**Julia Machine Learning Analysis**

**Car Evaluation Data Set**

The car evaluation data set consists of mostly categorical data including features about the car as well as a consumer rating which is used as the class. After loading the data the only preprocessing that was applied was encoding of the categorical labels and splitting the data into testing and training partitions. This was done using a 70 - 30 train-test-split. The models trained included: CategoricalNB, DecisionTreeClassifier, SVC, and MLPClassifier. All were left with default settings other than MLPClassifier where the max\_iterations was tuned up to 1000 in order to train a more accurate model. Benching showed that DecisionTreeClassifier had the highest accuracy and CategoricalNB had the smallest model size and fastest training time.

**Abalone Data Set**

The abalone data set includes features which specify attributes of the abalone mollusk. The rings attribute is used as the class. Most of the data was already numeric (and continuous) so only the sex feature was encoded. In order to tune the models to be more accurate, an extra step was added to the preprocessing phase. The rings class was refactored as age using an algorithm which calculates the mollusk age based on the number of rings. Then the data was split into testing and training partitions using a 70 - 30 train-test-split. The models trained included: GaussianNB, DecisionTreeClassifier, SVC, and MLPClassifier. All were left with default settings other than MLPClassifier where the max\_iterations was tuned up to 1000 in order to train a more accurate model. Benching showed that MLPClassifier had the highest accuracy, GaussianNB had the smallest model size, and DecisionTreeClassifier had the fastest training time.

**Madelon Data Set**

The madelon data set includes artificial data. The features are grouped into 32 clusters and labeled randomly, making the problem highly nonlinear. The dataset includes 500 features which left alone could lead to overfitting. To prevent this, SelectKBest was used in the preprocessing phase to remove 480 irrelevant columns and get more accurate predictions. hen the data was split into testing and training partitions using a 70 - 30 train-test-split. The models trained included: GaussianNB, DecisionTreeClassifier, SVC, and MLPClassifier. All were left with default settings other than MLPClassifier where the max\_iterations was tuned up to 1000 in order to train a more accurate model. Benching showed that SVC had the highest accuracy, GaussianNB had the smallest model size, and DecisionTreeClassifier had the fastest training time.

**KDD Cup 1999 Data Set**

The KDD data set was used during the International Conference on Knowledge Discovery and Data Mining to build predictive models for distinguishing between “bad” and “good” connections. After loading the data set, the categorical tcp, http, and SF features were encoded into a numeric format. Then the data was split into testing and training partitions using a 70 - 30 train-test-split. The models trained included: GaussianNB, DecisionTreeClassifier, SVC, and MLPClassifier. All were left with default settings other than MLPClassifier where the max\_iterations was tuned up to 1000 in order to train a more accurate model. Benching showed that DecisionTreeClassifier had the highest accuracy, GaussianNB had the smallest model size, and MLPClassifier had the fastest training time.