Data Science and Machine Learning Essentials

Lab 2A – Acquiring Data in Azure Machine Learning

By Stephen Elston and Graeme Malcolm

Overview

In this lab, you will learn how to upload data files to Microsoft Azure Machine Learning (Azure ML), and how to use the **Join** module to combine data from multiple sources.

Note: This lab builds on knowledge and skills developed in Lab 1: *Getting Started with Azure ML*. If you have little experience with Azure ML, and you did not complete Lab 1, you are advised to do so before attempting this lab.

What You'll Need

To complete this lab, you will need the following:

- An Azure ML account
- A web browser and Internet connection
- The lab files for this lab

Note: To set up the required environment for the lab, follow the instructions in the **Setup** document for this course. Then download and extract the lab files for this lab.

Uploading a Data File to Azure ML

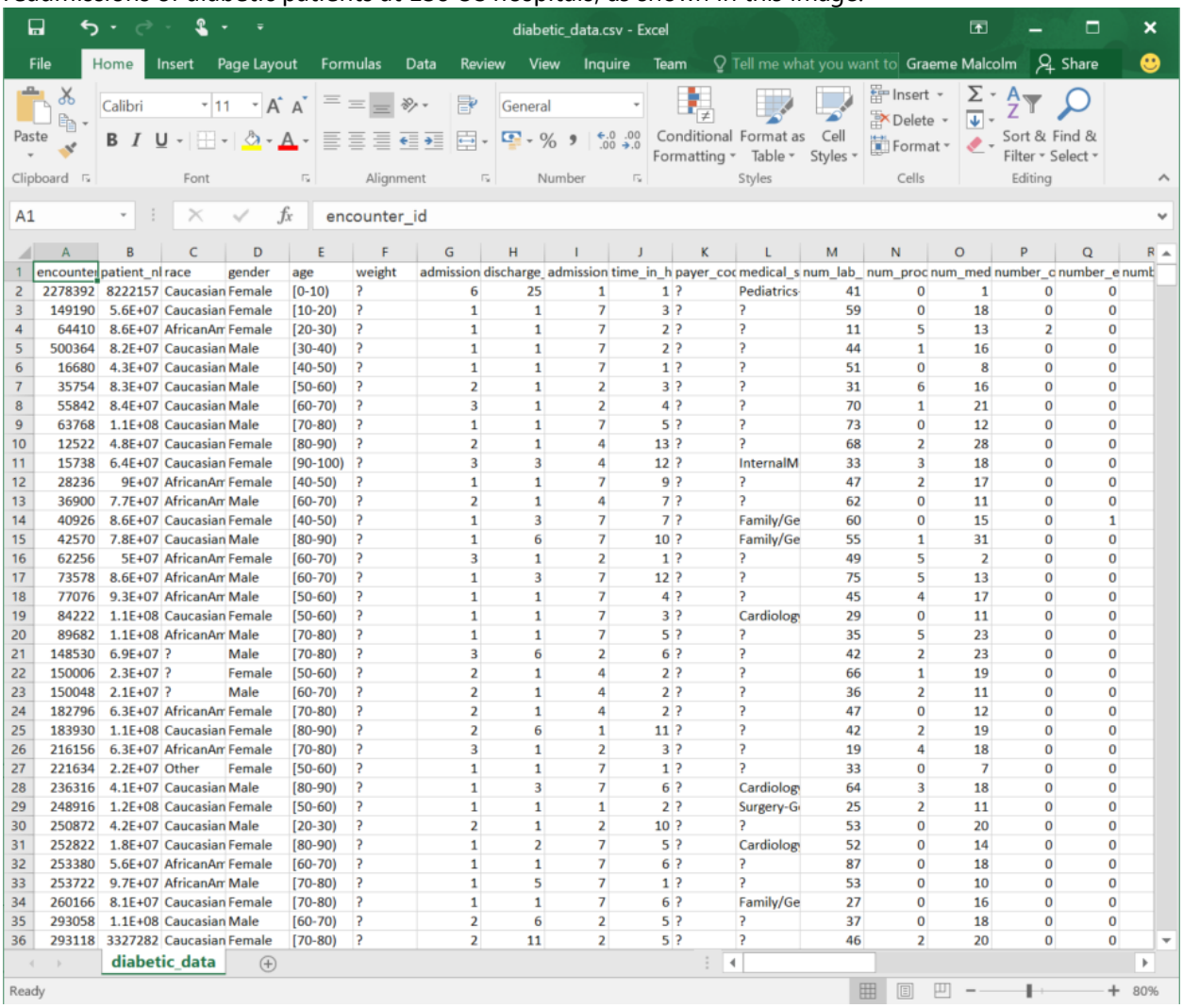
When you need to create an experiment based on your own data, or data you have obtained from a third-party, you must begin by uploading the data to Azure ML. In this exercise, you will upload a file containing data about diabetes patient appointments.

Note: The data used in this exercise was obtained from the University of California machine learning repository.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Examine the Data

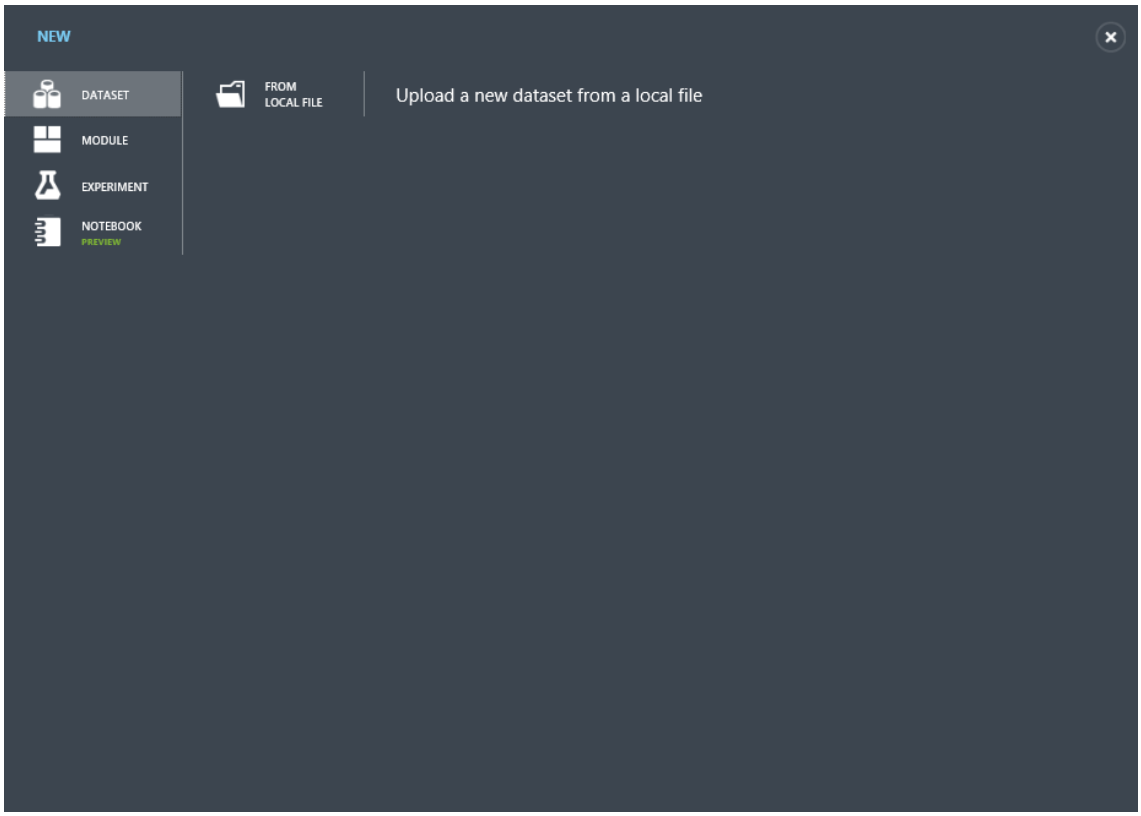
1. Open the **diabetic_data.csv** file in the folder where you extracted the lab files, using either a spreadsheet application such as Microsoft Excel, or a text editor such as Microsoft Windows Notepad.
2. View the contents of the file, noting that it contains data on over 101,000 admissions and readmissions of diabetic patients at 130 US hospitals, as shown in this image:



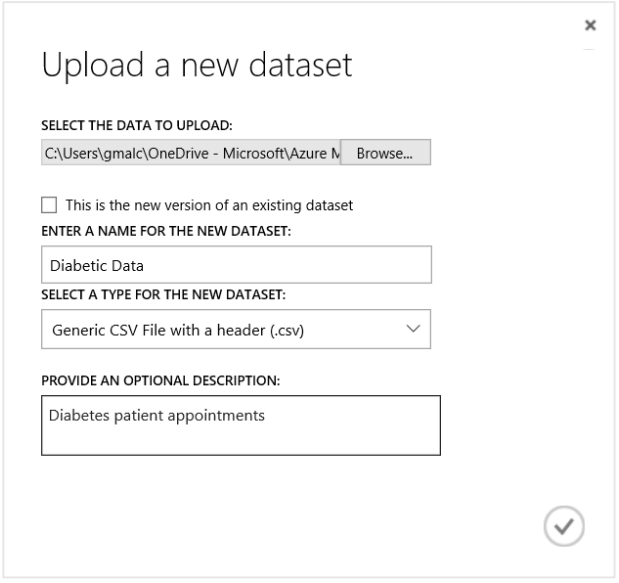
3. Close the data file without saving any changes.

Upload the Data File to Create a New Dataset in Azure ML

1. Open a browser and browse to https://studio.azureml.net. Then sign in using the Microsoft account associated with your Azure ML account.
2. Create a new blank experiment with the title **Diabetes Data**.
3. With the **Diabetes Data** experiment open, at the bottom left, click **NEW**. Then in the **NEW** dialog box, click the **DATASET** tab as shown in the following image.



4. Click **FROM LOCAL FILE**. Then in the **Upload a new dataset** dialog box, browse to select the **diabetic_data.csv** file from the folder where you extracted the lab files on your local computer and enter the following details as shown in the image below, and then click the OK icon.
    - **This is a new version of an existing dataset**: Unselected
    - **Enter a name for the new dataset**: Diabetic Data
    - **Select a type for the new dataset**: Generic CSV file with a header (.csv)
    - **Provide an optional description**: Diabetes patient appointments.



5. Wait for the upload of the dataset to be completed, and then on the experiment items pane, expand **Saved Datasets** and **My Datasets** to verify that the **Diabetic Data** dataset is listed.

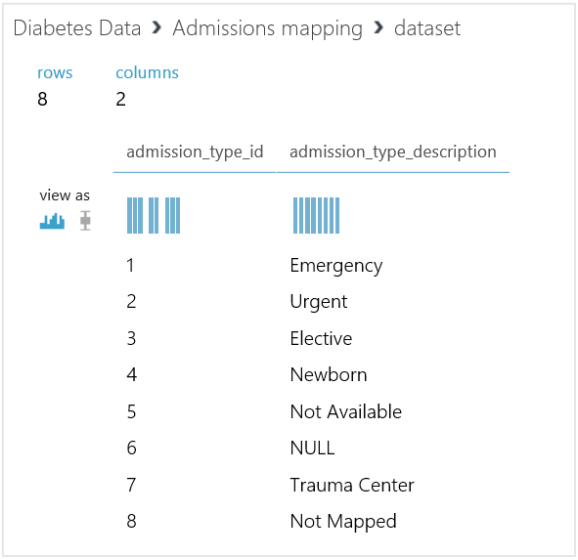
Visualize the Dataset in Azure ML

1. Drag the **Diabetic Data** dataset to the canvas for the **Diabetes Data** experiment.
2. Right-click the output port for the **Diabetic Data** dataset on the canvas and click **Visualize** to view the data in the dataset.
3. Verify that the dataset contains the data you viewed in the source file, and then close the dataset.

Using Reference Data

Occasionally, you may need to combine data from multiple data sources. In this case, the diabetic appointment data includes a code that indicates the admission type for the appointment, but it does not include a human readable description of the admission type. In this exercise, you will use the Azure ML **Join** module to combine rows in the **Diabetes Data** experiment with rows in a second dataset, which provides the admission type description values for the admission type codes.

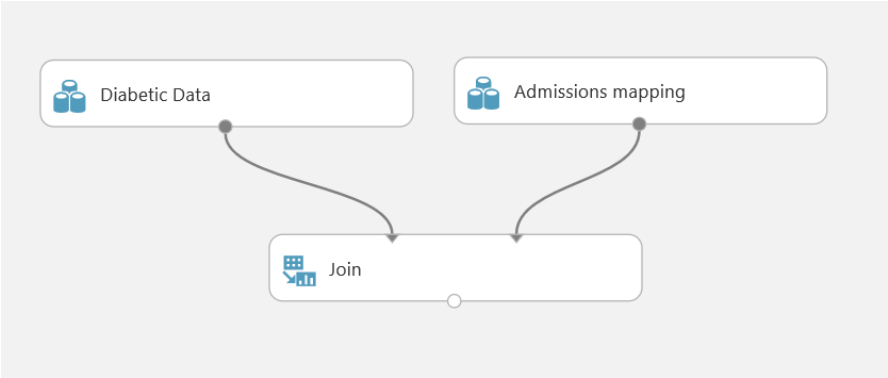
1. In Azure ML Studio, ensure that the **Diabetes Data** experiment you created in the previous exercise is open.
2. Add a new dataset by uploading the **admissions_mapping.csv** file in the folder where you extracted the lab files. Name the new dataset **Admissions Mapping**.
3. In the **Diabetes Data** Azure ML experiment, search for the **Admissions Mapping** dataset and drag it to the canvas, next to the **Diabetic Data** dataset. Do not connect it to any other modules at this point, but visualize the dataset output port and note that the dataset contains rows that map **admission_type_id** codes to **admission_type_description** text values as shown in the following image:



4. Close the dataset.
5. Search for the **Join** module, and drag it to the canvas below both the **Diabetic Data** dataset and the **Admission Mapping** dataset. Then connect the output ports from the **Diabetic Data** dataset and the **Admission Mapping** dataset to the **Dataset1** and **Dataset2** input ports of the **Join** module respectively.
6. Select the **Join** module, and in the **Properties** pane, set the following properties:
    a. **Join key columns for L**: Launch the column selector and select **admission_type_id**.
    b. **Join key columns for R**: Launch the column selector and select **admission_type_id**.
    c. **Join type**: Left Outer Join
    d. **Keep right key column**: *Unselected*

Note: The key columns in this example have the same name, but this is not a requirement to join columns from two datasets. By using a left outer join, The **Join** module will retain any rows in the left dataset (the diabetic data) that do not have a matching **admission_type_id** column in the right dataset (the admissions mapping data).

7. Verify that your experiment resembles the following image, and then save and run it.



8. When the experiment has finished running, visualize the **Results dataset** output port of the **Join** module, and verify that it contains the cleansed patient admission data from the original **Diabetic Data** dataset and the **admission_type_description** column from the **Admissions Mapping** dataset as shown in the following image.

mepiride-oglitazone	metformin-rosiglitazone	metformin-pioglitazone	change	diabetesMed	readmitted	admission_type_description
o	No	No	No	No	NO	NULL
o	No	No	Ch	Yes	>30	Emergency
o	No	No	No	Yes	NO	Emergency
o	No	No	Ch	Yes	NO	Emergency
o	No	No	Ch	Yes	NO	Emergency
o	No	No	No	Yes	>30	Urgent
o	No	No	Ch	Yes	NO	Elective
o	No	No	No	Yes	>30	Emergency
o	No	No	Ch	Yes	NO	Urgent
o	No	No	Ch	Yes	NO	Elective
o	No	No	No	Yes	>30	Emergency
o	No	No	Ch	Yes	<30	Urgent
o	No	No	Ch	Yes	<30	Emergency
o	No	No	No	Yes	NO	Emergency
o	No	No	No	Yes	>30	Elective
o	No	No	Ch	Yes	NO	Emergency
o	No	No	Ch	Yes	<30	Emergency
o	No	No	No	Yes	NO	Emergency

9. Close the results dataset.
10. If necessary, save the experiment. Then close your browser. The datasets you uploaded will remain available in your Azure ML account.

Summary

This lab was designed to teach you how to upload data to Azure ML and to use the Join module to combine data from multiple data sources.

Note: The experiment created in this lab is available in the Cortana Analytics library at http://gallery.cortanaanalytics.com/Collection/5bfa7c8023724a29a41a4098d3fc3df9.