A base de dados escolhida para o projeto A3 foi **“Employee dataset”**, está disponível no Kaggle: <https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset>

As colunas da base de dados estão divididas da seguinte forma:

* “education”: coluna referente a qualificação educacional dos funcionários (
* “joining\_year”: coluna referente ao ano em que o funcionário entrou na empresa (Tipo Int)
* “city”: coluna referente ao local ou cidade onde cada funcionário está (Tipo String)
* “payment\_tier”: coluna referente a categorização dos funcionários em diferentes níveis salarias (Tipo float)
* “age”: coluna referende a idade (Tipo Int)
* “gender”: coluna referente a identidade de gênero (Tipo String)
* “ever\_benched”: coluna que indica se um funcionário já esteve temporariamente sem trabalho atribuído (Tipo String)
* “experience”: coluna referente aos anos de experiencia do funcionário na área atual (Tipo String)
* “leaver\_or\_not”: coluna referente ao destino (Tipo String)

A variável target será payment\_tier, com objetivo de prever em qual nível salarial um funcionário pertence. Com isso será útil para entender os fatores que afeta se um funcionário vai receber mais ou menos dinheiro.

Uma variável que precisa ser modificada será a “experience”, ela será transformada em uma variável numérica, a convertendo os valores para números inteiros(int), removendo características não numéricas.

As colunas “joining\_year” e “age” são variáveis numéricas que ode estar em diferentes escalas. Com a padronização, ela pode ser aplicada para garantir que todas as variáveis numéricas tenham uma escala comparável.