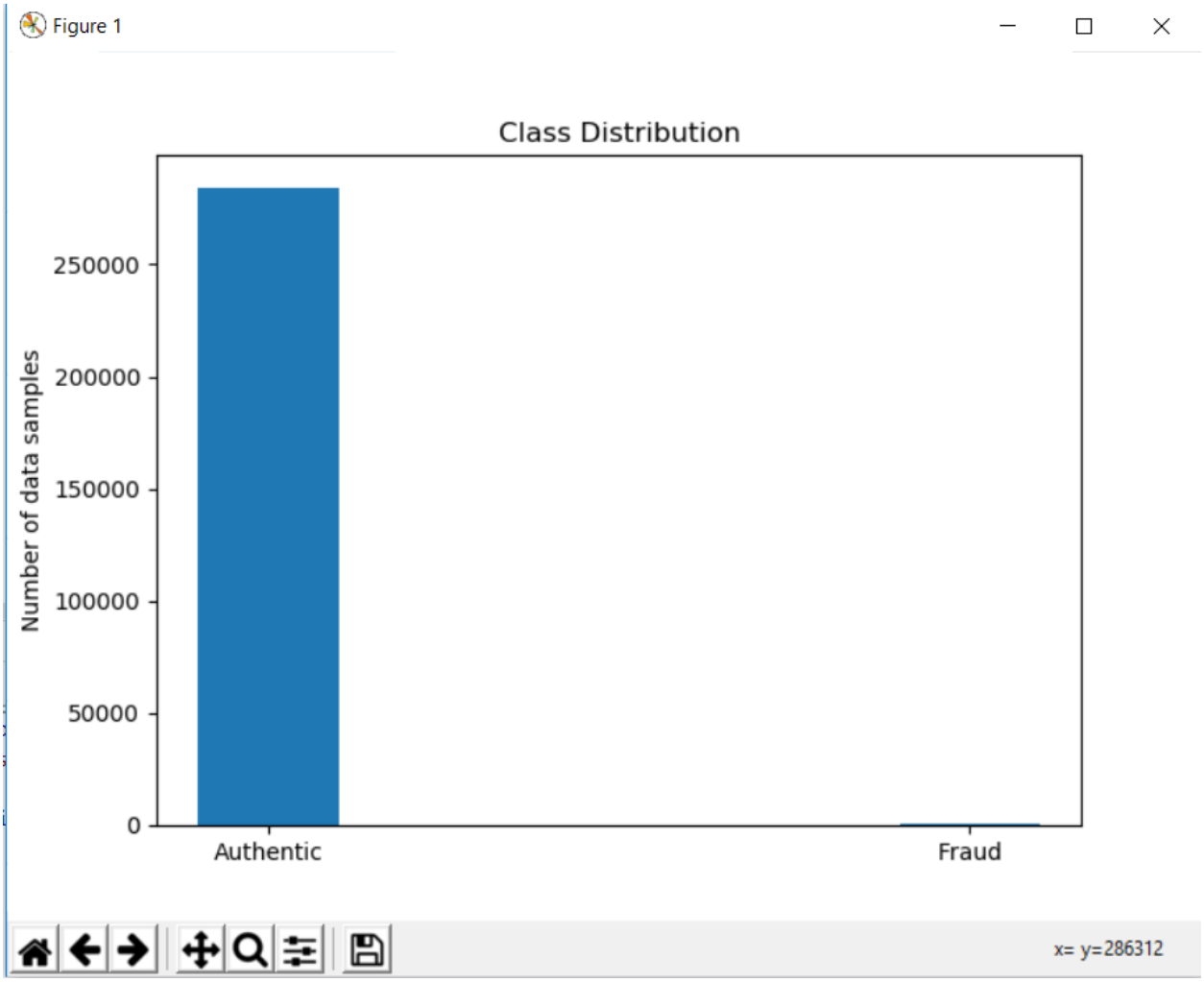
Name: Chaitanya Sardesai
UTA ID: 1001536420

To detect the credit card transaction is fraudulent or authentic.

Class sample distribution:



As we can see there is **significant difference in number of samples of each class**.

Procedure:

1. **Rescaling and splitting the data** between train and test data before under sampling.
    a. 'Time' and 'Amount' features are rescaled between range [0, 1]
    b. All class 1 samples are present at the end of the dataset, hence shuffle all the data to get a better split
    c. Train data = 80%, Test Data = 20%
2. Now **under sample the train data**, to get class sample distribution as [0.5, 0.5] in train data
    a. Pick all the class 0 samples and pick the same amount of class 1 samples at random.
3. Create a **forest of 5 trees**
    a. Reshuffle the train data so as remove class bias while dividing it into 5 subsets of data set.
    b. Divide train data to create 5 different datasets without overlapping of any data sample.

Name: Chaitanya Sardesai
UTA ID: 1001536420

4. **Build 5 different trees**, based on these different smaller datasets
    a. As all the features are continuous, need to pick best threshold and best attribute based on information gain.
    b. 50 different thresholds are generated based on feature's minimum and maximum value
    c. threshold = min_val + k * (max_val - min_val) / 51 … where k iterates from 1 to 50
    d. Create nodes till all the examples are exhausted.
5. **Test the model** using test data
    a. For every sample in test data, get the classification by each tree
    b. Take a vote, and classify the data to a class whichever class has higher votes (count)
6. **Compute the confusion matrix** and get the **Recall metric value**.
7. **Calculate accuracy** simply based on target class, predicted class.

Results:

Code is executed 10 times to check the performance of the algorithm.

Number of Trees	Confusion Matrix	Recall Metric	Accuracy
5	[[53929 2925] [17 91]]	0.8425925925925926	0.9483515326006812
5	[[54228 2632] [8 94]]	0.9215686274509803	0.9536533127348057
5	[[54562 2298] [9 93]]	0.9117647058823529	0.959499315333029
5	[[54858 1995] [12 97]]	0.8899082568807339	0.9647659843404375
5	[[54302 2538] [18 104]]	0.8524590163934426	0.95512798005688
5	[[54267 2590] [6 99]]	0.9428571428571428	0.9544257575225589
5	[[54169 2688] [10 95]]	0.9047619047619048	0.95263509006004
5	[[54838 2038] [6 80]]	0.9302325581395349	0.9641164284961904
5	[[54736 2130] [13 83]]	0.8645833333333334	0.9623784277237456
5	[[53665 3212] [5 80]]	0.9411764705882353	0.9435237526772234
5	[[54841 2025] [6 90]]	0.9375	0.9643446508198448

Average Recall: 0.903582, **Standard Deviation Recall:** 0.034706

Average Accuracy: 0.95662, **Standard Deviation Accuracy:** 0.007038

Using Random Forest, we have achieved good results by looking at the Recall and Accuracy values, their average and std for this classification problem.