Top 10 Machine Learning Algorithms

29 Jan 2016

Latest Update made on May 11, 2018

Machine Learning Algorithms Every Engineer Should Know

1. **Naïve Bayes Classifier Algorithm**
2. **K Means Clustering Algorithm**
3. **Support Vector Machine Algorithm**
4. **Apriori Algorithm**
5. **Linear Regression**
6. **Logistic Regression**
7. **Artificial Neural Networks**
8. **Random Forests**
9. **Decision Trees**
10. **Nearest Neighbours**

According to a recent study, machine learning algorithms are expected to replace 25% of the jobs across the world, in the next 10 years. With the rapid growth of big data and availability of programming tools like Python and R –machine learning is gaining mainstream presence for data scientists. Machine learning applications are highly automated and self-modifying which continue to improve over time with minimal human intervention as they learn with more data. For instance, Netflix's recommendation algorithm learns more about the likes and dislikes of a viewer based on the shows every viewer watches. To address the complex nature of various real world data problems, specialized machine learning algorithms have been developed that solve these problems perfectly. For beginners who are struggling to understand the basics of machine learning, here is a brief discussion on the top machine learning algorithms used by data scientists.

Machine Learning algorithms are classified as –

1) Supervised Machine Learning Algorithms

Machine learning algorithms that make predictions on given set of samples. Supervised machine learning algorithm searches for patterns within the value labels assigned to data points.

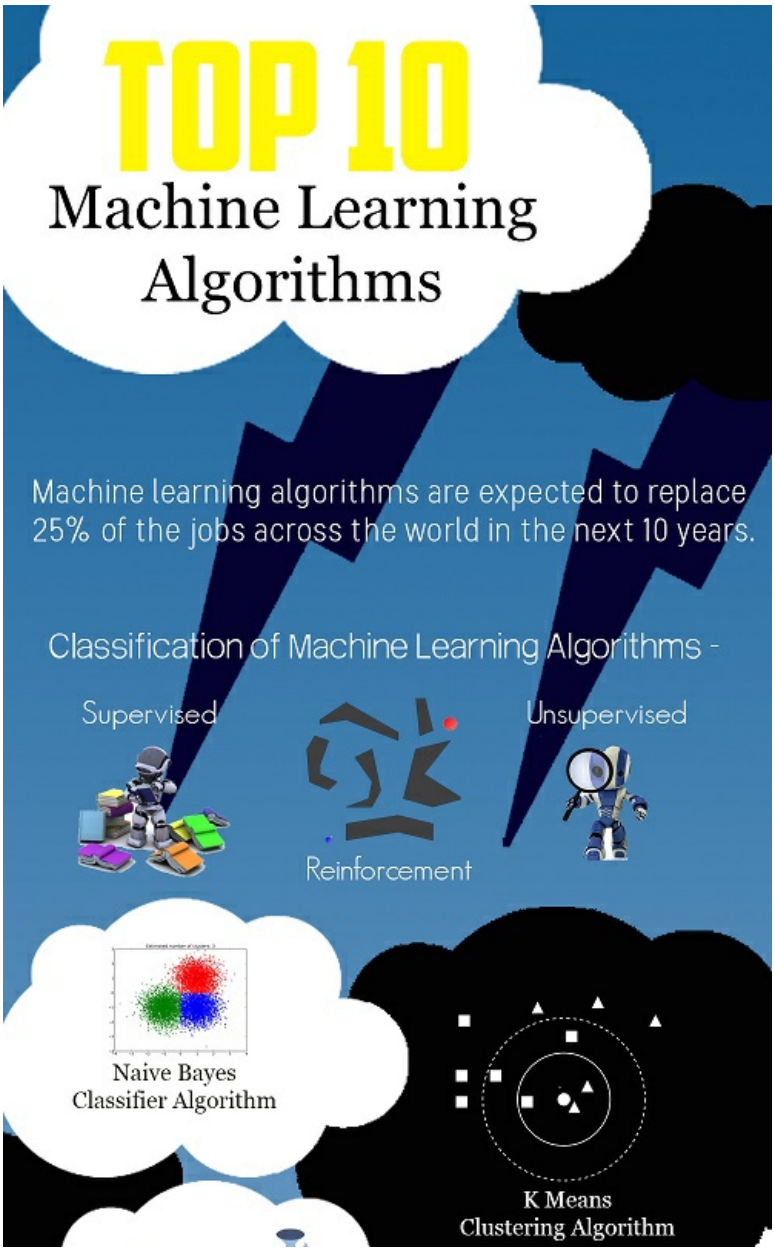
2) Unsupervised Machine Learning Algorithms

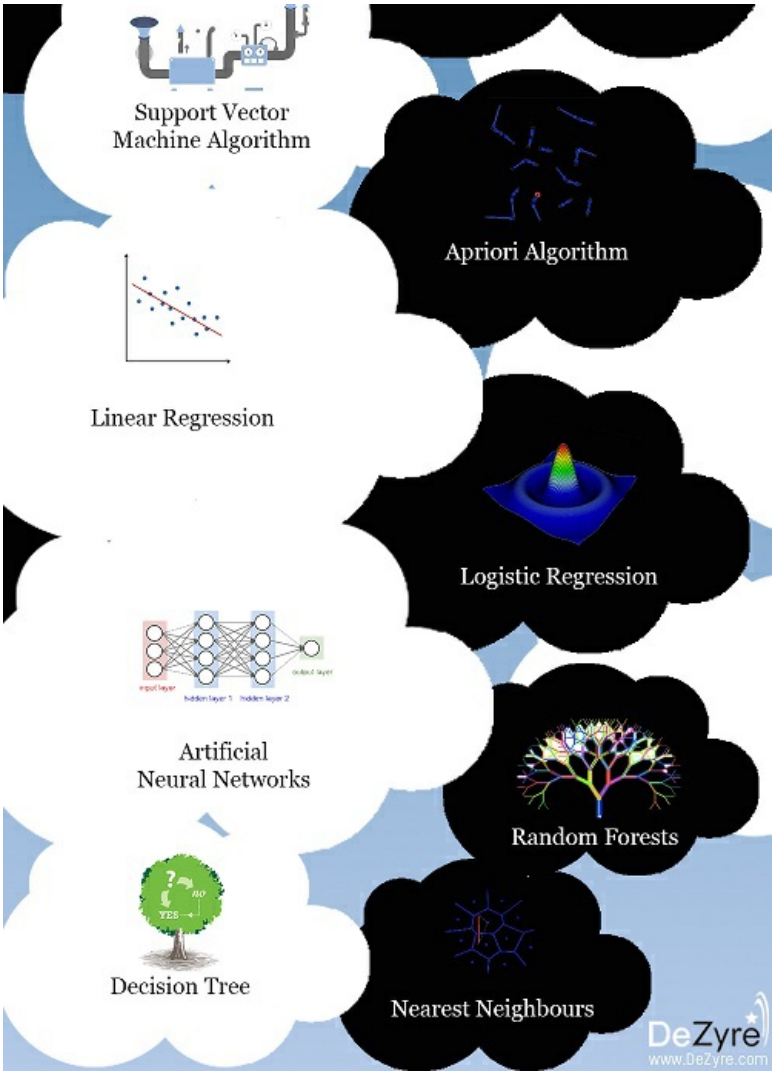
There are no labels associated with data points. These machine learning algorithms organize the data into a group of clusters to describe its structure and make complex data look simple and organized for analysis.

3) Reinforcement Machine Learning Algorithms

These algorithms choose an action, based on each data point and later learn how good the decision was. Over time, the algorithm changes its strategy to learn better and achieve the best reward.

Common Machine Learning Algorithms Infographic





Learn Data Science in Python and R to develop interesting machine learning applications!

What other machine learning algorithms do you think should have been on the list?

Enter your name here...

Write your answer here...

SUBMIT

---

1) Naïve Bayes Classifier Algorithm

It would be difficult and practically impossible to classify a web page, a document, an email or any other lengthy text notes manually. This is where Naïve Bayes Classifier machine learning algorithm comes to the rescue. A classifier is a function that allocates a population's element value from one of the available categories. For instance, Spam Filtering is a popular application of Naïve Bayes algorithm. Spam filter here, is a classifier that assigns a label "Spam" or "Not Spam" to all the emails.

Naïve Bayes Classifier is amongst the most popular learning method grouped by similarities, that works on the popular Bayes Theorem of Probability- to build machine learning models particularly for disease prediction and document classification. It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content.

When to use the Machine Learning algorithm - Naïve Bayes Classifier?

1. If you have a moderate or large training data set.
2. If the instances have several attributes.
3. Given the classification parameter, attributes which describe the instances should be conditionally independent.

Applications of Naïve Bayes Classifier

1. **Sentiment Analysis**- It is used at Facebook to analyse status updates expressing positive or negative emotions.
2. **Document Categorization**- Google uses document classification to index documents and find relevancy scores i.e. the PageRank. PageRank mechanism considers the pages marked as important in the databases that were parsed and classified using a document classification technique.
3. Naïve Bayes Algorithm is also used for classifying news articles about Technology, Entertainment, Sports, Politics, etc.
4. **Email Spam Filtering**-Google Mail uses Naïve Bayes algorithm to classify your emails as Spam or Not Spam

Advantages of the Naïve Bayes Classifier Machine Learning Algorithm

1. Naïve Bayes Classifier algorithm performs well when the input variables are categorical.
2. A Naïve Bayes classifier converges faster, requiring relatively little training data than other discriminative models like logistic regression, when the Naïve Bayes conditional independence assumption holds.
3. With Naïve Bayes Classifier algorithm, it is easier to predict class of the test data set. A good bet for multi class predictions as well.
4. Though it requires conditional independence assumption, Naïve Bayes Classifier has presented good performance in various application domains.

Data Science Libraries in Python to implement Naïve Bayes – Sci-Kit Learn

Data Science Libraries in R to implement Naïve Bayes – e1071

Access the Solution to Kaggle Data Science Challenge - Predict the Survial of Titanic Passengers

2) K Means Clustering Algorithm

K-means is a popularly used unsupervised machine learning algorithm for cluster analysis. K-Means is a non-deterministic and iterative method. The algorithm operates on a given data set through pre-defined number of clusters, k. The output of K Means algorithm is k clusters with input data partitioned among the clusters.

For instance, let's consider K-Means Clustering for Wikipedia Search results. The search term "Jaguar" on Wikipedia will return all pages containing the word Jaguar which can refer to Jaguar as a Car, Jaguar as Mac OS version and Jaguar as an Animal. K Means clustering algorithm can be applied to group the webpages that talk about similar concepts. So, the algorithm will group all web pages that talk about Jaguar as an Animal into one cluster, Jaguar as a Car into another cluster and so on.

Advantages of using K-Means Clustering Machine Learning Algorithm

- In case of globular clusters, K-Means produces tighter clusters than hierarchical clustering.
- Given a smaller value of K, K-Means clustering computes faster than hierarchical clustering for large number of variables.

Applications of K-Means Clustering

K Means Clustering algorithm is used by most of the search engines like Yahoo, Google to cluster web pages by similarity and identify the 'relevance rate' of search results. This helps search engines reduce the computational time for the users.

Data Science Libraries in Python to implement K-Means Clustering – SciPy, Sci-Kit Learn, Python Wrapper

Data Science Libraries in R to implement K-Means Clustering – stats

3) Support Vector Machine Learning Algorithm

Support Vector Machine is a supervised machine learning algorithm for classification or regression problems where the dataset teaches SVM about the classes so that SVM can classify any new data. It works by classifying the data into different classes by finding a line (hyperplane) which separates the training data set into classes. As there are many such linear hyperplanes, SVM algorithm tries to maximize the distance between the various classes that are involved and this is referred as margin maximization. If the line that maximizes the distance between the classes is identified, the probability to generalize well to unseen data is increased.

SVM's are classified into two categories:

- Linear SVM's – In linear SVM's the training data i.e. classifiers are separated by a hyperplane.
- Non-Linear SVM's- In non-linear SVM's it is not possible to separate the training data using a hyperplane. For example, the training data for Face detection consists of group of images that are faces and another group of images that are not faces (in other words all other images in the world except faces). Under such conditions, the training data is too complex that it is impossible to find a representation for every feature vector. Separating the set of faces linearly from the set of non-face is a complex task.

Advantages of Using SVM

- SVM offers best classification performance (accuracy) on the training data.
- SVM renders more efficiency for correct classification of the future data.
- The best thing about SVM is that it does not make any strong assumptions on data.
- It does not over-fit the data.

Applications of Support Vector Machine

SVM is commonly used for stock market forecasting by various financial institutions. For instance, it can be used to compare the relative performance of the stocks when compared to performance of other stocks in the same sector. The relative comparison of stocks helps manage investment making decisions based on the classifications made by the SVM learning algorithm.

Data Science Libraries in Python to implement Support Vector Machine –SciKit Learn, PyML , SVM^{Struct} Python , LIBSVM

Data Science Libraries in R to implement Support Vector Machine – klar, e1071

4) Apriori Machine Learning Algorithm

Apriori algorithm is an unsupervised machine learning algorithm that generates association rules from a given data set. Association rule implies that if an item A occurs, then item B also occurs with a certain probability. Most of the association rules generated are in the IF_THEN format. For example, IF people buy an iPad THEN they also buy an iPad Case to protect it. For the algorithm to derive such conclusions, it first observes the number of people who bought an iPad case while purchasing an iPad. This way a ratio is derived like out of the 100 people who purchased an iPad, 85 people also purchased an iPad case.

Basic principle on which Apriori Machine Learning Algorithm works:

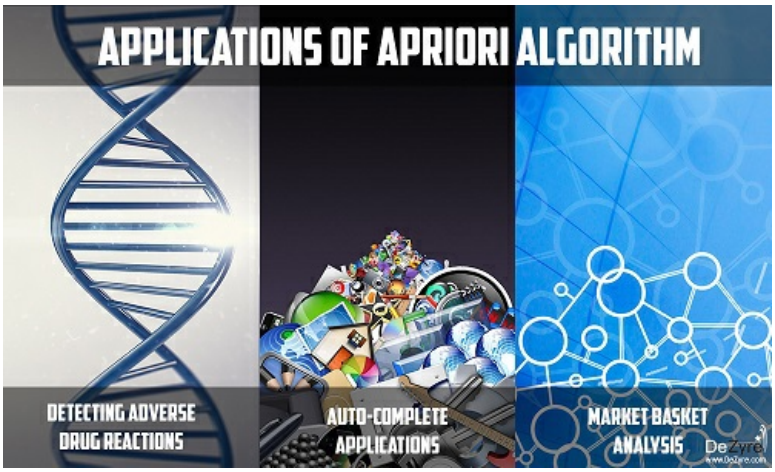
- If an item set occurs frequently then all the subsets of the item set, also occur frequently.
- If an item set occurs infrequently then all the supersets of the item set have infrequent occurrence.

Advantages of Apriori Algorithm

- It is easy to implement and can be parallelized easily.

- Apriori implementation makes use of large item set properties.

Applications of Apriori Algorithm



- **Detecting Adverse Drug Reactions**

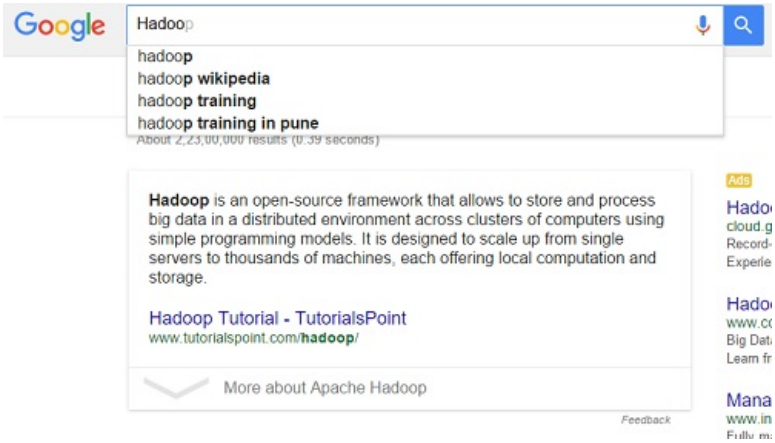
Apriori algorithm is used for association analysis on healthcare data like-the drugs taken by patients, characteristics of each patient, adverse ill-effects patients experience, initial diagnosis, etc. This analysis produces association rules that help identify the combination of patient characteristics and medications that lead to adverse side effects of the drugs.

- **Market Basket Analysis**

Many e-commerce giants like Amazon use Apriori to draw data insights on which products are likely to be purchased together and which are most responsive to promotion. For example, a retailer might use Apriori to predict that people who buy sugar and flour are likely to buy eggs to bake a cake.

- **Auto-Complete Applications**

Google auto-complete is another popular application of Apriori wherein - when the user types a word, the search engine looks for other associated words that people usually type after a specific word.



Data Science Libraries in Python to implement Apriori Machine Learning Algorithm – There is a python implementation for Apriori in PyPi

Data Science Libraries in R to implement Apriori Machine Learning Algorithm – arules

5) Linear Regression Machine Learning Algorithm

Linear Regression algorithm shows the relationship between 2 variables and how the change in one variable impacts the other. The algorithm shows the impact on the dependent variable on changing the independent variable. The independent variables are referred as explanatory variables, as they explain the factors the impact the dependent variable. Dependent variable is often referred to as the factor of interest or predictor.

Advantages of Linear Regression Machine Learning Algorithm

- It is one of the most interpretable machine learning algorithms, making it easy to explain to others.
- It is easy of use as it requires minimal tuning.
- It is the mostly widely used machine learning technique that runs fast.

Applications of Linear Regression

- **Estimating Sales**

Linear Regression finds great use in business, for sales forecasting based on the trends. If a company observes steady increase in sales every month - a linear regression analysis of the monthly sales data helps the company forecast sales in upcoming months.

Access the Solution to Kaggle Data Science Challenge - Walmart Store Sales Forecasting

- **Risk Assessment**

Linear Regression helps assess risk involved in insurance or financial domain. A health insurance company can do a linear regression analysis on the number of claims per customer against age. This analysis helps insurance companies find, that older customers tend to make more insurance claims. Such analysis results play a vital role in important business decisions and are made to account for risk.

Data Science Libraries in Python to implement Linear Regression – statsmodel and SciKit

Data Science Libraries in R to implement Linear Regression – stats

Explanations about the top machine learning algorithms will continue, as it is a work in progress. Stay tuned to our blog to learn more about the popular machine learning algorithms and their applications!!!

Learn Data Science in Python and R to solve a range of data science problems using machine learning!

6) Decision Tree Machine Learning Algorithm

DECISION TREE MACHINE LEARNING ALGORITHMS

Decision Tree Algorithm

A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision based on certain conditions.

Why use Decision Tree Algorithm?

1. 

Help make decision under uncertainty and help you improve communication as they present a visual representation of a decision situation.

2. 

Help a data scientist capture the idea that if a different decision was taken then how the operational nature of a situation or model would have changed intensely.

3. 

Make optimal decisions by allowing a data scientist to traverse through forward and backward calculation paths.

When to use Decision Tree Machine Learning Algorithm?

Training data contains errors.

Instances are represented by attribute value pairs.

Training data has missing values

The target function has discrete output values.

Advantages of Using Decision Tree Machine Learning Algorithms

Instinctual and can be explained to anyone with ease

Implicitly perform feature selection which is very important in predictive analytics

Help save data preparation time as they are not sensitive to missing values and outliers

Do not require making any assumption on the linearity in the data

Drawbacks of Using Decision Tree Machine Learning Algorithm

1. 

The outcomes may be based on expectations

2. 

You are making a weekend plan to visit the best restaurant in town as your parents are visiting but you are hesitant in making a decision on which restaurant to choose. Whenever you want to visit a restaurant you ask your friend Tyrion if he thinks you will like a particular place. To answer your question, Tyrion first has to find out, the kind of restaurants you like. You give him a list of restaurants that you have visited and tell him whether you liked each restaurant or not (giving a labelled training dataset). When you ask Tyrion that whether you will like a particular restaurant R or not, he asks you various questions like “Is “R” a roof top restaurant?” , “Does restaurant “R” serve Italian cuisine?”, “Does R have live music?”, “Is restaurant R open till midnight?” and so on. Tyrion asks you several informative questions to maximize the information gain and gives you YES or NO answer based on your answers to the questionnaire. Here Tyrion is a decision tree for your favourite restaurant preferences.

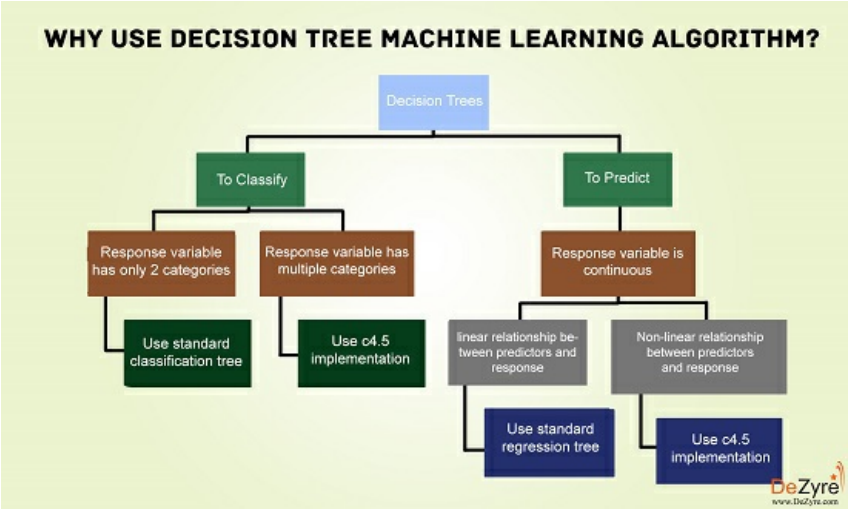
A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions. In a decision tree, the internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i.e. the decision made after computing all of the attributes. The classification rules are represented through the path from root to the leaf node.

Types of Decision Trees

Classification Trees- These are considered as the default kind of decision trees used to separate a dataset into different classes, based on the response variable. These are generally used when the response variable is categorical in nature.

Regression Trees-When the response or target variable is continuous or numerical, regression trees are used. These are generally used in predictive type of problems when compared to classification.

Decision trees can also be classified into two types, based on the type of target variable- Continuous Variable Decision Trees and Binary Variable Decision Trees. It is the target variable that helps decide what kind of decision tree would be required for a particular problem.



Why should you use Decision Tree Machine Learning algorithm?

- These machine learning algorithms help make decisions under uncertainty and help you improve communication, as they present a visual representation of a decision situation.
- Decision tree machine learning algorithms help a data scientist capture the idea that if a different decision was taken, then how the operational nature of a situation or model would have changed intensely.
- Decision tree algorithms help make optimal decisions by allowing a data scientist to traverse through forward and backward calculation paths.

When to use Decision Tree Machine Learning Algorithm

- Decision trees are robust to errors and if the training data contains errors- decision tree algorithms will be best suited to address such problems.
- Decision trees are best suited for problems where instances are represented by attribute value pairs.
- If the training data has missing value then decision trees can be used, as they can handle missing values nicely by looking at the data in other columns.
- Decision trees are best suited when the target function has discrete output values.

Advantages of Using Decision Tree Machine Learning Algorithms

- Decision trees are very instinctual and can be explained to anyone with ease. People from a non-technical background, can also decipher the hypothesis drawn from a decision tree, as they are self-explanatory.
- When using decision tree machine learning algorithms, data type is not a constraint as they can handle both categorical and numerical variables.
- Decision tree machine learning algorithms do not require making any assumption on the linearity in the data and hence can be used in circumstances where the parameters are non-linearly related. These machine learning algorithms do not make any assumptions on the classifier structure and space distribution.
- These algorithms are useful in data exploration. Decision trees implicitly perform feature selection which is very important in predictive analytics. When a decision tree is fit to a training dataset, the nodes at the top on which the decision tree is split, are considered as important variables within a given dataset and feature selection is completed by default.
- Decision trees help save data preparation time, as they are not sensitive to missing values and outliers. Missing values will not stop you from splitting the data for building a decision tree. Outliers will also not affect the decision trees as data splitting happens based on some samples within the split range and not on exact absolute values.

Drawbacks of Using Decision Tree Machine Learning Algorithms

- The more the number of decisions in a tree, less is the accuracy of any expected outcome.
- A major drawback of decision tree machine learning algorithms, is that the outcomes may be based on expectations. When decisions are made in real-time, the payoffs and resulting outcomes might not be the same as expected or planned. There are chances that this could lead to unrealistic decision trees leading to bad decision making. Any irrational expectations could lead to major errors and flaws in decision tree analysis, as it is not always possible to plan for all eventualities that can arise from a decision.
- Decision Trees do not fit well for continuous variables and result in instability and classification plateaus.

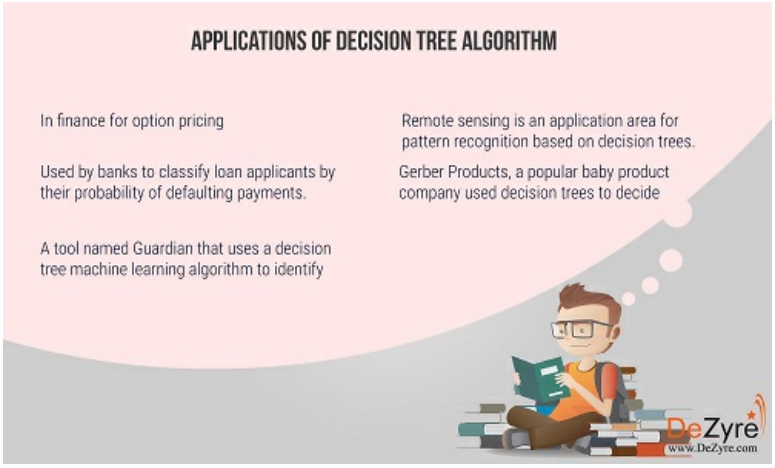
- Decision trees are easy to use when compared to other decision making models but creating large decision trees that contain several branches is a complex and time consuming task.
- Decision tree machine learning algorithms consider only one attribute at a time and might not be best suited for actual data in the decision space.
- Large sized decision trees with multiple branches are not comprehensible and pose several presentation difficulties.

Applications of Decision Tree Machine Learning Algorithm

- Decision trees are among the popular machine learning algorithms that find great use in finance for option pricing.
- Remote sensing is an application area for pattern recognition based on decision trees.
- Decision tree algorithms are used by banks to classify loan applicants by their probability of defaulting payments.
- Gerber Products, a popular baby product company, used decision tree machine learning algorithm to decide whether they should continue using the plastic PVC (Poly Vinyl Chloride) in their products.
- Rush University Medical Centre has developed a tool named Guardian that uses a decision tree machine learning algorithm to identify at-risk patients and disease trends.

The Data Science libraries in Python language to implement Decision Tree Machine Learning Algorithm are – SciPy and Sci-Kit Learn.

The Data Science libraries in R language to implement Decision Tree Machine Learning Algorithm is caret.



Random Forest Machine Learning Algorithm

Let’s continue with the same example that we used in decision trees, to explain how Random Forest Machine Learning Algorithm works. Tyrion is a decision tree for your restaurant preferences. However, Tyrion being a human being does not always generalize your restaurant preferences with accuracy. To get more accurate restaurant recommendation, you ask a couple of your friends and decide to visit the restaurant R, if most of them say that you will like it. Instead of just asking Tyrion, you would like to ask Jon Snow, Sandor, Bronn and Bran who vote on whether you will like the restaurant R or not. This implies that you have built an ensemble classifier of decision trees - also known as a forest.

You don’t want all your friends to give you the same answer - so you provide each of your friends with slightly varying data. You are also not sure of your restaurant preferences and are in a dilemma.You told Tyrion that you like Open Roof Top restaurants but maybe, just because it was summer when you visited the restaurant you could have liked it then. You may not be a fan of the restaurant during the chilly winters. Thus, all your friends should not make use of the data point that you like open roof top restaurants, to make their recommendations for your restaurant preferences.

By providing your friends with slightly different data on your restaurant preferences, you make your friends ask you different questions at different times. In this case just by slightly altering your restaurant preferences, you are injecting randomness at model level (unlike randomness at data level in case of decision trees). Your group of friends now form a random forest of your restaurant preferences.

Random Forest is the go to machine learning algorithm that uses a bagging approach to create a bunch of decision trees with random subset of the data. A model is trained several times on random sample of the dataset to achieve good prediction performance from the random forest algorithm.In this ensemble learning method, the output of all the decision trees in the random forest, is combined to make the final prediction. The final prediction of the random forest algorithm is derived by polling the results of each decision tree or just by going with a prediction that appears the most times in the decision trees.

For instance, in the above example - if 5 friends decide that you will like restaurant R but only 2 friends decide that you will not like the restaurant then the final prediction is that, you will like restaurant R as majority always wins.

Access the Solution to Kaggle Data Science Challenge - Expedia Hotel Recommendations

Why use Random Forest Machine Learning Algorithm?

- There are many good open source, free implementations of the algorithm available in Python and R.
- It maintains accuracy when there is missing data and is also resistant to outliers.
- Simple to use as the basic random forest algorithm can be implemented with just a few lines of code.
- Random Forest machine learning algorithms help data scientists save data preparation time, as they do not require any input preparation and are capable of handling numerical, binary and categorical features, without scaling, transformation or modification.
- Implicit feature selection as it gives estimates on what variables are important in the classification.

Advantages of Using Random Forest Machine Learning Algorithms

- Overfitting is less of an issue with Random Forests, unlike decision tree machine learning algorithms. There is no need of pruning the random forest.
- These algorithms are fast but not in all cases. A random forest algorithm, when run on an 800 MHz machine with a dataset of 100 variables and 50,000 cases produced 100 decision trees in 11 minutes.
- Random Forest is one of the most effective and versatile machine learning algorithm for wide variety of classification and regression tasks, as they are more robust to noise.
- It is difficult to build a bad random forest. In the implementation of Random Forest Machine Learning algorithms, it is easy to determine which parameters to use because they are not sensitive to the parameters that are used to run the algorithm. One can easily build a decent model without much tuning.
- Random Forest machine learning algorithms can be grown in parallel.
- This algorithm runs efficiently on large databases.
- Has higher classification accuracy.

Drawbacks of Using Random Forest Machine Learning Algorithms

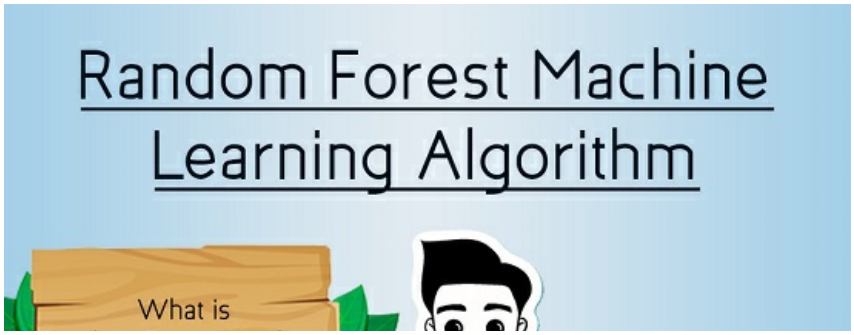
- They might be easy to use but analysing them theoretically, is difficult.
- Large number of decision trees in the random forest can slow down the algorithm in making real-time predictions.
- If the data consists of categorical variables with different number of levels, then the algorithm gets biased in favour of those attributes that have more levels. In such situations, variable importance scores do not seem to be reliable.
- When using RandomForest algorithm for regression tasks, it does not predict beyond the range of the response values in the training data.

Applications of Random Forest Machine Learning Algorithms

- Random Forest algorithms are used by banks to predict if a loan applicant is a likely high risk.
- They are used in the automobile industry to predict the failure or breakdown of a mechanical part.
- These algorithms are used in the healthcare industry to predict if a patient is likely to develop a chronic disease or not.
- They can also be used for regression tasks like predicting the average number of social media shares and performance scores.
- Recently, the algorithm has also made way into predicting patterns in speech recognition software and classifying images and texts.

Data Science libraries in Python language to implement Random Forest Machine Learning Algorithm is Sci-Kit Learn.

Data Science libraries in R language to implement Random Forest Machine Learning Algorithm is randomForest.



Random Forest Machine Learning Algorithm?

Random Forest uses a bagging approach to create a bunch of decision trees with random subset of the data.

The output of all the decision trees in the random forest, is combined to make the final prediction.

The final prediction of the random forest algorithm is derived –

By polling the results of each decision tree.

Just by going with a prediction that appears the most times in the decision trees.

Why use Random Forest Machine Learning Algorithm?

There are many good open source, free implementations of the algorithm available in Python and R.

It maintains accuracy when there is missing data and is also resistant to outliers.

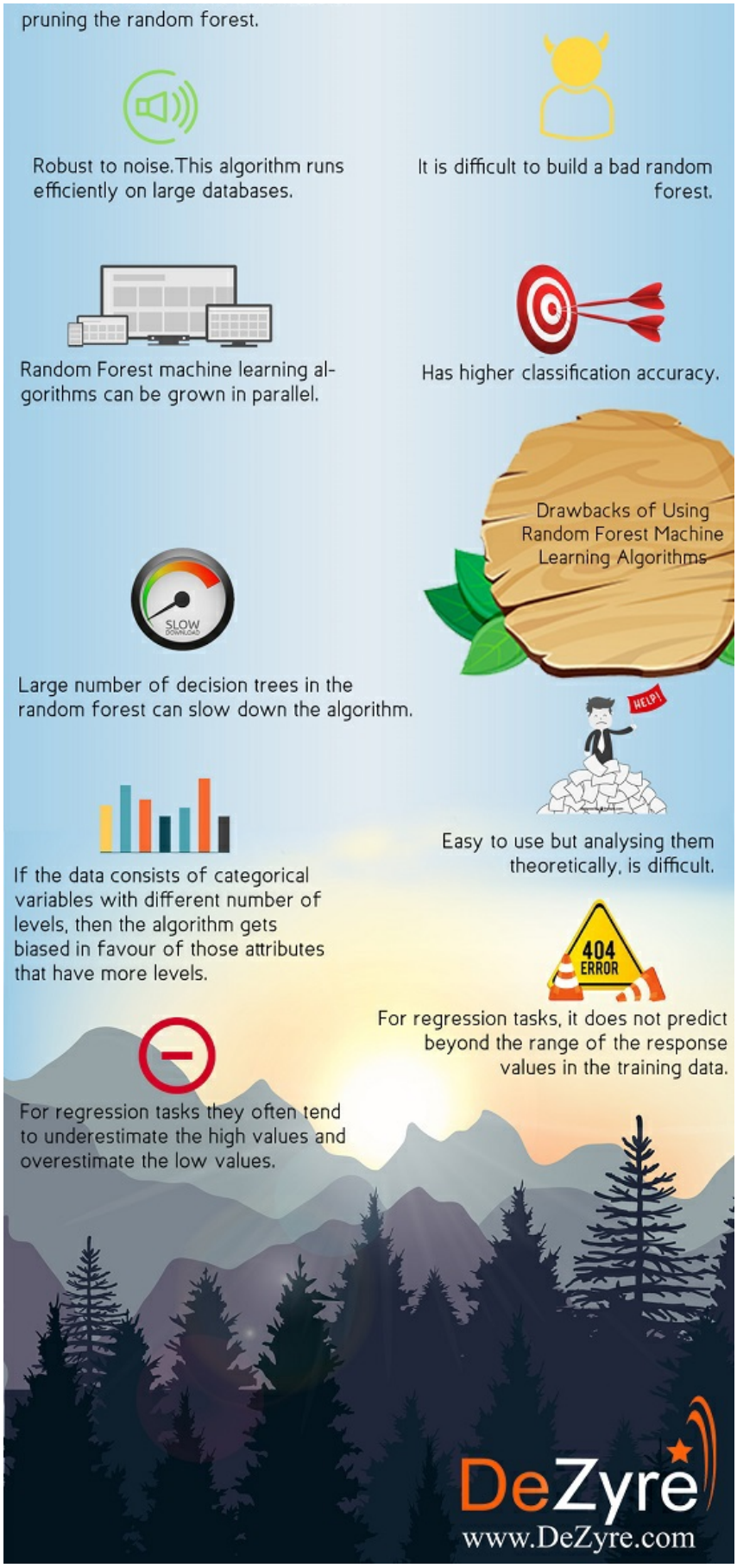
Help data scientists save data preparation time as it does do not require any input preparation.

Implicit feature selection as it gives estimates on what variables are important in the classification.

Advantages of Using Random Forest Machine Learning Algorithms

These algorithms are fast but not in all cases. A random forest algorithm, when run on an 800 MHz machine with a dataset of 100 variables and 50,000 cases produced 100 decision trees in 11 minutes.

Overfitting is less of an issue with Random Forests. There is no need of



Logistic Regression

The name of this algorithm could be a little confusing in the sense that Logistic Regression machine learning algorithm is for classification tasks and not regression problems. The name 'Regression' here implies that a linear model is fit into the feature space. This algorithm applies a logistic function to a linear combination of features to predict the outcome of a categorical dependent variable based on predictor variables.

The odds or probabilities that describe the outcome of a single trial are modelled as a function of explanatory variables. Logistic regression algorithms helps estimate the probability of falling into a specific level of the categorical dependent variable based on the given predictor variables.

Just suppose that you want to predict if there will be a snowfall tomorrow in New York. Here the outcome of the prediction is not a continuous number because there will either be snowfall or no snowfall and hence linear regression cannot be applied. Here the outcome variable is one of the several categories and using logistic regression helps.

Based on the nature of categorical response, logistic regression is classified into 3 types –

· **Binary Logistic Regression** – The most commonly used logistic regression when the categorical response has 2 possible outcomes i.e. either yes or not. Example –Predicting whether a student will pass or fail an exam, predicting whether a student will have low or high blood pressure, predicting whether a tumour is cancerous or not.

· **Multi-nominal Logistic Regression** - Categorical response has 3 or more possible outcomes with no ordering. Example- Predicting what kind of search engine (Yahoo, Bing, Google, and MSN) is used by majority of US citizens.

· **Ordinal Logistic Regression** - Categorical response has 3 or more possible outcomes with natural ordering. Example- How a customer rates the service and quality of food at a restaurant based on a scale of 1 to 10.

Let us consider a simple example where a cake manufacturer wants to find out if baking a cake at 160°C, 180°C and 200°C will produce a ‘hard’ or ‘soft’ variety of cake (assuming the fact that the bakery sells both the varieties of cake with different names and prices). Logistic regression is a perfect fit in this scenario instead of other statistical techniques. For example, if the manufactures produces 2 cake batches wherein the first batch contains 20 cakes (of which 7 were hard and 13 were soft) and the second batch of cake produced consisted of 80 cakes (of which 41 were hard and 39 were soft cakes). Here in this case if linear regression algorithm is used it will give equal importance both the batches of cakes regardless of the number of cakes in each batch. Applying a logistic regression algorithm will consider this factor and give the second batch of cakes more weightage than the first batch.

When to Use Logistic Regression Machine Learning Algorithm

- Use logistic regression algorithms when there is a requirement to model the probabilities of the response variable as a function of some other explanatory variable. For example, probability of buying a product X as a function of gender
- Use logistic regression algorithms when there is a need to predict probabilities that categorical dependent variable will fall into two categories of the binary response as a function of some explanatory variables. For example, what is the probability that a customer will buy a perfume given that the customer is a female?
- Logistic regression algorithms is also best suited when the need is to classify elements two categories based on the explanatory variable. For example-classify females into ‘young’ or ‘old’ group based on their age.

Advantages of Using Logistic Regression

- Easier to inspect and less complex.
- Robust algorithm as the independent variables need not have equal variance or normal distribution.
- These algorithms do not assume a linear relationship between the dependent and independent variables and hence can also handle non-linear effects.
- Controls confounding and tests interaction.

Drawbacks of Using Logistic Regression

- When the training data is sparse and high dimensional, in such situations a logistic model may over fit the training data.
- Logistic regression algorithms cannot predict continuous outcomes. For instance, logistic regression cannot be applied when the goal is to determine how heavily it will rain because the scale of measuring rainfall is continuous. Data scientists can predict heavy or low rainfall but this would make some compromises with the precision of the dataset.
- Logistic regression algorithms require more data to achieve stability and meaningful results. These algorithms require minimum of 50 data points per predictor to achieve stable outcomes.
- It predicts outcomes depending on a group of independent variables and if a data scientist or a machine learning expert goes wrong in identifying the independent variables then the developed model will have minimal or no predictive value.
- It is not robust to outliers and missing values.

Applications of Logistic Regression

- Logistic regression algorithm is applied in the field of epidemiology to identify risk factors for diseases and plan accordingly for preventive measures.
- Used to predict whether a candidate will win or lose a political election or to predict whether a voter will vote for a particular candidate.
- Used to classify a set of words as nouns, pronouns, verbs, adjectives.
- Used in weather forecasting to predict the probability of rain.
- Used in credit scoring systems for risk management to predict the defaulting of an account.

The Data Science libraries in Python language to implement Logistic Regression Machine Learning Algorithm is Sci-Kit Learn.

The Data Science libraries in R language to implement Logistic Regression Machine Learning Algorithm is stats package (glm () function)

Artificial Neural Networks Machine Learning Algorithm- Human Brain Simulator

Human brain has a highly complex and non-linear parallel computer which can organize the structural constituents i.e. the neurons interconnected in a complex manner between each other. Let us take a simple example of face recognition-whenever we meet a person, a person who is known to us can be easily recognized with his name or he works at XYZ place or based on his relationship with you. We may be knowing thousands of people, the task requires the human brain to immediately recognize the person (face recognition).Now, suppose instead of the human brain doing it, if a computer is asked to perform this task. It is not going to be an easy computation for the machine as it does not know the person. You have to teach the computer that there are images of different people. If you know 10,000 people then you have to feed all the 10,000 photographs into the computer. Now, whenever you meet a person you capture an image of the person and feed it to the computer. The computer matches this photograph with all the 10,000 photographs that you have already fed into the database. At the end of all the computations-it gives the result with the photograph that best resembles the person. This could take several hours or more depending on the number of images present in the database. The complexity of the task will increase with increase in the number of images in the database. However, a human brain can recognise it instantly.

What is the future of Machine Learning?

Enter your name here...

Write your answer here...

SUBMIT

PREVIOUS NEXT

Answers

Currently have 53answers

Q: What other machine learning algorithms do you think should have been on the list?

dilaw k-means
May 06 2019, 10:44 PM

Knn knn
May 06 2019, 08:43 PM

s.vadivukkarasi adaboost and stacking algorithms
Feb 04 2019, 07:42 PM

zooyoung lee XGboost
Jan 28 2019, 02:14 AM

SRINIVASA V Ensembling
Jan 23 2019, 04:30 PM

View 28 more answers

Q: What is the future of Machine Learning?

James Machine learning will continue to advance until nearly everything a human can do will be done better by a machine (but the machine won't "know" it, because self awareness is a long ways away)
Oct 23 2018, 03:31 PM

Nagham It is too big to be abbreviated with word
Jul 28 2018, 04:02 PM

Adam Irfan To predict and classify the persons characters by the training dataset.
Jun 26 2018, 03:44 PM

DQ recurrent neural network
Apr 24 2018, 01:31 PM

mahavir machines will be work on multiple level. communication between machine to machine is easily possible
Mar 22 2018, 05:06 PM

View 15 more answers

Big Data and Hadoop Training Courses in Popular Cities

- Hadoop Training in Texas
- Hadoop Training in California
- Hadoop Training in Dallas
- Hadoop Training in Chicago
- Hadoop Training in Charlotte
- Hadoop Training in Dubai
- Hadoop Training in Edison
- Hadoop Training in Fremont
- Hadoop Training in San Jose
- Hadoop Training in Washington

- Hadoop Training in New Jersey
- Hadoop Training in New York
- Hadoop Training in Atlanta
- Hadoop Training in Canada
- Hadoop Training in Abu Dhabi
- Hadoop Training in Detroit
- Hadoop Trainging in Germany
- Hadoop Training in Houston
- Hadoop Training in Virginia

Upcoming Live Machine Learning

02	**Sat and Sun (6 weeks)**	**$399**

Jun

7:00 AM - 10:00 AM PST

LEARN MORE

Relevant Courses

- Hadoop Online Training
- Apache Spark Training
- Data Science in Python Training
- Data Science in R Language Training
- Salesforce Certification Training
- NoSQL Database Training
- Hadoop Admin Training

You might also like

- Top 100 Hadoop Interview Questions and Answers 2017
- Pig Interview Questions and Answers
- Hive Interview Questions and Answers
- HBase Interview Questions and Answers
- MapReduce Interview Questions and Answers
- HDFS Interview Questions and Answers
- Real-Time Hadoop Interview Questions and Answers
- Hadoop Admin Interview Questions and Answers
- Basic Hadoop Interview Questions and Answers
- Apache Spark Interview Questions and Answers
- Data Analyst Interview Questions and Answers
- 100 Data Science Interview Questions and Answers (General)
- 100 Data Science in R Interview Questions and Answers

- 100 Data Science in Python Interview Questions and Answers
- Recap of Hadoop News for September 2018
- Introduction to TensorFlow for Deep Learning
- Recap of Hadoop News for August 2018
- AWS vs Azure-Who is the big winner in the cloud war?
- Top 5 Reasons to Learn AWS
- Top 50 AWS Interview Questions and Answers for 2018
- Recap of Hadoop News for July 2018
- Top 10 Machine Learning Projects for Beginners
- Recap of Data Science News for June 2018
- Recap of Apache Spark News for June 2018

Blog Categories

- Big Data
- CRM
- Data Science
- Live Courses
- Mobile App Development
- NoSQL Database
- Web Development

Tutorials

- Hadoop Online Tutorial – Hadoop HDFS Commands Guide
- MapReduce Tutorial–Learn to implement Hadoop WordCount Example
- Hadoop Hive Tutorial-Usage of Hive Commands in HQL
- Hive Tutorial-Getting Started with Hive Installation on Ubuntu
- Learn Java for Hadoop Tutorial: Inheritance and Interfaces
- Learn Java for Hadoop Tutorial: Classes and Objects
- Learn Java for Hadoop Tutorial: Arrays
- Apache Spark Tutorial–Run your First Spark Program
- PySpark Tutorial-Learn to use Apache Spark with Python
- R Tutorial- Learn Data Visualization with R using GGVIS
- Neural Network Training Tutorial
- Python List Tutorial

- MatPlotLib Tutorial
- Decision Tree Tutorial
- Neural Network Tutorial
- Performance Metrics for Machine Learning Algorithms
- R Tutorial: Data.Table
- SciPy Tutorial
- Step-by-Step Apache Spark Installation Tutorial
- Introduction to Apache Spark Tutorial
- R Tutorial: Importing Data from Web
- R Tutorial: Importing Data from Relational Database
- R Tutorial: Importing Data from Excel
- Introduction to Machine Learning Tutorial
- Machine Learning Tutorial: Linear Regression
- Machine Learning Tutorial: Logistic Regression
- Support Vector Machine Tutorial (SVM)
- K-Means Clustering Tutorial
- dplyr Manipulation Verbs
- Introduction to dplyr package
- Importing Data from Flat Files in R
- Principal Component Analysis Tutorial
- Pandas Tutorial Part-3
- Pandas Tutorial Part-2
- Pandas Tutorial Part-1
- Tutorial- Hadoop Multinode Cluster Setup on Ubuntu
- Data Visualizations Tools in R
- R Statistical and Language tutorial
- Introduction to Data Science with R
- Apache Pig Tutorial: User Defined Function Example
- Apache Pig Tutorial Example: Web Log Server Analytics
- Impala Case Study: Web Traffic
- Impala Case Study: Flight Data Analysis
- Hadoop Impala Tutorial
- Apache Hive Tutorial: Tables
- Flume Hadoop Tutorial: Twitter Data Extraction
- Flume Hadoop Tutorial: Website Log Aggregation

- Hadoop Sqoop Tutorial: Example Data Export
- Hadoop Sqoop Tutorial: Example of Data Aggregation
- Apache Zookepeer Tutorial: Example of Watch Notification
- Apache Zookepeer Tutorial: Centralized Configuration Management
- Hadoop Zookeeper Tutorial
- Hadoop Sqoop Tutorial
- Hadoop PIG Tutorial
- Hadoop Oozie Tutorial
- Hadoop NoSQL Database Tutorial
- Hadoop Hive Tutorial
- Hadoop HDFS Tutorial
- Hadoop hBase Tutorial
- Hadoop Flume Tutorial
- Hadoop 2.0 YARN Tutorial
- Hadoop MapReduce Tutorial
- Big Data Hadoop Tutorial for Beginners- Hadoop Installation

Online Courses

- Hadoop Training
- Spark Certification Training
- Data Science in Python
- Data Science inR
- Data Science Training

Salesforce Certifications - ADM 201 and DEV 401 (Platform App Builder)
AWS Solution Architect Associate Certification Training
Deep Learning Course with TensorFlow

One-on-One Training
Data Science in R Programming
Hadoop Administration
NoSQL Databases for Big Data
Salesforce Certifications - ADM 201 and DEV 401 (Platform App Builder)

Free Courses
Introduction to Data Science in Python
Java for Beginners by John Purcell

About DeZyre

About Us
Contact Us
FAQ
Pricing
Mini Projects
Online Hackathons
DeZyre Reviews
Blog
Tutorials
Webinar
Student Portfolios
Privacy Policy
Disclaimer

Connect with us