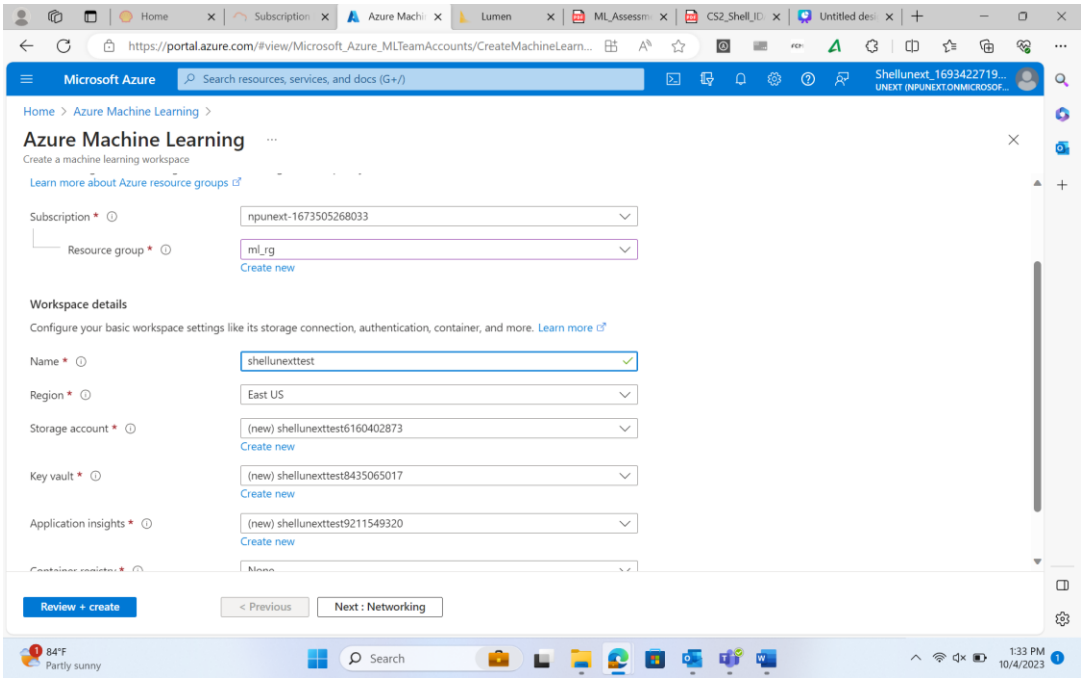
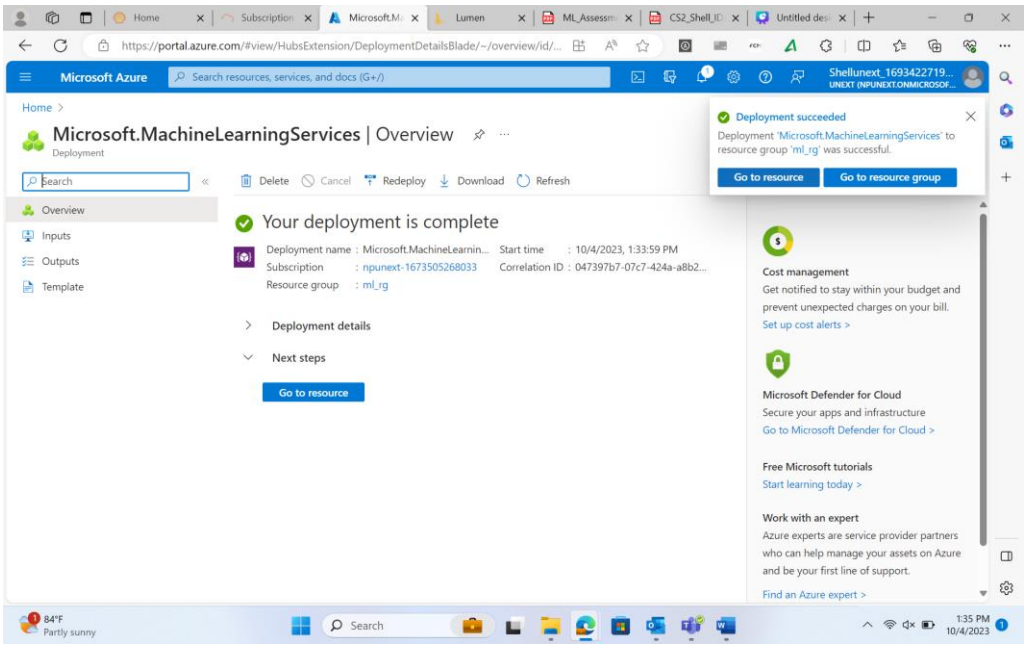
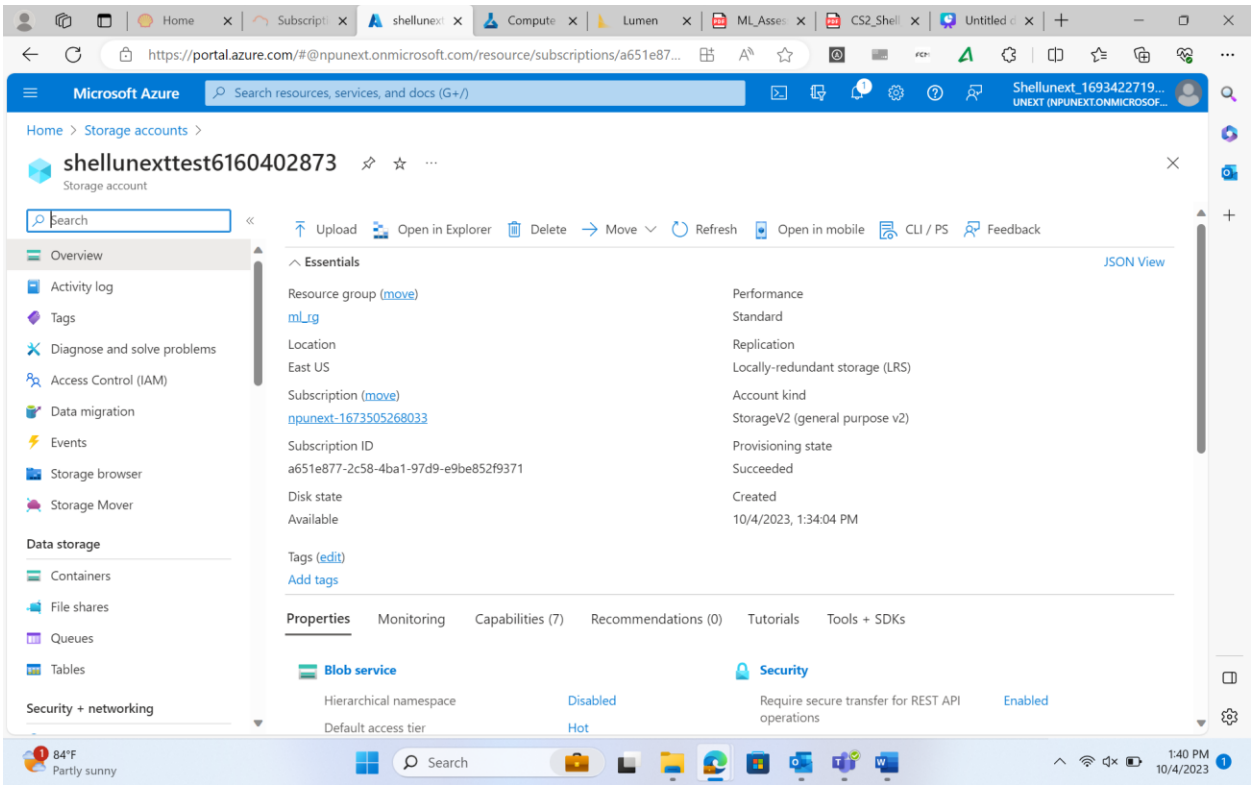
Name: Indushree Manjunatha Hegde

Email: Indushree.Hegde@shell.com

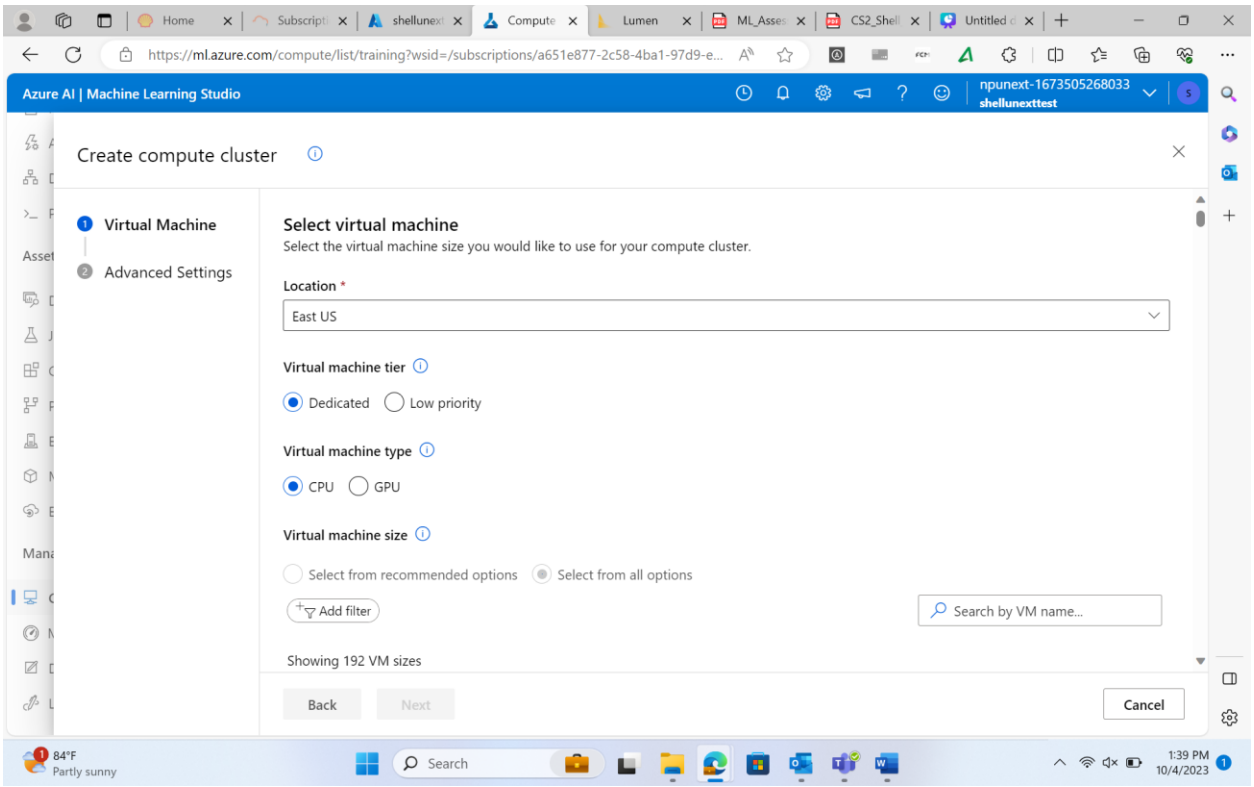
Creating Azure Machine Learning :

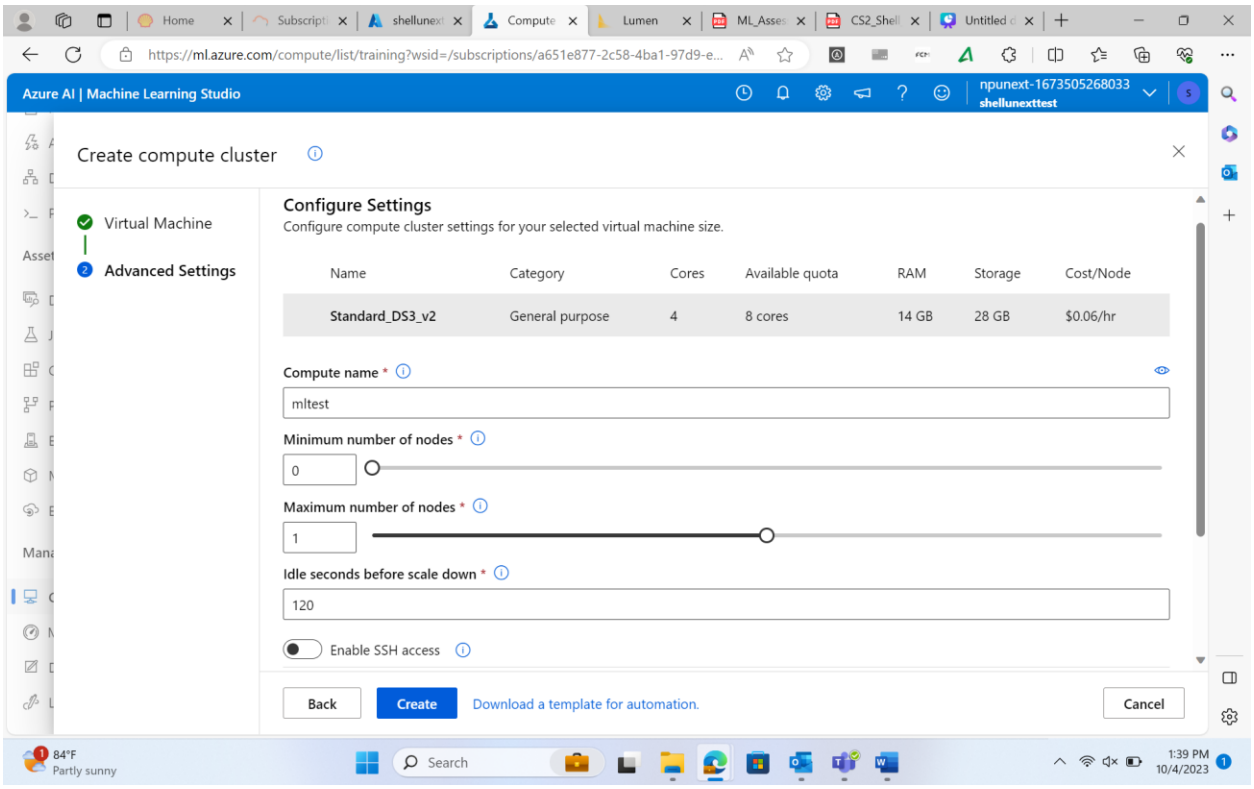


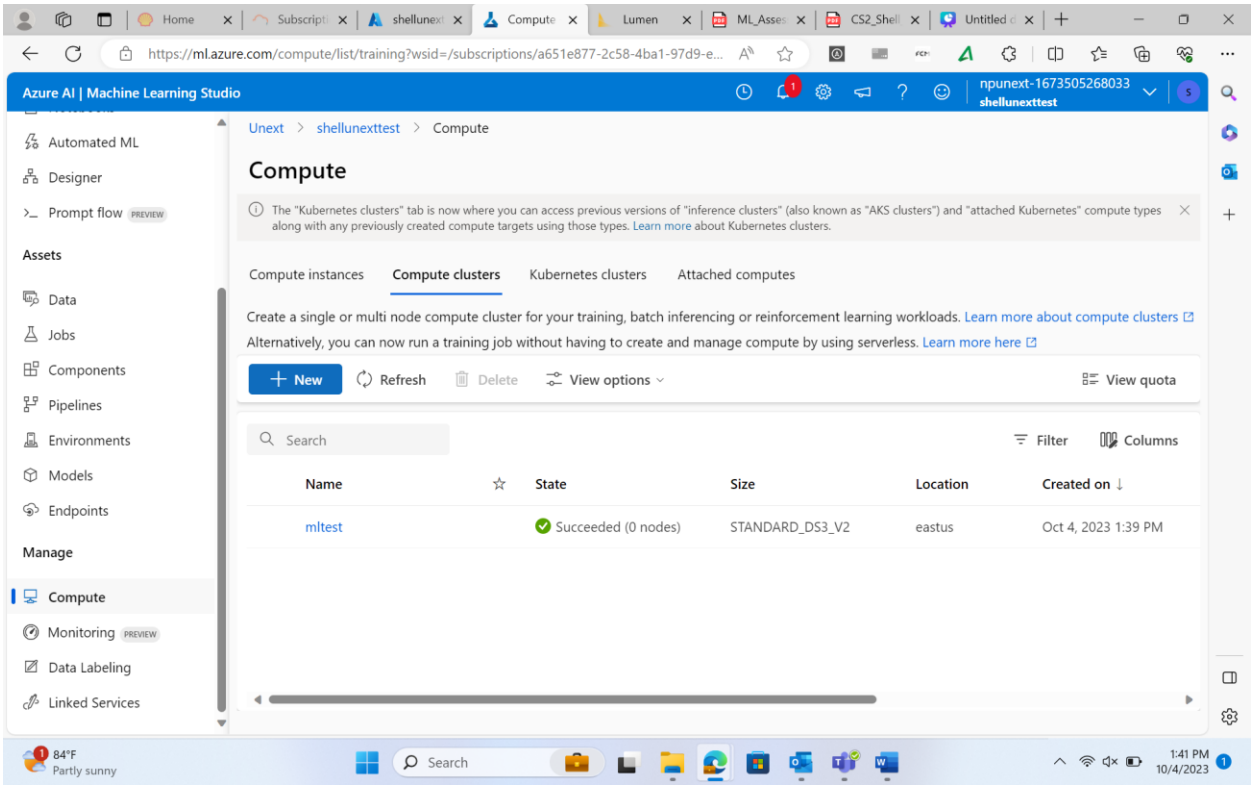




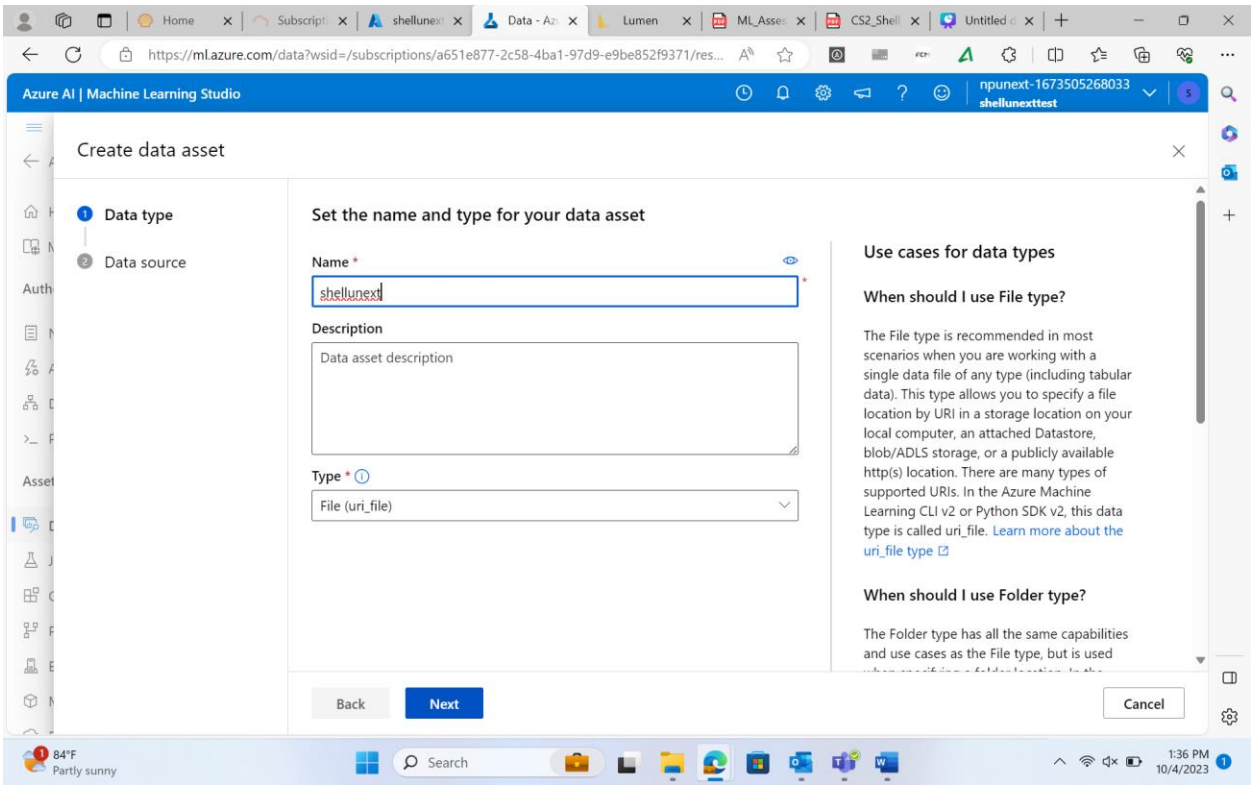
1.Data Preparation

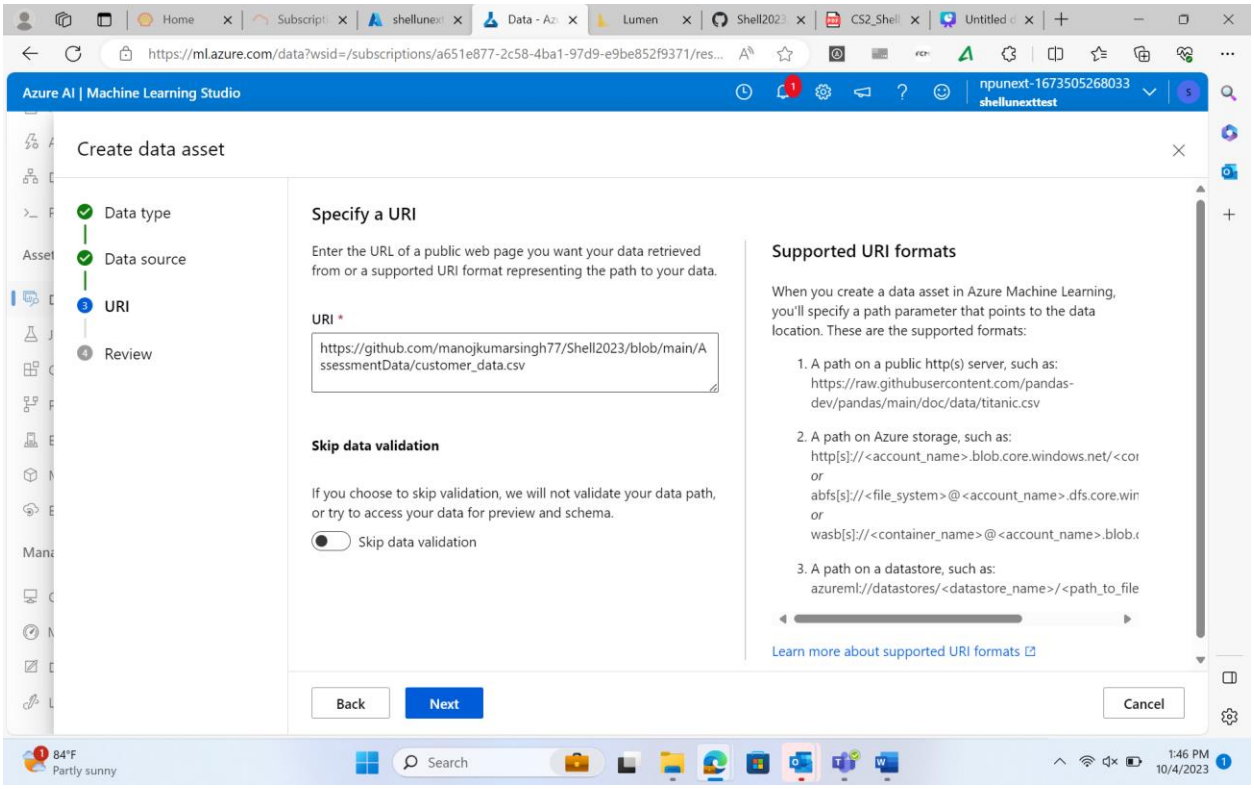


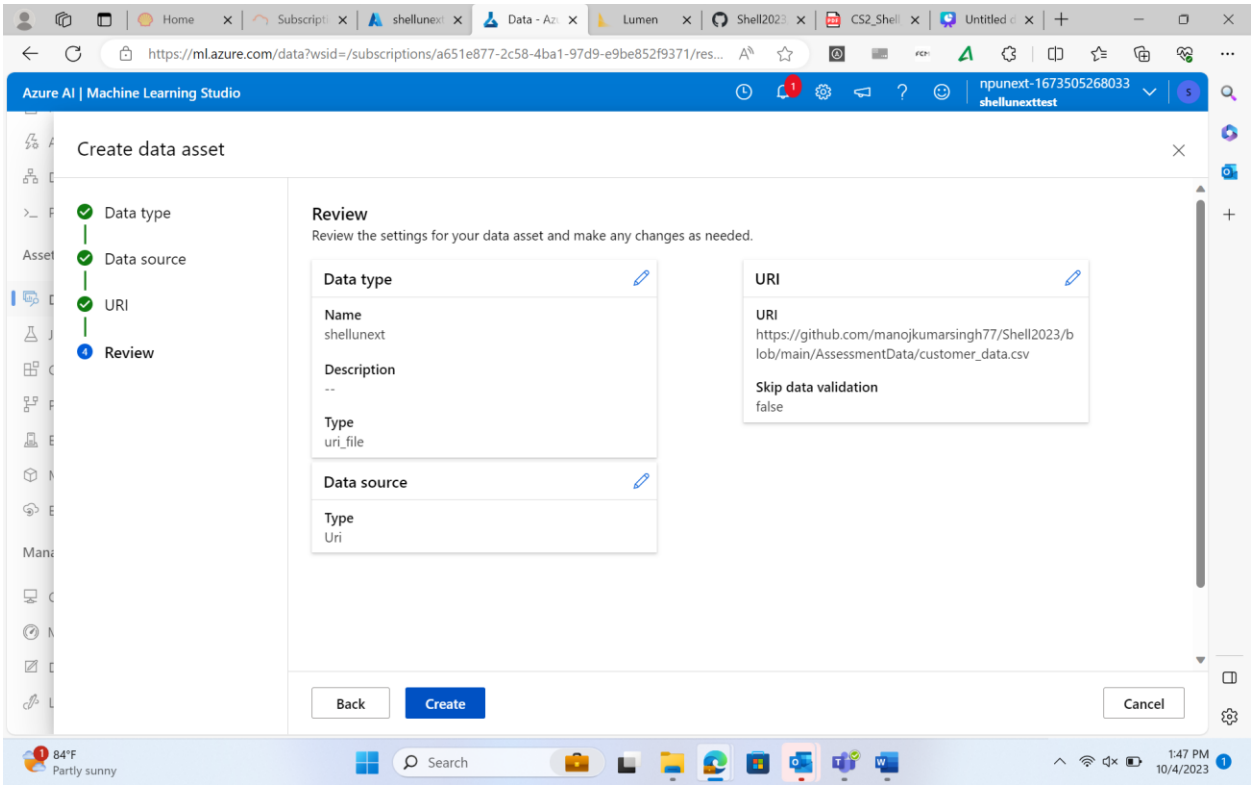


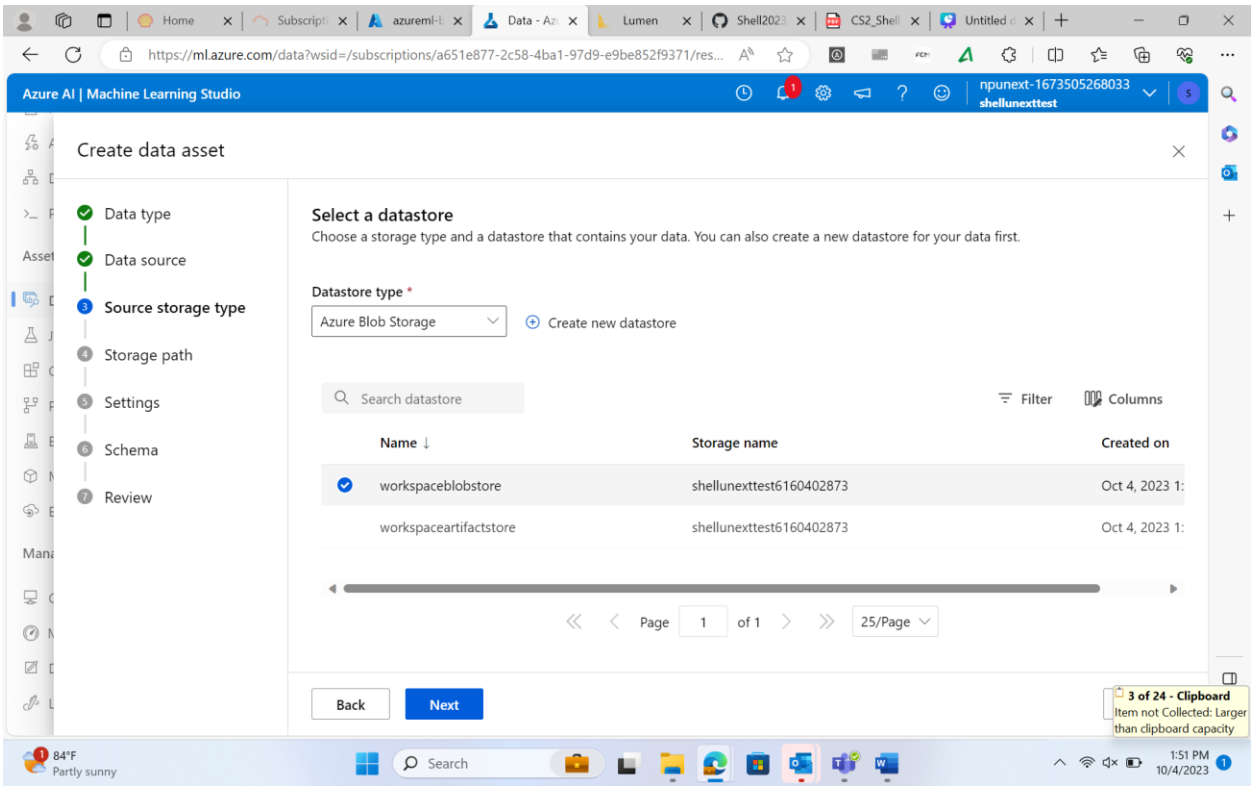


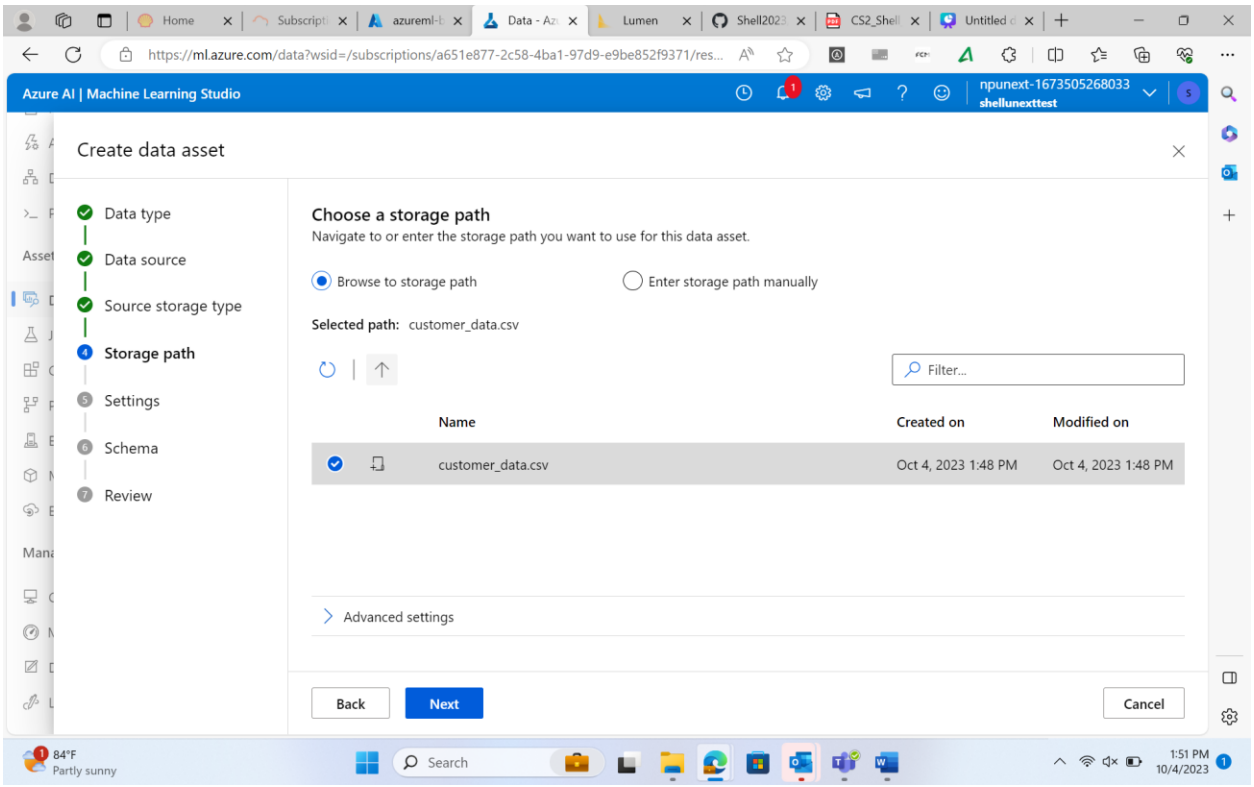
Creating the data assets:

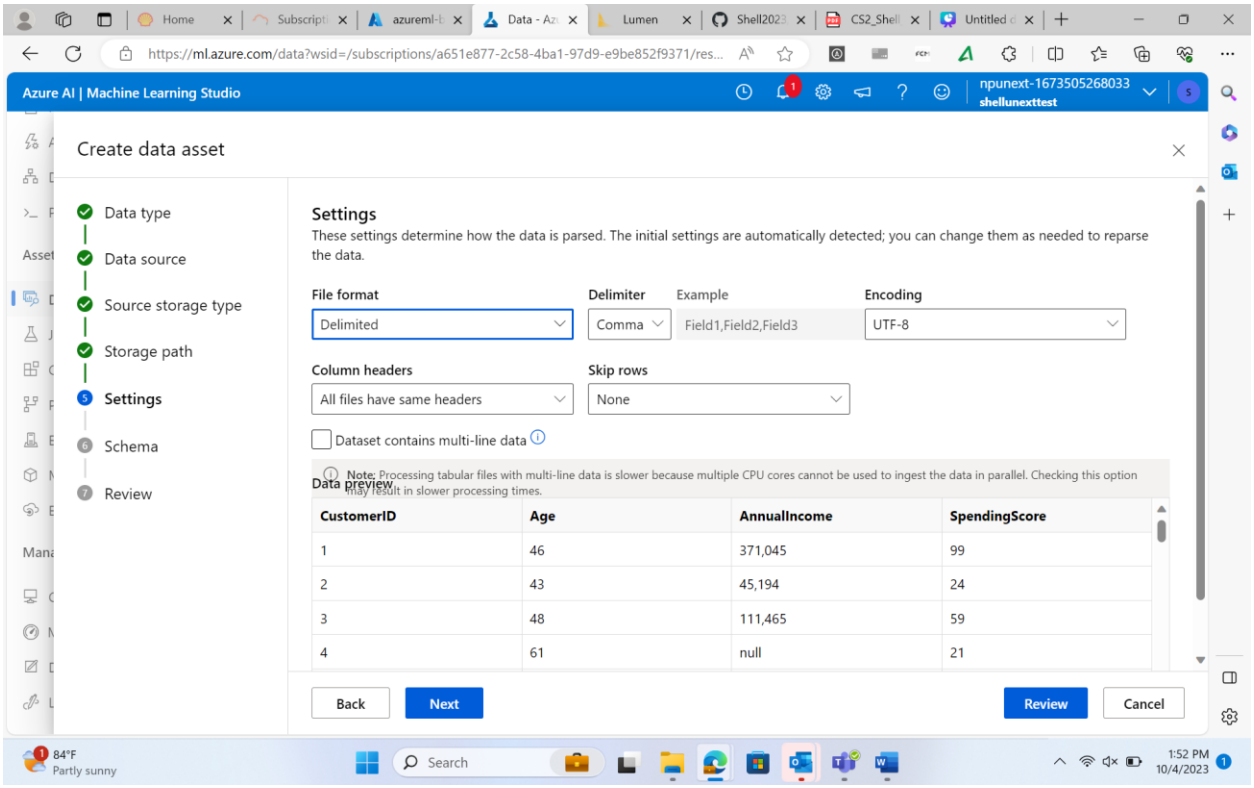

Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Create data asset
Data type
Data source
Set the name and type for your data asset
Name *
shellunext
Description
Data asset description
Type *
File (uri_file)
Use cases for data types
When should I use File type?
The File type is recommended in most scenarios when you are working with a single data file of any type (including tabular data). This type allows you to specify a file location by URI in a storage location on your local computer, an attached Datastore, blob/ADLS storage, or a publicly available http(s) location. There are many types of supported URIs. In the Azure Machine Learning CLI v2 or Python SDK v2, this data type is called uri_file. Learn more about the uri_file type
When should I use Folder type?
The Folder type has all the same capabilities and use cases as the File type, but is used
Back
Next
Cancel

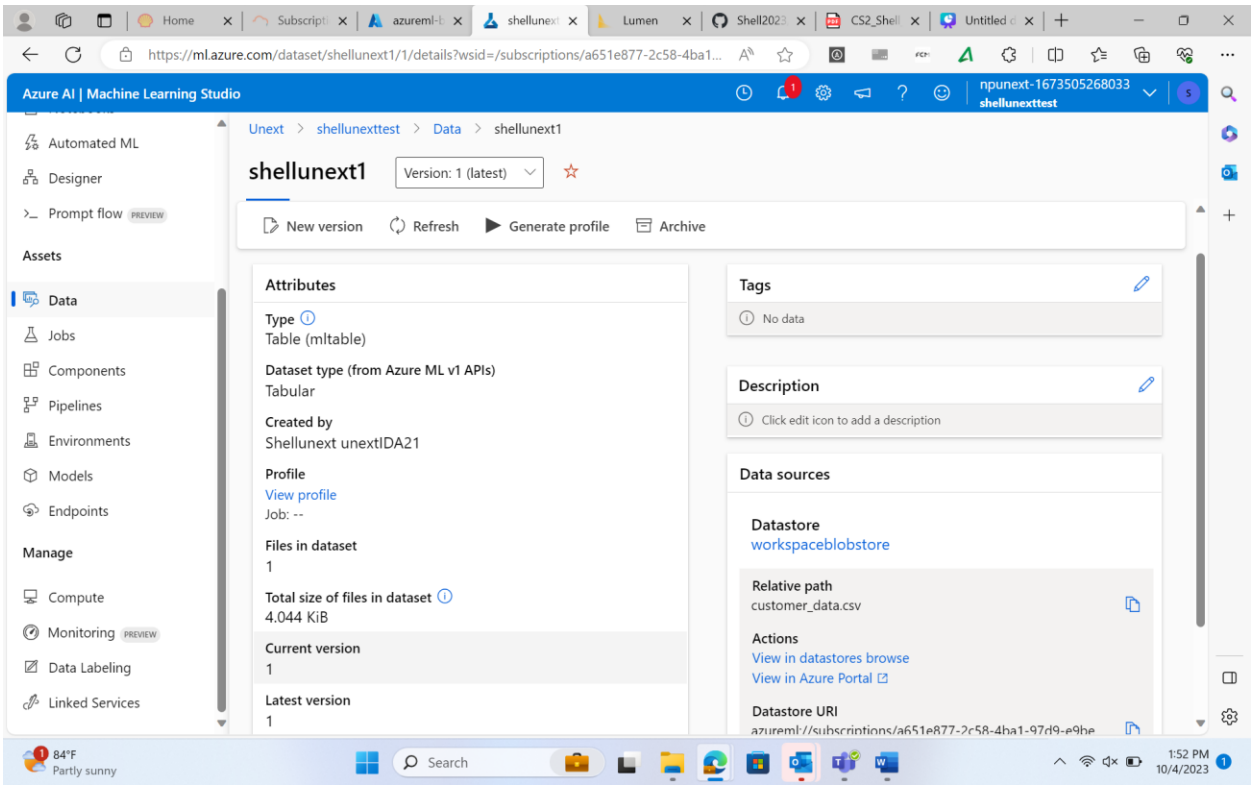

Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Create data asset
Data type
Data source
URI
Review
Specify a URI
Enter the URL of a public web page you want your data retrieved from or a supported URI format representing the path to your data.
URI *
https://github.com/manojkumarsingh77/Shell2023/blob/main/AssessmentData/customer_data.csv
Skip data validation
If you choose to skip validation, we will not validate your data path, or try to access your data for preview and schema.
Skip data validation
Supported URI formats
When you create a data asset in Azure Machine Learning, you'll specify a path parameter that points to the data location. These are the supported formats:
1. A path on a public http(s) server, such as: https://raw.githubusercontent.com/pandas-dev/pandas/main/doc/data/titanic.csv
2. A path on Azure storage, such as: http[s]://<account_name>.blob.core.windows.net/<co
or
abfs[s]://<file_system>@<account_name>.dfs.core.wi
or
wasb[s]://<container_name>@<account_name>.blob.
3. A path on a datastore, such as: azureml://datastores/<datastore_name>/<path_to_file
Learn more about supported URI formats
Back
Next
Cancel


Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Create data asset
Data type
Data source
URI
Review
Review the settings for your data asset and make any changes as needed.
Data type
Name
shellunext
Description
--
Type
uri_file
URI
URI
https://github.com/manojkumarsingh77/Shell2023/blob/main/AssessmentData/customer_data.csv
Skip data validation
false
Data source
Type
Uri
Back
Create
Cancel
84°F
Partly sunny
Search
1:47 PM
10/4/2023

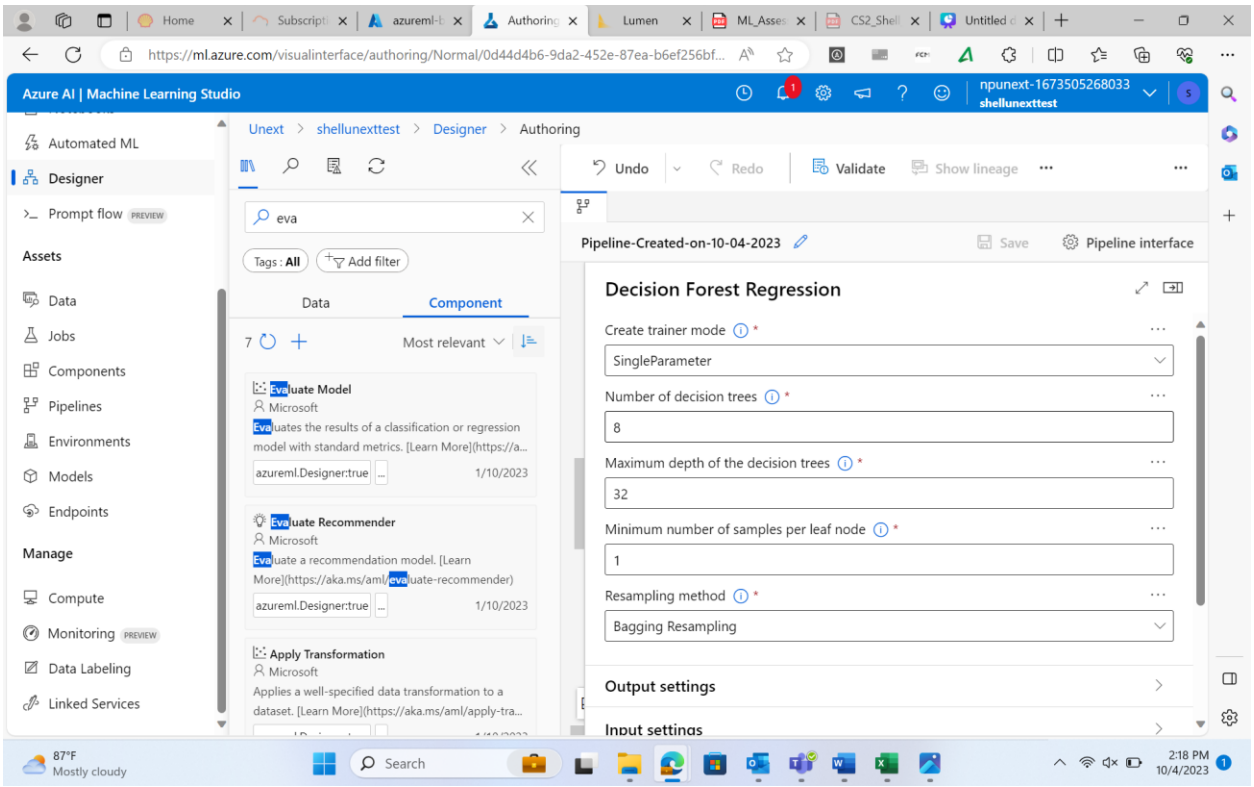

Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Create data asset
Data type
Data source
Source storage type
Storage path
Settings
Schema
Review
Select a datastore
Choose a storage type and a datastore that contains your data. You can also create a new datastore for your data first.
Datastore type *
Azure Blob Storage
Create new datastore
Search datastore
Filter
Columns
Name
Storage name
Created on
workspaceblobstore
shellunexttest6160402873
Oct 4, 2023 1:
workspaceartifactstore
shellunexttest6160402873
Oct 4, 2023 1:
Page
1
of 1
25/Page
Back
Next
3 of 24 - Clipboard
Item not Collected: Larger than clipboard capacity
84°F
Partly sunny
Search
1:51 PM
10/4/2023

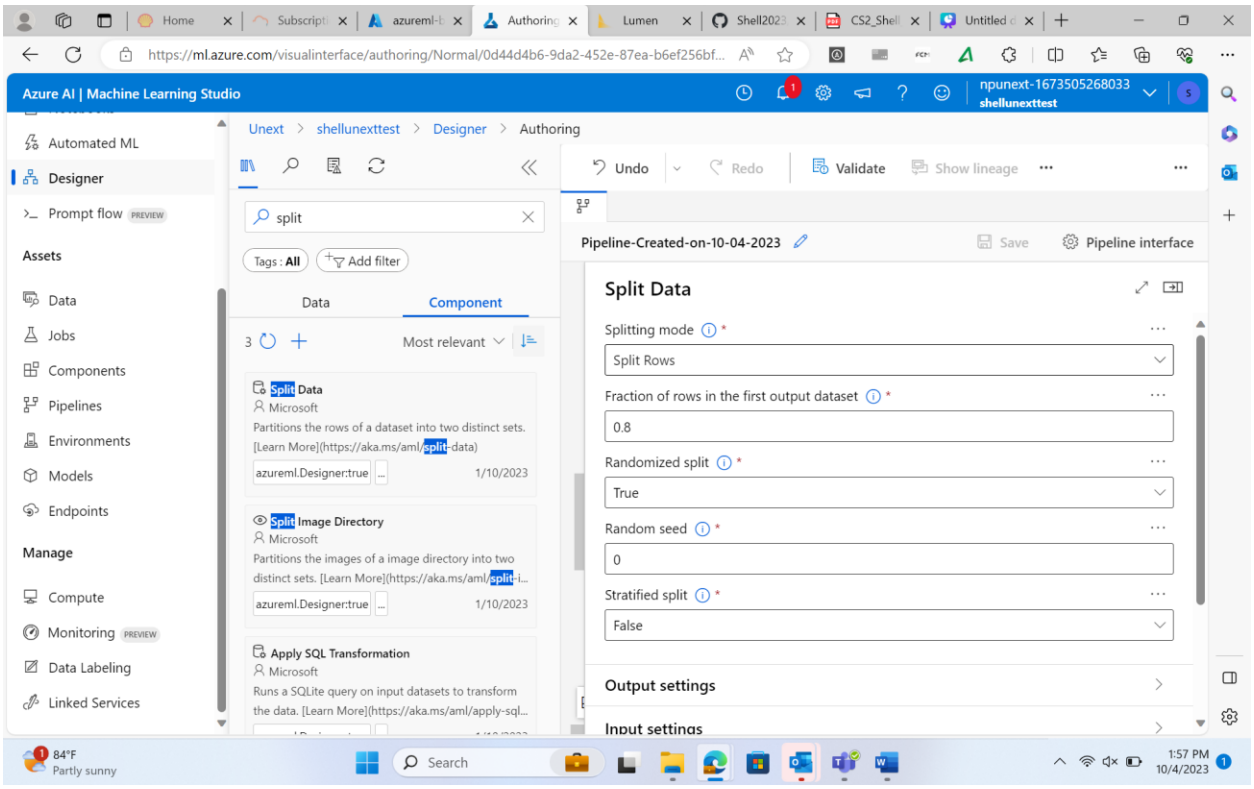

Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Create data asset
Data type
Data source
Source storage type
Storage path
Settings
Schema
Review
Choose a storage path
Navigate to or enter the storage path you want to use for this data asset.
Browse to storage path
Enter storage path manually
Selected path: customer_data.csv
Filter...
Name
Created on
Modified on
customer_data.csv
Oct 4, 2023 1:48 PM
Oct 4, 2023 1:48 PM
Advanced settings
Back
Next
Cancel

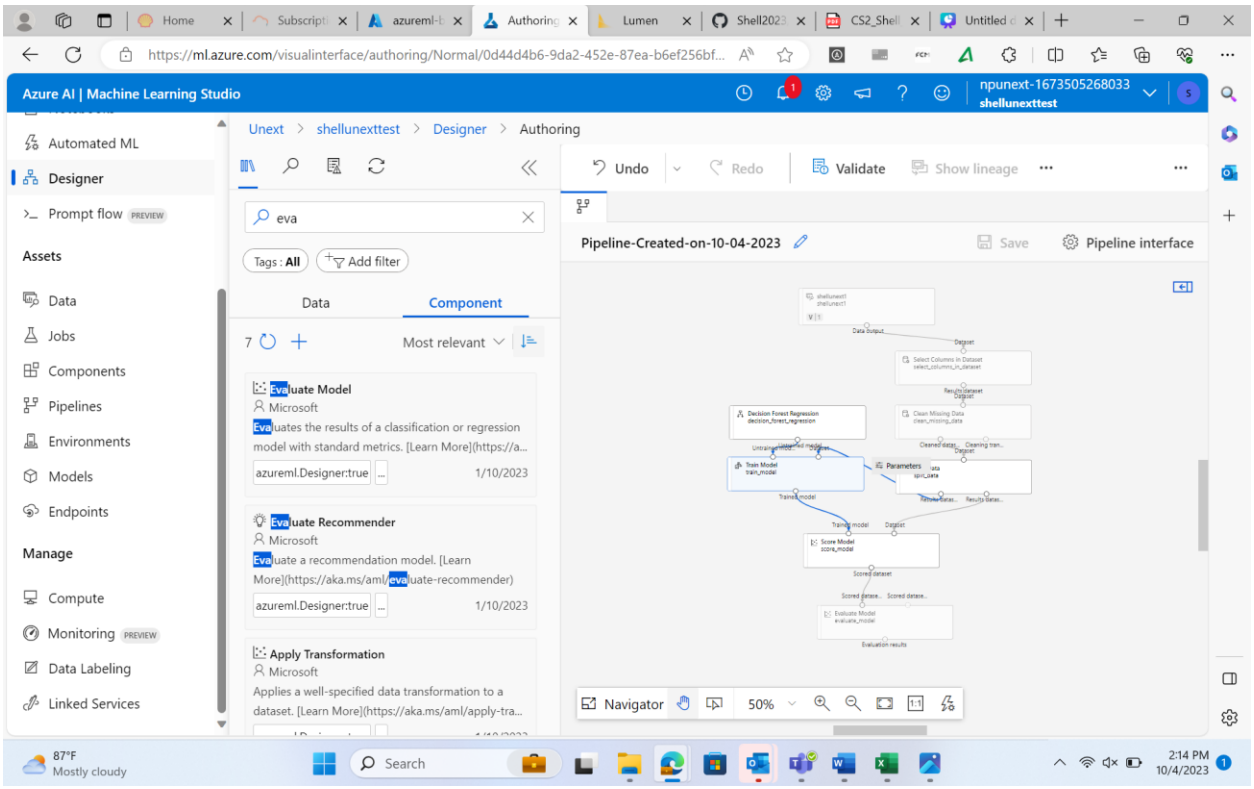

Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Create data asset
Data type
Data source
Source storage type
Storage path
Settings
Schema
Review
Settings
These settings determine how the data is parsed. The initial settings are automatically detected; you can change them as needed to reparse the data.
File format
Delimited
Delimiter
Comma
Example
Field1,Field2,Field3
Encoding
UTF-8
Column headers
All files have same headers
Skip rows
None
Dataset contains multi-line data
Note: Processing tabular files with multi-line data is slower because multiple CPU cores cannot be used to ingest the data in parallel. Checking this option may result in slower processing times.
Data preview
CustomerID
Age
AnnualIncome
SpendingScore
1
46
371,045
99
2
43
45,194
24
3
48
111,465
59
4
61
null
21
Back
Next
Review
Cancel

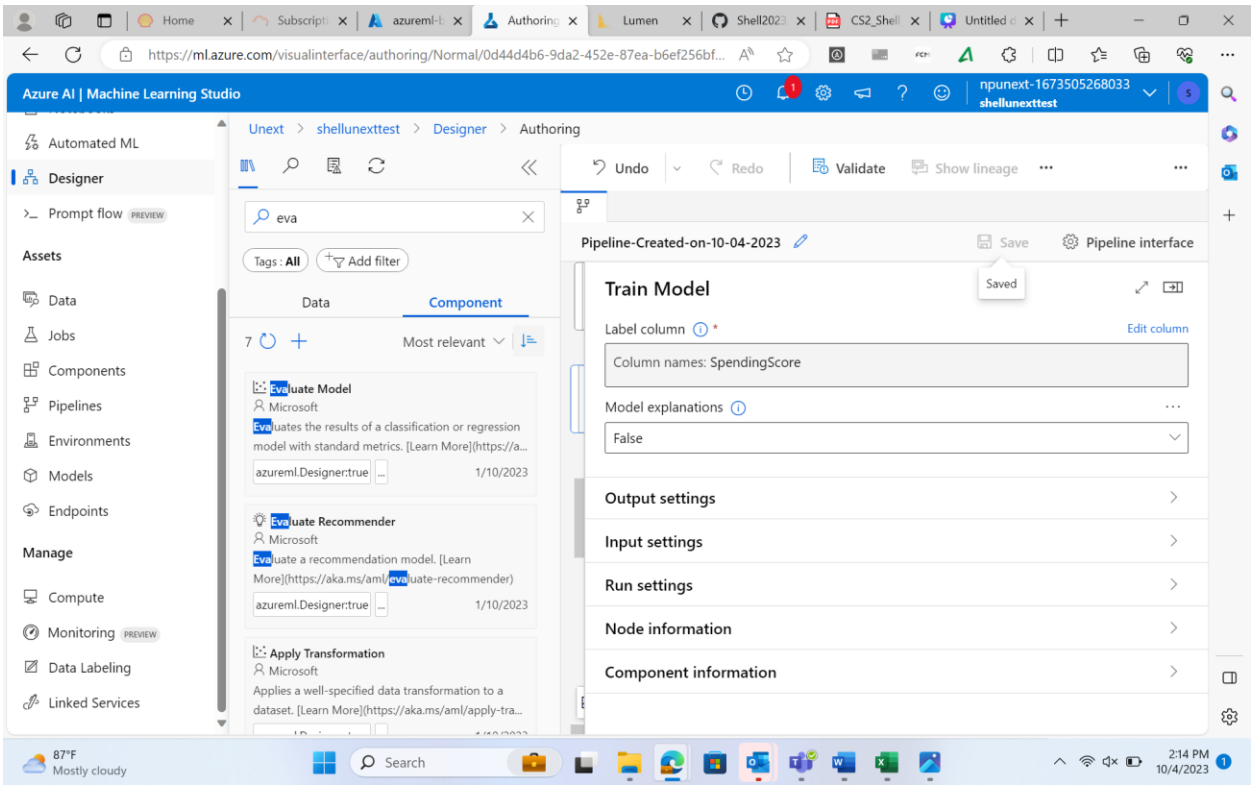


2.Model Development

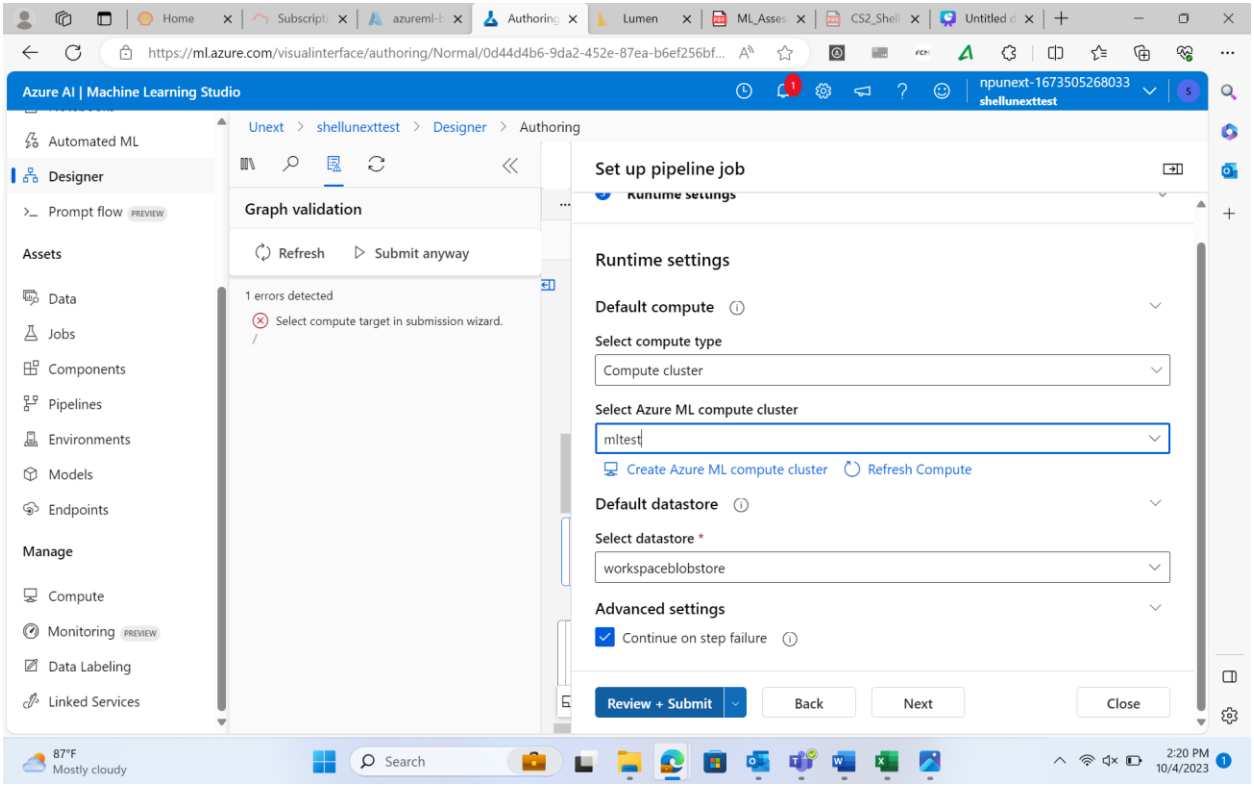


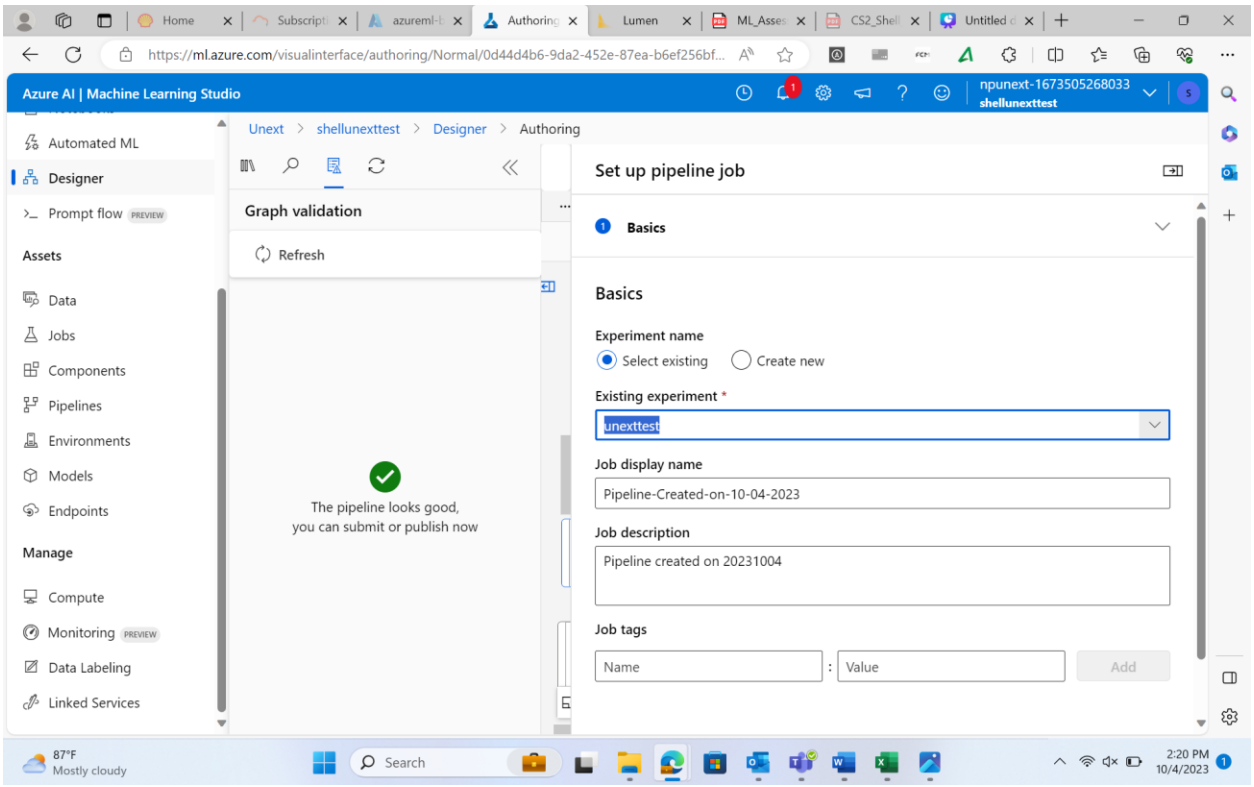

Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Unext > shellunexttest > Designer > Authoring
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Monitoring PREVIEW
Data Labeling
Linked Services
split
Tags : All
Add filter
Data
Component
3
Most relevant
Split Data
Microsoft
Partitions the rows of a dataset into two distinct sets. [Learn More](https://aka.ms/aml/split-data)
azureml.Designer:true
1/10/2023
Split Image Directory
Microsoft
Partitions the images of a image directory into two distinct sets. [Learn More](https://aka.ms/aml/split-i...
azureml.Designer:true
1/10/2023
Apply SQL Transformation
Microsoft
Runs a SQLite query on input datasets to transform the data. [Learn More](https://aka.ms/aml/apply-sql...
Undo
Redo
Validate
Show lineage
Pipeline-Created-on-10-04-2023
Save
Pipeline interface
Split Data
Splitting mode *
Split Rows
Fraction of rows in the first output dataset *
0.8
Randomized split *
True
Random seed *
0
Stratified split *
False
Output settings
Input settings
84°F
Partly sunny
Search
1:57 PM
10/4/2023

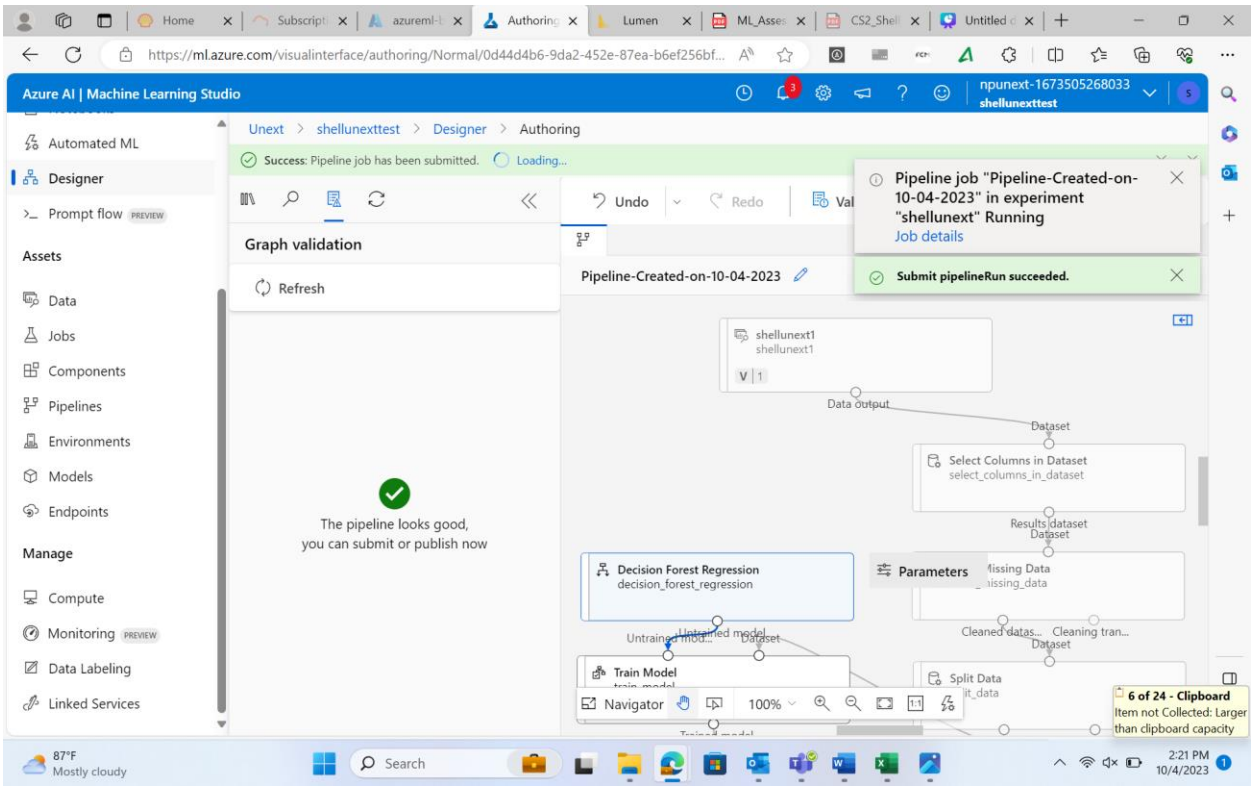

Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Unext > shellunexttest > Designer > Authoring
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Monitoring PREVIEW
Data Labeling
Linked Services
eva
Tags : All
Add filter
Data
Component
7
Most relevant
Evaluate Model
Microsoft
Evaluates the results of a classification or regression model with standard metrics. [Learn More](https://a...
azureml.Designer:true
1/10/2023
Evaluate Recommender
Microsoft
Evaluate a recommendation model. [Learn More](https://aka.ms/aml/evaluate-recommender)
azureml.Designer:true
1/10/2023
Apply Transformation
Microsoft
Applies a well-specified data transformation to a dataset. [Learn More](https://aka.ms/aml/apply-tra...
Undo
Redo
Validate
Show lineage
Pipeline-Created-on-10-04-2023
Save
Pipeline interface
Navigator
50%
87°F
Mostly cloudy
Search
2:14 PM
10/4/2023

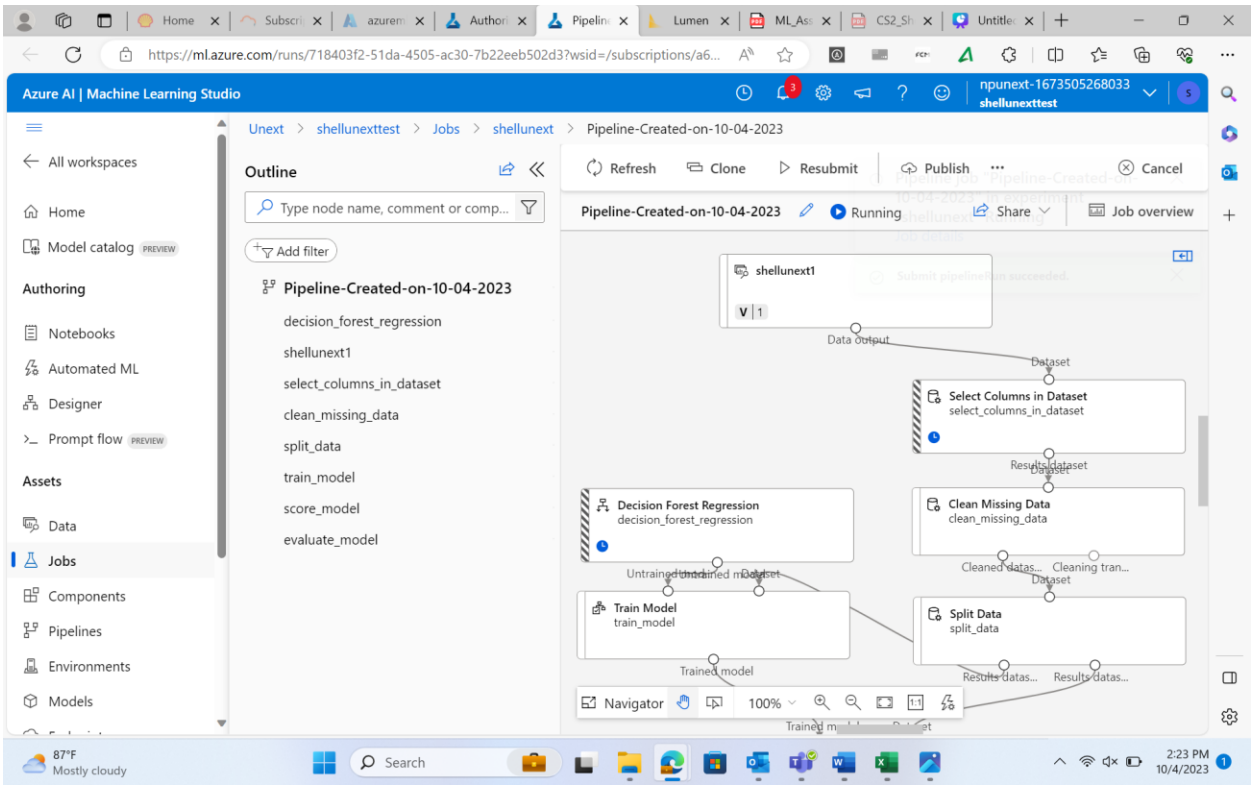


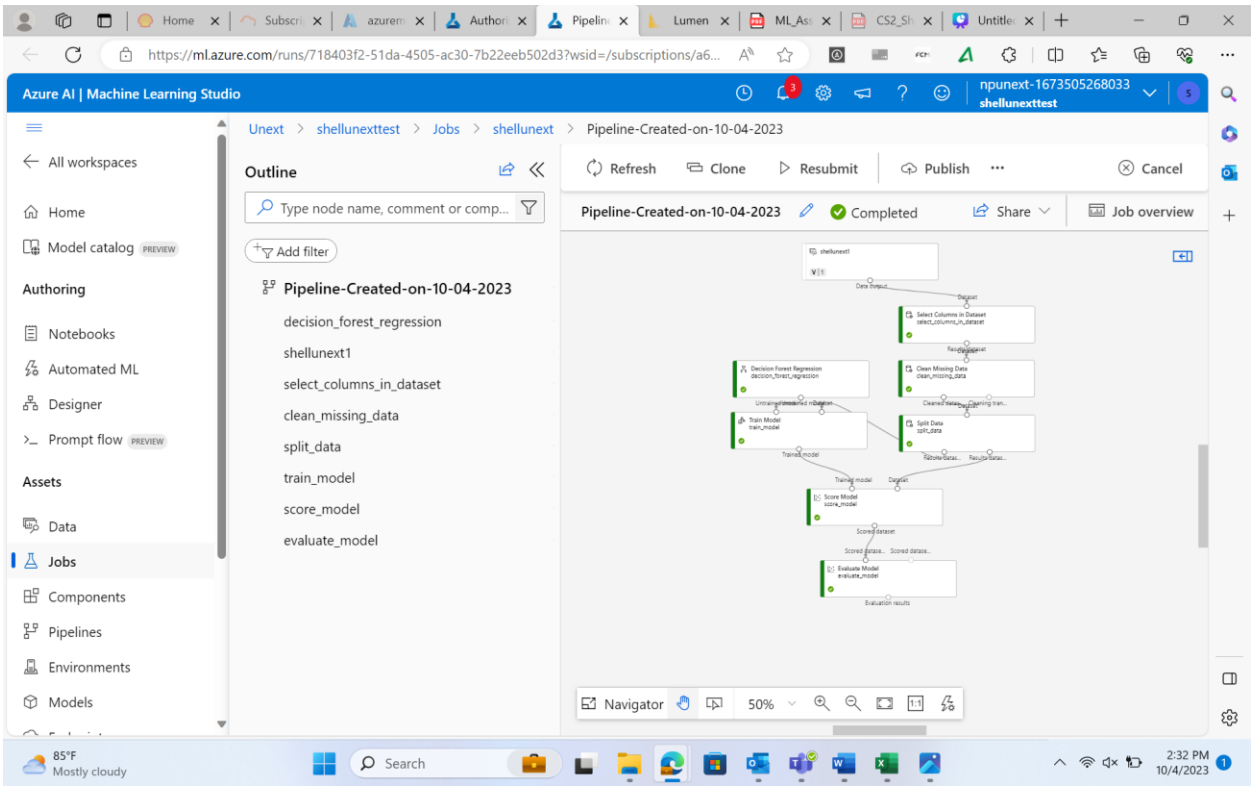
3.Hypertuning:




Azure AI | Machine Learning Studio
Unext > shellunexttest > Designer > Authoring
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Monitoring PREVIEW
Data Labeling
Linked Services
Graph validation
Refresh
The pipeline looks good, you can submit or publish now
Set up pipeline job
Basics
Experiment name
Select existing
Create new
Existing experiment *
unexttest
Job display name
Pipeline-Created-on-10-04-2023
Job description
Pipeline created on 20231004
Job tags
Name
Value
Add


Azure AI | Machine Learning Studio
Unext > shellunexttest > Designer > Authoring
Success: Pipeline job has been submitted.
Loading...
Undo
Redo
Graph validation
Refresh
Pipeline-Created-on-10-04-2023
Pipeline job "Pipeline-Created-on-10-04-2023" in experiment "shellunext" Running
Job details
Submit pipelineRun succeeded.
shellunext1
Data output
Dataset
Select Columns in Dataset
select_columns_in_dataset
Results dataset
Decision Forest Regression
decision_forest_regression
Parameters
Train Model
Split Data
Navigator
100%
6 of 24 - Clipboard
Item not Collected: Larger than clipboard capacity


Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Unext > shellunexttest > Jobs > shellunext > Pipeline-Created-on-10-04-2023
All workspaces
Home
Model catalog PREVIEW
Authoring
Notebooks
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Outline
Type node name, comment or comp...
Add filter
Pipeline-Created-on-10-04-2023
decision_forest_regression
shellunext1
select_columns_in_dataset
clean_missing_data
split_data
train_model
score_model
evaluate_model
Refresh
Clone
Resubmit
Publish
Cancel
Pipeline-Created-on-10-04-2023
Running
Share
Job overview
shellunext1
Data output
Dataset
Select Columns in Dataset
select_columns_in_dataset
Results dataset
Decision Forest Regression
decision_forest_regression
Clean Missing Data
clean_missing_data
Untrained model
Cleaned datas...
Cleaning tran...
Train Model
train_model
Split Data
split_data
Trained model
Results datas...
Navigator
100%
87°F
Mostly cloudy
Search
2:23 PM
10/4/2023


Azure AI | Machine Learning Studio
npunext-1673505268033
shellunexttest
Unext > shellunexttest > Jobs > shellunext > Pipeline-Created-on-10-04-2023
All workspaces
Home
Model catalog PREVIEW
Authoring
Notebooks
Automated ML
Designer
Prompt flow PREVIEW
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Outline
Type node name, comment or comp...
Add filter
Pipeline-Created-on-10-04-2023
decision_forest_regression
shellunext1
select_columns_in_dataset
clean_missing_data
split_data
train_model
score_model
evaluate_model
Refresh
Clone
Resubmit
Publish
Cancel
Pipeline-Created-on-10-04-2023
Completed
Share
Job overview
Navigator
50%
85°F
Mostly cloudy
Search
2:32 PM
10/4/2023

Answers for following questions:

1. What are the key steps involved in preparing the dataset for training a machine learning model using Azure Machine Learning? Briefly explain each step.

Ans:

1.Data gathering: Compile the required information from a variety of sources, including databases, CSV files, Azure Blob Storage, etc. Make sure the data is thorough and pertinent to the issue you are attempting to solve.

2. Data exploration: Recognize the layout, dimensions, and characteristics of the dataset. Examine each feature's distribution, look for any potential outliers, and judge the accuracy of the data. For this step, visualization tools may be useful.

3. Cleaning of Data:

Dealing with missing values Choose between removing missing data-containing rows and imputation of missing values using mean, median, or more sophisticated imputation algorithms.

Managing outliers Find and deal with outliers that could harm the performance of the model. Depending on the context, outliers can either be deleted or altered.

4. Engineering and Feature Selection:

Feature Choice: Engineering additional features could help the model by giving it useful information. This may entail changing variables, producing interaction terms, or generating new variables from the ones that already exist.

5. Encoding and Data Transformation:

Scaling numerical features to guarantee that they all contribute equally to the model is known as normalization or standardization. Min-Max scaling and z-score normalization are typical methods.

Convert category variables into numerical representations via categorical encoding. Depending on the type of categorical data, techniques may include one-hot encoding or label encoding.

Data splitting into training and testing sets is step six. The testing set is used to assess the machine learning model's performance after it has been trained using the training set. 80% for training and 20% for testing is a typical split ratio.

7. Uploading Data toAzure: Upload the preprocessed and cleaned dataset to the Azure Machine Learning workspace. This might be any Azure-based data storage option, including Azure SQL Database and Azure Blob Storage.

8. Data Versioning: Use dataset version control. Versioning capabilities are available through Azure Machine Learning, assuring the reproducibility and traceability of experiments performed on particular dataset versions.

9. Data Pipeline Configuration:

Utilize the pipeline features of Azure Machine Learning to set up data pipelines. As a result, data preprocessing may be automated and replicated, assuring consistency between experiments and deployments.

2. Why is it important to split the dataset into training and testing sets when developing a machine learning model? How does this help in model evaluation?
   For numerous reasons, dividing the dataset into training and testing sets is an essential step in creating machine learning models.

   1. Model Assessment: Unbiased Assessment: An objective assessment of the final model fit on the training dataset can be made using the testing set. It serves as a stand-in for the model's performance on hypothetical real-world data.
   Avoiding Overfitting: When a model learns the training data, including its noise and outliers, too well, it is unable to generalize to new data. To determine whether the model is overfitting, use a different testing set.
   Evaluation of Generalization Any machine learning model's main objective is to successfully generalize to new data. The benchmark used to gauge the model's generalizability is the testing set.
   For numerous reasons, dividing the dataset into training and testing sets is an essential step in creating machine learning models.

   Training data, 1.
   2. Hyperparameter Tuning: A number of hyperparameters (such as tree depth in decision trees or learning rate in neural networks) are changed during model construction to enhance performance. Analyzing the model's performance in various hyperparameter combinations is possible with the testing set.
   3. Preventing Data Leakage: Testing data gives the model a valid, independent dataset to work from. If training data from tests were included, the model might memorize them, which could result in data leakage and erroneous performance assessments.
   4. Decision-Making: The outcomes of testing sets assist stakeholders in choosing whether to use the model. It provides assurance that the model has undergone a thorough evaluation and is prepared for use in the actual world.

3. Describe a machine learning algorithm suitable for predicting customer purchasing behaviour in the given scenario. Explain why you chose this algorithm.
   For numerous reasons, dividing the dataset into training and testing sets is an essential step in creating machine learning models.

   Algorithm for Random Forests:

1. Ensemble Learning: The ensemble learning technique Random Forest mixes various decision trees to produce predictions. To obtain a more precise and reliable forecast, it constructs numerous decision trees and blends them.

Dealing with Non-Linearity:

Complex, non-linear interactions between numerous elements (such age, income, and perhaps other hidden patterns) frequently have an impact on purchasing behavior. Because Random Forests can recognize non-linear associations in the data, they are appropriate for challenging prediction tasks.

3. The Value of the Feature

Each feature in the dataset is assigned a feature relevance value using Random Forest. This is helpful in marketing scenarios because it reveals which variables—like age and income—contribute significantly to forecasting consumer behavior. Effective marketing strategies can be guided by this data.

4.What is hyperparameter tuning, and why is it important in machine learning? Explain a technique used for hyperparameter tuning and its benefits.

Finding the ideal set of hyperparameters for a machine learning algorithm is referred to as hyperparameter tuning. Hyperparameters are configuration options that are external to the model and cannot be determined by data learning. Although they have a substantial impact on the model's performance, their values are not changed throughout training. Examples include the neural network learning rate, the decision tree depth, or the quantity of clusters in the k-means algorithm.

Hyperparameter Tuning's Relevance:

Performance Optimization: Choosing the right hyperparameters can greatly improve the performance of the model. Underfitting (model is too simple) or overfitting (model is too sophisticated and retains the training data) may result from suboptimal hyperparameters. Generalization: Well-tuned models perform better when applied to unobserved data, increasing their dependability and utility in practical applications.