Projeto: O quanto experiências de vulnerabilidade incidem sobre a perda do desempenho acadêmico da matemática?

Autor: Henrique Augusto Torres Simplício

Setembro

2021

I - Exposição do problema

No mundo globalizado, o desempenho matemático tem-se mostrado uma variável de fundamental importância para o bem estar do indivíduo. Os benefícios em conhecer matemática ao longo da formação básica passam por diferentes áreas de modo que o aprendizado na disciplina é capaz de produzir maiores oportunidades no mercado de trabalho, ganho salarial e mesmo menor associação com psicopatologias.

Alguns estudos demonstram os riscos do baixo desempenho acadêmico na disciplina. Baixos conhecimentos matemáticos reduzem as oportunidades de emprego e o progresso no mercado de trabalho (Parsons & Binner, 1997). Ao analisar a renda direta na vida adulta, estima-se que 10% na melhoria das notas da matemática durante a educação básica possa incidir em 4,6% no incremento da renda futura do indivíduo (Cury & Filho, 2014). Outros resultados, obtidos através de estudos longitudinais, apontam também para a mesma relação entre o conhecimento da disciplina e o incremento futuro da renda (Ritchie & Bates, 2013). Ao compararmos os efeitos da baixa proficiência em numeracia com o de outras áreas (como a literacia), é possível encontrar evidências apontando para prejuízos ainda maiores decorrentes da perda do desempenho matemático, atuando em aspectos da vida como: auto-estima, conseguir um trabalho em tempo integral com maior nível de complexidade, e mesmo engajamento em temas como a política ou o voto (Parsons & Binner, 2005).

Estes resultados chamam a atenção para os possíveis efeitos capazes de reduzir o aprendizado da matemática. Conforme literatura especializada, experiências adversas e de vulnerabilidade estão associadas com uma série de limitações no desenvolvimento (Hughes, et. al. 2017). Alguns resultados apontam para a interação entre estas experiências de violência e/ou vulnerabilidade - como deter menores condições econômicas, morar na zona rural, e realizar trabalho infantil - e o prejuízo no desempenho acadêmico (Andrade & Laros, 2007; Berthelot, 2001; Palermo, 2014).

Com intuito de avaliar parte deste problema, o presente projeto tem por intuito analisar como experiências de vulnerabilidade podem incidir sobre o desempenho acadêmico da matemática. Considerando a variabilidade individual, econômica e cultural das pessoas que habitam o território nacional, seria possível levantar em que medida variáveis externas ligadas à experiências e aspectos de vulnerabilidade poderiam promover prejuízos no desempenho da matemática?

Para avaliar as experiências de vulnerabilidade analisaremos variáveis ligadas à disposição familiar como escolaridade materna (avaliando o nível de instrução), presença ou ausência da figura paterna, incentivo e apoio famíliar aos estudos, além de proxys de renda e acesso

a infraestrutura como trabalho infantil, morar em rua com acesso a iluminação, ter acesso a computador, água tratada, rua pavimentada, dentre outras.

O objetivo é de identificar como estas variáveis categóricas poderiam interferir no desempenho acadêmico da matemática (variável escalar).

No quadro a seguir, descrevo as Features e Targets com seus respectivos nomes nas colunas do banco de dados. Ao todo, detemos 16 Features para o target analisado.

Features (VI)				Target(VD)
Família	Escolaridade	• Escolaridade da Mãe • Escolaridade do Pai	RECOD_TX_RESP_Q004 RECOD_TX_RESP_Q005	Desempenho em matemática (PROFICIENCIA_MT_SAEB)
	Incentivo ao estudo/ apoio familiar aos estudos	• Conversar com o filho sobre o que acontece na escola • Incentivar a estudar • Incentivar:Fazer tarefa de casa • Incentivar:Comparecer às aulas • Ir às reuniões de pais na escola • Com que Idade que você entrou na escola	RECOD_TX_RESP_Q006A RECOD_TX_RESP_Q006B RECOD_TX_RESP_Q006C RECOD_TX_RESP_Q006D RECOD_TX_RESP_Q006E RECOD_TX_RESP_Q013	
	Presença familiar	• Mora com pai ou Padrasto • Mãe ou madrasta	RECOD_TX_RESP_Q003A RECOD_TX_RESP_Q003B	
SES Infraestrutura	Renda	• Com que frequência a família paga alguém para fazer faxina dentro de casa • Trabalho infantil • Carro	RECOD_TX_RESP_Q007 RECOD_TX_RESP_Q017E RECOD_TX_RESP_Q009g	
	Infraestrutura	• Rua pavimentada (asfalto ou calçamento). • Água tratada da rua. • Iluminação na rua. • Computador (ou notebook).	RECOD_TX_RESP_Q008A RECOD_TX_RESP_Q008B RECOD_TX_RESP_Q008C RECOD_TX_RESP_Q009c	

II - Coleta ou Importação dos dados

Com intuito de medir fatores capazes de interferir no desempenho da matemática, avaliamos o principal instrumento de aferição dos resultados da educação básica nacional: o Sistema de Avaliação da Educação Básica, (SAEB). A prova do Saeb (que leva o mesmo nome do sistema) foi criada ainda nos anos 90 com intuito de fornecer informações para implementar uma política de avaliação nacional da educação. Ao longo dos anos, esta avaliação foi passando por mudanças, sendo hoje aplicada anualmente em todo país.

O SAEB fornece resultados através de uma amostra significativa da educação nacional.

Como critério de avaliação, escolhemos os dados da prova de matemática do quinto ano do ensino fundamental. Estas informações foram coletadas através de dados públicos obtidos através do site governamental do INEP:

(https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb)

Desta forma, este projeto realizará análises através de dados secundários.

III - Preparação dos dados

Para preparação e análise dos dados foram usados 2 softwares de análise estatísticas: o python versão 3.0 (uso majoritário), através do JupyterLab, e o SPSS (Statistical Package for the Social Sciences).

O uso destas duas ferramentas de programação e análise de dados se dá com intuito de aproveitar o melhor de cada uma delas, analisando resultados através tanto de técnicas paramétricas como não paramétricas.
No Python, as principais bibliotecas para análise dos dados, visualização e produção de gráficos, além de técnicas e aprendizado de máquinas foram as apresentadas na imagem abaixo.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import r2_score
import statsmodels.api as sm
import warnings
from scipy import stats
from yellowbrick.regressor import CooksDistance
from yellowbrick.datasets import load_concrete
from sklearn.linear_model import LinearRegression
from yellowbrick.regressor import ResidualsPlot
import shap
shap.initjs()
from sklearn.ensemble import RandomForestClassifier
import scipy.stats as stats
from scipy.special import kolmogorov
from scipy.stats import kstwobign
import seaborn as sns
```

Para preparação dos resultados, foram recodificadas algumas variáveis do banco original transformando-as em numéricas para que as análises pudessem ser realizadas no SPSS.
O banco de dados do SAEB também apresentava muitos dados faltantes de variáveis que eram enquadradas pelo python como NaN. Na figura abaixo, através da função *isnull( )* é possível identificar à quantidade de valores faltantes em cada uma das features.

```
[472]: df.isnull().sum()

[472]: ID_UF                          0
       RECOD_TX_RESP_Q004       2057144
       RECOD_TX_RESP_Q005       2188750
       RECOD_TX_RESP_Q006A       982396
       RECOD_TX_RESP_Q006B       943730
       RECOD_TX_RESP_Q006C       952940
       RECOD_TX_RESP_Q006D       972936
       RECOD_TX_RESP_Q006E       978896
       RECOD_TX_RESP_Q013        644569
       RECOD_TX_RESP_Q003A       957897
       RECOD_TX_RESP_Q003B      1004148
       RECOD_TX_RESP_Q007        930647
       RECOD_TX_RESP_Q014        678874
       RECOD_TX_RESP_Q017E       949346
       RECOD_TX_RESP_Q008A       930910
       RECOD_TX_RESP_Q008B       958253
       RECOD_TX_RESP_Q008C       943302
       RECOD_TX_RESP_Q009c      1002908
       RECOD_TX_RESP_Q009g       922318
       PROFICIENCIA_MT_SAEB      773863
       dtype: int64
```

Estas faltas se apresentam como um problema na medida em que impossibilitam algumas técnicas de machine learning que serão usadas no *Python* posteriormente.

Desta forma, elas foram eliminadas do banco de dados através da função *dropna( )*. A imagem abaixo demonstra a descrição de cada uma das variáveis com seu respectivo N padronizado. Através dela, é possível observar o valor total da amostra (n = 378700)

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 378700 entries, 225703 to 2061123
Data columns (total 20 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   ID_UF                 378700 non-null  float64
 1   RECOD_TX_RESP_Q004    378700 non-null  category
 2   RECOD_TX_RESP_Q005    378700 non-null  category
 3   RECOD_TX_RESP_Q006A   378700 non-null  category
 4   RECOD_TX_RESP_Q006B   378700 non-null  category
 5   RECOD_TX_RESP_Q006C   378700 non-null  category
 6   RECOD_TX_RESP_Q006D   378700 non-null  category
 7   RECOD_TX_RESP_Q006E   378700 non-null  category
 8   RECOD_TX_RESP_Q013    378700 non-null  category
 9   RECOD_TX_RESP_Q003A   378700 non-null  category
 10  RECOD_TX_RESP_Q003B   378700 non-null  category
 11  RECOD_TX_RESP_Q007    378700 non-null  category
 12  RECOD_TX_RESP_Q014    378700 non-null  category
 13  RECOD_TX_RESP_Q017E   378700 non-null  category
 14  RECOD_TX_RESP_Q008A   378700 non-null  category
 15  RECOD_TX_RESP_Q008B   378700 non-null  category
 16  RECOD_TX_RESP_Q008C   378700 non-null  category
 17  RECOD_TX_RESP_Q009c   378700 non-null  category
 18  RECOD_TX_RESP_Q009g   378700 non-null  category
 19  PROFICIENCIA_MT_SAEB  378700 non-null  float64
dtypes: category(18), float64(2)
memory usage: 15.2 MB
```

Para analisar as variáveis categóricas deste modelo também foi usada à função *get_dummies( )* com intuito de transformá-las de forma adequada para implementar técnicas de regressão.

IV - Análise Exploratória

Nesta etapa, apresentaremos uma descrição das variáveis usadas para realização deste estudo. Nas imagens à seguir é possível identificar a distribuição de frequência das features através da função *value_counts( )*:

Analisando variabilidade da Variável dependente

```
[130]: df["RECOD_TX_RESP_Q004"].value_counts()/378700*100

[130]: Ensino Médio completo.                                31.640613
       Ensino Superior completo (faculdade ou graduação).    24.706628
       Ensino Fundamental completo.                          20.458146
       Ensino Fundamental, até o 5º ano.                     12.577766
       Não completou o 5º ano do Ensino Fundamental.         10.616847
       Name: RECOD_TX_RESP_Q004, dtype: float64

[147]: df["RECOD_TX_RESP_Q005"].value_counts()/378700*100

[147]: Ensino Médio completo.                                28.864008
       Ensino Superior completo (faculdade ou graduação).    22.457090
       Ensino Fundamental completo.                          19.724584
       Ensino Fundamental, até o 5º ano.                     15.125165
       Não completou o 5º ano do Ensino Fundamental.         13.829152
       Name: RECOD_TX_RESP_Q005, dtype: float64

[131]: df["RECOD_TX_RESP_Q006A"].value_counts()/378700*100

[131]: Sempre ou Quase sempre    49.935569
       De vez em quando          42.243200
       Nunca ou Quase nunca       7.821231
       Name: RECOD_TX_RESP_Q006A, dtype: float64

[133]: df["RECOD_TX_RESP_Q006B"].value_counts()/378700*100

[133]: Sempre ou quase sempre.    87.110906
       De vez em quando.          10.167415
       Nunca ou quase nunca.       2.721679
       Name: RECOD_TX_RESP_Q006B, dtype: float64

[134]: df["RECOD_TX_RESP_Q006C"].value_counts()/378700*100

[134]: Sempre ou quase sempre    81.616583
       De vez em quando          13.944547
       Nunca ou quase nunca       4.438870
       Name: RECOD_TX_RESP_Q006C, dtype: float64
```

```
[137]: df["RECOD_TX_RESP_Q006D"].value_counts()/378700*100

[137]: Sempre ou quase sempre    90.920518
       De vez em quando           6.169263
       Nunca ou quase nunca       2.910219
       Name: RECOD_TX_RESP_Q006D, dtype: float64

[138]: df["RECOD_TX_RESP_Q006E"].value_counts()/378700*100

[138]: Sempre ou quase sempre    61.427779
       De vez em quando          29.945075
       Nunca ou quase nunca       8.627145
       Name: RECOD_TX_RESP_Q006E, dtype: float64

[139]: df["RECOD_TX_RESP_Q007"].value_counts()/378700*100

[139]: Nunca ou quase nunca.                                          73.580935
       De vez em quando (uma vez por semana, a cada quinze dias etc.).    15.237919
       Sempre ou quase sempre (ex.: três ou mais dias por semana).        11.181146
       Name: RECOD_TX_RESP_Q007, dtype: float64

[140]: df["RECOD_TX_RESP_Q014"].value_counts()/378700*100

[140]: Somente em escola pública.                 75.740692
       Em escola pública e em escola particular.    16.967256
       Somente em escola particular.                 7.292052
       Name: RECOD_TX_RESP_Q014, dtype: float64

[141]: df["RECOD_TX_RESP_Q017E"].value_counts()/378700*100

[141]: Não uso meu tempo para isso    86.146818
       Mais de 2 horas.                5.514655
       Menos de 1 hora.                5.186427
       Entre 1 e 2 horas.              3.152099
       Name: RECOD_TX_RESP_Q017E, dtype: float64

[142]: df["RECOD_TX_RESP_Q008A"].value_counts()/378700*100

[142]: Sim    73.909691
       Não    26.090309
       Name: RECOD_TX_RESP_Q008A, dtype: float64

[143]: df["RECOD_TX_RESP_Q008B"].value_counts()/378700*100

[143]: Sim    78.54423
       Não    21.45577
       Name: RECOD_TX_RESP_Q008B, dtype: float64

[144]: df["RECOD_TX_RESP_Q008C"].value_counts()/378700*100

[144]: Sim    89.612622
       Não    10.387378
       Name: RECOD_TX_RESP_Q008C, dtype: float64

[145]: df["RECOD_TX_RESP_Q009c"].value_counts()/378700*100

[145]: Nenhum       39.781885
       1            38.728809
       2            15.153684
       3 ou mais     6.335622
       Name: RECOD_TX_RESP_Q009c, dtype: float64

[146]: df["RECOD_TX_RESP_Q009g"].value_counts()/378700*100

[146]: 1            43.982044
       Nenhum       36.073145
       2            14.531291
       3 ou mais     5.413520
       Name: RECOD_TX_RESP_Q009g, dtype: float64
```

Para avaliar a variável *target* usamos a função *.describe(*), além dos gráficos de *boxplot* e histograma com intuito de avaliar em que medida os dados da variável dependente poderiam ser distribuídos adequadamente ou estando ou não dentro de uma distribuição normal (curva gaussiana).

```
[153]: df.describe()
```

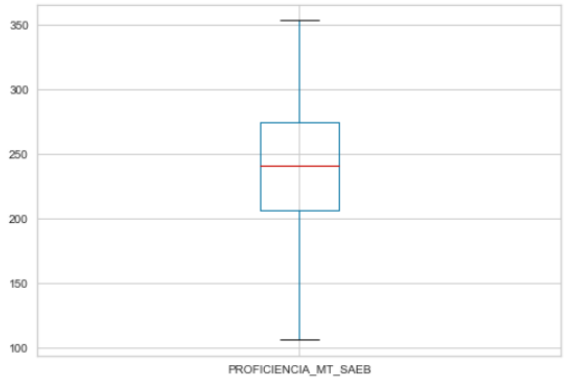
[153]:

	ID_UF	PROFICIENCIA_MT_SAEB
count	378700.000000	378700.000000
mean	32.314434	240.309836
std	10.042947	47.445980
min	11.000000	105.998488
25%	26.000000	206.471025
50%	33.000000	240.752487
75%	35.000000	274.357414
max	53.000000	353.089638

Conforme apresentado pela imagem acima é possível perceber identificar na variável Target “Proficiência_MT_SAEB” uma próximidade entre média (mean = 240.309839) e mediana (median =240.752487) (percentil 50) com uma relativa variabilidade dos dados (std = 47.4459).

```
[88]: df["PROFICIENCIA_MT_SAEB"].plot(kind="box")
```

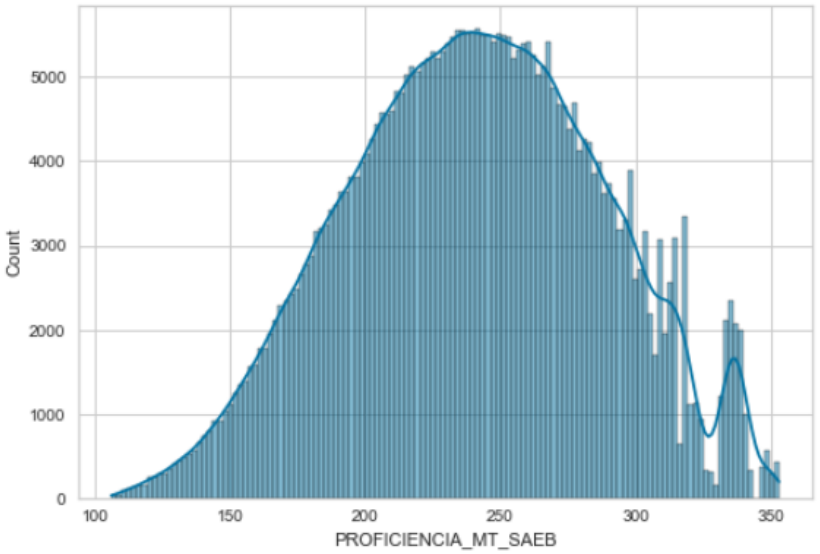
[88]: <AxesSubplot:>



Usando o gráfico de *boxplot* é possível visualizar que a construção do gráfico não identificou a incidência de outliers.

```
[161]: sns.histplot(data=df, x="PROFICIENCIA_MT_SAEB",kde = True)

[161]: <AxesSubplot:xlabel='PROFICIENCIA_MT_SAEB', ylabel='Count'>
```



Através da visualização do histograma, é possível avaliar uma tendência de desajuste da curva normal no extremo dos alunos com desempenho acima de 300. Esse desvio pode se apresentar como um problema para a realização de testes paramétricos.

Na próxima etapa, para tentar compreender a relação entre vulnerabilidade e desempenho acadêmico, realizaremos uma regressão linear. Através desta regressão, será observado se a distribuição dos resíduos se organiza de forma normal.

V - Modelagem

Para realizar a modelagem de análise de dados, escolhemos as seguintes variáveis: Escolaridade da Mãe; Escolaridade do Pai; Conversar com o filho sobre o que acontece na escola; incentivos: aos estudos, fazer tarefa de casa, comparecer às aulas, ir às reuniões de pais na escola; com que idade a criança entrou na escola; possui pai ou padrasto (sim/não); mãe ou madrasta(sim/não); com que frequência a família paga alguém para fazer faxina dentro de casa; trabalho infantil; possuem carro; rua pavimentada (asfalto ou calçamento); agua tratada da rua; iluminação na rua; Possuem computador (ou notebook).

Conforme demonstrado mais abaixo, estas variáveis foram escolhidas com base em construtos teóricos que avaliam tanto experiências de vulnerabilidade na família, quanto de *proxies* de renda e infraestrutura.

Features (VI)		
Construto teórico	**Descrição da Variável**	**Coluna**
Família	• Escolaridade da Mãe • Escolaridade do Pai	RECOD_TX_RESP_Q004 RECOD_TX_RESP_Q005
	• Conversar com o filho sobre o que acontece na escola • Incentivar a estudar • Incentivar:Fazer tarefa de casa • Incentivar:Comparecer às aulas • Ir às reuniões de pais na escola • Com que Idade que você entrou na escola	RECOD_TX_RESP_Q006A RECOD_TX_RESP_Q006B RECOD_TX_RESP_Q006C RECOD_TX_RESP_Q006D RECOD_TX_RESP_Q006E RECOD_TX_RESP_Q013
	• Mora com pai ou Padrasto • Mãe ou madrasta	RECOD_TX_RESP_Q003A RECOD_TX_RESP_Q003B
SES **Infraestrutura**	• Com que frequência a família paga alguém para fazer faxina dentro de casa • Trabalho infantil • Carro	RECOD_TX_RESP_Q007 RECOD_TX_RESP_Q017E RECOD_TX_RESP_Q009g
	• Rua pavimentada (asfalto ou calçamento). • Água tratada da rua. • Iluminação na rua. • Computador (ou notebook).	RECOD_TX_RESP_Q008A RECOD_TX_RESP_Q008B RECOD_TX_RESP_Q008C RECOD_TX_RESP_Q009c

Para analisar e representar os dados através de modelos de regressão, usamos principalmente as bibliotecas: *yellowbrick*, *scikit-learn*,*stats.*

Definimos o modelo de predição geral através de um test-size que incorpore cerca de ⅔ (67%) dos dados. Com intuito de estabelecer critérios claros de reprodutibilidade, definimos um valor = 5 para o random-state.

```
[164]: X = df[["RECOD_TX_RESP_Q004","RECOD_TX_RESP_Q006A",
               "RECOD_TX_RESP_Q006B","RECOD_TX_RESP_Q006C","RECOD_TX_RESP_Q006D","RECOD_TX_RESP_Q006E","RECOD_TX_RESP_Q013","RECOD_TX_RESP_Q003A",
               "RECOD_TX_RESP_Q003B","RECOD_TX_RESP_Q007","RECOD_TX_RESP_Q014","RECOD_TX_RESP_Q017E","RECOD_TX_RESP_Q008A","RECOD_TX_RESP_Q008B",
               "RECOD_TX_RESP_Q008C","RECOD_TX_RESP_Q009c","RECOD_TX_RESP_Q009g", "RECOD_TX_RESP_Q005"]]

[165]: X = pd.get_dummies(X)

[166]: y = df["PROFICIENCIA_MT_SAEB"]

[168]: X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = .67, random_state = 5)
```

Considerando a grande quantidade de fatores que incidem no desempenho acadêmico, é possível identificar um r2 de treino e teste variando entre 17-18%.

```
[181]: print(r2_score(y_test, y_pred))
       print(r2_score(y_train, y_pred_train))

       0.1776621969619191
       0.17890206445275225
```

Ao avaliar o resultado através da biblioteca *statsmodel* é possível perceber que o modelo possui elevado nível de significância (F-statistics = 0.00) com um valor de R-quadrado ajustado = 0.179. Também foi identificado na biblioteca um valor de Durbi-Watson

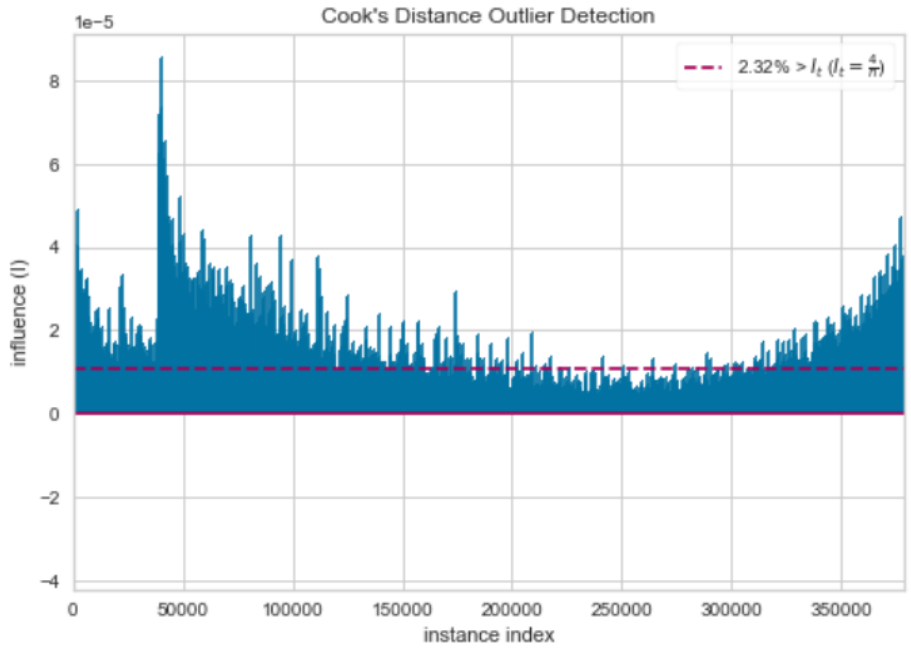
satisfatório. Este teste avalia a possibilidade de homocedasticidade da amostra através do modelo, valores situados entre um e 2 são considerados preferíveis. Nosso modelo encontrou o valor = 1.988. Contudo, alguns resultados chamam a atenção novamente para problemas com a distribuição, o valor de Prob(Omnibus) = 0.00 indicam problemas com o resíduo da distribuição.

OLS Regression Results

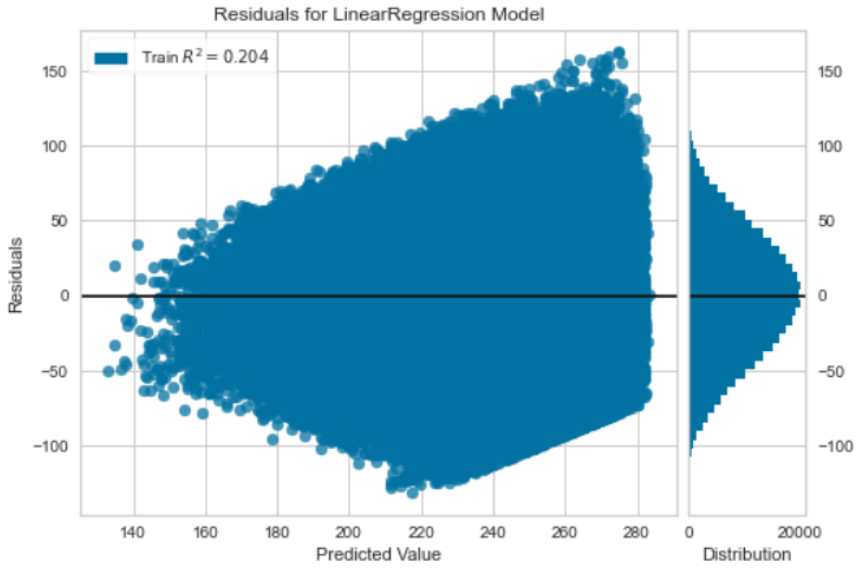
Dep. Variable:	PROFICIENCIA_MT_SAEB	R-squared:	0.179
Model:	OLS	Adj. R-squared:	0.179
Method:	Least Squares	F-statistic:	699.6
Date:	Tue, 14 Sep 2021	Prob (F-statistic):	0.00
Time:	17:04:28	Log-Likelihood:	-6.4723e+05
No. Observations:	124970	AIC:	1.295e+06
Df Residuals:	124930	BIC:	1.295e+06
Df Model:	39		
Covariance Type:	nonrobust		

Omnibus:	394.531	Durbin-Watson:	1.988
Prob(Omnibus):	0.000	Jarque-Bera (JB):	305.861
Skew:	0.019	Prob(JB):	3.83e-67
Kurtosis:	2.761	Cond. No.	2.85e+16

Para retirar a dúvida, analisamos graficamente duas medidas que podem nos ajudar na avaliação do modelo: a distribuição dos resíduos e a distância de *Cook's*. Além destas medidas, calculamos a distribuição normal dos resíduos através do teste de *Kolmogorov-Smirnov*.

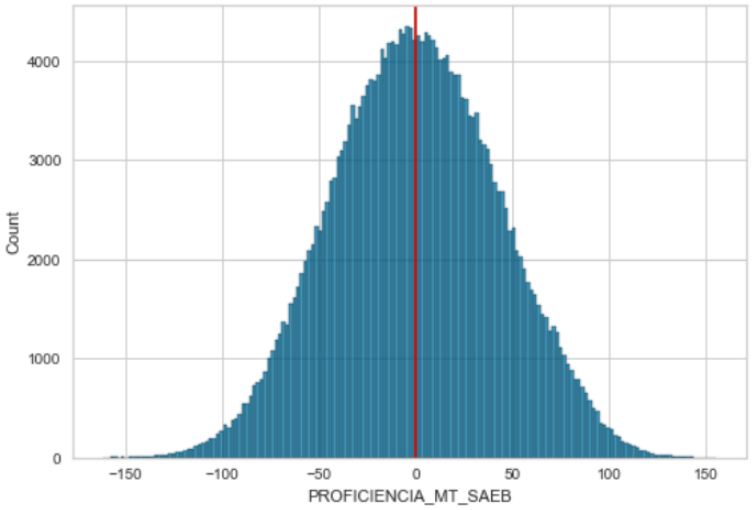


A distância de Cook's avalia segundo cada um dos casos da amostra o quanto eles poderiam estar influenciando na regressão linear. Valores acima de 1 demandam maiores atenções. Conforme apresentado, pelo gráfico, nosso modelo apresenta uma grande quantidade de valores que excedem esse valor. Um dos fatores que podem interferir neste processo é justamente problemas relacionados à distribuição dos resíduos.



[195]:
```
residuals = y_test - y_pred
sns.histplot(x=residuals)
plt.axvline(x=0, c='r')
D, p = stats.kstest(residuals, 'norm')
print(f'O valor do teste Kolmogorov é: {D} com um p-value = {p}')
```

```
O valor do teste Kolmogorov é: 0.4758388412483871 com um p-value = 0.0
```



Avaliando graficamente a distribuição dos resíduos, segundo a representação acima seria possível questionar ainda se a distribuição dos resíduos é ou não é normal. Para tanto, realizamos o teste Kolmogorov-Smirnov onde foi detectado que o valor de p de significância é menor que 0,05 (p=0,001). Este valor assume que nossa distribuição não é normal.

Com intuito de estabelecer uma análise que se adapte melhor à disposição do banco de dados e do modelo, foi implementado um modelo de regressão random forest regression. Este modelo possui uma melhor receptividade de dados não paramétricos.

Adotamos o mesmo critério de treino e teste realizado na Regressão Linear.

[189]:
```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = .67, random_state = 5)
```

Através desta regressão alcançamos um valor de r-quadrado com valores muito próximos (~0.178)

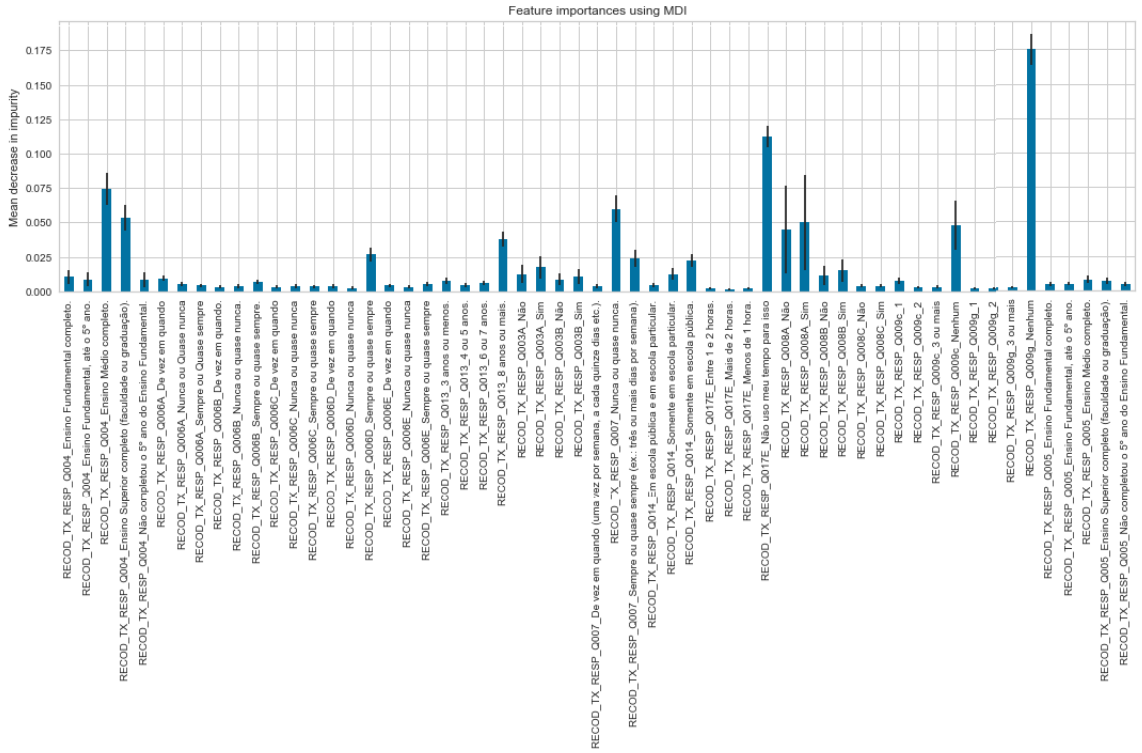
[201]:
```
print(r2_score(y_test, y_pred))
print(r2_score(y_train, y_pred_train))
```

```
0.17881950924970924
0.17890206445275225
```

[59]:
```
print(mape(y_test,y_pred))
print(mape(y_train,y_pred_train))
```

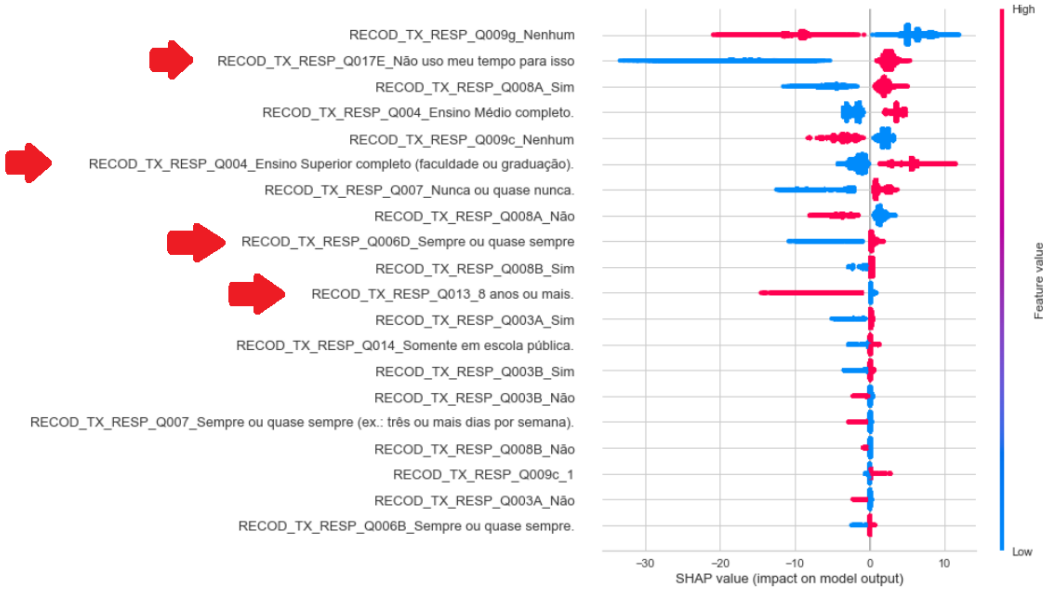
```
0.16055010191563165
0.15490643312501745
```

Através do gráfico de Features Importance é possível identificar a prevalência do peso de algumas features do modelo, comparativamente às demais.



Não possuir carro, a criança não trabalhar ao longo do quinto ano, pavimentação da rua, e grau de instrução materno surgiram como as principais features explicativas do desempenho da matemática.

Para analisar o peso de cada uma destas variáveis criamos um gráfico de shap usando a biblioteca que carrega o mesmo nome.



Através deste gráfico, é possível calcular o quanto cada feature (eixo vertical) impacta no desempenho acadêmico de acordo com a pontuação (eixo horizontal da reta). O gráfico apresenta as cores azuis representando o valor de 0 (onde o fenômeno da feature não está presente) e com a vermelha (valor = 1) onde a característica encontra-se presente.
Neste gráfico, é possível perceber que alunos que respondem negativamente trabalho infantil chegam a perder mais de 30 pontos nos testes de desempenho em matemática. Possuir nenhum carro possui uma grande dispersão de resultados. Embora o prejuízo chegue a uma perda de 20 pontos, há muitos casos onde alunos chegam a ganhar mais de 10 pontos. No campo da família, a mãe possuir ensino superior completo atua principalmente de forma positiva no desempenho dos estudantes (ganhos até 15 pontos). Outros resultado que merece atenção, encontram-se no atraso e o incentivo em comparecer às aulas. é o aluno começar a escola com treze anos ou mais na escola. Alunos com este atraso escolar chegam a perder até 15 pontos na avaliação. Pais que não incentivam a criança a comparecer sempre ou quase sempre na escola também indicam uma perda de pontuação que chega a 10 pontos.

VI - Considerações

O presente projeto teve por principal intuito avaliar através de dados transversais sociodemográficos como experiências de vulnerabilidade incidem sobre o desempenho acadêmico da matemática. Para tanto, adotamos técnicas e procedimentos estatísticos que demonstraram que o banco de dados governamental não obedece a uma distribuição normal. Desta forma, ao adotar o modelo *random forest regression*, mais afeito a distribuições não-paramétricas, constatou-se que o modelo adotado explicou cerca de 17% da variância dos resultados com *features* que poderiam incidir na perda de quase 10% do desempenho da prova de matemática. Contudo, é preciso destacar algumas limitações deste empreendimento. Como principais limitações, é possível destacar:

- O valor de mape do Randomforest (~0,15);
- Não avaliação da multicolinearidade do modelo;
- O uso acentuado de variáveis sendo categorizadas de acordo com o critério dummy;
- Não adoção de variáveis neuropsicológicas, mais ligadas ao indivíduo, na interpretação do fenômeno do desempenho escolar da matemática.;

Este projeto não leva em conta variáveis pertencentes a atributos cognitivos do indivíduo como inteligência, memória de trabalho, personalidade, dentre outras variáveis que costumam explicar a maior parcela da variância dos resultados (Herrnstein & Murray, 2010). Infelizmente, não existe até a presente data uma política nacional do estado de cômputo

destas informações. O registro delas demandaria uma coleta especializada que, por consequência, demandaria maiores esforços e investimentos por parte do Estado.

Entretanto, é possível afirmar que estes resultados são expressivos na medida em que o poder explicativo destas variáveis ultrapassa mesmo a influência da escola na avaliação do desempenho acadêmico quando comparamos com outros resultados internacionais (Berthelot, 2001).

Creio ser possível destacar que os resultados aqui apresentados reiteram a importância de pesquisas translacionais capazes de apontar para o risco que experiências de vulnerabilidade podem promover no desempenho acadêmico da matemática. A integração de diferentes áreas na tentativa de promover uma educação baseada em evidências parece ter um efeito promissor em uma política que tente reduzir a perda do desempenho acadêmico (Simplício, et al, 2020). Considerando a alta dependência da matemática na implementação de ferramentas dentro do mundo globalizado, parece imprescindível avaliar os fatores que poderiam incidir diretamente sobre a perda de desempenho na disciplina.

Referências

Berthelot, J. M., Ross, N., & Tremblay, S. (2001). Factors affecting Grade 3 student performance in Ontario: A multilevel analysis. Education Quarterly Review, 7(4), 25.
Geary, D. C., & Geary, D. C. (2007). Educating the evolved mind. Educating the evolved mind, 1-99.

Curi, A. Z. (2014). *A relação entre o desempenho escolar e os salários no Brasil* (Doctoral dissertation, Universidade de São Paulo).
Herrnstein, R. J., & Murray, C. (2010). The bell curve: Intelligence and class structure in American life. Simon and Schuster.

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., ... & Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. The Lancet Public Health, 2(8), e356-e366.

Palermo G., Silva, D., Novellino, M..(2014). Fatores associados ao desempenho escolar: uma análise da proficiência em matemática dos alunos do 5o ano do ensino fundamental da rede municipal do Rio de Janeiro. Revista Brasileira de Estudos de População, 31(2), 367-394.

Parsons, S., & Bynner, J. (2005). Does numeracy matter more?.National Research and Development Centre for Adult Literacy and Numeracy

Parsons, S., & Bynner, J. (1997). Numeracy and employment. *Education+ Training*.

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological science*, *24*(7), 1301-1308.

Simplicio, H., Gasteiger, H., Dorneles, B. V., Grimes, K. R., Haase, V. G., Ruiz, C., ... & Moeller, K. (2020). Cognitive research and mathematics education—How can basic research reach the classroom?. Frontiers in psychology, 11, 773.