**What are Different types of learning in machine learning ?**

* Supervised learning
* Unsupervised learning
* Reinforcement learning
* Semi supervised learning

**What are different applications of machine learning ?**

* Bioinformatics
* Brain–machine interfaces
* Computer Networks
* Computer vision
* Credit-card fraud detection
* Financial market analysis
* Handwriting recognition
* Information retrieval
* Insurance
* Internet fraud detection
* Medical diagnosis
* Optimization
* Recommender systems
* Search engines
* Sentiment analysis
* Sequence mining
* Speech recognition
* Time series forecasting
* User behavior analytics

**What is Supervised learning ?**

When the data is labeled in during training process it’s calling supervised learning

**What is Unsupervised learning ?**

When the data is not labeled in during training process it’s calling supervised learning.

**What is Reinforcement learning ?**

Reinforcement learning is an important type of Machine Learning where an agent learn how to behave in a environment by performing actions and seeing the results.

[Readmore](https://medium.freecodecamp.org/an-introduction-to-reinforcement-learning-4339519de419)

**What is Semi supervised learning ?**

Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning and supervised learning

[Readmore](https://en.wikipedia.org/wiki/Semi-supervised_learning)

**What is Anomaly detection ?**

Anomaly detection is the identification of data points, items, observations or events that do not conform to the expected pattern of a given group. These anomalies occur very infrequently but may signify a large and significant threat such as cyber intrusions or fraud.

[Readmore](https://www.techopedia.com/definition/30297/anomaly-detection)

**What is Bias, Variance and Trade-off ?**

Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.

[Readmore](https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/)

**What are best algorithms for supervised learning ?**

* Linear Regression
* Logistic Regression
* Decision Tree
* Random Forest
* KNN
* Naive bayes

**What are best algorithms for unsupervised learning ?**

* K-means
* Hierarchical Clustering
* t-SNE Clustering
* DBSCAN Clustering
* Principal Component Analysis (PCA)
* Anomaly detection

**What are best algorithms for Reinforcement learning ?**

* Q-Learning
* SARSA  -State-Action-Reward-State-Action
* DQN – Deep Q Network
* DDPG – Deep Deterministic Policy Gradient

**What is regression ? When we will use this method ?**

A technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with, and depends on, a change in one or more independent variables.

[Read more](http://www.businessdictionary.com/definition/regression.html)

**What is classification ? When we will use this method ?**

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

[Read more](https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623)

**What is clustering ? When we will use this method ?**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups

[Read more](https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/https:/www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/)

**What is regularization ?**

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

[Readmore](https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a)

**What is Difference between l1 regularization and l2 regularization ?**

Mathematically speaking, it adds a regularization term in order to prevent the coefficients to fit so perfectly to overfit. The difference between the L1and L2 is just that L2 is the sum of the square of the weights, while L1 is just the sum of the weights.

[Read more](http://www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/)

**What ensemble learning ?**

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem.

[Read more](http://www.scholarpedia.org/article/Ensemble_learning)

**What are ensemble techniques ?**

Basic : max voting, averaging, weighted averages

Advanced : Stacking, Blending, Bagging, Boosting

**What is bagging ?**

Bagging is an abbreviation for “bootstrap aggregating”. It’a meta-algorithm, which takes M subsamples (with replacement) from the initial dataset and trains the predictive model on those subsamples. The final model is obtained by averaging the “bootstrapped” models and usually yields better results.

[Read more](https://www.quora.com/What-is-bagging-in-machine-learning)

**What is boosting ?**

Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones.

[Read more](https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/)

**What are the best bagging algorithms ?**

* Bagging Meta-estimator
* Random Forest

**What are the best boosting algorithms ?**

* Adaboost
* Gradient boosting algorithm (GBM)
* Extreme gradient boosting (XBM)
* Light GBM
* Catboost

**What is dimensionality reduction ?**

In statistics, machine learning, and information theory, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

[Read more](https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/)

**What are best dimensionality reduction algorithms ?**

* Missing Value Ratio
* Low Variance Filter
* High Correlation Filter
* Random Forest
* Backward Feature Elimination
* Forward Feature Selection
* Factor Analysis
* Principal Component Analysis
* Independent Component Analysis
* t-Distributed Stochastic Neighbor Embedding (t-SNE)
* UMAP
* [readmore](https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/)

**What is recommendation system ?**

A recommender system or a recommendation system is a subclass of information filtering system that seeks to predict the “rating” or “preference” a user would give to an item.

[Reference](https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/)

**What are best techniques for recommendation system ?**

1. Content based filtering
2. Collaborative filtering

**What is Content based filtering ?**

When a friend asks you for a book recommendation, it’s pretty natural to ask what kinds of books they like. From there, you could think of a few titles that are similar to the things they’ve liked in the past. This process, of recommending content based on its characteristics, is at the heart of content-based filtering, the technology behind Netflix and Pandora’s recommendation engines.

[Read more](https://www.upwork.com/hiring/data/what-is-content-based-filtering/)

**What is collaborative filtering ?**

Collaborative filtering is a technique used by recommender systems. Collaborative filtering has two senses, a narrow one and a more general one

[Read more](https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0)

**What is overfitting ?**

Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points. Overfitting the model generally takes the form of making an overly complex model to explain idiosyncrasies in the data under study

Overfitting occurs when the model is working well with test data, fail at test data.

[Reference](https://elitedatascience.com/overfitting-in-machine-learning)

**How can you overcome from it  ?**

Cross-Validation : Cross Validation in its simplest form is a one round validation, where we leave one sample as in-time validation and rest for training the model. But for keeping lower variance a higher fold cross validation is preferred.

Early Stopping : Early stopping rules provide guidance as to how many iterations can be run before the learner begins to over-fit.

Pruning : Pruning is used extensively while building CART models. It simply removes the nodes which add little predictive power for the problem in hand.

Regularization : This is the technique we are going to discuss in more details. Simply put, it introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term.

[Reference](https://www.analyticsvidhya.com/blog/2015/02/avoid-over-fitting-regularization/)

**What is underfitting ?**

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Intuitively, underfitting occurs when the model or the algorithm does not fit the data well enough. Specifically, underfitting occurs if the model or algorithm shows low variance but high bias.

[**Reference**](https://chemicalstatistician.wordpress.com/2014/03/19/machine-learning-lesson-of-the-day-overfitting-and-underfitting/)

**How can you improve model performance ?**

You can improve the model performance by using following methods

1. Add more data
2. Treat missing and Outlier values
3. Feature Engineering
4. Feature Selection
5. Multiple algorithms
6. Algorithm Tuning
7. Ensemble methods
8. Cross Validation

[**Readmore**](https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/)

**How can you deploy your model ?**

We have 2 libraries for model deployment in python pickle and joblib.

You can easily deploy your model using flask api.

[Reference](https://www.analyticsvidhya.com/blog/2017/09/machine-learning-models-as-apis-using-flask/)

**What is Time series modeling ?**

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. … Time series forecasting is the use of a model to predict future values based on previously observed values.

Reference

1. [Time series using R](https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/)
2. [Time Series using Python](https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b?gi=e264c5451bf7)

**What is Stationary data  and non stationary data ?**

A stationary (time) series is one whose statistical properties such as the mean, variance and autocorrelation are all constant over time. Hence, a non-stationary series is one whose statistical properties change over time.

[Readmore](https://www.quora.com/What-is-Stationary-series-and-non-Stationary-series)

**What are different types of Times series algorithms ?**

* Naive Approach
* Simple average
* Moving average
* Single Exponential smoothing
* Holt’s linear trend method Method
* Holt’s Winter seasonal method Method
* ARIMA (Autoregressive and moving average)
* [Readmore](https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/)

**What is difference between statistical model and machine learning model ?**

Machine Learning is an algorithm that can learn from data without relying on rules-based programming.

Statistical Modelling is formalization of relationships between variables in the form of mathematical equations.

[Reference](https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/)

**Metrics :**

**Regression metrics:**

**What is mean square error ?**

In statistics, the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss

[Read more](https://www.statisticshowto.datasciencecentral.com/mean-squared-error/)

**What is root mean square error ?**

The root-mean-square deviation or root-mean-square error is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.

[Read more](https://www.statisticshowto.datasciencecentral.com/rmse/)

**What is R-Square ?**

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model.

Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

0% indicates that the model explains none of the variability of the response data around its mean.

100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits your data.

[Read more](http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit)

**Classification Metrics:**

**What is Confusion Matrix ?**

it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

[Read more](https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62)

**What is Type I Error  and Type 2 Error ?**

In statistical hypothesis testing, a type I error is the rejection of a true null hypothesis (also known as a “false positive” finding), while a type II error is the failure to reject a false null hypothesis (also known as a “false negative” finding). More simply stated, a type I error is to falsely infer the existence of something that is not there (confirming to common belief with false information), while a type II error is to falsely infer the absence of something that is present (going against the common belief with false information).

[Read more](https://keydifferences.com/difference-between-type-i-and-type-ii-errors.html)

**What is Difference between type 1 and type 2 error ?**

[Read more](https://keydifferences.com/difference-between-type-i-and-type-ii-errors.html)

**What is Precision and Recall ?**

In pattern recognition, information retrieval and binary classification, precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances

[Read more](https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c)

What is ROC curve ? when you will use it ?

As the area under an ROC curve is a measure of the usefulness of a test in general, where a greater area means a more useful test, the areas under ROC curves are used to compare the usefulness of tests. The termROC stands for Receiver Operating Characteristic.

[Read more](https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used)

**What is F1 Score ? when you will use it ?**

The f1-score is one of the most popular performance metrics. From what I recall this is the metric present in sklearn. In essence f1-score is the harmonic mean of the precision and recall.

[Read more](https://stackoverflow.com/questions/45963174/what-is-f1-score-and-what-its-value-indicates)

**Programming :**

**What is best programming language for machine learning Python or R or Spark or Sas ? Why ?**

Python is best option for machine learning, If you want understand why[read more](https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/)

**What is dummy variables ?**

Dummy variables (sometimes called indicator variables) are used in regression analysis and Latent Class Analysis. As implied by the name, these variables are artificial attributes, and they are used with two or more categories or levels. It’s used when you want to work with categorical variables which have no quantifiable relationship with each other.

[Read more](https://www.statisticshowto.datasciencecentral.com/dummy-variables/)

**What is one-hot encoding ?**

In digital circuits and machine learning, one-hot is a group of bits among which the legal combinations of values are only those with a single high bit and all the others low. A similar implementation in which all bits are ‘1’ except one ‘0’ is sometimes called one-cold

[Read more](https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f)

**How can you handle missing data in your data set ?**

In statistical language, if the number of the cases is less than 5% of the sample, then the researcher can drop them. In the case of multivariate analysis, if there is a larger number of missing values, then it can be better to drop those cases (rather than do imputation) and replace them.

[Read more](https://www.statisticssolutions.com/missing-values-in-data/)

**How can you handle duplicate values in your data set ?**

You can remove those values by using pandas dataframe with df.duplicates() function.

[Read more](https://www.reddit.com/r/MachineLearning/comments/2rmfx7/why_remove_duplicate_samples/)

**How can you handle outliers values  in your data set ?**

1. Univariate method:
2. Multivariate method
3. Minkowski error

[Read more](https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.html)

**What is data preprocessing ?**

**What are best libraries of data preprocessing ?**

Numpy, Pandas

**What are best libraries of machine learning ?**

Scikitlearn, StatsModels, Tensorflow, Keras, Catboost, Xgboost, Light Boost

**What are best libraries of data visialization ?**

Matplotbit, seaborn

**What is Rescaling ?**

**What is Feature scaling ?**

**What are different types of function in machine learning for Feature scaling ?**

StandardScaler, MinMaxScaler, Normalizer

[Readmore](https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e)

**What is Data Normalization ?**

It is also called as feature scaling.

What’s the difference between Normalization and Standardization?

**Machine learning Algorithms interview questions**

When to apply L1 regression ?

When to apply L2 regression ?

What is difference between KNN and K Means ?

How to choose k value in KNN ?

How the treen will be pruned in decision trees ?

How the tree will be split in decision trees ?

How can you divide the data to train set and test set ? Is it 50: 50 or 70: 30 ?

How can you do sentiment analysis ?

**What is PCA  and How it will work ?**

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It’s often used to make data easy to explore and visualize. We can also use PCA to Reduce dim

PCA is the simplest of the true eigenvector-based multivariate analyses.