UFRJ

Universidade Federal do Rio de Janeiro
Instituto de Computação
Professor: Geraldo Xexéo
MAB602: Data Warehousing no Suporte à Tomada de Decisão - 2021/1
Prova Individual
Prova: A

Nome: _____

DRE: _____

Data de entrega: 3 de novembro de 2021 até às 23:59 no horário do Moodle.

Esta prova é para execução individual. A prática de cola/plágio resultará na nota zero. Cada aluno ou aluna deve buscar uma solução própria para a prova. Os alunos podem se consultar e discutir sobre formas de resolver as questões (principalmente porque não posso evitar isso).

Para executar a prova, os alunos deverão utilizar um computador pessoal e os softwares que desejarem, sempre identificando.

O resultado dos problemas especificados na prova devem ser apresentados em um relatório em PDF.

Além disso, o PDF deve indicar um repositório GITHUB, onde tudo que foi feito deve estar depositado.

A prova é um processo completo de Data Warehouse e Análise de Dados. Pode ser necessário fazer algumas iterações para obter uma prova consistente, porém o relatório deve ser entregue com a versão final do processo.

Sugiro que a prova seja feita com o KNIME em seu computador ou em Python, no Google Colab, na rede, caso seu computador seja lento.

Para o relatório, recomendo o uso de ferramentas como o Overleaf, o Jupiter, a própria impressão com o Google Colab, o Word ou o Google Docs.

Caso seja feita uma impressão direto das ferramentas, como o Google Colab, não é necessário se preocupar com quebras de página no texto, **porém as imagens não podem ficar quebradas**.

Deve ser identificado o sistema operacional usado (Windows, Linux, etc...).

Como a resposta da prova deve ser um relatório em PDF, ele pode ficar grande demais para ser aceito no Moodle. Não se preocupe com o tamanho do PDF antes da hora. Se no final ficar com um PDF muito grande para colocar como resposta, tente reduzi-lo em sites que fazem isso na rede. Caso contrário, coloque o relatório completo no GitHub e coloque um menor no Moodle, apontando para o relatório completo.

As questões podem ser resolvidas com uma composição de ferramentas, por exemplo criando um programa Python que gere um arquivo CSV que vai ser lido mais tarde por uma planilha para fazer gráficos. Apesar de não ser penalizado por isso na nota, pode não ser a solução ideal, já que é fácil usar gráficos em Python.

Só serão tiradas dúvidas sobre a prova no grupo do Whatsapp, não serão dadas respostas em mensagens pessoais

No caso do uso de uma ferramenta Low Code ou No Code, como o KNIME, **o relatório deve conter as imagens dos diagramas criados e imagens das configurações ou apenas tabelas mostrando os principais valores usados nas configurações**. No Github deve ser colocado o fluxo/arquivo/script usado. No KNIME é possível **exportar** um diagrama.

Boa Sorte!

Parte 1. Questões

1. **(1,5 pontos)** Realize a coleta de dados, obtendo os **microdados do ENADE dos anos de 2019, 2018 e 2017**, no site do INEP, facilmente encontrados na rede.

   Deve ser feito um script/programa/fluxo KNIME ou em outra ferramenta que automatize a recuperação desses dados na rede. Esse programa deve ser apresentado no relatório.

   O relatório final deve indicar onde esses dados podem ser coletados, que arquivos são obtidos, e uma análise textual de que tipo de informação eles possuem.

2. **(2,5 pontos)** Crie um modelo dimensional (estrela) a partir dos dicionários de dados encontrados para os 3 anos.

   Esse modelo dimensional, para cada dimensão, deve abarcar os dados disponíveis nos 3 anos analisados, mesmo que haja diferença entre os anos na dimensão.

   O número de dimensões não deve ultrapassar 15 e não ser menor que 10. Podem ser criadas dimensões formadas pela combinação de colunas, se for interessante para ser usada nas questões 5 e 6 dessa prova.

   Caso sejam identificadas mais de uma tabela fato, pode ser usado mais de um modelo.

   Pode também ser necessário incluir outros dados que não estão nos arquivos de dados, por exemplo, a substituição de códigos por significados, de acordo com o dicionário de dados, ou de acordo com outra base disponibilizada. Se for necessário baixar outra base, isto deve estar descrito na seção anterior.

   Devem ser indicadas as ferramentas usadas para análise e desenho do modelo de dados.

3. **(0,5 ponto)** Crie a base de dados do Data Warehouse em um banco de dados relacional, relatando o *script* SQL usado para isso.

   Pode ser usado qualquer banco relacional, ou consultável via SQL. Quem usar Python ou outra linguagem de programação pode usar um banco de dados simples, como o SQLite.

4. **(1,0 ponto)** Nesta questão é feita a carga de dados. Todos os dados devem ser alimentados em um banco de dados **relacional**, a sua escolha, local ou on-line, de acordo com o modelo dimensional planejado.

   Deve ser detalhadamente descrita a forma de carga de dados. Configurações de ferramentas de carga ou programas feitos devem ser listados. Fluxos do KNIME devem ser descritos por seu desenho e por seu XML.

   Diferença entre as 3 bases que causem problemas nos dados, como dados faltantes, devem ser indicadas.

   Serão aceitas soluções que criem as tabelas em arquivos .CSV, mas apenas no caso de alunos com máquinas muito fracas. Nesse caso, deve ser feita a descrição da máquina no relatório e isso será julgado. *A priori*, nesse caso, é melhor usar o Google Colab com SQLite, por exemplo.

5. **(2,0 pontos)** Nesta questão deve ser feita a análise de dados.

   Devem ser propostas **5 perguntas** que demonstrem alguma característica interessante sobre o resultado do ENADE e criadas, a partir de programas/scripts/fluxos KNIME ou outra ferramenta, pelo menos 1 tabela **e** 1 gráfico que esclareça o questionamento feito.

   Respostas típicas incluem *box-plots*, *scatter-plots* com linhas de regressão, etc. As perguntas esperadas podem questionar se há alguma diferença de resultado o ENADE (fato) em função de uma ou mais dimensões criadas. Dimensões como características demográficas.

   Devem ser apresentados no relatório todos os programas utilizados e a saída proposta.

6. **(2,0 pontos)** Nessa questão deve ser feita uma tentativa de aprender algo a partir dos dados.

   Para isso pode ser usado qualquer programa/biblioteca/módulo de aprendizado.

   Deve ser feita uma proposta de aprendizado, por exemplo: "descobrir a nota do ENADE em função das variáveis X,Y,Z com o algoritmo W". Essa proposta deve ser implementada e seus resultados apresentados. Não é necessário que seja aprendido algo relacionado ao fato, pode ser aprendido, por exemplo, que tipo de aluno acha a prova fácil.

   É obrigatório que haja algum aprendizado, para isso será considerado uma Acurácia maior que 60% e Precisão maior que 60%. Devem ser feitas pelo menos 2 tentativas de aprender a mesma coisa. Caso em 2 tentativas não se alcance a acurácia desejada, isso deve ser relatado. As tentativas devem incluir mudanças significativas nos dados ou nos algoritmos.

7. **(0,5 ponto)** Liste todas as ferramentas utilizadas, indicando o motivo da escolha. Inclua um link de referência para cada ferramenta.