Trabalho Prático - Introdução à Ciência dos Dados (CCF425)
Preparação dos Dados

Amanda Oliveira (ER05149)[2], Germano dos Santos (EF03873)[1]
Otávio Gomes (EF03890)[1], Thalita Mendonça (ER05141)[2]

[1] Universidade Federal de Viçosa - Campus Florestal

[2]Universidade Federal de Viçosa - Campus Rio Paranaíba

1. Introdução

Com o tema Taxas de Rendimento Escolar no Brasil, foi feita uma preparação dos dados coletados, no qual constam as taxas de rendimento dos alunos, a taxa de abandono escolar, o número de horas-aula diária, a média de alunos por turma, o percentual de docentes com nível superior, a taxa de adesão ao EAD em 2020 e censo escolar, com o objetivo de agrupar, organizar, entender e viabilizar a análise da variação na taxa de rendimento escolar e os fatores que a influenciaram. Para isso, na primeira parte desse trabalho foram elencadas 20 perguntas base para o entendimento desses dados. Na presente etapa, foi feita a análise e a preparação dos dados com o objetivo de entender suas características, agrupar os dados que serão utilizados, assim como limpar possíveis dados desnecessários.

2. Descrição dos Dados

Os dados analisados são:

- A base de dados relacionada a Rendimento, conta com, Taxa de Aprovação, Taxa de Reprovação e Taxa de Abandono, fazendo um agrupamento dos últimos 5 anos, com dados de todos os anos do Ensino Fundamentos e do Ensino Médio, para ambas as taxas, com especificações para escolas públicas, privadas, municipais e estaduais e também urbana e rural.
- Número de alunos por turma, no qual foram agrupados todos os dados referentes ao número de alunos por ano escolar, dependência administrativa (se é dado referente à escola pública estadual, federal ou municipal ou ainda escola particular) e localização (se é rural ou urbana) por município entre os anos de 2016 e 2020.
- Docentes com superior completo, no qual foram agrupados todos os dados por ano escolar (creche, pré-escola, fundamental I, fundamental II, médio, EJA e educação especial), dependência administrativa (se é dado referente à escola pública estadual, federal ou municipal ou ainda escola particular) e localização (se é rural ou urbana) por município entre os anos de 2016 e 2020.
- Número de horas aula, no qual foram agrupados todos os dados referentes ao número de horas aula por ano escolar, dependência administrativa (se é dado referente à escola pública estadual, federal ou municipal ou ainda escola particular) e localização (se é rural ou urbana) por município entre os anos de 2016 e 2020.
- Censo Escolar, que descreve as instalações das escolas como: existência de esgoto sanitário, existência de internet, computadores, etc. Essa base de dados será utilizada para investigar a associação entre a infraestrutura das escolas das regiões com a taxa de rendimento das escolas.
- Atributos das regiões, colhidos pelo IBGE, como PIB serão utilizados para associação de renda e densidade populacional à taxa de rendimento das escolas.

3. Preparação dos Dados

Para preparar os dados acima descritos, foi determinada uma janela temporal na qual todos os dados seriam avaliados, a janela escolhida foi dos últimos 5 anos. Esse tempo foi determinado, pois com ele será possível observar algum tipo de tendência nos dados, mas sobretudo será possível avaliar se houve alguma mudança devido à pandemia de Covid-19 no ano de 2020 com relação aos outros anos. Também foi definido que o escopo espacial ficará limitado à região sudeste do Brasil dividido em mesorregiões. Dessa forma, será possível avaliar os dados de maneira mais específica e com mais níveis de detalhes.

Tendo em vista essa preparação de dados, nas bases de dados sobre Taxas de Rendimentos, Número de alunos por turma, Docentes com superior completo e Número de horas aulas, foi feita uma filtragem, para que assim fosse selecionada apenas a região sudeste, como já mencionada, e também um agrupamento de todos os últimos 5 anos, para melhor avaliação dos dados. Além disso, foi feita a organização dos dados faltantes, que na tabela original contava com (–), alterando para Not a Namber (NaN). Como há um grande número de dados divididos em muitas partes, bem como suas agregações, por exemplo, taxa de rendimento por localização (Urbana, Rural e Total), esses números NaN não exercerão impacto muito grande nos resultados, pois ora serão suplantados pela existência de valores agregados, ora serão pouco relevantes para o agrupamento total.

Sobre o Censo Escolar, em [SOARES et al.] é criada uma escala de infraestrutura a partir dos dados do Censo. Esse trabalho modela a infraestrutura a partir do modelo logístico que possui como parâmetros variáveis dicotômicas relacionadas às instalações das escolas. Para tanto, a preparação dos dados do Censo Escolar, em [SOARES et al.], foi realizada de tal forma que permite a reprodução em nosso estudo. Portanto, selecionamos as mesmas variáveis que o trabalho citado, que descrevem: água consumida pelos alunos, abastecimento de água, abastecimento de energia elétrica, esgoto, sanitário, sala de diretoria, sala de professor, laboratório de informática, laboratório de ciências, sala de atendimento especial, quadra de esportes coberta/descoberta, cozinha, biblioteca, parque infantil, berçário, sanitário fora ou dentro do prédio, sanitário para educação infantil, sanitário para deficientes físicos, dependências para deficientes físicos, TV, DVD, copiadora, impressora, computadores, internet.

Para a coleta das informações do IBGE foi realizado o desenvolvimento de um cliente para a API de dados agregados do IBGE[1]. Assim, para cada informação coletada foi criada uma base em formato *csv*, logo 4 arquivos: densidade e pib das mesorregiões e munícipios da região Sudeste. A densidade obtida é resultado do censo IBGE 2010 e o PIB obtido é de 2018, portanto, é importante ressaltar que existirá uma defasagem em termos estatísticos que não há como contornar, pois o censo IBGE de 2020 ainda não foi realizado. Além disso, essas duas variáveis foram resumidas em índice GINI das mesorregiões da região Sudeste conforme mostra a figura 1.

[1] https://servicodados.ibge.gov.br/api/docs/agregados

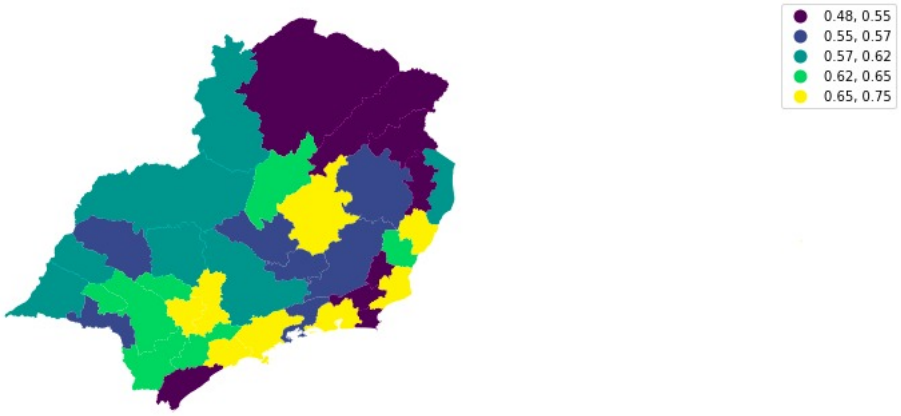

Figura 1. Índice GINI

4. Conclusão

Feita a análise inicial dos dados, assim como a elaboração de limites para a nossa análise, podemos observar que a grande maioria dos dados é numérica, com textos demarcando apenas o município a qual pertence cada informação, bem como a dependência administrativa e a localização. Conjuntos de dados foram agrupados para facilitar o manuseio dos dados, bem como para tornar os dados e nossos limites claros. Dessa forma, consideramos que os dados estão devidamente organizados e prontos para que sejam feitas as análises para responder as 20 perguntas inicialmente elaboradas.

Referências

SOARES, J. J., RIBEIRO, G., AKEMI, C., and FRANCISCO, D. Uma escala para medir a infraestrutura escolar. *Est. Aval. Educ.*, 24(54):78–99.