Tratamento e Preparação dos Dados para Análise

1 Conhecendo a base

O diabetes é uma condição crônica grave que compromete a capacidade do organismo de regular os níveis de glicose no sangue de forma eficaz, o que pode resultar em uma diminuição da qualidade de vida e da expectativa de vida.

O Sistema de Vigilância de Fatores de Risco Comportamentais (BRFSS) é uma pesquisa de saúde conduzida anualmente por telefone pelo Centro de Controle e Prevenção de Doenças dos Estados Unidos (CDC). Esse levantamento coleta informações de milhares de americanos sobre comportamentos de risco à saúde, condições crônicas e o uso de serviços preventivos.

Neste projeto, foi utilizado o conjunto de dados do BRFSS referente ao ano de 2015, disponível no

Kaggle: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

1.1 Dicionário de Dados:

- **Diabetes_binary:** 0 = sem diabetes, 1 = com diabetes
- **HighBP:** 0 = sem pressão alta, 1 = com pressão alta
- **HighChol:** 0 = sem colesterol alto, 1 = com colesterol alto
- **CholCheck:** 0 = não fez exame de colesterol na vida, 1 = fez exame de colesterol alguma vez
- **BMI:** Índice de Massa Corporal (IMC)
- **Smoker:** 0 = não fumante, 1 = fumante
- **Stroke:** 0 = sem histórico de AVC, 1 = com histórico de AVC
- **HeartDiseaseorAttack:** 0 = sem histórico de doença cardíaca ou ataque cardíaco, 1 = com histórico de doença cardíaca ou ataque cardíaco
- **PhysActivity:** 0 = não pratica atividade física, 1 = pratica atividade física
- **Fruits:** 0 = não consome frutas, 1 = consome frutas
- **Veggies:** 0 = não consome vegetais, 1 = consome vegetais
- **HvyAlcoholConsump:** 0 = não consome álcool em altas quantidades, 1 = consome álcool em altas quantidades
- **AnyHealthcare:** 0 = não tem plano de saúde, 1 = tem plano de saúde

- **NoDocbcCost:** 0 = não foi ao médico por questões financeiras, 1 = foi ao médico por questões financeiras (últimos 12 meses)
- **GenHlth:** Saúde geral (1 a 5) - 1 = Excelente, 2 = Muito boa, 3 = Boa, 4 = Aceitável, 5 = Ruim
- **MentHlth:** Nos últimos 30 dias, quantos dias a saúde mental não foi boa (0 a 30)
- **PhysHlth:** Nos últimos 30 dias, quantos dias a saúde física não foi boa (0 a 30)
- **DiffWalk:** 0 = não tem dificuldade para caminhar, 1 = tem dificuldade para caminhar
- **Sex:** 0 = feminino, 1 = masculino
- **Age:** Idade em faixas 1 = 18-24; 2 = 25-29; 3 = 30-34; 4 = 35-39; 5 = 40-44; 6 = 45-49; 7 = 50-54; 8 = 55-59; 9 = 60-64; 10 = 65-69; 11 = 70-74; 12 = 75-79; 13 = 80+
- **Education:** Níveis de ensino 1 = nunca frequentou a escola; 2 = escola primária; 3 = escola secundária incompleta ; 4 = escola secundária; 5 = faculdade incompleta ou curso técnico; 6 = completou faculdade ou diplomas superiores
- **Income:** Renda anual (dólares) em faixas 1 = < 10.000; 2 = 10.000-14.999; 3 = 15.000-19.999; 4 = 20.000-24.999; 5 = 25.000-34.999; 6 = 35.000-49.999; 7 = 50.000-74.999; 8 = 75.000+

2 Tratamento dos dados e preparação para análise

2.1 Traduzindo os nomes das colunas do banco de dados como apresentado abaixo.

```
Data columns (total 22 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Diabetes                70692 non-null  float64
 1   Pressao_Alta            70692 non-null  float64
 2   Colesterol_Alto         70692 non-null  float64
 3   Colesterol_Exame        70692 non-null  float64
 4   IMC                     70692 non-null  float64
 5   Fumante                 70692 non-null  float64
 6   AVC                     70692 non-null  float64
 7   Problema_Cardiaco       70692 non-null  float64
 8   Atividade_Fisica        70692 non-null  float64
 9   Come_Frutas             70692 non-null  float64
 10  Come_Legumes            70692 non-null  float64
 11  Consumo_Bebida_Alcoolica  70692 non-null  float64
 12  Plano_Saude             70692 non-null  float64
 13  Sem_Dinheiro_Consultas  70692 non-null  float64
 14  Saude_Geral             70692 non-null  float64
 15  Dias_Problemas_Mentais  70692 non-null  float64
 16  Dias_Problemas_Fisicos  70692 non-null  float64
 17  Dificuldade_Andar       70692 non-null  float64
 18  Genero                  70692 non-null  float64
 19  Faixa_Idade             70692 non-null  float64
 20  Ensino                  70692 non-null  float64
 21  Faixa_Renda             70692 non-null  float64
dtypes: float64(22)
memory usage: 11.9 MB
```

2.2 Transformando o tipo das colunas binárias de float para categórico como observa na tabela abaixo

```
Data columns (total 22 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Diabetes                  70692 non-null  category
 1   Pressao_Alta              70692 non-null  category
 2   Colesterol_Alto           70692 non-null  category
 3   Colesterol_Exame          70692 non-null  category
 4   IMC                       70692 non-null  float64
 5   Fumante                   70692 non-null  category
 6   AVC                       70692 non-null  category
 7   Problema_Cardiaco         70692 non-null  category
 8   Atividade_Fisica          70692 non-null  category
 9   Come_Frutas               70692 non-null  category
 10  Come_Legumes              70692 non-null  category
 11  Consumo_Bebida_Alcoolica  70692 non-null  category
 12  Plano_Saude               70692 non-null  category
 13  Sem_Dinheiro_Consultas    70692 non-null  category
 14  Saude_Geral               70692 non-null  category
 15  Dias_Problemas_Mentais    70692 non-null  float64
 16  Dias_Problemas_Fisicos    70692 non-null  float64
 17  Dificuldade_Andar         70692 non-null  category
 18  Genero                    70692 non-null  category
 19  Faixa_Idade               70692 non-null  category
 20  Ensino                    70692 non-null  category
 21  Faixa_Renda               70692 non-null  category
dtypes: category(19), float64(3)
memory usage: 2.9 MB
```

2.3 Verificando se as colunas numéricas do tipo float podem ser convertidas para o tipo inteiro.

```
Data columns (total 22 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Diabetes                  70692 non-null  category
 1   Pressao_Alta              70692 non-null  category
 2   Colesterol_Alto           70692 non-null  category
 3   Colesterol_Exame          70692 non-null  category
 4   IMC                       70692 non-null  int8
 5   Fumante                   70692 non-null  category
 6   AVC                       70692 non-null  category
 7   Problema_Cardiaco         70692 non-null  category
 8   Atividade_Fisica          70692 non-null  category
 9   Come_Frutas               70692 non-null  category
 10  Come_Legumes              70692 non-null  category
 11  Consumo_Bebida_Alcoolica  70692 non-null  category
 12  Plano_Saude               70692 non-null  category
 13  Sem_Dinheiro_Consultas    70692 non-null  category
 14  Saude_Geral               70692 non-null  category
 15  Dias_Problemas_Mentais    70692 non-null  int8
 16  Dias_Problemas_Fisicos    70692 non-null  int8
 17  Dificuldade_Andar         70692 non-null  category
 18  Genero                    70692 non-null  category
 19  Faixa_Idade               70692 non-null  category
 20  Ensino                    70692 non-null  category
 21  Faixa_Renda               70692 non-null  category
dtypes: category(19), int8(3)
memory usage: 1.5 MB
```

2.4 Salvando a base de dados após o tratamento de dados no formato parquet. Podemos observar que após os tratamentos o espaço em memória original de 11,9MB passou para 1,5MB uma redução de 90,4% no espaço em memória utilizado.