The MNIST (Modified National Institute of Standards and Technology) dataset contains 70,000 small images of handwritten digits. Each digit image is labeled with the digit it represents. The dataset is used for machine learning classification algorithms, and this assignment will evaluate the classification of the digits and the impact of Principal Component Analysis (PCA) on the data.

The dataset was downloaded, and the data variables and target were saved to X and y objects, respectively. The X object contains 70,000 handwritten digits with 784 dimensions or features (28 x 28 images). The y object is the true value of the handwritten digit. The data was split so there were 60,000 observations in the training set and 10,000 observations in the test set. The X data was processed to convert the data type, reshape the data, and normalize it by dividing by 255. The y data was processed to convert the data type.

A Random Forest Classifier was prepared using the full set of 784 explanatory variables and the training set of 60,000 observations. The time to fit the model and evaluate with the test data set was recorded. Classification performance was assessed with the F1 score, which is a weighted average of the precision and recall. The F1 score can range between 0 and 1, and a higher value is preferred.

Principal component analysis was used to reduce dimensionality and determine the proportion of the variance explained by each feature. In this instance, PCA was conducted on the full set of 70,000 digits and identified 154 principal components that represent 95 percent of the variability in the explanatory variables. The training and test sets for the X variables were transformed with the reduced dimensions. The time to conduct this analysis was recorded. A graph representing the PCA cumulative summary was prepared and visualizes the effect of PCA reduced dimensionality and explained variance.

The reduced training and test data sets were used to build another Random Forest Classifier. The time to fit the model and evaluate with the test data set was recorded, the classification performance was assessed with the F1 score, and the information was compared to the first Random

Forest Classification model. Even though PCA reduced the complexity of the MNIST data set by reducing the number of components from 784 to 154, the total time needed to fit a Random Forest Classifier and evaluate the data increased. The reduced complexity of the data set is making it harder for the Random Forest Classifier to generate trees. The Random Forest Classifier generates a split by looking for a feature and threshold that results in the purest subsets. The criteria to evaluate this split is the Gini index, which is a criterion to minimize the probability of misclassification. The training data with the reduced components contains less information for the Random Forest Classifier to optimally operate, therefore it takes more time. The F1 score for the Random Forest Reduced model is less than the Random Forest model, but this is expected as we fit the PCA analysis to explain only 95% of the variance.

Given this information, the experiment was repeated utilizing a Logistic Regression classifier, first with the training and test data sets containing the full set of 784 explanatory variables and then with the training and test data sets containing the reduced 154 explanatory variables from the PCA. The time to fit the models and evaluate with the test data set was recorded, the classification performance was assessed with the F1 score, and the information was compared to both Random Forest Classification models.

Using the Logistic Regression Classifier with reduced complexity training data obtained from PCA reduced the calculation time over 50% and resulted in almost the same F1 score (0.923 vice 0.926). The Logistic Regression Classifier has a lower F1 score compared to Random Forest Classification that used the full set of 784 explanatory variables. The performance of the Random Forest Classifiers would be higher if the number of n_estimators was increased, as only 10 trees were used. The decision to use one classification algorithm verses another will depend on whether time or accuracy is more important. The decision to use PCA will depend on the type of algorithm used and not all algorithms benefit from a dataset with reduced dimensionality.