I’m planning to use a project online from Kaggle.com: <https://www.kaggle.com/c/forest-cover-type-prediction>

I’m planning to use logistic regression and LDA for this project. If possible, I’ll try to use random forest, which I need to learn.

In this competition you are asked to predict the forest cover type (the predominant kind of tree cover) from strictly cartographic variables (as opposed to remotely sensed data). Each observation is a 30m x 30m patch. You are asked to predict an integer classification for the forest cover type. The seven types are:

1 - Spruce/Fir

2 - Lodgepole Pine

3 - Ponderosa Pine

4 - Cottonwood/Willow

5 - Aspen

6 - Douglas-fir

7 - Krummholz

The training set (15120 observations) contains both features and the Cover\_Type.

Data Fields

Elevation - Elevation in meters

Aspect - Aspect in degrees azimuth

Slope - Slope in degrees

Horizontal\_Distance\_To\_Hydrology - Horz Dist to nearest surface water features

Vertical\_Distance\_To\_Hydrology - Vert Dist to nearest surface water features

Horizontal\_Distance\_To\_Roadways - Horz Dist to nearest roadway

Hillshade\_9am (0 to 255 index) - Hillshade index at 9am, summer solstice

Hillshade\_Noon (0 to 255 index) - Hillshade index at noon, summer solstice

Hillshade\_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice

Horizontal\_Distance\_To\_Fire\_Points - Horz Dist to nearest wildfire ignition points

Wilderness\_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation

Soil\_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation

Cover\_Type (7 types, integers 1 to 7) - Forest Cover Type designation