02

Instalação e um pouco sobre o explorer

Transcrição

Vamos supor que nós fazemos parte de uma empresa de *Data Science* e essa empresa foi contratada por um banco.

Esse banco quer que os clientes façam mais depósitos na conta. Então, eles submeteram os clientes a uma campanha de marketing. Com isso, conseguiram coletar dados de clientes, além da informação se esses clientes realizaram depósitos depois da campanha ou não.

Para estudar esse problema, usaremos a ferramenta chamada **Weka**. Ela é uma ferramenta bastante visual, escrita em Java, e que necessita do Java instalado na máquina para rodar.

Nesta página https://www.cs.waikato.ac.nz/ml/weka/ (http://) a Weka poderá ser instalada. A última versão para Windows no momento da gravação do curso, por exemplo, já vem com o Java para o caso do Java não estar instalado na máquina.

Mas também existem versões sem ele, tanto para o Windows, Mac e Linux.

O processo para fazer o download e instalação do programa no Windows será tranquilo, e só precisaremos num primeiro momento dar dois cliques no link na página para baixar. Com a instalação, ele aparecerá na aba de programas e poderemos abrir o Weka 3.8, versão com a qual trabalharemos.

Caso apareça alguma mensagem de erro, ela fará referência ao menu de ferramentas, em que poderemos instalar mais pacotes, mas o ignoraremos por enquanto. Teremos, portanto, a interface gráfica do Weka, e serão mostradas diversas aplicações.

Uma delas será o "*Explorer*", explorador de dados. Também haverá um experimentador de dados, o "*Knowledge Flow*", fluxo de conhecimento, a bancada de trabalho ou "*Workbench*" e a interface básica simples, com um terminal.

Nesse curso, focaremos sobretudo no "*Explorer*", então clicaremos nessa aba. Abrindo o "*Explorer*"< não conseguiremos fazer nada se não carregarmos dados. Então, precisaremos carregar a base de dados do banco com o qual trabalharemos.

Há alguns arquivos de banco na máquina, mas teremos salvo em "Weka > Data". Buscaremos por arquivos `csv`, formato em que encontraremos muitas bases de dados. Abriremos esse formato e veremos que é possível trabalhar em Weka dessa forma.

Porém, esse formato não será o mais recomendável. O padrão do Weka será o `.arff`, e ele funcionará melhor para todas as funções. Algumas funções precisarão de propriedades desse arquivo `.arff`, então trabalharemos sempre com eles.

Para observar a diferença entre os formatos, abriremos o Notepad++, semelhante ao Notepad do Windows, mas mais voltado para a programação, então abrirá arquivos de códigos por exemplo. Já teremos aberto o arquivo "bank.csv", um arquivo `.csv` normal em que a primeira linha terá os principais atributos da nossa base de dados, como idade, trabalho e estado civil.

Cada uma das linhas será um cliente, ou seja, as informações de cada cliente. Haverá o arquivo `.arff` reescrito com os nomes em português, para facilitar nosso entendimento. Ele será quase igual ao arquivo `.csv`, mas nele, as

propriedades dos clientes aparecerão um pouco abaixo, e antes delas conseguimos acrescentar algumas informações a mais.

Conseguimos colocar comentários usando o símbolo de porcentagem `%`, então conseguimos escrever, por exemplo, do que se trata a base de dados, como `%base do banco modificada`.

Com `@relation` conseguimos dar um nome para essa base de dados, `@relation Bank`. Com o termo `@attribute` conseguimos dizer o nome de cada coluna. A cada linha será nomeada uma coluna.

Por fim, colocaremos um `@data` e então começarão efetivamente nossos dados.

Poderíamos ter apenas copiado todas as linhas do arquivo `.csv`, exceto pela primeira linha, e passado para o arquivo `.arff`, para então escrever essas informações.

Depois de idade, teremos o termo `numeric`, pois idade é um atributo numérico.

Já no trabalho, teremos algumas classes definidas com os tipos de trabalhos que as pessoas podem ter. Aí sim teríamos que especificar, com "administração", "técnico", "serviços", ou seja, o tipo de trabalho, e teremos que colocar todos os tipos de trabalho que aparecerem, pois isso nos ajudará a trabalhar.

O arquivo `.arff` estará disponível para ser baixado e salvo. Com ele, conseguiremos começar a trabalhar.