Exercise 26: Predictive Modeling with Random Forest Machine Learning

This exercise was created by Shobha Yadav, PhD student in the Department of Geology and Geography at WVU.

This exercise describes how to generate a predictive model using the Random Forest machine learning algorithm as implemented in ArcGIS Pro. The Random Forest algorithm is a popular supervised machine learning method used for both classification and regression. It allows you to build predictive models using variables derived from tabular attributes, distance-based features, and raster grids.

In this exercise, the percentage of households without internet access will be predicted using other county-level characteristics that may correlate with lack of internet access.

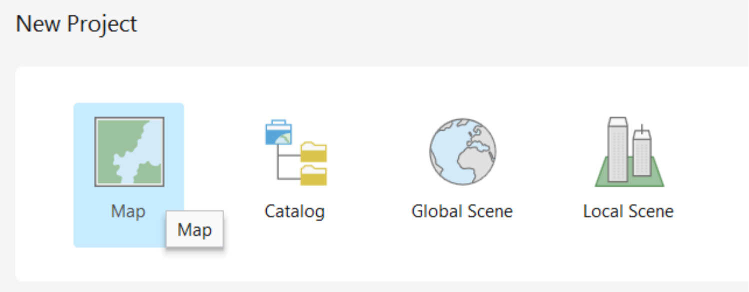
Topics covered in this exercise include:

1. Create training and validation data partitions.
2. Random forest classification in ArcGIS Pro.
3. Model evaluation.

Step 1. Create and Prepare a New Project

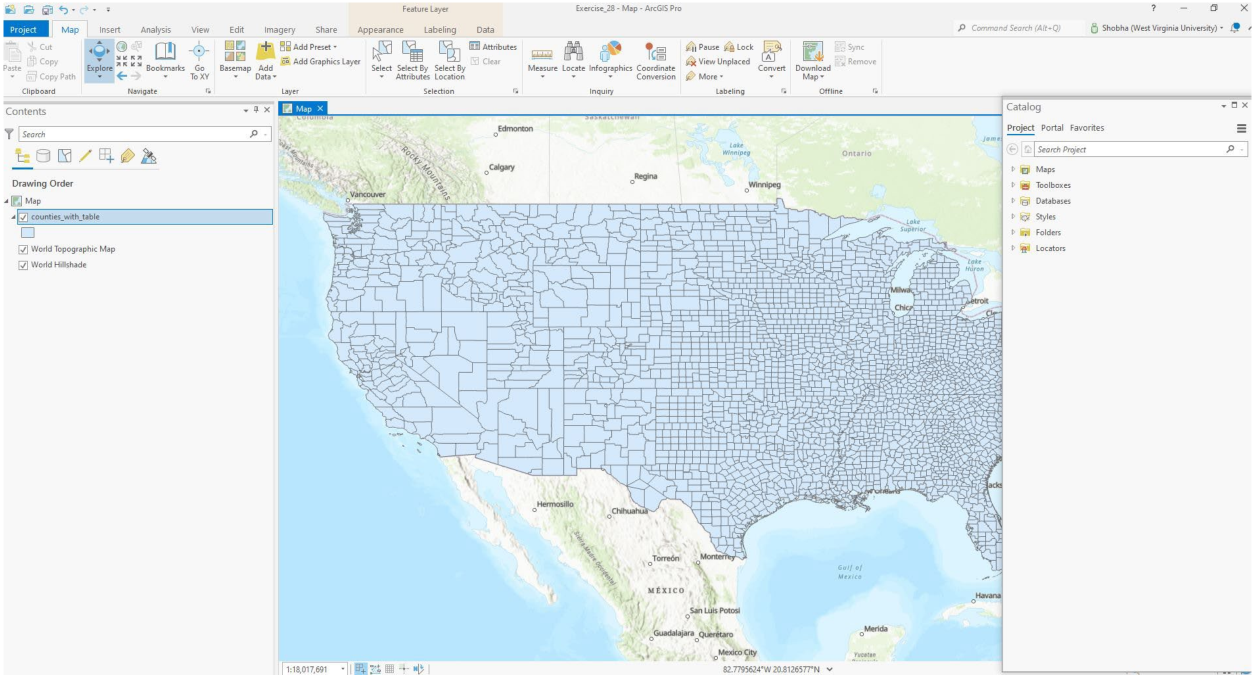
You will begin the analysis by creating a new project to work within.

- □ Open ArcGIS Pro.
- □ Once ArcGIS Pro launches, select **Map** under New Projects.
- □ In the Create a New Project Dialog Box, name your new project **Exercise_26** and save it to your personal folder. You can leave the "Create a new folder for this project" option selected.



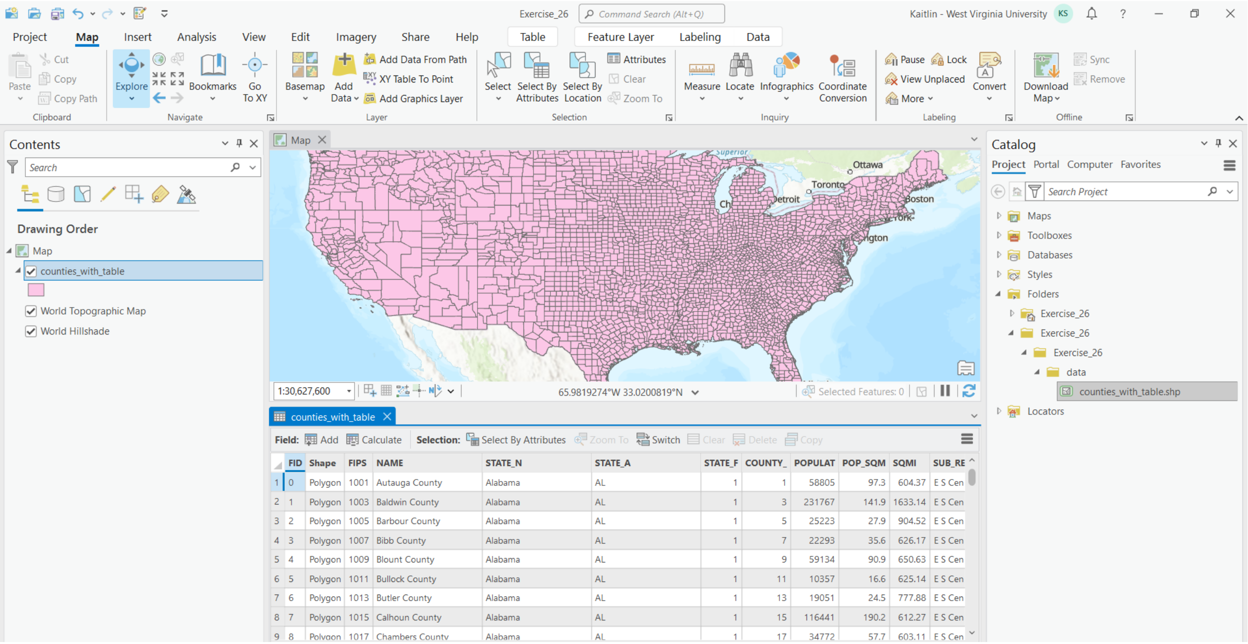
- □ Download the **Exercise_26** data from https://www.wvview.org/. All lab materials are available on the course webpage and linked to the exercise. You will need to extract the compressed files and save them to the location of your choosing.

- □ Click on the Add Data Button. Navigate to your copy of the lab data. Navigate to the downloaded **Data** folder. Add the **counties_with_table.shp** file.



Step 2. Explore Dataset

The dataset contains county-level data for the contiguous United States. To explore the dataset, right-click on **counties_with_table.shp** and then click on the attribute table. Take a moment to learn about the data.



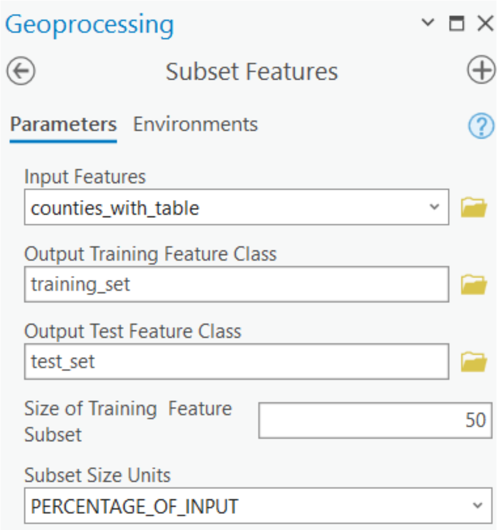
	FID	Shape	FIPS	NAME	STATE_N	STATE_A	STATE_F	COUNTY_	POPULAT	POP_SQM	SQMI	SUB_RE
1	0	Polygon	1001	Autauga County	Alabama	AL	1	1	58805	97.3	604.37	E S Cen
2	1	Polygon	1003	Baldwin County	Alabama	AL	1	3	231767	141.9	1633.14	E S Cen
3	2	Polygon	1005	Barbour County	Alabama	AL	1	5	25223	27.9	904.52	E S Cen
4	3	Polygon	1007	Bibb County	Alabama	AL	1	7	22293	35.6	626.17	E S Cen
5	4	Polygon	1009	Blount County	Alabama	AL	1	9	59134	90.9	650.63	E S Cen
6	5	Polygon	1011	Bullock County	Alabama	AL	1	11	10357	16.6	625.14	E S Cen
7	6	Polygon	1013	Butler County	Alabama	AL	1	13	19051	24.5	777.88	E S Cen
8	7	Polygon	1015	Calhoun County	Alabama	AL	1	15	116441	190.2	612.27	E S Cen
9	8	Polygon	1017	Chambers County	Alabama	AL	1	17	34772	57.7	603.11	E S Cen

Question 1. What is the projection system used for this map and data layer? (2 Points)

Step 3. Create Training and Validation data

Before you can train the Random Forest algorithm, you need separate, non-overlapping training and validation sets. To achieve this, you must split the data into two subsets, training, and validation, using the **Subset Features Tool**.

- □ Navigate to the Analysis ribbon and click on Tools. In the Geoprocessing Pane search for “subset features.”
- □ Double-click on the **Subset Features Tool** (this tool is in the Geostatistical Analyst Toolbox). In the tool window, use the original data as the Input Features. Name the Output Training Feature Class “**training_set**” and the Test Training Feature Class “**test_set**.” 50% of the counties will be randomly selected to train the model, and 50% will be randomly selected to validate the model. It will look like this:



Input Features: counties_with_table

Output Training Feature Class: training_set

Output Test Feature Class: test_set

Size of Training Feature Subset: 50

Subset Size Units: PERCENTAGE_OF_INPUT

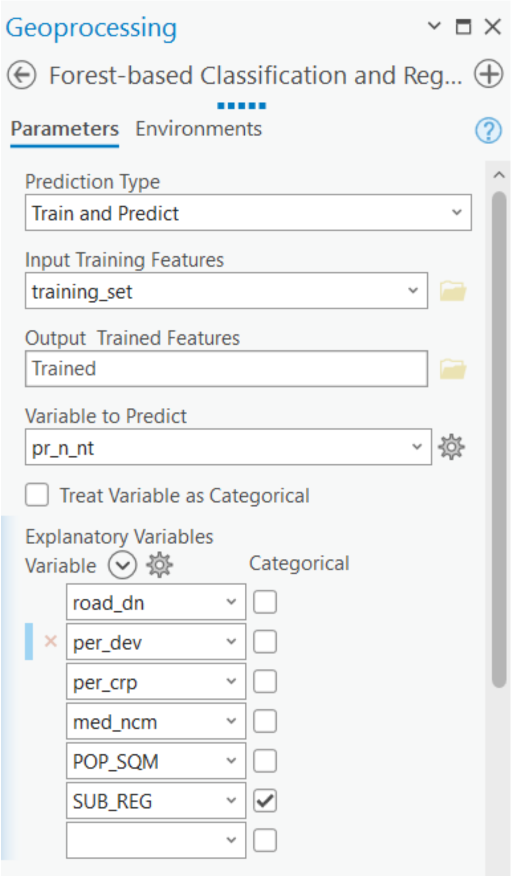
- □ Now, click Run to execute the tool. It will take a few seconds to run. Once the tool executes, you will see two more layers added to your Contents Pane. You should have two different partitions and they should not overlap.

Question 2. Why is it necessary to have separate, non-overlapping datasets to train and assess a machine learning model? (4 Points)

Step 4. Train the Algorithm and Use the Model

Once you have training and validation samples, we can run the tool.

- □ In the Geoprocessing Pane, navigate to the Spatial Statistics Tools Toolbox.
- □ Navigate to Modeling Spatial Relationship. Click on the **Forest-Based Classification and Regression Tool**.
- □ Set the Prediction Type to "Train and Predict". The Input Training Features should be set to the **Training_Set** layer, the Output Trained Features to **Trained**, and the Variable to Predict should be set to the **pr_n_nt** field. This represents the



Prediction Type: Train and Predict

Input Training Features: training_set

Output Trained Features: Trained

Variable to Predict: pr_n_nt

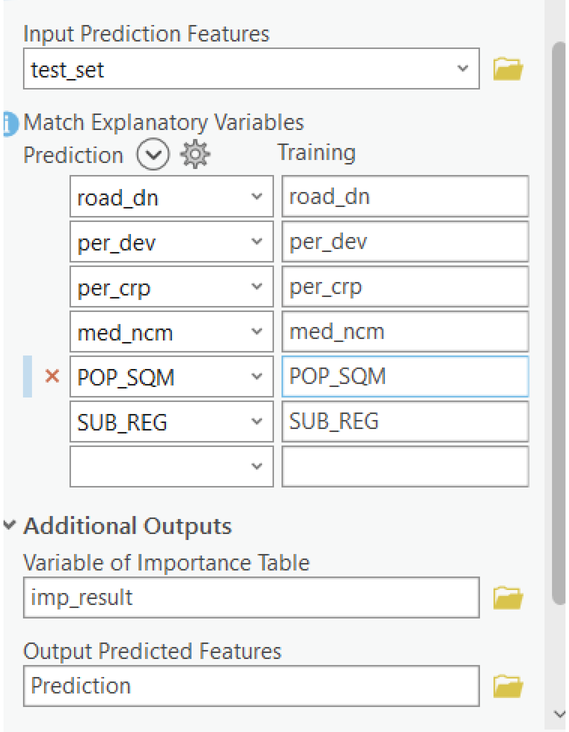
□ Treat Variable as Categorical

Explanatory Variables

Variable	Categorical
road_dn	□
per_dev	□
per_crp	□
med_ncm	□
POP_SQM	□
SUB_REG	☑
	□

estimated percentage of households in the county that do not have internet access.

- □ Next, your Input Prediction Features should be set to **test_set**.
- □ For Explanatory Training Variables, you need to select a few continuous variables, namely: road density (**road_dn**), percent developed (**per_dev**), percentage crop (**per_crp**), household median income (**med_ncm**), and population density (**POP_SQM**). You will also include one categorical variable: the sub-region of the country in which the county occurs. Make sure that this variable is treated as a categorical predictor. (**SUB_REG**).



Input Prediction Features

test_set

Match Explanatory Variables

Prediction	Training
road_dn	road_dn
per_dev	per_dev
per_crp	per_crp
med_ncm	med_ncm
POP_SQM	POP_SQM
SUB_REG	SUB_REG

Additional Outputs

Variable of Importance Table

imp_result

Output Predicted Features

Prediction

- □ Under Additional Outputs, name the Variable of Importance Table "**imp_result**"
- □ Name the Output Predicted Features "**prediction**."
- □ You do not need to change any of the other settings. Click Run to execute the tool. The output should be automatically placed on your map.

Note: The output layer contains predictions for the percent of households without internet access for the withheld test counties. The variable importance estimates have also been added to the Contents Pane as a stand-alone table.

- □ Right-click on the "**imp_result**" table Contents Pane and click open.

Note: Each student will get a different number because the split was randomly selected, and the random forest model is stochastic.

Question 3. What percentage of importance was contributed to median income? (4 Points)

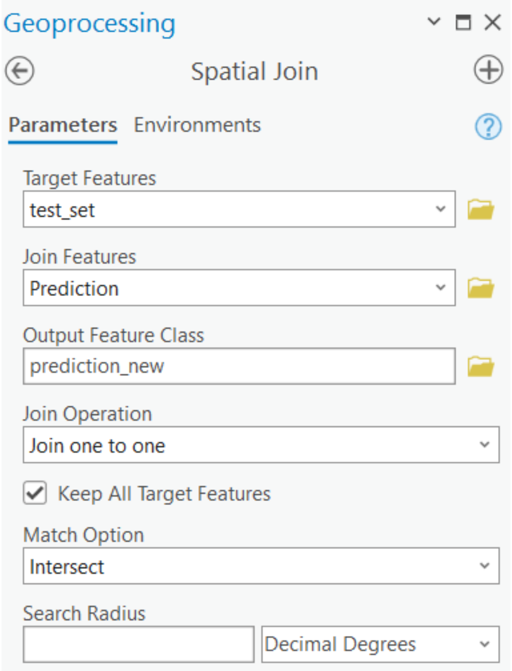
Question 4. What percentage of importance was contributed to population density? (4 Points)

Question 5. Which variable has the lowest percentage of importance in the model? (4 Points)

Step 5. Calculate Root Mean Square Error (RMSE)

In this step, you will calculate RMSE based on the prediction.

- □ To calculate RMSE, you first need to join the **prediction** layer with the **test_set** layer.
- □ In the Geoprocessing Pane, search for "Spatial Join" to find the **Spatial Join Tool**.
- □ In a new window, set the Target Features to **test_set** and the Join Features to **prediction**.
- □ Name the Output Feature Class **prediction_new**.
- □ You do not need to change any of the other settings.
- □ Click Run to execute the tool. The output should automatically be added to your map.



Geoprocessing

Spatial Join

Parameters Environments

Target Features: test_set

Join Features: Prediction

Output Feature Class: prediction_new

Join Operation: Join one to one

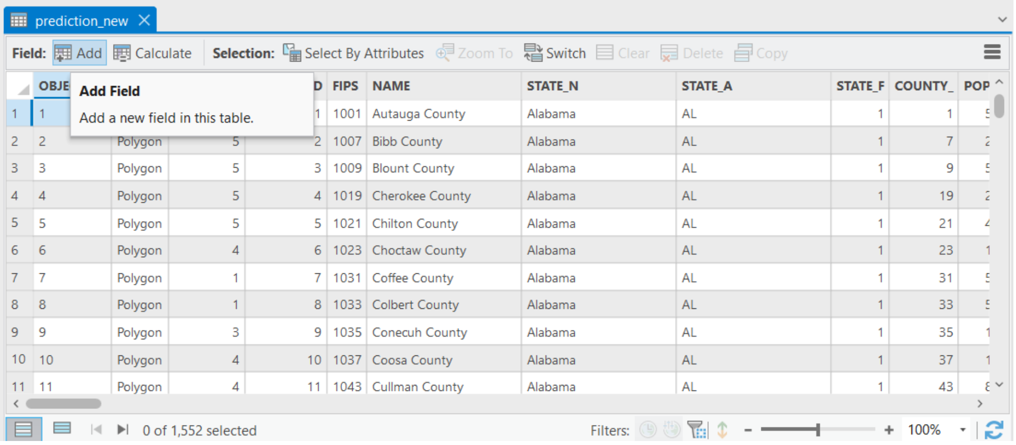
☑ Keep All Target Features

Match Option: Intersect

Search Radius: Decimal Degrees

Now, you can use the new prediction to calculate RMSE.

- □ Right-click on the **prediction_new** layer in the Contents Pane and open the attribute table.
- □ In the attribute table, click Add.


prediction_new
Field: Add Calculate
Selection: Select By Attributes Zoom To Switch Clear Delete Copy
Filters:
100%

Add Field

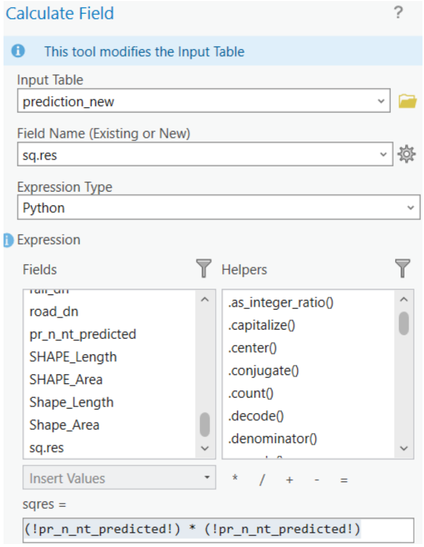
Add a new field in this table.

	OBJE			D	FIPS	NAME	STATE_N	STATE_A	STATE_F	COUNTY_	POP
1	1			1	1001	Autauga County	Alabama	AL	1	1	5
2	2	Polygon	5	2	1007	Bibb County	Alabama	AL	1	7	2
3	3	Polygon	5	3	1009	Blount County	Alabama	AL	1	9	5
4	4	Polygon	5	4	1019	Cherokee County	Alabama	AL	1	19	2
5	5	Polygon	5	5	1021	Chilton County	Alabama	AL	1	21	4
6	6	Polygon	4	6	1023	Choctaw County	Alabama	AL	1	23	1
7	7	Polygon	1	7	1031	Coffee County	Alabama	AL	1	31	5
8	8	Polygon	1	8	1033	Colbert County	Alabama	AL	1	33	5
9	9	Polygon	3	9	1035	Conecuh County	Alabama	AL	1	35	1
10	10	Polygon	4	10	1037	Coosa County	Alabama	AL	1	37	1
11	11	Polygon	4	11	1043	Cullman County	Alabama	AL	1	43	8

0 of 1,552 selected

- □ In the new window, name the new field **sqres,** the Alias **sq.res** and set the type to "double."
- □ Click Save. You will see a new column added to your attribute table once you navigate back to it.
- □ Right-click on your new column and click Calculate Field.
- □ In the formula bar, type:

  (!pr_n_nt_predicted!) *
  (!pr_n_nt_predicted!)
- □ Click Apply and your new field will be populated with data.
- □ Right-click on the **sq.res** column and then click on Explore Statistics.
- □ A new pane will show statistics calculated for your new column. Now you need to calculate RMSE by hand using the sum of the residuals (Sum) and the number of samples (Count). Divide the sum by the count then take the square root to obtain RMSE.



This tool modifies the Input Table

Input Table: prediction_new

Field Name (Existing or New): sq.res

Expression Type: Python

Fields:
road_dn
pr_n_nt_predicted
SHAPE_Length
SHAPE_Area
Shape_Length
Shape_Area
sq.res

Helpers:
.as_integer_ratio()
.capitalize()
.center()
.conjugate()
.count()
.decode()
.denominator()

sqres =

```
(!pr_n_nt_predicted!) * (!pr_n_nt_predicted!)
```

Note: The unit of a root means square error is always the same as what you are evaluating in your model.

Question 6. What RMSE did your model yield? (4 Points)

Question 7. Explain what the RMSE metric represents? (4 Points)

Question 8. What are the units of RMSE for this prediction or problem? (4 Points)

Question 9. Why should you calculate RMSE using the withheld test or validation data as opposed to the training data? (4 Points)

Step 6. Repeat the Model Using Only Median Income

Lastly, replicate this process but only use the median income variable as a predictor in the model. You need to repeat Steps 4 and 5.

Question 10. Discuss and compare the results obtained using the two different sets of predictor variables? How do the RMSE values compare for predicting the withheld test or validation data? Does including more variables, other than just median income, improve the model performance? (8 Points)

END OF EXERCISE