**CLOUDALAB**



**Fraud Detection using Self Organizing Maps**

**(Machine Learning, Deep Learning)**

**Sourav Tiwari, Arijit Kundu, Ritwik Dhara, Amayeeka Behera**

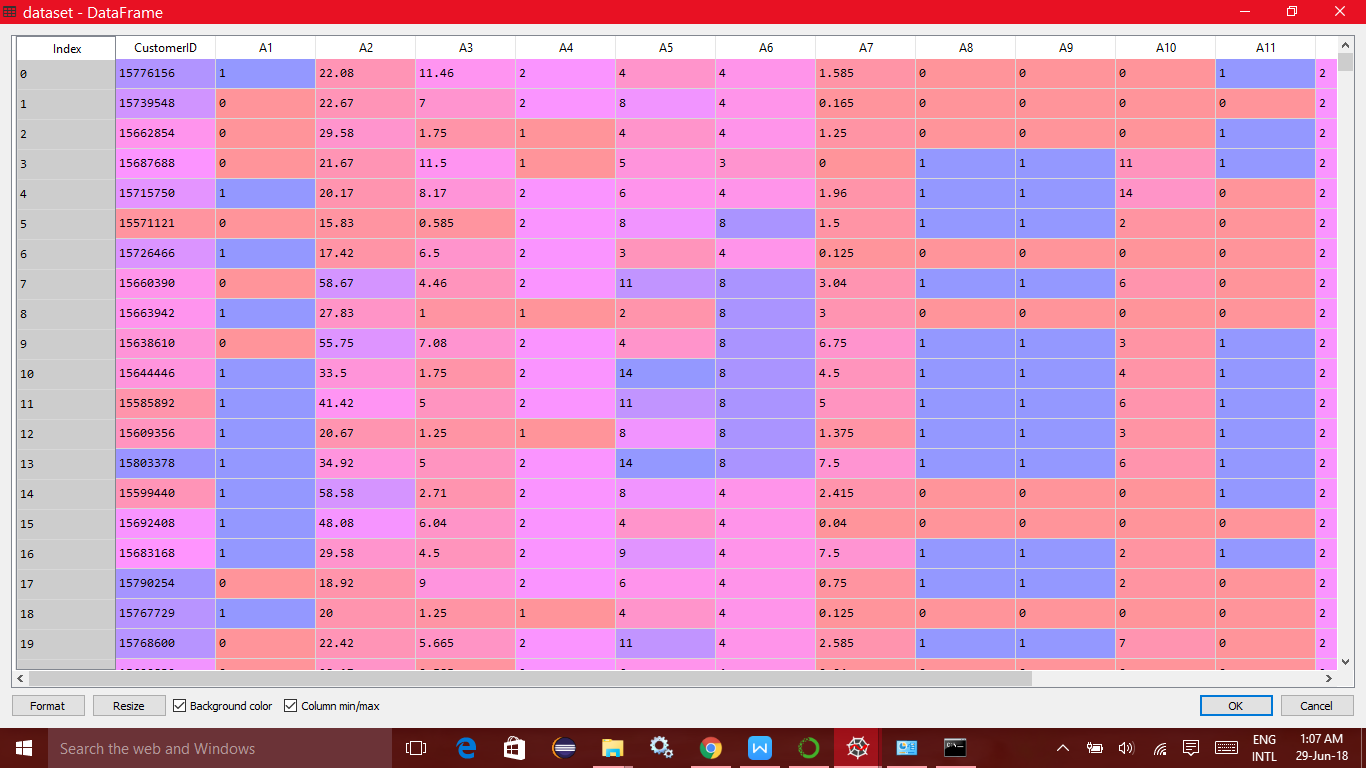
**30/06/2018**

**Fraud in Banking :**

*The bank had issued credit cards to its customers in past few years and the issuance was based on credit scores, applications etc.*

*There are chances that some fraudulent applications got approved (due to human error or something). We have to find some possible frauds so that they can be further investigated by the bank, thus saving the bank from possible losses in the future.*

*For this, we are using Machine learning, Deep learning concepts & data from the Credit\_Card\_Applications.csv file.*

**

*A Screenshot of the dataset taken from Spyder Variable Explorer*

**

**What is Machine learning ?**

*Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.*

*The name machine learning was coined in 1959 by* ***Arthur Samuel****. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs.*

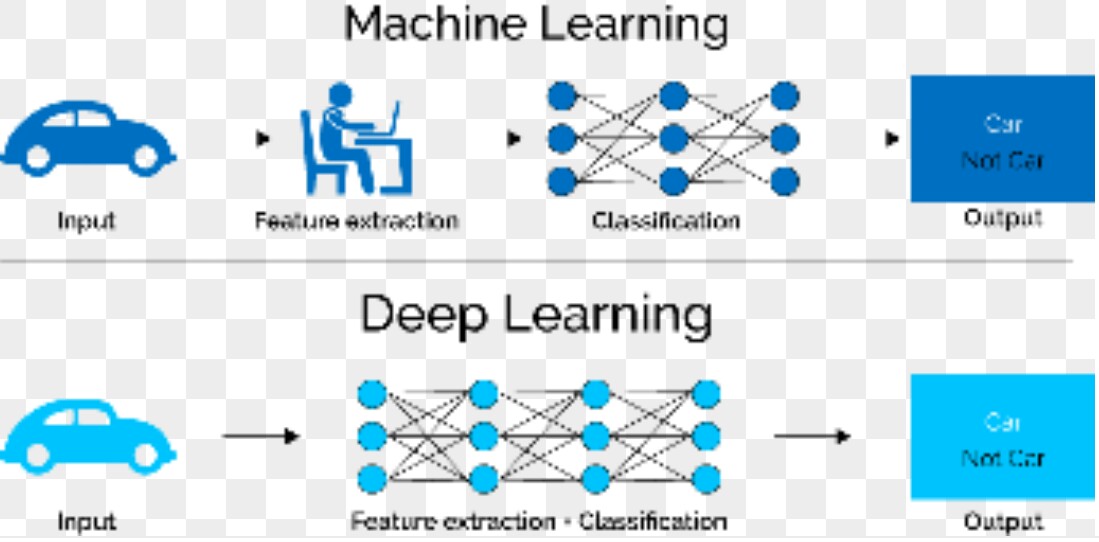
*Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible.*

*Example applications include email filtering, detection of network intruders, malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision.*

*Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers.*

*It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field.*

*Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in data*

**

*. *

**What is Deep Learning ?**

*Deep learning is a class of machine learning algorithms that:*

* *Use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.*
* *Learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners.*
* *Learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.*

*In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation.*

*In an image recognition application, the raw input may be a matrix of pixels*

*The first representational layer may abstract the pixels and encode edges.*

*The second layer may compose and encode arrangements of edges.*

*The third layer may encode a nose and eyes; and the fourth layer may recognize that the image contains a face. Importantly, a deep learning process can learn which features to optimally place in which level on its own. (Of course, this does not completely obviate the need for hand-tuning)*

*Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design and board game programs, where they have produced results comparable to and in some cases superior to human experts*

**Business Problem & Solution Approach :**

*Class = 0 reflects that customer didn’t get credit card from bank and Class = 1 indicates that customer got his/her credit card application approved from the bank.*

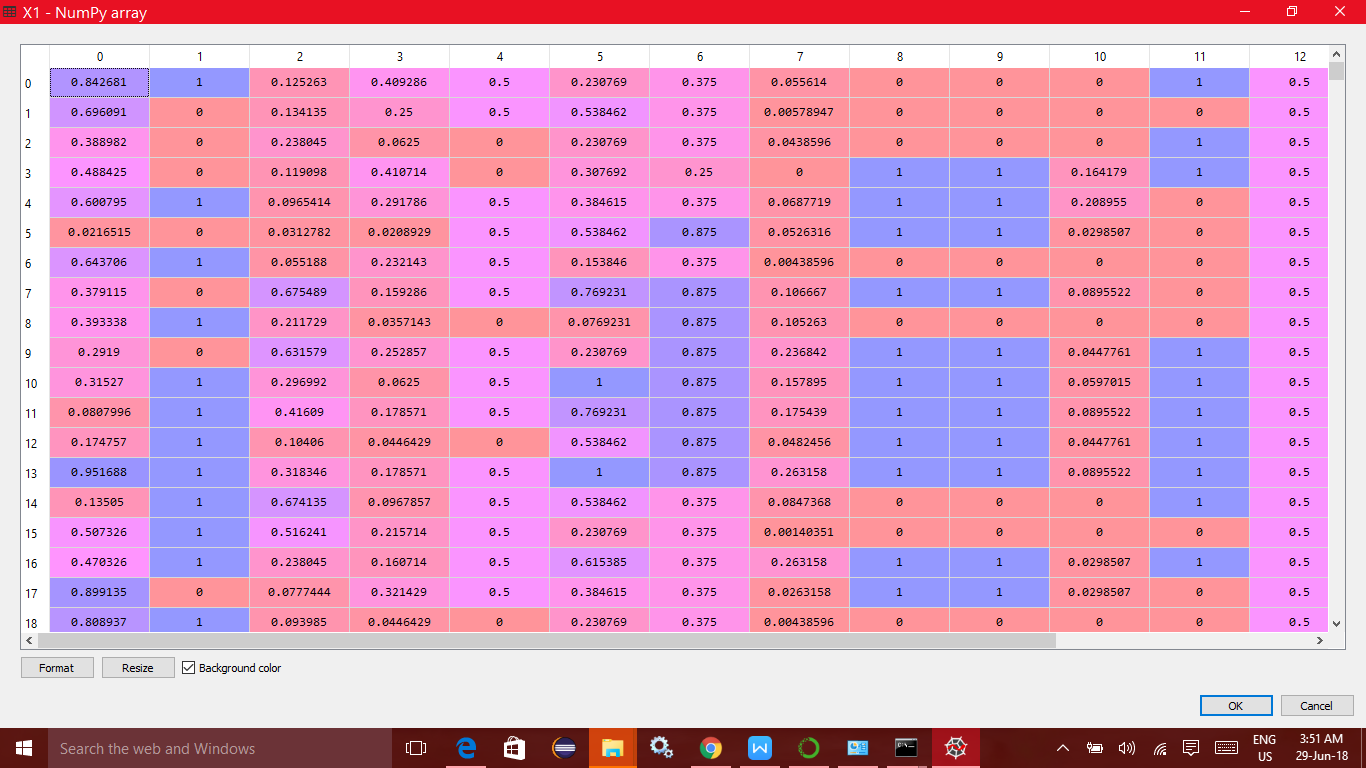
*Plan of Action:*

*Now that we have looked at the dataset and know our task as well (to find possible frauds so that bank can investigate further) lets define a path on which we shall move to get the desired results. Lets go step by step.*

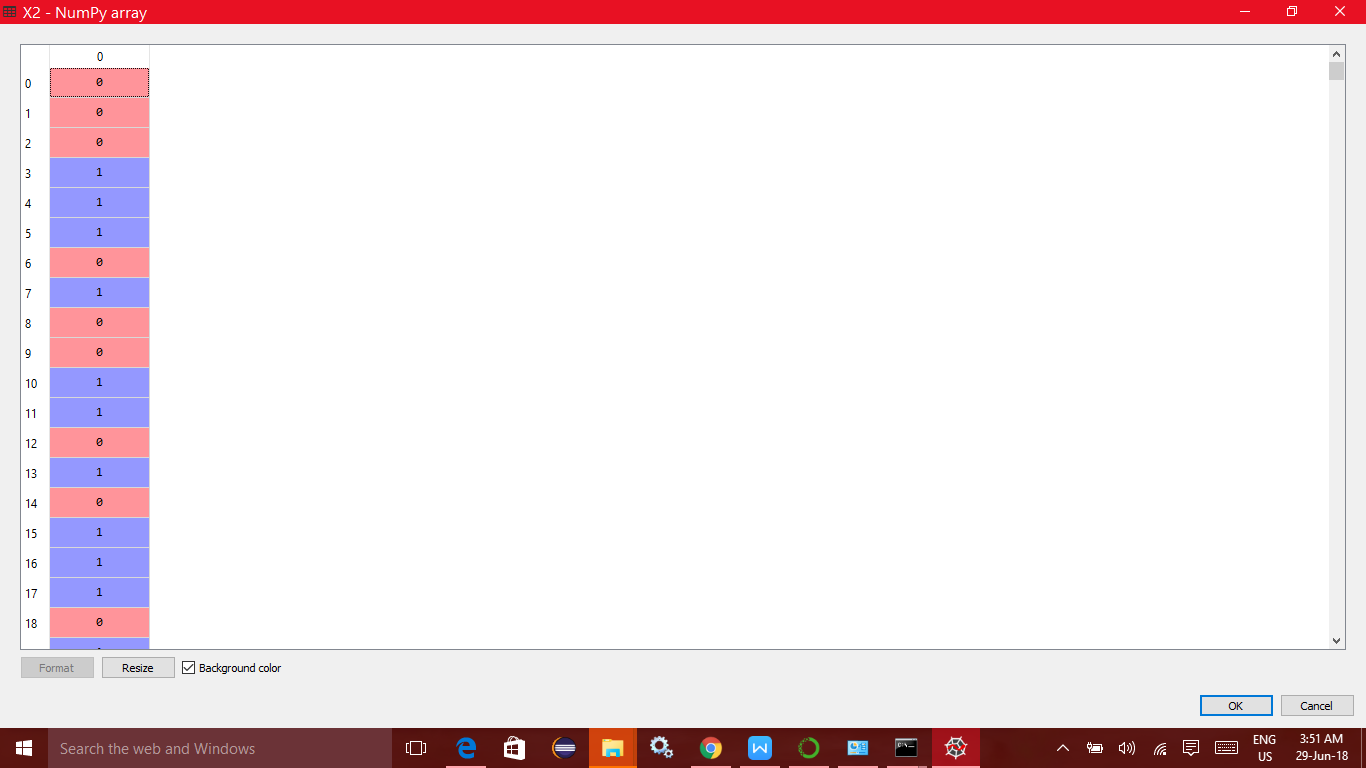
***Step 1:***

*Apply a Self Organizing Map (SOM) on the dataset ignoring the class column. We leverage Class column later. Thus we feed 15 columns (CustomerID + 14 Attributes) to the SOM.*

*Fig 3: A SOM. Data from input layers is mapped on to a 2D lattice.*

**

*The dataset to be fed to SOM*

**

*Each customer’s attribute is fitted to one of the neuron (Winner neuron for that row) out of 15 x 15 nodes on the 2D map. Thus entire dataset is converged on the map with each neuron/node containing many customers having some correlations in their attributes.*

*Fig 4: Snap of Variable explorer showing node/ neuron co-ordinates on the 2D SOM map and their size.*

**

*The co-ordinate (0,0) on the map denotes first neuron or node. As seen in the image, the size of that node is 4. This means that 4 customers have been linked to this node as they had some similar attributes and their correlations which got identified by the SOM.*

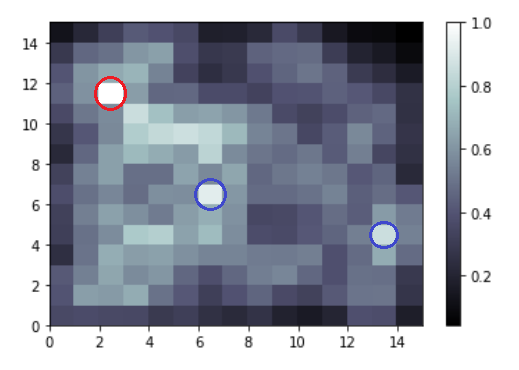
*Similarly, notice co-ordinate (1,14) having size 27. That means 27 customers out of 690 had same correlations and hence got grouped into one node. The list continues for 225 (15 x 15) nodes and customers gets mapped to each of the nodes.*

***Step 2:***

*Once we have mapped each row of our dataset to each node on the 15x15 SOM map, we then move to find outliers. This means, we resort to find out as to which neuron/node appears strange from the rest of the lot. Once we find outlier nodes/ neurons, we can then find the customers grouped under them and they will be the most probable frauds.*

*To find outliers, We utilize mean interneuron distance (MID). If MID for a node/ neuron is high, it indicates that the node/ neuron is an outlier and that it appears different from the rest of the nodes. The customers grouped in those nodes/ neurons are then our highest probable frauds.*

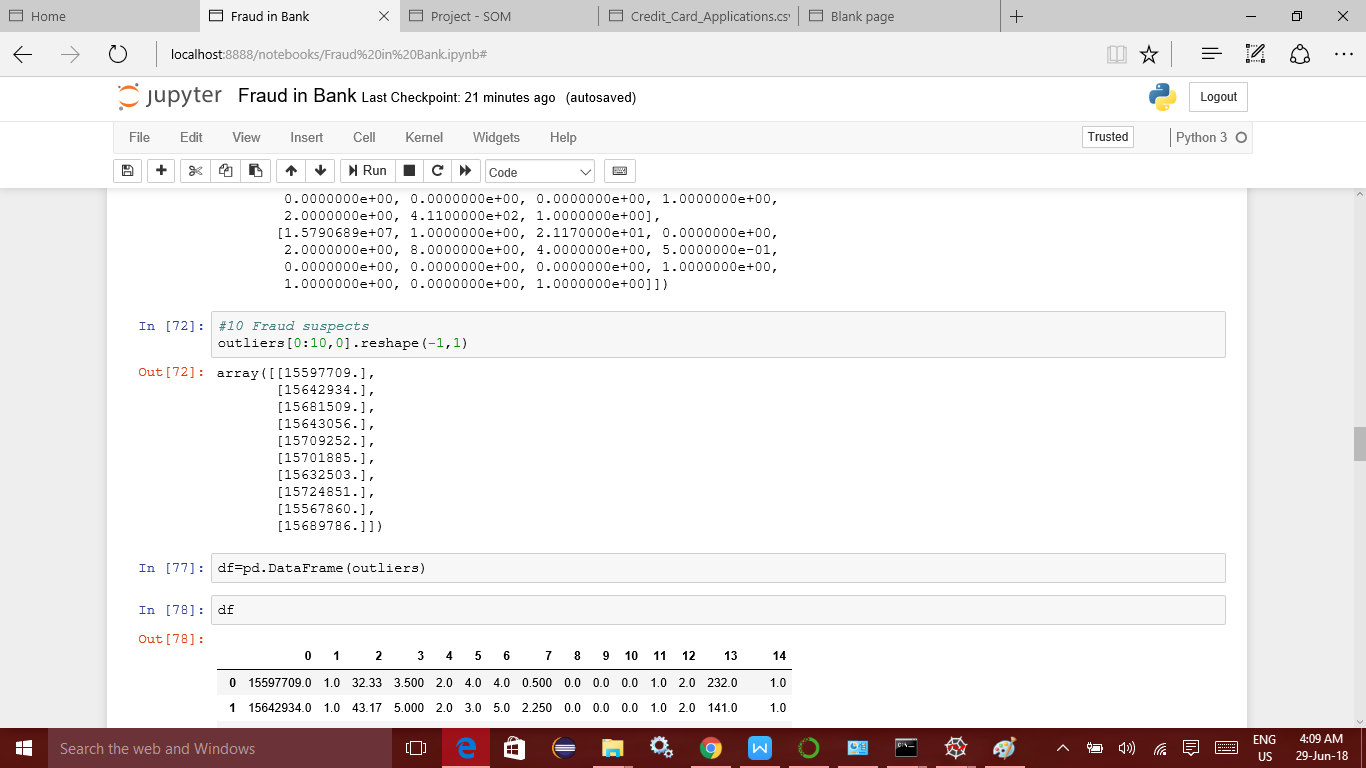
*Fig 5: Plot of nodes showing the outliers. The one circled in red is the node with highest distance, shown in white. The slightly less significant ones are circles in blue.*

**

*As seen in Fig 5, the node circled in red is surely an outlier and the customers grouped in that neurons are our most probable frauds.*

***Step 3:***

*Now we extracted the customer ids of the customers*

**

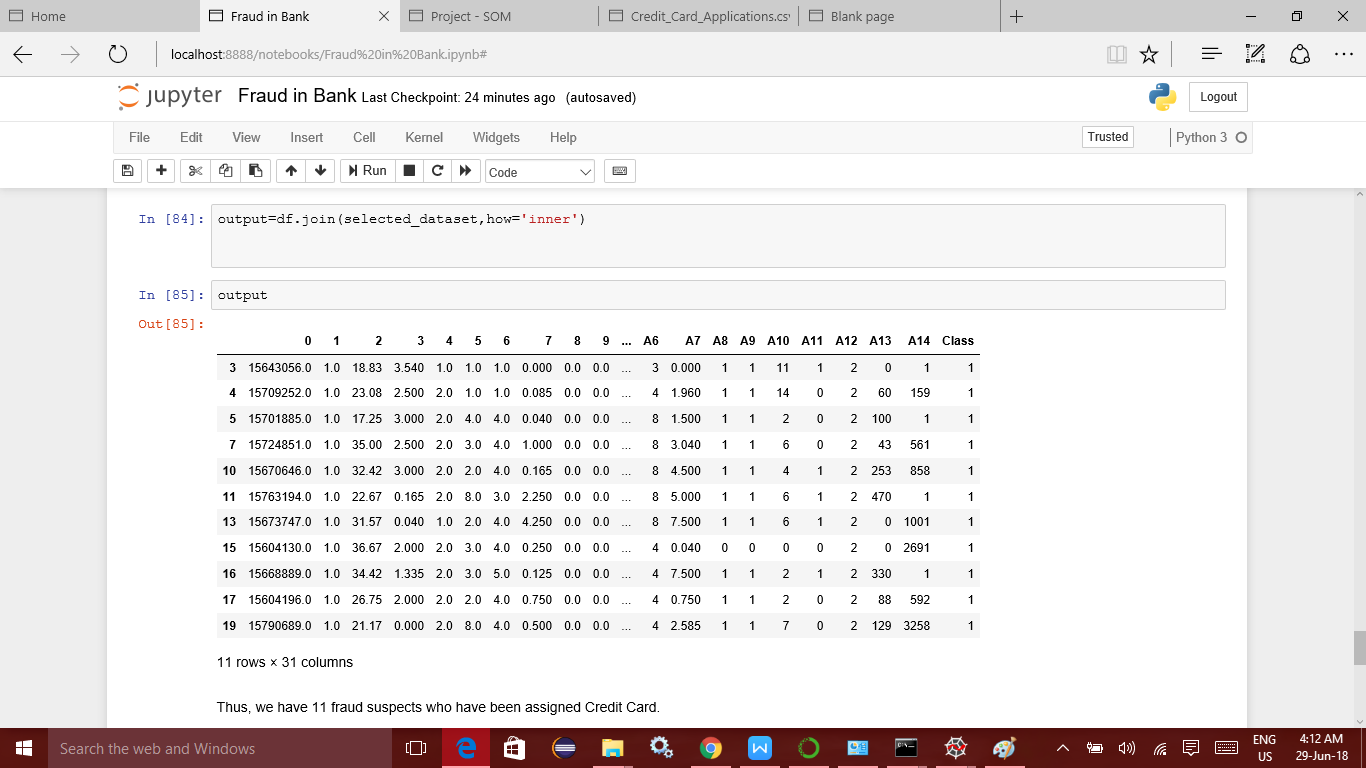
*These extracted customers are our highest probable frauds. For an extended list, we can consider frauds in nodes which were circled in blue in the previous step.*

***Step 4:***

*Now check if these customers were issued credit card or not. For this step, you will utilize the last column of the dataset, i.e. the ‘class’.*

*Class = 0 means credit card was not issued. Hence the bank is safe.*

*Class = 1 means credit card was issued. And if any of these customers got it issued, then bank should investigate it, as they are probable frauds.*

**

**Solution & Link:**

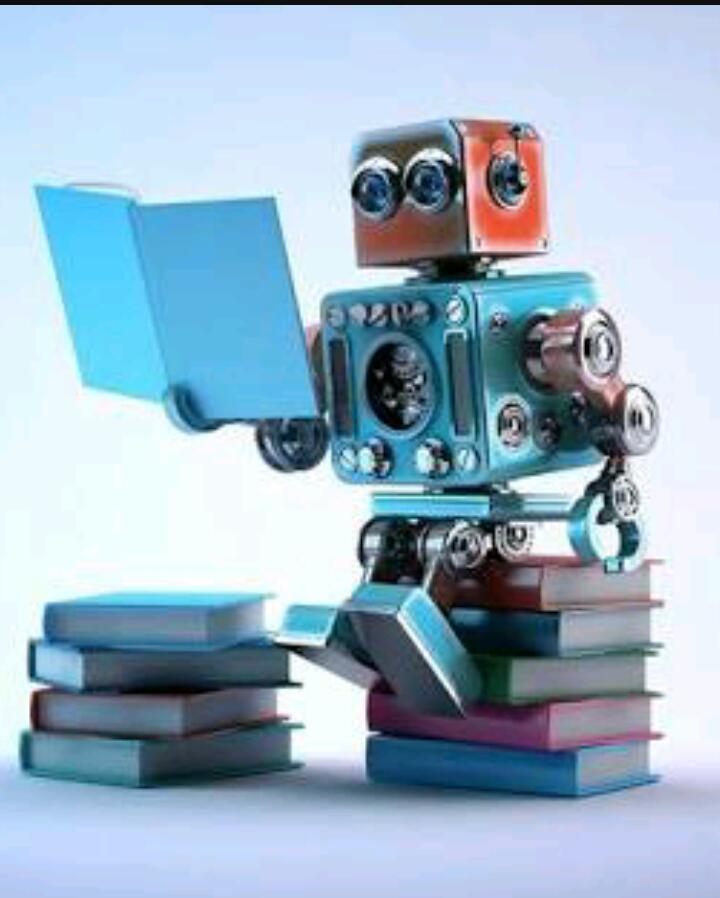
*Thus we detected 11 customers who are detected to be frauds using Self Organizing Maps (Unsupervised Machine Learning) and who have been approved to have a credit card. Thus, these should be investigated by the Bank to prevent any possibility of future losses by the bank.*

***Link to Code executed in Jupyter Notebook:***

**

**Conclusion :**

*Thus, using Self Organizing Maps (Unsupervised Machine Learning) we could detect frauds and the bank should investigate the file of these customers and cross check if the customer is indeed a fraud or not and take necessary actions.*

**