Sobre o conjunto de dados “income.csv”

Imagine que trabalhamos no departamento de marketing de uma empresa de planejamento financeiro e gostaríamos de identificar potenciais clientes em um banco de dados que compramos. Nosso cliente-alvo é qualquer pessoa com uma renda anual acima de $50.000, mas geralmente não temos informações sobre a renda de um novo cliente em potencial. Portanto, gostaríamos de desenvolver um modelo que analise outros fatores para nos ajudar a prever se um cliente em potencial tem uma renda acima do limite de $50.000.

Para resolver esse problema, são fornecidos dados de 32.560 clientes em potencial. As seguintes são as variáveis em nosso conjunto de dados:

- **age (idade):** é a idade autodeclarada do cliente.
- **workClassification (classificação do trabalho):** é o tipo de empregador para o qual o cliente trabalha. Exemplos incluem *Private (Privado)*, *Local-gov (Governo Local)*, *Federal-gov (Governo Federal)*, etc.
- **EducationLevel (nível de educação):** é o nível mais alto de educação alcançado pelo cliente em potencial. Exemplos dos valores incluem *Bachelors (Bacharelado)*, *HS-grad (Ensino Médio)*, *Masters (Mestrado)*, etc.
- **EducationYears (anos de educação):** é o número de anos de educação que um cliente possui.
- **maritalStatus (estado civil):** é a designação do estado civil do cliente. Exemplos incluem Divorced (Divorciado), Separated (Separado), Never-married (Nunca casado), etc.
- **occupation (ocupação):** é o tipo de trabalho que o cliente exerce. Exemplos incluem Adm-clerical (Administrativo), Sales (Vendas), Tech-support (Suporte Técnico), etc.
- **relationship (relacionamento):** é o relacionamento relatado entre o cliente e seu gerente designado.
- **Race (raça):** é a identidade racial autodeclarada do cliente.
- **gender (gênero):** é a identidade de gênero autodeclarada – seja Male (Masculino) ou Female (Feminino).
- **WorkHours (horas de trabalho):** é o número de horas em uma semana que o cliente normalmente trabalha.
- **nativeCountry (país de origem):** é a nação de origem do cliente em potencial.
- **income (renda):** é a classe que estamos tentando prever e tem valores: <=50K e >50K.