CPS844 Assignment 1

Xiaoxin Zhou 501072188
Chaoyu Wang 500815335

Introduction:
The dataset for this project is from the UCL Machine learning repository: Adult data set [1]. This data extraction was done by Barry Becker from the 1994 Census database. Used to prediction task is to determine whether a person makes over 50K a year or not. In our assignment, we will try five or more different classification methods to predict the salary and compare the results of the methods. This project will use the For Loop method to find the most important attributes for our classification method and use the best combination of attributes to classify the results.

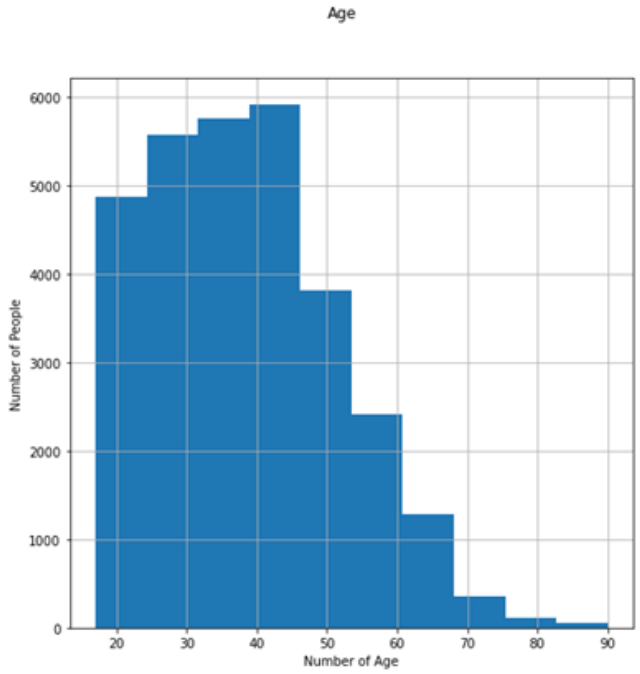
The Attribute information for the Adult data set:

- **age:** continuous.
- **workclass:** Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt:** continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num:** continuous.
- **marital-status:** Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation:** Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship:** Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race:** White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex:** Female, Male.
- **capital-gain:** continuous.
- **capital-loss:** continuous.
- **hours-per-week:** continuous.
- **native-country:** United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
- **Income:** continuous.

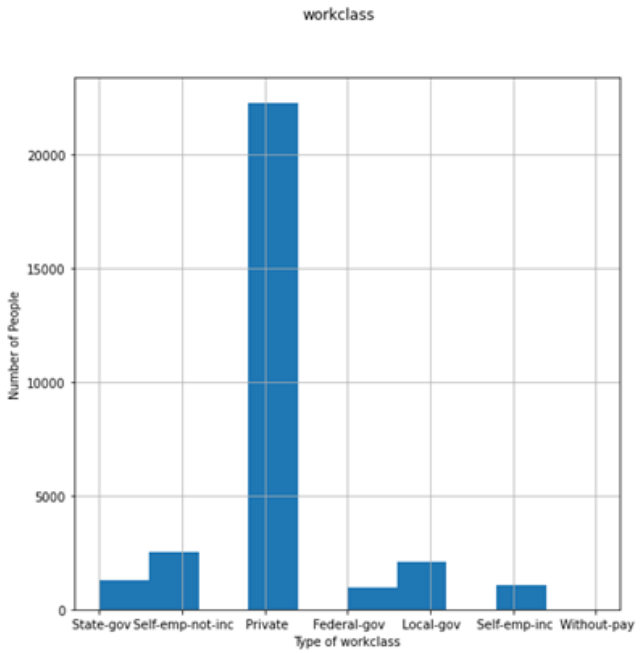
Visual Data:

The next step will be to visualize the Adult data set and what is included in the data set '?' Value and empty Value for conversion and deletion. Make the dataset less error-prone during training and testing.

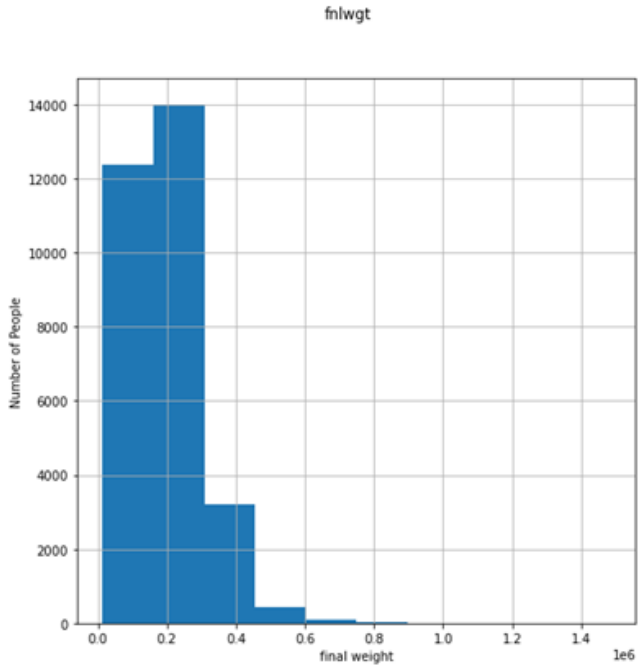
Age:



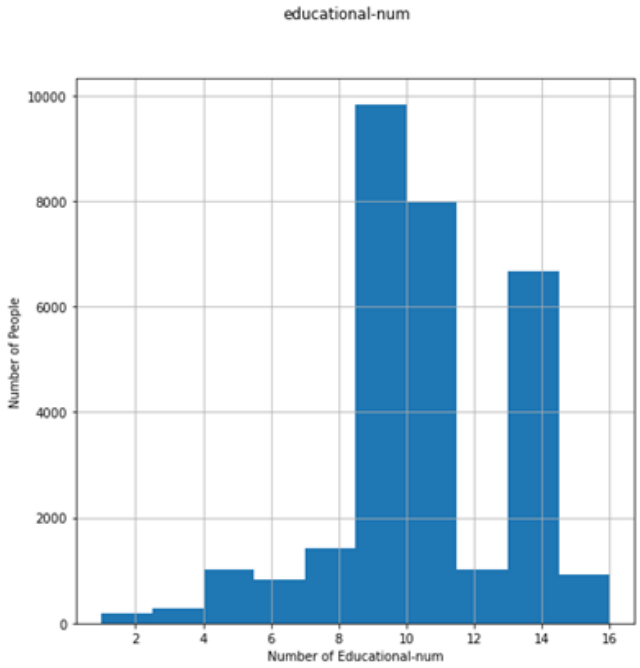
Workclass:



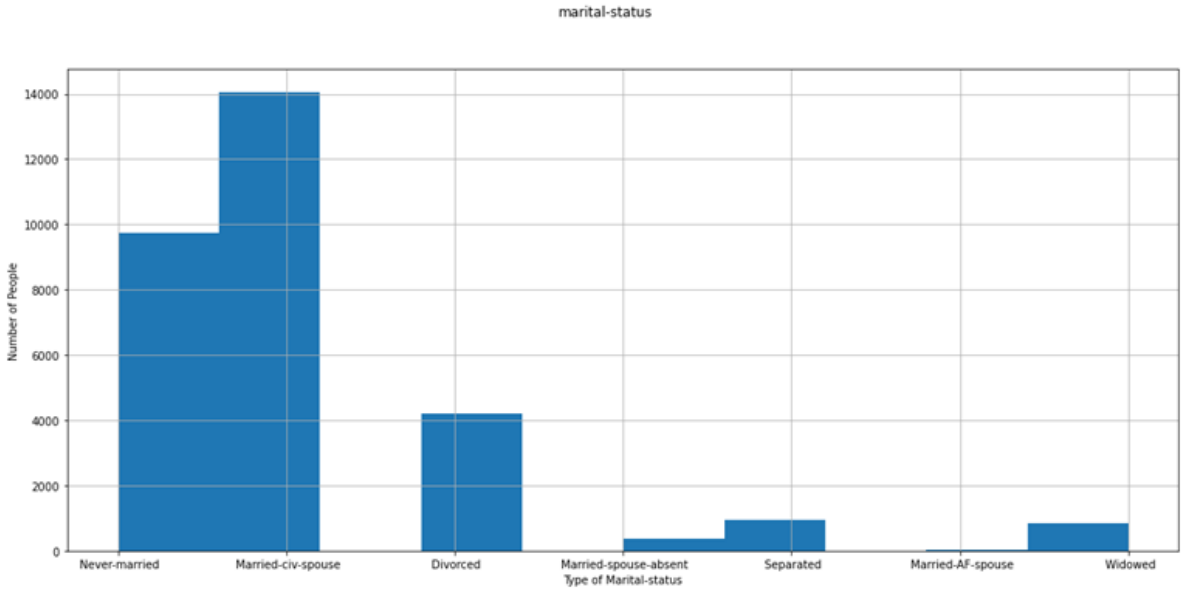
Fnlwgt:



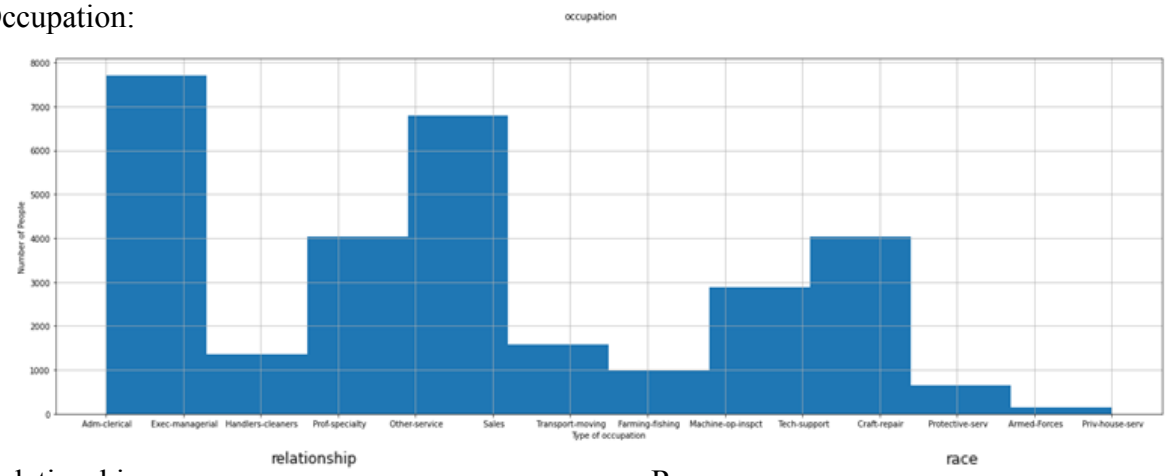
Education-num:



Marital-status:



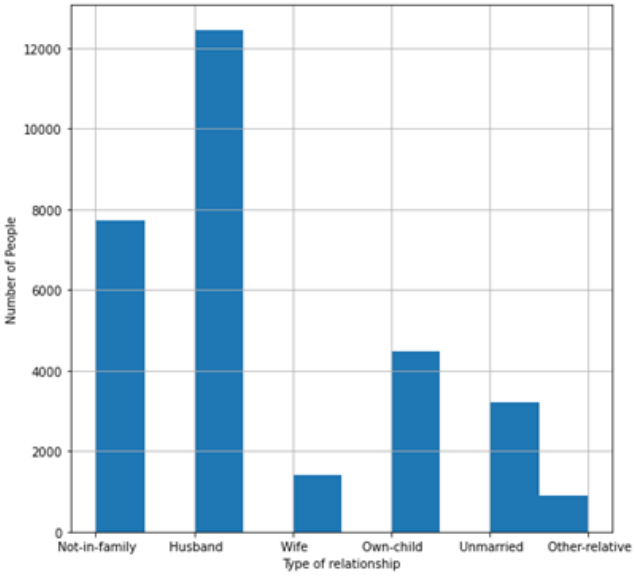
Occupation:



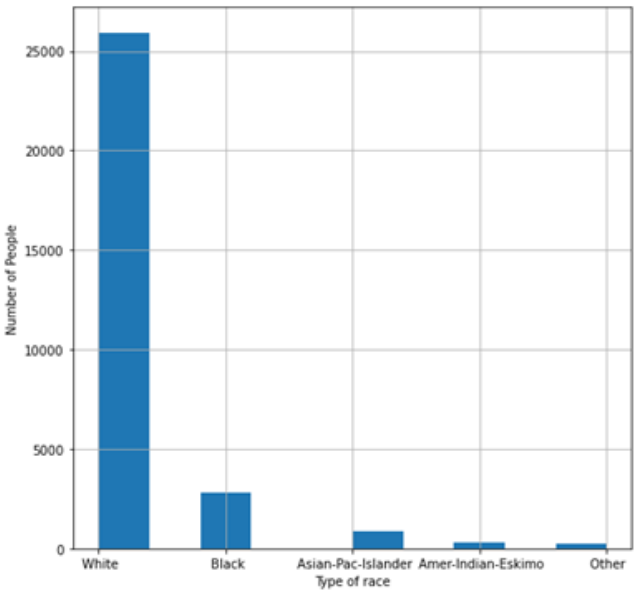
relationship

race

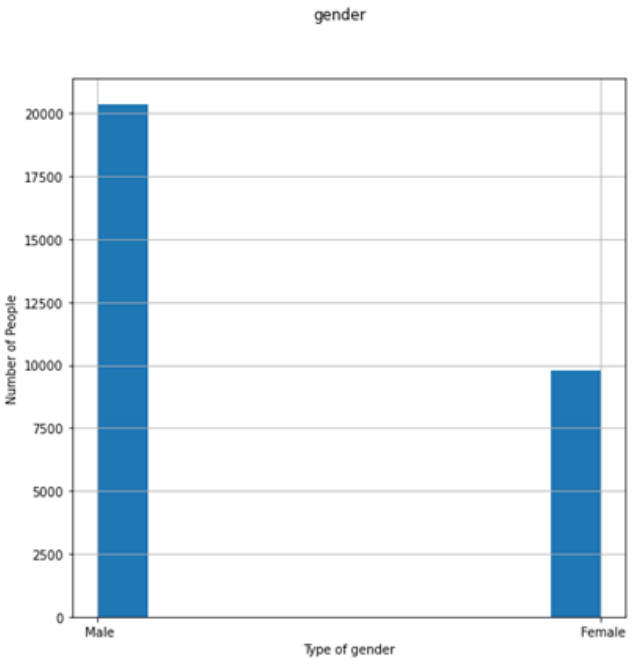
Relationship:



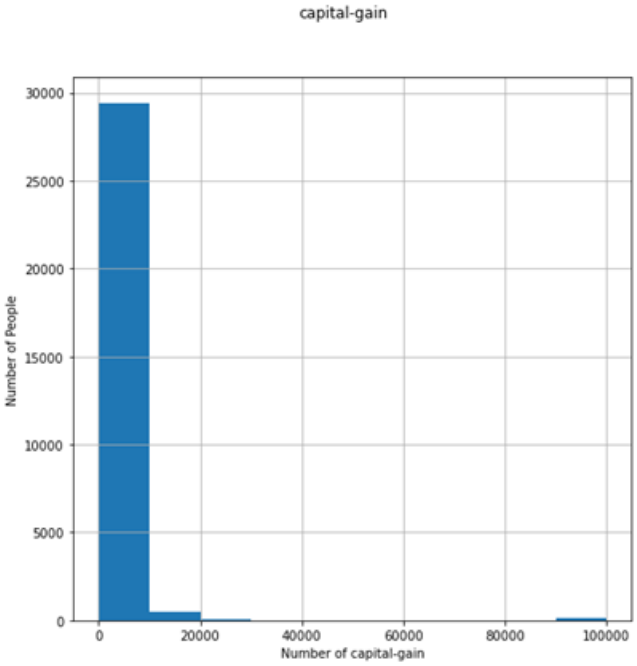
Race:



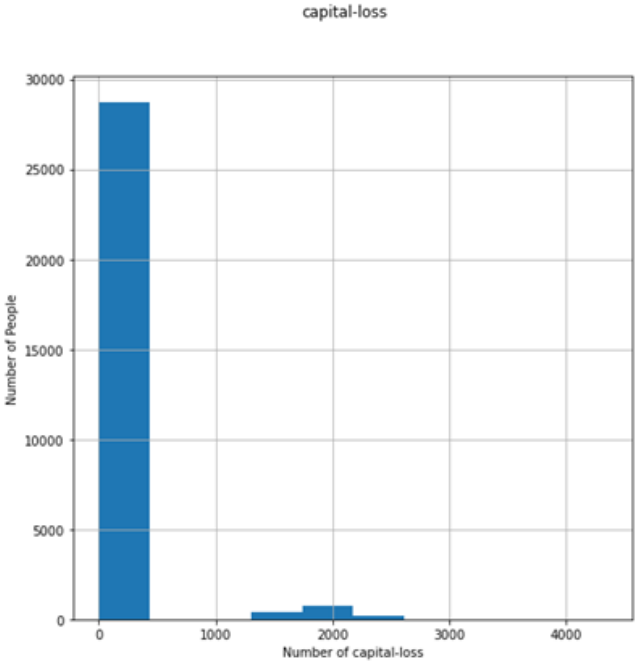
Gender:



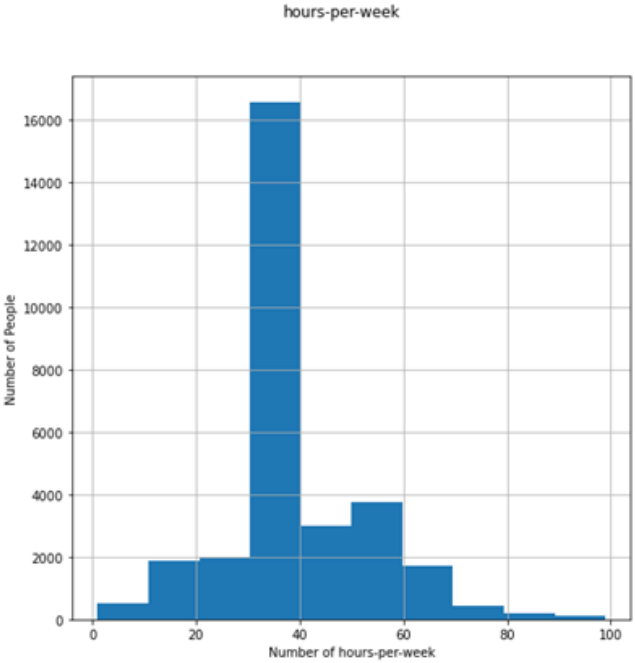
Capital_gain:



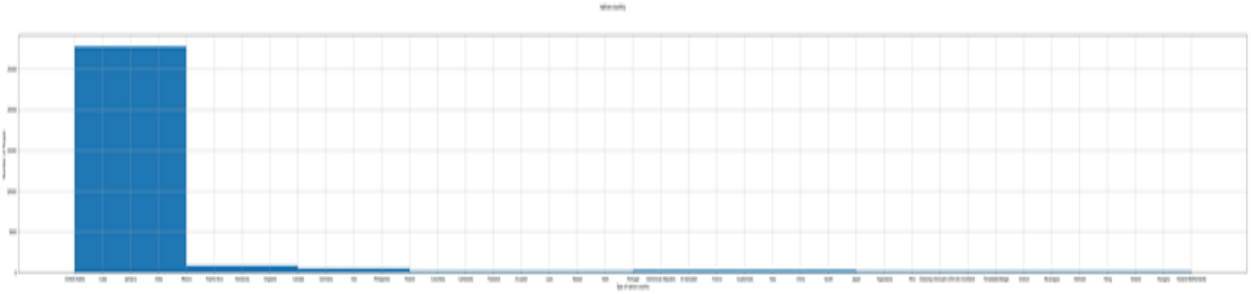
Capital_loss:



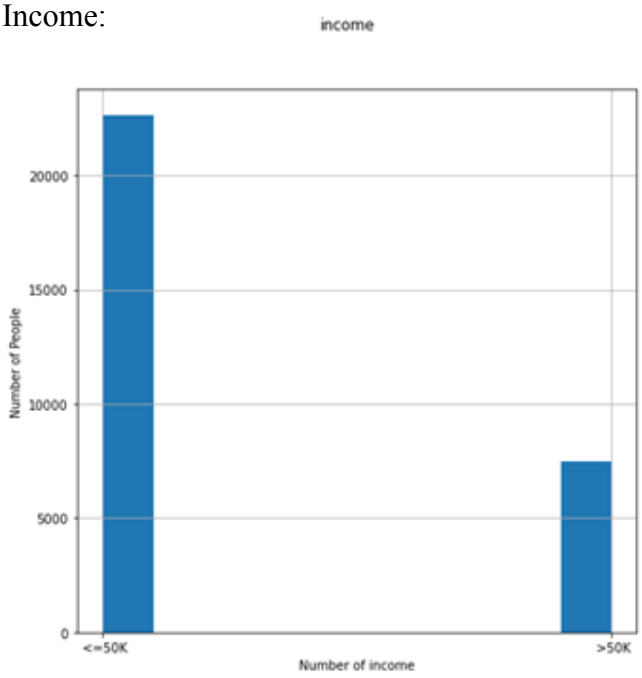
Hours-per-week



Native-country:

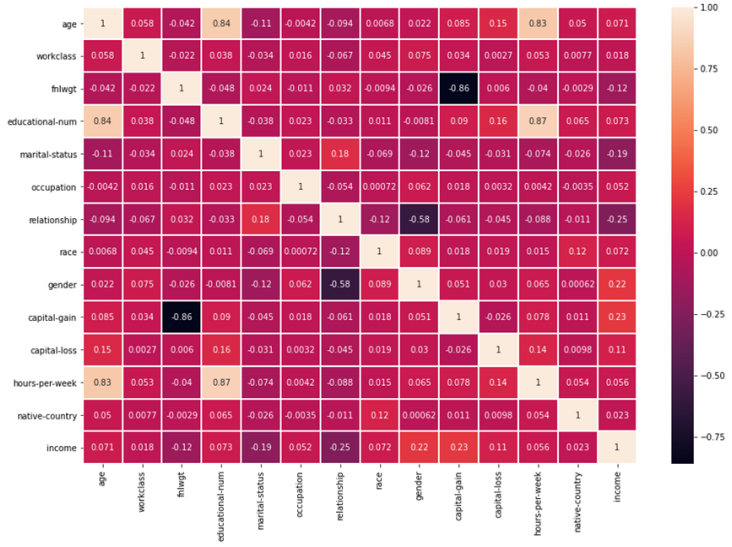


Income:



Heatmap and Normalized:

Use heatmap to further visualize 14 complex data with different colours. Data were normalized using methods from ML101 k-NN before heatmap [2]. The Heatmap shows that few data have a strong relationship (we set >|0.70| as a strong relationship).



Data processing:

The project removes income from the data and uses it as the Y for testing, and uses the remaining data as the initial X. All tests for this project split the data for 80% training and 20% for testing.

The best combination of the data X:
This step uses For Loop and itertools.combinations to combine data X (which has drop income) and bring it into Log Regression for classification. Thus find the best combination as a new X and start retesting with five different classifiers. By testing the best combination of the data X is ['age', 'workclass', 'educational-num', 'marital-status', 'relationship', 'gender', 'capital-gain', 'capital-loss' , 'hours-per-week', 'native-country'] (Drop ['fnlwgt', 'occupation', 'race','income'])
The following tests use the best combination of the data X

Five different classification methods:
The project uses five different classification methods, with different parameters for all algorithms to get the best result.

The first is Logistic regression, we have changed the solver of Logistic regression to ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']. After testing lbfgs got the best results. But the output gap between different solvers is not big. It is worth mentioning that newton-cg will fail to fit in some combinations, and will not be able to output results.

- The accuracy of the Logistic Regression newton-cg is 0.7745731808387204
- The accuracy of the Logistic Regression lbfgs is 0.7747389358528095
- The accuracy of the Logistic Regression liblinear is 0.7734128957400961
- The accuracy of the Logistic Regression sag is 0.7742416708105421
- The accuracy of the Logistic Regression saga is 0.7730813857119178

The second method uses Gaussian Naive Bayes for classification.

- The accuracy of Gaussian Naive Bayes is 0.7898226421349246

The third method uses Quadratic Discriminant Analysis for classification.

- The accuracy of Quadratic Discriminant Analysis is 0.7957898226421349

The fourth method uses Support Vector Classification, We transform the SVM kernel, which includes ['linear', 'poly', 'rbf', 'sigmoid']. One of the best is linear. Among them, the poly cannot be fitted and the program has been running and cannot be stopped, so it can be speculated that the poly kernel of the SVM type is not suitable for this data set.

- The accuracy of Support Vector Classification linear is 0.7705950605005801
- The accuracy of Support Vector Classification rbf is 0.7614785347256755
- The accuracy of Support Vector Classification sigmoid is 0.7079396651748715

The fifth method is Random Forests, we have adjusted the features, which are ['auto', 'sqrt', 'log2']. The results obtained by this method are the best results among the five methods. However, in this method, different features did not lead to particularly large differences in the results, they all floated around 0.81. But log2 is the best result of this method.

- The accuracy of the Random Forest Model auto is 0.8160119343610144
- The accuracy of the Random Forest Model sqrt is 0.8165091994032819

- The accuracy of the Random Forest Model log2 is 0.8178352395159953

Through the above research, the random forests method characterized by log2 obtained the best result in this project is 0.8178352395159953.

Voting Classifier:

In this step, VotingClassifier[3] is used to put five classifiers into this module for voting classification. And use hard and soft two methods for voting classification to compare the results of the five classifiers to find the best method. The Hard Voting Classifier determines the final result based on the minority obeying the majority, while the Soft Voting Classifier uses the average value of the probability that all models predict a sample to be a certain category as the standard, and the corresponding type with the highest probability is the final prediction result. All five classifiers in this method use the optimal parameters derived above. The only difference is that the probability needs to be changed to True for the SVM in the soft voting classifier.

- The accuracy of the Hard Voting classifier is 0.7946295375435107
- The accuracy of the Soft Voting classifier is 0.7941322725012432

It can be seen from the results that although the algorithms of the two methods are different, the soft and hard voting classifiers also do not have an excessive impact on the output. In contrast, the Hard voting classifier has better results.

Conclusion:

The project uses five methods of Logistic regression, Gaussian Naive Bayes, Quadratic Discriminant Analysis, Support Vector Classification and Random Forests for testing, and uses Voting Classifier for additional testing. Prior to this, we expected that Voting Classifier might achieve better scores than the other five methods, but this is not the case from the results. So Random Forests got the best score in the processed data as 0.8178352395159953

Reference

[1] Kohavi,R & Becker,B. 1996. Adult Data set. UCL Machine learning repository.

https://archive.ics.uci.edu/ml/datasets/Adult

[2] Burak Celal Akyuz. 2021. ML101 k-NN. Kaggle.

https://www.kaggle.com/iambca/ml101-k-nn

[3] sklearn.ensemble.Voting.Classifer.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html