

Mestrado em Informática Aplicada

US Health Insurance – Regression Models

Inteligência Artificial Aplicada

Guilherme Ribeiro de Oliveira – 200100084

Santarém, 24 de junho de 2024

Índice

[Dimensão Pedagógica 3](#_Toc170043134)

[Dimensão Técnica 3](#_Toc170043135)

[Dimensão Comparativa 4](#_Toc170043136)

[Dimensão Aplicativa 4](#_Toc170043137)

[Dimensão de Dados 5](#_Toc170043138)

[Dimensão de Autodescoberta 5](#_Toc170043139)

[Bibliografia 5](#_Toc170043140)

# Dimensão Pedagógica

O projeto tem como objetivo criar um fluxo de trabalho claro e compreensível para a previsão de custos dos seguros médicos usando Modelos de Regressão. Os principais aspetos são:

#### Documentação

* **Comentários e Explicações**: O código inclui comentários detalhados e explicações para cada função, tornando mais fácil para outros entenderem os processos.
* **Carregamento dos Dados**: Cada etapa do carregamento dos dados é comentada, incluindo o tempo de execução, para fornecer uma visão clara de como os dados são manipulados desde o início.
* **Pré-processamento**: As etapas de pré-processamento, como a normalização dos dados e a conversão de variáveis categóricas, são explicadas em detalhes, facilitando a compreensão e a replicação do fluxo de trabalho.

**Clareza**

* **Estrutura Lógica**: O código é estruturado de maneira lógica, com uma separação clara entre o carregamento de dados, pré-processamento, treinamento de modelos, avaliação de resultados visualização dos dados e salvamento dos dados.

# Dimensão Técnica

O projeto visa prever os custos médicos de indivíduos com base nas suas características demográficas e de saúde. Para esse objetivo as opções preferidas foram:

* **Bibliotecas**:
  + ***pandas*:** para manipulação de dados
  + ***numpy*:** para cálculos numéricos
  + ***sklearn*:** para *Machine* *Learning*
  + ***matplotlib*:** para visualização de dados
  + ***joblib*:** para guardar os modelos
* **Modelo**: Foram utilizados um modelo de regressão linear simples, um modelo *ridge*, um modelo *lasso* e um modelo *decission tree regressor* para prever os custos seguros médicos.
* **Avaliação do modelo:** Os modelos foram avaliados usando as seguintes métricas:
  + **R^2:** mede a proporção da variância dos dados explicada pelo modelo.
  + **MSE:** mede o erro quadrático médio entre os valores previstos e os valores reais.
  + **MAE:** mede o erro absoluto médio entre os valores previstos e os valores reais.

# Dimensão Comparativa

**Contraste com outras abordagens:**

**Modelos de Regressão**: O código utiliza diferentes modelos de regressão para prever os valores dos custos de seguros médicos nos EUA, comparando os seus desempenhos.

* **Regressão Linear Simples**: Utiliza *LinearRegression* do *Scikit-learn*, que é um modelo de regressão linear básico, simples de interpretar e rápido para treinar, adequado para situações onde há uma relação linear clara entre as variáveis independentes e a variável dependente.
* **Regressão Ridge**: Utiliza *Ridge* do *Scikit-learn*. Esta abordagem é útil para evitar sobreajuste em modelos de regressão linear, particularmente quando há multicolinearidade entre as variáveis independentes.
* **Regressão Lasso**: Utiliza *Lasso* do *Scikit-learn*. Este método não só ajuda a evitar sobreajuste, mas também pode eliminar variáveis irrelevantes, tornando o modelo mais interpretável e eficiente.
* **Árvore de Decisão**: Utiliza *DecisionTreeRegressor* do *Scikit-learn*, um modelo não linear que captura interações complexas entre variáveis. As árvores de decisão são intuitivas e fáceis de visualizar, mas podem ser propensas a sobreajuste se não forem adequadamente ajustadas.

# Dimensão Aplicativa

**Módulos Reutilizáveis**

* **Estrutura Modular**: O código foi estruturado de maneira que possa ser reutilizado e modificado para diferentes tarefas de processamento de dados e modelagem. Por exemplo, as funções de pré-processamento, avaliação de modelos e visualização são genéricas e podem ser aplicadas a outros conjuntos de dados e modelos.
* **Facilidade de Modificação**: O código permite que novos modelos possam ser integrados facilmente ao pipeline existente, como foi feito com a inclusão de modelos de Regressão Ridge, Lasso e Decision Tree. Além disso, novos métodos de avaliação e métricas podem ser adicionados sem a necessidade de reescrever grandes partes do código.

**Visualização**

* **Ferramentas Adicionais de Visualização**: O código inclui funções para criar gráficos de resíduos e comparações entre valores reais e previstos, usando a biblioteca matplotli. Estas visualizações são essenciais para entender o desempenho dos modelos e identificar possíveis áreas onde é possível melhorar.

**Operação**

* **Execução Local e na Nuvem**: A aplicação inicialmente foi projetada para ser executada localmente, utilizando *Python* para o desenvolvimento e avaliação dos modelos. No entanto, com pequenas modificações, a aplicação foi adaptada para implantação em nuvem, mais especificamente no *Google Colab*.

# Dimensão de Dados

**Natureza dos Dados**:

* Os dados utilizados são do ficheiro *medical\_insurance.csv*, ficheiro esse que foi retirado do website *kaggle* e possui informações sobre cerca de 2772 utentes.

# Dimensão de Autodescoberta

**Implantação em Nuvem**:

* A aplicação foi adaptada para ser implementada em serviços de nuvem, como *Google Colab*. Permitindo assim inserir *datasets* em ambientes *cloud* e ainda utilizar ferramentas do *Google Colab* como guardar ficheiros pretendidos pelo usuário na sua conta google drive, como foi feito no código.

# Bibliografia

Projeto no *Colab* - <https://colab.research.google.com/drive/1Bg_Z--C_uvUHzEIulo9aot9_vCJO_lXG#scrollTo=G3g9m_lYBH6R>

*Dataset* - <https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset>

*Google Colab* - <https://colab.google>

*Scikit-learn* - <https://scikit-learn.org/stable/>