**Avaliação A1 - Nano Degree - Big Data Analytics e Processamento de Dados Massivos**

**Professor: Ramon Pereira**

**2020/2**

**Contexto**

O seu grupo foi selecionado para realizar uma análise de dados sobre Síndrome Respiratória Aguda no Brasil. Vocês deverão responder as questões abaixo relacionadas a doença utilizando uma base de dados histórica do ano de 2018.

O grupo deverá preparar uma apresentação descritiva da base de dados e com as respostas encontradas nas questões abaixo. A apresentação será realizada no final da aula.

Além da apresentação, o grupo deverá entregar esse documento preenchido com as respostas.

**Análise Descritiva**

1. Assumindo o tamanho da base de dados atual, qual ferramenta e plataforma vocês recomendariam para o processamento destes dados e por quê?
2. Caso a quantidade de registros dessa base fosse aumentado em 100 vezes, a recomendação seria a mesma? Por que? Se não, qual seria a nova recomendação e por quê?
3. De posse da tabela, faça a identificação dos tipos das colunas dessa base de dados
4. Qual a quantidade de pessoas que estão acima do IMC ideal?
5. Faça um histograma da variável CLASSI\_FIN. O que esse histograma nos diz?
6. Quantas pessoas realizaram PCR do tipo convencional?
7. Faça o histograma da distribuição geográfica por unidade federativa (UF) dos pacientes com desta base de dados.
8. Faça um histograma por mês de internação dos pacientes
9. É possível identificar o número de mortes nesta base de dados? Se sim, coloque o número de mortos por SRAG em 2018 na sua frequência absoluta e na sua frequência relativa. Se não, qual estratégia o grupo utilizaria para conseguir estes números de forma real ou por uma aproximação boa?
10. Qual a quantidade média de internações por SRAG no ano de 2018?

**Análise Preditiva**

1. Enumere os tipos de tarefas preditivas que podem ser realizadas com esta base de dados e dê exemplos com as colunas cada uma delas.
2. Seguindo os princípios da metodologia CRISP-DM, escolha uma das tarefas respondidas na questão número um e crie um ciclo de um modelo de análise preditiva de forma textual para ser explicado.