INFO526 Programming Assignment 1
Meng Wu

Tools: Python
Codes: https://github.com/wum5/INFO526_codes/tree/master/hw1

Methods:
I implemented all the four algorithms (i.e. Logistic Regression, Naïve Bayes, Decision tree and Nearest Neighbor) on the adult data. In comparison with the number of examples in the entire set, the number of examples with missing data is relatively small. Thus, I decided to remove all the entries with missing values. For Nearest Neighbor, I scaled down all the continuous features between 0 and 1, because some insignificant variable with larger range would be dominating the objective function in this classifier. I also converted all categorical variables to numerical variables (e.g. Male=1; Female=0), because these categorical features have string values cannot be fed into Logistic Regression. For each algorithm, I used the 10-fold cross-validation on the training examples (split into training set and tuning set) to better evaluate the model's performance. For Logistic Regression, I combined the cross-validation with recursive feature elimination to find optimal number of features used for subsequent learning and prediction. This step might be important because removing irrelevant features could enhance generalization by reducing overfitting. With the cross-validation, I could also search for the optimal depth of the Decision Tree and the value of *k* in Nearest Neighbors. This step is important because 1) larger *k* produces bias while smaller *k* produce variance; 2) higher depth produces variance while lower depth produce bias. I used both accuracy and ROC curve to evaluate the performance of predictions. In this case, ROC was more appropriate and informative due to the skewed data structure.

Table 1. Summary of prediction accuracy using four different algorithms

Algorithms	Naïve Bayes	Nearest Neighbor	Decision Tree	Logistic Regression
Accuracy	78.90%	80.60%	84.90%	82.00%

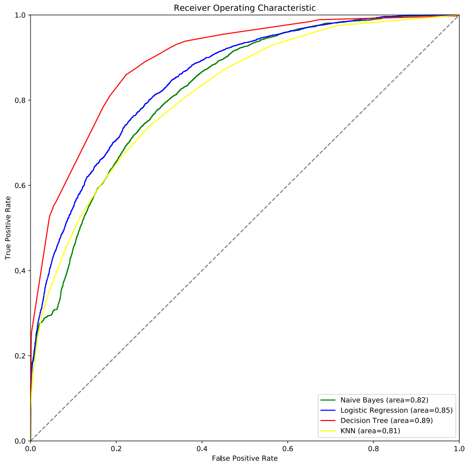

Figure 1. ROC Curve of four different algorithms

Results and Discussion:
For each algorithm, I just showed the outputs coming from the model with the best performance after tuning the hyper-parameters. From both the accuracy score and ROC curve, all the four algorithms showed descent prediction performance (accuracy >78% and ROC area >0.8) and Decision Tree outperformed the other three algorithms. I hypothesized that the Decision Tree gave a better prediction here is due to the reason that it is more flexible, robust to outliers and noisy data, insensitive to Monotone transformation of features and irrelevant inputs.