Autor: José Norberto Fagundes Isaias
Matrícula: 19202785
DEC0014 - Inteligência Artificial e Computacional

MODELO DE PREVISÃO DE CHUVA COM ÁRVORE DE DECISÃO

Para este trabalho, tive de fazer diversas presunções. A primeira presunção é a de que a coluna “RainTomorrow” era uma previsão feita de acordo com os dados das outras células da mesma linha através da coluna “RISK_MM”, já que se observarmos esta, sempre que o seu valor é maior que 1,1 “RainTomorrow” é “Yes”. Outra presunção foi de que sempre iremos querer saber se irá chover no dia seguinte, por isso, usei a coluna “RainTomorrow” como *target* no treinamento do modelo. Presumi que, se quiséssemos saber se iria chover no dia atual, teríamos que saber diversos parâmetros essenciais para a previsão que possivelmente não saberíamos ainda, já que as medições ainda não teriam sido feitas. Além disso, para este caso, não poderíamos usar a coluna “Rainfall” já que, obviamente, se ela for maior 0, significa que já choveu no dia e se for 0 significa que não choveu, e então a árvore usaria somente este parâmetro para determinar se irá chover ou não no dia. Por fim, outra presunção feita, foi a de que as colunas “WindGustDir” e “WindGustSpeed” são médias da direção e velocidade do vento durante o dia.

Com isto em mente, após ler o csv com os dados meteorológicos obtidos usando a biblioteca Pandas usando um DataFrame, escolhi remover as colunas de “Date”, “Evaporation”, “Sunshine”, “WindDir9am”, “WindDir3pm”, “WindSpeed9am”, “WindSpeed3pm” e “RISK_MM”. Estas colunas foram removidas porque as julguei inúteis para o modelo, ou que iriam apenas complicá-lo demais sem muito benefício de precisão no resultado, que iriam atrapalhar na formação da árvore ou simplesmente porque estavam vazias no arquivo csv dado. Depois, fiz a codificação das colunas “Location”, “WindGustDir”, “RainToday” e “RainTomorrow” para que o classificador conseguisse entender estes dados, já que eles eram em formato de texto, passando-os para um número inteiro. Em seguida, criei um DataFrame com os dados da coluna “RainTomorrow” para o *target* do modelo para fazer o treinamento e então removi esta coluna do DataFrame que contém todos os dados meteorológicos.

Depois do tratamento dos dados, criei uma Árvore de Decisão usando o critério de entropia e a treinei com os dados tratados. Usei o critério de entropia pois observei que, entre este e o critério gini, este fornecia uma árvore mais alta e menos larga, o que a torna mais fácil de visualizar e possivelmente menos custosa computacionalmente. Então, apenas criei um gráfico com a árvore criada e fiz com que este fosse exportado no formato png. O resultado foi uma árvore enorme devido a quantidade de parâmetros observados. Contudo, percebi que, dependendo das colunas que eu usasse para o treinamento e da coluna usada como *target* a árvore ficaria simples demais, basicamente um único if-else.