1. * 1. Learning the application domain: get familiar with the application
     2. Creating a target data set: data selection
     3. Data cleaning and preprocessing: data purification
     4. Data reduction and transformation: data purification
     5. Choosing functions of data mining: depends on the goals of mining
     6. Choosing the mining algorithm(s): suitable algorithms can be better and faster
     7. Data Mining and Analytics: search for patterns of interest
     8. Pattern evaluation and knowledge presentation: show the result
     9. Use of discovered knowledge: use the knowledge
     10. Statistics: Statistics is at the core of sophisticated machine learning algorithms, capturing and translating large amounts of structured and unstructured data patterns into actionable evidence.
     11. Mathematics: The concepts of mathematics especially in Linear Algebra and Calculus are useful in identifying trends and patterns in data and thereby assist in the creation of algorithms.
     12. Coding: A thorough understanding of data structures and algorithms is necessary to create efficient code that can analyze large sets of data.
   1. Supervised learning: Supervised learning is a purpose-based training method, it needs to label data and you know what you get. It usually be used to regression and classification.
   2. Unsupervised learning: Unsupervised learning is a training method with no clear purpose, it doesn’t need to label data. You can't know in advance what the result is. It usually be used to clustering (simply put, it is an automatic classification method) and dimensionality reduction (to reduce the complexity of the data while keeping the relevant structure as much as possible).
   3. 1. Root Node: The first node of the tree that contains all data in dataset.
      2. Leaf Node: A root that does not have any children.
      3. Decision Node: A node that has two or more children.
      4. Splitting: The process of partitioning the data in a decision node. Each partition gets sent to a child of the decision node. Splitting method is a statistical method that chooses a feature (field) of data and calculates splitting threshold for the chosen feature.
      5. Pruning: The process of removing a decision node and all its sub-tree.
      6. C4.5 and CART
      7. Differences: C4.5 developed by computer scientists reduces the information disorder using entropy. CART developed by statisticians reduces the statistical disorder using Gini impurity. CART must be a binary tree, C4.5 can be a multinomial tree.
   4. Accuracy: ratio of valid answers to total number of data rows.
   5. Confusion Matrix: The matrix that represents the performance of a model with True Positives (number of positive cases that the model answers correctly), False Positives (number of negative cases that the model returns as positive), False Negatives (number of positive cases that the model returns as negative), True Negatives (number of negative cases that the model answers correctly).
   6. Out of sample data or test data is a dataset that the supervised learning model hasn't seen and used to test the performance of it by used these two methods in conjunction.
   7. 1. Underfitting is a model’s accuracy on training data is low.
      2. Overfitting is a model’s accuracy on training data is high but accuracy drops on out of sample (test) data.
   8. Bias-Variance Tradeoff is a central problem in supervised learning paradigm. High Bias Model means underfitting. High Variance Model means overfitting. Ensemble methods are used to balance the bias and variance.
   9. Ensemble method: Train multiple models and then combine their results to have better predictive performance. Two ensemble methods: Bagging and Boosting.
      1. Bagging: Random Forest is a supervised learning algorithm that uses bagging technique to create an ensemble of decision trees. It has two steps: First, sample with replacement from X and Y, called Xb and Yb. Second, train a decision tree with feature bagging on Xb and Yb.
      2. Boosting: GBM is a supervised learning algorithm that uses boosting technique to create an ensemble of decision trees. First, given Model Fm, create a new training set with output values calculated based on the difference between outputs of Fm and correct outputs. These differences are called pseudo-residuals. Then a new model Hm that minimizes the pseudo-residuals is trained. The result model of this step is F(m+1) = Fm + Hm.