The Report of Forest Cover Type Prediction and Description

Abstract of Report:

This report mainly explores the data which got from the US Geological Survey(USGS) and US Forest Service(USFS) through the prediction model. The purpose of this report is to assist managers in supporting the decision-making process by predicting the results of the model — for example, the Forest cover type prediction and so on through the characteristics of the data.

This study used two prediction models, Decision Tree (supervised) and Clustering (unsupervised). Among them, Decision Tree will predict the Forest Cover Type and evaluate the classification accuracy under various strategies. Clustering clusters similar data features to try to find the cluster and the corresponding Forest Cover Type features.

This report is divided into four parts: The first is a data dictionary, which describes the definition and description of the materials and feature elements used and provides a clear explanation and introduction of the content of the data that will be used in the subsequent follow-up. This is followed by Data cleansing and pre-processing. In order to make model calculations faster and better predictive results, data preprocessing is necessary and essential. These include combining attributes, detecting missing data, sampling strategies, feature selection strategies, and Data Balancing strategies and so on.

As can be seen from the results of Decision Tree, the model through feature selection (93.075%) is better than the model without feature selection (91.913 %), and the model after data balancing (93.054%) is not superior to the model with unbalance data. On the other hand, Clustering can also find the relationship of some corresponding Forest Cover Types according to the optimal clusters (k=7). There is a cluster that can match Type 7 and the other one can match Type 2, and a cluster with type 3, 4, and 6 features.

Finally, there is a figure to detail the reliable data characteristics pattern of Forest Cover Type 4 (Cottonwood/Willow) to assist in the detection and provision of decision support for managing the forest type.

Section 1 – Data Dictionary

- Attribute information:

Attribute name and a brief description of the data are provided in the table. The sort in the table is the same as the sort provided for data analysis.

Column Name	Description	Data Type
Elevation	Elevation in meters	quantitative
Aspect	Aspect in degrees azimuth	quantitative
Slope	Slope in degrees	quantitative
H_Dist_Hyd	Horz Dist to nearest surface water features	quantitative
V_Dist_Hyd	Vert Dist to nearest surface water features	quantitative
H_Dist_Road	Horz Dist to nearest roadway	quantitative
Shade_9	Hillshade index at 9am, summer solstice	quantitative
Shade_12	Hillshade index at noon, summer soltice	quantitative
Shade_3	Hillshade index at 3pm, summer solstice	quantitative
H_Dist_FP	Horz Dist to nearest wildfire ignition points	quantitative
Soil_Type	Soil Type designation(40 types)	qualitative
Wildeness_Designation	Wilderness area designation(4 types)	qualitative
Cover_type	Forest Cover Type designation(7 types)	integer

- Soil Types: 1 to 40 based on the USFS Ecological Landtype Units (ELUs)
- Wilderness Areas: 1: Rawah 2: Neota 3: Comanche Peak 4: Cache la Poudre
- Forest Cover Type: 1: Spruce/Fir 2: Lodgepole Pine 3: Ponderosa Pine
- 4: Cottonwood/Willow 5: Aspen 6: Douglas-fir 7: Krummholz
- Number of instances(records): 581,012
- Number of attributes: 12 measures
- Missing attribute percentage: 0%
- Target (class) attribute: Cover_type (= 1 to 7)
- Target (class) distribution:

1.Spruce-Fir	36.46%	5. Aspen	1.63%
2. Lodgepole Pine	48.76%	6. Douglas-fir	2.99%
3. Ponderosa Pine	6.15%	7. Krummholz	3.53%
4.Cottonwood/Willow	0.47%		

- Data Source:

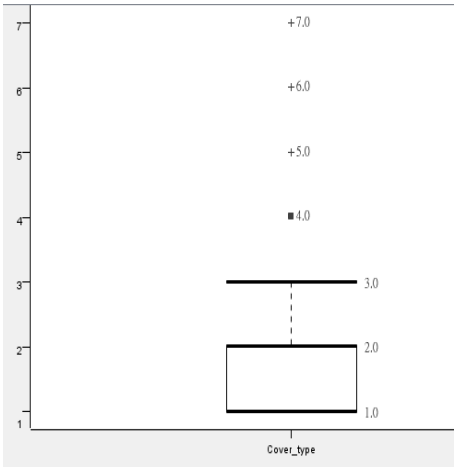
Forest cover type: US Forest Service (USFS) Region 2 Resource Information System (RIS) data

Other Independent variables: US Geological Survey (USGS) and USFS data.

Section 2 – Data Cleansing and Pre-processing

For the boost analysis efficiency, the Wideness area designation and Soil Type of the original data have changed to its corresponding number from the binary value. (i.e Rawah Wilderness Areas denoted by 1 and Neota Wilderness Areas denoted by 2) Please see Appendices (figure 1) for the process details.

I split the raw data randomly into a training dataset and a test dataset through **stratified sampling**, which composed 80% and 20% respectively (training data has 464,809; test data has 116,203). The train data was used to build the model and the test data was used to evaluate the model. The reason to use stratified sampling because cover_type is unevenly distributed among the types, and cover_type is split into sub-populations according to different types by stratified sampling, to avoid pulling out too many specific types.



i. Data Sampling:

Then, evaluate and select feature combinations with higher accuracy and reflects the structure of the database but using less data toward getting a better result. The feature selection strategy is Forward Feature Selection(FFS) which is added to the feature set starting with the best individual feature until exploring the different subset.

Next, selects 80% of training data by **drawing random** which we split before to be used in the classifier that follows and 20% data for validation (training data has 371,847; validation data has 92,962). Finally, using Decision Tree classifier to the best accuracy feature combinations.

At the same time, Backward Feature Elimination(BFE) was also implemented to compare the best accuracy of features with Forward Feature Selection. However, the Forward Feature Selection has a better performance.

	Accuracy	Nr. of features	Selected features
FFS	93.938%	7	Elevation, Aspect, H_Dist_Hyd, H_Dist_Road, H_Dist_FP, Soil_Type and Wilderness_Designation
BFE	92.4%	9	Elevation, Slope, H_Dist_Hyd, V_Dist_Hyd, H_Dist_Road, Shade_3,

			H_Dist_FP, Soil_Type and Wilderness_Designation

It can be seen that Backward Feature Elimination selected more features, but the accuracy is lower than the Forward Feature Selection. That the reason why chooses Forward Feature Selection as the feature selection strategy (higher accuracy and fewer features to save the compute time).

Next, to avoid that the variability of the data was not captured after feature selection, hence, the values of the mean and standard deviation were checked to be identical with the original data.

ii. Data Balancing Strategies:

Balancing data is necessary because minority or exceptional classes in classification will cause misclassification or ignore the minority class. Here I use two balancing methods to compare the result.

a. Equal Size Sampling: Downsizing the data set at random by removes rows until the minority class has achieved the necessary balance. The values in the cover_type column will be equally distributed.

Category	**Total Values**	**Category**	**Total Values**
Type 1	2,197	Type 5	2,197
Type 2	2,197	Type 6	2,197
Type 3	2,197	Type 7	2,197
Type 4	2,197	**Sum:**	**15,379**

b. SMOTE: The algorithm oversamples the training data to enrich it. It can help to equal cover_type class distribution to get excellent classification performance.

Category	**Total Values**	**Category**	**Total Values**
Type 1	226,641	Type 5	226,641
Type 2	226,641	Type 6	226,641
Type 3	226,641	Type 7	226,641
Type 4	226,641	**Sum:**	**1,586,487**

c. The Mean and Standard Deviation of the Sample

After sampling, it must be confirmed whether the sample has the variability of capturing the entire sample. Therefore, the standard deviation and the mean can be regarded as the standard to be inspected.

Row Data	Mean	Variance	SD
Elevation	2,959	78,401	280
Aspect	156	12,513	112
H_Dist_Hyd	270	45,138	212
H_Dist_Road	2,350	2,431,942	1,559
H_Dist_FP	1,980	1,756,289	1,325
Soil_Type	24	90	9
Wildemess_Designation	2	1	1

Row Data Statistic Table

After Feature Selection	Mean	Variance	SD
Elevation	2,959	78,391	280
Aspect	156	12,525	112
H_Dist_Hyd	269	45,177	213
H_Dist_Road	2,350	2,431,276	1,560
H_Dist_FP	1,980	1,753,493	1,324
Soil_Type	24	90	9
Wildemess_Designation	2	1	1

Statistic Data After Feature Selection

It can be seen from the above tables that the mean and SD of sub-sample after Feature selection are almost the same as raw data, and the values after data balancing are similar too.

Section 3 – Decision Tree Model

According to the purpose of this project is to identify the forest type. The model determines the type of forest for a particular region. That is to say; it's easy to provide different management for the forest depending on its type.

The Decision Tree(DT) model can solve such classification problems, and the forest type can be predicted based on its new cells' measurements or relevant data. In order to find out the best Decision Tree model, the performance of different strategy models will be tested through the test set. Finally, the best model is determined by comparing the predicted value with the actual value for classification.

i. There are four different **Prediction Models and Pre-processing Strategies**:

a. Feature selection and without feature selection (Figure 2)

1) Model 1: DT model **without** feature selection
2) Model 2: DT model **with** feature selection

b. Two different way to balance train data(Figure 4)
   3) Model 3: DT model with feature selection and balancing data by **Equal Size Sampling**(Use exact sampling)
   4) Model 4: DT model with feature selection and balancing data by **SMOTE**(Oversample minority classes):

The algorithm of all Decision Trees used Gain ratio without pruning because of the better performance. After training prediction models, evaluating the classification model performance is the next phase. The test set was used to evaluate performance, and the result of scorer means accuracy rate of prediction.

	Feature Selection	**Data Balancing**	**Scorer**
Model 3	YES	Equal Size Sampling	66.891%
Model 4	YES	SMOTE	93.054%
Model 2	YES	NA	93.075 %
Model 1	NO	NA	91.913 %

It can be seen that the best performance is the Model 2 (93.075%). It is followed by model 4 which did data balancing (93.054%).

ii. **Tune Parameter:** After selecting the appropriate model and pre-processing strategy, it also needs to find the best parameters to fine-tune the model.

I set a loop to Start values at 2, with each step size =1, and loop to 15. Based on the maximum accuracy of the Scorer, find the best Parameter. Fortunately, the default parameter at the beginning is the best parameter, so it does not need to be changed.

	Scorer	**Default Parameter**	**Best Parameter**
Model 4	93.054%	2	2
Model 2	93.075%	2	2

iii. Model Evaluation

In addition to verifying the accuracy of test data through the scorer, the Receiver Operating Characteristic(ROC) Curve is also drawn to evaluate the appropriate model.

When the model is entirely worthless, the sensitivity is equal to the false positive rate, which is a diagonal line from the origin to the upper right corner. This line is called the opportunity line. The more the ROC curve deviates from the opportunity line (offset to the upper left), the larger the area under the ROC curve, the better the authenticity of the model, indicating the higher the accuracy of the prediction.

Similarly, the Area Under the Curve (AUC) can be used to discriminate the discriminating power of the ROC curve. The AUC value ranges from 0 to 1, and the larger the value, the better. However, If the model 2 is compared with the model 4 through AUC, it is difficult to distinguish the more suitable model because the results of the seven cover types are different. Finally, model2 will be recommended as the optimal model based on the results of Scorer and its precision.

Section 4 – Clustering Model

Now, we will adopt Clustering for the Forest Cover Type dataset to find clusters based on similarity measurement of data features. In simple terms, Clustering divides similar objects into different subsets, so that members of the same subset have similar properties. The idea is to cluster the database to obtain different clusters for the various types of cover.

The first step is to normalise the raw data so that all data is between 0~1, which makes it easier to perform cluster distance operations. Next, we have to decide the most appropriate k value.

i. k selection

These k values represent how many clusters should be in the data set, and then all points are assigned to the nearest centroid. After that recalculate the centroid of each cluster until the centroids are no longer changed. I set a loop to let k go from 2 to 8 to find the best result. The result of the best configuration occurs with nine clusters(k=7).

Next, we give the Cluster different colours and go straight de-normalised. In the visual results, the following findings were found:

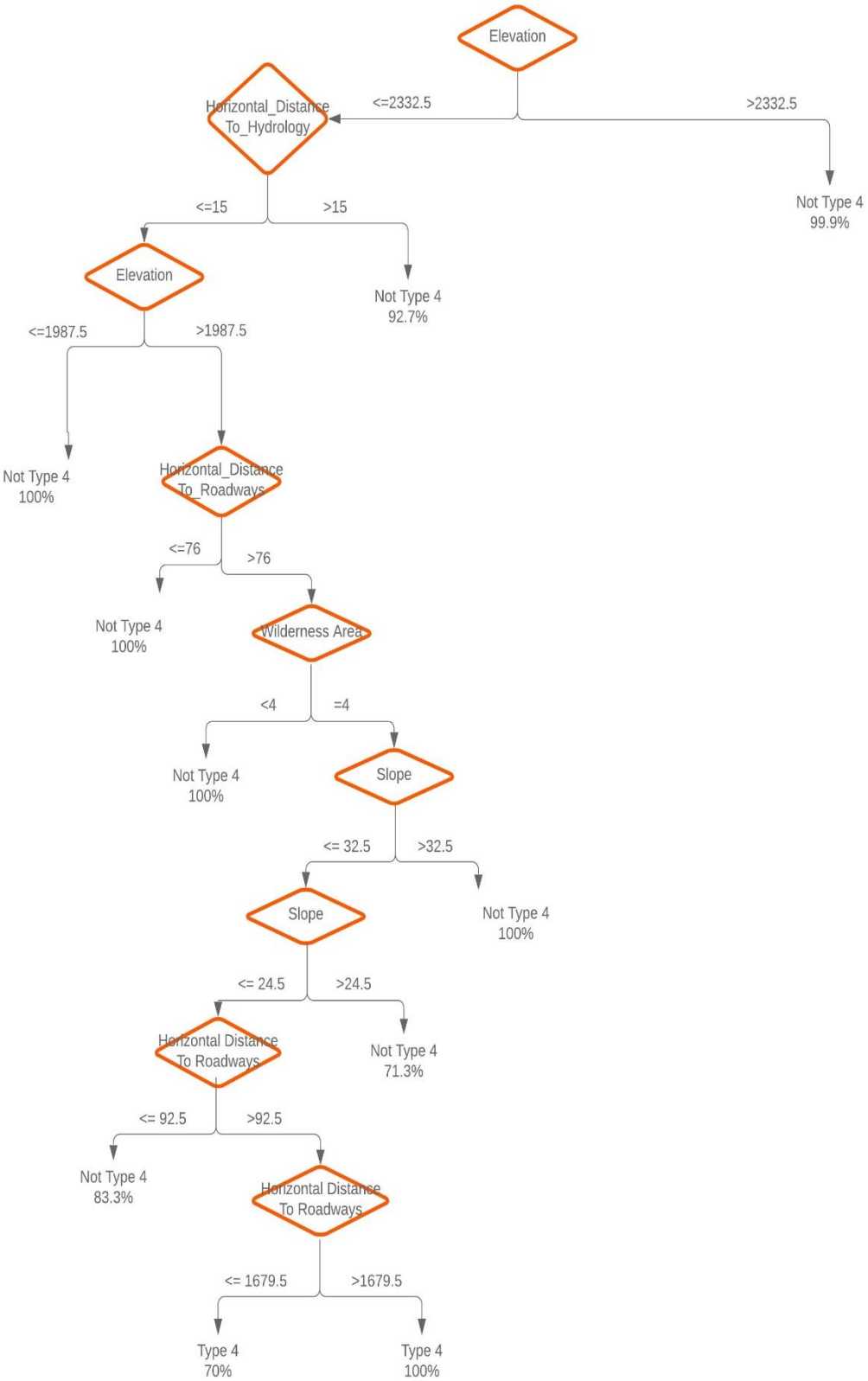
Cluster	Wilderness Area	Note
Cluster 2	area 3 & area 4	The Lowest elevation
Cluster 1	area 1 & area 2	The elevation is between 3159 to 3359; and the Aspect is between 0 to 225
Cluster 6	area 2 & area 3	The highest elevation
Cluster 5	area 2 & area 3	Has the farthest hydrological vertical distance
Cluster 4	area 1 & area 2	The elevation is between 3109 to 3359; and the Aspect is between 0 to 175 (similar with cluster 1)

From the above information, we can find the forest cover type corresponding to some clusters:

Cluster	Cover Type	Note
Cluster 2	Type 3 & 4 & 6	1. Type 4 only grows in area 4 2. Type 3 and Type6 are also only grown in areas 3 and 4 3. Type 3&4&6 has a lowest elevation
Cluster 6	Type 7	1. Has the highest elevation 2. Almost all of them are in area 2 & 3
Cluster 5	Type 2	1. Has the farthest hydrological vertical distance

Section 5 – Description of Forest Cover Type 4

Finally, since Forest Cover Type 4 (Cottonwood/Willow) requires special management, type4 patterns will be found in the following ways. If Type 4 is treated as a class and other types are treated as another class, we can find a reliable conclusion about the Type 4 pattern by the Decision Tree model.


Elevation
<=2332.5
>2332.5
Horizontal_Distance To_Hydrology
Not Type 4 99.9%
<=15
>15
Elevation
Not Type 4 92.7%
<=1987.5
>1987.5
Not Type 4 100%
Horizontal_Distance To_Roadways
<=76
>76
Not Type 4 100%
Wilderness Area
<4
=4
Not Type 4 100%
Slope
<= 32.5
>32.5
Slope
Not Type 4 100%
<= 24.5
>24.5
Horizontal Distance To Roadways
Not Type 4 71.3%
<= 92.5
>92.5
Not Type 4 83.3%
Horizontal Distance To Roadways
<= 1679.5
>1679.5
Type 4 70%
Type 4 100%

Appendices

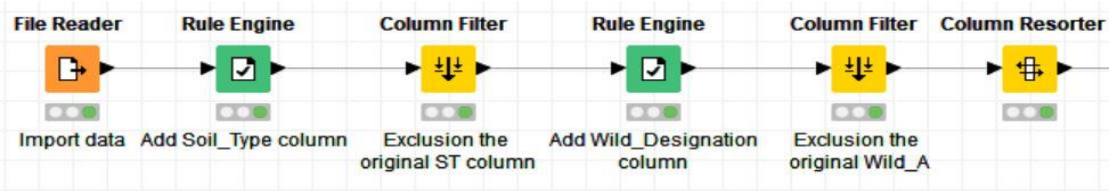


Figure 1: Data Pre-processing

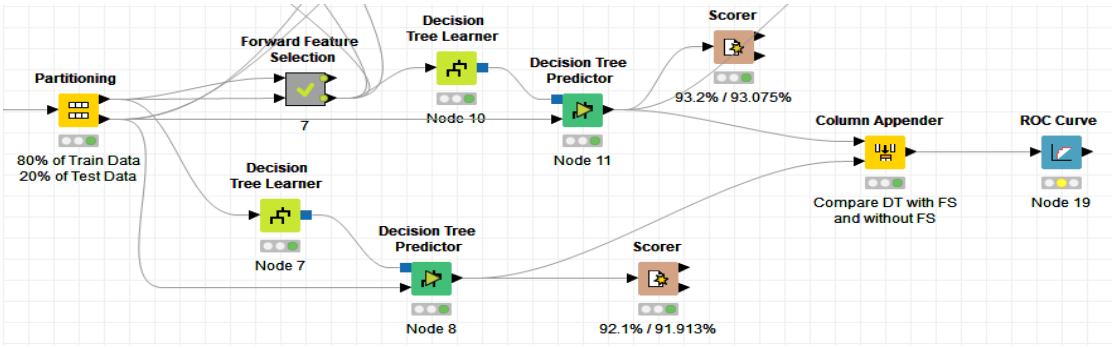


Figure 2: FS vs. no FS

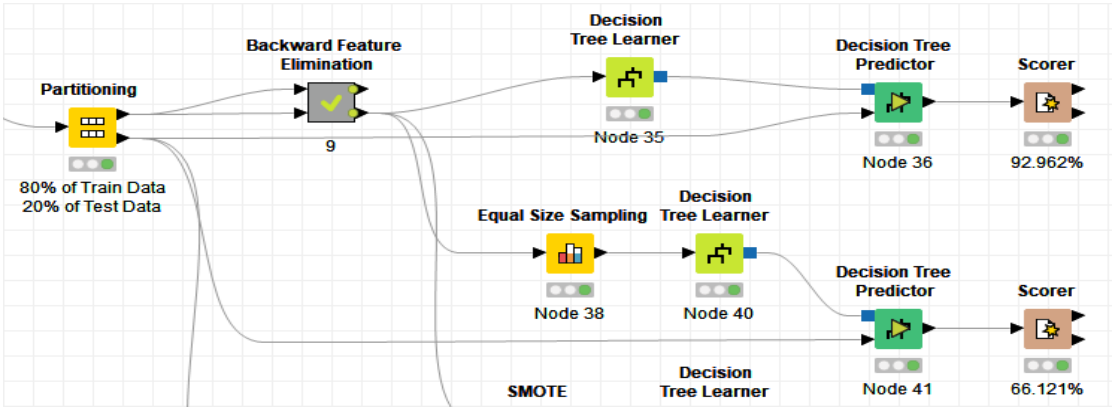


Figure 3: Backward Feature Elimination

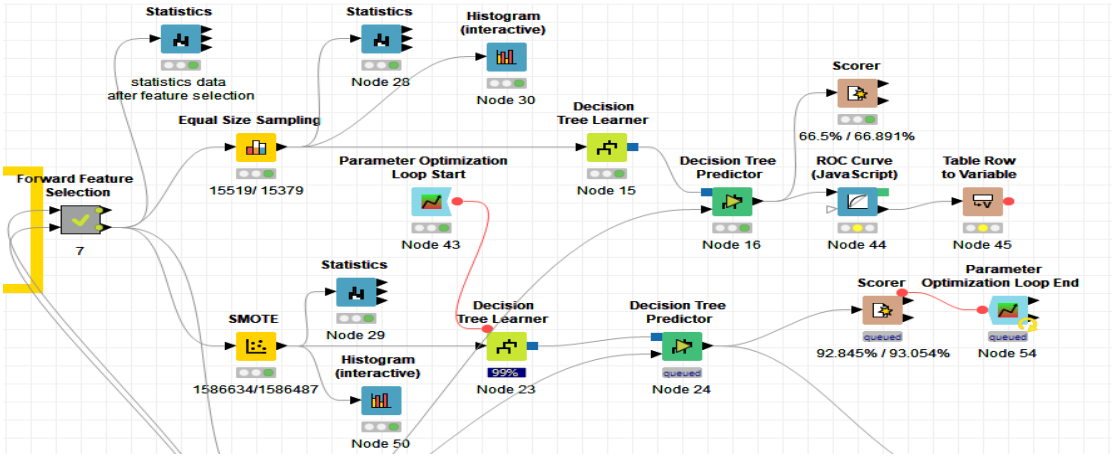


Figure 4: Equal Size Sampling and SMOTE

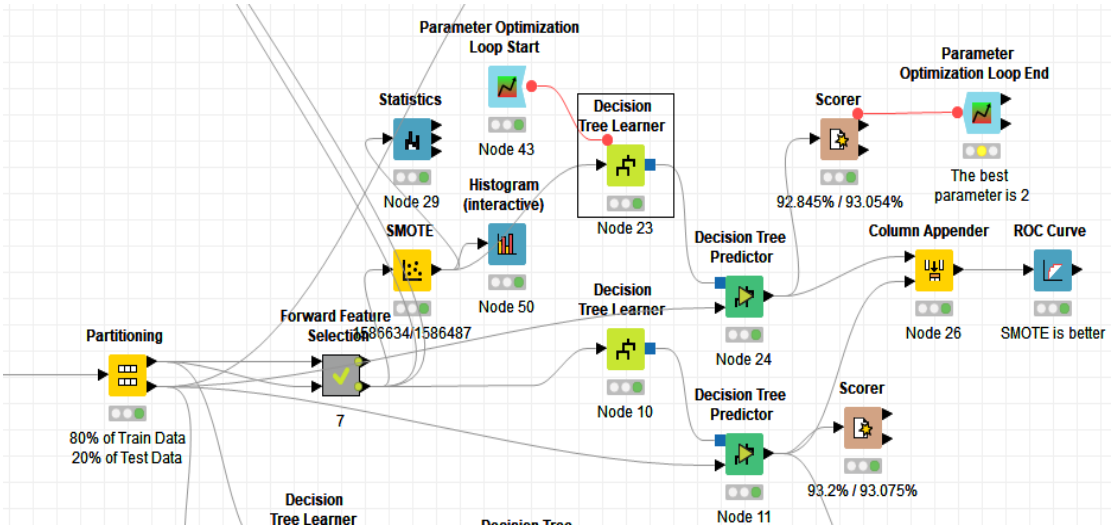


Figure 5: Compare SMOTE results with unbalancing results

Cover_type \ Prediction (Cover_type)	5	2	1	7	3	6	4
5	1649	195	25	0	21	9	0
2	449	53037	2650	37	257	227	3
1	69	2558	39443	284	3	11	0
7	0	23	133	3946	0	0	0
3	23	146	5	0	6526	337	114
6	6	98	9	0	272	3060	28
4	0	0	0	0	53	27	470

Correct classified: 108,131 Wrong classified: 8,072

Accuracy: 93.054 % Error: 6.946 %

Cohen's kappa (K) 0.889

Figure 6: Scorer Table

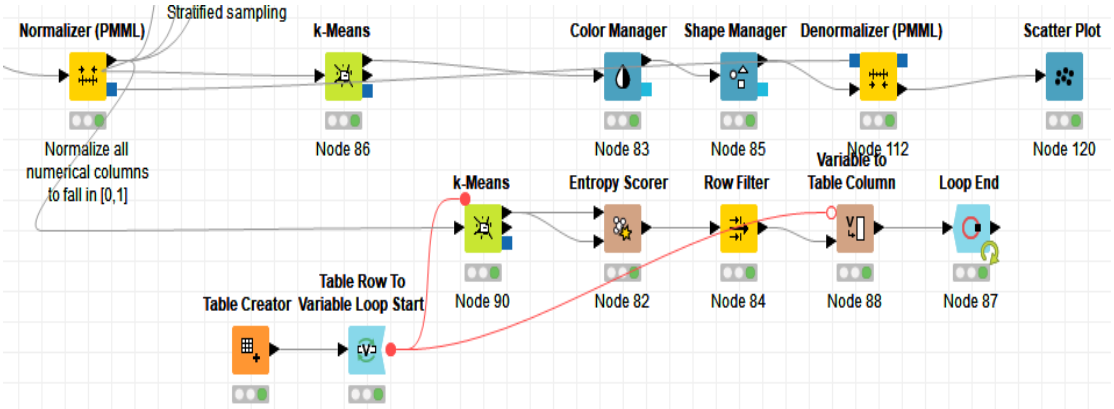


Figure 7: Clustering workflow of All features