EXAMPLE &TUTORIAL

HOW TO USE «***SEMI-AUTOMATIC FEATURE ENGINEERING TOOL***» FOR PRODUCING LSM

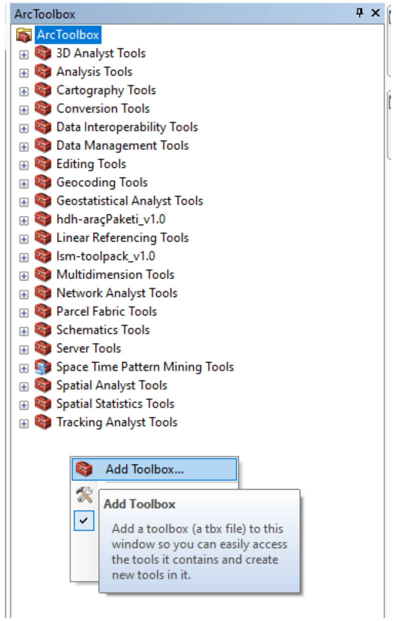
Requirements

- ArcGIS 10.3.1 or later or ArcGIS Pro 1.1 or later (don't have it? try trial edition)
- R Statistical Computing Software, 3.3.2 or later (What is R?). If you're experiencing issues when using a new version of R (i.e.g, R 4.0 and upper) Recommended Version v3.6.3
- ArcGIS R-Bridge -- Recommended Version v1.0.1.239
- Java Runtime Environment for FSelector package
- If you are setting up modules for the first time, you will need an internet connection to install the R-packages on the depository.

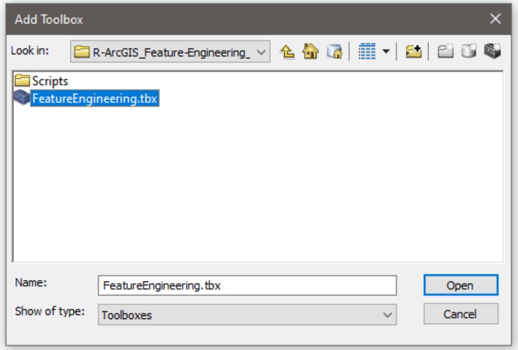
Installation

- Download this repository and unzip ‘R-ArcGIS_Feature-Engineering_ToolPack-main.zip’ into the folder C:*targetfolder*.
  - https://github.com/emrehanks/R-ArcGIS_Feature-Engineering_ToolPack
- Open your ArcGIS application:

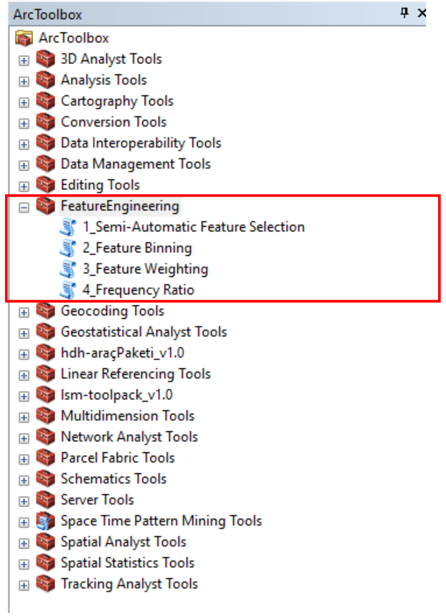
1. Navigate ArcToolBox panel



2. Add Toolbox on ArcMap

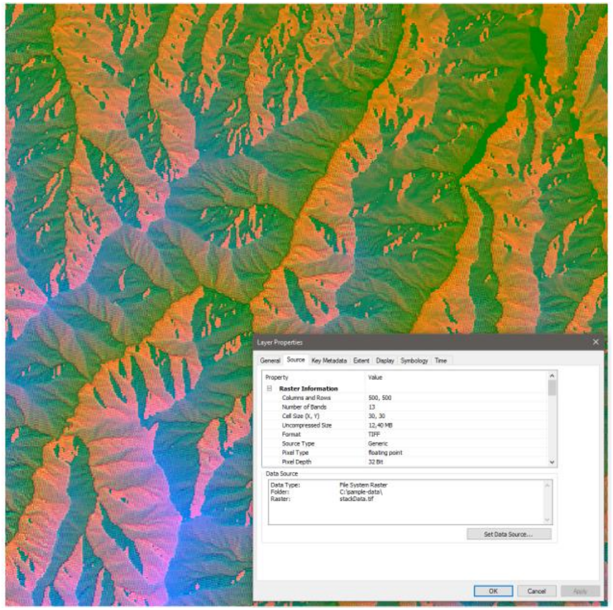


3. Be sure to toolbox added on ArcToolBox panel.

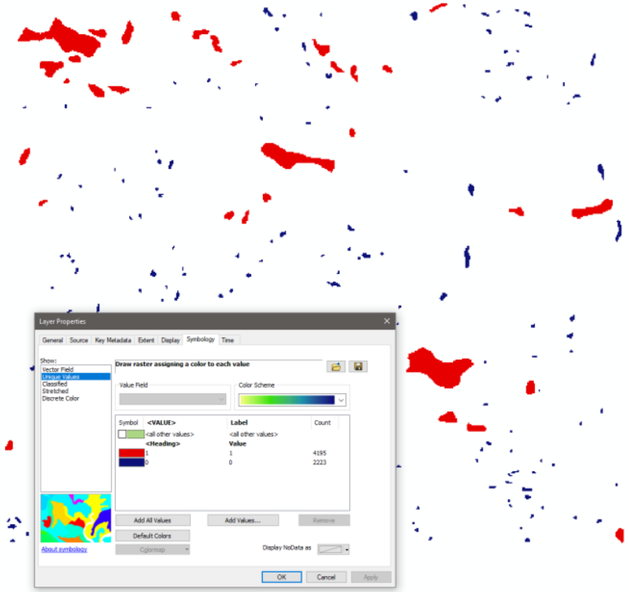


Sample Data

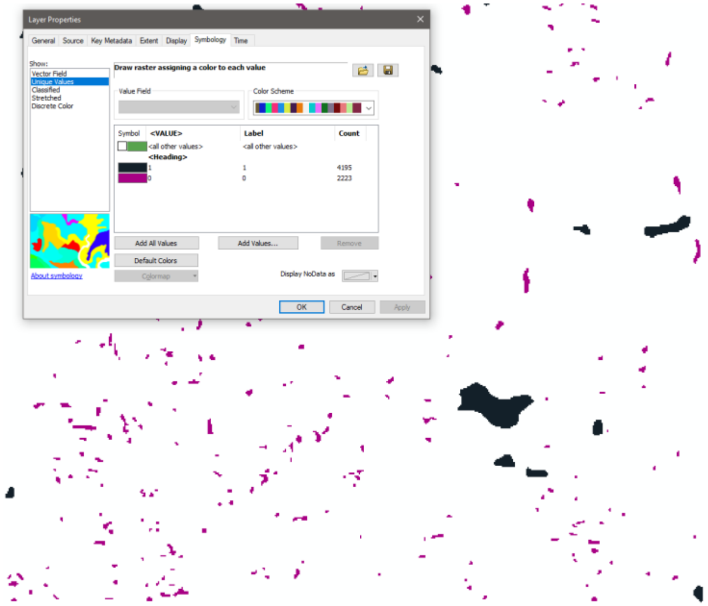
- Download and and unzip ‘sample-data.zip’ into the folder C:*targetfolder*.
- This data contains:



stackData.tif -> Multi-bands geographic image (13 bands, 500*500 and 30m pixel size)
Not: band or factor names was included in this folder as name as «stackNames.csv»



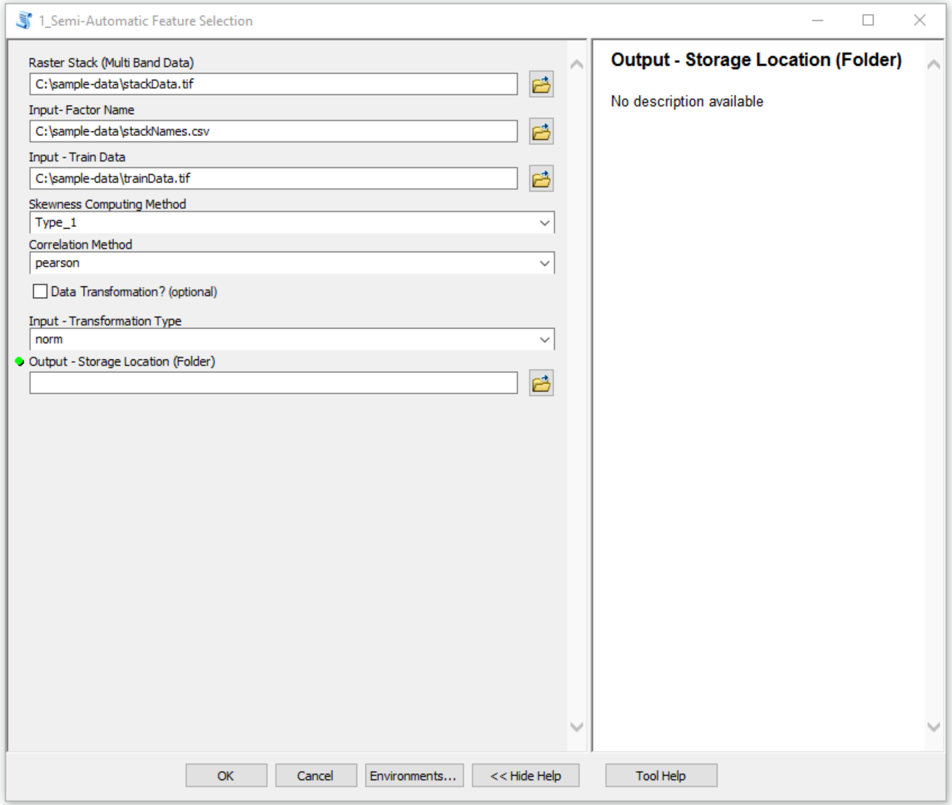
trainData.tif-> Single-bands geographic image (30m pixel size)

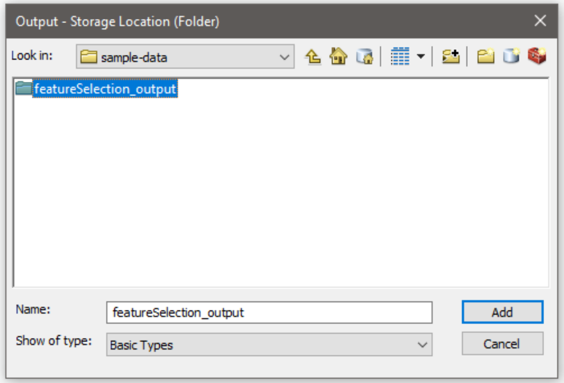


validationData.tif-> Single-bands geographic image (30m pixel size)

1_Semi-Automatic Feature Selection

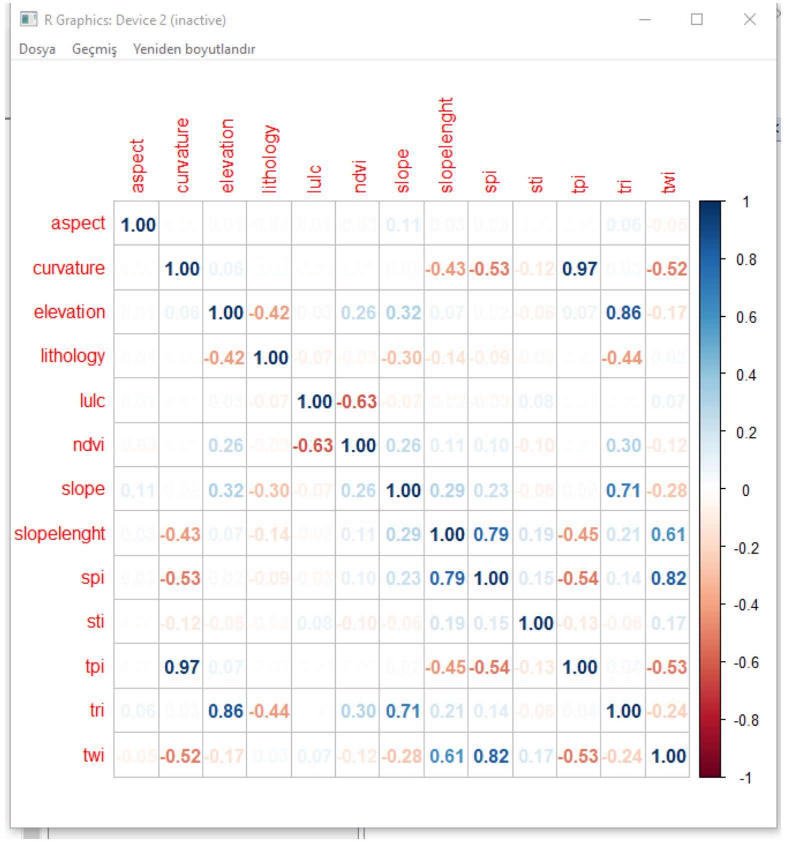
1. Import stackData.tif
2. If available, import stackNames.csv.
3. Import train image file **Not:** Landslide area must be labeled as'1' and non-landslide areas must be labeled as '0'.
4. Select type of skewness function.
5. Select type of Correlation method
6. Give the output location.
7. Select a folder and click Add
8. Click «OK»

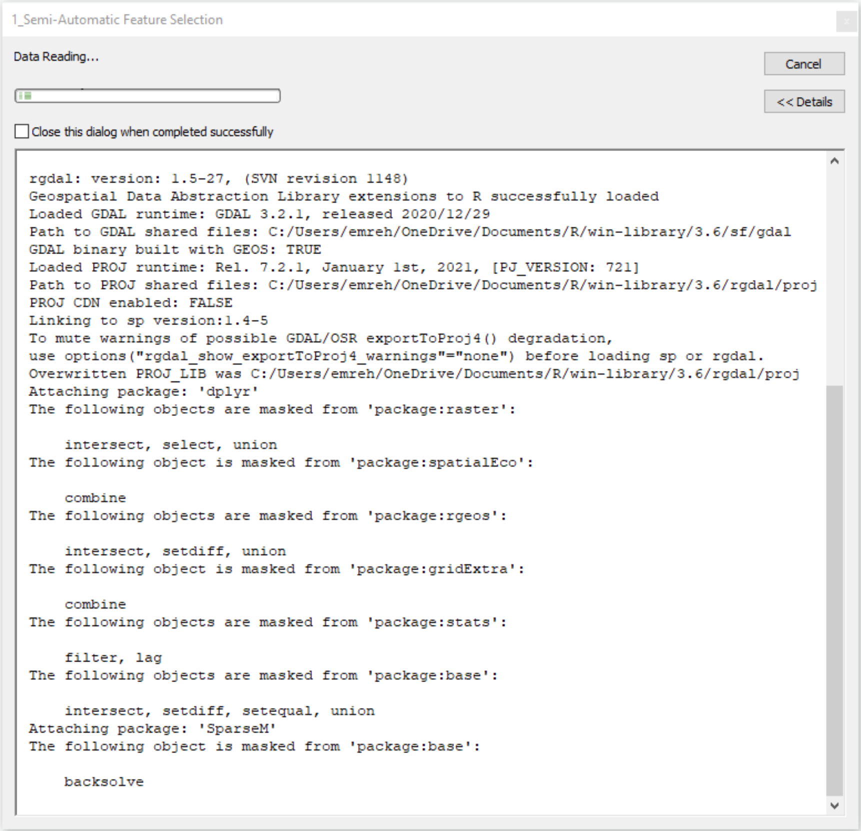


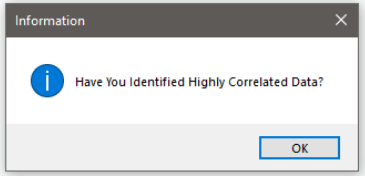


1_Semi-Automatic Feature Selection

- Note that the data with correlation and click Ok on information window
- In this tutorial; The data namely **curvature, tri, and spi** were identified as the most correlated features.

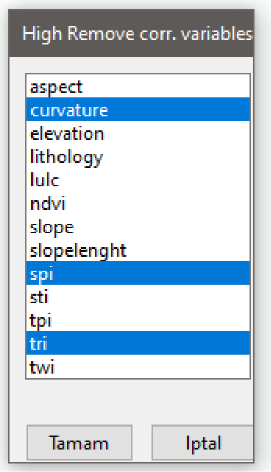




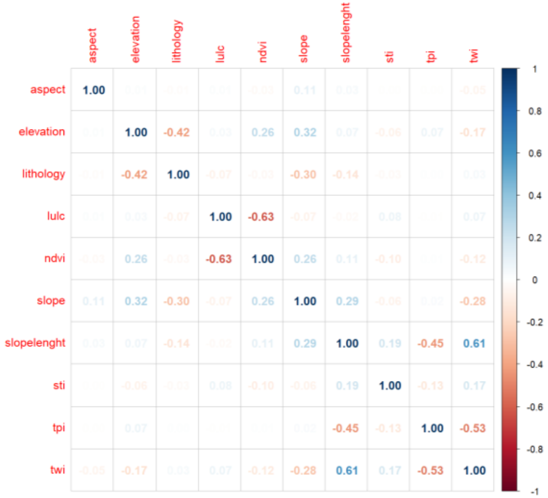


1_Semi-Automatic Feature Selection

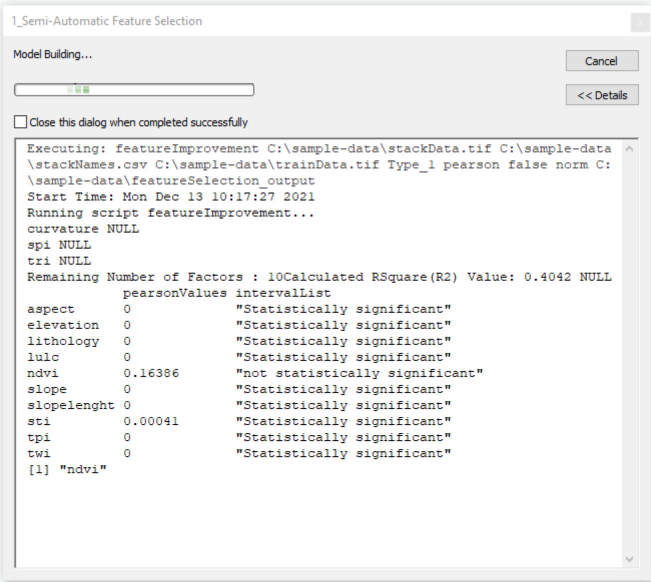
1. You can select multiple features by «ctrl and click».

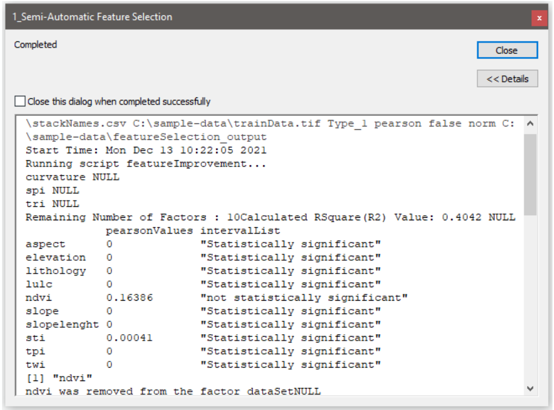


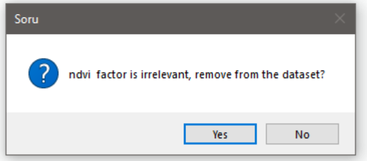
2. The second information graph about correlation statistic was shown as below:



3. Pearson Correlation reported that the NDVI feature is not statistically significant. The tool is awaiting confirmation from us whether to delete this factor or not.

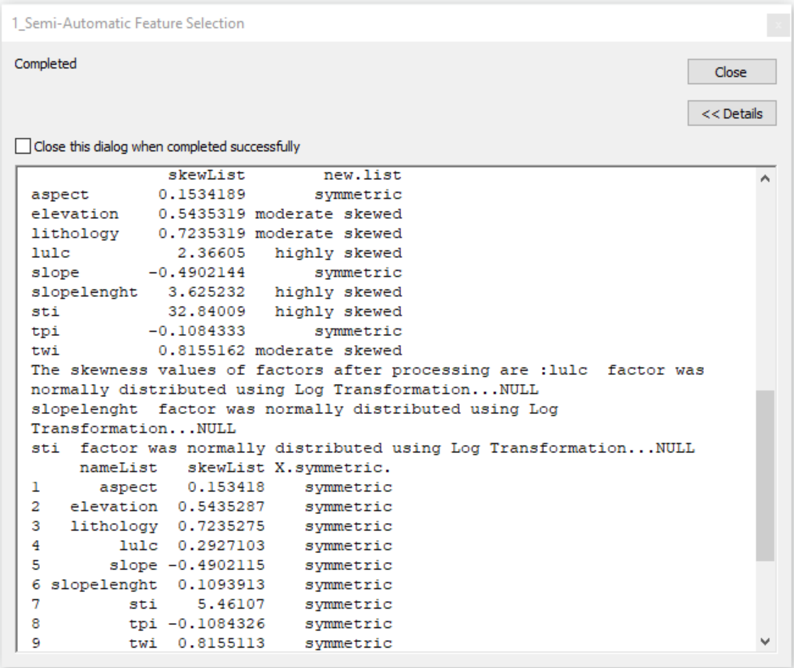




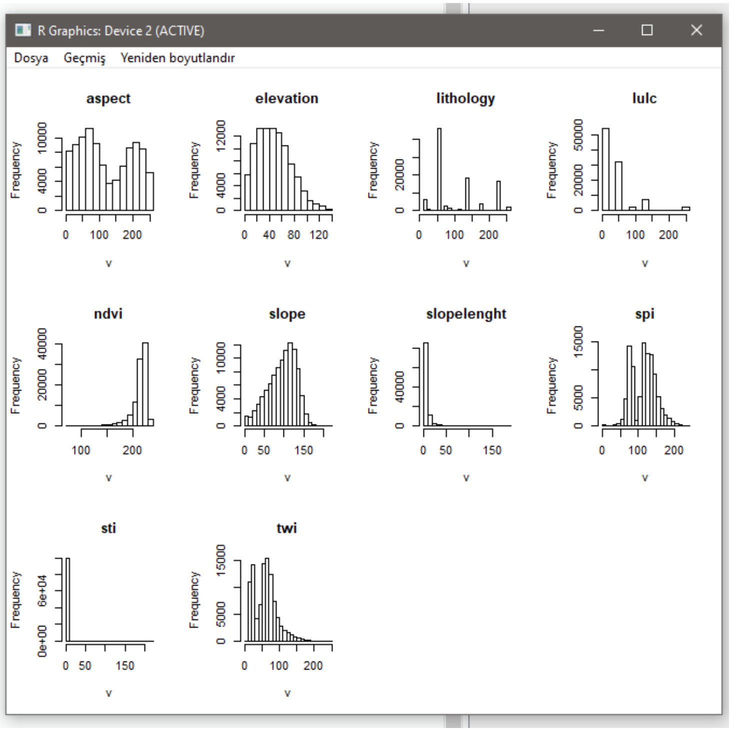


1_Semi-Automatic Feature Selection

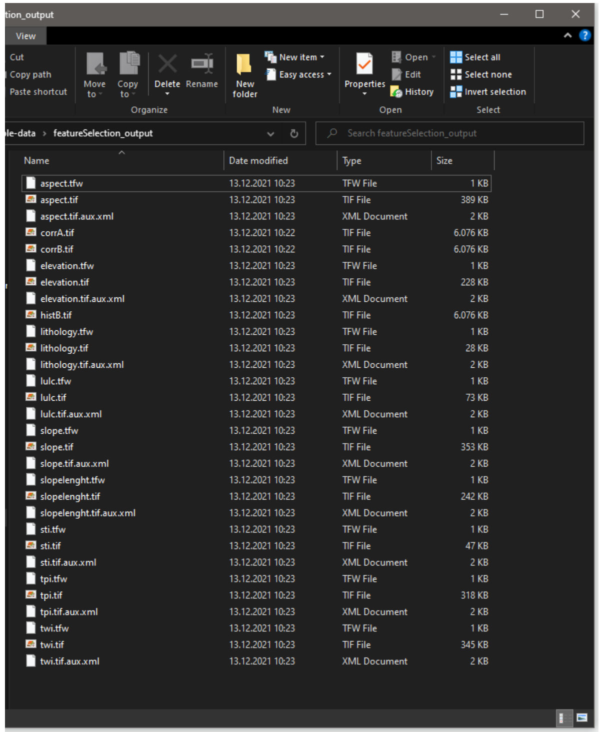
1. The before and after analyzed of the skewness score report is also shown in the dialog panel: **Not**: *Do not fill the check bar (i.e., Close this dialog when completed successfully) if you want to note and keep all scores for future use.*



2. Histogram plots of features after skewness computing method processing.

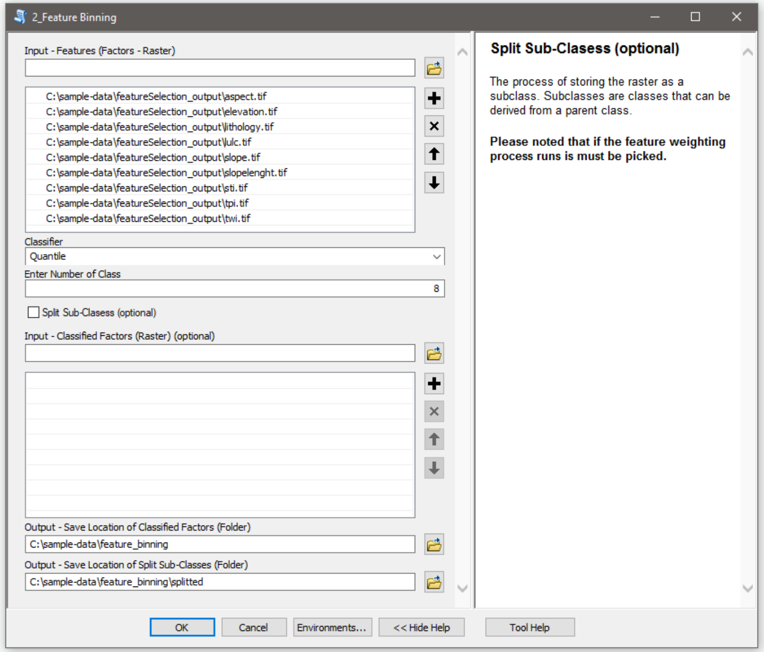


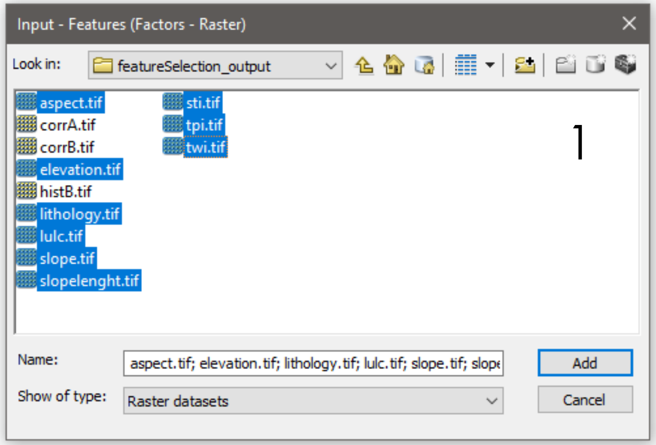
2. All feature selection processes successfully completed. Users also can access all process output from the selected folder.

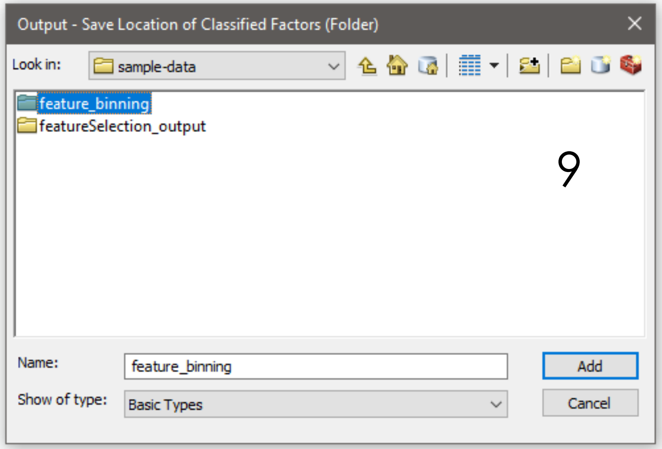


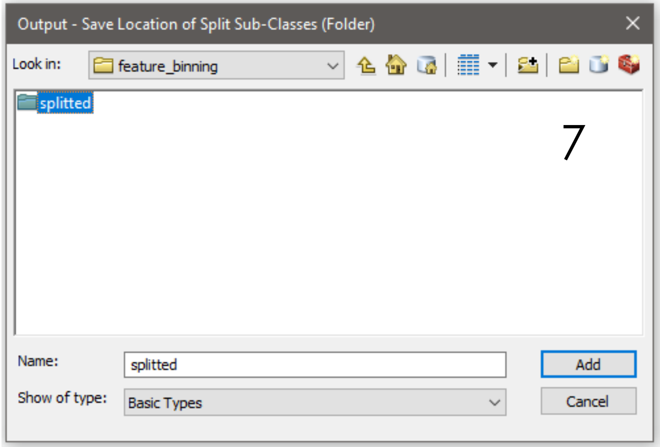
2_Feature Binning

1. Import only countinuos feature in (Factors-Raster) tab.
2. Select Classifier method.
3. Enter number of class
4. If the user wants to process the feature weighting for future work, the pick as "**Split Sub-Classes**" must be approved
5. Import only features that categorical data type such as lithology and LULC. ***However, since both lithology and lulc log are transformed by the Log Transformation process in this study, therefore both data should be selected in the first input area.***
6. Give the output location for the classified features.
7. Select a folder and click Add
8. Give the output location for split features
9. Select a folder and click Add
10. Click «OK»



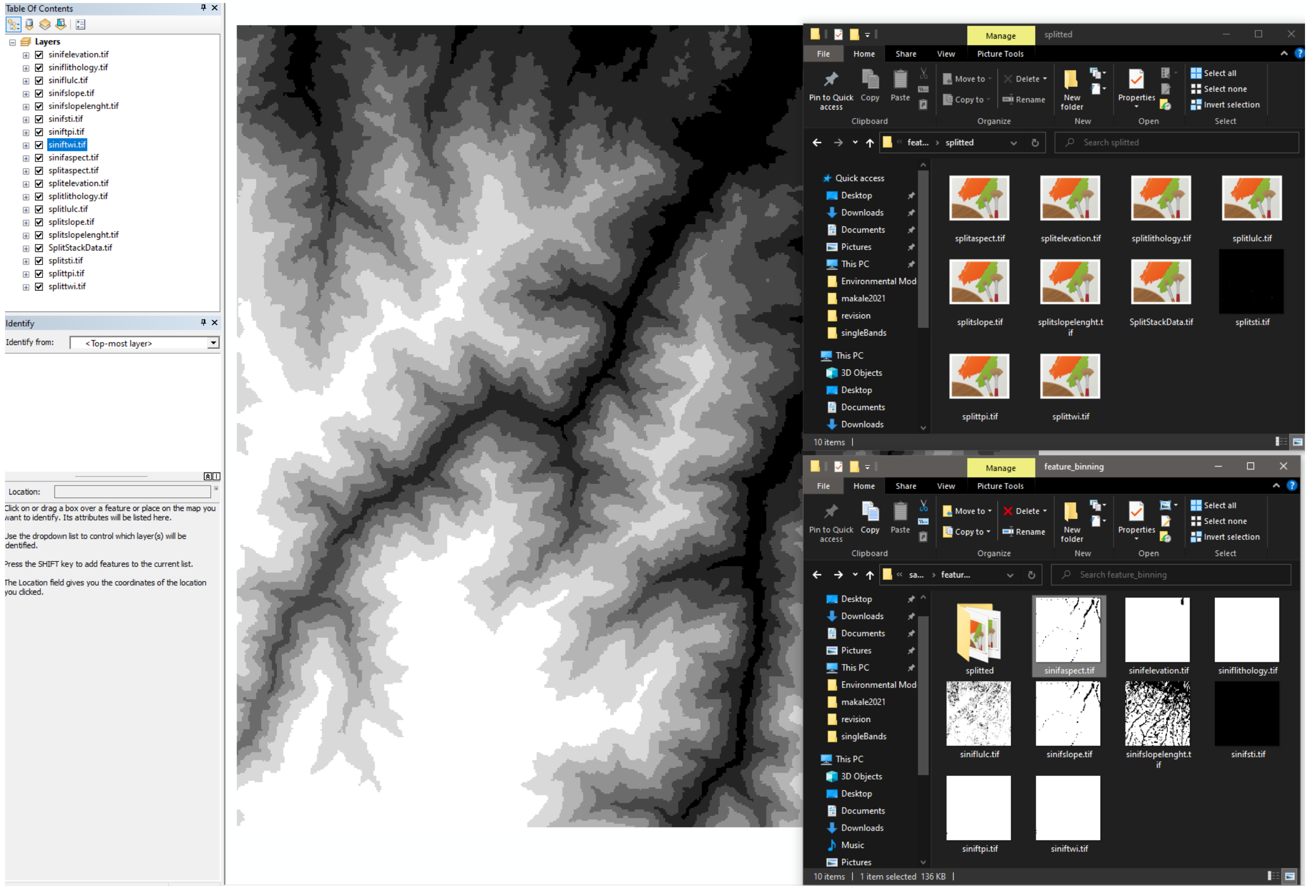






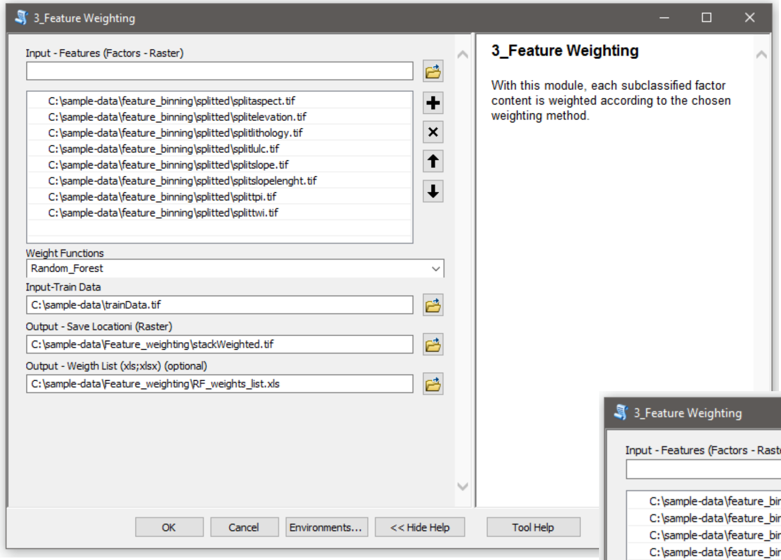
2_Feature Binning

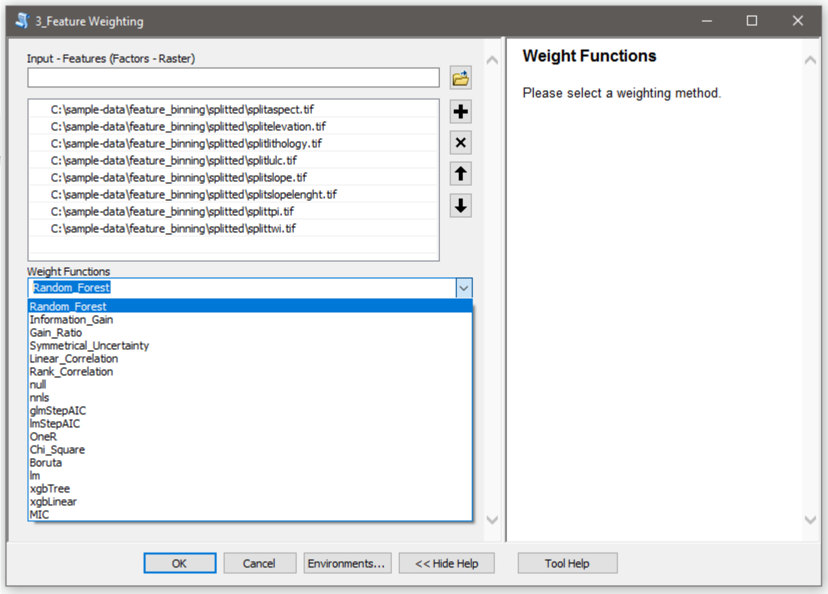
- Feature Binning process is completed.
- Classified and split features are stored in the selected folder after the process.



3_Feature Weighting

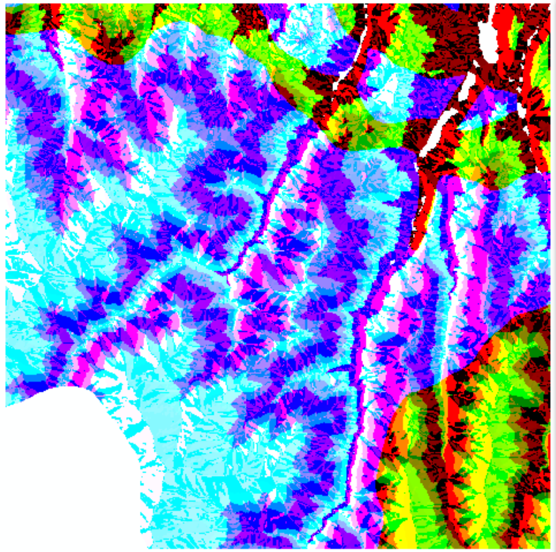
1. Import only split feature in (Factors-Raster) tab. Not: For this study data, the SPI factor was eliminated because of some problem about spi data type. Therefore, only 8 factors were selected for the weighing process.
2. Select Weight method. Suer selected several method for this process. However, not guaranteed to work every method including in this tab. **Recommended to user find the most suitable method by trial and error.**
3. Import train data.
4. Give the output location for the weighted features as a raster stack data.
5. Select a folder and give raster data and clik OK
6. Give the output location for weight score of features
7. Select a folder and give XLS or XLSX document name
8. Click «OK»

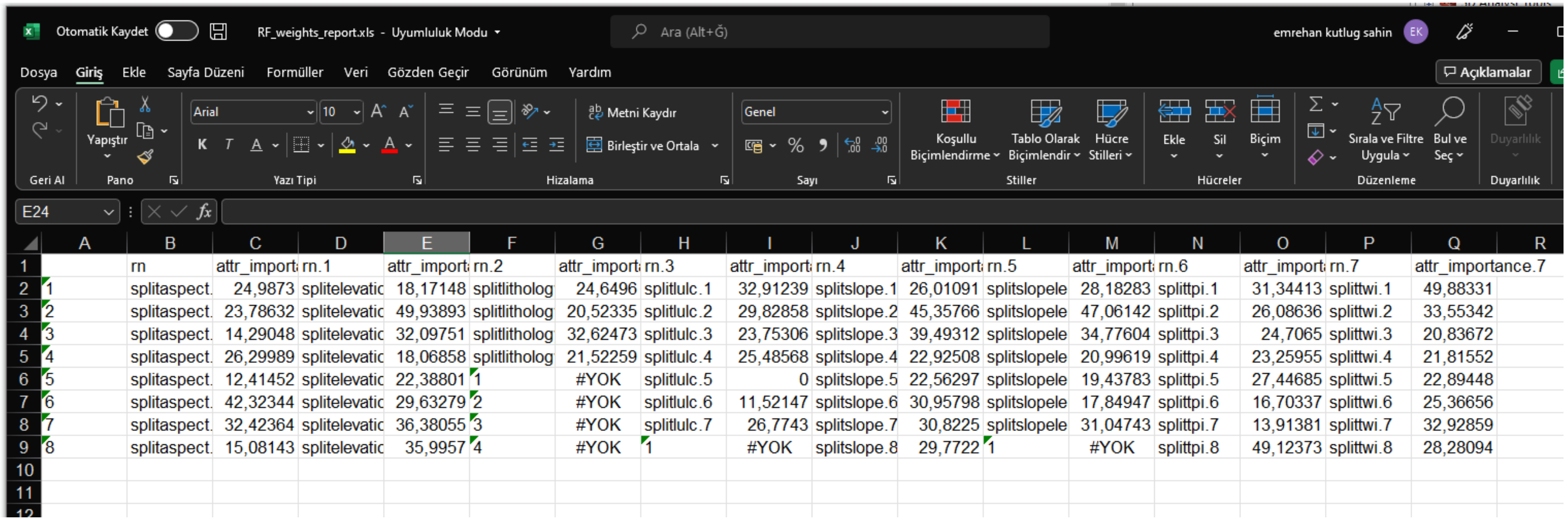




3_Feature Weighting

- Raster stack data and lists of weight scores are located in the selected folder.
- Finally, the necessary raster stack data was obtained to generate the Landslide Susceptibility Mapping.

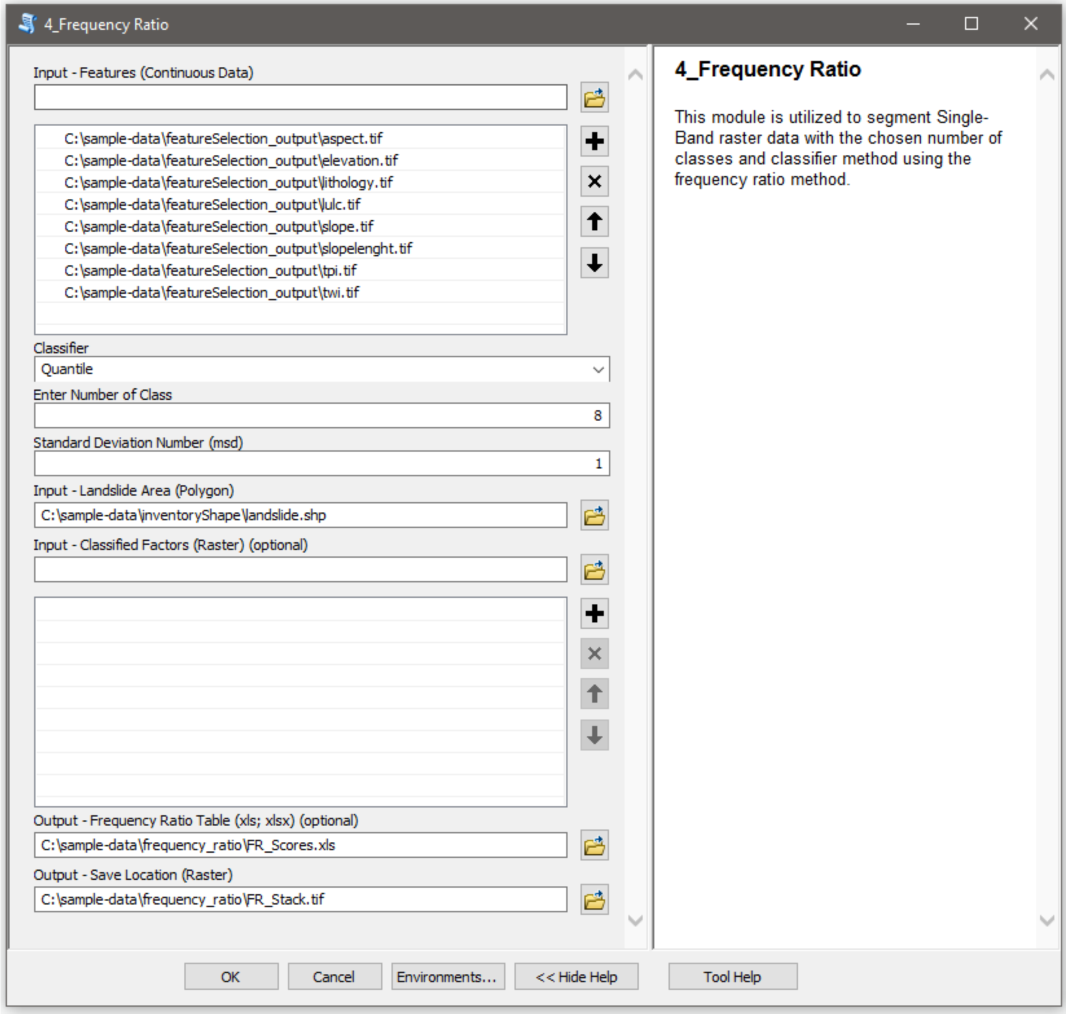




	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		rn	attr_import	rn.1	attr_import	rn.2	attr_import	rn.3	attr_import	rn.4	attr_import	rn.5	attr_import	rn.6	attr_import	rn.7	attr_importance.7
2	1	splitaspect.	24,9873	splitelevatic	18,17148	splitlitholog	24,6496	splitlulc.1	32,91239	splitslope.1	26,01091	splitslopele	28,18283	splittpi.1	31,34413	splittwi.1	49,88331
3	2	splitaspect.	23,78632	splitelevatic	49,93893	splitlitholog	20,52335	splitlulc.2	29,82858	splitslope.2	45,35766	splitslopele	47,06142	splittpi.2	26,08636	splittwi.2	33,55342
4	3	splitaspect.	14,29048	splitelevatic	32,09751	splitlitholog	32,62473	splitlulc.3	23,75306	splitslope.3	39,49312	splitslopele	34,77604	splittpi.3	24,7065	splittwi.3	20,83672
5	4	splitaspect.	26,29989	splitelevatic	18,06858	splitlitholog	21,52259	splitlulc.4	25,48568	splitslope.4	22,92508	splitslopele	20,99619	splittpi.4	23,25955	splittwi.4	21,81552
6	5	splitaspect.	12,41452	splitelevatic	22,38801	1	#YOK	splitlulc.5	0	splitslope.5	22,56297	splitslopele	19,43783	splittpi.5	27,44685	splittwi.5	22,89448
7	6	splitaspect.	42,32344	splitelevatic	29,63279	2	#YOK	splitlulc.6	11,52147	splitslope.6	30,95798	splitslopele	17,84947	splittpi.6	16,70337	splittwi.6	25,36656
8	7	splitaspect.	32,42364	splitelevatic	36,38055	3	#YOK	splitlulc.7	26,7743	splitslope.7	30,8225	splitslopele	31,04743	splittpi.7	13,91381	splittwi.7	32,92859
9	8	splitaspect.	15,08143	splitelevatic	35,9957	4	#YOK	1	#YOK	splitslope.8	29,7722	1	#YOK	splittpi.8	49,12373	splittwi.8	28,28094

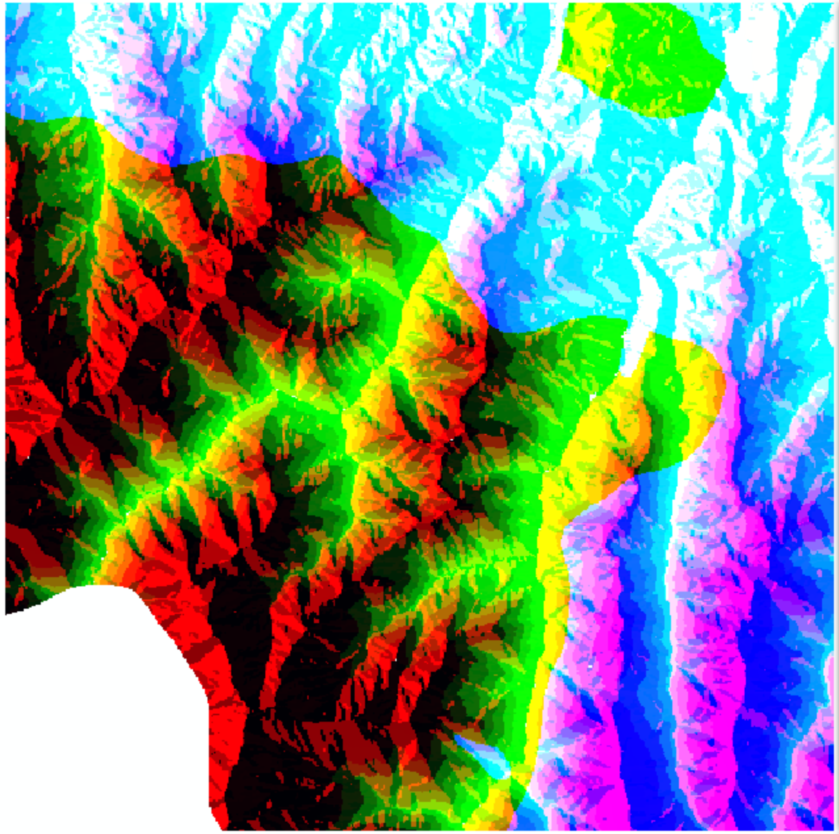
4_Frequency Ratio

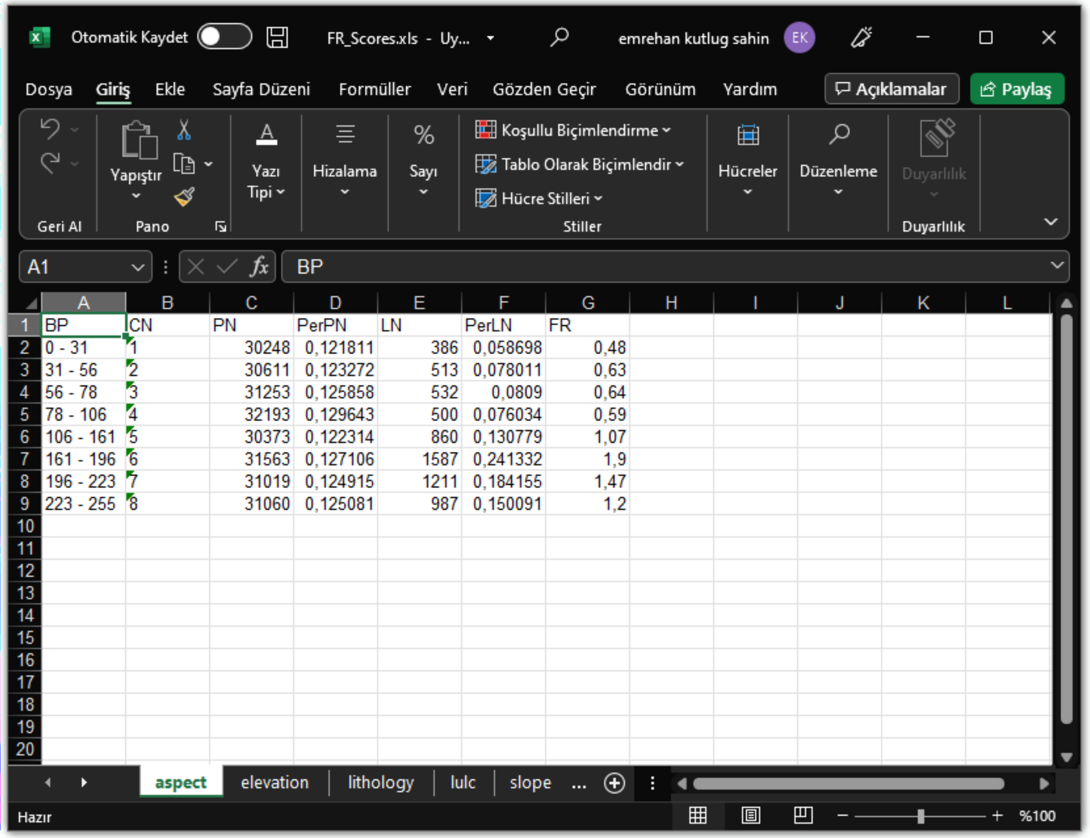
1. Import only feature produced with the Feature Selection module. Not: For this study data, the SPI factor was eliminated because of some problem about spi data type. Therefore, only 8 factors were selected for the weighing process.
2. Select the Classifier method.
3. Give number of class.
4. If «*msd*» method was selected as a classifier method, the user should be given a standard variation score.
5. Import only Landslide data with a polygon vector.
6. Select a folder and give XLS or XLSX document name
7. Give the output location for the raster stack data.
8. Click «OK»



4_Frequency Ratio

- Raster stack data and lists of frequency scores for each feature are located in the selected folder.
- Finally, the necessary raster stack data was obtained to generate the Landslide Susceptibility Mapping.

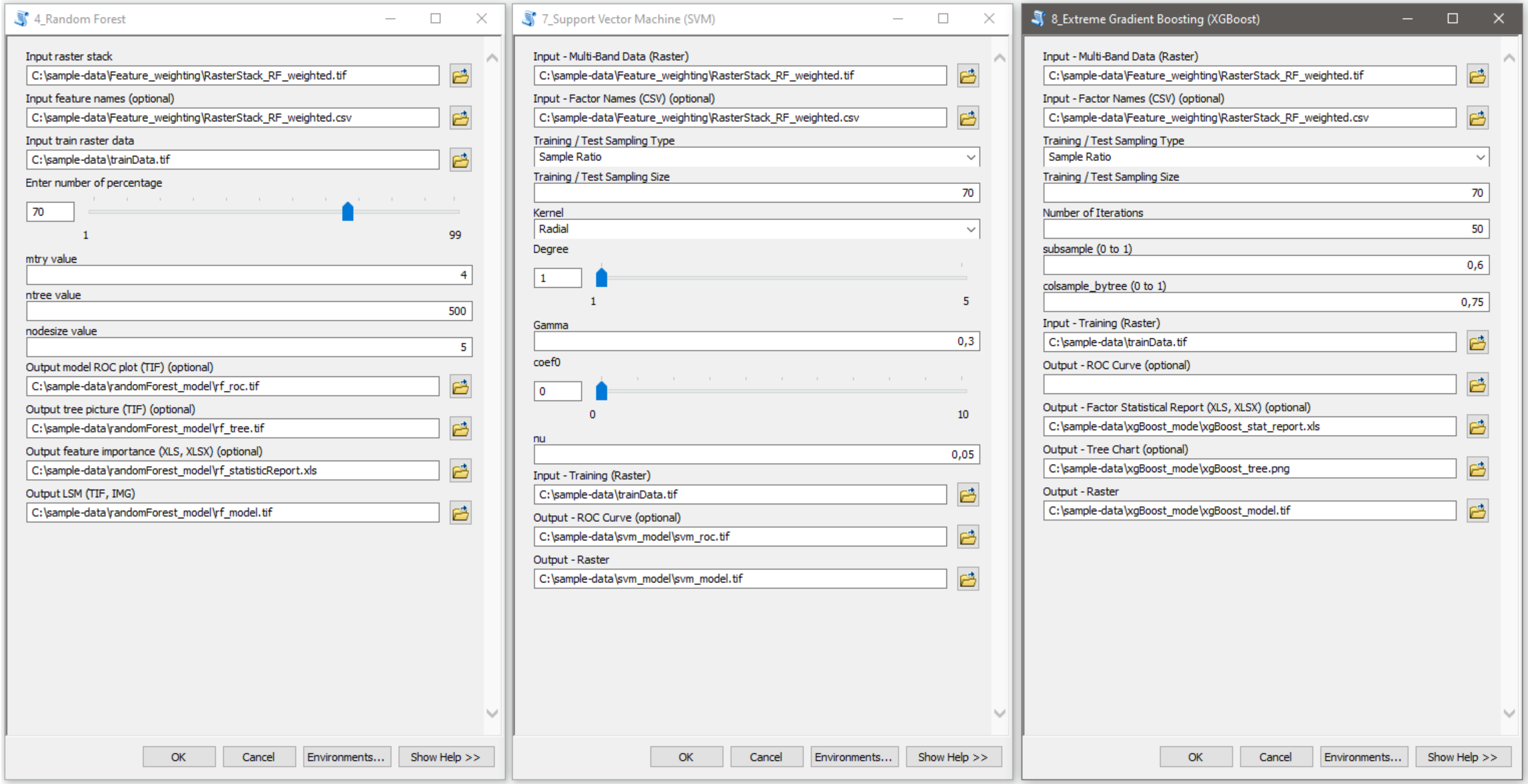




BP	CN	PN	PerPN	LN	PerLN	FR
0 - 31	1	30248	0,121811	386	0,058698	0,48
31 - 56	2	30611	0,123272	513	0,078011	0,63
56 - 78	3	31253	0,125858	532	0,0809	0,64
78 - 106	4	32193	0,129643	500	0,076034	0,59
106 - 161	5	30373	0,122314	860	0,130779	1,07
161 - 196	6	31563	0,127106	1587	0,241332	1,9
196 - 223	7	31019	0,124915	1211	0,184155	1,47
223 - 255	8	31060	0,125081	987	0,150091	1,2

Producing LSM using LSM Tool Pack v1.0

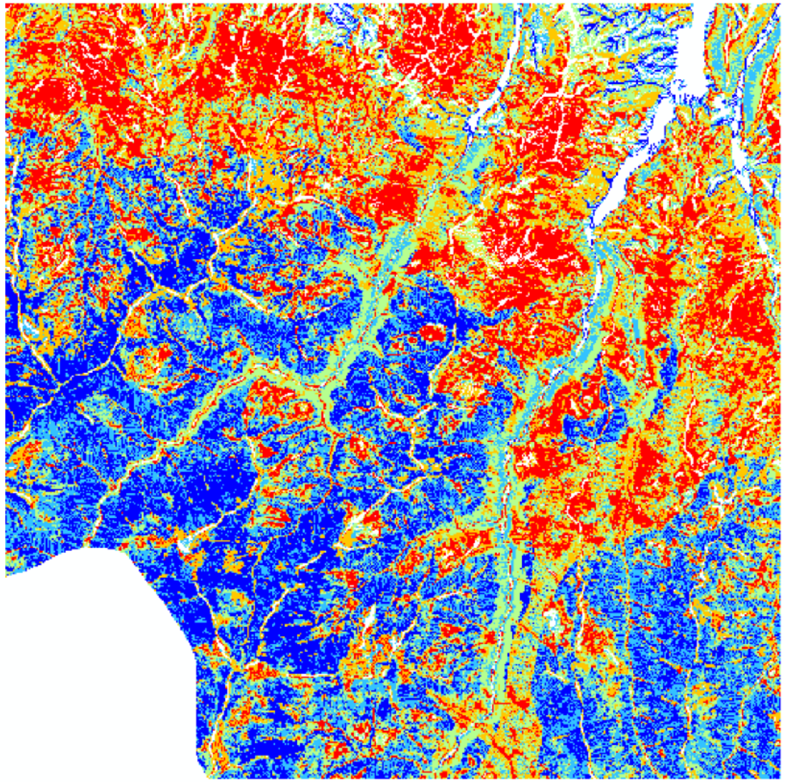
- Three basic ML methods namely Extreme Gradient Boosting, Random Forest, and Support Vector Machine can be used for modeling.
- Users easily can be reached at LSM Tool Pack v1.0 on the given Github webpage. https://github.com/emrehanks/R-ArcGIS-LSM_ToolPack



Producing LSM using LSM Tool Pack v1.0

- Output of *4_Random Forest* module results

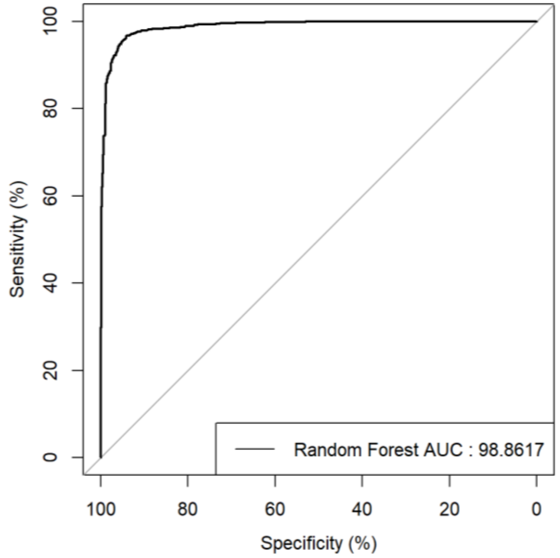
LS Map



Scores of feature importance

	IncNodePurity
splitaspect	46,69777
splitelevation	104,6985
splitlithology	158,4259
splitlulc	43,22014
splitslope	317,4535
splitslopelenght	27,36402
splittpi	36,46575
splittwi	94,80457

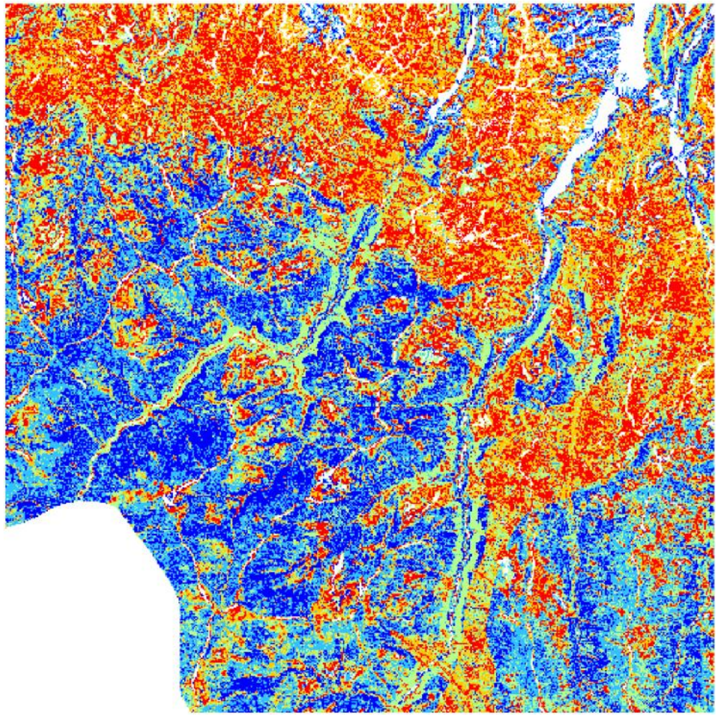
AUC-ROC



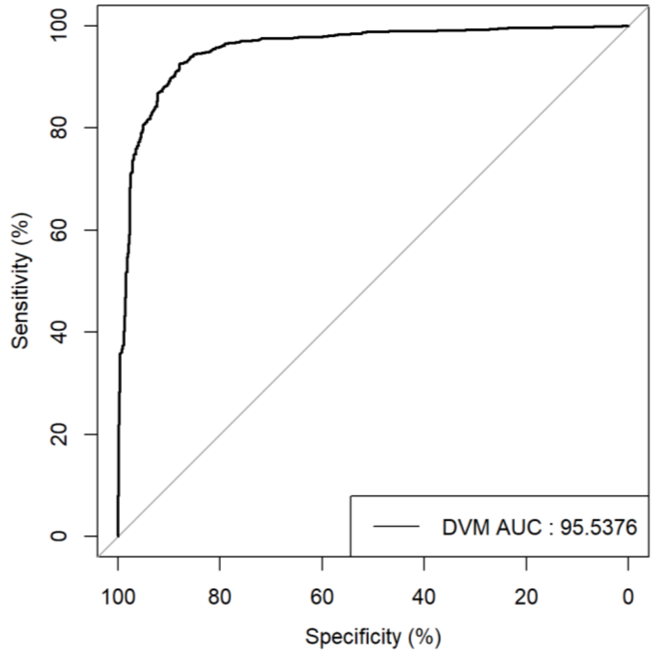
Producing LSM using LSM Tool Pack v1.0

- Output of *7_Support Vector Machine* (SVM) module results

LS Map



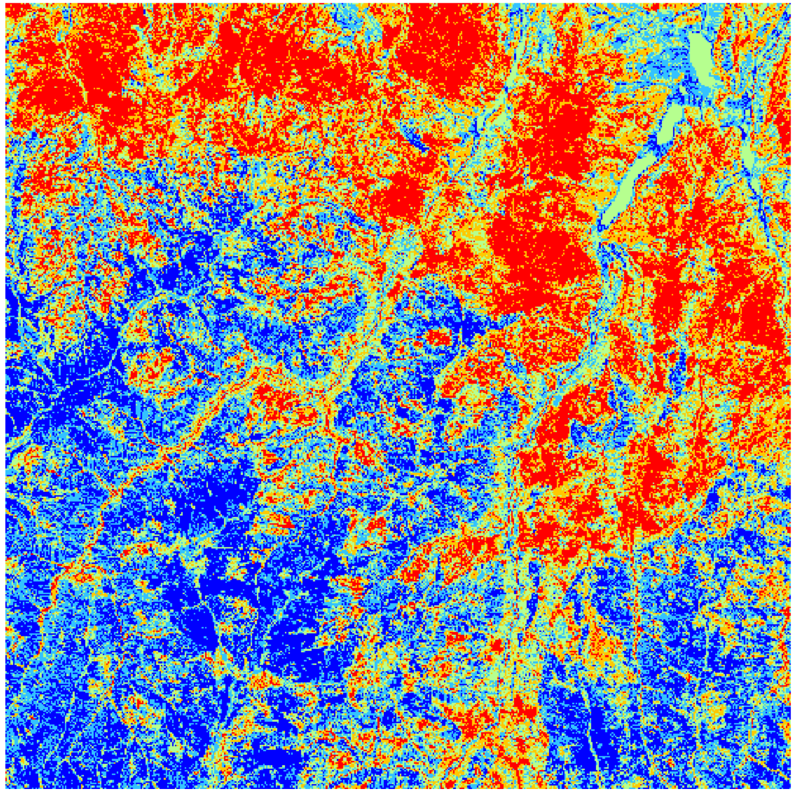
AUC-ROC



Producing LSM using LSM Tool Pack v1.0

- Output of *8_Extreme Gradient Boosting (XGBoost)* module results

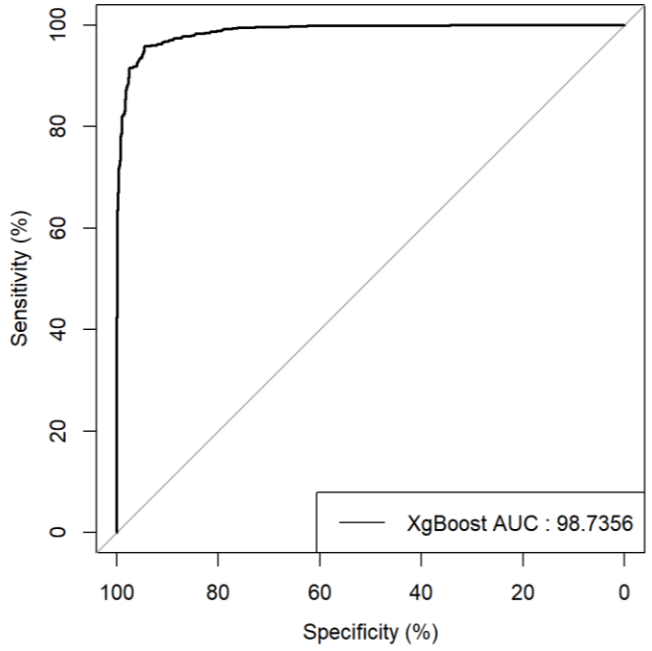
LS Map



Scores of feature importance

	Feature	Gain	Cover	Frequency
1	splitslope	0,314797967	0,179756685	0,125998771
2	splitlithology	0,175911772	0,114584788	0,077443147
3	splitelevation	0,162048064	0,170299626	0,175169023
4	splittwi	0,127587305	0,152884603	0,118623233
5	splitaspect	0,06697848	0,121317078	0,170866626
6	splitlulc	0,059982692	0,105898053	0,094652735
7	splittpi	0,049694773	0,091197581	0,145666872
8	splitslopelenght	0,042998946	0,064061585	0,091579594

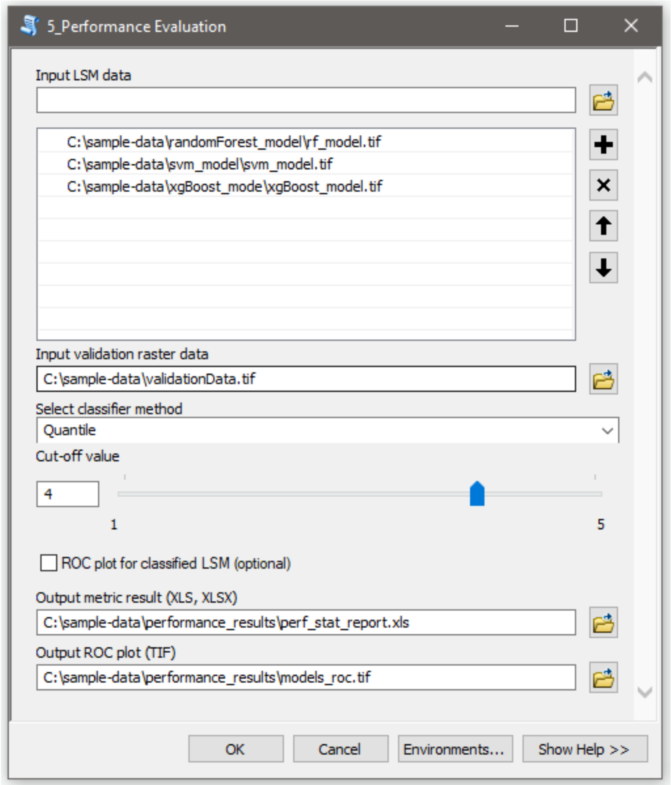
AUC-ROC



Producing LSM using LSM Tool Pack v1.0

- Output of *5_Performance Evaluation* module results

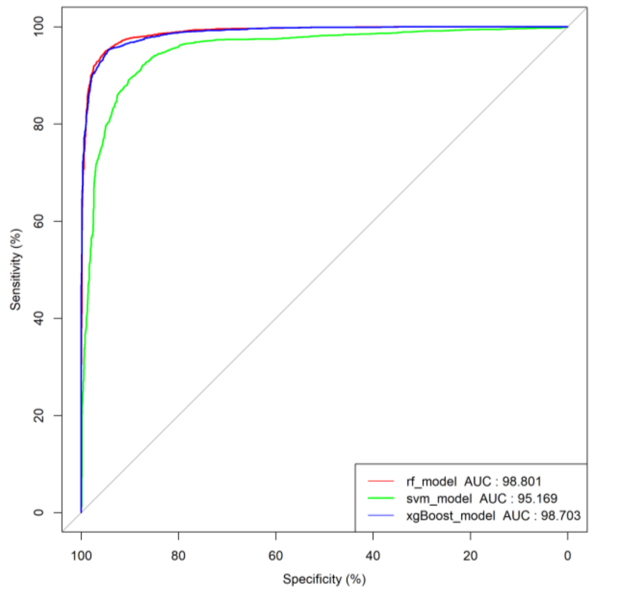
5_Performance Evaluation Module GUI



Performance results of three models

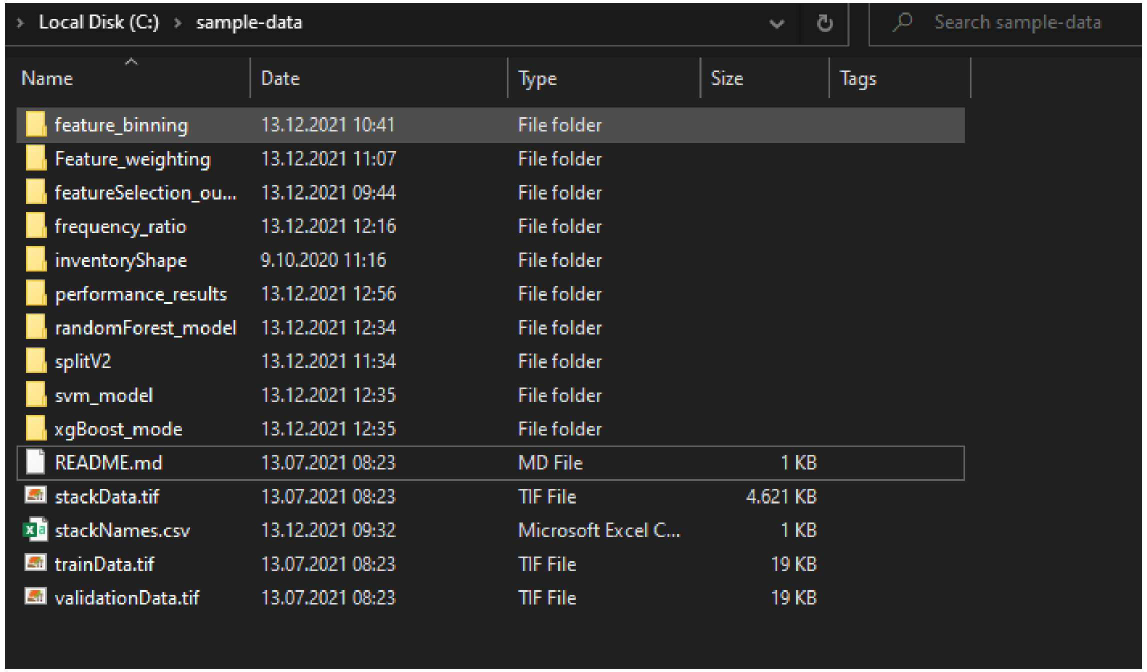
	Accuracy	AUC.Class	AUC.NonC	MAE	RMSE	Kappa	Precision	Recall	F1
xgBoost_model	**0,909594**	0,925583	0,987035	0,090406	0,300677	0,811683	0,987911	0,870542	0,92552
rf_model	**0,890765**	0,91232	0,988009	0,109235	0,330507	0,775588	0,991235	0,838119	0,90827
svm_model	**0,8552**	0,873667	0,951688	0,1448	0,380526	0,702629	0,959145	0,810097	0,878343

Comparison ROC graph of models



Results

- For this sample study, XGBoost model constructed with RF-selected subset was superior to other models.
- All oputput was published on article supplementry data folder.



Link:

[LSM Tool Pack v1.0](https://github.com/emrehanks/R-ArcGIS-LSM_ToolPack)
https://github.com/emrehanks/R-ArcGIS-LSM_ToolPack

[Semi-Automatic Feature Engineering Toolbox: FE.LSM Tool Pack](https://github.com/emrehanks/R-ArcGIS_Feature-Engineering_ToolPack)
https://github.com/emrehanks/R-ArcGIS_Feature-Engineering_ToolPack

Emrehan Kutlug Sahin, Ismail Colkesen, Suheda Semih Acmali, Aykut Akgun, Arif Cagdas Aydinoglu, Developing comprehensive geocomputation tools for landslide susceptibility mapping: LSM tool pack, Computers & Geosciences, 2020, 104592, ISSN 0098-3004, https://doi.org/10.1016/j.cageo.2020.104592. (http://www.sciencedirect.com/science/article/pii/S009830042030577X)