nName \_\_\_\_\_\_\_Mao, Keying Linxia Li\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_Dec 07, 2022\_\_\_\_\_\_\_

(last name, first name)

NYU ID: \_\_\_\_\_\_ km5580 ll4764\_\_\_\_\_\_\_\_

Course Section Number: \_\_DS-GA-2433-001\_\_\_\_\_

**Project Part 3**

**Total in points** (100 points total): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Professor’s Comments:**

Affirmation of my independent effort:

Keying Mao

Linxia Li

**1. Chronic disease prediction**

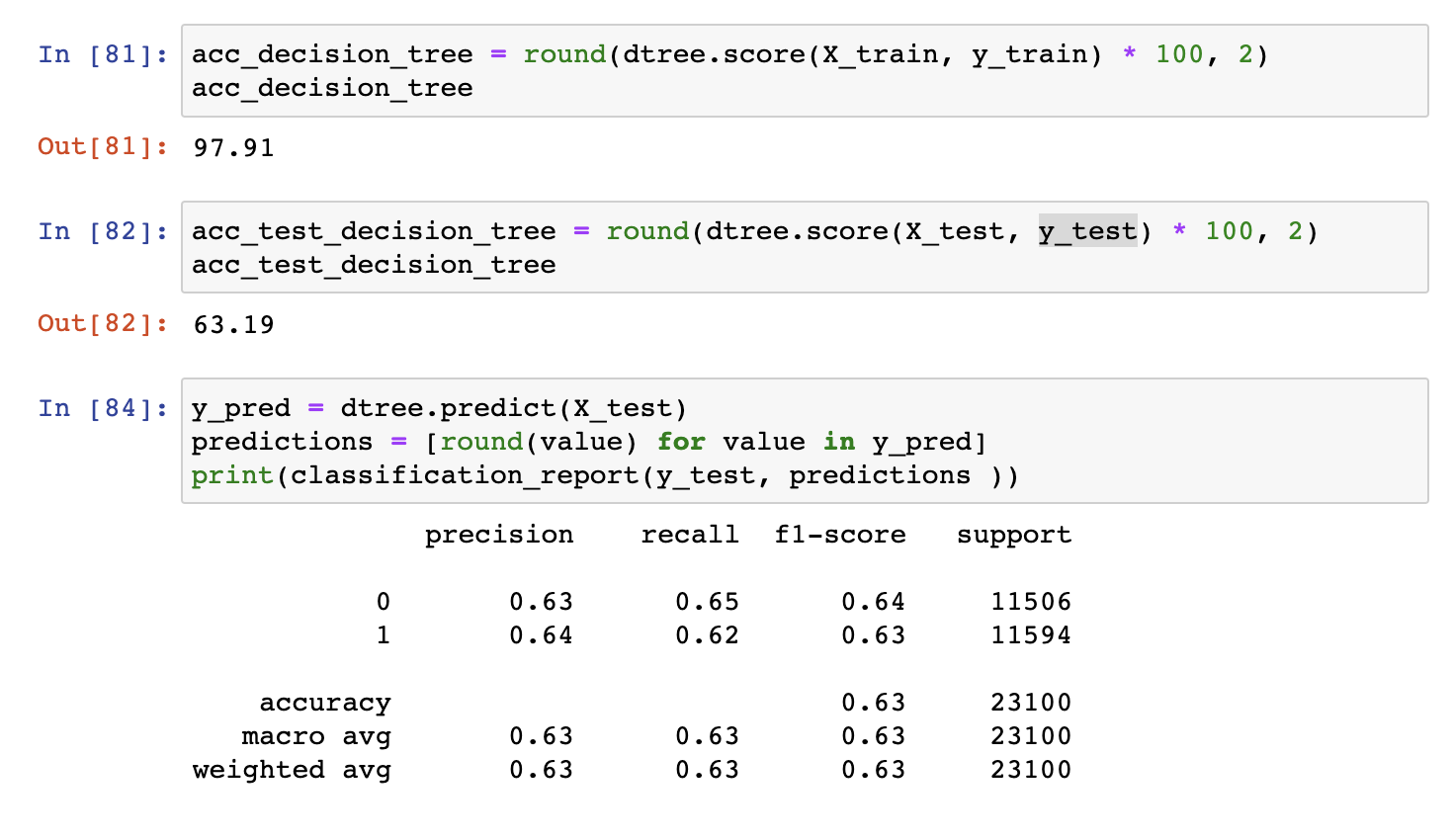
**1.1. Dataset Introduction & Pre-process**

We are using the chronic disease data from Kaggle called cardiovascular disease dataset. The link is <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>. Basically, there are factors such as age, gender, height, weight, cholesterol, glucose, smoke, alcohol, etc. that can affect cardiovascular disease. In order to forecast cardiovascular disease, the first thing we are cleaning and reformatting the data. Firstly, we dropped the unused column such as id. Secondly, we turned the age column from born days to the actual age by dividing 365.25. Then we change the gender column to dummies as 1 and 0 instead of 2 and 1. Finally, we get the categorical variables like cholesterol and gluc into dummy variables.

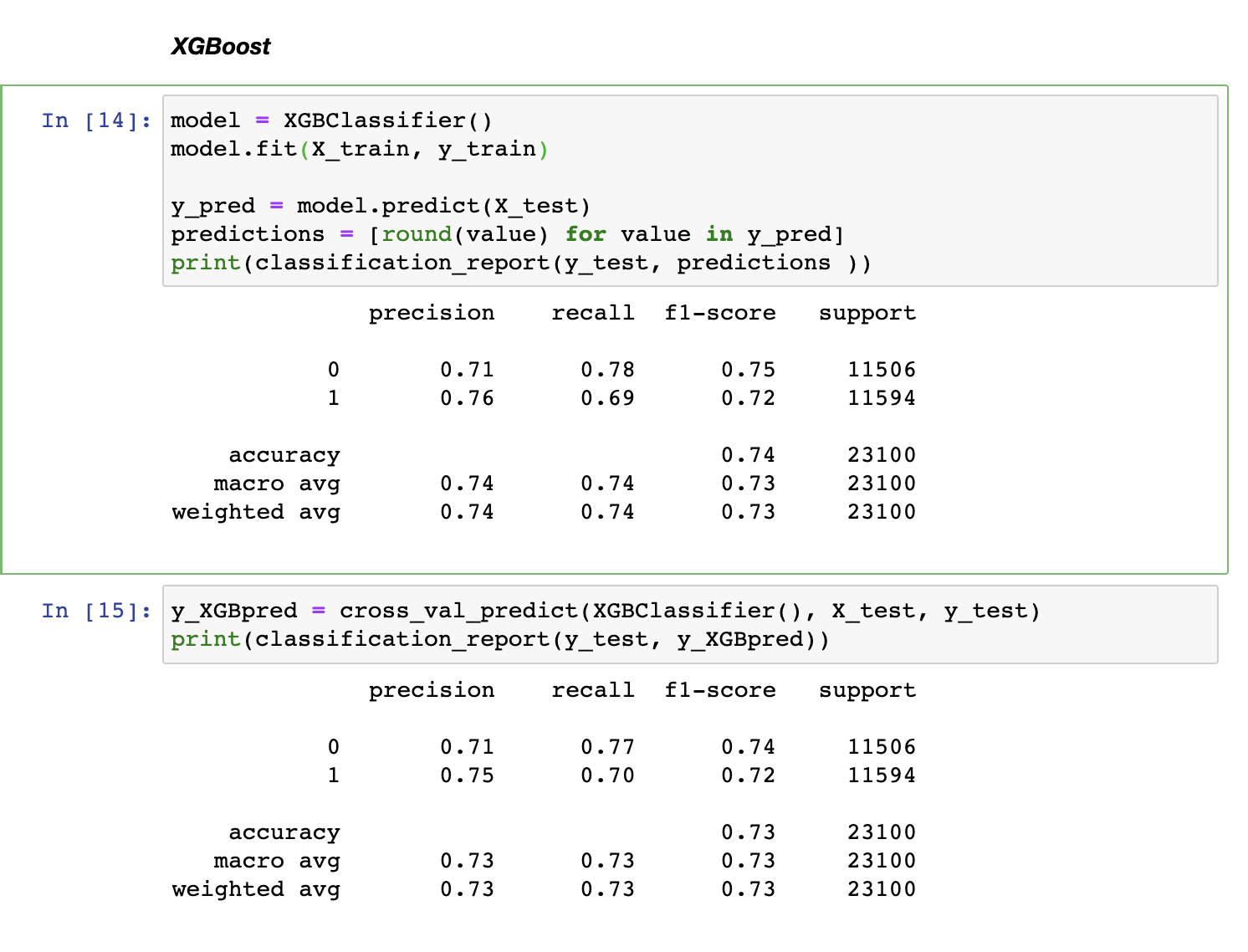
**1.2. Modeling & Result**

We are using machine learning algorithms to predict the probabilities that customers developing cardiovascular disease. So our target variable is cardio column and the rest of the columns would be the x variable.

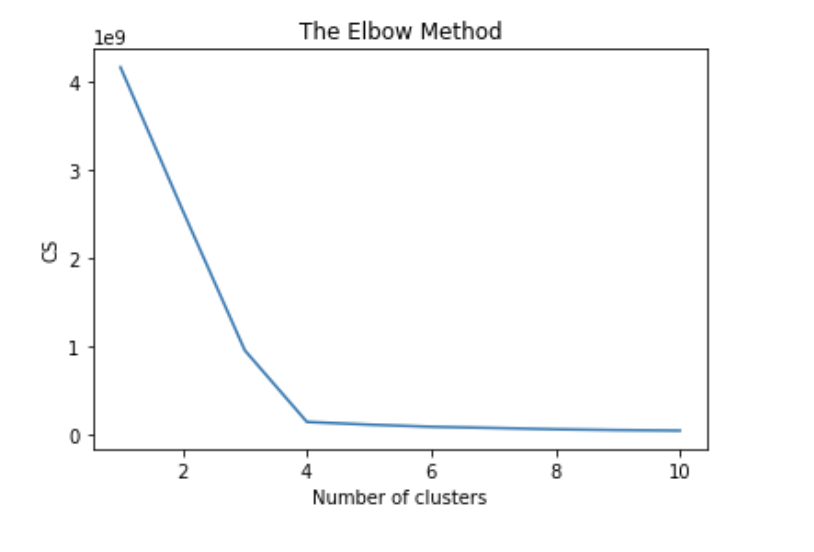
We trained and tested in three models: **decision tree, XGBoost, and K-means**. As for the evaluation of the decision tree model, we can see the accuracy for 1 is 0.64 and recall is 0.62.

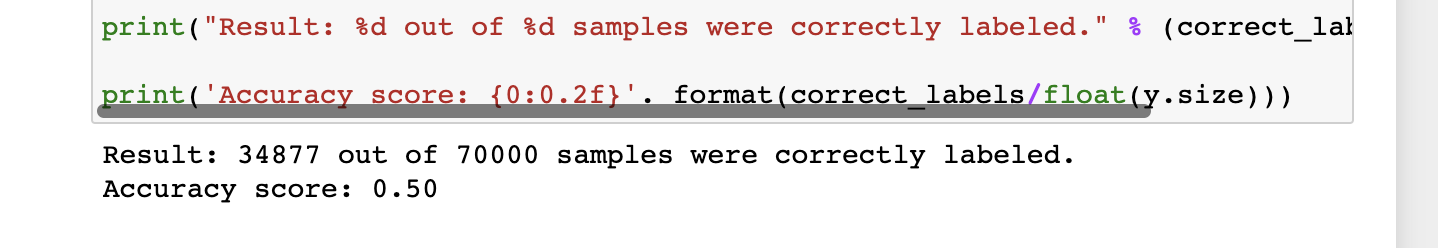


From the XGBoost classification report, before the cross validation, we can see the precision is 0.76 and recall is 0.69. But after cross validation, the precision for 1 drop 1% and recall for 1 increase 1%.



For Kmeans, I normalized the x variables so that each of the variables is between 0 and 1. Then I used the Elbow method to select the best cluster number for K-means. As a result, cluster = 4 would be the optimal number for data that may be clustered. The result shows that accuracy would be 0.50. And 34877 out of 70000 samples were correctly labeled.





**1.3. Impact on Business**

Compared to the above three models, we found that XGBoost performs the best to predict cardiovascular disease. In this way, when a customer wants to buy a cardiovascular-related insurance product, we could adjust a higher price when the model shows customers a sign of high risk of cardiovascular disease. For those people that have the sign of extremely high risk, the insurance company should further confirm their medical history and confirm whether their eligibility to apply for such product type.

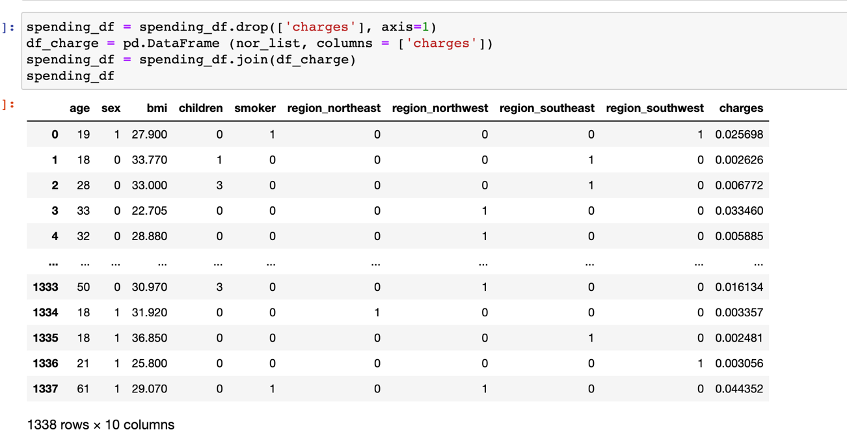
**2. Insurance Expense Prediction**

**2.1. Data Introduction & Pre-process**

We get the dataset “insurance.csv” from Kaggle with the link<https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset>. The dataset has 1338 observations and focuses on US people’s health insurance expenses and their personal information and status. The dataset includes features- age, sex, BMI, number of children, smoker (if smoke), and region, which may impact the Individual medical costs billed by health insurance- charges.

Some features are categorical with two categories, such as sex (female:1, male:0) and smoker(yes:1, no:0), then we encode them with 1 and 0. “Region” has more than 2 categories, so we got dummies with this feature and adopted binary values in it.

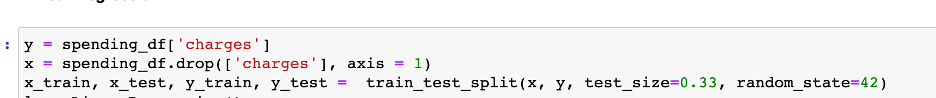
A wide range can be found in people’s health insurance expenses. In order to convenient subsequent data processing. We adopted normalization before building prediction models. The following graph shows the code and final dataset.



**2.2 Modeling and result**

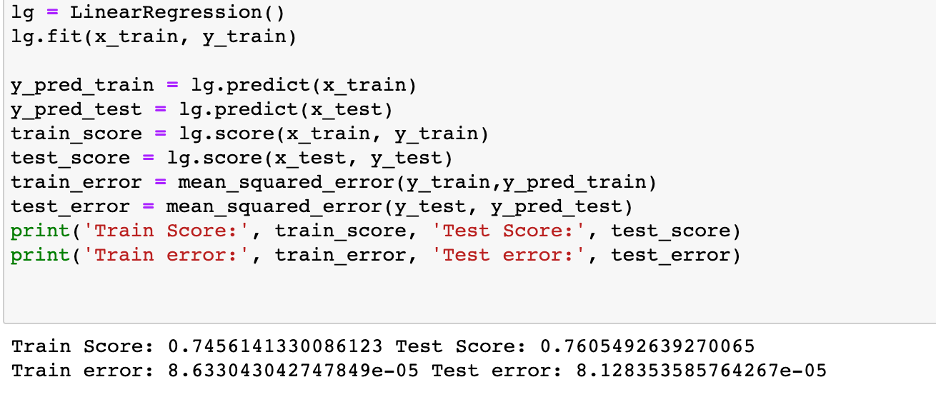
Our target is predicting the health insurance expense by adopting machine learning algorithms methods in features. First, we split the dataset into a training set and a testing set with a ratio of 0.33, the testing set will be used to evaluate the performance of each method. Second, we use Linear regression, Ridge Regression, and Lasso regression to train models. Finally, we choose R-square and MSE as metrics to evaluate regression performance. Linear Regression has the best performance since it has the highest R-square and lowest MSE in both training and testing.

**split train-test**



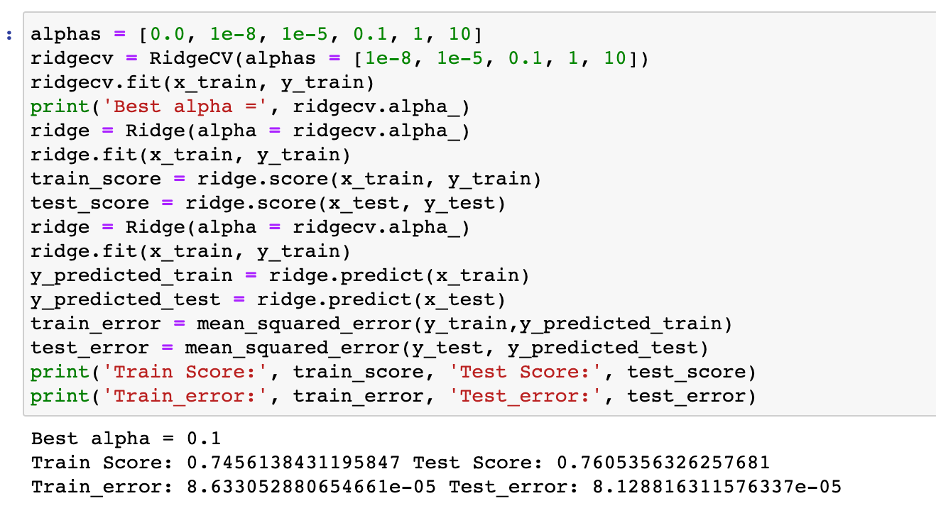
**Linear Regression**

Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. With Linear Regression, the prediction can cover 75% of observations in the training set and 76% in the testing set. MSE in the training set is 8.63\*e-5 while it is 8.13\*e-5 in the testing set



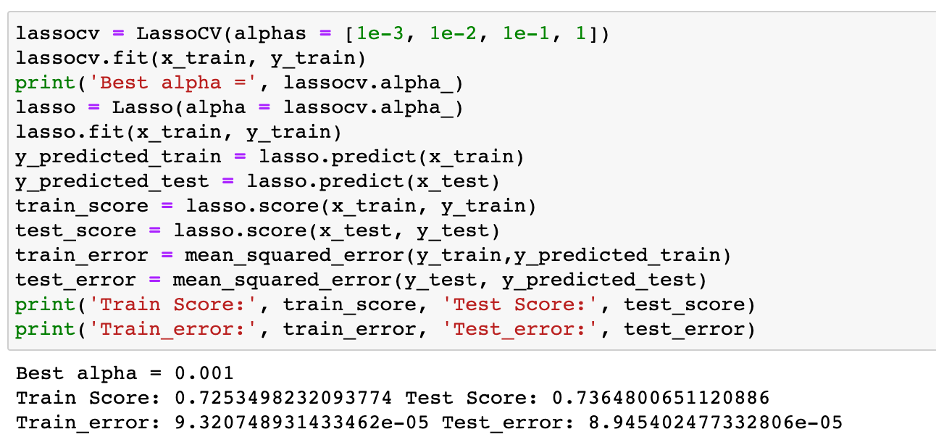
**Ridge Regression**

Ridge regression is a model-tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. With Ridge Regression, the best performance appears when alpha = 0.1. the prediction can cover 75% of observations in the training set and 76% in the testing set. MSE in the training set is 8.63\*e-5 while it is 8.1\*e-5 in the testing set



**Lasso Regression**

Lasso regression is a type of linear regression that uses shrinkage. Lasso regression performs L1 regularization. With Lasso Regression, the best performance appears when alpha = 0.001. the prediction can cover 73% of observations in the training set and 74% in the testing set. MSE in the training set is 9.32\*e-5 while it is 8.95\*e-5 in the testing set



**2.3 Impact on Business**

Linear Regression has the best performance since it has the highest R-square and lowest MSE in both training and testing. We will use it to predict insurance expenses in the future. Companies can use this model to predict a user's Individual medical costs billed by health insurance. For those with high charges, companies can take steps such as raising the price of insurance, reducing some benefits, or imposing restrictions.

**3. Google Cloud**

