Writeup for RTI CDS Analytics Exercise 01

Lucas McCabe

All initial data exploration, analysis, and processing in a Jupyter Notebook (Exploration_and_Processing.ipynb) to keep track of and explain any observations and process. After cleaning and processing the data, there were over 100 features. These features and their correlations with one another are shown in Figure 1 (figure of choice to fulfill Step 8). This figure informed my choice of features moving forward. The 24 selected features were the ones most strongly-correlated with whether an individual earned greater than $50,000. We can summarize the topics of these 24 features with the following general descriptors: age, education, hours worked, net capital, occupation, marital status, family relationship, and sex.

Two models were attempted. The first was a decision tree classifier, which resembles a flow chart in form and function, and the model achieved 81.75% accuracy in testing. The second was a simple multilayer perceptron, which is a type of artificial neural network, and the model achieved 84.19% accuracy in testing. Certainly, the multilayer perceptron model outperformed the decision tree classifier, but training took almost twice as long. Because the dataset is not extraordinarily large, the difference in training time is not prohibitive, so the multilayer perceptron model appears to be the best choice. We have shown that we can reliably predict with well above 80% accuracy whether an individual in the provided dataset earned more than $50,000.

	Decision Tree Model	Multilayer Perceptron Model
F1-Score	0.5995	0.6362
Accuracy	81.75%	84.19%

Table 1: Evaluation metrics of the two final models.

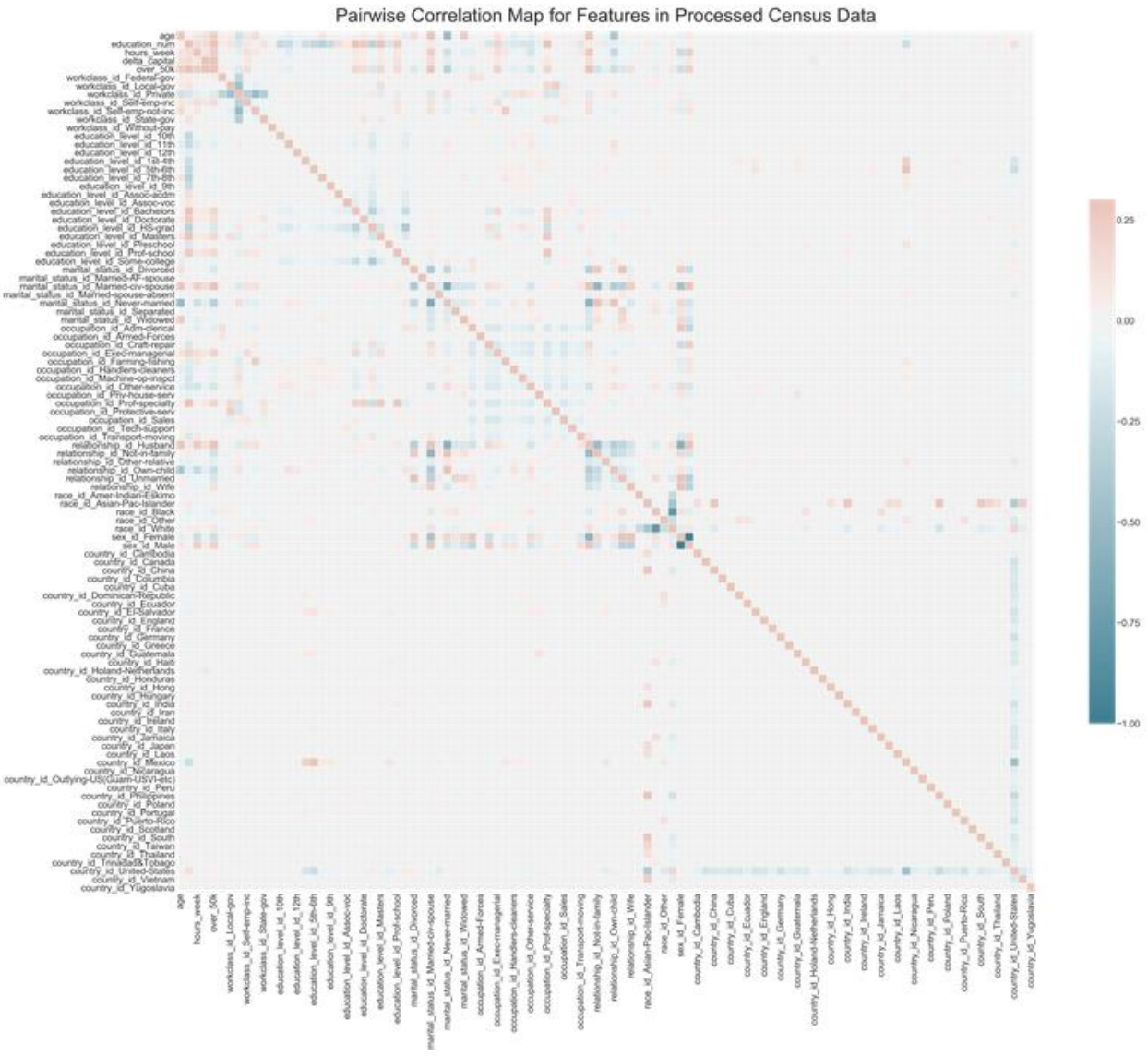

Figure 1: Pairwise correlation map for features in processed census data. A separate figure for pairwise correlation with the target variable over_50k only was also producd, but does not display well in Microsoft Word. It is available in Figures/correlation_figure_over_50k.jpg. A version of that figure that only includes the 24 features used in the models is available at Figures/consolidated_correlation_figure_over_50k.jpg.

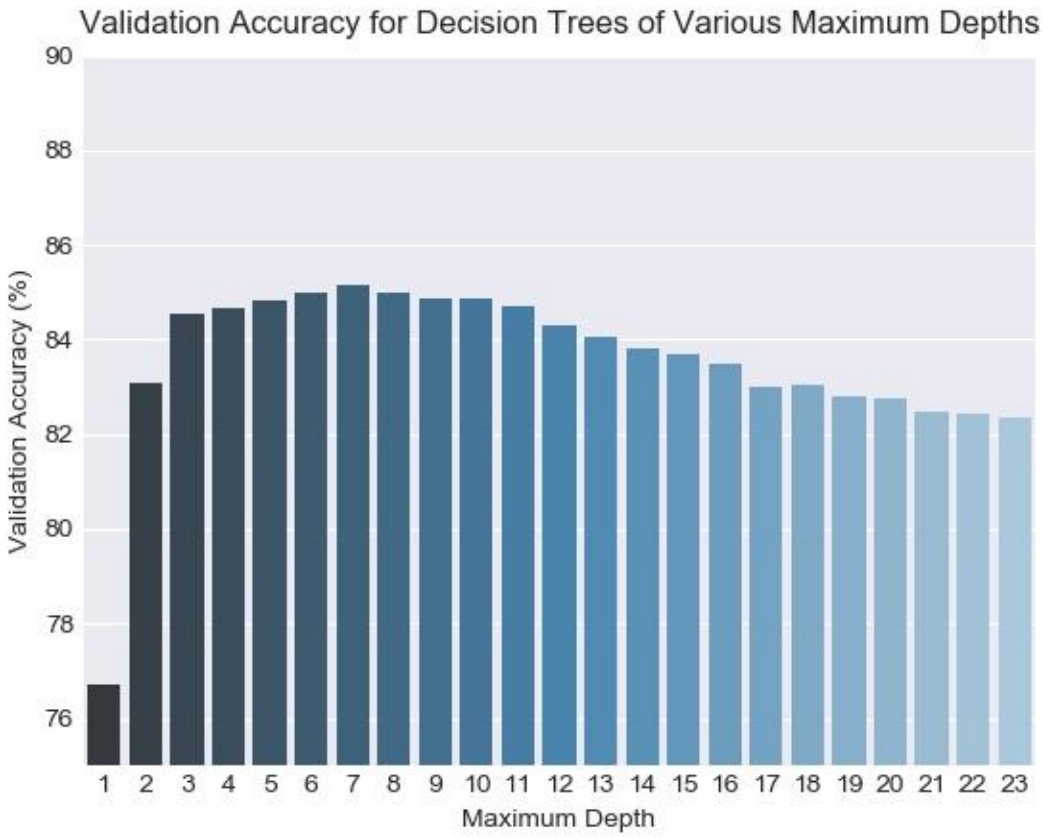

Figure 2: Validation accuracy for decision trees of various maximum depths. A maximum depth of seven was ultimately selected.