Aprendizado de máquina

INF01017 Arthur Zachow, 00277951
Prof. Mariana Recamonde Mendoza Data: 30/08/2021

Relatório de atividade prática sobre árvores de decisão

Instruções de execução

As instruções para execução do código podem ser encontradas no pacote que contém este arquivo. Todos os resultados foram gerados a partir do uso da linguagem de programação Python3 juntamente com o uso das bibliotecas *scikit learn*, *pandas*, *numpy* e *graphviz*. Os dados para os gráficos de perfomance foram obtidos através da execução de 100 vezes cada classificador com diferentes divisões dos dados.

Experimento A

Os resultados gerados nas figuras 2, 3 e 1 foram gerados a partir dos dados produzidos pelo código presente na seção do "Experimento A"do notebook que acompanha este pacote.

A comparação entre as árvores de decisão, de acordo com 2, 3, e 1, deixa claro que a estrutura de ambas as árvores é similar e que o desempenho igualmente é bem similar. O atributo mais relevante para ambas as árvores é o nível de glicose pois este é o atributo escolhido para ambas as medidas como raíz da árvore. É importante notar que mesmo que as árvores sejam similares em sua estrutura a escolha de atributos muda a partir do primeiro nível das árvores. É notável que os classificadores construídos são relativamente sensíveis à divisão dos dados.

Experimento B

Os resultados gerados nas figuras 4, 5 e 6 foram gerados a partir dos dados produzidos pelo código presente na seção do "Experimento B"do notebook que acompanha este pacote.

Os parâmetros utilizados foram 2 valores para a profundidade máxima da árvores, sendo estes 5 e 10. Estes valores foram escolhidos pois são menores que a profundidade máxima da árvore baseline 2 e por isto tornariam a árvore mais genérica em suas decisões. Pela mesma razão foram escolhidos os valores de 5 e 10 para o número mínimo de dados nas folhas.

Através da figura 4 é possível notar que houve uma melhora relevante da precisão dos classificadores com as limitações propostas. É especialmente notável que os melhores resultados foram os que limitavam a profundidade máxima da árvore em 5. Valores maiores de profundidade até geram melhora nos resultados mas esta melhora é mais tênue. Podemos observar que nestas situações os classificadores também continuam relativamente sensíveis à divisão dos dados.

Nas figuras 5 e 6, podemos notar que as diferenças entre as duas árvores se encontra mais próxima às folhas. É também possível observar na diferença das estruturas que a razão pela qual a altura das árvores parece ser um fator mais impactante na precisão, isto se deve pois a adaptação à limitação de altura preserva regras relevantes e descarta regras mais ajustadas aos dados. O impacto do número mínimo de samples nas folhas tem um impacto mais localizado na estratura da árvore, ficando mais contido localmente nos nós próximos às folhas primariamente.

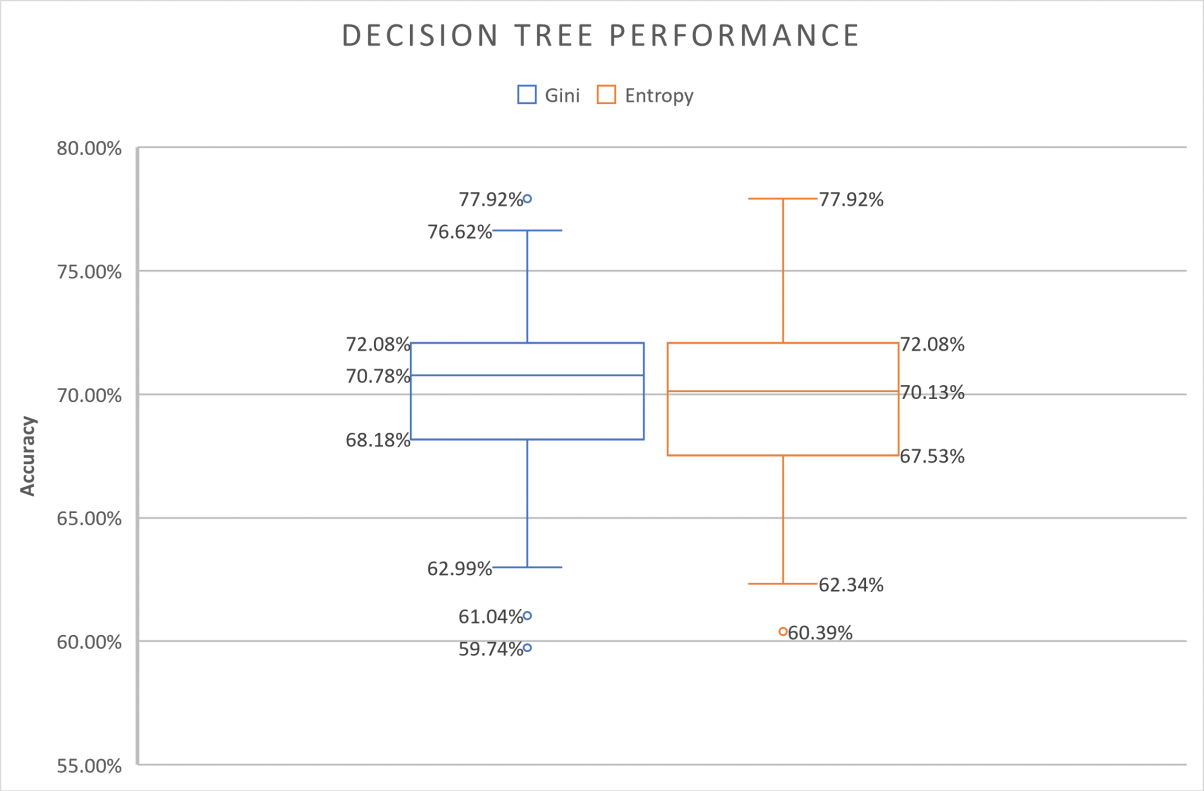

Figura 1: Comparação de performance das árvores.

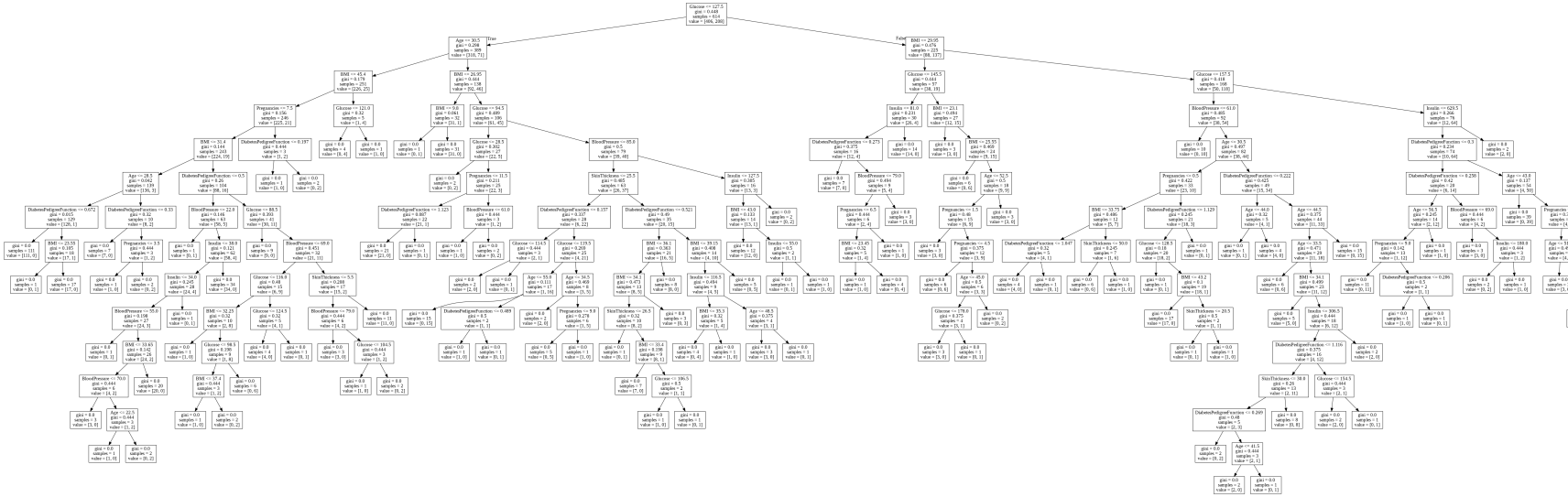

Figura 2: Árvore de decisão para a medida Gini.

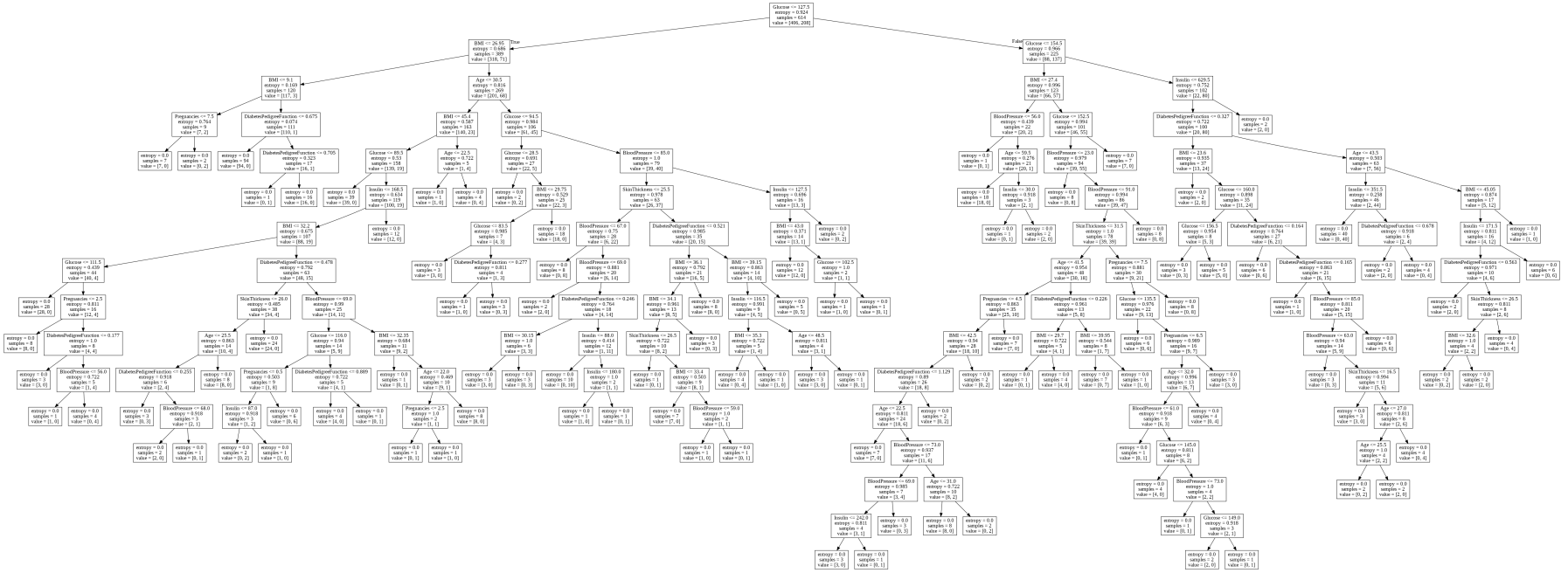

Figura 3: Árvore de decisão para a medida de entropia.

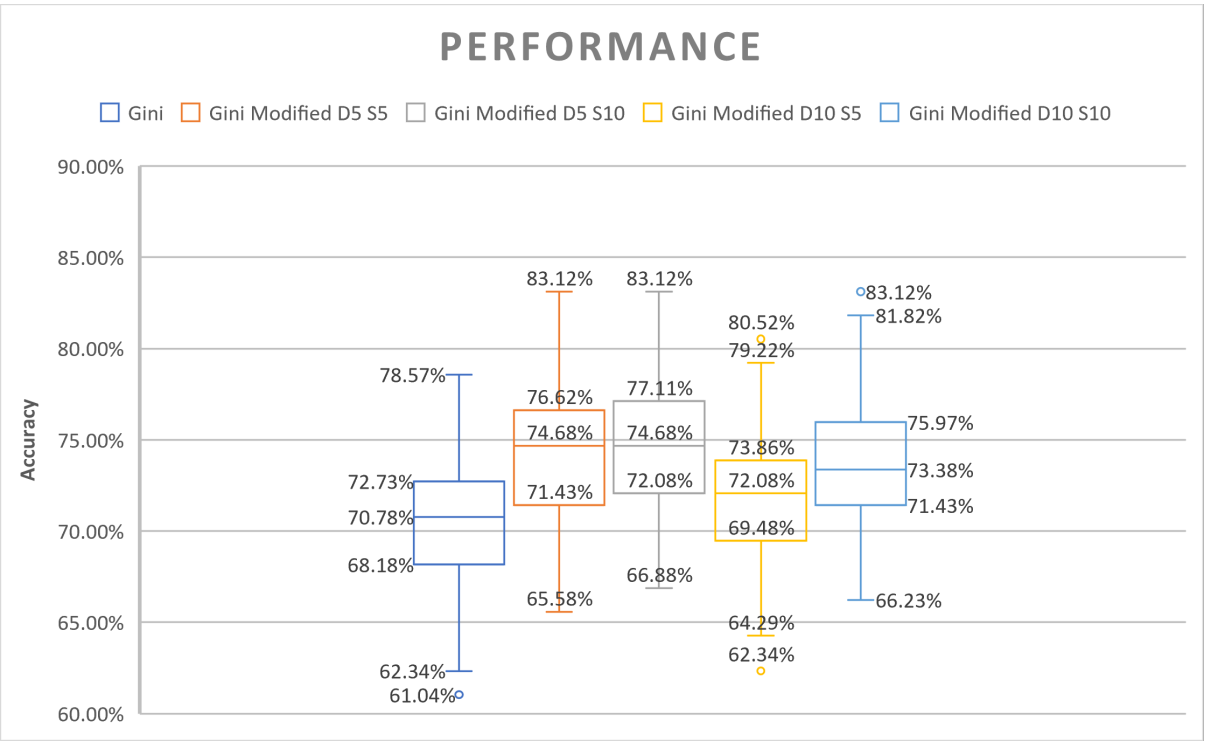

Figura 4: Comparação de performance das árvores modificadas.

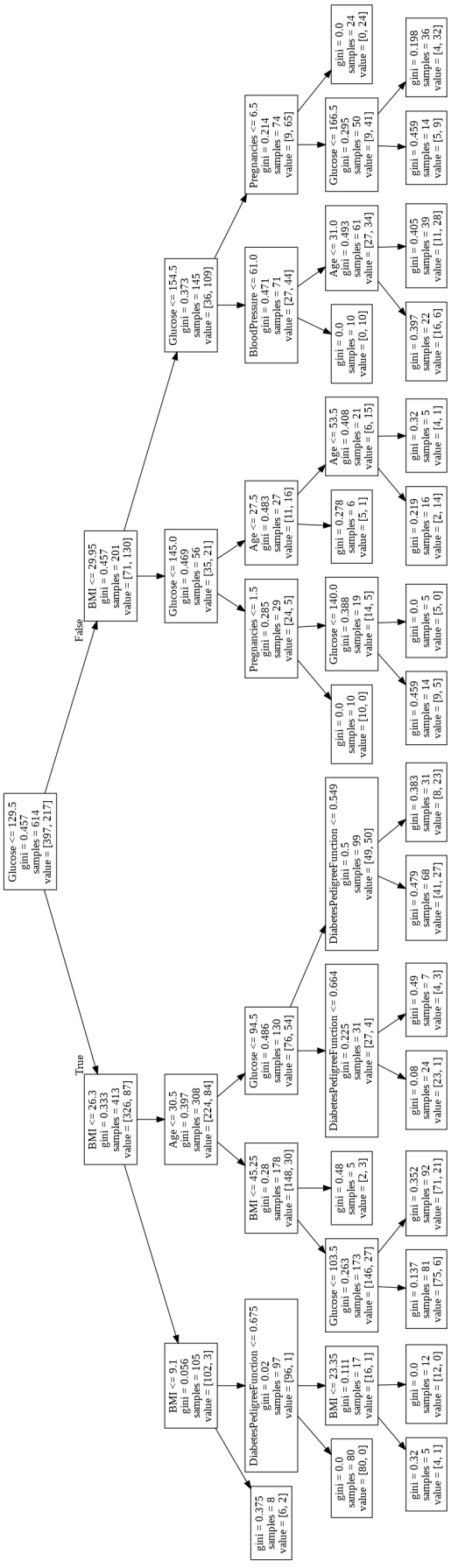

Figura 5: Árvore de decisão com profundidade máxima = 5 e número mínimo de dados em cada folha = 5.

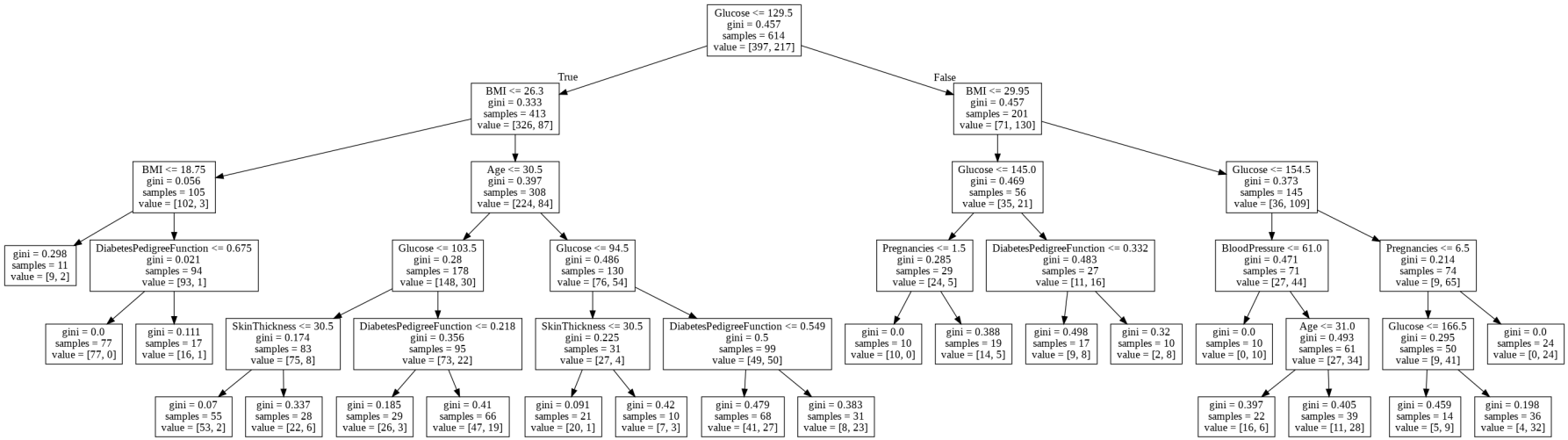

Figura 6: Árvore de decisão com profundidade máxima = 5 e número mínimo de dados em cada folha = 10.