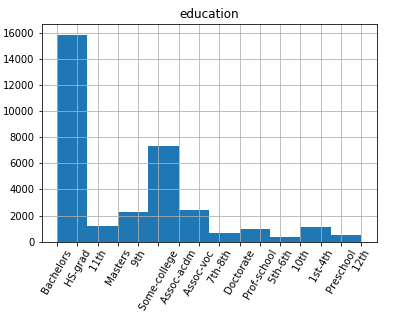
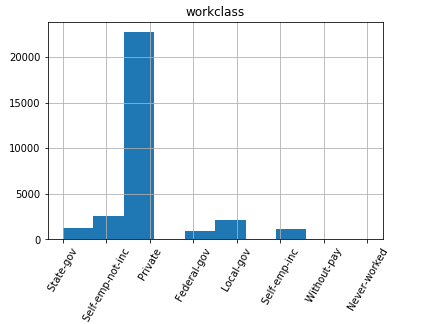
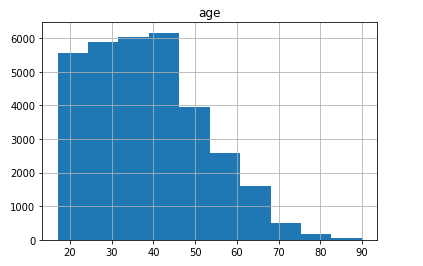
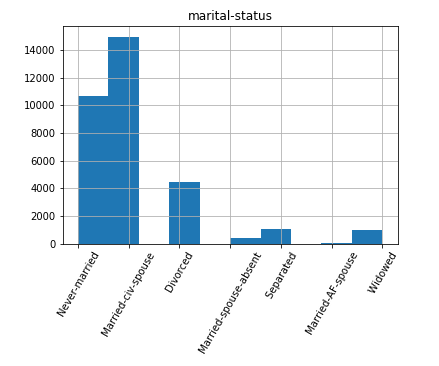
# Introduction

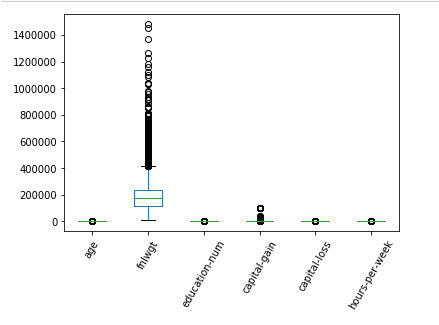
The algorithm should be of a similar level of sophistication as the ID3 decision tree building algorithm. Machine learning stems from artificial intelligence and statistics. After two decades of active and calm periods, the first International Forum on Machine Learning, held in 1980, marked the full recovery of machine learning. Machine learning has begun to affect all walks of life in civil society, such as self-driving in the consumer sector. The data was extracted from the US 1994 Census database and can be used to predict whether household income exceeds 50K$/year. The data set class variable is whether the annual income exceeds 50k$. The attribute variable contains important information such as age, job type, education, occupation, and ethnicity. It is worth mentioning that there are 7 category variables among the 14 attribute variables.

# Exploration

The data set characteristics are multivariate. It contains 48,842 instances. Data is from American society. The data types are categorical and integer. There are 14 attributes in total. And it contains missing values. The following picture shows our EDA analysis of the data.







We plan to use a variety of machine learning models to fit the data. Such as ID3 decision Tree, KNN, and random forest algorithms, etc.

The main algorithm of symbolism is the Decision Tree. As a kind of method that imitates the human brain to deal with decision-making problems in daily life, the decision tree has a binary tree structure like the two classification tasks such as “whether” or “good or bad”. In the process of getting the conclusion of the problem, that is, the final decision, all the accumulation of the judgments of the sub-decisions of each feature in the data makes the scope of the solution shrink. Generally speaking, a complete decision tree consists of a root node containing data and feature ensembles, several process nodes that represent feature decisions, and corresponding leaf nodes that represent decision results. Among them, one of the core problems in decision tree learning is feature partitioning. Three classical partitioning methods form three representative algorithms for decision trees: ID3, C4.5, and CART. The ID3 algorithm is a greedy algorithm proposed by J. Ross Quinlan in 1975 to construct a decision tree. It is based on the "Occam razor", that is, the smaller the decision tree, the better the decision tree. According to the feature selection and the information gain evaluation in the ID3 algorithm, the feature with the largest information gain is selected as the branching criterion each time. ID3 algorithm can segment nominal data sets without pruning. To eliminate excessive data matching, adjacent leaf nodes that do not generate a large amount of information gain can be merged (for example, by setting an information gain threshold). One disadvantage of using information gain is that it favors properties with a large number of values -- that is, in a training set, the more a property accepts a different value, the more likely it is to use it as a segmentation property. Sometimes it makes no sense. In addition, ID3 cannot handle continuously distributed data characteristics.

Occam razor illustrates a concept of information entropy:



This ID3 algorithm can be to sum up as follows: Using all unused features and calculate the sample entropy associated with it. Attributes with the smallest entropy will be left behind. Generating a node with this attribute.

The KNN algorithm is a famous artificial intelligence statistical learning algorithm, which accounts for a large proportion in machine learning, which is one of the simplest machine learning algorithms simultaneously. The K Nearest Neighbor algorithm, also known as the KNN algorithm, is a classic distance classifier algorithm in machine learning.

There are supervised learning methods and unsupervised learning methods in machine learning field according to different training samples. Supervised learning, also known as teacher learning, requires knowledge of which category each sample in the training sample set belongs to; unsupervised learning, also known as non-teacher learning, only knows each training sample in the training sample set, I don't know the category to which each training sample belongs, and even in some cases the number of categories is unknown. Distance classifiers are the primary method in supervised learning methods that rely primarily on the distance (or degree of similarity) between patterns to construct classifiers, sometimes referred to as "template matching." The distance classifier is characterized by simple algorithm and high efficiency of classifier learning and recognition. It is still one of the main methods to solve the identification problem, especially the application scenarios where the number of certain categories is relatively large and the recognition speed is relatively fast.

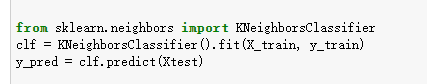
The representative of unsupervised learning is the cluster analysis method. These methods still use the distance to measure the similarity between the patterns, and then divide the different subsets according to the similarity of the samples in the unsupervised sample set to achieve unsupervised samples. Clustering.

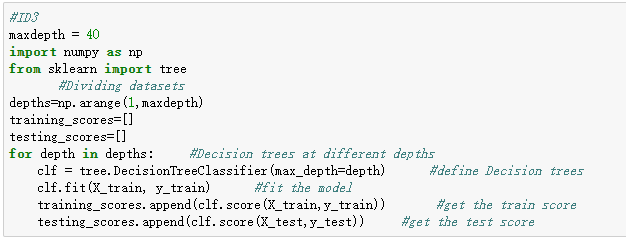
Random forest is an integrated algorithm (integrated learning), belonging to the bag type. By combining several weak classifiers and voting or averaging the final results, the results of the whole model have higher accuracy and generalization performance. It achieved satisfactory results mainly because of "randomness" and "forest", which made it resist overfitting and which made it more accurate.

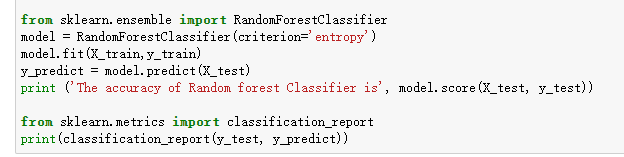
Bagging, also known as bootstrap aggregating, is an integrated technique for training classifiers by re-selecting k new data sets on the original dataset. It uses the set of trained classifiers to classify the new samples, and then uses the majority vote or the mean of the output to count the classification results of all classifiers. The highest category is the final tag. Such algorithms can effectively reduce bias and reduce variance. Random Forest (RF) is a Bagging algorithm. In fact, after the introduction of the Bagging algorithm, the random forest is almost ready. RF is just some details about the details.

# Methodology

We first use several pre-designed machine learning models to fit the data. And divide the age into 4 parts according to 'Teenager', 'young man', 'elder', 'old man'. Then we use the train\_test\_split in sklearn.model\_selection to separate the training set from the test set. Because the dataset contains data of many category types, we use LabelEncoder in sklearn.preprocessing to convert categorical data into numeric data and increase the data dimension.

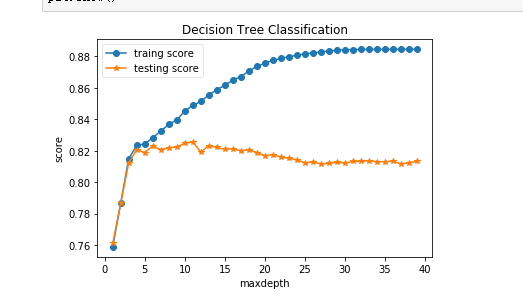




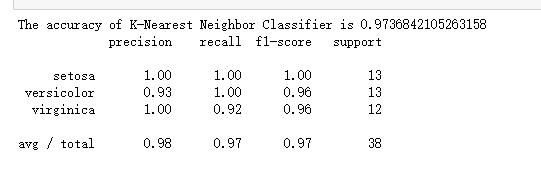


# Evaluation

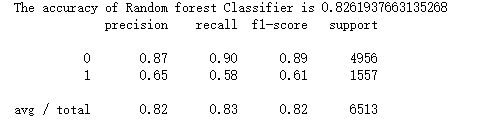
Finally, we get the results of the data on each model and the training test curve. The following figure shows the training test score chart of the ID3 algorithm.



The following figure shows the model evaluation of the KNN algorithm.



The following figure is a model evaluation of random forest algorithm.



# Conclusion

From the above experimental data, we can see that the random forest has many shortcomings. Due to the increased disturbance of the attributes, the performance of the base learner in the random forest is reduced, which makes the performance of the random forest at the beginning poor, but with the base learner Increased, random forests usually converge to lower generalization errors, compared to Bagging. It is sensitive to the number of attributes considered for each partition. The random forest approach has proven to be very good for large data sets and items that have large and sometimes unrelated features. It is easy to overfit when the noise is large. Of course, random forests have many advantages. Reduce over-fitting: the introduction of two randomness makes it difficult for random forest to fall into over-fitting and has good anti-noise ability. Since only a few attributes are considered in each partition, it may be faster than Bagging and Boosting, so they are very effective on large databases. So compared to Bagging, the computational cost is smaller and the training efficiency is higher. Data preprocessing: adaptability to data, can handle discrete and continuous, no need to standardize. Missing value processing: There are good ways to fill in missing values and still maintain high accuracy even if a large portion of the data is missing. An intrinsic estimate of the importance of the variable is given, which balances the error for unbalanced sample classification. Can calculate the intimacy of each instance, useful for data mining, detecting outliers, and data visualization.

Of course, the KNN model also has many shortcomings. The computational complexity of the classification is proportional to the total number of samples in the training set. The algorithm itself is simple and effective, and the classifier does not need to use the training set for training. For points of unknown category, calculating its distance from points in the known category dataset

Sort by distance

Select the K points with the smallest distance from the current point

Determine the probability of occurrence of the category of the first k points

Returns the category with the highest frequency of the first k points as the current point prediction classification

Significant shortcomings of the ID3 algorithm are below:

Overfitting: Over-fitting problems are prone to occur, resulting in poor prediction of data outside the training set

Ignore feature associations: Ignore the correlation of attributes in a dataset. When there are many intricate relationships between numerical variables, such as financial data analysis, it is not a good choice.

Model sensitivity: The model is not robust enough, and small changes in one node may cause the entire tree to be very different.

So as the old saying goes, the quality of the model depends to a large extent on the quality of the feature engineering. The feature engineering works well and the model fits the data better.