1. Introdução

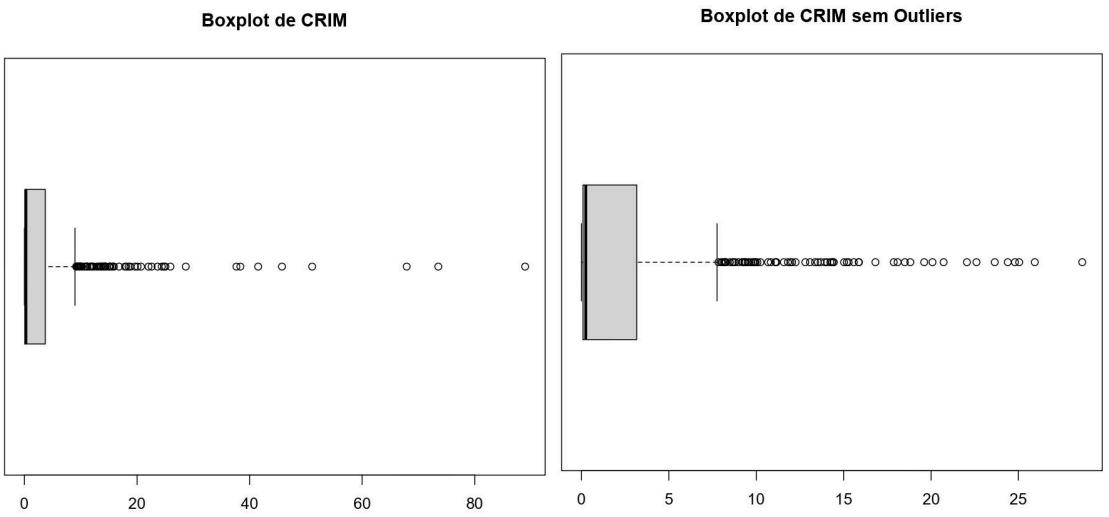
O objetivo deste relatório é analisar e predizer o valor médio da variável medv. O clássico conjunto de dados do Boston Housing Dataset é um clássico para o aprendizado de modelos de regressão, tal data possui 14 variáveis e 504 registros. Para realizar a tarefa foi realizada uma análise inicial exploratória de dados, modelagem de modelos diferentes e após avaliar os resultados dos mesmo foi escolhido um melhor modelo.

2. Exploração dos dados

Após confirmar que os dados não possuem valores nulos foi a vez de analisar resumos estatísticos das variáveis bem como suas correlações com a variável algo.

```
> summary(dados)
      crim                zn              indus            chas        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
      nox               rm              age              dis        
 Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      rad              tax           ptratio          black       
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
     lstat            medv      
 Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 6.95   1st Qu.:17.02  
 Median :11.36   Median :21.20  
 Mean   :12.65   Mean   :22.53  
 3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :37.97   Max.   :50.00  
```

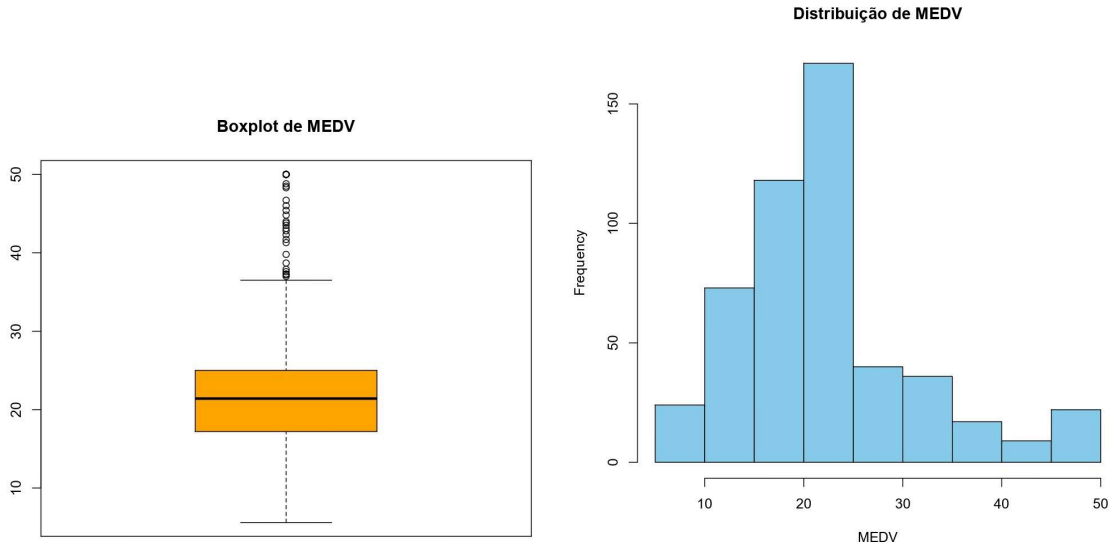
Logo de cara se notou que a variável “crim” estava muito provavelmente enviesada com outliers, os quais deveriam ser tratados. Para isso se gerou um "boxplot" e se utilizou da regra dos 3 desvios padrões para tratar os prováveis outliers devido sua simplicidade e eficácia, mesmo reconhecendo que os dados da variável em questão não seguem uma distribuição normal, ela ainda é útil em distribuições enviesadas. Assim passamos do primeiro boxplot para o segundo após o tratamento. Após o tratamento os dados ficaram com o tamanho de 498 registros, o que não foi considerado uma grande perda de informação.



E com base nos resultados da função “Summary” esta era a maior anomalia a ser resolvida, então partimos para avaliar o comportamento da variável algo, bem como sua correlação com o restante das variáveis independentes.

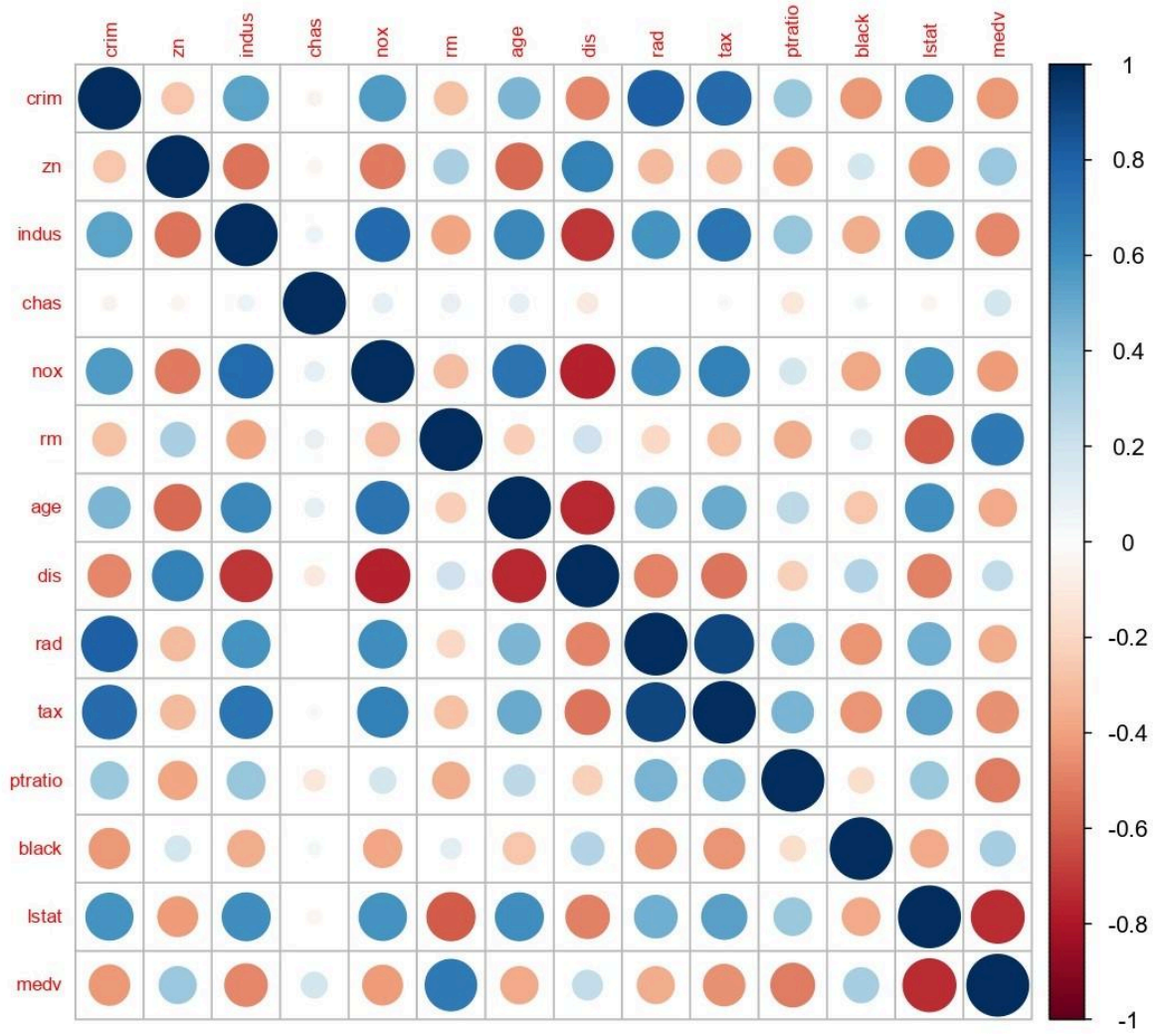
3. **Análise da Variável-Alvo**

Os valores obtidos para média, mediana e desvio padrão foram, respectivamente, 22.75 de média, 21.4 de mediana o que indica que metade das casas possui valor abaixo disso, e com uma dispersão de 9.09 pontos. Logo parti-se para analisar visualmente a variável independente obtendo o seguinte boxplot e histograma.



Onde o boxplot mesmo tendo ponto acima da linha de limite superior, não foram considerados outliers pois não ficaram maiores que 3 desvio padrão. E o histograma de medv nos mostrou que ela não segue uma distribuição normal, o que pode impactar em modelos estatísticos que assumem normalidade.

e em relação a correlação da variável com as demais, foram encontradas apenas 2 fortes sendo uma positiva (rm) e outra negativa (lstat), como pode ser visto na matriz de correlação abaixo. E que as demais variáveis apresentam baixa correlação e algumas baixíssimas, como iremos analisar mais adiante.



4. Modelagem

Como metodologia para os treinamentos, os dados foram divididos em 75% dados de treino e 25% dados de teste. E foram escolhidos 4 modelos para serem treinados, sendo eles: Regressão Linear, Árvore de Decisão, Lasso, Random Forest. E todos foram treinados usando um valor de k = 5 para cross-validation. Abaixo estão os resultados por modelo.

```
set.seed(42)
train_index = createDataPartition(dados$medv, p = 0.75, list = FALSE)
train_data = dados[train_index, ]
test_data = dados[-train_index, ]
```

1. Regressão Linear

RMSE	Rsquared	MAE
4.882465	0.7466223	3.455272

2. Árvore de Decisão

cp	RMSE	Rsquared	MAE
0.02960339	5.188192	0.7093988	3.578608
0.03049594	5.188192	0.7093988	3.578608
0.11235075	5.716108	0.6396668	3.946278
0.14082763	6.878174	0.4972571	4.936520
0.48357933	8.423148	0.3714310	6.119792

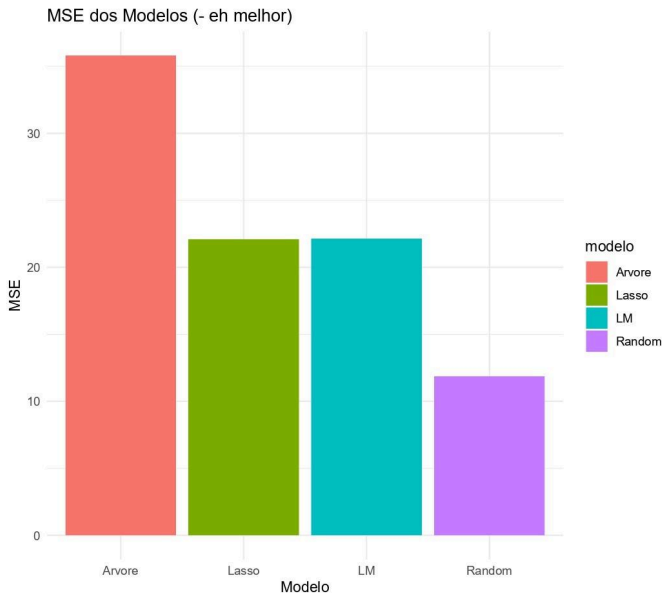
3. Lasso

lambda	RMSE	Rsquared	MAE
0.00100000	4.881019	0.7468787	3.445421

4. Random Forest

mtry	RMSE	Rsquared	MAE
2	3.715845	0.8674095	2.549227
4	3.417620	0.8788575	2.320023
7	3.395524	0.8764357	2.308280
10	3.423163	0.8737024	2.335394
13	3.484034	0.8687267	2.364482

Comparação dos MSE dos modelos



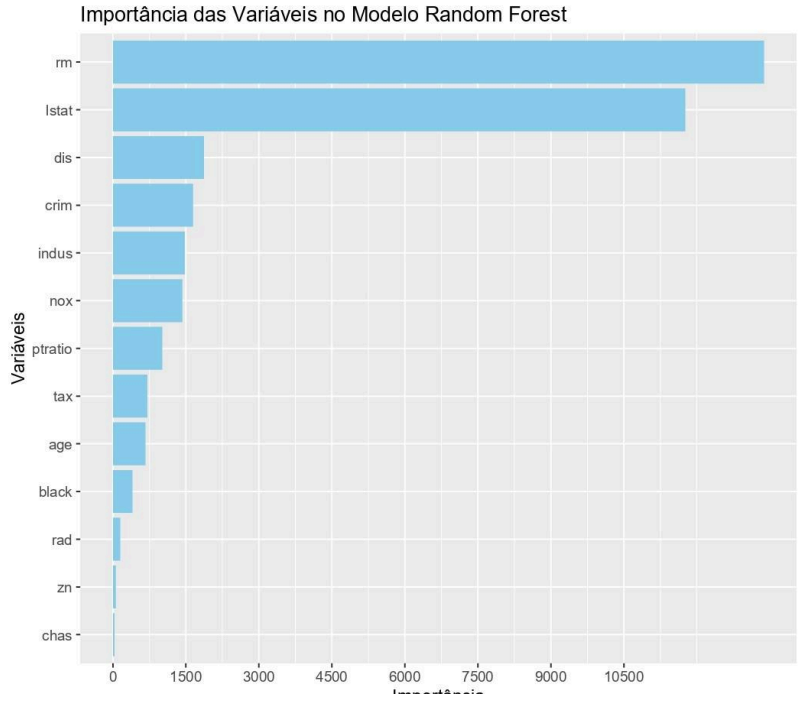
Com base no RMSE o modelo Random Forest foi o escolhido, porém para não usar como métrica isolada, será usado MSE e a correlação dos modelos, que são as métricas das quais o projeto deve se basear.

	Regressão Linear	Árvore de Decisão	Lasso	Random Forest
MSE	22.13	35.80	22.09	11.86
Correlation	0.81	0.69	0.81	0.90

E novamente o modelo Random Forest se saiu superior, obteve um MSE de 11.86, o que indica que o erro médio quadrático das previsões foi 11.86. Em termos mais intuitivos, o erro médio absoluto pode ser estimado pelo RMSE, que é aproximadamente 3.4 pontos no valor-alvo. E também apresenta uma correlação de 90% linear entre predições e valores reais, o que indica que as previsões do modelo tem uma forte associação linear com os valores reais. Com base nestes dados o modelo Random Forest foi escolhido para resolver o problema.

5. Feature Importance

Com o modelo escolhido, passamos a fazer de entender o que estava influenciando suas decisões, e o que poderia estar atrapalhando ou apenas usando recursos computacionais desnecessários. Para isso foi realizado um “feature importance” do modelo.



Onde pode-se perceber que as duas variáveis com forte correlação fazem quase todo o trabalho, seguindo de 4 variáveis que ajudam em quase 1500 pontos. E as demais estão apenas usando recursos computacionais. Claro em nosso caso como se trata de uma base de dados pequena, isto não influencia, porém pensando em problemas reais e com dados grande, as 7 variáveis com menos importância estaria usando mais recursos computacionais do que propriamente ajudando o modelo a predizer.

Feature Importance do Modelo escolhido

mtry	RMSE	Rsquared	MAE
2	3.330351	0.8820773	2.341339
3	3.408887	0.8737909	2.373182
4	3.499657	0.8655406	2.413332
5	3.519545	0.8634610	2.451967
6	3.605736	0.8561613	2.506550

```
$MSE
[1] 13.48907

$Correlation
[1] 0.8904334
```

O modelo com feature importance reduziu ligeiramente o RMSE o que indica um erro ligeiramente menor ao prever os valores, Teve uma melhora muito sutil em R^2 indicando que explica uma porção ligeiramente maior da variação, o aumento de MSE uma das principais métricas que devem ser levadas em consideração neste projeto aumentou dando indícios que este novo modelo tem mais dificuldade em lidar com casos mais complexos, e a correlação também teve um ligeiro decréscimo.

6. Conclusão

Neste projeto, o objetivo foi modelar uma solução para previsão do valor médio das casas ocupadas pelo proprietário medv. A análise seguiu um pipeline estruturado, iniciado a exploração dos dados, tratamento de outliers, análise do comportamento da variável alvo e avaliação de diferentes modelos. Durante a análise, identificamos outliers, tratados pela regra dos 3 desvios padrão, e observamos que a variável medv apresentou distribuição assimétrica, o que reforçou a necessidade de modelos robustos.

Quatro modelos foram avaliados, e o Random Forest apresentou o melhor desempenho, com MSE de 11,86 e correlação de 90%. Embora uma análise de feature importance tenha indicado que algumas variáveis têm menor relevância, o modelo foi descartado já que, apesar dos pequenos ganhos oferecidos pelo feature importance, neste caso com estes dados de pequeno tamanho, não é justificável usar este segundo modelo, já que das duas métricas principais, MSE e Correlação, ele perdeu poder. Então o modelo escolhido é o modelo Random Forest que usa todas as variáveis dos dados.