Instituto de Ciências Matemáticas e de Computação – USP

Disciplina: SCC0244 — Mineração a partir de Grandes Bases de Dados

1º Trabalho Prático

Profa. Dra. Agma J. M. Traina, Prof. Dr. Caetano T. Jr.,
Christian C. Bones, João V. O. Novaes

15 de Setembro de 2020

Definição do Trabalho:

Este trabalho deve ser feito em grupos de **3 ou 4** alunos e entregue até o dia **07 de dezembro de 2020**. Todos os participantes devem ser identificados pelo nome, Nº USP e e-mail.

Sobre a entrega:

O trabalho está dividido em duas partes, e ambas devem ser entregues no TIDIA da disciplina em um único arquivo formato em .zip. A primeira parte corresponderá às questões da seção 2. Essas questões serão avaliadas como pontos extras e serão utilizadas caso o aluno necessite. Essa entrega deve conter os seguintes arquivos:

- um (único) arquivo em formato .sql, com os comandos em SQL da resolução do exercício;
- um arquivo em formato .pdf, contendo:
  - A resposta específica de cada questão do exercício. Por exemplo, um print screen das 10 primeiras tuplas retornadas pela consulta.
  - Uma breve explicação de como cada comando funciona.

Já a segunda parte do trabalho corresponde às questões da seção 3. A nota deste trabalho será definida, principalmente, pela apresentação dos resultados dessa seção. Deverão ser entregues os arquivos com o código-fonte utilizado no arquivo .sql. Para a composição da nota seá também avaliada uma apresentação de 15 minutos (dividida entre apresentação e seção de perguntas) que deverá ser feita nos dias 08 e 15 de dezembro de 2020. Cada grupo deve preparar um apresentação com slides, onde deve-se destacar/apresentar os resultados obtidos, as dificuldades encontradas e as escolhas tomadas durante o desenvolvimento do trabalho.

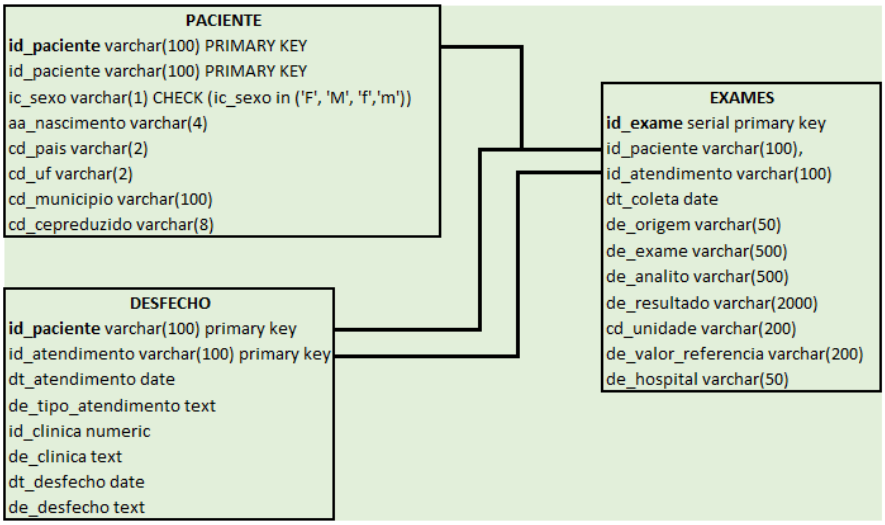
Em caso de dúvidas sobre o trabalho, contatar via e-mail o Pós-Doc Christian Bones (christian.bones@usp.br) ou o aluno PAE João Novaes (novaes.jvo@usp.br).

1 Descrição da Atividade

Será utilizada uma base de dados sobre os registros de pacientes relacionados ao COVID-19 disponibilizado pela FAPESP em:
https://repositoriodatasharingfapesp.uspdigital.usp.br/

A Figura 1 apresenta as tabelas que serão utilizadas neste trabalho. A tabela paciente contém todos os pacientes que receberam atendimento no hospital com suspeita de Covid-19. A tabela exame contém todos os exames requisitados para um determinado paciente e os respectivos resultados. E a tabela desfecho contém quais foram as conclusões obtidas para o paciente.

Figure 1: Diagrama das tabelas do banco



Para facilitar o trabalho, uma base de dados está disponível em uma máquina virtual (**VirtualBox**). Para instalar a maquina virtual, foi disponibilizado o manual de instalação no TIDIA da disciplina.

2 Questões

Nessa seção são apresentadas algumas questões para recuperar informações sobre os dados presentes na base de dados, úteis para cada um tomar um conhecimento geral dos dados disponibilizados.

1. Qual a quantidade de pacientes presente na base de dados?
   Quantos são homens e quantos são mulheres?

2. Baseado na questão anterior:
   Qual é faixa etária dos pacientes homens e mulheres?
   Qual a distribuição dos quartis dentro de cada faixa?
   Qual a distribuição em cada gênero por década de vida?

3. Qual a maior quantidade de exames solicitados para um único paciente ?
4. Qual é a média de exames pedidos para homens e para mulheres?
5. Quantos exames de Coronavírus (2019-nCoV) foram solicitados? Quantos deles apresentam resultado positivo?
6. Para cada idade, mostre os resultados dos exames de Coronavírus (2019-nCoV).
7. Qual é o desfecho para a maioria dos casos registrados? E para cada distribuição por gênero e por década de vida?
8. Considerando as tabelas e as consultas solicitadas anteriormente, escreva/projete uma consulta para extrair algum conhecimento da base de dados que não foi descoberto pelas consultas anteriores. Apresente uma breve justificativa do objetivo da consulta e, por que esse objetivo é relevante.

3 Trabalhos

Nessa Seção são apresentados dois trabalhos que deverão ser entregues em 07 de dezempro e apresentados entre os dias 08 e 15 de dezembro de 2020.

1. Realizar o download dos dados de Covid-19 da FAPESP e criar uma nova base de dados. O modelo poderá ser o mesmo apresentado anteriormente. Deverão ser apresentadas as soluções para cada um dos problemas encontrados durante a carga dos dados, justificando o motivo das decisões tomadas.
2. Criar uma Árvore de Decisão para classificar os dados para prever qual será o resultado do exame "NOVO CORONAVÍRUS 2019 (SARS-CoV-2), DETECÇÃO POR PCR". Um dos pontos mais importantes dessa questão é selecionar quais atributos serão utilizados durante a classificação, lembrando que dependendo do número e da entropia dos atributos, o tempo de construção e a qualidade da Árvore de Decisão podem mudar.

   Seguem abaixo algumas instruções sobre está questão:

   - Deverão ser apresentados os atributos escolhidos e as justificativas do por que eles foram escolhidos. Também deve ser apresentado um gráfico que mostra a precisão da Árvore de Decisão.
   - A fim de verificar a qualidade da Árvore de Decisão, as tuplas da base deverão ser divididas em dois conjuntos: treinamento e teste. A quantidade de tuplas em cada um dos conjuntos deve ser definida pelos alunos e corretamente justificada.
   - Nessa questão devem ser utilizadas bibliotecas externas para montar a Árvore de Decisão e classificar os dados. Recomenda-se utilizar a biblioteca scikit-learn do python (https://scikit-learn.org/stable/modules/tree.html).
   - Os dados que serão utilizados durante a classificação deverão ser recuperados por um comando em SQL. Esse tutorial (https://wiki.python.org.br/BancosDeDadosSql) explica como realizar as consultas via python.

Referências

Este trabalho utiliza dados disponibilizados pelo repositório COVID-19 Data Sharing/BR, disponível em: https://repositoriodatasharingfapesp.uspdigital.usp.br