Machine Learning Project report **Jiyan Jonsdotter**
ML **Massimo Innocentini**
Dr. Norman Hendrich
TA: Philipp Ruppel Due Date: 30/06/2018

Introduction

The dataset that we decided to analyse is from the Forest Cover Type competition. The goal of the competition is to predict the type of forest, given a set of features, hence It is a multi-class classification problem. Kaggle provides a training set of data with the forest cover type, in total there are 15120 observations, and a test data set with only the observed features with 565892 samples. There are 7 type of forest cover used in the competition, each assigned an integer number:

1. `Spruce/Fir`
2. `Lodgepole Pine`
3. `Ponderos Pine`
4. `Cottonwood/Willow`
5. `Aspen`
6. `Douglas-fir`
7. `Krummholz`

There are 13 type of features which describe a 30m x 30m area in each sample, the features are:

- `Elevation`:
- `Aspect`:
- `Slope`:
- `Horizontal_Distance_To_Hydrology`:
- `Vertical_Distance_To_Hydrology`:
- `Horizontal_Distance_to_Roadways`:
- `Hillshade_9am`:
- `Hillshade_Noon`:
- `Hillshade_3pm`:
- `Horizontal_Distance_To_Fire_Points`:
- `Wilderness_Area`:
- `Soil_Type`:
- `Cover_Type` :

Few of those terms are clear like elevation, however there some of them need further explanation. The `Slope` and `Aspect` identify respectively the land inclination and the direction of the inclination in degrees. The hydrology distance reports the distance from the closest water source. The hill shade is a grayscale representation value of the illumination of the surface, which takes into account the position of the sun at different times. The values returned range from 0 to 255.

`Wilderness area` is divided into 4 groups: Rawah , Neota, Comanche Peak and Cache la Poudre. For simplicity in the dataset each is assigned an integer number. As the name suggests wilderness area are reservations which are untouched by humans in order to prevent natural conditions and wildlife. Finally the `Soil type` is also divided into subgroups, the data identifies 40 of them. Every feature is of integer type.

Data Analysis

The initial analysis performed on the data showed that few of the feature resemble a normal distribution, like `Hillshade_3pm` or the `Slope` distribution, plus the variance for those features is also closer to 1. After computing the mean of each column we noticed that there is definitely an imbalance among the values of the columns. For instance elevation seems to be measured in meters, while the slope if the land is measured in degrees, in this case one has a mean of 2700 while the other of 16 accordingly. We believe this will cause bias during the training phase and the data should be normalised.

Another imbalance noticed in the features is that even though there are 40 types of soil type, type 10 occurrence is way higher than the others, moreover there seems to be redundant information while plotting the forest cover type with the type of soil. The redundancy was noticed with the wilderness areas too. Finally some feature like `Hillshade_Noon` seems to be constant among all the samples hence not really representing anything useful.

Interesting Features

We tried plotting most features against the forest cover type in order to view if there was correlation. There were few that showed potential influence on the resulting forest as can be seen from plots Fig 1, 2, 3, 4. The plot in Fig 4 shows how the forest type changes based on the distance from water. The plot merge both the vertical and horizontal distance from water and each colour in the plot indicated a kind of forest cover. The idea was taken from the Kernel of skillsmuggler [1], where the author plot both distances from the water but uses elevation as a colour. From the plot It does look like it can be an interesting feature.

The other two features which seemed interested during the exploration of the data, were the soil type and maybe the wilderness area. In particular during the analysis we noticed that a certain kind of forest type was linked to few soil types, like Fig 5 and 6 shows. However It is necessary to mention as said before that we had redundancy information and similar soil type were also found with the same forest cover type.

Data Processing Findings

We tried to run a simple clustering algorithm using kmeans with k= [3,4,5,6,7,8,9,10] however the result were not promising, the silhouette score mean was around 0. That means that the clusters found were not actually indicating representative features.

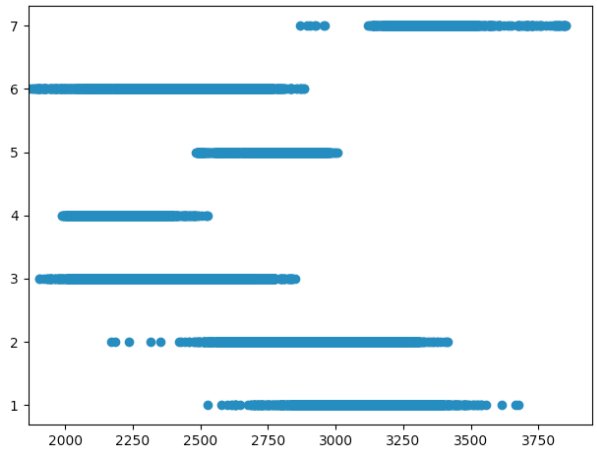

Figure 1: Elevation / forest cover type

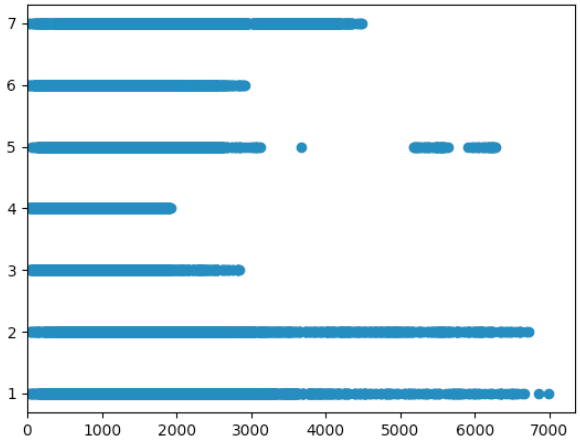

Figure 2: Fire points distance / forest cover type

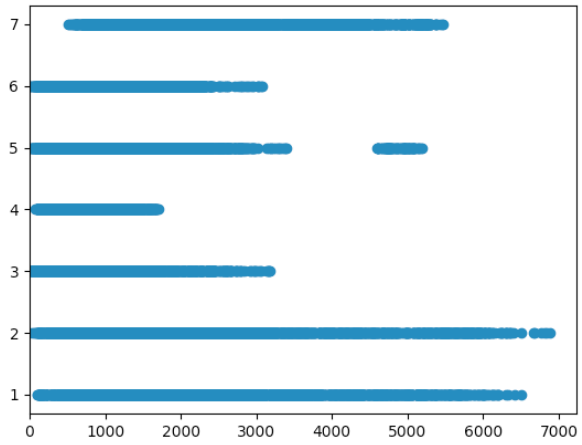

Figure 3: Roads distance / forest cover type

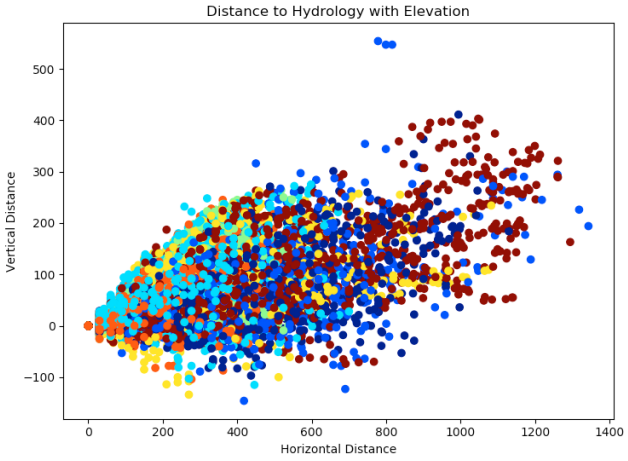

Figure 4: Water distance / forest cover type

Moreover we applied PCA just to check the results are shown in Fig 7, we followed the sklearn example to plot the result according to the difference classes we had in our dataset. The result of the components are also not helpful since there is not a single component which held a high variance. Which means there does not seems to be a subgroup of feature which overwhelm the others. Anyway the dataset provided does not have too many features so the the training of the model should still be efficient. The only extra feature which should be reduce are the soil types and wilderness area since now they are split in multiple columns. One way is to either use a range to split to merge them or at least compress them into a single column like [1] does. The former seems to be more plausible and the least likely to lose important features.

Finally an interesting feature of the data is the wilderness area, which at the moment we don't know how to consider with the respect to the future model. As we described in the introduction, those areas are uncontaminated by human, hence in a way their different from areas not protected because there humans can affect the forest in addition natural characteristics of the land. For instance as you can see from Fig 8, when the zone is considered a wilderness area of type 4 when there is a forest cover of either type 3 or 4 with strong probability given the occurrences.

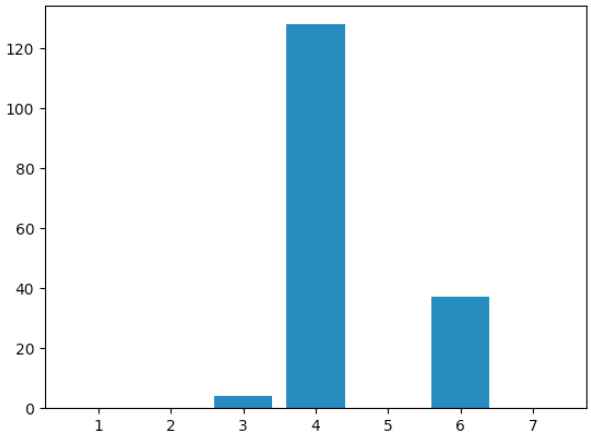

Figure 5: Soil type 14 / forest cover type

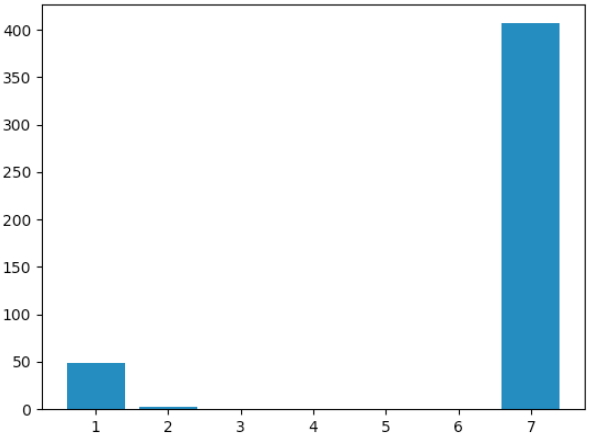

Figure 6: Soil type 40 / forest cover type

Conclusion

In conclusion we believe that the features which will play an important role are the distances from: fire, water and roads, the elevation of the area will also quite likely have an impact on the type of forest to be found as we have seen from Fig 1. The other features seems to be less representative, except for the soil type which we will need to asses during training. Since this is a multi-class classification problem, we will probably try to use the basic K-nearest-neighbors at the beginning to see the results, then move on to SVM and if we have enough time, implement a deep neural network which should perform better given the nature of the problem.

References

[1] skillsmuggler. https://www.kaggle.com/skillsmuggler/eda-and-dimension-reduction

[2] Scikit-learn tutorials. http://scikit-learn.org/stable/tutorial/index.html

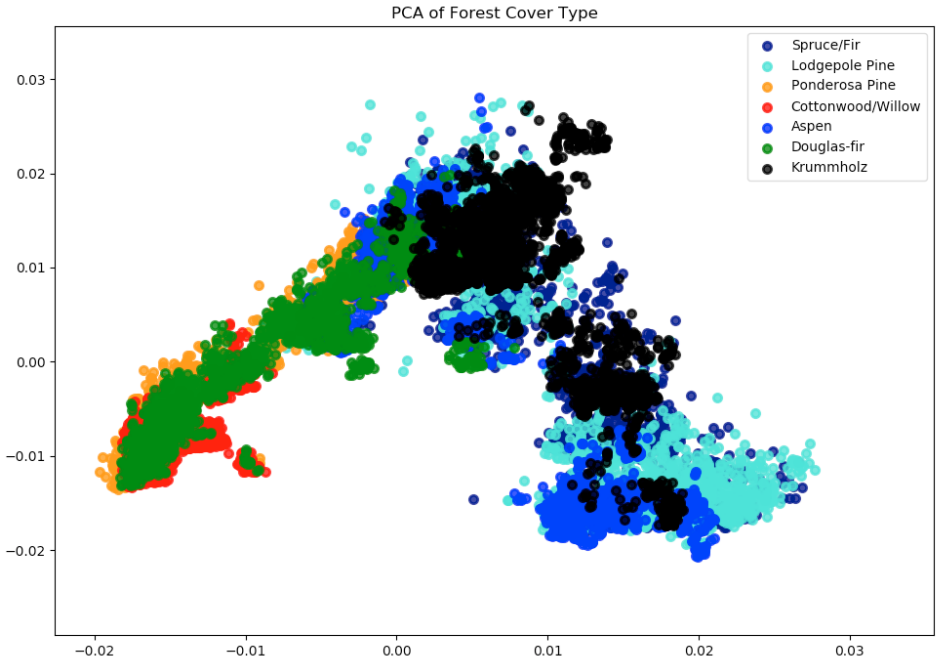

Figure 7: PCA applied on the dataset

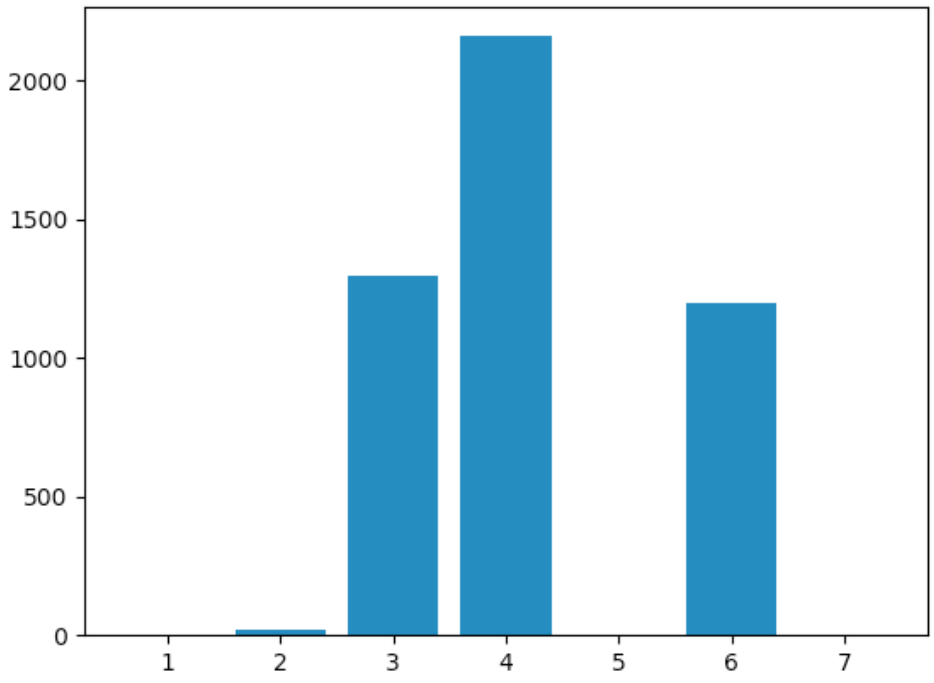

Figure 8: Wilderness area 4 in respect to forest cover type.