RELATÓRIO

Daniel Bayerl Vieira 117234580

Repositório do GitHub

1-Fluxo de coleta e carga dos dados.

1.1- Coleta

O fluxo e coleta dos dados é feito a partir de um notebook chamado “Fluxo.ipynb”, que realiza a coleta dos dados utilizando as bibliotecas (requests, io e zipfile), que são utilizadas para realizar o download e salvar os dados necessários, que são:

-Dados do Enade de 2017,2018 e 2019, todos disponíveis aqui.

-Dados do Grupo, Curso e da Instituição para poder popular as tabelas do modelo dimensional, que estão disponíveis aqui.

1.2- Tratamento

Após os dados serem coletados, são necessários alguns tratamentos que foram feitos utilizando o (pandas e numpy), pois as bases de 2018,2018 e 2019 não são exatamente iguais, com algumas pequenas diferenças em códigos de algumas colunas, para isso foram tratadas:

-CO_MODALIDADE que é o código que diz se o aluno tem o modelo de ensino a distância ou presencial, o ano de 2018 tinha os códigos 0 e 1, enquanto os demais tinham os códigos 1 e 2, logo foi feito um ajuste nos dados de 2018 para igualar as codificações.

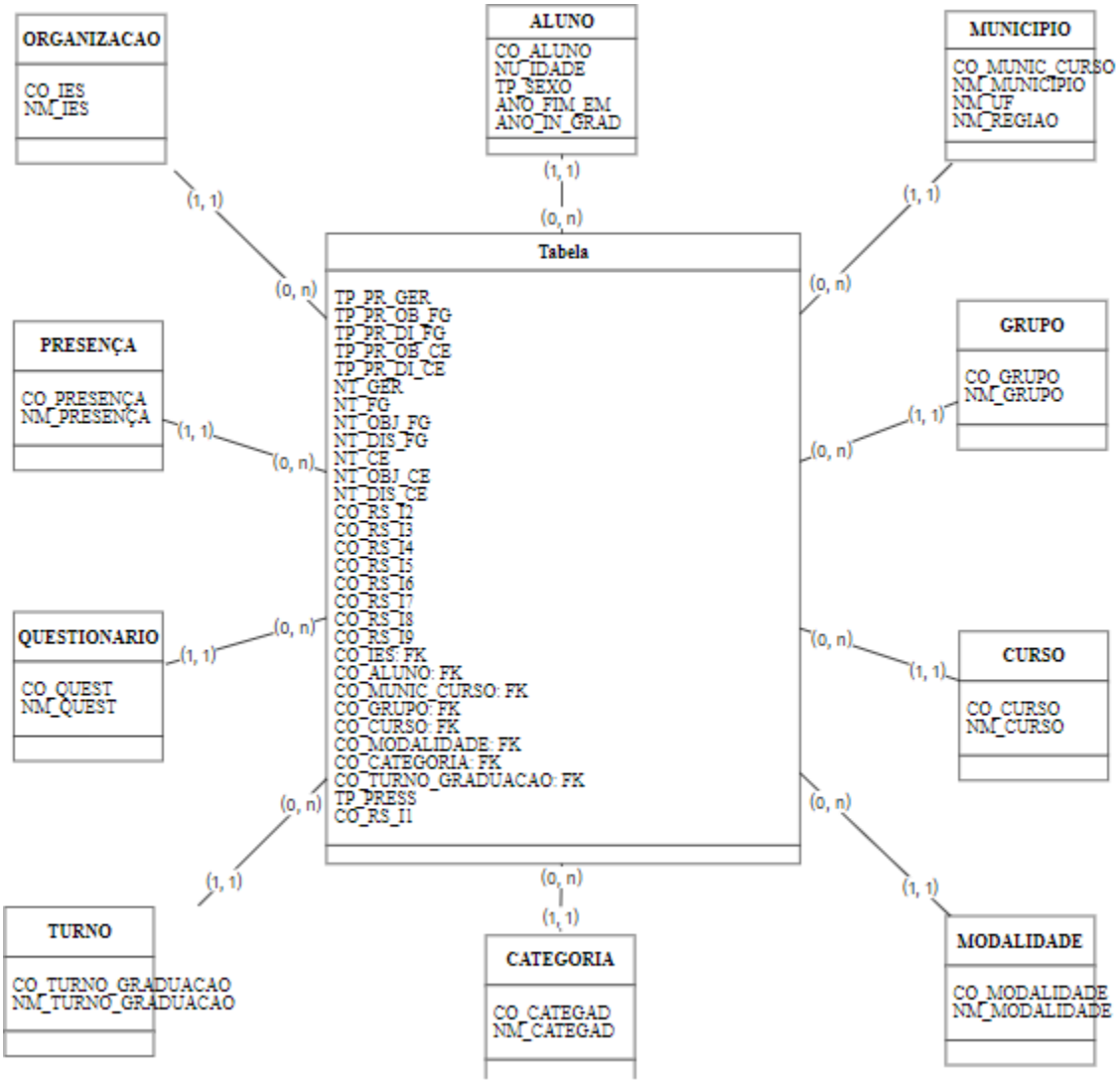
-CO_CATEGAD que é o código da categoria da instituição, como pública federal,privada com fins lucrativos, entre outros, só que nos anos de 2018 e 2019 as categorias foram divididas em várias que eram parecidas, então achei melhor unificar para o padrão 2017, que é mais intuitivo e mais direto.

Além disso, os campos de Notas estavam como strings, então foi feito um tratamento para serem armazenados como float.

As bases auxiliares que foram mencionadas anteriormente, não tinham problemas de codificação diferente, então apenas foram filtradas as colunas de interesse, que eram os códigos e nomes para Grupo e Instituição, o Curso ficou como uma concatenação de Grupo, Instituição e Local, já que não foi encontrada uma base com uma definição para o mesmo.

1.3- Carga

A carga foi efetuada em dois locais diferentes, um deles é no banco de dados que foi criado utilizando o seguinte modelo criado na ferramenta (BrModelo):



As tabelas MODALIDADE, CATEGORIA, TURNO, PRESENCA e QUESTIONARIO, foram criadas dentro do próprio código, pois seus códigos e definições estão presentes nos arquivos baixados do Enade, no caminho (‘DATA/201X/1.LEIA-ME/), aqui em todos os anos temos a tabela dicionário, que foi utilizada para montar as tabelas mencionadas.

As tabelas GRUPO, CURSO e ORGANIZACAO, foram montadas a partir dos dados auxiliares que foram baixados e estão localizados dentro de (‘DATA/201X’), para cada um dos anos.

A tabela NOTAS e ALUNO, foram preenchidas com os dados das tabelas baixadas do Enade de todos os anos.

A outra maneira de carga dos dados é salvar todas as tabelas que estão presentes no banco na pasta (‘tabelas’) em formato csv, garantindo assim que caso ocorra algum problema

em salvar os dados, eles possam ser acessados pelo arquivo de análise, que vai ser mencionado posteriormente.

Na segunda célula do arquivo (‘Fluxo.ypynb’), estão localizados alguns parâmetros que devem ser alterados para que a carga ocorra no banco de dados, que é o usuário e senha de um usuário do banco de dados mysql que estiver instalado no seu computador, e a variável save_sql, que por padrão estará definida como false, deve ser modificada para true, caso seja desejado que os dados sejam salvos no banco de dados. O programa já cria a database enade, então somente os parâmetros acima são necessários, é importante que não tenha uma database enade já criada, caso contrário ela será dropada.

É recomendado não trabalhar com o banco de dados, o que é recomendado já que vai adiantar muito o tempo de execução do programa, que pode levar mais de 20 minutos, por apresentar uma quantidade expressiva de dados, logo apenas rode o programa sem alterar nenhuma variável, que ele irá funcionar corretamente, e demorar apenas alguns minutos para fazer todos os processos descritos até o momento.

As bibliotecas necessárias para o funcionamento do código são instaladas automaticamente quando for executada a primeira célula do notebook.

2-Leitura e Análise dos dados.

2.1- Leitura

A Leitura assim como a carga, podem ser feitas de duas maneiras, diretamente do banco, ou dos arquivos csv criados no processo de carga, para isso haverá uma variável read_sql na segunda célula do código que vai determinar se vai ser lido do banco de dados ou não, ela só deve ser alterada para True, caso seja realizado com sucesso a carga dos dados no banco de dados.

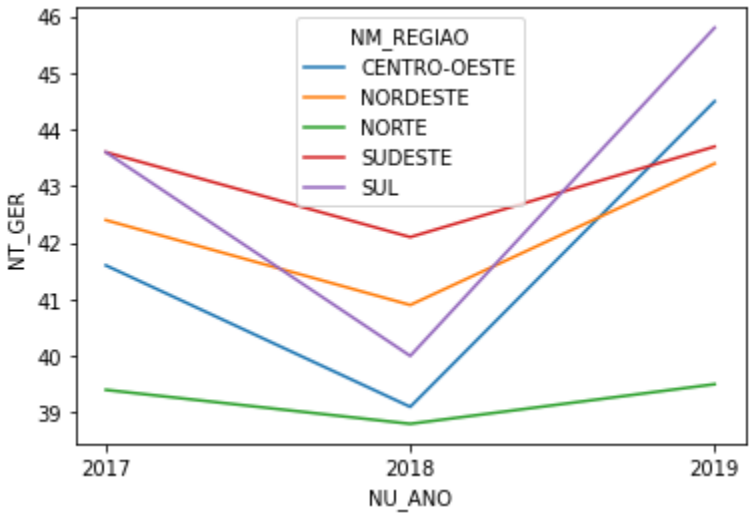
Esse início do código basicamente ele carrega nos DataFrames do pandas, os dados que foram carregados no arquivo (‘Fluxo.ipynb’).

2.2- Analises

Nessa parte do código foram realizadas 5 análises com um gráfico e uma tabela, todos eles apresentando o ano no eixo X e a nota geral no eixo Y, avaliados em cima de algumas métricas para comparação.

2.2.1 - Quais regiões tem as maiores notas?

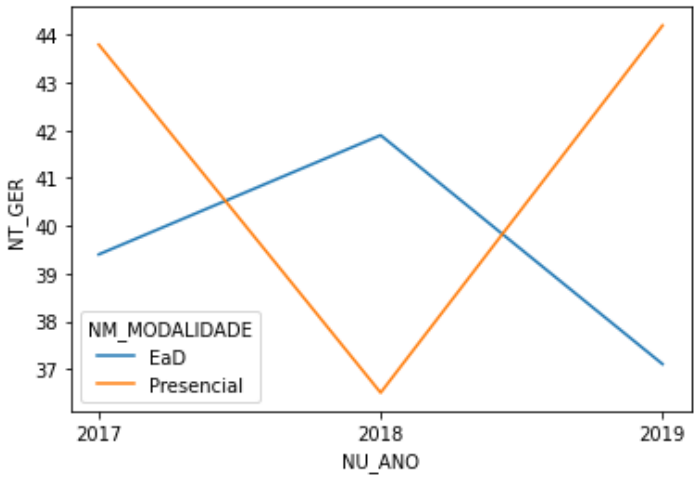
	NM_REGIAO	NU_ANO	NT_GER
0	CENTRO-OESTE	2017	41.6
1	CENTRO-OESTE	2018	39.1
2	CENTRO-OESTE	2019	44.5
3	NORDESTE	2017	42.4
4	NORDESTE	2018	40.9
5	NORDESTE	2019	43.4
6	NORTE	2017	39.4
7	NORTE	2018	38.8
8	NORTE	2019	39.5
9	SUDESTE	2017	43.6
10	SUDESTE	2018	42.1
11	SUDESTE	2019	43.7
12	SUL	2017	43.6
13	SUL	2018	40.0
14	SUL	2019	45.8)



Podemos ver que os alunos do Sudeste apresentaram um melhor rendimento até 2018, porém foram ultrapassados tanto pelo Centro-Oeste como o Sul, no ano de 2019, e o Norte apresenta o pior rendimento em todos os anos.

2.2.2 - Qual modalidade de ensino apresenta os melhores resultados?

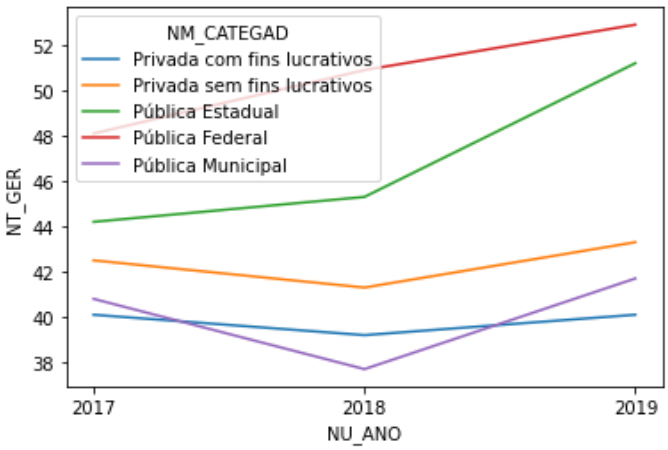
	NM_MODALIDADE	NU_ANO	NT_GER
0	EaD	2017	39.4
1	EaD	2018	41.9
2	EaD	2019	37.1
3	Presencial	2017	43.8
4	Presencial	2018	36.5
5	Presencial	2019	44.2)



Podemos ver que o presencial e o EaD tiveram uma troca de liderança em 2018 e uma volta ao patamar de 2017 em 2019, com o Presencial se mostrando melhores resultados, e uma média dos 3 anos, superior.

2.2.3 - Qual Categoria de instituição tem notas mais altas?

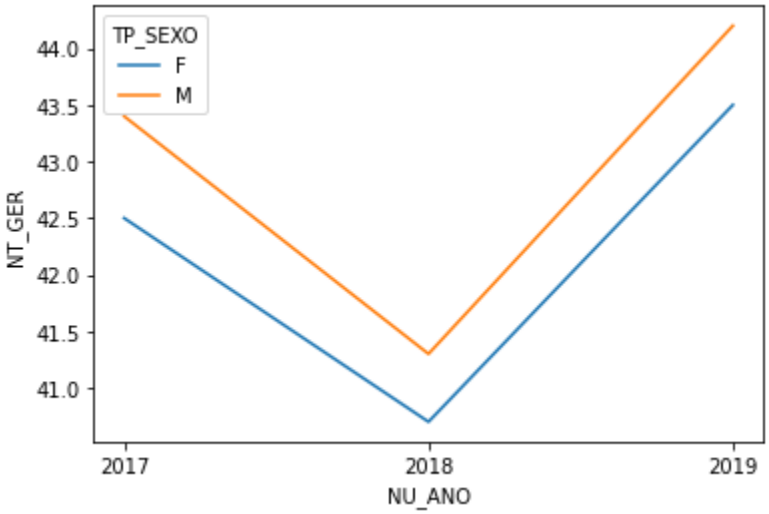
	NM_CATEGAD	NU_ANO	NT_GER
2	Privada com fins lucrativos	2017	40.1
3	Privada com fins lucrativos	2018	39.2
4	Privada com fins lucrativos	2019	40.1
5	Privada sem fins lucrativos	2017	42.5
6	Privada sem fins lucrativos	2018	41.3
7	Privada sem fins lucrativos	2019	43.3
8	Pública Estadual	2017	44.2
9	Pública Estadual	2018	45.3
10	Pública Estadual	2019	51.2
11	Pública Federal	2017	48.1
12	Pública Federal	2018	50.9
13	Pública Federal	2019	52.9
14	Pública Municipal	2017	40.8
15	Pública Municipal	2018	37.7
16	Pública Municipal	2019	41.7)



O que se mostra bem claro é o destaque das universidades públicas federais, seguida pelas estaduais e depois com uma diferença bem maior, as privadas e a pública municipal mais distantes das duas primeiras.

2.2.4 - Qual sexo apresenta melhores resultados?

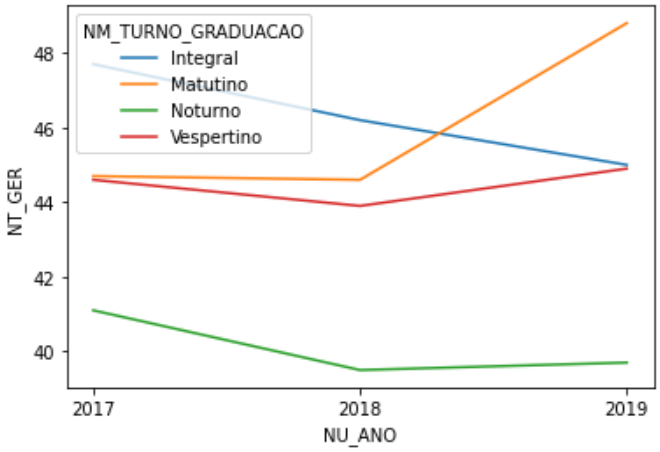
	TP_SEXO	NU_ANO	NT_GER
0	F	2017	42.5
1	F	2018	40.7
2	F	2019	43.5
3	M	2017	43.4
4	M	2018	41.3
5	M	2019	44.2)



Podemos ver que os estudantes do sexo masculino mantém uma diferença constante das estudantes do sexo feminino, porém não apresentam grandes diferenças, geralmente estando entre 0.5 e 1 ponto apenas.

2.2.5 - Estudantes de qual turno apresentam melhores notas?

	NM_TURNO_GRADUACAO	NU_ANO	NT_GER
0	Integral	2017	47.7
1	Integral	2018	46.2
2	Integral	2019	45.0
3	Matutino	2017	44.7
4	Matutino	2018	44.6
5	Matutino	2019	48.8
6	Noturno	2017	41.1
7	Noturno	2018	39.5
8	Noturno	2019	39.7
9	Vespertino	2017	44.6
10	Vespertino	2018	43.9
11	Vespertino	2019	44.9)



Podemos ver que os estudantes do turno integral e matutino trocaram a liderança entre os anos de 2017 e 2019, seguidos pelo turno vespertino não muito distante, e o noturno com um desempenho mais distante dos demais.

2.3- Aprendizagem

2.3.1- Método

Para a aprendizagem foi utilizado o método de random forest da biblioteca (sklearn) do python, e para isso foi apenas necessário passar as colunas que vão ser utilizadas para aprender a informação, e a coluna com a informação que desejamos que seja aprendida, depois disso basta dividir os dados que vão ser usados para treinamento e teste, para essa aprendizagem foi utilizado 0.8 para teste e 0.2 para treino, mas mesmo utilizando 0.5 para teste e 0.5 para treino, foram obtidos quase os mesmos resultados, mostrando que os dados tem uma boa constância nas suas informações, não precisando de um grande volume de dados para aprender.

Após isto, basta chamar a função RandomForestRegressor do (sklearn), e utilizar o seu resultado nos valores separados para teste.

2.3.2- Objetivo e resultados.

Foi realizada uma aprendizagem que já retornou um bom resultado na primeira tentativa.

A primeira foi tentar estimar a nota com base no turno, modalidade, categoria, tipo de presença e a instituição. Com isso, primeiramente foi passado os valores da nota para inteiros antes de realizar o processo de aprendizagem. Com isso a validação foi feita através da comparação do teste com os resultados gerados pelo modelo, e definindo um acerto como uma diferença menor ou igual a 1 entre os dados, já que poderia ocorrer algumas diferenças por arredondamento. Com isso obtivemos um resultado de aproximadamente 61% de acurácia.

Ferramentas utilizadas

Modelagem:
- BrModelo - https://app.brmodeloweb.com/

Banco de Dados:
- MySQL - https://www.mysql.com

Coleta, Tratamento, Carga e Análises:
- Python - https://www.python.org
- Pandas - https://pandas.pydata.org
- Numpy - https://numpy.org
- Sklearn - https://scikit-learn.org/stable/
- Seaborn - https://seaborn.pydata.org
- Requests - https://pypi.org/project/requests/
- ZipFile - https://docs.python.org/3/library/zipfile.html
- Io - https://docs.python.org/3/library/io.html
- Sqlalchemy - https://www.sqlalchemy.org