou abandono irá analisar a correlação com a variável de localização ou escolaridade, sendo hipóteses diferentes. A segunda comunidade analisará a correlação da taxa de aprovação com as regiões. Por último, a terceira busca correlação entre o índice de abandono com o pib per capita do do estado. Para observar as relações dentro das comunidades o grupo pretende utilizar algoritmo de regressão logística e analisar se a hipótese realmente se concretiza.

* + 1. **Visualizações dos dados resultantes**

A visualização dos dados será importante para auxiliar a análise dos resultados, por este motivo, o grupo pretende plotar gráficos com relacionando diferentes atributos a fim de ver o comportamento dos dados na relacao: região relacionada com taxa de aprovação, taxa de reprovação/abandono com o PIB per capita, taxa de reprovação com escolaridade e taxa de reprovação/abandono com a localização com o auxílio da biblioteca bokeh.