**Florestas Aleatórias**

Melhoria de performance através das árvores de decisão



Olá aficcionados por inteligência artificial, preparados para uma expedição pela selva!?

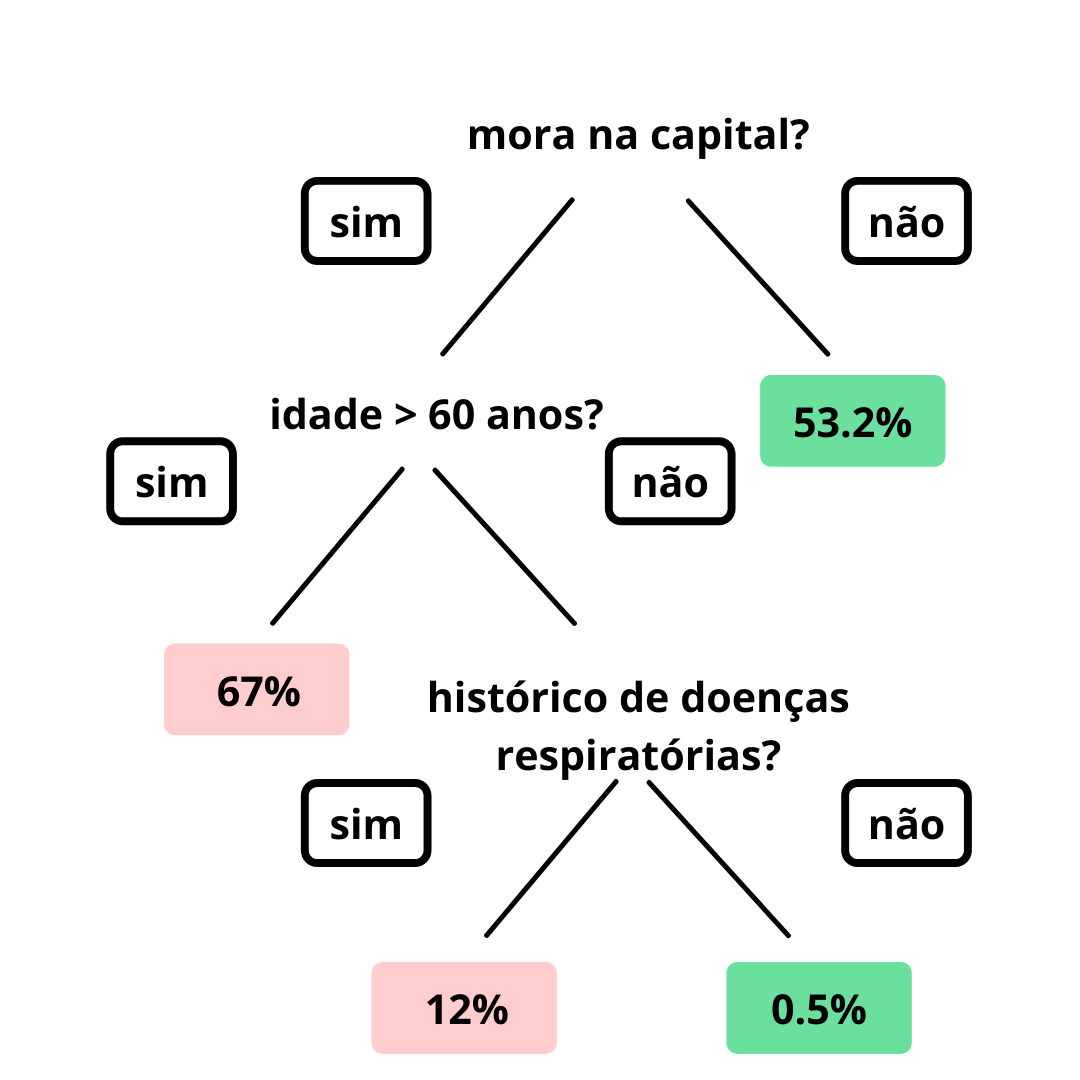
Hoje iremos falar sobre um famoso algoritmo de aprendizado de máquina: as Florestas Aleatórias. Nosso objetivo é entender sua estrutura básica, a ideia intuitiva de seu funcionamento, e como a aplicação destas pode incrementar o desempenho de nossas análises. Então, calce suas botas, pegue sua garrafa de água e vamos desbravar essa floresta!

**Árvore de decisão**

Antes de adentrar nas florestas aleatórias, nada mais justo que entender a estrutura de uma única árvore, não é mesmo!? Basicamente, a**s árvores de decisão são constituídas de uma estrutura hierárquica de condições.** Tais condições são estipuladas com base nas características dos dados e, através da verificação delas, estes são divididos por meio de condicionais “se/senão”.[1]

Para exemplificar, imagine que você tenha em mãos dados de pacientes de um hospital, dos quais irá considerar se o paciente reside na capital, se é do sexo feminino e se ele já foi acometido com alguma doença respiratória. Seu objetivo é identificar quais pacientes têm maior risco de desenvolver sintomas mais graves quando acometidos com um problema de saúde qualquer.

A estrutura da árvore de decisão para o nosso exemplo pode ser visualizada na imagem abaixo**.** Baseado nos **dados de treinamento**, se você mora na capital e tem idade superior a 60 anos, existe uma probabilidade de 0.67 de você desenvolver sintomas mais graves quando adquirir um problema de saúde qualquer. Vale ressaltar que esses dados foram gerados para entendermos a estrutura de uma árvore de decisão, ou seja, são dados sintéticos.



Árvore de decisão construída com dados sintéticos inspirados em um exemplo

encontrado em [5]. Imagem criada pelo autor no Canvas.

Para cada novo paciente, podemos submeter suas características as perguntas estipuladas pelo nosso modelo e verificar se ele atende ou não as condições, seguindo as linhas de acordo com a resposta “sim/não”. No final da árvore, nas folhas, teremos as probabilidades de pertencimento do dado a cada classe. No nosso exemplo, o vermelho representa o desenvolvimento de sintomas mais graves.

**Floresta Aleatória**

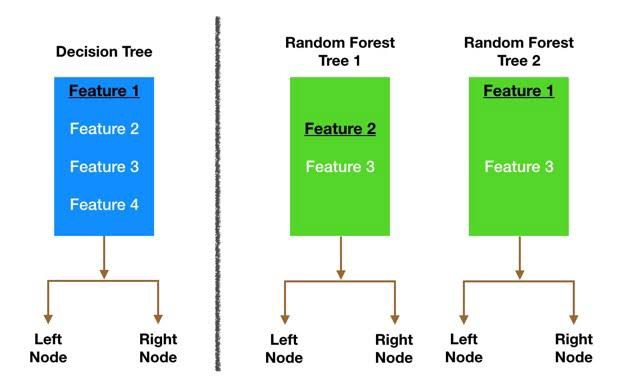
Agora que já entendemos a estrutura básica das árvores de decisão, vamos começar nossa discussão a respeito das florestas aleatórias. Basicamente **uma floresta aleatória é constituída de um conjunto de árvores de decisão**, daí o nome de “floresta”.[3]

A ideia por trás das florestas aleatórias é utilizar o conhecimento do conjunto de árvores que a constitui para melhorar o desempenho. Esse conceito é conhecido como **método de aprendizado em conjunto**, no qual **utiliza-se de múltiplos modelos de aprendizado para alcançar um desempenho melhor em relação a um único modelo.** E a pergunta que fica é: como isso aumenta o desempenho de uma única árvore de decisão?[3, 4]

Simples, a combinação de modelos de diversas árvores de decisão aumenta o escopo de conhecimento do nosso modelo no geral. É como se utilizássemos o conhecimento de diversos médicos para avaliar o desfecho clínico dos pacientes do exemplo anterior, o que certamente resultaria em uma avaliação muito mais precisa.

Voltando às árvores de decisão, outro questionamento que devemos fazer é: quais características utilizar e em qual ordem elas devem aparecer para a construção das condições que irão dividir nossos dados. Médicos diferentes podem ponderar as características dos pacientes de diversas maneiras e, através disso, chegar a conclusões um pouco diferentes. Todavia, quando considerarmos muitos médicos, esperasse que haja um consenso no diagnóstico. Então, como aplicar isso à nossa floresta? Através da aleatoriedade![3]

Ao invés de uma estrutura estática, cada árvore de decisão dentro da floresta considera **um subconjunto aleatório de características durante a formulação das condições,** técnica conhecida como bagging.[2, 4]



Exemplificação da seleção aleatória de características, extraída de [6].

A construção de diversas árvores de decisão com conjuntos de características distintas e randomizadas irá possibilitar a construção de diversos “pontos de vista” a respeito do mesmo problema e, no final, muitas árvores de decisão, assim como muitos médicos, tenderão a um consenso.[4]

**Conclusão**

Em suma, as florestas aleatórias são estruturas construídas através da combinação de diversas árvores de decisão e, seu principal objetivo é aumentar a confiabilidade dos resultados e melhorar a performance através de métodos de aprendizado em conjunto. No final, elas utilizam do conhecimento adquirido pelas árvores de decisão individualmente para construir modelos mais acurados.

Com isso, finalizamos nossa expedição por essa selva de conhecimentos. Fiquem ligados e até a próxima!

Para saber mais, acesse:

Árvores de decisão

1. Rokach L., Maimon O. (2005) Decision Trees. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X\_9

Ensemble methods

2. OPITZ, D.; MACLIN, R.. Popular Ensemble Methods: an empirical study. Journal Of Artificial Intelligence Research, [S.L.], v. 11, p. 169-198, 1 ago. 1999. AI Access Foundation. http://dx.doi.org/10.1613/jair.614.

Random Forest

3. Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

4. MÜLLER, Andreas C.; GUIDO, Sarah. **Introduction to machine learning with python**. O'Reilly.

Imagens

5.https://en.wikipedia.org/wiki/Decision\_tree

6.https://www.kdnuggets.com/2020/01/random-forest-powerful-ensemble-learning-algorithm.html