CIS 3200 Term Project Tutorial

Authors: Mario Arguello Jr., Brianne Phillips, Michael Brumwell, and Christian Rodriguez

Instructor: Jongwook Woo

Date: 12/14/2019

Lab Tutorial

Mario Arguello Jr (marguel3@calstatela.edu)

Brianne Phillips (bphill@calstatela.edu)

Michael Brumwell (mbrumwe@calstatela.edu)

Christian Rodriguez (crodr214@calstatela.edu)

12/14/2019

Patient Influenza Vaccinations with Hospital Names from California Department of Public Health (Bayesian Linear Regression with Microsoft Azure Machine Learning Studio)

Objectives

In this hands-on lab, you will learn how to:

- Get data manually by downloading it from the CS Dept of Public Health website.
- Create MS Azure ML Studio Experiment.

- Use Bayesian Regression model to find insights.
- Use predictive models to find how many California flu patients may get vaccinated in 2020.
- Visualizations in Score and Evaluation models.

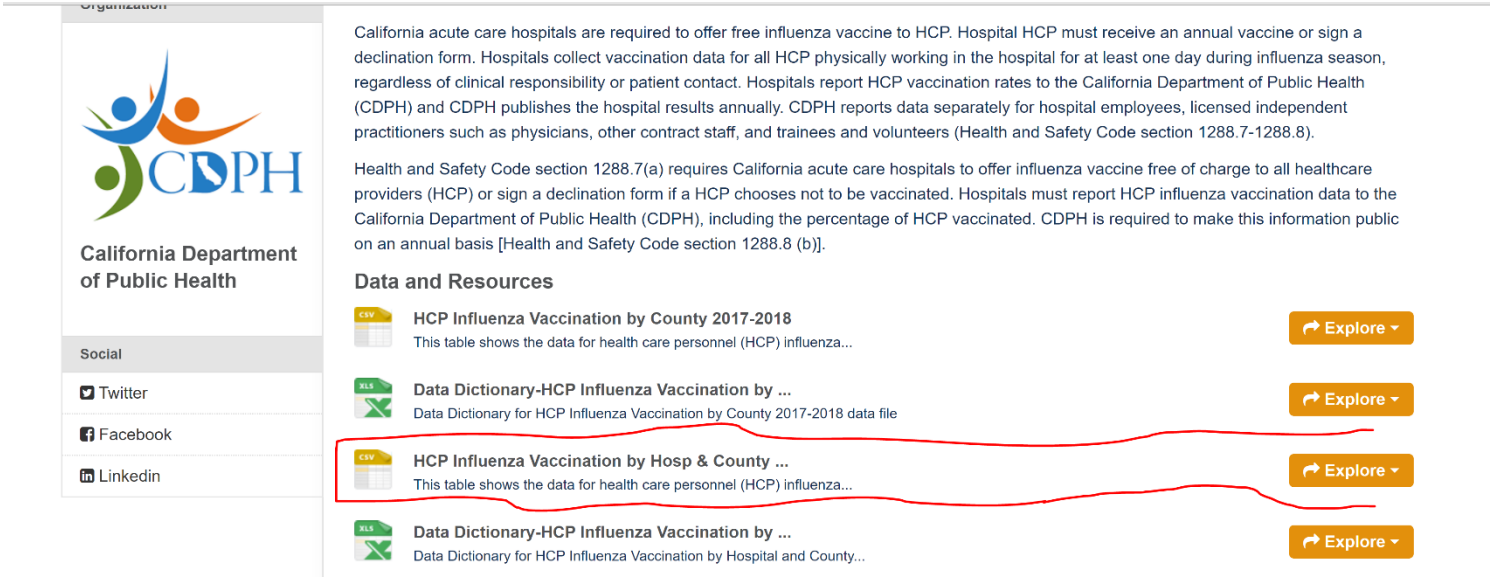
Platform Spec

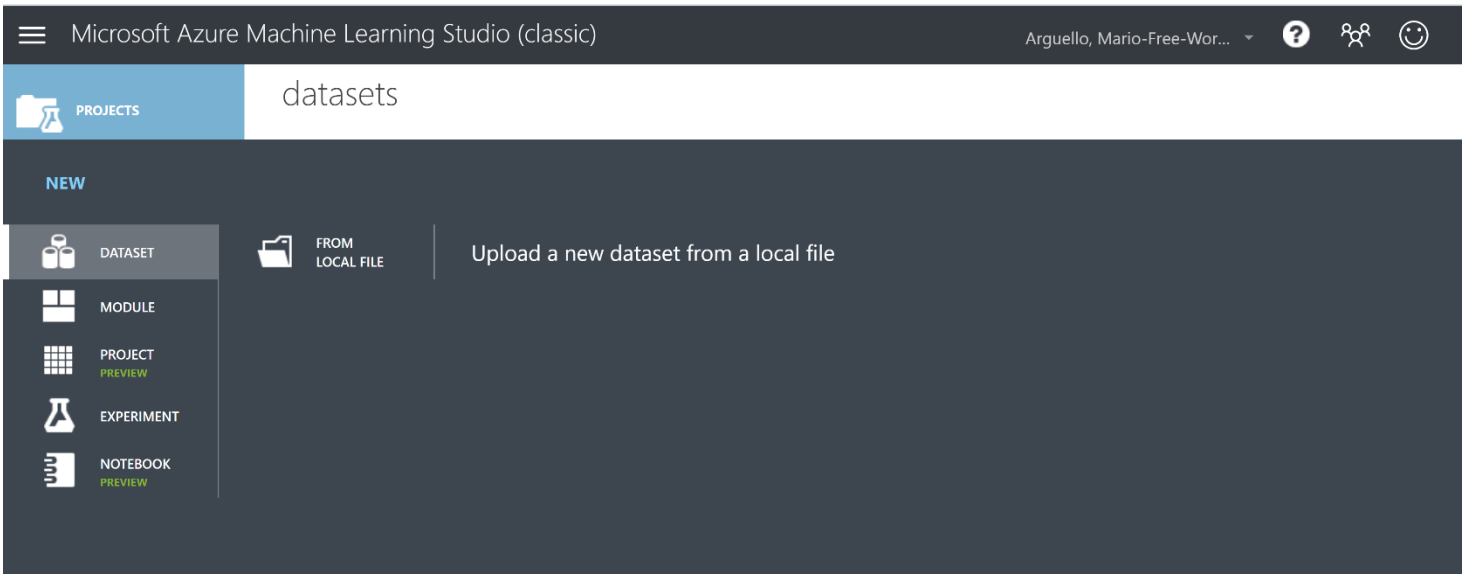
- Data Set size: 81.9 Kilobytes

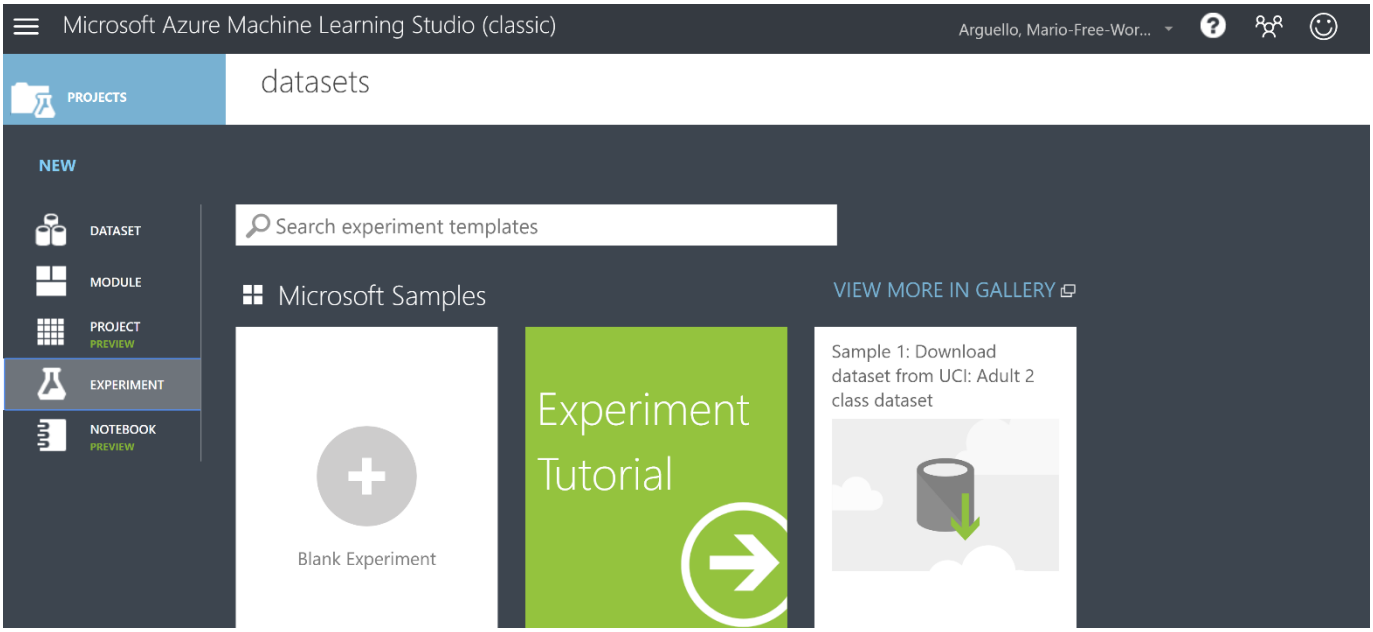
Step 1: Get data manually by downloading via website

This step is to get data manually form the California Department of Public Health called HCP Flu Vaccination by Hospital and County.

1. Go to website: https://data.ca.gov/dataset/health-care-personnel-influenza-vaccination.
2. Download csv file titled HCP Influenza Vaccinations by Hosp & County 2017-2018, see below.
3. Sign into you MS Azure ML Studio account.
4. Upload the data set you downloaded by clicking on “from local file.”
5. Create an Experiment by clicking “+ New” icon again and name the experiment “HCP Flu Vaccination Project.



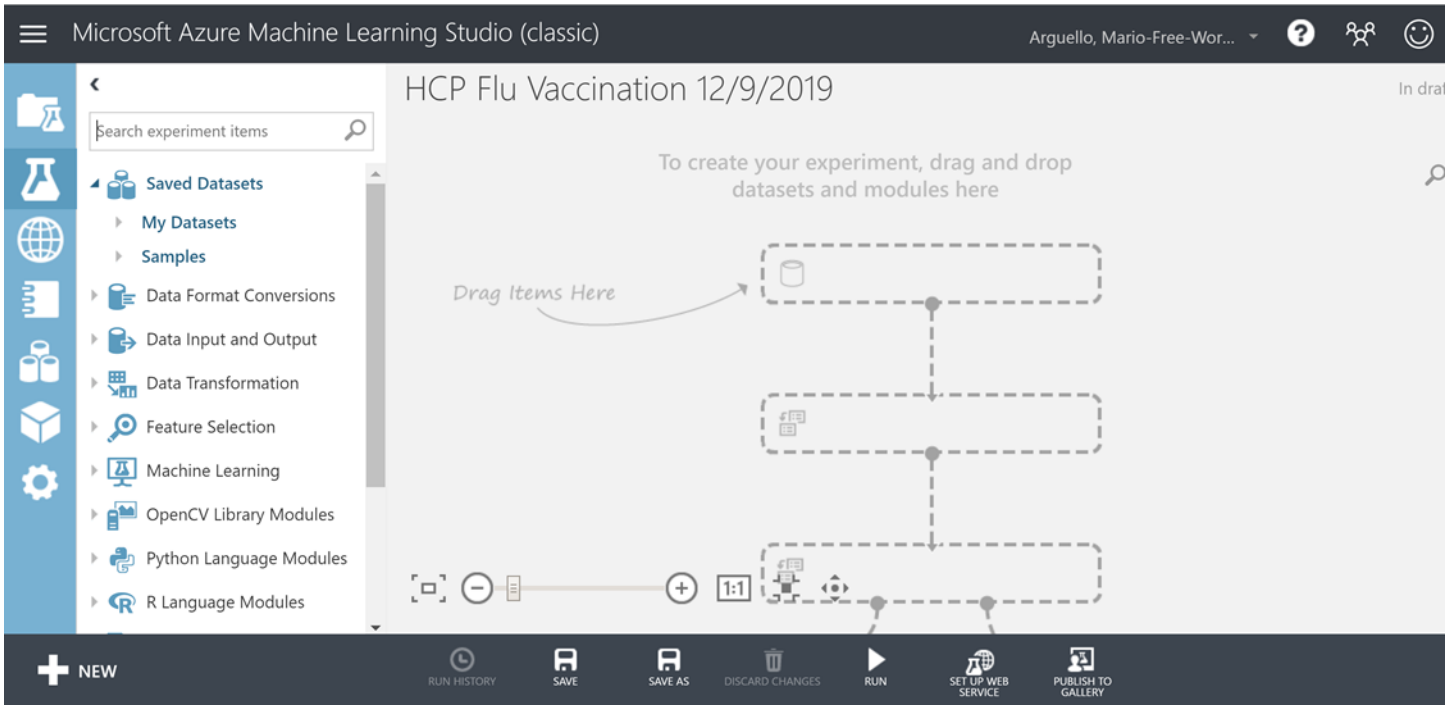




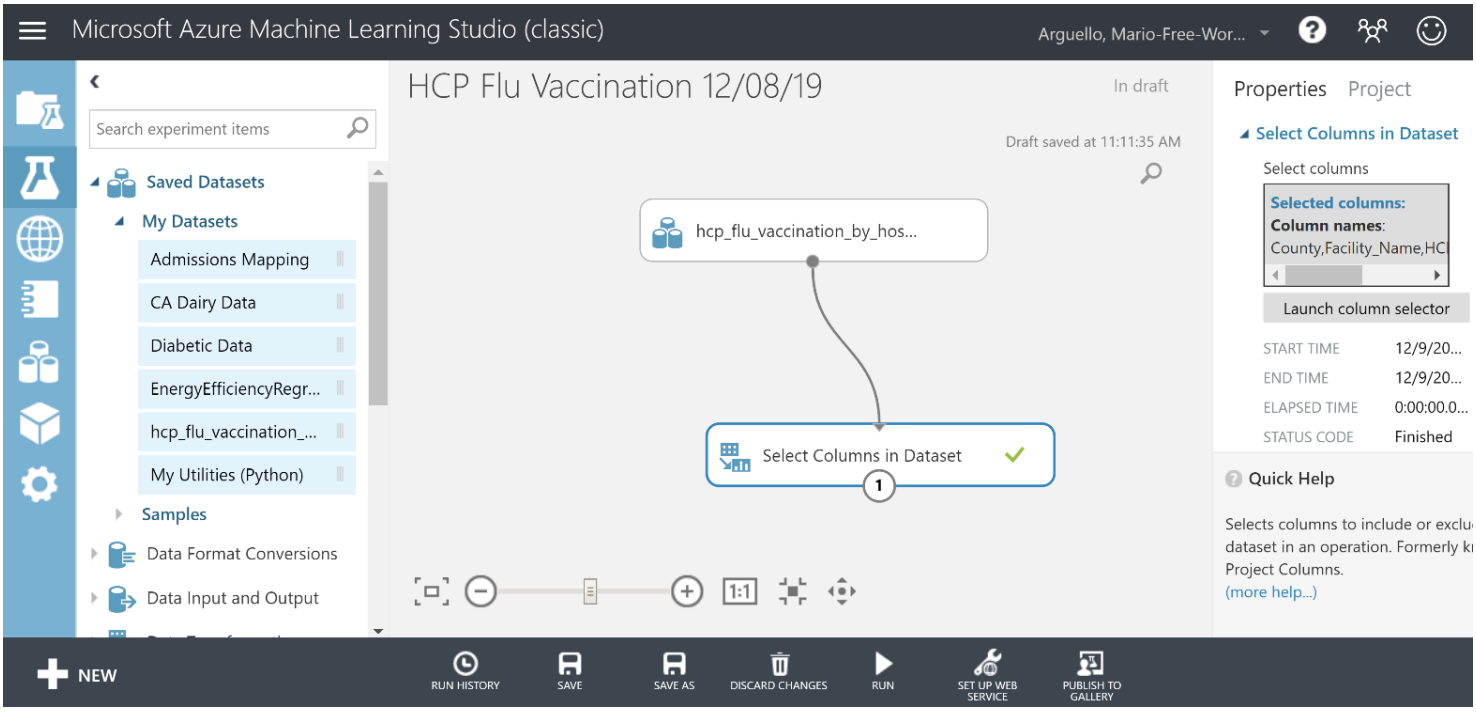
Step 2: Drag Modules and using Bayesian Linear Regression

In this step you will drag the data set and drag the modules that will be used to find the linear regression and look for prediction insights of numbers of patients who received flu vaccines in every county from California during the flu season of 2017-2018.

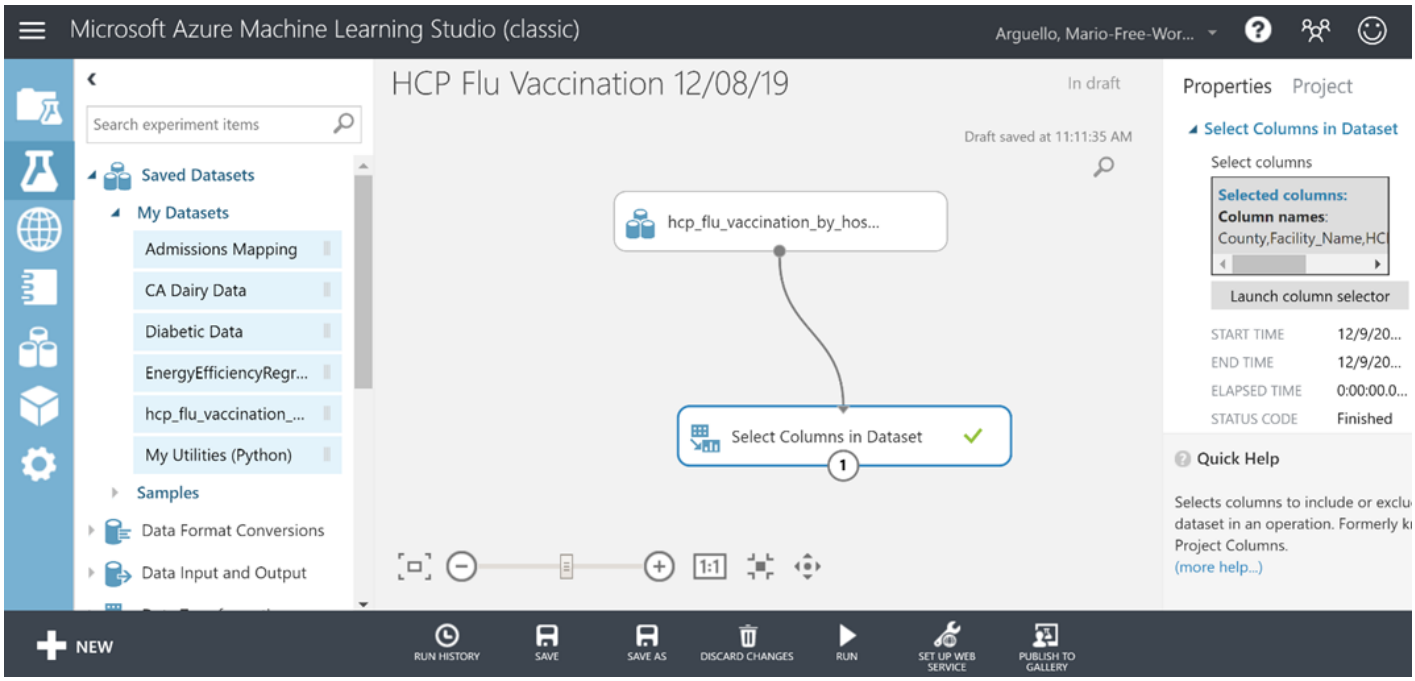
6. Search for the data set name “HCP Flu…” module on the top left in a search bar and drag it to the center on the experiment you created called “HCP Flu Vaccination Project.”



7. Search “Select Columns in Data Set” and drag it to the center, like the first module you dragged for the data set. Click “launch the column selector”, by all columns, click by name, select County, Facility Name, and HCP_Percent_Vaccinated.



8. Connect from HCP flu data set module to the Select Columns Data Set. Save and run the experiment.

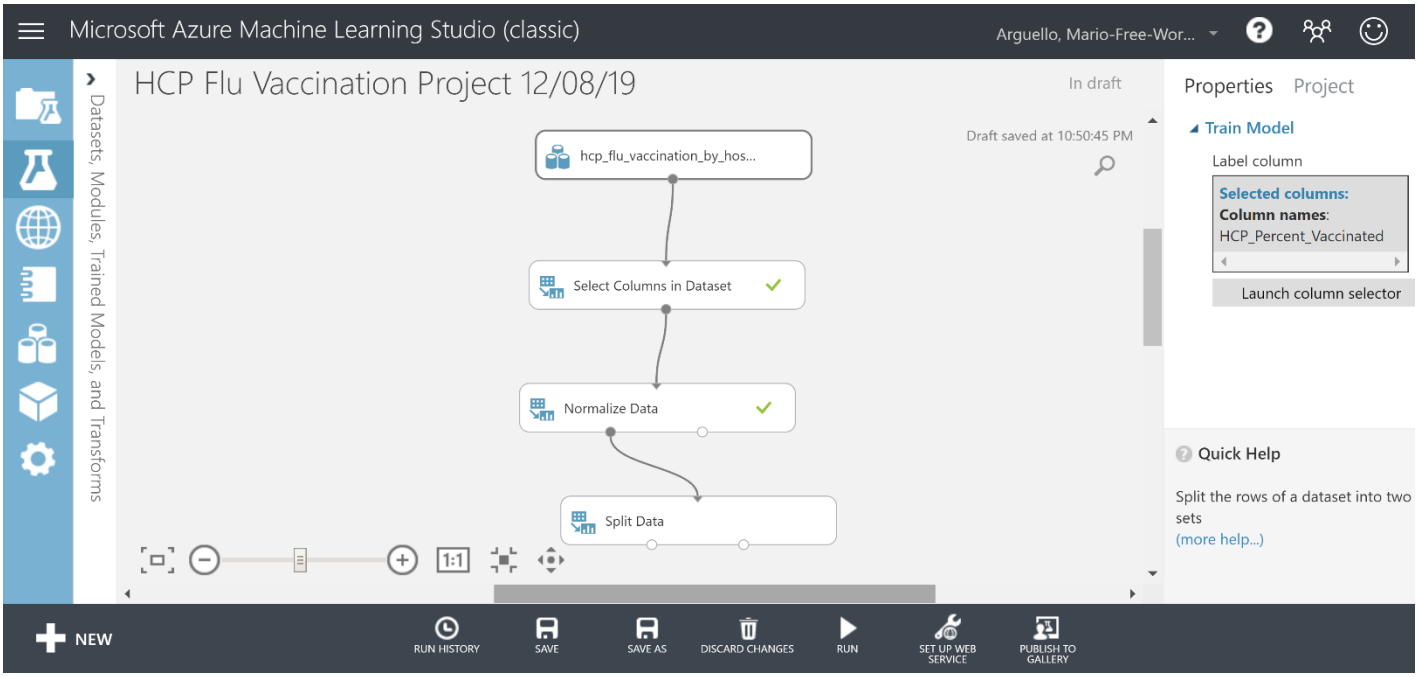


9. Search “Normalize Data” and “Split Data” and drag them to the center pane.

10. On the properties pane of “Normalize Data,” on transformation method select logistic and click launch selector, select no columns, and only select Facility ID.

11. On Split data make sure the following from the properties pane stays the same: Splitting Mode: Split rows, fraction of rows: 0.5, random speed: 0, and stratified split: false.

12. Connect the modules from “Select Data Set” to the “Normalize data.” Then connect from “Normalize Data” module to the “Spilt Data” module. Save and run the experiment.

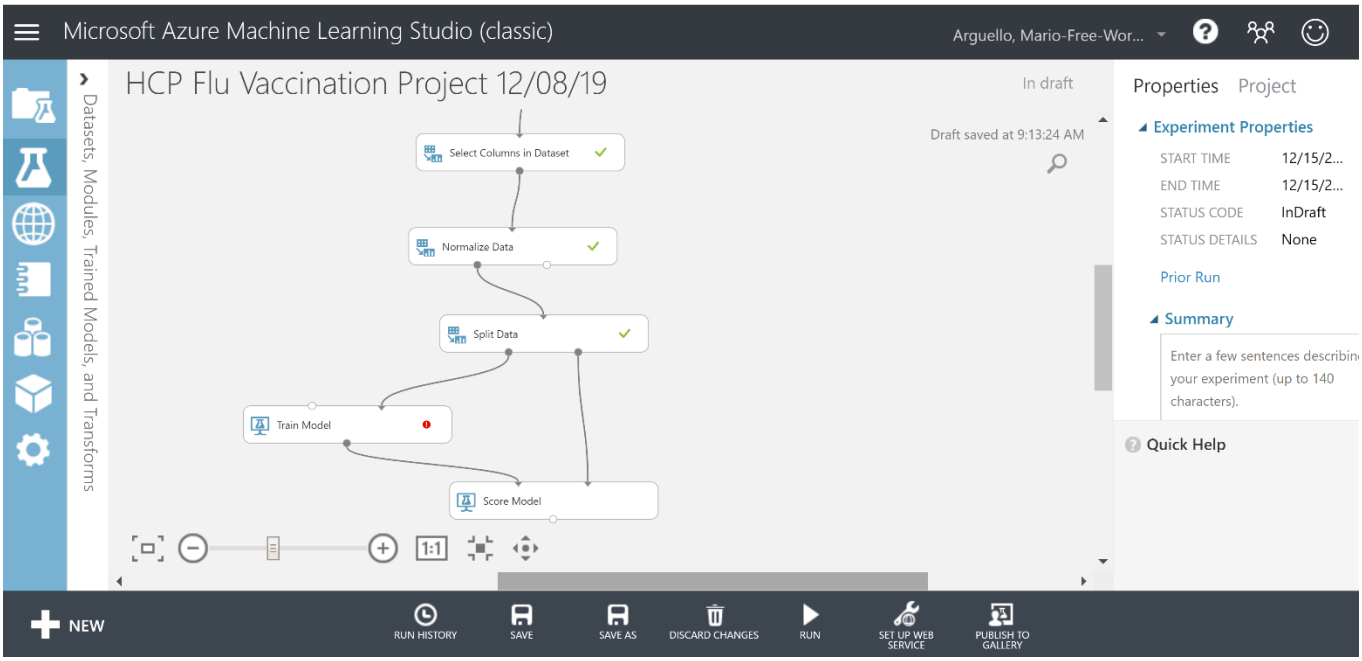


13. Search for “Train Model” and “Score Model” modules and drag them to the center pane.

14. On the properties pane of “Train Model” launch column selector, click by rules, include column name and search for “HCP_Percent_Vaccinated.”

15. Select the “Score Model” module and on the properties pane the “append score columns…” output should be checked.

16. From the “Split Data” module you will connect to the right most dot called “data set.” From “Split Data” you will connect to the right most dot of the “Score Model” module called “data set.” Save and run the experiment. Your image should now look like the following below.



17. The reason you see a red error dot on the “Train Model” module is because you need a linear regression module in order to train the model. The error says that the input port untrained model is unconnected.

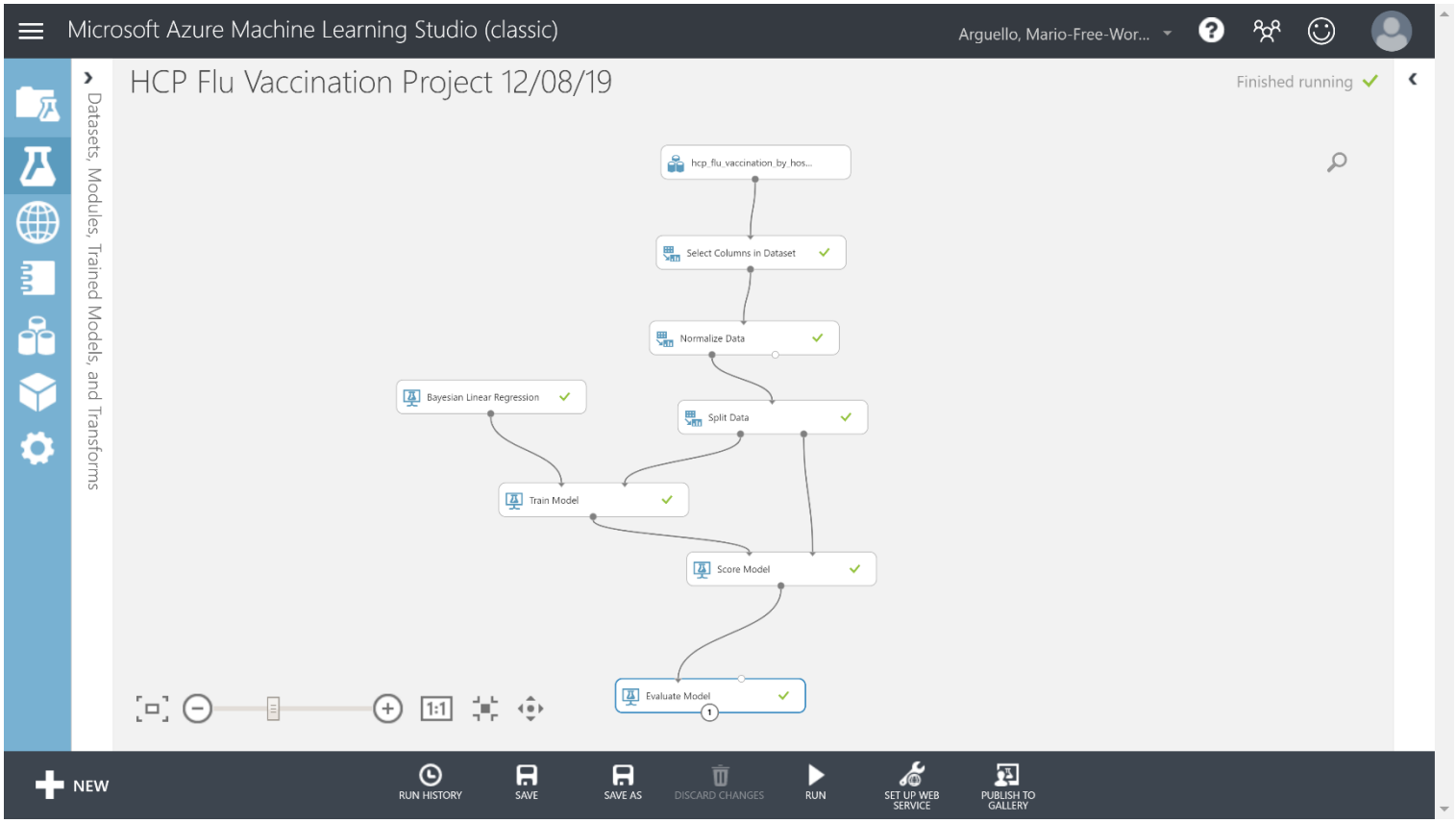
18. Search “Bayesian Linear Regression” model and drag it above the “Train Model” module.

19. On the properties pane make sure that the “regularization weight” is set to 1 and “allow unknown…” is checked.

20. Now connect the “Bayesian Linear Regression” module to the left most dot of the “Train Model” called “untrained model.”

21. Finally, search the “Evaluate Model” module and drag it under the “Score Model” module. The properties pane should say that there is no parameters.

22. From the “Score Model” you will connect to the left most dot of the evaluate model called “scored data set” Save and run the experiment. Your experiment should finally look like the image below.

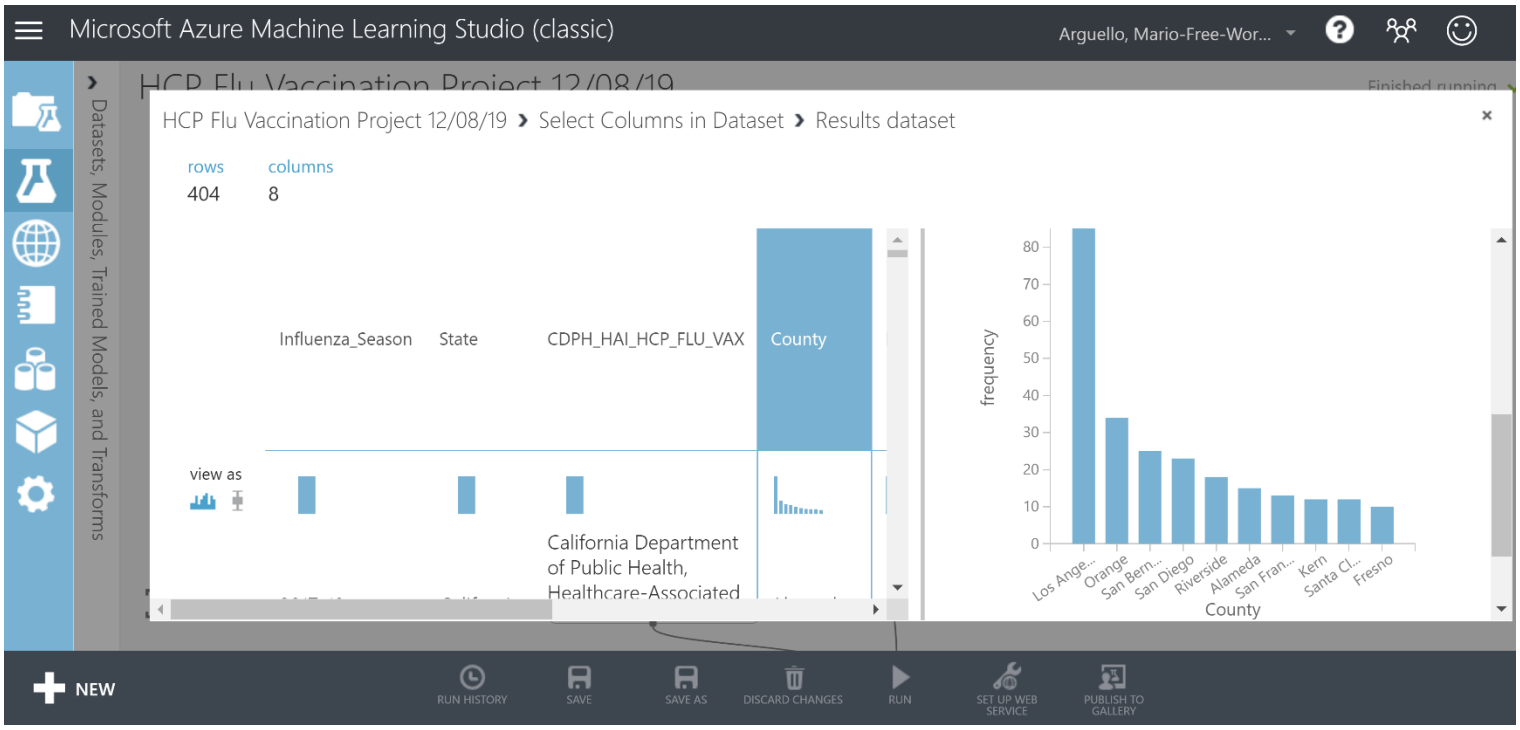


Step 3: Visualizations

This step is to add visualizations and insights form the Bayesian linear regression from the experiment you created and hopefully to accurately make predictions based on the data set.

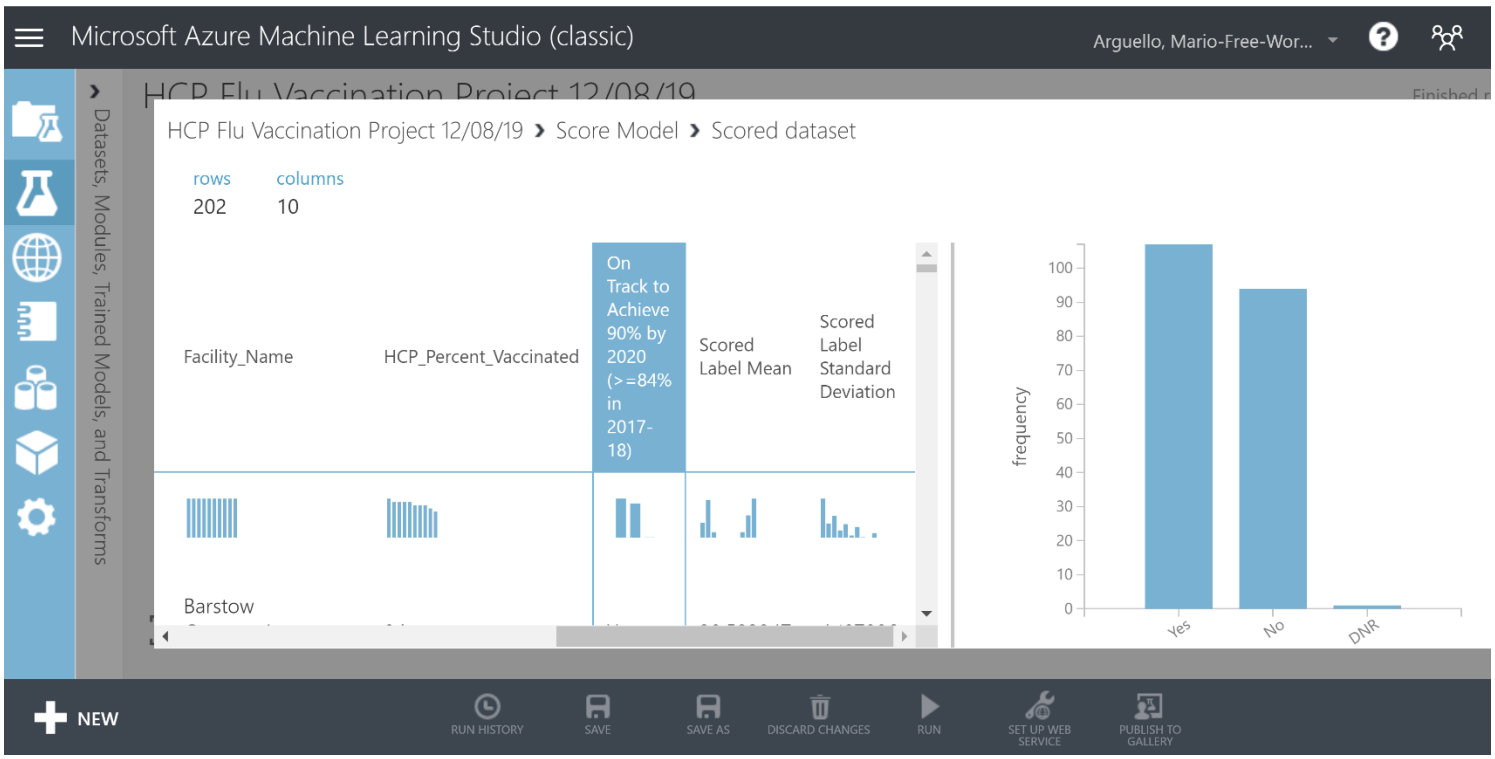
23. Please select and right click “Select Columns in Data Set” module. Go to “results data set” and click visualize. You will see some simple statistics of the number of patients who got vaccinated in every county for the flu season 2017-2018.

24. If you click on the county column you can see percentages of the patients getting vaccinated in a histogram chart below.



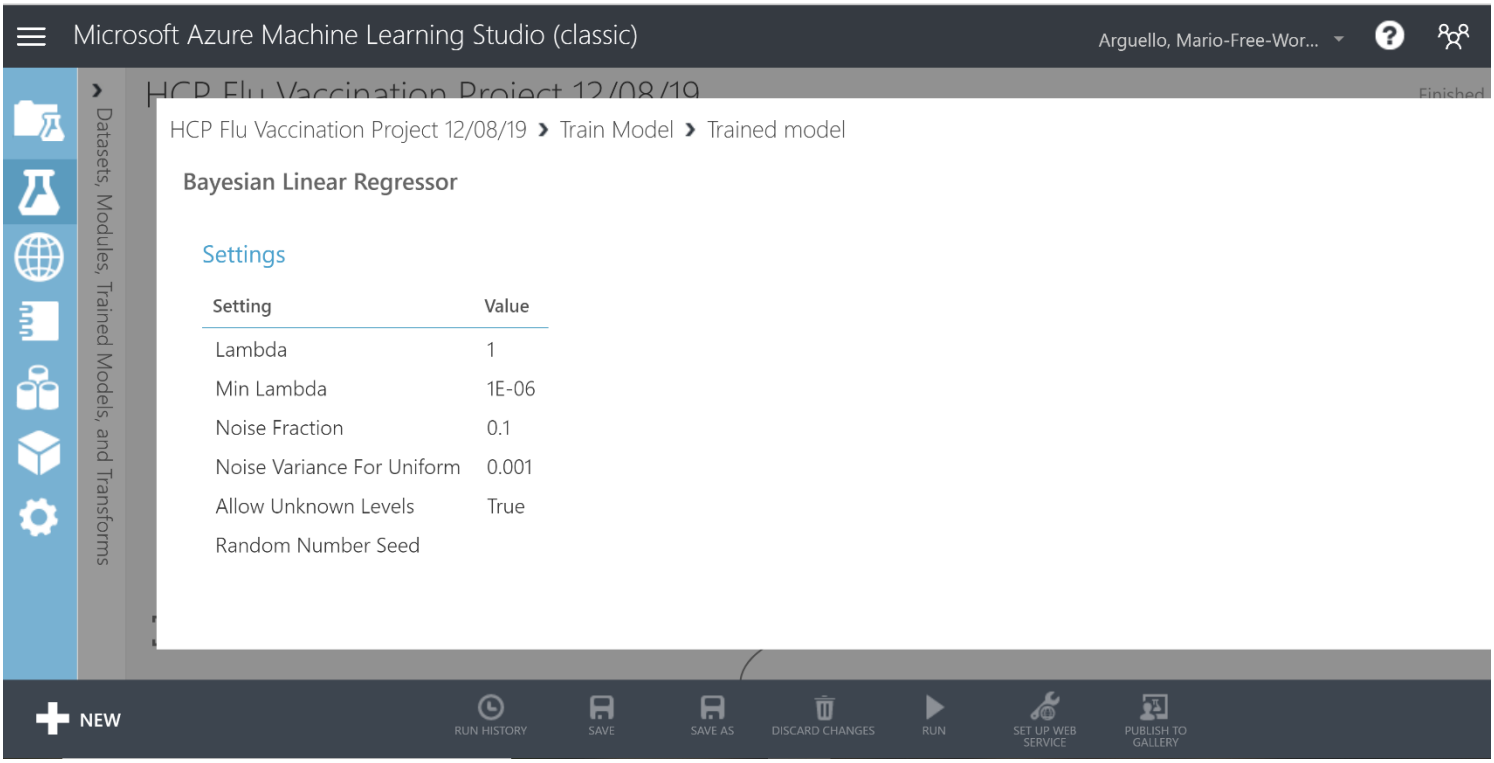
25. Select “Score Model” module, right click, scored data set and click visualize.

26. Click the column “on track to achieve 90% 2020” and you will see the surveys of the patients who answered yes or no in the histogram to the right. DNR means the hospital did not report.



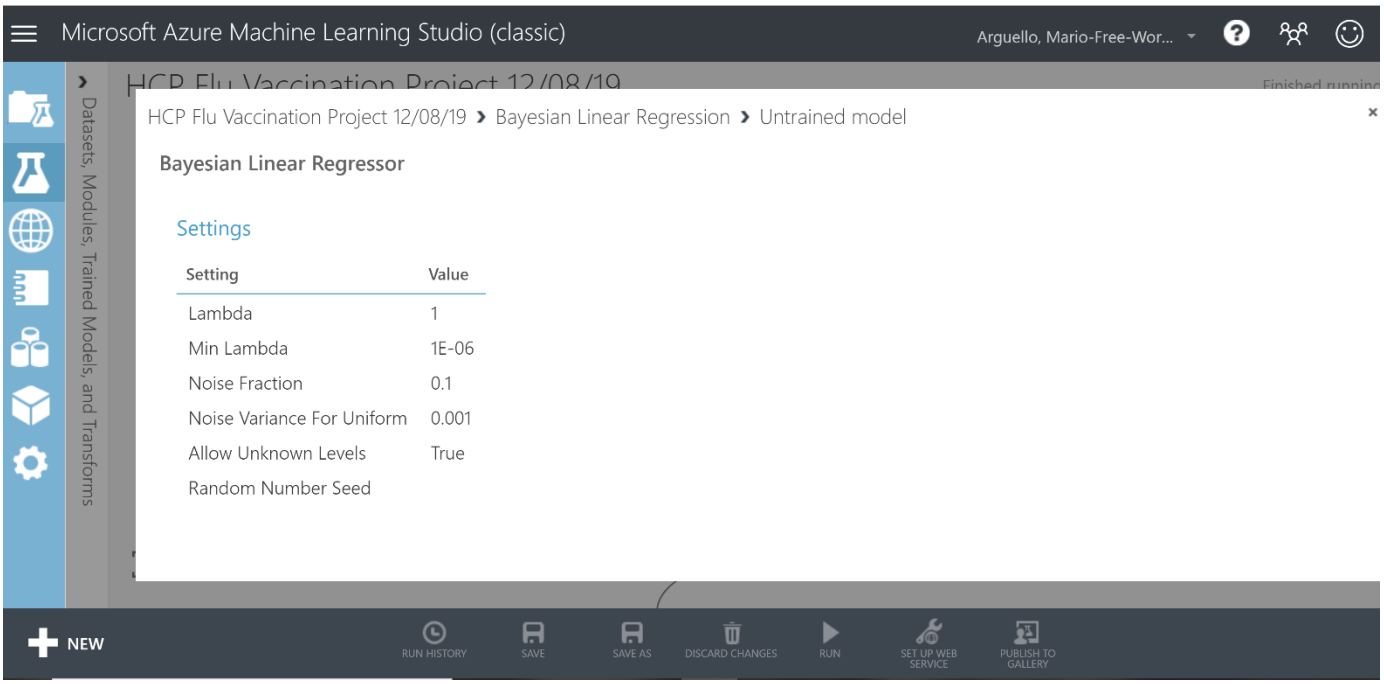
27. Select the “Train Model” module, right click, and on trained model click visualize.

28. You will see some numbers under the Bayesian Linear Regressor in the trained model form.



29. Now select “Bayesian Linear Regression” module, right click and on “untrained model” click visualize.

30. Now you will see regression prediction numbers as an untrained model form. In the previous step you saw the regression prediction statistics in trained model form. The image you see should resemble ours below.



31. Lastly, click the “Evaluate Model” module, right click, and select evaluation results and click visualize.

32. Now you see the evaluation result of this experiment. If you click the “negative likelihood” column you can see the number of percentages of incorrect predictions. You can also see the mean absolute error, relative absolute error, and the coefficient of determination of this experiment you and I created from all patients who will get vaccinated in 2020 just from gathering the data set from every hospital in every county in California.

33. This concludes this lab experiment. Feel free to check more insights from this experiment and you will see the progress of how his data set was analyzed and predicted using the Bayesian Linear Regression analysis in data science.

References

1. URL of Data Source, https://data.ca.gov/dataset/health-care-personnel-influenza-vaccination

2. GitHub URL, https://github.com/bulldog19999/DataProcessing_FluVaccinations

3. Bayesian Linear Regression Article, https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/bayesian-linear-regression

4. California Department of Public Health Website, https://data.ca.gov/dataset

5. Microsoft Azure Machine Learning Studio tool used to create this tutorial, https://studio.azureml.net/