ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Python for Data Analysis

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Statlog (Landsat Satellite) Data Set

Data Exploration

Exploring the Data Set & Clean Up

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Data Set Information

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Database

The database consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the classification with the central pixel in each neighbourhood

Data Set Characteristics

- Number of instances : 6435
- Number of features : 36
- Missing Values : N/A
- Associated Tasks : Classification

Target Value Classes :

- 1 : Red Soil
- 2 : Cotton Crop
- 3 : Grey Soil
- 4 : Damp Grey Soil
- 5 : Soil with Vegetation Stubble
- 6 : Mixture Class
- 7 : Very Damp Grey Soil

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Data Visualization

Find links between features and target variable

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Data Visualization

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

✓ Spectrum Visualization

We use spectrum visualization to plot the number the cardinality of the different pixel values. This visualization let us saw that the train and test set had a very similar repartition of the classes.

✓ Bar Plot of classes

With Bar Plot, we use matplotlib library to plot a graph representing the category of data with rectangular bars with lengths and heights that is proportional to the number of individuals representing a specific label.

✓ Principal Component Analysis (PCA)

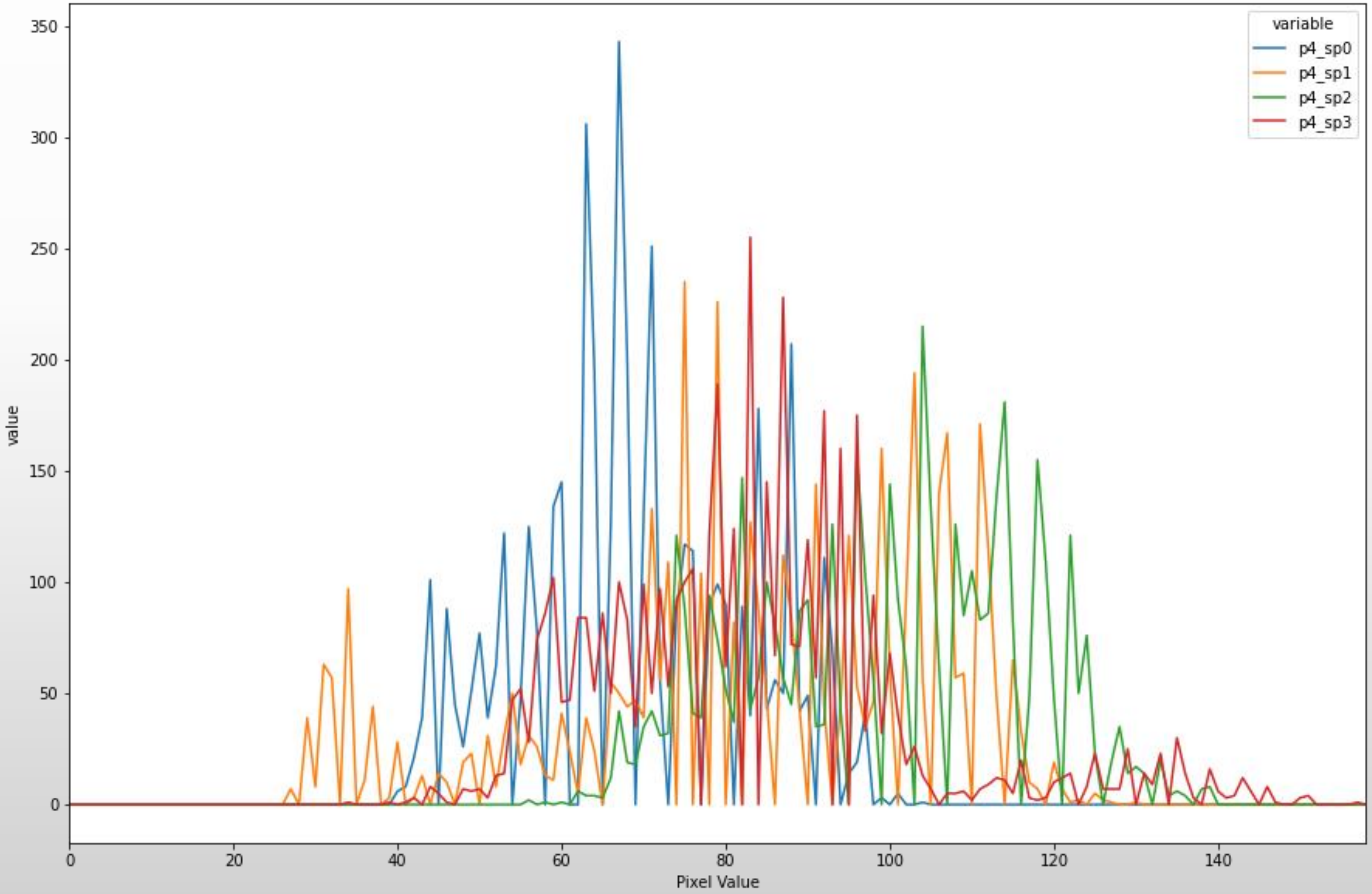
Principal Component Analysis is the process of computing the principal components and using them to perform a change of basis on the data. We find the first few principal components and ignore the rest

✓ Line plot of pixel value depending on pixel position and spectrum

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Data Visualization

- ✓ Spectrum Visualization

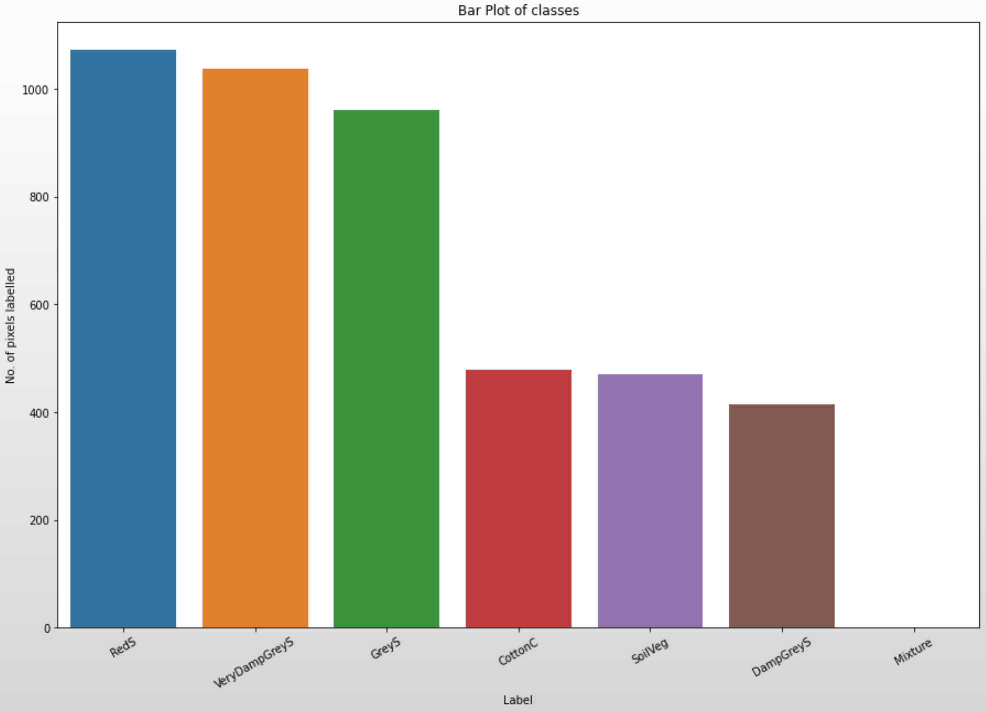


Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Data Visualization

✓ Bar Plot of classes

In the description of the dataset, it is indicated that we don't have values for the 6th label 'mixture classes', we see that on the bar plot, we indeed don't have values for the 6th label. We can, thanks to this plot, get the proportion of pixels for each labels for our observations.



We see on the plot on the left that 3 labels are largely more represented that the 3 others :

- Red Soil
- Very damp grey soil
- Greil Soil

The following 3 labels are less represented in our dataset but are still significant :

- Cotton crop
- Soil with vegetation stubble
- Damp grey soil

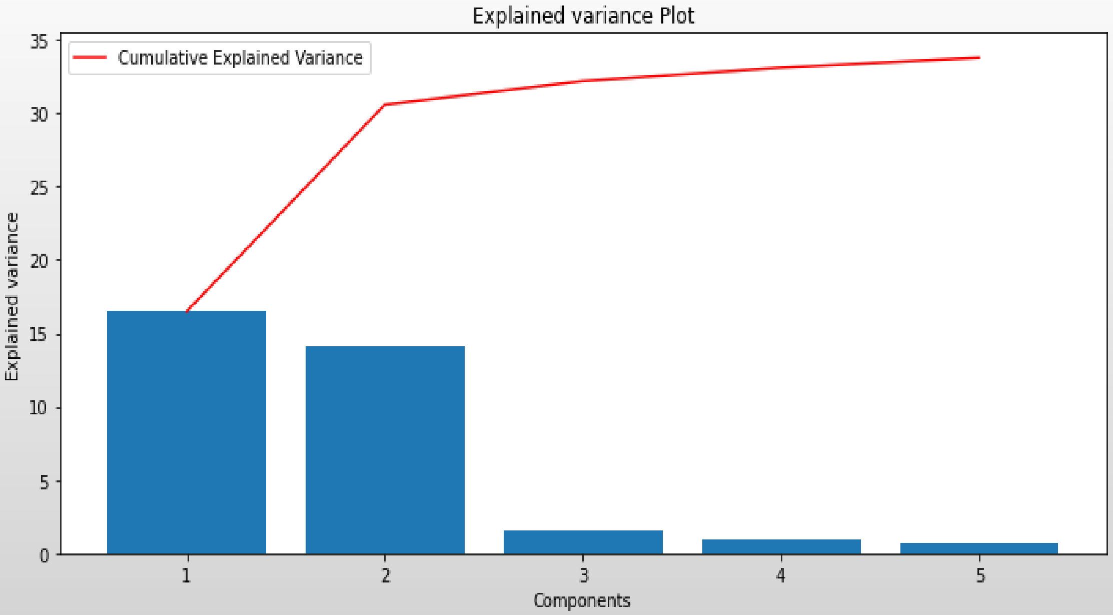
Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

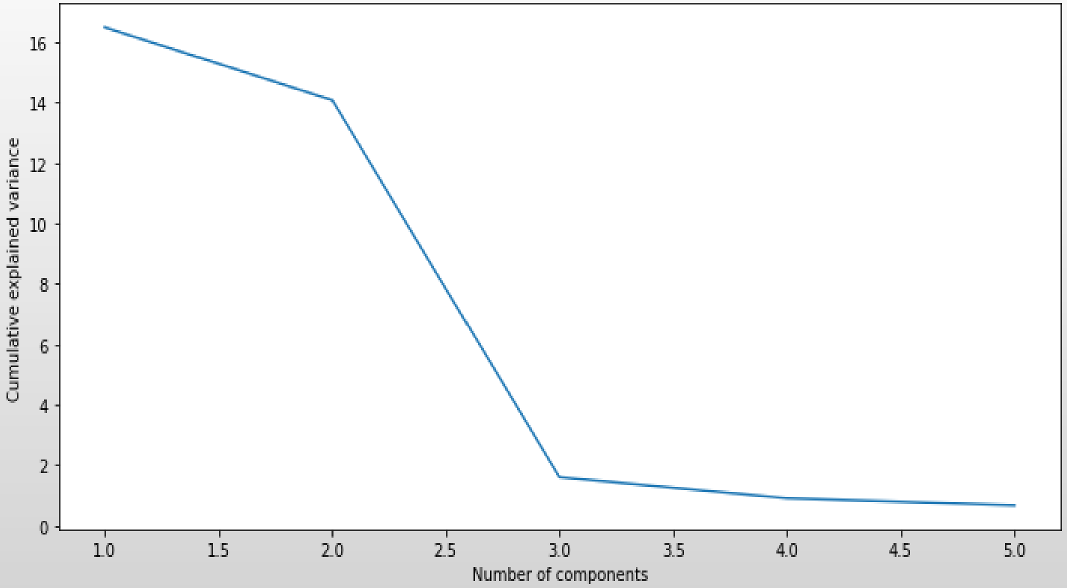
Data Visualization

✓ Principal Component Analysis (PCA)

Before applying PCA on our data, we standardize our initial variables so that each one of them contributes equally to the analysis.

With this analysis we find out that the two first principal components store most of the information. We see this on the scree plot below which plot the percentage of explained variance depending on the number of variables that we consider.





Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Classification Models

Predict the label to which our observations belong

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Classification Models

✓ Logisitic Regression

Logistic Regression is used as our target value is categorical. Logistic regression models the probability that the response Y belongs to one of our 7 cateogries.

✓ Decision Tree

Decision Tree builds a tree based on data then splits the data, based on the "best feature" in the dataset, into subsets that contain possible values for the beast feature. Algorithm then recursively generate new tree nodes until we have optimised maximum accuracy.

✓ Random Forest

Random Forest operates by constructing a multitude of decision trees but correct for decision tree's habit to overfit the data. It trains the different decision trees, and each individual tree predicts the value for target value.

✓ Random Forest Bagging

Bagging repeatedly selects a random sample of the training set and fits trees to these samples. It trains individual models in a parallel way. Each model is trained by a random subset of the data.

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH

Regression Models

Results of our models

Thanks to cross-validation, we compared our different classifiers with default values for the hyperparameters. We found the following results on accuracy on the test dataset :

Classifier	Accuracy
BaggingClassifier	0.886
RandomForestClassifier	0.893
DecisionTreeClassifier	0.8485
LogisticRegression	0.8395

We then took the two best models (Bagging Classifier and RandomForestClassifier) while modifying the hyperparameters and obtained these results showing that the RandomForestClassifier is slightly more performant.

Classifier	Accuracy
BaggingClassifier(n_estimators=15)	0.8855
RandomForestClassifier(n_estimators=15)	0.8945

Charly Kyan ALIZADEH ASLAN | Jacky KUOCH