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).

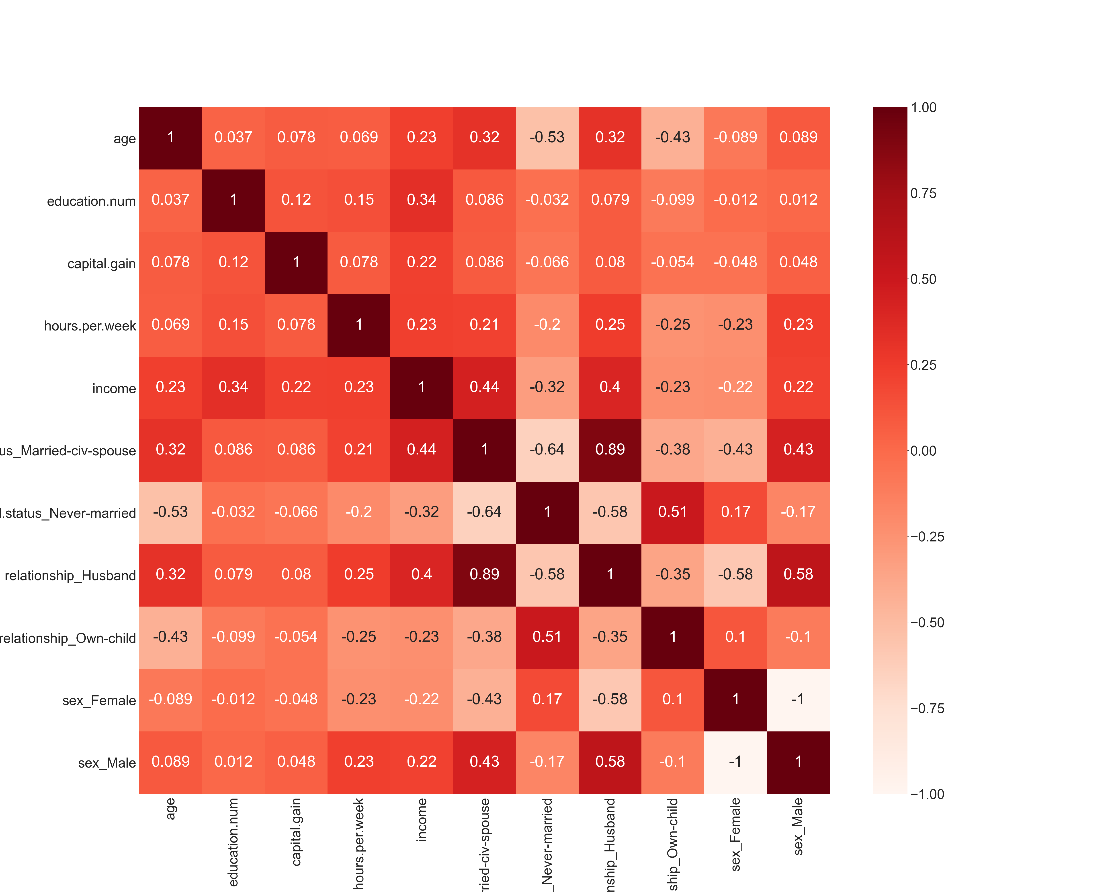
Inicialmente notou-se que a base possuía dados faltantes em 3 das 14 colunas, eram estas “workclass” que aparecia em cerca de 5.64% das linhas, “occupation” que aparecia em cerca de 5.66% das linhas e “native.country” que aparecia em cerca de 1.79% das linhas. Para simplificar o problema foi decidido 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.

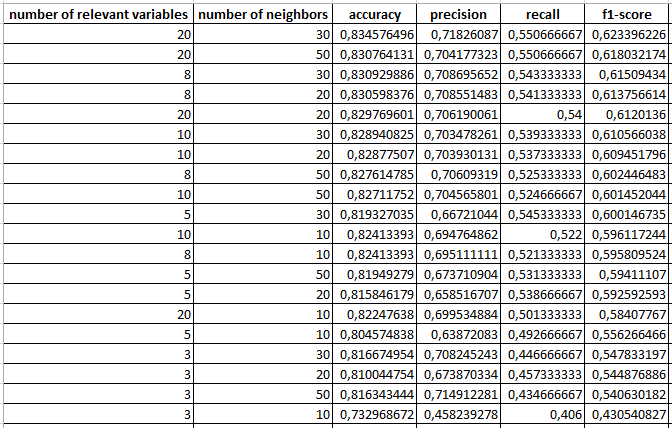
Após a decomposição e normalização foi feita um mapa de calor com a correlação entre cada uma das colunas e a coluna “income” com o objetivo de escolher as colunas com maior relevância para criar o modelo classificador. Um mapa de calor contendo as 10 variáveis mais significativas é apresentado no verso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

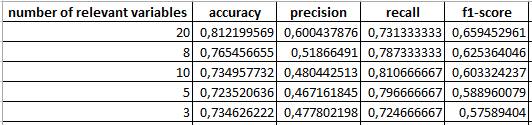
O melhor resultado obtido para estes dois algoritmos foi utilizando o KNN com 30 vizinhos e 20 variáveis mais significativas, foi-se obtida uma acuracidade de 83.5% o que está bem próximo dos melhores resultados obtidos pela competição no Kagle. Infelizmente não foi encontrado dados sobre a precisão,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



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



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



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