Insurance **dataset**

Link in Kaggle:

<https://www.kaggle.com/datasets/mirichoi0218/insurance>

**Dataset Description:**

## **Context**

Machine Learning with R by Brett Lantz is a book that provides an introduction to machine learning using R. As far as I can tell, Packt Publishing does not make its datasets available online unless you buy the book and create a user account which can be a problem if you are checking the book out from the library or borrowing the book from a friend. All of these datasets are in the public domain but simply needed some cleaning up and recoding to match the format in the book.

## **Content**

****Columns****

age: age of primary beneficiary

sex: insurance contractor gender, female, male

bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,  
objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

children: Number of children covered by health insurance / Number of dependents

smoker: Smoking

region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

charges: Individual medical costs billed by health insurance

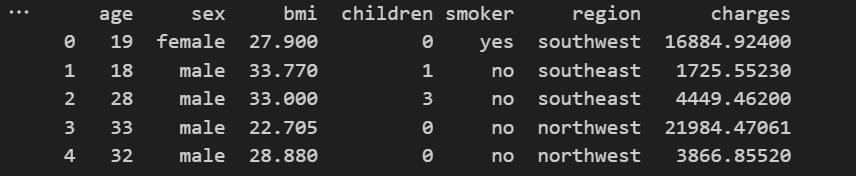
## **Acknowledgements**

The dataset is available on GitHub [here](https://github.com/stedy/Machine-Learning-with-R-datasets" \t "https://www.kaggle.com/datasets/mirichoi0218/_blank).

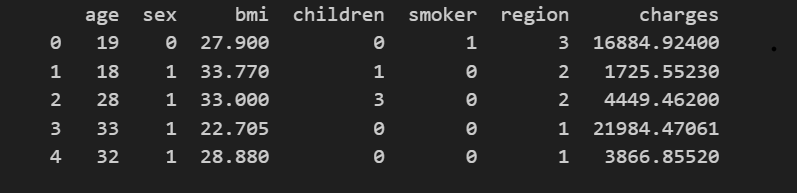
## **Inspiration**

Can you accurately predict insurance costs?

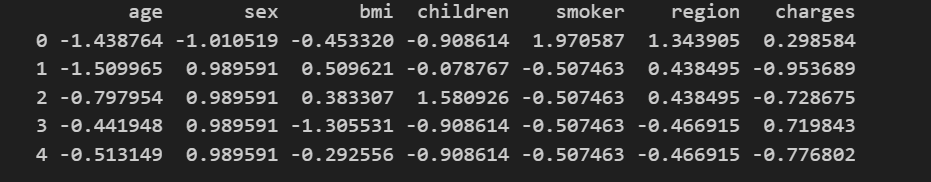
There is no null values or noise data so I did not have to cleaning data , but the are text columns like



**After making a label encoder to change the values to binary column , become:**



Then I have scaled the data using StandardScaler



Then our used features

are: ['age','sex','bmi','children','smoker','region']

And there shape are: (1338, 7) While our target is[charges ] And its shape is: (1338,)

Then splitting the data level:

I have used cross validation (k folds) with k=10 The assign the variables with train\_test\_split Function,I give it parameters:test\_size=20%

Because I noticed that increasing the score of testing while training remains constant at increasing or decreasing the size

and in target variable (y) I assigned to it

[charges] column which my target to make the model predict these values

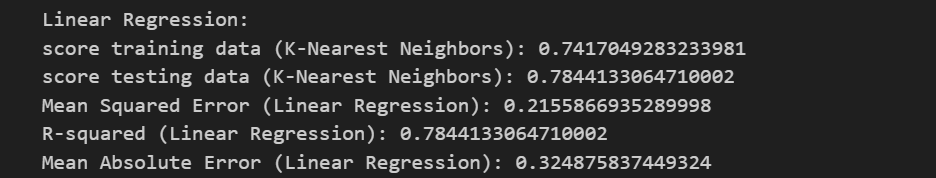
and in in features(x) I assigned to it all remaining columns except [charge].

Then I scaled the data after splitting at: X\_train,x\_test and

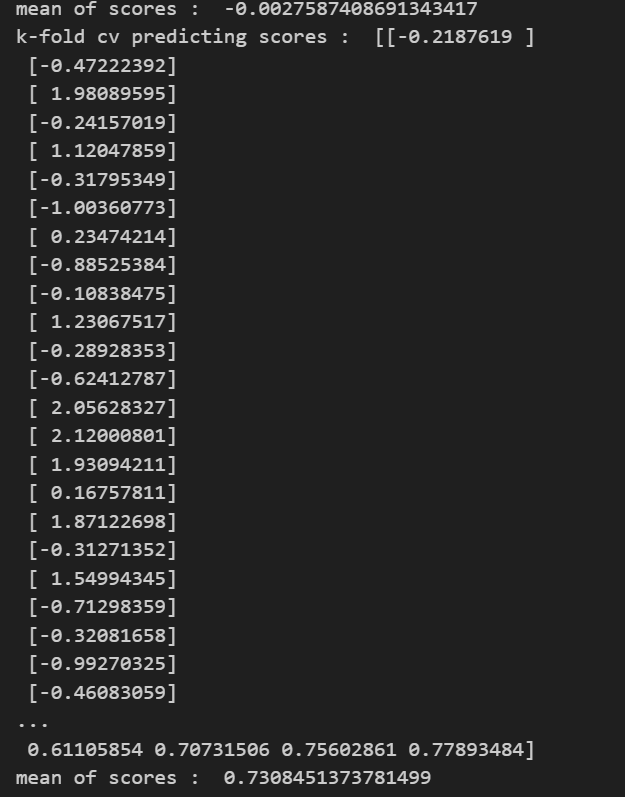
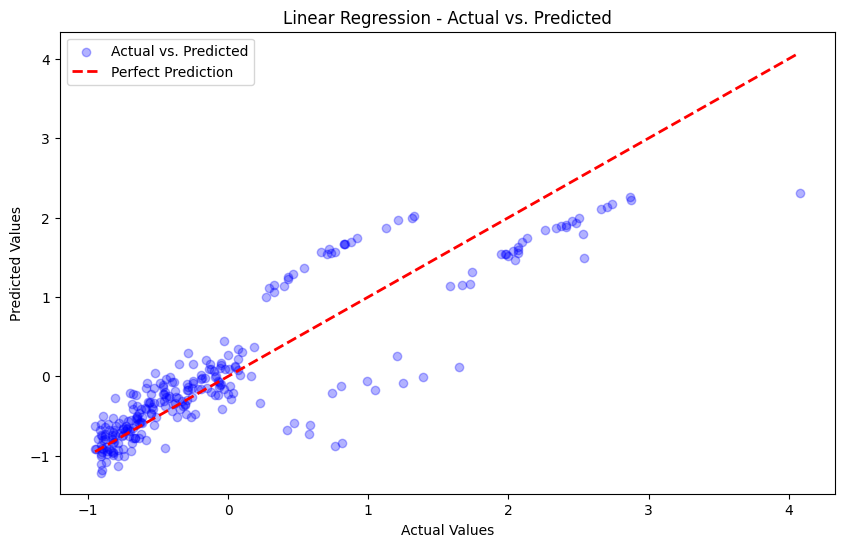
Y\_train,y\_test

The the data is ready to be in the models.

1. **Linear Regression**
2. **Goal : predict charges**
3. **Code : used linear regression from sklearn.lineae\_model**
4. **Cross validation :** "k-fold cross-validation" with k=10
5. **Result :**



And this the plotting of the model:

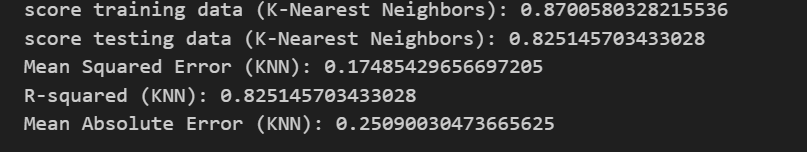


K-fold for 10 of cv

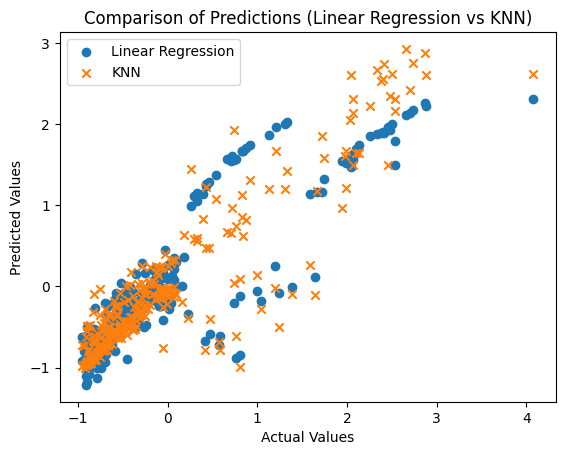
**K Nearest Neighbors (KNN)**

**Goal : predict charges**

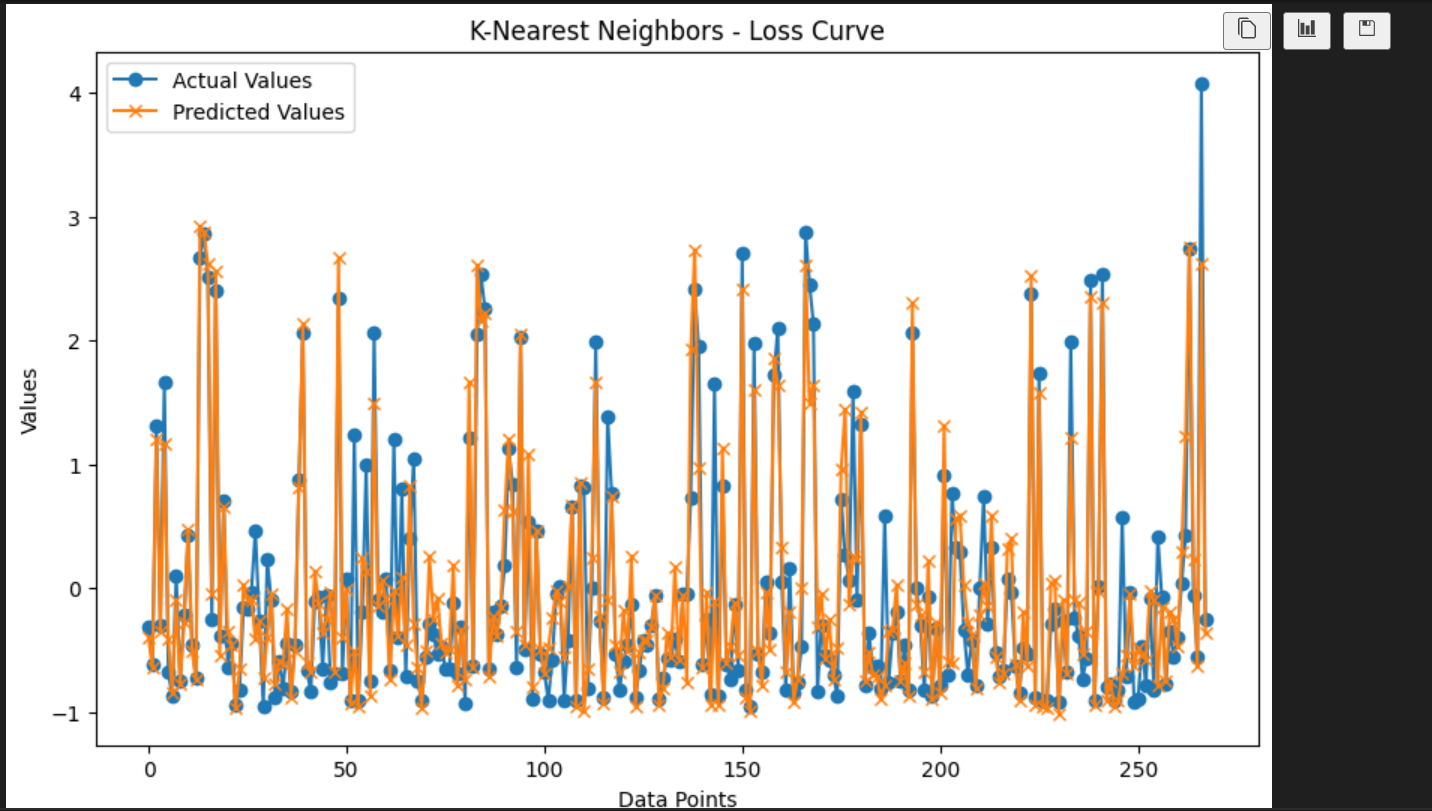
**Cross validation :** "k-fold cross-validation" with k=10

**Result :**

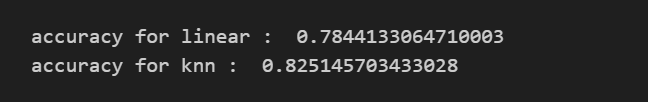
And this the plotting of the model:



Loss curve :



Accuracy :



***Image dataset :***

Link :

*[Fashion MNIST](https://github.com/zalandoresearch/fashion-mnist)*

**Number of Classes and Labels:**

There are 10 classes in Fashion MNIST , used 3 of them

**Total Number of Samples:**

Fashion MNIST typically contains 60,000 training samples and 10,000 testing samples.

**Size of Each Image**:

The size of each image is 28x28 pixels.

**Number of Samples Used:**

**Training Set:**

A subset of the original training set is used based on the specified class labels and further split into training and validation sets.

**Testing Set:**

A subset of the original testing set is used based on the specified class labels.  
  
**b. Implementation Details:**

Feature Extraction Phase:

Number of Features Extracted: The number of features extracted is 3 of 10 but HOG (Histogram of Oriented Gradients) features are extracted for each image.

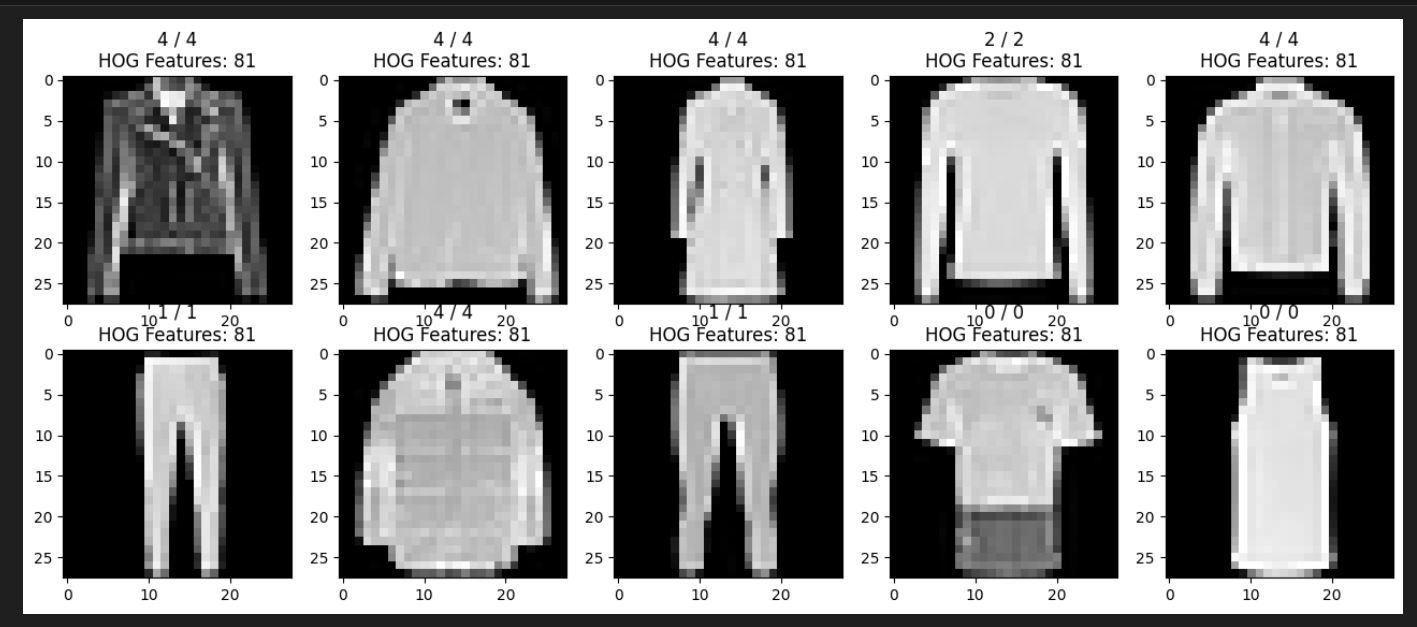
Names of Features: HOG features.

Dimension of Resulted Features: The dimension of the HOG features is not explicitly mentioned, but it's typically a one-dimensional array.

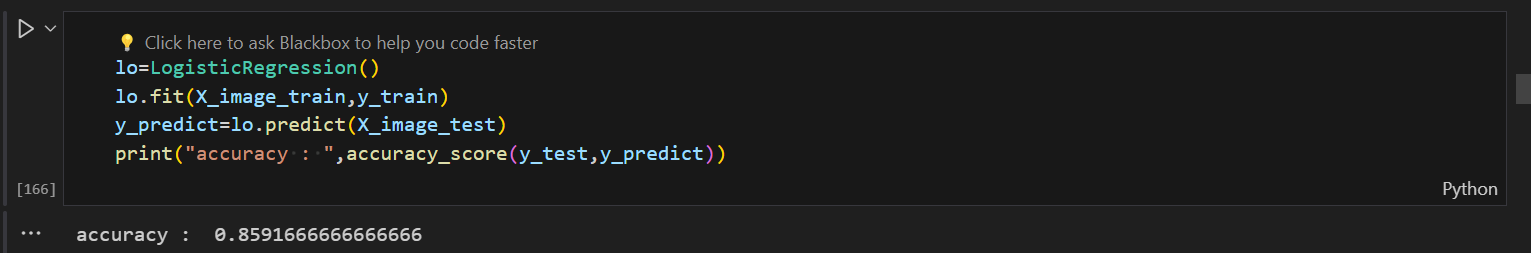
Cross-Validation:

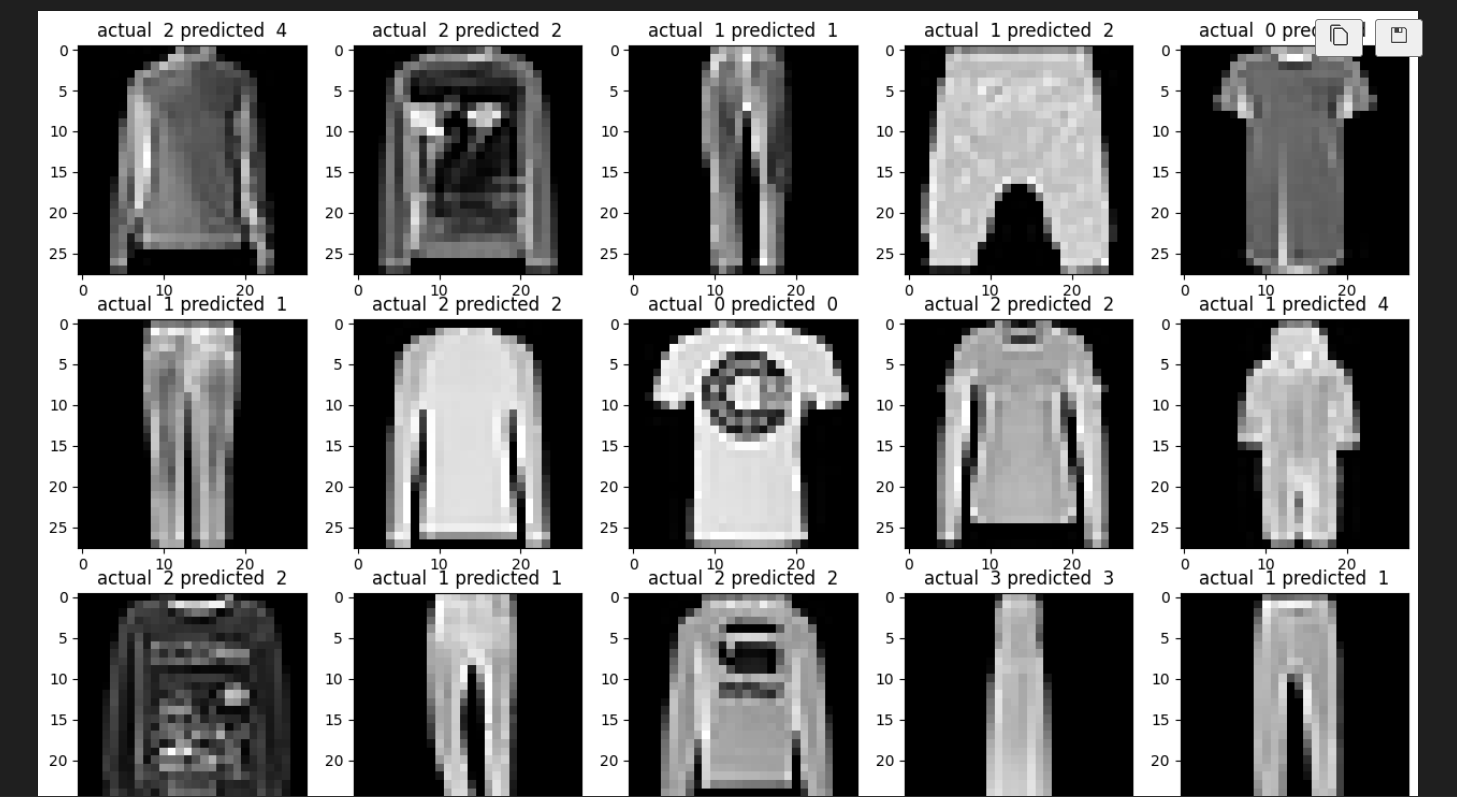
Cross-validation is not explicitly used in this code. The dataset is split into training and testing sets.

After hog and feature extraction:

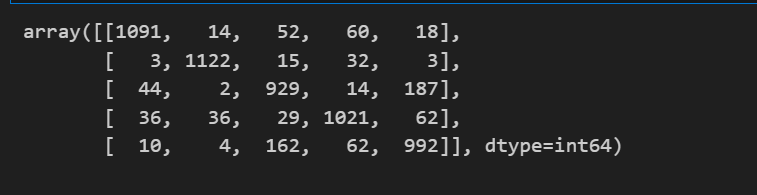


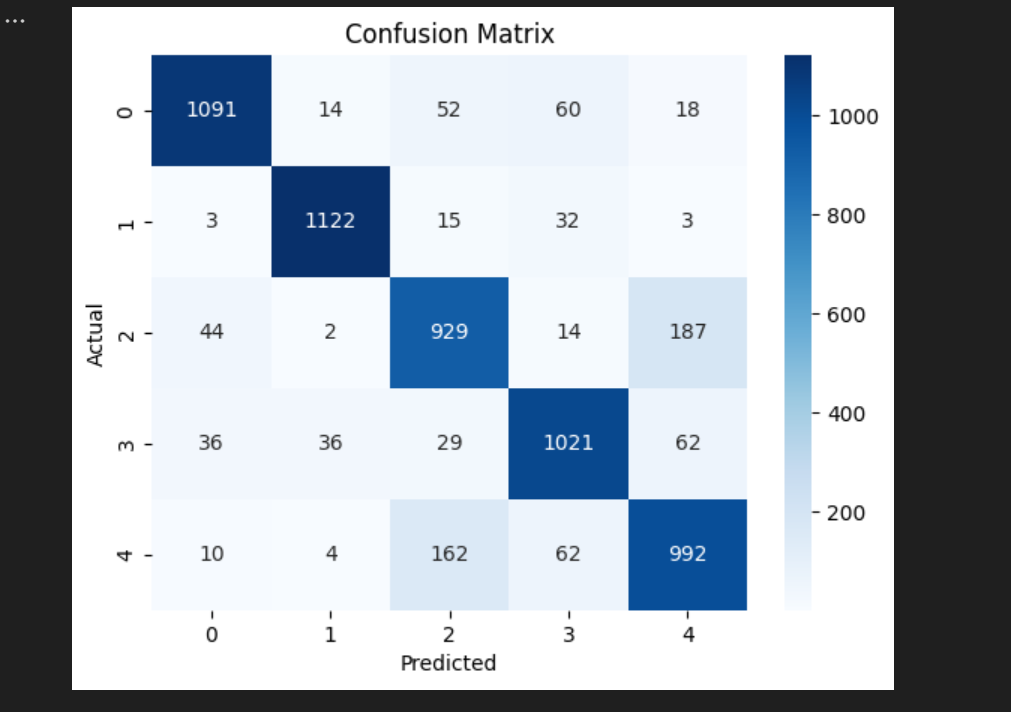
Accuracy for logistic regression :





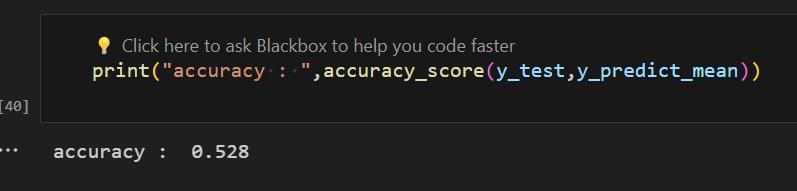
confusion\_matrix(y\_test,y\_predict) :

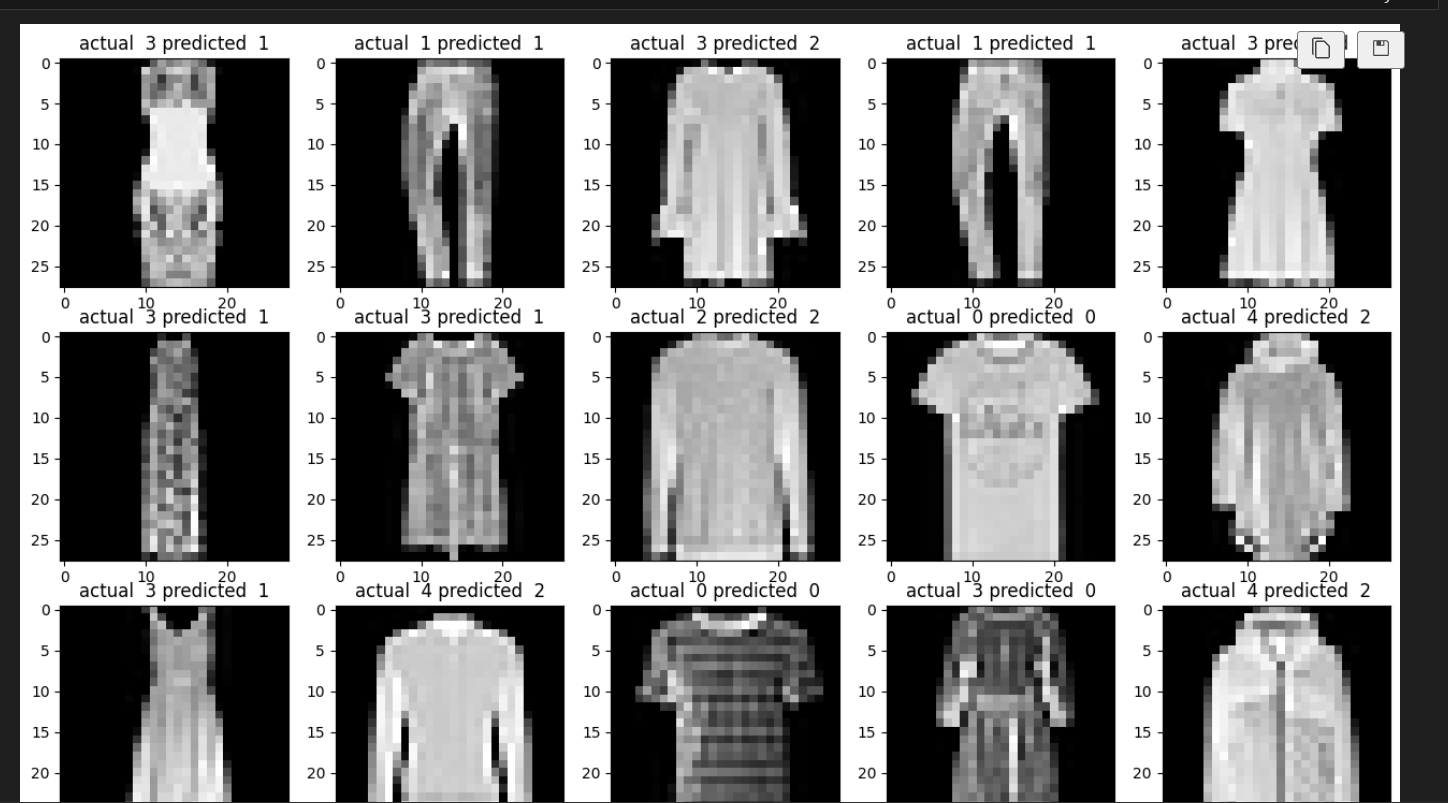




1. mean:

Accuracy :





Confusion\_matrix :

