PCS5024 Atividade 2

Previsão de receita de um norte-americano baseado na base adult

Aluno: Filipe Assis Mourão

Número Usp : 8988914

Neste segundo exercício foi pedido para se analisar a base de dados “Adult Census Income” que possui 15 colunas, 14 sendo features como idade, nível de escolaridade e estado civil, além de uma coluna dizendo se o descrito cidadão americano ganhava mais ou menos de 50 mil dólares anualmente.

A base de dados está amplamente disponível no site de competições kaggle, onde também é possível discutir possíveis soluções para o problema. As melhores acuracidades para este problema variam entre 84% e 88% , utilizando algoritmos para redução de dimensionalidade como pca e algoritmos como random forest. Para este exercício nos foi solicitado que utilizássemos apenas os algoritmos k-nearest neighbors (KNN) e Naive Bayes(NB).

Inicialmente notou-se que a base possuía dados faltantes em 3 das 14 colunas, eram estas “workclass” que aparecia em cerca de 5.64% das linhas, “occupation” que aparecia em cerca de 5.66% das linhas e “native.country” que aparecia em cerca de 1.79% das linhas. Para simplificar o problema foi decidido que todas as linhas que possuíam dados faltantes seriam excluídas do problema, nisso o número total de 32561 linhas se reduziu para 30162 uma perda de 7.4% dos dados disponíveis.

Após a remoção dos dados a coluna “income” foi convertida em uma coluna com dados binários em que 1 significaria que o cidadão descrito pela linha ganha mais de 50 mil dólares e 0 que ele não ganha. Após isso, foi notado que 7 colunas possuíam valores discretos, dessa forma, foi necessário decompor essas colunas em colunas com indicadores, através da função get_dummies() da biblioteca pandas. Em seguida, foi feita uma normalização de todas as colunas entre os valores 0 e 1, utilizando a função preprocessing.MinMaxScaler() da biblioteca sklearn, para que a ordem de grandeza dos valores não influenciasse na construção do nosso modelo classificador.

Após a decomposição e normalização foi feita um mapa de calor com a correlação entre cada uma das colunas e a coluna “income” com o objetivo de escolher as colunas com maior relevância para criar o modelo classificador. Um mapa de calor contendo as 10 variáveis mais significativas é apresentado no verso desta folha.

Foram então finalmente criados diferentes modelos utilizando um diferente número de variáveis significativas (3,5,10 e 20), os dados foram divididos em 80% para treinamento e 20% para teste e randomizados usando a seed 42. Os melhores resultados obtidos tanto para o NB quanto o KNN estão apresentados no verso desta folha, com a ressalva que para o algoritmo KNN também foram variados o número de vizinhos próximos para se fazer uma classificação.

O melhor resultado obtido para estes dois algoritmos foi utilizando o KNN com 30 vizinhos e 20 variáveis mais significativas, foi-se obtida uma acuracidade de 83.5% o que está bem próximo dos melhores resultados obtidos pela competição no Kagle. Infelizmente não foi encontrado dados sobre a precisão, recall e f1-score para comparação.

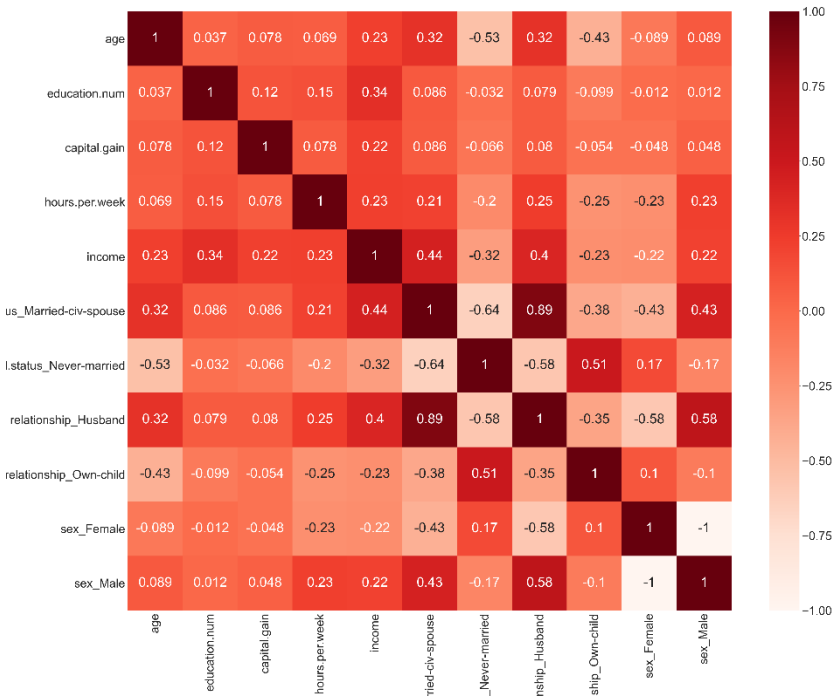

Imagem 1: Mapa de calor com a correlação das 10 variáveis mais relevantes

number of relevant variables	number of neighbors	accuracy	precision	recall	f1-score
20	30	0,834576496	0,71826087	0,550666667	0,623396226
20	50	0,830764131	0,704177323	0,550666667	0,618032174
8	30	0,830929886	0,708695652	0,543333333	0,61509434
8	20	0,830598376	0,708551483	0,541333333	0,613756614
20	20	0,829769601	0,706190061	0,54	0,6120136
10	30	0,828940825	0,703478261	0,539333333	0,610566038
10	20	0,82877507	0,703930131	0,537333333	0,609451796

Imagem 2: Resultados obtidos com o algoritmo KNN variando alguns parâmetros

number of relevant variables	accuracy	precision	recall	f1-score
20	0,812199569	0,600437876	0,731333333	0,659452961
8	0,765456655	0,51866491	0,787333333	0,625364046
10	0,734957732	0,480442513	0,810666667	0,603324237
5	0,723520636	0,467161845	0,796666667	0,588960079
3	0,734626222	0,477802198	0,724666667	0,57589404

Imagem 3: Resultados obtidos com o algoritmo NB variando o número de variáveis principais