通过观察数据集，我们发现正负样本的数量存在比较大的差异，比例大约在4:1。而对模型而言，不均衡数据构建的模型会更愿意偏向于多类别样本的标签，实际应用价值较低。因此为了规避数据不平衡带来的风险，我们寻找了一些常见解决数据不平衡的方法，包括欠采样和过采样方法。由于本数据集总样本数量较少，因此我们最后决定尝试采用过采样的SMTOE、Borderline SMOTE和ADASYN来进行数据集的处理。

**SMOTE**（Synthetic Minority Oversampling Technique）是在随机[采样](https://so.csdn.net/so/search?q=%E9%87%87%E6%A0%B7&spm=1001.2101.3001.7020)的基础上改进的一种过采样算法。SMOTE实现简单，但其弊端也很明显，由于SMOTE对所有少数类样本一视同仁，并未考虑近邻样本的类别信息，往往出现样本混叠现象。

**Borderline SMOTE**是在SMOTE基础上改进的过采样算法，该算法仅使用边界上的少数类样本来合成新样本，从而改善样本的类别分布。其采样过程是将少数类样本分为3类，分别为**Safe**、**Danger**和**Noise**，具体说明如下。最后，仅对表为Danger的少数类样本过采样。

**ADASYN** （adaptive synthetic sampling）与Borderline SMOTE相似，对不同的少数类样本赋予不同的权重，从而生成不同数量的样本。

By observing the data set, we found that there is a relatively large difference in the number of positive and negative samples, with a ratio of about 4:1. As far as the model is concerned, the model constructed with unbalanced data will be more willing to be biased towards the labels of multi-category samples, and the actual application value is low. Therefore, to avoid the risks caused by data imbalance, we have found some common methods to solve data imbalance, including under-sampling and over-sampling methods. Due to the small number of total samples in this dataset, we finally decided to try to use over-sampled SMTOE, Borderline SMOTE and ADASYN to process the dataset.

SMOTE (Synthetic Minority Oversampling Technique) is an oversampling algorithm improved based on random sampling. SMOTE is simple to implement, but its disadvantages are also obvious. Since SMOTE treats all minority samples equally and does not consider the category information of neighboring samples, sample aliasing often occurs.

Borderline SMOTE is an improved oversampling algorithm based on SMOTE. This algorithm only uses the minority class samples on the border to synthesize new samples, thereby improving the class distribution of samples. The sampling process is to divide the minority class samples into three categories, namely Safe, Danger and Noise, as described below. Finally, only the minority class samples whose table is Danger are oversampled.

ADASYN (adaptive synthetic sampling) is like Borderline SMOTE, assigning different weights to different minority samples, thereby generating different numbers of samples.

Random forests are a type of machine learning algorithm that can be used for classification and regression tasks. The random forest algorithm works by creating a large number of decision trees, each of which is trained on a random subset of the training data. The predictions made by the individual decision trees are then combined to make the final prediction, which is typically more accurate than the predictions made by any individual tree.

One of the key advantages of random forests is that they are able to handle large and complex datasets, and they can also provide estimates of the importance of each feature in the dataset. This makes them a popular choice for many applications, including image and text classification, as well as predictive modeling. In addition, the random forest algorithm has good anti-noise ability and can run efficiently on large data sets. Especially in the classification of breast cancer data sets, it performs well, with high accuracy and fast training speed. In addition, there are not many parameters in the random forest, and it is easy to find the best parameters.

Although the random forest algorithm is fast enough, when the number of decision trees in the random forest is large, the space and time required for training will be large, which will cause the model to slow down. Therefore, in practical applications, if you encounter a situation with high real-time requirements, it is best to choose other algorithms.

As for other model algorithms, such as logistic regression, ANN and SVM, there are generally problems such as low accuracy and too many parameters to be adjusted in the kernel function, especially in the classification of breast cancer data sets, the performance is not ideal, so we use Randomforest for training.