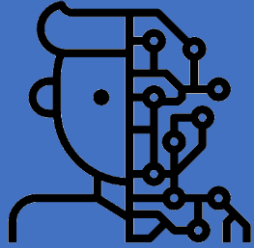


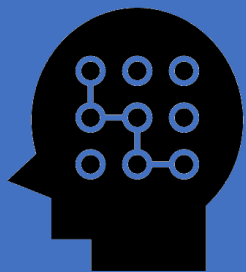


What is Machine Learning



## What machine learning does?

- Find patterns in data
- Uses those patterns to predict the future
- Examples:
  - Detecting credit card Fraud
  - Determine whether a customer is likely to switch to a competitor
  - Deciding when to do preventive maintenance on a factory robot



What does it mean to learn?

- How did you learn to read?
  - Learning requires:
    - Identity patterns
    - Recognizing those patterns when you see them again
- “This is what machine learning does”*

## Finding Patterns: A Simple Example

Name	Amount	Fraudulent
Amit	\$3200	No
Rahul	\$1300	Yes
Ramesh	\$5700	Yes
Vinay	\$5700	No

**What's the pattern for fraudulent transactions?**

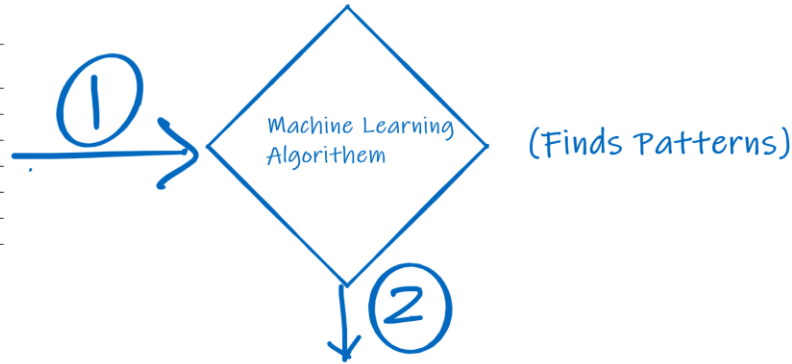
## Finding Patterns: Another Example

Name	Amount	Where Issued	Where Used	Age	Fraudulent
Sameer	\$2500	USA	USA	23	No
Payal	\$2394	USA	RUS	27	Yes
Piyush	\$1009	USA	RUS	25	Yes
Amit	\$8488	FRA	USA	63	No
Peter	\$298	AUS	JAP	59	No
Jones	\$3150	USA	RUS	43	No
Harry	\$8155	USA	RUS	25	Yes
Mark	\$7475	UK	GER	31	No
Nancy	\$550	USA	RUS	26	No
Eli	\$7345	USA	RUS	19	Yes

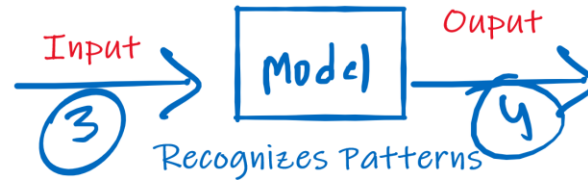
# Machine Learning in a Nutshell

(Contain Pattern)

Name	Amount	Where Issued	Where Used	Age	Fraudulent
Shameer	\$2500	USA	USA	23	No
Payal	\$2394	USA	RUS	27	Yes
Piyush	\$1009	USA	RUS	25	Yes
Amit	\$8488	FRA	USA	63	No
Peter	\$298	AUS	JAP	59	No
Jones	\$3150	USA	RUS	43	No
Harry	\$8155	USA	RUS	25	Yes



Name	Amount	Where Issued	Where Used	Age	Fraudulent
Mark	\$7475	UK	GER	31	?
Nancy	\$550	USA	RUS	26	?
Eli	\$7345	USA	RUS	19	?



Name	Amount	Where Issued	Where Used	Age	Fraudulent
Mark	\$7475	UK	GER	31	No
Nancy	\$550	USA	RUS	26	No
Eli	\$7345	USA	RUS	19	Yes

# Relationship between AI and ML

## Artificial Intelligence

An umbrella term, which we can loosely describe as:  
“all the various techniques we might use to make a computer do something smart”

## Machine Learning

A successful subset of AI that focuses on  
**learning from data**, not on programming explicit rules to follow



Rule based programming

vs



Machine Learning



# Rule based Programming vs ML



We write explicit rules  
to follow

We provide data to  
learn from

# Rule based Analysis



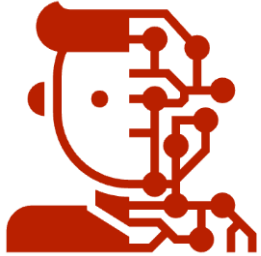
Problem statement is fairly simple

Rules are straightforward and can be easily codified

Rules can change frequently

Few problem instances to train ML model

# ML based Analysis



Problem statement is reasonably complex

Hard to find patterns using visualizations and other exploratory tools

Decision variables sensitive to data, need to change as new information is received

Large corpus available to train models

# ML based and Rule based Models

## ML - based

- Dynamic – alter output based on patterns in data
- Expert skill not needed, need an intuition for how models work
- To update model, update corpus
- Large, high-quality data corpus
- Can not operate on a single problem instance

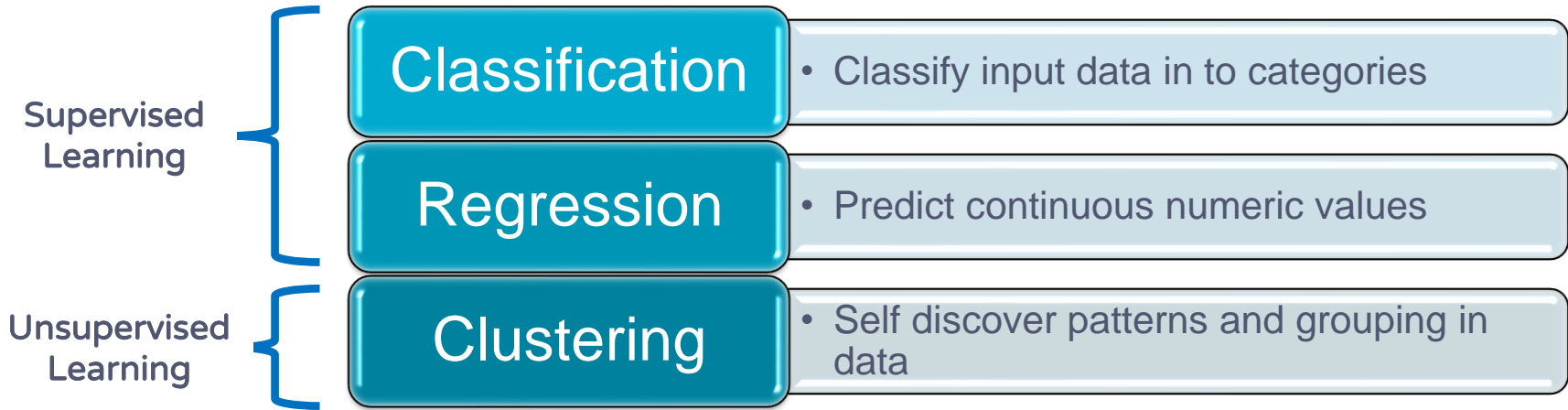
## Rule - based

- Static – rules are applied independent of data
- Experts vital for formulating rules, experts based on problem
- To update model, need to update rules i.e. record model
- No corpus required
- Can operate on isolated problem instances



# Machine Learning types

# Machine Learning Types



# Classification

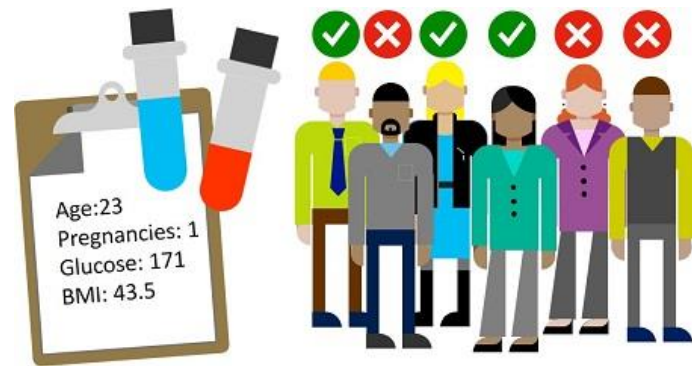
- Classify input data in to categories
- Make predictions in a non-continuous form
- Classification is binary (either A or B) or multiclass (A or B or C ...etc.)
- Supervised learning

## Examples: Binary classification

- Email: Spam or Not Spam?
- Identify sentiment as positive or negative.
- Determine whether a patient's lab sample is cancerous.

## Examples: Multiclass classification

- Email filters as spam, junk, or good.
- Stocks: Buy, sell or hold?
- Images: Cat, dog or mouse?
- Positive, negative or neutral sentiments?



# Classification

## Features

## Label

Id	PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPr...	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic
1	1354778	0	171	80	34	23	43.50972593	1.213191354	21	0
2	1147438	8	92	93	47	36	21.24057571	0.158364981	23	0
3	1640031	7	115	47	52	35	41.51152348	0.079018568	23	0
4	1883350	9	103	78	25	304	29.58219193	1.282869847	43	1
5	1424119	1	85	59	27	35	42.60453585	0.549541871	22	0
6	1619297	0	82	92	9	253	19.72416021	0.103424498	26	0
7	1660149	0	133	47	19	227	21.94135672	0.174159779	21	0
8	1458769	0	67	87	43	36	18.2777226	0.23616494	26	0
9	1201647	8	80	95	33	24	26.62492885	0.443947388	53	1
10	1403912	1	72	31	40	42	36.88957571	0.103943637	26	0
11	1943830	1	88	86	11	58	43.22504089	0.230284623	22	0
12	1824483	3	94	96	31	36	21.29447943	0.259020482	23	0
13	1848869	5	114	101	43	70	36.49531966	0.079190164	38	1
14	1669231	7	110	82	16	44	36.08929341	0.281276159	25	0
15	1683688	0	148	58	11	179	39.19207553	0.160829008	45	0
16	1738587	3	109	77	46	61	19.84731197	0.204345272	21	1
17	1884264	3	106	64	25	51	29.0445728	0.589188017	42	1
18	1485251	1	156	53	15	226	29.78619164	0.203823525	41	1
19	1536832	8	117	39	32	164	21.23099598	0.089362745	25	0
20	1438701	3	102	100	25	289	42.18572029	0.175592826	43	1



# Regression

- Predict a numeric value, typically in a continuous form
- learning from labeled historical data to predict or forecast new values
- Supervised learning

## Examples:

- Given past stock data predict price tomorrow
- Given location and attributes of a home predict price



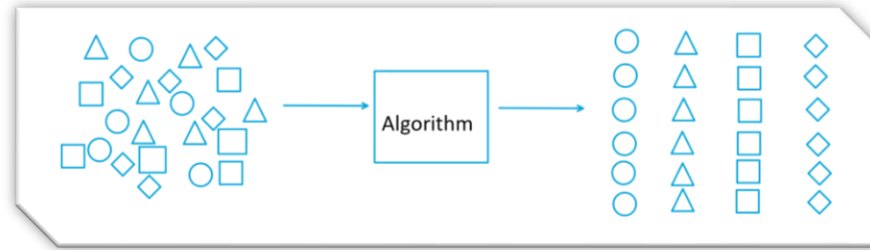
# Regression

## Features

## Label

symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	length	width	height	curb-weight	engine-type	num-of-cylinders	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price
3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	13495
3	NaN	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9	111	5000	21	27	16500
1	NaN	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	9	154	5000	19	26	16500
2	164	audi	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.4	10	102	5500	24	30	13950
2	164	audi	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.4	8	115	5500	18	22	17450
2	NaN	audi	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	15250
1	158	audi	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	17710
1	NaN	audi	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136	mpfi	3.19	3.4	8.5	110	5500	19	25	18920
1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131	mpfi	3.13	3.4	8.3	140	5500	17	20	23875
0	NaN	audi	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053	ohc	five	131	mpfi	3.13	3.4	7	160	5500	16	22	NaN
2	192	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16430
0	192	bmw	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108	mpfi	3.5	2.8	8.8	101	5800	23	29	16925
0	188	bmw	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164	mpfi	3.31	3.19	9	121	4250	21	28	20970

# Clustering

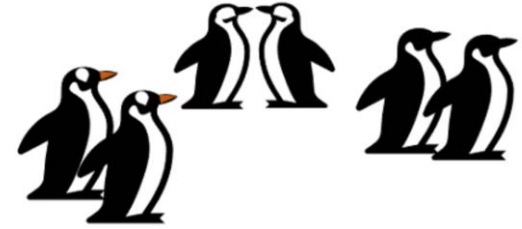


- Find cases with similar characteristics in an unlabeled dataset and group them together
- Unsupervised learning







## Examples:

- Document discovery – find all documents related to homicide cases
- Social media ad targeting – find all users who are interested in sports

# Clustering



## Features

 Id	 CulmenLength	 CulmenDepth	 FlipperLength	 BodyMass	 Species
1	39.1	18.7	181	3750	0
2	39.5	17.4	186	3800	0
3	40.3	18	195	3250	0
4	null	null	null	null	0
5	36.7	19.3	193	3450	0
6	39.3	20.6	190	3650	0
7	38.9	17.8	181	3625	0
8	39.2	19.6	195	4675	0
9	34.1	18.1	193	3475	0
10	42	20.2	190	4250	0
11	37.8	17.1	186	3300	0
12	37.8	17.3	180	3700	0
13	41.1	17.6	182	3200	0
14	38.6	21.2	191	3800	0
15	34.6	21.1	198	4400	0
16	36.6	17.8	185	3700	0
17	38.7	19	195	3450	0
18	42.5	20.7	197	4500	0
19	34.4	18.4	184	3325	0
20	46	21.5	194	4200	0

# Supervised vs Unsupervised Learning



Classification

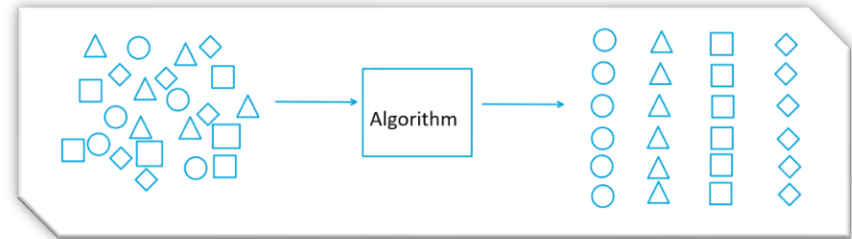


Regression



Supervised Learning

Unsupervised Learning



Clustering

## Figure out right machine learning type?

- Predicting the online sales volume for the next financial quarter?
- Analyzing X-ray images to detect whether a person has pneumonia?
- Grouping together online shoppers with similar traits for targeted marketing?
- Forecasting stock market index values based on macro economic changes?
- Processing new tweets to categorize them as positive or negative?
- Predict ice cream sales based on the weather forecast?
- Determine the amount of credit to give to a customer?
- Determine if a social media post has positive or negative sentiment?
- Approve or reject a customer's application for credit?

# Questions: Figure out right machine learning type?

- Predicting the online sales volume for the next financial quarter? -> **Regression**
- Analyzing X-ray images to detect whether a person has pneumonia? -> **Classification**
- Grouping together online shoppers with similar traits for targeted marketing? -> **Clustering**
- Forecasting stock market index values based on macro economic changes? -> **Regression**
- Processing new tweets to categorize them as positive or negative? -> **Classification**
- Predict ice cream sales based on the weather forecast? -> **Regression**
- Determine the amount of credit to give to a customer? -> **Regression**
- Determine if a social media post has positive or negative sentiment? -> **Classification**
- Approve or reject a customer's application for credit? -> **Classification**



# Feature Engineering and Selection



# Feature Selection

## Features

## Label

Id	PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPr...	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic
1	1354778	0	171	80	34	23	43.50972593	1.213191354	21	0
2	1147438	8	92	93	47	36	21.24057571	0.158364981	23	0
3	1640031	7	115	47	52	35	41.51152348	0.079018568	23	0
4	1883350	9	103	78	25	304	29.58219193	1.282869847	43	1
5	1424119	1	85	59	27	35	42.60453585	0.549541871	22	0
6	1619297	0	82	92	9	253	19.72416021	0.103424498	26	0

Name	Degree	Certification	Age	Experience	Location	Phone	height	Salary

# Feature Selection

**Feature Selection** is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

- Feature selection has a huge impact on the performance of the model in XL
- You need to identify and remove irrelevant or partially relevant features
- **Principle** – Minimum redundancy and maximum relevance

## **Benefits of selecting features?**

- **Reduces Overfitting:** Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy:** Less misleading data means modeling accuracy improves.
- **Reduces Training Time:** fewer data points reduce algorithm complexity and algorithms train faster.

# Feature Engineering

- **Feature engineering** is the process of creating new features from raw data to increase the predictive power of the machine learning model.
- Engineered features capture additional information that is not available in the original feature set.
- Examples of feature engineering are aggregating data, calculating a moving average, and calculating the difference over time

# Question: Feature Engineering

Question: Splitting the address field into country, city and street number can be appropriately mapped to \_\_\_\_\_ machine learning task.

- A. Feature engineering
- B. Feature selection

Question: You need to map the right Learning task for a given scenario?

“Picking temperature and pressure to train a weather model”

- A. Feature engineering
- B. Feature selection

Training

vs

Validation  
dataset

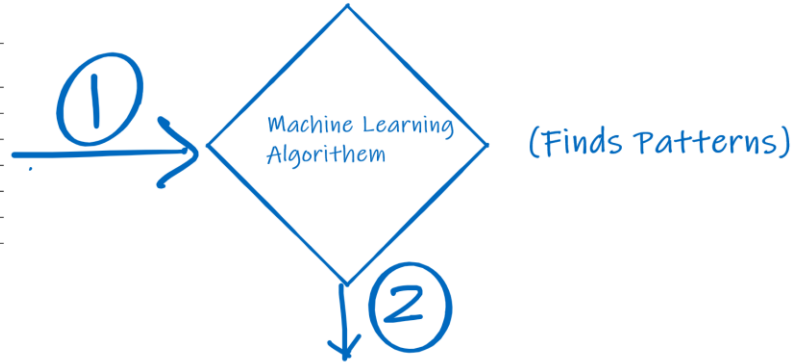
vs

testing

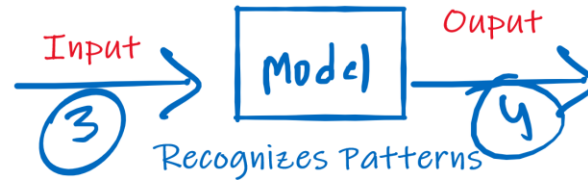
# Machine Learning in a Nutshell

(Contain Pattern)

Name	Amount	Where Issued	Where Used	Age	Fraudulent
Shameer	\$2500	USA	USA	23	No
Payal	\$2394	USA	RUS	27	Yes
Piyush	\$1009	USA	RUS	25	Yes
Amit	\$8488	FRA	USA	63	No
Peter	\$298	AUS	JAP	59	No
Jones	\$3150	USA	RUS	43	No
Harry	\$8155	USA	RUS	25	Yes



Name	Amount	Where Issued	Where Used	Age	Fraudulent
Mark	\$7475	UK	GER	31	?
Nancy	\$550	USA	RUS	26	?
Eli	\$7345	USA	RUS	19	?



Name	Amount	Where Issued	Where Used	Age	Fraudulent
Mark	\$7475	UK	GER	31	No
Nancy	\$550	USA	RUS	26	No
Eli	\$7345	USA	RUS	19	Yes



# Training vs Validation vs Testing Dataset

- The **training dataset** is the sample of data used to train the model. It is the largest sample of data used when creating a machine learning model.
- The **validation dataset** is a second sample of data used to provide an evaluation of the model to see if the model can correctly predict, or classify, using data not seen before. The validation dataset is used to tune the model. . It helps to get an unbiased evaluation of the model while tuning its hyperparameters.
- A **testing dataset** is a set of data used to provide a final unbiased evaluation of the model. A test dataset is an independent sample of data and is used once a model has been completely trained with the training and validation datasets.



## Question

Which two datasets do you use to build a machine learning model? Each correct answer presents part of the solution. Choose the correct answers

1. Training dataset
2. Azure Open Dataset
3. Validation dataset
4. Testing dataset



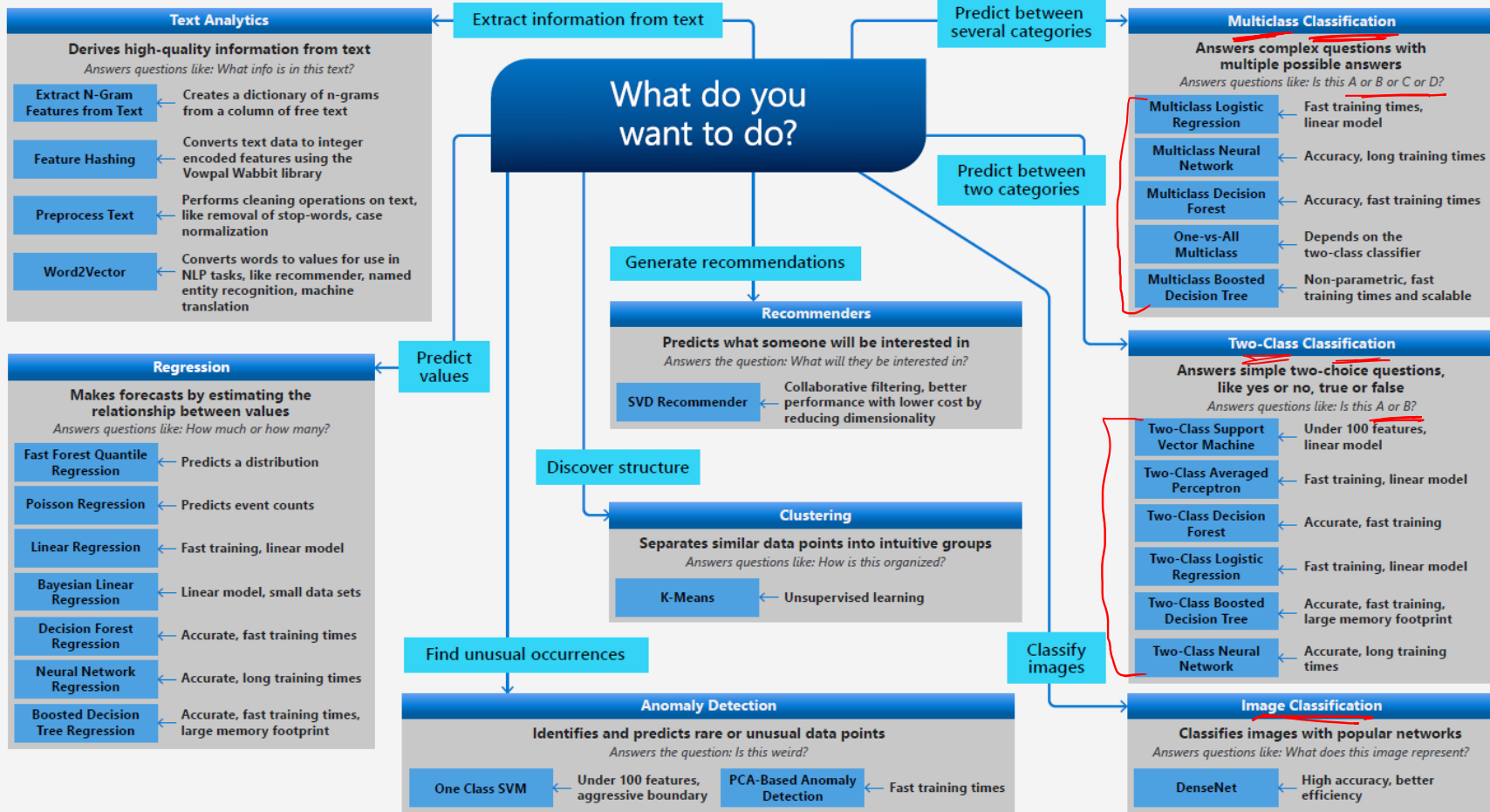
How to select Algorithm

---



# Microsoft Azure Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.



# How to select machine learning algorithms

What do you want to do with your data?

Algorithm Cheat Sheet

Additional requirements

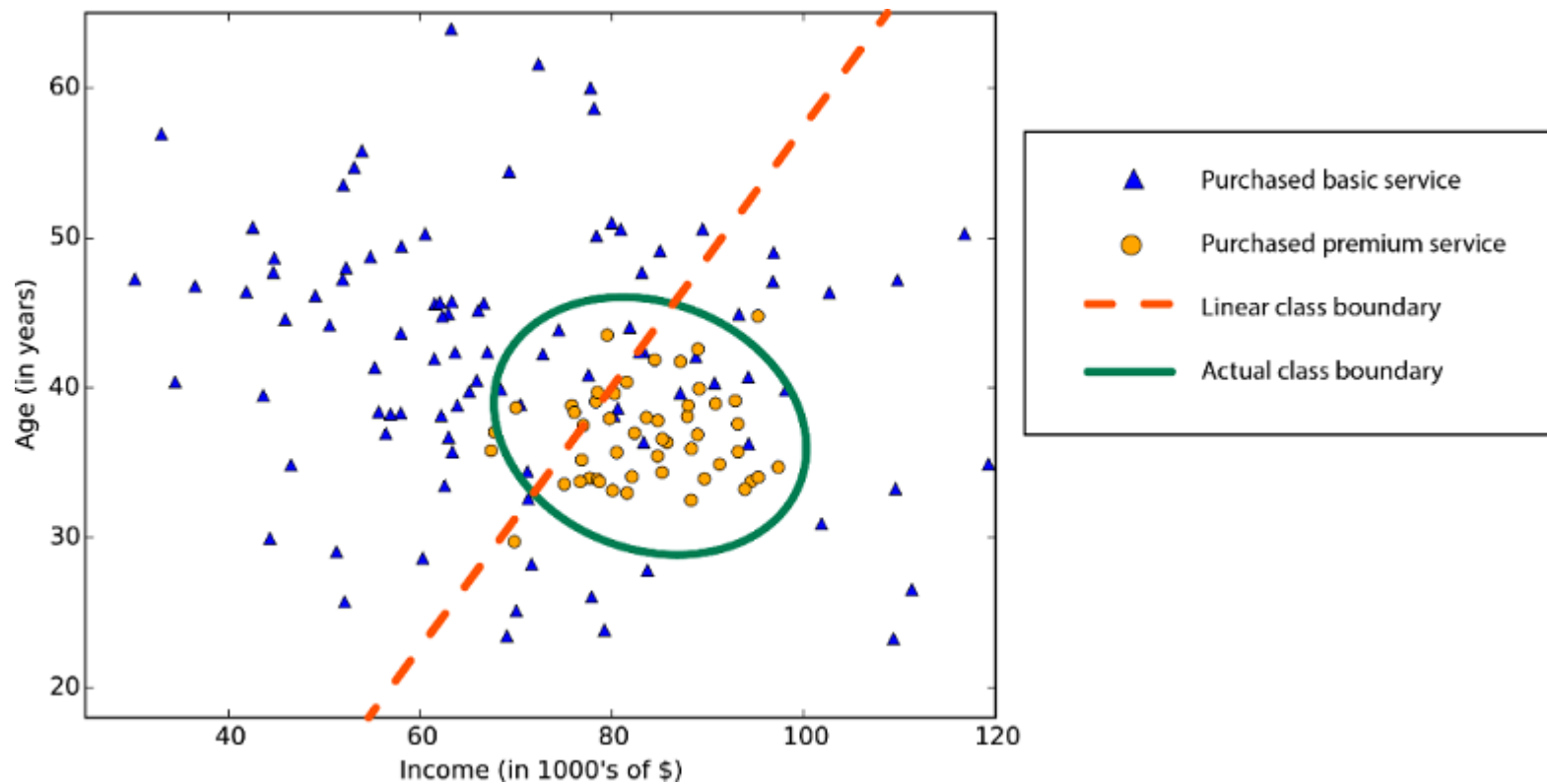
Accuracy

Training time

Linearity

Number of parameters

Number of features



Demo: Creating a ML workspace



## Workspace: Associated resources

- **Azure Storage account:** Is used as the default datastore for the workspace. Jupyter notebooks that are used with your Azure Machine Learning compute instances are stored here as well.
- **Azure Container Registry:** Registers docker containers that you use during training and when you deploy a model.
- **Azure Application Insights:** Stores monitoring information about your models.
- **Azure Key Vault:** Stores secrets that are used by compute targets and other sensitive information that's needed by the workspace.

Demo: Designing a pipeline  
Regression Model

The text is written in a white, handwritten style on a solid blue background. The words "Designing a pipeline" are enclosed in a large, hand-drawn white oval. The words "Regression Model" are enclosed in a smaller, hand-drawn white oval. A dashed white arrow points from the top right towards the "Designing a pipeline" oval. Another dashed white arrow points from the left towards the "Regression Model" oval.



# Machine Learning Studio

- **Experiments** are training runs you use to build your models.
- **Pipelines** are reusable workflows for training and retraining your model.
- **Datasets** aid in management of the data you use for model training and pipeline creation.
- Once you have a model you want to deploy, you create a registered **model**.
- Use the registered model and a scoring script to create a **deployment endpoint**.
- **Compute targets** are used to run your experiments.
  - ✓ **Compute Instances:** A compute instance is used as a compute target for authoring and training models for development and testing purposes.
  - ✓ **Compute Clusters:** Scalable clusters of virtual machines for on-demand processing of experiment code, for running batch inference on large amounts of data.
  - ✓ **Inference Clusters:** Deployment targets for predictive services that use your trained models.
  - ✓ **Attached Compute:** Links to existing Azure compute resources, such as Virtual Machines or Azure Databricks clusters.

# Metric: Regression

- **Mean Absolute Error (MAE):** The average difference between predicted values and true values. This value is based on the same units as the label, in this case dollars. The lower this value is, the better the model is predicting.
- **Root Mean Squared Error (RMSE):** The square root of the mean squared difference between predicted and true values. The result is a metric based on the same unit as the label (dollars). When compared to the MAE (above), a larger difference indicates greater variance in the individual errors (for example, with some errors being very small, while others are large).
- **Relative Squared Error (RSE):** A relative metric between 0 and 1 based on the square of the differences between predicted and true values. The closer to 0 this metric is, the better the model is performing. Because this metric is relative, it can be used to compare models where the labels are in different units.
- **Relative Absolute Error (RAE):** A relative metric between 0 and 1 based on the absolute differences between predicted and true values. The closer to 0 this metric is, the better the model is performing. Like RSE, this metric can be used to compare models where the labels are in different units.
- **Coefficient of Determination (R<sup>2</sup>):** Coefficient of determination is a measure of the variance from the mean in its predictions. Its value varies between 0 and 1, where 1 typically indicates a perfectly fit model, while 0 indicates a random one. This metric is more commonly referred to as R-Squared.

Demo: Deploying a ML model  
Regression Model

The text is written in a white, handwritten style on a solid blue background. The word "Deploying" is circled with a white line. The words "Regression Model" are underlined with a white line and also circled with a white line. A dashed white arrow points from the top right towards the word "Deploying". Another dashed white arrow points from the left towards the underlined "Regression Model".

# Questions: ML Studio

**Question:** A dataset contains attributes that have values in different units with different ranges of values. Which data preprocessing method is used to transform the values into a common scale?

- Normalization
- Binning
- Substitution
- Sampling

**Question:** Split data

- You can divide a dataset using regular expression.
- You can split a dataset for training/testing by rows.
- You can split a dataset for training/testing by columns.

**Question:** You need to hold back a dataset from model training so it can be used to estimate a model's prediction error while tuning its hyperparameters. Which dataset should you use?

- Training dataset
- Testing dataset
- Raw data
- Validation dataset

# Questions: ML Studio

**Question:** What should you do to measure the accuracy of a trained machine learning model?

- Score the model.
- Summarize the data.
- Normalize the data.
- Create features.

**Question:** What should you do to measure the accuracy of the predictions and assess model fit?

- Evaluate the model.
- Detect languages.
- Score the model.
- Evaluate the probability function.



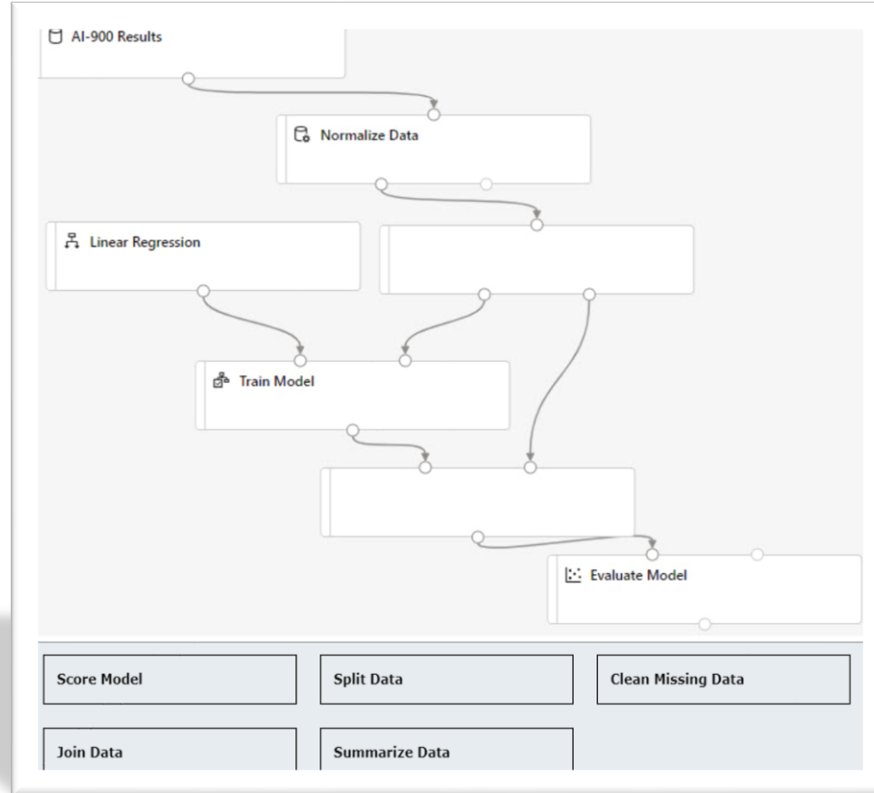
The image shows a blue background with a vertical crease down the center. The text "Delete Resources" is written in white, cursive script. The word "Resources" is circled with a white oval. A dashed white arrow points from the top left towards the word "Delete". Another dashed white arrow points from the top right towards the word "Resources". Below the text, there are several horizontal white lines of varying lengths, suggesting a list or a table.

Delete Resources



Demo: Classification Model

**Question:** Which modules should you use to complete the pipeline? To answer, drag the appropriate module to the relevant slots. A module may be used once, more than once, or not at all.





# Metric: classification

- **Accuracy:** The ratio of correct predictions (true positives + true negatives) to the total number of predictions. In other words, what proportion of diabetes predictions did the model get right?
- **Precision:** Precision is a measure of the correct positive results. Precision is the number of true positives divided by the sum of the number of true positives and false positives. Precision is scored between 0 and 1. Values closer to 1 are better.
- **Recall:** The fraction of the cases classified as positive that are actually positive (the number of true positives divided by the number of true positives plus false negatives). In other words, out of all the patients who actually have diabetes, how many did the model identify?
- **F1 Score:** F1 score is a measure combining precision and recall. F1 score is the weighted average of precision and recall (the number of true positives divided by the sum of true positives and false negatives). F-score is scored between 0 and 1. Values closer to 1 are better.
- **AUC:** measures the area under a curve that represents true positive rate over true negative rate. AUC ranges between 0 and 1. Values closer to 1 indicate that the model is performing better. AUC value of 0.4 means that the model is performing worse than a random guess. AUC values range between 0 and 1. The higher the value, the better the performance of the classification model.

## Metric: classification

**Question:** Which two metrics can you use to evaluate classification machine learning models? Each correct answer presents a complete solution. Choose the correct answers

- A. Mean Absolute Error (MAE)
- B. Precision
- C. Recall
- D. Average Distance to Cluster Center

**Question:** In the evaluation of the classification model you get a value of 0.3 for the area under the curve (AUC) metrics. What does it mean?

- A. 40 percent of data is allocated to training and 60 percent to testing.
- B. 60 percent of data is allocated to training and 40 percent to testing.
- C. The model is performing worse than a random guess.
- D. The model is performing better than a random guess.

## Metric: classification

**Question:** Which two metrics can you use to evaluate a classification machine learning model? Each correct answer presents a complete solution.

- A. Coefficient of determination
- B. Maximal Distance to Cluster Center
- C. F-score
- D. Root mean squared error (RMSE)
- E. Precision

# Metric: classification

**Question:** You evaluate a machine learning model and it generates the matrix shown in the exhibit?

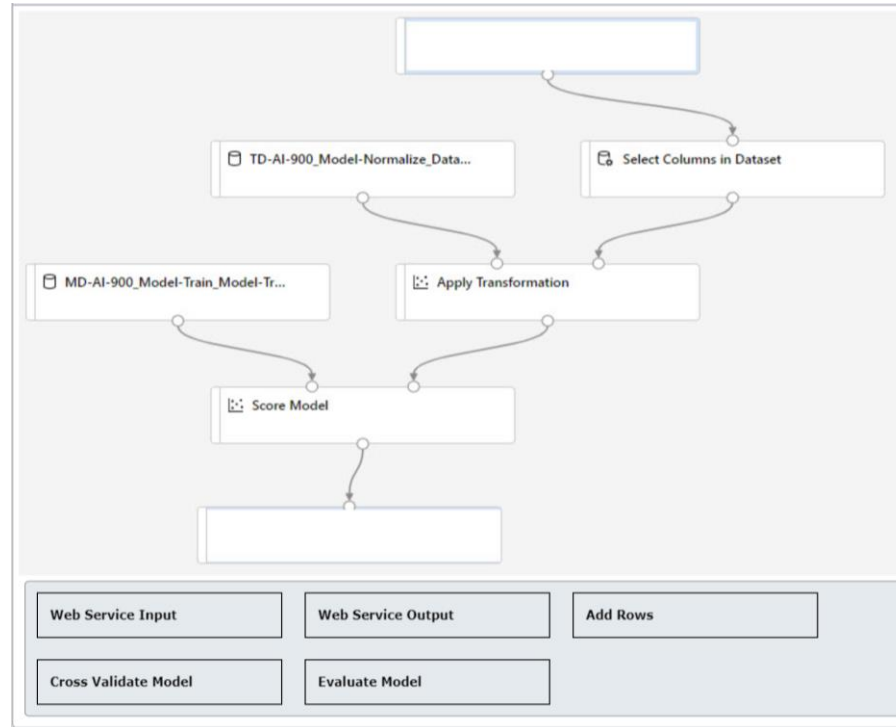
For each of the following statements, select Yes if the statement is true. Otherwise, select No

- A. The machine learning model is a classification model
- B. There are 879 false positives
- C. There are 2015 true negatives

		Actual	
		1	0
Predicted	1	95	11
	0	879	2015

**Question:** You create a real-time inference pipeline from a training pipeline in Azure Machine Learning designer. You need to complete the inference pipeline.

Which modules should you use to complete the pipeline? To answer, drag the appropriate module to the relevant slots. A module may be used once, more than once, or not at all.





Automated Machine Learning

The image shows the phrase "Automated Machine Learning" written in a white, handwritten-style font on a blue background. The word "Automated" is circled with a white line, and the word "Learning" is also circled with a white line. Two dashed white arrows point towards these circles: one from the upper left pointing to "Automated", and another from the upper center pointing to "Learning". The entire phrase is underlined with a double white line.

