Algoritmo Árvore de Decisão

O algoritmo de Árvores de Decisão gera uma estrutura de árvore que ajuda na classificação e predição das amostras desconhecidas. Com base nos registros do conjunto de treinamento, uma árvore é montada e, a partir desta árvore, pode-se classificar a amostra desconhecida sem necessariamente de testar todos os valores dos seus atributos. O algoritmo de classificação por árvores de decisão é considerado um algoritmo supervisionado, pois é necessário saber quais são as classes de cada registro do conjunto de treinamento.

A mineração de modelos de classificação em bases de dados é um processo composto por duas fases: aprendizado e teste. Na fase de aprendizado, um algoritmo classificador é aplicado sobre um conjunto de dados de treinamento. Como resultado, obtêm-se a construção do classificador propriamente dito. Tipicamente, o conjunto de treinamento corresponde a um subconjunto de observações selecionadas de maneira aleatória a partir da base de dados que se deseja analisar. Cada observação do conjunto de treinamento é caracterizada por dois tipos de atributo: o **atributo classe**, que indica a classe a qual a observação pertence; e os **atributos preditivos**, cujos valores serão analisados para que seja descoberto o modo como eles se relacionam com o atributo classe.

Para exemplificar estes conceitos, considere o conjunto de dados de treinamento apresentado abaixo. Neste exemplo, o conjunto de dados é composto por observações selecionadas a partir de uma base hipotética de informações, onde teremos um atributo classe Inglês, para saber se os atributos preditivos Tipo Escola (Pública ou Privada) e Graduação (Sim, Não), influenciam no fato de uma pessoa saber falar inglês ou não.

Nome	Tipo Escola	Graduação	Inglês
João	Publica	Sim	Sim
Paulo	Publica	Nao	Nao
Carlos	Publica	Sim	Nao
José	Privada	Sim	Sim

Maria	Privada	Sim	Sim
Carla	Privada	Sim	Sim
Cris	Privada	Nao	Sim
Janaína	Privada	Nao	Sim

É importante observar que uma árvore de decisão pode ser utilizada com duas finalidades: previsão (exemplo: descobrir se um cliente será um bom pagador em função de suas características) e descrição (fornecer informações interessantes a respeito das relações entre os atributos preditivos e o atributo classe numa base de dados).

Para analisarmos a árvore de decisão dos dados acima, utilizaremos a ferramenta Open Source Weka.