PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Herlis Gomes Peixoto Junior

RECURSOS ESCOLARES E A QUALIDADE DA EDUCAÇÃO BÁSICA NO BRASIL

Belo Horizonte
2019

Herlis Gomes Peixoto Junior

RECURSOS ESCOLARES E A QUALIDADE DA EDUCAÇÃO BÁSICA NO BRASIL

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2019

SUMÁRIO

1. Introdução

1.1. Contextualização

A qualidade da educação pública no Brasil é motivo de preocupação, pois o país, apesar de em 2019 ser uma das maiores economias do mundo, não figura na mesma posição em rankings internacionais quando se trada da qualidade da sua educação pública. Essa deficiência na educação acaba por formar uma população com baixo capital humano, o que por consequência cria obstáculos para o desenvolvimento econômico-social do país.

Um dos fatores que impede um maior nível na qualidade da educação pública é o baixo valor de investimento por aluno. Somado a essa dificuldade, há a restrição fiscal a qual o governo brasileiro está sujeito e que impede maiores volumes de investimento público na educação. Assim, uma forma de obter ganhos de qualidade na educação é buscar uma melhoria na relação custo-benefício dos gastos com educação, priorizando os recursos educacionais que trazem maiores retornos para o aprendizado dos alunos.

1.2. O problema proposto

Dado a restrição fiscal que o Estado brasileiro possui e que dificulta maiores investimentos públicos na educação, o trabalho se propõe a identificar os recursos escolares que mais impactam na qualidade da educação básica, o que poderia permitir uma otimização na alocação dos recursos públicos disponíveis e consequentemente uma melhoria na qualidade da educação.

A fim de se mensurar a qualidade da educação, o estudo utiliza as notas obtidas na Prova Brasil e que fazem parte do Sistema de Avaliação da Educação Básica - Saeb. Utilizando uma amostra de escolas públicas, os dados mensuram a qualidade da educação por meio das notas em matemática e português dos alunos de escolas públicas do Ensino Fundamental I (1º ao 5º ano). Para mensurar os recursos escolares disponíveis nas escolas, utiliza-se o Censo Escolar que é realizado com a participação de todas as escolas públicas brasileiras. Todos esses

dados são obtidos e divulgados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP. A fim de tratar o problema de registros omissos em algumas variáveis empregadas no estudo, utilizou-se dados do Produto Interno Bruto *per capita* de 2016 dos municípios brasileiros, os quais são divulgados pelo Instituto Brasileiro de Geografia e Estatística – IBGE.

O objetivo do estudo é mensurar quais os recursos escolares que tem maior impacto no aprendizado dos alunos e que poderiam ser priorizados no gasto público, com o intuito de melhorar a qualidade da educação.

Para restringir o escopo de análise, o estudo se concentra nas notas do Saeb do 5º ano do Ensino Fundamental das escolas públicas brasileiras no ano de 2017. Quanto aos dados dos recursos escolares, utiliza-se o Censo Escolar do Brasil de 2017.

2. Coleta de Dados

Os dados com as notas do Saeb de 2017 foram obtidos no site do INEP no seguinte link: http://download.inep.gov.br/microdados/microdados_saeb_2017.zip, em 24 de setembro de 2019.

Os dados do Censo Escolar de 2017 também foram obtidos no site do INEP no link: http://download.inep.gov.br/microdados/micro_censo_escolar_2017.zip, em 24 de setembro de 2019.

Os dados do Produto Interno Bruto *per capita* de 2016 dos municípios brasileiros foram obtidos no site do IBGE no link: ftp://ftp.ibge.gov.br/Pib_Municipios/2016/base/base_de_dados_2010_2016_xls.zip, em 08 de outubro de 2019.

A descrição dos dados utilizados se encontra no quadro abaixo:

Saeb 2017			
Variável	**Tipo**	**Descrição**	**Código de Pre-enchimento**
ID_UF	Num	Código da Unidade da Federação	11-RO
			12-AC
			13-AM
			14-RR
			15-PA
			16-AP

			17-TO
			21-MA
			22-PI
			23-CE
			24-RN
			25-PB
			26-PE
			27-AL
			28-SE
			29-BA
			31-MG
			32-ES
			33-RJ
			35-SP
			41-PR
			42-SC
			43-RS
			50-MS
			51-MT
			52-GO
			53-DF
ID_MUNICIPIO	Num	Código do Município	
ID_ESCOLA	Num	Código da Escola	
ID_DEPENDENCIA_ADM	Num	Dependência Administrativa	1 - Federal
			2 - Estadual
			3 - Municipal
			4 - Privada
ID_LOCALIZACAO	Num	Localização	1 - Urbana
			2 - Rural
PC_FORMACAO_DOCENTE_INICIAL	Num	Indicador de Adequação da Formação Docente - Percentual de professores com formação em Licenciatura em Pedagogia (ou Bacharelado com complementação pedagógica)	
NIVEL_SOCIO_ECONOMICO	Char	Indicador de Nível Socioeconômico (Inse) - Calculado a partir do nível de escolaridade dos pais e da posse de bens e contratação de serviços pela família dos alunos.	1- Muito Baixo
			2 - Baixo
			3 - Médio Baixo
			4 - Médio
			5 - Médio Alto
			6 - Alto
			7 - Muito Alto
MEDIA_5EF_LP	Num	Média em Língua Portuguesa 5º ano	
MEDIA_5EF_MT	Num	Média em Matemática 5º ano	

Fonte: INEP

Censo Escolar 2017			
Variável	**Tipo**	**Descrição**	**Código de Pre-enchimento**
CO_ENTIDADE	Num	Código da Escola	
TP_SITUACAO_FUNCIONAMENTO	Num	Situação de funcionamento	1 - Em Atividade 2 - Paralisada 3 - Extinta (ano do Censo) 4 - Extinta em Anos Anteriores
IN_LOCAL_FUNC_PREDIO_ESCOLAR	Num	Local de funcionamento da escola - Prédio Escolar	0 - Não 1 - Sim
IN_AGUA_REDE_PUBLICA	Num	Abastecimento de água - Rede pública	0 - Não 1 - Sim
IN_AGUA_INEXISTENTE	Num	Abastecimento de água - Inexistente	0 - Não 1 - Sim
IN_ENERGIA_REDE_PUBLICA	Num	Abastecimento de energia elétrica - Rede pública	0 - Não 1 - Sim
IN_ENERGIA_INEXISTENTE	Num	Abastecimento de energia elétrica - Inexistente	0 - Não 1 - Sim
IN_ESGOTO_REDE_PUBLICA	Num	Esgoto sanitário - Rede pública	0 - Não 1 - Sim
IN_ESGOTO_INEXISTENTE	Num	Esgoto sanitário - Inexistente	0 - Não 1 - Sim
IN_LIXO_COLETA_PERIODICA	Num	Destinação do lixo - Coleta periódica	0 - Não 1 - Sim
IN_LIXO_OUTROS	Num	Destinação do lixo - Outros	0 - Não 1 - Sim
IN_SALA_DIRETORIA	Num	Dependências existentes na escola - Sala de Diretoria	0 - Não 1 - Sim
IN_SALA_PROFESSOR	Num	Dependências existentes na escola - Sala de professores	0 - Não 1 - Sim
IN_LABORATORIO_INFORMATICA	Num	Dependências existentes na escola - Laboratório de informática	0 - Não 1 - Sim
IN_LABORATORIO_CIENCIAS	Num	Dependências existentes na escola - Laboratório de ciências	0 - Não 1 - Sim
IN_QUADRA_ESPORTES	Num	Dependências existentes na escola - Quadra de esportes coberta ou descoberta	0 - Não 1 - Sim
IN_COZINHA	Num	Dependências existentes na escola - Cozinha	0 - Não 1 - Sim
IN_BIBLIOTECA_SALA_LEITURA	Num	Dependências existentes na escola - Biblioteca e/ou Sala de leitura	0 - Não 1 - Sim
IN_PARQUE_INFANTIL	Num	Dependências existentes na escola - Parque infantil	0 - Não 1 - Sim
IN_BERCARIO	Num	Dependências existentes na escola - Berçário	0 - Não 1 - Sim
IN_BANHEIRO_EI	Num	Dependências existentes na escola - Banheiro adequado à educação infantil	0 - Não 1 - Sim
IN_SECRETARIA	Num	Dependências existentes na escola - Sala de secretaria	0 - Não 1 - Sim

IN_BANHEIRO_CHUVEIRO	Num	Dependências existentes na escola - Banheiro com chuveiro	0 - Não 1 - Sim
IN_REFEITORIO	Num	Dependências existentes na escola - Refeitório	0 - Não 1 - Sim
IN_ALMOXARIFADO	Num	Dependências existentes na escola - Almoxarifado	0 - Não 1 - Sim
IN_AUDITORIO	Num	Dependências existentes na escola - Auditório	0 - Não 1 - Sim
IN_PATIO_COBERTO	Num	Dependências existentes na escola - Pátio Coberto	0 - Não 1 - Sim
IN_AREA_VERDE	Num	Dependências existentes na escola - Área Verde	0 - Não 1 - Sim
IN_LAVANDERIA	Num	Dependências existentes na escola - Lavanderia	0 - Não 1 - Sim
IN_EQUIP_TV	Num	Equipamentos existentes na escola - Aparelho de televisão	0 - Não 1 - Sim
IN_EQUIP_DVD	Num	Equipamentos existentes na escola - DVD	0 - Não 1 - Sim
IN_EQUIP_COPIADORA	Num	Equipamentos existentes na escola - Copiadora	0 - Não 1 - Sim
IN_EQUIP_RETROPROJETOR	Num	Equipamentos existentes na escola - Retroprojetor	0 - Não 1 - Sim
IN_EQUIP_IMPRESSORA	Num	Equipamentos existentes na escola - Impressora	0 - Não 1 - Sim
IN_EQUIP_SOM	Num	Equipamentos existentes na escola - Aparelho de som	0 - Não 1 - Sim
IN_EQUIP_MULTIMIDIA	Num	Equipamentos existentes na escola - Projetor Multi-mídia (Datashow)	0 - Não 1 - Sim
IN_EQUIP_FOTO	Num	Equipamentos existentes na escola - Máquina foto-gráfica/Filmadora	0 - Não 1 - Sim
IN_COMPUTADOR	Num	Equipamentos existentes na escola - Computador	0 - Não 1 - Sim
IN_INTERNET	Num	Acesso à Internet	0 - Não 1 - Sim
IN_BANDA_LARGA	Num	Internet Banda Larga	0 - Não possui 1 - Possui
IN_ALIMENTACAO	Num	Alimentação escolar para os alunos	0 - Não oferece 1 - Oferece

Fonte: INEP

3. Processamento e Tratamento de Dados

3.1. Dados Saeb

Os dados do Saeb de 2017 possuem 73.674 registros. Os dados não apresentam registros duplicados.

No tocante às informações ausentes, a variável *media_5ef_total* possui 25.844 registros ausentes. Como essa variável é a variável resposta, efetuou-se a

exclusão dos registros ausentes. As três variáveis *pc_formacao_docente_inicial*, *nu_matriculados_censo_5ef* e *taxa_participacao_5ef* possuem respectivamente 197, 58 e 58 registros ausentes. Como a quantidade de registros é pequena em relação ao tamanho da amostra, decidiu-se por excluir esses registros ausentes. Também foram excluídos os registros em que a taxa de participação dos alunos da escola no Saeb foi igual a zero, pois nesses casos não foi possível mensurar o desempenho da escola.

A variável *nivel_socio_economico* possui 10.016 registros ausentes. A quantidade de registros ausentes é representativa com relação ao tamanho da amostra de dados. Além disso, devido à correlação entre a variável resposta e a variável *nivel_socio_economico*, omitir essa variável do modelo pode trazer viés ao modelo a ser estimado. Dessa forma, decidiu-se por utilizar um modelo de árvore de decisão para estimar os registros faltantes para a variável em questão. Para estimar as categorias dos registros faltantes utilizou-se o Produto Interno Bruto *per capita* de 2016 do município em que a escola está localizada.

3.2. Dados Censo Escolar

O Censo Escolar de 2017 possui 185.925 registros. Os dados não apresentam registros em duplicidade.

O número de registros do Censo Escolar é maior que o número de registros do Saeb devido ao fato de que o Saeb é realizado sobre uma amostra das escolas brasileiras, enquanto que o Censo Escolar é realizado com todas as escolas do Brasil. Dessa forma, utilizou-se apenas os registros do Censo Escolar de 2017 em que a escola também participou do Saeb no respectivo ano.

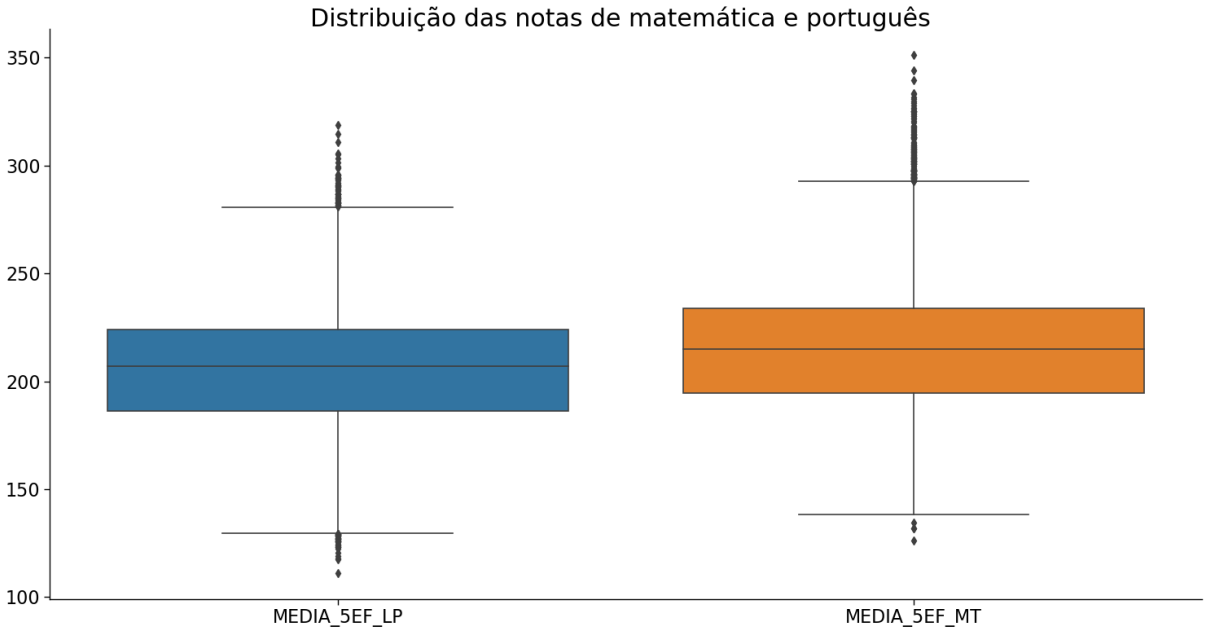
Com relação aos registros faltantes, as variáveis utilizadas no estudo não apresentam dados faltantes.

4. Análise e Exploração dos Dados

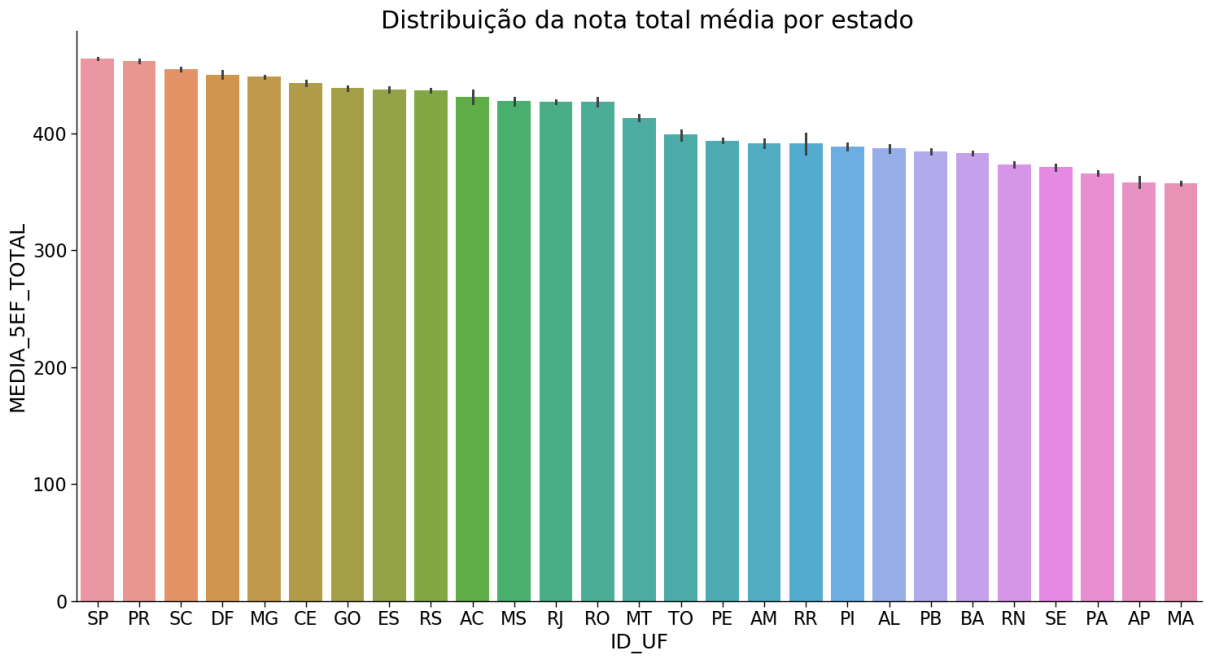
De modo a analisar e explorar os dados em estudo, utilizou-se gráficos, tabelas cruzadas e matriz de correlação com as variáveis disponíveis nos dados.

4.1. Análise dos dados do Saeb

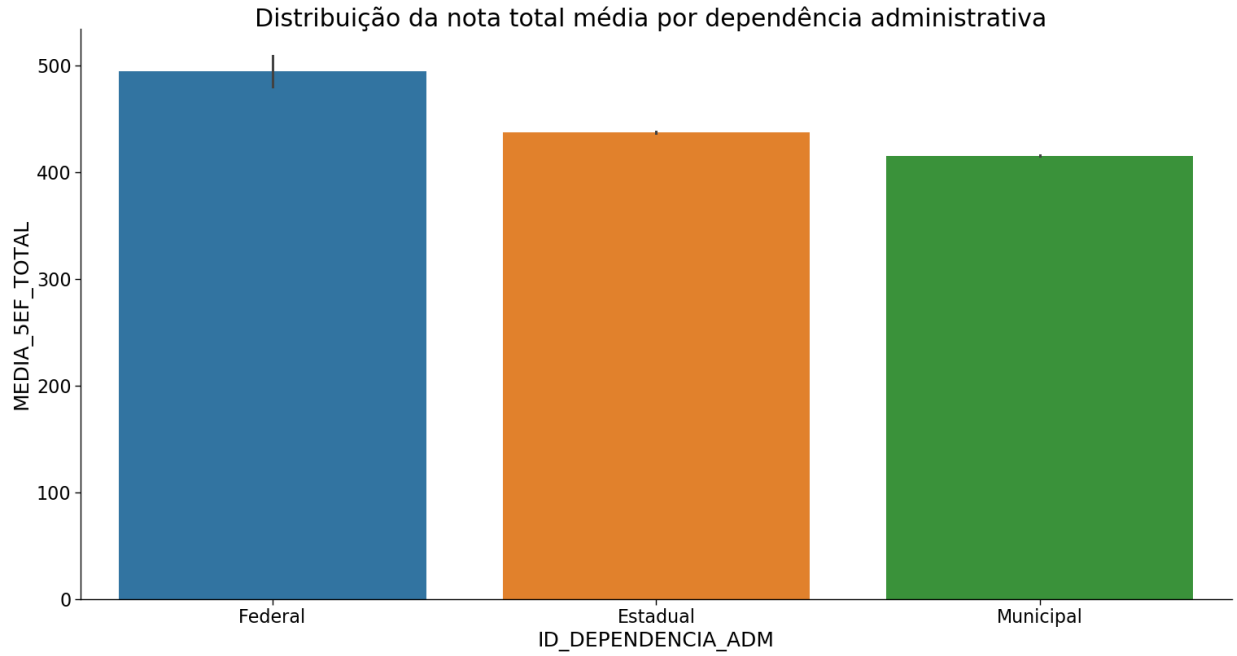
A primeira análise dos dados do Saeb nos mostra que a nota média em português é de 204,82, em matemática é de 214,66. Com base no gráfico de *boxplot* abaixo, as notas apresentam uma distribuição simétrica em torno da mediana, porém, há muitos valores extremos para ambas as variáveis, principalmente para notas maiores. Com isso, pode-se concluir que há varias escolas que se destacam das demais e que há algumas com déficits no ensino de matemática e português.



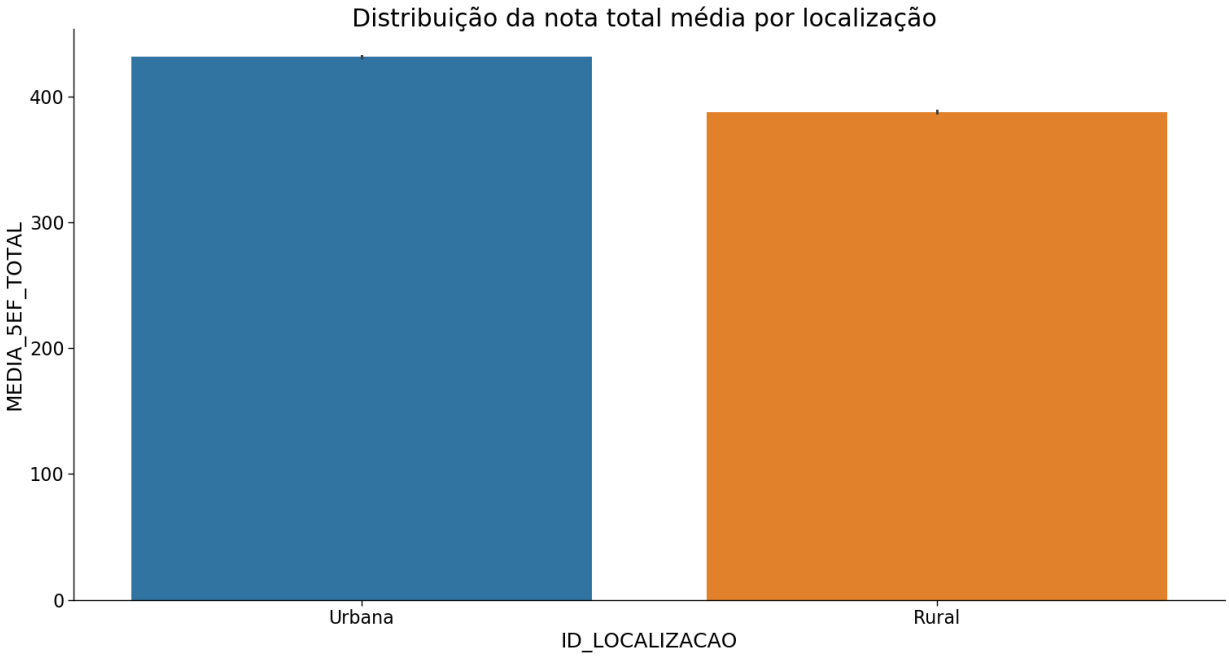
A fim de facilitar a análise das notas, criou-se a variável *media_5ef_total* que é a soma das notas médias em português e matemática das escolas avaliadas pelo Saeb e que será utilizada como *proxy* para a qualidade da educação oferecida nas escolas. Uma primeira análise dessa variável distribuída pelos estados brasileiros demonstra que há uma discrepância na educação oferecida entre as unidades da federação do estado brasileiro, uma vez que a diferença entre a maior nota média e a menor é de aproximadamente 100 pontos, conforme gráfico abaixo. Como os estados brasileiros apresentam uma grande diferença de renda e consequentemente de orçamento público direcionado à educação, o gráfico indica que a diferença na qualidade da educação entre os estados pode estar relacionada aos investimentos em recursos educacionais nas escolas.



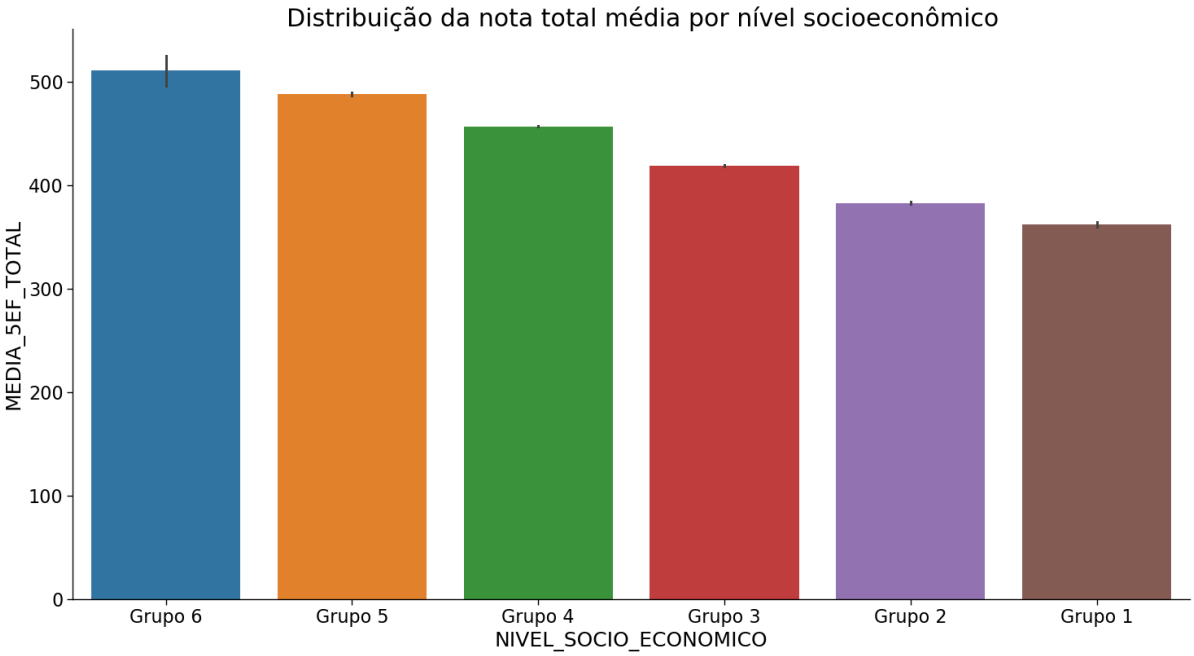
Com relação ao ente federativo que é responsável pela escola, pode-se notar que as escolas federais possuem o melhor desempenho, ao passo que as estaduais são melhores que as municipais. Essa diferença pode estar relacionada à diferença no nível de investimento nas escolas entre os entes da federação, como observado no gráfico abaixo.



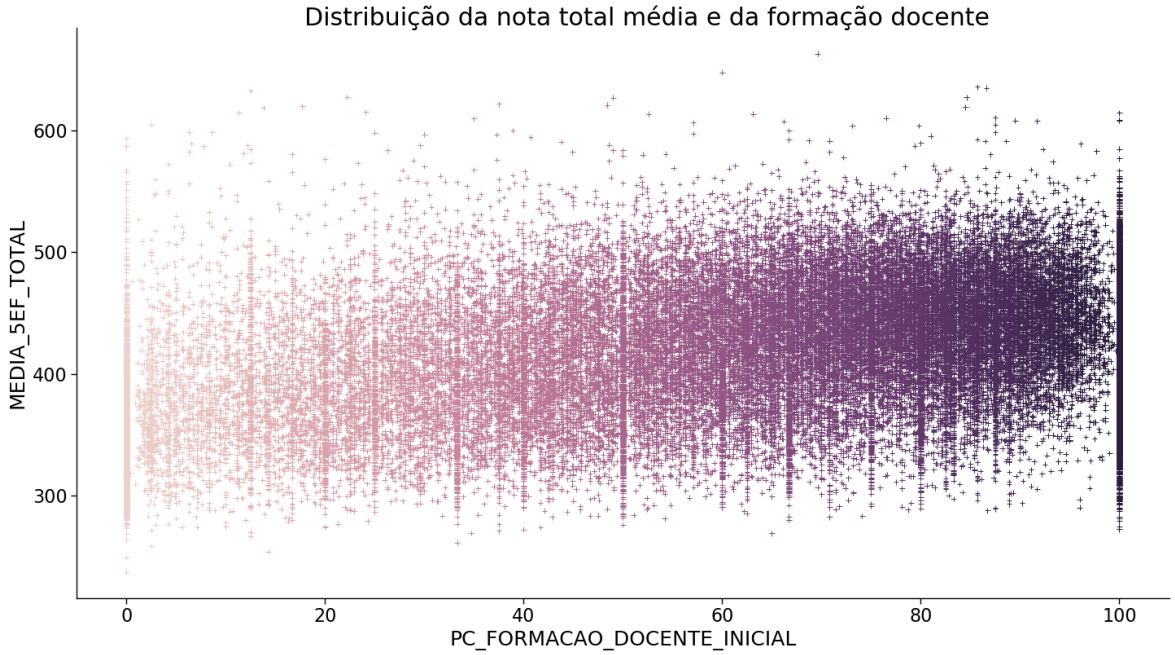
No tocante à localização da escola entre urbana e rural, pode-se notar no gráfico abaixo que as escolas de zona rural possuem uma nota média menor do que as localizadas em áreas urbanas. A diferença pode se dar por precariedade das escolas rurais ou por algum fator relacionado ao ambiente rural, como a distância entre a casa dos alunos e a escola.



A partir do gráfico que mostra a relação entre a nota total média da escola e a classificação socioeconômica predominante entre os seus alunos, demonstra-se que as condições socioeconômicas a que os alunos estão sujeitos tem grande impacto sobre o seu rendimento escolar, uma vez que há uma diferença de aproximadamente 150 pontos entre a categoria de mais baixo nível socioeconômico e a de maior nível.



Por fim, ao analisarmos a correlação entre a nota média das escolas e o percentual de professores da escola que possuem nível superior observa-se uma correlação positiva entre as variáveis, apesar de que os dados possam apresentar algum erro de medida, uma vez que para valores extremos da formação docente a correlação parece não existir.

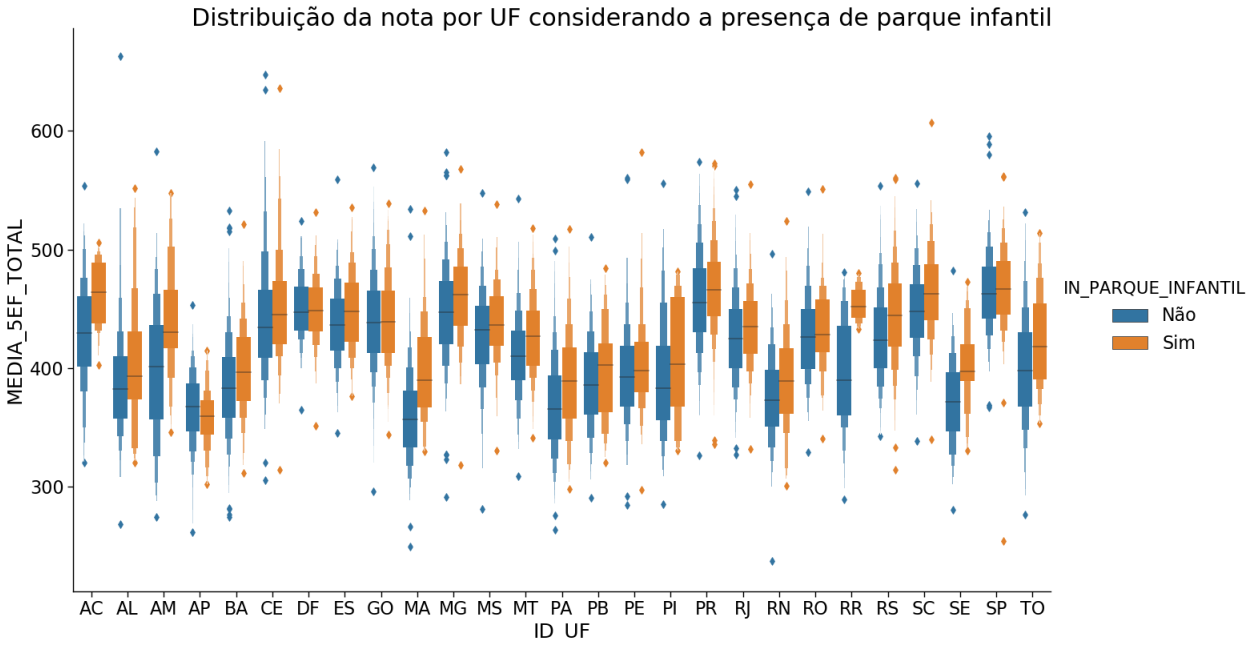


4.2 Análise dos dados do Censo Escolar

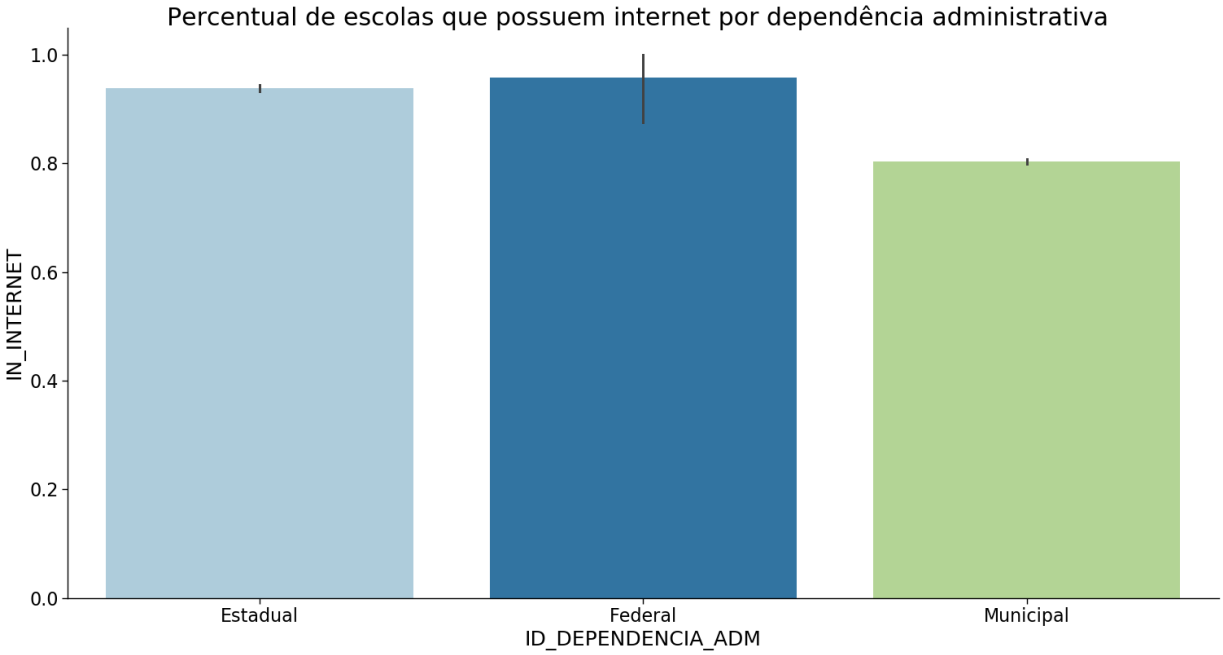
Os dados do Censo Escolar apresentam dificuldade de análise devido à quantidade de variáveis, que somam o total de 166 variáveis. Diante disso, pode-se

constatar que muitas variáveis são informações complementares de outras. Assim, nesse estudo procurou-se manter as variáveis principais para análise, excluindo aquelas que apenas adicionam informações complementares, a fim de simplificar a análise dos dados.

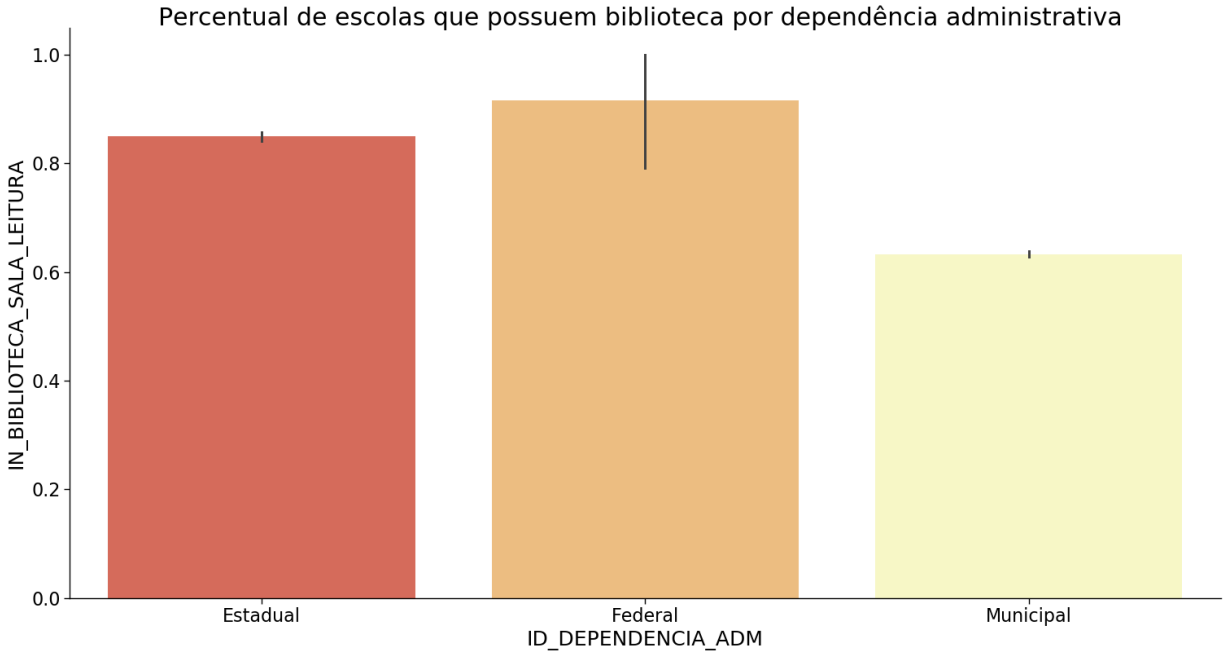
Como identificado na seção anterior, a nota média das escolas pode estar relacionada à presença de recursos educacionais, como pode ser visto no gráfico abaixo, que demonstra que a nota média das escolas dentro de um mesmo estado é maior quando essa escola possui parque infantil. Nesse gráfico fica evidente que a distribuição das notas das escolas tende a ser maior nas escolas que possuem parque infantil.



Quando se analisa a presença de internet nas escola, pode-se notar que as escolas federais possuem maior cobertura, seguidas pelas estaduais e por fim pelas municipais. Essa diferença embasa a hipótese da seção anterior de que a diferença entre os entes federados pode estar relacionada aos níveis de investimento diferentes entre eles, como apresentado no gráfico abaixo.

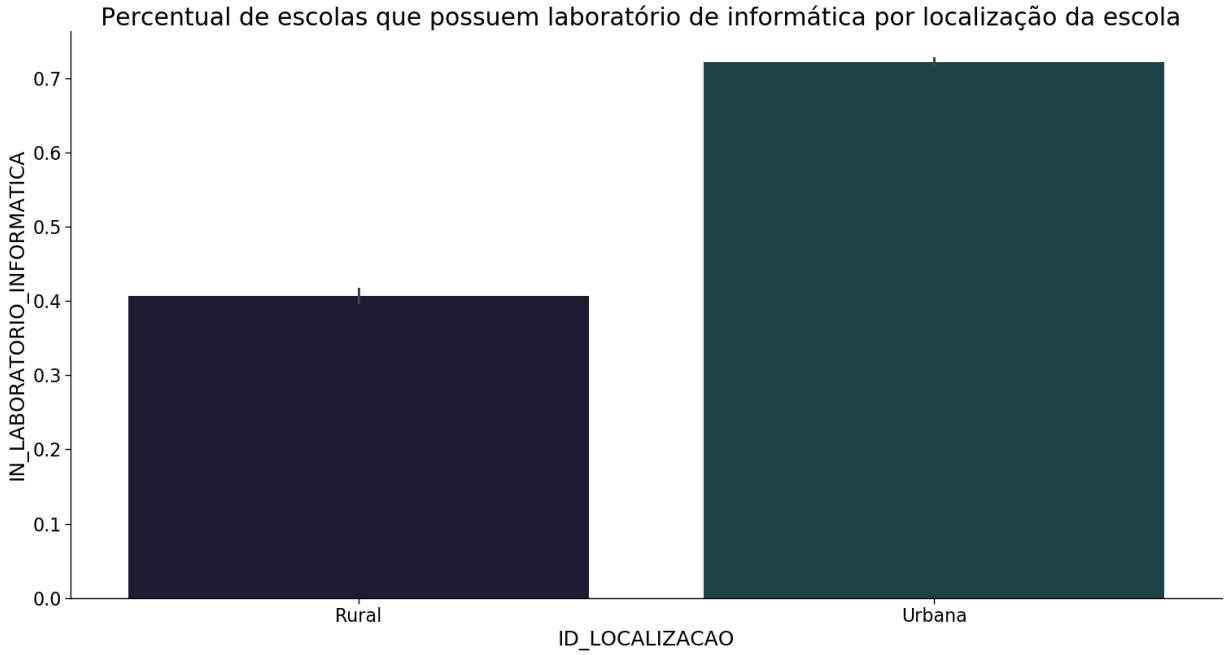


Com relação à presença de bibliotecas ou salas de leitura nas escolas, os dados mostram que as diferenças do nível de investimento nas escolas entre União, estados e municípios se mantêm, conforme gráfico abaixo.

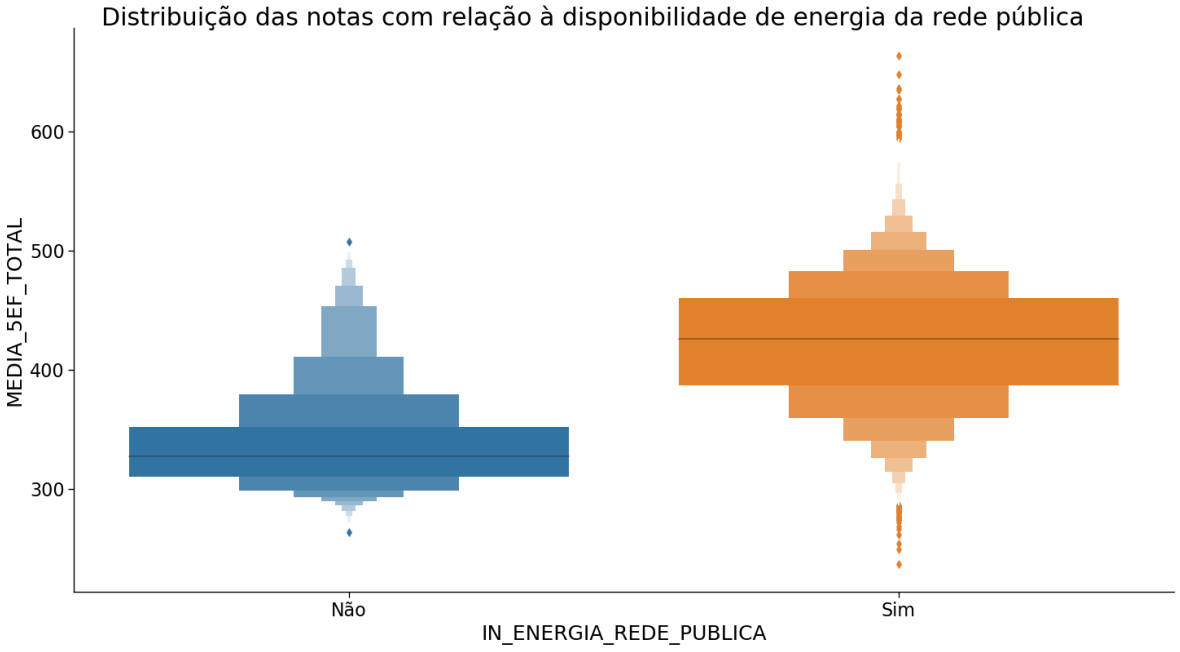


No tocante à diferença de notas entre as escolas da zona rural e urbana, com base nos dados do censo, a diferença pode ser resultado da menor quantidade de recursos disponíveis nas escolas rurais. Como exemplo, o gráfico a seguir mostra a

menor frequência de laboratórios de informática nas escolas rurais, quando comparado com as escolas urbanas.



De forma a demonstrar que a falta de infraestrutura nas escolas também pode impactar na qualidade da educação, com base no gráfico abaixo, observa-se que a nota média é menor nas escolas que não são atendidas pela rede pública de energia.



5. Criação de Modelos de Machine Learning

Utilizou-se no estudo modelos de regressão linear múltipla a fim de verificar se os recursos educacionais apresentam impacto sobre as notas das escolas no Saeb. O estudo procurou demonstrar também o valor do impacto individual de cada variável do censo escolar sobre as notas das escolas, com o objetivo de identificar os recursos que trazem a maior contribuição para a qualidade da educação.

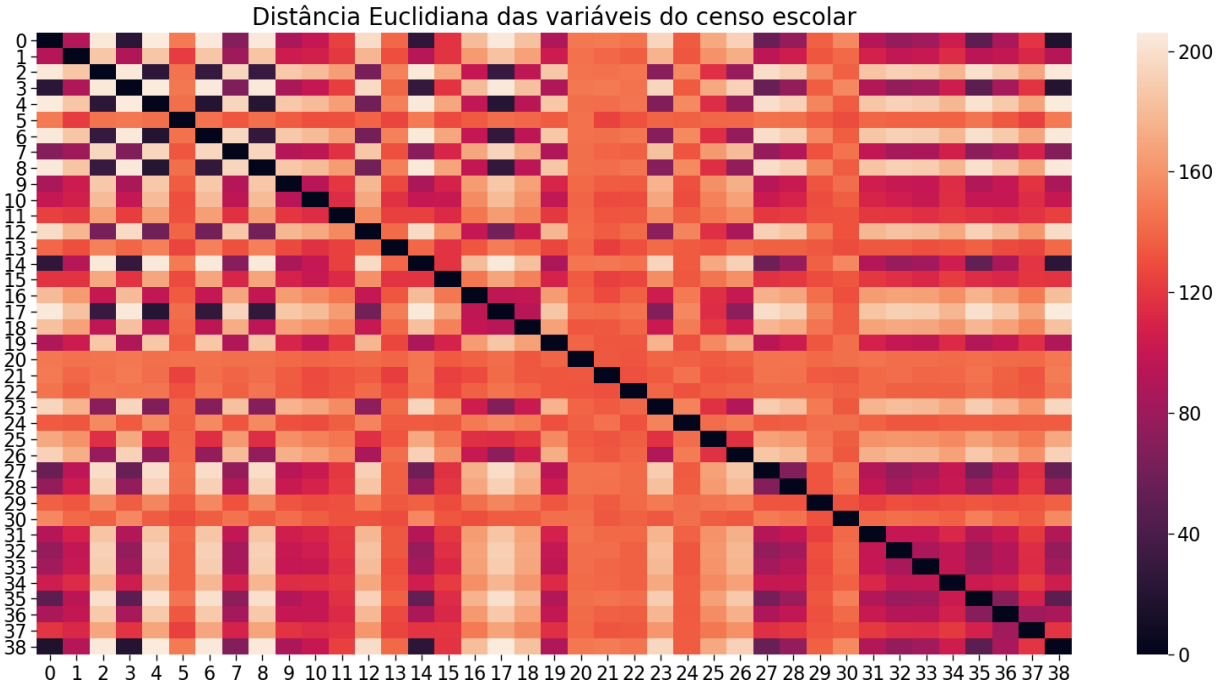
O primeiro modelo desenvolvido utilizou apenas as variáveis disponíveis na base de dados do Saeb. O modelo demonstra que a nota escolar depende do estado em que a escola está localizada, do ente federativo que é responsável pela escola e da localização entre rural e urbana. Esse resultado confirma que a nota das escolas pode estar relacionada ao nível de investimento e recursos educacionais disponíveis nas escolas.

```
#Criar dummys para UF, Dependência Adm e Localização
dummies = pd.get_dummies(saeb[['ID_UF', 'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO']], dro
p_first= True)
saeb_dummies = pd.concat([saeb, dummies], axis= 1)
#Modelo OLS 01
saeb_TOTAL = saeb_dummies[['MEDIA_5EF_TOTAL']]
saeb_exog = saeb_dummies.drop(columns = ['ID_PROVA_BRASIL', 'ID_UF', 'ID_MUNICIPIO',
'ID_ESCOLA', 'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO', 'NIVEL_SOCIO_ECONOMICO', 'NU_MAT
RICULADOS_CENSO_5EF', 'NU_PRESENTES_5EF', 'MEDIA_5EF_LP', 'MEDIA_5EF_MT', 'MEDIA_5EF_
TOTAL'])
saeb_exog = sm.add_constant(saeb_exog, prepend= False)
modelo01 = sm.OLS(saeb_TOTAL, saeb_exog)
resultado01 = modelo01.fit()
resultado01.summary()
```

O segundo modelo calculou o impacto dos recursos educacionais presentes nas escolas sobre as suas respectivas notas. Foram utilizadas 38 variáveis explicativas no modelo. Entretanto, com base na estatística t de significância, nem todas as variáveis se mostraram significativas.

```
censo_escolas = pd.read_csv('pucMinas/dados/ESCOLAS.csv', sep = '|', encoding='latin'
, dtype={"DT_ANO_LETIVO_INICIO": object, "DT_ANO_LETIVO_TERMINO": object})
#Selecionar variáveis de interesse
censo_escolas = censo_escolas.iloc[:,[1, 4, 26, 40, 44, 45, 48, 49, 51, 52, 57, 58, 5
9, 60, 61, 65, 66, 69, 70,
71, 74, 77, 78, 79, 81, 82, 83, 87, 88, 92, 94, 96, 97, 98, 100, 101, 103, 104, 120,
121, 123]]
censo_escolas = censo_escolas[censo_escolas['TP_SITUACAO_FUNCIONAMENTO'] == 1]
saeb_censo = pd.merge(left=saeb, right=censo_escolas, how='inner', left_on='ID_ESCOLA
', right_on ='CO_ENTIDADE', sort=False)
#Modelo OLS 2 - Com dados do Censo Escolar
saeb_TOTAL02 = saeb_censo[['MEDIA_5EF_TOTAL']]
saeb_exog02 = saeb_censo.drop(columns=['NIVEL_SOCIO_ECONOMICO', 'MEDIA_5EF_TOTAL', 'I
D_UF', 'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO', 'PC_FORMACAO_DOCENTE_INICIAL'])
saeb_exog02 = sm.add_constant(saeb_exog02, has_constant= 'add')
modelo02 = sm.OLS(saeb_TOTAL02, saeb_exog02)
resultado02 = modelo02.fit()
resultado02.summary()
```

As variáveis explicativas podem possuir alta correlação entre si, pois muitas variáveis trazem informações semelhantes, como é o caso das variáveis *in_agua_rede_publica* e *in_agua_inexistente. Diante disso*, calculou-se a distância euclidiana entre elas para mensurar o grau de correlação. A escolha pela distância euclidiana se deu pelo fato de que as variáveis explicativas são do tipo *dummy* ou dicotômica. Para visualizar as distâncias euclidianas entre as variáveis utilizou-se um gráfico de calor, conforme abaixo:



A partir da análise das distâncias euclidianas, desenvolveu-se um terceiro modelo. Nesse terceiro modelo foram eliminadas as variáveis que não possuíam significância estatística no modelo anterior e também foram eliminadas as variáveis que possuíam uma baixa distância euclidiana, pois isso indicaria alta correlação entre variáveis dependentes, o que poderia causar viés no modelo. Os resultados mostram que os recursos educacionais disponíveis nas escolas explicam 33% do desempenho escolar, além disso, variáveis de infraestrutura de serviços públicos como *in_energia_rede_publica*, *in_esgoto_rede_publica* e *in_lixo_coleta_periodica* são as que tem maior impacto sobre o rendimento escolar. Outras variáveis como *in_parque_infantil*, *in_*alimentacao e in_*internet* são as que tem maior impacto sobre o desempenho escolar do ensino fundamento I.

```
#Modelo OLS 3 -
 Seleção de dados do Censo Escolar, com base na significância e dist. euclidiana
dist_alto = xpd[xpd<=60]
dist_alto.columns = dados_heat.columns.values
dist_alto.index = dados_heat.columns.values
sns.heatmap(dist_alto, xticklabels=True, yticklabels=True)
plt.show()
saeb_TOTAL03 = saeb_censo[['MEDIA_5EF_TOTAL']]
saeb_exog03 = saeb_censo.drop(columns=['NIVEL_SOCIO_ECONOMICO', 'MEDIA_5EF_TOTAL', 'ID_UF',
 'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO', 'PC_FORMACAO_DOCENTE_INICIAL',
'IN_COZINHA', 'IN_ENERGIA_INEXISTENTE', 'IN_BERCARIO', 'IN_LIXO_OUTROS', 'IN_COMPUTADOR', '
IN_SECRETARIA', 'IN_BANHEIRO_CHUVEIRO',
'IN_LOCAL_FUNC_PREDIO_ESCOLAR', 'IN_AGUA_INEXISTENTE', 'IN_SALA_DIRETORIA', 'IN_LABORATORIO
_CIENCIAS', 'IN_EQUIP_DVD', 'IN_EQUIP_COPIADORA',
'IN_BANHEIRO_EI', 'IN_AUDITORIO'])
saeb_exog03 = sm.add_constant(saeb_exog03, has_constant= 'add')
modelo03 = sm.OLS(saeb_TOTAL03, saeb_exog03)
resultado03 = modelo03.fit()
resultado03.summary()
resultado03.params.sort_values()
```

Um quarto modelo foi construído de modo a identificar se as variáveis disponíveis no Saeb e que foram utilizadas no primeiro são significativas mesmo quando as variáveis do Censo Escolar estão presentes. Uma primeira análise demonstrou que a variável *id_uf* distorce o modelo e não possui mais a significância encontrada no primeiro modelo. Porém, as variáveis *id_dependencia_adm* e *id_localizacao* ainda são significativas para explicar o desempenho escolar. A hipótese que se propõe é que variáveis como administração escolar, que estariam

representadas por *id_dependencia_adm,* e particularidades do ambiente rural impactam o desempenho das escolas.

```
#Modelo OLS 4 - Seleção de dados do Censo Escolar e Dados adicionais do Saeb
dummies_censo = pd.get_dummies(saeb_censo[['ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO']], drop_f
irst= True)
saeb_censo_dummies = pd.concat([saeb_censo, dummies_censo], axis= 1)
saeb_TOTAL04 = saeb_censo_dummies[['MEDIA_5EF_TOTAL']]
saeb_exog04 = saeb_censo_dummies.drop(columns=['NIVEL_SOCIO_ECONOMICO', 'MEDIA_5EF_TOTAL',
'ID_UF', 'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO',
'IN_COZINHA', 'IN_ENERGIA_INEXISTENTE', 'IN_BERCARIO', 'IN_LIXO_OUTROS', 'IN_COMPUTADOR', '
IN_SECRETARIA', 'IN_BANHEIRO_CHUVEIRO',
'IN_LOCAL_FUNC_PREDIO_ESCOLAR', 'IN_AGUA_INEXISTENTE', 'IN_SALA_DIRETORIA', 'IN_LABORATORIO
_CIENCIAS', 'IN_EQUIP_DVD', 'IN_EQUIP_COPIADORA',
'IN_BANHEIRO_EI', 'IN_AUDITORIO'])
saeb_exog04 = sm.add_constant(saeb_exog04, has_constant= 'add')
modelo04 = sm.OLS(saeb_TOTAL04, saeb_exog04)
resultado04 = modelo04.fit()
resultado04.summary()
```

A fim de mensurar o impacto que as condições socioeconômicas dos alunos possuem sobre o desempenho escolar e considerar uma possível correlação dessa variável com os recursos escolares, um quinto modelo foi construído, em que se incluiu a variável *nivel_socio_economico*. Entretanto, nesse modelo foram excluídas 8.314 observações ausentes. O modelo consegue explicar 46,2% da variação do desempenho escolar. Além disso, ele demonstra que as variáveis *id_dependencia_adm* e *id_localizacao* não possuem significância quando as condições socioeconômicas são incluídas ao modelo.

```
#Modelo OLS 5 - Censo Escolar, Dados adicionais e Nível Socioeco. sem dados faltantes
saeb_censo_socio = saeb_censo[saeb_censo.loc[:,'NIVEL_SOCIO_ECONOMICO'].notnull()]
saeb_censo_socio.shape
dummies_censo_socio = pd.get_dummies(saeb_censo_socio[['ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO
', 'NIVEL_SOCIO_ECONOMICO']], drop_first= True)
saeb_censo_dummies_socio = pd.concat([saeb_censo_socio, dummies_censo_socio], axis= 1)
saeb_censo_dummies_socio.shape
saeb_TOTAL05 = saeb_censo_dummies_socio[['MEDIA_5EF_TOTAL']]
saeb_exog05 = saeb_censo_dummies_socio.drop(columns=['NIVEL_SOCIO_ECONOMICO', 'MEDIA_5EF_TOT
AL', 'ID_UF', 'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO',
'IN_COZINHA', 'IN_ENERGIA_INEXISTENTE', 'IN_BERCARIO', 'IN_LIXO_OUTROS', 'IN_COMPUTADOR', 'I
N_SECRETARIA', 'IN_BANHEIRO_CHUVEIRO',
'IN_LOCAL_FUNC_PREDIO_ESCOLAR', 'IN_AGUA_INEXISTENTE', 'IN_SALA_DIRETORIA', 'IN_LABORATORIO_
CIENCIAS', 'IN_EQUIP_DVD', 'IN_EQUIP_COPIADORA',
'IN_BANHEIRO_EI', 'IN_AUDITORIO'])
saeb_exog05 = sm.add_constant(saeb_exog05, has_constant= 'add')
modelo05 = sm.OLS(saeb_TOTAL05, saeb_exog05)
resultado05.summary()
```

Como a quantidade de observações que foram excluídas no modelo anterior é alto, desenvolveu-se um modelo de árvore de decisão para prever as observações ausentes da variável *nivel_socio_economico*. O modelo utiliza as variáveis id*_uf, id_dependencia_adm* e *id_localizacao*, além de utilizar a variável *Produto Interno Bruto per capita*, que mensura o PIB per capita do município em que a escola está localizada. O modelo apresentou um baixo índice de acurácia de 58,87%.

```
#Dados train e test para modelo de previsão de NIVEL_SOCIO_ECONOMICO
saeb_notnull = saeb_pib[(saeb_pib['Produto Interno Bruto per capita\n(R$ 1,00)'].notnull()
& saeb_pib['NIVEL_SOCIO_ECONOMICO'].notnull())]
saeb_null = saeb_pib[saeb_pib['NIVEL_SOCIO_ECONOMICO'].isnull()]
train, test = train_test_split(saeb_notnull, test_size = 0.2)
trainX = train.drop(columns = ['ID_PROVA_BRASIL', 'ID_UF', 'ID_MUNICIPIO', 'ID_ESCOLA',
'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO', 'PC_FORMACAO_DOCENTE_INICIAL',
'NIVEL_SOCIO_ECONOMICO', 'NU_MATRICULADOS_CENSO_5EF',
'NU_PRESENTES_5EF', 'TAXA_PARTICIPACAO_5EF', 'MEDIA_5EF_LP', 'MEDIA_5EF_TOTAL',
'MEDIA_5EF_MT', 'NO_MUNICIPIO', 'NO_UF', 'Chave',
'Nome da Unidade da Federação', 'Nome do Município',])
testX = test.drop(columns = ['ID_PROVA_BRASIL', 'ID_UF', 'ID_MUNICIPIO', 'ID_ESCOLA',
'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO', 'PC_FORMACAO_DOCENTE_INICIAL',
'NIVEL_SOCIO_ECONOMICO', 'NU_MATRICULADOS_CENSO_5EF',
'NU_PRESENTES_5EF', 'TAXA_PARTICIPACAO_5EF', 'MEDIA_5EF_LP', 'MEDIA_5EF_TOTAL',
'MEDIA_5EF_MT', 'NO_MUNICIPIO', 'NO_UF', 'Chave',
'Nome da Unidade da Federação', 'Nome do Município',])
trainY = train[['NIVEL_SOCIO_ECONOMICO']]
testY = test[['NIVEL_SOCIO_ECONOMICO']]
#Modelo e estimativa de NIVEL_SOCIO_ECONOMICO com Decision Tree
arvore = tree.DecisionTreeClassifier()
arvore = arvore.fit(trainX, trainY)
arvore.score(testX, testY)
```

Por fim, com as observações previstas pelo modelo de árvore de decisão, o modelo final utilizou as variáveis do Censo Escolar selecionadas anteriormente e informações das condições socioeconômicas e da formação docente. O modelo explica 44,6% da variação do desempenho escolar, o que demonstra que investimentos nos recursos escolares podem trazer grande impacto para a qualidade da educação. Além disso, o modelo também demonstra que as condições socioeconômicas dos alunos tem forte impacto sobre o desempenho escolar, em que a diferença entre os grupos 1 e 6 é de 79,56 pontos, o que representa mais 25% da nota média. Outra variável que se mostrou relevante no desempenho escolar é a qualificação profissional, pois escolas que possuem todos os professores com formação em ensino superior possuem nota com 19,44 pontos a mais das que não possuem nenhum professor com formação superior.

```
#Modelo OLS 6 Dados do Censo Escolar, Dados adicionais e Nível Socioeconômico COM dados fal
tantes
saeb_predito02 = saeb_predito[['ID_PROVA_BRASIL', 'ID_UF', 'ID_MUNICIPIO', 'ID_ESCOLA', 'ID
_DEPENDENCIA_ADM', 'ID_LOCALIZACAO',
'PC_FORMACAO_DOCENTE_INICIAL', 'NIVEL_SOCIO_ECONOMICO', 'NU_MATRICULADOS_CENSO_5EF', 'NU_PR
ESENTES_5EF', 'TAXA_PARTICIPACAO_5EF',
'MEDIA_5EF_LP', 'MEDIA_5EF_MT', 'MEDIA_5EF_TOTAL']]
saeb_censo_predito = pd.merge(left=saeb_predito02, right=censo_escolas, how='inner', left_o
n='ID_ESCOLA', right_on ='CO_ENTIDADE', sort=False)
saeb_censo_predito = saeb_censo_predito.drop(columns = ['ID_PROVA_BRASIL', 'ID_ESCOLA', 'ID
_MUNICIPIO', 'NU_MATRICULADOS_CENSO_5EF',
'NU_PRESENTES_5EF', 'TAXA_PARTICIPACAO_5EF', 'MEDIA_5EF_LP', 'MEDIA_5EF_MT', 'CO_ENTIDADE',
 'TP_SITUACAO_FUNCIONAMENTO'])
saeb_censo_predito.shape
dummies_censo_socio02 = pd.get_dummies(saeb_censo_predito[['NIVEL_SOCIO_ECONOMICO']], drop_
first= True)
saeb_censo_dummies_pred = pd.concat([saeb_censo_predito, dummies_censo_socio02], axis= 1)
saeb_censo_dummies_pred.shape
saeb_TOTAL06 = saeb_censo_dummies_pred[['MEDIA_5EF_TOTAL']]
saeb_exog06 = saeb_censo_dummies_pred.drop(columns=['NIVEL_SOCIO_ECONOMICO', 'MEDIA_5EF_TOT
AL', 'ID_UF', 'ID_DEPENDENCIA_ADM', 'ID_LOCALIZACAO',
'IN_COZINHA', 'IN_ENERGIA_INEXISTENTE', 'IN_BERCARIO', 'IN_LIXO_OUTROS', 'IN_COMPUTADOR', '
IN_SECRETARIA', 'IN_BANHEIRO_CHUVEIRO',
'IN_LOCAL_FUNC_PREDIO_ESCOLAR', 'IN_AGUA_INEXISTENTE', 'IN_SALA_DIRETORIA', 'IN_LABORATORIO
_CIENCIAS', 'IN_EQUIP_DVD', 'IN_EQUIP_COPIADORA',
'IN_BANHEIRO_EI', 'IN_AUDITORIO'])
saeb_exog06 = sm.add_constant(saeb_exog06, has_constant= 'add')
modelo06 = sm.OLS(saeb_TOTAL06, saeb_exog06)
resultado06 = modelo06.fit()
resultado06.summary()
```

Os parâmetros, estatística e testes com relação ao modelo são apresentados abaixo.

OLS Regression Results

Model:	OLS	Adj. R-squared:	0.446
Method:	Least Squares	F-statistic:	1145.
Date:	Tue, 15 Oct 2019	Prob (F-statistic):	0.00
Time:	16:19:24	Log-Likelihood:	-2.1672e+05
No. Observations:	42630	AIC:	4.335e+05
Df Residuals:	42599	BIC:	4.338e+05
Df Model:	30		

	coef	std err	t	P>|t|	[0.025	0.975]
const	299.5037	5.178	57.838	0.000	289.354	309.653
PC_FORMACAO_DOCENTE_INICIAL	0.1945	0.008	25.635	0.000	0.180	0.209
IN_AGUA_REDE_PUBLICA	2.0532	0.571	3.598	0.000	0.935	3.172
IN_ENERGIA_REDE_PUBLICA	22.2205	2.281	9.740	0.000	17.749	26.692
IN_ESGOTO_REDE_PUBLICA	8.2235	0.459	17.899	0.000	7.323	9.124
IN_ESGOTO_INEXISTENTE	-4.3923	2.014	-2.181	0.029	-8.340	-0.445
IN_LIXO_COLETA_PERIODICA	8.9318	0.776	11.516	0.000	7.412	10.452
IN_SALA_PROFESSOR	2.4385	0.543	4.491	0.000	1.374	3.503
IN_LABORATORIO_INFORMATICA	1.7074	0.449	3.806	0.000	0.828	2.587
IN_QUADRA_ESPORTES	4.6681	0.448	10.419	0.000	3.790	5.546
IN_BIBLIOTECA_SALA_LEITURA	1.8392	0.483	3.808	0.000	0.893	2.786
IN_PARQUE_INFANTIL	4.8512	0.504	9.633	0.000	3.864	5.838
IN_REFEITORIO	0.8525	0.435	1.958	0.050	-0.001	1.706
IN_ALMOXARIFADO	3.3420	0.410	8.153	0.000	2.539	4.145
IN_PATIO_COBERTO	2.2934	0.397	5.780	0.000	1.516	3.071
IN_AREA_VERDE	2.0092	0.430	4.672	0.000	1.166	2.852
IN_LAVANDERIA	3.9488	0.587	6.722	0.000	2.797	5.100
IN_EQUIP_TV	2.2768	0.807	2.823	0.005	0.696	3.858
IN_EQUIP_RETROPROJETOR	-1.3280	0.418	-3.180	0.001	-2.147	-0.509
IN_EQUIP_IMPRESSORA	2.2055	0.508	4.342	0.000	1.210	3.201
IN_EQUIP_SOM	6.0793	0.623	9.760	0.000	4.858	7.300
IN_EQUIP_MULTIMIDIA	4.7531	0.596	7.978	0.000	3.585	5.921
IN_EQUIP_FOTO	4.6439	0.487	9.534	0.000	3.689	5.599
IN_INTERNET	10.4801	0.696	15.065	0.000	9.117	11.844
IN_BANDA_LARGA	-1.5029	0.548	-2.743	0.006	-2.577	-0.429
IN_ALIMENTACAO	12.6679	4.676	2.709	0.007	3.504	21.832
NIVEL_SOCIO_ECONOMICO_Grupo 2	3.9123	0.892	4.388	0.000	2.165	5.660
NIVEL_SOCIO_ECONOMICO_Grupo 3	21.4473	0.933	22.990	0.000	19.619	23.276
NIVEL_SOCIO_ECONOMICO_Grupo 4	49.2542	1.029	47.846	0.000	47.236	51.272
NIVEL_SOCIO_ECONOMICO_Grupo 5	76.2272	1.260	60.491	0.000	73.757	78.697
NIVEL_SOCIO_ECONOMICO_Grupo 6	79.5652	11.321	7.028	0.000	57.376	101.755

Omnibus:	4654.045	Durbin-Watson:	1.172
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9658.771
Skew:	0.695	Prob(JB):	0.00
Kurtosis:	4.873	Cond. No.	4.15e+03

6. Apresentação dos Resultados

Conforme o modelo desenvolvido demonstra, as variáveis externas ao aluno pode ter forte impacto sobre o seu desempenho escolar. Nesse sentido, o estudo procurou identificar quais os recursos escolares que possuem maior impacto sobre o desempenho escolar no ensino fundamental I, a fim de sugerir quais os recursos deveriam ser priorizados, gerando assim uma maior eficiência das políticas públicas para educação em um cenário de restrição orçamentária.

Nesse sentido, o quadro abaixo demonstra as variáveis em ordem de decrescente de impacto sobre as notas escolares.

Variáveis	Impacto
IN_ENERGIA_REDE_PUBLICA	22,22
IN_ALIMENTACAO	12,67
IN_INTERNET	10,48
IN_LIXO_COLETA_PERIODICA	8,93
IN_ESGOTO_REDE_PUBLICA	8,22
IN_EQUIP_SOM	6,08
IN_PARQUE_INFANTIL	4,85
IN_EQUIP_MULTIMIDIA	4,75
IN_QUADRA_ESPORTES	4,67
IN_EQUIP_FOTO	4,64
IN_LAVANDERIA	3,95
IN_ALMOXARIFADO	3,34
IN_SALA_PROFESSOR	2,44
IN_PATIO_COBERTO	2,29
IN_EQUIP_TV	2,28
IN_EQUIP_IMPRESSORA	2,21
IN_AGUA_REDE_PUBLICA	2,05
IN_AREA_VERDE	2,01
IN_BIBLIOTECA_SALA_LEITURA	1,84
IN_LABORATORIO_INFORMATICA	1,71
IN_REFEITORIO	0,85

7. Links

Vídeo de Apresentação

https://github.com/herlisjunior/pucMinas/video

Scripts e base de dados

https://github.com/herlisjunior/pucMinas