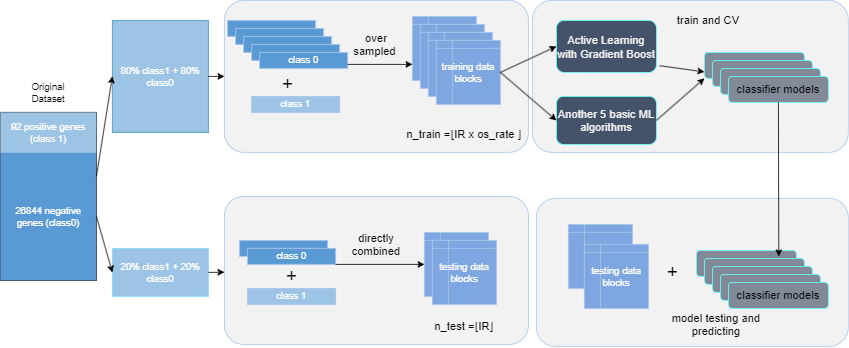
During the work *predicting genes related to RNA methylation pathways,* we found the data imbalance problem in the original dataset is prominent. With imbalance factor (IR) of over 200, which is higher than the imbalance factor among most public datasets. We believe this problem is solvable, we tried Active Learning, data oversampling algorithm SMOTE, and basic machine learning classifiers (SVM, GB, GNB, RF, LR).

**Data collecting and preprocessing**:

Original dataset collected from Harmonizome website, 26936 genes included in, each gene owns 1517 features. The value of each feature is discrete 0, 1, -1. We initially picked 92 positive genes out (known genes associated with RNA methylation pathways) while the other 26844 genes are assumed to be negative samples. This is because most genes in the human body serve other functions, which means only few genes are related to the RNA methylation pathway. Therefore, the number of false negatives in training data should be very small.

We took 80% positive and negative samples for training and cross validation，other 20% data remaining for testing. Next, calculated the imbalance factor IR, it is equal to the ratio of the number of majority classes to the number of minority classes in the training sample. Since the dataset’s IR exceeds most public datasets, reaching 290, it is necessary to oversample training data. We applied basic SMOTE algorithm to oversample. Assuming parameter OSRATE, a decimal number between 0 and 1 used to control the oversample rate. In our experiment, this parameter is set to 0.6. We uniformly divided the majority training samples (negative training samples) into NTRAIN = IR\* OSRATE (round down) pieces, combining each negative sample piece with the same positive sample to get NTRAIN copies of data blocks. Then each data block was balanced by SMOTE. Features in generated samples still get discrete value range of 0, -1, 1.

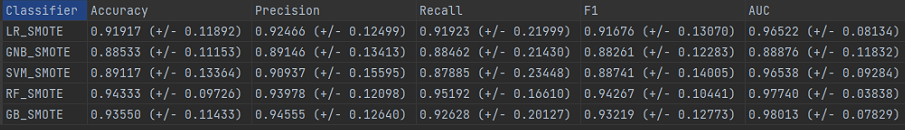
For test set, also calculate its IR, which should be equal to the training set’s. Then directly divide the test dataset into IR(rounded down) pieces, each data piece combined with the same positive test sample. Then, the test-set containing NTEST = IR data blocks is obtained. As shown in FIG 1.



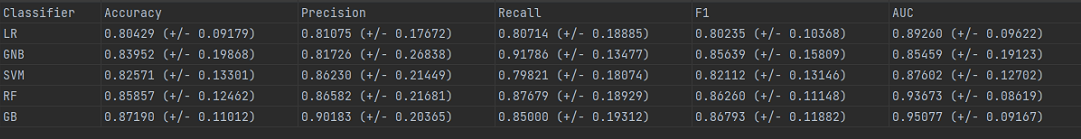
【FIG1】

**Model training and cross-validation**:

We applied active-learning algorithm (based on gradient-boost and uncertainty-sampling) and 5 machine-learning classifiers: SVM, GB, GNB, RF, LR. Trained on NTRAIN=173 data blocks. Notice: Hyperparameters in GB based active-learning model are default assigned while gird-search technique is used to select best (accuracy) hyperparameters among given parameter combinations in 3 machine-learning classifiers (SVM, GB, RF). As for cross-validation, active-learning method hits average accuracy 89.74% without SMOTE, reaching 91% accuracy after oversampling. GB hits 87% accuracy in ML methods (Figure 3) and this value reached 93% of best after oversampling (Figure 2). It should be noted that the final datasets obtained after oversampling are not the same between AL and ML, this is because the training scripts for AL and ML are different.



【FIG2】



【FIG3】