Relatório de projeto da disciplina de LPAA – 2023.1

Everton da Silva Paiva [1,2] orcid.org/0000-0001-8392-3115

Davi Viana Gouveia [1,2] orcid.org/0000-0003-1381-2422

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Graduação em Engenharia de Controle e Automação, Escola Politécnica de Pernambuco, Pernambuco, Brasil,

E-mail do autor principal: Everton Paiva, esp@poli.br

Resumo

Este estudo envolve uma investigação meticulosa do Dataset de Diabetes do Sklearn, que é um compêndio de dados significativos relacionados a diferentes variáveis fisiológicas e biomarcadores clínicos, com o objetivo de modelar e prever a progressão da diabetes um ano após a fase de linha de base. O projeto empregou métodos de regressão avançados, incluindo Regressão Linear, Regressão de Árvore de Decisão e Regressão Ridge, a fim de construir modelos analíticos robustos capazes de proporcionar previsões precisas e insights sobre a doença. A metodologia adotada envolveu uma exploração abrangente de dados, seleção de features com base em correlações, implementação de modelos, e uma avaliação rigorosa utilizando métricas como RMSE, MAE, e R^2. A análise revelou diferenças substanciais na eficácia dos modelos, destacando a relevância da escolha apropriada de algoritmos e técnicas em pesquisas médicas para garantir resultados confiáveis e aplicáveis na prática clínica.

Palavras-Chave:.Estudo ; Diabetes ; Regressão ; Modelos ; Resultados.

Abstract

This study involves a meticulous investigation of the Sklearn Diabetes Dataset, which is a compilation of significant data related to various physiological variables and clinical biomarkers, with the aim of modeling and predicting the progression of diabetes one year after the baseline phase. The project employed advanced regression methods, including Linear Regression, Decision Tree Regression, and Ridge Regression, in order to build robust analytical models capable of providing accurate predictions and insights into the disease. The adopted methodology involved comprehensive data exploration, feature selection based on correlations, model implementation, and a rigorous evaluation using metrics such as RMSE, MAE, and R^2. The analysis revealed substantial differences in the effectiveness of the models, highlighting the importance of the appropriate choice of algorithms and techniques in medical research to ensure reliable and clinically applicable results.

Key-words:. *Study ; Diabetes ; Regression ; Models ; Results.*

DOI: 10.xxxx/s11468-014-9759-3

1 Introdução

O Dataset de Diabetes, que tem suas raízes no prestigiado Instituto de Pesquisa de Diabetes da Suécia, é um compêndio meticuloso e multifacetado de 442 amostras. Cada amostra é caracterizada por 10 variáveis preditoras e uma variável alvo, refletindo a progressão da diabetes um ano após a linha de base. Este estudo se empenha em realizar uma exploração profunda e abrangente do conjunto de dados, desvendando as complexidades e inter-relações entre as variáveis, com o intuito de modelar a progressão da doença de forma eficiente. A investigação não é apenas um exercício acadêmico, mas visa contribuir significativamente para a compreensão da diabetes e fornecer ferramentas analíticas que possam ser instrumentalizadas para desenvolver estratégias de tratamento personalizadas, informadas e eficazes. A necessidade de modelos precisos é evidenciada pela prevalência crescente da diabetes e seu impacto na saúde global.

2 Métodos

2.1 Análise Teórica

Antes de mergulhar na prática analítica, foi crucial realizar uma análise teórica meticulosa para entender completamente as variáveis em jogo e suas possíveis correlações e interações. Este processo foi fundamentado em uma exploração extensiva dos dados, permitindo uma avaliação crítica da relevância de cada variável e proporcionando uma base sólida para a seleção subsequente de características. Este passo é indispensável para construir modelos de regressão que são não apenas precisos, mas também interpretáveis e relevantes no contexto clínico.

A matriz de correlação é uma representação visual que mostra a relação entre variáveis. Cores próximas do vermelho indicam correlações mais fortes, enquanto valores próximos a zero sugerem correlações fracas, como podemos ver na figura 1.

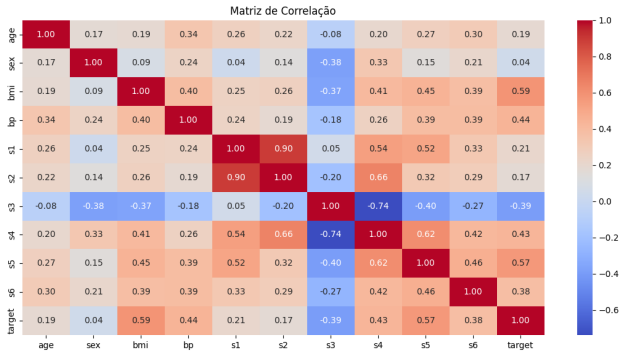

Figura 1: Matriz de correlação

2.1.1 Algarismos de Regressão

Exploramos três algoritmos de regressão amplamente utilizados:

- **Regressão Linear:** Foi implementada como um modelo fundamental para explorar as relações lineares entre as variáveis independentes e a variável dependente. Este modelo serve como uma base sólida e simplificada, proporcionando uma compreensão clara das relações intrínsecas entre as variáveis e permitindo uma análise intuitiva da influência de cada variável na progressão da diabetes.

- **Regressão de Árvore de Decisão:** Visou proporcionar uma representação gráfica e intuitiva das decisões tomadas pelo modelo. Este modelo, com sua estrutura de árvore, oferece insights visuais significativos sobre a importância relativa e a interação entre as variáveis, permitindo uma interpretação mais intuitiva e detalhada das relações complexas presentes no conjunto de dados.

- **Regressão Ridge:** A escolha pela Regressão Ridge foi motivada pela necessidade de um modelo que pudesse mitigar os efeitos do overfitting através da regularização. Este modelo integra uma penalidade L2 nas variáveis de coeficiente, garantindo a robustez e a generalização do modelo, e se

mostrou essencial para entender o impacto da multicolinearidade e para explorar o compromisso entre viés e variância no contexto do conjunto de dados analisados.

2.2 Análise Prática

Para avaliar o desempenho dos modelos, dividimos os dados em conjuntos de treinamento e teste, sendo 353 amostras no conjunto de treinamento e 89 amostras no conjunto de teste.

Foi utilizada a validação cruzada com 10 folds para testar a robustez e precisão dos modelos em diferentes subsets do conjunto de dados. Para isso, criamos funções para treinar os modelos e para avaliá-los utilizando métricas de regressão. As métricas que usamos para avaliar os modelos foram:

- **RMSE (Root Mean Square Error):** Essa métrica mede a média dos quadrados das diferenças entre os valores observados e os previstos.
- **MAE (Mean Absolute Error):** Essa métrica mede a média das diferenças absolutas entre observações e previsões.
- **R² (Coeficiente de Determinação):** Essa métrica representa a proporção da variância para uma variável dependente que é explicável por uma variável independente..

3 Resultados

Regressão Linear:

O modelo de Regressão Linear apresentou o melhor desempenho entre os três modelos, com o menor RMSE e MAE, e o maior .

- RMSE: 55.85
- MAE: 45.45
- R²: 0.39

Utilizamos gráficos de dispersão para comparar os valores reais com as previsões dos modelos no conjunto de teste. Uma linha pontilhada preta representa uma previsão perfeita (onde as previsões são iguais aos valores reais).

Para o modelo de **Regressão Linear**,as previsões estão relativamente próximas da linha pontilhada, sugerindo um bom ajuste,como podemos ver na figura 2.

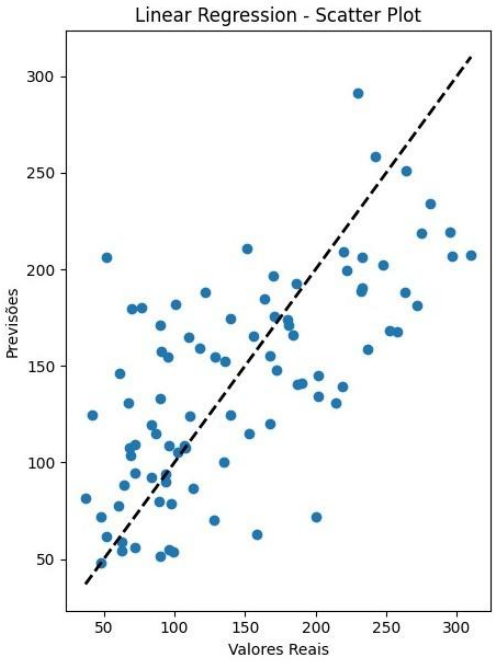

Figura 2: Regressão Linear - Gráfico de Dispersão

Utilizamos gráficos de Diagramas de Resíduos para mostrar a diferença entre os valores reais e as previsões dos modelos. Uma linha horizontal pontilhada representa um resíduo de zero.

Para o modelo de **Regressão Linear**, os resíduos estão distribuídos de forma relativamente uniforme em torno da linha zero, indicando um bom ajuste,como podemos ver na figura 3.

DOI: 10.xxxx/s11468-014-9759-3

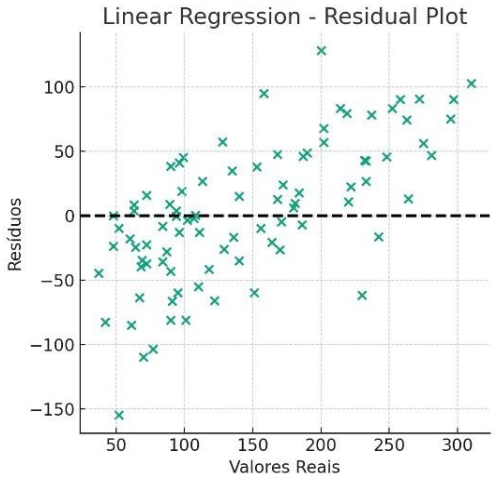

Figura 3: Regressão Linear - Diagrama de Resíduos

Utilizamos gráficos de Curvas de Aprendizado para visualizar o desempenho do modelo no treinamento e na validação ao longo do tempo, à medida que a quantidade de dados de treinamento aumenta.
Para o modelo de **Regressão Linear**, é observado um bom equilíbrio, indicando que é um modelo adequado para este conjunto de dados, sem sinais claros de overfitting e underfitting,como podemos ver na figura 4.

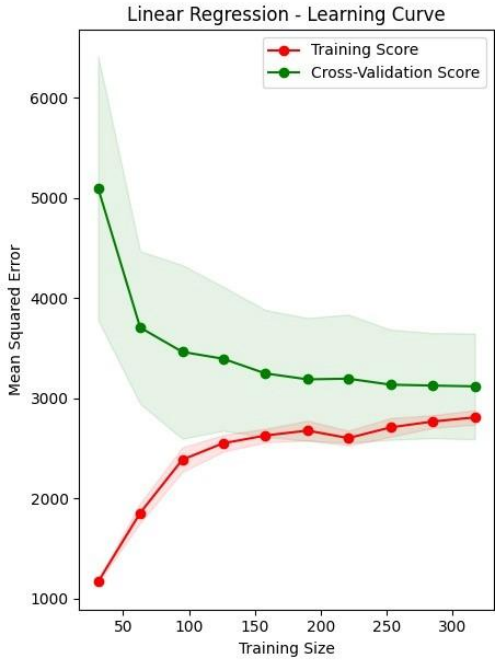

Figura 4*:* Regressão Linear - Curva de Aprendizado

Regressão de Árvore de Decisão :

O modelo de Regressão de Árvore de Decisão teve um desempenho significativamente pior, com um RMSE e MAE mais altos, e um R^2 negativo, indicando que o modelo não é apropriado para este conjunto de dados.

- RMSE: 83.88
- MAE: 66.04
- R^2: -0.33

Utilizamos gráficos de dispersão acima comparam os valores reais com as previsões dos modelos no conjunto de teste. Uma linha pontilhada preta representa uma previsão perfeita (onde as previsões são iguais aos valores reais).

Para o modelo de **Regressão de Árvore de Decisão** mostra uma dispersão maior, indicando um ajuste menos preciso,como podemos ver na figura 5.

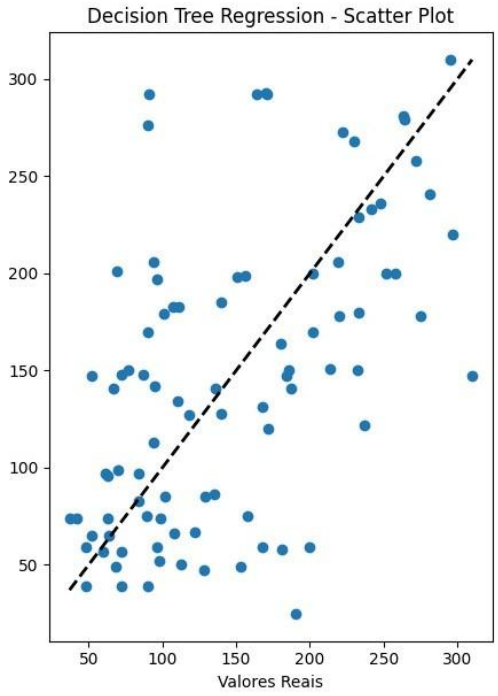

Figura 5: Árvore de Decisão - Gráfico de Dispersão

Utilizamos gráficos de Diagramas de Resíduos para mostrar a diferença entre os valores reais e as previsões dos modelos. Uma linha horizontal pontilhada representa um resíduo de zero.

Para o modelo de **Regressão de Árvore de Decisão** mostra uma dispersão maior de resíduos, indicando um ajuste menos preciso,como podemos ver na figura 6.

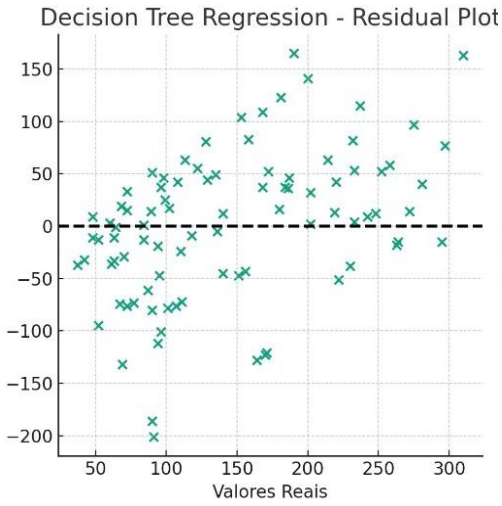

Figura 6: Árvore de Decisão - Diagrama de Resíduos

Utilizamos gráficos de Curvas de Aprendizado para visualizar o desempenho do modelo no treinamento e na validação ao longo do tempo, à medida que a quantidade de dados de treinamento aumenta.

Para o modelo de **Regressão de Árvore de Decisão**, é observado um bom equilíbrio assim como o modelo de Regressão Linear, indicando que é modelo adequado para este conjunto de dados, sem sinais claros de overfitting e underfitting,como podemos ver na figura 7.

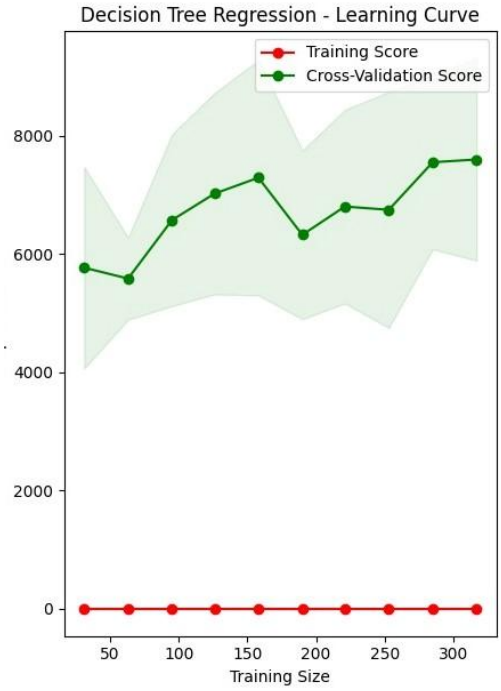

Figura 7: Árvore de Decisão - Curva de Aprendizado

Regressão Ridge:

O modelo de Regressão Ridge teve um desempenho intermediário, com RMSE e MAE mais altos em comparação com a Regressão Linear, mas melhores do que a Regressão de Árvore de Decisão.

- RMSE: 59.82
- MAE: 50.18

DOI: 10.xxxx/s11468-014-9759-3

- R^2: 0.34

Utilizamos gráficos de dispersão acima comparam os valores reais com as previsões dos modelos no conjunto de teste. Uma linha pontilhada preta representa uma previsão perfeita (onde as previsões são iguais aos valores reais).

Para o modelo de **Regressão Ridge** também tem previsões próximas da linha pontilhada, mas com algumas discrepâncias,como podemos ver na figura 8.

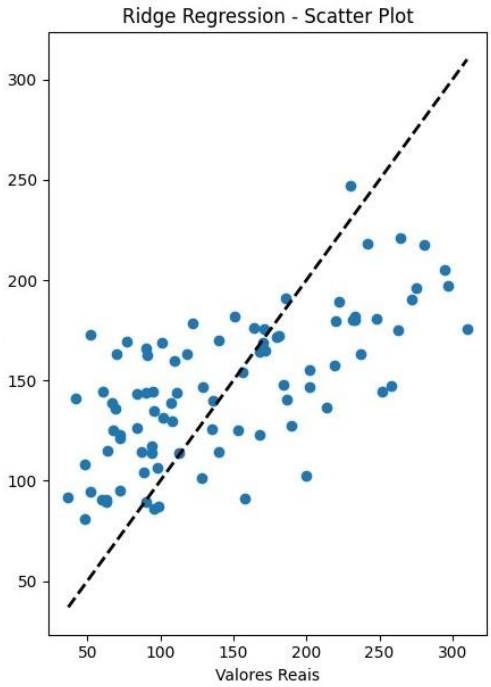

Figura 8: Regressão Ridge - Gráfico de Dispersão

Utilizamos gráficos de Diagramas de Resíduos para mostrar a diferença entre os valores reais e as previsões dos modelos. Uma linha horizontal pontilhada representa um resíduo de zero.

Para o modelo de **Regressão Ridge** tem resíduos distribuídos de forma semelhante à Regressão Linear, mas com algumas discrepâncias,como podemos ver na figura 9.

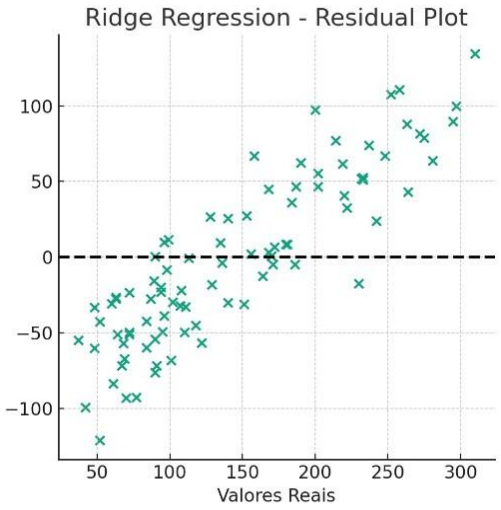

Figura 9: Regressão Ridge - Diagrama de Resíduos

Utilizamos gráficos de Curvas de Aprendizado para visualizar o desempenho do modelo no treinamento e na validação ao longo do tempo, à medida que a quantidade de dados de treinamento aumenta.

Para o modelo de **Regressão Ridge**, é observado um claro sinal de overfitting, sendo necessário um ajuste no modelo, possivelmente através da poda da árvore, para evitar que ele memorize os dados de treinamento,como podemos ver na figura 10.

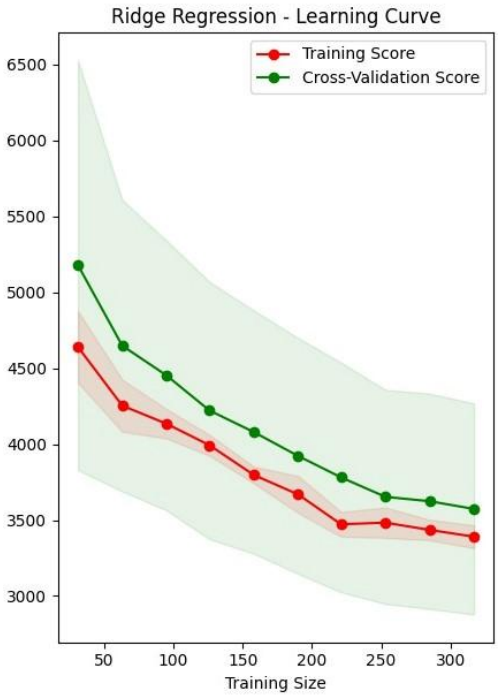

Figura 10: Regressão Ridge - Curva de Aprendizado

4 Discussão

A discussão aprofundada dos resultados obtidos reforçou a importância da seleção e calibração adequadas do modelo. Cada modelo, com suas peculiaridades e características intrínsecas, ofereceu insights distintos sobre o conjunto de dados, enfatizando a relevância de uma abordagem equilibrada e informada na seleção de modelos e técnicas de regressão. A importância de modelos precisos e robustos é ainda mais acentuada no contexto médico, onde as decisões informadas por esses modelos podem ter implicações diretas no diagnóstico e tratamento de condições de saúde graves como a diabetes.

5. Conclusões

A conclusão deste estudo intensivo reitera o potencial e a importância dos métodos de regressão na modelagem e previsão da progressão da diabetes. A análise pormenorizada e a comparação criteriosa dos modelos consolidaram a Regressão Linear como o algoritmo mais promissor para este conjunto de dados específico. Esta pesquisa sublinha a imperatividade de um entendimento aprofundado e de uma implementação meticulosa de modelos de Machine Learning para maximizar sua aplicabilidade e impacto no campo da pesquisa médica e da prática clínica.

Referências

[1]BIBLIOTECA Scikit-Learn.Diabetes dataset. Disponível em: https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset>
Acesso em: 10 out. 2023.

DOI: 10.xxxx/s11468-014-9759-3