

## **Predicting Multi-Class Forest Cover-Type**

**By**

**Ayman Amer Abdulhafed Alkubati**

In this report, using the CRISP-DM methodology, a dataset taken from the UCI Machine Learning repository forest cover type dataset will be analyzed and processed, developing four different models. These are the Decision Tree model (DT), Random Forest (RF) model, Neural Network (MLP type), and Logistic Regression (LR) model, using Knime analytics platform. The CRISP-DM methodology consists of six steps, namely: business understanding, data understanding, data preparation, model building, testing and evaluation, and deployment

## Business Understanding:

Having a dataset containing forest cover type based on attributes like elevation, aspect, slope, hillshade, and soil type, the goal is to predict the forest cover type (7 types) based on these attributes. Factors like elevation, slope, distance to hydrology, and soil type can be analyzed to determine their effect on the forest cover type. This can help gain insights into the composition and distribution of forest cover types.

## Data Understanding:

The dataset contains 581,012 rows and 55 columns consisting of quantitative variables like elevation in meters, slope in degrees, and hillshade index, as well as qualitative binary variables for wilderness areas and soil types. The target variable is an integer from 1 to 7 representing the forest cover type. We need to understand features like the number of missing values in each column, outliers, and skewness of the data. This can help determine appropriate data preparation techniques for modeling.

Numeric    Nominal    Data Preview											
Search: <input type="text"/>											
Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings
Elevation	<input type="checkbox"/>	1863	3849	2745.723	402.333	161871.760	0.103	-0.940	44983185	0	0
Aspect	<input type="checkbox"/>	0	360	155.461	111.062	12334.808	0.463	-1.136	2546924	138	0
Slope	<input type="checkbox"/>	0	61	16.614	8.522	72.623	0.535	-0.077	272179	7	0
Horizontal_Distance_To_Hydrology	<input type="checkbox"/>	0	1343	219.205	205.205	42109.260	1.563	3.145	3591242	1698	0
Vertical_Distance_To_Hydrology	<input type="checkbox"/>	-146	554	49.543	59.755	3570.691	1.588	3.704	811655	2018	0
Horizontal_Distance_To_Roadways	<input type="checkbox"/>	0	6890	1778.801	1337.381	1788587.755	1.114	0.653	29142097	3	0
Hillshade_9am	<input type="checkbox"/>	0	254	212.389	30.546	933.047	-1.054	1.110	3478712	1	4
Hillshade_Noon	<input type="checkbox"/>	99	254	218.413	22.946	526.503	-0.912	0.951	3577597	0	3
Hillshade_3pm	<input type="checkbox"/>	0	248	134.809	45.314	2053.361	-0.350	0.000	2207904	96	5
Horizontal_Distance_To_Fire_Points	<input type="checkbox"/>	0	7173	1871.670	1645.756	2708513.539	1.652	2.031	30663572	2	0
WildArea1	<input type="checkbox"/>	0	1	0.297	0.457	0.209	0.890	-1.207	4860	11523	0
WildArea2	<input type="checkbox"/>	0	1	0.030	0.172	0.030	5.465	27.872	499	15884	0
WildArea3	<input type="checkbox"/>	0	1	0.388	0.487	0.237	0.462	-1.787	6349	10034	0
WildArea4	<input type="checkbox"/>	0	1	0.285	0.452	0.204	0.951	-1.096	4675	11708	0
SoilType1	<input type="checkbox"/>	0	1	0.022	0.146	0.021	6.571	41.184	355	16028	0
SoilType2	<input type="checkbox"/>	0	1	0.038	0.191	0.037	4.831	21.343	623	15760	0
SoilType3	<input type="checkbox"/>	0	1	0.059	0.235	0.055	3.754	12.097	962	15421	0

Figure 1: Numeric data on Data explorer.

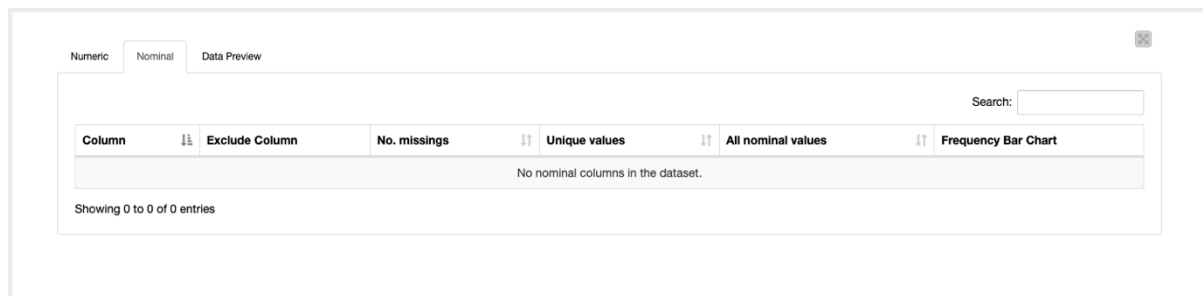


Figure 2: Starting with no defined nominal data.

The screenshot shows the 'Data Preview' tab in the Data Explorer. It displays a table with 8 columns: 'Row ID', 'Elevation', 'Aspect', 'Slope', 'Horizontal\_Distance\_To\_Hydrology', 'Vertical\_Distance\_To\_Hydrology', 'Horizontal\_Distance\_To\_Roadways', and 'Hillshade\_9am'. The table contains 18 rows of data, from Row0 to Row17. At the bottom right, there are buttons for 'Reset', 'Apply', and 'Close'.

Row ID	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hillshade_9am
Row0	2596	51	3	258	0	510	221
Row1	2590	56	2	212	-6	390	220
Row2	2804	139	9	268	65	3180	234
Row3	2785	155	18	242	118	3090	238
Row4	2595	45	2	153	-1	391	220
Row5	2579	132	6	300	-15	67	230
Row6	2606	45	7	270	5	633	222
Row7	2605	49	4	234	7	573	222
Row8	2617	45	9	240	56	666	223
Row9	2612	59	10	247	11	636	228
Row10	2612	201	4	180	51	735	218
Row11	2886	151	11	371	26	5253	234
Row12	2742	134	22	150	69	3215	248
Row13	2609	214	7	150	46	771	213
Row14	2503	157	4	67	4	674	224
Row15	2495	51	7	42	2	752	224
Row16	2610	259	1	120	-1	607	216
Row17	2517	72	7	85	6	585	228

Figure 3: Data Preview on the Data explorer node.

## Data Preparation:

We filtered out 8 rows that had missing values for the CoverType target variable, as we do not want to impute or infer those target values. For the remaining columns, we replaced any missing values with the median value for that column. This allows us to retain more rows for modeling without skewing the data.

To make the cover type more interpretable, we converted the integer target variable to a categorical string variable with 7 types. We also added a colored layer column indicating the cover type with a distinct color for each of the 7 types. This helps visually distinguish between the cover types, as seen in Figure 4.

The multiple binary columns for wilderness area and soil type were concatenated into a single column for each using the Many to One node in Knime. This appended the binary columns into a single categorical variable containing all the wilderness areas and soil types respectively. This simplifies modeling with a single variable for wilderness area and soil type rather than multiple binary columns.

The data is now ready for modeling with filtering of missing data, imputation, categorical encoding, and concatenating the binary variables into single columns. The next step will be model building and evaluation.

File Edit Hilitte Navigation View

Table "default" - Rows: 16375 Spec - Columns: 13 Properties Flow Variables

Row ID	D Elevati...	D Aspect	D Slope	D Horiz...	D Vertic...	D Horiz...	D Hillsh...	D Hillsh...	D Horiz...	S Cover...	S WildAr...	S SoilType
Row16	2,610	259	1	120	-1	607	216	239	161	6,096	5	WildArea1 SoilType29
Row17	2,517	72	7	85	6	595	228	227	133	5,607	5	WildArea1 SoilType18
Row18	2,504	0	4	95	5	691	214	232	156	5,572	5	WildArea1 SoilType18
Row19	2,503	38	5	85	10	741	220	228	144	5,555	5	WildArea1 SoilType18
Row20	2,501	71	9	60	8	767	230	223	126	5,547	5	WildArea1 SoilType18
Row21	2,880	209	17	216	30	4,986	206	253	179	4,323	2	WildArea1 SoilType30
Row22	2,768	114	23	192	82	3,339	252	209	71	5,972	5	WildArea1 SoilType30
Row23	2,511	54	8	124	0	638	225	222	130	5,569	5	WildArea1 SoilType18
Row24	2,507	22	9	120	14	732	215	221	143	5,534	5	WildArea1 SoilType18
Row25	2,492	135	6	0	0	860	229	237	142	5,494	5	WildArea1 SoilType18
Row26	2,489	163	10	30	-4	849	230	243	145	5,486	5	WildArea1 SoilType18
Row27	2,962	148	16	323	23	5,916	240	236	120	3,395	2	WildArea1 SoilType29
Row28	2,811	135	1	212	30	3,670	220	238	154	5,643	2	WildArea1 SoilType12
Row29	2,739	117	24	127	53	3,281	253	210	71	6,033	5	WildArea1 SoilType30
Row30	2,703	122	30	67	27	3,191	254	201	52	6,123	5	WildArea1 SoilType30
Row31	2,522	105	7	120	1	595	233	231	130	5,569	5	WildArea1 SoilType18
Row32	2,519	102	6	124	4	616	230	233	137	5,559	5	WildArea1 SoilType18
Row33	2,516	23	6	150	4	658	216	227	147	5,541	5	WildArea1 SoilType18
Row34	2,515	41	9	162	4	680	221	220	133	5,532	5	WildArea1 SoilType18
Row35	2,900	45	19	242	20	5,199	221	195	100	4,115	2	WildArea1 SoilType29
Row36	2,709	125	28	67	23	3,224	253	207	61	6,094	5	WildArea1 SoilType30
Row37	2,511	92	7	182	18	722	231	229	131	5,494	5	WildArea1 SoilType18
Row38	2,749	98	30	124	53	3,316	252	183	36	6,005	5	WildArea1 SoilType30
Row39	2,686	354	12	0	0	3,167	200	219	157	6,155	5	WildArea1 SoilType30
Row40	2,699	347	3	0	0	2,096	213	234	159	6,853	1	WildArea1 SoilType20
Row41	2,570	346	2	0	0	331	215	235	158	5,745	2	WildArea1 SoilType29
Row42	2,533	71	9	150	-3	577	230	223	126	5,552	5	WildArea1 SoilType18
Row43	2,703	330	27	30	17	3,141	146	197	184	6,186	5	WildArea1 SoilType16
Row44	2,678	128	5	95	23	1,660	229	236	141	6,546	2	WildArea1 SoilType12
Row45	2,529	68	8	210	-5	666	228	225	130	5,484	5	WildArea1 SoilType18
Row46	2,524	94	7	212	-4	684	232	229	130	5,474	5	WildArea1 SoilType18
Row47	2,536	99	6	234	0	659	230	232	136	5,475	5	WildArea1 SoilType18
Row48	2,498	66	6	95	7	900	227	227	135	5,357	5	WildArea1 SoilType18
Row49	2,489	100	7	85	13	810	232	231	131	5,334	5	WildArea1 SoilType18
Row50	2,713	117	30	60	17	3,297	254	198	48	6,039	5	WildArea1 SoilType30
Row51	2,739	323	25	85	43	3,118	149	205	192	6,219	1	WildArea1 SoilType29
Row52	2,696	72	2	30	0	3,271	222	234	149	6,071	1	WildArea1 SoilType30
Row53	2,510	79	14	192	19	891	237	215	106	5,325	5	WildArea1 SoilType18
Row54	2,502	81	7	175	11	912	230	227	129	5,316	5	WildArea1 SoilType18
Row55	2,722	315	24	30	19	3,216	148	212	200	6,132	1	WildArea1 SoilType16
Row56	2,500	74	11	190	9	930	233	219	116	5,279	5	WildArea1 SoilType18
Row57	2,486	68	5	180	-4	870	225	230	139	5,262	5	WildArea1 SoilType18
Row58	2,489	11	4	175	13	840	216	232	153	5,254	5	WildArea1 SoilType18
Row59	2,489	42	6	162	13	810	221	227	141	5,247	5	WildArea1 SoilType18
Row60	2,490	75	5	134	17	810	227	230	137	5,218	5	WildArea1 SoilType18
Row61	2,952	107	11	42	7	5,845	239	226	116	3,509	2	WildArea1 SoilType29
Row62	2,705	90	8	134	22	2,023	232	228	129	6,615	2	WildArea1 SoilType12
Row63	2,507	40	7	153	10	930	221	224	138	5,221	5	WildArea1 SoilType18

Figure 4: Colored Rows based on CoverType values.

Numeric Nominal Data Preview

Search:

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
CoverType	<input type="checkbox"/>	0	7	2, 1, 5, 3, 4, 6, 7	
WildAreaType	<input type="checkbox"/>	0	4	WildArea3, WildArea1, WildArea4, WildArea2	
SoilType	<input type="checkbox"/>	0	38	SoilType10, SoilType29, SoilType30, SoilType3, [...], SoilType9, SoilType36, SoilType28, SoilType25, SoilType8	

Showing 1 to 3 of 3 entries

Figure 5: Nominal data.

## Model Building and Testing:

### A- Decision Tree Model:

After understanding the business problem, exploring and transforming the data, according to the CRISP-DM methodology we can continue with the fourth step which is working on building our models. We start with partitioning the data into two groups, one for training which is 70% of the whole dataset, and a testing set of 30%, stratified on CoverType column, while using a random seed value of 12345.

In the next step, we built a Decision Tree model, having the target variable being the column named CoverType. To evaluate the model and check its accuracy, Knime offers various tools, in this analysis we can use a tool called Scorer and also plotting it on the ROC curve. Based on the variable importance scores from the decision tree model, the most predictive variables for distinguishing between the seven forest cover types appear to be soil type and elevation. Figure 7 shows the plotted ROC curve derived from this decision tree model. Lastly, the Decision Tree graphical model of the first three levels is shown in Figure 8.

CoverType...	5	2	1	7	3
5	530	75	16	0	25
2	79	681	164	2	15
1	16	142	462	67	1
7	0	9	75	563	1
3	13	13	0	0	438
6	10	12	0	0	156
4	0	0	0	0	37

Correct classified: 3,714	Wrong classified: 1,198
Accuracy: 75.611%	Error: 24.389%
Cohen's kappa (κ): 0.714%	

Figure 6: Decision Tree Scorer.

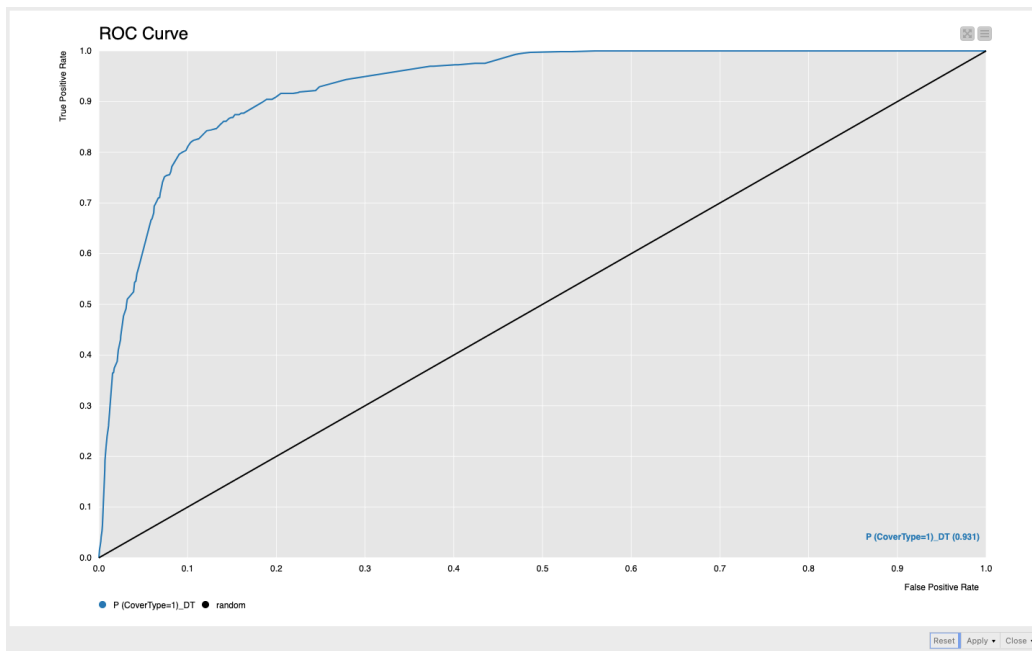


Figure 7: Decision Tree ROC Curve.

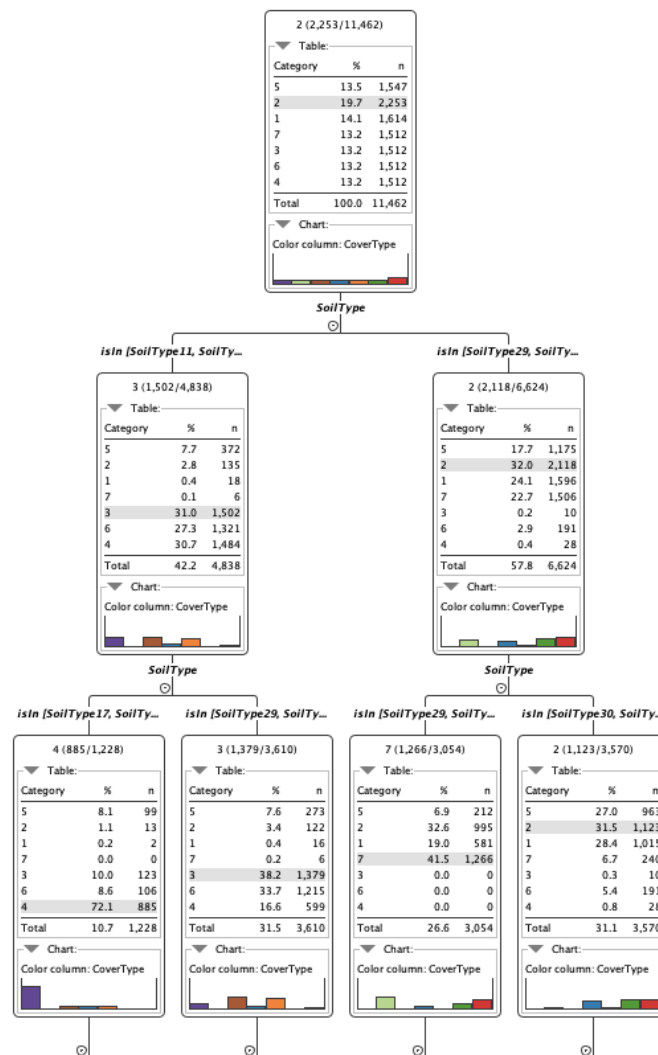
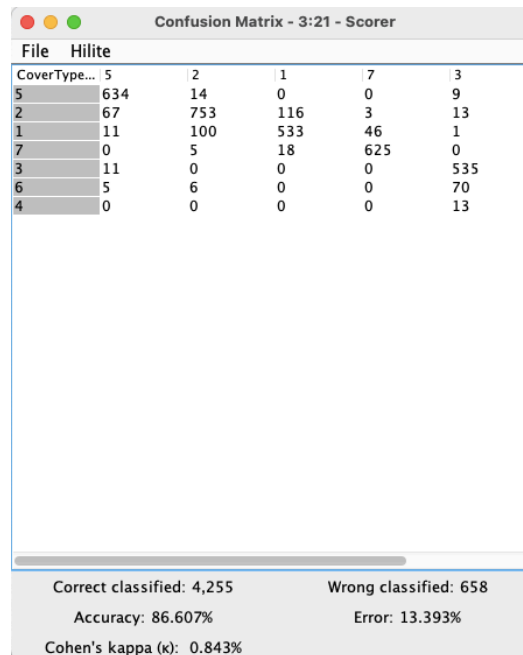


Figure 8: Three levels Decision Tree.

## B- Random Forest Model:

Similar to how we created the Decision Tree model, we can use the exact same settings we used to create the Random Forest model to be able to compare the accuracy that each model produce. And as the fifth step of the CRISP-DM methodology focuses on testing and evaluating the output of the newly built models, we will use the two tools we used previously, the Scorer and ROC Curve, displayed in Figure 9 and 10 respectively.



A screenshot of a software window titled "Confusion Matrix - 3:21 - Scorer". It displays a confusion matrix for a classification task with 5 classes. The matrix is a 5x5 grid where the first column represents the actual classes (CoverType) and the subsequent columns represent the predicted classes (Hilite). The values in the matrix are: Row 1: 634, 14, 0, 0, 9; Row 2: 67, 753, 116, 3, 13; Row 3: 11, 100, 533, 46, 1; Row 4: 0, 5, 18, 625, 0; Row 5: 11, 0, 0, 0, 535. Below the matrix, summary statistics are provided: Correct classified: 4,255; Wrong classified: 658; Accuracy: 86.607%; Error: 13.393%; Cohen's kappa (κ): 0.843%.

File	Hilite	5	2	1	7	3
CoverType...	5	634	14	0	0	9
5	2	67	753	116	3	13
2	1	11	100	533	46	1
1	7	0	5	18	625	0
7	3	11	0	0	0	535
3	6	5	6	0	0	70
6	4	0	0	0	0	13

Correct classified: 4,255      Wrong classified: 658  
Accuracy: 86.607%      Error: 13.393%  
Cohen's kappa (κ): 0.843%

Figure 9: Random Forest Scorer.

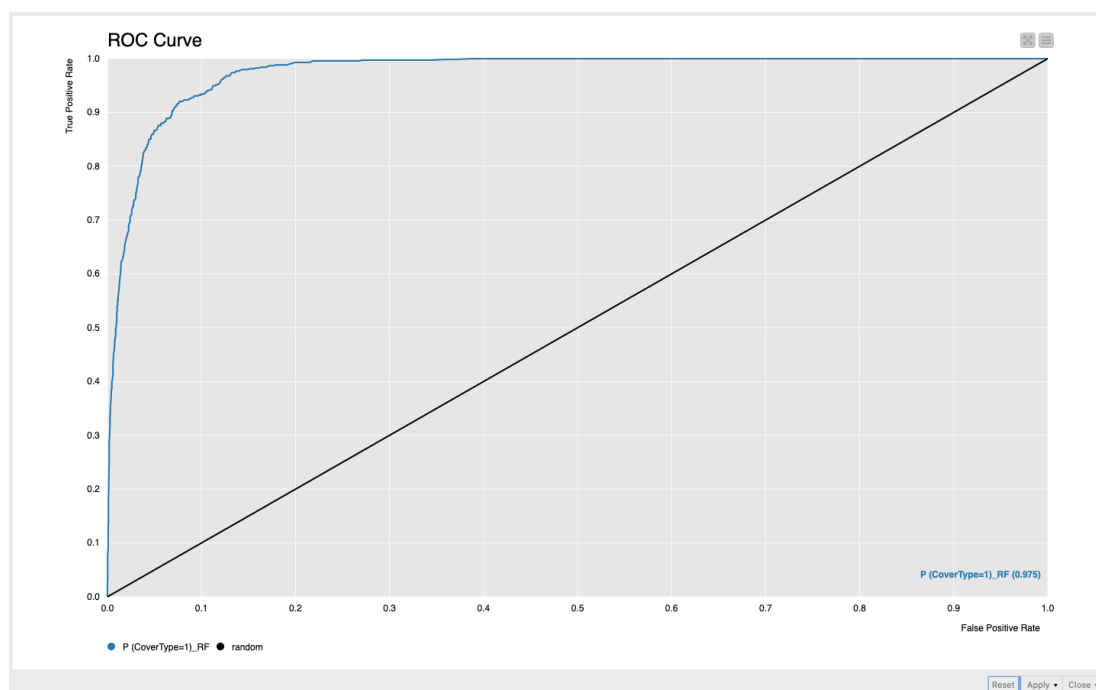
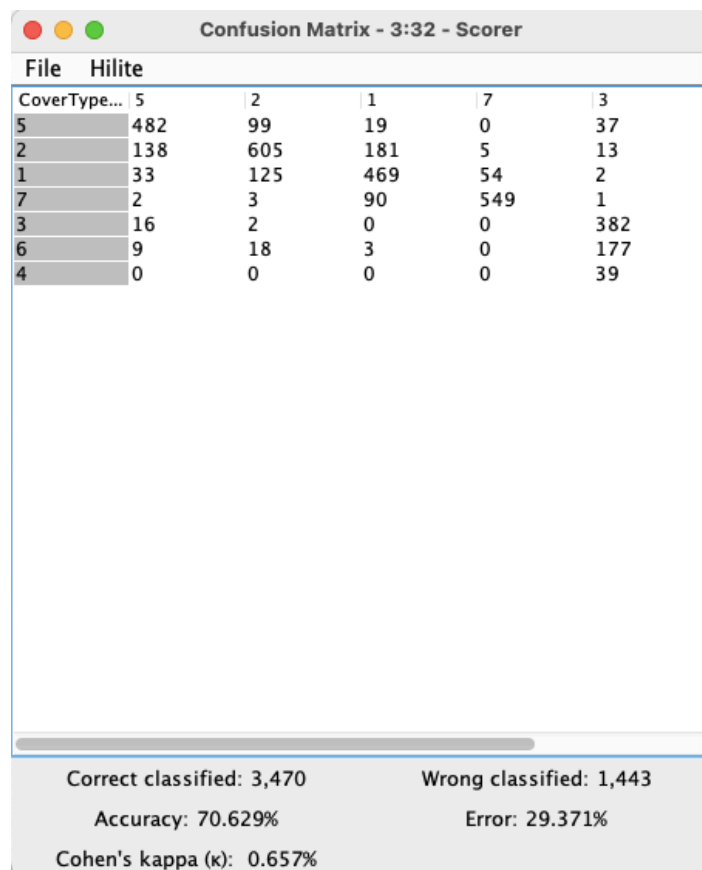


Figure 10: Random Forest ROC Curve.

### C- Neural Network (MLP type) Model:

The third model utilizes a neural network (MLP), where the data passes through a normalization node to standardize it in a range between 0 and 1 before feeding it into the machine learning algorithm. After preprocessing, the data goes into partition and MLP learner nodes, which have the CoverType column set as the target variable. Accuracy metrics are generated by ROC Curve and Scorer nodes in KNIME. The full workflow with intermediate outputs is visualized in the accompanying figures.



File	Hilite
CoverType...	5
5	482
2	138
1	33
7	2
3	16
6	9
4	0

5	482	99	19	0	37
2	138	605	181	5	13
1	33	125	469	54	2
7	2	3	90	549	1
3	16	2	0	0	382
6	9	18	3	0	177
4	0	0	0	0	39

Correct classified: 3,470

Wrong classified: 1,443

Accuracy: 70.629%

Error: 29.371%

Cohen's kappa (κ): 0.657%

Figure 11: Neural Network Scorer.



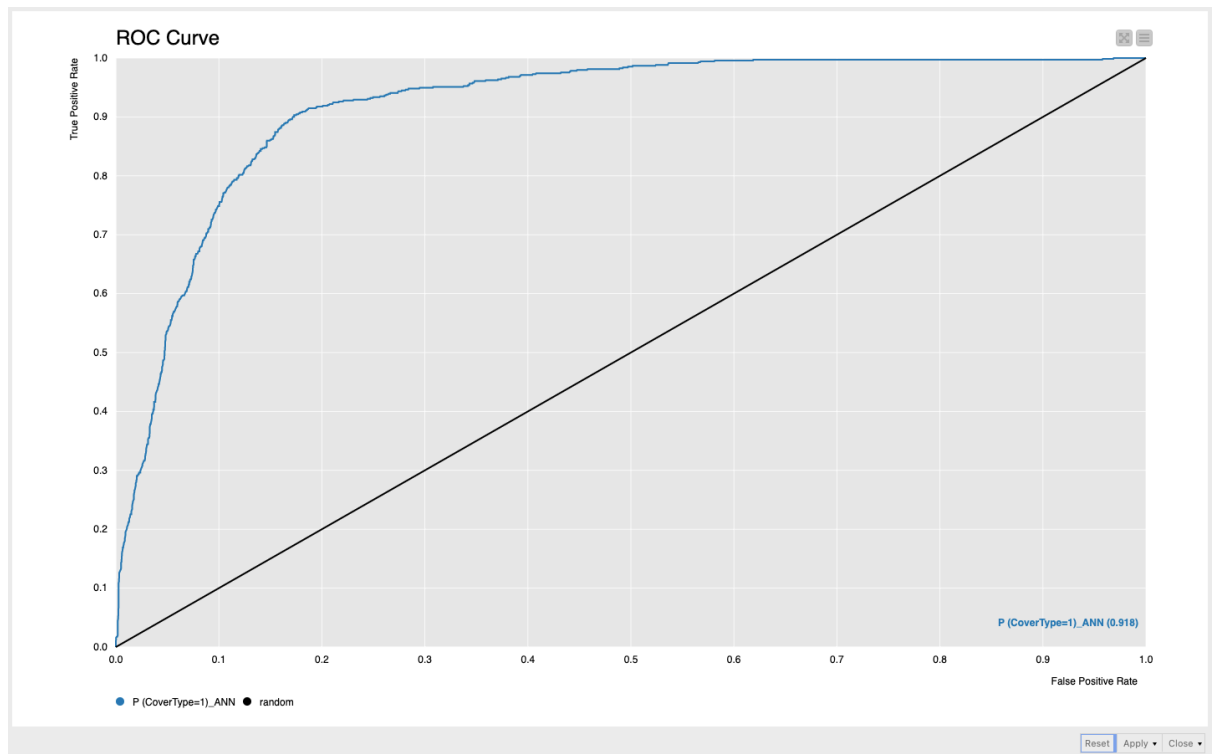


Figure 12: Neural Network ROC Curve.

#### D- Logistic Regression (LR) Model:

The final model employs logistic regression (LR), structured in KNIME with a series of nodes for data preprocessing before the machine learning algorithm. As in the prior neural network model, the data gets normalized between 0 and 1 first. The processed data is partitioned and fed into the LR learner, with the CoverType set again as target output variable. Model accuracy is evaluated by routing the LP model predictions into ROC Curve and Scorer nodes native to KNIME. These provide numeric metrics and visual evaluation of how well the cover type is predicted. The complete workflow - data transformations, LP model building, accuracy checking - is visualized in the accompanying figures within the KNIME interface.

Confusion Matrix - 3:41 - Scorer					
File	Hilite				
CoverType...	5	2	1	7	3
5	513	67	19	0	21
2	174	561	193	9	6
1	35	95	484	74	1
7	1	1	75	571	0
3	32	1	0	0	354
6	19	3	0	0	147
4	0	0	0	0	38

Correct classified: 3,482      Wrong classified: 1,431  
 Accuracy: 70.873%      Error: 29.127%  
 Cohen's kappa ( $\kappa$ ): 0.66%

Figure 13: Logistic Regression Scorer.

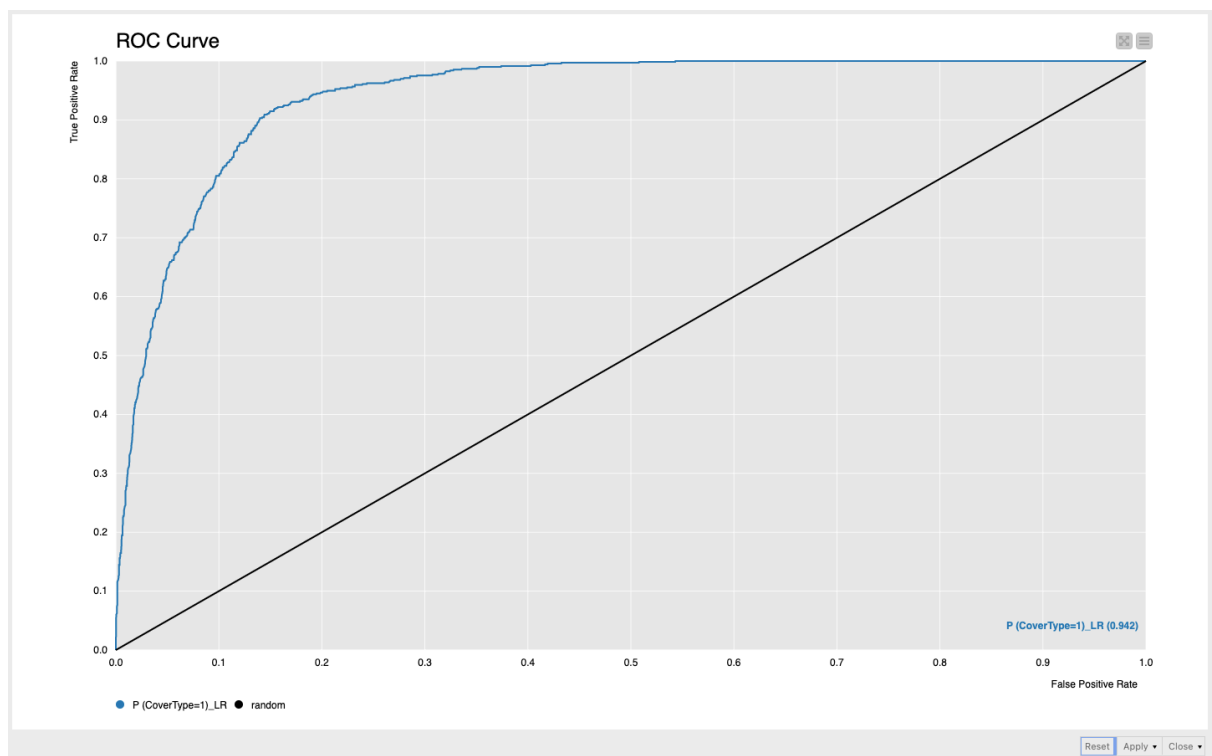


Figure 14: Logistic Regression ROC Curve.

### Testing and Evaluation:

To compare the performances of all the models we used in this report, Decision Tree, Random Forest, Neural Network, and Logistic Regression, we created an ROC curve that combines the output of each into one chart, Figure 15. The Random Forest model performed better than the other models with 0.975 accuracy. On the other hand, the model that yielded the lowest prediction results is Neural Network (MLP) with 0.918 accuracy. Therefore, the Random Forest model is preferred to the other models we built for having more accuracy and less error rate.

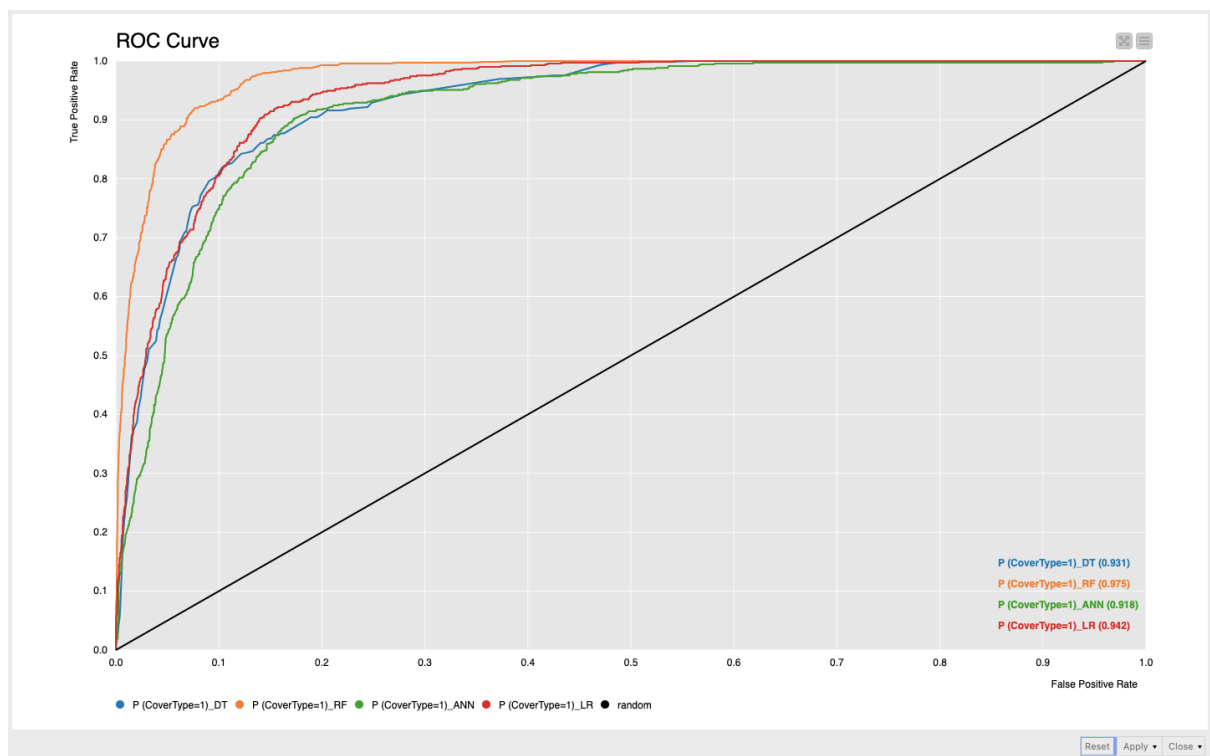


Figure 15: ROC Curve combining all the models.

### Deployment:

In the sixth and final step of the CRISP-DM methodology, deployment of the superior model can take place to support decision making processes. In the graph below, Figure 16 shows the Variable Importance chart which shows how much the model is utilizing each variable to make accurate predictions. In our Random Forest model used in this analysis, soil type was the most important, followed by elevation and wild area type, while the slope, hill shade and aspect had the least influence.

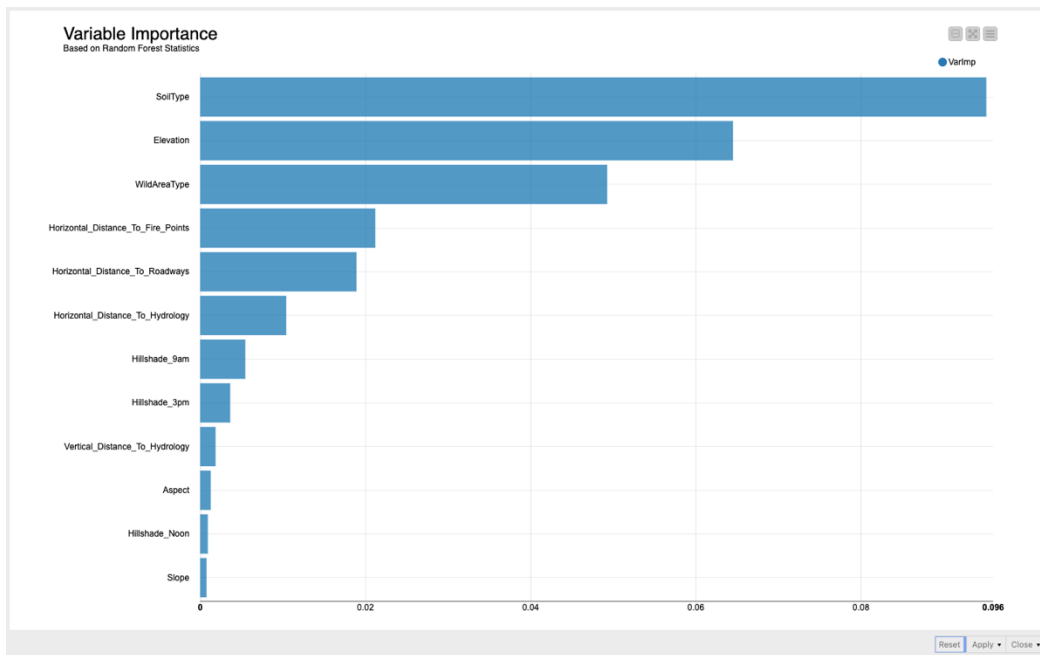


Figure 16: Random Forest Importance Graph

In conclusion, this report utilized the CRISP-DM methodology to develop and evaluate models for predicting forest cover type. Four models were built and tested: Decision Tree, Random Forest, Neural Network, and Logistic Regression. The process followed the phases of CRISP-DM - establishing the business need, understanding the data, preparing the data, training and tuning models, and evaluating performance.

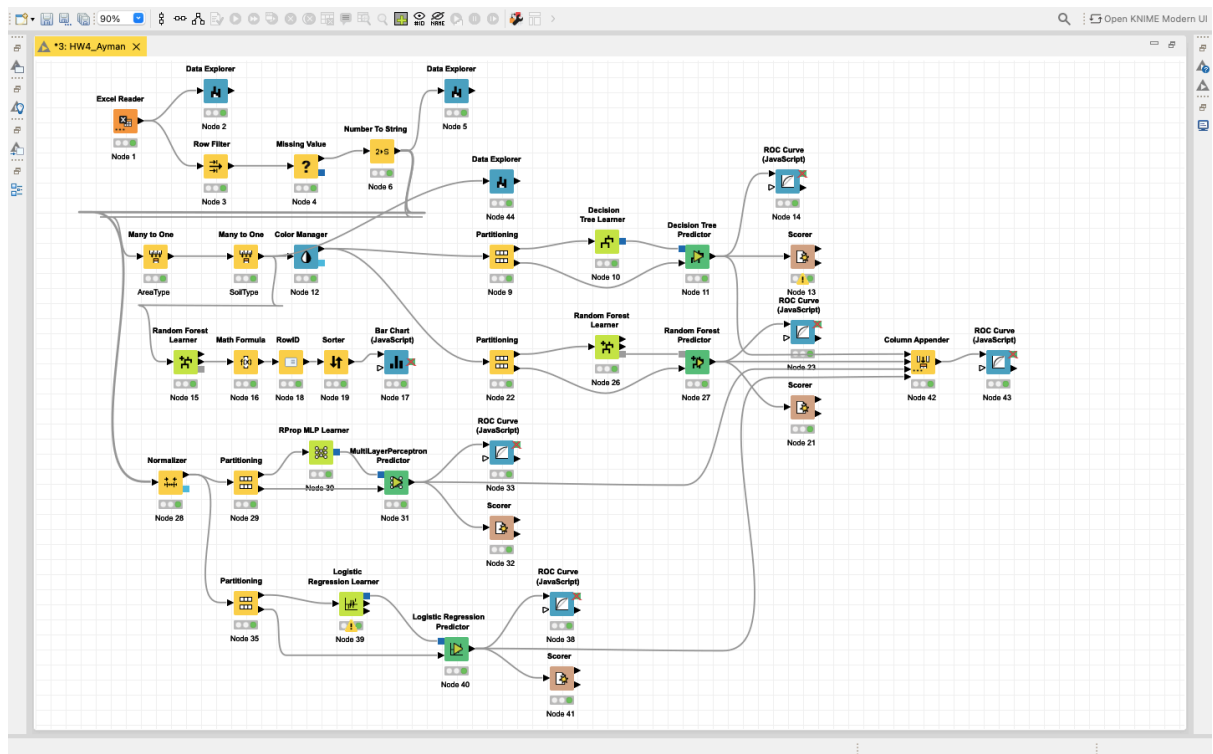


Figure 17: Knime Workflow.