# 1 Random Forest

Grande parte do aprendizado de máquina refere-se a classificação de observações em grupos e classes. Existem diversos algoritmos para tal, como árvores de decisão, regressão logística e *random forest*.

*Random Forest* é um algoritmo de classificação que combina muitas árvores de decisão individuais, que irão trabalham em conjunto (ensemble). Cada uma dessas árvores irá retornar uma previsão de classe, e aquela com mais votos torna-se a previsão do modelo.

Para que esse modelo possa funcionar, é importante que essas árvores de decisão tenham baixa correlação entre elas, de modo que seu trabalho conjunto seja superior ao trabalho individual de cada uma delas, ou seja, produzindo previsões mais precisas que qualquer uma da previsões individuais.

Isso faz com que as árvores se protejam de seus erros individuais, enquanto algumas podem estar erradas, outras estarão certas, e assim, como um grupo, elas tem maior chance de se mover na direção correta. Nesse sentido, a chance de fazer previsões corretas aumentam com o número de árvores não correlacionadas pertencentes ao modelo.

Existem duas formas de verificar se os comportamentos de cada árvore de decisão não são correlacionados, ou têm baixa correlação entre si: *bagging (agregação de bootstrap)* e *feature randomness (característica de aleatoriedade.*

- Bagging (Agregação de Bootstrap) (Ensacamento): As árvores de decisão são muito sensíveis aos dados em que são treinadas - pequenas alterações no conjunto de treinamento podem resultar em estruturas de árvore significativamente diferentes. Isso é uma vantagem para *random forest*, pois permite que cada árvore individual faça uma amostragem aleatória do conjunto de dados com substituição, considerando sempre o mesmo tamanho de amostra e resultando em árvores diferentes, processo conhecido como *bagging*.

  Por exemplo, se temos uma amostra de tamanho N, ao invés de selecionarmos subconjuntos dessa amostra, selecionamos outras amostras de tamanho N a partir da amostra original, mas com reposição. Se os dados dados de treinamento fossem [1, 2, 3, 4, 5, 6], uma das árvores de decisão poderia receber a lista [1, 2, 2, 3, 6, 6].

- Feature Randomness (Aleatoriedade de recursos): Para dividir um nó de uma árvore de decisão normal, de todos os recursos possíveis, é considerado aquele que produz a maior separação entre as observações no nó esquerdo versus aqueles no nó direito. Cada árvore em uma floresta aleatória pode escolher apenas um subconjunto aleatório de recursos, utilizando recursos diferentes na tomada de decisão. Essa é uma forma de forçar mais variação entre as árvores no modelo, produzindo menor correlação entre as árvores e mais diversificação.

Ademais, as *features (recursos)* que selecionamos e os hiperparâmetros que escolhemos afetarão as correlações finais também.