CS432-Project-3

Ammar Ahmad: 19100176
Rabeez Riaz: 19100165
Tooba Mukhtar: 19100210

Important Preprocessing Steps

- Feature scaling was used to standardize the range of all the numeric variables of data. We normalized the the range of all features so that each feature contributes approximately proportionately to the final model.
- Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling as compared to without it.
- We used feature engineering in the preprocessing phase which showed us the importance of our features relative to each other so before training the models using various classifiers we removed the three least important features from our data.The following are the removed features in addition to the categorical variables:
    - Wilderness Area
    - Soil_Type
    - Slope
    - Aspect

Classification

We used Decision Tree as a classifier since they can handle both the numerical and categorical data and our data contains both types of variables. Besides this they are extremely fast as well.

We later on used Random Forest as a classifier because RF overcome several problems with decision trees, including :

- Reduction in overfitting: Since RF generates a lot of trees so their results can be averaged hence there is a significantly lower risk of overfitting.

The data was also classified using Naive Bayes classifier. We used this since a NB classifier will converge quicker as compared to other models hence less training data is required to train the model.

The following are results of the models based on our data:

Decision Tree

Results

Accuracy = 67 % Max_depth = 5

Accuracy = 74 % Max_depth = 10

Accuracy = 83 % Max_depth = 15

Accuracy = 89.7 % Max_depth = 20

Accuracy = 92.1 % Max_depth = 30

Analysis

- As we increase the depth of the decision tree the accuracy of the model increases because the training error reduces.
- The model was fitted after dropping the less important features such as Wilderness Area, Slope, Aspect however this did not affect the overall accuracy of the model significantly. The change in accuracy was less than 1%.
- The model was fitted after dropping the categorical variables from the data which are wilderness area and soil type. This reduced the accuracy of the model by around 3% since soil_type was an important feature in our data.

The accuracy of the model is high since a DT can easily handle feature interactions between the data.

Random Forest

Results

Accuracy = 94.01 %

Analysis

- The model was fitted after dropping less important features such as Wilderness Area, Slope and Aspect however this did not affect the overall accuracy of the model.The change was less than 1%.
- The model was also fitted after dropping the categorical variables from the data which are wilderness area and soil type. This reduced the accuracy of the model by around 3% since soil_type was an important feature in our data.

The accuracy with this model is the highest since RF deals with the problem of overfitting and by using multiple trees the chance of stumbling across a classifier that doesn't perform well is less.

Naive Bayes

Results

Accuracy = 60.01 %

Analysis

- The model was fitted after dropping less important features such as Wilderness Area, Slope and Aspect however this did not affect the overall accuracy of the model.The change was less than 1%.
- The model was also fitted after dropping the categorical variables from the data which are wilderness area and soil type. This reduced the accuracy of the model by around 3% since soil_type was an important feature in our data.

The low accuracy of the model might be because NB does not learn the interactions between the features in the data.

Association Rules

The categorical variables Soil_Type, Wilderness and Cover_Type were used to find the association rules. We calculated Support, Confidence, Lift, Imbalance ratio, Chi-square, Jaccard,kulczynski and maximum confidence as part of our association rules.

Minimum support = 0.02

Minimum confidence = 0.75

The following results represents the data without binning.

This resulted in a total of 25 association rules.

The following are the top 10 association rules from the data based on the Lift measure:

- {Cover_Type=2,Soil_Type=13} => {Wilderness_Area=3}

  Support = 0.023

  Confidence = 1

  Lift = 2.29

- {Cover_Type=2,Soil_Type=31} => {Wilderness_Area=3}

  Support = 0.023

  Confidence = 1

  Lift = 2.29

- {Soil_Type=13} => {Wilderness_Area=3}

  Support = 0.029

  Confidence = 0.98

  Lift = 2.26

- {Soil_Type=31} => {Wilderness_Area=3}

  Support = 0.043

  Confidence = 0.98

Lift = 2.26

- {Cover_Type=2,Soil_Type=33} => {Wilderness_Area=3}

  Support = 0.042

  Confidence = 0.97

  Lift = 2.24

- {Soil_Type=12} => {Wilderness_Area=1}

  Support = 0.051

  Confidence = 1

  Lift = 2.23

- {Soil_Type=30} => {Wilderness_Area=1}

  Support = 0.052

  Confidence = 1

  Lift = 2.23

- {Cover_Type=2,Soil_Type=12} => {Wilderness_Area=1}

  Support = 0.047

  Confidence = 1

  Lift = 2.23

- {Cover_Type=2,Soil_Type=30} => {Wilderness_Area=1}

  Support = 0.035

  Confidence = 1

  Lift = 2.23

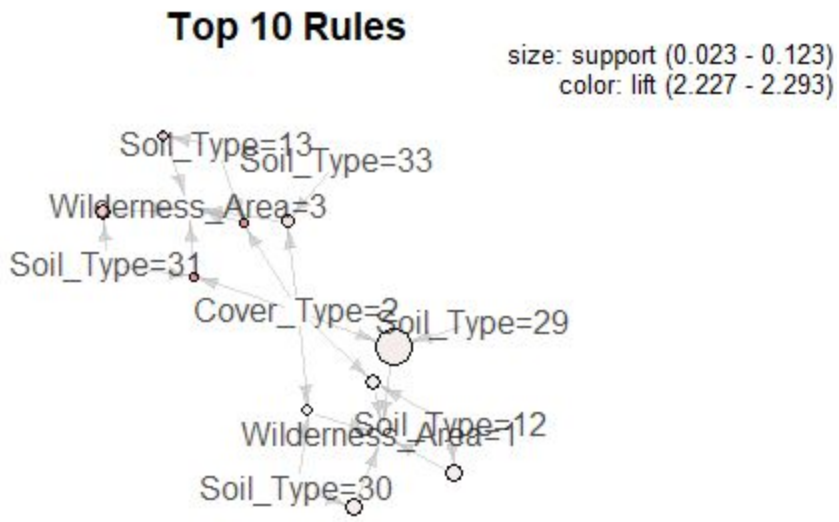
- {Cover_Type=2,Soil_Type=29} => {Wilderness_Area=1}

  Support = 0.123

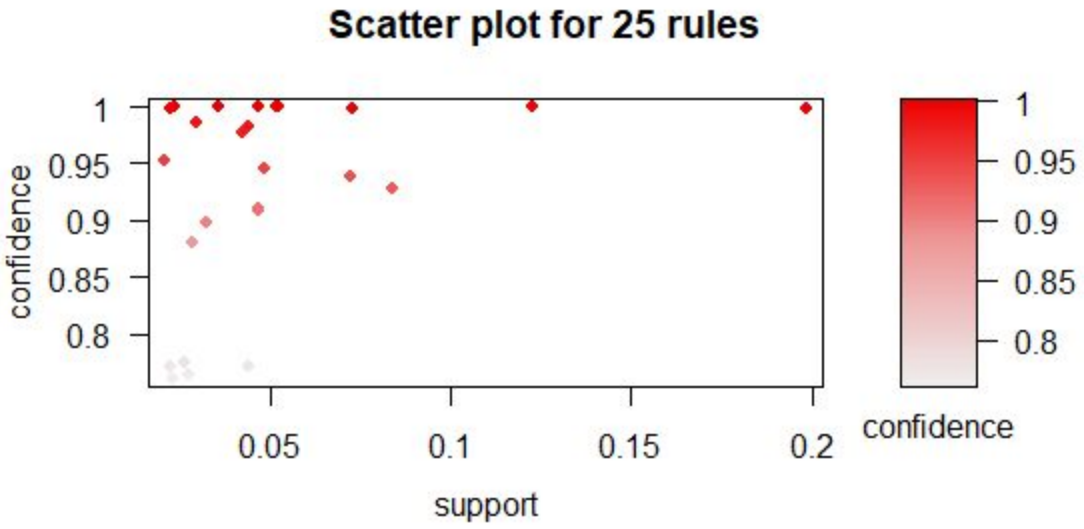
  Confidence = 0.99

  Lift = 2.23

The following diagram illustrates the top 10 rules:



The following is a scatter plot of all the association rules from the data:



As a result of the association rules we only found forest cover type Spruce/Fir and Lodgepole Pine .on the R.H.S of those rules. The following is a description of the association rules according to the Cover Type.

Association Rules for Cover_Type 1 = Spruce/Fir

- {Wilderness_Area=1,Soil_Type=22} => {Cover_Type=1}

  Support = 0.026

  Confidence = 0.78

  Lift = 2.13

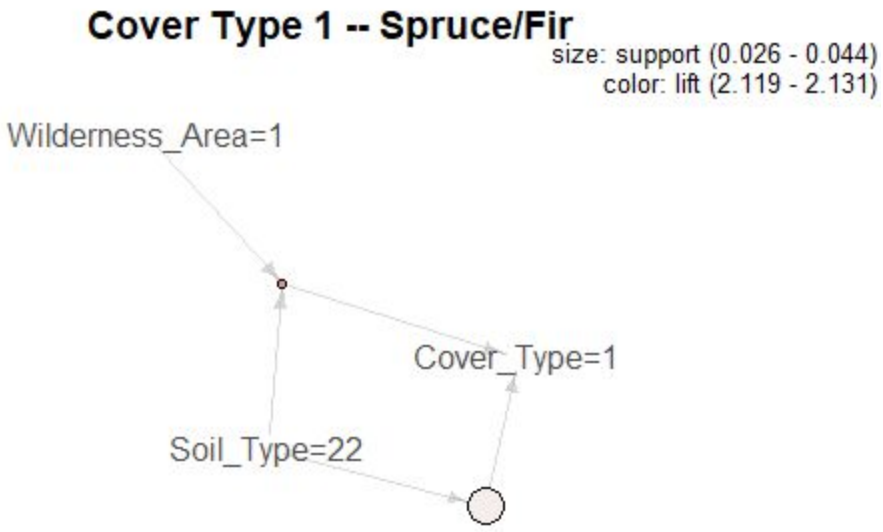
- {Soil_Type=22} => {Cover_Type=1}

  Support = 0.044

  Confidence = 0.77

  Lift = 2.11

The following diagram shows 2 association rules.



Association Rules for Cover_Type 2 = Lodgepole Piner

- {Wilderness_Area=1,Soil_Type=12} => {Cover_Type=2}

  Support = 0.047

  Confidence = 0.91

  Lift = 1.87

- {Wilderness_Area=3,Soil_Type=13} => {Cover_Type=2}

  Support = 0.023

  Confidence = 0.77

  Lift = 1.58

- {Soil_Type=13} => {Cover_Type=2}

  Support = 0.023

  Confidence = 0.76

  Lift = 1.56

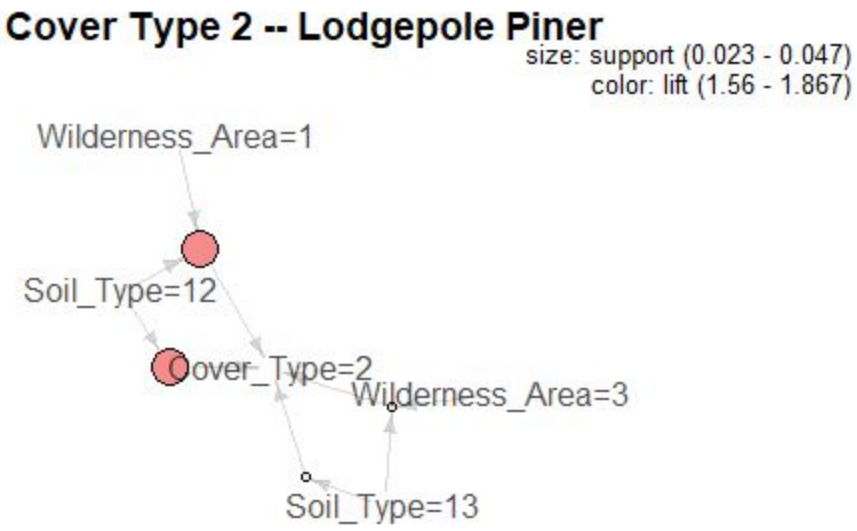
- {Soil_Type=12} => {Cover_Type=2}

  Support = 0.047

  Confidence = 0.91

  Lift = 1.87

The following diagram shows 4 association rules.



The following results represents the data with binning.

- {Soil_Type=11} => {Wilderness_Area=3}

  Support = 0.020

  Confidence = 0.95

  Lift = 12.81

- {Soil_Type=38} => {Elevation=10}

  Support = 0.020

  Confidence = 0.76

  Lift = 12.53

- {Cover_Type=6} => {Elevation=1}

  Support = 0.024

  Confidence = 0.80

  Lift = 12.03

- {Soil_Type=13} => {Wilderness_Area=3}

  Support = 0.030

  Confidence = 0.99

  Lift = 9.96

- {Soil_Type=13} => {Cover_Type=2}

  Support = 0.023

  Confidence = 0.76

  Lift = 9.94

- {Cover_Type=7} => {Elevation=10}

  Support = 0.029

  Confidence = 0.83

  Lift = 9.92

- {Soil_Type=24} => {Wilderness_Area=3}

  Support = 0.028

  Confidence = 0.76

  Lift = 9.90

- {Soil_Type=31} => {Wilderness_Area=3}

  Support = 0.043

  Confidence = 0.98

  Lift = 9.90

- {Soil_Type=12} => {Wilderness_Area=1}

  Support = 0.05

  Confidence = 1

  Lift = 9.89

- {Soil_Type=12} => {Cover_Type=2}

  Support = 0.05

Confidence = 0.91

Lift = 9.59

The results of the association rules after binning and before binning are approximately the same since both of them include forest cover type Spruce/Fir and Lodgepole Pine on their R.H.S.

Boosting Techniques

The boosting techniques work on the weak classifiers and apply classification on the results of these weak classifiers however our dataset does not give a good accuracy with weak classifiers so we haven't used those. As a result of this the boosting technique does not improve the accuracy of our model.
We used 2 boosting techniques. The following are the results:

Adaboost

Accuracy = 59.2%

Report:

	precision	recall	f1-score	support
1	0.65	0.62	0.64	42575
2	0.72	0.62	0.66	56536
3	0.41	0.92	0.56	7103
4	0.00	0.00	0.00	514
5	0.00	0.00	0.00	1895
6	0.00	0.00	0.00	3457
7	0.08	0.21	0.12	4123
avg / total	0.61	0.59	0.59	116203

Gradient Boosting

Accuracy = 75.1%

Report:

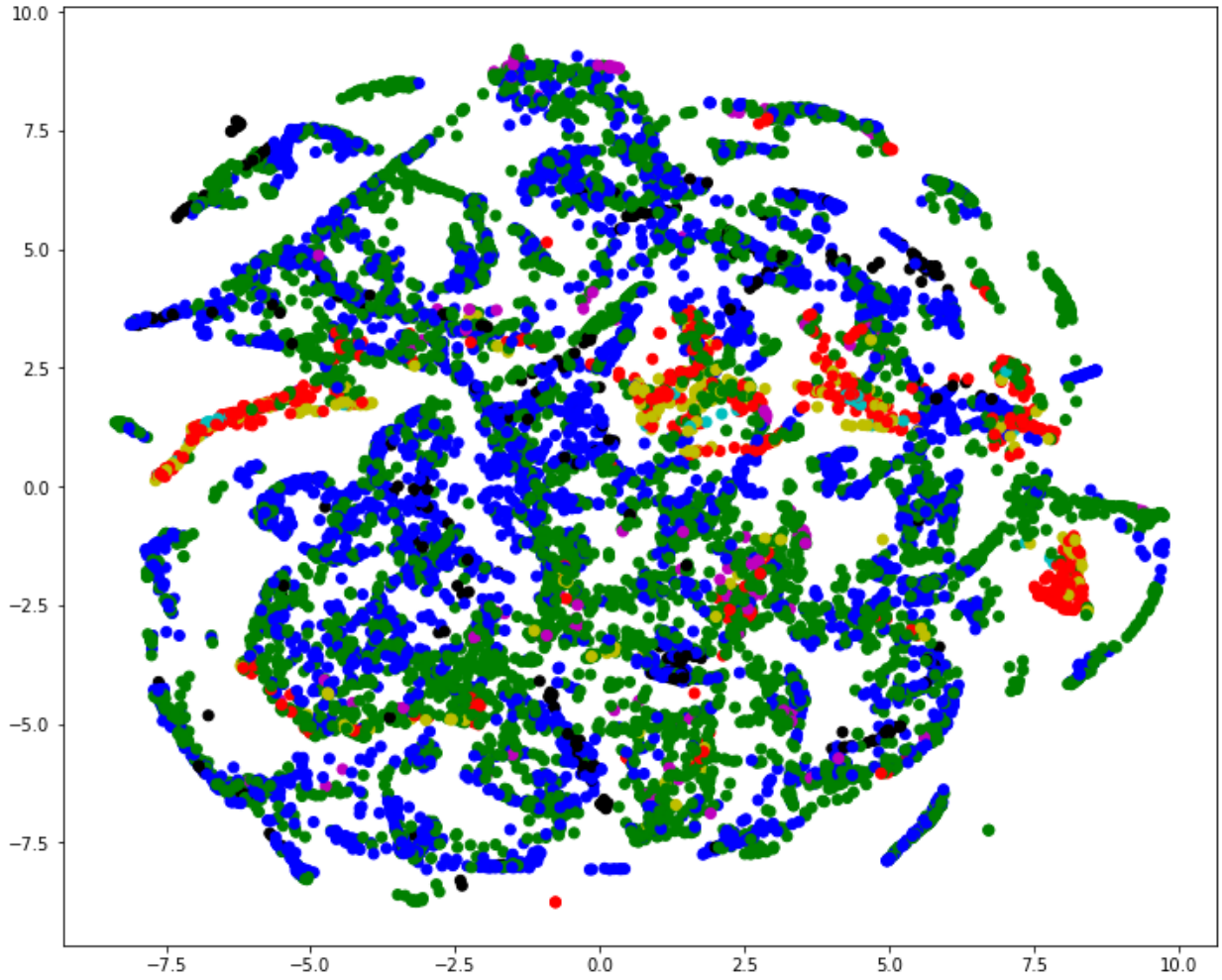
	precision	recall	f1-score	support
1	0.73	0.74	0.74	42575
2	0.76	0.82	0.79	56536
3	0.74	0.76	0.75	7103
4	0.81	0.65	0.72	514
5	0.77	0.16	0.26	1895
6	0.64	0.35	0.46	3457
7	0.84	0.52	0.64	4123
avg / total	0.75	0.75	0.74	116203

Clustering

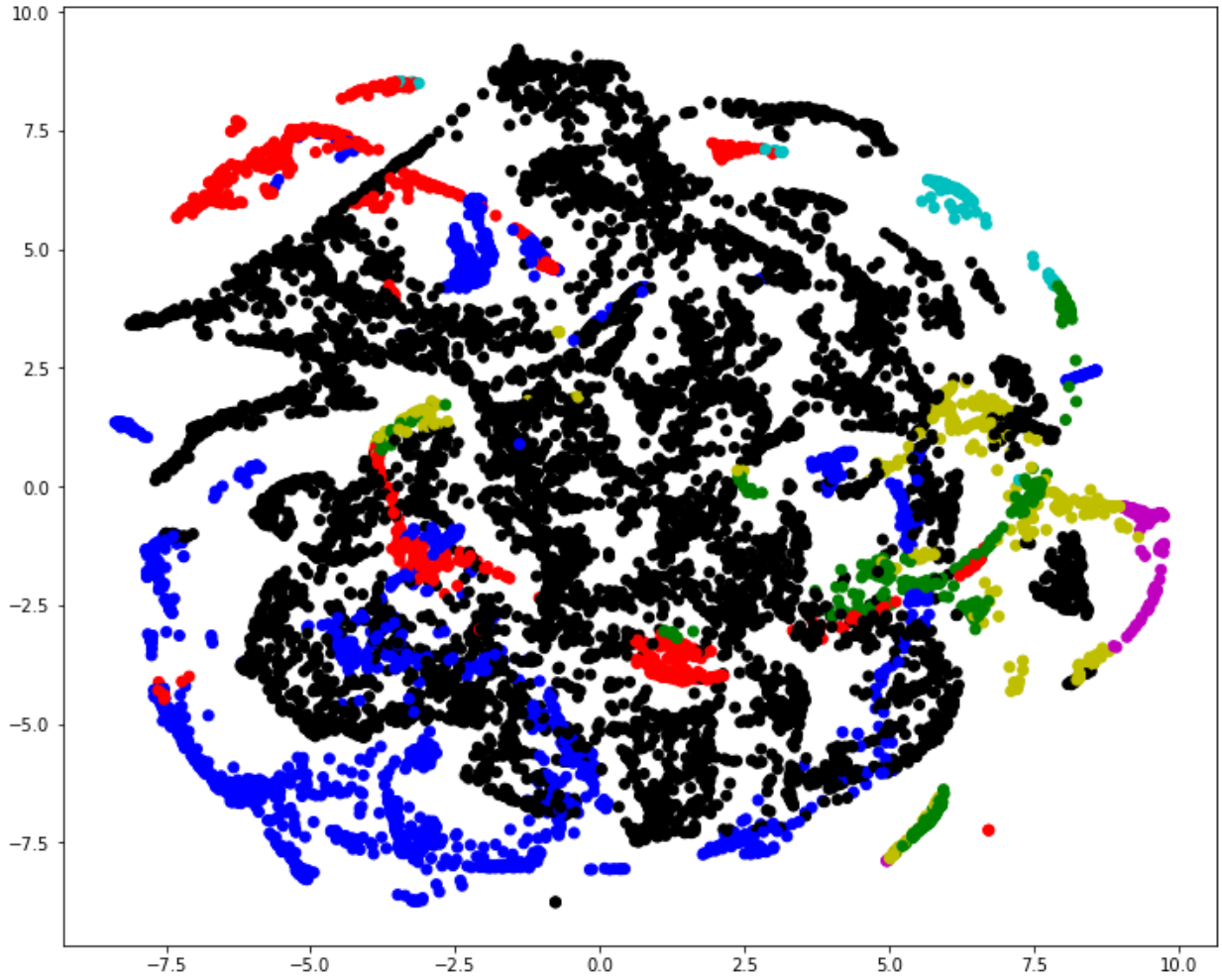
We removed the categorical variables from our data before applying the clustering technique. We used K-Mean clustering on our data. The clustering was done with different bandwidths. Each bandwidth resulted in different number of clusters. The following are the results:

- Bandwidth = 500 Number of Clusters = 34
- Bandwidth = 700 Number of Clusters = 7
- Bandwidth = 1000 Number of Clusters = 2

We considered the cluster with bandwidth 700 since it results in 7 clusters which represent the number of Cover_Types in our data.



There was a visualization issue with the clusters because of the 10 dimensional feature vector in our data so dimensionality reduction with 2 components was done using t-Distributed Stochastic Neighbor Embedding.This was fitted on 10000 examples from the data. After the dimensionality reduction there was no visible difference between the clusters hence no cluster labels.


10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Observations

The Lodgepole Pine trees are the most common in the Roosevelt National Forest.
The Lodgepole Pine trees belong to Soil_Type 12 and Soil_Type 13 according to the following association rules :

- {Soil_Type=13} => {Cover_Type=2}
- {Soil_Type=12} => {Cover_Type=2}

The Spruce/Fir trees belong to Soil_Type 22 according to the following association rules :

- {Soil_Type=22} => {Cover_Type=1}

According to the association rules Lodgepole Pine trees belongs to the wilderness Area 1 and 3.