Data Warehouse no Suporte à Tomada de Decisão
Relatório da Prova Individual

Arthur Moreira de Albuquerque
DRE 114146877

Análise de microdados do ENADE dos anos de 2017, 2018, 2019

O trabalho foi feito no sistema operacional Windows 11. O Python Notebook precisa ser executado no VS Code (não pode ser executado no Colab) com as extensões 'Jupyter' e 'Jupyter Notebook Renderers' para baixar os dados na pasta Downloads do Windows. Isso é necessário pois a pasta Downloads é o endereço no qual o 'CSV reader' do Workflow KNIME vai pegar os dados. Depois disso é só abrir o Workflow KNIME e executar para ver a parte que contém a análise de dados e os resultados dos algoritmos de aprendizado.

Link do relatório (Google Doc):
https://docs.google.com/document/d/1Hb3K8dJsjC-m_pmD9Pycb-HtMgij0AFV5hV3at5bBVE/edit?usp=sharing

Link do repositório no GitHub:
https://github.com/troclaux/dw_trabalho_final_individual

Questão 1

Neste trabalho achei melhor usar as mesmas ferramentas que foram utilizadas no trabalho em grupo. Primeiro se executa um Python Notebook para baixar os dados. Para isso, foi necessário importar a biblioteca GoogleDriveDownloader, que baixa os dados que estão salvos na minha conta do Google Drive na mesma pasta em que o Python Notebook se localiza. O próximo passo consiste em criar o banco de dados com a ajuda do SQLite. Finalmente, o programa em Python lê os arquivos .txt que contém os dados e os insere no banco de dados.

Depois que os arquivos foram baixados e o banco de dados foi preenchido, achei mais prático construir um workflow do knime para aplicar o algoritmo de aprendizado, particionar os dados e sintetizar representações visuais para responder as 5 perguntas propostas.

Os dados são armazenados no Google Drive, mas estão disponíveis originalmente neste link. Conforme o enunciado, serão utilizados os dados dos anos 2017, 2018 e 2019. Entre os arquivos obtidos estão:

- MICRODADOS_ENADE_2017.txt
- microdados_enade_2018.txt
- microdados_enade_2019.txt

Felizmente todos os arquivos .txt contém os dados no mesmo formato, com a primeira linha contendo os atributos separados por ponto e vírgula (;) e depois temos as tuplas com os dados, que também são separados por ponto e vírgula.

Observando o dicionário de dados que vem junto com os dados podemos ter maior clareza do conteúdo que foi baixado.

Questão 2

Depois de analisar o conjunto de dados no dicionário de dados, foi feita a escolha para organizar o modelo no seguinte formato:

- Tabela fato:
  - CLASSIFICACAO_ENADE
- Tabelas dimensão:
  - TEMPO
  - SITUACAO_QUEST_DISCURSIVA
  - IES
  - PROVA_OBJETIVA
  - PERCEPCAO_PROVA
  - QUESTIONARIO_ESTUDANTE
  - ESTUDANTE
  - FEEDBACK
  - VETOR_RESP_ESTUDANTE
  - VETOR_GABARITO

Foi utilizado o site https://www.diagrams.net/ para desenhar o modelo de dados e para executar a análise de dados, preferimos o KNIME.

Questão 3

Como o trabalho já estava sendo feito em Python, foi utilizado o SQLite para criar a base de dados do Data Warehouse. Para fazer isso, se cria um objeto Connection que representa o banco de dados usando a função connect() do módulo sqlite3. Com isso, o arquivo dw_prova_individual.db é criado. Segundamente, é preciso criar as tabelas com os respectivos atributos. Foi necessário escrever strings com os comandos para criar a tabelas conforme o exemplo abaixo:

```
sql_create_estudante_table = """ CREATE TABLE IF NOT EXISTS estudante (
                                    estudante_id integer PRIMARY KEY,
                                    nu_idade integer,
                                    tp_sexo text,
                                    co_turno_graduacao text,
                                    tp_inscricao_adm integer,
                                    tp_inscricao integer
                                );"""
```

Finalmente se cria uma função create_table() que aceita o objeto Connection juntamente com a string que contém os comandos SQL. Dentro da função, chamamos o método execute() do objeto Cursor para executar a declaração de criação de tabela. Com isso, o SQLite cria a base de dados.

```
1  def create_table(conn, create_table_sql):
2      try:
3          c = conn.cursor()
4          c.execute(create_table_sql)
5      except Error as e:
6          print(e)
7
```

Questão 4

Para inserir os dados foi necessário usar a biblioteca pandas. Com ela, é possível ler os dados de arquivos csv com a função to_sql(), que permite a leitura apenas das colunas (atributos) desejados, o que facilitava a inserção dos dados de cada tabela.

Ao implementar esse banco de dados, percebi que os atributos qe_i69 até o qe_i81 estavam presentes apenas nos dados do ENADE 2017. Foi necessário adaptar a inserção de dados para inserir esses atributos apenas para 2017.

O processo de preenchimento consistiu nas seguintes etapas:

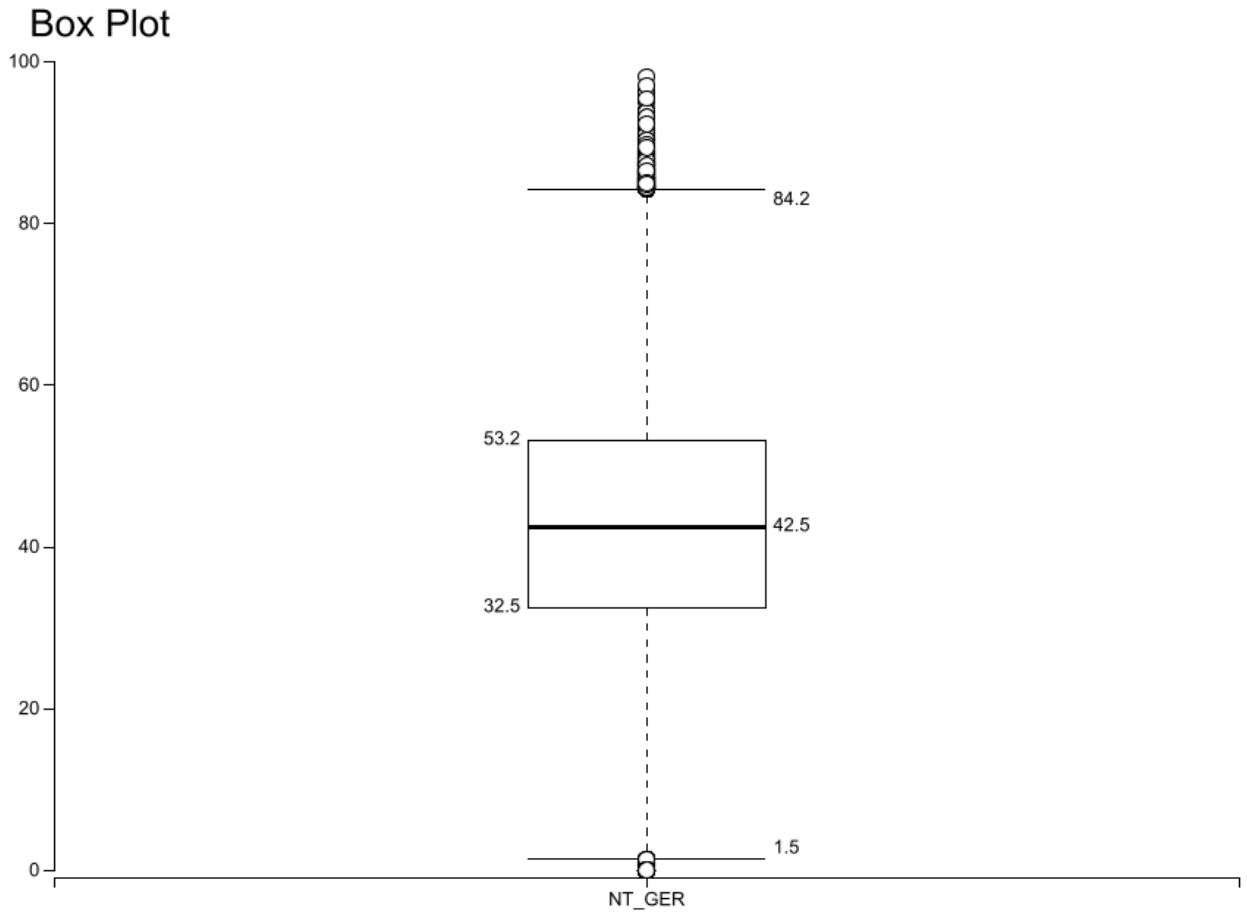
1. Declaração das listas de atributos de cada tabela
2. Declaração de strings que contém o caminho para os microdados ENADE de cada ano
3. Declaração da função que vai criar a conexão com o banco de dados SQLite
4. Declaração da função que vai inserir os dados de 2017, 2018, 2019 no banco de dados
5. Finalmente se utiliza as funções e bibliotecas importadas para ler as colunas corretas de cada tabela e inserir os dados com a função to_sql()

Questão 5

Para essa parte do trabalho foi utilizado o KNIME por conta da interface intuitiva e porque ele facilita a manipulação dos dados. Abaixo vamos ter a análise de dados com as respostas para 5 perguntas propostas:

1) *Qual é a mediana do desempenho bruto dos participantes do ENADE?*

O melhor jeito de responder a pergunta foi por meio de um boxplot que analisa os valores da variável nt_ger. Essa variável representa a nota bruta da prova, que consiste da média ponderada da formação geral (25 %) e componente específico (75%). A nota varia de 0 até 98,1.



Conforme podemos observar pelo gráfico, a nota geral possui mediana 42.5, com quartil inferior 32,5 e quartil superior 53,2. O limite inferior é 1,5 enquanto o limite superior é 84,2.

Ao gerar a tabela do atributo no KNIME, foi obtido uma tabela com 1000 linhas, já que os valores podem ser decimais. Por conta disso não vou incluir a tabela nessa pergunta, mas ela ainda está presente e disponível para consulta no workflow do KNIME.

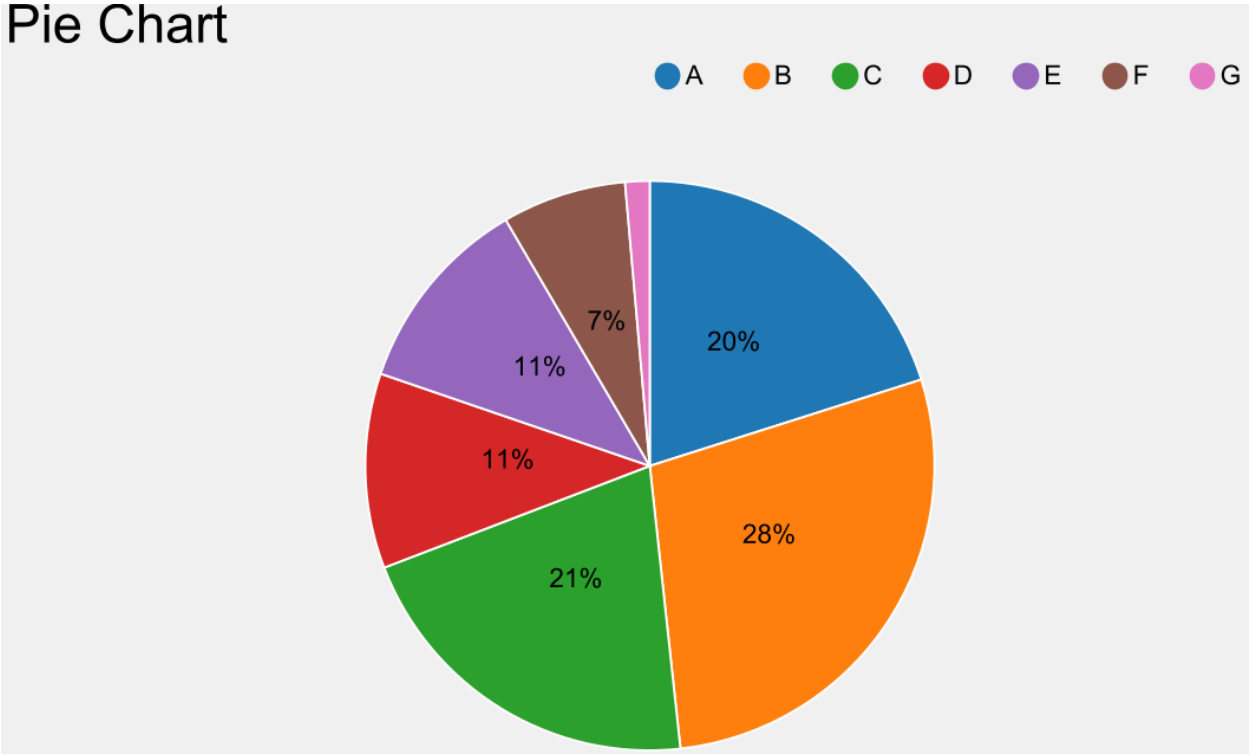
Esperava um desempenho melhor dos graduandos do ENADE, mas imagino que essas notas são um reflexo sobre o sistema de educação brasileiro, que deveria receber maior investimento.

1) *Qual é a distribuição dos participantes por renda familiar?*

A resposta para a primeira pergunta está na coluna do atributo qe_i08, que armazena as respostas dos participantes no questionário do ENADE. A pergunta é "Qual a renda total de sua família, incluindo os seus rendimentos?". Para interpretar os dados, é necessário consultar o dicionário de dados, que nos fornece a seguinte legenda:

- A = Até 1,5 salário mínimo (até R$ 1.405,50)
- B = De 1,5 a 3 salários mínimos (R$ 1.405,51 a R$ 2.811,00)
- C = De 3 a 4,5 salários mínimos (R$ 2.811,01 a R$ 4.216,50)
- D = De 4,5 a 6 salários mínimos (R$ 4.216,51 a R$ 5.622,00)
- E = De 6 a 10 salários mínimos (R$ 5. 622,01 a R$ 9.370,00)
- F = De 10 a 30 salários mínimos (R$ 9.370,01 a R$ 28.110,00)
- G = Acima de 30 salários mínimos (mais de R$ 28.110,00)

A	273666
B	382981
C	284374
D	150636
E	154494
F	96110
G	18777



A partir do gráfico podemos notar que a maioria da população está na faixa de 1,5 a 3 salários mínimos. Listando cada faixa por tamanho, obtemos:

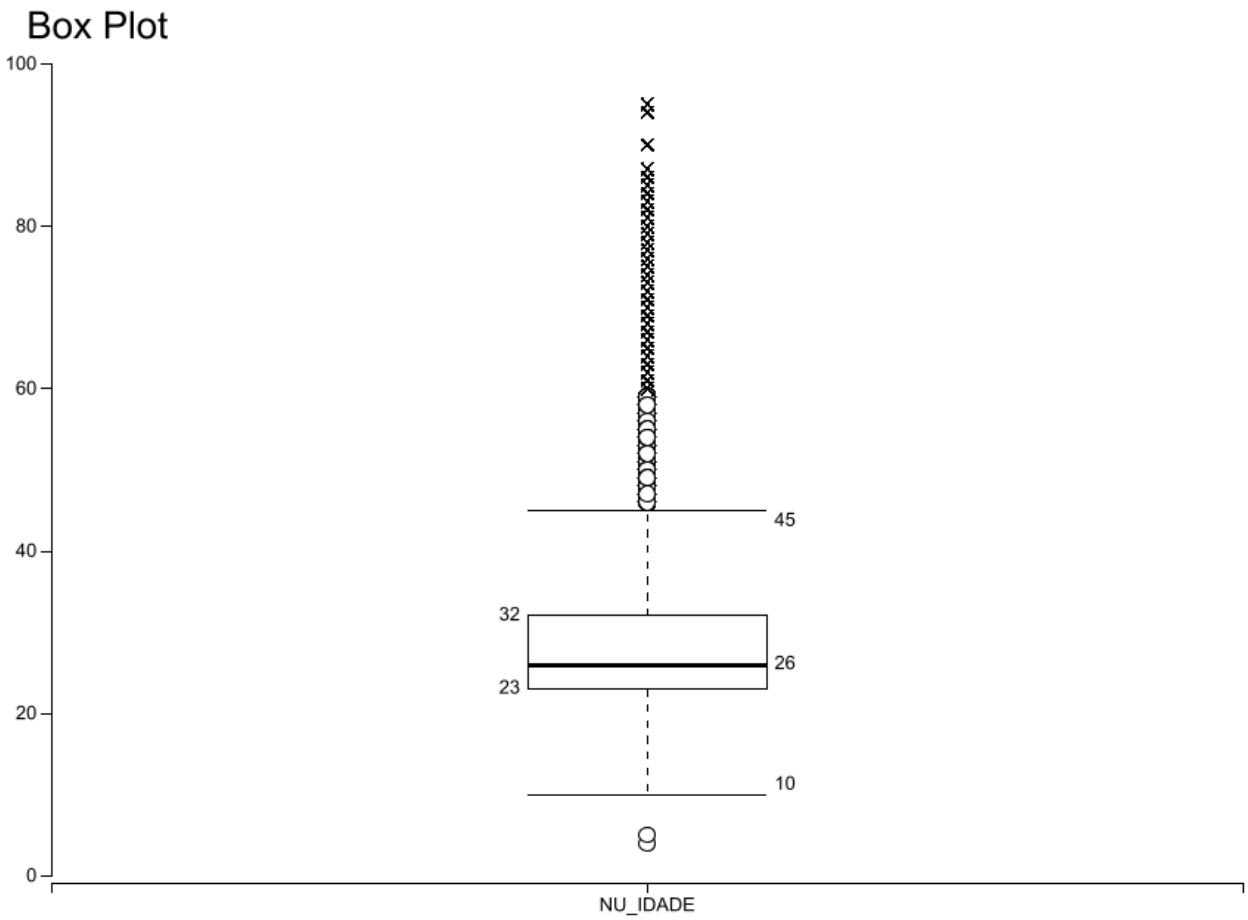
1,5 a 3	Maior parcela da população
3 a 4,5	
Até 1,5	
4,5 a 6 ≅ 6 a 10	
10 a 30	
Acima de 30	Menor parcela da população

Os resultados estão de acordo com o esperado, onde participantes com renda acima de 30 salários mínimos são uma parcela ínfima da população, assim como a maioria da população cair nos grupos de faixa salarial mais humilde (por exemplo 1,5 a 3 salários mínimos).

3) *Qual é a distribuição de idade dos participantes do ENADE?*

Novamente se utiliza o boxplot para responder, a variável analisada é nu_idade.

Row ID	count
4	1
5	1
10	2
11	3
12	1
14	1
16	1
17	4
18	80
19	3179
20	17484
21	61409
22	152747
23	188451
24	160372
25	126527
26	99646
27	79912
28	67731
29	57856
30	50919
31	46055
32	40810
33	36528
34	34024
35	31673
36	29904
37	27257
38	24501
39	21728
40	19532
41	17132
42	15177
43	13497
44	11943
45	10865
46	9486
47	8382
48	7537
49	6626
50	5668
51	5148
52	4747
53	4134
54	3549
55	3107
56	2587
57	2194
58	1813
59	1488
60	1259
61	1039
62	805
63	653
64	535
65	403
66	329
67	265
68	187
69	149
70	124
71	77
71	77
72	54
73	52
74	44
75	24
76	20
77	10
78	11
79	7
80	5
81	4
82	1
83	7
84	1
85	2
86	2
87	1
90	1
94	2
95	1



Conforme o esperado, a mediana da idade dos graduandos é 26, valor que segue o que observo na vida real nos meus amigos que estão terminando o curso. Além disso, temos quartil inferior 23, quartil superior 32, limite inferior 10 e limite superior 45.

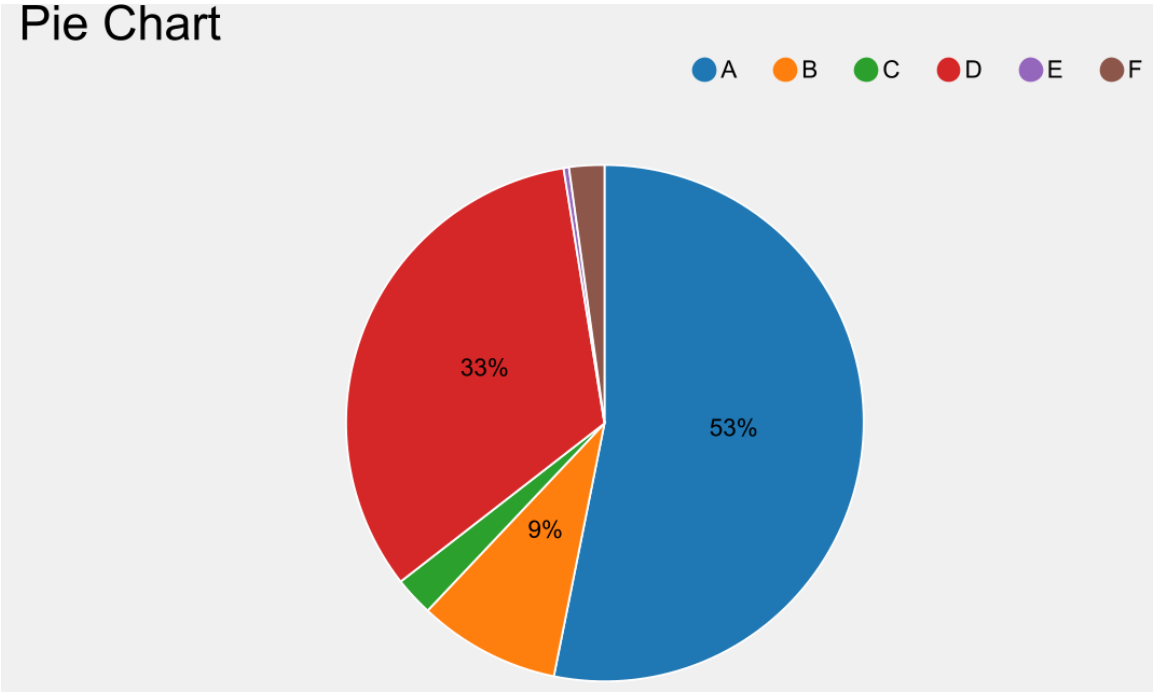
4) *Qual é a cor/raça mais comum entre os participantes do ENADE?*

Nesse atributo, os rótulos vieram como caracteres únicos, mas basta olhar para o dicionário de variáveis para compreender o significado de cada letra.

- A = Branca
- B = Preta
- C = Amarela
- D = Parda
- E = Indígena
- F = Não quero declarar

Conhecendo a legenda, podemos interpretar a tabela e o Pie Chart:

A	723696
B	121027
C	32921
D	448817
E	4750
F	29828



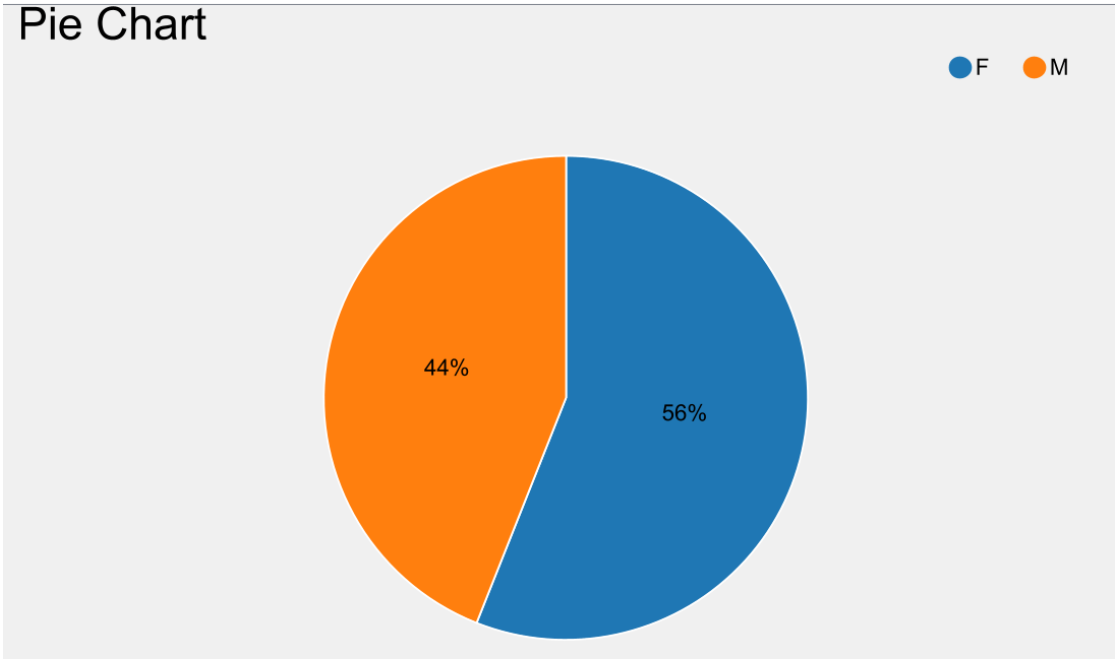
Conforme podemos observar, a maioria esmagadora dos participantes se identificou como cor branca, consistindo em 53% do total. Abaixo temos cada cor/raça, ordenada de maior para menor porcentagem do total de participantes.

Branca > Parda > Preta > Amarela > Indígena

5) *Qual é a porcentagem de homens e mulheres participando do ENADE?*

Para isso, basta utilizar o nó "Pie/Donut Chart" do KNIME, que seleciona a coluna tp_sexo do arquivo .csv e automaticamente constrói o gráfico que mostra a proporção de homens e mulheres participando do ENADE.

Row ID	count
F	850965
M	668528

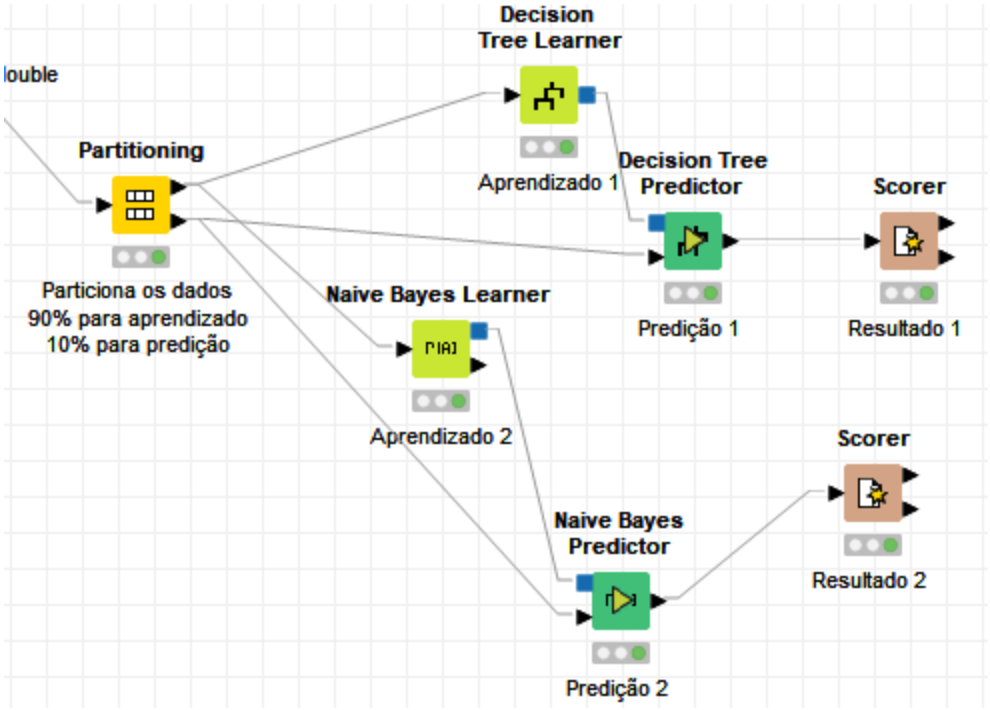


Nesse gráfico, o 'M' representa a parcela dos homens e o 'M' representa a das mulheres. Conforme podemos ver, existe uma diferença substancial entre as porcentagens. Com 44% dos participantes pertencendo ao sexo masculino e 56% pertencendo ao sexo feminino.

Questão 6

Para o aprendizado, novamente foi utilizado o KNIME, que já vem com os módulos de aprendizado e predição necessários. Nesse exercício, foram utilizados os algoritmos de aprendizado Naive Bayes e Decision Tree.

Primeiro os dados são particionados. Nesse workflow foi separado 90% dos dados para aprendizado e 10% para predição. Depois da partição, os dados são inseridos nos módulos de aprendizado e de predição, conforme podemos ver abaixo:



O atributo qe_i12 estava no questionário para os alunos do ENADE e armazena os resultados da seguinte pergunta:

"Ao longo da sua trajetória acadêmica, você recebeu algum tipo de auxílio permanência? No caso de haver mais de uma opção, marcar apenas a bolsa de maior duração".

Os participantes podiam marcar as seguintes alternativas:

A. Nenhum
B. Auxílio moradia
C. Auxílio alimentação
D. Auxílio moradia e alimentação
E. Auxílio permanência
F. Outro tipo de auxílio

Abaixo estão listados os resultados da primeira tentativa. Por precaução rodei mais 2 vezes para me certificar e felizmente os resultados foram consistentes. No final do processamento dos dados foi possível encontrar os seguintes resultados em cada algoritmo ao se analisar a acurácia e a precisão do atributo qe_i12:

Decision Tree:

- Acurácia de 0.9
- Precisão de 0.946

Naive Bayes:

- Acurácia de 0.886
- Precisão de 0.96

Questão 7

Recapitulando as ferramentas utilizadas, temos:

VS Code
Editor de texto da Microsoft, que foi utilizado por conter muitas funcionalidades, extensões e suporte para otimizar o desenvolvimento de software. Ele permite edição e execução de Python Notebooks com as extensões 'Jupyter' e 'Jupyter Notebook Renderers'. O VS Code também é o meu editor de texto favorito para programar.

Python
Escolhi essa linguagem porque existem várias bibliotecas que se especializam em manipulação de dados nela, como também o fato dela ser uma linguagem de fácil compreensão.

KNIME
Esse programa foi utilizado porque possui ferramentas (nós) que permitem a leitura, manipulação e representação gráfica de dados. Além disso, o KNIME possui nós para diversos algoritmos de aprendizado, que utilizei para responder a Questão 6.

SQLite
Essa ferramenta foi uma recomendação do professor e foi essencial para inserir os dados lidos no arquivo.bd. Por meio dela, foi possível criar a conexão com o banco de dados e inserir cada instância. Outro motivo para o uso do SQLite no trabalho foi o tutorial compreensível que encontrei na Web, que ensinava o básico para o manuseio de um banco de dados.

Pandas
Biblioteca mais importante para o trabalho, pois viabiliza a leitura tanto de arquivos .txt como .csv. Muito útil por permitir a leitura de colunas específicas, o que facilita o processo de inserir dados em cada tabela do banco de dados.

GoogleDriveDownloader
Escolhemos usar essa biblioteca para baixar os dados e o Workflow KNIME a partir da minha conta do Google Drive no computador. Esses arquivos são baixados na mesma pasta que o Python Notebook e os dados são copiados para pasta Downloads(que é onde o workflow KNIME vai buscar os dados para leitura).

DB Browser
Esse programa serviu mais para verificar se os dados estavam sendo corretamente adicionados no banco de dados. Por meio dele é possível ver as tabelas e cada dado que foi inserido no arquivo .db de forma intuitiva.

Excel

Foi usado esse programa para ler os arquivos .csv ocasionalmente, assim como para ler o dicionário de dados. O motivo pela escolha do programa foi pela conveniência, já que ele já estava instalado no meu computador.

https://www.diagrams.net/

O site foi utilizado por possuir interface intuitiva e por possuir todas funcionalidades necessárias para construir modelo estrela.