Aprendizado de Máquina Aplicado a Problemas

Educação

Authors:

Vítor Augusto Paiva de Brito

N° USP :

13732303

2024

Contents

1 Introdução

Neste documento, foi feita a análise de correlação entre os atributos de uma base de dados de taxas educacionais e de *IDH*. Para extrair informação do conjunto de dados, os atributos serão segmentados por nível de escolaridade, ou seja, dados do ensino médio, fundamental e infantil serão analisados de forma a se obter uma relação entre os próprios conjuntos e entre os demais atributos, incluindo o atributo alvo.

1.1 Tratamento de valores faltantes

Dado que existem valores numéricos faltantes de certos atributos na base, faz-se a média aritmética dos valores da mesma coluna dos objetos pertencentes ao mesmo estado. Assim, é feita uma aproximação coerente do valor real para análise.

2 Resultados

2.1 Análise geral do conjunto

Analisando o conjunto de dados de maneira holística na Figura 1 é possível observar a distribuição dos objetos do conjunto de acordo com o IDH municipal.

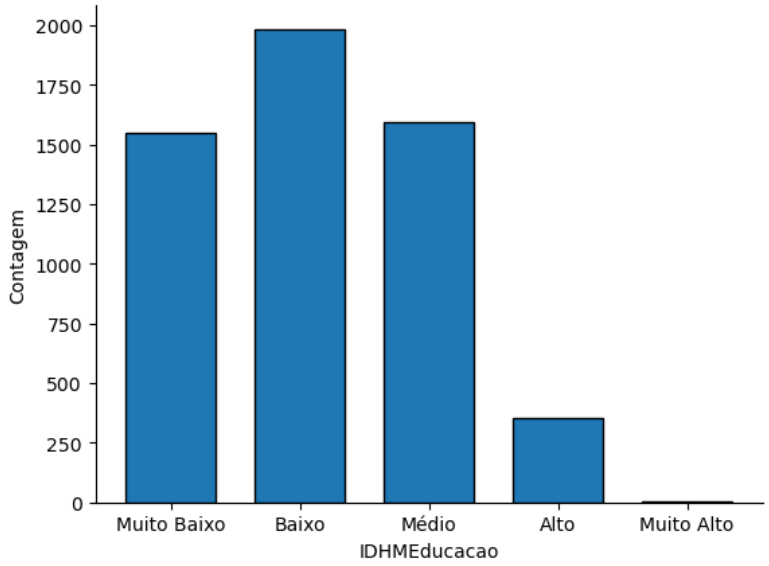

Figure 1: Histograma dos valores de IDH.

Desse modo, infere-se que há uma concentração significativa de registros nas categorias inferiores do índice, portanto, vê-se que os municípios, em geral, não têm um IDH bom.

Para tornar a análise menos complexa, dado que a quantidade de objetos com atributo alvo da classe 'Muito Alto' é da ordem de unidades, os municípios dessa categoria serão integrados à classe 'Alto', sendo analisado conjuntamente.

2.2 Análise geográfica do IDH dos municípios.

A fim de analisar como o IDH dos municípios de cada região estão distribuídos, faz-se uso do gráfico de barras da Figura 2.

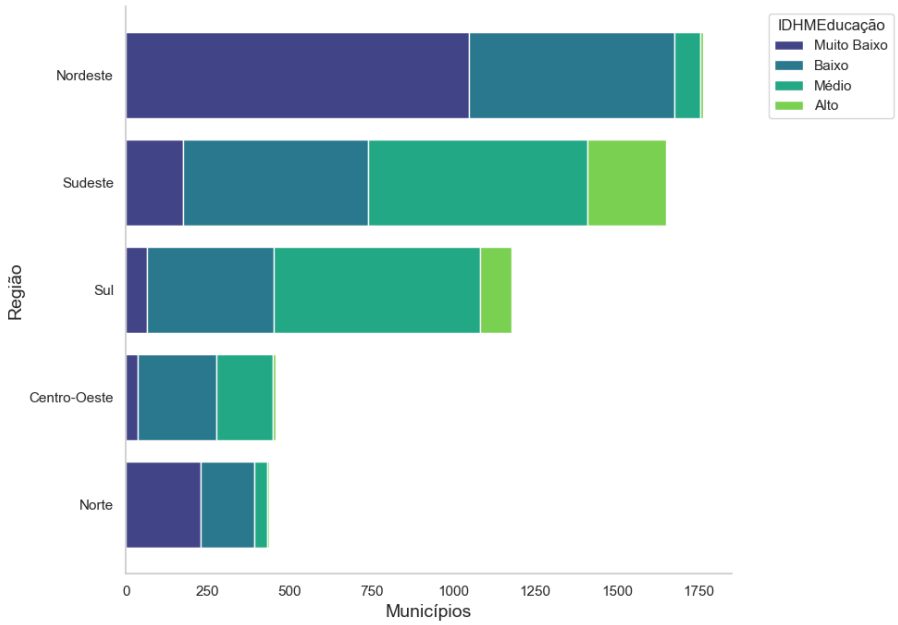

Figure 2: Gráfico de barras da quantidade de municípios com a porcentagem de registros pertencentes a cada classe de IDH.

Segundo a ilustração, infere-se que, devido à maior quantidade e à maior proporção de municípios com IDH abaixo da média, os objetos da região Nordeste representam a maior parte dos registros com IDH muito baixo ou baixo. Assim, possíveis políticas públicas podem ser mais efetivas caso aplicadas com enfoque nesses municípios.

Ademais, vê-se que a maior parte dos municípios com IDH na média ou acima são das regiões Sul e Sudeste, fator que pode estar relacionado ao elevados índice educacionais das regiões. Além disso, observa-se que as regiões Centro-Oeste e Norte apresentam mais da metade dos municípios com IDH abaixo da média, o que exige uma dose de atenção por parte do estado e da comunidade para amenizar tais indicadores.

Posteriormente, para analisar a representação relativa de cada região nas categorias do atributo alvo, utiliza-se o gráfico de barras da Figura 3.

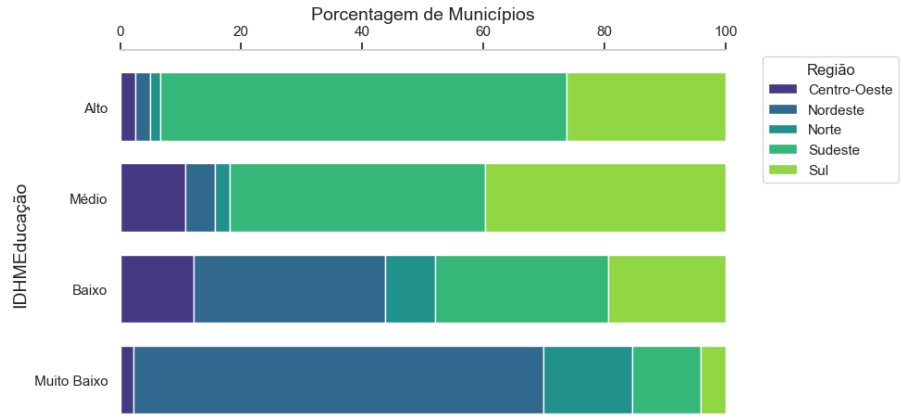

Figure 3: Gráfico de barras de participação relativa das regiões nas categorias de IDH.

Com base na figura, reitera-se a análise de que grande parte da quota de municípios com IDH na média ou acima pertence às regiões Sul e Sudeste e que os objetos com indicador abaixo da média é do Nordeste, com significante parcela da região Centro-Oeste nas categorias 'Baixo' e 'Médio' e com participação considerável de registros da região Norte entre os indicadores alvo mais baixos.

2.3 Análise relacional entre os atributos preditivos.

2.3.1 Mapas de calor.

Primeiramente, a fim de analisar a correlação dos atributos numéricos, lança-se mão da tabela de correlação, analisada com dois métodos distintos, com coeficiente de Pearson e de Kendall, e do mapa de calor correspondente, disponibilizados nas Figuras 4 e 5.

O atributo alvo foi transformado em uma ordem numérica de 1 a 5, de 'Muito Baixo' para 'Alto' para que a correlação entre a coluna e o restante das variáveis pudesse ser calculada.

A partir da imagem da correlação de Pearson, é possível inferir que o atributo alvo tem alta dependência com a taxa de disfunção idade-série e com as taxas referentes ao ensino fundamental, como aprovação, reprovação e evasão, fator que evidencia uma alta relação desses fatores com o IDH e suas importâncias. Além disso, em relação à taxa de disfunção idade-série, vê-se uma relação consideravelmente alta com as taxas do ensino fundamental.

Consonante à correlação com coeficiente de Kendall, é possível observar relações análogas à análise de Pearson, reiterando as dependências entre os atributos do conjunto de dados.

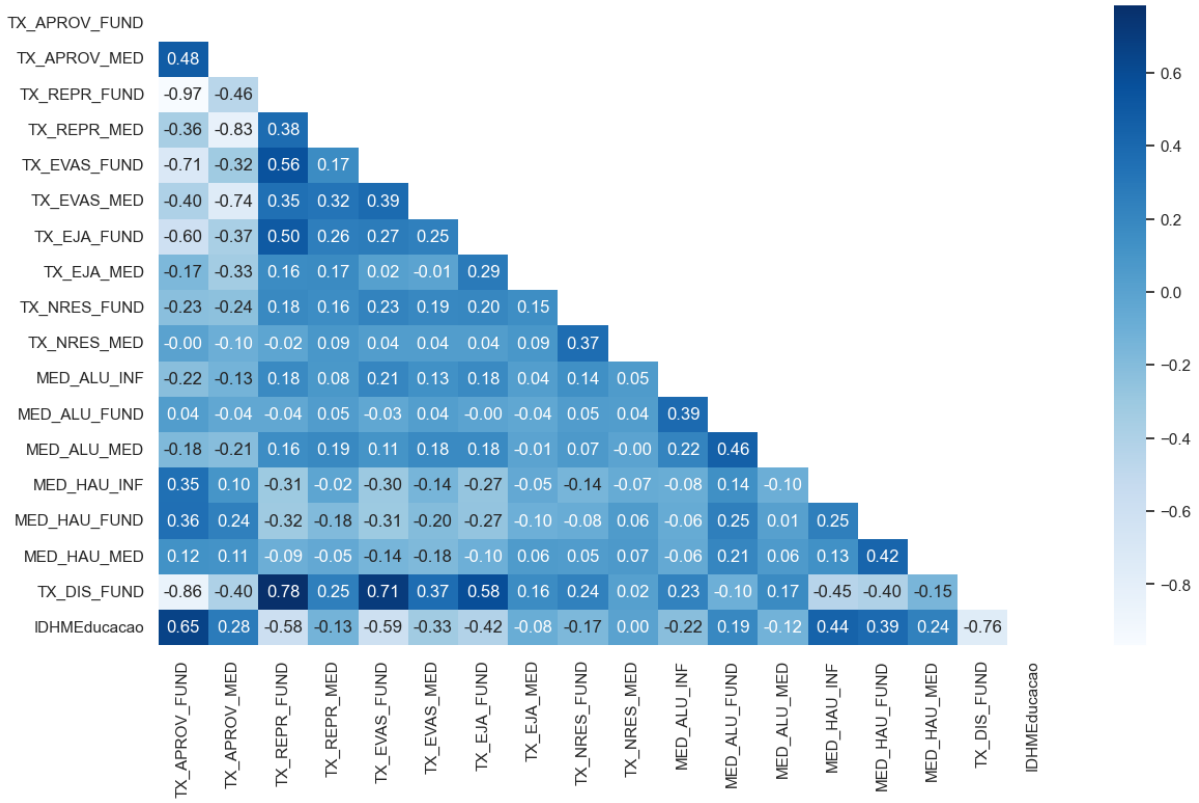

Figure 4: Mapas de calor das correlações de Pearson dos atributos.

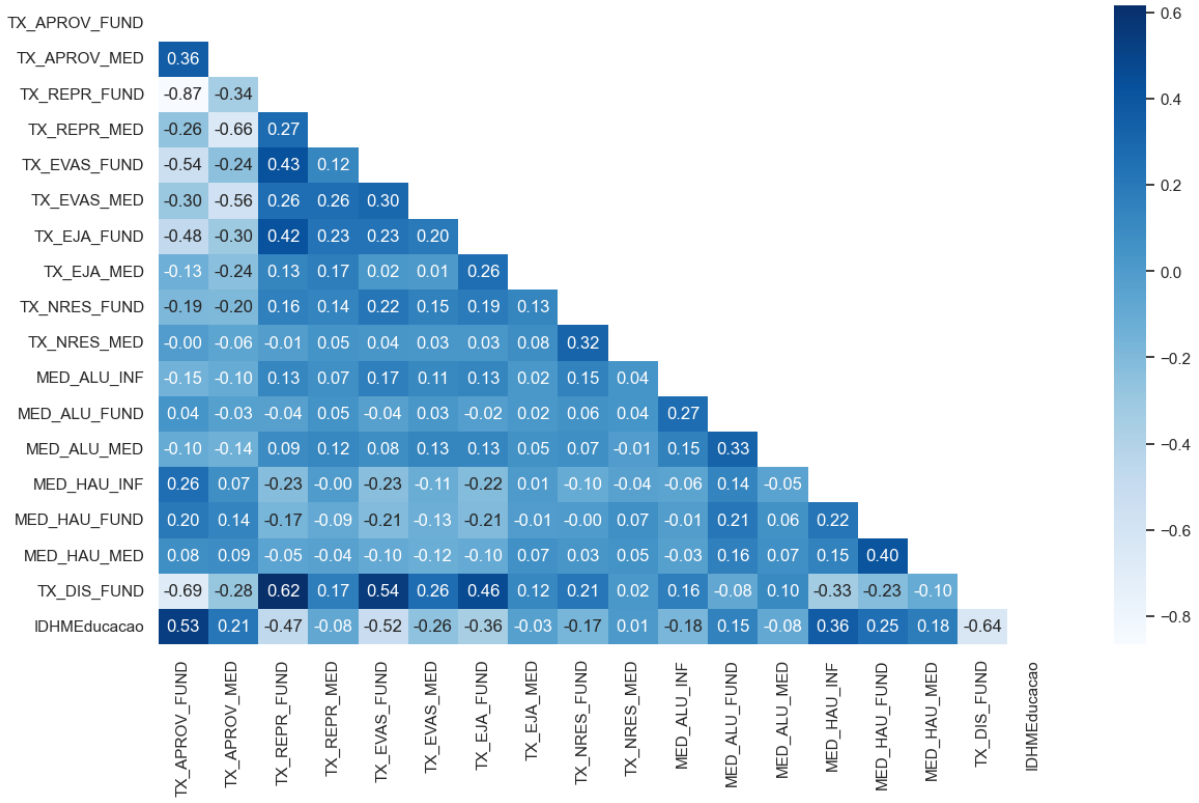

Figure 5: Mapas de calor das correlações de Kendall dos atributos.

2.3.2 Gráficos de dispersão

Primeiramente, observa-se os mapas de dispersão das taxas referentes ao ensino médio e ao ensino fundamental nas Figuras 6 e 7.

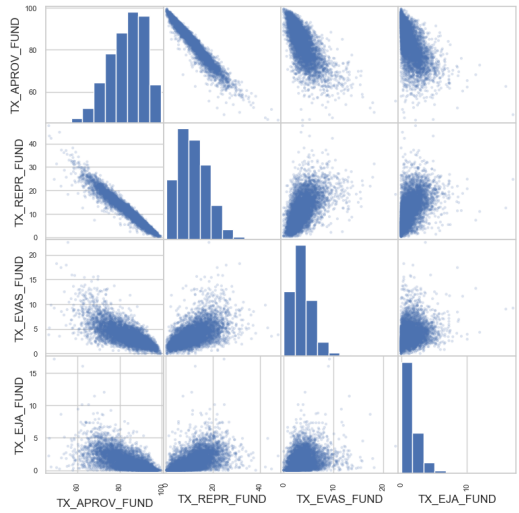

Figure 6: Gráficos de dispersão e histogramas das taxas do ensino fundamental.

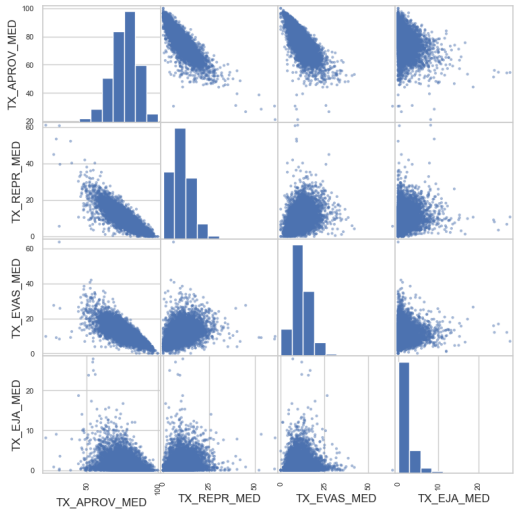

Figure 7: Gráficos de dispersão e histogramas das taxas do ensino médio.

Ao se considerar os mapas de dispersão dos atributos dos níveis de ensino fundamental e médio, infere-se que, principalmente entre os atributos do ensino médio, verifica-se uma fraca correlação linear entre os atributos enquanto os atributos do ensino fundamental se mostram ligeiramente mais relacionados, principalmente nos atributos de taxa de promoção e de repetência. Assim, além dos dois primeiros atributos, conclui-se que os atributos de cada nível não estão fortemente correlacionados entre si.

Subsequentemente, a fim de se verificar a relação entre os atributos dos níveis médio e fundamental com a taxa de distorção respectivas, lança-se mão das Figuras 8 e 9.

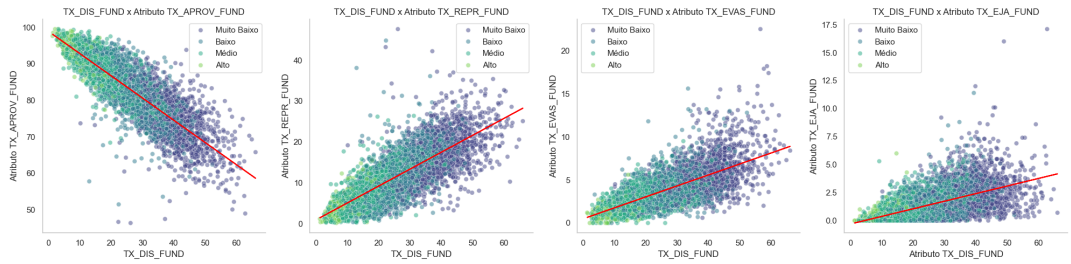

Figure 8: Relação das taxas do ensino fundamental da base de dados com a taxa de distorção idade-série no ensino fundamental.

Assim, vê-se que o ensino fundamental demonstrou relação mais forte das taxas com a distorção.

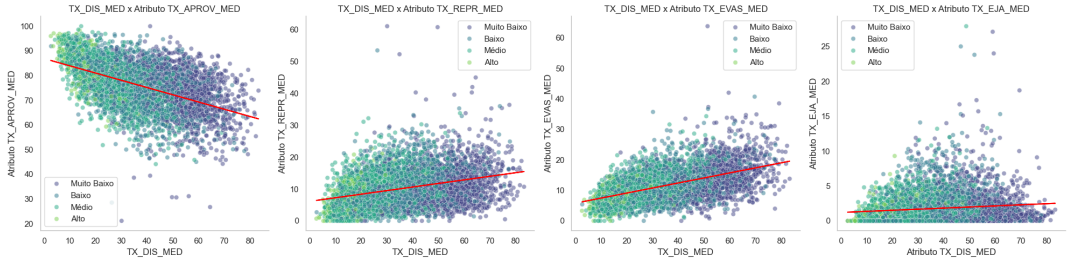

Figure 9: Relação das taxas do ensino médio da base de dados com a taxa de distorção idade-série no ensino médio.

A fim de se verificar a hipótese de que a média de alunos e de aula-hora no ensino infantil impacta as etapas posteriores de educação, verifica-se os mapas de dispersão em função das taxas de distorção de idade nas Figuras 10 e 11.

Figure 10: Relação entre a média de alunos no ensino infantil com a distorção idade-série no ensino fundamental.

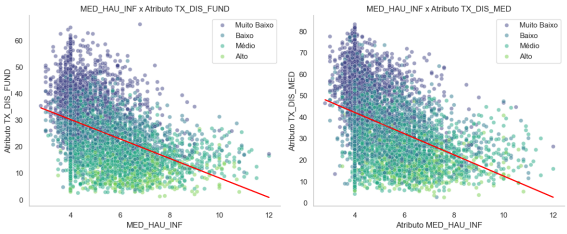

Figure 11: Relação entre a média de horas-aula no ensino infantil com a distorção idade-série no ensino fundamental..

Após a visualização, é razoável concluir que a relação é fraca pelo comportamento pouco compátivel com a linearidade, demonstrando, assim, pouca proporção causa-consequência dos atributos.

2.4 Análise relacional entre os atributos preditores e o atributo alvo.

2.4.1 Boxplots.

A priori, com intuito de verificar a correlação do atributo alvo com as quatro taxas iniciais de cada nível de ensino, faz-se uso do *boxplot* nas Figuras 12 e 13.

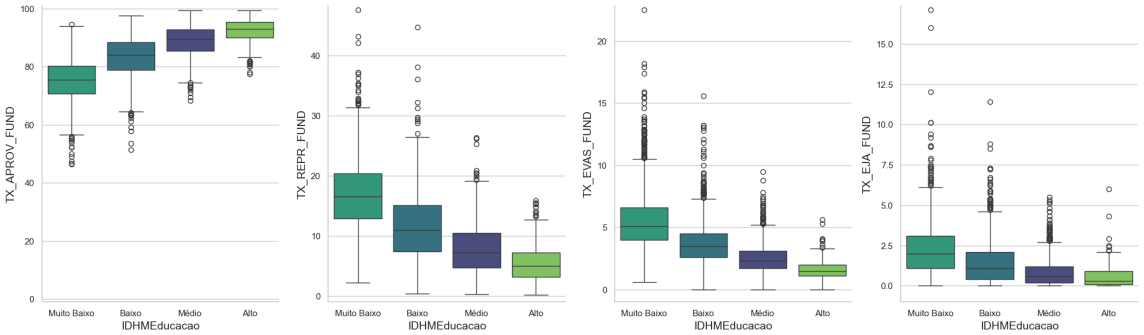

Figure 12: Boxplots das taxas relacionadas ao ensino fundamental em cada categoria de IDH.

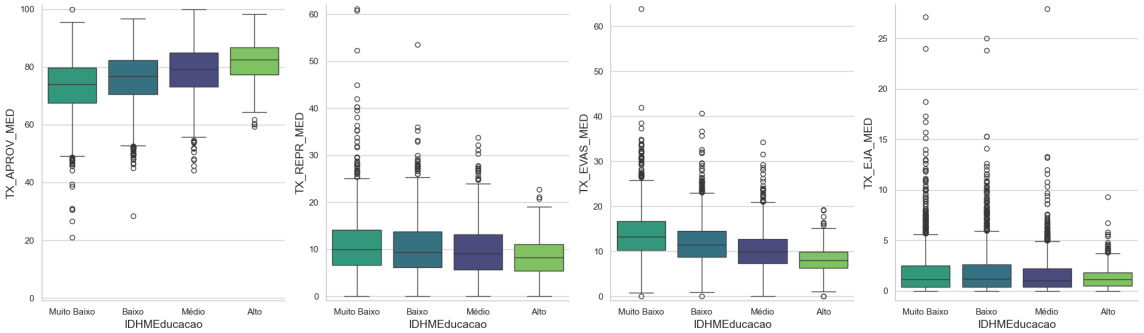

Figure 13: Boxplots das taxas relacionadas ao ensino médio em cada categoria de IDH.

Após visualização, infere-se que o *IDH* de um município tem relação mais próxima com as taxas do ensino fundamental em comparação ao ensino médio graças à maior proporcionalidade entre os atributos e menos inconsistência, o que legitima o mapa de calor abordado no tópico 2.3.1.

Após, a fim de se verificar a relação das médias de alunos e de horas-aula, Figuras 14 e 15, respectivamente, e do atributo alvo, vê-se que ambos não apresentam relação forte com o *IDH*, porém, é razoável concluir que a média de horas-aula é mais relevante para o *IDH* em comparação à média de alunos por turma nos três níveis de ensino, sendo negativo em relação á média de alunos, exceto no ensino fundamental, e positiva em relação às horas-aula.

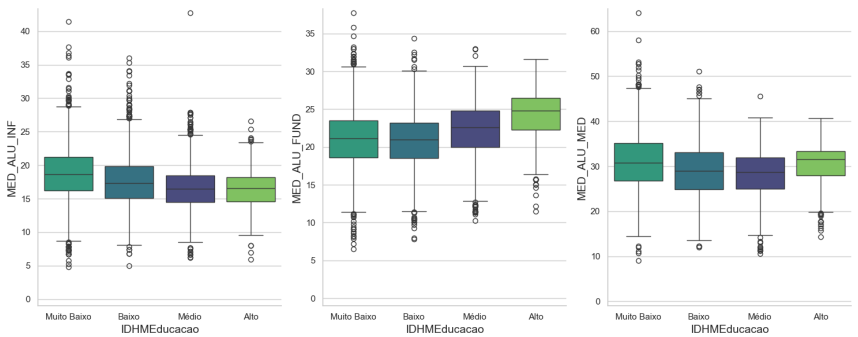

Figure 14: Boxplots da média de alunos por turma em relação a cada categoria de IDH.

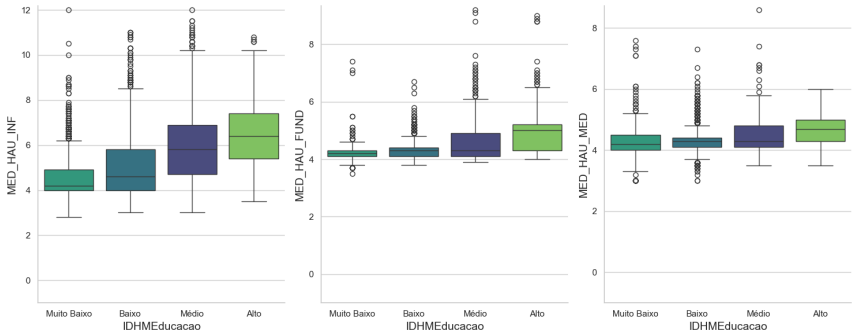

Figure 15: Boxplots da média de horas-aula em relação a cada categoria de IDH.

Por fim, verifica-se a relação entre a taxa de distorção idade-série e o *IDH* do município na Figura 16.

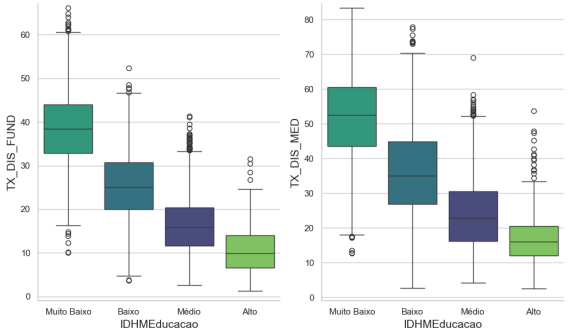

Figure 16: Boxplots das taxas de distorção idade-série em relação a cada categoria de IDH.

Assim, confirmando o fato representado nos mapas de calor no tópico 2.3.1, infere-se que ambas as taxas de distorção idade-série, tanto no ensino fundamental, quanto no ensino médio, tem forte correlação negativa com o atributo alvo.

3 Conclusão

Portanto, foi possível inferir que, em um conjunto de dados desbalanceado, em que as classes com *IDH* menor tem número significativamente superior de representantes.

Em relação à análise geográfica, observa-se que a região Nordeste se evidencia como a mais carente em função do atributo, seguida das regões Norte e Centro-Oeste. Opostamente, as regiões Sudeste e Sul apresentaram uma proporção maior de municípios com IDH médio ou acima.

Na questão das correlações dos atributos preditivos, percebe-se que as quatro taxas iniciais dos dois níveis de ensino, as quais não são fortemente correlacionadas entre si, são relevantes

para a taxa de distorção idade-série, a qual se configurou como atributo preditivo fortemente relacionado ao atributo alvo, além de que, as taxas que tangem à educação infantil não apresentaram tanta relevância para estimar a taxa de distorção idade-série. Além disso, nos boxplots, foi possível concluir que as taxas do par de níveis escolares, fundamental e médio, são ambas significantes para estimar a categoria de *IDH* de um município, dando ênfase às taxas do ensino fundamental. Ademais, as médias de alunos e de horas-aula demonstraram ligeira relevância de predição em comparação às taxas de promoção, repetência, evasão e migração, porém a média de horas-aula possui relação mais próxima à linearidade. Além disso, viu-se que as taxas de distorção idade-série tem elevada correlação com o atributo alvo. Por fim, é razoável enfatizar que as relações ilustradas no mapa de calor foram legitimadas.