- **Challenge:** Sprint 3
- **Descrição da Entrega:** Análise do banco de dados sobre diabetes e doenças relacionadas
- **Banco de Dados:** 'Diabetes.csv'

Contexto

Após a análise de um tipo de enfermidade incidente sobre a população (COVID-19), o próximo passo para um cientista de dados é o entendimento preliminar de outra enfermidade e um segundo passo em direção à previsão de enfermidades com algoritmos de machine learning. Para isto, é preciso que o cientista entenda as principais variáveis determinantes da diabetes e a inter-relação entre estas variáveis.

Diabetes está entre as doenças crônicas mais prevalentes no mundo, impactando milhões de pessoas a cada ano e exercendo uma carga financeira significativa na economia. O diabetes é uma doença crônica grave em que os indivíduos perdem a capacidade de regular eficazmente os níveis de glicose no sangue, o que pode levar à redução da qualidade de vida e da expectativa de vida. Após a digestão dos alimentos, os açúcares resultantes são liberados na corrente sanguínea, sinalizando ao pâncreas para liberar insulina. A insulina ajuda a permitir que as células do corpo utilizem esses açúcares presentes no sangue como fonte de energia. O diabetes é geralmente caracterizado pelo corpo não produzir insulina suficiente ou não conseguir usar a insulina produzida de maneira eficaz como necessário.

Variáveis:

- **Diabetes_012:** 0 = sem diabetes 1 = pré-diabetes 2 = diabetes
- **HighBP:** 0 = sem pressão alta 1 = pressão alta
- **HighChol:** 0 = sem colesterol alto 1 = colesterol alto
- **CholCheck:** 0 = não fez verificação de colesterol nos últimos 5 anos 1 = fez verificação de colesterol nos últimos 5 anos
- **BMI:** Índice de Massa Corporal
- **Smoker:** Você já fumou pelo menos 100 cigarros em toda a sua vida? [Nota: 5 maços = 100 cigarros] 0 = não 1 = sim
- **Stroke:** (Alguma vez lhe disseram que) você teve um derrame. 0 = não 1 = sim
- **HeartDiseaseorAttack:** Doença cardíaca coronariana (DCC) ou infarto do miocárdio (IM) 0 = não 1 = sim
- **PhysActivity:** Atividade física nos últimos 30 dias - excluindo trabalho 0 = não 1 = sim
- **Fruits:** Consumo de frutas 1 ou mais vezes por dia 0 = não 1 = sim
- **Veggies:** Consumo de vegetais 1 ou mais vezes por dia 0 = não 1 = sim
- **HvyAlcoholConsump:** Bebedores excessivos (homens adultos que consomem mais de 14 bebidas por semana e mulheres adultas que consomem mais de 7 bebidas por semana) 0 = não 1 = sim
- **AnyHealthcare:** Possui algum tipo de cobertura de saúde, incluindo seguro de saúde, planos pré-pagos como HMO, etc. 0 = não 1 = sim
- **NoDocbcCost:** Houve algum momento nos últimos 12 meses em que você precisou consultar um médico, mas não pôde devido ao custo? 0 = não 1 = sim
- **GenHlth:** Você diria que, em geral, sua saúde é: escala de 1 a 5 1 = excelente 2 = muito boa 3 = boa 4 = regular 5 = ruim

- **MentHlth:** Agora pensando sobre sua saúde mental, que inclui estresse, depressão e problemas emocionais, por quantos dias nos últimos 30 dias sua saúde mental não esteve boa? escala de 1 a 30 dias
- **PhysHlth:** Agora pensando sobre sua saúde física, que inclui doença física e lesões, por quantos dias nos últimos 30 dias sua saúde física não esteve boa? escala de 1 a 30 dias
- **DiffWalk:** Você tem grande dificuldade para caminhar ou subir escadas? 0 = não 1 = sim
- **Sex:** 0 = feminino 1 = masculino
- **Age:** Categoria de idade em 13 níveis (_AGEG5YR ver código) 1 = 18-24 9 = 60-64 13 = 80 anos ou mais
- **Education:** Nível de educação escala de 1 a 6 1 = Nunca frequentou a escola ou apenas jardim de infância 2 = Da 1ª à 8ª série (Ensino Fundamental) 3 = Da 9ª à 11ª série (Algum Ensino Médio) 4 = 12ª série ou GED (Ensino Médio completo) 5 = Faculdade de 1 a 3 anos (Alguma faculdade ou escola técnica) 6 = Faculdade de 4 anos ou mais (Graduado)
- **Income:** Escala de renda escala de 1 a 8 1 = menos de $10,000 5 = menos de $35,000 8 = $75,000 ou mais

Objetivo

No entendimento de possíveis problemas de saúde que podem acometer a população, o cientista deve analisar o conjunto de dados de forma a entender o inter-relacionamento entre as variáveis. Para isso, ele deverá fazer uma entrega da análise do banco de dados, seguindo os seguintes passos:

- **Valor Total da Entrega (100 pontos)**
- **Documentação (80 pontos)**
    - Análise, limpeza e formatação dos dados: Remoção de duplicatas, identificação de valores nulos e tratamento de outliers (20 pontos)
    - Estatísticas descritivas das variáveis (20 pontos)
    - Gráficos: entre 3 e 4 gráficos, focando na distribuição e inter-relacionamento entre as variáveis (10 pontos)
    - Análise de correlação e testes de hipóteses: Cálculo da matriz de correlação e execução de pelo menos dois testes de hipótese que faça sentido entre os dados (20 pontos)
    - Organização (10 pontos)
- **Apresentação (20 pontos)**
    - Descrição resumida do projeto, das variáveis e dos tratamentos realizados na base (5 pontos)
    - Comentários sobre as estatísticas descritivas das variáveis e comportamento dos principais gráficos (5 pontos)
    - Conclusões obtidas através dos testes de correlação e testes de hipótese (5 pontos)
    - Organização (5 pontos): Cinco lâminas de apresentação executiva. Primeiro slide com objetivo e resumo executivo, 3 slides com principais gráficos, análise das estatísticas descritivas, com título e breve texto (parágrafo) com análise, slide final com conclusão geral da análise de dados. Este material (PDF) é utilizado para apresentação executiva do projeto de dez minutos. Os detalhes ficam no Jupyter notebook.

Entrega: O arquivo com a documentação deverá ser entregue em formato ‘.ipynb’ (Notebook), enquanto o arquivo da apresentação deverá ser entregue em formato ‘.pdf’ (Portable Document Format). A não entrega de algum destes documentos ocasiona na diminuição da pontuação. A entrega deverá ser feita exclusivamente pelo Teams.