Análise da Taxa de Rendimento Escolar no Brasil
Introdução à Ciência dos Dados (CCF425)

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. Foi realizada 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. Foram realizadas também as análises de diferentes métricas para avaliação da qualidade do ensino ao longo do tempo e do espaço. Posteriormente foi realizada a análise de todos os dados, extraindo informações relevantes acerca de diversas características relacionadas à taxa de rendimento escolar, em específico da região sudeste. Por fim, fizemos uma análise preditiva buscando observar como essas taxas devem se comportar no futuro.

O restante deste trabalho está organizado da seguinte forma. A seção 2 apresenta a descrição dos dados recolhidos. Na seção 3 nós apresentamos a preparação realizada com os dados, agrupamentos, limpezas e organização geral. Na seção 4 discutimos e apresentamos os resultados obtidos através da análise dos dados colhidos. Na seção 5 mostramos a análise preditiva implementada. Na seção 6 apresentamos uma conclusão geral do relatório. Por fim, na seção 7 apresentamos os apêndices referenciados no texto.

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.

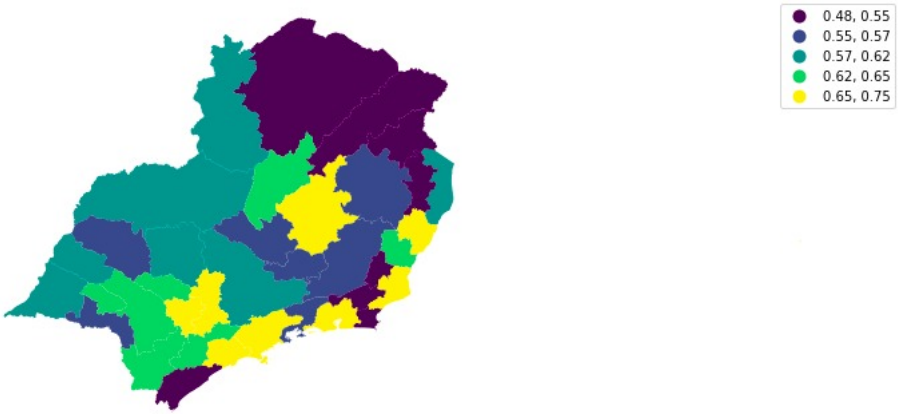

Figura 1. Índice GINI

4. Avaliação e Discussão

4.1. Análise de Taxas de Rendimento

As estatísticas educacionais, disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) nos mostram as taxas de rendimento escolar que são de extrema importância para o acompanhamento e averiguação dos dados referentes às escolas, além de ser um insumo no cálculo do Índice de Desenvolvimento da Educação Básica (Ideb), o indicador de qualidade educacional. As taxas de rendimento são divididas entre alguns conceitos, dentre eles a taxa de aprovação, taxa de reprovação e taxa de abandono, que foram utilizados no presente trabalho conforme mostrado em 7.1.

A taxa de abandono escolar consiste na porcentagem de alunos que pararam de frequentar a escola desde a data estabelecida como "data referência"pelo Censo. Analisando os dados pode-se notar que esta taxa, conforme mostrado em 7.2, que se encontrava com 0,1% em 2016, caiu para 0,08% no ano seguinte, 2017. Após a baixa, nos três próximos anos a taxa seguiu em aumento, com 0,15% em 2018, 0,17% em 2019 e 0,31% em 2020. Apesar das redes de ensino, em 2020 por evidente influência da pandemia de COVID-19 e como forma de adaptação para este cenário, receberem recomendações do Conselho Nacional de Educação (CNE) entre outras entidades, para que fosse adaptado os critérios de avaliação dos alunos levando em consideração os fins de aprendizagem de fato cumpridos na tentativa de se minimizar o abandono escolar, o mesmo não teve efeito visto que a taxa quase dobrou em comparação com o ano de 2019 e triplicou se avaliado os últimos cinco anos.

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

Comparadas as taxas de abandono escolar entre escolas públicas e privadas é discrepante, o índice da rede pública é sempre superior ao da rede privada, mesmo a última tendo sofrido um pequeno aumento nos últimos anos. Uma queda simbólica dessa diferença pode ser notada entre os anos de 2018 com a rede pública sinalizando uma evasão de 0,97% enquanto a rede privada 0,15% e 2019 com 0.62% de evasão na rede pública e 0,17% da rede privada. No ano de 2020, a rede pública marcou uma taxa de 1,17 % acentuando novamente as diferenças com a rede privada que possuiu 0,31% de abandono, tendo também sofrido um aumento considerável se comparado com o ano anterior.

A taxa de aprovação consiste na porcentagem de alunos que atingem as especificações mínimas para a conclusão da etapa de ensino em que estavam no fim do ano letivo. Fazendo uma análise geral dos dados de todas as escolas da região sudeste dos últimos cinco anos, temos em 2016 uma taxa de 93,63% de aprovação, 94,1% no ano de 2017, 94,27% no ano de 2018, no ano de 2019 teve 95,14% e 2020, 98,5%.

A taxa de reprovação consiste na porcentagem de alunos que não atingem as especificações mínimas para a conclusão da etapa de ensino em que estavam no fim do ano letivo. Analisando os dados pode-se notar que ao contrário da taxa de abandono, a taxa de reprovação vem obtendo uma queda constante com o passar dos anos. Nota-se que a taxa se encontrava com 5,37% em 2016, caindo para 4,98% em 2017, 4,86% em 2018, 4,29% em 2019 e 0,44% em 2020.

Fazendo uma análise mais específica sobre as escolas públicas e privadas da região sudeste conforme mostrado em 7.3, incluindo as de zona rural e urbana, quanto a taxa de aprovação, fazendo uma comparação entre ambas, em 2016 percebemos uma diferença significativa entre elas, com 92,7% de taxa de aprovação para pública e 98,3% para a privada. Já no ano de 2017, a escola pública teve um aumento para 93,3% e a escola privada se manteve com 98,5%. No ano de 2018 ambas não tiveram mudanças muito significativas, continuando com os dados aproximados do ano de 2017. No ano de 2019 a escola pública teve novamente um aumento para 94,4% e a privada com 98,6%. Em 2020, as escolas públicas e privadas tiveram um aumento significativo, principalmente a escola pública, que teve um salto para 98,3% de aprovação, e a privada, com 99,1%.

Continuando com as análises das escolas públicas e privadas, sobre a taxa de reprovação dos últimos 5 anos, em relação ao ano de 2016, a escola pública tinha uma taxa de 6,06% totalizando zona urbana e rural, e para a escola privada, 1,5%. No ano de 2017, podemos notar uma queda para ambas, com 1,3% para a escola privada e 5,5% para escola pública. Assim como na taxa de aprovação, a taxa de reprovação do ano de 2018 teve diferença de apenas 0,1% a menos do que a taxa de 2017. Já em 2019, a escola privada tinha uma taxa de 1,1% de reprovação e a escola pública 4,8%. O ano de 2020 teve uma queda na taxa de reprovação muito expressivo, com 0,5% de reprovação para escola privada e 0,4% para escola pública.

Comparando as escolas públicas e privadas em relação às taxas de reprovação e aprovação, podemos perceber que os três primeiros anos, ambas as taxas se mantiveram constantes, começando a ter uma mudança maior nos anos de 2019 e 2020, com valores maiores para o último ano.

Analisamos as taxas de aprovação e reprovação dos últimos cinco anos para todas

as escolas conforme mostrado na tabela 7.4, dos anos finais escolares, sendo eles, 9º do ensino fundamental e o 3ª do ensino médio. Para ambas as taxas, percebemos que houve um aumento ao longo dos anos da taxa de aprovação das duas séries, e consequentemente, uma diminuição da taxa de reprovação. O 9º ano do ensino fundamental tem como 90% de taxa de aprovação em 2016, e 6,99% de taxa de reprovação, se mantendo constante nos anos de 2017, e 2018, com pequenas alterações. No ano de 2019, a taxa de aprovação vai para 93,40% e reprovação cai para 5,20%, e no ano de 2020, também tem aumento nas taxas, com 97,66% de aprovação e 0,51% de reprovação. Para a 3ª série, a taxa de aprovação dos anos 2016, 2017 e 2018, tiveram em 93%, e reprovação variou entre 4,02% e 3,83%. No ano de 2019, a taxa de aprovação foi de 95,03%, e reprovação 2,97%, ambas não sofreram muita alteração em 2020.

4.2. Análise do Número de Alunos por Turma

A avaliação da mudança no número de alunos por turma ao longo do tempo se justifica, pois pode gerar um impacto direto na qualidade do ensino, uma vez que turmas muito cheias ou muito vazias podem gerar um impacto no aprendizado de cada estudante. Dessa forma, fizemos a análise dessa variável em diferentes níveis de avaliação. Fizemos uma análise considerando a região sudeste como um bloco e também análises individuais de cada estado. Observamos a variação do número de alunos por turma no ensino infantil, fundamental e médio separadamente.

Para facilitar o entendimento, serão mostrados nesta seção apenas os gráficos que representam toda a região sudeste, uma vez que a distribuição dos valores é bastante similar entre os estados, conforme pode ser observado nas Figuras 2, 3 e 4.

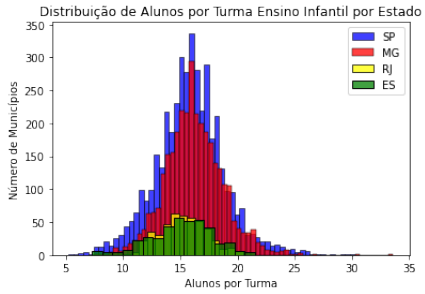

Figura 2. Distribuição do Número de Alunos por Turma - Infantil

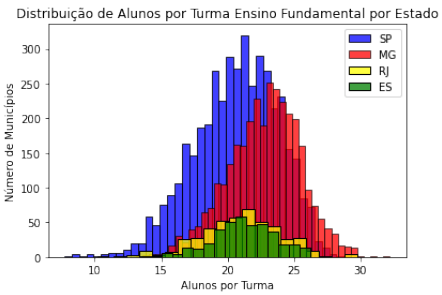

Figura 3. Distribuição do Número de Alunos por Turma - Fundamental

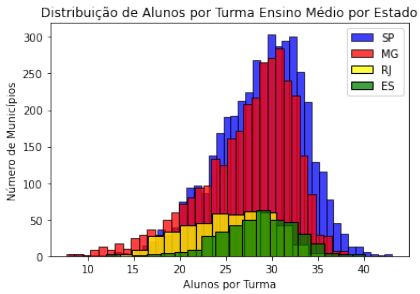

Figura 4. Distribuição do Número de Alunos por Turma - Medio

Nos gráficos apresentados nas Figuras 5, 6 e 7, são mostradas as variações do número de alunos por turma para os três níveis de ensino na região Sudeste entre os anos de 2016 e 2020.

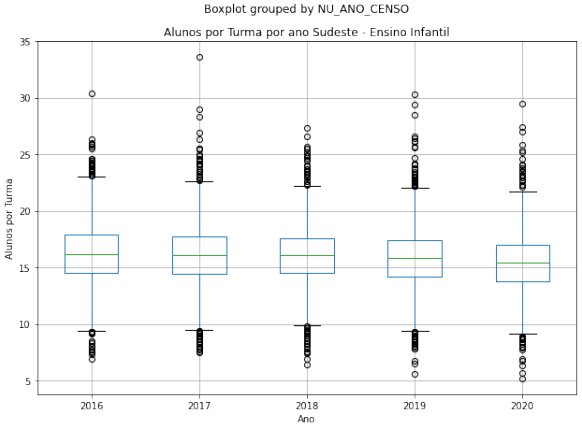

Figura 5. Alunos por Turma - Ensino Infantil, Sudeste

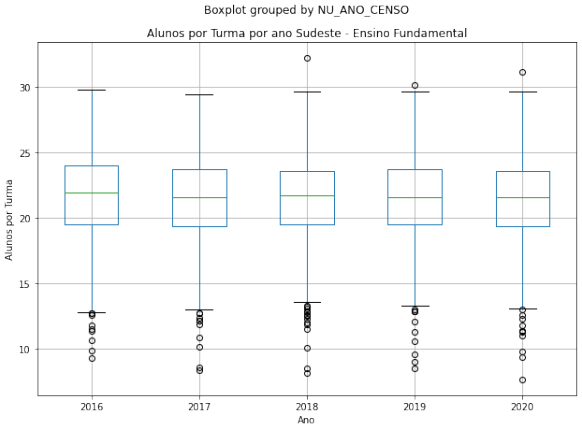

Figura 6. Alunos por Turma - Ensino Fundamental, Sudeste

Após plotar os gráficos, passamos então a um processo de análise, no qual foi feito um Teste de Hipótese com o objetivo de observar se, no ano de 2020, devido à pandemia de Covid-19, bem como outros fatores associados, o número de alunos por turma havia se alterado significativamente. É importante que aqui sejam levados alguns fatores em

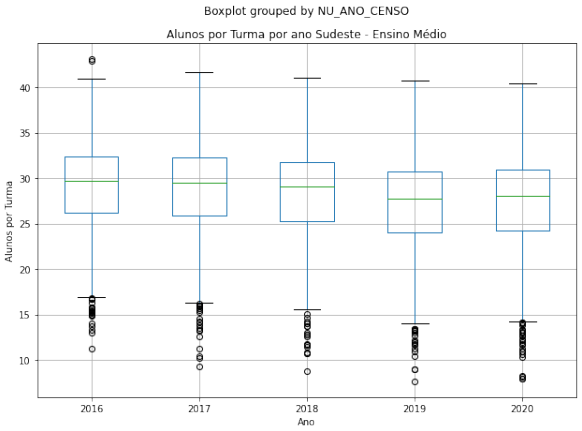

Figura 7. Alunos por Turma - Ensino Médio, Sudeste

consideração: primeiro, o número de crianças nascidas no Brasil vem caindo ao longo do tempo então com isso, espera-se que ocorra uma queda no número de alunos por turma ao longo do tempo, considerando que a maioria das escolas continuarão abertas no curto prazo, mesmo com a queda no número de alunos; segundo não foram analisados o número de escolas que oferecem cada nível de ensino, portanto podem haver mais escolas que oferecem o ensino fundamental do que médio, por exemplo, o que faz com que o número de alunos em cada um desses períodos escolares sejam diferentes. Por esse motivo, não foi avaliado o número de alunos por turma em si, mas sim sua variação ao longo do tempo.

Realizado o teste de hipótese, considerando um nível de significância de 5%, extraímos as seguintes conclusões: para o ensino infantil, houve uma queda significativa do número de alunos por turma, sendo que, observando a figura 8 podemos observar que há uma tendência de queda na quantidade de alunos por turma desde de 2016, porém 2020 apresentou a maior queda nesses anos. Para o ensino fundamental, apesar de ter havido

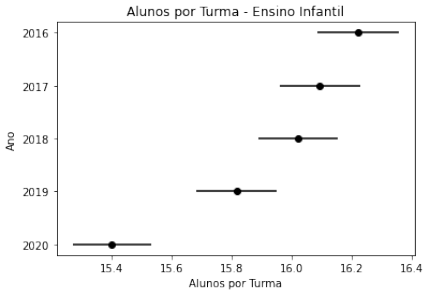

Figura 8. Alunos por Turma - Ensino Infantil, Sudeste

uma queda no número de alunos por turma essa queda não foi significativa, considerando um nível de significância de 5%. Observe na Figura 9 que a variação do número de alunos por turma não representa uma queda ou um aumento consistente ao longo dos anos analisados. Para o ensino médio podemos observar uma queda considerável no número de alunos em 2019, valor que não apresentou uma alta relevante em 2020, dessa forma, pelo teste de hipóteses observamos uma queda significativa para 2020 considerando o período de 2016 a 2019 como um todo, porém podemos observar pela Figura 10 que a queda do número de alunos se deu de maneira mais significativa em 2019.

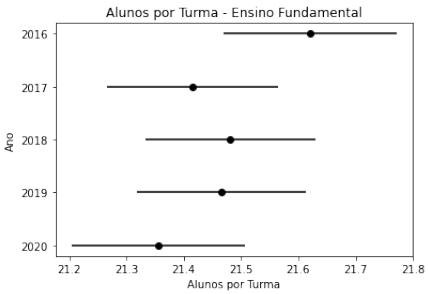

Figura 9. Alunos por Turma - Ensino Fundamental, Sudeste

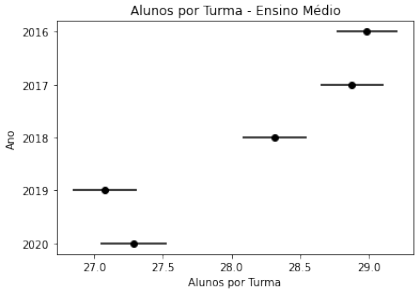

Figura 10. Alunos por Turma - Ensino Medio, Sudeste

Dessa forma, concluímos que, para a região sudeste, houve uma queda no número de alunos por turma no período entre 2016 e 2020 para os níveis infantil e médio, porém para o ensino fundamental esse número se manteve estável no período. Com os dados, não é possível dizer com certeza que a pandemia de Covid-19 afetou o número de alunos por turma, mas certamente esse valor está em queda, uma vez que menos crianças nascem por ano no país. Além disso, é importante mencionar que, com a queda natural do número de alunos há uma tendência de fechamento de salas e escolas, o que pode fazer com que essa métrica se estabilize ou mude muito pouco ao longo dos anos devido à política educacional de remanejamento dos estudantes.

4.3. Análise do Número de Horas Aula

Analisando agora o número de horas aula por dia ao longo do tempo, fizemos algumas observações interessantes.

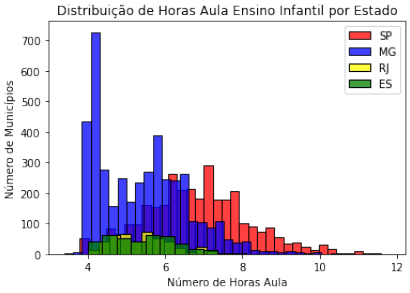

Figura 11. Distribuição do Número de Horas Aula - Infantil

Podemos observar nas Figuras 11, 9 e 10 que os municípios do estado de Minas Gerais têm um número de horas aula menor do que estado de São Paulo, enquanto os esta-

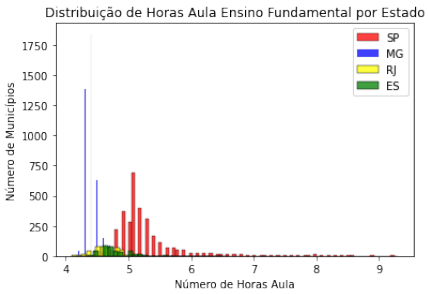

Figura 12. Distribuição do Número de Horas Aula - Fundamental

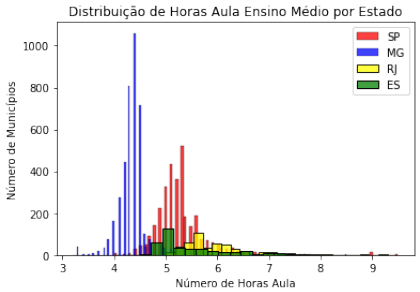

Figura 13. Distribuição do Número de Horas Aula - Medio

dos do Rio de Janeiro e Espírito Santo estão distribuídos de diferentes formas a depender do ano de ensino.

Da mesma forma como na Análise de Alunos por Turma, vamos apresentar aqui as estatísticas referentes à região sudeste como um todo, enquanto no github encontram-se os gráficos referentes a cada estado em particular.

Nos gráficos apresentados nas Figuras 14, 15 e 16, são mostradas as variações do número de horas aula diária para os três níveis de ensino na região Sudeste entre os anos de 2016 e 2020.

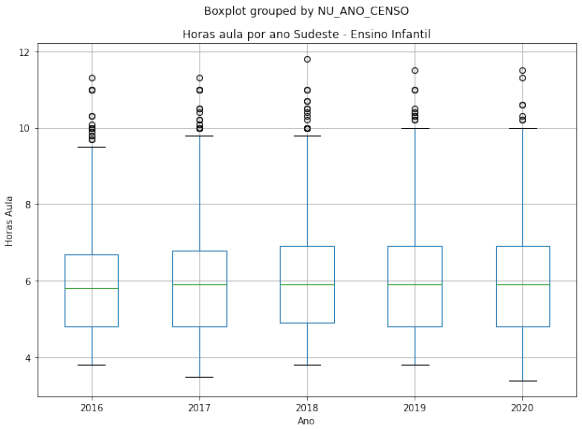

Figura 14. Horas Aula Diária - Ensino Infantil, Sudeste

Para se analisar os dados apresentados nesses gráficos foi realizado um Teste de Hipóteses, principalmente para avaliar se no ano de 2020 houve uma mudança signifi-

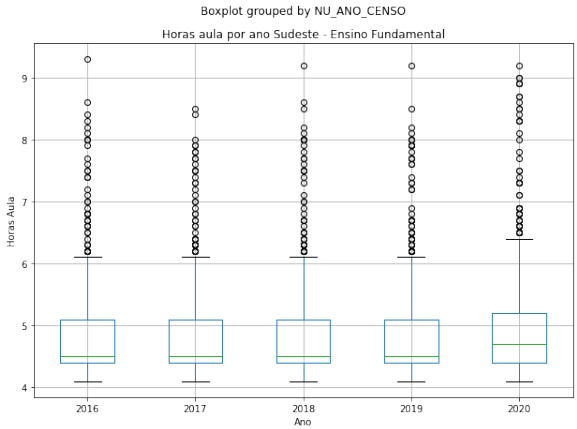

Figura 15. Horas Aula Diária - Ensino Fundamental, Sudeste

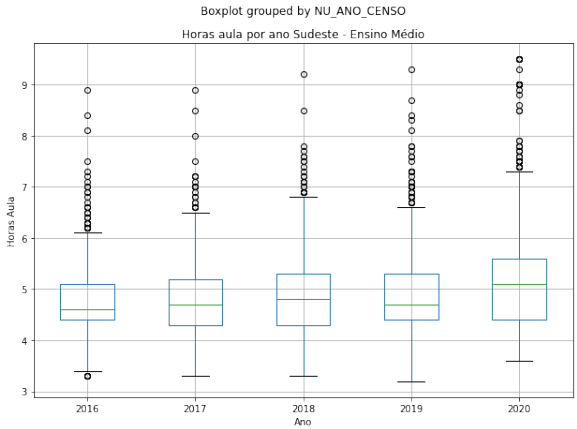

Figura 16. Horas Aula Diária - Ensino Médio, Sudeste

cativa nesse valor. Dessa forma, considerando um nível de significância de 5%, para o ensino infantil, não foi possível observar uma mudança significativa no número de horas aula no período entre 2016 e 2019 e o ano de 2020. Porém, observando o gráfico apresentado na Figura 17 podemos observar que o número de horas aula no ensino infantil vem aumentando ao longo dos alunos de maneira consistente. Para o ensino fundamen-

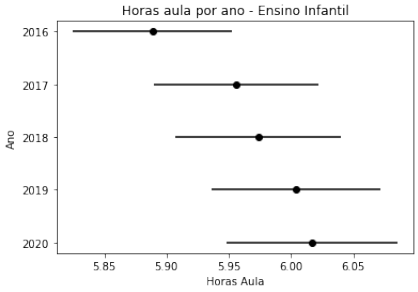

Figura 17. Horas Aula Diária - Ensino Infantil

tal também considerando um nível de significância de 5%, podemos observar que houve uma variação significativa se comparado o período de 2016 a 2019 e o ano de 2020. Observe no gráfico da Figura 18 que no ano de 2020 esse aumento foi diferente de tudo

que ocorria nos anos anteriores. Por fim, para o ensino médio também podemos observar

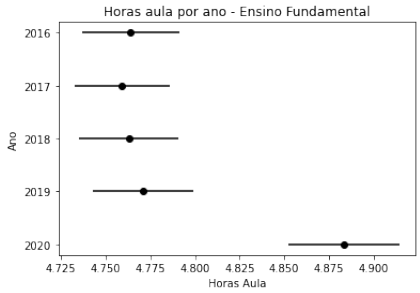

Figura 18. Horas Aula Diária - Ensino Fundamental

uma variação significativa no número de horas aula diária, também havendo um aumento diferente da tendência observada nos anos anteriores. Observe a Figura 19

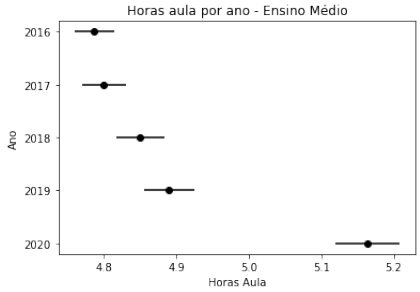

Figura 19. Horas Aula Diária - Ensino Médio

Dessa forma, concluímos que, para os níveis fundamental e médio, há uma grande chance da pandemia de Covid-2019 ter provocado um aumento no número de horas aula computado, o que não ocorreu para o ensino infantil. No entanto, devido ao fato de as aulas terem sido na modalidade online em sua grande maioria, não é possível determinar se realmente houveram mais horas aula ou se apenas devido à forma como foram computadas as horas aula nesse ano houve causou esse aumento.

4.4. Análise da Porcentagem de Docentes com Nível Superior

Analisando agora a variação da porcentagem de docentes com nível superior no período entre 2016 e 2020, observamos um aumento constante ao longo do tempo. Observe as Figuras 20, 21 e 22.

Assim, para avaliar de forma concreta essa variação foi realizado um teste de hipóteses com nível de significância de 5%. Para todos os níveis de ensino foi possível observar um aumento constante ao longo do tempo do número de docentes com nível superior. Para todos eles houve um aumento significativo considerando o período de 2016 a 2019 e o ano de 2020. É possível observar essa variação e a tendência de aumento nos gráficos das Figuras 23, 24 e 25.

Assim, concluímos que está em curso um processo no qual os professores cada dia mais necessitam ter um curso de graduação para conseguirem trabalhar, uma vez que, mesmo para o ensino infantil em média, mais de 80% dos docentes já possuem nível superior.

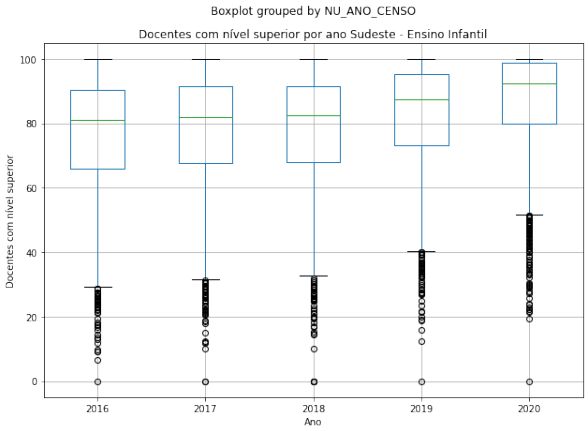

Figura 20. Docentes com Nível Superior - Infantil

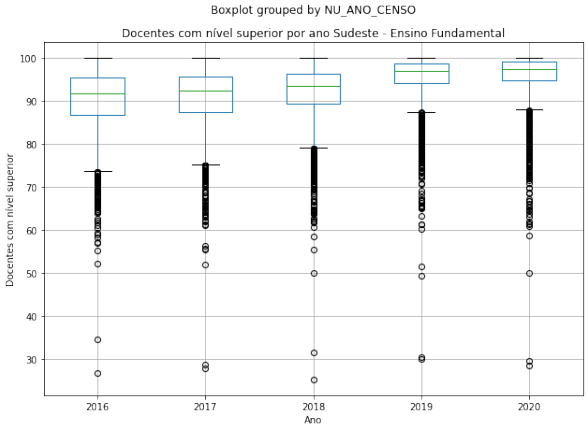

Figura 21. Docentes com Nível Superior - Fundamental

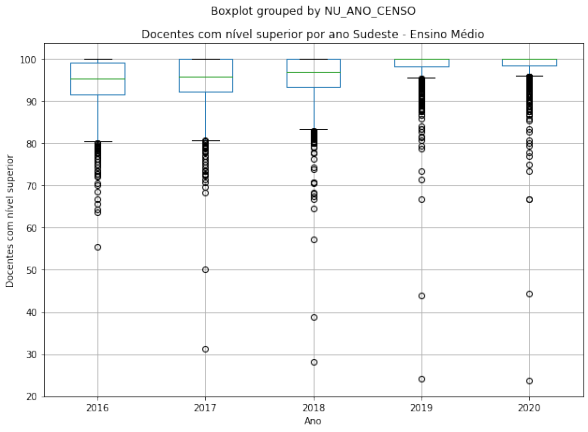

Figura 22. Distribuição de Docentes com Nível Superior - Médio

4.5. Análise de Atributos Geodemográficos

Segundo Cerqueira e Giviez [CERQUEIRA and GIVISIEZ]: ”A compreensão dos fenômenos demográficos, tanto em seus aspectos estáticos como dinâmicos, tem uma importância crucial na investigação das características educacionais de uma população”.

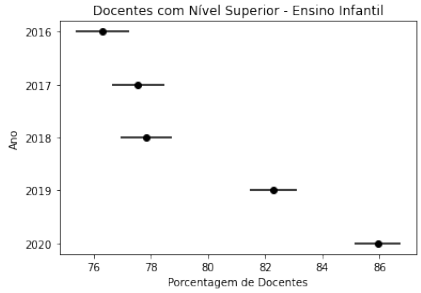

Figura 23. Porcentagem de Docentes com Nível Superior - Ensino Infantil

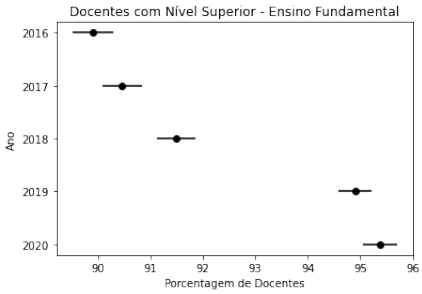

Figura 24. Porcentagem de Docentes com Nível Superior - Ensino Fundamental

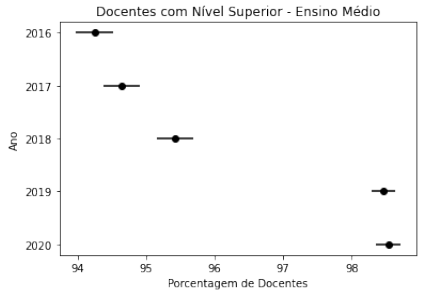

Figura 25. Porcentagem de Docentes com Nível Superior - Ensino Médio

Portanto, foi necessário analisar atributos disponibilizados pelo IBGE com o objetivo de associá-los com a taxa de rendimento escolar das mesorregiões da região Sudeste. Três atributos principais foram selecionados: pib, densidade e além da área quadrada de cada município. A partir da densidade e área quadrada, foi possível obter a população de cada município, logo o *pib per capita*. As variáveis socioeconômicas foram resumidas na taxa GINI, agrupando os municípios em mesorregiões como já mostrado na 1.

A partir da análise do GINI, é possível levantar hipóteses: a taxa GINI correlaciona-se com a taxa de rendimento? A taxa GINI correlaciona-se com a quantidade de escolas?

Para responder essas perguntas, foram feitas agrupamentos com a base de escolas, selecionando as regiões de interesse para contar as quantidades de escolas por mesorregião. Para visualização, geramos um gráfico de cloropleto conforme visto 26.

Analisando o gráfico acima, podemos nos perguntar se a taxa GINI relaciona com a quantidade de escolas das mesorregiões. É perceptível que essa correlação vista na

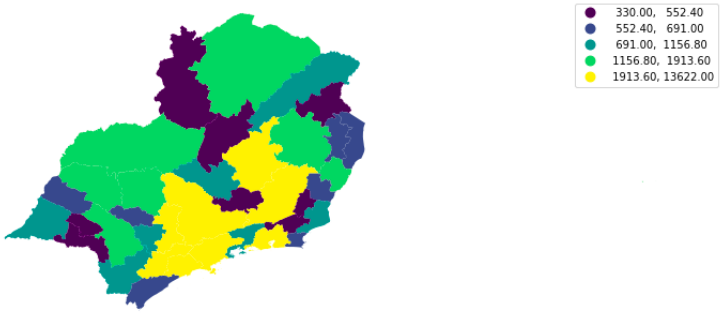

Figura 26. Quantidade de Escolas por Mesorregião

matriz 27 é suficiente para afirmar que existe uma relação entre as duas variáveis. O coeficiente positivo, nos indica que quanto mais uma região for igualitária, GINI maior, mais escolas existirão naquela região.

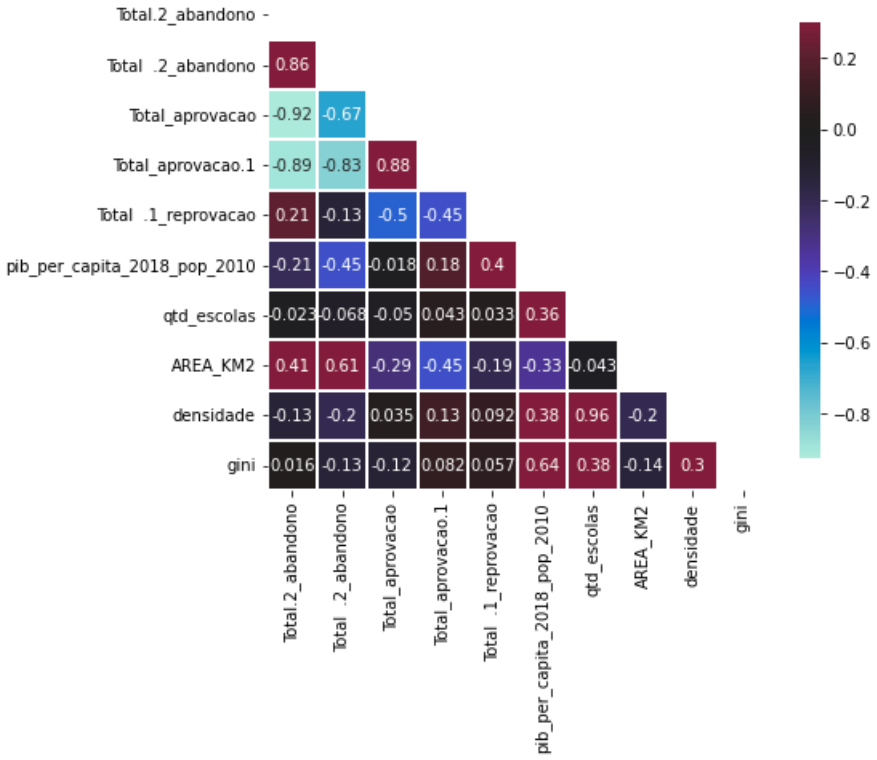

Figura 27. Matriz de correlação das variáveis por Mesorregião

Ainda analisando a matriz de correlação conseguimos associar a variável *AREA_KM2* que descreve a área da mesorregião em quilômetros quadrados com as variáveis de taxas de abandono: *Total.2_abandono* e *Total. 2_abandono*, que descrevem, a taxa de abandono no ensino fundamental e no ensino médio, respectivamente.

4.6. Associação da infraestrutura e taxa de rendimentos

A infraestrutura das escolas é um ponto importante quando almejamos analisar as taxas de rendimentos de cada escola. Portanto, verificamos se existe alguma associação entre

as variáveis do censo escolar de 2020 selecionadas mencionadas na seção 3 e as taxas de rendimento do mesmo ano. A associação foi feita utilizando a métrica lift, que tem por objetivo calcular se probabilidade de a escola que pertence a um grupo de rendimento alto possuir internet aumenta ou diminui, por exemplo.

Para realizar essa associação, foi preciso analisar a distribuição das taxas de aprovação em cada escola conforme visto em 28. É notável que muitas escolas possuem um taxa acima de 80% de aprovação, enquanto outras possuem uma taxa baixa, porém dispersa.

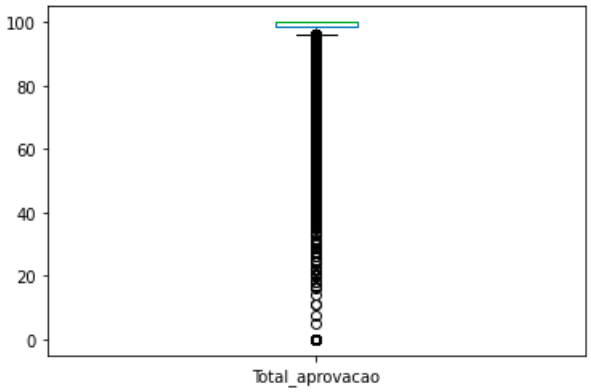

Figura 28. Distribuição de Taxa de Aprovação por Escolas

Decidimos, assim, separar as escolas em quatro classes: rendimento muito alto, acima de 99%; alto, acima de 80; médio, acima de 60 e baixo, abaixo de 60. As escolas ficaram distribuídas de acordo com a figura 29.

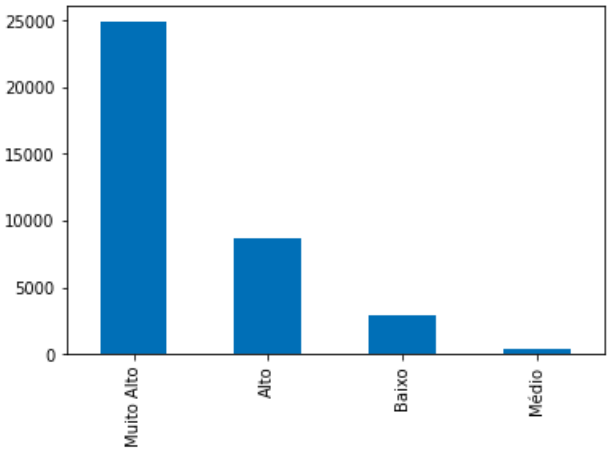

Figura 29. Distribuição por classes de Taxa de Aprovação por Escolas

Com essa separação é possível calcular a métrica lift da infraestrutura para cada classe conforme mostra as figuras 7.5, 7.6, 7.7, 7.8. Analisando essas figuras é possível perceber que as escolas que apresentam um rendimento considerado 'baixo' em nosso estudo, apresentam a métrica lift indicando uma melhor infraestrutura. Esse comportamento pode ser explicado pela cobrança das escolas, já que possuem uma infraestrutura

melhor que as demais escolas oferecem um nível de ensino melhor e maior cobrança para com seus alunos.

Essa hipótese pode ser verificada com a tabela apresentada em 30. É notável que as escolas federais se concentram no rendimento considerado como Baixo e há poucas escolas municipais, o que indica uma associação entre as taxas de rendimento e a dependência administrativa que a escola pertence, sendo a Municipal menos defasada em relação às federais.

Dependência Administrativa	Estadual	Federal	Municipal	Privada
aprovacao_classes				
Alto	2870	4	2506	3281
Baixo	1717	168	51	1022
Muito Alto	5736	6	13489	5656
Médio	134	0	76	173

Figura 30. Tabela de classes e dependência por Taxa de Aprovação por Escolas

5. Análise Preditiva

A análise preditiva foi baseada em uma regressão linear na taxa de aprovação, reprovação e abandono do ensino fundamental por município da região Sudeste[2]. Foi escolhido essa faixa etária por simplicidade para esse trabalho. É importante notar que o modelo criado para o ensino fundamental é muito parecido com o modelo preditivo para o ensino médio.

Utilizamos para o modelo atributos criados durante a análise deste trabalho: área total em quilômetros quadrados do município, pib, densidade, população de 2010 e o pib per capita do município, densidade, pib, área em quilômetros quadrados da mesorregião que o município pertence. Além disso, utilizamos o índice gini da mesorregião, o ano do censo, a quantidade de horas aula e alunos turma para cada município. Ainda calculamos a média da infraestrutura das escolas para cada município.

Os modelos não obtiveram um bom resultado. A métrica R^2 é de 0.42, 0.37 e 0.22 para a taxa de aprovação, reprovação e abandono, respectivamente. Esses valores podem ser explicados pelo fato de os dados não relacionarem diretamente com o ano, visto que os dados de infraestrutura é relativo apenas ao ano de 2020 e foram replicados para os outros anos (2016, 2017, 2018, 2019). Além disso, os atributos socioeconômicos como pib, população, densidade, estão defasados e não apresentam também uma relação direta com os anos.

Ademais, ressaltamos que a análise de rendimento escolar pode ser difícil de ser modelada apenas utilizado critérios quantitativos e pode requerer uma análise qualitativa importante, uma vez que características não mensuráveis, como suporte oferecido por pais e professores aos alunos, por exemplo, não podem ser mensurados de forma numérica.

Assim, mesmo realizando uma separação de variáveis, para aquelas que tinham um p-valor maior que 0.05, o modelo, como esperado, não obteve melhoras significativas.

[2] `https://nbviewer.org/gist/gegen07/0cd27344d8eeeaf3fa567df68270c952#An\%C3\%A1lise-Preditiva`

Para trabalhos futuros, é importante ter e analisar os dados respectivos aos anos para conseguir relacionar a geodemografia com as taxas de rendimento escolar, bem como buscar observar se existem mais métricas relacionadas à taxa de rendimento escolar que possam impactar os resultados. Por fim, é importante manter em mente a possibilidade de essa métrica ser influenciada de maneira significativa por critérios não numéricos, como discutido.

6. 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.

Após esse processo, realizada a análise dos dados, foram tiradas inúmeras conclusões, apresentadas na seção de Análise e Discussão. Essas conclusões parciais mostram que o banco de dados escolhido para ser trabalhado, nomeadamente, a taxa de rendimento dos alunos, apresenta muitas visões diferentes e proporciona a avaliação de diversos valores para se observar o que tem causado a mudança nessa taxa, bem como como ela deve se comportar no futuro. Por fim, realizamos uma análise preditiva por meio da regressão linear e observamos que os modelos obtidos não tiveram um bom desempenho.

Dessa forma, consideramos que o desenvolvimento deste trabalho propiciou uma abrangência grande acerca dos conteúdos aprendidos na disciplina de maneira aplicada a um tema específico. O tema se mostrou bastante complexo e cheio de nuances e requer um aprofundamento ainda maior para se poder chegar ao ponto de desenvolver uma análise preditiva que gere resultados interessantes. Apesar disso, todo o processo foi bastante proveitoso e consideramos que este trabalho serviu ao seu propósito de colocar em prática o que foi estudado ao longo da disciplina.

Referências

CERQUEIRA, C. A. and GIVISIEZ, G. H. N. Conceitos básicos em demografia e dinâmica demográfica brasileira. *Associação Brasileira de Estudos Populacionais*, pages 13–44.

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

7. Apêndices

7.1. Figura 1.

	Total_aprovacao			Total.1_reprovacao			Total.2_abandono		
	mean	std	median	mean	std	median	mean	std	median
Ano									
2016	93.634153	5.725787	95.1	5.373509	5.016473	4.0	0.992338	1.512900	0.5
2017	94.108793	5.527262	95.5	4.981213	4.856013	3.7	0.909994	1.391296	0.4
2018	94.277827	5.383617	95.7	4.862441	4.691158	3.6	0.859731	1.477050	0.3
2019	95.143750	4.931760	96.7	4.292698	4.488208	2.9	0.563552	1.119966	0.2
2020	98.514904	2.816622	99.6	0.444804	1.564477	0.0	1.040292	2.371663	0.0

Figura 31.

7.2. Figura 2.

			Total.2_abandono		
			mean	std	median
Ano	Dependência Administrativa	Localização			
2016	Privada	Rural	0.482456	1.663486	0.00
		Total	0.100787	1.000420	0.00
		Urbana	0.082267	0.933642	0.00
	Pública	Rural	0.528986	0.975072	0.00
		Total	1.142266	1.067121	0.90
		Urbana	1.223921	1.198869	0.90
2017	Privada	Rural	0.896610	2.225322	0.00
		Total	0.086825	0.509136	0.00
		Urbana	0.040805	0.206284	0.00
	Pública	Rural	0.465982	0.939436	0.00
		Total	1.051319	1.040951	0.80
		Urbana	1.127578	1.153324	0.80
2018	Privada	Rural	0.754098	2.520025	0.00
		Total	0.154331	1.950279	0.00
		Urbana	0.117403	1.887341	0.00
	Pública	Rural	0.442126	0.926606	0.00
		Total	0.976918	1.012620	0.70
		Urbana	1.045144	1.119364	0.70
2019	Privada	Rural	0.710606	1.819055	0.00
		Total	0.175953	2.200045	0.00
		Urbana	0.133378	2.161468	0.00
	Pública	Rural	0.334708	0.700836	0.00
		Total	0.629736	0.697508	0.40
		Urbana	0.660671	0.755655	0.40
2020	Privada	Rural	1.470588	3.912439	0.00
		Total	0.311549	1.116298	0.00
		Urbana	0.260695	0.749533	0.00
	Pública	Rural	0.512841	1.602032	0.00
		Total	1.173261	1.965746	0.25
		Urbana	1.267086	2.164726	0.20

Figura 32.

7.3. Figura 3.

			Total_aprovacao			Total.1_reprovacao		
			mean	std	median	mean	std	median
Ano	Dependência Administrativa	Localização						
2016	Privada	Rural	95.040351	5.163535	96.00	4.477193	4.731091	3.60
		Total	98.304593	2.492415	98.70	1.594619	2.282145	1.30
		Urbana	98.394933	2.387831	98.70	1.522800	2.210842	1.20
	Pública	Rural	95.076425	5.451199	96.70	4.394589	5.125750	2.70
		Total	92.789748	4.625183	93.60	6.067986	4.245105	5.20
		Urbana	92.453897	4.941350	93.45	6.322182	4.496824	5.30
2017	Privada	Rural	95.633898	4.943251	96.70	3.469492	3.863322	2.90
		Total	98.540580	2.436954	98.90	1.372596	2.315210	1.00
		Urbana	98.648725	2.216562	99.00	1.310470	2.189607	1.00
	Pública	Rural	95.385826	5.138893	96.90	4.148192	4.911532	2.60
		Total	93.360192	4.473618	94.30	5.588489	4.075196	4.70
		Urbana	93.037650	4.739909	93.90	5.834772	4.278407	4.80
2018	Privada	Rural	95.913115	4.910651	97.50	3.332787	3.786411	2.30
		Total	98.588583	2.462174	99.00	1.257087	1.431331	0.90
		Urbana	98.694511	2.295300	99.10	1.188086	1.281694	0.90
	Pública	Rural	95.515354	5.055661	97.10	4.042520	4.749482	2.50
		Total	93.526199	4.403973	94.50	5.496882	4.018080	4.50
		Urbana	93.222542	4.675866	94.25	5.732314	4.228876	4.70
2019	Privada	Rural	96.025758	6.436705	98.05	3.263636	5.951202	1.25
		Total	98.638502	3.809321	99.10	1.185545	3.083492	0.80
		Urbana	98.764611	3.494330	99.20	1.102011	2.784431	0.80
	Pública	Rural	96.051207	4.637537	97.60	3.614085	4.433040	2.15
		Total	94.496882	4.195587	95.50	4.873381	3.910239	3.90
		Urbana	94.280815	4.422619	95.30	5.058513	4.109286	4.05
2020	Privada	Rural	97.080882	5.349020	99.35	1.448529	4.041519	0.00
		Total	99.167060	2.654212	99.70	0.521391	2.395194	0.20
		Urbana	99.233957	2.515403	99.70	0.505348	2.380195	0.20
	Pública	Rural	99.061173	2.482838	100.00	0.425986	1.842531	0.00
		Total	98.393106	2.283814	99.30	0.433633	1.097459	0.00
		Urbana	98.302578	2.452833	99.30	0.430336	1.116615	0.00

Figura 33.

7.4. Figura 4.

	9º Ano_aprovacao			9º Ano.1_reprovacao			3ª série_aprovacao			3ª série.1_reprovacao		
	mean	std	median	mean	std	median	mean	std	median	mean	std	median
Ano												
2016	90.644313	7.523743	91.70	6.995402	6.265759	6.0	93.016727	6.542008	94.3	4.023286	4.921557	2.7
2017	90.905970	7.791354	92.20	6.484714	6.048709	5.3	93.392412	6.235282	94.7	3.836455	4.637079	2.5
2018	91.717688	7.279736	93.10	6.159598	6.108140	5.0	93.360655	6.432771	94.7	3.626001	4.445038	2.3
2019	93.402649	6.564116	94.80	5.200049	5.563439	3.9	95.300822	5.674890	96.9	2.975895	4.312794	1.6
2020	97.661621	4.190155	99.65	0.518309	1.877360	0.0	95.294196	7.666275	98.7	2.661019	5.701714	0.0

Figura 34.

7.5. Figura 5.

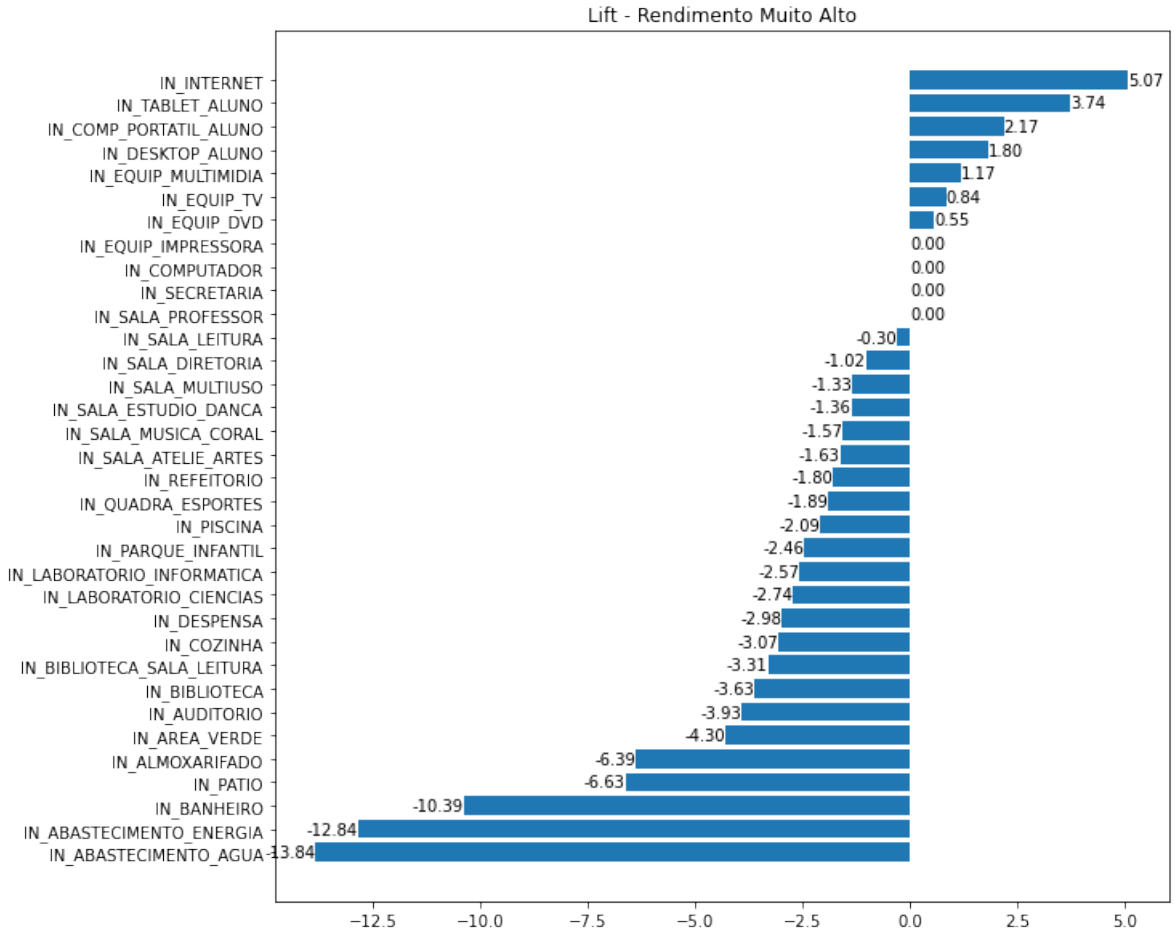

Figura 35.

7.6. Figura 6.

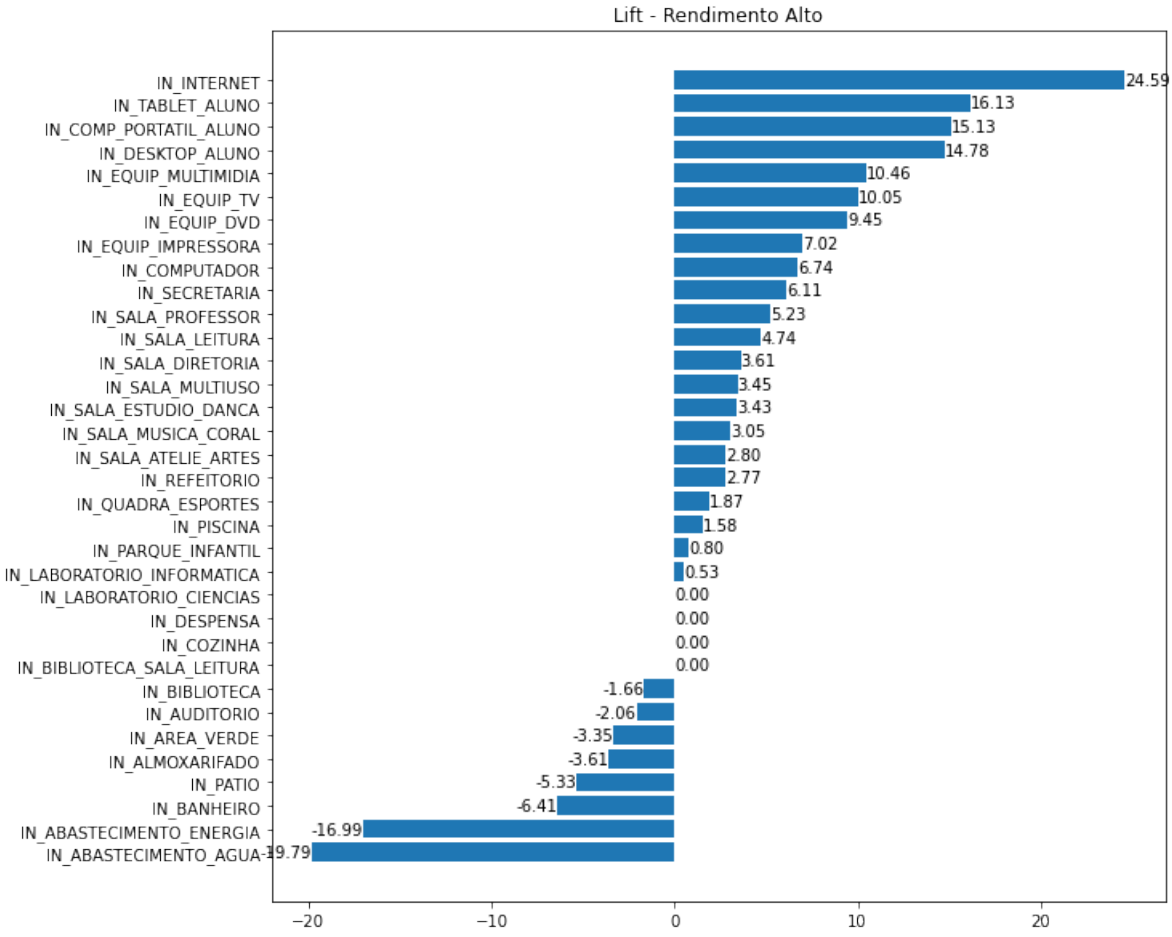

Figura 36.

7.7. Figura 7.

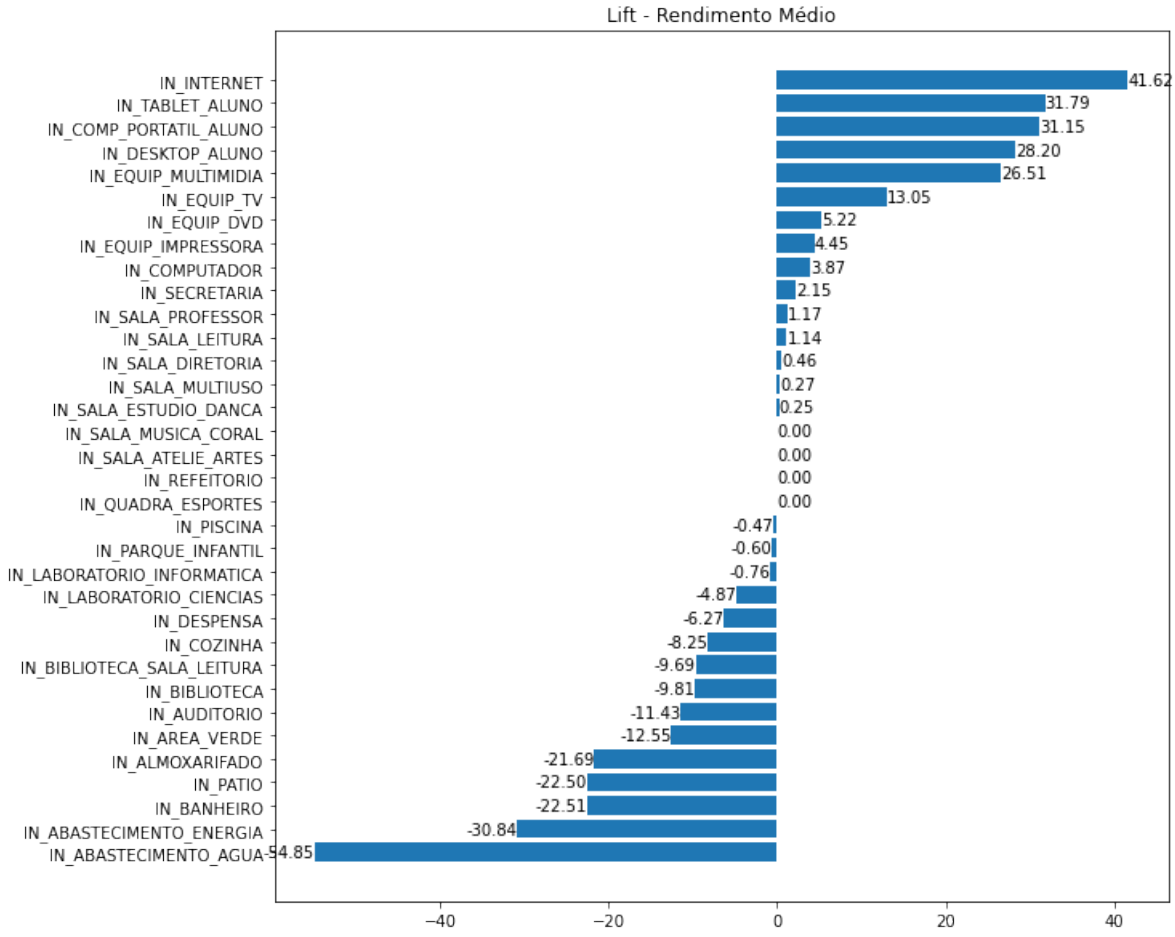

Figura 37.

7.8. Figura 8.

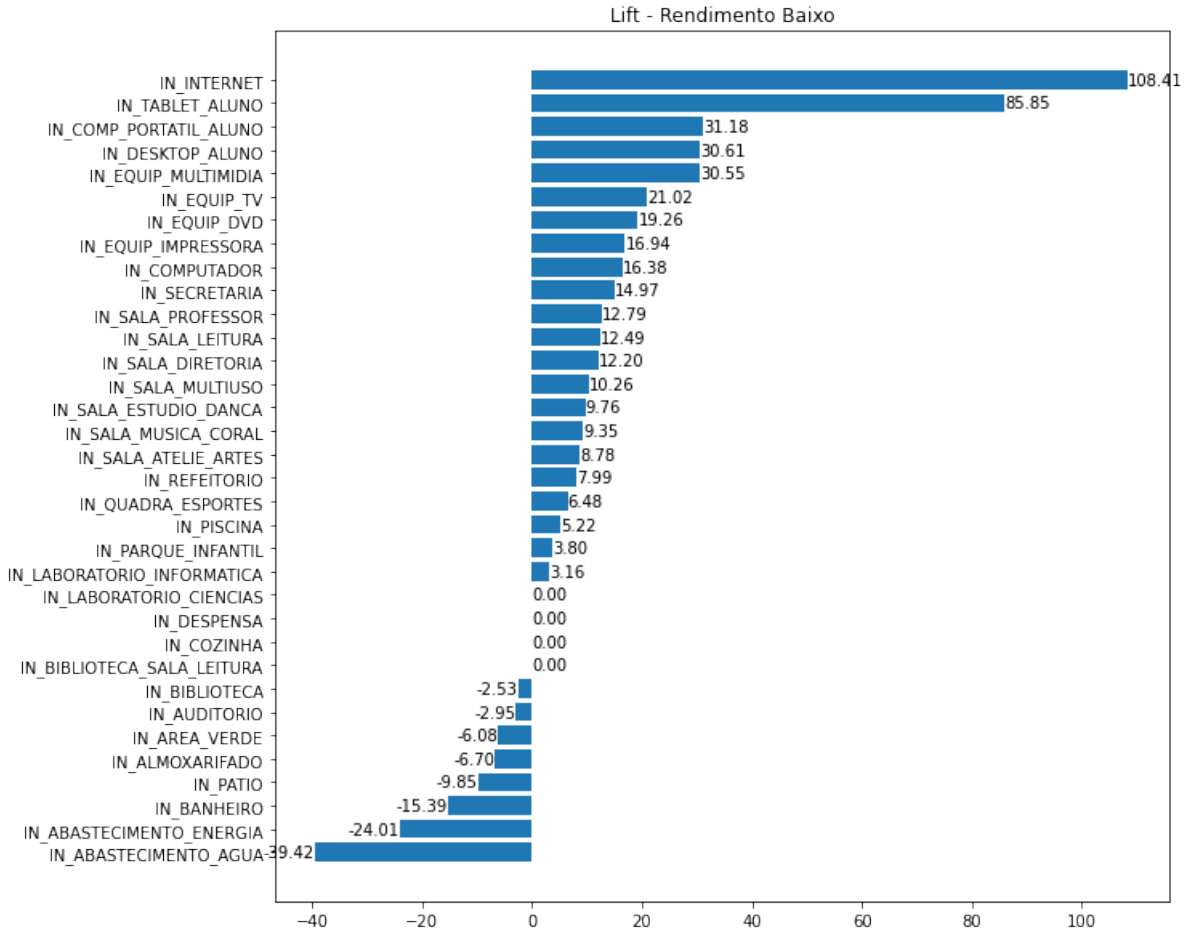

Figura 38.