ISSN: 2319-0124
josif 2023
15ª JORNADA CIENTÍFICA E TECNOLÓGICA
12º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

Uso da Regressão Linear Múltipla para Determinação de Gastos Médicos em Seguros de Saúde

Isaias S. de SOUZA[1]; Maria Jaqueline dos S. SILVA[2]; Diego SAQUI[3]

RESUMO

A Inteligência Artificial (IA) e o Aprendizado de Máquina (AM) são áreas da Computação que vem ganhando destaque em vários aspectos, dentre eles, a área Atuarial, que é responsável por estudar as nuances ao redor da precificação de seguro ou plano de saúde dado um conjunto de características de um grupo ou indivíduo segurado. Nesse sentido, quando lidamos com desafios relacionados à determinação de uma variável que depende de outros fatores (variáveis) para sua explicação, um método comum utilizado para esse tipo de problema é o da Regressão Linear Múltipla. Visto que, esse tem como objetivo, trazer uma quantidade significativa de variáveis independentes que possam explicar com melhor precisão o resultado final da variável dependente. Sendo assim, esse trabalho tem como objetivo a aplicação do modelo de regressão linear múltipla, utilizando do método mínimos quadrados ordinários (MQO), sobre um conjunto de dados de segurados a fim de determinar os gastos totais sobre um contrato de seguro de saúde.

Palavras-chave: Aprendizado de Máquina; Inteligência Artificial; Regressão Linear; Grupo Segurado.

1. INTRODUÇÃO

O seguro pode ser definido como um contrato entre uma pessoa para com outra, onde uma das partes se denomina segurador e se obriga mediante o recebimento de um prêmio, a indenizar a outra parte, denominado segurado, em caso de prejuízo resultante de risco futuro previsto ocorrido após determinado período (“SUSEP”, 2023). Dentro do contexto de seguros, há também o destinado a cobrir riscos possíveis que ocorrem contra pessoas, como morte, invalidez, doenças e até mesmo atendimentos médicos. Sendo que para a determinação do preço justo e operacional, pago pelo segurado, é necessário a avaliação de vários aspectos sobre o segurado ou grupo segurado, quando o seguro for coletivo. As características comumente determinantes são a idade, doenças pré existentes e condições insalubres (CHAVES, 2022).

Situado dentro do ecossistema da Inteligência Artificial, o Aprendizado de Máquina (AM) é uma área que busca aprender com base em conjunto de dados estabelecidos e com isso, tomar decisões cada vez mais precisas. Deste modo, é de extrema importância que se tenham grandes quantidades de dados de amostra para que esse aprendizado seja cada vez mais coerente e alinhado com a realidade (LUDERMIR, 2021).

Com o passar dos anos, o volume de dados tende a aumentar de maneira exponencial (LOBO, 2017). Dessa forma, a computação está sempre em evolução buscando métodos para manipulação desses dados de forma otimizada, identificando padrões e obtendo novas perspectivas,

[1]Isaias Santos de Souza, IFSULDEMINAS – *Campus* Muzambinho. E-mail: 12171002079@muz.ifsuldeminas.edu.br.
[2]Maria Jaqueline dos Santos Silva, IFSULDEMINAS – *Campus* Muzambinho. E-mail: 12171002078@muz.ifsuldeminas.edu.br.
[3]Diego Saqui, IFSULDEMINAS – *Campus* Muzambinho. E-mail: diego.saqui@muz.ifsuldeminas.edu.br.

os chamados *insights*. Outro aspecto importante, é o fato de trazer informações significativas sobre os dados, sendo esses muitas vezes de baixo nível, devendo passar por um filtro antes de serem apresentados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Por esse motivo, para a definição do método que deve ser aplicado a um conjunto de dados com determinadas características, utiliza-se como apoio o processo de Análise e Extração do Conhecimento, do inglês, *Knowledge Discovery in Databases (KDD),* que, por meio das etapas de seleção, processamento e transformação, a base de dados é validada e passada para o estágio de mineração, que faz a incorporação do método verificado, finalizando o processo com a interpretação e descoberta de novas percepções sobre o resultado obtido. Nesse sentido, a computação divide o Aprendizado de Máquina (AM) em três categorias: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço.

Dada a complexidade de determinação do preço final de um contrato de seguro, uma forma de garantir uma boa explicabilidade dessa variável dependente Y, quando há duas ou mais variáveis X que possam influenciar em seu resultado, pode ser investigada por meio do modelo de Regressão Linear Múltipla. Portanto, este trabalho tem como finalidade a aplicação de um modelo de AM, Regressão Linear Múltipla, para análise, conhecimento e previsão da precificação de gastos médicos com seguro de um indivíduo ou grupo segurado.

2. MATERIAL E MÉTODOS

A metodologia foi norteada pelas principais etapas: Seleção da base de dados, Pré-processamento, Transformação, Mineração de dados, Avaliação e Interpretação (AI Magazine, 1996). A ferramenta utilizada foi a linguagem Python na plataforma Google Colaboratory, que pode ser acessado no link no rodapé[4].

A base de dados do presente trabalho, *Medical insurance price prediction*, foi retirada do repositório de *datasets* Kaggle[5], caracterizada por ter 2772 observações ou instâncias de 7 variáveis. As variáveis observadas são dadas por vários fatores que podem afetar as despesas médicas, incluindo idade, sexo, Índice de Massa Corpórea (IMC), tabagismo, número de filhos e região. A despesa médica aqui é uma das variáveis e foi predita a partir das informações das outras seis, ou seja, esta será a variável resposta.

Durante a fase de seleção de dados foram retirados todos os dados considerados *outliers* de todas colunas numéricas. Selecionando o valores presentes no intervalo interquartil (IQR) descrito pelos 50% valores médios quando ordenados do mais baixo para o mais alto. Com esta seleção a base de dados teve a quantidade de instâncias reduzidas para 2440 (DHADSE, 2021). Houve

[4] https://colab.research.google.com/drive/1r4tGIc85jLN_jsbePc8oCNy51X7klZX_?usp=sharing.
[5] https://www.kaggle.com/datasets/harishkumardatalab/medical-insurance-price-prediction

também a separação entre dados de treino e dados de teste, em uma proporção de 30% de teste e 70% para treinamento para a estimação das variáveis do modelo e verificação dos resultados.

Na fase de pré-processamento, os dados tiveram verificações a respeito de valores não preenchidos, ou seja, nulos. Neste trabalho, não foram identificados valores nestas condições.

Na etapa de transformação foram tratadas as variáveis categóricas para conversão em variáveis *dummies,* que consiste em criar mais variáveis para cada estado possível dada a coluna selecionada. Após este estágio a base de dados passou a ter onze variáveis exógenas (BLANKMEYER, 2022).

Já na etapa de mineração de dados foi analisada a correlação entre as variáveis independentes e a variável explicada, onde foram constatadas 27% das variáveis exógenas com correlação moderada. Foi construído o modelo de regressão utilizando para isto a função *OLS* da biblioteca *statsmodels* do Python.

Para a avaliação e interpretação dos resultados foram utilizadas métricas como coeficiente de determinação ajustado (R^2), Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE) (JÚNIOR, 2023).

3. RESULTADOS E DISCUSSÕES

O modelo resultado é dado pela equação, da Equação 1.

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \quad (1)$$

Onde os parâmetros do modelo foram estimados de acordo com o Quadro 1.

k	α	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
11	1858,35	235,18	78,05	389,52	1028,29	830,07	-6915,23	8773,59	1246,59	831,04	-104,51	-114,76

Quadro 1. Parâmetros encontrados

As métricas de desempenho utilizadas para avaliação do modelo de regressão tiveram seus valores obtidos para os dados de teste conforme o Quadro 2.

R^2 Ajustado	MAE - Erro Médio Absoluto	MSE - Erro Quadrático Médio	RMSE - Raiz do Erro Quadrático Médio
0,600062	2691,56	22081232,78	4699,07

Quadro 2. Métricas de desempenho obtidas

De acordo com as métricas obtidas, pode-se concluir que:

- 60% das variabilidades da coluna endógena pode ser explicada pelas variáveis do modelo.
- As métricas para regressão MAE, MSE e RMSE, levando em base o desvio padrão (7435,54) proporcionalmente, indicou que o modelo tem um bom desempenho em relação à

variabilidade dos dados.

Portanto, com base nos resultados analisados a partir das métricas mais comumente utilizadas para modelos de regressão, pode-se concluir que o modelo de regressão desenhado para os dados trabalhados tem um desempenho razoável.

4. CONCLUSÕES

O presente estudo procurou predizer os gastos de seguros de saúde utilizando a técnica de regressão linear pelo método de Mínimos Quadrados Ordinários (MQO). Para tal base de dados, podem existir métodos de aprendizado de máquina com maior eficácia, como Redes Neurais Artificiais e XGBoost, que podem reduzir os erros com resultados mais robustos, porém exigem um custo maior em questão de desempenho e de conhecimento para melhores ajustes em seus hiperparâmetros. Sendo que para resultados razoáveis com dados não fortemente correlacionados a regressão linear múltipla pode ser uma solução fácil de ser implementada e interpretada.

5. REFERÊNCIAS

BLANKMEYER, E. **How robust is linear regression with dummy variables ?** Rochester, NY, 15 jul. 2022. Disponível em: https://papers.ssrn.com/abstract=4194218. Acesso em: 2 jul. 2023.

CHAVES, R. H. S. **Mercado de seguros no Brasil: funcionalidades no movimento de reprodução ampliada do capital.** 7 mar. 2022.

DHADSE, A. **Removing Outliers. Understanding How and What behind the Magic.** Analytics Vidhya, 24 abr. 2021. Disponível em: https://medium.com/analytics-vidhya/removing-outliers-understanding-how-and-what-behind-the-magic-18a78ab480ff. Acesso em: 2 jul. 2023.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases.** AI Magazine, v. 17, n. 3, p. 37–37, 15 mar. 1996.

JÚNIOR, C. DE O. **Métodos para Regressão: Entendendo as métricas R^2, MAE, MAPE, MSE e RMSE**. Data Hackers, 10 fev. 2023. Disponível em: https://medium.com/data-hackers/prevendo-n%C3%BAmeros-entendendo-m%C3%A9tricas-de-regress%C3%A3o-35545e011e70. Acesso em: 2 jul. 2023.

LOBO, L. C. **Inteligência Artificial e Medicina.** Revista Brasileira de Educação Médica, v. 41, p. 185–193, jun. 2017.

LUDERMIR, T. B. **Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências.** Estudos Avançados, v. 35, n. 101, p. 85–94, abr. 2021.

SUSEP. **S - T – SUSEP - Superintendência de Seguros Privados.** SUSEP. Disponível em: https://www.gov.br/susep/pt-br/conteudo-do-glossario/s-t. Acesso em: 1 jul. 2023.