UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Maria Bianca Sousa Costa

1. Introdução

Com a função *shape* no python, foi obtido o tamanho do dataset, tendo 100 linhas e 3 colunas. Sendo as colunas “number_courses”, “time_study’ e “Marks” que foi traduzida com a função *rename*.

Número de matérias: Representa o número de disciplinas que o aluno está cursando. Esse atributo pode ser relevante para entender a carga de estudo do aluno e seu envolvimento acadêmico. Sendo assim, uma variável independente.

Tempo Médio de estudo por dia em horas: Indica a quantidade de tempo que o aluno dedica aos estudos. Esse atributo pode ser considerado relevante para avaliar o esforço e a dedicação do aluno no aprendizado. Também sendo uma variável independente.

Notas: Representa as notas obtidas pelos alunos nas disciplinas. Esse é o atributo alvo da regressão para inferir sobre ele com base nas informações dos outros atributos. Caracterizando uma variável dependente.

Abaixo se encontra uma tabela com os cinco primeiros valores de cada coluna. Obtida com a função head.

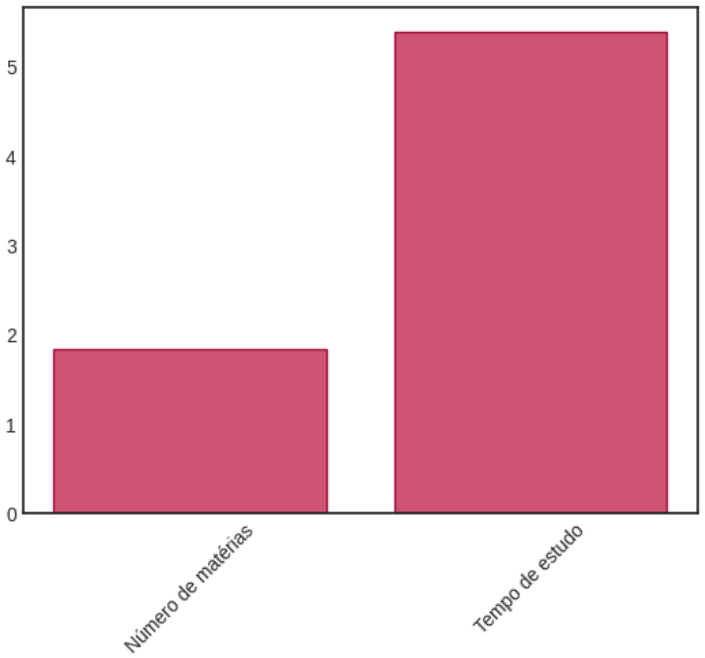
Número de matérias	Tempo de estudo	Notas
3	4.508	19.202
4	0.096	7.734
4	3.133	13.811
6	7.909	53.018
8	7.811	55.299

2. Observações

3. Resultados e discussão

Na primeira questão foi usado o modelo de Lasso que é um método de regularização usado na regressão linear. É uma técnica que adiciona uma penalidade à função de perda durante o treinamento do modelo, a fim de reduzir a complexidade e evitar overfitting. O score da regressão foi de 0.94 e o erro quadrático médio foi de 3.76. A figura abaixo mostra o atributo evidenciado com Lasso que foi o tempo de estudo.

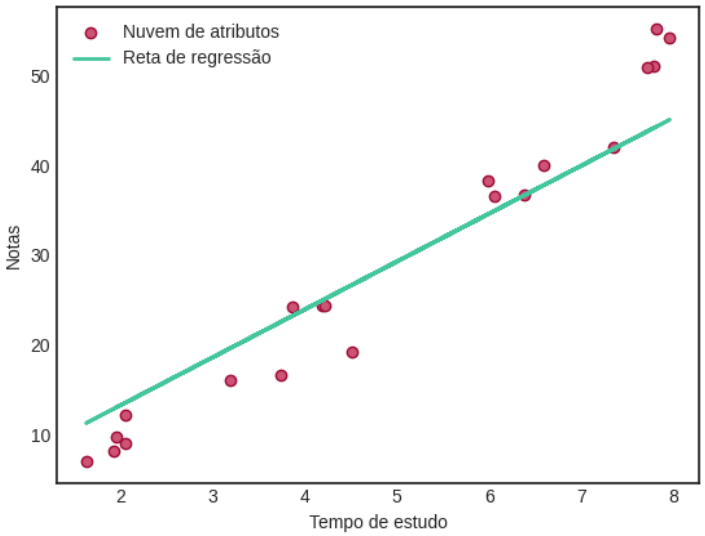
Figura 01: Atributo mais relevante usando Lasso



Fonte: Produzido pelo autor, 2023

Na segunda questão, foi feita a regressão linear que é uma técnica estatística utilizada para modelar a relação entre uma variável dependente contínua e uma ou mais variáveis independentes. Ela assume uma relação linear entre essas variáveis e tenta encontrar a reta que melhor se ajusta aos dados observados. Essa técnica é frequentemente usada para prever ou estimar valores futuros com base em padrões históricos.

Figura 02: Influência do tempo de estudo na nota dos alunos



Fonte: Produzido pelo autor, 2023

De acordo com o gráfico, quanto maior o tempo dedicado ao estudo maior será a nota do aluno.

Na terceira questão foi usado o regressor Lasso e Ridge para compará-los e obter o regressor mais eficiente, o resultado obtido está na tabela abaixo:

Tabela 01: Melhores parâmetros e Scores

Melhor parâmetro para Lasso (α):	Melhor score Lasso:	Melhor parâmetro Ridge (α):	Melhor score Ridge:
0.1	0.93	10	0.93

Fonte: Produzido pelo autor, 2023

O regressor Ridge é uma extensão da regressão linear que lida com o problema de multicolinearidade, que ocorre quando há alta correlação entre as variáveis independentes. Ela adiciona um termo de regularização à função objetivo da regressão linear para penalizar os coeficientes das variáveis independentes. Isso ajuda a reduzir a magnitude dos coeficientes e, assim, reduz o impacto de variáveis altamente correlacionadas. Ele é útil quando há multicolinearidade e quando é importante controlar o ajuste excessivo (overfitting) do modelo.

O Grid Search consiste em definir uma grade de valores para os hiperparâmetros e avaliar o desempenho do modelo treinado para cada combinação desses valores. Assim, é possível encontrar a combinação de hiperparâmetros que resulta no melhor desempenho do modelo.

O K-fold é uma técnica utilizada na validação cruzada de modelos de regressão. A validação cruzada é uma abordagem para avaliar a capacidade de generalização de um modelo, ou seja, quão bem ele se sairá em dados não vistos. A ideia é dividir o conjunto de dados em partições, treinar e testar o modelo k vezes, onde cada uma das partições é usada como conjunto de teste e as outras partições são usadas como conjunto de treinamento. O modelo é treinado nos dados de treinamento e avaliado nos dados de teste. Esse processo é repetido muitas vezes, de modo que cada partição seja utilizada como conjunto de teste uma vez.

A seguir os scores obtidos na quarta questão:

Tabela 02: Scores médios

Score médio para Lasso:	Score médio para Ridge:
0.92	0.92

Fonte: Produzido pelo autor, 2023

4. Conclusões

A primeira questão foi mostrou que dentre as três colunas do dataset as a variável independente mais importante foi o Tempo de estudo tendo como única variável dependente a coluna de Notas. Já na segunda questão, a regressão se mostrou satisfatória, mostrando a relação entre Tempo de estudo e as Notas, já como os melhores parâmetros para Lasso. A terceira questão mostrou os acertos dos regressores que foram similares.

A quarta questão mostrou que tanto Lasso quanto Ridge possuem o mesmo desempenho considerando a base de dados estudada, portanto foi decidido permanecer com Lasso, visto que foi a primeira técnica utilizada e com o melhor parâmetro de alpha (α) para essa técnica.

5. Próximos passos

A regressão logística é aplicada quando a variável dependente é binária. Ela estima os coeficientes de regressão que quantificam o efeito das variáveis independentes sobre a probabilidade de ocorrência de uma determinada classe. Os coeficientes estimados são usados para calcular as probabilidades preditas e, com base nelas, é feita a classificação. Portanto, mesmo com os resultados satisfatórios, é relevante considerar fazer uma regressão logística com esse dataset, transformando a coluna notas em aprovados e não aprovados e em seguida comparar os dois tratamentos dos dados.