Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Santa Helena
Bacharelado em Ciência da Computação

Disciplina:	Introdução a Ciência de Dados
Estudantes:	Angela Favretto Pastorello - 1196243 Bruna Oenning Amador - 2241463 Daniel de Viveiros Inácio - 1923820 Matheus Biscaya Gutierrez - 1657470
Equipe:	Dados Educação

RESULTADOS FINAIS

1 INTRODUÇÃO

O ENEM de 2020, comparado com os dados de anos anteriores à pandemia (2015 até 2019)[1], foi o ano que obteve a maior taxa de desistência, mesmo com um menor número de inscrições comparado aos anos de 2015 e 2016, por exemplo, conforme apresentado no Gráfico 1, que demonstra lado a lado o número absoluto de inscrições dos anos de 2015 até 2020 e o número de desistências de cada ano.

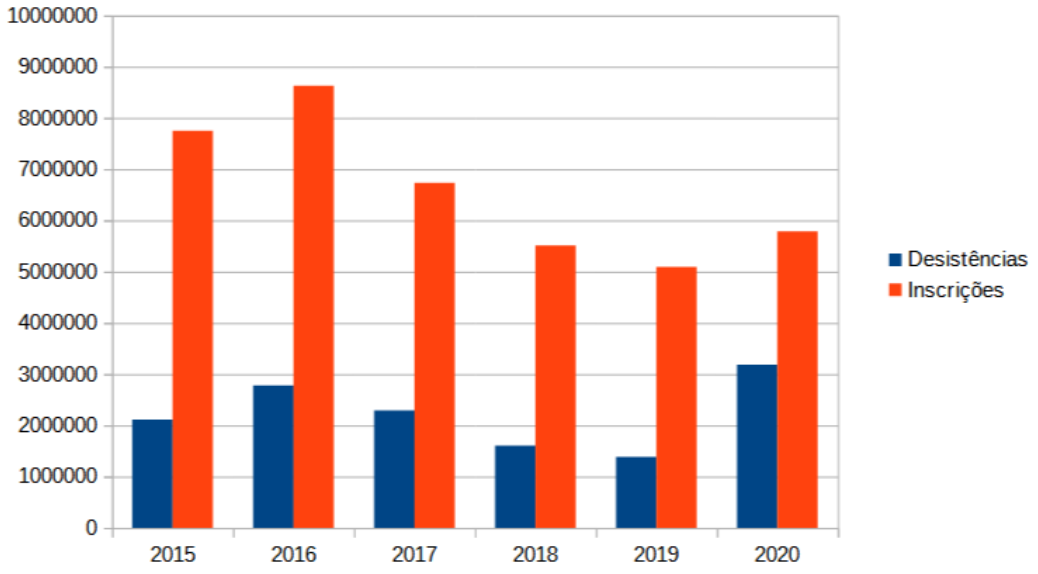

Gráfico 1 - Inscrições e desistências por ano

[1] Para a comparação com outros anos, foram utilizados os dados de todas as pessoas participantes e foi considerada desistência somente com relação aos dados do segundo dia de aplicação.

Da mesma forma, o Gráfico 2 demonstra a comparação entre a porcentagem de desistência em relação ao número de inscrições ao longo dos anos de 2015 até 2020 e é notável que houve um aumento na quantidade de estudantes que desistiram da prova no ano de pandemia (2020). Tendo em vista isto, o tema do projeto é 'Analisar o impacto da pandemia na educação através da análise das desistências do ENEM', sendo o objetivo principal avaliar o quanto a pandemia impactou neste aumento de desistências do ENEM, em comparação com o ano de 2019, e buscar a relação dos resultados com os aspectos socioeconômicos, identificando quais obtiveram maior influência.

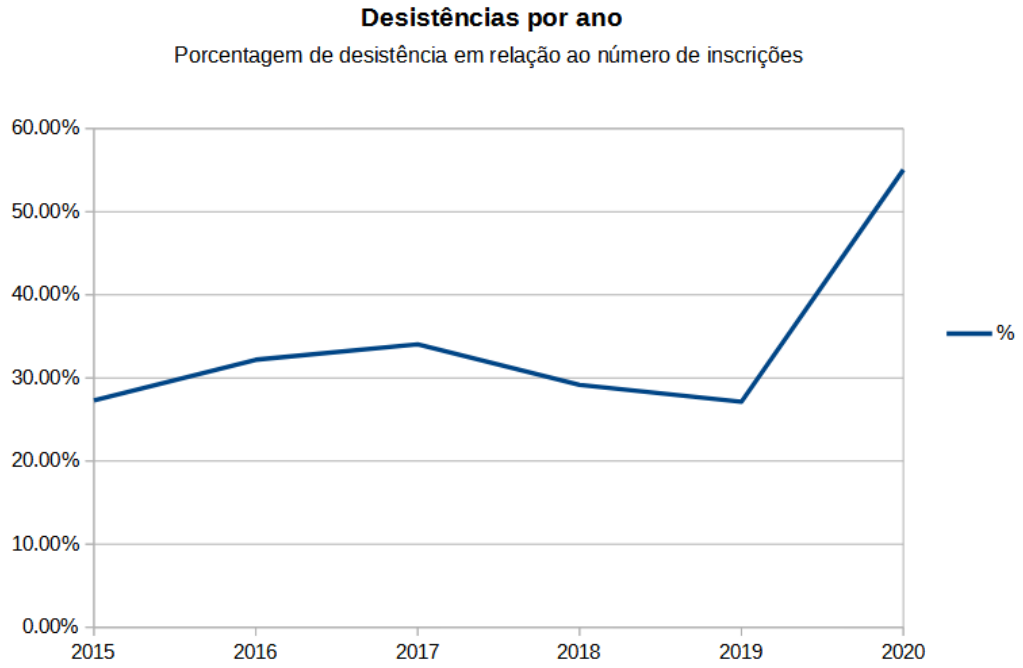

Gráfico 2 - Percentagem de desistências em relação ao número de inscrições por ano

Com relação às perguntas de pesquisa, foram identificados três aspectos principais. Primeiramente, o Questionário Socioeconômico apresentado nos Dados do ENEM, o qual foi questionado sobre 'Qual a influência dos fatores socioeconômicos do indivíduo sobre as desistências?'. Por outro lado, tendo em vista a pandemia de COVID-19, foi disposta a pergunta inicial 'A situação da pandemia no mês antes da prova está relacionada com as desistências no

município?'. Por fim, para a análise com dados do COVID-19 relacionados com os aspectos socioeconômicos dos municípios, tais como o PIB, IDH e densidade demográfica, foi questionado ‘Qual a influência da situação da pandemia durante o período antes da prova em conjunto com os aspectos socioeconômicos dos municípios sobre as desistências?’.

Como resultados, foram obtidos que os fatores que influenciaram o indivíduo a desistir foram principalmente a raça/cor e ser do sexo feminino, além de existir uma maior chance de indígenas faltarem a prova. As explicações discutidas no presente trabalho são de que os indígenas são os mais afetados por conta da região que habitam, de baixa densidade demográfica, dificultando sua locomoção até o local de prova. Por outro lado, existem pesquisas sobre o quanto a pandemia afetou as mulheres, sendo identificado que elas passaram a cuidar de alguém na pandemia, o que pode ter resultado na diminuição do tempo para se dedicar aos estudos e realização da prova e, consequentemente, maior número de desistências.

Por último, com relação a influência da situação da pandemia durante o período antes da prova em conjunto com os aspectos socioeconômicos dos municípios sobre as desistências, foi percebido que o maior problema identificado pelo modelo de regressão linear, conforme disposto na Seção 3.1, foram as condições sociais dos participantes. Vale complementar que existe maior influência do IDHM sobre a variável de resposta, ou seja, o aumento das desistências. O que também foi percebido como principal padrão na clusterização, que em estados de maior IDH, o impacto causado pelo COVID-19 nas desistências foi menor, discutido na Seção 3.3. Apesar do COVID-19 não ter influência direta sobre o aumento das desistências no modelo, a variável pode ter impactado indiretamente nos fatores socioeconômicos.

2 PROCESSAMENTO DE DADOS

Serão utilizados os seguintes dados para responder às perguntas, utilizando os processamentos descritos em cada tópico:

- Dados do Enem: Dados do INEP[2] dos anos de 2019 e 2020. Para a análise do questionário socioeconômico, foram utilizados os dados de renda(que foram convertidos para salários mínimos, a raça/cor do participante e a quantidade de determinados itens (carro, televisão, banheiros). Também, foram utilizados dados de presença do indivíduo, ano de formação, cidade e estado. Como processamento para realização da análise, foi aplicado um filtro de público/amostra, visto que foram considerados apenas estudantes que terminaram o Ensino Médio no ano de realização da prova, a transformação de campos e a diferença entre a taxa de abstenção de 2020 e 2019;
- Dados sobre o COVID-19: Dados do Ministério da Saúde[3] até o dia da prova (17 de Janeiro de 2021). Foi realizada a contagem total de casos e óbitos, considerando a população dos municípios, obtendo casos totais per capita, as mortes totais per capita, casos totais per capita um mês antes da prova e as mortes totais per capita um mês antes da prova;
- Produto Interno Bruto (PIB): Dados do IBGE[4] de 2010. Foram utilizados o PIB per capita de cada município e os seus respectivos códigos. Como processamento para realização da análise, foram alterados os nomes das colunas do PIB para facilitar a visualização, além de ser necessária a transformação do PIB per capita em *float*;
- Índice de Desenvolvimento Humano Municipal (IDHM): Dados do Programa das Nações Unidas para o Desenvolvimento[5] (PNUD) de 2010. Foram utilizados os dados do IDHM, IDHM Educação, IDHM Renda, IDHM

[2] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem
[3] www.gov.br/saude/pt-br
[4] www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html
[5] www.br.undp.org/content/brazil/pt/home/idh0.html

Longevidade e o código do município. Como processamento para realização da análise, foi necessário transformar o código do município em String;

- Densidade demográfica: Dados foram fornecidos pelo professor. Foram utilizados os dados de densidade demográfica e código dos municípios. Como processamento para realização da análise, foram alterados os nomes das colunas para facilitar a visualização;
- Desempenho escolar: Dados do INEP[6] de 2019 e 2020. Foi utilizada a diferença do rendimento escolar entre os anos de 2019 e 2020 por município.

Vale ressaltar que foi realizada a junção da tabela de desistências do ENEM com os dados socioeconômicos pelo código do município para a análise municipal, sendo necessário retirar o dígito verificador como padronização dos dados. Além de que foram considerados apenas municípios com pelo menos cinco alunos inscritos para a análise da regressão linear, conforme detalhado na Seção 3.1.

[6] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais

3 RESULTADOS

Os modelos escolhidos pela equipe para a realização do projeto foram gráficos de mapa e distribuição dos dados, correlação e diagrama de dispersão para a visualização, detalhados e discutidos na Seção 3.1, regressão linear e regressão logística para a análise estatística, presentes na Seção 3.2, e, por fim, o classificador para a aplicação de conceitos de *machine learning*, trabalhado na Seção 3.3.

3.1 VISUALIZAÇÃO

Para a etapa de visualização, foi aplicada a distribuição dos dados, para auxiliar nos testes estatísticos, obtendo como resultados que o maior número de desistências foi no estado do Amazonas, seguido de Rondônia, Goiás, Tocantins e Ceará, conforme apresentado no Gráfico 3.

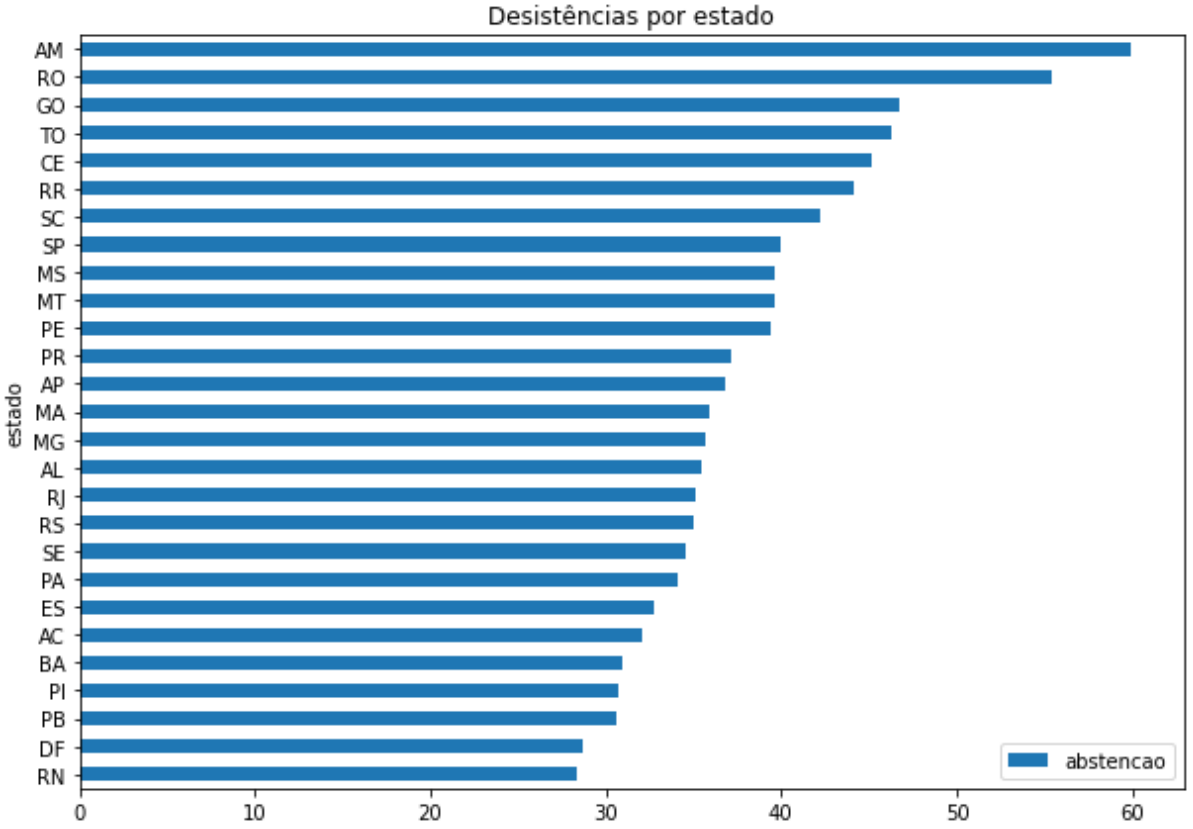

Gráfico 3 - Desistências por estado

Para auxiliar nas análises geográficas, foi aplicado um gráfico de mapa, com auxílio da biblioteca matplotlib, geopandas e o pacote geobr.

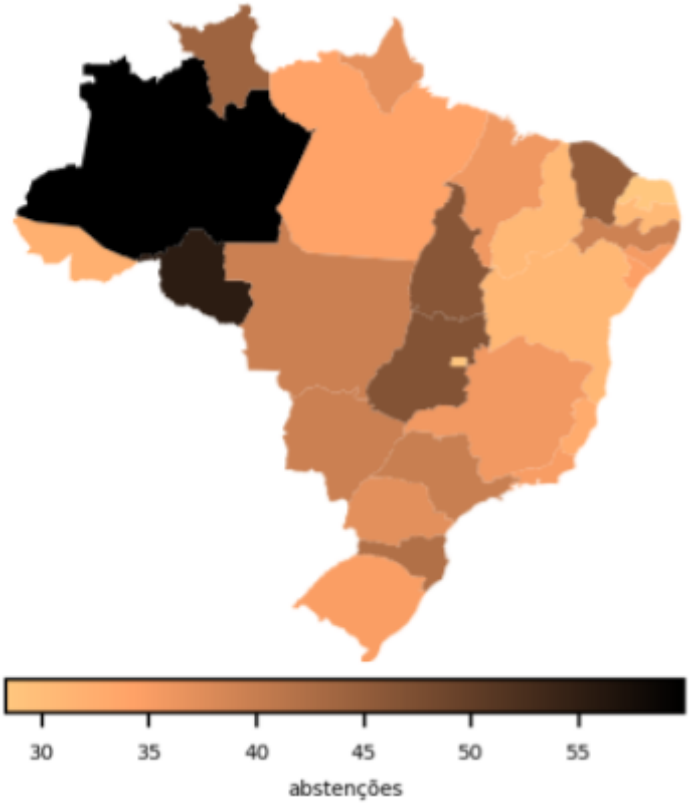

Gráfico 4 - Desistências por estado visualizadas no mapa do Brasil

É notável, tendo como base o Gráfico 4, uma concentração maior de desistências na região Norte e Centro-Oeste, por outro lado, a região Nordeste é a que possui os menores índices de desistência.

Também, foi utilizado o diagrama de dispersão para verificar a associação entre as variáveis (CELSO, 2022), apresentado no Diagrama 1, e tem-se que municípios com menor número de estudantes cadastrados resultam em diferenças extremas entre os anos. Posto isso, apenas municípios com pelo menos 5 alunos inscritos foram considerados para a análise na Seção 3.2, resultando em uma redução de 5016 municípios para 4774.

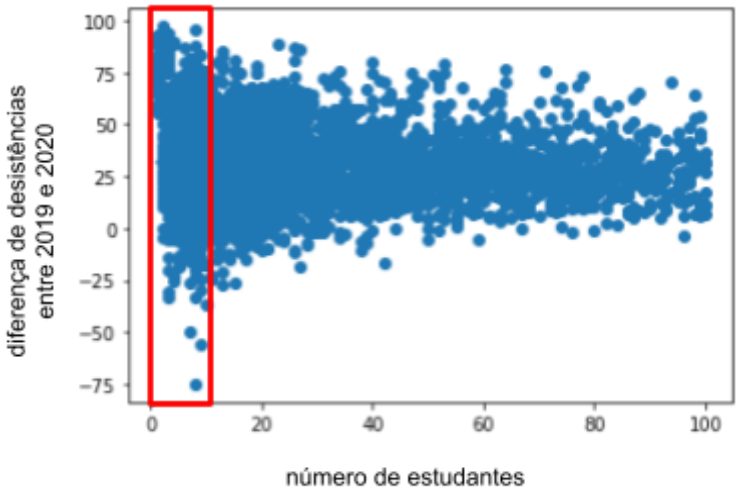

Diagrama 1 - Diagrama de dispersão entre número de estudantes e a diferença de desistências

A análise de correlação é uma forma descritiva que mede se há e qual o grau de dependência entre variáveis, ou seja, o quanto uma variável interfere em outra, lembrando que essa relação de dependência pode ou não ser causal (GUIMARÃES, 2021). Essa medida de grau de relação é medida através de coeficientes, conforme pode ser visto na Figura 1, em que foram analisadas todas as variáveis propostas para o presente trabalho.

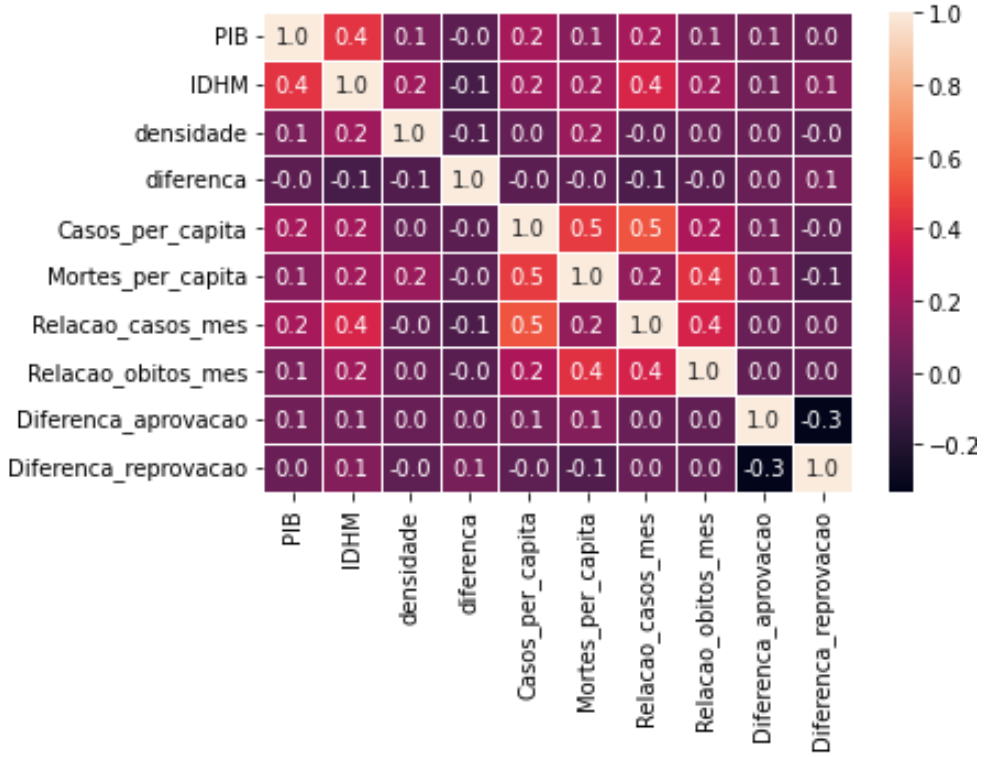

Figura 1 - Correlação entre as variáveis estudadas

Visto que este coeficiente assume apenas valores entre -1 e 1, as variáveis que possuem maior correlação são o IDHM, escolhido para representar o desenvolvimento do município, e o PIB, que representa a renda do município, e as relações de casos e mortes, sendo escolhido os casos per capita total para ser utilizado no modelo da Seção 3.2. Também, a diferença de aprovação e reprovação possuem uma correlação negativa, então foi utilizada apenas a reprovação.

3.2 ANÁLISE ESTATÍSTICA

Para a etapa de análise estatísticas, foram utilizadas as análises de regressão, que são muito usadas para se entender a relação entre variáveis (CELSO, 2022). Tendo isso em vista, foi utilizado o modelo de regressão linear e o modelo de regressão logística, apresentados na Seção 3.2.1 e na Seção 3.2.2 respectivamente para responder às perguntas de pesquisas iniciais.

3.2.1 Regressão linear

O modelo de regressão linear, realizado em nível municipal, visou verificar a relação entre os fatores socioeconômicos e os dados da pandemia com o aumento das desistências de 2020 em relação a 2019, para responder às perguntas propostas na Seção 1. Inicialmente, para responder a pergunta: ‘A situação da pandemia no mês antes da prova está relacionada com as desistências no município?’ foi aplicado o modelo de regressão linear.

Dessa forma, considerando os aspectos da pandemia durante a construção do modelo, foram consideradas quatro variáveis para analisar se a situação da pandemia estava relacionada com as desistências no município, são elas: Casos per capita, mortes per capita, casos per capita 30 dias antes do exame e mortes per capita 30 dias antes do exame. Ao realizar regressão linear simples utilizando cada variável para explicar a desistência nos municípios, foi observado que cada uma das variáveis resultou em um coeficiente negativo, ou seja, quanto maiores forem a ocorrência total de casos e mortes causados pela pandemia de COVID-19 nos municípios, menores serão as desistências, sugerindo que a ocorrência total de casos e mortes causados pela pandemia está pouco relacionada com as desistências que ocorreram nos municípios. De qualquer forma, as variáveis demonstraram p-valores significativos.

Em seguida, para entender ‘O quão relacionada estava a situação da pandemia durante o período antes da prova com os aspectos socioeconômicos dos municípios?’, foi utilizada a correlação entre as variáveis, apresentada na Seção 3.1.

Tendo isso em vista, o modelo de regressão busca entender os fatores que influenciam na diferença entre a taxa de desistência de 2020, em relação ao ano de 2019. Em termos estatísticos, as variáveis socioeconômicas, assim como as variáveis de COVID-19, são as variáveis dependentes, ou seja, que dependem dos valores das outras, já a diferença é chamada de variável independente.

Para a análise da pandemia, foi escolhida a variável que representa os casos totais per capita até o dia do exame dado que ela representa adequadamente todos os municípios, ao contrário das outras variáveis apresentadas, por existirem municípios onde não houveram mortes até o dia do exame e nem mortes durante o período de 30 dias antes do exame. Além disso, a variável possui menor relação com as variáveis que representam as características socioeconômicas do município quando comparada a variável que representa os casos 30 dias antes do exame, resultando em menor risco de colinearidade.

Dep. Variable:	diferenca	**R-squared:**	0.150
Model:	OLS	**Adj. R-squared:**	0.145
Method:	Least Squares	**F-statistic:**	27.90
Date:	Thu, 23 Jun 2022	**Prob (F-statistic):**	2.68e-143
Time:	10:32:45	**Log-Likelihood:**	-19880.
No. Observations:	4774	**AIC:**	3.982e+04
Df Residuals:	4743	**BIC:**	4.002e+04
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
IDHM	-55.9346	5.708	-9.799	0.000	-67.126	-44.743
PIB	2.929e-05	9.95e-06	2.945	0.003	9.79e-06	4.88e-05
Casos_per_capita	-29.3222	11.864	-2.471	0.013	-52.582	-6.063
Diferenca_reprovacao	0.1415	0.041	3.490	0.000	0.062	0.221

Figura 2 - Resultado da regressão linear múltipla

Utilizando as variáveis IDHM, PIB, casos de COVID-19 per capita e diferença de reprovação, vale ressaltar que a densidade com as outras variáveis não deu um valor estatisticamente significativo (p-value alto). Dado os resultados da Figura 2, foram identificadas correlações negativas nas variáveis IDHM e casos per capita, isto significa que possuem correlação, mas quando uma variável cresce a outra decresce, ou vice-versa (GUIMARÃES, 2021). Sendo assim, pode-se interpretar que quanto menor o IDHM, maior o aumento de desistências, indicando que os moradores de municípios com o IDHM maior estão mais protegidos. Em relação aos casos per capita, quanto menor os casos per capita, maior o aumento de desistências. Dessa maneira, tem-se que o COVID-19 não teve impacto direto nas desistências.

Por outro lado, o PIB e a diferença do índice de reprovação resultaram em correlações positivas. Posto isso, pode-se dizer que quanto maior o PIB, maior o aumento de desistências, concluindo que as pessoas mais ricas foram mais afetadas. Da mesma forma, mesmo existindo uma diminuição no número de reprovações no ano de 2020, comparado com 2019, quanto maior esta diferença de reprovação entre os anos, maior o aumento de desistências. Logo, sobre a formação do participante em 2020, é provável que exista uma menor motivação para fazer a prova caso o estudante tenha reprovado no ano de realização da prova.

Portanto, o maior problema identificado pelo modelo foram as condições sociais dos participantes, ressaltando a influência do IDHM para o aumento das desistências e, vale evidenciar que, apesar do COVID-19 não ter influência direta sobre o aumento das desistências, a variável pode ter impactado indiretamente nos fatores socioeconômicos, o que não foi possível explicar através do modelo.

Na análise por estado, os coeficientes mais altos significam que existem outros fatores que também influenciam o aumento das desistências no estado e não são explicados e controlados pelo modelo, conforme apresentado na Figura 3.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	68.3062	4.138	16.506	0.000	60.193	76.419
C(estado, Treatment(reference='São Paulo'))[T.Acre]	-11.1232	3.597	-3.092	0.002	-18.175	-4.072
C(estado, Treatment(reference='São Paulo'))[T.Alagoas]	-5.9801	1.977	-3.024	0.003	-9.857	-2.104
C(estado, Treatment(reference='São Paulo'))[T.Amapá]	4.7534	4.171	1.140	0.255	-3.424	12.931
C(estado, Treatment(reference='São Paulo'))[T.Amazonas]	10.5587	2.424	4.356	0.000	5.807	15.311
C(estado, Treatment(reference='São Paulo'))[T.Bahia]	-13.0865	1.440	-9.086	0.000	-15.910	-10.263
C(estado, Treatment(reference='São Paulo'))[T.Ceará]	-1.6296	1.513	-1.077	0.282	-4.596	1.337
C(estado, Treatment(reference='São Paulo'))[T.Distrito Federal]	-5.9071	15.653	-0.377	0.706	-36.594	24.780
C(estado, Treatment(reference='São Paulo'))[T.Espírito Santo]	-3.6058	1.983	-1.818	0.069	-7.493	0.282
C(estado, Treatment(reference='São Paulo'))[T.Goiás]	15.1105	1.279	11.818	0.000	12.604	17.617
C(estado, Treatment(reference='São Paulo'))[T.Maranhão]	-7.8791	1.565	-5.036	0.000	-10.946	-4.812

Figura 3 - Resultado da regressão linear múltipla utilizando o Estado como variável categórica

Vale ressaltar que São Paulo é o estado utilizado como referência para análise, então os coeficientes positivos possuem maior impacto em relação ao estado de São Paulo. Para melhor visualização, o Gráfico 5 apresenta os estados que foram mais e menos afetados segundo seus coeficientes.

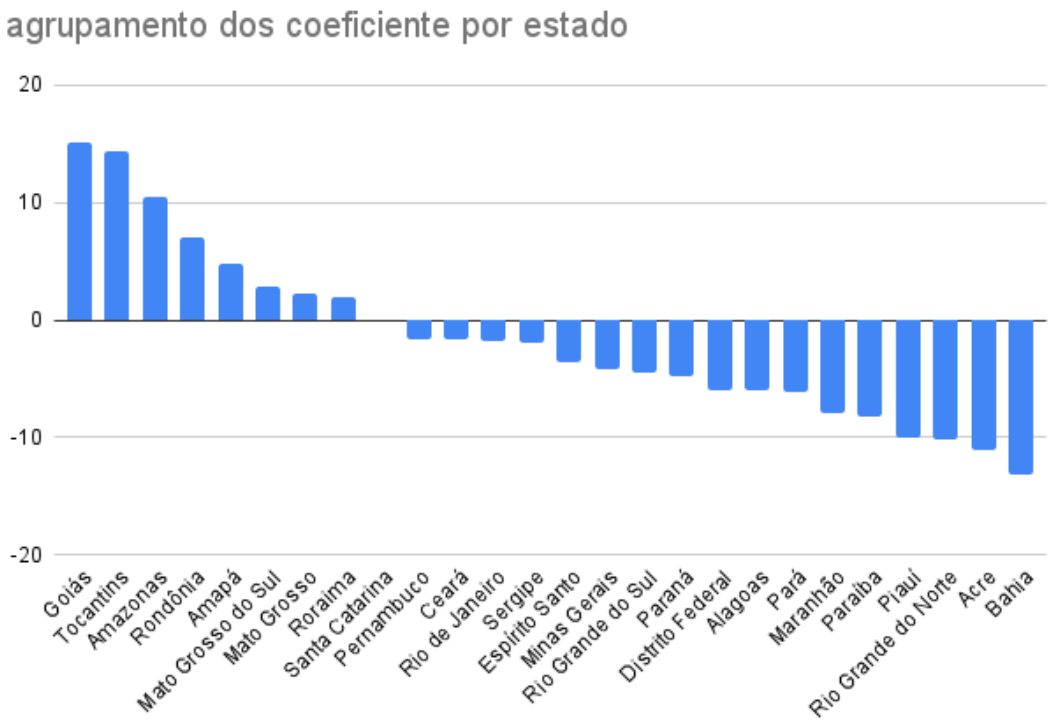

Gráfico 5 - Agrupamento dos coeficientes por estado

É possível perceber que os estados de Goiás, Tocantins e Amazonas foram os mais afetados, por outro lado, o Rio Grande do Norte, Acre e Bahia foram os menos afetados, ambos com relação ao estado de São Paulo.

3.2.2 Regressão logística

A regressão logística foi utilizada para analisar os fatores socioeconômicos que influenciaram os indivíduos a desistirem do ENEM no ano de 2020 em comparação ao ano de 2019, em resposta à pergunta 'Qual a influência dos fatores socioeconômicos do indivíduo sobre as desistências?'. Os resultados para o ano de 2019 são apresentados no Gráfico 6.

Logit Regression Results

Dep. Variable:	falta	No. Observations:	1432936
Model:	Logit	Df Residuals:	1432925
Method:	MLE	Df Model:	10
Date:	Wed, 22 Jun 2022	Pseudo R-squ.:	0.02090
Time:	22:38:44	Log-Likelihood:	-6.3537e+05
converged:	True	LL-Null:	-6.4893e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.0676	0.029	-70.284	0.000	-2.125	-2.010
C(TP_SEXO)[T.M]	0.0244	0.005	5.298	0.000	0.015	0.033
C(TP_COR_RACA)[T.2]	0.0655	0.008	8.574	0.000	0.051	0.080
C(TP_COR_RACA)[T.3]	0.1290	0.005	24.421	0.000	0.119	0.139
C(TP_COR_RACA)[T.4]	0.0658	0.015	4.304	0.000	0.036	0.096
C(TP_COR_RACA)[T.5]	0.3905	0.025	15.902	0.000	0.342	0.439
C(automovel)[T.sim]	-0.3260	0.005	-66.409	0.000	-0.336	-0.316
salario	-0.0982	0.001	-90.931	0.000	-0.100	-0.096
IDHM	0.0108	0.000	25.220	0.000	0.010	0.012
PIB	0.0020	9.78e-05	20.647	0.000	0.002	0.002
densidade	-2.26e-05	9.3e-07	-24.309	0.000	-2.44e-05	-2.08e-05

Gráfico 6 - Modelo de regressão logística do ano de 2019

Todas as variáveis tiveram significância estatística alta e, podem ser transformados os valores para razão de possibilidade, conforme o Gráfico 7.

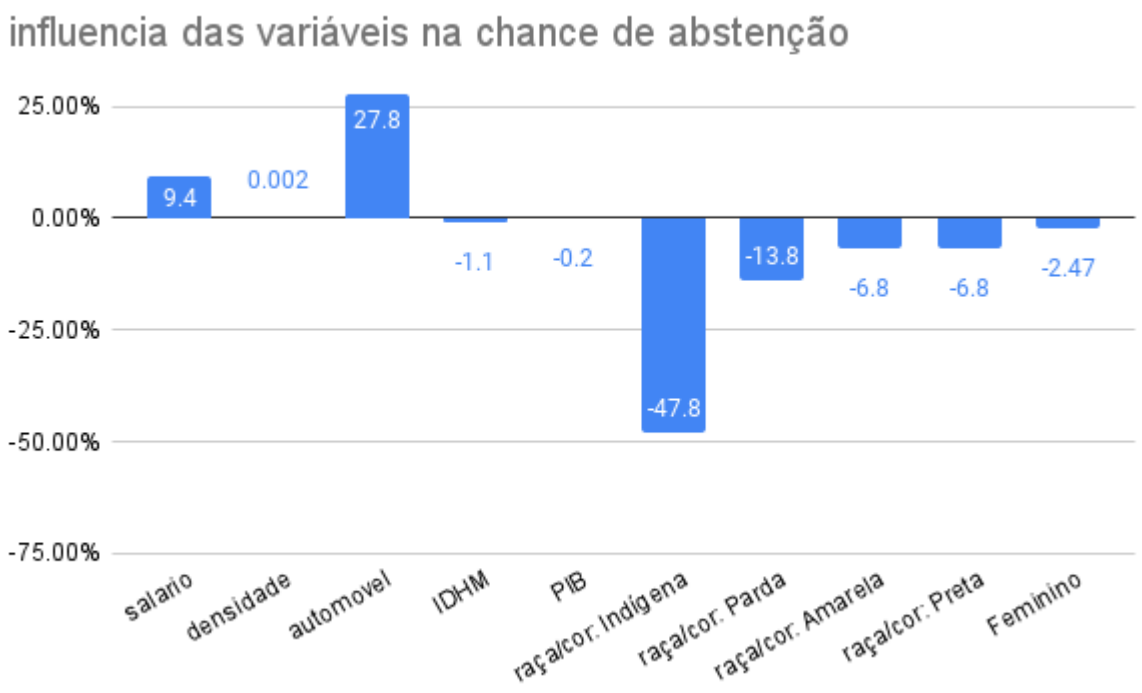

Gráfico 7 - Razão de possibilidades das variáveis de 2019

Indivíduos os quais as variáveis categóricas referem-se à raça/cor Indígena, Parda, Amarela e Preta, têm maior chance de faltar quando comparado com a categoria base (raça/cor Branca). Posto isso, pode-se concluir que os indígenas são os mais afetados provavelmente pela região pouco densa onde moram, ressaltando que as pessoas com carro possuem 27.8% de chance a mais de não faltar do que as que não possuem.

Na análise das variáveis econômicas é interessante notar que aumentar o salário aumenta a chance da pessoa ir, porém, para o PIB, a situação é oposta. Por outro lado, dado o questionário socioeconômico aplicado no ENEM de 2020, tem-se o resultado da regressão logística apresentado no Gráfico 8.

Logit Regression Results

Dep. Variable:	falta	No. Observations:	1339171
Model:	Logit	Df Residuals:	1339160
Method:	MLE	Df Model:	10
Date:	Wed, 22 Jun 2022	Pseudo R-squ.:	0.02384
Time:	22:34:18	Log-Likelihood:	-8.8528e+05
converged:	True	LL-Null:	-9.0690e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.3704	0.023	-15.891	0.000	-0.416	-0.325
C(TP_SEXO)[T.M]	0.1481	0.004	40.744	0.000	0.141	0.155
C(TP_COR_RACA)[T.2]	0.0938	0.006	15.152	0.000	0.082	0.106
C(TP_COR_RACA)[T.3]	0.1769	0.004	42.564	0.000	0.169	0.185
C(TP_COR_RACA)[T.4]	0.0936	0.012	7.627	0.000	0.070	0.118
C(TP_COR_RACA)[T.5]	0.5102	0.021	24.360	0.000	0.469	0.551
C(automovel)[T.sim]	-0.3567	0.004	-90.967	0.000	-0.364	-0.349
salario	-0.0676	0.001	-107.329	0.000	-0.069	-0.066
IDHM	0.0027	0.000	7.996	0.000	0.002	0.003
PIB	0.0022	8.23e-05	26.629	0.000	0.002	0.002
densidade	-2.336e-05	7.59e-07	-30.787	0.000	-2.49e-05	-2.19e-05

Gráfico 8 - Modelo de regressão logística do ano de 2020

É notável que todas as variáveis tiveram significância estatística alta e, os valores podem ser transformados para razão de possibilidade, conforme o Gráfico 9.

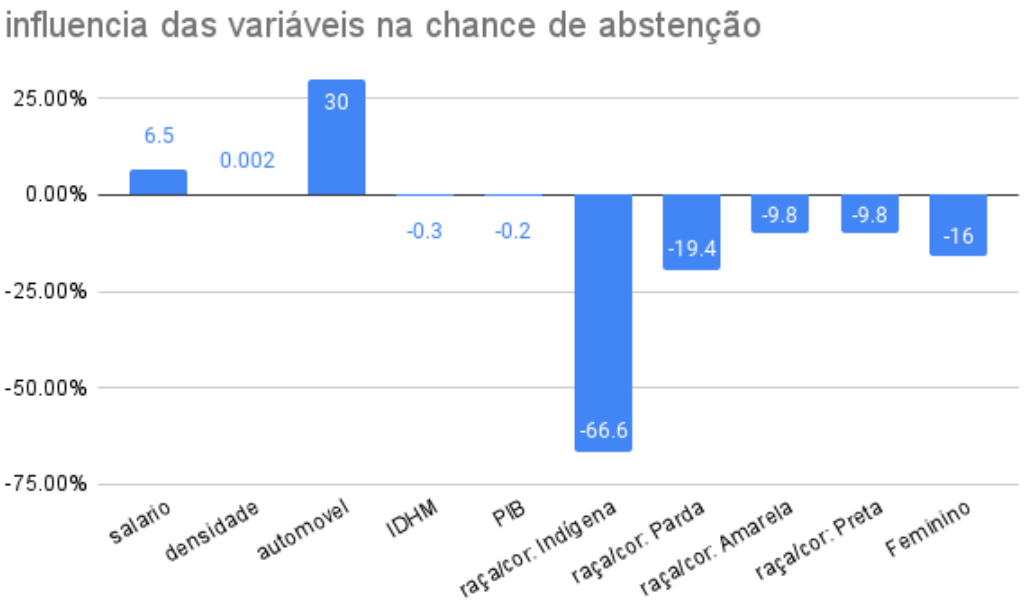

Gráfico 9 - Razão de possibilidades das variáveis de 2020

Em 2020, as raças/cor foram afetadas em uma proporção similar ao ano de 2019, obtendo como resultado que os indígenas foram os mais afetados. Contudo, não houve uma diferença significativa entre os resultados obtidos em 2019 para o ano de 2020, se comparado às outras raças/cor.

Em seguida, ressalta-se que ser do sexo feminino aumenta as chances em 16% de falta, aumento significavo quando comparado ao ano de 2019, que era 2.5%. Um fator que podem ter influenciado este aumento é, por exemplo, segundo a pesquisa “Sem parar: o trabalho e a vida das mulheres na pandemia”, da Gênero e Número e da Sempreviva Organização Feminista[7], que as mulheres brasileiras passaram a cuidar de alguém na pandemia, resultando na diminuição do tempo para se dedicar aos estudos e realização da prova.

Por fim, a renda que aumentava em 9.4% a chance de não faltar para cada salário mínimo em 2019, em 2020, a chance é de 6.5%, ou seja, apesar de o fator continuar influenciando nas desistências, seu impacto na pandemia foi menor conforme o modelo utilizado.

[7] https://mulheresnapandemia.sof.org.br/

3.3 *MACHINE LEARNING*

O modelo de clusterização foi utilizado na etapa de *machine learning* para encontrar um padrão nos fatores que pudessem influenciar nas desistências por estados. Além dos dados da diferença das desistências entre os anos de 2019 e 2020, foram incluídos os dados referentes ao PIB e ao IDH e também dados de casos e mortes de COVID-19 per capita. O principal padrão notado foi que em estados de maior IDH o impacto causado pelo COVID-19 nas desistências foi menor. Também foi notada uma divisão por regiões nos agrupamentos.

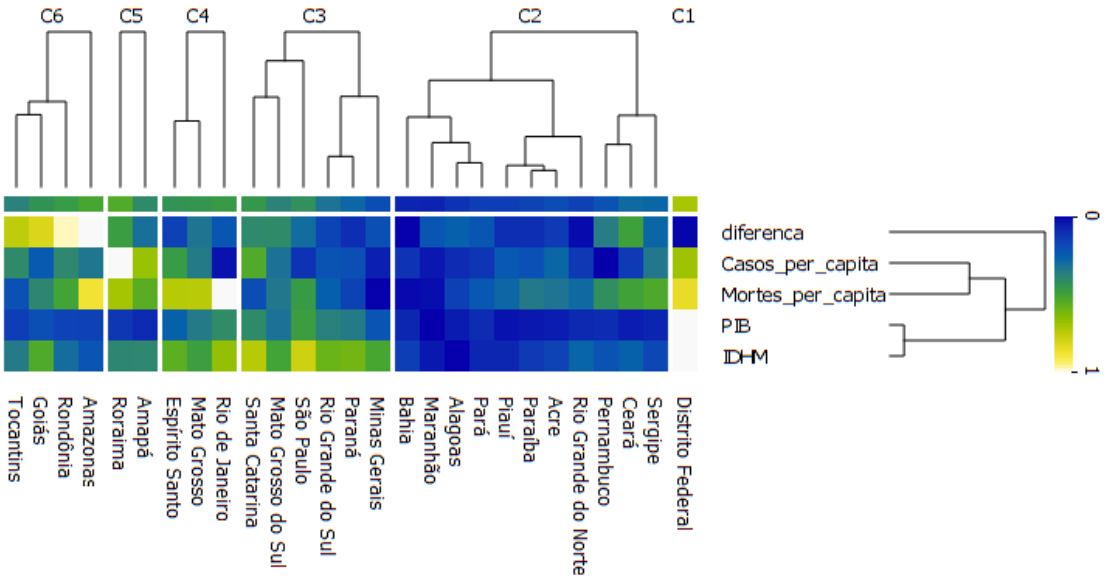

Gráfico 10 - Clusterização representada por mapa de calor

O primeiro *Cluster* é o Distrito Federal, um *outlier* que possui valores de PIB e IDH extremamente mais altos que o restante do país, além de altas taxas de morte e casos per capita e a menor diferença nas desistências.

O *Cluster* 2 (C2) possui baixos valores em todas as categorias e inclui os estados do Nordeste. Por outro lado, o *Cluster* 3 (C3), que possui o segundo maior IDH, teve baixas taxas de casos e mortes por COVID-19 per capita e teve leve impacto nas desistências. O *Cluster* 4 (C4) possui o IDH próximo ao do *cluster*

anterior e valores altos de morte per capita, teve um resultado das diferenças semelhante ao C2. Vale ressaltar que os estados do Sul estão no C3 e os do Sudeste no C3 e C4.

O *Cluster* 5 possui os maiores valores de mortes em estados de baixo IDH e a maior quantidade de casos per capita, sendo o segundo mais afetado pelas desistências. O *Cluster* 6 é o grupo com os valores de diferença mais alto e possui o segundo IDH mais baixo. Os estados do Norte estão principalmente nesses dois últimos *clusters* e é possível notar que embora haja uma relação entre o IDH e o impacto do COVID-19 nas desistências, os valores dessas categorias não serão sempre proporcionais pois existem outros fatores responsáveis por essa diferença que não foram analisados aqui.

Os gráficos *boxplot* a seguir comparam os valores medianos dos *clusters* em cada variável analisada:

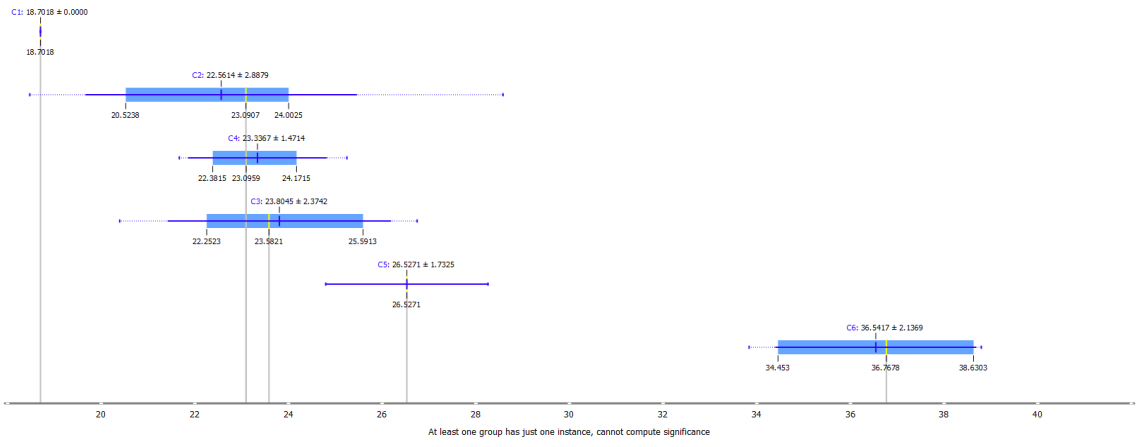

Gráfico 11 - *Boxplot* diferenças

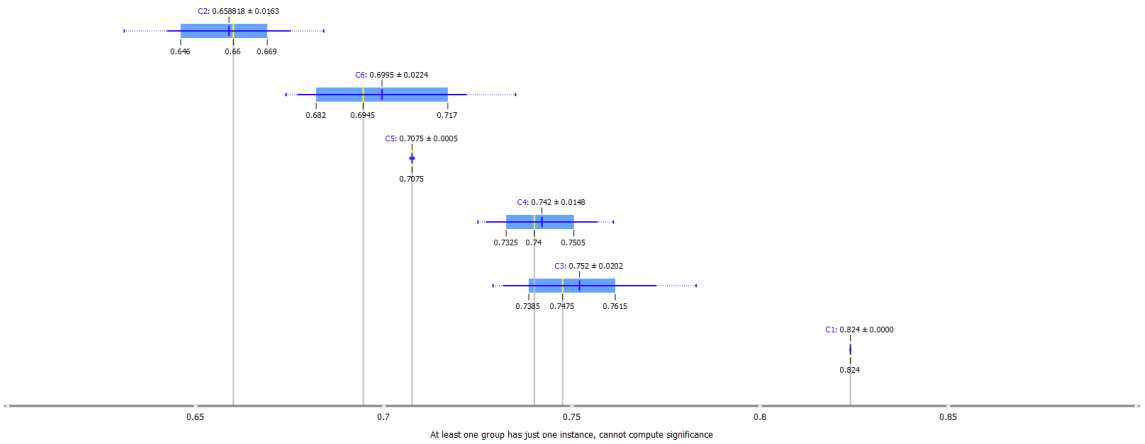

Gráfico 12 - *Boxplot* IDHM

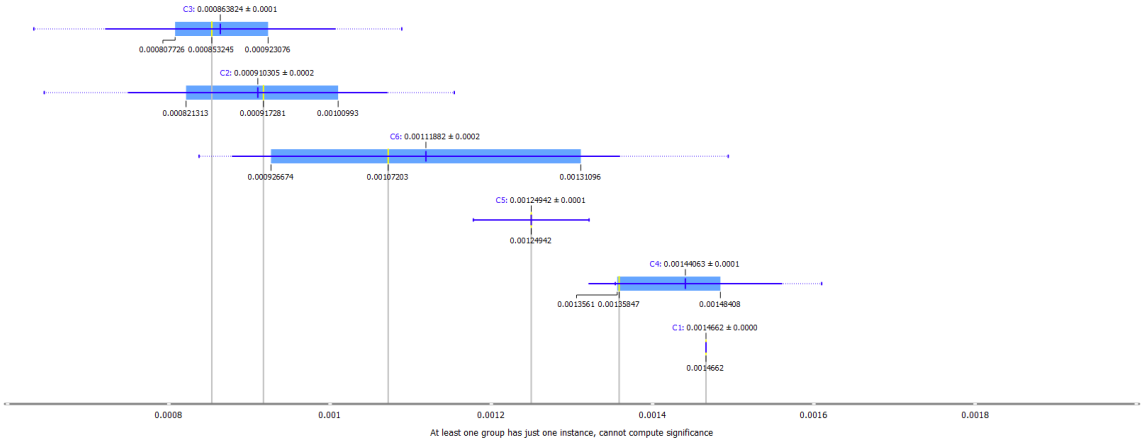

Gráfico 13 - *Boxplot* mortes per capita

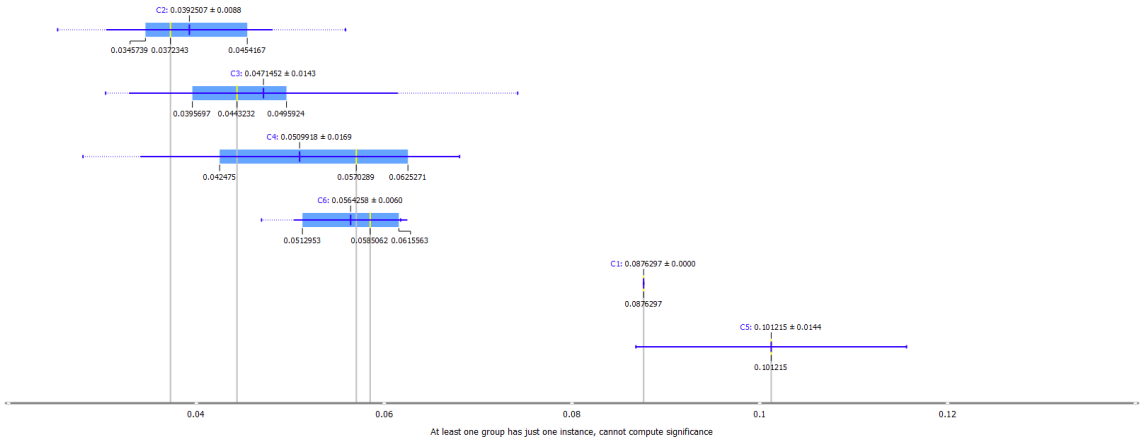

Gráfico 14 - *Boxplot* casos per capita

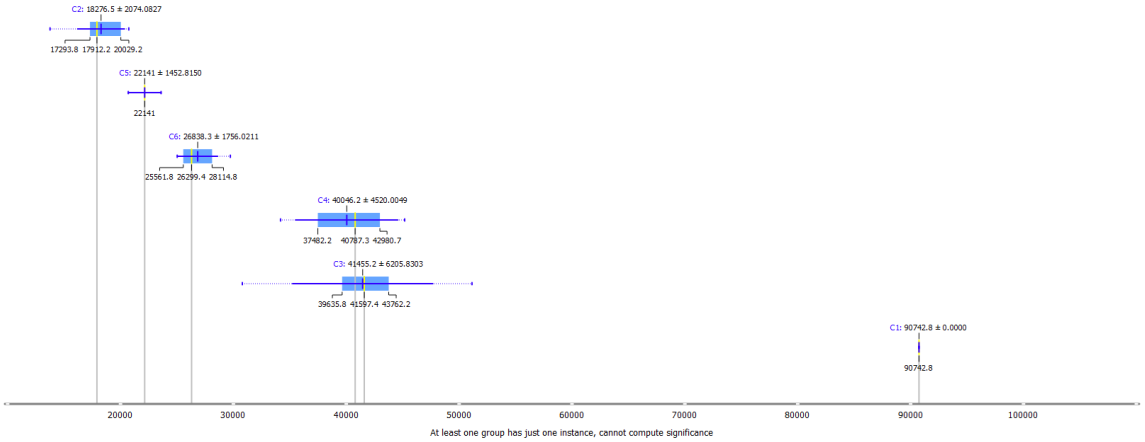

Gráfico 15 - *Boxplot* PIB

4 CONCLUSÃO

Durante a realização do trabalho, foram encontradas limitações, como o fato de que não foi possível encontrar os dados econômicos por município em 2020, como desemprego e desigualdade, que poderiam ajudar a explicar o modelo. Também, não foram encontrados dados socioeconômicos de 2020, o que também poderia trazer melhores resultados. Também, a equipe precisou trabalhar com uma grande quantidade de dados. Por fim, foi percebido durante a elaboração dos resultados que os fatores que influenciam o indivíduo a desistir não são os mesmos que influenciam na análise municipal.

Contudo, a equipe gerou resultados que podem auxiliar pesquisas mais aprofundadas sobre o estudo do aumento das desistências no ENEM durante a pandemia. Portanto, como trabalhos futuros poderá ser realizada a categorização dos municípios que aplicaram ou não aplicaram a prova para a análise, além de acrescentar outras variáveis que possam explicar melhor as descobertas do presente trabalho.

REFERÊNCIAS

CELSO, Luiz (2022). Introdução à Ciência de Dados: Tutoriais. Gitlab. Disponível em: <https://gitlab.com/introcienciadedados/tutoriais>. Acesso em: 27 jun. 2022.

GUIMARÃES, Amanda Munari (2022). Estatística: análise de correlação usando Python e R. Medium. Disponível em: <https://medium.com/omixdata/estat%C3%ADstica-an%C3%A1lise-de-correla%C3%A7%C3%A3o-usando-python-e-r-d68611511b5a>. Acesso em: 27 jun. 2022.