**INTRODUCTION**

* A Bank, where the customers are closing down their accounts due to certain number of factors and we don’t know which factors are most influencing ones.
* In this project we tried to investigate that which factors are responsible for the bank losing customers using ANN and experimented using different techniques so that after comparison we can get the best output possible.
* The different techniques used are:
  + Back Propagation
  + Resilient Back Propagation
  + Naive Bayes
  + Stochastic gradient descend algorithm (ADAM)

**DATA SET**

We have a Database consisting of 10000 entries. The different fields in the dataset are:

Screen Shot 2018-06-21 at 2.19.59 PM.png

* **Customer ID:**

A unique identification number, generated at time of account opening

* **Exited**:

Its our response variable and tells us whether the Customer has exited or not.

* **Estimated Salary:**

Its the estimated salary of the customer.

* **Tenure:**

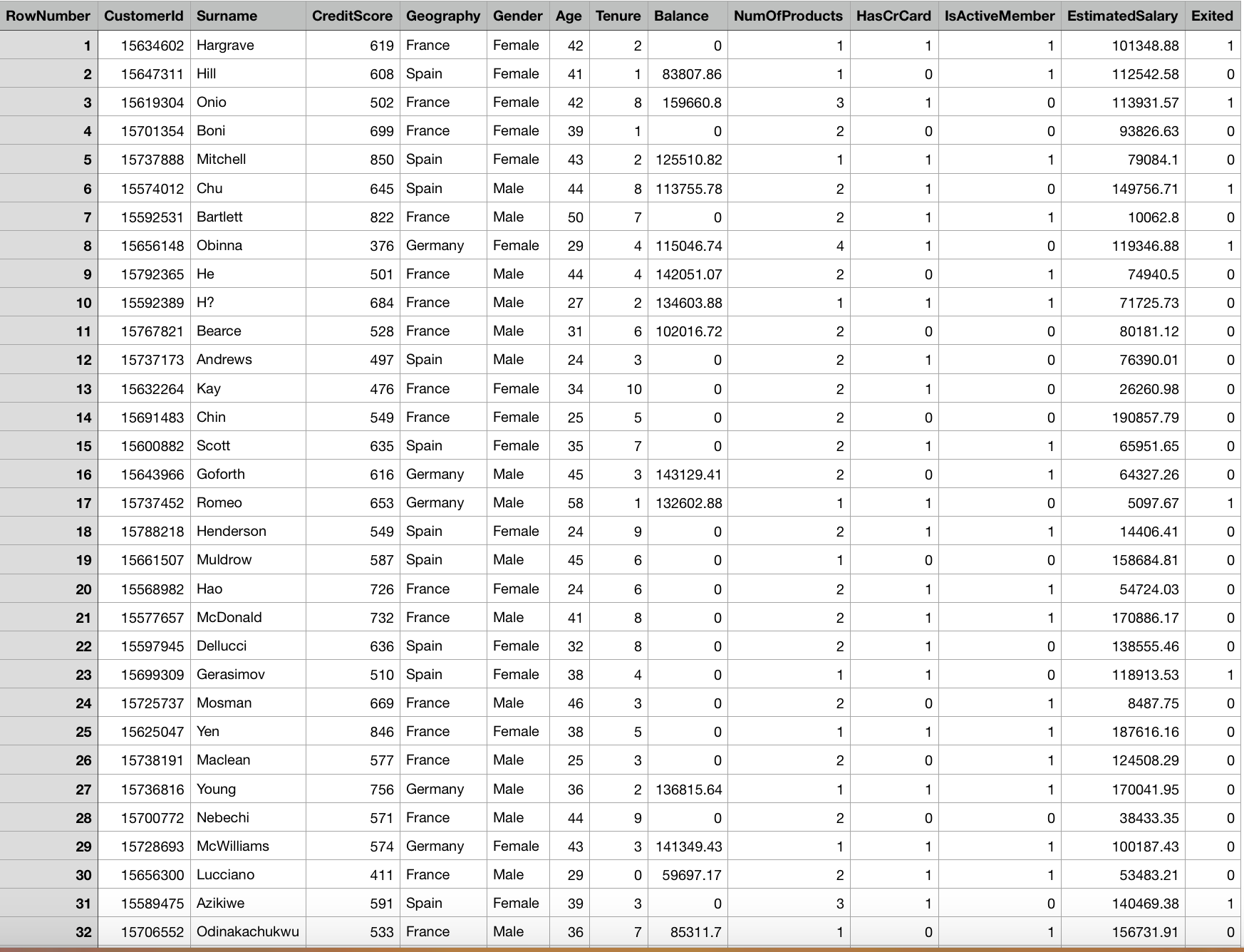
For how long the customer has been linked with the bank.

* **Has r Card:**

Whether the customer has credit card or no.

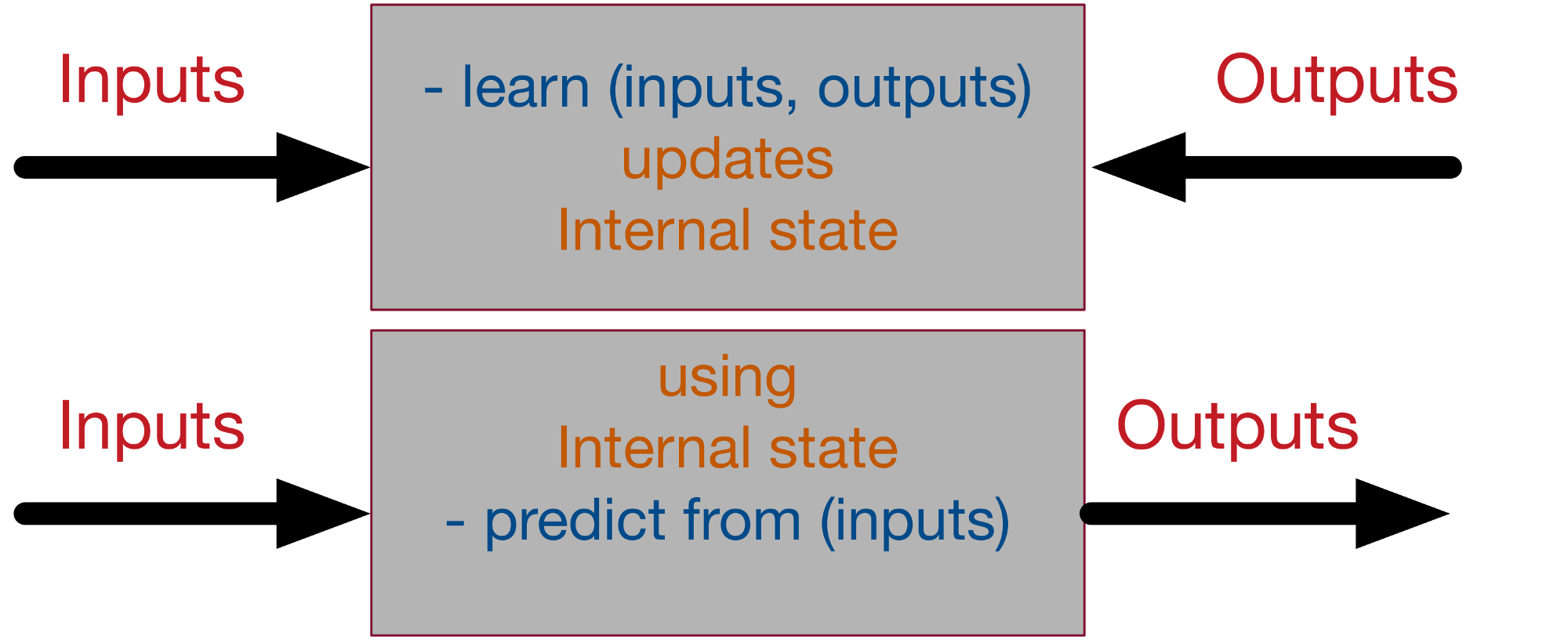
* **Number of Products:**

How many services did he avail from the bank?



**ANALYSIS PROCESS**

* Supervised learning is the Data mining task of inferring a function from labelled training data, that can be presented as a black box with 2 methods learn and predict as following;



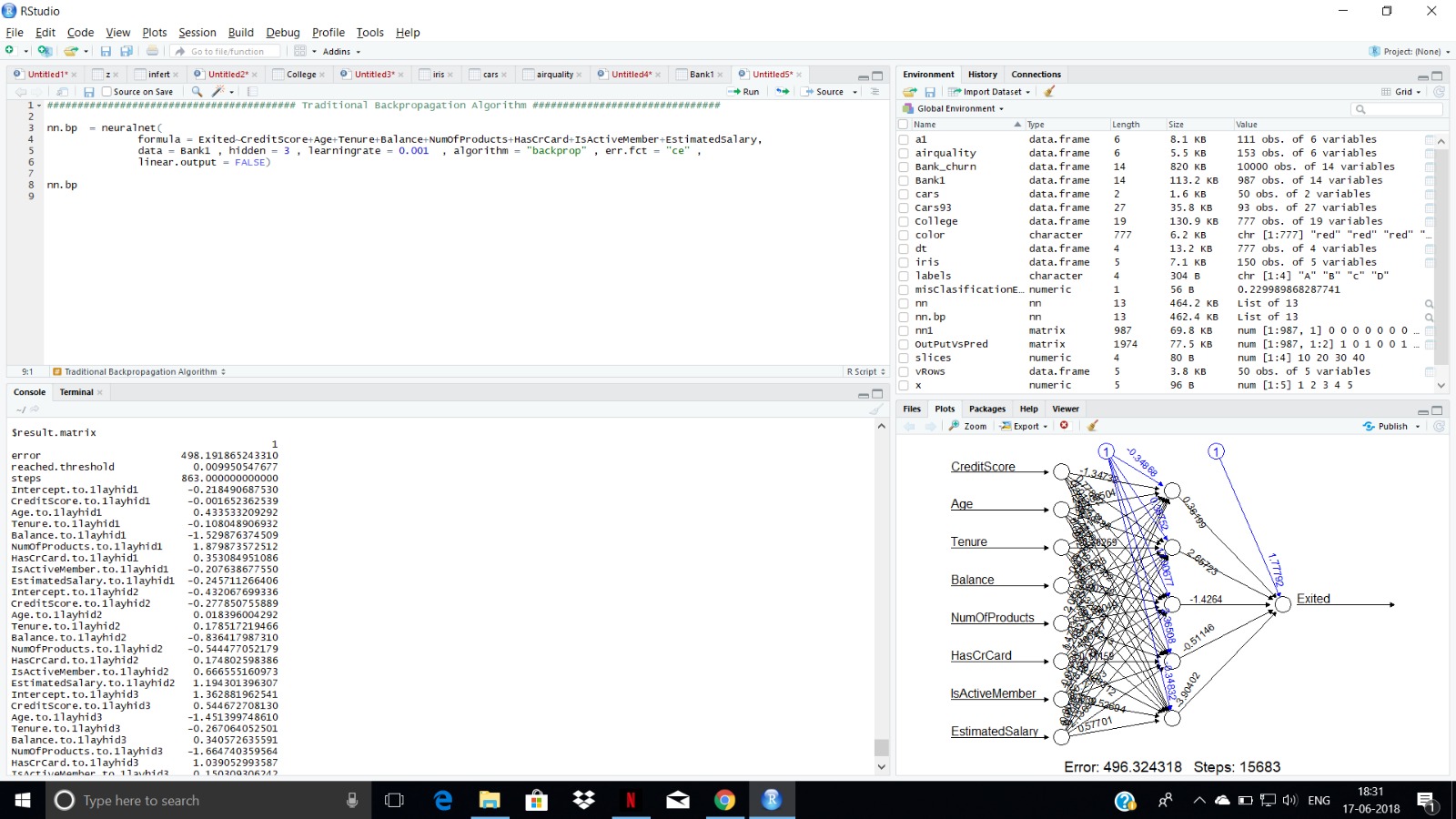
* **Classification** is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data
* input object (typically a vector) and a desired output value (also called the supervisory signal).

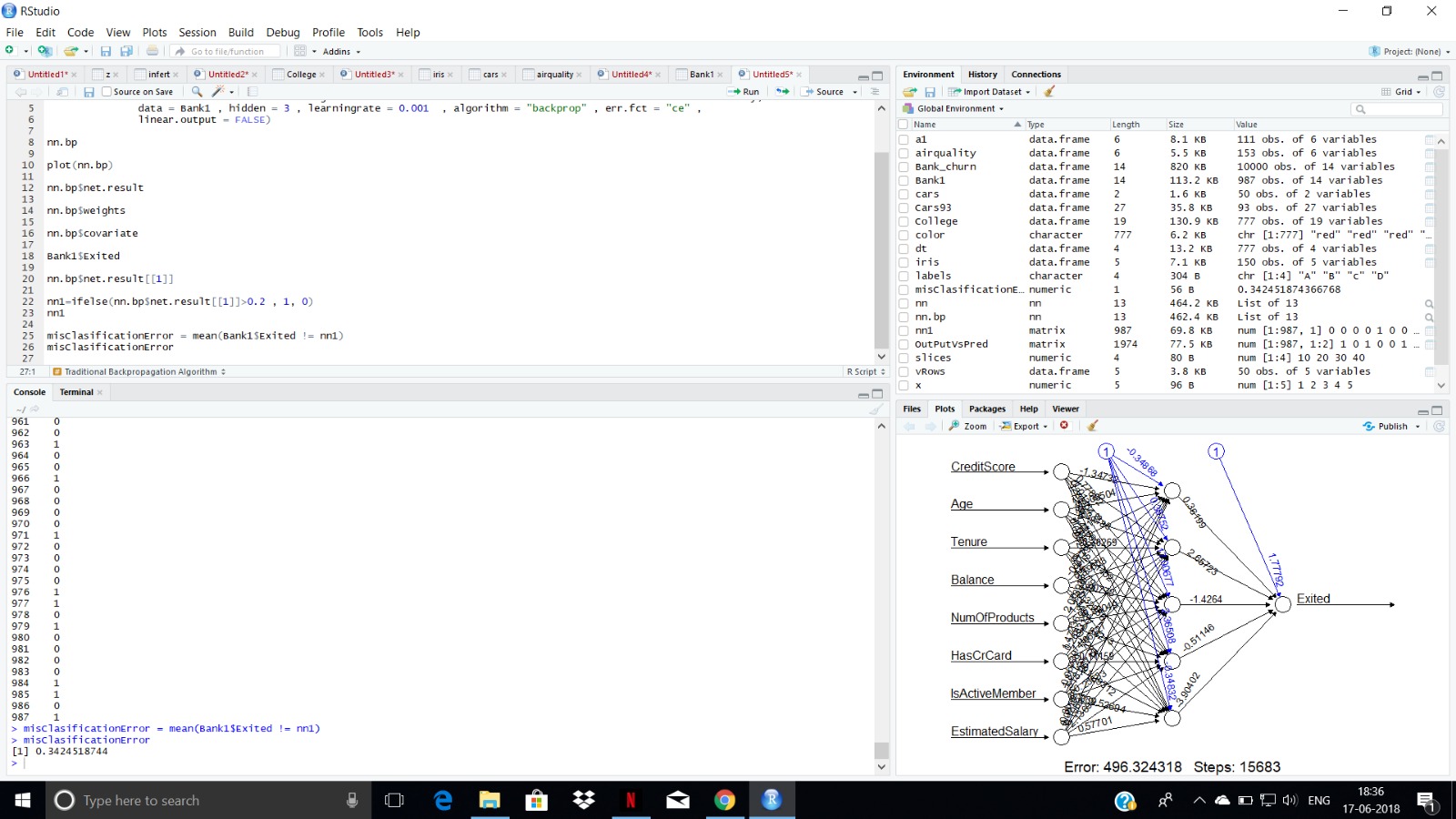
**DATA PREPARATION**

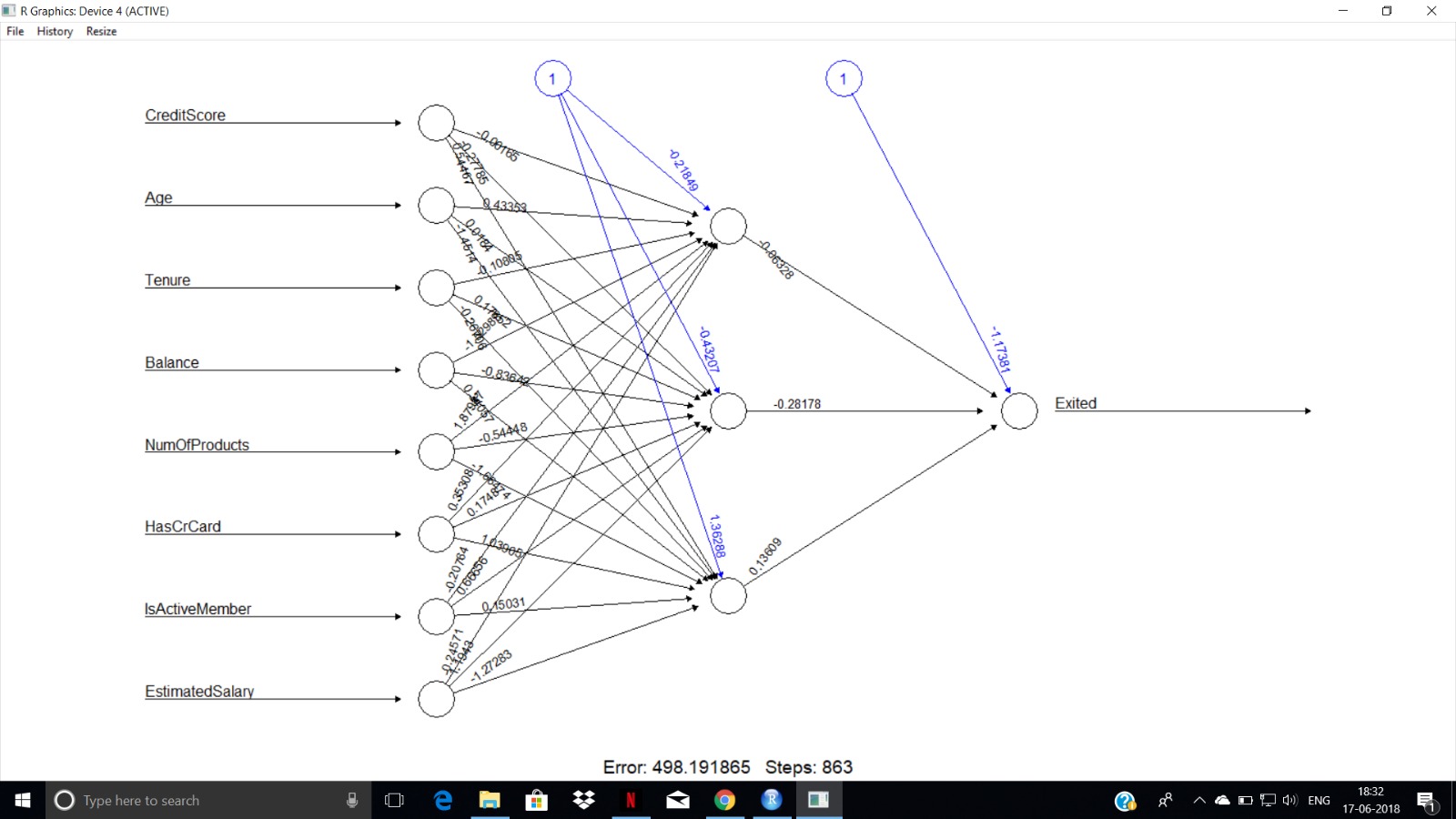
* LABEL ENCODER: Label encoding is simply converting each value in a column to a number.
* ONE HOT ENCODER: In this the integer encoded variable is removed and a new binary variable is added for each unique integer value.
* STANDARD SCALING: Gaussian with zero mean and unit variance.
* DATA SPLITTING: Is done into 80-20. 80% as the training set and 20% is used for validation and testing.

**BACK PROPOGATION**

* Back propagation is a method used in ANN to calculate a gradient that is needed in the calculation of the weights to be used in the network. It is commonly used to train deep neural network.
* Back propagation is commonly used by the gradient descent optimization algorithm to adjust the weight of neurons by calculating the gradient. This technique is also sometimes called backward propagation of errors.

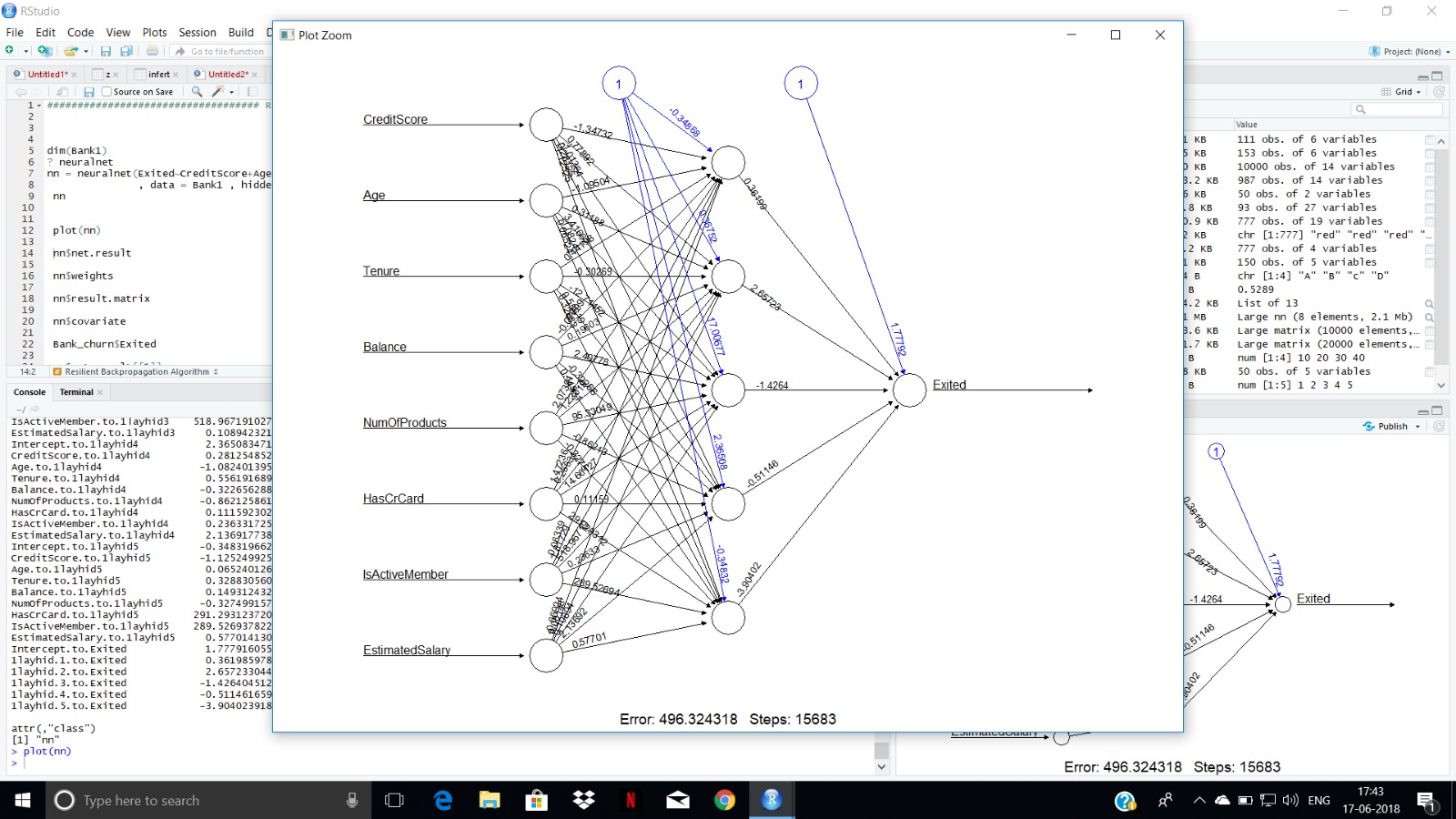


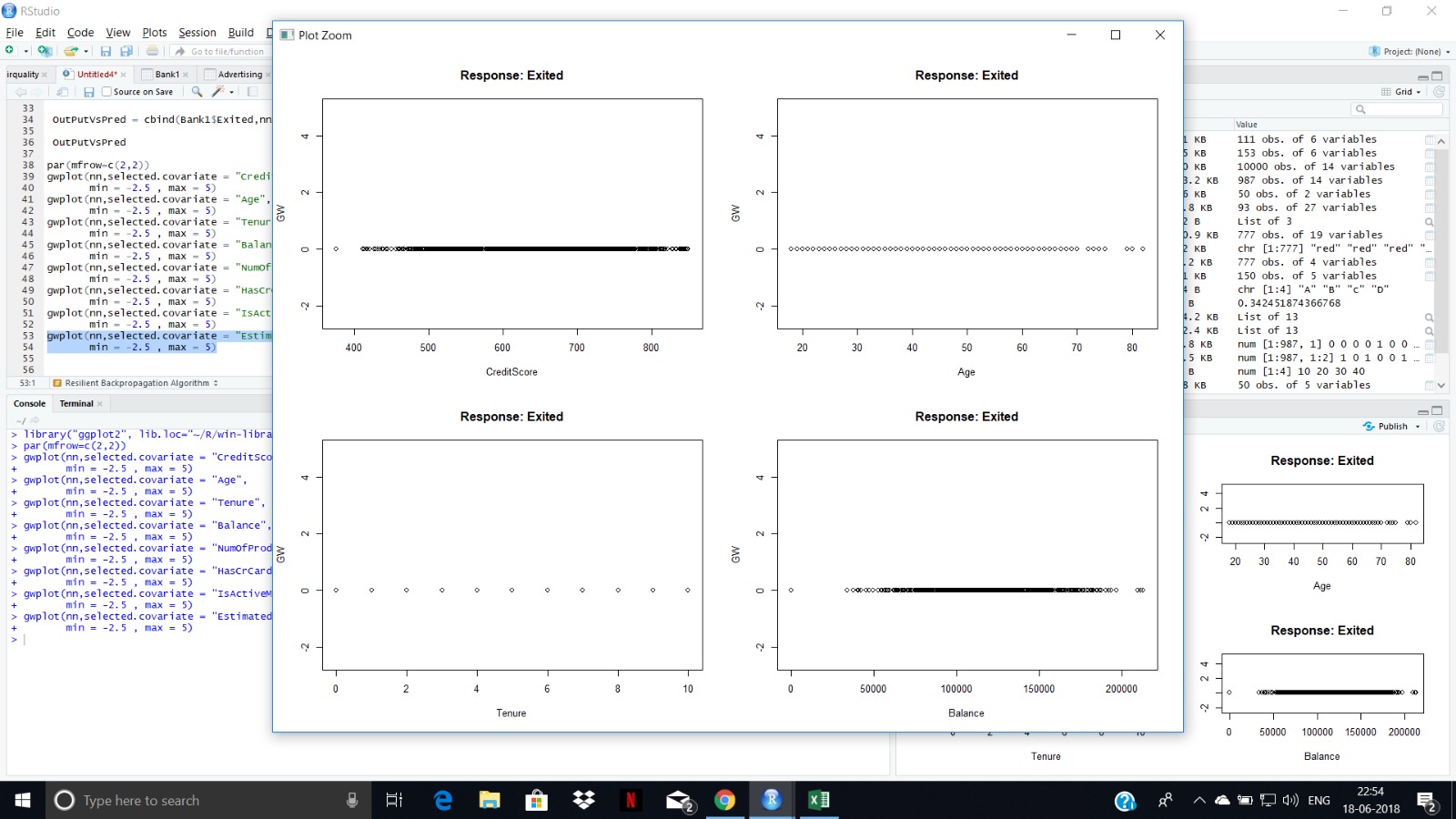




**RESILLIENT BACK PROPOGATION**

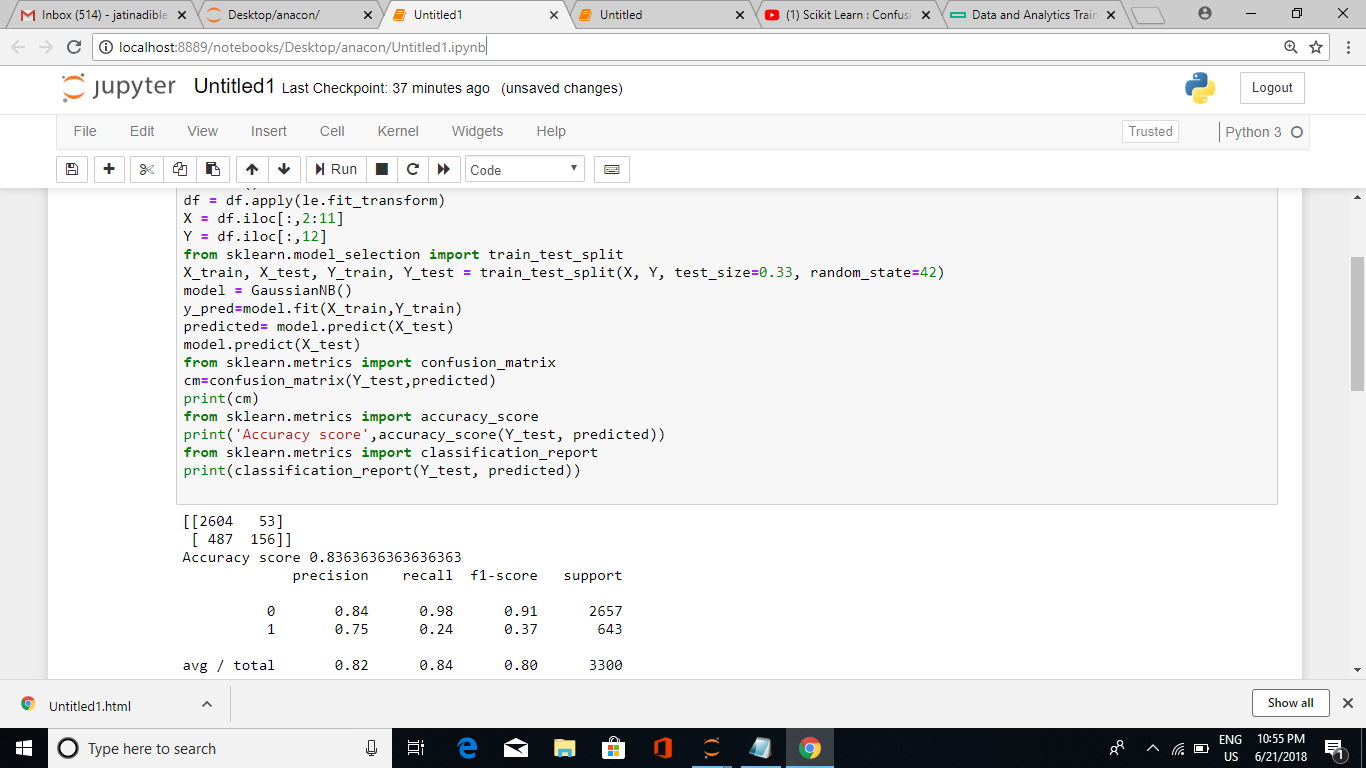
* Resilient back propagation (Rprop), an algorithm that can be used to train a neural network, is similar to the more common (regular) back-propagation.
* But it has two main advantages over back propagation:
  + - First, training with Rprop is often faster than training with back propagation.
    - Second, Rprop doesn't require you to specify any free parameter values.

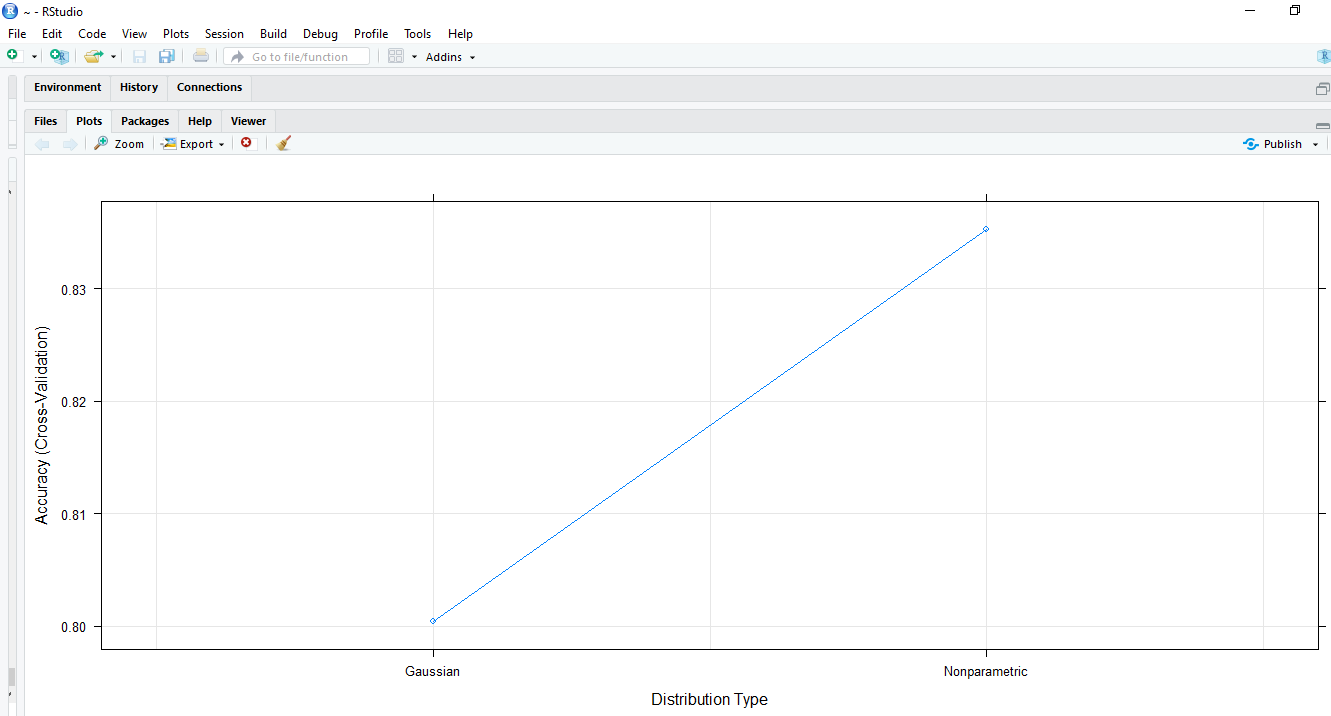


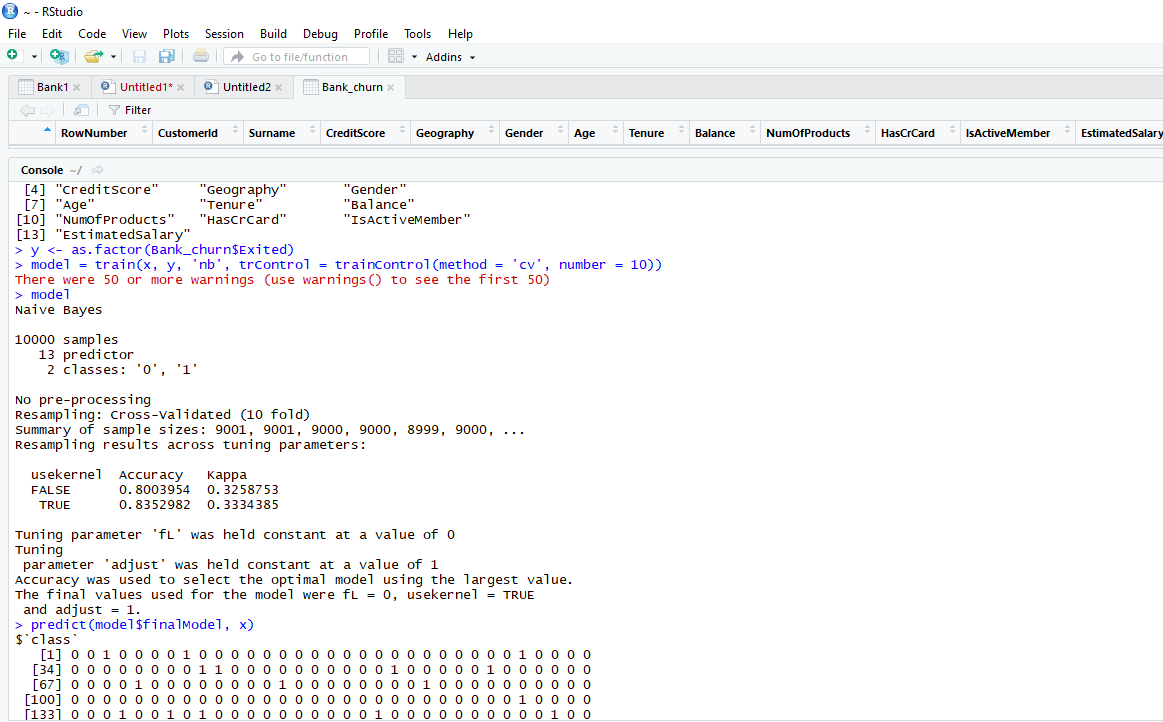


**NAÏVE BAYES**

* The Naive Bayes Classifier technique is based on the so-called Bayesian theorem.
* The model is very naively believing that the effect of the occurrence of any of the events is completely independent of the occurrence of other events.
* Naive Bayes is known to provide results at par and sometimes even better than highly complex and computationally expensive classification models.
* Super simple, you're just doing a bunch of counts and need less training data.

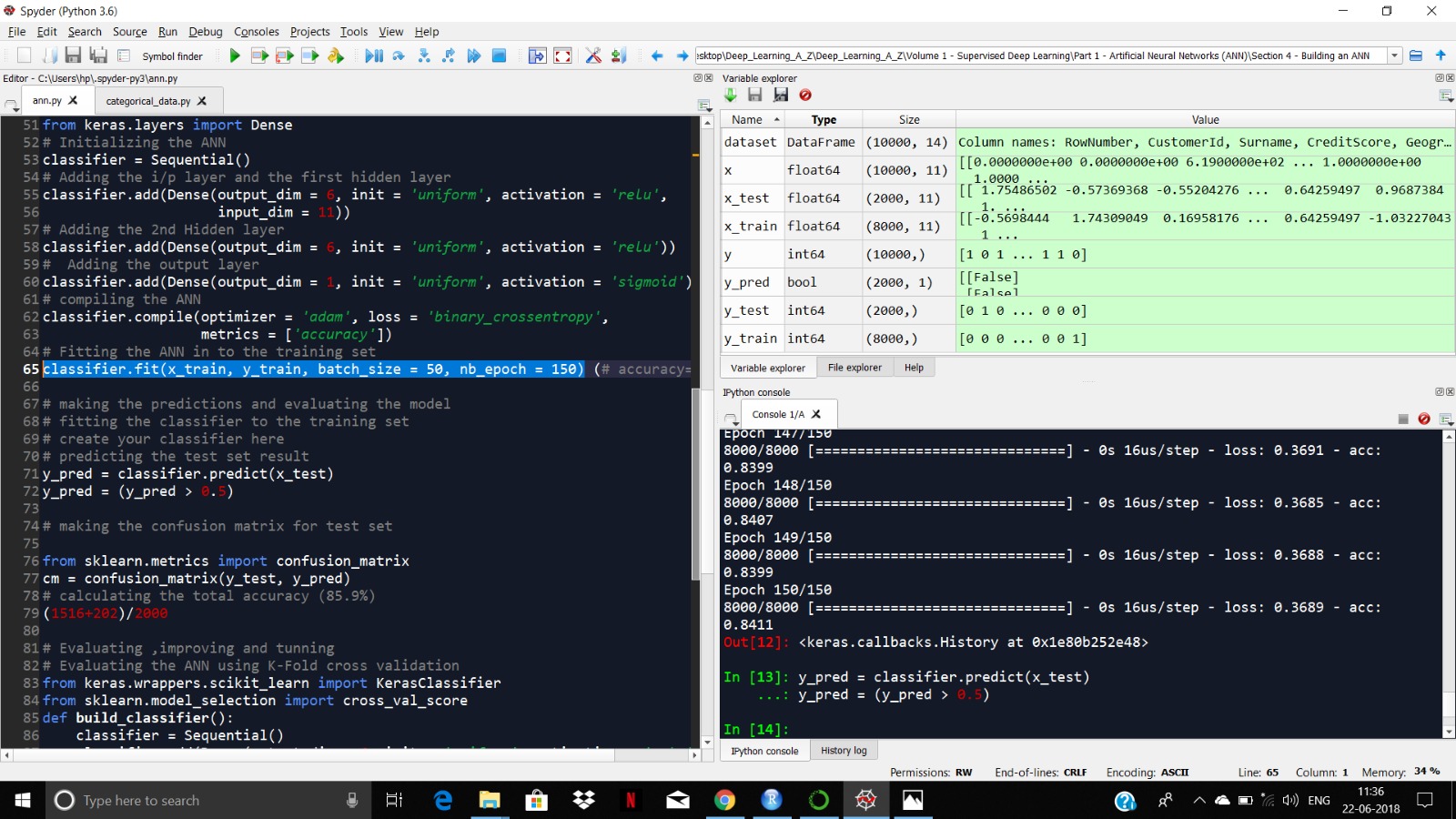


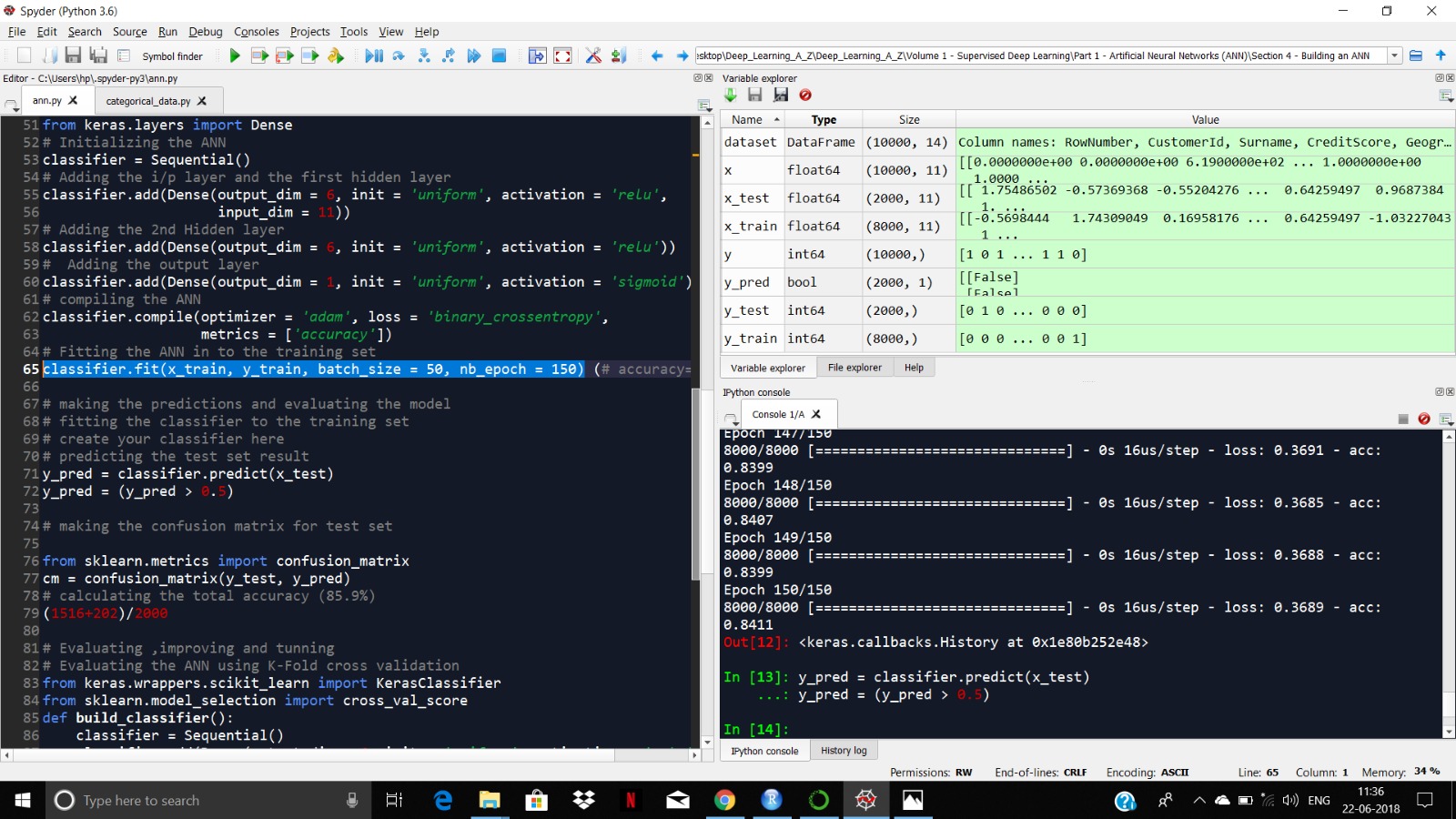




**STOCHASTIC GRADIENT DESCENT ALGORITHM**

* Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data.
* Adam is different to classical stochastic gradient descent.
* Stochastic gradient descent maintains a single learning rate (termed alpha) for all weight updates and the learning rate does not change during training.
* Adam realizes the benefits of both AdaGrad and RMSProp.





**PROJECT CONCLUSION**

* The most efficient and accurate result was given by ADAM optimiser with accuracy of 84.40%.
* Followed by Naive Bayes, which gave an accuracy of 83.63%.
* We didn’t get much precise results from Resilient Back Propagation and Back Propagation.

**BUSINESS INSIGHTS**

* This model can help the banks reduce the attrition rate of the customers leaving.
* Since number of customers and account holders add up to be the major revenue generators, any model helping in increasing the number of customers would be better for the bank.
* Bank would be knowing where its services are lacking and therefore can take the required steps thus attracting more customers.
* More customers would lead to greater revenue.

**LEARNING OUTCOME**

During my internship with HPE and NUS, I gained a lot of knowledge and experience on machine learning, Big data and Artificial neural networks. I learnt to set up various cluster, node on Linux system especially Hadoop, Hive and working live on these Hadoop and Hive clusters and using them for SQL and database learning. I was able to learn various neural networks algorithm for data analytics and machine learning that helped me gain skill set and exposure of working and learning from various faculty and students which has enhanced my knowledge and skill more.