The United States Forest Service (USFS) administers the forest and grasslands of the United States, which spans more than 193 million acres. One data point the USFS regularly maintains is the type of forest cover of a particular parcel of land. One way to classify the forest cover type of a particular parcel would be to visit every parcel of land, and manually determine what kind of cover type the forest maintains. However, this is not feasible considering the enormous span of the USFS responsibilities.

While visiting every possible parcel of land is not possible, the USFS believes that thousands of man hours could be saved if there was a machine learning solution to this classification problem. Even though the USFS has not manually determined the forest cover type for these parcels of land, extensive algorithms and satellite technologies have been developed by the USGS for cartographic features. These features include elevation, distance to water, soil type, and distance to roadways. The data mining question is can supervised machine learning be applied to classify forest cover types by strictly using cartographical features?

The data set has been supplied by Kaggle and consists of 15.1k rows, 55 predicting features, and 1 target feature, being the cover type. The rows in this instance are 30x30 meter parcels that were human classified by humans in a study of four wilderness areas located in the Roosevelt National Forest of northern Colorado. The 55 different predicting features are really grouped into nine different categories: elevation, aspect, slope, distance to hydrology, distance to roadways, hillshade (index for sun at certain times of the day), distance to fire points, wilderness areas (four wilderness areas), and the soil type (40 different binary soil type features). The response variable is the forest cover type, and there are seven different cover types.

Some initial strategies for this multi-class classification problem would be to create an x, y component of the parcel from the aspect, which could provide some insight into slopes facing particular ways. Because this is a fairly large training set, k-fold cross validation may take too long, therefore I will control the model using the leave one out method to evaluate the model's performance. I will also ensure that the training and test sets are representative using a stratified sampling strategy.

This multiclass-classification problem will be looked at using multiple learning algorithms. First, I will look at both hierarchical and k-means clustering to see if the features can be used in a semi-supervised classification. I will also look into using the decision tree algorithm for a transparent solution. I will also look into the black-box models of the naïve bayes and random forest supervised learning algorithms.

The best model chosen will be based on the accuracy of the model on the test set. The test set will be 20% of the dataset.