Project Document

Credit Card Fraud Detection Analysis

Harish Chandra Jyoshi
16262845

GitHub link: https://github.com/HarishChandra95/ISL/tree/master/project/Source

Summary

- The datasets consists of transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- It contains only numerical input variables which are the result of a PCA transformation. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Approaches Used:

- **Data Preprocessing:**
    - Amount and Time columns in the dataset are scaled
    - Created a subsample of dataframe in order to have equal amount of fraud and non-fraud transactions
    - Using the original dataset may cause overfitting
    - Used Undersampling technique to create a new dataset comprising of an equal representation from both classes.
    - The main goal is to fit the data on the undersampled data and test on original dataset.
- **For predicting the job classifications:**
    - KNN approach
    - Decision Tree Model
    - Random Forest

- Support Vector Machine
- Logistic Regression Model

For each approach generated a model using training data. Evaluated the generated models on predicted versus actual values using confusion matrix. Computed the accuracy and the recall value for each model. The Logistic Regression approach has less error when compared to other approaches like KNN, decision trees etc.

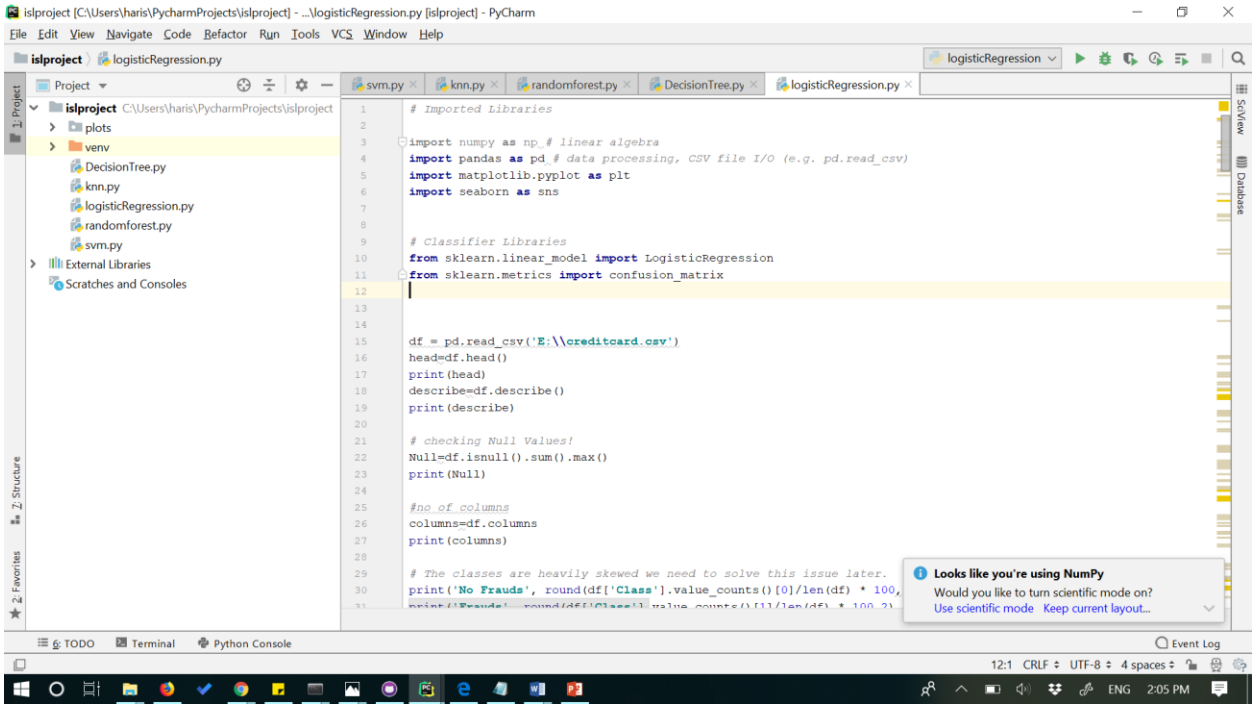
Logistic Regression classifier is more accurate than the other three classifiers in most cases. GridSearchCV is used to determine the parameters that gives the best predictive score for the classifier. Logistic regression model is tested against the full dataset and hit an accuracy of 97.9%.

The credit card fraud detection analysis done using the model created from Logistic Regression approach are accurate. The main issue with "Random Under-Sampling" is that we run the risk that our classification models will not perform as accurate as we would like to since there is a great deal of information loss. In our under-sample data, our model is unable to detect for a large number of cases non-fraud transactions correctly and instead, misclassifies those non fraud transactions as fraud cases. predictions and accuracies may be subjected to change since I implemented data shuffling on both types of data-frames. The main thing is to see if our models are able to correctly classify no fraud and fraud transactions.

Technical Appendix

- **Steps:**

a) Imported the required libraries



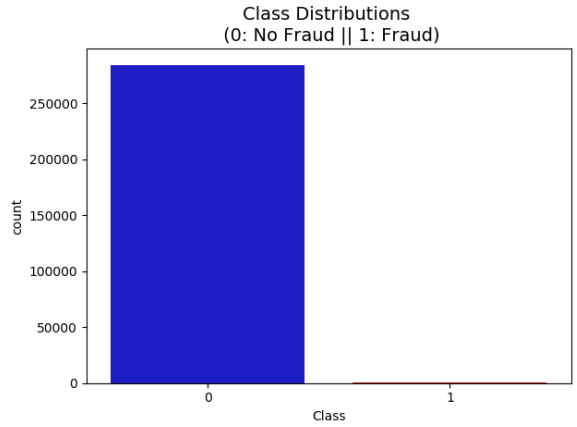
b) Loading the dataset

```
#loading the dataset
df = pd.read_csv('E:\\creditcard.csv')
head=df.head()
print(head)
describe=df.describe()
print(describe)
```

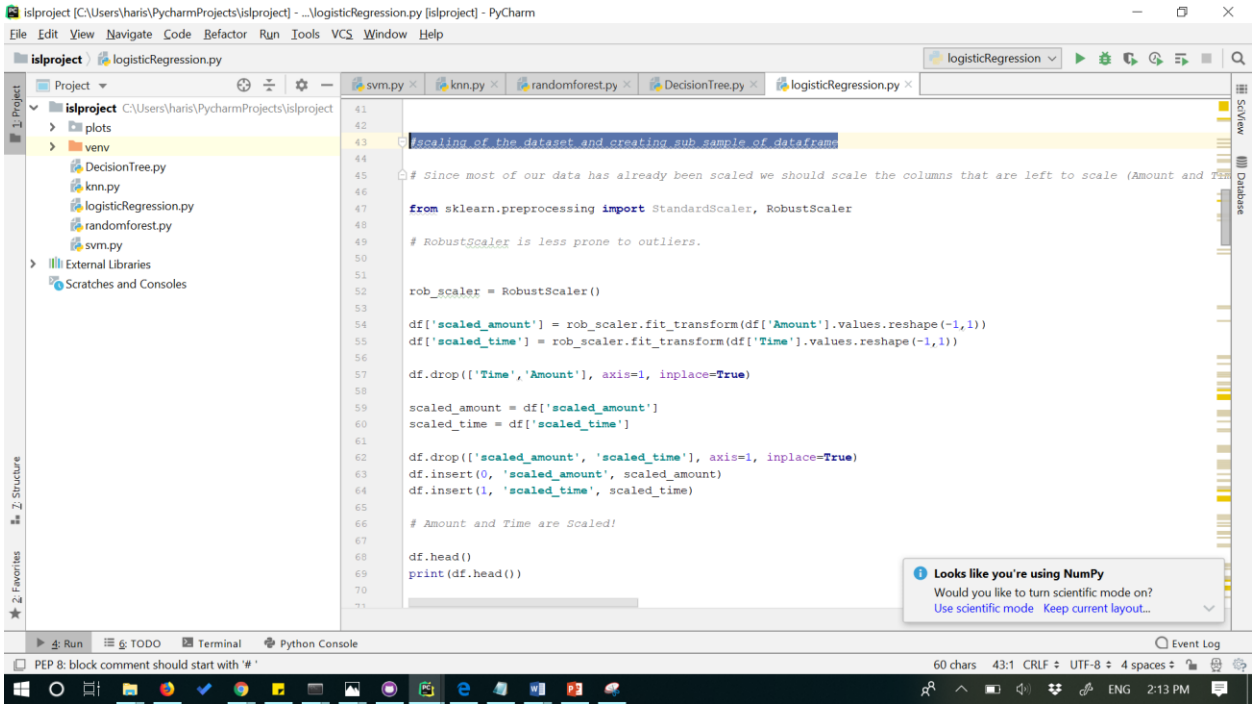
c) Initial Data Structure

```
colors = ["#0101DF", "#DF0101"]

sns.countplot('Class', data=df, palette=colors)
plt.title('Class Distributions \n (0: No Fraud || 1: Fraud)', fontsize=14)
plt.show()
```

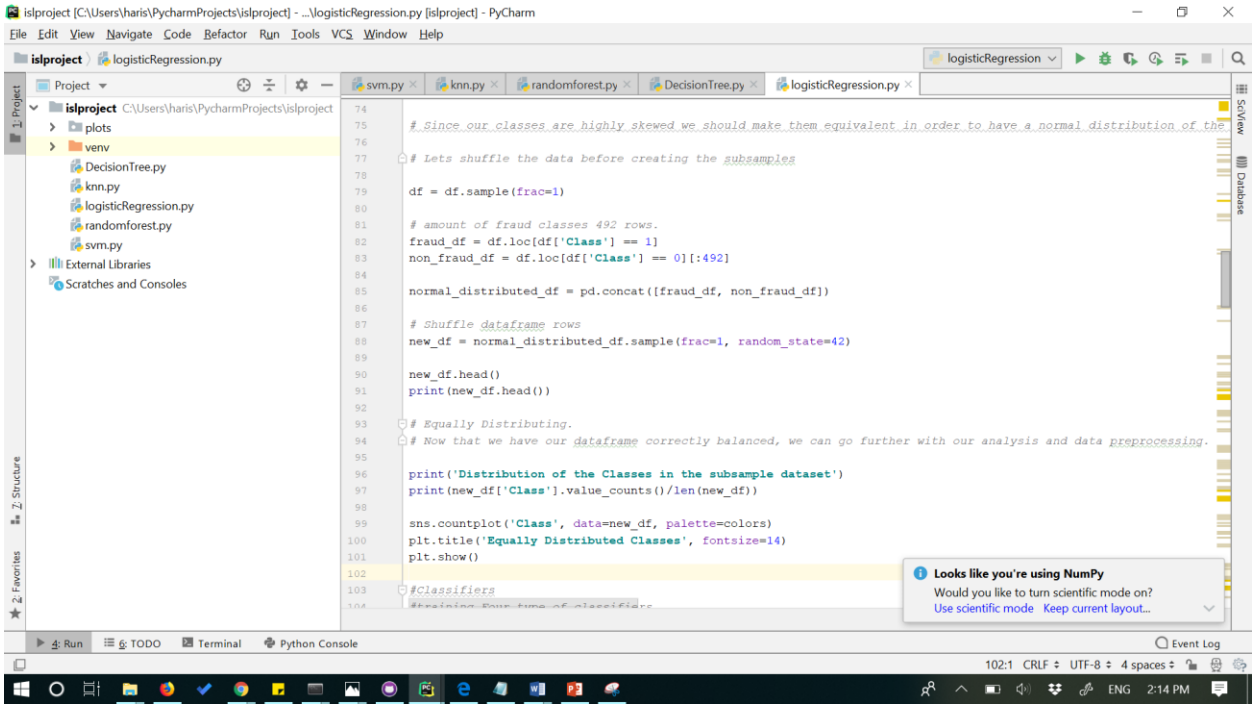


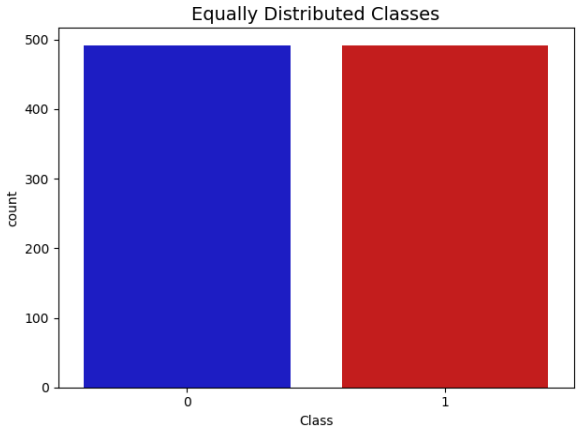
d) Scaling of the dataset and creating sub sample of dataframe



e) Random Undersampling of the data.

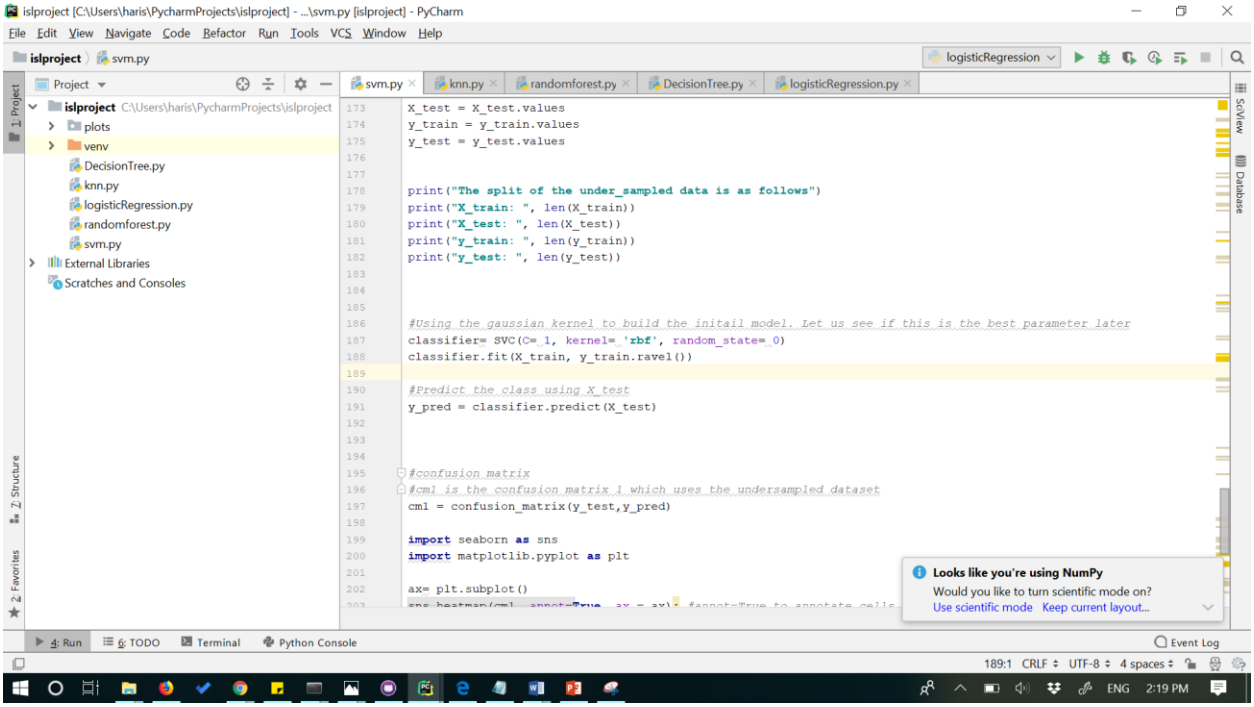
Since our classes are highly skewed we should make them equivalent in order to have a normal distribution of the classes.



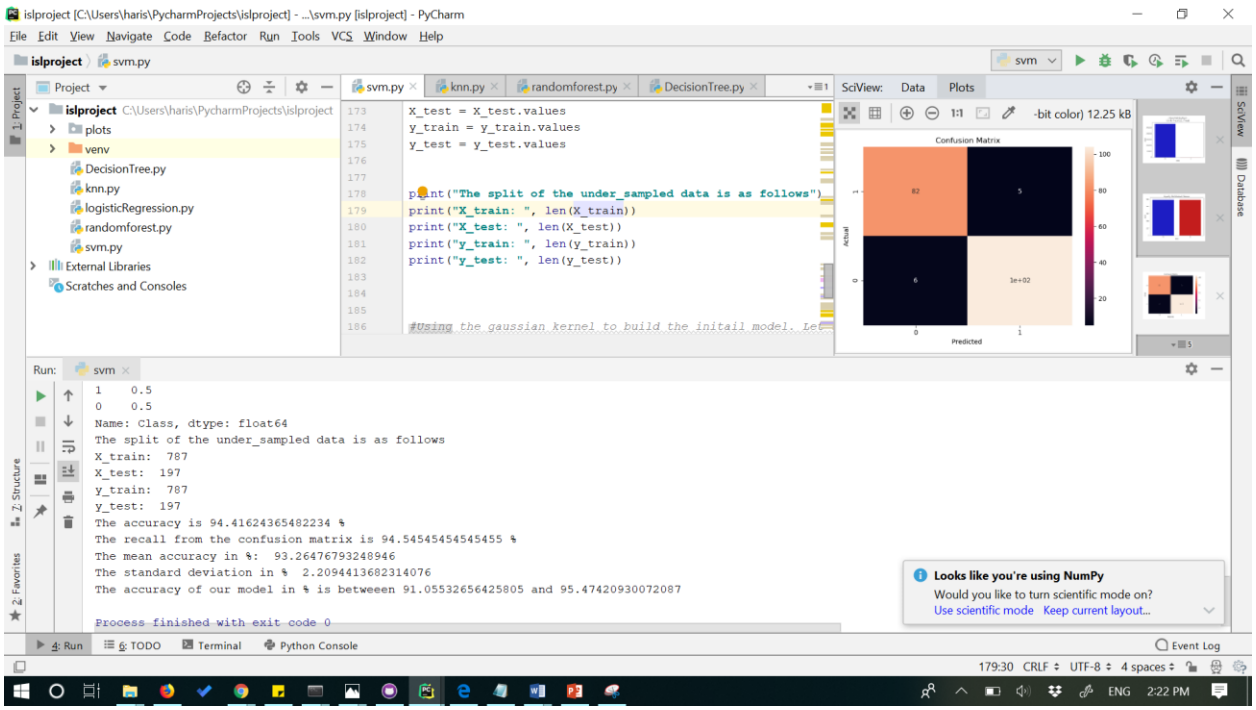


f) Different Classifiers and their outputs

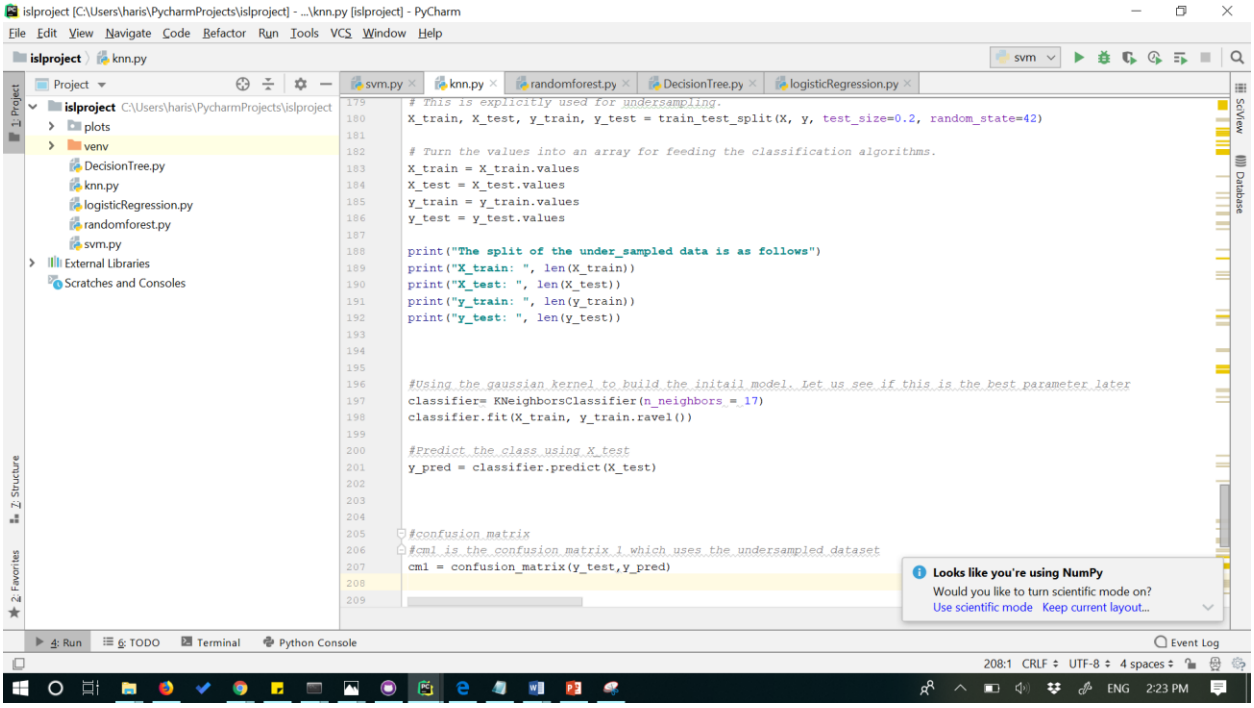
- **Support vector machine:**



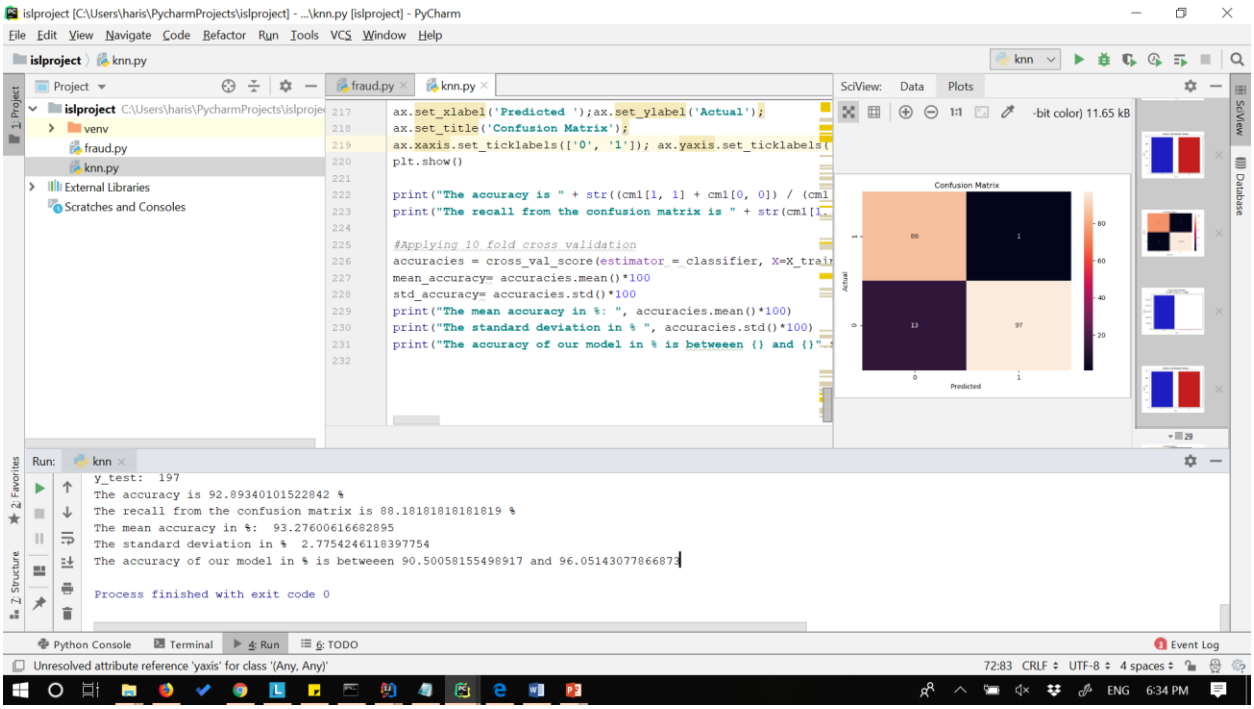
Accuracy and recall along with the confusion matrix:



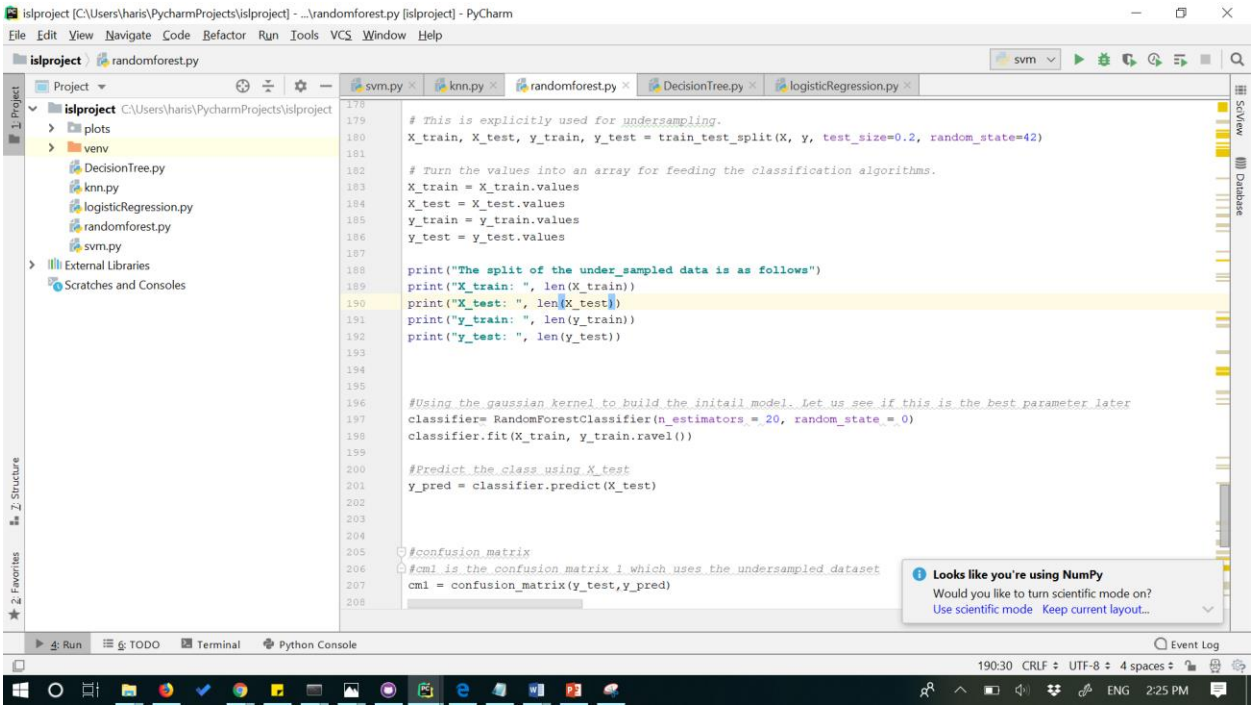
- **KNN approach:**



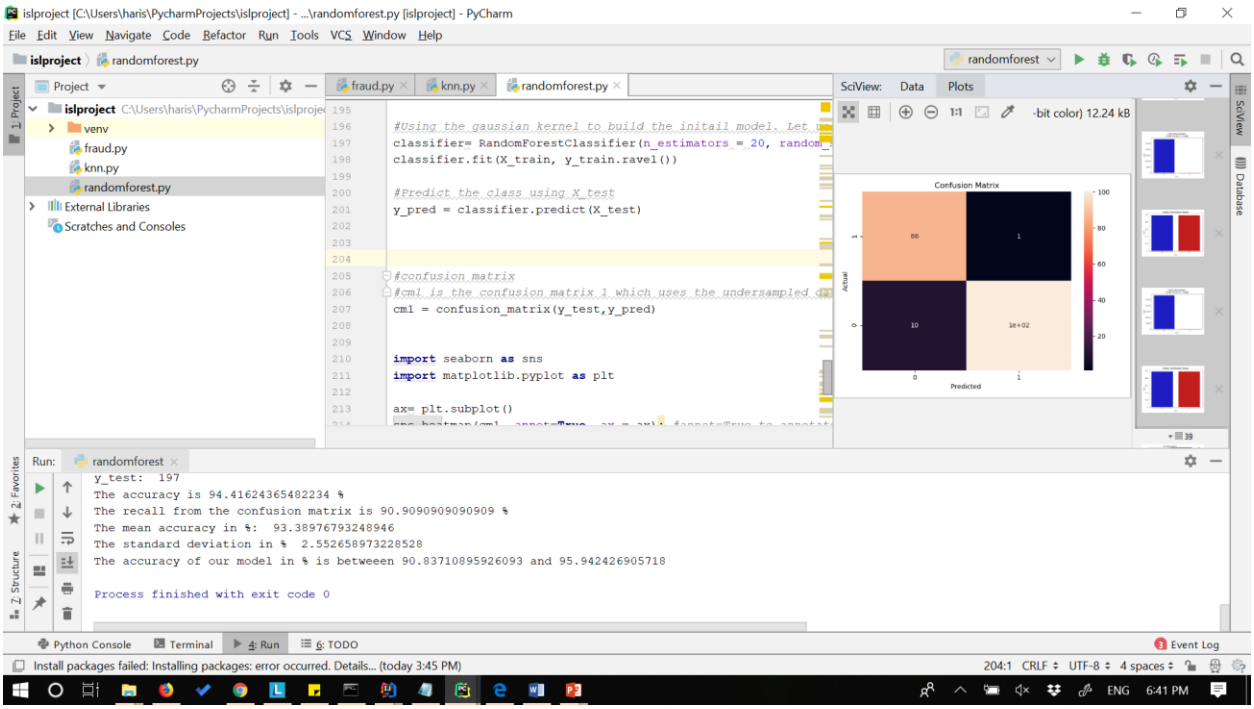
Accuracy and recall along with confusion matrix:



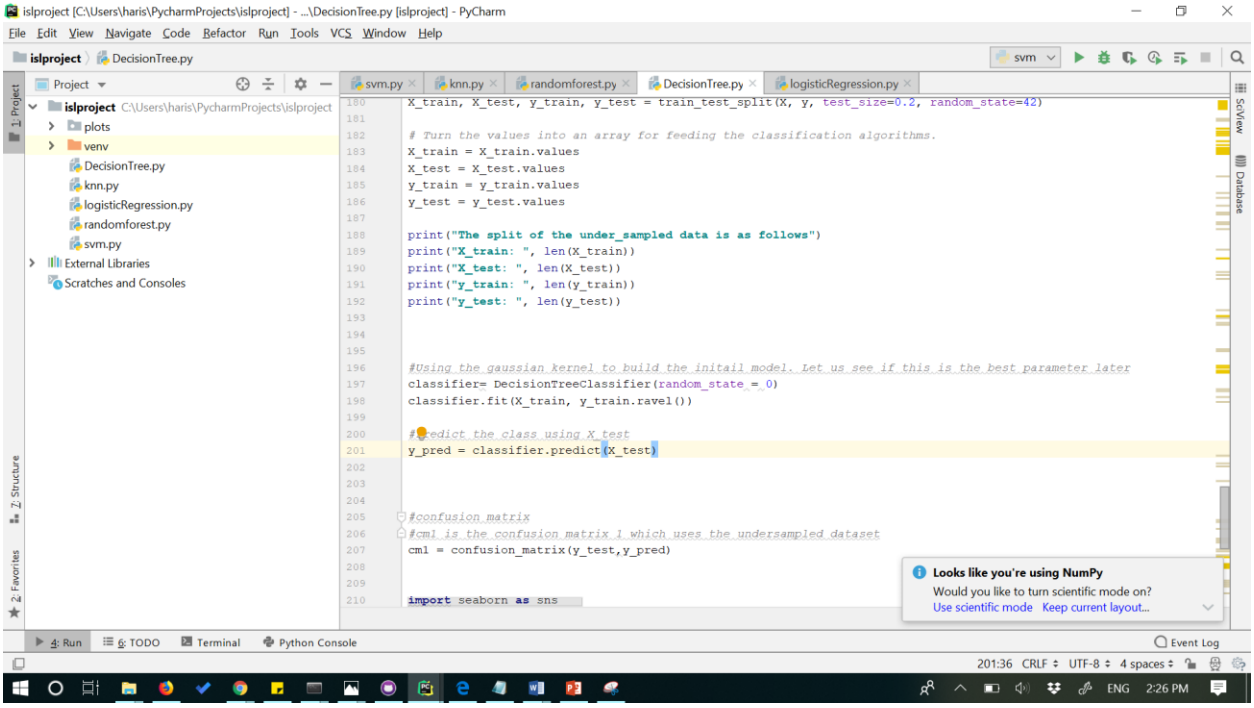
- **Random Forest Approach:**



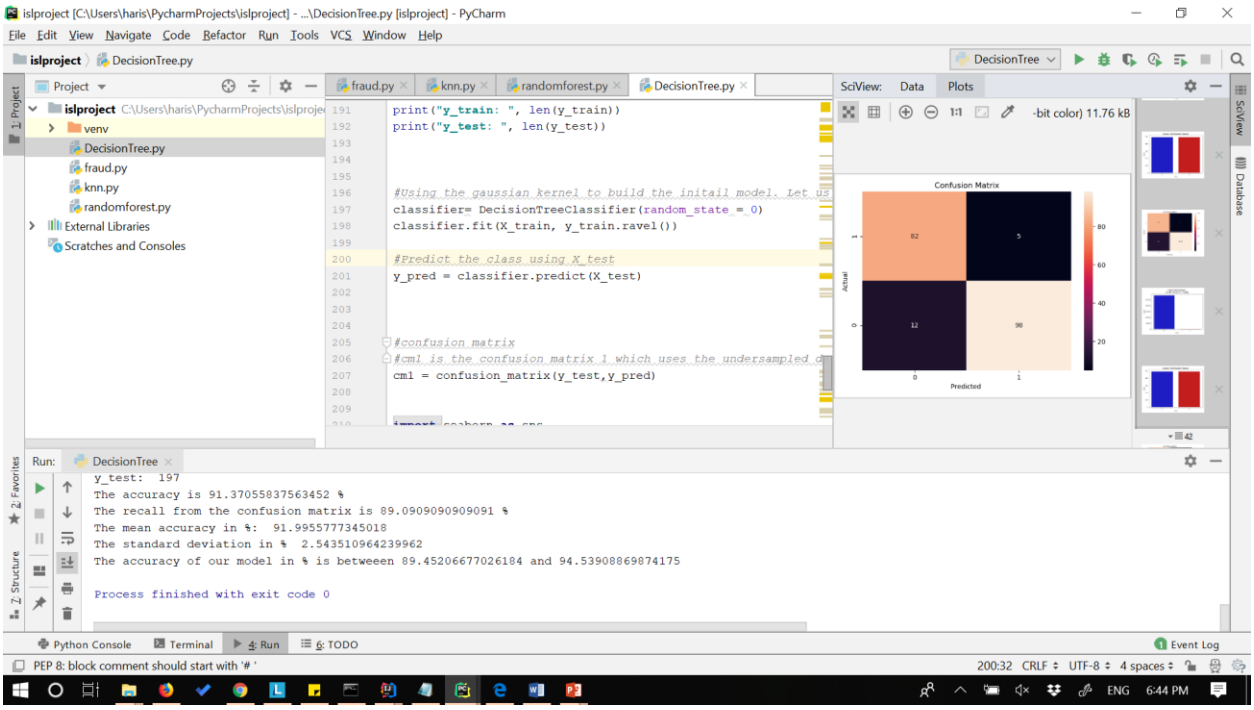
Accuracy and Recall along with confusion matrix:



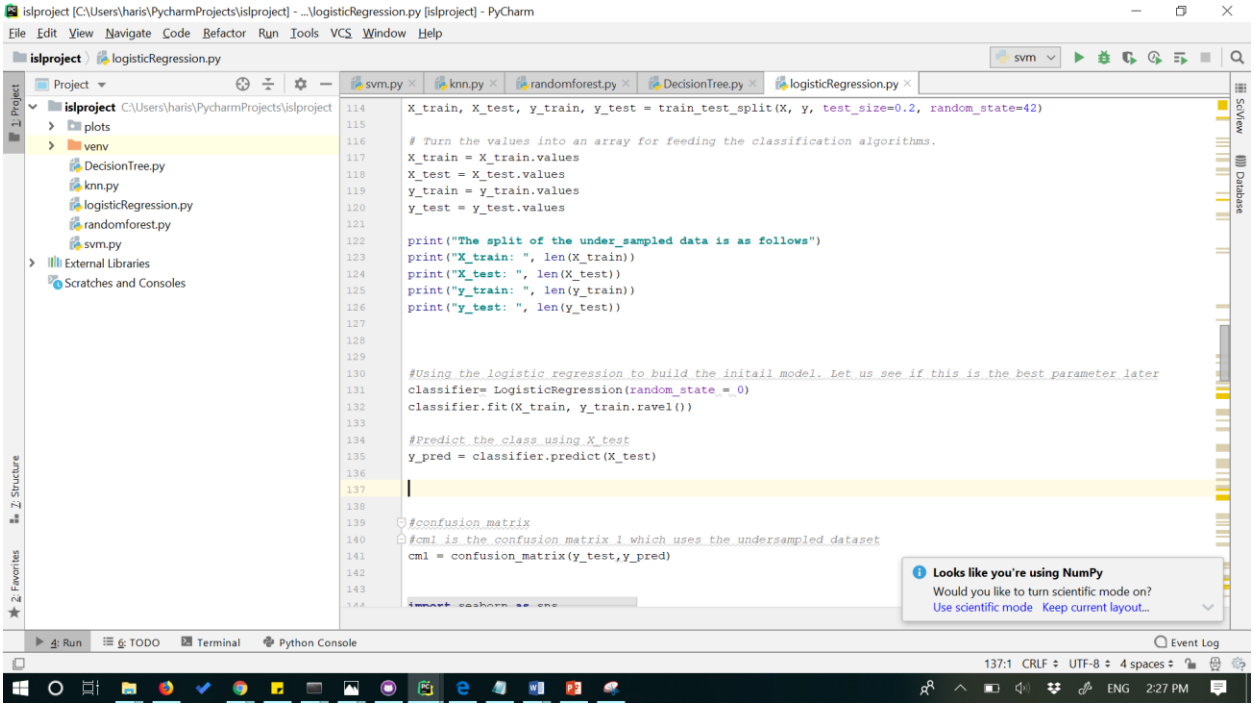
- **Decision Tree approach:**



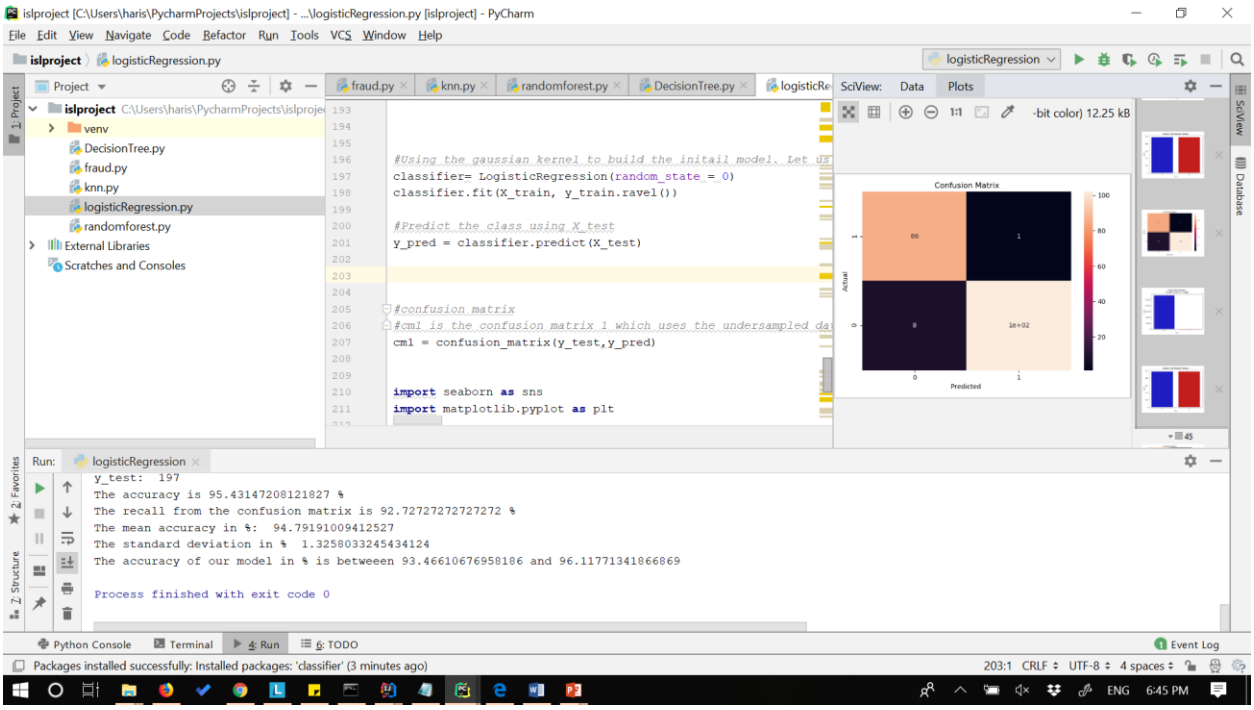
Accuracy and Recall along with the confusion matrix:



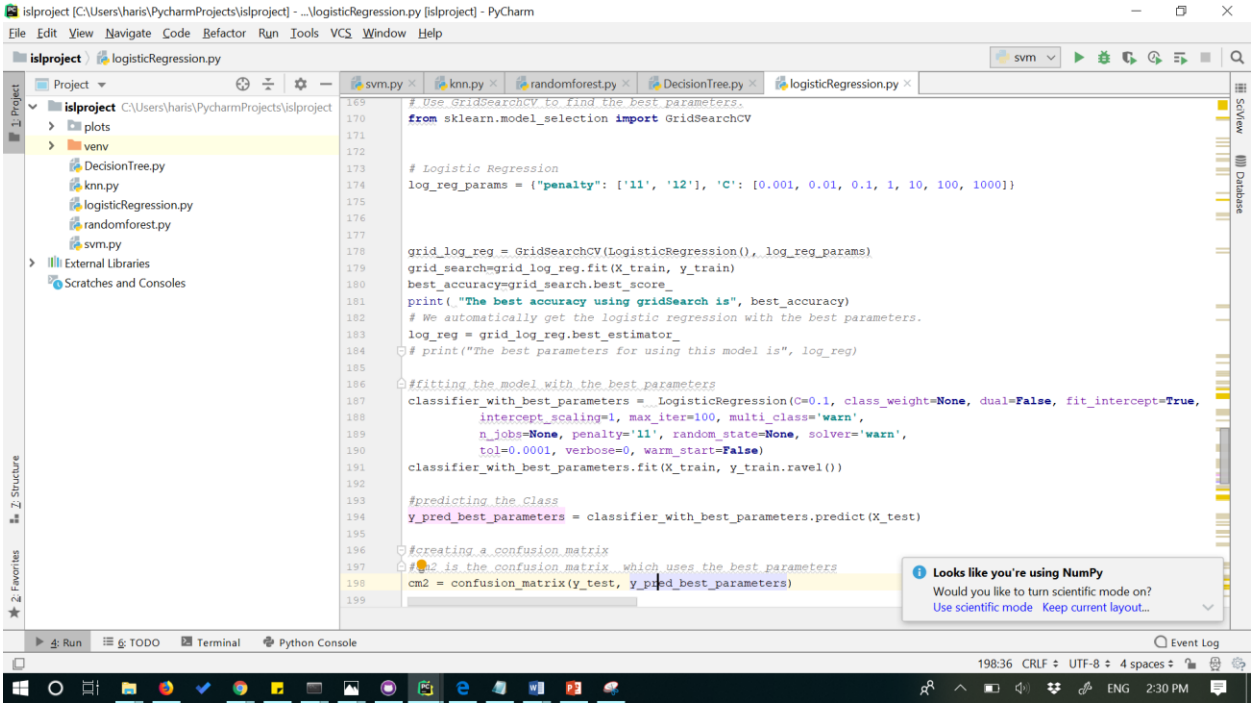
- **Logistic Regression approach:**



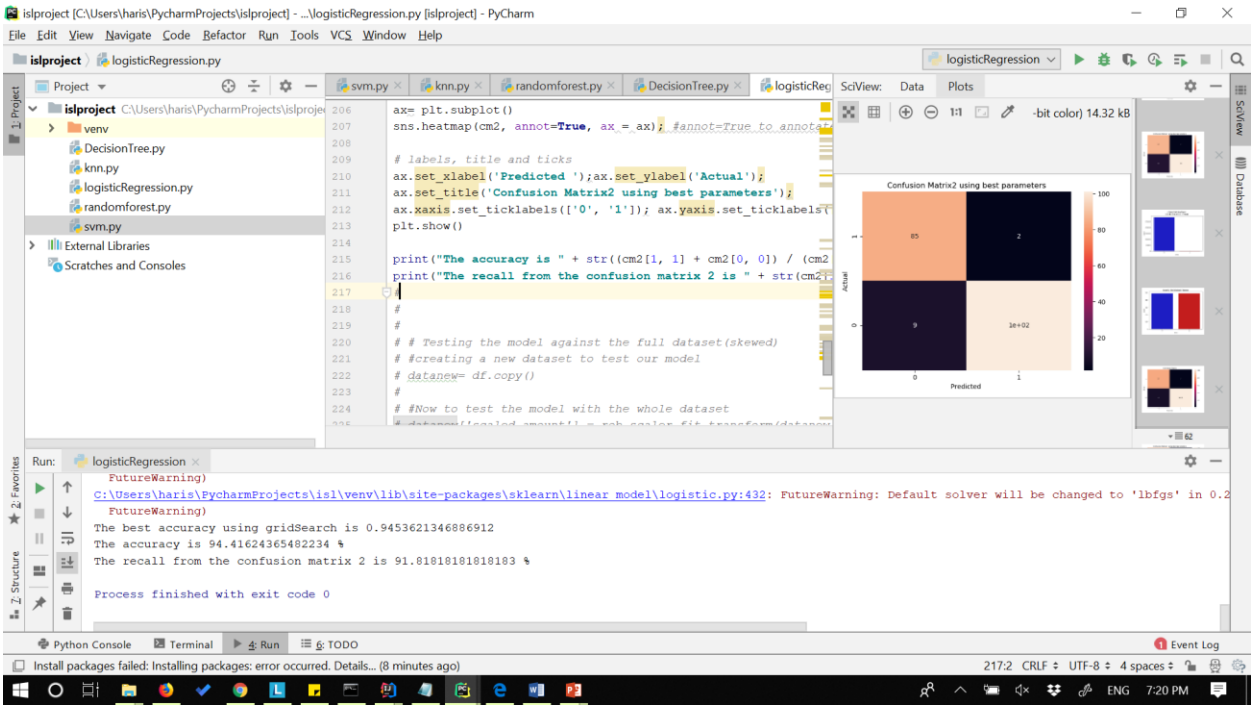
Accuarcy and recall using confusion matrix:



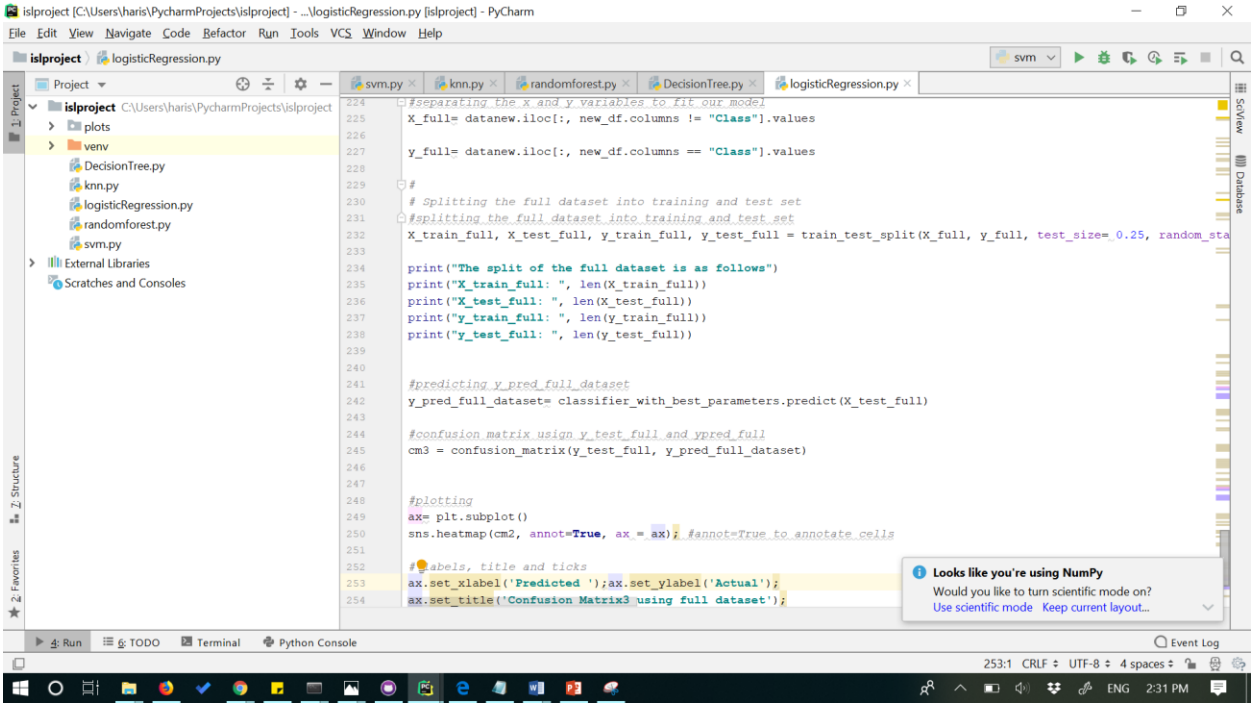
- **Logistic Regression when performed with best parameters Using grid search:**



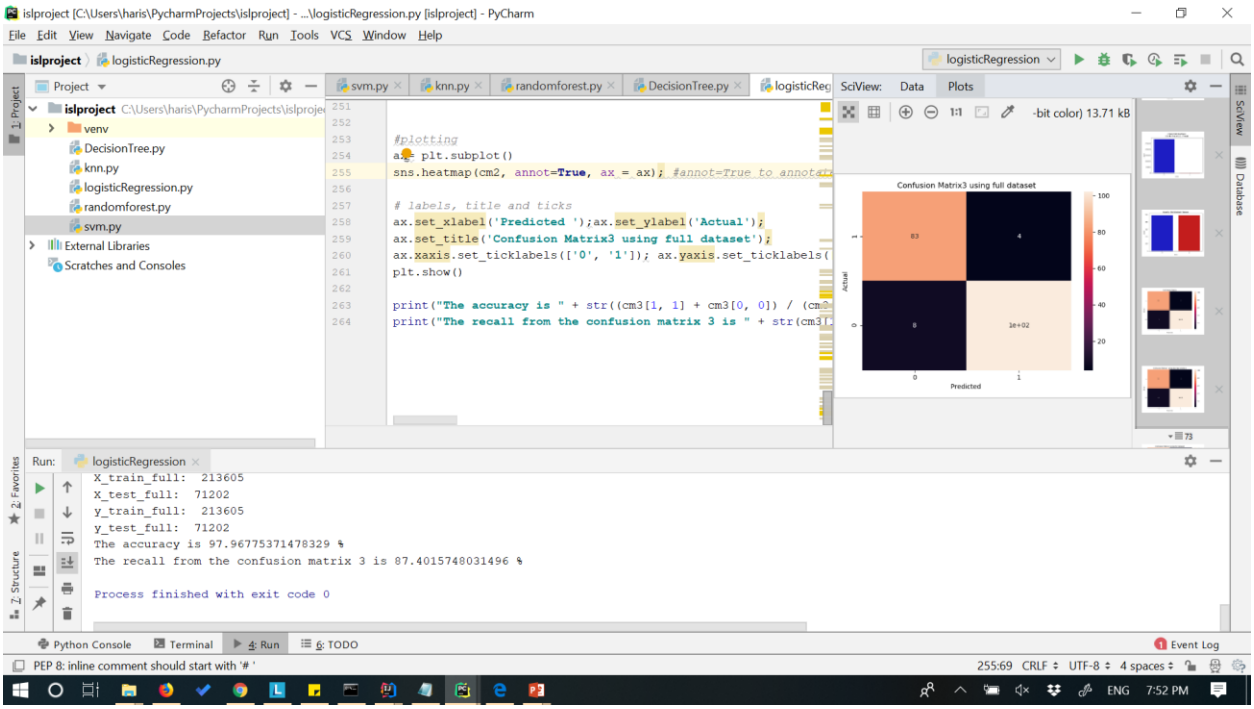
Accuracy and recall along with confusion matrix:



- **Logistic Regression with the test on the actual Dataset.**



Accuracy and recall along with confusion matrix.



References:

1. **https://www.kaggle.com/metetik/classification-algorithms-comparison/notebook**
2. **https://www.kaggle.com/mlg-ulb/creditcardfraud**