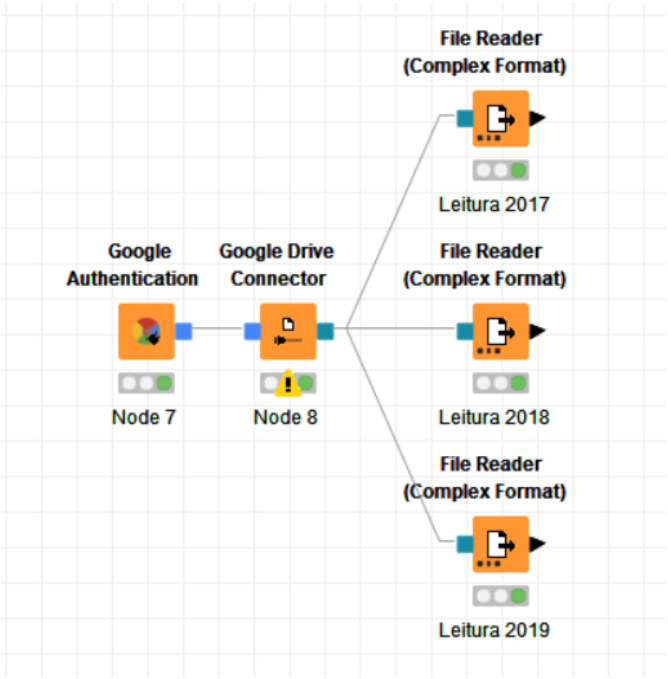
Prova Individual

Thiago Outeiro Pereira Damasceno
DRE: 116038363

1. Coleta de dados

Os microdados do ENADE podem ser encontrados no site:
https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade
Além disso para auxilio da turma, foi disponibilizado pelo professor um repositório no github, sendo ele: https://github.com/LINE-PESC/Datasets, porém ao tentar recuperar os dados, obtive alguns problemas com o repositorio do professor, e utilizei os dados baixados direto do site e os adicionei em meu googleDrive no link:
https://drive.google.com/drive/folders/1ULCZak3HH6qQ66pWU6CurbvCUOGiYPxn?usp=sharing.
Ao baixar os arquivos é possível identificar 3 pastas, “Leia-me”, “Inputs” e “Dados”, na pasta Leia-me é onde existe um dicionário das terminologias utilizadas nos arquivos de dados e legendas sobre como ele é classificado, além disso também existem dois exemplos dos questionários feitos para os estudantes, na pasta Inputs temos arquivos que parecem ser para a leitura dos dados, mas estes não são relevantes, e na pasta Dados é onde temos os dados de fato.

Para a coleta de dados, utilizei o KINME utilizando o seguinte fluxo:



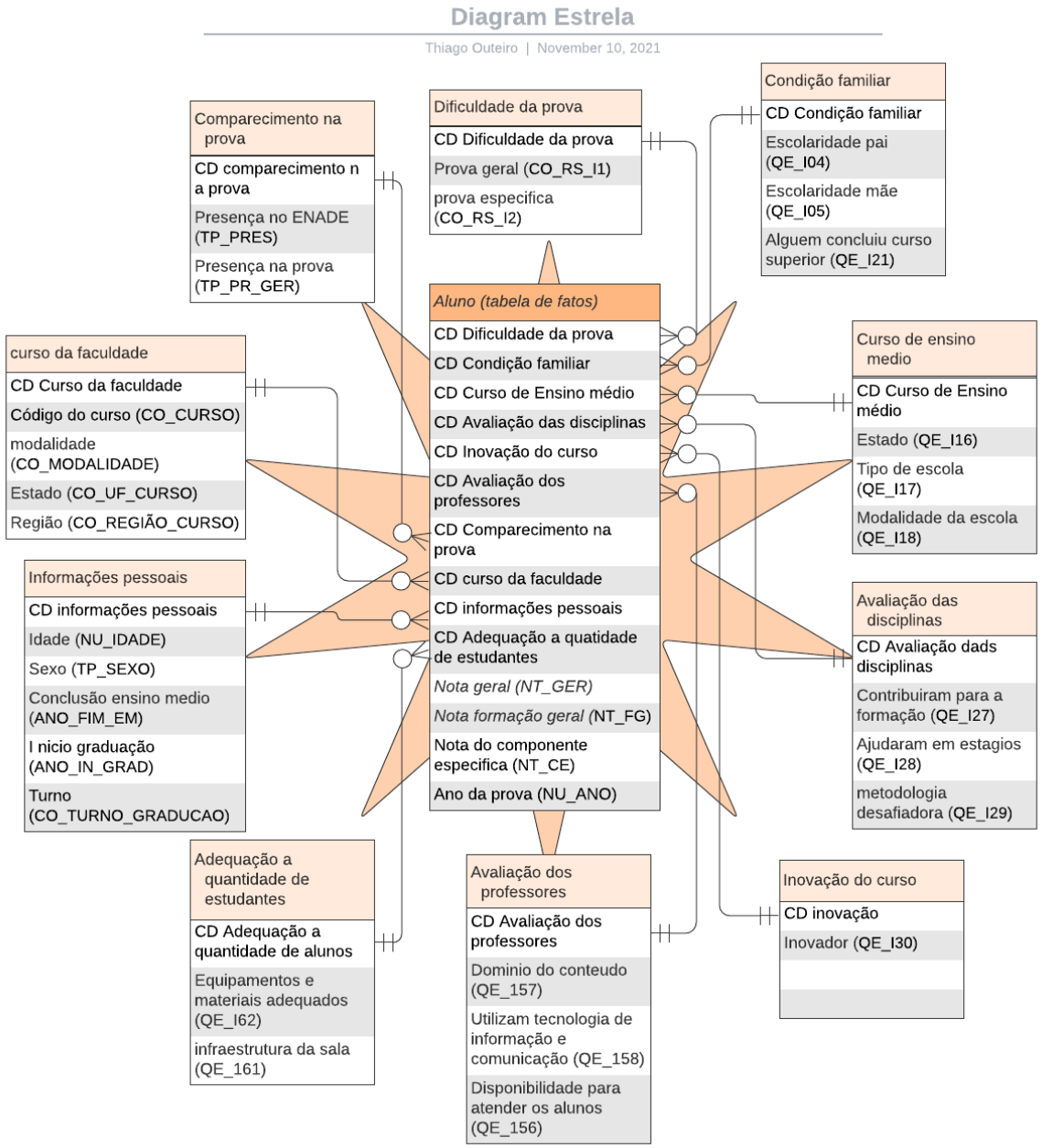
Utilizando os módulos do google drive foi possível facilmente realizar a coleta dos dados. Já o FileReader foi necessário utilizar o Complex Format, pois utilizando o fileReader normal ou o CSV obtive o seguinte erro:

ERROR CSV Reader 0:16 Execute failed: Row with ID 'Row190726' can't be converted to the configured data types. Increasing the number of scanned rows or changing the target

Porém utilizando o complex format as tabelas foram lidas de forma correta

2. Diagrama Estrela

Para criar o modelo dimensional estrela, foi utilizado os arquivos dentro da pasta leia-me para identificar quais colunas seriam utilizadas, e foi utilizado o site Lucidchart para criar o modelo. Obtendo o seguinte resultado:



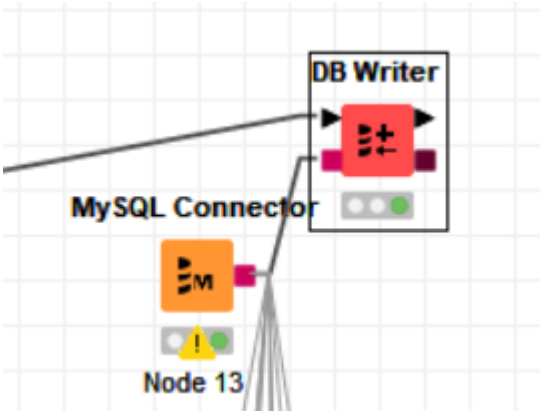
Para a escolha das dimensões foram escolhidos valores que podem ajudar a avaliar os fatos, e foram agrupados, a maioria deles vieram das perguntas do formulário do ENADE, mas algumas também se tratam do curso, ou de informações do aluno.

3. Criação do banco de dados

Para a criação do banco foi utilizado o aplicativo mySQL Workbench, e criado uma conection na própria IDE.

Para a criação de tabelas foi utilizado o KNIME, no momento da carga dos dados, utilizando os nós MySQL Connector e DBWrite:

1. **MySQL Connector:** Utilizado para se conectar com o banco criado no MySQL WorkBench, sua saida é o banco que irá se conectar na entrada no DB Writer.
2. **DB Writer:** Utilizado para escrever no banco de dados recebido do MySQL Connector, irá criar a tabela passada nas configurações caso ela ainda não exista.



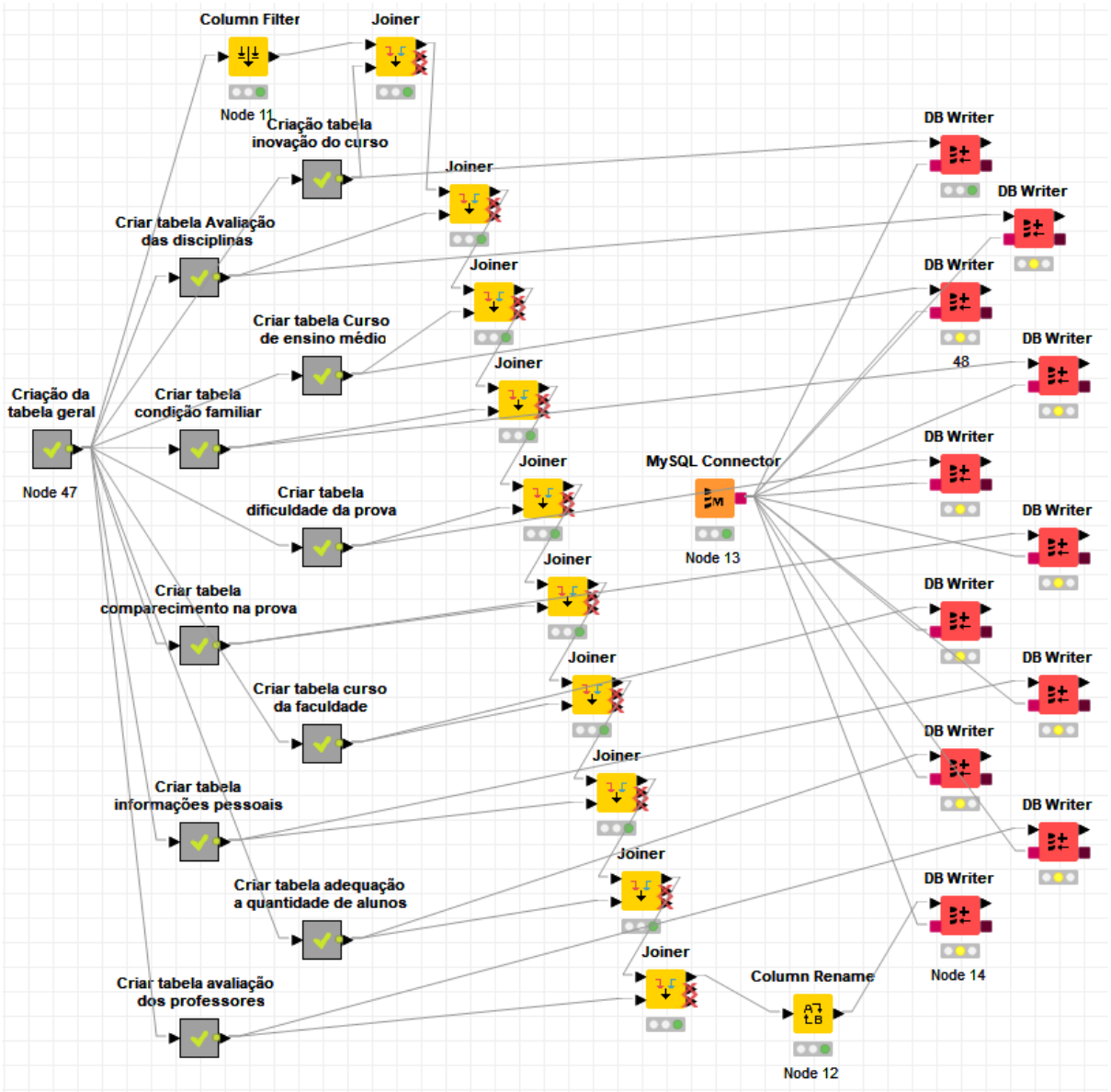
Para configurar esses nós é necessário que o banco seja criado no mySql Workbench e o schema seja criado com o seguinte script:

```
2 •   create schema provadw;
3 •   use provadw;
```

Após isso no node de Connector será necessário selecionar o banco (no caso foi utilizado o próprio main) e o no writer será selecionado o schema e a tabela a ser criada.

4. Carregamento dos dados

Para o carregamento dos dados foi utilizado o banco criado na questão 3, em conjunto com o KNIME obtendo o seguinte fluxo:

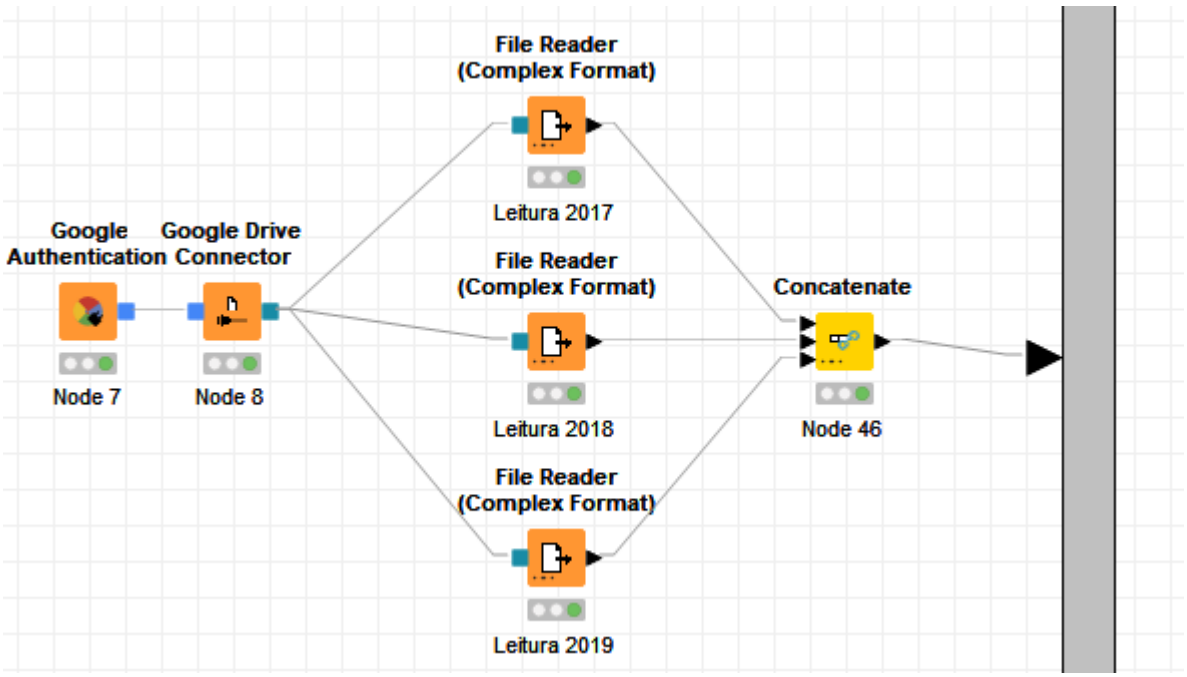


Além disso é possível ver diversos metanodes. estes são de dois tipos:

1. **Criação da tabela geral:** A criação da tabela geral é o primeiro metanode mais à esquerda do fluxo, ele irá recuperar os dados dos 3 anos e junta-los em uma única tabela.
2. **Criação das tabelas de dimensões do modelo:** Todos os outros metanodes são para criação das tabelas de cada dimensão do modelo.

Tabela geral

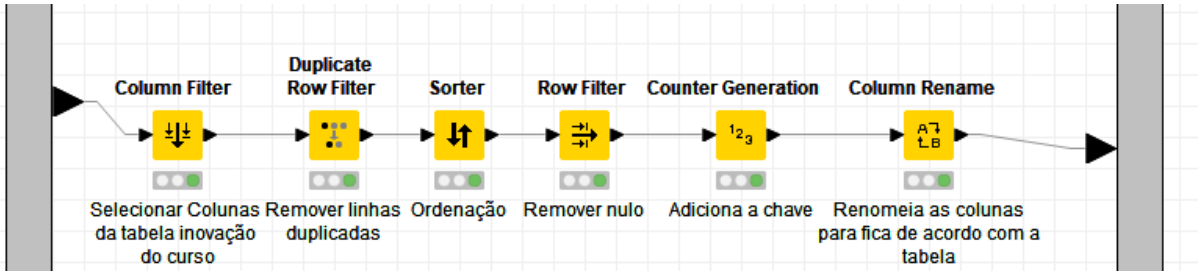
A criação da tabela geral foi suprimida em um metanode para facilitar a apresentação do resultado, ela se da pelo seguinte fluxo:



Como mostrado na questão 1, a captura é feita pelos nós do google em conjunto com o file reader (complex format), em conjunto com um nó de concatenação para juntar todos os dados em uma única tabela. Caso existam colunas presentes em um ano que não existam em outro, o KNIME apenas irá setar o valor ? (nulo) para os valores desta coluna nas linhas que não a tiverem.

Tabela de Dimensões:

A criação das tabelas de dimensões são muito parecidas, para exemplificar será mostrado o fluxo da tabela “curso de ensino médio” abaixo:



São utilizados os seguintes nós:

1. Column Filter: Utilizado para pegar apenas as colunas necessárias para a dimensão, no exemplo mostrado será pego as colunas: “QE_I16”, “QE_I17” e “QE_I18”
2. Duplicate row filter: Irá remover as linhas duplicadas deixando apenas as linhas com valores únicos, no exemplo por serem 3 colunas, foram 727 linhas:

Table "default" - Rows: 727 | Spec - Columns: 3 | Properties | Flow

Row ID	I QE_I16	S QE_I17	S QE_I18
Row0	51	B	A
Row1	51	A	B
Row3	11	A	A
Row5	51	A	A
Row19	15	A	A
Row23	50	D	A
Row31	11	B	A
Row38	11	E	A
Row41	51	A	D
Row43	51	A	E
Row44	31	A	C
Row49	51	A	C
Row51	51	E	A
Row56	51	D	D
Row59	35	A	A
Row62	51	D	B
Row77	51	D	A
Row78	31	A	A
Row85	?	?	?
Row99	51	B	D
Row101	50	B	A
Row118	50	E	A
Row131	33	E	A
Row152	35	B	A
Row153	29	B	A
Row159	32	B	A
Row178	29	A	A
Row218	41	A	E
Row240	41	A	A
Row245	13	A	A
Row251	99	A	A

3. Sorter: Este node não era necessário, mas facilitou o entendimento e a visualização, então o mantive, ele irá ordenar os dados da tabela.

Table "default" - Rows: 727 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	I QE_I16	S QE_I17	S QE_I18
Row85	?	?	?
Row245786_...	?	A	B
Row3	11	A	A
Row40907	11	A	B
Row11250	11	A	C
Row2717	11	A	D
Row40902	11	A	E
Row31	11	B	A
Row85992	11	B	B
Row295578	11	B	C
Row45740	11	B	D
Row218871	11	B	E
Row298662_...	11	C	A
Row40914	11	D	A
Row85862	11	D	B
Row86243	11	D	C
Row85970	11	D	D
Row172674	11	D	E
Row38	11	E	A
Row92806	11	E	B
Row85904	11	E	C
Row85909	11	E	D
Row168138_...	11	E	E
Row376032	11	F	A
Row4473	12	A	A
Row49439	12	A	B
Row36501	12	A	C

4. Row filter: Este nó é utilizado para remover os valores nulos, em algumas dimensões foi necessário utilizar este nó varias vezes (a quantidade de colunas da tabela) pois o nó avalia o valor por coluna. No exemplo apenas duas linhas foram removidas, foram de 727 para 725 linhas.

Table "default" - Rows: 725 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	QE_I16	QE_I17	QE_I18
Row3	11	A	A
Row40907	11	A	B
Row11250	11	A	C
Row2717	11	A	D
Row40902	11	A	E
Row31	11	B	A
Row85992	11	B	B
Row295578	11	B	C
Row45740	11	B	D
Row218871	11	B	E
Row298662_...	11	C	A
Row40914	11	D	A
Row85862	11	D	B
Row86243	11	D	C
Row85970	11	D	D
Row172674	11	D	E
Row38	11	E	A
Row92806	11	E	B
Row85904	11	E	C
Row85909	11	E	D
Row168138_...	11	E	E
Row376032	11	F	A
Row4473	12	A	A
Row49439	12	A	B

5. Counter generation: Este nó irá adicionar o ID da dimensão

Table "default" - Rows: 725 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	QE_I16	QE_I17	QE_I18	Counter
Row3	11	A	A	0
Row40907	11	A	B	1
Row11250	11	A	C	2
Row2717	11	A	D	3
Row40902	11	A	E	4
Row31	11	B	A	5
Row85992	11	B	B	6
Row295578	11	B	C	7
Row45740	11	B	D	8
Row218871	11	B	E	9
Row298662_...	11	C	A	10
Row40914	11	D	A	11
Row85862	11	D	B	12
Row86243	11	D	C	13
Row85970	11	D	D	14
Row172674	11	D	E	15
Row38	11	E	A	16
Row92806	11	E	B	17
Row85904	11	E	C	18
Row85909	11	E	D	19
Row168138_...	11	E	E	20
Row376032	11	F	A	21
Row4473	12	A	A	22
Row49439	12	A	B	23
Row36501	12	A	C	24
Row49427	12	A	D	25
Row49453	12	A	E	26

6. Column rename: Este nó irá modificar os nomes das colunas para que fique de acordo com o modelo

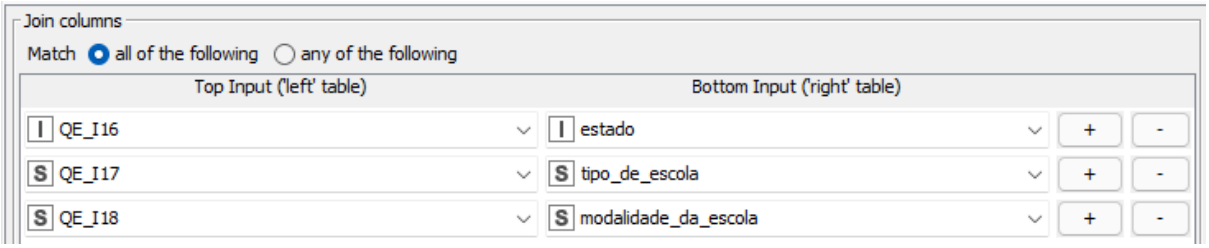
Table "default" - Rows: 725 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	estado	tipo_de_escola	modalidade_da_escola	CD_curso_de_ensino_medio
Row3	11	A	A	0
Row40907	11	A	B	1
Row11250	11	A	C	2
Row2717	11	A	D	3
Row40902	11	A	E	4
Row31	11	B	A	5
Row85992	11	B	B	6
Row295578	11	B	C	7
Row45740	11	B	D	8
Row218871	11	B	E	9
Row298662_...	11	C	A	10
Row40914	11	D	A	11
Row85862	11	D	B	12
Row86243	11	D	C	13
Row85970	11	D	D	14
Row172674	11	D	E	15
Row38	11	E	A	16
Row92806	11	E	B	17
Row85904	11	E	C	18
Row85909	11	E	D	19
Row168138_...	11	E	E	20
Row376032	11	F	A	21
Row4473	12	A	A	22
Row49439	12	A	B	23

Tabela de fatos

Para a criação da tabela de fatos foi necessário utilizar o nó Joiner, que tem por objetivo realizar um join (neste caso foi utilizado o inner join).

Ainda utilizando o exemplo da dimensão anterior, após passar pelo filtro de colunas, serão filtradas apenas as colunas que não estão presentes em nenhuma dimensão, deixando por exemplo as colunas “QE_I16”, “QE_I17” e “QE_I18”, chamaremos esta de pré tabela de fatos. Na configuração do joiner você irá escolher quais colunas participarão do joiner, neste exemplo serão as seguintes:



Sendo a coluna da esquerda as colunas da pré tabela de fatos e a da esquerda a tabela da minha dimensão. Isso criará uma nova coluna na pré tabela de fatos que fará com que para cada linha, tenha o valor do id (neste caso CD_curso_de_ensino_medio). Além disso nas configurações do joiner é selecionado que estas colunas que participarão do joiner serão excluídas, ficando apenas a coluna da chave.

Após passar por todos os joiners este será o resultado final:

Table "default" - Rows: 1230932 | Spec - Columns: 14 | Properties | Flow Variables

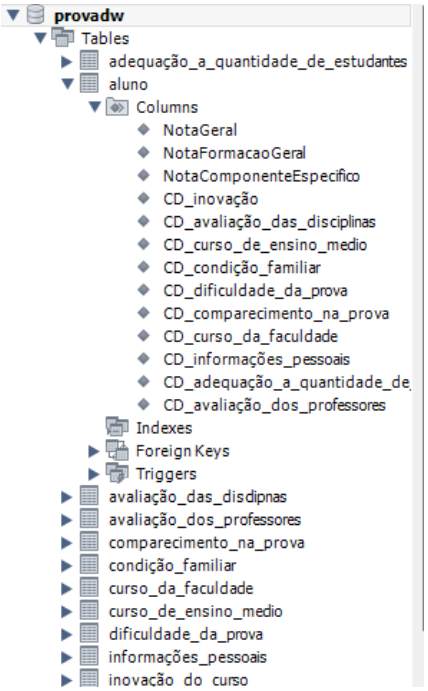
Row ID	NU_ANO	NT_GER	NT_FG	NT_CE	CD_ino...	CD_av...	CD_cur...	CD_con...	CD_difi...	CD_co...	CD_cur...	CD_inf...	CD_ad...	CD_av...
Row0_Row0_...	2017	52,6	64,1	48,7	2	362	620	46	33	8	2	3623	10	209
Row1_Row0_...	2017	63,1	82,1	56,7	2	284	616	54	41	8	2	1538	29	299
Row2_Row2_...	2017	33,7	52,5	27,4	0	356	620	58	38	8	2	1818	9	9
Row3_Row3_...	2017	58,1	72,6	53,2	3	355	0	42	25	8	2	1818	35	227
Row4_Row0_...	2017	46,3	73,2	37,3	2	204	620	46	32	8	2	2515	9	145
Row5_Row0_...	2017	64,2	80,7	58,7	2	269	615	14	33	8	2	1814	28	153
Row6_Row0_...	2017	41,4	43,0	40,8	2	213	615	42	33	8	2	475	11	153
Row7_Row7_...	2017	48,8	45,1	50,0	4	361	620	44	32	8	2	1814	2	160
Row8_Row8_...	2017	52,9	66,9	48,2	5	442	616	58	40	8	2	1477	44	363
Row9_Row0_...	2017	62,7	82,3	56,2	2	284	620	42	32	8	2	880	27	159
Row10_Row0...	2017	51,9	61,6	48,7	2	352	615	30	33	8	2	876	18	345
Row11_Row2...	2017	58,4	60,6	57,7	0	204	620	54	32	8	2	4070	11	201
Row12_Row3...	2017	77,7	59,3	83,8	3	363	616	44	40	8	2	2551	27	234
Row13_Row1...	2017	37,5	30,0	40,0	1	361	615	44	32	8	2	2138	27	290
Row14_Row1...	2017	66,4	88,7	58,9	1	354	620	42	33	8	2	902	9	289
Row15_Row1...	2017	71,5	75,6	70,1	1	290	620	68	32	8	2	2507	28	290
Row16_Row3...	2017	57,8	70,6	53,5	3	91	620	70	32	8	2	9247	45	291

Apenas estarão nas tabelas as notas e as chaves.

Por se tratar de um innerJoin, os valores nulos das chaves são descartados, já que todas são chaves primárias para a tabela de fatos, nenhuma pode ser nula, o que faz com que saiamos inicialmente de 1519493 linhas para 1230932, ou seja, cerca de 81% dos dados iniciais.

Essa diferença se dá primariamente por alunos que estavam inscritos no ENADE mas não preencheram os questionários.

Como resultado, no MySQL workbench tivemos as tabelas:



Realizando um select a tabela de “Aluno” como exemplo, o resultado ficou da seguinte forma:

```
3 • use provaDW;
4 • select * from aluno;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

NotaGeral	NotaFormacaoGeral	NotaComponenteEspecifico	CD_inovação	CD_avaliação_das_disciplinas	CD_curso_de_ensino_medio	CD_condição_familiar	CD_dificuldade_da_prova	CD_comparecimento_na_prova	CD_curso_da_faculdade	CD_informações_pessoais	CD_adequação_a_quantidade_de
52,6	64,1	48,7	2	290	620	46	33	8	2	3623	10
63,1	82,1	56,7	2	218	616	54	41	8	2	1538	29
33,7	52,5	27,4	0	284	620	58	38	8	2	1818	9
58,1	72,6	53,2	3	283	0	42	25	8	2	1818	35
46,3	73,2	37,3	2	144	620	46	32	8	2	2515	9
64,2	80,7	58,7	2	203	615	14	33	8	2	1814	28
41,4	43,0	40,8	2	153	615	42	33	8	2	475	11
48,8	45,1	50,0	4	289	620	44	32	8	2	1814	2
52,9	66,9	48,2	5	364	616	58	40	8	2	1477	44
62,7	82,3	56,2	2	218	620	42	32	8	2	880	27
51,9	61,6	48,7	2	280	615	30	33	8	2	876	18
58,4	60,6	57,7	0	144	620	54	32	8	2	4070	11
77,7	59,3	83,8	3	291	616	44	40	8	2	2551	27
37,5	30,0	40,0	1	289	615	44	32	8	2	2138	27
66,4	88,7	58,9	1	282	620	42	33	8	2	902	9
71,5	75,6	70,1	1	224	620	68	32	8	2	2507	28
57,8	70,6	53,5	3	43	620	70	32	8	2	9247	45
62,7	75,3	58,5	2	275	0	58	33	8	2	664	28
75,1	78,0	74,1	1	217	615	44	32	8	2	1437	20
50,5	51,2	50,3	3	291	95	44	39	8	2	1814	36
46,1	75,6	36,3	2	291	620	56	40	8	2	902	28
44,1	68,3	36,0	0	127	620	54	32	8	2	1814	0
47,9	67,5	41,3	1	72	620	56	32	8	2	1814	8
55,1	81,5	46,3	5	289	603	27	33	8	2	2507	8
73,5	96,4	65,9	4	217	620	44	26	8	2	2547	44
60,4	70,1	57,2	2	280	616	56	32	8	2	2515	3
62,7	92,8	52,7	1	217	620	68	17	8	2	2138	27

5. Análise de dados

As 5 propostas de perguntas são:

1. A escolaridade da família influencia na nota do ENADE?

2. A modalidade do curso da faculdade influencia na nota do ENADE?
3. A adequação a quantidade de alunos influencia na nota do ENADE?
4. O aluno que cursa o ensino médio em um estado diferente da faculdade influencia na nota do ENADE?
5. A diferença de tempo entre o fim do ensino médio e o início da graduação afeta a nota do ENADE?

Para responder a pergunta 1 irei utilizar o KNIME utilizando o nó de barChart.
Utilizando a dimensão de condição familiar para os valores de escolaridade da mãe e do pai, temos os seguintes valores:

- A = Nenhuma.
- B = Ensino Fundamental: 1º ao 5º ano (1ª a 4ª série).
- C = Ensino Fundamental: 6º ao 9º ano (5ª a 8ª série).
- D = Ensino Médio.
- E = Ensino Superior - Graduação.
- F = Pós-graduação.

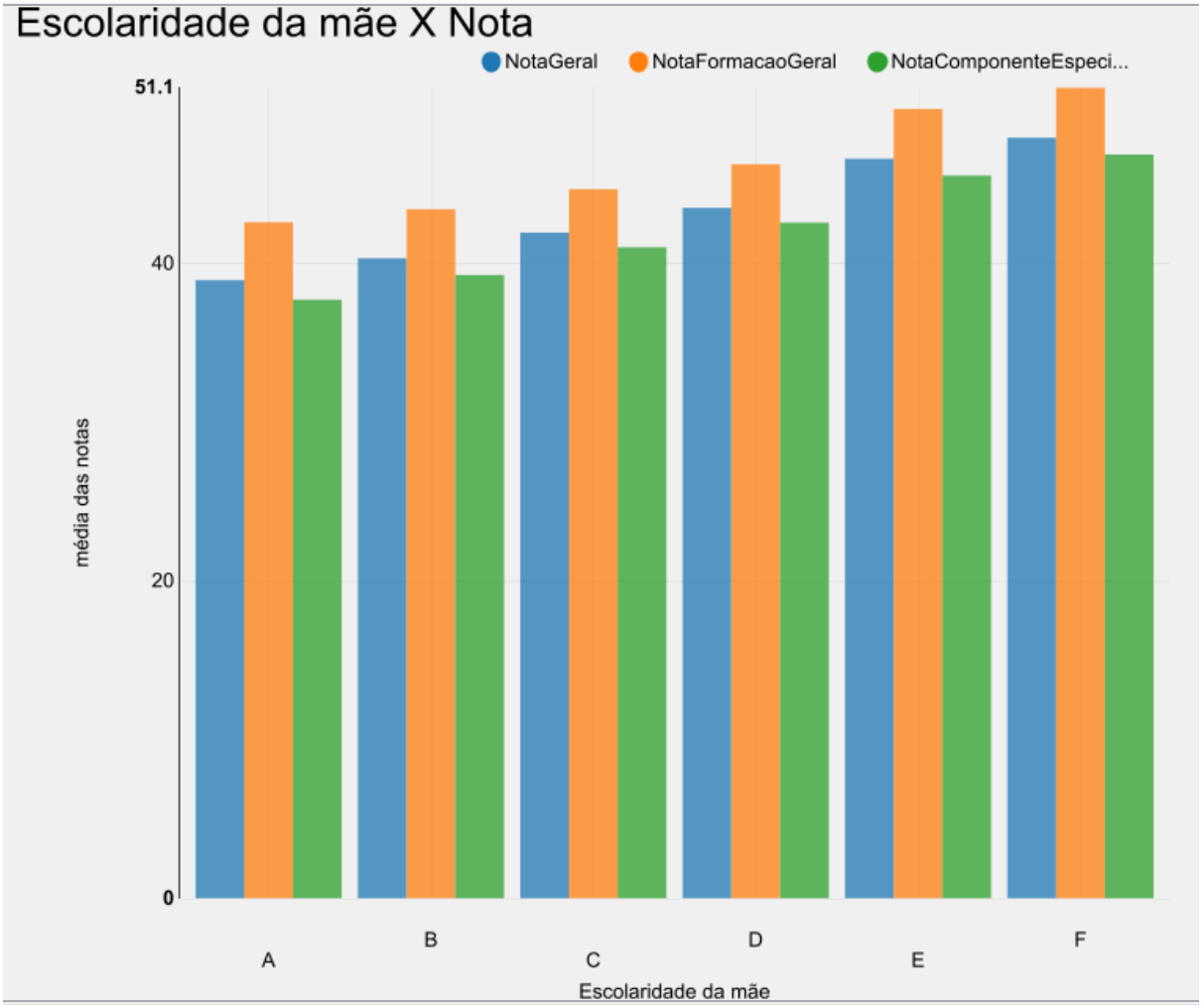
Já os valores de “alguém se formou em sua familia?” temos

- A = sim
- B = não

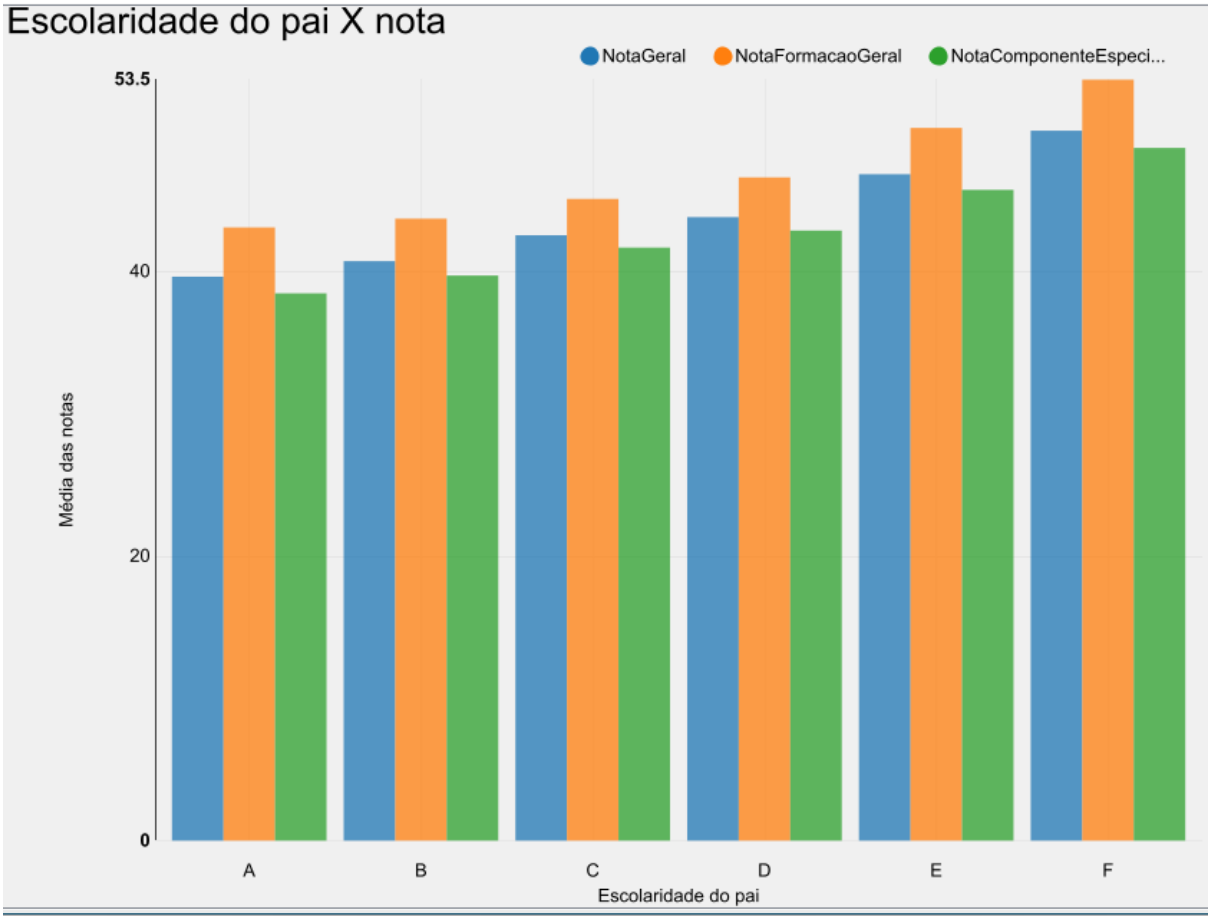
Para avaliar serão apresentados 2 gráficos para cada valor da dimensão, um mostrando a média das notas por opção assinalada no questionário e outro mostrando a quantidade de pessoas que assinalaram a opção no questionário

Média das notas em relação à escolaridade:

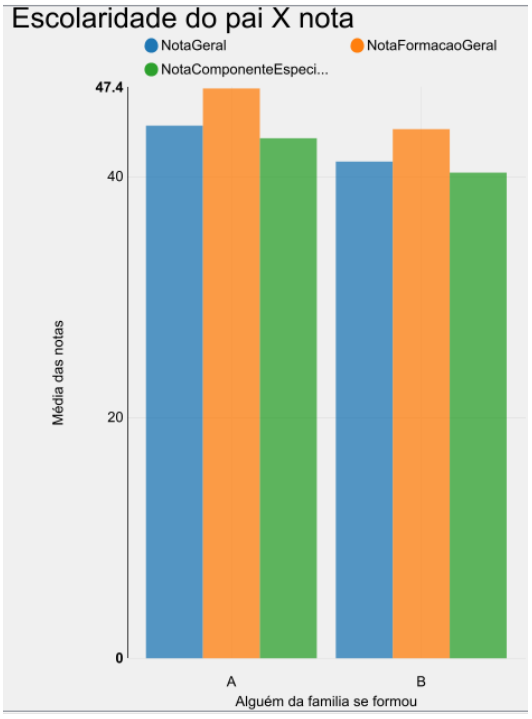
Escolaridade da mãe:



Escolaridade do pai:

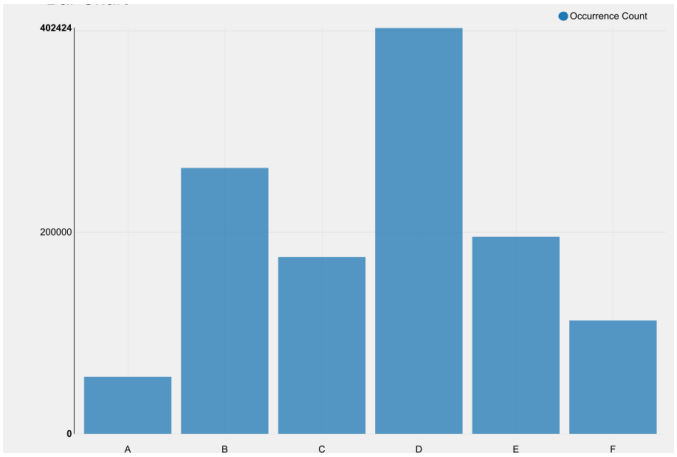


Alguém da família se formou na faculdade:

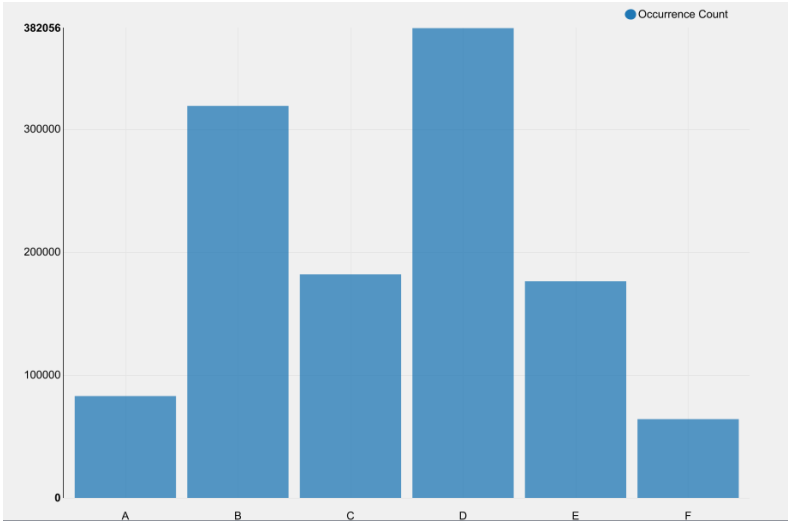


Quantidade de pessoas que marcaram as opções em:

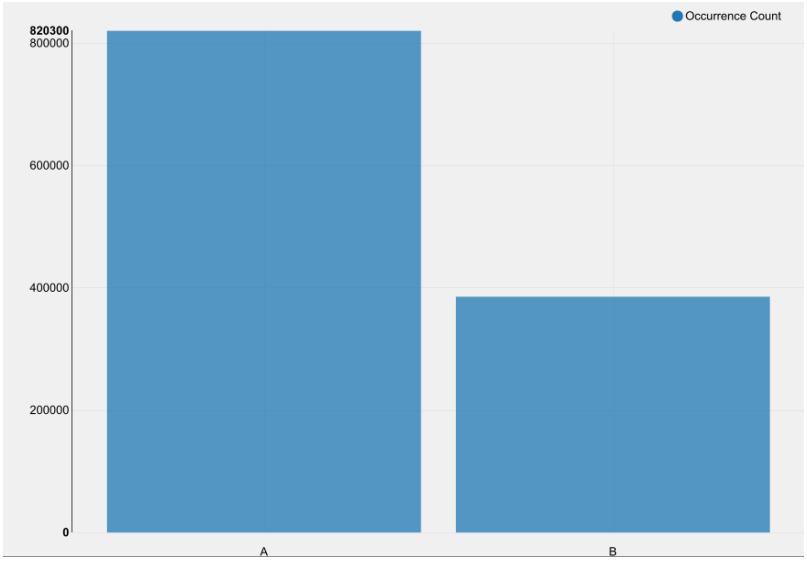
Escolaridade da mãe:



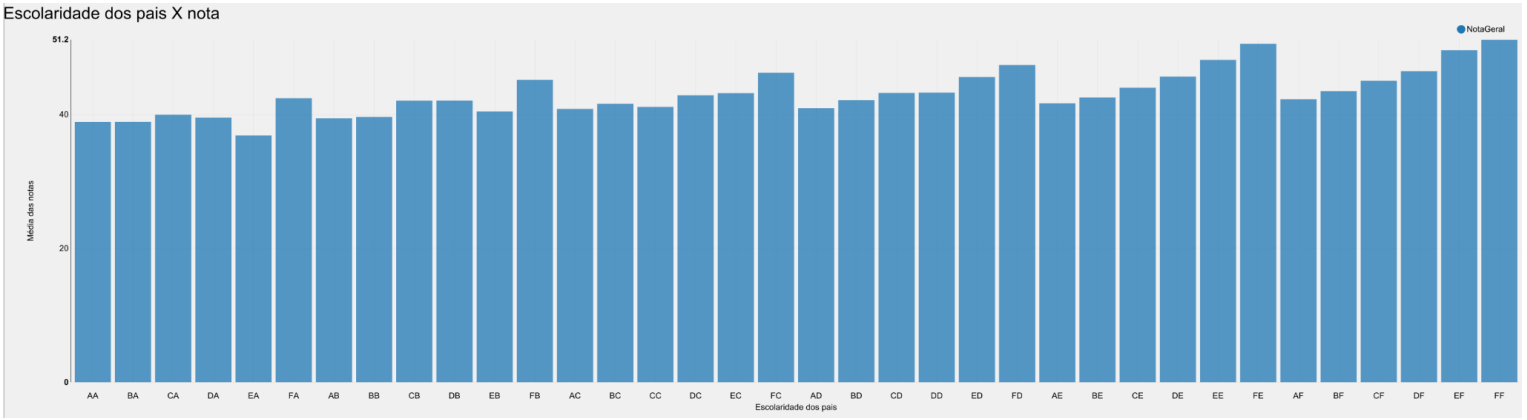
Escolaridade do pai:



Alguém da família se formou na faculdade:



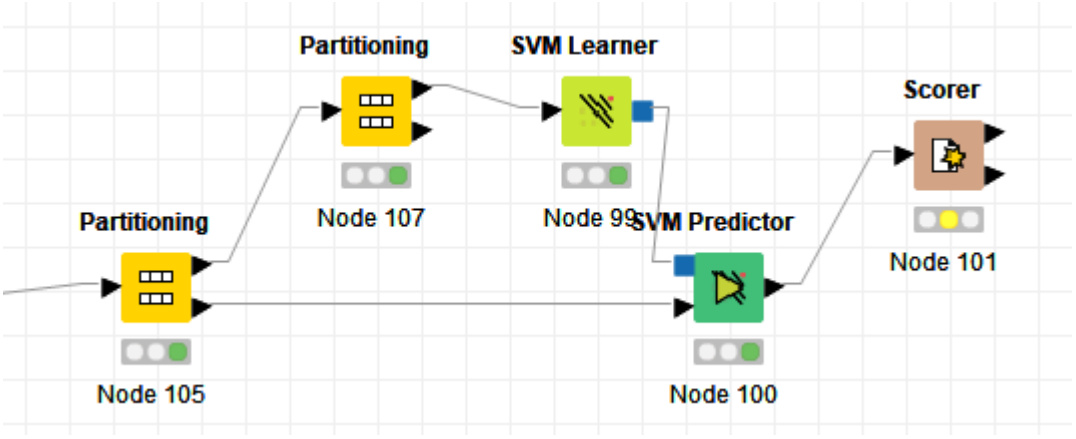
Os resultados mostram que a condição familiar influencia sim na nota do ENADE visto que tanto para o pai quanto para a mãe, quanto maio o nivel de escolaridade de um dos dois, maior a médias das notas. Além disso, o fato de alguém ter se formado no ensino superior também faz com que a nota seja maior, visto que a média de quem marcou esta opção como Sim (opção A) é maior do que a média de quem marcou não.
Mais um gráfico que pode ajudar a avaliar é o gráfico abaixo que mostra apenas a nota Geral:



Nele foram concatenados as opções marcadas no questionário para escolaridade do pai e escolaridade da mãe, formando por exemplo AC indicando que a opção de escolaridade do pai era A e da mãe C
Com isso é possível perceber a grande diferença entre as notas gerais conforme o gráfico anda para o lado esquerdo (ou seja, a escolaridade aumenta). É possível perceber que a maior média entre os que têm a escolaridade da mãe igual sempre é quando a escolaridade do pai é a maior (F) e toda a média do gráfico vai subindo conforme a escolaridade de mãe aumenta também, onde no caso AA a média geral não chega em 40, mas no FF ela é 51,2.

6. Aprendizado de máquina

Para realizar o aprendizado também foi utilizado o KNIME, utilizando o nós ensinados em aula.



1. **Partitioning:** Usado para criar uma separação em duas tabelas com as mesmas colunas, mas com linhas diferentes, para que uma parte seja usada para o aprendizado e outra para a predição
2. **SVM Learner:** Nó de aprendizado, irá aprender baseado em um coluna de string
3. **Predictor:** Irá utilizar o que o learner aprender para tentar prever o valor da coluna de string selecionada no learner
4. **Scorer:** Utilizado para fazer a “pontuação” da predição.

Foram utilizados 2 partitioning, separando 80% no primeiro para o learning e 20% para a predição, o segundo foi utilizado um valor absoluto de 10000 linhas para o learning, pois como havia mais de 1 milhão de linhas, a learning estava demorando demais (mais de 2 horas).

Esta parte foi feita totalmente baseada na aula, porém ao colocar os valores, a predição não estava funcionando corretamente, foram tentadas as seguintes predições:

- Baseado nas notas qual o sexo da pessoa
- Baseado na nota especifica qual a dificuldade que a pessoa achou da prova específica
- Baseado na nota geral qual o estado do curso da pessoa
- Baseado na nota geral qual a modalidade do curso da pessoa
- Baseado nas notas qual a modalidade do curso da pessoa
- Baseado na idade qual a modalidade do curso da pessoa

Porém em todas as tentativas a previsão retornava sempre o mesmo valor para todas as predições. No exemplo da nota geral baseada no curso, ele predizia 2 para todos os valores, indiferente da partição.

Com isso infelizmente não foi possível realizar a predição para nenhuma das perguntas tentadas.

7. Referências

- **Github:** Github com os arquivos do trabalho: https://github.com/topd97/prova_DW_2021
- **Knime:** Para a maioria das questões foi utilizado o KNIME, mostrado em aula e que pode ser encontrado em: https://www.knime.com/
- **MySQL workbench:** Fora isso também foi utilizado o MySQL workbench que pode ser encontrado em: https://www.mysql.com/products/workbench/