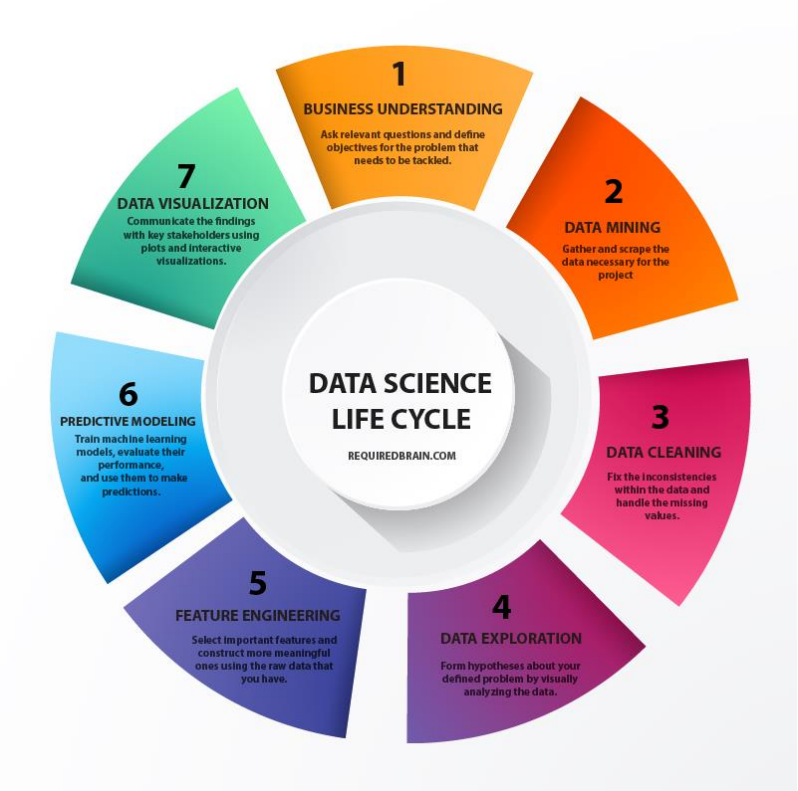
# Aplicação 2
# Introdução a *Data Science* com Estruturas de Dados I [1]
# Análise de dados de COVID19

Considerando as etapas do ciclo de vida de *Data Science*, ilustradas na Figura 1 (AWARI, 2022), temos a etapa *Data Exploration* (Exploração dos Dados ou podemos dizer Compreensão dos Dados) na qual o cientista de dados faz uso de diversas técnicas de análise de dados, estatística e visualização para explorar o conjunto de dados obtido para melhor compreendê-lo e posterior realização da engenharia de atributos e criação da predição propriamente dita.

Figura 1.: Ciclo de Vida de Ciência de Dados




Fonte: https://awari.com.br/tudo-sobre-ciencia-de-dados/. Data da Consulta: 03/03/2022.

Segundo Facelli (2021), "a análise das características presentes em um conjunto de dados permite a descoberta de padrões e tendências que podem fornecer informações valiosas que ajudem a compreender o processo que gerou os dados".

Assim, na Aplicação 2, faremos o uso da estrutura de dados **lista encadeada** ou **lista duplamente encadeada** para realizar algumas análises no conjunto de dados (*dataset*) "COVID-19 *Mexico Patient Health Dataset*" (dfcovid.xlsx). Sugestão: ao obter o *dataset*, fazer uso do excel para converter de "xlsx" para "csv" para fazer

[1] Baseado no trabalho de semestres anteriores da profa. Dra. Valéria Farinazzo Martins (Mackenzie).

a leitura em C++ de cada objeto (registro) como linha de um arquivo texto ou, se preferir, leia como “xlsx” mesmo.

Originalmente esse *dataset* consiste em 95839 casos que são formados por 19 atributos e registrados pelo governo mexicano entre 15 de janeiro de 2020 e 3 de maio de 2020 para dados sobre a doença de Covid-19.

O estudo de Yavuz e Dudak (2022) realizou a anáslie do *dataset* fazendo uso da ferramenta Weka (*Waikato Environment for Knowledge Analysis*) com os algoritmos de classificação do Aprendizado de Máquina Supervisionado: Naive Bayes, *k-Nearest Neighbor*, *Support Vector Machine* e *Decision Tree* (Random Forest)

Os 20 atributos do dataset são apresentados nos Quadros 1 e 2.

Quadro 1 – Atributos do Dataset df_covid.xslx – Parte I

| Nome do Atributo | Tipo de dado | Intervalo | Descrição |
|---|---|---|---|
| sexo | Numérico | 0-1 | 1-Mulher, 2-Homem |
| tipo_paciente | Numérico | 1-2 | Tipo 1, Tipo 2 |
| intubado | Numérico | 1-99 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL<br>99 = NÃO DISPONÍVEL |
| pneumonia | Numérico | 1-99 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL<br>99 = NÃO DISPONÍVEL |
| idade | Numérico | 0-113 | Idade do paciente no intervalor de 0 a 113 |
| diabetes | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| copd<br>(doença pulmonar obstrutiva crônica) | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| asma | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| imunossupressao | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| hipertensao | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |

Fonte: (COVID19, 2019)

Quadro 2 – Atributos do Dataset df_covid.xlsx – Parte II

| Nome do Atributo | Tipo de dado | Intervalo | Descrição |
|---|---|---|---|
| outras_doencas | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| cardiovascular | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| obesidade | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| icu<br>(insuficiência renal crônica) | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| teste_covid | Numérico | 1-3 | 1 = COVID-19 Positivo<br>2 = COVID-19 Negativo<br>3 = NÃO APLICÁVEL |
| fumante | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| outro_caso | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| gravidez | Numérico | 1-98 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL |
| uci<br>(unidade de terapia intensiva) | Numérico | 1-99 | 1 = SIM<br>2 = NÃO<br>98/97 = NÃO APLICÁVEL<br>99 = NÃO DISPONÍVEL |
| obito | Numérico | 0-1 | 1 = SIM, 0 = NÃO |

Fonte: (COVID19, 2019)

Tendo por base o apresentado, elaborar um programa contendo opções de um menu para:

1. Leitura dos dados: no qual o arquivo original deve ter sido mapeado para um arquivo contendo os atributos relevantes para o grupo ler e montar a **lista encadeada ou duplamente encadeada**. Pensar na melhor forma de montar a estrutura de dados e o que carregar nela.
2. Cinco opções contendo métodos para análise de dados, como: contar a quantidade de mulheres grávidas com COVID19; total de homens fumantes que faleceram; e/ou outras questões pertinentes para análise de dados. Cada grupo irá planejar as questões que deseja responder sobre os dados mapeados na estrutura montada. Com os resultados obtidos em cada um dos cinco métodos, apresentar uma descrição textual e/ou criar tabelas/gráficos ou outros recursos que julgar pertinentes para demonstrar esses resultados.

3. Encerra a Aplicação: os dados alocados são liberados e a aplicação desenvolvida é finalizada.

**Observações:**
1. O trabalho pode ser feito por grupos de até 5 pessoas.

2. Um único aluno do grupo deverá publicar o trabalho no Moodle.

3. Deverá ser entregue um relatório com os resultados da **"Atividade Aplicação 2"** deste projeto com base no Template disponibilizado contendo:

- Dados dos integrantes do grupo (nome e TIA).
- Decisões relativas ao *dataset*, por exemplo: remoção de objetos (motivo), eliminação de colunas (atributos - motivos), outros.
- Informações e detalhes sobre as cinco opções selecionadas pelo grupo para análise.
- **Printscreen de testes de execução das opções do menu. Ao menos 2 testes de cada opção, se for permitido, caso contrário basta um único teste da opção**. ESSENCIAL!
- O relatório deve conter ao seu final um Apêndice contendo o código fonte desenvolvido (separado por arquivos, se for o caso). Em cada arquivo inserir:
    - um cabeçalho (comentário) com as identificações completas de todos os membros do grupo.
    - Documentação adequada e inclusão de comentários úteis e informativos.

4. Junto ao relatório também devem ser entregues os códigos fontes em C++ e o *dataset* utilizado (contendo as modificações realizadas pelo grupo). A entrega deve ser realizada na data limite 01 de junho até as 23h59min.

5. Deverá ser realizada uma apresentação do projeto no dia 02 de junho no horário da aula:

- O grupo deverá apresentar o processo de construção da solução do seu projeto, resultados obtidos e testes do menu de opções no tempo máximo de 6 (seis) minutos.

O projeto será avaliado de acordo com os seguintes critérios:

- Completude, clareza e ausência de erros de linguagem no relatório;
- Funcionamento correto da Aplicação;
- O trabalho deve ser desenvolvido na **linguagem C++** e será testado usando o compilador do **Dev C++**.
- O quão fiel é o programa quanto à descrição do enunciado;
- Indentação, comentários e legibilidade do código;
- Clareza na nomenclatura de variáveis e funções;
- Apresentação realizada com clareza, conhecimento e cumprimento do tempo estabelecido.

Para auxiliar na documentação do código e entendimento do que é um programa com boa legibilidade siga as dicas apresentadas nas páginas abaixo:

- http://www.ime.usp.br/~pf/algoritmos/aulas/layout.html
- http://www.ime.usp.br/~pf/algoritmos/aulas/docu.html

Como este trabalho pode ser feito em **grupo**, evidentemente você pode "*discutir*"' o problema dado com outros **grupos**, inclusive as "*dicas*"' para chegar às soluções, mas você deve ser responsável pela solução final e pelo desenvolvimento da sua aplicação.

Um vídeo ilustrativo sobre a leitura de um arquivo texto e montagem em uma estrutura em C++ pode ser encontrado em AlgoritmosAZ - C/C++ (2019).

O *dataset* obtido segue dentro da nossa disciplina no Moodle para seu uso e modificação.

**Referências**

AlgoritmosAZ - C/C++ Ler arquivo e armazenar em uma struct em C++. 2019. Vídeo encontrado em: https://www.youtube.com/watch?v=UTFw9SY42bY. Data da consulta: 03/03/2022.

AWARI, Tudo sobre Ciência de Dados: o que é, como funciona e qual sua importância. Fevereiro, 2022. Endereço: https://awari.com.br/tudo-sobre-ciencia-de-dados/. Data da Consulta: 03/03/2022.

*COVID-19. Mexico Patient Health Dataset.* (2020, 05 19). Disponível em Kaggle.com: https://www.kaggle.com/riteshahlawat/covid19-mexico-patient-health-dataset . Data da consulta: 15 de março de 2021.

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. 2ª. Edição. Rio de Janeiro: LTC- Livros Técnicos e Científicos. 2012. Endereço da biblioteca do Mackenzie: https://app.minhabiblioteca.com.br/reader/books/9788521637509/epubcfi/6/2[%3Bvnd.vst.idref%3Dcover]!/4/2/2%4051:3. Data da consulta: 08/03/2022.

YAVUZ, Ü. N. A. L.; DUDAK, Muhammed Nuri. Classification of Covid-19 Dataset with Some Machine Learning Methods. **Journal of Amasya University the Institute of Sciences and Technology**, v. 1, n. 1, p. 30-37. Disponível em: https://dergipark.org.tr/en/pub/jauist/issue/55760/748667. Data da Consulta: 08/03/2022.