Documentação do Trabalho de Machine Learning

Maria Mello e Carolina Cruz

1. Nome e Link do Dataset

Nome do dataset: *winequality-red.csv*
Link para acesso: *Red Wine Quality*

2. Número de Registros e Variáveis

- Número de registros: *1599 registros*
- Número de variáveis: *12 variáveis*

3. Justificativa para a Escolha do Dataset

Este dataset foi escolhido por atender aos seguintes critérios:

- **Variáveis numéricas contínuas**: O dataset contém variáveis numéricas adequadas para a análise de regressão linear, que são essenciais para modelar relações contínuas.
- **Variáveis categóricas binárias**: O dataset também apresenta uma variável binária, *quality* (no caso iremos transformar em binária) que é crucial para aplicar técnicas de regressão logística.
- **Tamanho adequado**: O número de registros e variáveis é suficiente para treinar modelos robustos, sem ser excessivamente grande, o que facilitaria o processamento em um tempo razoável.

Além disso, a qualidade dos dados é razoavelmente boa, com poucos valores ausentes ou inconsistentes, o que facilita a preparação para análise e modelagem.

Interpretação dos resultados

1. Testes de Normalidade e seus Valores-p

- **Teste de Shapiro-Wilk**: O valor-p obtido foi **menor que 0.05** (p-value < 2.2e-16), indicando que a distribuição dos dados não segue uma distribuição normal. Isso significa que a hipótese nula de normalidade é rejeitada.
- **Teste de Kolmogorov-Smirnov (Lilliefors)**: O valor-p também foi **menor que 0.05** (p-value < 2.2e-16), o que reforça a conclusão de que os dados não seguem uma distribuição normal.

2. Interpretação dos Valores-p

- **Se p > 0.05**: Não rejeitamos a hipótese nula de normalidade. Ou seja, a variável é normalmente distribuída.

- **Se p ≤ 0.05**: Rejeitamos a hipótese nula de normalidade. A variável não é normalmente distribuída e, portanto, pode não ser adequada para técnicas que exigem normalidade, como a regressão linear.

No caso dos seus testes, como ambos os valores-p são extremamente pequenos ($p < 0.05$), podemos concluir que a variável "alcohol" não segue uma distribuição normal.

3. Adequação da Variável para Regressão Linear

A regressão linear geralmente assume que as variáveis independentes são normalmente distribuídas. Como a variável "alcohol" não apresenta normalidade, isso pode afetar a validade do modelo de regressão linear. Embora a regressão linear possa ser robusta a desvios da normalidade em algumas situações, seria prudente considerar transformações para lidar com a falta de normalidade.

4. Possíveis Transformações

Caso a normalidade não seja verificada, é comum aplicar transformações nas variáveis para aproximá-las de uma distribuição normal. Algumas opções incluem:

- **Transformação logarítmica (log)**: A transformação logarítmica pode ser útil para variáveis com assimetria positiva (valores mais altos mais dispersos).
- **Transformação de raiz quadrada (√)**: Usada quando os dados possuem uma distribuição com assimetria moderada.
- **Transformação de Box-Cox**: Esta transformação ajusta uma potência para a variável, podendo resolver tanto a falta de normalidade quanto a heterocedasticidade.

Cada uma dessas transformações pode ajudar a melhorar a adequação dos dados à regressão linear. A escolha da transformação depende da análise da distribuição dos dados após cada tentativa.

Se você desejar aplicar uma dessas transformações, posso ajudar a implementar e visualizar os resultados.

A interpretação dos resultados dos testes de correlação realizados é a seguinte:

Correlação de Pearson:

- **Coeficiente de correlação**: 0.4762
- **Força da correlação**: Moderada, pois o valor do coeficiente está entre 0.3 e 0.7.
- **Direção**: Positiva, o que significa que à medida que o valor do álcool aumenta, o valor da qualidade também tende a aumentar.
- **Significância estatística**: O valor-p é < 2.2e-16, indicando que a correlação é estatisticamente significativa, ou seja, é altamente improvável que essa correlação tenha ocorrido por acaso.
- **Implicações para o modelo de regressão**: A correlação moderada e positiva sugere que o álcool é um fator que influencia a qualidade, mas há variabilidade

suficiente para que outros fatores também sejam relevantes. O modelo de regressão pode usar essa correlação como uma base, mas outras variáveis devem ser consideradas.

Correlação de Spearman (com jitter):

- **Coeficiente de correlação**: 0.4327
- **Força da correlação**: Moderada, similar à correlação de Pearson, pois o valor do coeficiente está na mesma faixa.
- **Direção**: Positiva, semelhante à correlação de Pearson.
- **Significância estatística**: O valor-p é < 2.2e-16, mostrando que a correlação é altamente significativa.
- **Implicações para o modelo de regressão**: A correlação de Spearman sugere uma relação monotônica entre álcool e qualidade. A correlação sendo moderada indica que, embora o álcool tenha algum efeito sobre a qualidade, outros fatores podem desempenhar um papel significativo. A escolha entre Pearson e Spearman depende de se as relações entre as variáveis são lineares ou monotônicas.

Matriz de Correlação:

A matriz de correlação para múltiplas variáveis mostra a relação entre "fixed.acidity", "volatile.acidity", "citric.acid" e "alcohol". O uso de uma matriz de correlação pode ajudar a entender melhor como essas variáveis se relacionam entre si e com a qualidade, e pode ser útil para a seleção de variáveis no modelo de regressão.

Conclusão:

Tanto a correlação de Pearson quanto a de Spearman indicam uma relação moderada e positiva entre o álcool e a qualidade, com alta significância estatística. Essas informações são valiosas para entender como a variável "alcohol" pode ser usada em modelos de regressão para prever "quality", mas outros fatores também devem ser considerados para melhorar a precisão do modelo.