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 correlation with this attribute.

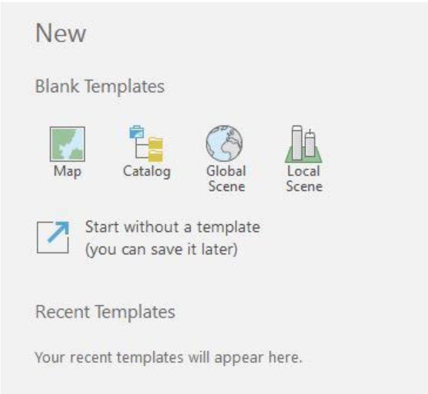
<u>Topics covered in this exercise include</u>:

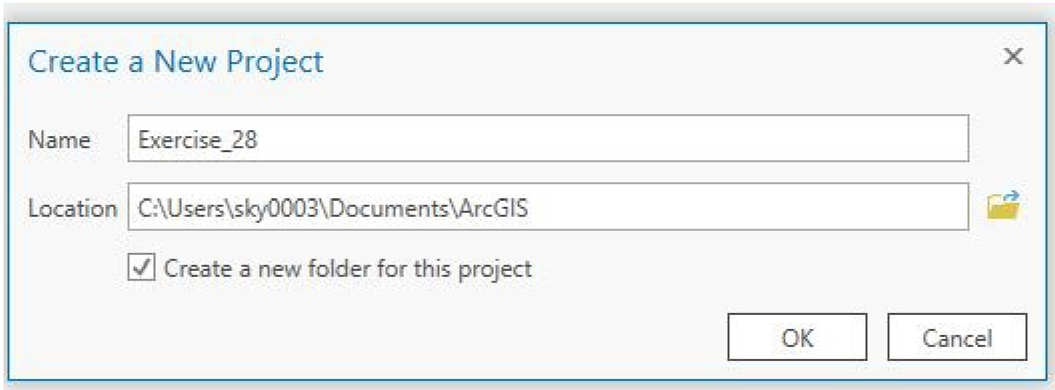
1. Create training and validation data partitions
2. Random forest classification in ArcGIS Pro
3. Model evaluation

Step 1. Create and Prepare a New Project

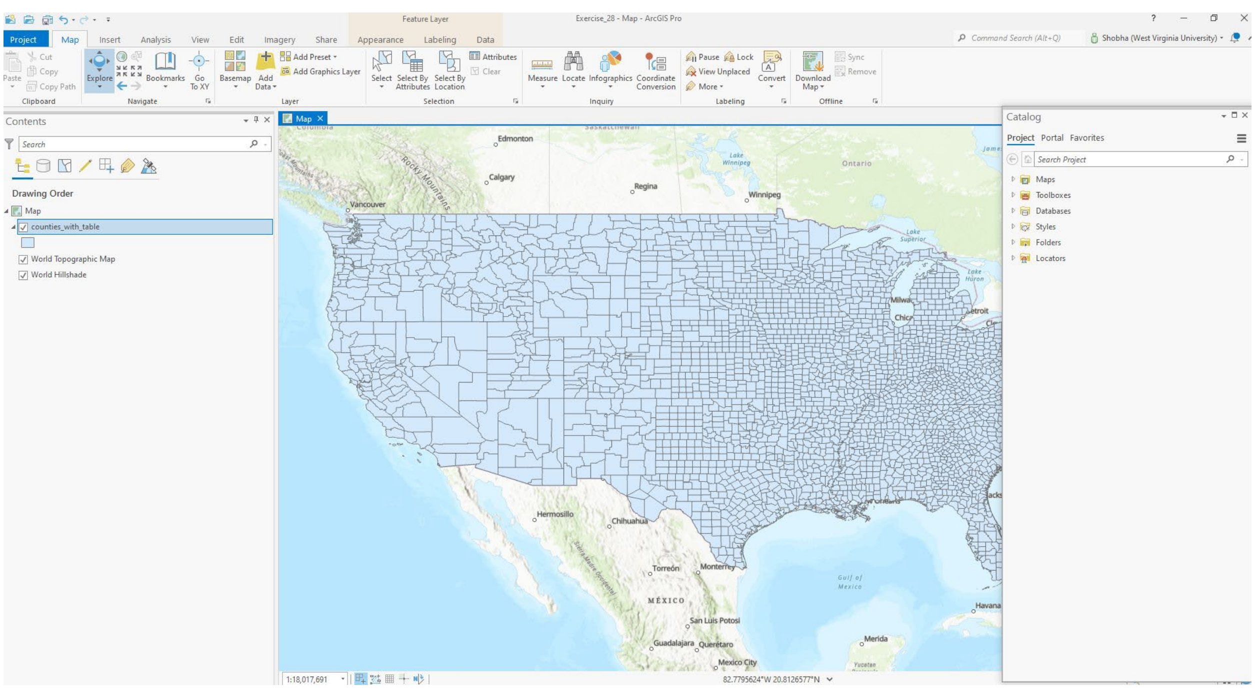
You will begin the analysis by creating a new project to work within.

- Open ArcGIS Pro. This can be done by navigating to All Apps followed by the ArcGIS Folder. Within the ArcGIS Folder, select ArcGIS Pro. Note that you can also use a Task Bar or Desktop shortcut if they are available on your machine.
- Once ArcGIS Pro launches, select **Map** under New Blank Templates.
- 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.



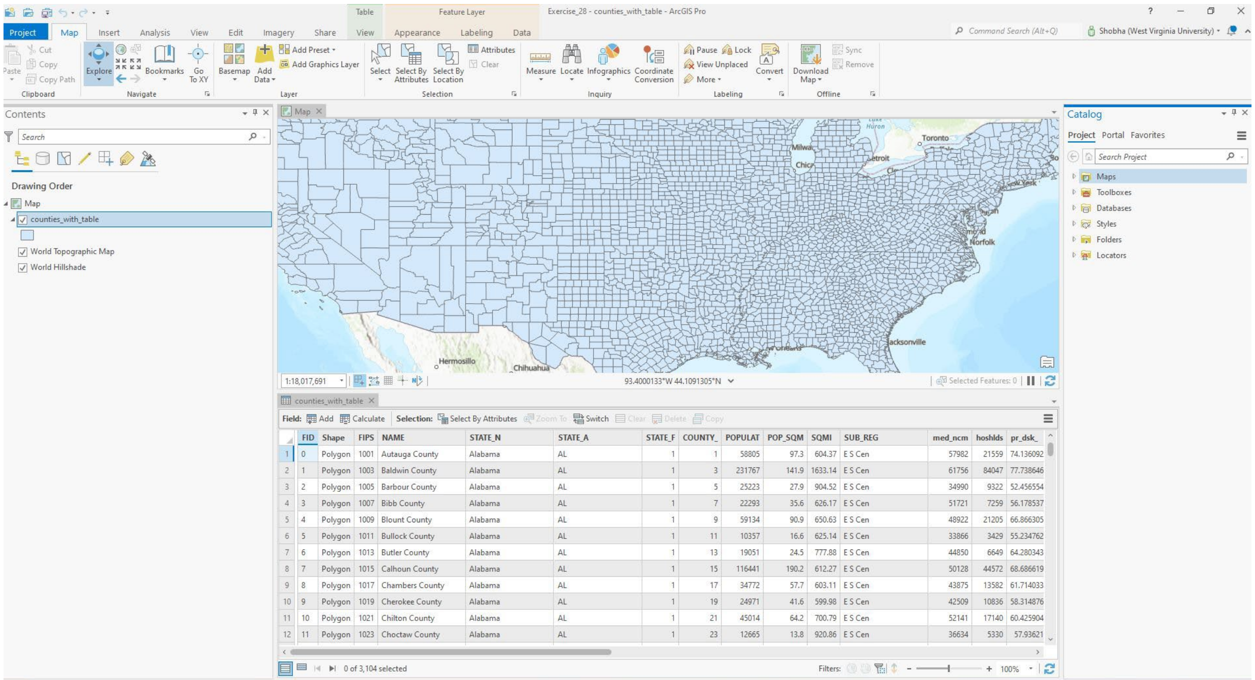


- Download the **Exercise_26** data from https://www.wvview.org/. All lab materials are available on the course webpage and linked to the exercise. You will need to extract the compressed files and save it to the location of your choosing.
- Click on the Add Data Button. Navigate to your copy of the 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-clicking on **counties_with_table.shp** and then click on the attribute table. Take a moment to learn about these data.

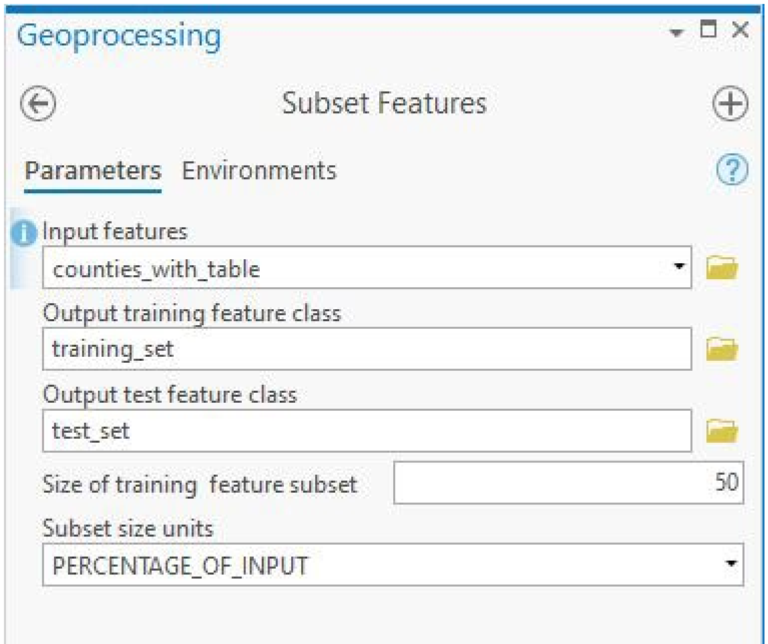


Question 1. What is the projection system used to project this map and data layer? (2 Points)

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:



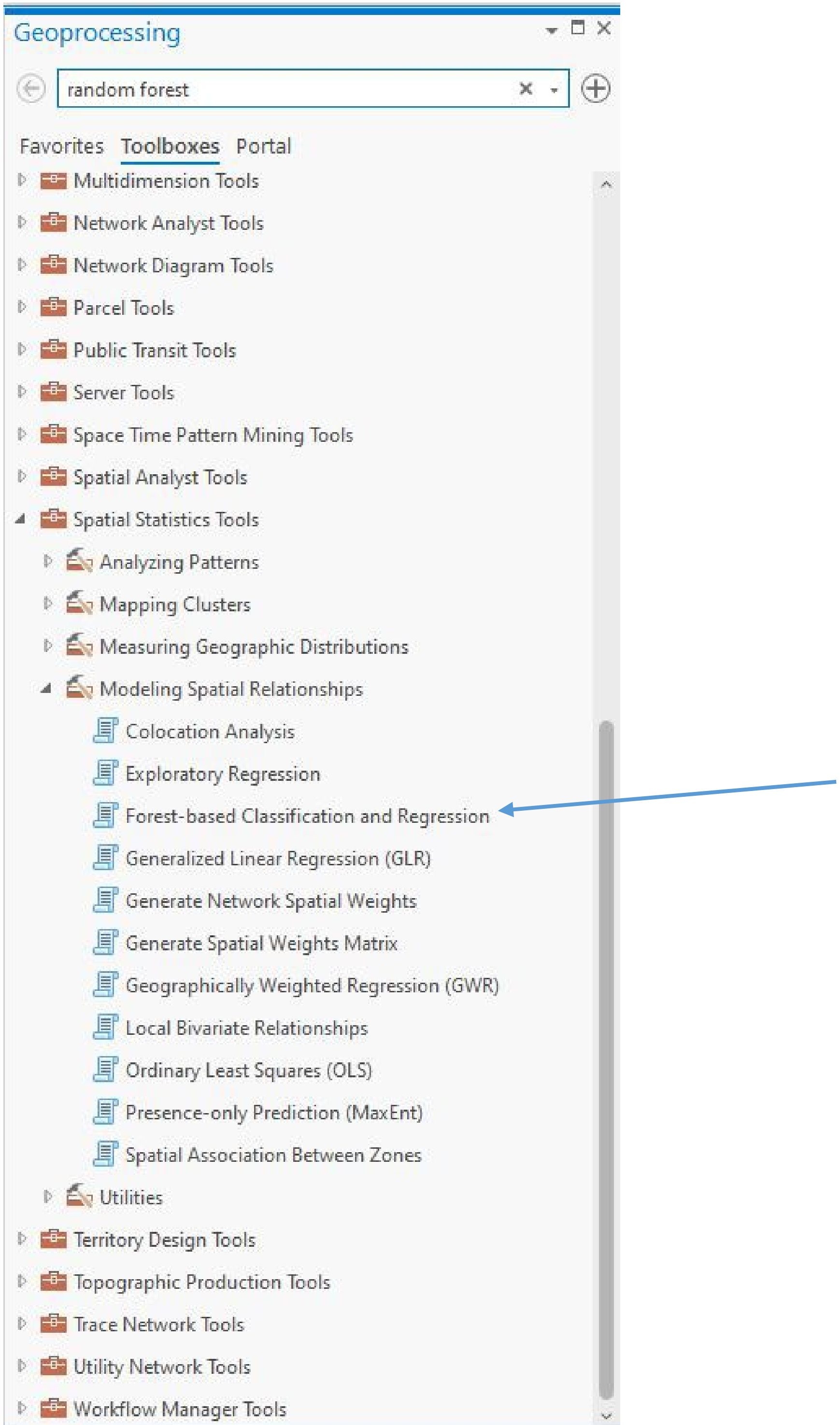
- Now, click Run to execute the tool. It will take a few seconds to run. Once the tool executes, you will see two more layers added to your Contents Pane. You should have two different partitions and they should not be overlapping.

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

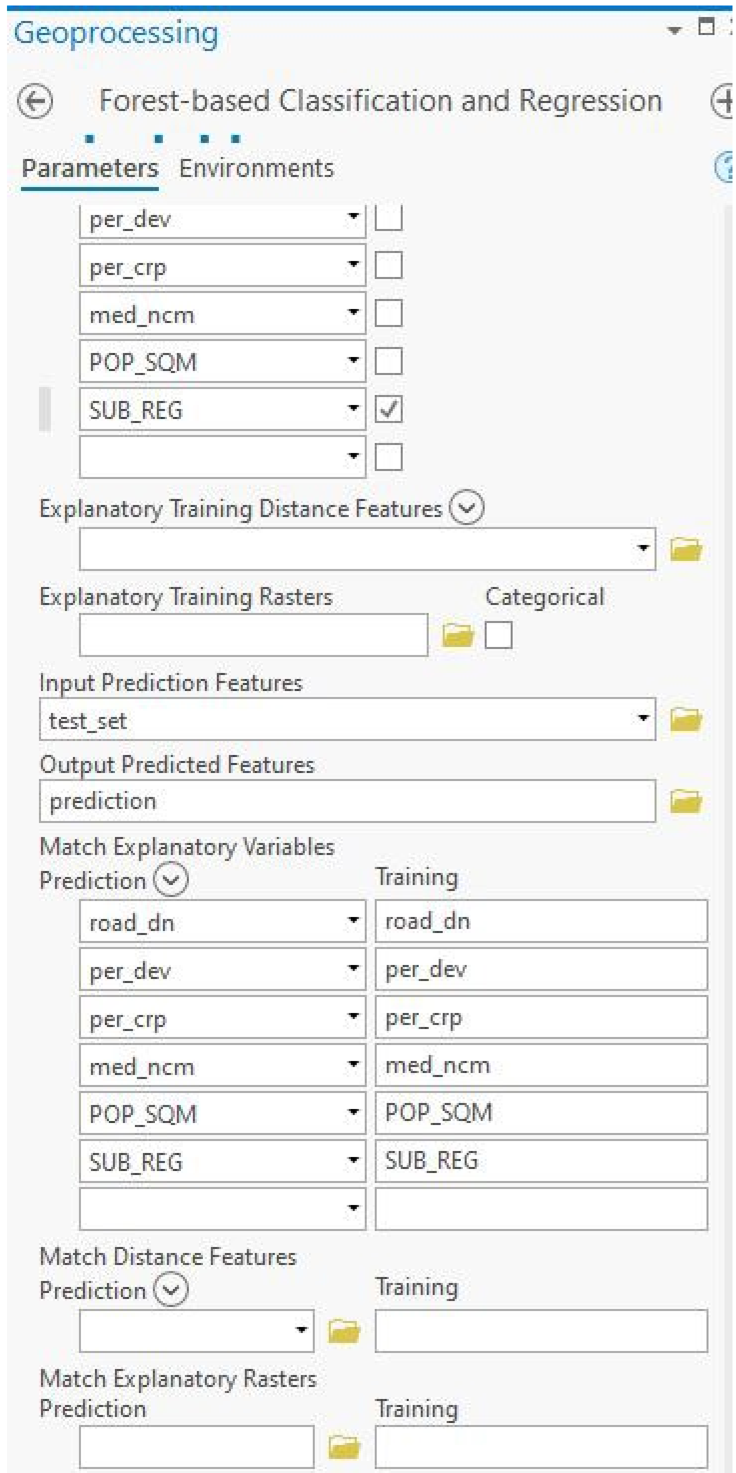
Once you have training and validation samples, we can run the tool.

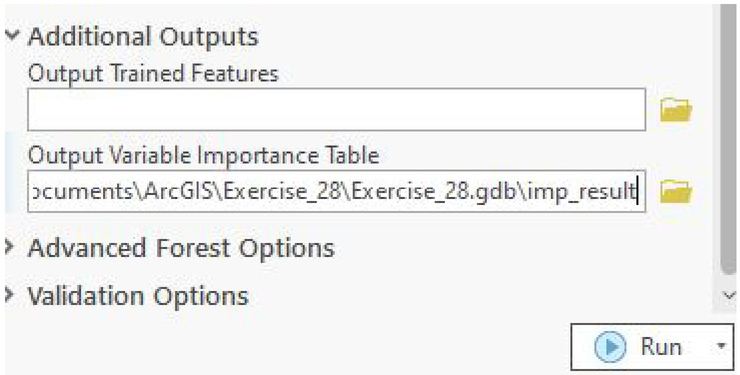
- In the Geoprocessing Pane, navigate to the Spatial Statistics Tools Toolbox.
- Navigate to Modeling Spatial Relationship. Click on the **Forest-Based Classification and Regression Tool**.



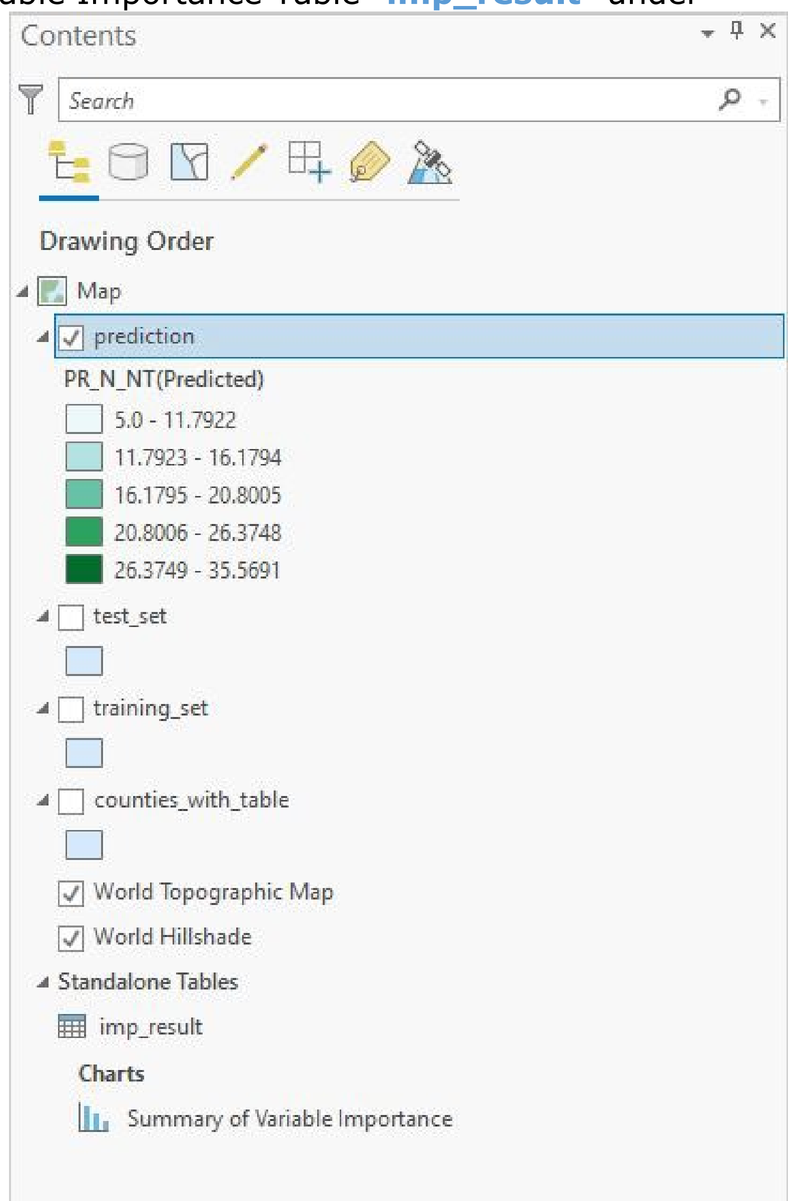
- Set the Prediction Type to the “Predict to Features” type. The Input Training Features should be set to the **Training_Set** layer and the Variable to Predict should be set to the **pr_in_nt** field. This field represents the estimated percent of households in the county that do not have internet access.

- 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**).





- Name the Output Predicted Features "**prediction**."
- Name the Output Variable Importance Table "**imp_result**" under Additional Outputs.
- 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.



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 and a bar chart has been generated from these data.

- Right-click on the chart and click open.
- The chart shows the importance of variables in the model. You will see that the median

income and population density are the most important variables in the model.

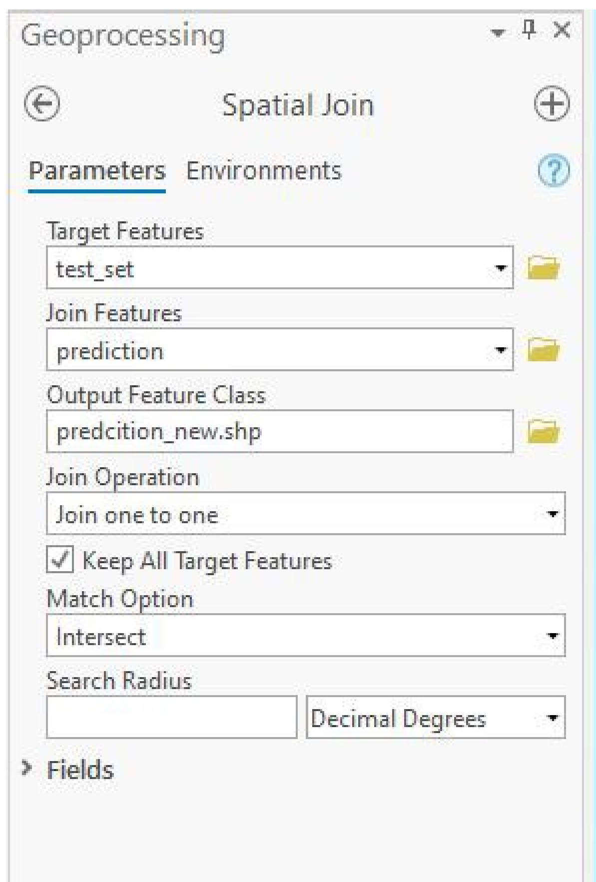
- Again right-click on the "**imp_result**" table above chart in the Content Pane and click open.

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)

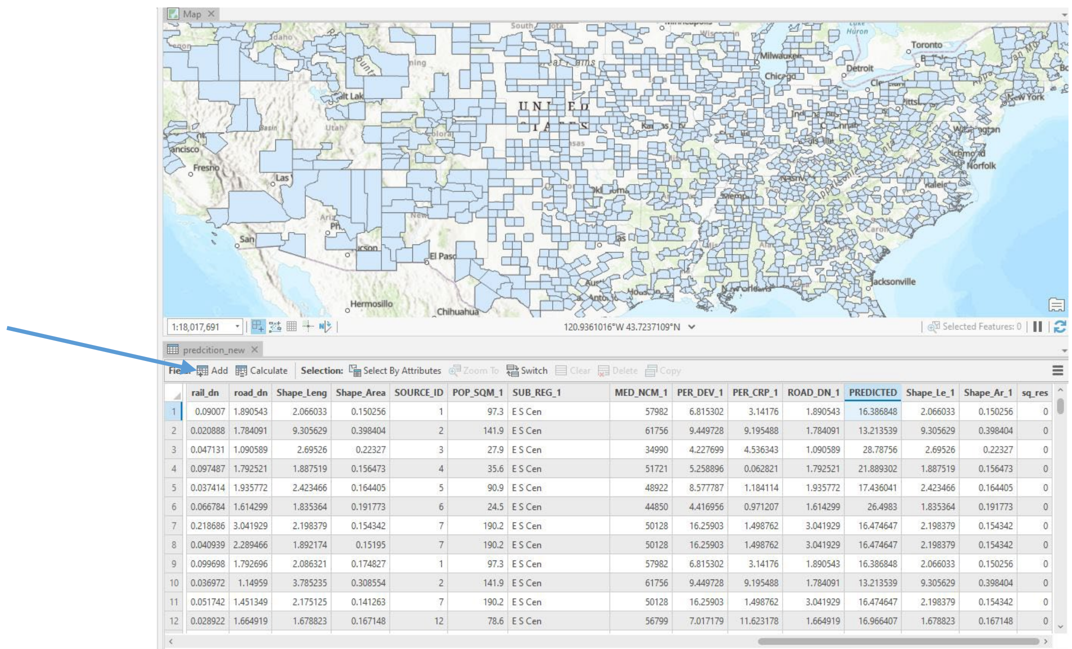
In this step, you will calculate RMSE based on the prediction.

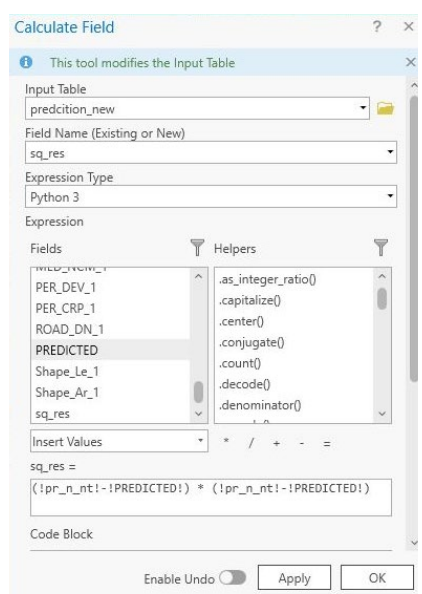
- Right-click on the **prediction** layer in the Contents Pane, then select Attribute Table.
- To calculate RMSE, you first need to join the **prediction** layer with the **test_set** layer.
- In the Geoprocessing Pane, search for "Spatial Join" to find the **Spatial Join Tool**.
- In a new window, set 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.

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.





- In the new window, name the new field **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 Statistics.

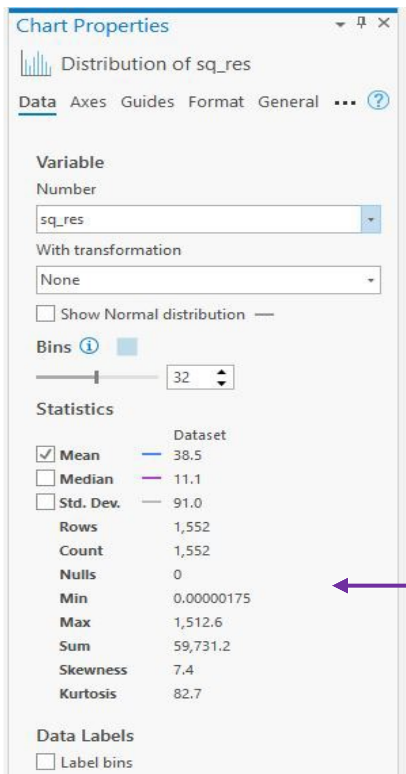
- A new pane will show statistics calculated for your new column. Now you need to calculate RMSE by hand using the sum of the residuals (Sum) and the number of samples (Count). Divide the sum by the count then take the square root to obtain RMSE.

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)



This is just example. You might get different answer

Step 6. Repeat the Model Using Only Median Income

Lastly, replicated this process but only use the median income variable as a predictor in the model. You need to repeat Step 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