Modelo Preditivo

Sistemas de Gerenciamento de Banco de Dados

Carolina Fração, Pâmela Mendonça e Thomas Pozzer

Novembro de 2019

Introdução

Este trabalho busca apresentar uma alternativa de solução ao segundo problema da disciplina de Sistemas de Gerenciamento de Banco de Dados do quarto semestre do curso de Engenharia de Software.

1 Escolha da Base de Dados

Para o desenvolvimento do trabalho, nosso primeiro passo foi a escolha do banco de dados a ser utilizado. A principio a escolha do grupo foi de um banco de dados indicado no moodle pelo professor. Porém, devido a falta de dados úteis para suprir a previsão anteriormente escolhida, chegamos ao concenso que antes da escolha dos dados devemos analisá-los melhor. Deste modo, resolvemos analizar os datasets da ferramenta Orange, foi quando nos deparamos com o banco de dados Iris. Esse banco possui dados sobre 3 tipos diferentes de flores iris, esses dados são a largura e o comprimento das sépalas e pétalas.

2 Objetivo

O objetivo da escolha da base de dados é fazer um modelo preditivo que consiga, baseado em dados de entrada, prever o tipo da flor. A escolha desse objetivo foi feita pois, durante a análise dos dados, percebemos que o banco de dados Iris possuiam pouco ruído vindo com dados simples e compreensíveis. Além disso o mesmo utiliza de todas as colunas de dados para suprir com o objetivo.

3 Etapas Executadas

Na execução de um modelo preditivo existem diversas etapas que devem ser executadas, segue abaixo as etapas que nosso grupo executou:

3.1 Técnicas de pré-processamento

Devido ao nosso dataset nao conter ruídos decidimos alterá-lo manualmente para que fosse possível a execução da etapa de pré-processamento. Deste modo deixamos algumas colunas com dados faltantes e inválidos. Para resolvermos estes problemas utilizamos da

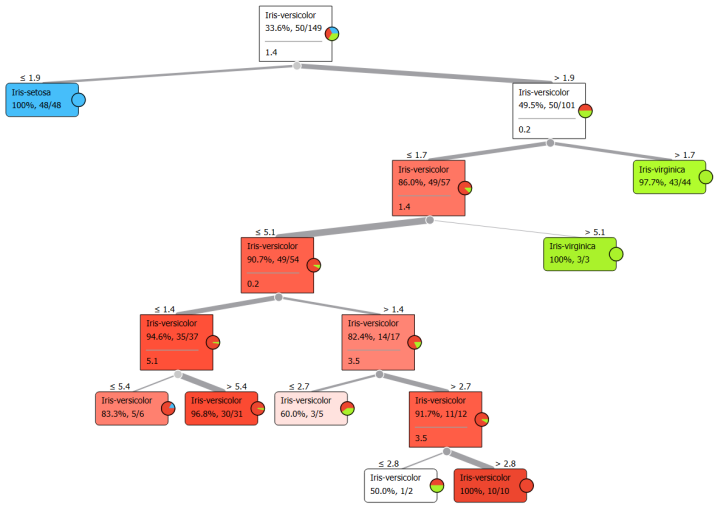
funcionalidade *Input Missing Values*, a qual calcula a média dos dados já presentes em determinada colula e atribui aos dados faltantes.

3.2 Escolha do classificador

A escolha do classificador é também um passo muito importante para construção do modelo preditivo pois é onde é feita a previsão. Então para que possamos comparar, decidimos utilizar dois tipos de classificadores para analisarmos se chegaram na mesma resposta. Segue abaixo quais classificadores foram escolhidos e suas respectivas descrições:

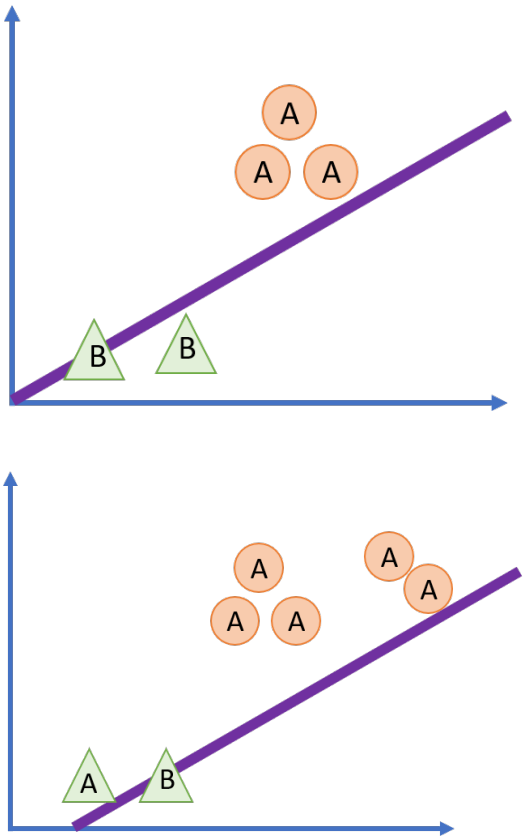
3.2.1 Árvore

Um dos classificadores escolhidos foi a árvore, a qual pode ser usada tanto para dados categóricos quanto numéricos. A árvore representa um fluxograma, em que adquire conhecimento simbólico a partir de dados de treinamento. Decidimos usar esse classificador pois é um dos métodos de aprendizagem mais práticos disponiveis, e que tem preferência por árvores menores, que é justamente gerada a partir de nosso dataset. Para os hyper paramêtros escolhidos para a árvore, colocamos o número mínimo de instâncias das folhas como 2, que a divisão de subset não fosse menor que 7,e que a profundidade máxima fosse 100. A vizualição da árvore ficou da seguinte forma:



3.2.2 Regressão Logística

Outro classificador utilizado foi a regressão logística. Esse classificador foi escolhida para auxiliar a resolver os problemas de classificação. Um dos problemas encontrados na classificação binária, é quando ocorre uma grande adição de dados de determinada classe, o que (com a representação de um grafo) deslocaria a hipótese para a direita. Esse deslocamento gera conflito, pois os itens que eram classificados como B, por exemplo, acabam sendo classificados como A.



Além disso, ela é indicada para problemas que requerem estimação das probablidades e a utilização de variáveis métricas e não métricas simultâneamente.

3.3 Resultados Obtidos

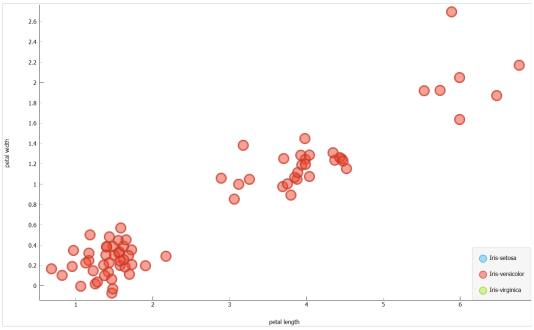
Usamos a prediction e Scatter plot para obtermos os resultados desejados, no caso, as predictions para sabermos o que os dois classificadores que usamos, para fazer previsões, obteram como resultado, e o quanto eles tem certeza desse resultado. Já o Scatter plot usamos para sabermos, baseado nas medidas das petalas, a onde se agrupam as predições dos tipos de flores. carol

3.3.1 Prediction

A prediction mostra a diferença entre o quanto a árvore e a regresssão lógica tem certeza da previsão obtida. A arvore demonstra mais certeza nas suas previsões, e inclusive, quando ele obtem como resultado o tipo de flor como iris-setosa, ele tem 100% de certeza sobre esse resultado. Já os resulta obtidos pela regressão lógica mostram uma duvida maior sobre o resultado, por exemplo, para as previsões que resultavam na iris-virginica, ela mostrava que ficava mais em dúvida entre esse resultado e outro tipo de flor.

3.3.2 Scatter Plot

O scatter plot gera uma gráfico que coloca cada resultado como um ponto numa posição baseada na altura e largura das pétalas das flores, com seus tipos previstos pela árvore, e pela regressão logística. É evidente que os pontos com mesmas cores ficam aglomerados junto com outros pontos com essa mesma características numa certa localização do gráfico, apesar de que os pontos que representam a iris-virginica ficam menos aglomerados com outros pontos do seu grupo de tipo, do que os outros grupos de pontos diferentes.



3.3.3 Matriz Confusão

A matriz de confusão gera uma matriz que nos ajuda a saber se o modelo previu bem, ou seja, se ele atribui a classe desejada, assim tendo como medir o quanto nosso modelo acerta.

Predicted

Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	48	1	0	**49**
Iris-versicolor	0	45	5	**50**
Iris-virginica	1	4	45	**50**
Σ	**49**	**50**	**50**	**149**

4 Lições aprendidas

Após o desenvolvimento do trabalho, conseguimos observar que tivemos diversas lições aprendidas. Desde coisas simples como por exemplo o quão importante é o pré processamento para que um modelo preditivo seja mais acertivo. Até coisas mais complexas, como realmente escolher algum método para a previsão de dados.

5 Problemas encontrados

Acreditamos que o maior problema encontrado pelo nosso grupo foi de realmente entender como funcionaria um modelo preditivo e o que iríamos prever, por exemplo, decidimos logo um banco de dados que íriamos usar, sem nem pensar o que iriamos predizer, foi quando percebemos que era melhor primeiro analisarmos o conteúdo que dava os dados, assim podendo os analisar melhor e saber o que predizer, foi quando trocamos o banco e decidimos por um dataset do orange(iris) onde poderiamos treinar e executar todas etapas do trabalho.

6 Conclusão

Diante do problema apresentado pudemos notar que o objetivo principal é de expandir nossos conhecimentos através do desenvolvimento deste trabalho, sendo então possível dizer que com certeza o objetivo foi alcançado, assim pudemos garantir que conseguimos compreender melhor como funciona um modelo preditivo e o quão importante é essa área no futuro e que cada vez mais será utilizado. Logo, conclui-se que o modelo preditivo faz o que é desejado.