

## Relatório Projeto 3

*Gabriel Zezze, Pedro Luiz da Costa, Rafael Libertini Argondizo*

Em nosso projeto exploramos uma base de dados com mais de 50M de arquivos vetoriais representando desenhos em 350 diferentes categorias. Frente ao desafio de criar um classificador para reconhecer desenhos dentro dessas categorias usamos alguns métodos de machine learning oferecidos pela biblioteca scikit-learn do python.

O primeiro método avaliado foi o Random Forest, um método baseado em árvores de decisão. Para entendermos como esse método funciona, precisamos entender o conceito de uma árvore de decisão.

Um bom exemplo para abstrair esse conceito é a previsão da temperatura máxima de amanhã na cidade de São Paulo, dia 15 de Janeiro. Se não soubermos em qual mês estamos, é razoável definir um escopo de estimativas de 15°C a 38°C, sabendo que essas são as mínimas e máximas anuais. Podemos fazer alguma pergunta para melhorar essa estimativa, como por exemplo, “Em que estação estamos?”. Se estamos no verão, nosso escopo diminui para 22°C a 38°C. Perguntando, “Qual a temperatura máxima histórica nos dia 15 de Janeiro” (31°C) diminuimos para 27°C a 37°C. Estamos chegando perto, porém, para fazer uma melhor previsão, podemos nos perguntar “Qual foi a temperatura máxima hoje?”, sabendo que hoje a máxima foi de 29°C podemos pensar que o ano foi um pouco mais frio que a média histórica, finalizando nossa predição dizendo que amanhã a máxima será de 30°C.

Poderíamos seguir fazendo questões ad infinitum, porém, a partir desse ponto, o retorno só irá diminuir, portanto podemos seguir com nossa estimativa atual sem grandes perdas de confiança no resultado e sem desperdiçar esforço. Resumindo o exemplo, para chegar em uma predição fizemos uma série de perguntas, cada uma reduzindo o escopo de possíveis resultados, até chegarmos em um resultado que estávamos confiantes (isso tudo sem gastar esforço desnecessário). Esse processo ainda não é a árvore de decisão, porém, nos auxiliará no seu entendimento. Podemos resumir nossa predição em um gráfico.

Fizemos nossa predição com base em um modelo mental de pergunta-e-resposta, que nada mais é que uma árvore de decisões simplificada.

As duas principais diferenças entre nosso modelo e a árvore de decisões são, primeiramente, o fato de termos negligenciado os ramos alternativos, ou seja, a predição



que teríamos chegado se a resposta às questões fosse diferente (ex: Se a resposta às estações fosse inverno.), e por último, uma característica muito diferente é que a árvore de decisões só aceita perguntas de sim e não (True ou False), portanto para a pergunta da temperatura hoje, o método perguntará todas as possíveis temperaturas, e alguma delas terá resposta True, enquanto as outras tem False.

Portanto um diagrama de árvore de decisões bem simplificado ficará parecido com a figura abaixo:

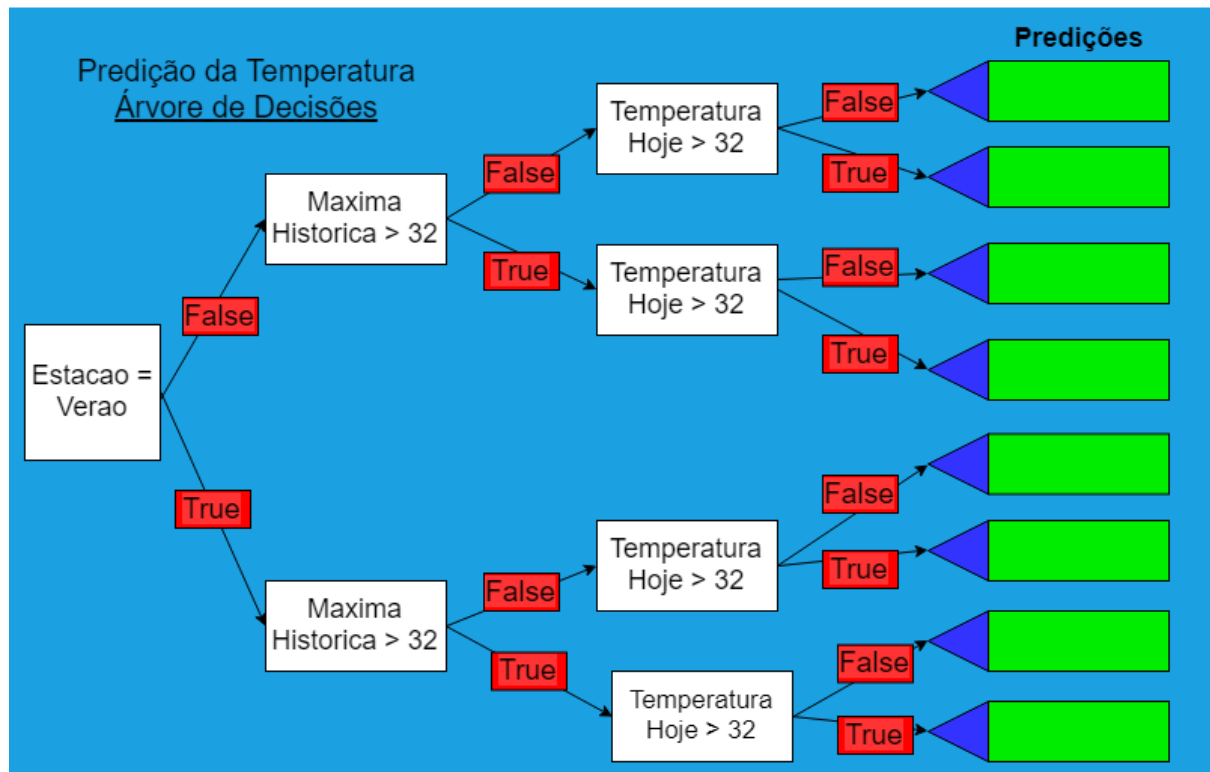


fig. 2 Árvore de Decisões (incompleta)

Esse diagrama pode parecer muito diferente do nosso modelo mental, porém ele é apenas uma expansão dele. Nosso modelo está contido nessa árvore, basta tomar o caminho mais abaixo e seguir por nossas respostas e chegamos a mesma predição.

O modelo Random Forest, em sua essência, é uma árvore de decisões, que consegue decidir as “perguntas” que irá fazer, partindo dos dados que colocamos, e assim, montar uma árvore que consiga englobar o máximo de casos para fazer uma previsão coerente.

Apesar de simples nosso modelo já serve para fazer algumas predições, porém ele não tem nenhum conhecimento prévio do assunto, assim ele não sabe que se Inverno = Mais Frio, ele tem que aprender todas as relações com os dados do problema que jogamos nele. Chegamos agora a parte de Machine Learning, onde vamos treinar nosso Random Forest para que ele possa efetivamente fazer predições. Temos que então separar um pedaço de nossos dados, geralmente 80%, para alimentar o programa. Damos ao Machine Learning todos os dados relevantes ao problema, e dizemos qual o resultado desses dados.

Seguindo nosso exemplo da temperatura, precisamos de uma base de dados com informações relevantes para a previsão de tempo (conhecidas como “Features”) e dos valores da temperatura naqueles dias (valor que nós queremos prever “target”). Assim, ele “entenderá” que fatores levam a um específico resultado. Essa é chamada a fase de treino, que deve ser feita antes de qualquer previsão.

Com nosso Random Forest treinado, podemos agora usar o que sobrou dos nossos dados como base de Teste. Utilizando os mesmos tipos de variáveis que o modelo treinou, nós damos um cenário para o modelo e ele nos retorna um valor de temperatura, baseado no que ele sabe sobre dias com essas características. Um exemplo muito simplificado é o de treinarmos com uma base que em janeiro a temperatura é sempre maior que 30 graus, portanto, se nós tentarmos prever a temperatura em um dia de janeiro, ele nos retorna uma temperatura maior que 30 graus.