**People Analytics and Employee Retention**

**Project Report**

Submitted by

**Arpit Khurana (15BCE0353)**

**Ishas Prasad Diskshit (15BCE0436)**

**Yash Goil (15BCE0467)**

In fulfillment for the project of

**Artificial Intelligence**

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING





**School of Computer Science and Engineering**

**DECLARATION**

We hereby declare that the project entitled People Analytics and Employee Retention submitted by us to the School of Computer Science and Engineering, VIT University, Vellore-14 in fulfillment of the requirements for the award of the project of Artificial Intelligence in Computer Science and Engineering is a record of bonafide work carried out by us under the supervision of Geraldine Bessie Amali D, Associate Professor. We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other project of this institute or of any other institute or university.

**Arpit Khurana (15BCE0353)**

**Ishas Prasad Diskshit (15BCE0436)**

**Yash Goil (15BCE0467)**



**School of Computer Science and Engineering**

**CERTIFICATE**

The project report entitled People Analytics and Employee Retention is prepared and submitted by Arpit Khurana (Register No: 15BCE0353) Ishas Prasad Diskshit (15BCE0436) and Yash Goil (15BCE0467). It has been found satisfactory in terms of scope, quality and presentation as partial fulfillment of the requirements for the award of the project of Artificial Intelligence in Computer Science and Engineering in VIT University, India.

Geraldine Bessie Amali D

Associate Professor

**Guide**

**Index**

|  |  |  |
| --- | --- | --- |
| **S No.** | **TOPIC** | **Page No** |
| 1. | Abstract | 5 |
| 2. | Introduction | 5 |
| 3. | Hardware/Software Requirements | 6 |
| 4. | Algorithm and Approach | 6 |
| 5. | Existing Models | 7 |
| 5. | Developed Model | 9 |
| 6. | Results and Discussion | 11 |
| 7. | Conclusion | 11 |
| 8. | References | 12 |
| 9. | Appendix | 13 |

**1. Abstract**

Employee attrition is very important. Companies spend lots of resources, money and valuable time to train their employees to get the best results and hence if the employee leaves the company due to some reason it is a great loss of human resource for the company. Therefore, understanding why employees leave is very important for businesses to maximize efficiency and reduce training cost which translates into greater profit. The following question is the one we would like answered: Can we predict future terminations? If so, how well can we predict?

**Dataset:**

Link of dataset: (https://www.kaggle.com/ludobenistant/hr-analytics/discussion/26565)

**2. Introduction**

Organizations spend heaps of assets, cash and profitable time to prepare their representatives to get the best outcomes and thus if the worker leaves the organization because of some reason it is an extraordinary loss of human asset for the organization. In this manner, understanding why representatives leave is vital for organizations to expand proficiency and lessen preparing cost which converts into more prominent benefit. This additionally helps in making a friendlier and vivacious work space for representatives to work in which brings about better efficiency and imparts a feeling of gratefulness for what work they are doing. Through this examination, we might want to think of the most ideal orders which could be utilized to anticipate if a specific representative will take off. From our analysis we will find what type of people leave their job. And what are the values of different parameters for these employees like they have high salary or low. With the help of our data companies can take steps to prevent their employees from leaving the company.

HR truly needs to start thinking outside of its traditional thinking and methodologies to powerfully address the HR challenges and issues in the future. On a personal level we like to think of People Analytics as when the artificial intelligence process is applied to HR information.

**3. Hardware/Software Requirements**

This requires the following hardware

CPU: Core 2 Duo/Athlon X2 or better

RAM: 1.5GB

Graphic Card: 512MB of Graphics Memory

Storage: 12GB

The software requirements for a modeling and simulation of AI needed to do this include:

**Language:** Python

**Version:** Python 2.7

**Libraries:-**

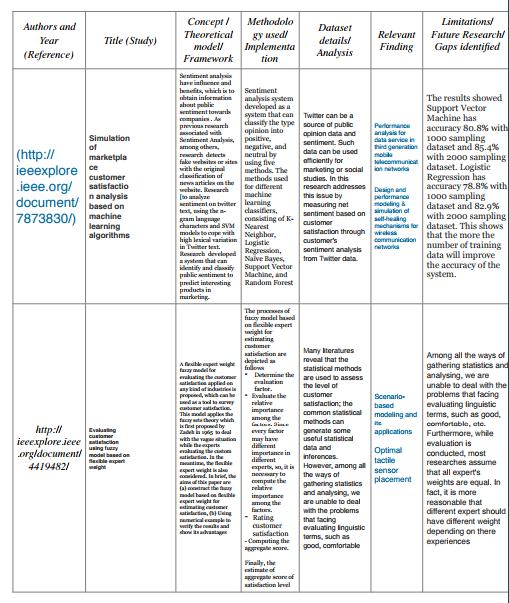
* numpy
* pandas
* seaborn
* matplotlib.pyplot
* sklearn
* sklearn.metrics
* sklearn.tree
* sklearn.ensemble
* csv
* keras.models
* keras.layers
* random
* math

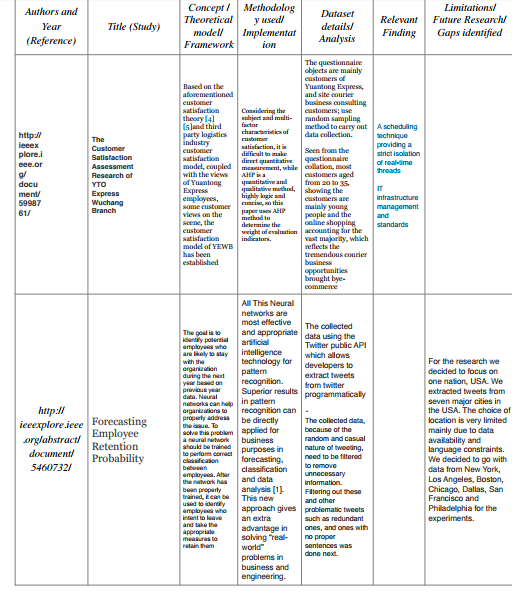
**4. Algorithms and Approach:**

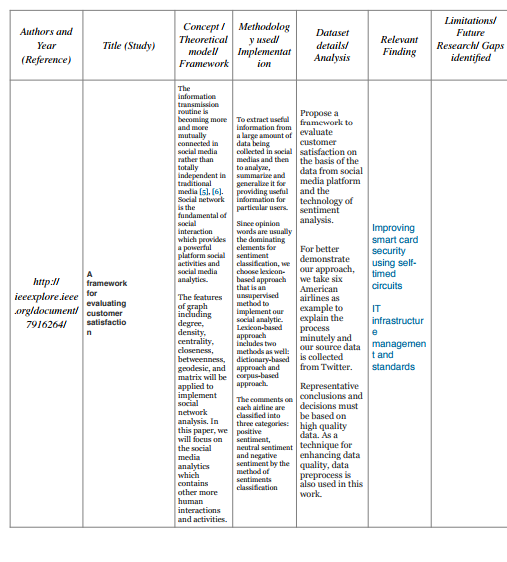
In our given dataset, we have data about the relations of every employee with each other. So, we first plot a graph which helped us to understand the different communities or group that has been built within the organizations. In our graph, the circles would represent people, and an edge between two vertices signifies that those two individuals are related to each other. This graph helped us understand if any employee feels left alone or if some employee is working with those people who are not within his domain of work. Then we also have the data of the employee about their satisfaction level, Last evaluation, Time since last performance evaluation (in Years), Number of projects completed while at work, Average monthly hours at workplace, Number of years spent in the company, Whether the employee had a workplace accident, Whether the employee left the workplace or not, Whether the employee was promoted in the last five years, Department in which they work for. We have used pandas to load the data in tabular form.

Then we obtained the number of employees who have left the company. We plotted this in a pie chart. Then we found the satisfaction levels of employee and plot it on bar graph. We also, plotted the number of work accidents by users. We plotted the number of employees who were going to be promoted and who decided to leave their department. From, these graphs and plots we concluded what type of people left their job. And what are the values of different parameters for these employees like they have high salary or low. Finally, from this data, companies can take steps to prevent their employees from leaving the company.

**5. Existing Models**







**5. Developed Model (Design and Module Wise Description)**

**I) SVM**

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

* It doesn’t perform well, when we have large data set because the required training time is higher
* It also doesn’t perform very well, when the data set has more noise i.e. target classes are overlapping
* SVM doesn’t directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is related SVC method of Python scikit-learn library.

**II) Decision Tree**

* ID3 algorithm begins with the original set S as the root node.
* On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H(S) (or information gain IG(S)) of that attribute.
* It then selects the attribute which has the smallest entropy (or largest information gain) value.
* The set S is then split by the selected attribute (e.g. age is less than 50, age is between 50 and 100, age is greater than 100) to produce subsets of the data.
* The algorithm continues to recur on each subset, considering only attributes never selected before.
* Recursion on a subset may stop, when
* All the elements in the class belong to same class
* All instances does not belong to same class but there is no attribute to select
* There is no example in the subset
* Steps in ID3
* Calculate the entropy of every attribute using the data set S
* Split the set S into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or, equivalently, information gain is maximum)
* Make a decision tree node containing that attribute
* Recurs on subsets using remaining attributes.

**III) Random Forest:**

### Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This [oob (out-of-bag) data](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm" \l "ooberr) is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

After each tree is built, all of the data are run down the tree, and [proximities](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox) are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

**IV) Artificial Neural Network (Back Propagation):**

Back Propagation has 2 phases

Forward pass phase: Computes ‘functional signal’, feed forward propagation of input pattern signals through network

Backward pass phase: Computes ‘error signal’, propagates the error backwards through network starting at output units (where the error is the difference between actual and desired output values)

* Step 1: Initialize all weights to small random values.
* Step 2: Choose an input-output training pair.
* Step 3: Calculate the actual output from each neuron in a layer by propagating the signal forward through the network layer by layer (forward propagation).
* Step 4: Compute the error value and error signals for output layer.
* Step 5: Propagate the errors backward to update the weights and compute the error signals for the preceding layers.
* Step 6: Check whether the whole set of training data have been cycled once, yes – go to step 7; otherwise go to step 2.
* Step 7: Check whether the current total error is acceptable; yes- terminate the training process and output the field weights, otherwise initiate a new training epoch by going to step 2.

**6. Results and Discussion**

In this project we have used various algorithms to analyse why employees leave. We have used SVC, Random Forest, Decision Tree and Artificial Neural Network. With SVC we obtained an accuracy of 95.6 %, with Random Forest, an accuracy of 98.8 %, with Decision Tree 98.16 % and with the Artificial Neural Network an accuracy of 96.62 %. Random Forest was the most accurate of all the algorithms.

Analysis of given Dataset

• Draw graphs for feature extraction

• Train model on the given dataset using SVC and Random forest Classifiers

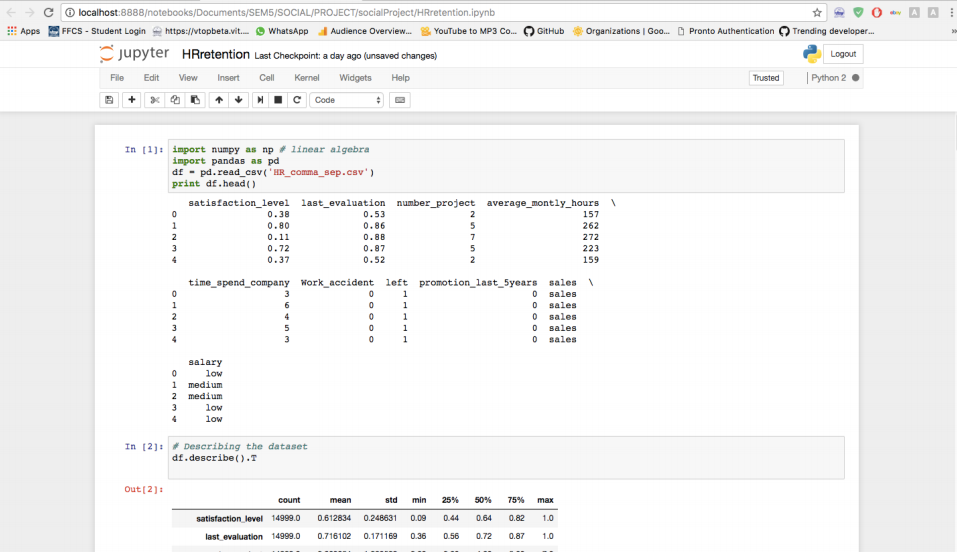
• Predict whether the employee is going to leave job or not, based on our trained model.

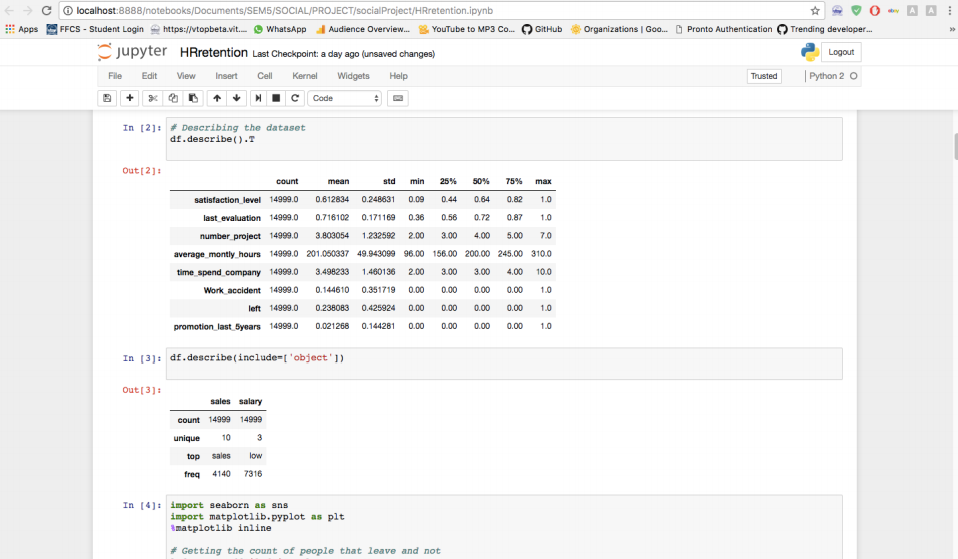
• Identified what may be the consequences, he is facing, which will be useful for employee retention.

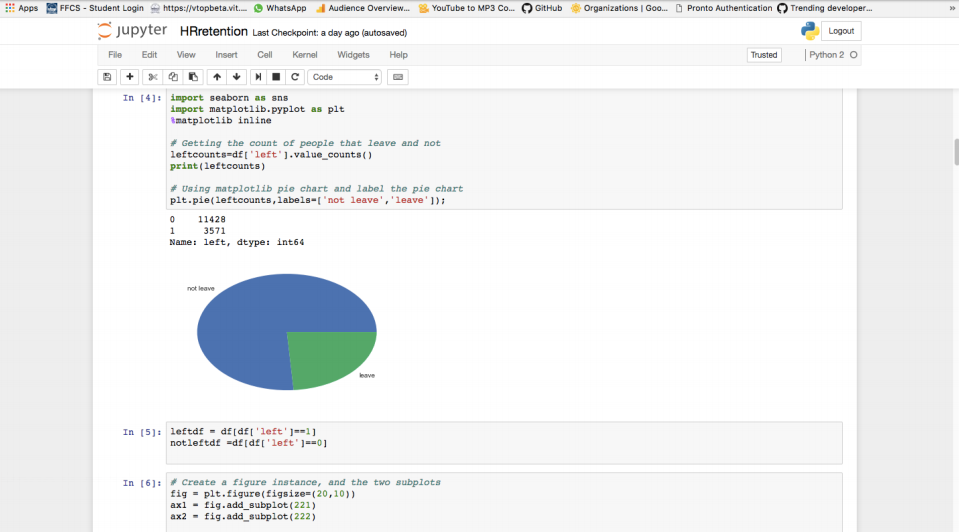
**7.** **Conclusion**

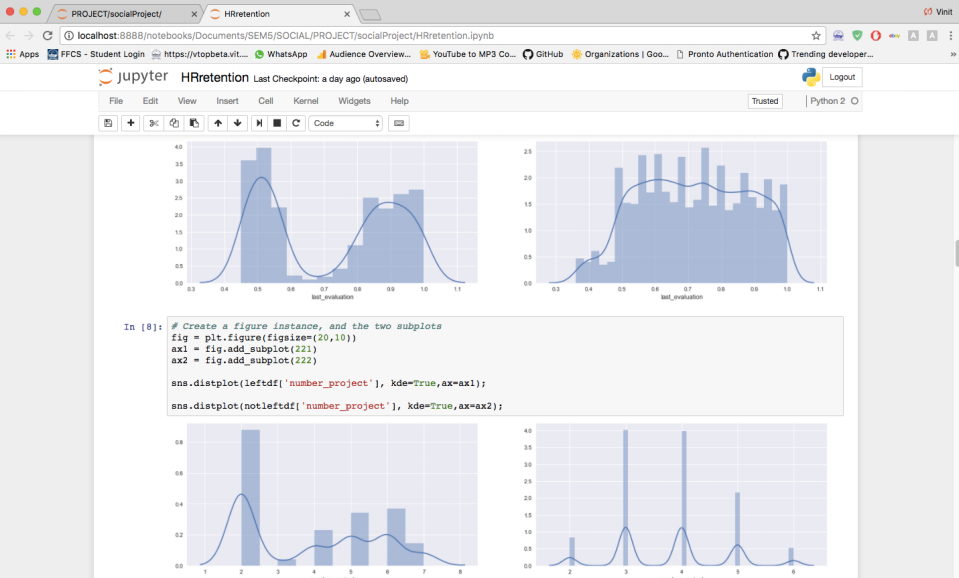
From our analysis we found what type of people leave their job. And what are the values of different parameters for these employees like they have high salary or low. With the help of data companies can take steps to prevent their employees from leaving the company using algorithms in machine learning like SVC, Decision Tree, Random Forest and Artificial Neural Network.

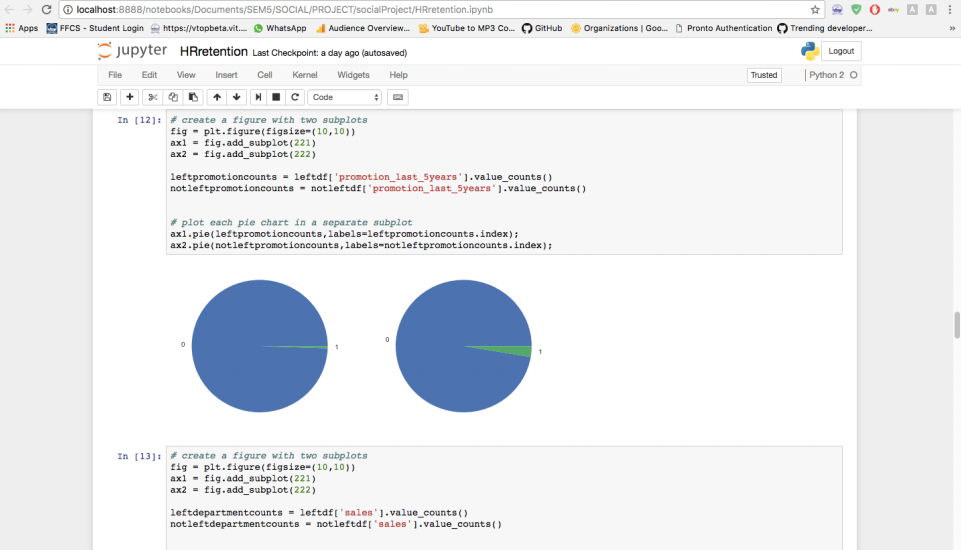
**Appendix**

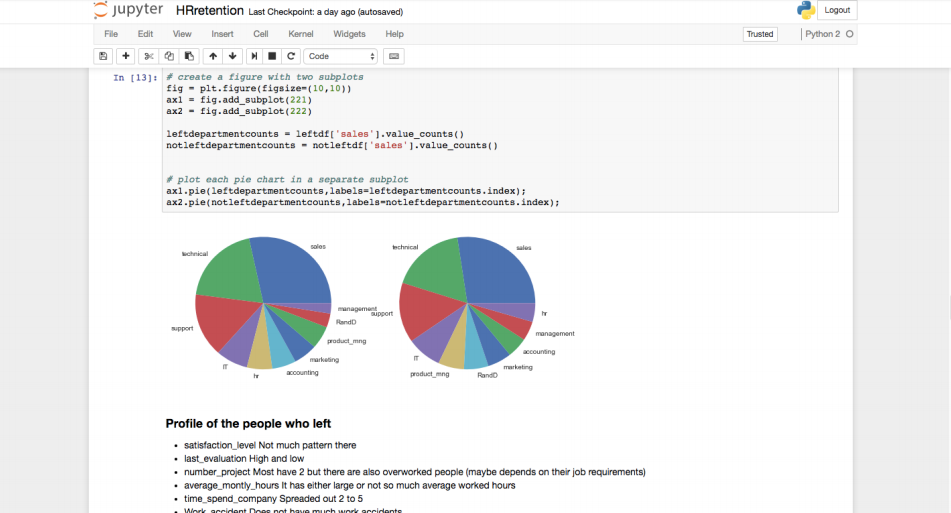


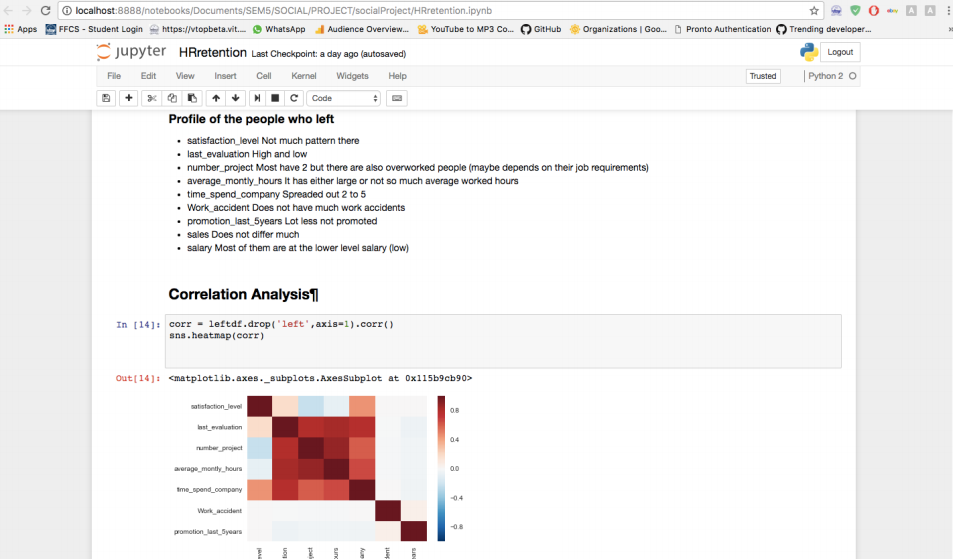


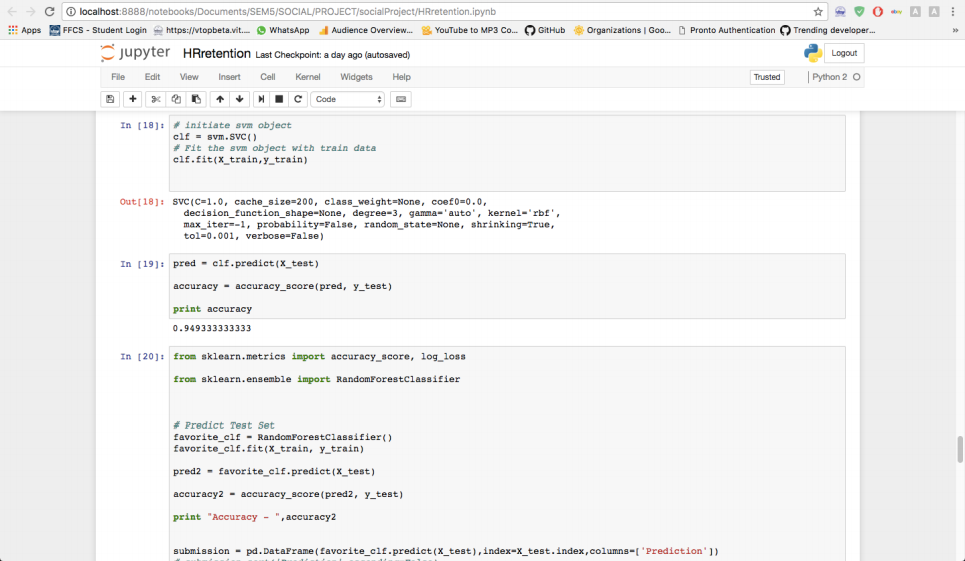


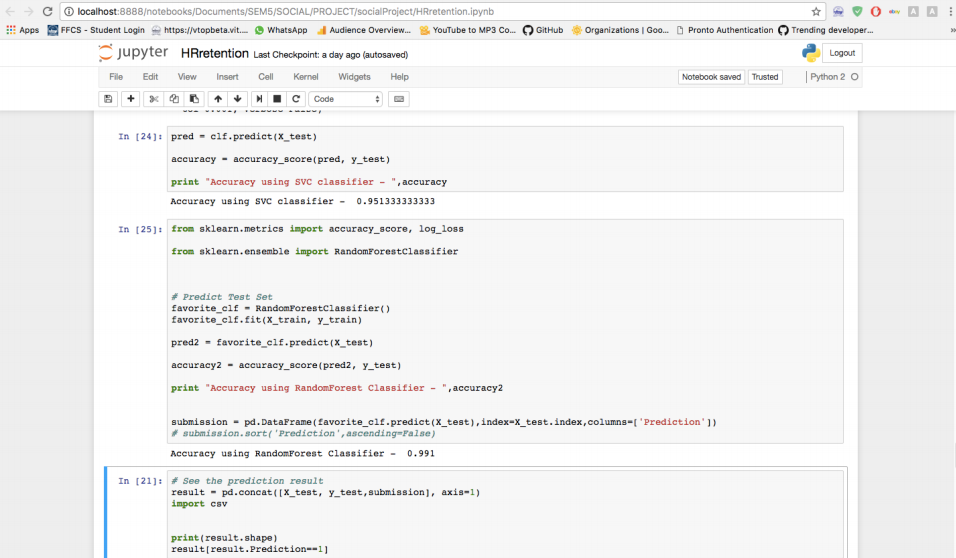


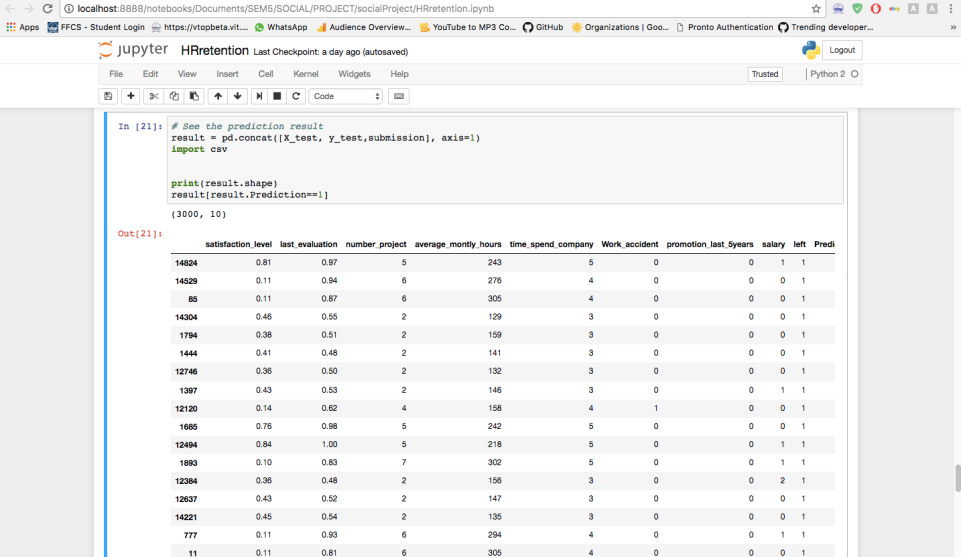


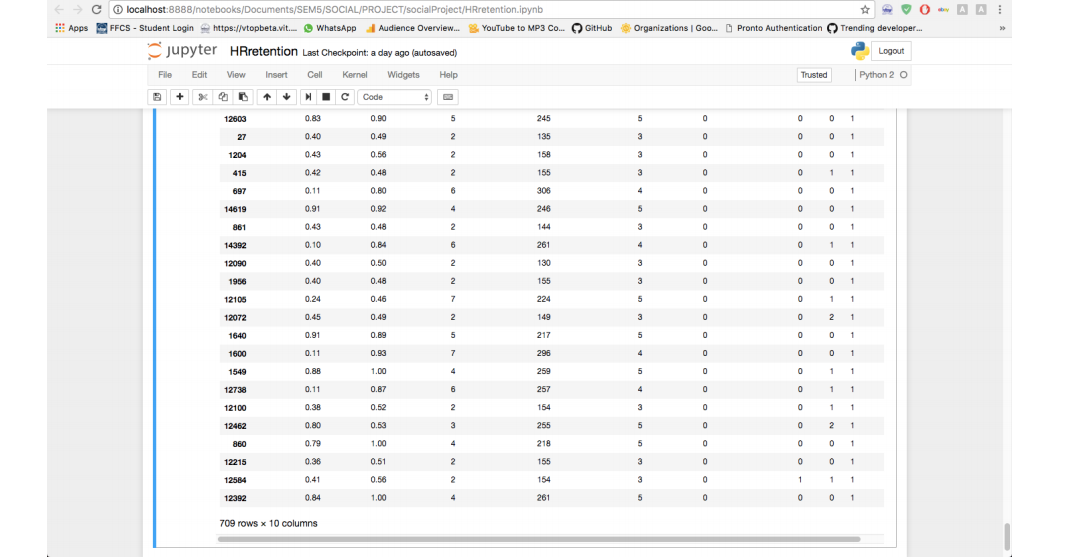




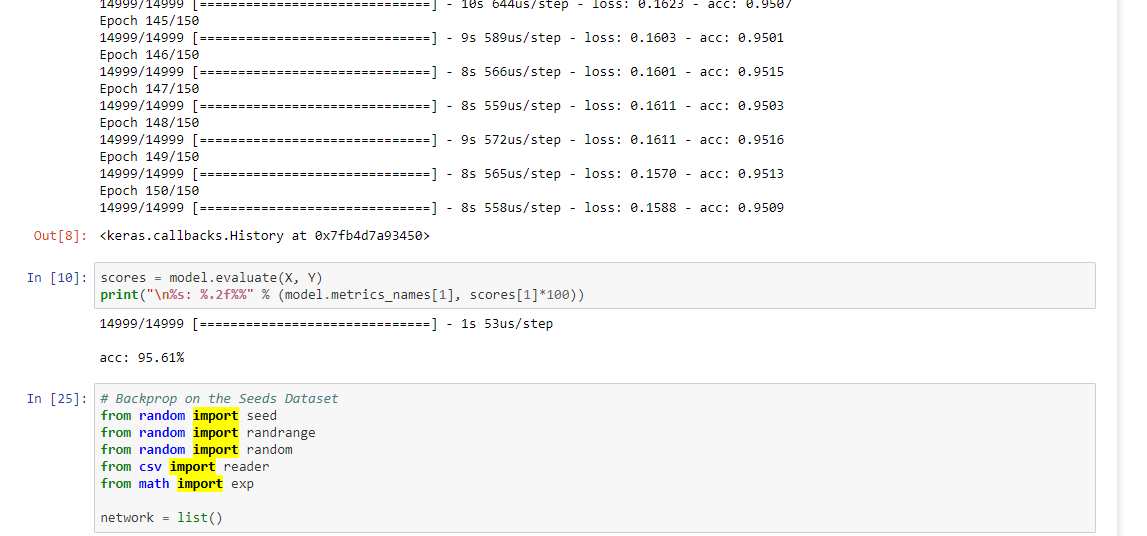












****

****

****

**References**

i) Serrat, Olivier. "Employee performance analysis." Knowledge solutions . Springer Singapore, 2017. 39-43.

ii) Marcus, S., Melanie Moy, and Thayne Coffman. "Employee performance analysis." Mining graph data (2007): 443-467.

iii) McGloin, Jean Marie, and David S. Kirk. "HR Retension analysis." Handbook of quantitative criminology . Springer New York, 2010. 209-224.

iv) <https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/>

v) https://medium.freecodecamp.org/building-a-3-layer-neural-network-from-scratch-99239c4af5d3