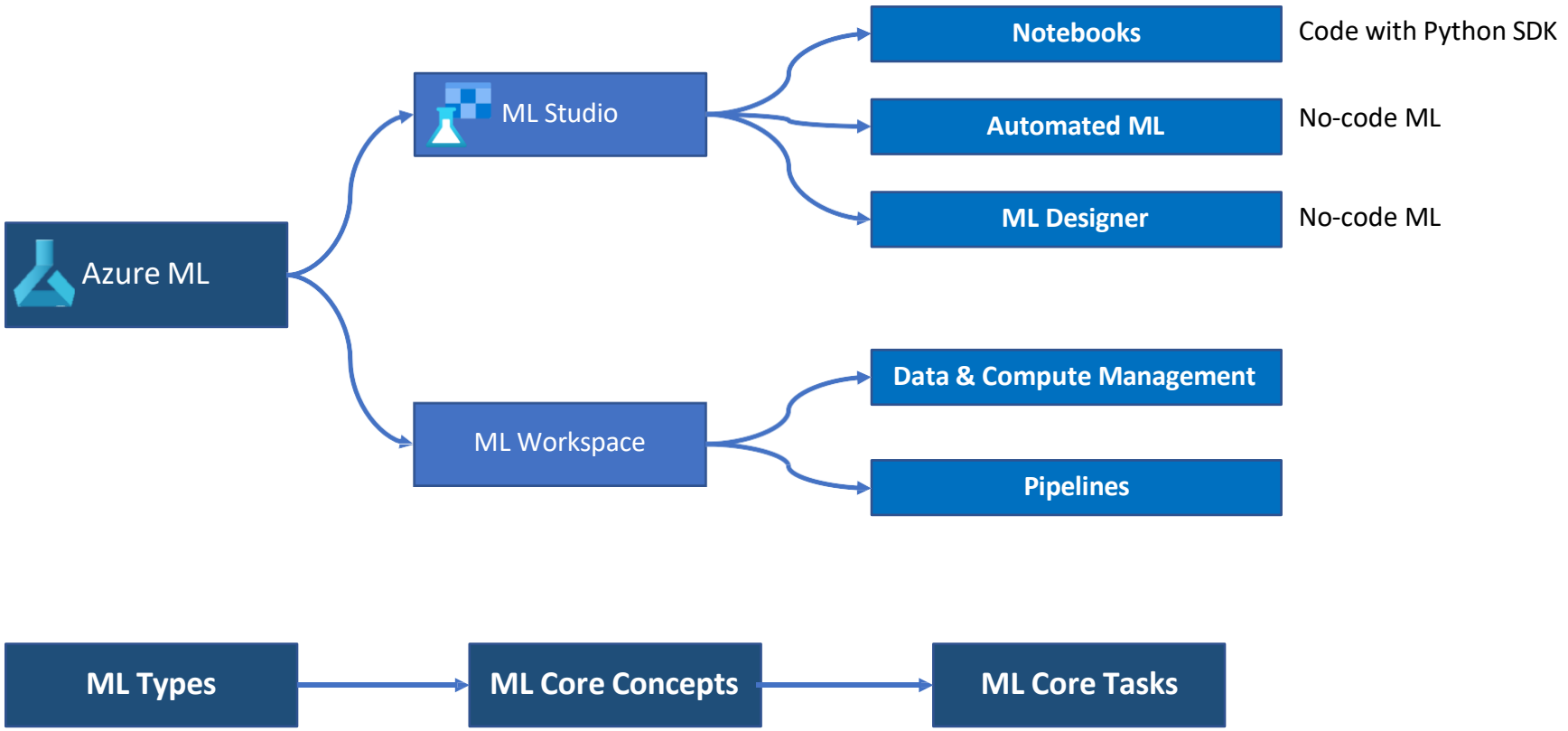
Section 3

Introduction to Azure Machine Learning

Azure Machine Learning – Azure ML Service

- **Azure Machine Learning** service - A cloud-based platform for creating, managing, and publishing machine learning models.



Common Machine Learning Types

Machine Learning Types – Supervised ML Models

- **Supervised ML Model**: training data includes known label values

Classification predicting one of several categories or classes	**Regression** predicting numeric values	**Time-series forecasting** predicting numeric values based on time
• Whether a patient is at risk of diabetes • whether tumor is benign or malignant • Categorize loan applications as low-risk or high-risk • Whether a loan will be repaid • Fraud detection • Object detection	• The number of cars that will be sold next month • The price of a pre-owned car based on its maker, model, engine size, and mileage • The number of future car rentals • Predicting how many minutes a flight will be delayed based on weather information • The Carbon Dioxide Emissions by Energy Consumption Use in an area	• Heights of ocean tides • Weather forecasting • Global temperature • Monthly counts of sunspot • Pollution levels • Blood pressure tracking • Heart rate monitoring • Population • Unemployment rates

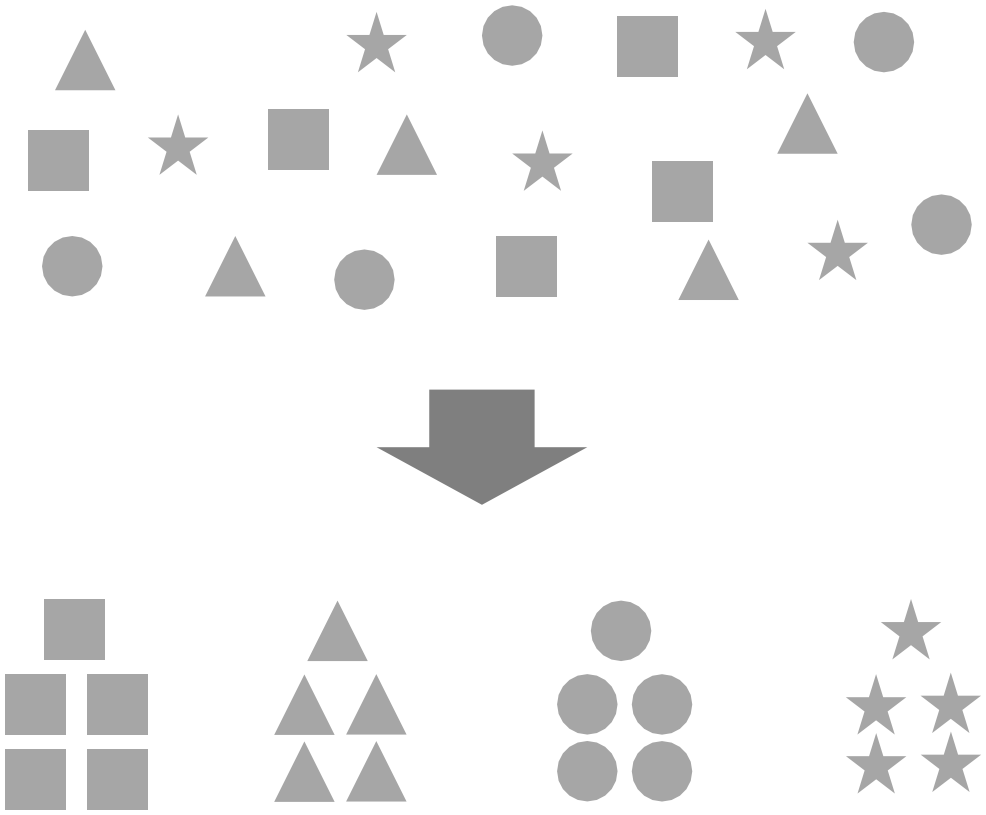
Machine Learning Types – Unsupervised ML Models

- **Unsupervised ML Model**: to separate inputs into clusters based purely on their characteristics, or *features* without a known label value

Clustering

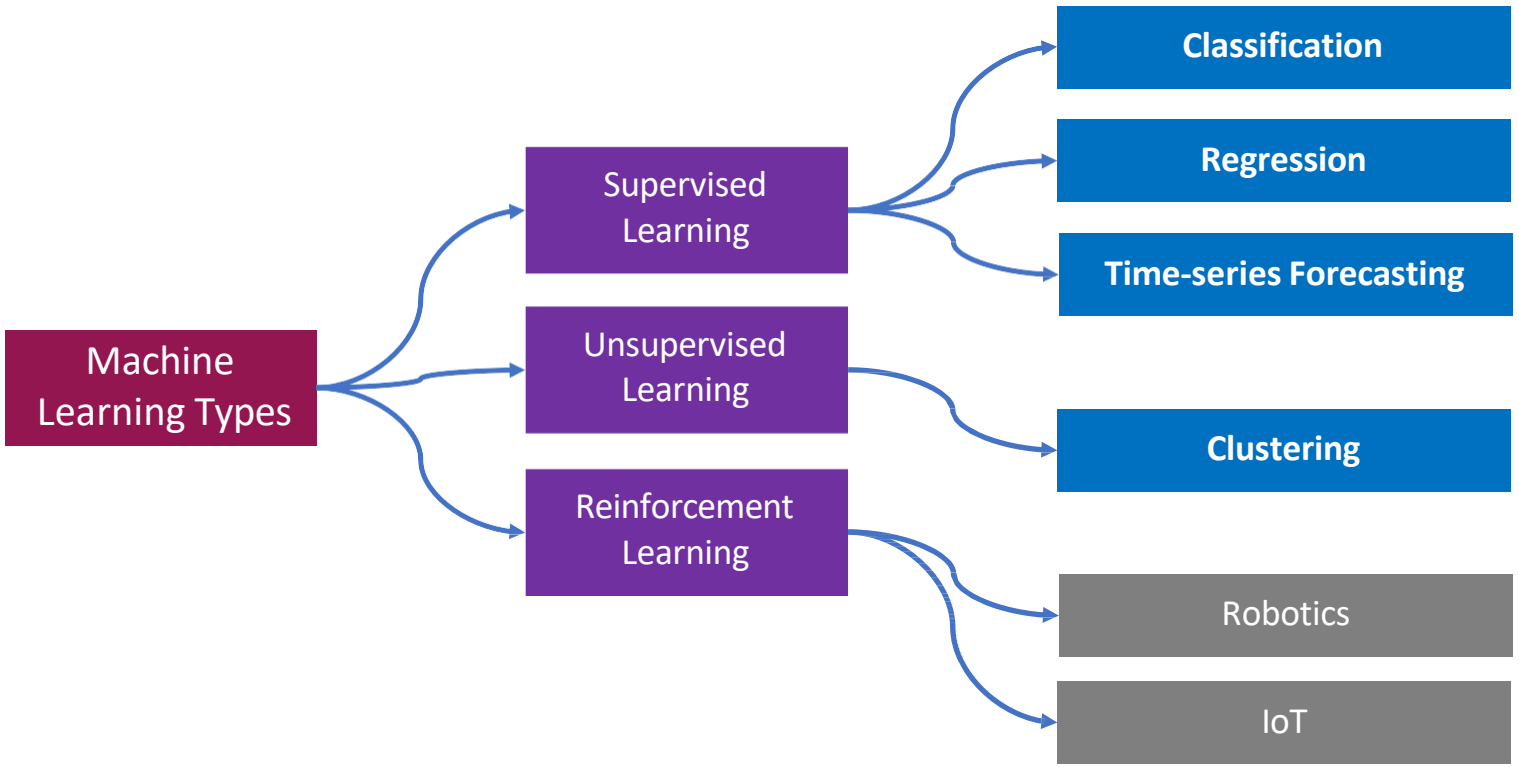
group inputs into clusters based on the similarity of their features

- Segment customers into different groups
- Group cars of similar size or weight
- Group penguins based their similarities
- Image classification - group images that have similar characteristics
- Face grouping – divide faces into groups



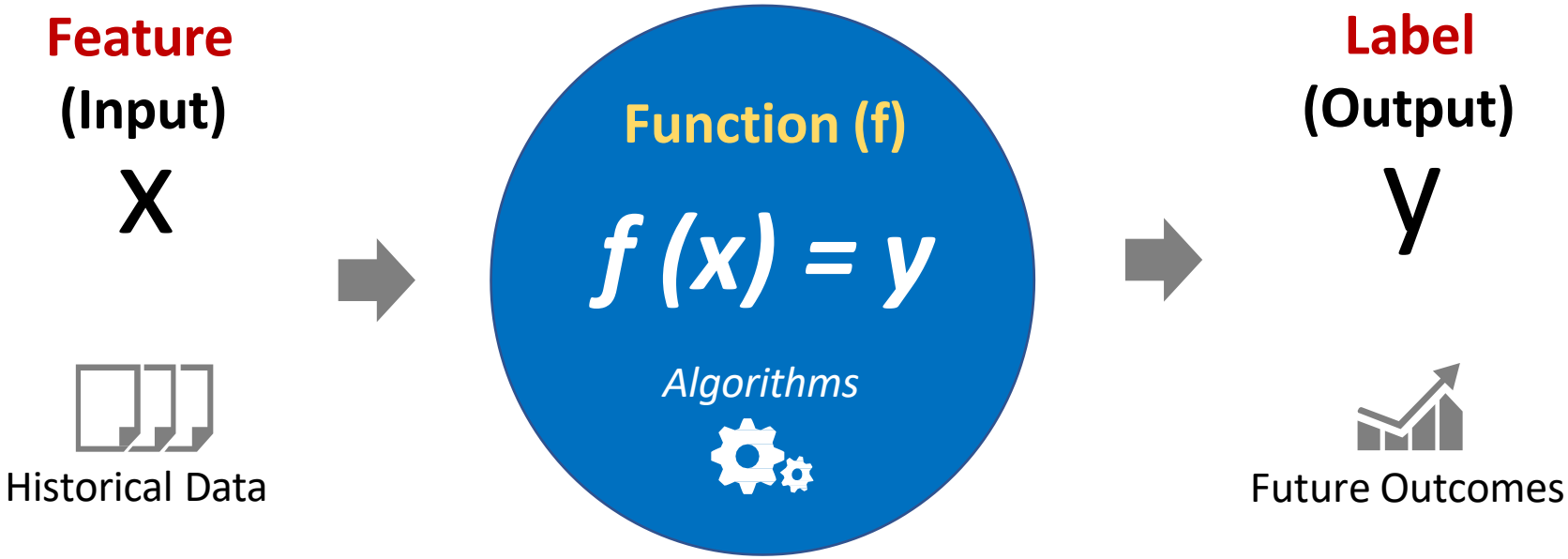
Feature: Shape
The minimum number of features: 1

Machine Learning Types – Summary



Core Machine Learning Concepts

Core ML Concepts – Features & Labels



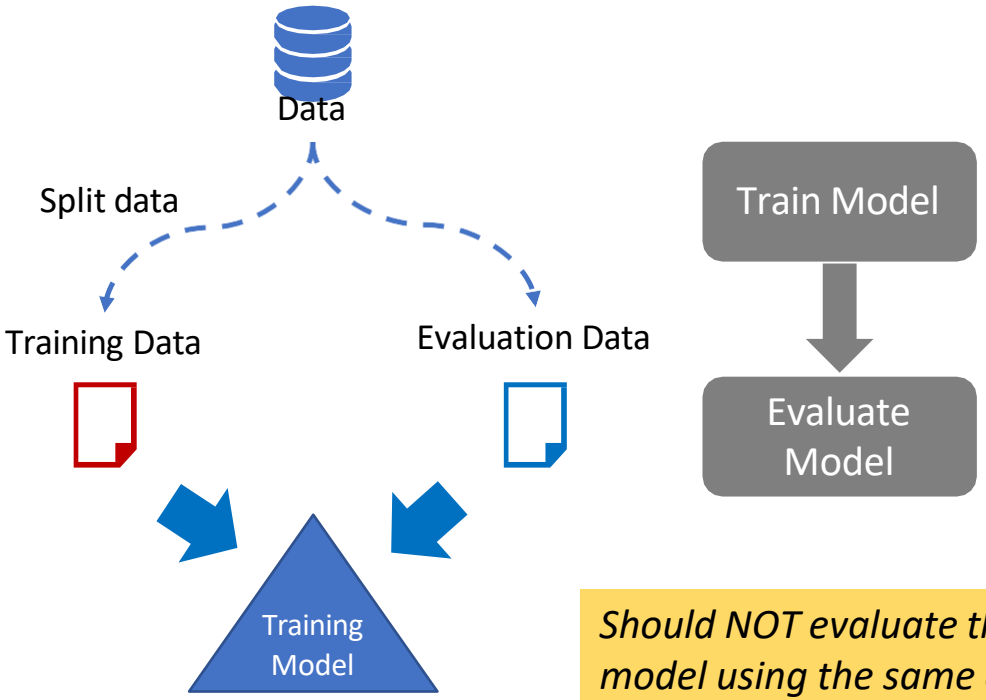
- Feature: Data values that influence the prediction of a model
- Labelling: Process of tagging training data with known values

Core ML Concepts – Training and Validation Datasets

- **Split Data**

  Split Rows (one of the common techniques)

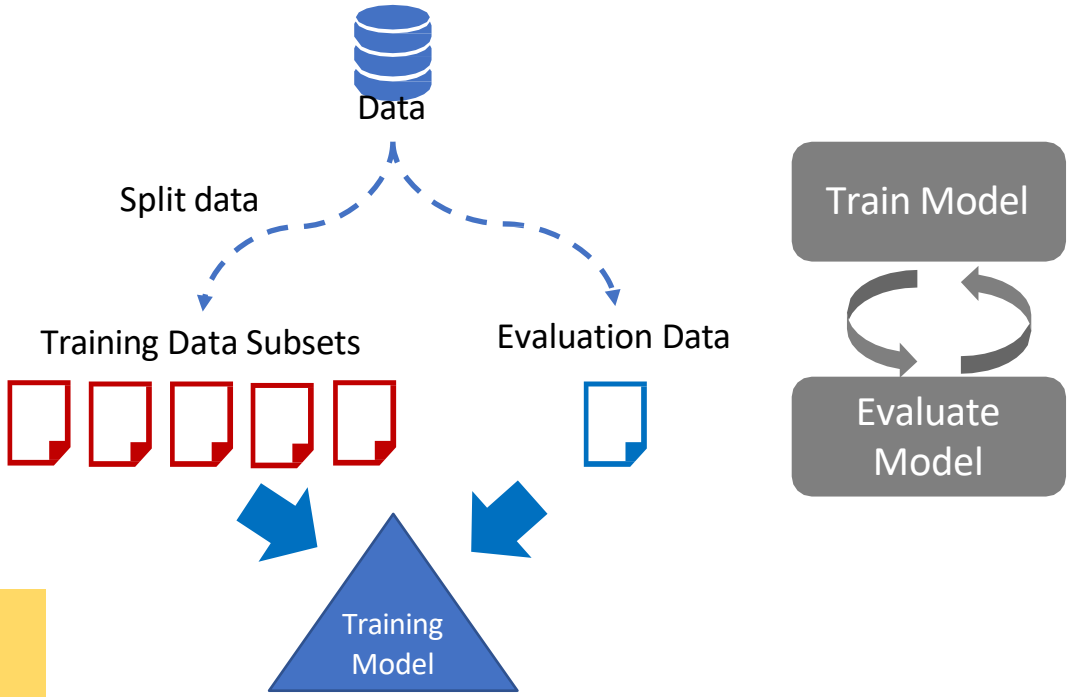
  - randomly split the data into rows for training and rows for evaluation/testing



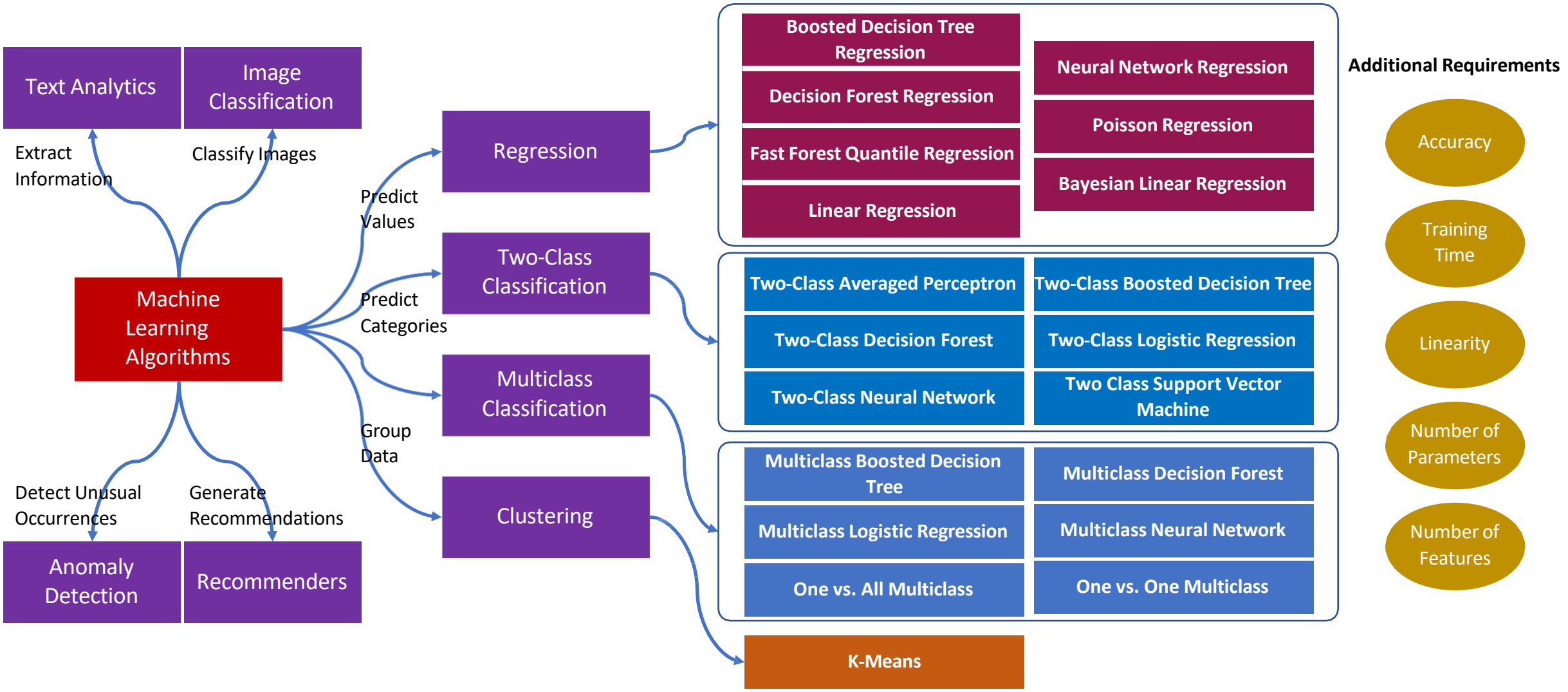
Should NOT evaluate the model using the same data that was used for training

- **Cross-validation**

  - Iteratively test the trained model with data it wasn't trained with and compare the predicted value with the actual known value



Core ML Concepts – Machine Learning Algorithms



Core ML Concepts – Evaluation Metrics for Classification

Confusion Matrix

Predicted \ Actual	1	0
1	True Positives	False Positives
0	False Negatives	True Negatives

What's the proportion of correctly classified instances?

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

What's the percentage of the class predictions made by the model were correct?

Precision

$$\frac{TP}{TP + FP}$$

What's the percentage of a certain class correctly identified?

Recall

$$\frac{TP}{TP + FN}$$

The weighted average of precision and recall
Closer to 1 the better

F1 Score

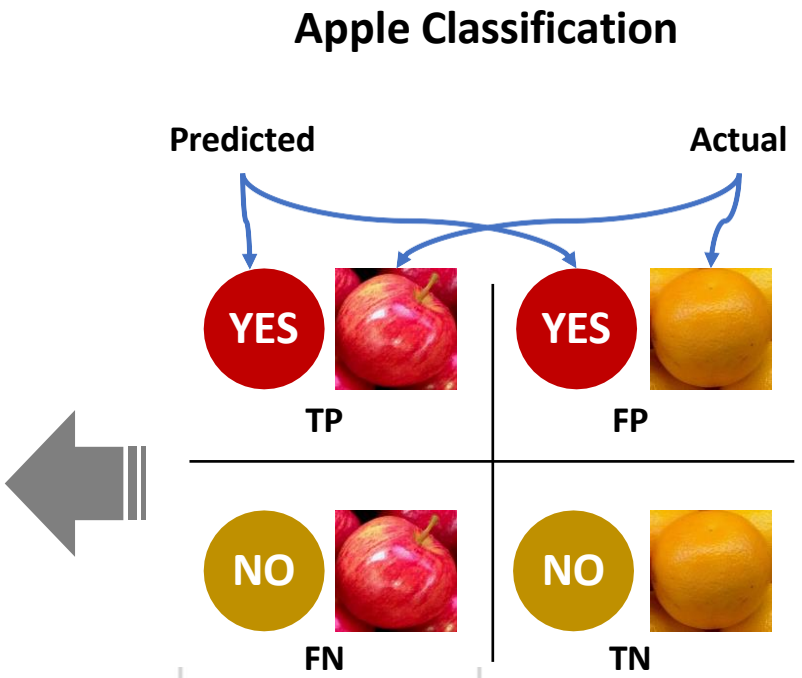
$$2 * \frac{Precision * Recall}{Precision + Recall}$$

$$=$$

$$\frac{2 * TP}{2 * TP + FP + FN}$$

Core ML Concepts – Evaluation Metrics for Classification Example

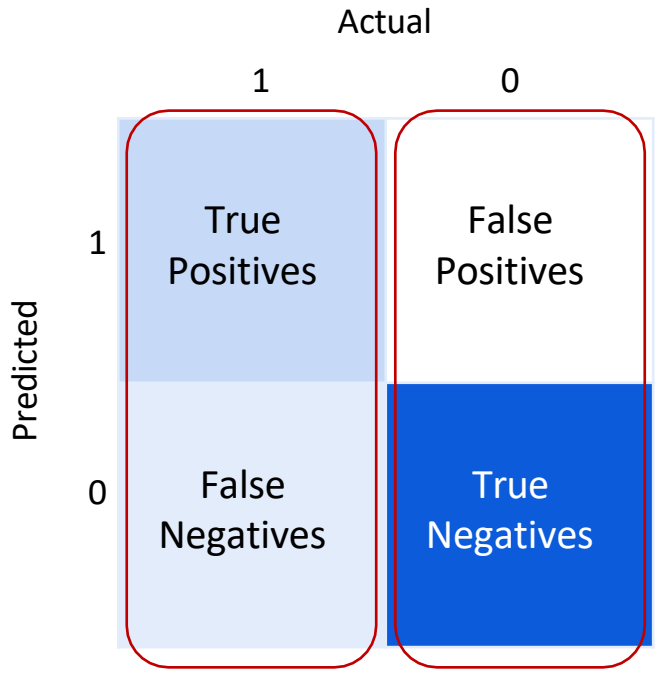
	A	B	C	D
1			Actual	
2			1	0
3	Predicted	1	10	2
4		0	5	3



5					Formula
6	TP		10	Accuracy	0.65 (C6+C7)/SUM(C6:C9)
7	TN		3	Precision	0.833333333 C6/(C6+C8)
8	FP		2	Recall	0.666666667 C6/(C6+C9)
9	FN		5	F1 Score	0.740740741 2*C6/(2*C6+C8+C9)

Core ML Concepts – Evaluation Metrics for Classification – ROC Curve & AUC

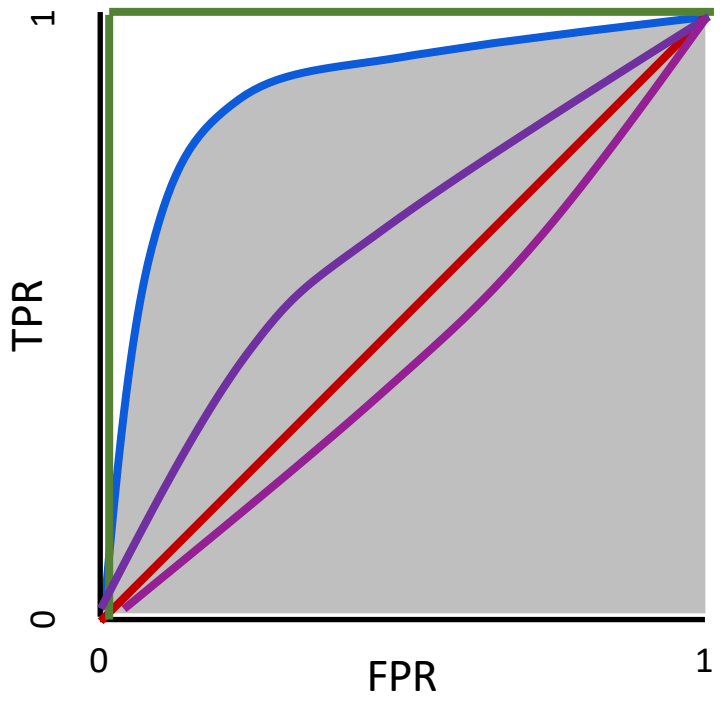
Confusion Matrix



$\frac{TP}{TP+FN}$	True Positive Rate (TPR)	False Positive Rate (FPR)	$\frac{FP}{FP+TN}$

Recall

ROC Curve
(receiver operating characteristic curve)



Ideal
Random Classifier
ROC Curve Example
Bad Model Example

AUC
(area under the curve)

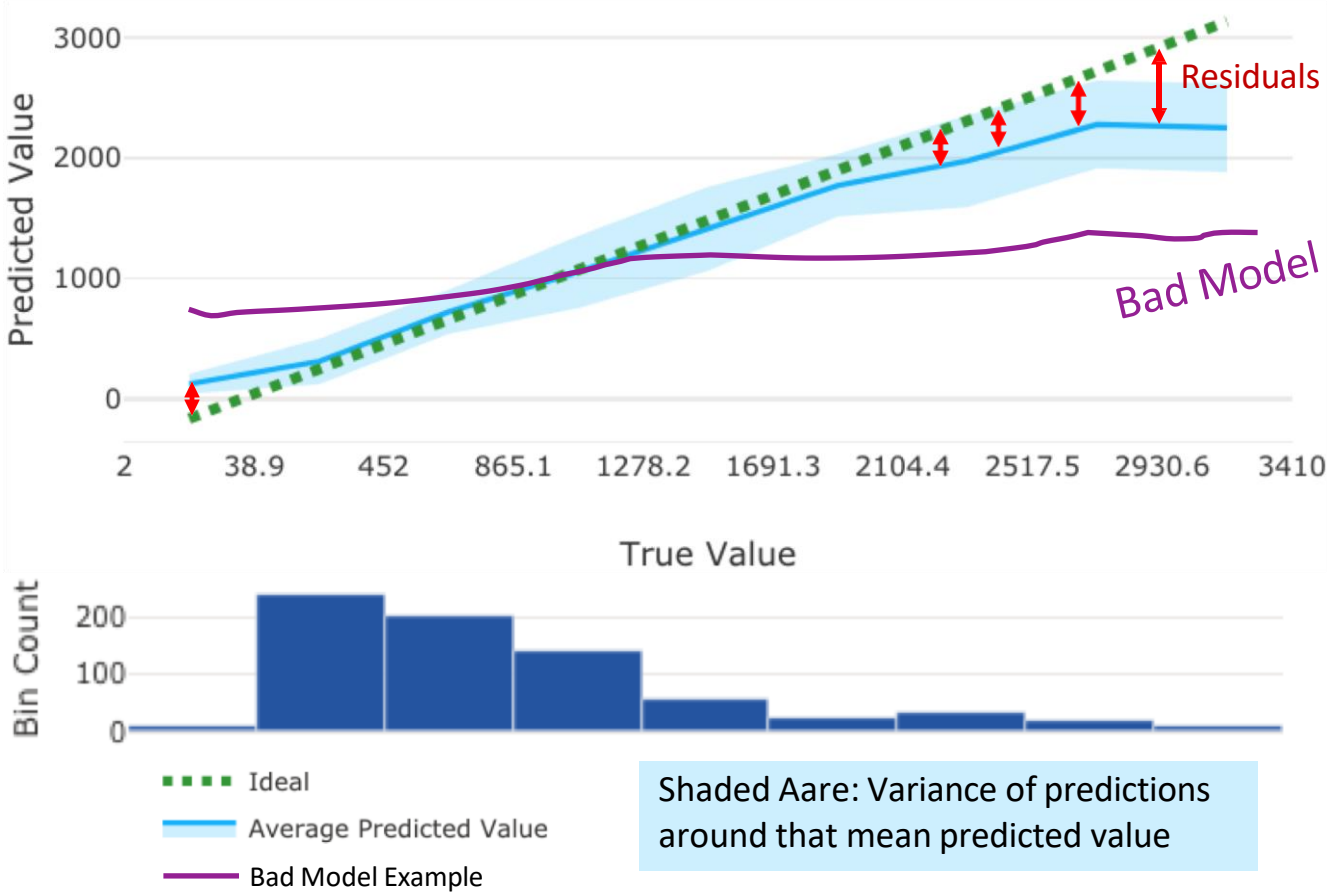
AUC = 1 **Ideal**

AUC = 0.5 **Random guessing (Y = X)**

AUC < 0.5 **Worse than random guessing**

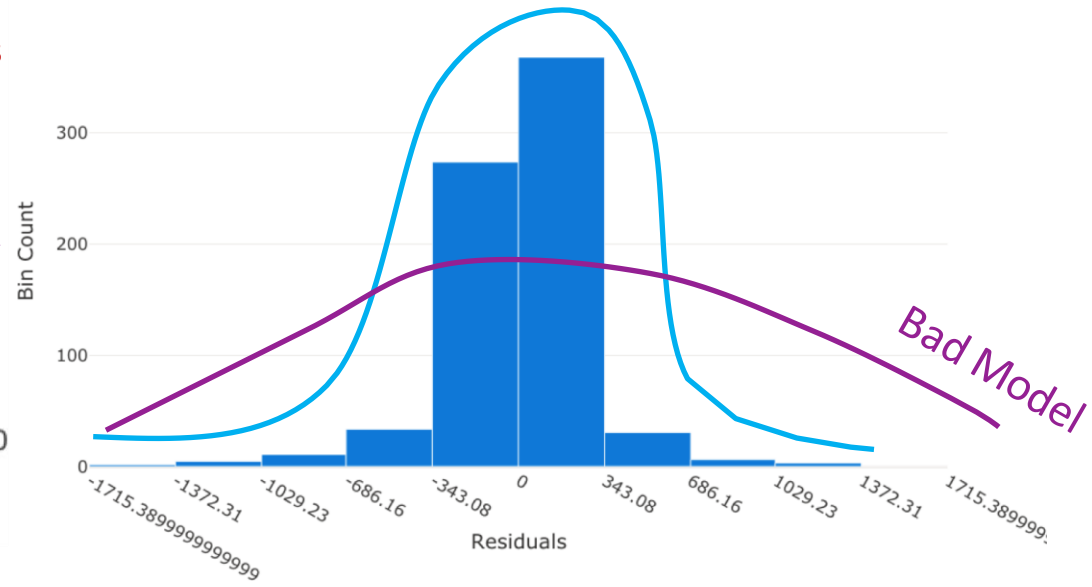
Core ML Concepts – Evaluation Metrics - Predicted vs. True

Predicted vs. True



Shaded Aare: Variance of predictions around that mean predicted value

Residual Histogram



The closer of frequently occurring residual values to 0, more accurate the ML model.

- **Residuals** (Errors) - the difference between the predicted and actual value (label)
- **Residual Histogram** – a diagram shows the frequency of residual value ranges

Core ML Concepts – Evaluation Metrics for Regression

Mean Absolute Error (MAE)	The average of the absolute errors (difference between predicted values and true values)	$MAE = \frac{\sum_{i=1}^{n} \lvert P_i - T_i \rvert}{n}$
Root Mean Squared Error (RMSE)	The square root of the mean squared errors (difference between predicted and true values)	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - T_i)^2}{n}}$
Relative Absolute Error (RAE)	The relative absolute difference between predicted and true values • compare models where the labels are in different units	$RAE = \frac{\sum_{i=1}^{n} \lvert P_i - T_i \rvert}{\sum_{i=1}^{n} \lvert \overline{T} - T_i \rvert} \quad \overline{T} = \frac{\sum_{i=1}^{n} T_i}{n}$
Relative Squared Error (RSE)	The relative squared difference between predicted and true values • compare models where the labels are in different units	$RSE = \frac{\sum_{i=1}^{n} (P_i - T_i)^2}{\sum_{i=1}^{n} \left(\overline{T} - T_i\right)^2}$
Coefficient of Determination (R2)	Represents the predictive power of the model as a value between 0 and 1	$R^2 = 1 - \frac{\sum_{i=1}^{n} (P_i - T_i)^2}{\sum_{i=1}^{n} \left(\overline{T} - T_i\right)^2} = 1 - RSE$

Core ML Concepts – Evaluation Metrics for Regression Example

	A	B	C	D	E	F	G	H
1	Number (i)	Predicted Value (Pi)	True Value (Ti)	Absolute Difference (|Pi-Ti|)	ABS Difference Square (|Pi-Ti|)^2	ABS Difference to the Mean (|Ti- (mean of Ti)|)	ABS Mean Square (|Ti- (mean of Ti)|)^2	Formula
2	1	-1	-2	1	1	5.4	29.16	
3	2	1	2	1	1	1.4	1.96	
4	3	2	1.5	0.5	0.25	1.9	3.61	
5	4	3	2.5	0.5	0.25	0.9	0.81	
6	5	3.5	4	0.5	0.25	0.6	0.36	
7	6	4	3	1	1	0.4	0.16	
8	7	5	6.5	1.5	2.25	3.1	9.61	
9	8	5.5	5.5	0	0	2.1	4.41	
10	9	6	5	1	1	1.6	2.56	
11	10	7	6	1	1	2.6	6.76	
12		SUM	34	8	8	20	59.4	SUM(row2:row11)
13		Mean	3.4					C12/A11
14			MAE	0.8				D13/A11
15			RMSE		0.894427191			SQRT(E12/A11)
16			RAE			0.4		D12/F12
17			RSE				0.134680135	E12/G12
18			R-Squared				0.865319865	1-G17

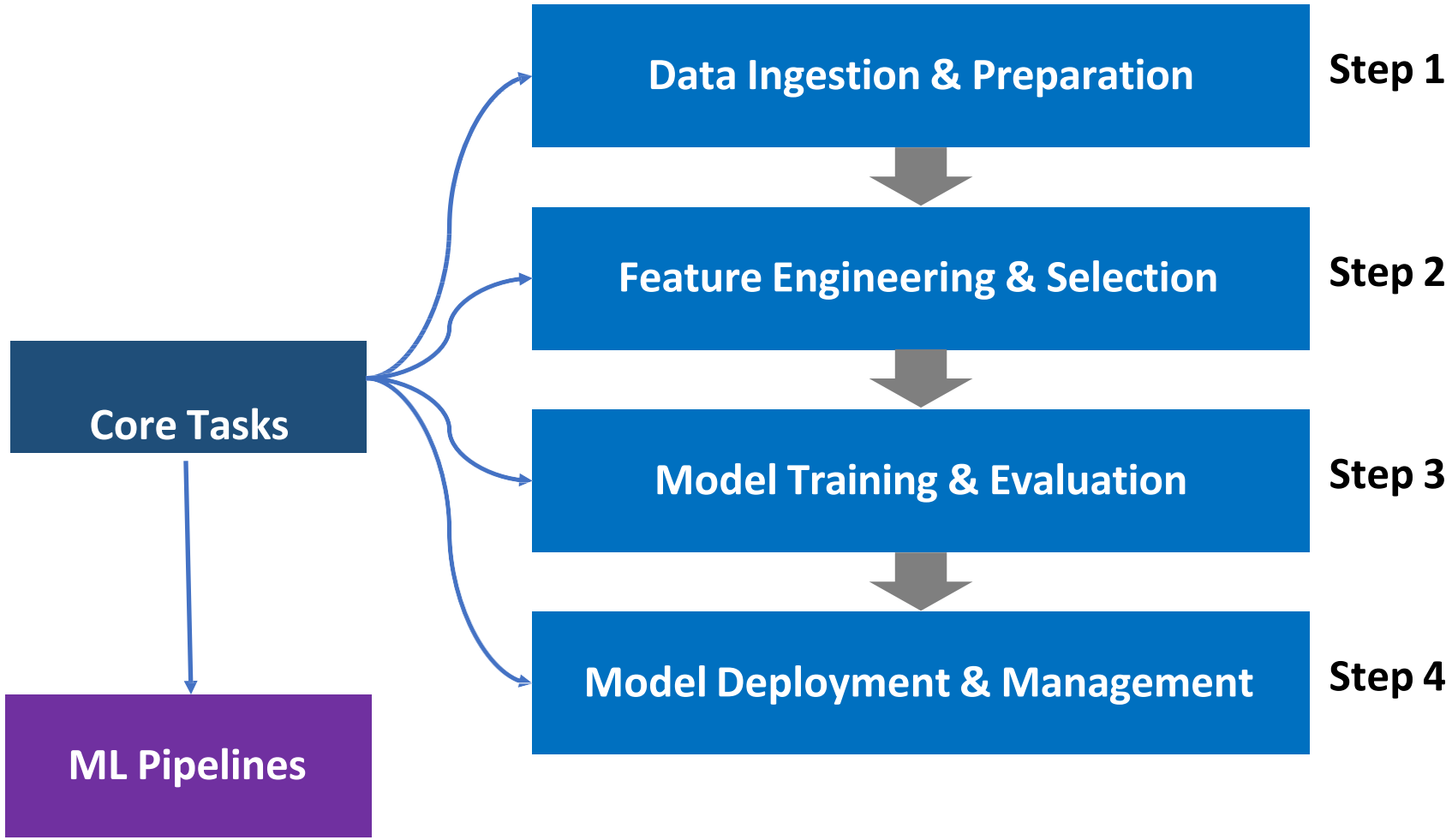
Lower the value the better

Closer to 0 the better

Closer to 1 the better

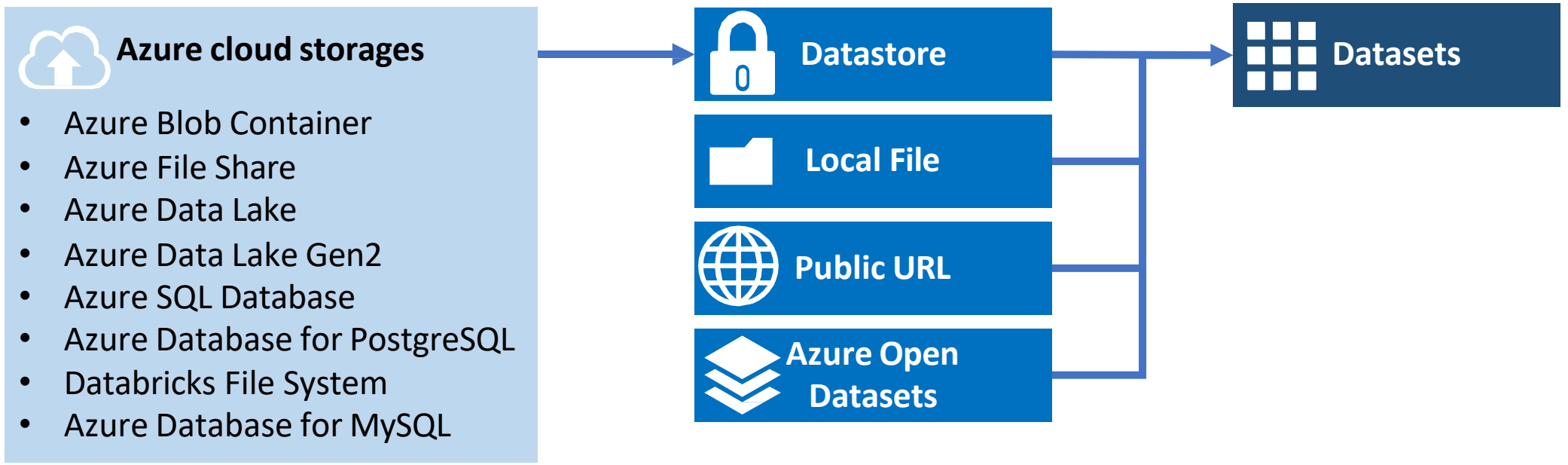
Machine Learning Solution Core Tasks

ML Solution Core Tasks Overview



ML Solution Core Tasks – Data Ingestion

- **Create Datasets** in Azure Machine Learning Designer **(Recommended)**



- **Import Data module** – import data from Azure cloud storages or public URLs
- **Enter Data Manually module** – create a small dataset by manually entering values

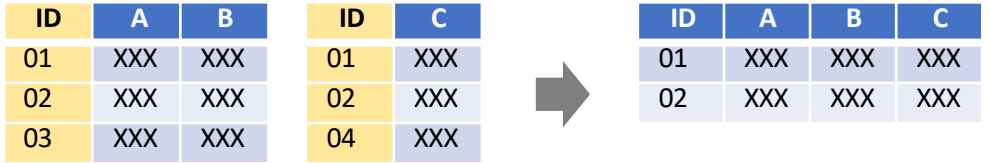
ML Solution Core Tasks – Data Preparation Common Tasks

Data Transformation

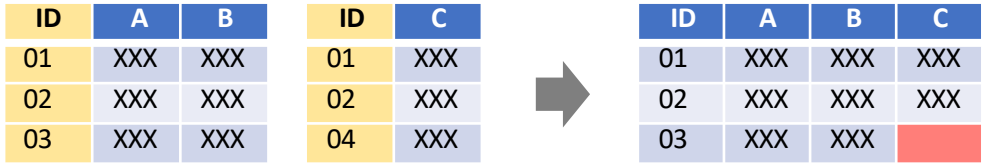
- Select Columns in Dataset
  - Exclude unwanted features (columns) from the dataset
- Add Columns
  - Combine all columns from two or more datasets
- Join Data
  - Merge two datasets using a database-style join operation based on the key column, or Composite keys using multiple columns.

  Examples:

  - **Inner Join** – combined rows when the values of key column match



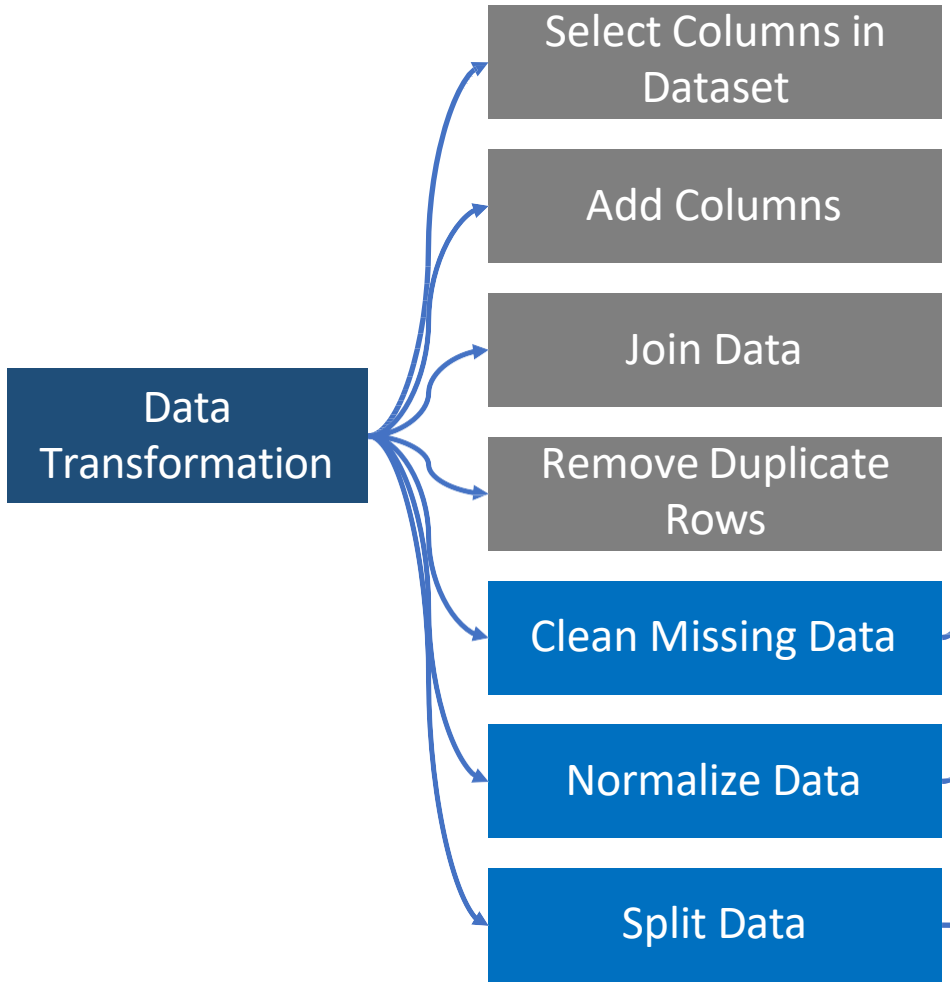
  - **Left Outer Join** – keep left table



  Key Column Missing value/ Replacement value

- Remove Duplicate Rows
  - Remove potential duplicates from a dataset
- Clean Missing Data
- Normalize Data
- Split Data

ML Solution Core Tasks – Data Preparation Common Tasks Cont.



Ignore, replace, or infer missing values

Examples:

- **Minimum missing value ratio** - [0-1, default 0] the minimum portion of missing values to be operated
- **Maximum missing value ratio** - [0-1, default 1] the maximum portion of missing values to be operated
- Custom substitution value
- Replace with mean / median / mode
- Remove entire row / column

Change the values of numeric columns to be on a common scale

Examples:

- **MinMax** - Rescaling to the [0,1]

Significantly larger values may lead to higher bias

length	width	height	curb-weight
168.8	64.1	48.8	2548
168.8	64.1	48.8	2548
171.2	65.5	52.4	2823

→

length	width	height	curb-weight
0.413433	0.316667	0.083333	0.411171
0.413433	0.316667	0.083333	0.411171
0.449254	0.433333	0.383333	0.517843

Separate data into training and validation datasets

Examples:

- **Split Rows** - [0-1, default 0.5] the fraction of rows in the first output dataset
  For example, if the value is 0.7, then the dataset is divided 70/30

ML Solution Core Tasks – Feature Engineering & Selection

- **Feature Engineering**
  - The process of creating new features from original data to help machine learning algorithms to learn better

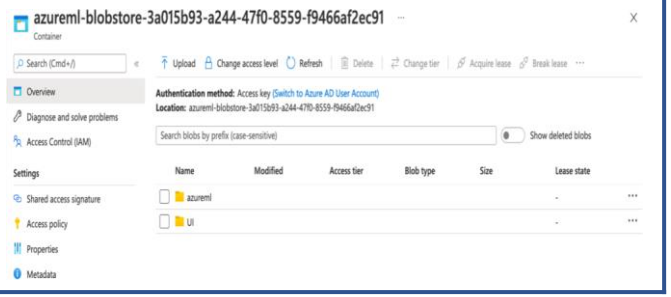
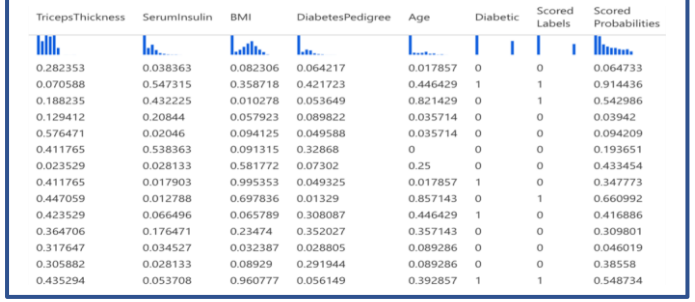
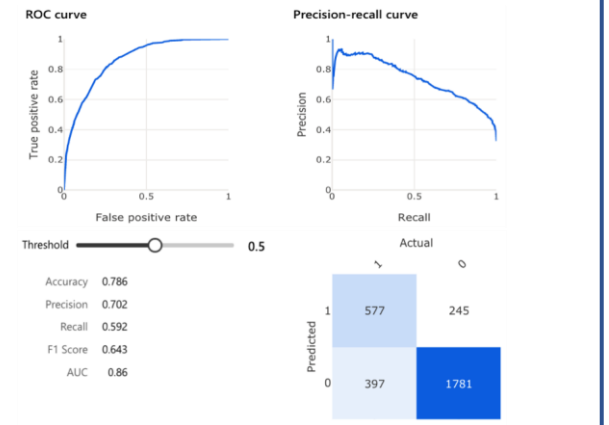
Features (Columns)

Observation / Record (Rows)

	A	B	C	D
1	**Date**	**Auctions Reported**	**Sold**	**Clearance Rate**
2	1-Apr-21	798	750	0.9398
3	2-Apr-21	841	800	0.9512
4	3-Apr-21	675	666	0.9867
5	4-Apr-21	987	780	0.7903
6	5-Apr-21	888	850	0.9572
7	6-Apr-21	767	721	0.94
8	7-Apr-21	879	801	0.9113
9	8-Apr-21	977	905	0.9263

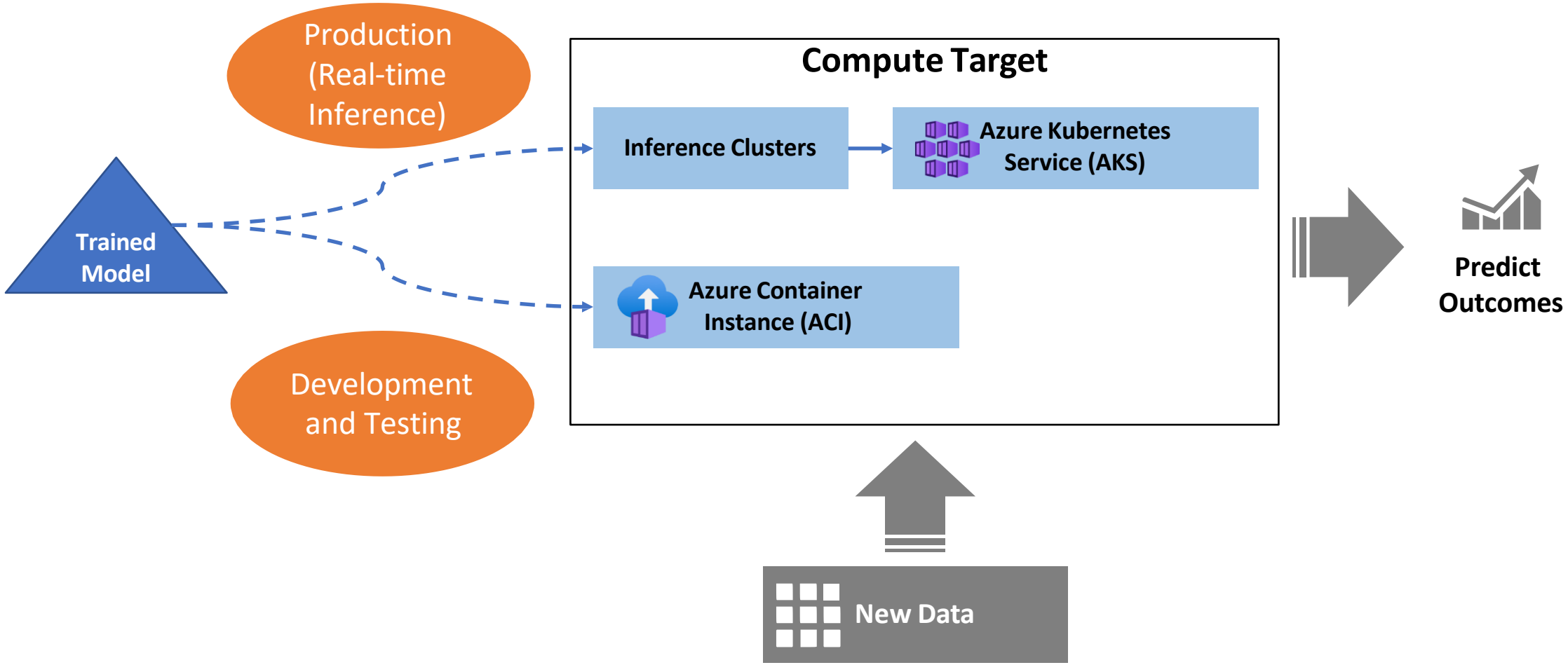
- **Automation – Featurization**
  - Automatically pre-process the features before training
  - Feature Type: DateTime
  - Engineer Features:
    - *Year*
    - *Month*
    - *Day*
    - *Day of week*
    - *Day of year*
    - *Quarter*
    - *Week of the year*
    - *Hour*
    - *Minute*
    - *Second*
- **Feature Selection**
  - The purpose is to narrow the features by selecting a subset of relevant, useful features of original data to reduce noise and improve training performance
  - The process of applying statistical tests to the input datasets to determine which columns are more predictive of the output

ML Solution Core Tasks – Model Training & Evaluation

	Train Model	Score Model	Evaluate Model
Purpose	**Train a ML model**	**Generate predictions on new actual data**	**Measures the accuracy of a trained model**
Azure ML Module	• **Train Model module** – for a regression or classification model • **Train Clustering Model module** – for a clustering model	• **Score Model module** – for a regression or classification model • **Assign Data to Clusters module** – for a clustering model	• **Evaluate Model module** – for a regression or classification or clustering model
Input / Output	• **Input**: dataset for training and algorithm • **K-means clustering** **Number of centroids** parameter – the number of clusters for the model 	• **Input**: dataset for validation (Split Data) 	• **Output**: evaluation metrics for comparison of the performance of multiple training models 

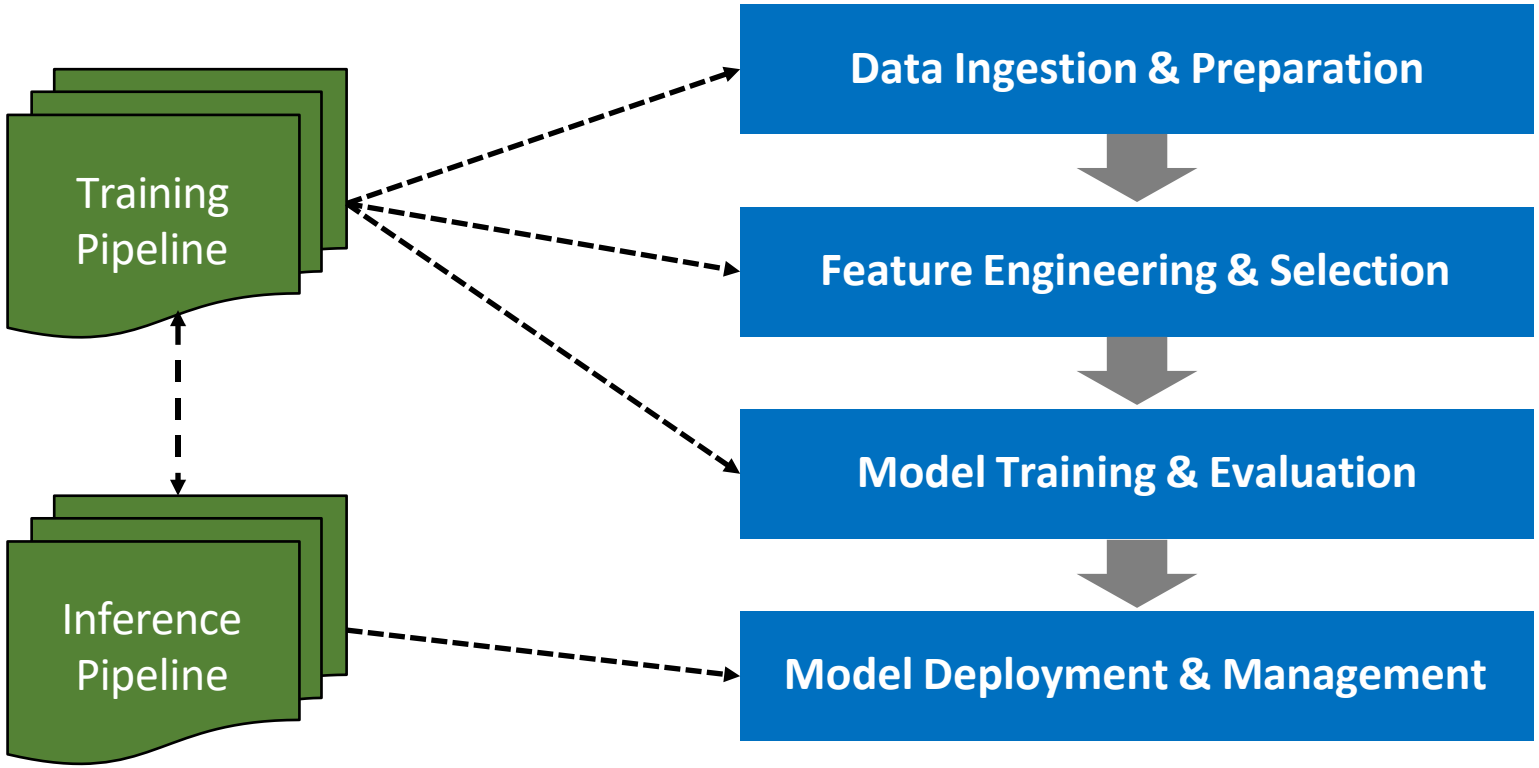
TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic	Scored Labels	Scored Probabilities
0.282353	0.038363	0.082306	0.064217	0.017857	0	0	0.064733
0.070588	0.547315	0.358718	0.421723	0.446429	1	1	0.914436
0.188235	0.432225	0.010278	0.053649	0.821429	0	1	0.542986
0.129412	0.20844	0.057923	0.089822	0.035714	0	0	0.03942
0.576471	0.02046	0.094125	0.049588	0.035714	0	0	0.094209
0.411765	0.538363	0.091315	0.32868	0	0	0	0.193651
0.023529	0.028133	0.581772	0.07302	0.25	0	0	0.433454
0.411765	0.017903	0.995353	0.049325	0.017857	1	0	0.347773
0.447059	0.012788	0.697836	0.01329	0.857143	0	1	0.660992
0.423529	0.066496	0.065789	0.308087	0.446429	1	0	0.416886
0.364706	0.176471	0.23474	0.352027	0.357143	0	0	0.309801
0.317647	0.034527	0.032387	0.028805	0.089286	0	0	0.046019
0.305882	0.028133	0.08929	0.291944	0.089286	0	0	0.38558
0.435294	0.053708	0.960777	0.056149	0.392857	1	1	0.548734

ML Solution Core Tasks – Model Deployment & Management

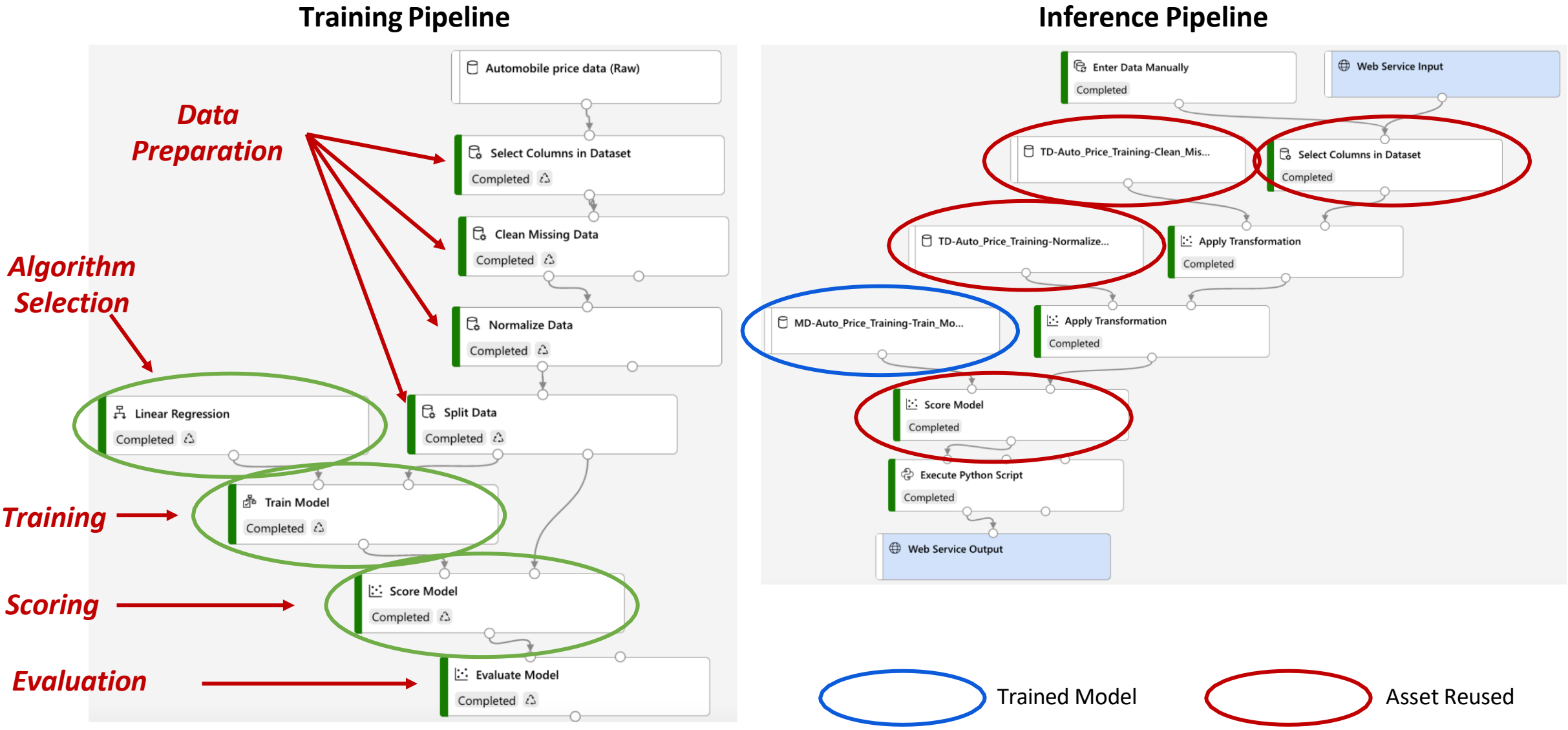


ML Solution Core Tasks – Building Azure ML Pipelines

- **Azure Machine Learning Pipelines**
  - Orchestrate data preparation, model training, deployment, and tasks management
  - Help automate machine learning workflows

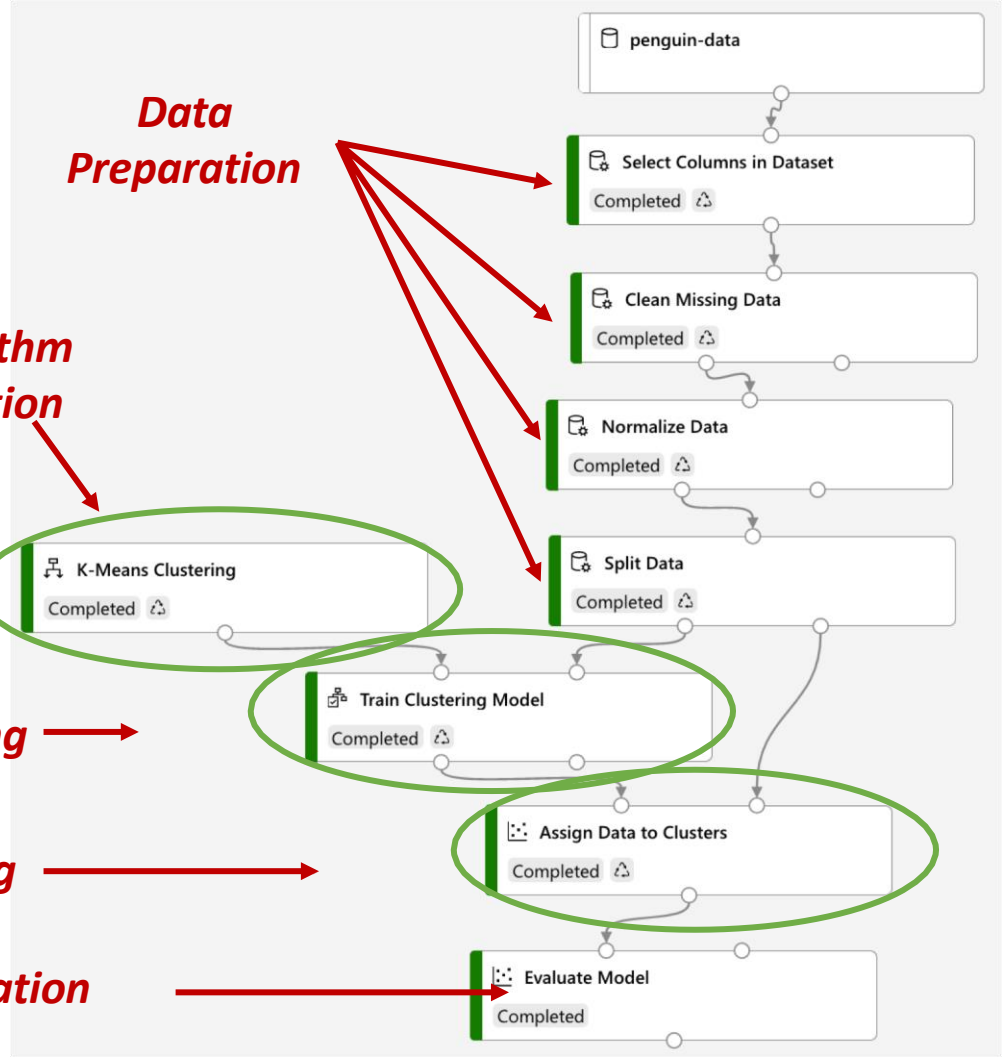


ML Solution Core Tasks – Creating Pipelines – Regression Model

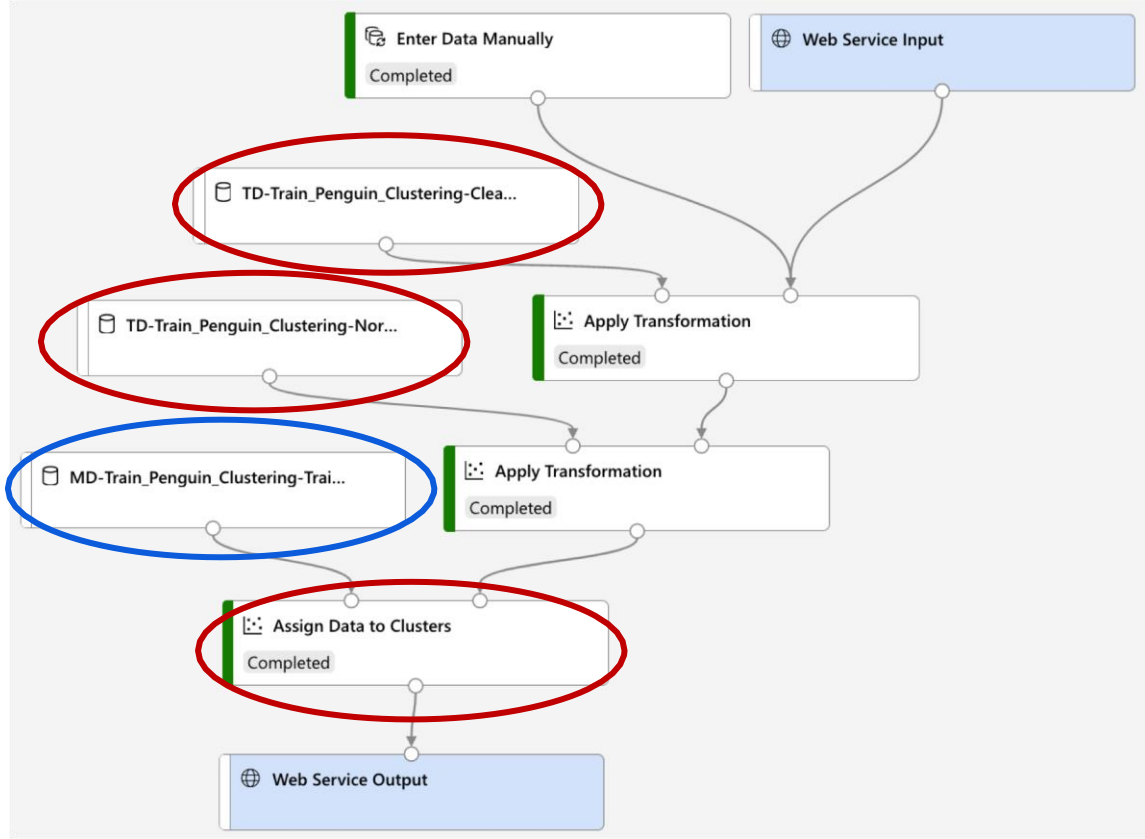


ML Solution Core Tasks – Creating Pipelines – Classification Model

Training Pipeline



Inference Pipeline



Trained Model

Asset Reused

Azure Machine Learning Workspace

Azure Machine Learning – Workspace

- **Azure Machine Learning Workspace**

A centralized place to manage data, compute resources, code, models, and other artefacts related to your machine learning workloads.

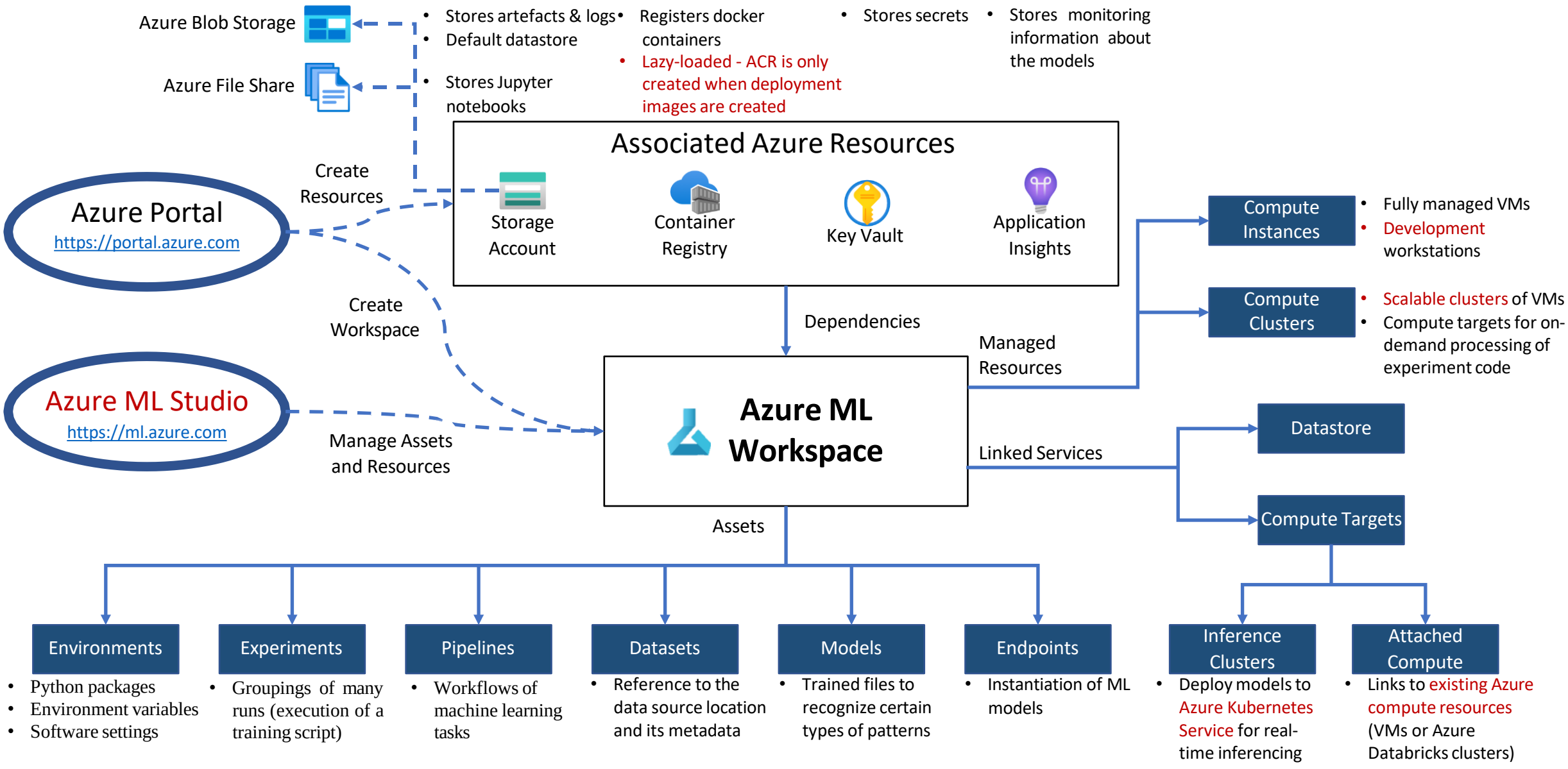
- **Workspace Management**

Business/Data Experts — **No Code**

Programmer — **Coding**

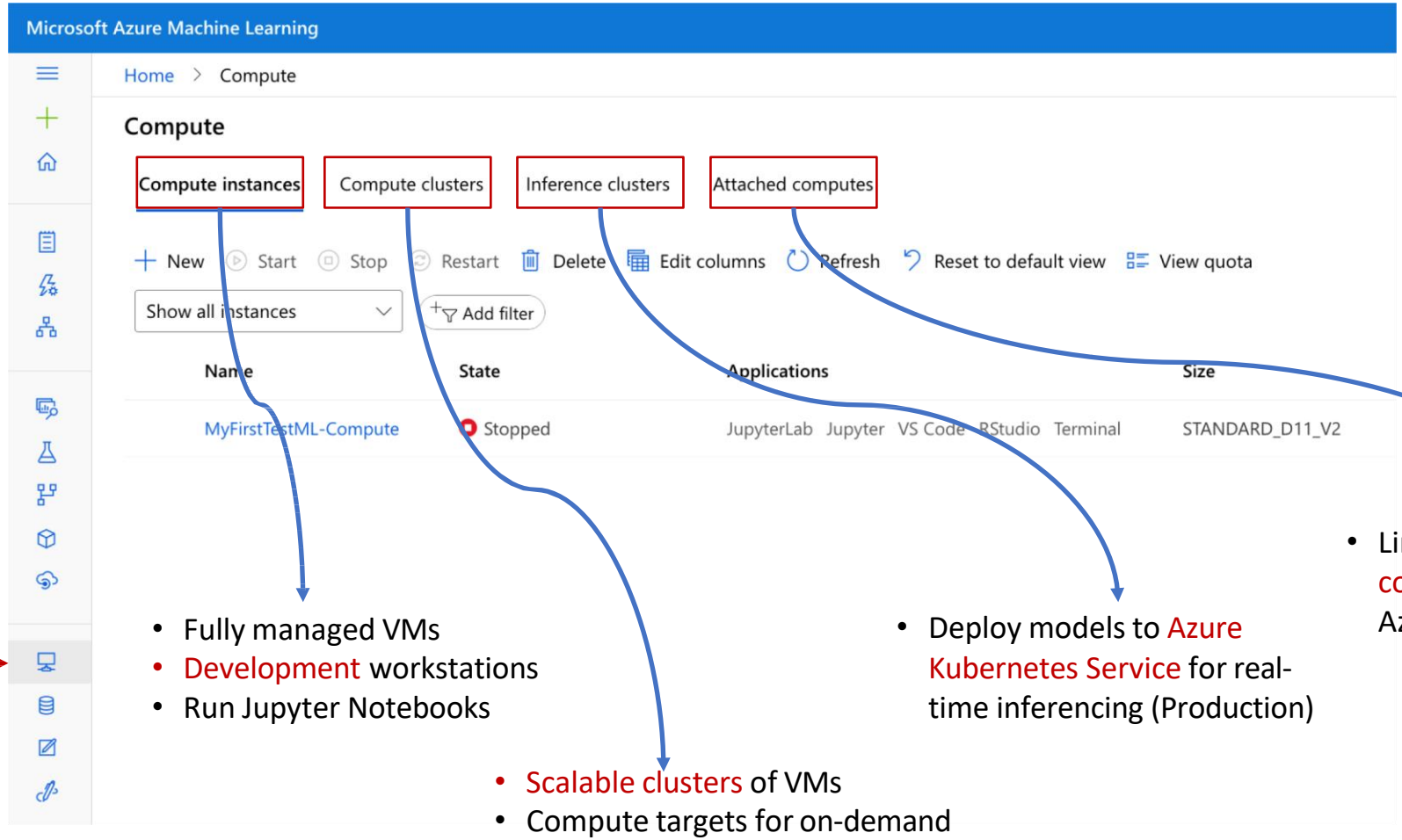
Task	Azure Portal	Azure ML Studio	Python SDK / R SDK	Azure ML CLI	Visual Studio Code
Create a workspace	X		X	X	X
Manage workspace access	X			X	
Create and manage compute resources	X	X	X	X	
Create a Notebook VM		X			

Azure Machine Learning – Workspace Architecture



Azure Machine Learning – Compute Resource

Compute



- Fully managed VMs
- Development workstations
- Run Jupyter Notebooks

- Scalable clusters of VMs
- Compute targets for on-demand processing of experiment code
  - Run Training Pipelines
  - Run Inference Pipelines

- Deploy models to Azure Kubernetes Service for real-time inferencing (Production)

- Links to existing Azure compute resources (VMs or Azure Databricks clusters)

Demo

Creating an Azure ML Workspace

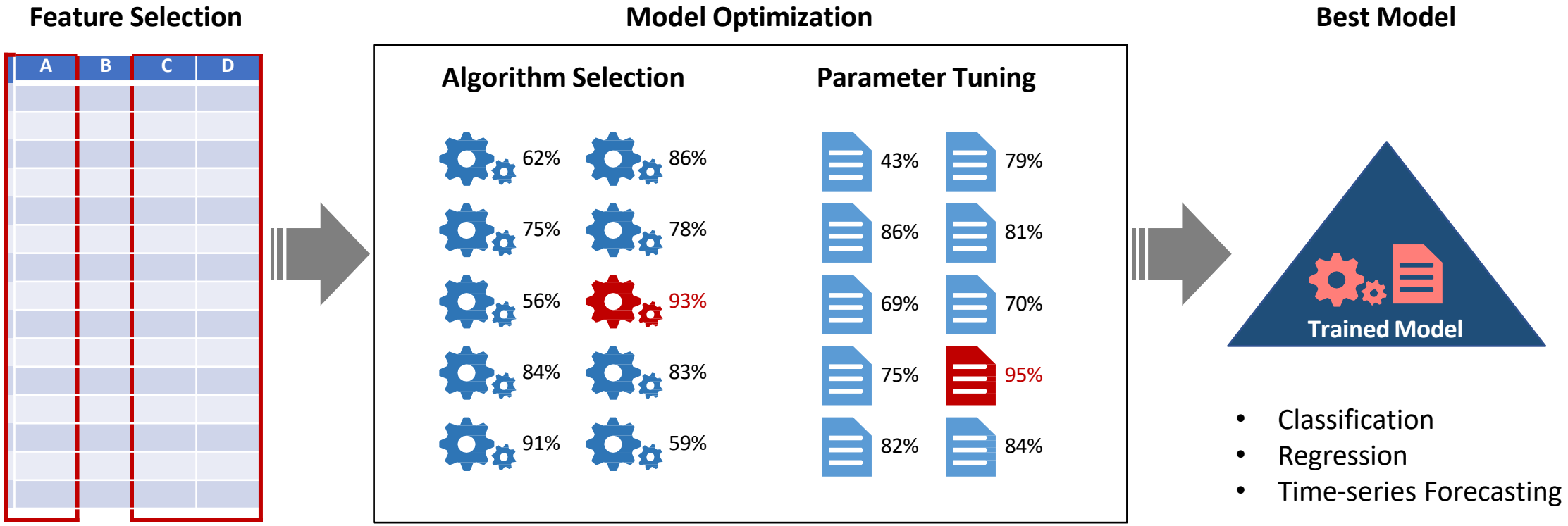
in

Azure Portal

No-code Machine Learning with Azure ML Studio #1 AutoML

Azure Machine Learning – Automated ML (AutoML)

- **AutoML** - the process of automating the time-consuming, iterative tasks of machine learning model development



- Classification
- Regression
- Time-series Forecasting

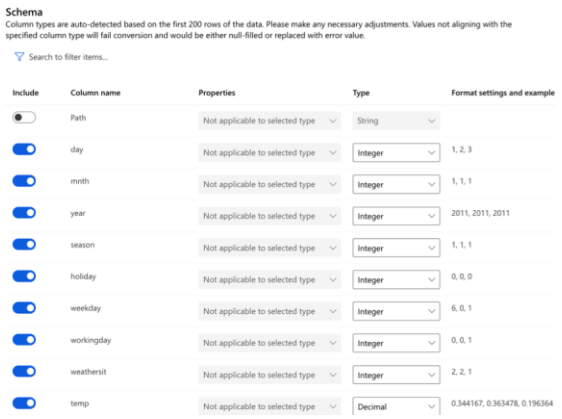
Azure Machine Learning – AutoML Model Training Steps

Create Dataset	Create Experiment	Train Model	Evaluate Model	Deploy Model

Upload dataset

temp	atemp	hum	windspeed	rentals
0.344	0.364	0.806	0.16	331
0.363	0.354	0.696	0.249	131
0.196	0.189	0.437	0.248	120
0.2	0.212	0.59	0.16	108
0.227	0.229	0.437	0.187	82
0.204	0.233	0.518	0.09	88
0.197	0.209	0.499	0.169	148
0.165	0.162	0.536	0.267	68
0.138	0.116	0.434	0.362	54

Feature Selection



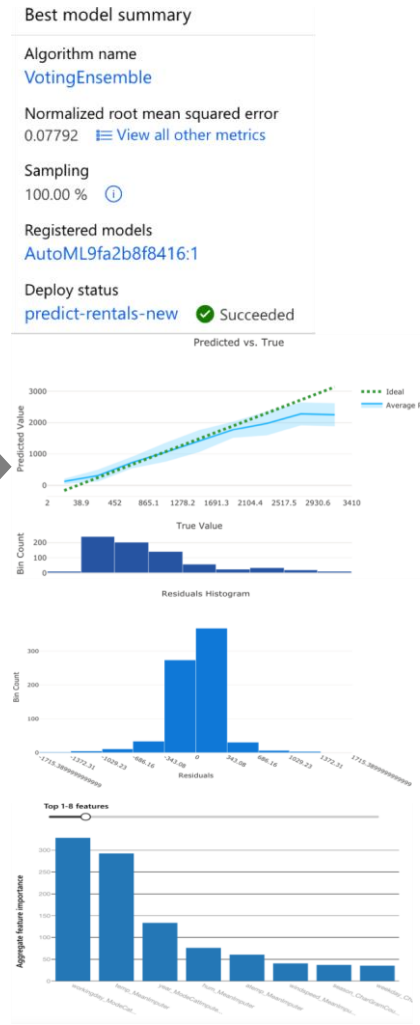
Model Selection

- Classification
- Regression
- Time series forecasting

Settings

- Algorithm selection
- Parameter setting
- Featurization

Automated Training



Production (Real-time Inference)

Azure Kubernetes Service (AKS)

Development & Testing

Azure Container Instance (ACI)

Hands-on Lab

Training a Regression Model

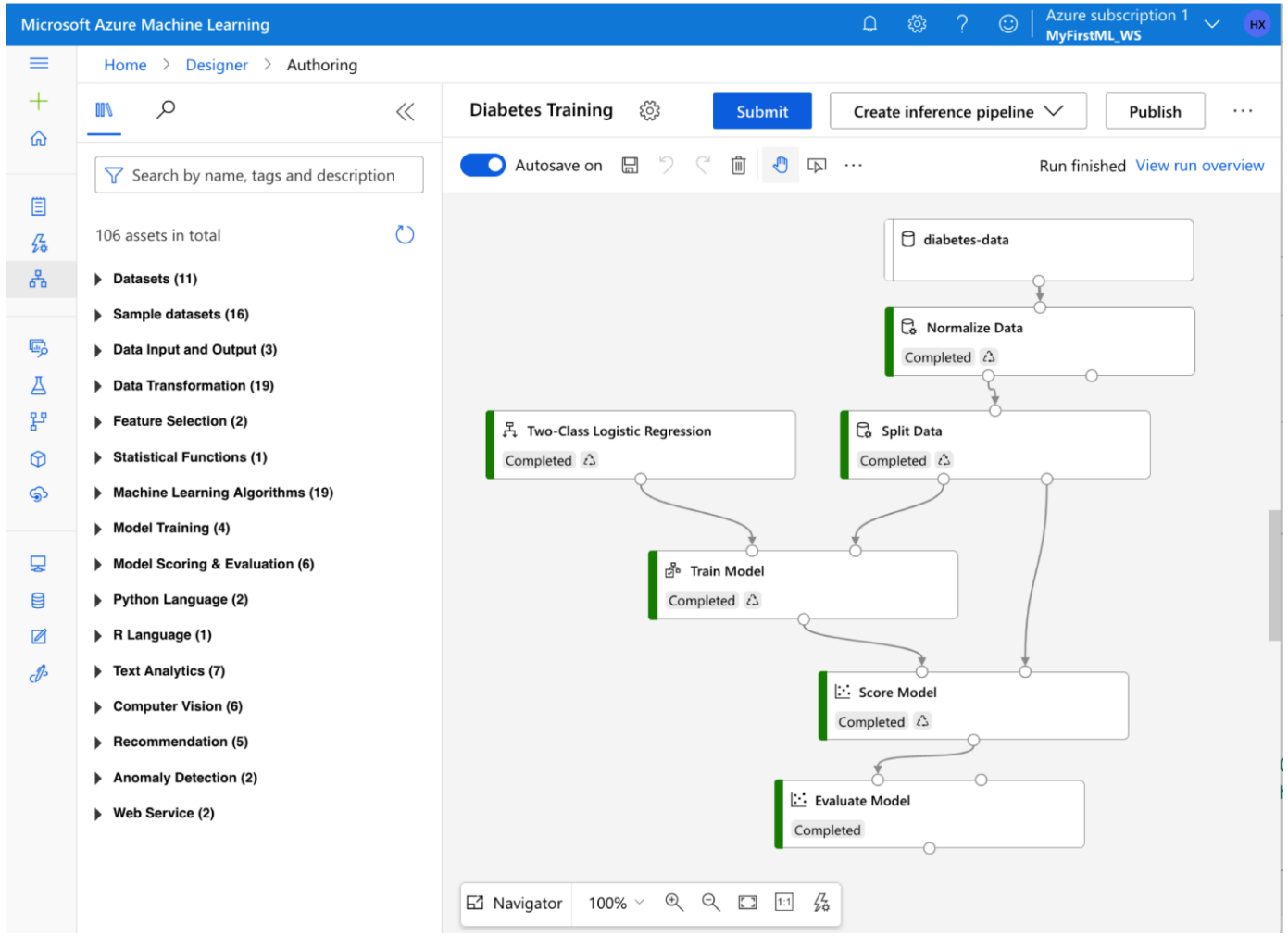
with

Azure Automated ML

No-code Machine Learning with Azure ML Studio #2 Designer

Azure Machine Learning – ML Designer

- **Azure ML designer** provides a drag-and-drop visual canvas to build, test, and deploy ML models.
  - Datasets
  - Modules
  - Pipelines
- **Supported open-source frameworks**
  - MLflow
  - Kubeflow
  - ONNX (Open Neural Network Exchange)
  - PyTorch
  - TensorFlow
- **Supported programming languages**
  - Python and R



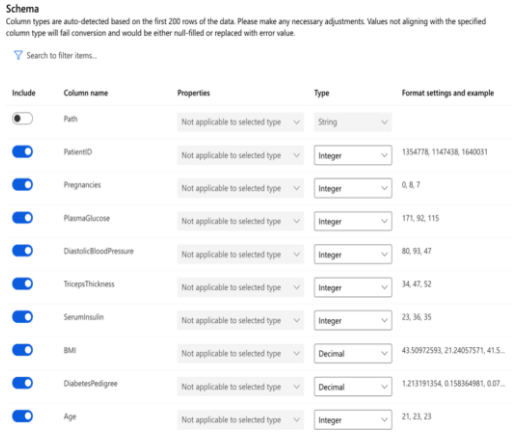
Azure Machine Learning – ML Designer Model Training Steps

Create Dataset	Create Training Pipeline	Evaluate Model	Create Inference Pipeline	Deploy Model

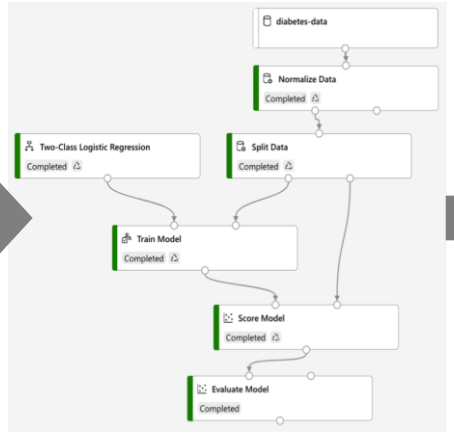
Upload dataset

Id	PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPr...
1	1354778	0	171	80
2	1147438	8	92	93
3	1640031	7	115	47
4	1883350	9	103	78
5	1424119	1	85	59
6	1619297	0	82	92
7	1660149	0	133	47
8	1458769	0	67	87
9	1201647	8	80	95

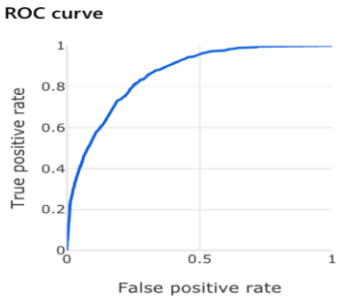
Feature Selection



Feature Engineering & Selection
Model Selection
Algorithm selection

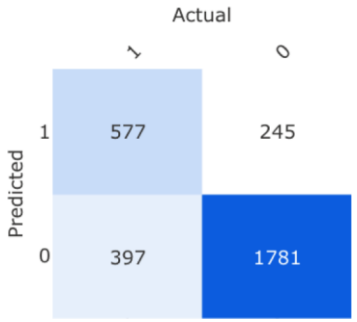


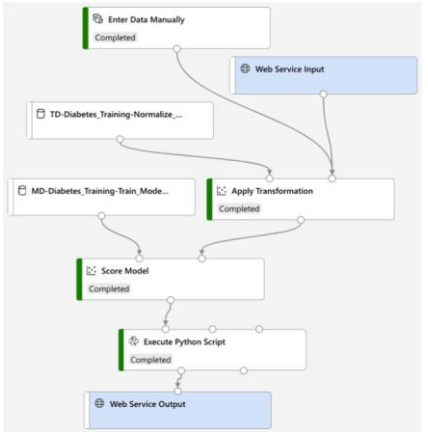
Run the training Pipeline



Threshold 0.5

Accuracy	0.786
Precision	0.702
Recall	0.592
F1 Score	0.643
AUC	0.86





Run the Inference Pipeline

PatientID	DiabetesPrediction	Probability
1882185	1	0.703411
1662484	1	0.852884
1228510	1	0.85252

Production (Real-time Inference)

Azure Kubernetes Service (AKS)

Development & Testing

Azure Container Instance (ACI)

Hands-on Lab

Training a Classification Model

with

Azure ML Designer

Azure Machine Learning Summary

Azure Machine Learning Summary