

Name: Indushree Manjunatha Hegde

Email: [Indushree.Hegde@shell.com](mailto:Indushree.Hegde@shell.com)

## Creating Azure Machine Learning :

The screenshot shows the 'Create a machine learning workspace' page in the Microsoft Azure portal. It includes fields for Subscription (npunext-1673505268033) and Resource group (ml\_rg). Below these, 'Workspace details' are configured with Name (shellunexttest), Region (East US), Storage account ((new) shellunexttest6160402873), Key vault ((new) shellunexttest8435065017), Application insights ((new) shellunexttest9211549320), and Container registry (None). At the bottom, there are 'Review + create' and 'Next : Networking' buttons.

The screenshot shows the 'Microsoft.MachineLearningServices | Overview' page. A deployment named 'Deployment' is listed with status 'Deployment succeeded'. Deployment details include: Deployment name: Microsoft.MachineLearnin..., Start time: 10/4/2023, 1:33:59 PM, Subscription: npunext-1673505268033, Correlation ID: 047397b7-07c7-424a-a8b2..., and Resource group: ml\_rg. To the right, there are links for 'Go to resource' and 'Go to resource group'. The page also features sections for 'Cost management', 'Microsoft Defender for Cloud', 'Free Microsoft tutorials', and 'Work with an expert'.

The screenshot shows the Microsoft Azure Storage account overview page for the account 'shellunexttest6160402873'. The left sidebar contains navigation links for Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Storage Mover, and Data storage (Containers, File shares, Queues, Tables). The main content area displays the following details:

Essentials	
Resource group (move)	ml_rg
Location	East US
Subscription (move)	npunext-1673505268033
Subscription ID	a651e877-2c58-4ba1-97d9-e9be852f9371
Disk state	Available
Tags (edit)	
Add tags	
Properties	
Hierarchical namespace	Disabled
Default access tier	Hot
Blob service	
Require secure transfer for REST API operations	Enabled

The status bar at the bottom indicates the weather as 84°F Partly sunny and the date/time as 10/4/2023 1:40 PM.

## 1. Data Preparation

The screenshot shows the 'Create compute cluster' wizard in the Azure AI | Machine Learning Studio. The left sidebar lists 'Virtual Machine' and 'Advanced Settings'. The main form is titled 'Select virtual machine' and asks to choose a virtual machine size. It includes fields for 'Location' (set to East US), 'Virtual machine tier' (set to Dedicated), 'Virtual machine type' (set to CPU), and 'Virtual machine size'. There are two radio buttons: 'Select from recommended options' and 'Select from all options'. A search bar 'Search by VM name...' is also present. At the bottom are 'Back', 'Next', and 'Cancel' buttons.

**Create compute cluster**

**Configure Settings**  
Configure compute cluster settings for your selected virtual machine size.

Name	Category	Cores	Available quota	RAM	Storage	Cost/Node
Standard_DS3_v2	General purpose	4	8 cores	14 GB	28 GB	\$0.06/hr

**Compute name \*** mltest

**Minimum number of nodes \*** 0

**Maximum number of nodes \*** 1

**Idle seconds before scale down \*** 120

Enable SSH access

**Back** **Create** Download a template for automation. **Cancel**

**Compute**

The "Kubernetes clusters" tab is now where you can access previous versions of "inference clusters" (also known as "AKS clusters") and "attached Kubernetes" compute types along with any previously created compute targets using those types. [Learn more about Kubernetes clusters](#)

**Compute instances** **Compute clusters** **Kubernetes clusters** **Attached computes**

Create a single or multi node compute cluster for your training, batch inferencing or reinforcement learning workloads. [Learn more about compute clusters](#) Alternatively, you can now run a training job without having to create and manage compute by using serverless. [Learn more here](#)

**New** **Refresh** **Delete** **View options** **View quota**

Name	State	Size	Location	Created on
mltest	Succeeded (0 nodes)	STANDARD_DS3_V2	eastus	Oct 4, 2023 1:39 PM

Creating the data assets:

Screenshot of the Azure AI | Machine Learning Studio "Create data asset" wizard, Step 1: Data type.

The "Name" field is filled with "shellunext".

The "Type" dropdown is set to "File (uri\_file)".

On the right, there is a sidebar titled "Use cases for data types" with sections for "When should I use File type?" and "When should I use Folder type?".

Screenshot of the Azure AI | Machine Learning Studio "Create data asset" wizard, Step 2: URI.

The "URI" field contains "https://github.com/manojkumarsingh77/Shell2023/blob/main/AssessmentData/customer\_data.csv".

A "Skip data validation" toggle switch is present.

On the right, there is a sidebar titled "Supported URI formats" listing three types of URIs:

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/<container_name> or  
abfs[s]://<file_system>@<account_name>.dfs.core.windows.net or  
wasbs[s]://<container_name>@<account_name>.blob.core.windows.net`
3. A path on a datastore, such as:  
`azureml://datastores/<datastore_name>/<path_to_file>`

Azure AI | Machine Learning Studio

Create data asset

Review

Review the settings for your data asset and make any changes as needed.

Data type

Name: shellunext

Description: --

Type: uri\_file

Data source

Type: Uri

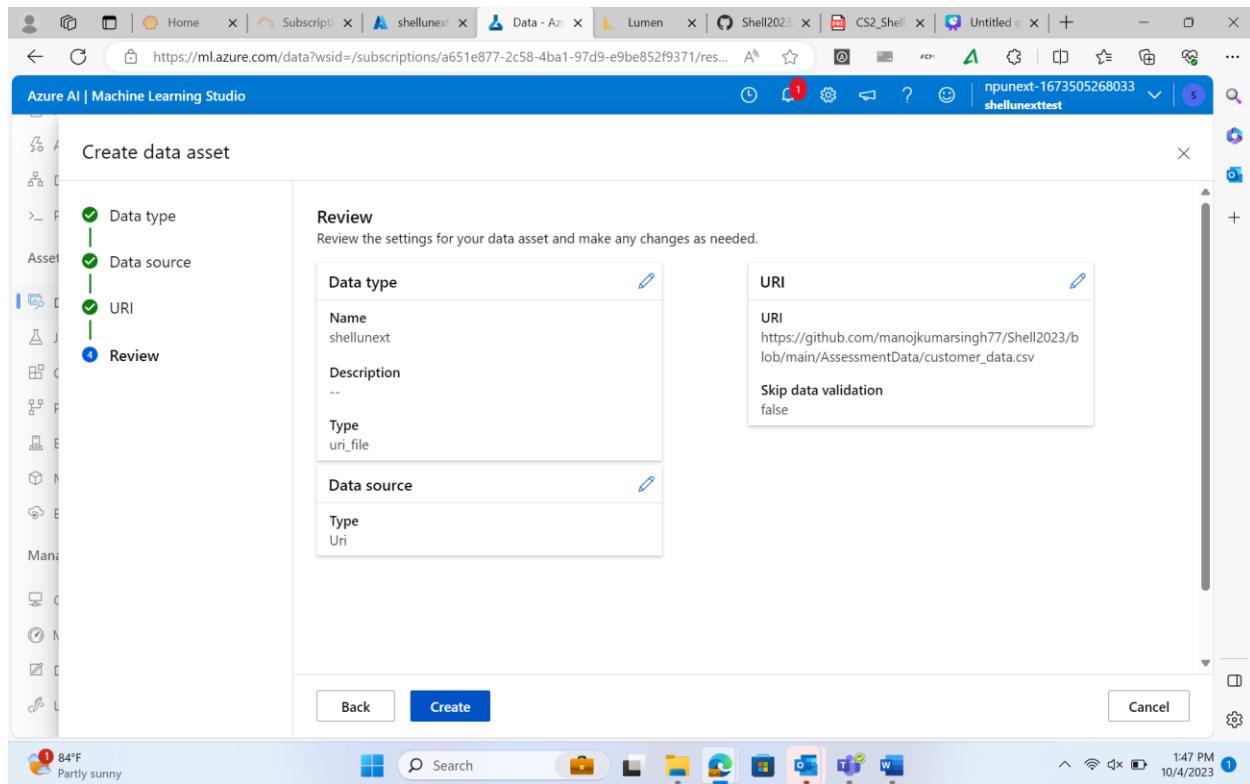
URI

URI: https://github.com/manojkumarsingh77/Shell2023/blob/main/AssessmentData/customer\_data.csv

Skip data validation: false

Back Create Cancel

84°F Partly sunny 1:47 PM 10/4/2023



Azure AI | Machine Learning Studio

Create data asset

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

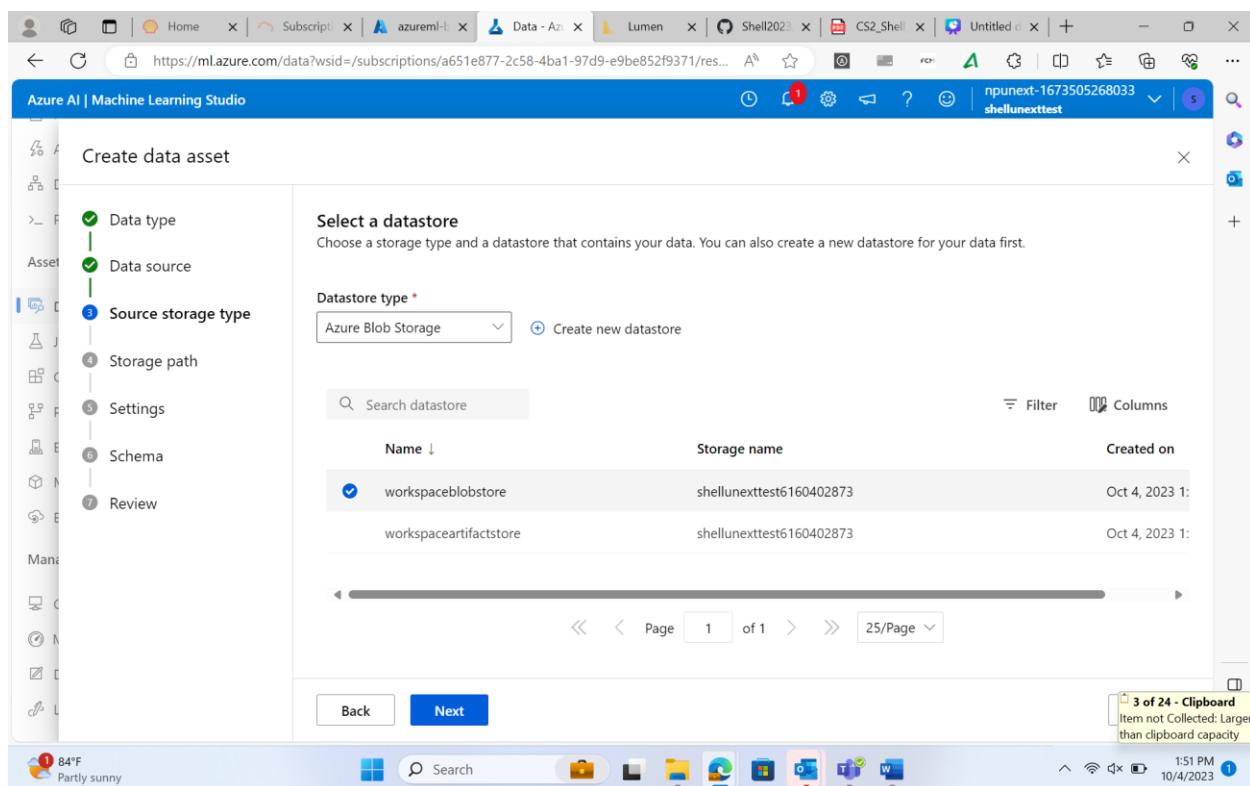
Name ↓	Storage name	Created on
workspaceblobstore	shellunexttest6160402873	Oct 4, 2023 1:
workspaceartifactstore	shellunexttest6160402873	Oct 4, 2023 1:

Search datastore Filter Columns

Back Next

3 of 24 - Clipboard Item not Collected: Larger than clipboard capacity

84°F Partly sunny 1:51 PM 10/4/2023



Azure AI | Machine Learning Studio

Create data asset

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

84°F Partly sunny 1:51 PM 10/4/2023

Azure AI | Machine Learning Studio

Create data asset

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 Data preview will result in slower processing times.

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

84°F Partly sunny 1:52 PM 10/4/2023

The screenshot shows the Azure AI | Machine Learning Studio interface. On the left, there's a navigation sidebar with sections like 'Automated ML', 'Designer', 'Prompt flow (PREVIEW)', 'Assets' (Data, Jobs, Components, Pipelines, Environments, Models, Endpoints), and 'Manage' (Compute, Monitoring (PREVIEW), Data Labeling, Linked Services). The main area displays a dataset named 'shellunext1'. The dataset details include:

- Type: Table (mhtable)
- Dataset type (from Azure ML v1 APIs): Tabular
- Created by: Shellunext unext!DA21
- Profile: View profile (Job: --)
- Files in dataset: 1
- Total size of files in dataset: 4.044 KIB
- Current version: 1
- Latest version: 1

On the right, there are sections for 'Tags' (No data), 'Description' (Click edit icon to add a description), and 'Data sources'. Under 'Data sources', it shows a 'Datastore workspaceblobstore' with a relative path 'customer\_data.csv' and actions for 'View in datastores browse' and 'View in Azure Portal'. The status bar at the bottom shows the date and time as 10/4/2023 1:52 PM.

## 2. Model Development

The screenshot shows the Azure AI | Machine Learning Studio interface, specifically the 'Designer' section. The left sidebar is identical to the previous screenshot. The main area shows a pipeline titled 'Pipeline-Created-on-10-04-2023'. The pipeline interface includes tabs for 'Data' and 'Component'. A search bar at the top of the component list shows 'eva'. The component list contains three items:

- Evaluate Model: Evaluates the results of a classification or regression model with standard metrics. [Learn More](https://aka.ms/ml-evaluate-model) (azureml.Designertrue) - 1/10/2023
- Evaluate Recommender: Evaluate a recommendation model. [Learn More](https://aka.ms/ml-evaluate-recommender) (azureml.Designertrue) - 1/10/2023
- Apply Transformation: Applies a well-specified data transformation to a dataset. [Learn More](https://aka.ms/ml-apply-transformation) (azureml.Designertrue) - 1/10/2023

The right side of the screen shows the configuration details for the 'Evaluate Model' component, which is selected. The configuration fields include:

- Create trainer mode: SingleParameter
- Number of decision trees: 8
- Maximum depth of the decision trees: 32
- Minimum number of samples per leaf node: 1
- Resampling method: Bagging Resampling

The status bar at the bottom shows the date and time as 10/4/2023 2:18 PM.

**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

84°F Partly sunny

split

Tags : All + Add filter

Component

Split Data

Partitions the rows of a dataset into two distinct sets. [Learn More](https://aka.ms/aml/split-data)

azureml.Designerttrue 1/10/2023

Split Image Directory

Partitions the images of a image directory into two distinct sets. [Learn More](https://aka.ms/aml/split-i...)

azureml.Designerttrue 1/10/2023

Apply SQL Transformation

Runs a SQLite query on input datasets to transform the data. [Learn More](https://aka.ms/aml/apply-sql...)

Split Data

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

Save Pipeline interface

Split Data

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

1:57 PM 10/4/2023

**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

87°F Mostly cloudy

eva

Tags : All + Add filter

Component

Evaluate Model

Evaluates the results of a classification or regression model with standard metrics. [Learn More](https://aka.ms/aml/evaluate-model)

azureml.Designerttrue 1/10/2023

Evaluate Recommender

Evaluate a recommendation model. [Learn More](https://aka.ms/aml/evaluate-recommender)

azureml.Designerttrue 1/10/2023

Apply Transformation

Applies a well-specified data transformation to a dataset. [Learn More](https://aka.ms/aml/apply-tra...)

Evaluate Model

Evaluates the results of a classification or regression model with standard metrics. [Learn More](https://aka.ms/aml/evaluate-model)

Decision Forest Regression

decision\_forest\_regression

Train Model

train\_model

Parameters

scored\_dataset

Score Model

scored\_dataset

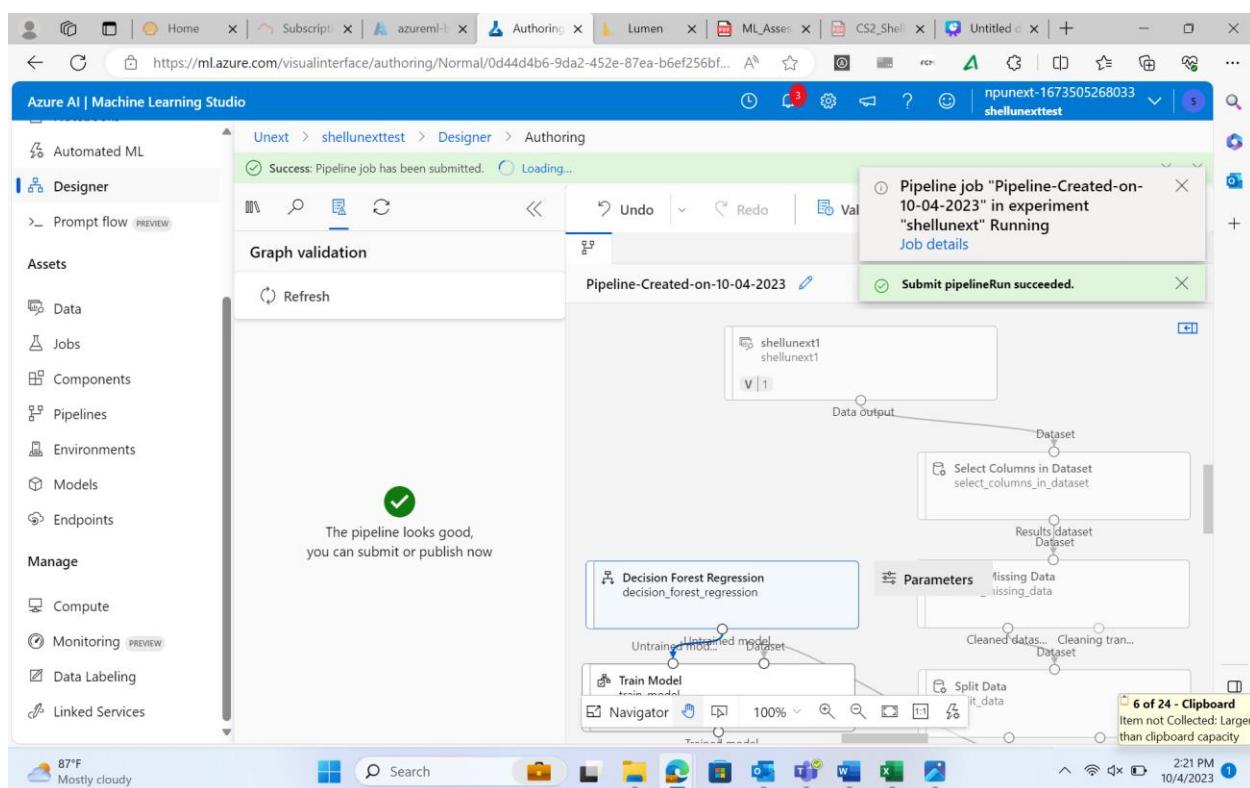
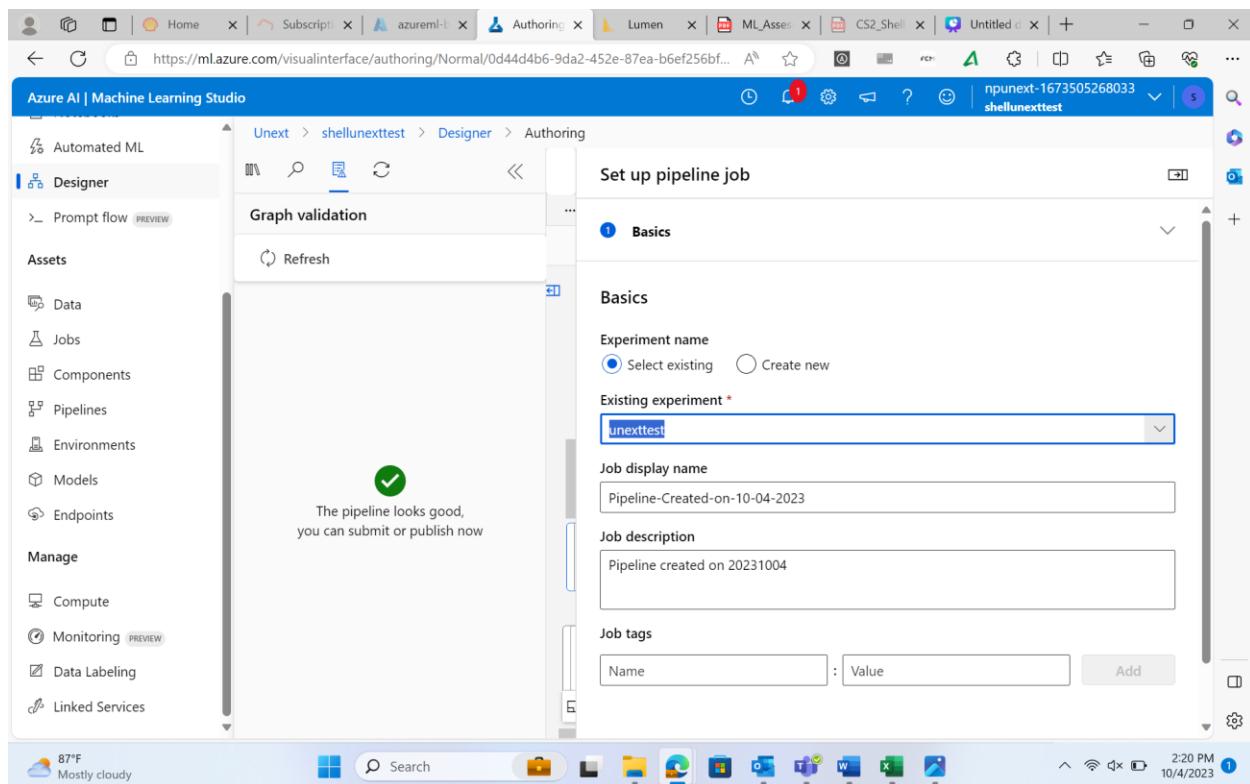
Evaluation results

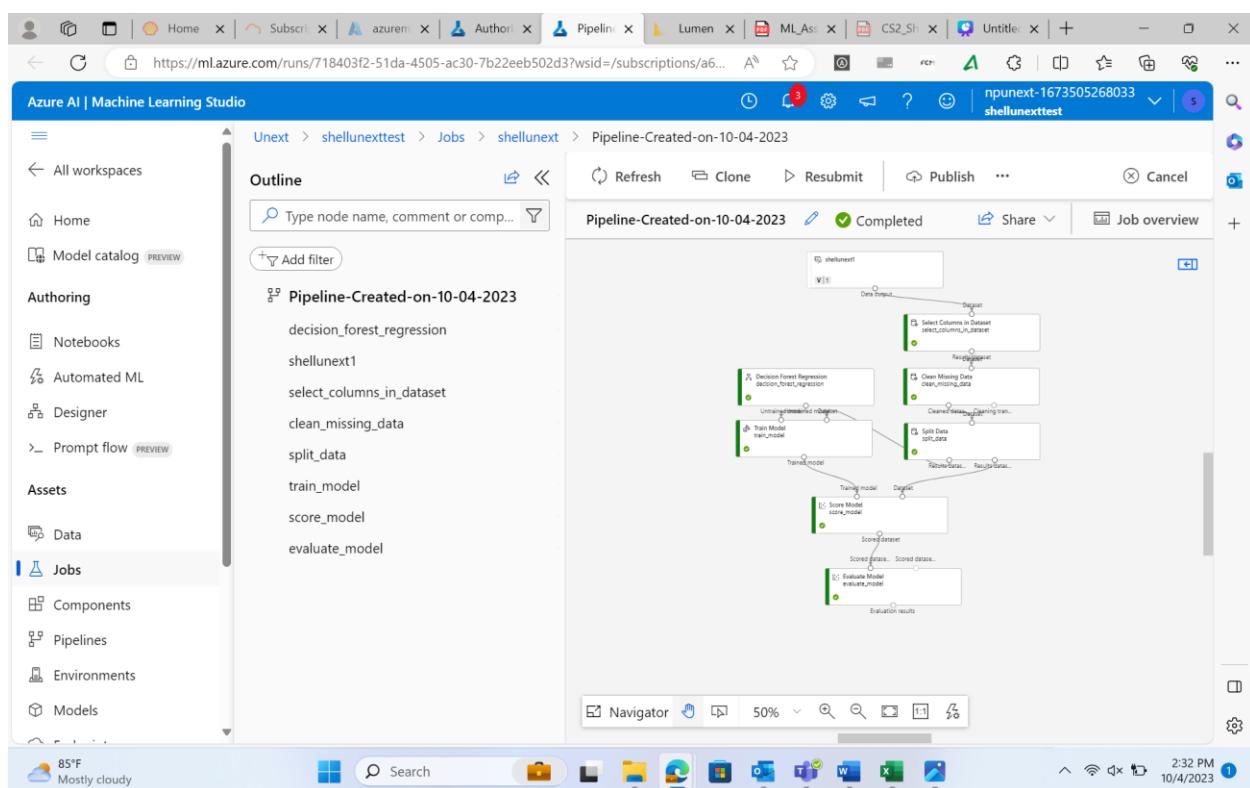
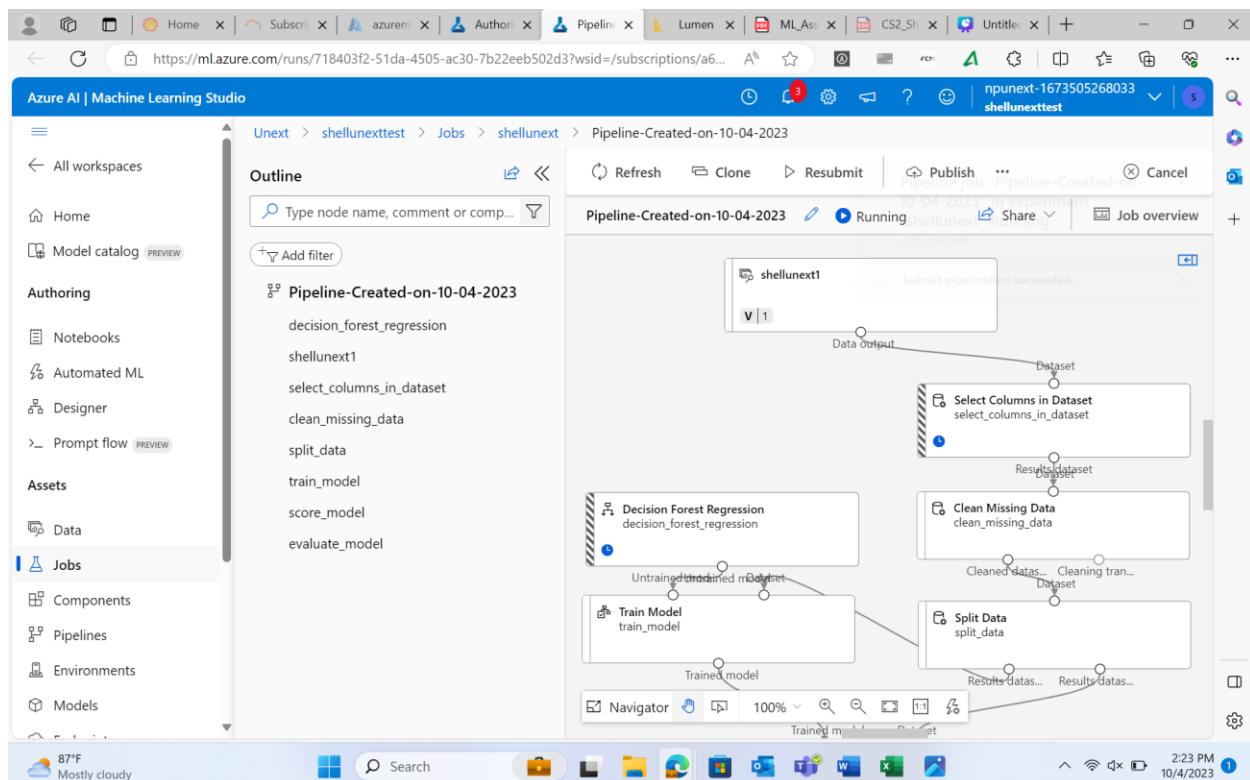
Navigator 50% 2:14 PM 10/4/2023

The screenshot shows the Azure AI | Machine Learning Studio interface. On the left, there's a sidebar with sections like 'Automated ML', 'Designer', 'Assets', 'Data', 'Jobs', 'Components', 'Pipelines', 'Environments', 'Models', 'Endpoints', 'Manage', 'Compute', 'Monitoring', 'Data Labeling', and 'Linked Services'. The main area is titled 'Train Model' under 'Pipeline-Created-on-10-04-2023'. It includes fields for 'Label column' (set to 'SpendingScore'), 'Model explanations' (set to 'False'), and sections for 'Output settings', 'Input settings', 'Run settings', 'Node information', and 'Component information'. At the bottom, there's a toolbar with various icons and a status bar showing '214 PM 10/4/2023'.

### 3. Hypertuning:

The screenshot shows the 'Set up pipeline job' configuration page in Azure AI | Machine Learning Studio. The left sidebar is identical to the previous screenshot. The main area is titled 'Set up pipeline job' and includes sections for 'Runtime settings', 'Default compute' (set to 'Compute cluster'), 'Select Azure ML compute cluster' (set to 'mitest'), 'Default datastore' (set to 'workspaceblobstore'), and 'Advanced settings' (with 'Continue on step failure' checked). Buttons at the bottom include 'Review + Submit', 'Back', 'Next', and 'Close'. The status bar at the bottom shows '2:20 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 to Azure: 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 as 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.