UNIVERSIDADE DO ESTADO DO AMAZONAS

Engenharia de Computação

VICTOR BRASIL DE PINA

RELATÓRIO DO PROJETO DE MACHINE LEARNING

Manaus - AM

2019

RESUMO

Neste relatório apresentarei os resultados do projeto de machine learning que possibilita a máquina classificar uma espécie de iris baseada em seus parâmetros. Explicarei o tipo do conjunto de dados, o pré-processamento necessário desses dados e o método e o modelo escolhidos. Logo após apresento os resultados obtivos.

METODOLOGIA

O primeiro passo foi identificar o conjunto de dados. Analisando-o, nota-se a presença de atributos alvos, que seriam a classificação da espécie de cada objeto (flor) presente no dataset, então o conjunto de dados é do tipo preditivo de classificação.

Após classificar meu conjunto, eu pude começar o pré-processamento de dados. Meu pré-processamento consistiu em verificar se não havia nenhum dado nulo e se havia alguma classe majoritária. Confirmado que não havia nehum dos dois eu pude separar os atributos e as classificações em dois arrays diferentes e substituir o nome das espécies por números para que o processamento seja mais leve.

Utilizando o método de árvore de decisão para e classificar os dados, usei o modelo de validação cruzada, ou “K-fold”, que consiste em dividir o dataset em K subconjuntos iguais e trinar o programa em K-1 subconjuntos e prever o subconjunto restante. Esse processo é repetido K vezes utilizando outro subconjunto para predição, dessa forma todo o conjunto é testado. O desempenho final, então é a predição de todo o conjunto de dados, dando assim uma predição mais confiável.

RESULTADO

Os resultados variam pois os subconjuntos formados do modelo de K-fold são formados de forma aleatória. Porém há uma consistência de acurácia de 93% à 95%. Isso significa que o programa conseguiu prever corretamente até 95% de todo o conjunto de dados.

O programa cria um gráfico da matriz de confusão demonstrando as predições falsas e verdadeiras, como pode analisar na imagem abaixo:

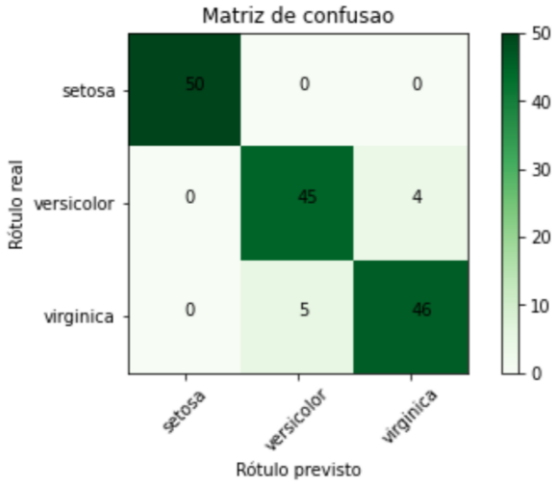

Imagem 1

De acordo com o gráfico 100% das setosas foram previstas corretamente. As outras duas, versicolor e virginica, tiveram resultados parecidos, onde 4 das 50 versicolors foram previstas como virginicas e 5 virginicas foram previstas como versicolor. Isso acontece pelos atributos das duas espécies serem similares, portanto a maquina erra cerca de 5% à 7% das predições dos objetos.

CONCLUSÃO

Apesar das semelhanças entre as versicolors e as virginicas, nota-se que o programa funciona com 100% de acurácia em identificar setosas e entre 93% e 95% no geral o tornando confiável na predição de espécies de iris.

BIBLIOGRAFIA

- FACELLI, Katti; LORENA, Ana Carolina
  Inteligência Artificial: uma abordagem de aprendizado de máquina. 2011.