Universidade Federal do Rio de Janeiro

UFRJ

Análise de Dados – ENADE

Data Warehouse - 2021.1

Letícia Tavares da Silva 117210390

Sumário

1. Introdução

O escopo do projeto é dividido em três partes. Primeiramente, é construído um Data Warehouse seguindo uma modelagem dimensional estrela com os microdados do ENADE de anos de 2017, 2018 e 2019 disponibilizados no site do INEP. A partir de consultas ao Data Warehouse elaborado, é então realizada uma Análise Exploratória para por fim construir um modelo de aprendizagem de máquina com a finalidade de predizer o desempenho de um aluno com base em informações preenchidas nas provas como dados da instituição, características e perfil socioeconômico do aluno. As bases de dados importadas possuem 169 variáveis, mas no presente projeto algumas foram descartadas. Todo o trabalho está disponível em um repositório do Github [1].

2. Microdados ENADE

O ENADE, sigla para Exame Nacional de Desempenho de Estudantes, é uma exame anual que tem como objetivo avaliar o conhecimento obtido em relação aos conteúdos previstos na grade curricular do respectivo curso de graduação dos estudantes acadêmicos que estão prestes a se formar. Dessa forma, todo ano, alunos concluintes do ensino superior são submetidos a uma prova de conhecimento geral e específico composta por questões objetivas e discursivas. Além das questões para avaliar o conhecimento do estudante, as provas também apresentam questionários para obter informações socioeconômicas e experiências acadêmicas do estudante, além de seu preparo para a prova, percepção sobre a mesma e avaliação da Instituição. Todas essas informações obtidas em cada ano são disponibilizadas pelo INEP como microdados.

2.1. Obtenção dos Dados

Para o desenvolvimento deste trabalho foram utilizados os microdados do Enade disponibilizados pelo site do INEP [2]. A pesquisa se baseou em dados de todos os estudantes que realizaram o exame no ano de 2017, 2018 e 2019. Dentre os dados disponíveis estão:

- Informações do Curso, da Instituição de Ensino Superior e da sua localização
- Informações Sociais do Estudante como gênero e idade
- Avaliação - Formação Geral e Componente Específico
- Nota Geral e todas as notas que a compõem
- Presença do Aluno no Enade e nas seções das provas
- Gabarito das questões objetivas e as respostas marcadas pelo aluno
- Questionário do Estudante, com questões socioeconômicas e questões sobre sua vida acadêmica

Totalizando 182 atributos na base de 2017 e 169 em cada uma das outras bases. Essa diferença se dá por variáveis que estão presentes apenas na base de 2017 referentes às questões específicas do curso de licenciatura. Dessa forma, como esses atributos são específicos para um grupo de cursos e não estão presentes em todas as bases, dificultando assim uma integração para uma análise exploratória, foram descartados neste trabalho.

As informações são disponibilizadas no formato TXT acompanhadas de um dicionário de variáveis no formato XLS que apresenta uma aba com a descrição do que significa cada atributo e seus possíveis valores e outra aba com atributos identificadores de cada código de cidade disponibilizados pelo IBGE, já que os microdados apresentam apenas esses códigos na identificação do local do curso.

Além dos dados descritos acima, foram também utilizados neste trabalho as bases também fornecidas pelo INEP que apresentam o Conceito ENADE de cada ano e podem ser encontradas na página do INEP [3]. Essas bases foram utilizadas com a finalidade de obter atributos referentes a alguns códigos que aparecem no microdados do ENADE como nome da instituição.

Para realizar extração de todas os dados listados anteriormente de forma automatizada, utilizou-se o script "00_Download_Dados" desenvolvido em um notebook jupyter com linguagem python que ao ser executado, realiza o download dos microdados do enade e dos dados do Conceito enade e os salva em uma pasta denominada de "Dados" no mesmo diretório em que se encontra o script.

3. Modelagem de Dados

No presente trabalho foi utilizada a modelagem Estrela (Star Schema) caracterizada por poucas tabelas e relacionamentos, onde todas as tabelas de Dimensão se relacionam direta e unicamente com a tabela Fato, sendo assim um modelo simples e eficiente.

3.1. Tratamento de Dados

Para construção do modelo, o dataset passou foi avaliado e passou por tratamento de dados. O dataset apresenta algumas linhas que não apresentam notas das provas, escolhidas para serem fatos da modelagem e target do modelo, assim foi decidido pela remoção desses casos. Também foi optado pela remoção de nulos da coluna "QI_01" já que foi observado que essas amostras também apresentavam valores nulos em quase todas as outras colunas pertencentes ao questionário. Uma outra remoção foi a de amostras com idades duvidosas como 4 e 11 anos, assim foi criado um filtro para só incluir linhas que apresentam idade maior ou igual à 16. Assim, após todas as remoções feitas, o dataset final ficou com 1.291.772 linhas.

Outro tratamento feito foi a substituição de valores pelos seus significados apresentados no dicionário de dados nas colunas. Algumas colunas possuem valores diferentes em cada ano que possuíam o mesmo significado, nesses casos, foram obtidas informações de todos os dicionários para que todos os valores fossem substituídos, ocasionando numa mesma substituição para valores diferentes.

Todas essas manipulações podem ser observadas no script python "01_Criação_Banco_de_Dados.ipynb" na sessão onde as tabelas de dimensão e fato são construídas.

3.2. Modelo Estrela do ENADE

O modelo elaborado neste trabalho foi desenvolvido com o intuito de ser centrado nas notas das provas, objetivando fornecer dados para análise de fatores que influenciam no desempenho de cada

aluno que presta a avaliação. Dessa forma, o modelo apresenta como fato a prova que possui como métrica a nota final e as notas de formação geral e componente específico que a compoem, além de 11 tabelas de dimensão. A modelagem pode ser observada no diagrama abaixo que foi elaborado com a ferramenta *Visual Paradigm*:

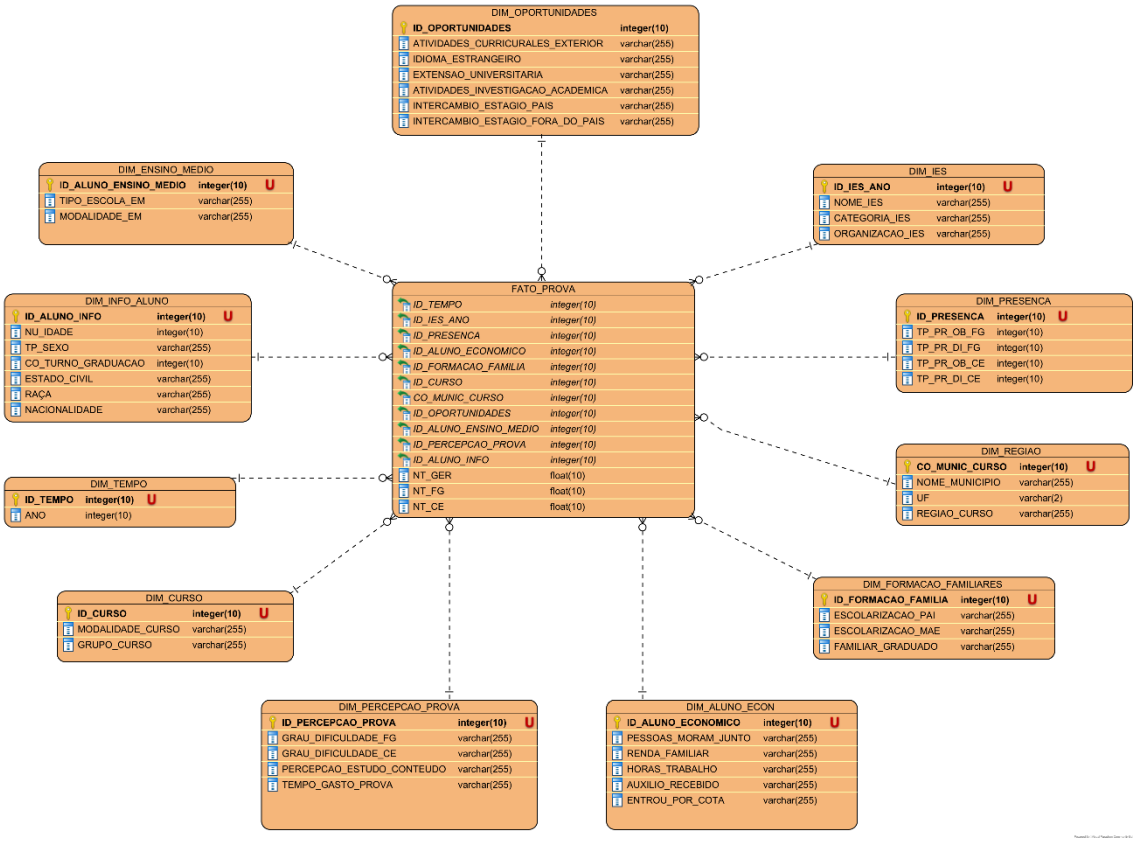

Figura 1: Modelagem dimensional Estrela Microdados ENADE [4]

Para selecionar as variáveis que compõem as tabelas, primeiro foram descartadas aquelas que não estavam presentes em todas as bases, sendo assim removidas 13 variáveis presentes apenas nos microdados do ano de 2017. Das 169 variáveis restantes, foram selecionadas as que mais pareciam estar relacionadas à nota do estudante, como dados da instituição de ensino do curso, dados econômicos do estudante, percepção da prova, entre outros. A seguir se encontra uma breve descrição de cada tabela presente no modelo elaborado e das variáveis que cada uma possui.

- **DIM_ALUNO_ECON**: possui variáveis importantes associadas a indicadores econômicos do estudante.

ID_ALUNO_ECONOMICO	Numérico	Chave Primária
PESSOAS_MORAM_JUNTO	Categórico	Quantidade de pessoas que moram junto com o estudante. Adaptação da coluna **QE_I07** onde os códigos foram substituídos pelos significados presente no dicionário.
RENDA_FAMILIAR	Categórico	Categoria de renda familiar do estudante. Adaptação da coluna **QE_I08** onde os códigos foram substituídos pelos significados presente no dicionário.
HORAS_TRABALHO	Categórico	Tempo que o estudante trabalha em média por semana. Adaptação da coluna **QE_I10** onde os códigos foram substituídos pelos significados presente no dicionário.

AUXILIO_RECEBIDO	Categórico	Indicação se o estudante recebe auxílio. Adaptação da coluna **QE_I12** onde os códigos foram substituídos pelos significados presente no dicionário.
ENTROU_POR_COTA	Categórico	Indicação se o estudante entrou por cota. Adaptação da coluna **QE_I15** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_TEMPO:** possui uma variável que descreve o tempo na granularidade de ano.

ID_TEMPO	Numérico	Chave Primária
ANO	Numérico	Ano da prova. A coluna foi renomeada, o nome original é **NU_ANO**.

- **DIM_IES:** possui variáveis referentes à Instituição em que o estudante cursa a graduação. A tabela foi construída utilizando também os dados da base Conceito Enade.

ID_IES_ANO	Numérico	Chave Primária
NOME_IES	Categórico	Nome da Instituição de Ensino Superior. Coluna foi construída com os dados presentes na base Conceito Enade.
CATEGORIA_IES	Categórico	Categoria Administrativa da Instituição de Ensino Superior. Adaptação da coluna **CO_CATEGAD** onde os códigos foram substituídos pelos significados presente no dicionário.
ORGANIZACAO_IES	Categórico	Organização Acadêmica da Instituição de Ensino Superior. Adaptação da coluna **CO_ORGACAD** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_ENSINO_MEDIO:** possui variáveis referentes ao ensino médio cursado pelo estudante. As variáveis fazem parte do questionário do estudante e os valores foram substituídos pelos significados apresentados no dicionário.

ID_ALUNO_ENSINO_MEDIO	Numérico	Chave Primária
TIPO_ESCOLA_EM	Categórico	Tipo de escola em que o estudante cursou o ensino médio. Adaptação da coluna **QE_I17** onde os códigos foram substituídos pelos significados presente no dicionário.
MODALIDADE_EM	Categórico	Modalidade do Ensino Médio cursado pelo aluno. Adaptação da coluna **QE_I18** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_FORMACAO_FAMILIARES:** possui variáveis associadas à escolarização de familiares do estudante. As variáveis fazem parte do questionário do estudante e os valores foram substituídos pelos significados apresentados no dicionário.

ID_FORMACAO_FAMILIA	Numérico	Chave Primária

ESCOLARIZACAO_PAI	Categórico	Escolarização do pai do estudante. Adaptação da coluna **QE_I04** onde os códigos foram substituídos pelos significados presente no dicionário.
ESCOLARIZACAO_MAE	Categórico	Escolarização do mãe do estudante. Adaptação da coluna **QE_I05** onde os códigos foram substituídos pelos significados presente no dicionário.
FAMILIAR_GRADUADO	Categórico	Existência de algum familiar graduado. Adaptação da coluna **QE_I21** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_CURSO:** possui variáveis referentes ao curso em que o estudante está se graduando.

ID_CURSO	Numérico	Chave Primária
MODALIDADE_CURSO	Categórico	Modalidade de ensino do curso. Adaptação da coluna **CO_MODALIDADE** onde os códigos foram substituídos pelos significados presente no dicionário.
GRUPO_CURSO	Categórico	Área de Enquadramento do curso. Adaptação da coluna **CO_GRUPO** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_REGIAO:** possui variáveis associadas à região a qual a Instituição em que o aluno presta o curso pertence em várias granularidades de divisões territoriais. Foi construída a partir do dicionário dos Microdados.

CO_MUNIC_CURSO	Numérico	Chave Primária
NOME_MUNICIPIO	Categórico	Município da Instituição. Construída a partir da aba **"MUNICÍPIOS"** presente na planilha do **Dicionário.**
UF	Categórico	UF referente ao Estado da Instituição. Construída a partir da aba **"MUNICÍPIOS"** presente na planilha do **Dicionário**.
REGIAO_CURSO	Categórico	Região da Instituição. Adaptação da coluna **CO_UF_CURSO** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_PERCEPCAO_PROVA:** possui variáveis associadas à percepção que o aluno teve após concluir a prova. As variáveis fazem parte do questionário de percepção da prova e os valores foram substituídos pelos significados apresentados no dicionário.

ID_PERCEPCAO_PROVA	Numérico	Chave Primária
GRAU_DIFICULDADE_FG	Categórico	Indicação do estudante sobre a dificuldade que apresentou na sessão de formação geral. Adaptação da coluna **CO_RS_I1** onde os códigos foram substituídos pelos significados presente no dicionário.
GRAU_DIFICULDADE_CE	Categórico	Indicação do estudante sobre a dificuldade que apresentou na sessão de componente específico.

		Adaptação da coluna **CO_RS_I2** onde os códigos foram substituídos pelos significados presente no dicionário.
PERC_EST_CONTEUDO	Categórico	Percepção do estudante sobre o estudo do conteúdo cobrado. Adaptação da coluna **CO_RS_I8** onde os códigos foram substituídos pelos significados presente no dicionário.
TEMPO_GASTO_PROVA	Categórico	Estimativa de tempo que o aluno levou para concluir a prova. Adaptação da coluna **CO_RS_I9** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_PRESENCA:** possui variáveis associadas à presença do aluno nas seções da Prova que indicam se aluno teve uma participação válida ou zerada por algum motivo como ausência ou anulação.

ID_PRESENCA	Numérico	Chave Primária
TP_PR_OB_FG	Booleano	Indicação de presença do estudante na parte objetiva da formação geral. Os valores da coluna foram alterados, onde os códigos que representavam algum tipo de participação no dicionário foram substituídos por 1 e todo o resto por 0.
TP_PR_DI_FG	Booleano	Indicação de presença do estudante na parte discursiva da formação geral. Os valores da coluna foram alterados, onde os códigos que representavam algum tipo de participação no dicionário foram substituídos por 1 e todo o resto por 0.
TP_PR_OB_CE	Booleano	Indicação de presença do estudante na parte objetiva do componente específico. Os valores da coluna foram alterados, onde os códigos que representavam algum tipo de participação no dicionário foram substituídos por 1 e todo o resto por 0.
TP_PR_DI_FG	Booleano	Indicação de presença do estudante na parte discursiva do componente específico. Os valores da coluna foram alterados, onde os códigos que representavam algum tipo de participação no dicionário foram substituídos por 1 e todo o resto por 0.

- **DIM_OPORTUNIDADES** possui variáveis associadas à opinião do estudante sobre oportunidades durante a graduação. As variáveis fazem parte do questionário do estudante e os valores foram substituídos pelos significados apresentados no dicionário.

ID_OPORTUNIDADES	Numérico	Chave Primária
ATIVIDADES_CURRICURALES_EXTERIOR	Categórico	Experiência em atividades extracurriculares no exterior. Adaptação da coluna **QE_I14** onde os códigos foram substituídos pelos significados presente no dicionário.
IDIOMA_ESTRANGEIRO	Categórico	Experiência em curso de idioma

		estrangeiro na Instituição. Adaptação da coluna **QE_I24** onde os códigos foram substituídos pelos significados presente no dicionário.
EXTENSAO_UNIVERSITARIA	Categórico	Oportunidades para participação em atividades de extensão na Universidade. Adaptação da coluna **QE_I43** onde os códigos foram substituídos pelos significados presente no dicionário.
ATIVIDADES_INVESTIGACAO_ACADEMICA	Categórico	Oportunidades para participação em projetos como IC na Instituição. Adaptação da coluna **QE_I44** onde os códigos foram substituídos pelos significados presente no dicionário.
INTERCAMBIO_ESTAGIO_PAIS	Categórico	Oportunidades para intercâmbio e/ou estágio no país. Adaptação da coluna **QE_I52** onde os códigos foram substituídos pelos significados presente no dicionário.
INTERCAMBIO_ESTAGIO_FORA_DO_PAIS	Categórico	Oportunidades para intercâmbio e/ou estágio fora do país Adaptação da coluna **QE_I53** onde os códigos foram substituídos pelos significados presente no dicionário.

- **DIM_INFO_ALUNO:** possui variáveis importantes associadas a indicadores sociais do aluno.

ID_ALUNO_INFO	Numérico	Chave Primária
NU_IDADE	Numérico	Idade do estudante.
TP_SEXO	Categórico	Gênero do estudante.
CO_TURNO_GRADUACAO	Numérico	Turno da graduação do estudante.
ESTADO_CIVIL	Categórico	Estado civil do estudante. Adaptação da coluna **QE_I01** onde os códigos foram substituídos pelos significados presente no dicionário.
RAÇA	Categórico	Raça autodeclarada do estudante. Adaptação da coluna **QE_I02** onde os códigos foram substituídos pelos significados presente no dicionário.
NACIONALIDADE	Categórico	Nacionalidade do estudante. Adaptação da coluna **QE_I03** onde os códigos foram substituídos pelos significados presente no dicionário.

- **FATO_PROVA:** Tabela fato do modelo que apresenta as notas da prova obtidas pelo estudante.:

ID_PROVA	Numérico	Chave Primária

ID_ALUNO_ECONOMICO	Numérico	Chave Estrangeira
ID_CURSO	Numérico	Chave Estrangeira
ID_ALUNO_ENSINO_MEDIO	Numérico	Chave Estrangeira
ID_FORMACAO_FAMILIA	Numérico	Chave Estrangeira
ID_IES_ANO	Numérico	Chave Estrangeira
ID_ALUNO_INFO	Numérico	Chave Estrangeira
ID_OPORTUNIDADES	Numérico	Chave Estrangeira
ID_PERCEPCAO_PROVA	Numérico	Chave Estrangeira
ID_PRESENCA	Numérico	Chave Estrangeira
CO_MUNIC_CURSO	Numérico	Chave Estrangeira
ID_TEMPO	Numérico	Chave Estrangeira
ID_ALUNO_INFO	Numérico	Chave Estrangeira
NT_ GER	Numérico	Nota geral da prova.
NT_ CE	Numérico	Nota do componente específico da prova.
NT_ FG	Numérico	Nota da formação geral da prova.

4. Banco de Dados

4.1. Alimentação do Banco de Dados

Para criação do banco de dados relacional deste trabalho foi escolhido a utilização do sistema de gerenciamento *sqlite* por possuir um gerenciamento mais simples através do python. Para alimentar o banco de dados com as tabelas criadas pela modelagem dimensional, foi utilizado o script "01_Criação_Banco_de_Dados.ipynb" em python que utiliza a biblioteca "sqlite3" para realizar a comunicação com o banco. Para isso, o script primeiro verifica se há o banco de dados "Enade_DW.sqlite" em seu diretório, se não , o cria. Em seguida, realiza todas as manipulações no dataset para criação das tabelas.

Com a conexão ao banco de dados já estabelecida e as tabelas criadas, para cada tabela, o script verifica se a tabela já existe no banco de dados, se não, a importa. Assim, se o script for executado com o banco de dados já alimentado, nada será importado.

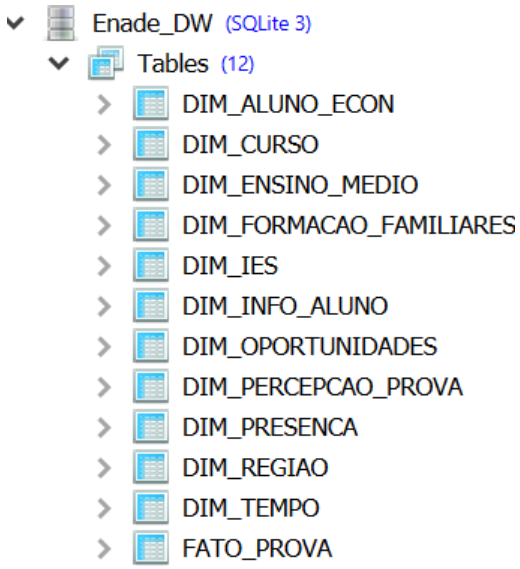

Figura 2: Visualização do Banco de Dados construído no SQLiteStudio versão 3.3.3

5. Análise Exploratória

Foi realizada uma Análise de Dados a partir do Data Warehouse construído. Nessa etapa, alguns dos dados foram utilizados para a realização de análises com o objetivo de avaliar a possível influência de alguns fatores socioeconômicos e geográficos na nota de um aluno.

A análise foi realizada no notebook "02_Analise_Exploratória_Dados.ipynb" utilizando *python* versão 3.8.8 como linguagem de programação onde foram utilizadas as bibliotecas *sqlite3* (para conexão com o banco de dados), *researchpy* (para gerar tabela de análise), *plotly* (para visualização). O notebook se encontra com as linhas que geram plots comentadas, pois o arquivo fica muito grande e exige muito de ram se forem descomentadas.

5.1. Nota por Gênero

	N	Mean	SD	SE	95% Conf.	Interval
TP_SEXO						
Feminino	738509	42.879	14.170	0.017	42.846	42.911
Masculino	553262	43.449	14.597	0.020	43.410	43.487

Figura 3: tabela com número de casos, média e desvio padrão da nota geral por grupo da coluna "TO_SEXO" que especifica o gênero do estudante.

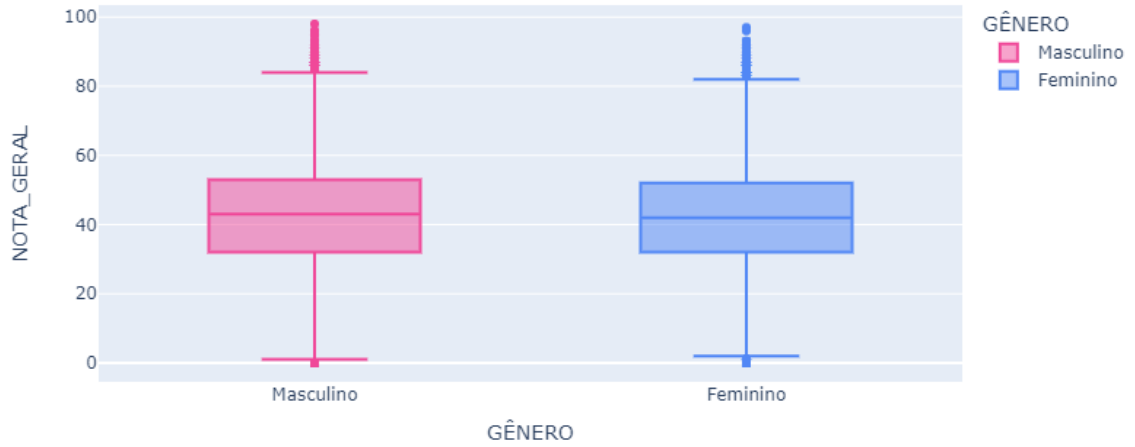

Figura 4: Boxplot apresentando a variação de notas para os grupos da coluna "TO_SEXO" que especifica o gênero do estudante. É possível ver que a maioria há variação significativa entre os grupos, já que apresentam média, dispersão e outliers muito similares.

Foi feita a seguinte pergunta "Será que o gênero dos estudantes possui correlação com os desempenhos deles na prova?", analisando a tabela e gráfico vemos que o número de inscritos do gênero feminino é consideravelmente maior, sendo 732.416 mil estudantes do gênero femino contra 540.806 do gênero masculino como podemos observar na tabela. No entanto, na análise da nota por gênero, vemos que há um desempenho muito semelhante dos estudantes de ambos os gêneros, com médias e desvio padrão muito próximos, assim como os outliers. Dessa forma, podemos concluir que o gênero não tem influência relevante no desempenho do aluno e assim não há evidência significativa para uma falta de investimento na educação em algum gênero específico.

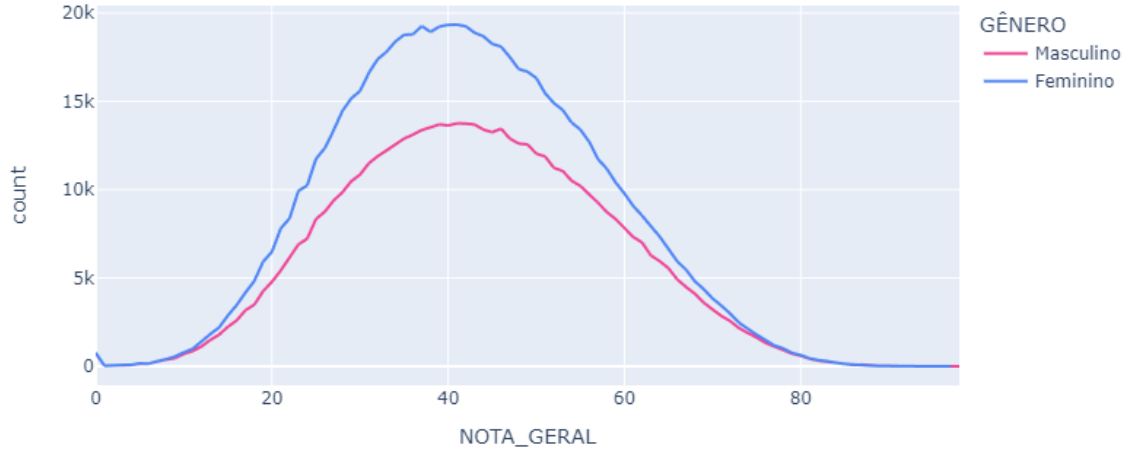

Figura 5: gráfico apresentando a distribuição de notas para os grupos da coluna "TO_SEXO" que especifica o gênero do estudante. Podemos notar que há mais estudantes inscritos do gênero feminino do que do sexo masculino, contudo a distribuição de nota é muito semelhante.

5.2. Nota por Raça

RAÇA	N	Mean	SD	SE	95% Conf.	Interval
Amarela	183666	42.946	14.252	0.033	42.881	43.011
Branca	124741	39.814	13.875	0.039	39.737	39.891
Indígena	98275	43.603	14.280	0.046	43.514	43.693
Não quero declarar	233147	41.517	13.942	0.029	41.461	41.574
Parda	127042	41.724	14.177	0.040	41.647	41.802
Preta	524900	44.933	14.482	0.020	44.894	44.972

Figura 6: tabela com número de casos, média e desvio padrão da nota geral por grupo da coluna "RAÇA" que específica a raça autodeclarada do estudante.

Para responder a pergunta "A raça ou etnia dos estudante impacta no desempenho na prova?", foi analisada a tabela acima onde vemos que as pessoas brancas constituem a raça com a maior parcela de inscritos nos anos analisados, onde cerca de metade dos inscritos se autodeclaram brancos. Tal representatividade é refletida quando analisamos as notas. No boxplot abaixo vemos alunos da raça branca tirarem, em média, notas maiores que todas as outras raças, e é a raça que mais apresenta outliers com notas altas. Uma explicação para esse resultado é que as raças parda, preta e indígena são na maior parte pertencentes à classe social baixa, que impacta diretamente no investimento na educação desses estudantes e assim, infelizmente, a raça do estudante tem impacto em seu desempenho.

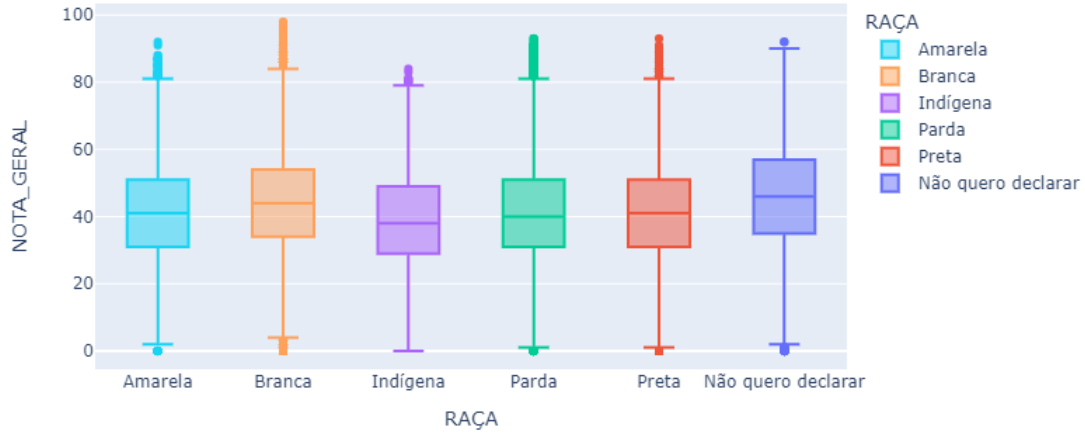

Figura 7: Boxplot apresentando a variação de notas para os grupos da coluna "RAÇA" que especifica a raça autodeclarada do estudante. É possível notar que há uma variação no desempenho do aluno dependendo de sua raça autodeclarada, com alunos autodeclarados brancos apresentando melhor desempenho.

5.3. Nota por Idade

	NU_IDADE
count	1291771.000
mean	28.466
std	7.661
min	16.000
25%	23.000
50%	26.000
75%	32.000
max	87.000

Figura 8: Descrição coluna Idade

Temos que os inscritos possuem idades bem variadas indo desde 16 anos até 87, todavia, 75% possuem até 31 anos, o que indica que poucos são os graduandos concluites acima da faixa dos 30. A fim de responder o questionamento "A idade dos estudante possui impacto nos desempenhos do Enade?", foi elaborado uma melhor visualização da relação da Idade com a nota, ao invés de plotar cada amostra no gráfico

de dispersão, os dados foram agrupados por idade e assim calcular a nota média de cada idade para então serem plotados. Vemos então pelo gráfico que a Idade possui uma correlação negativa, onde a medida que a idade vai aumentando, a nota média vai tendendo a cair, com estudantes entre 35 e 55 anos apresentando médias muito próximas, porém os inscritos com com 56 ou mais apresentam médias bem mais dispersas e mais baixa em sua maioria.

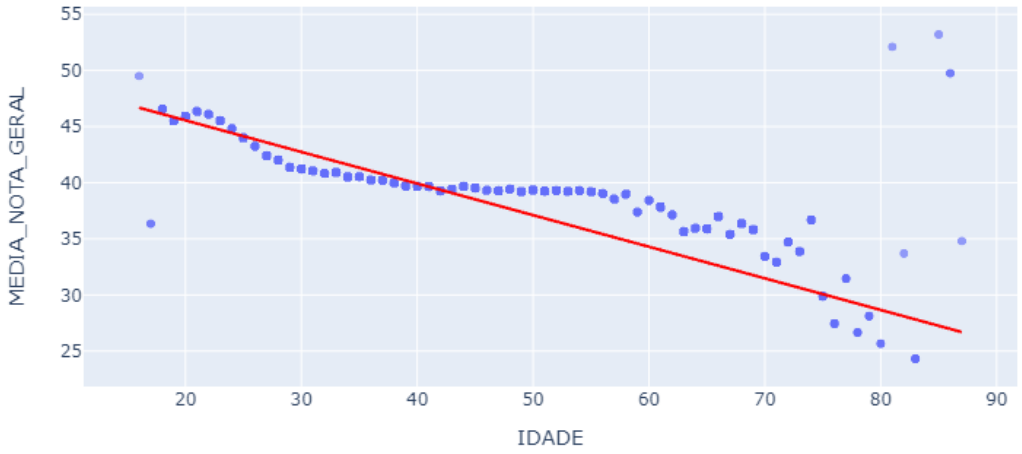

Figura 9: Correlação entre as variáveis presentes no dataset.

A tabela de correlação abaixo confirma a correlação negativa observada e avaliada em -0.149.

	NU_IDADE	NT_GER
NU_IDADE	1.000	-0.151
NT_GER	-0.151	1.000

Figura 10: Correlação entre a variável idade e a nota geral

5.4. Nota por Região

REGIAO_CURSO	N	Mean	SD	SE	95% Conf.	Interval
Centro-Oeste	101054	42.307	14.500	0.046	42.217	42.396
Nordeste	245348	42.866	14.362	0.029	42.809	42.923
Norte	78716	40.111	13.608	0.049	40.016	40.206
Sudeste	595015	43.644	14.356	0.019	43.607	43.680
Sul	271643	43.390	14.387	0.028	43.336	43.444

Figura 11: tabela com número de casos, média e desvio padrão da nota geral por grupo da coluna "REGIAO_CURSO" que especifica a região em que o curso que o Aluno realiza é oferecido.

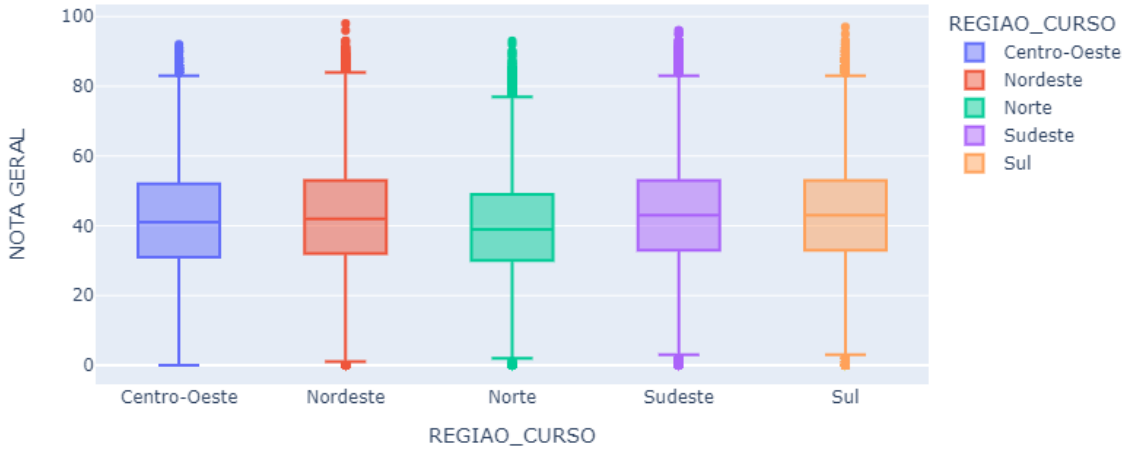

Figura 12: Boxplot apresentando a variação de notas para os grupos da coluna "REGIAO_CURSO" que especifica a região em que o curso que o aluno realiza é oferecido. Podemos visualizar que o desempenho dos alunos tende a variar de acordo com a Região.

Analisando a tabela e o gráfico de boxplot acima, podemos responder a questão "A região onde o curso é oferecido possui alguma influência nas notas das provas?" onde vemos uma diferença no desempenho dos estudantes das regiões, em que estudantes das regiões Sudeste e Sul apresentam melhores resultados e os da região Norte não vão tão bem na prova.

Ao aumentar a granularidade dos dados e analisar as notas por estado, vemos que a nota média por estado, podemos notar que se destacam os estados da Região Sudeste, Rio Grande do Sul, Distrito Federal com melhores desempenhos na nota. Tal observação sugere que boa parcela dos investimentos sejam direcionados para os estados dessa região.

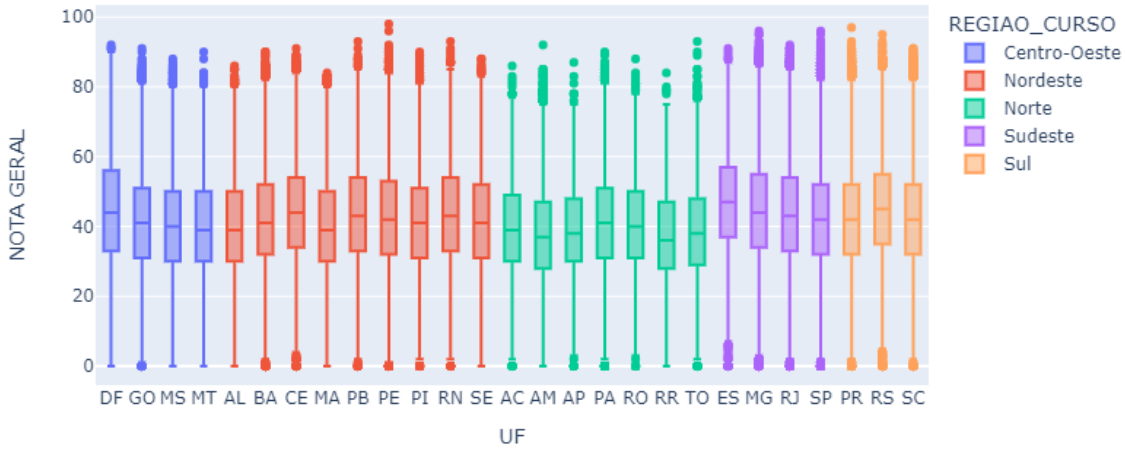

Figura 13: Boxplot apresentando a variação de notas para os grupos da coluna "UF" que especifica o Estado em que o curso que o aluno realiza é oferecido. Podemos visualizar que o desempenho dos alunos também tende a variar de acordo com a UF, não só com a Região, já que estados de uma mesma região apresentam variação nas notas diferentes.

5.5. Nota por Renda Familiar

RENDA_FAMILIAR	N	Mean	SD	SE	95% Conf.	Interval
Até 1,5 salário mínimo	258862	39.906	13.807	0.027	39.853	39.959
De 1,5 a 3 salários mínimos	364140	41.404	13.753	0.023	41.359	41.449
De 3 a 4,5 salários mínimos	269942	43.074	13.992	0.027	43.021	43.127
De 4,5 a 6 salários mínimos	143022	44.627	14.245	0.038	44.553	44.700
De 6 a 10 salários mínimos	146657	46.688	14.580	0.038	46.613	46.763
De 10 a 30 salários mínimos	91318	49.725	14.797	0.049	49.629	49.821
Acima de 30 salários mínimos	17830	50.460	15.203	0.114	50.236	50.683

Figura 14: tabela com número de casos, média e desvio padrão da nota geral por grupo da coluna "RENDA_FAMILIAR" que especifica a Renda Familiar média declarada pelo aluno.

Por último, foi feito o questionamento "Será que a Renda Familiar do estudante possui correlação com sua nota?". Através da tabela acima, podemos ver que a maior parcela dos inscritos possuem renda familiar de até 4.5 salários mínimos. Contudo, ao olharmos para o desempenho desses grupos no boxplot vemos uma clara tendência de aumento na nota à medida que a renda familiar aumenta. Uma possível explicação para essa situação é o fato de que quanto maior é o poder aquisitivo, maior são as oportunidades de investir em educação. Dessa forma, podemos afirmar que a renda familiar é um fator que possui uma correlação positiva com a nota no ENADE que também pode ser observada no boxplot a seguir.

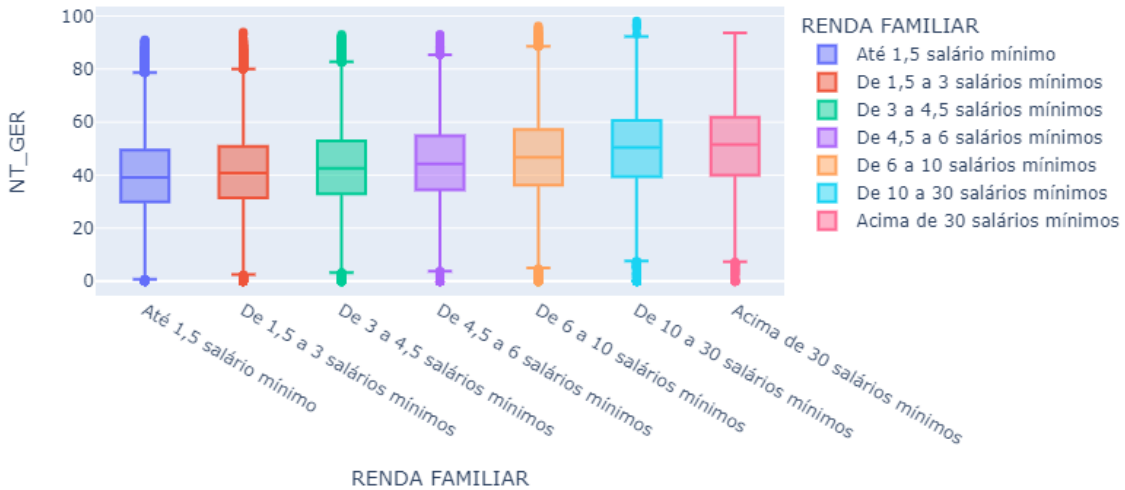

Figura 15: Boxplot apresentando a variação de notas para os grupos da coluna "RENDA_FAMILIAR" que específica a Renda Familiar média declarada pelo aluno. Podemos ver que a renda familiar do aluno é um fator que influencia seu desempenho, em que no geral, quanto maior a renda, melhor é o desempenho.

6. Modelo de Predição

Para esse trabalho foi elaborado um modelo de predição que com base alguns dos dados fornecidos pelo estudante (todos os selecionados para compor as dimensões do modelo

estrela desenvolvido) classifique se o estudante tirou uma nota acima da nota média do ano em que prestou a prova. Primeiramente, calculou-se a nota média de cada ano utilizando toda a base, com o conhecimento desses valores, criou-se a coluna target que indica se a nota geral do estudante é maior do que a nota média do ano, assim, se o estudante obteve uma nota maior do que a nota geral média do ano em que realizou a prova, a variável será igual à 1, caso contrário, será 0. O modelo tem então o objetivo de classificar essa variável como 1 ou 0 para cada estudante, classificando assim seu desempenho.

O modelo foi implementado na linguagem python e pode ser visto no script "03_Aprendizado_Random_Forest_Nota.ipynb". O algoritmo escolhido para compor o modelo foi o Random Forest com parâmetros iguais a máxima profundidade = 12, número de árvores = 5. O modelo foi treinado com 90% dos dados separados de forma aleatória e testado com os outros 10%. As estatísticas e matriz de confusão da predição podem ser vistas nas imagens abaixo.

Como grande parte das variáveis eram categóricas e todas as variáveis de entrada para o modelo necessitam ser numéricas, foi feita uma conversão dos dados onde variáveis que só assumem valores "Sim" ou "Não" foram transformadas em booleana, variáveis que apresentavam alguma informação numérica ou níveis de intensidade de forma categórica foram interpretadas numericamente como a coluna de Renda Familiar e por último, todas as as que sobraram foram convertidas utilizando o método *Mean Encoding*.

	precision	recall	f1-score	support
False	0.69	0.74	0.71	66696
True	0.70	0.65	0.67	62482
accuracy			0.69	129178
macro avg	0.69	0.69	0.69	129178
weighted avg	0.69	0.69	0.69	129178

Figura 16: Estatísticas da predição realizada pelo Modelo de Random Forest.

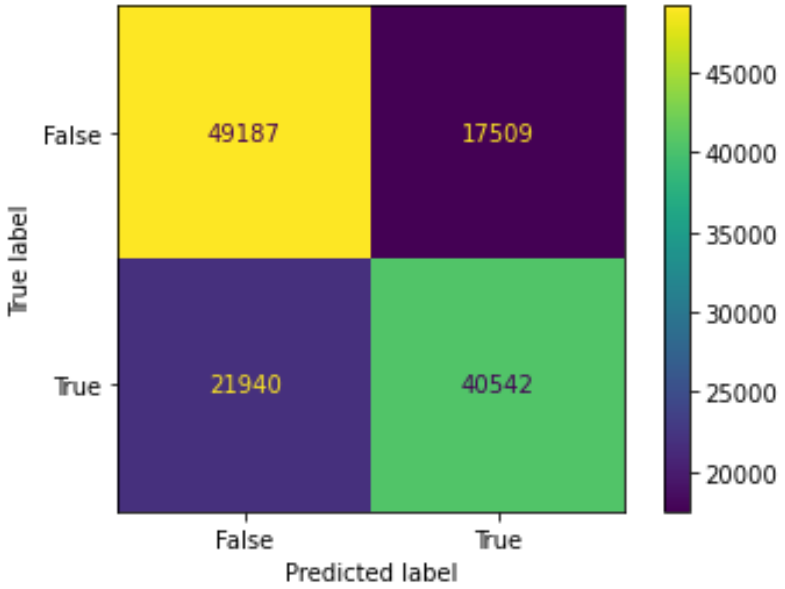

Figura 17: Matriz de Confusão do Modelo de Random Forest.

Vemos que o modelo obteve uma precisão de 70% e uma acurácia de 69% o que implica que o desempenho do estudante pode ser explicado em boa parte pelas informações acadêmicas, socioeconômicas e geográficas do estudante. A fim de observar o quanto cada variável impacta no desempenho do estudante, foi elaborado o gráfico abaixo que apresenta a importância de cada feature para o modelo.

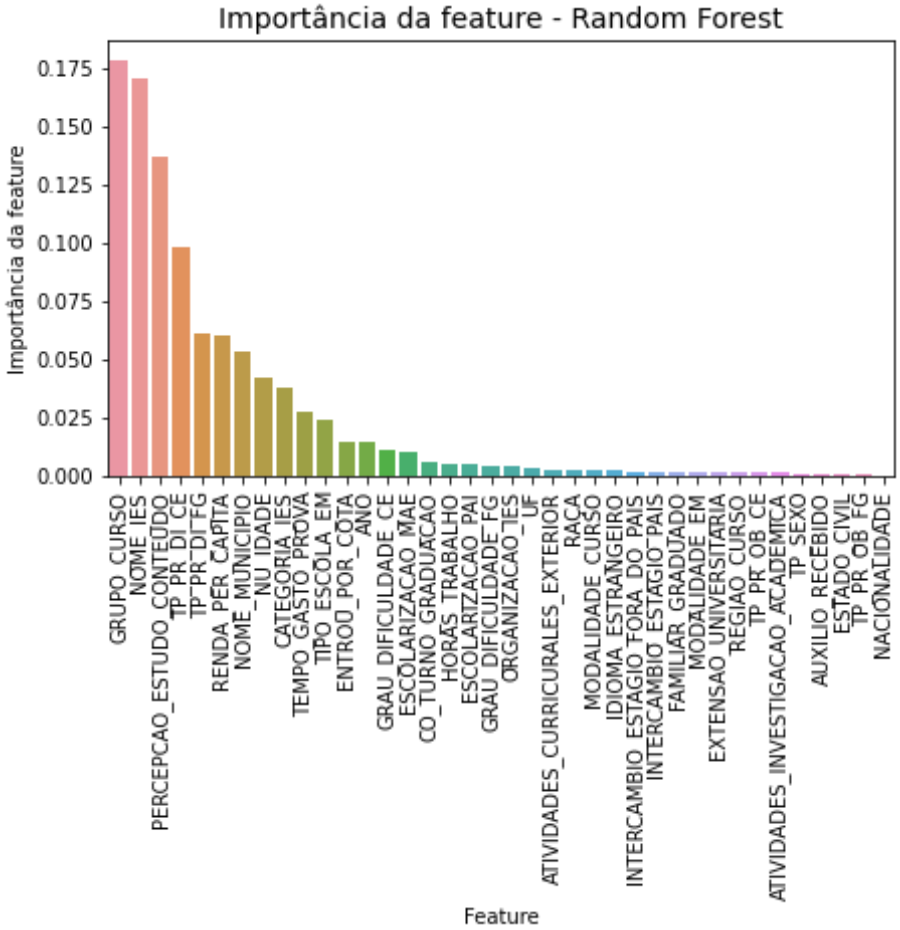

Figura 18: Gráfico que apresenta a importância de cada feature para o modelo de Random Forest construído.

A partir da análise do gráfico de barras acima, vemos o grupo a qual o curso do estudante pertence é o fator que mais impacta na sua classificação com base na nota final. Notamos também que o Instituição, o município em que o curso é oferecido e a Renda familiar do estudante também influencia diretamente no resultado. O grau de escolaridade dos pais também é um fator que pode determinar o desempenho do estudante.

7. Conclusão

Podemos concluir a partir da análise da importância das features para o modelo e de todas as análises feitas anteriormente que está relacionado com sua situação socioeconômica, a instituição em que realiza o curso, a região em que vive, as oportunidades que teve durante a graduação e nível de escolaridade de seus familiares. Vemos então que há uma necessidade de investimento na educação principalmente de minorias pertencentes à famílias menos favorecidas e em estados no Norte e Nordeste.

8. Ferramentas Utilizadas

Na tabela abaixo se encontram as ferramentas utilizadas na elaboração desse trabalho e um endereço eletrônico de referência de onde podem ser encontradas.

Ferramenta	Versão	Motivo Escolha
Jupyter Notebook	6.3.0	Foi a ferramenta escolhida para implementação pois oferece a opção de desenvolvimentos de notebooks e acesso de arquivos locais.
Python	3.8.8	Foi a linguagem escolhida por exigir menos do computador e pelo fato da estudante ter maior domínio com modelagem e visualização de dados com a linguagem já que lida com isso no estágio.
SQlite	3.36.0	Foi escolhida por ser um sistema que possui um gerenciamento mais simples através do python.
SQliteStudio	3.3.3	Foi escolhida por ser uma interface de fácil manipulação de um banco de dados em Sqlite.
Git bash	2.33.0	Foi escolhida pois fornece um jeito rápido e eficiente de manipular repositórios no Github.
Visual Paradigm	16.3	Foi escolhida pois foi uma ferramenta bastante utilizada pelo professor durante o curso.

9. Referências

1 – Repositório no Github contendo o trabalho. Disponível em:
< https://github.com/leticiatavaresds/DW_ENADE> Último Acesso em: 29 de nov. de 2021.

2 - Microdados do Exame Nacional de Desempenho dos Estudantes. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade> Último Acesso em: 29 de nov. de 2021.

3 - Dados Conceito Exame Nacional de Desempenho dos Estudantes. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/indicadores-de-qualidade-da-educacao-superior/outros-documentos> Último Acesso em: 29 de nov. de 2021.

4 – Diagrama em png do modelo dimensional estrala elaborado. Disponível em: <https://raw.githubusercontent.com/leticiatavaresds/DW_ENADE/main/Modelo%20Dimensional%20Estrela/ENADE.png> Último Acesso em: 29 de nov. de 2021.

5 - sklearn.ensemble.RandomForestClassifier Documentação. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> Último Acesso em: 29 de nov. de 2021.

6 - Improve your classification models using Mean /Target Encoding. Disponível em: <https://medium.datadriveninvestor.com/improve-your-classification-models-using-mean-target-encoding-a3d573df31e8> Último Acesso em: 29 de nov. de 2021.

7 - Um guia completo para modelagem dimensional. Disponível em: <https://www.astera.com/pt/tipo/blog/guia-de-modelagem-dimensional/> Último Acesso em: 29 de nov. de 2021.

8 - Slite3 Documentação. Disponível em: < https://docs.python.org/3/library/sqlite3.html> Último Acesso em: 29 de nov. de 2021.