Empirical Question: Can we use past Education Census data to predict whether someone will make a yearly income greater than $50,000? We will use Naive Bayes, Logistic Regression, Decision Tree, k-nearest neighbor, support vector machine, random forest and adaboost machine learning classifiers to help answer this question.

Preprocessing the Data

- Renamed column having numerical value for number of years of Education to 'Education 1'
- Encoded categorical features/discretized continuous text features (age, Education1, Capital-gain, etc) from column data frames as numbers.
- Deleted cells from table with non-numerical values such as ? or n/a
- Scaled the features with a mean of 0 and variance of 1 using Standard Scaler in sci-kit library.
- The data was preprocessed using the same method for all machine learning classifiers.
- A cross validation of 10 was used for all models.
- The data was explored visually to see which feature was correlated to predicting whether Income would be greater than $50,000 US. The variable 'Education' was selected as the feature to be evaluated by each of the models to predict the dependent variable 'Income.'

Hyperparameter Tuning

1. All classifiers had a cross validation of 10 and training size of 0.80 unless otherwise noted. These values were chosen based on standards in the machine learning literature.
2. Accuracy was used because it is the most commonly used evaluation metric for classification problems and is the number of correct predictions made as a ratio of all predictions made.

Why does changing c affect test data set?

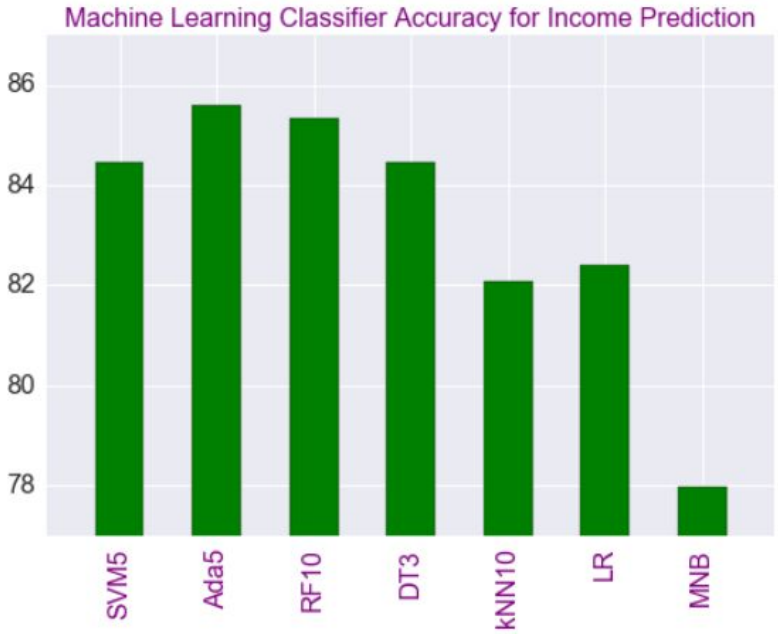
The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. For very tiny values of C, you should get misclassified examples, often even if your training data is linearly separable.

Why does changing tree depth affect test data set?

More trees is always better with diminishing returns. Deeper trees are almost always better subject to requiring more trees for similar performance. The above two points are directly a result of the bias-variance tradeoff. Deeper trees reduces the bias; more trees reduces the variance. The most important hyper-parameter is how many features to test for each split. The more useless features there are, the more features you should try.

ML Classifier Performance	Accuracy
Support Vector Machine, c = 1 Support Vector Machine, c = 5 Support Vector Machine, c = 10	0.8446 0.8447 0.8414
Adaboost w/ tree depth = 1 Adaboost w/ tree depth = 5	0.8506 0.8561
Random Forest w/ tree depth = 1 Random Forest w/ tree depth = 10 Random Forest w/ tree depth = 15	0.7657 0.8535 0.8504
k-Nearest neighbor, k = 3 k-Nearest neighbor, k = 10 k-Nearest neighbor, k = 20	0.8005 0.8207 0.8205
Decision tree, depth = 3, leaf size = 5, cv =10 Decision tree, depth = 3, leaf size = 5, cv = 5	0.8446 0.8442
Logistic regression	0.8239
Multinomial Naive Bayes	0.7796

The top 7 most accurate classifiers were plotted visually, changing the y-axis and labels to be aesthetically pleasing. The y-axis was changed from 77 to 87 to show the slight variations in the accuracy data and was shown as a percent. The axis scale was changed to this range since the data itself ranged from 77.96 to 85.61.



Why did the models perform differently? Compare and contrast the classifiers.
In our case, the Adaboost, Random Forest and Support Vector Machine were the top three performing classifiers in terms of accuracy of whether a person with a certain amount of education would have an income greater than $50,000. Other research suggests Adaboost is also one of the best out-of-the-box machine learning classifiers. Adaboost is the most accurate because

the model is run multiple times on reweighted training data. According to research by Fernandez-Delgado et al., in general random forest classifiers performed best from 179 classifiers on the entire UCI machine learning educational data set. Our random forest model was also very accurate since it's not overly sensitive to the specific hyper-parameters in the data set used. Support Vector machine was also accurate because we selected a variable (Education) that impacted the resulting income within the relatively small data set. Since the instances of data were linearly separable (Income > 50K or Income <50K), the SVM method had a high accuracy.

Empirical Economic Aid Question

In contrast with the preprocessed income data set, we wanted to answer a different question about economic aid data. *Can machine learning be used to predict whether using historical USAID economic aid data could predict future values*? To our knowledge, no research combining machine learning with economic/development aid forecasting has been conducted as of April 2017. We chose to use USAID data since USAID is the largest US government entity providing foreign aid in the largest number of countries.

Preprocessing the Data

We took the USAID by country freely available CSV file (https://explorer.usaid.gov/data.html) and processed the data to only include economic aid disbursement amounts for 2014-2016 for each of 176 countries. Other attributes were removed. We decided to use actual aid amounts in our classifier and converted the data to a matrix. We initially thought about including World Bank Gross Domestic Product (GDP) annual % growth in our models, but decided against this. We removed null values and had a cleaned data set of 153 rows x 4 columns. The '2016' values were used as the training data and the '2015' values were used as test data with an 80/20 size split.

Classifier Performance

After further data exploration, we noticed the aid amounts varied substantially from year to year. Twenty percent of the countries (36) that received economic aid in 2015 received 0 economic aid in 2016. Eighty-one percent (143) countries had aid amounts decrease by more than 15% in 2016. According to the World Bank's calculations, aid remittance to developing countries decreased by 2% cumulatively in 2016. A similar trend could have happened at USAID in that same time period we studied. A domain expert would needed for possible explanations.

We performed linear and logistic regression classifiers and measured the accuracies as we did for the Income data set described above. Neither of the models performed well. In fact, they both had a 0% accuracy so we did not plot them visually. Even when we changed the training data to '2015' and the test data to '2014', we had similar poor performing results. Since the actual aid amounts performed so poorly, perhaps future research might be to see if there is a correlation between the aid amount and a different dependent 'y' variable such as GDP.