1. ****what is Automated machine learning?****

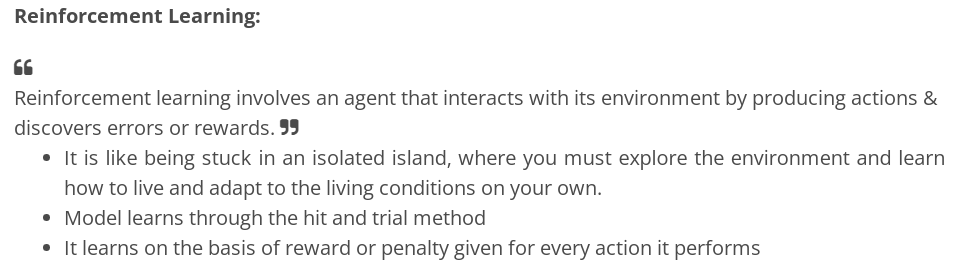
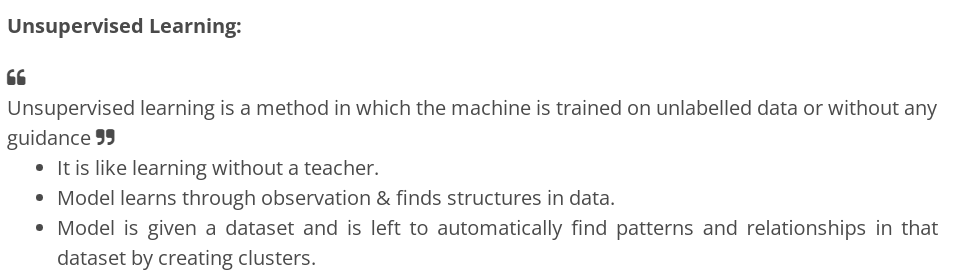
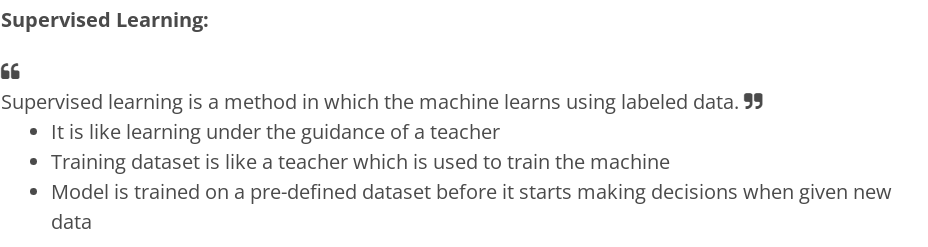


****Answer:****Automated machine learning (AutoML) is the process of automating the tasks of applying machine learning to real-world problems. AutoML potentially includes every stage from beginning with a raw dataset to building a machine learning model ready for deployment. AutoML was proposed as an artificial intelligence-based solution to the growing challenge of applying machine learning.[1][2] The high degree of automation in AutoML aims to allow non-experts to make use of machine learning models and techniques without requiring them to become experts in machine learning. Automating the process of applying machine learning end-to-end additionally offers the advantages of producing simpler solutions, faster creation of those solutions, and models that often outperform hand-designed models.

2.****What are the different types of Machine Learning?****

****Answer:****There are 3 different types of machine learning:

1. ****Supervised Learning****
2. ****Unsupervised Learning****
3. ****Reinforcement Learning****

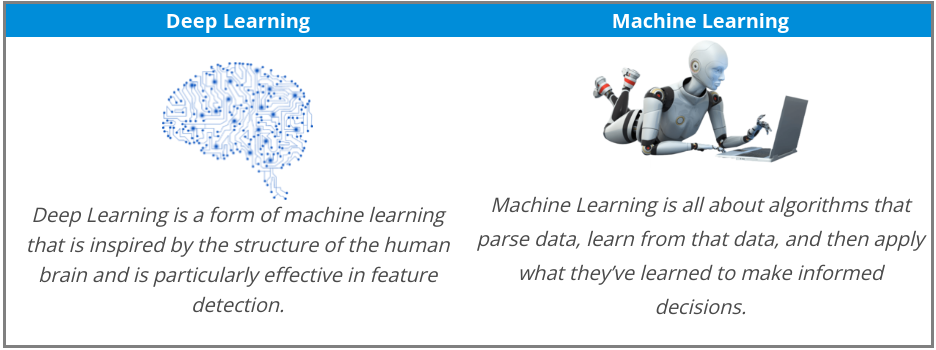
3. ****How would you teach Machine Learning to a kid?****

****Answer:****

* Suppose your friend invites you to his party where you meet total strangers. Since you have no idea about them, you will mentally classify them on the basis of gender, age group, dressing, etc.
* In this scenario, the strangers represent unlabeled data and the process of classifying unlabeled data points is nothing but unsupervised learning.
* Since you didn’t use any prior knowledge about people and classified them on-the-go, this becomes an unsupervised learning problem.

4. ****In what manners does Deep Learning differ from Machine Learning?****

****Answer:****



5. ****What do you understand by Precision and Recall?****

****Answer:****

* Imagine that, your girlfriend gave you a birthday surprise every year for the last 10 years. One day, your girlfriend asks you: ‘Sweetie, do you remember all the birthday surprises from me?’
* To stay on good terms with your girlfriend, you need to recall all the 10 events from your memory.Therefore, recall is the ratio of the number of events you can correctly recall, to the total number of events.
* If you can recall all 10 events correctly, then, your recall ratio is 1.0 (100%) and if you can recall 7 events correctly, your recall ratio is 0.7 (70%)  
  ****However, you might be wrong in some answers.****
* For example, let’s assume that you took 15 guesses out of which 10 were correct and 5 were wrong. This means that you can recall all events but not so precisely
* Therefore, precision is the ratio of the number of events you can correctly recall, to the total number of events you can recall (mix of correct and wrong recalls).
* From the above example (10 real events, 15 answers: 10 correct, 5 wrong), you get 100% recall but your precision is only 66.67% (10 / 15)

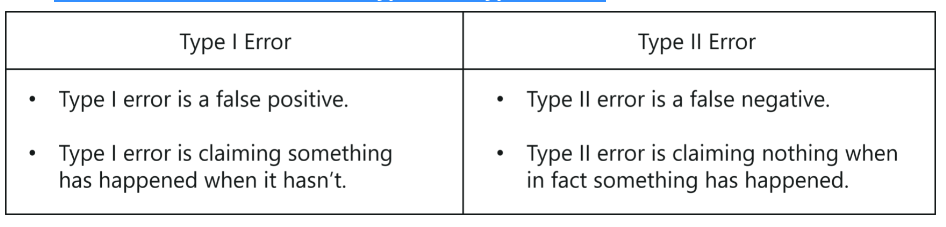
6. What do you understand by selection bias?

****Answer:****

* It is a statistical error that causes a bias in the sampling portion of an experiment.
* The error causes one sampling group to be selected more often than other groups included in the experiment.
* Selection bias may produce an inaccurate conclusion if the selection bias is not identified.

7. ****What’s the difference between Type I and Type II errors****?

****Answers****:

8. ****What is the difference between Gini Impurity and Entropy in a Decision Tree?****

****Answer:****

* Gini Impurity and Entropy are the metrics used for deciding how to split a Decision Tree.
* Gini measurement is the probability of a random sample being classified correctly if you randomly pick a label according to the distribution in the branch.
* Entropy is a measurement to calculate the lack of information. You calculate the Information Gain (difference in entropies) by making a split. This measure helps to reduce the uncertainty about the output label.

9. ****What is the difference between Entropy and Information Gain?****

****Answer:****

* Entropy is an indicator of how messy your data is. It decreases as you reach closer to the leaf node.
* The Information Gain is based on the decrease in entropy after a dataset is split on an attribute. It keeps on increasing as you reach closer to the leaf node.

10. ****What are collinearity and multicollinearity?****

****Answer:****

* Collinearity occurs when two predictor variables (e.g., x1 and x2) in a multiple regression have some correlation.
* Multicollinearity occurs when more than two predictor variables (e.g., x1, x2, and x3) are inter-correlated.

11. ****What is Cluster Sampling?****

****Answer:****

* It is a process of randomly selecting intact groups within a defined population, sharing similar characteristics.
* A cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.
* For example, if you’re clustering the total number of managers in a set of companies, in that case, managers (samples) will represent elements and companies will represent clusters.

12.****What is the difference between Data Mining and Machine Learning?****

****Answer:****

****Data mining**** can be described as the process in which the structured data tries to abstract knowledge or interesting unknown patterns. During this process, machine learning algorithms are used.

****Machine learning**** represents the study, design, and development of the algorithms which provide the ability to the processors to learn without being explicitly programmed.

13. ****What is the meaning of Overfitting in Machine learning?****

****Answer:**** Overfitting can be seen in machine learning when a statistical model describes random error or noise instead of the underlying relationship. Overfitting is usually observed when a model is excessively complex. It happens because of having too many parameters concerning the number of training data types. The model displays poor performance, which has been overfitted.

14. ****Why does overfitting occur?****

****Answer:**** The possibility of overfitting occurs when the criteria used for training the model is not as per the criteria used to judge the efficiency of a model.

15. ****What is the method to avoid overfitting?****

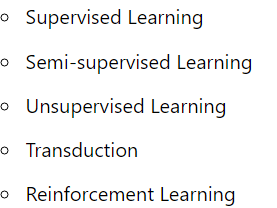
****Answer:**** Overfitting occurs when we have a small dataset, and a model is trying to learn from it. By using a large amount of data, overfitting can be avoided. But if we have a small database and are forced to build a model based on that, then we can use a technique known as cross-validation. In this method, a model is usually given a dataset of known data on which the training data set is run and a dataset of unknown data against which the model is tested. The primary aim of cross-validation is to define a dataset to “test” the model in the training phase. If there is sufficient data, ‘Isotonic Regression’ is used to prevent overfitting.

16. ****How is KNN different from k-means?****

****Answer:****KNN or K nearest neighbors is a supervised algorithm that is used for classification purposes. In KNN, a test sample is given as the class of the majority of its nearest neighbors. On the other side, K-means is an unsupervised algorithm that is mainly used for clustering. In k-means clustering, it needs a set of unlabeled points and a threshold only. The algorithm further takes unlabeled data and learns how to cluster it into groups by computing the mean of the distance between different unlabeled points.

17. ****What are the different types of Algorithm methods in Machine Learning?****

****Answer:****



18. ****What do you understand by the Reinforcement Learning technique?****

****Answer:**** Reinforcement learning is an algorithm technique used in Machine Learning. It involves an agent that interacts with its environment by producing actions & discovering errors or rewards. Reinforcement learning is employed by different software and machines to search for the best suitable behavior or path it should follow in a specific situation. It usually learns on the basis of reward or penalty given for every action it performs.

19. ****What is the trade-off between bias and variance?****

****Answer****: Both ****bias****and ****variance**** are errors. Bias is an error due to erroneous or overly simplistic assumptions in the learning algorithm. It can lead to the model under-fitting the data, making it hard to have high predictive accuracy and generalize the knowledge from the training set to the test set.

****Variance****is an error due to too much complexity in the learning algorithm. It leads to the algorithm being highly sensitive to high degrees of variation in the training data, which can lead the model to overfit the data.

To optimally reduce the number of errors, we will need to tradeoff bias and variance.

20. ****What do you mean by ensemble learning?****

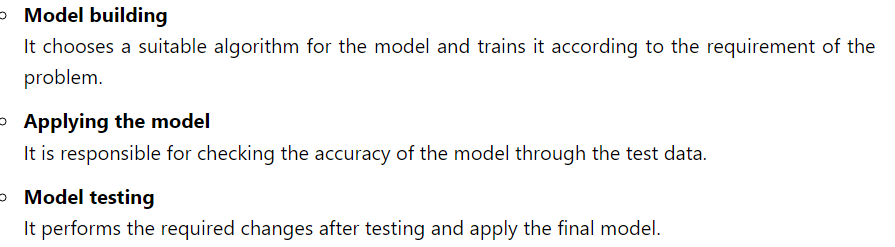
****Answer****: Numerous models, such as classifiers are strategically made and combined to solve a specific computational program which is known as ensemble learning. The ensemble methods are also known as committee-based learning or learning multiple classifier systems. It trains various hypotheses to fix the same issue. One of the most suitable examples of ensemble modeling is the random forest trees where several decision trees are used to predict outcomes. It is used to improve the classification, function approximation, prediction, etc. of a model.

21. ****What is a model selection in Machine Learning?****

****Answer****: The process of choosing models among diverse mathematical models, which are used to define the same data is known as Model Selection. ****Model learning**** is applied to the fields of ****statistics****, ****data mining****, and ****machine learning.****

22. ****What are the three stages of building the hypotheses or model in machine learning?****

****Answer:****



23. ****Describe ‘Training set’ and ‘training Test’.****

****Answer****: In various areas of information of machine learning, a set of data is used to discover the potentially predictive relationship, which is known as ****‘Training Set’****. The training set is an example that is given to the learner. Besides, the ****‘Test set’****is used to test the accuracy of the hypotheses generated by the learner. It is the set of instances held back from the learner. Thus, the training set is distinct from the test set.

24. ****What do you understand by ILP?****

****Answer:**** ****ILP**** stands for ****Inductive Logic Programming.**** It is a part of machine learning which uses logic programming. It aims at searching patterns in data that can be used to build predictive models. In this process, the logic programs are assumed as a hypothesis.

25. ****Describe Precision and Recall?****

****Answer:****

****Precision**** and****Recall b****oth are the measures that are used in the information retrieval domain to measure how well an information retrieval system reclaims the related data as requested by the user.

****Precision****can be said as a positive predictive value. It is the fraction of relevant instances among the received instances.

On the other side****, recall****is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. The recall is also known as sensitivity.

26. ****What are the functions of Supervised Learning?****

****Answer:****

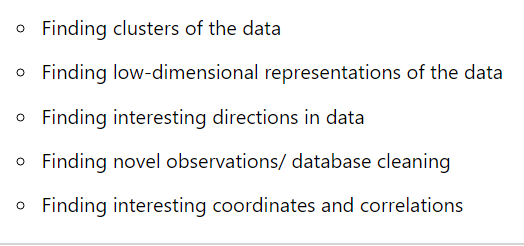
* Classification
* Speech Recognition
* Regression
* Predict Time Series
* Annotate Strings

27. ****What do you understand by algorithm-independent machine learning?****

Answer: Algorithm independent machine learning can be defined as machine learning, where mathematical foundations are independent of any particular classifier or learning algorithm.

28. ****What are the functions of Unsupervised Learning?****

****Answer:****

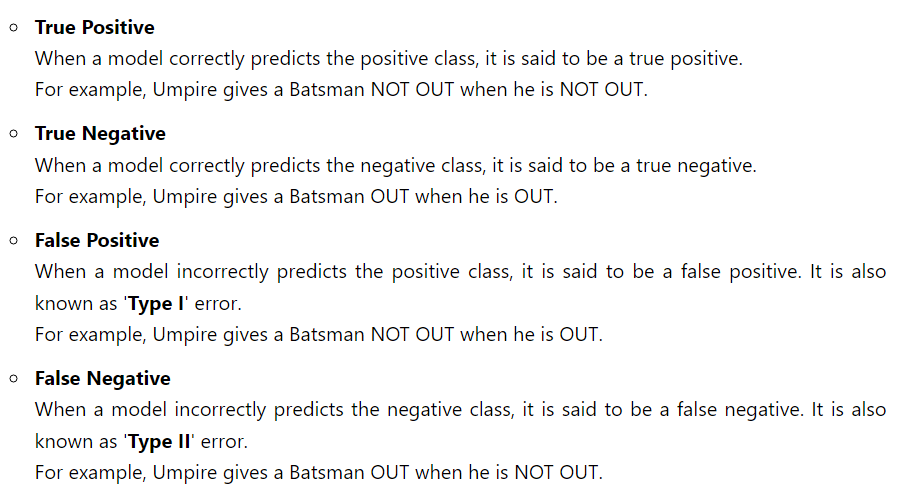


29. ****What do you mean by Genetic Programming?****

****Answer: Genetic Programming (GP)**** is almost similar to an Evolutionary Algorithm, a subset of machine learning.****Genetic programming**** software systems implement an algorithm that uses random mutation, a fitness function, crossover, and multiple generations of evolution to resolve a user-defined task. The genetic programming model is based on testing and choosing the best option among a set of results.

30. ****Explain True Positive, True Negative, False Positive, and False Negative in Confusion Matrix with an example.****

****Answer:****



30. ****What is Bagging and Boosting?****

****Answer:****

* Bagging is a process in ensemble learning which is used for improving unstable estimation or classification schemes.
* Boosting methods are used sequentially to reduce the bias of the combined model.

31. ****Which are the two components of the Bayesian logic program?****

****Answer:**** A Bayesian logic program consists of two components:

****Logical****  
\*It contains a set of Bayesian Clauses, which capture the qualitative structure of the domain.  
****Quantitative****  
\*It is used to encode quantitative information about the domain.

32. ****Describe dimension reduction in machine learning.****

****Answer:****

* Dimension reduction is the process that is used to reduce the number of random variables under consideration.
* Dimension reduction can be divided into feature selection and extraction.

33. ****Why instance-based learning algorithm sometimes referred to as a Lazy learning algorithm?****

****Answer:****In machine learning, lazy learning can be described as a method where induction and generalization processes are delayed until classification is performed. Because of the same property, an instance-based learning algorithm is sometimes called a lazy learning algorithm.

34. ****What are the Recommended Systems?****

****Answer:****Recommended System is a sub-directory of information filtering systems. It predicts the preferences or rankings offered by a user to a product. According to the preferences, it provides similar recommendations to a user. Recommendation systems are widely used in movies, news, research articles, ****products,**** ****social tips,**** ****music,****etc.

35. ****What do you understand by Underfitting?****

****Answer:**** Underfitting is an issue when we have a low error in both the training set and the testing set. Few algorithms work better for interpretations but fail for better predictions.

36. ****What is Regularization? What kind of problems does regularization solve?****

****Answer:****

****Regularization**** is a form of regression, which constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, it discourages learning a more complex or flexible model to avoid the risk of overfitting. It reduces the variance of the model, without a substantial increase in its bias.

****Regularization**** is used to address overfitting problems as it penalizes the loss function by adding a multiple of an L1 (LASSO) or an L2 (Ridge) norm of weights vector w.

37. ****Do you think that treating a categorical variable as a continuous variable would result in a better predictive model?****

****Answer:****For a better predictive model, the categorical variable can be considered as a continuous variable only when the variable is ordinal in nature.

38.****What do you understand by the F1 score?****

****Answer:****The F1 score represents the measurement of a model’s performance. It is referred to as a weighted average of the precision and recall of a model. The results tending to 1 are considered as the best, and those tending to 0 are the worst. It could be used in classification tests, where true negatives don’t matter much.

39. ****What is the meaning of Variance Error in ML algorithms?****

****Answer:****Variance error is found in machine learning algorithms that are highly complex and pose difficulties in understanding them. As a result, you can find a greater extent of variation in the training data. Subsequently, the machine learning model would overfit the data. In addition, you can also find excessive noise for training data which is completely inappropriate for the test data.

40. ****What is the ROC curve, and how does it work?****

****Answer:**** The receiver Operating Characteristic (ROC) curve provides a pictorial representation of the contrast level between false-positive rates and true positive rates. The estimates of true and false positive rates are taken at multiple thresholds. The ROC is ideal as a proxy for measuring trade-offs and sensitivity associated with a model. According to the measurements of sensitivity and trade-off, the curve can trigger false alarms.

41. ****What is precision, and what is a recall****?

****Answer:**** The recall is the number of true positive rates identified for a specific total number of datasets. Precision involves predictions for positive values claimed by a model as compared to the number of actually claimed positives. You can assume this as a special case for probability with respect to mathematics.

42. ****What are Naive Bayes?****

****Answer:****Naive Bayes is ideal for practical application in text mining. However, it also involves an assumption that it is not possible to visualize in real-time data. Naive Bayes involves the calculation of conditional probability from the pure product of individual probabilities of different components. The condition in such cases would imply complete independence for the features that are practically not possible or very difficult. Candidates should expect this type of follow-up machine learning interview questions.

43. ****How is the generative model different from the discriminative model?****

****Answer:**** The generative model will review the data categories. However, a discriminative model would review the difference between various data categories. Generally, discriminative models have better performance than generative models in classification tasks.

44. ****In which situation is classification better than regression?****

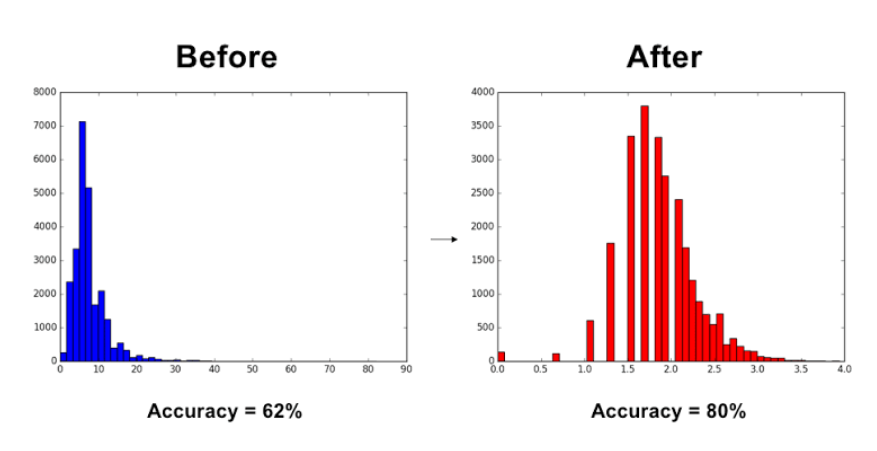
****Answer:**** Classification results in the generation of discrete values and dataset according to specific categories. On the other hand, regression provides continuous results with better demarcations between individual points. Classification is better than regression when you need the results to reflect the presence of data points in explicit categories in the dataset. Classification is better if you just want to find whether a name is female or male. Regression is ideal if you want to find out the correlation of the name with male and female names.

45. ****Name the five most popular machine learning algorithms.****

Answer: The most popular machine learning algorithms are decision ****trees****, ****probabilistic**** networks, and ****neural networks****. The other two popular ML algorithms are support vector machines and ****neural networks**** or ****backpropagation networks.****

46. ****What is the use of Box-Cox transformation?****

****Answer:**** Box-Cox transformation is a type of powerful transformation that helps in the transformation of data for normalizing distribution. Box-Cox transformation is also ideal for stabilizing variance by eliminating heteroskedasticity.



47. ****What is the difference between inductive and deductive learning?****

****Answer:****Inductive learning involves using observations for reaching conclusions. Deductive learning involves referring to conclusions for developing observations.

48. ****What are the different components of relational evaluation techniques?****

****Answer:****The significant components of relational evaluation techniques include data acquisition, query type, significance test, and scoring metric. The other important components include a cross-validation technique and ground truth acquisition.

49. ****What is a recommendation system?****

****Answer:****Answer: A recommendation system is a subclass in the information filtering system for predicting preference that a user would assign to an item. The best techniques for recommendation systems are collaborative filtering and content-based filtering.

50. ****Can you explain about bagging?****

****Answer:****Bagging is the short-form for bootstrap aggregating. Bagging is actually a meta-algorithm that takes M subsamples from the initial dataset as inputs. Subsequently, the algorithm trains a predictive model on the subsamples. The final model is a product of averaging bootstrapped models and provides better results.

# **What is automated machine learning (AutoML)?**

* Article
* 11/01/2022
* 10 minutes to read
* 25 contributors

Feedback

****APPLIES TO****: IMG_256 [Python SDK azure-ai-ml](https://aka.ms/sdk-v2-install)**[v2 (current)](https://aka.ms/sdk-v2-install)**

Select the version of the Azure Machine Learning Python SDK you are using:

v2 (current version)

Automated machine learning, also referred to as automated ML or AutoML, is the process of automating the time-consuming, iterative tasks of machine learning model development. It allows data scientists, analysts, and developers to build ML models with high scale, efficiency, and productivity all while sustaining model quality. Automated ML in Azure Machine Learning is based on a breakthrough from our [Microsoft Research division](https://www.microsoft.com/research/project/automl/).

* For code-experienced customers, [Azure Machine Learning Python SDK](https://aka.ms/sdk-v2-install). Get started with [Tutorial: Train an object detection model (preview) with AutoML and Python](https://learn.microsoft.com/en-us/azure/machine-learning/tutorial-auto-train-image-models).

## **How does AutoML work?**

During training, Azure Machine Learning creates a number of pipelines in parallel that try different algorithms and parameters for you. The service iterates through ML algorithms paired with feature selections, where each iteration produces a model with a training score. The better the score for the metric you want to optimize for, the better the model is considered to "fit" your data. It will stop once it hits the exit criteria defined in the experiment.

Using ****Azure Machine Learning****, you can design and run your automated ML training experiments with these steps:

****Identify the ML problem**** to be solved: classification, forecasting, regression, computer vision or NLP.

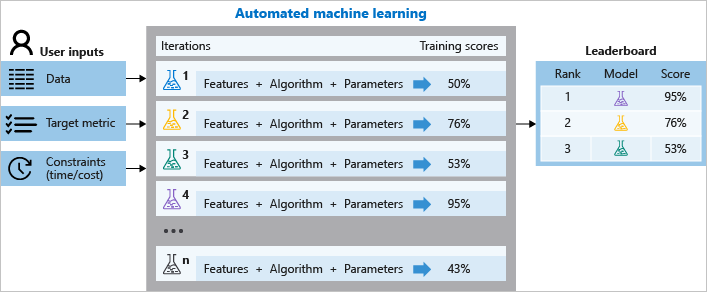
****Choose whether you want to a code-first experience or a no-code studio web experience****: Users who prefer a code-first experience can use the [AzureML SDKv2](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train) or the [AzureML CLIv2](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-train-cli). Get started with [Tutorial: Train an object detection model with AutoML and Python](https://learn.microsoft.com/en-us/azure/machine-learning/tutorial-auto-train-image-models). Users who prefer a limited/no-code experience can use the [web interface](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-automated-ml-for-ml-models) in Azure Machine Learning studio at [https://ml.azure.com](https://ml.azure.com/). Get started with [Tutorial: Create a classification model with automated ML in Azure Machine Learning](https://learn.microsoft.com/en-us/azure/machine-learning/tutorial-first-experiment-automated-ml).

****Specify the source of the labeled training data****: You can bring your data to AzureML in [many different ways](https://learn.microsoft.com/en-us/azure/machine-learning/concept-data).

****Configure the automated machine learning parameters**** that determine how many iterations over different models, hyperparameter settings, advanced preprocessing/featurization, and what metrics to look at when determining the best model.

****Submit the training job.****

****Review the results****

The following diagram illustrates this process. 

You can also inspect the logged job information, which [contains metrics](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-understand-automated-ml) gathered during the job. The training job produces a Python serialized object (.pkl file) that contains the model and data preprocessing.

While model building is automated, you can also [learn how important or relevant features are](https://learn.microsoft.com/en-us/azure/machine-learning/v1/how-to-configure-auto-train-v1" \l "explain) to the generated models.

## **When to use AutoML: classification, regression, forecasting, computer vision & NLP**

Apply automated ML when you want Azure Machine Learning to train and tune a model for you using the target metric you specify. Automated ML democratizes the machine learning model development process, and empowers its users, no matter their data science expertise, to identify an end-to-end machine learning pipeline for any problem.

ML professionals and developers across industries can use automated ML to:

* Implement ML solutions without extensive programming knowledge
* Save time and resources
* Leverage data science best practices
* Provide agile problem-solving

### **Classification**

Classification is a type of supervised learning in which models learn using training data, and apply those learnings to new data. Azure Machine Learning offers featurizations specifically for these tasks, such as deep neural network text featurizers for classification. Learn more about [featurization options](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "data-featurization). You can also find the list of algorithms supported by AutoML [here](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "supported-algorithms).

The main goal of classification models is to predict which categories new data will fall into based on learnings from its training data. Common classification examples include fraud detection, handwriting recognition, and object detection.

See an example of classification and automated machine learning in this Python notebook: [Bank Marketing](https://github.com/Azure/azureml-examples/blob/main/sdk/python/jobs/automl-standalone-jobs/automl-classification-task-bankmarketing/automl-classification-task-bankmarketing.ipynb).

### **Regression**

Similar to classification, regression tasks are also a common supervised learning task. AzureML offers featurization specific to regression problems. Learn more about [featurization options](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "data-featurization). You can also find the list of algorithms supported by AutoML [here](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "supported-algorithms).

Different from classification where predicted output values are categorical, regression models predict numerical output values based on independent predictors. In regression, the objective is to help establish the relationship among those independent predictor variables by estimating how one variable impacts the others. For example, automobile price based on features like, gas mileage, safety rating, etc.

See an example of regression and automated machine learning for predictions in these Python notebooks: [Hardware Performance](https://github.com/Azure/azureml-examples/blob/main/sdk/python/jobs/automl-standalone-jobs/automl-regression-task-hardware-performance/automl-regression-task-hardware-performance.ipynb).

### **Time-series forecasting**

Building forecasts is an integral part of any business, whether it's revenue, inventory, sales, or customer demand. You can use automated ML to combine techniques and approaches and get a recommended, high-quality time-series forecast. You can find the list of algorithms supported by AutoML [here](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "supported-algorithms).

An automated time-series experiment is treated as a multivariate regression problem. Past time-series values are "pivoted" to become additional dimensions for the regressor together with other predictors. This approach, unlike classical time series methods, has an advantage of naturally incorporating multiple contextual variables and their relationship to one another during training. Automated ML learns a single, but often internally branched model for all items in the dataset and prediction horizons. More data is thus available to estimate model parameters and generalization to unseen series becomes possible.

Advanced forecasting configuration includes:

* holiday detection and featurization
* time-series and DNN learners (Auto-ARIMA, Prophet, ForecastTCN)
* many models support through grouping
* rolling-origin cross validation
* configurable lags
* rolling window aggregate features

See an example of forecasting and automated machine learning in this Python notebook: [Energy Demand](https://github.com/Azure/azureml-examples/blob/main/sdk/python/jobs/automl-standalone-jobs/automl-forecasting-task-energy-demand/automl-forecasting-task-energy-demand-advanced.ipynb).

### **Computer vision**

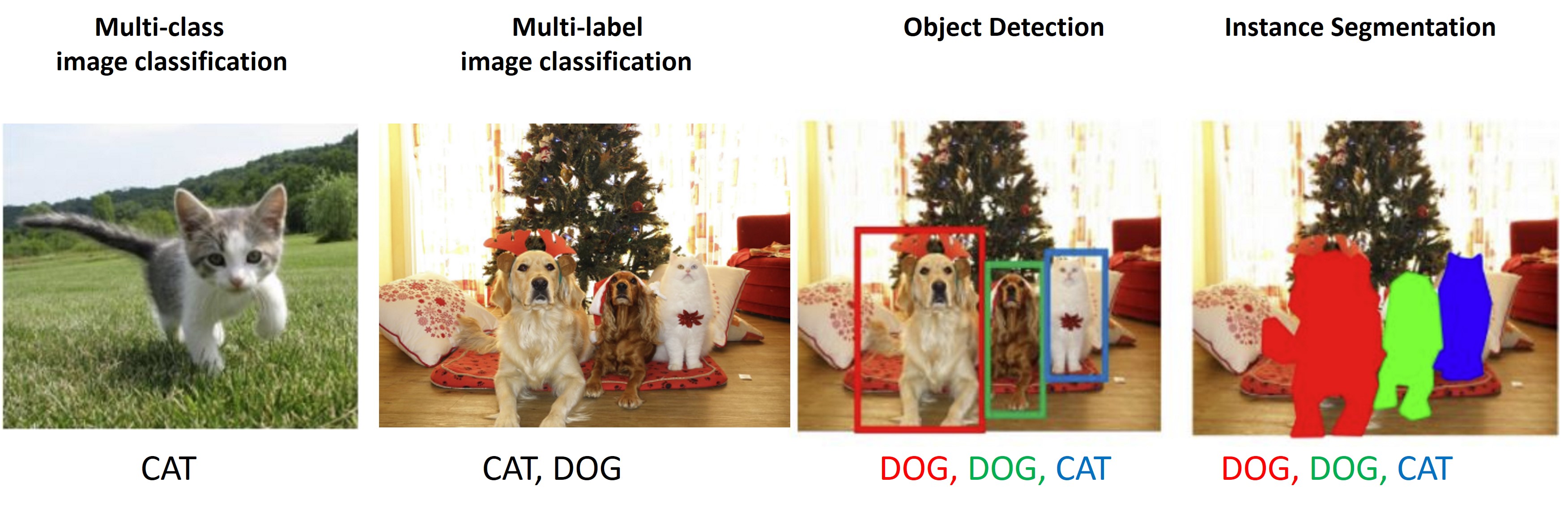
Support for computer vision tasks allows you to easily generate models trained on image data for scenarios like image classification and object detection.

With this capability you can:

* Seamlessly integrate with the [Azure Machine Learning data labeling](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-create-image-labeling-projects) capability
* Use labeled data for generating image models
* Optimize model performance by specifying the model algorithm and tuning the hyperparameters.
* Download or deploy the resulting model as a web service in Azure Machine Learning.
* Operationalize at scale, leveraging Azure Machine Learning [MLOps](https://learn.microsoft.com/en-us/azure/machine-learning/concept-model-management-and-deployment) and [ML Pipelines](https://learn.microsoft.com/en-us/azure/machine-learning/concept-ml-pipelines) capabilities.

Authoring AutoML models for vision tasks is supported via the Azure ML Python SDK. The resulting experimentation jobs, models, and outputs can be accessed from the Azure Machine Learning studio UI.

Learn how to [set up AutoML training for computer vision models](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-auto-train-image-models).

 Image from: <http://cs231n.stanford.edu/slides/2021/lecture_15.pdf>

Automated ML for images supports the following computer vision tasks:

| **Task** | **Description** |
| --- | --- |
| Multi-class image classification | Tasks where an image is classified with only a single label from a set of classes - e.g. each image is classified as either an image of a 'cat' or a 'dog' or a 'duck' |
| Multi-label image classification | Tasks where an image could have one or more labels from a set of labels - e.g. an image could be labeled with both 'cat' and 'dog' |
| Object detection | Tasks to identify objects in an image and locate each object with a bounding box e.g. locate all dogs and cats in an image and draw a bounding box around each. |
| Instance segmentation | Tasks to identify objects in an image at the pixel level, drawing a polygon around each object in the image. |

### **Natural language processing: NLP**

Support for natural language processing (NLP) tasks in automated ML allows you to easily generate models trained on text data for text classification and named entity recognition scenarios. Authoring automated ML trained NLP models is supported via the Azure Machine Learning Python SDK. The resulting experimentation jobs, models, and outputs can be accessed from the Azure Machine Learning studio UI.

The NLP capability supports:

* End-to-end deep neural network NLP training with the latest pre-trained BERT models
* Seamless integration with [Azure Machine Learning data labeling](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-create-text-labeling-projects)
* Use labeled data for generating NLP models
* Multi-lingual support with 104 languages
* Distributed training with Horovod

Learn how to [set up AutoML training for NLP models](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-auto-train-nlp-models).

## **Training, validation and test data**

With automated ML you provide the ****training data**** to train ML models, and you can specify what type of model validation to perform. Automated ML performs model validation as part of training. That is, automated ML uses ****validation data**** to tune model hyperparameters based on the applied algorithm to find the combination that best fits the training data. However, the same validation data is used for each iteration of tuning, which introduces model evaluation bias since the model continues to improve and fit to the validation data.

To help confirm that such bias isn't applied to the final recommended model, automated ML supports the use of ****test data**** to evaluate the final model that automated ML recommends at the end of your experiment. When you provide test data as part of your AutoML experiment configuration, this recommended model is tested by default at the end of your experiment (preview).

**Important**

Testing your models with a test dataset to evaluate generated models is a preview feature. This capability is an **[experimental](https://learn.microsoft.com/en-us/python/api/overview/azure/ml/" \l "stable-vs-experimental)** preview feature, and may change at any time.

Learn how to [configure AutoML experiments to use test data (preview) with the SDK](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "training-validation-and-test-data) or with the [Azure Machine Learning studio](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-automated-ml-for-ml-models" \l "create-and-run-experiment).

## **Feature engineering**

Feature engineering is the process of using domain knowledge of the data to create features that help ML algorithms learn better. In Azure Machine Learning, scaling and normalization techniques are applied to facilitate feature engineering. Collectively, these techniques and feature engineering are referred to as featurization.

For automated machine learning experiments, featurization is applied automatically, but can also be customized based on your data. [Learn more about what featurization is included](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-features" \l "featurization) and how AutoML helps [prevent over-fitting and imbalanced data](https://learn.microsoft.com/en-us/azure/machine-learning/concept-manage-ml-pitfalls) in your models.

**Note**

Automated machine learning featurization steps (feature normalization, handling missing data, converting text to numeric, etc.) become part of the underlying model. When using the model for predictions, the same featurization steps applied during training are applied to your input data automatically.

### **Customize featurization**

Additional feature engineering techniques such as, encoding and transforms are also available.

Enable this setting with:

Azure Machine Learning studio: Enable ****Automatic featurization**** in the ****View additional configuration**** section [with these steps](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-automated-ml-for-ml-models" \l "customize-featurization).

Python SDK: Specify featurization in your [AutoML Job](https://learn.microsoft.com/en-us/python/api/azure-ai-ml/azure.ai.ml.automl) object. Learn more about [enabling featurization](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "data-featurization).

## **Ensemble models**

Automated machine learning supports ensemble models, which are enabled by default. Ensemble learning improves machine learning results and predictive performance by combining multiple models as opposed to using single models. The ensemble iterations appear as the final iterations of your job. Automated machine learning uses both voting and stacking ensemble methods for combining models:

* ****Voting****: predicts based on the weighted average of predicted class probabilities (for classification tasks) or predicted regression targets (for regression tasks).
* ****Stacking****: stacking combines heterogenous models and trains a meta-model based on the output from the individual models. The current default meta-models are LogisticRegression for classification tasks and ElasticNet for regression/forecasting tasks.

The [Caruana ensemble selection algorithm](http://www.niculescu-mizil.org/papers/shotgun.icml04.revised.rev2.pdf) with sorted ensemble initialization is used to decide which models to use within the ensemble. At a high level, this algorithm initializes the ensemble with up to five models with the best individual scores, and verifies that these models are within 5% threshold of the best score to avoid a poor initial ensemble. Then for each ensemble iteration, a new model is added to the existing ensemble and the resulting score is calculated. If a new model improved the existing ensemble score, the ensemble is updated to include the new model.

See the [AutoML package](https://learn.microsoft.com/en-us/python/api/azure-ai-ml/azure.ai.ml.automl) for changing default ensemble settings in automated machine learning.

## **AutoML & ONNX**

With Azure Machine Learning, you can use automated ML to build a Python model and have it converted to the ONNX format. Once the models are in the ONNX format, they can be run on a variety of platforms and devices. Learn more about [accelerating ML models with ONNX](https://learn.microsoft.com/en-us/azure/machine-learning/concept-onnx).

See how to convert to ONNX format [in this Jupyter notebook example](https://github.com/Azure/azureml-examples/tree/main/v1/python-sdk/tutorials/automl-with-azureml/classification-bank-marketing-all-features). Learn which [algorithms are supported in ONNX](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train" \l "supported-algorithms).

The ONNX runtime also supports C#, so you can use the model built automatically in your C# apps without any need for recoding or any of the network latencies that REST endpoints introduce. Learn more about [using an AutoML ONNX model in a .NET application with ML.NET](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-automl-onnx-model-dotnet) and [inferencing ONNX models with the ONNX runtime C# API](https://onnxruntime.ai/docs/api/csharp-api.html).

## **Next steps**

There are multiple resources to get you up and running with AutoML.

### **Tutorials/ how-tos**

Tutorials are end-to-end introductory examples of AutoML scenarios.

****For a code first experience****, follow the [Tutorial: Train an object detection model with AutoML and Python](https://learn.microsoft.com/en-us/azure/machine-learning/tutorial-auto-train-image-models)

****For a low or no-code experience****, see the [Tutorial: Train a classification model with no-code AutoML in Azure Machine Learning studio](https://learn.microsoft.com/en-us/azure/machine-learning/tutorial-first-experiment-automated-ml).

How-to articles provide additional detail into what functionality automated ML offers. For example,

Configure the settings for automatic training experiments

* + [Without code in the Azure Machine Learning studio](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-automated-ml-for-ml-models).
  + [With the Python SDK](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-train).

Learn how to [train computer vision models with Python](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-auto-train-image-models).

Learn how to [view the generated code from your automated ML models](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-generate-automl-training-code).

### **Jupyter notebook samples**

Review detailed code examples and use cases in the [GitHub notebook repository for automated machine learning samples](<https://github.com/Azure/azureml-examples/tree/main/sdk/python/jobs/automl-standalone-jobs>.

### **Python SDK reference**

Deepen your expertise of SDK design patterns and class specifications with the [AutoML Job class reference documentation](https://learn.microsoft.com/en-us/python/api/azure-ai-ml/azure.ai.ml.automl).

<https://en.wikipedia.org/wiki/Automated_machine_learning>

https://www.automl.org/automl/