HR Analytics Project

# Contents

1. [Problem Definition](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#c099)
2. [Data Analysis](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#6270)
3. [EDA Concluding Remarks](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#030a)
4. [Pre-processing Pipeline](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#dd40)
5. [Building Machine Learning Models](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#8341)
6. [Concluding Remarks](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#13e2)

Let’s start with the problem definition or a short introduction on the project that I have chosen to elaborate and why it was made in the first place.

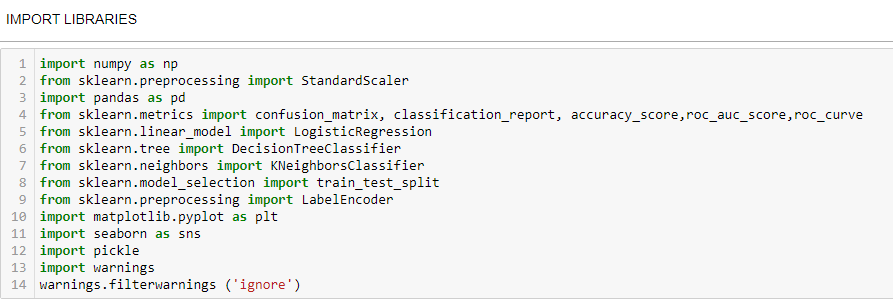
**1.ProblemDefination**

The Project I have opted to write over is HR Analytics Project that is basically fictional Project and It was created by Data Scientist In IBM. You can find the whole dataset in kaggle . Basically there so many companies who hired so many new employee every year and organise training program inside the company to the new comers and also for old employee to enhance the efficiency. But my concern through this project is where does HR fit in all of these.

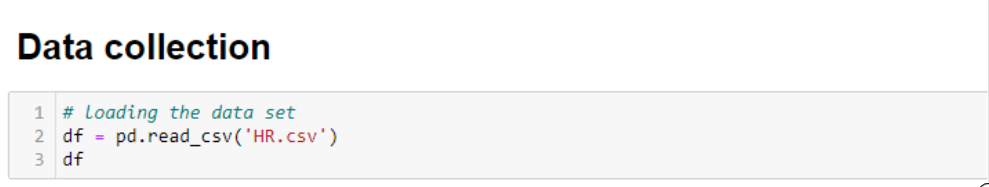
Human Resource Analytics (HR Analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on

Attrition in human resources refers to the gradual loss of employees’ overtime. In general, relatively high attrition is problematic for any company. HR professionals often assume a leadership role in designing company compensation programs, work culture and motivation systems that help the organization retain top employees. How does attrition affect companies and how does HR Analytics help in analysing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.  
  
Attrition affecting companies is a major problem since high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork and new hire trainings are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits an organization from increasing its collective knowledge base and experience over time. Errors and issues are more likely if you constantly have new workers too.  
  
Therefore the major goal of this project is to identify the “Attrition” rate as a simple Yes

No tag making this to be a classification problem!

****

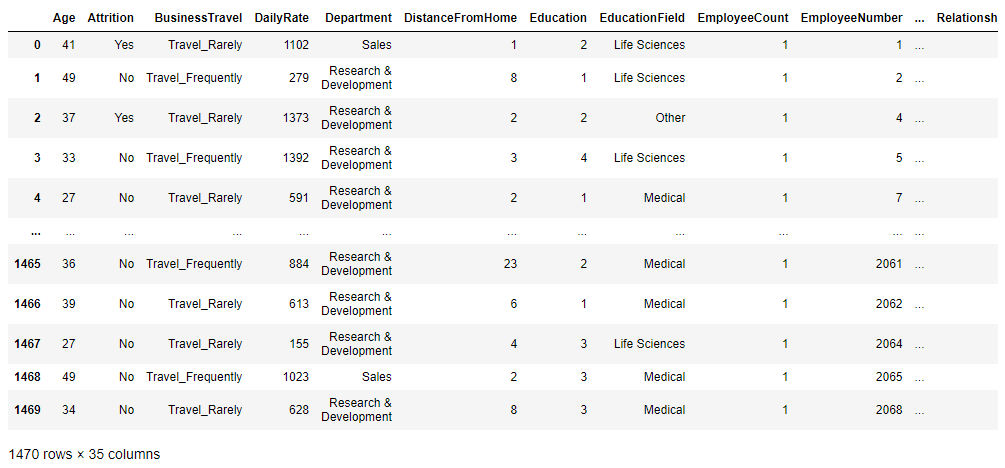
First of all I have imported all the necessary libraries or tools that is sure going to come into help of building best machine learning model. I have to import whole the dataset into single jupyter notebook that we can apply upon various method and machine learning Algorithm.



From the above read\_csv method we have imported entire dataset into jupyter notebook and stored into single variable name as df.

**2.Data Analysis-**

In this section of our project we will look at every columns of our data set and we will try to figure out which column is truly related to solve of problem statement or not .



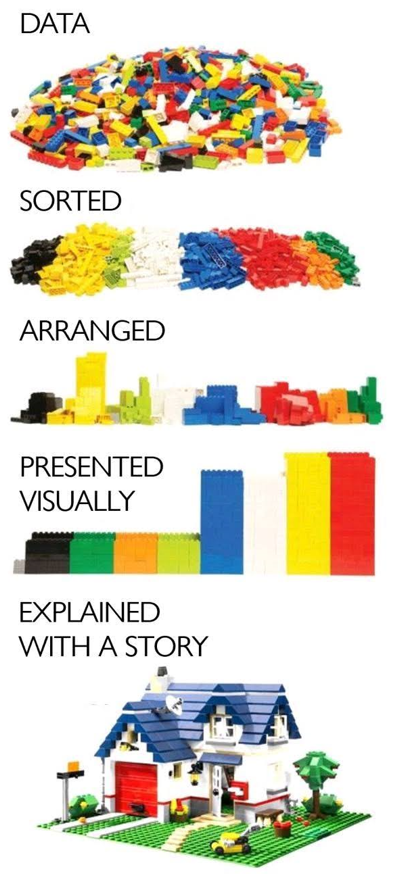
As we have dataset having 1470 rows and 35 columns below .in this 35 columns one is our target means Attrition and rest are features.

**3.EDA**

Without any hesitation we can say that EDA, that is Exploratory Data Analysis is heart for building of machine learning model. This is the utmost important part that every dataset must be go through it. EDA is process use various of procedure or method to make our dataset into appropriate format so we can achieve our real target. In this EDA we do entire dataset analysis via using various using tools and python libraries.

From the below picture, i can explain:

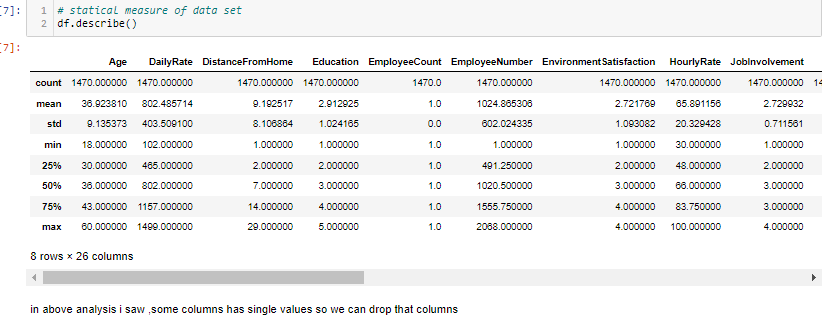
1. For any sort of Machine learning Model data is the most important thing that you must have. There are so many to collect the data. I will not go into depth here but I can say primary and secondary there two are the main sources of data collection
2. Then I sorted the entire data as per their features and Arrange into some format
3. By the different means of visualization technique, I have visualized the data and find a way to explain the whole story.



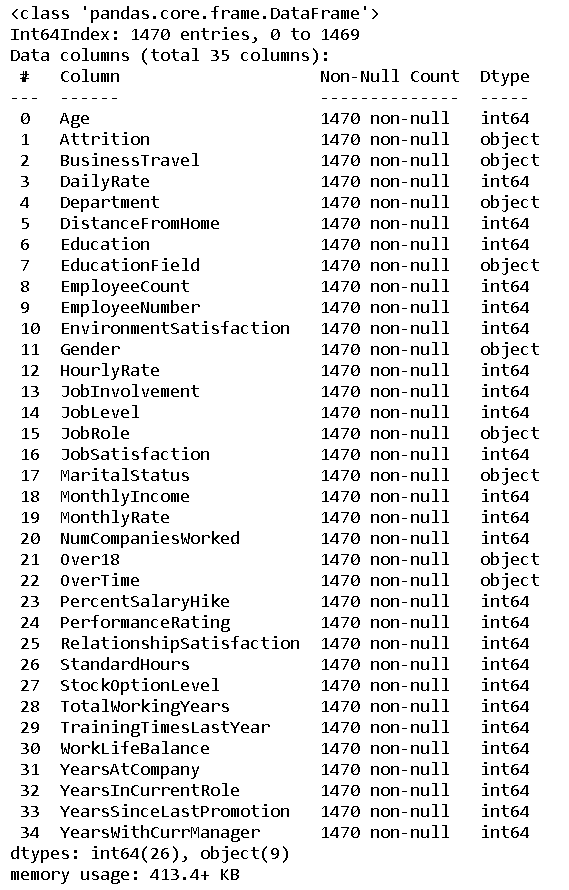
here is null() method with sum function allow us to figure out the missing value present in our dataset .



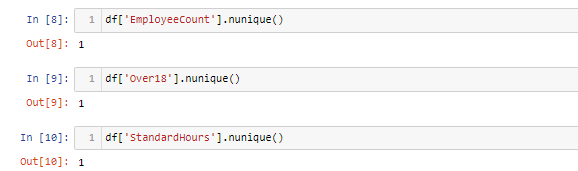
Now, fortunately we don’t have any null value present in dataset

from the above describe transpose method it is displaying that there is no missing data as every columns have same count (1470) And maybe columns like Monthly income, Total working years, Years At Company ,Years in Current Role ,Year Since Last Promotion , Year With CurrManager has huge difference in their 75 % and 100% so they must have some amount of outlier that is need to be treated.

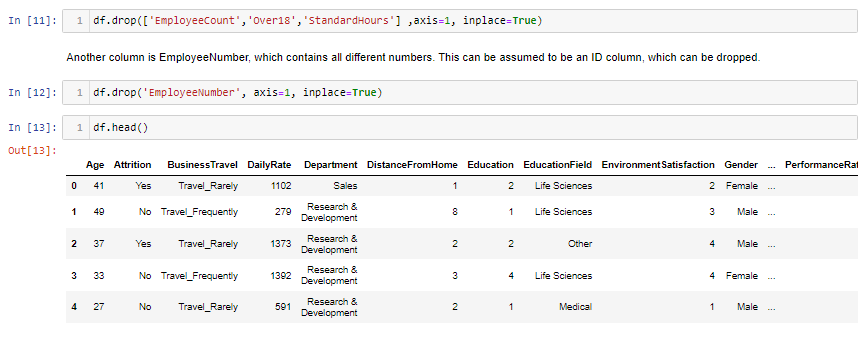




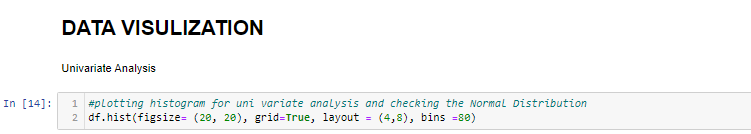
This df.info () method gave us all information regarding of types of our dataset.

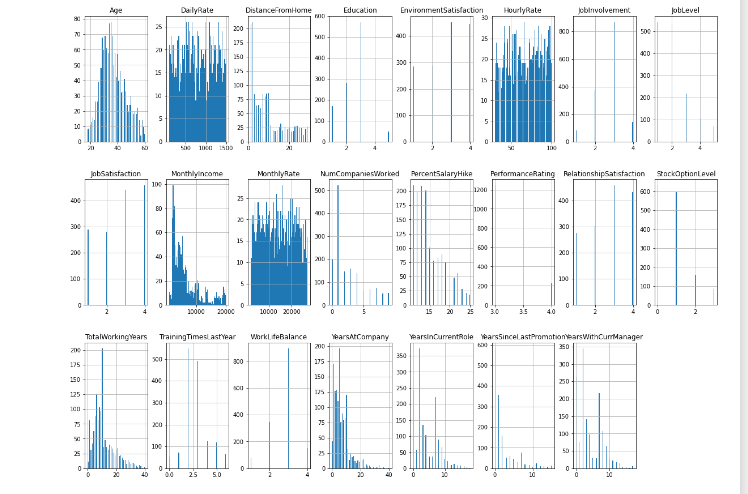


This df.nunique method has given us the unique value present into that columns of our dataset which we want.



Here we drop that columns which we do not want

Now we are going to plot UNIVARIATE ANALYSIS by plotting hist plot



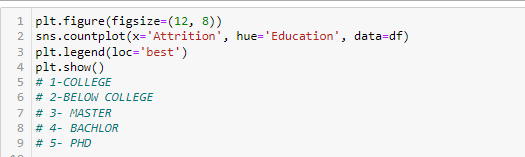
By hist plot we can check data distribution and skewness

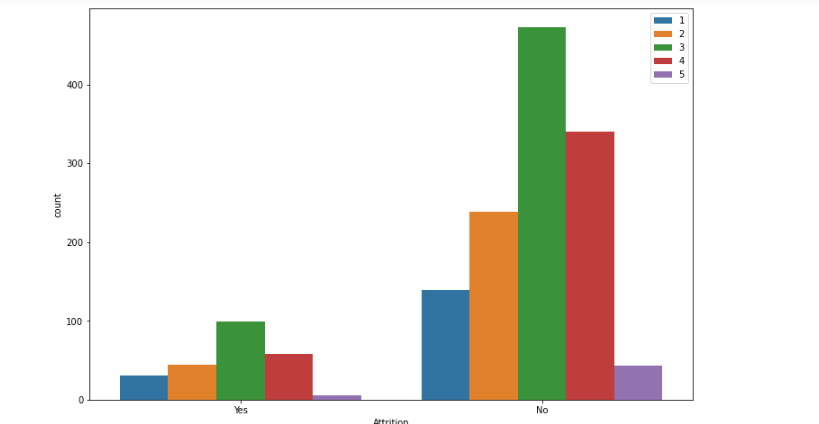
Now I am going to multivarirate analysis using bar plot



SO WE CAN SEE THAT IS JOB SATISFACTION IS HIGH ATTRITION IS LOW

Now I am going to check What level of education has higher chance to leave the job?



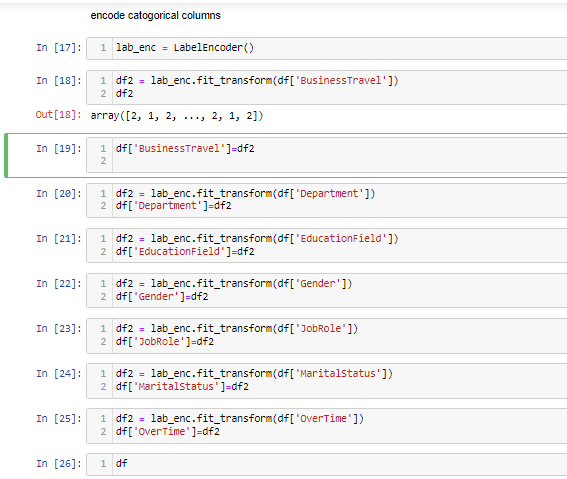


Look likes those who are not graduated from university are likely to leave. Phd employees seem not to look for a new role, this is properly they already have very high rank in the organization.

**4.Pre-Processing Pipeline**

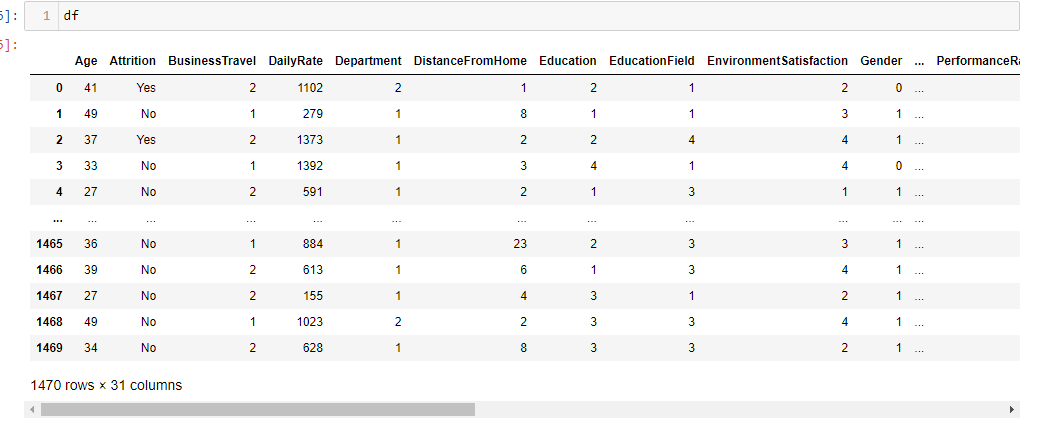
Data processing is a process of preparing the raw data and making it suitable for it a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating machine learning project it is note always a case that become across the clean and formatted data and while doing any operation with data it is a mandatory to clean it and put in formatted way. So, for this we use data pre-processing task.

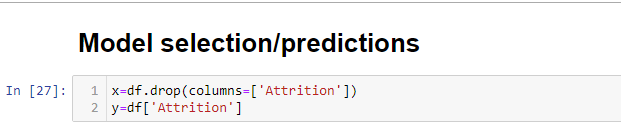


We can use here any of the technique of encoding but I am here using LABEL ENCODING TECHNIQUE

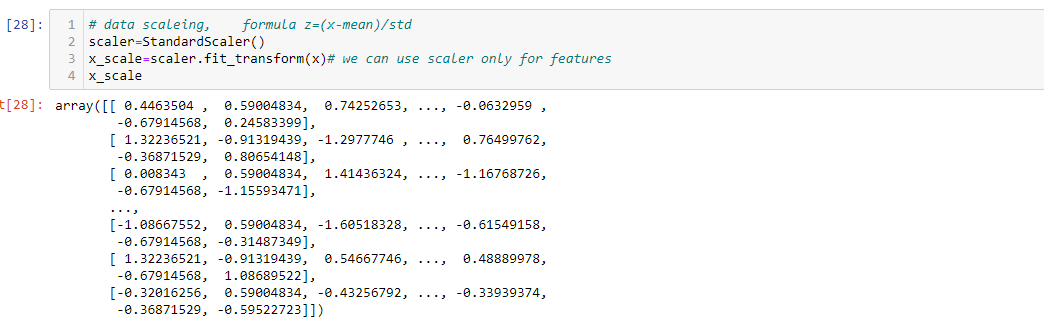
Bcoz it is easy to understand and simpl to learn.



now my data looks good and ready to perform other operations.

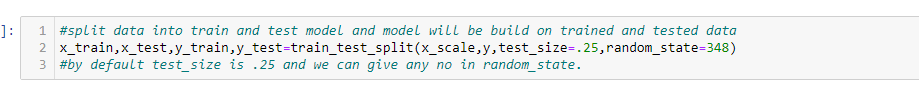


Above we can see , we break data into x and y x belong to features and y belongs to label



From the abovestandardScaler method we can scale that all the columns of our dataset so our dataset refrain from biasing for any particular column.

To find the best Random State for our Machine Learning Model .from the above I have sent 25% data for testing and 75% data for training. Actually, It’s up to you how precisely you want to make your model is.

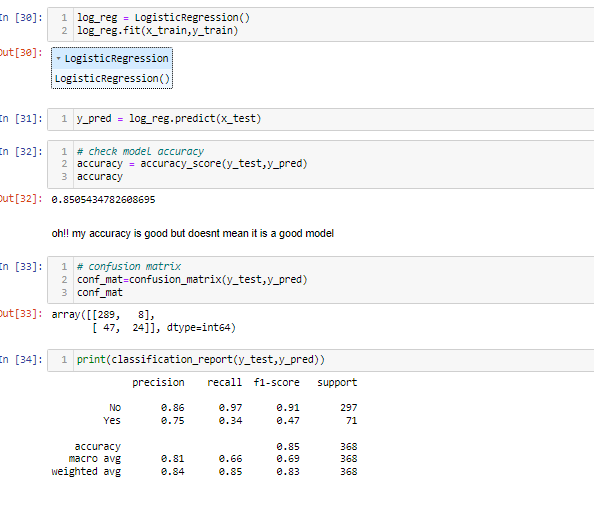


Before building the proper Machine learning model we have to decide how much data do you want to send as test and how much data do you want for training .It totally up to you.

**5.Building Machine Learning Models**

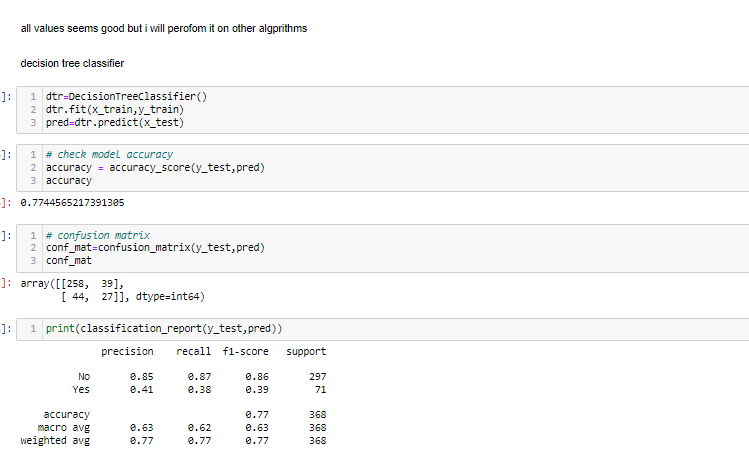
In order to build a classification method I have imported the necessary libraries and created a function that contains our entire machine learning model creation and its evaluation metrics steps. This makes our job easier since later on we just need to feed the model’s name and get the result without repeating/rewriting the same code again and again.

Firstly I am going to build model with logistic regression algorithm



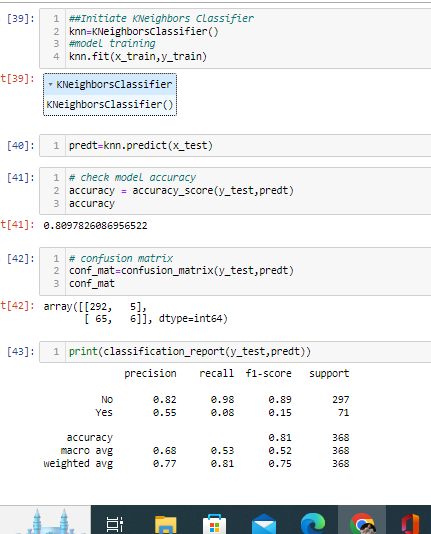
as we can see that logistic reggression has good results till now logistic reggresion is good but i will try another algoritm

**decision tree classifier**



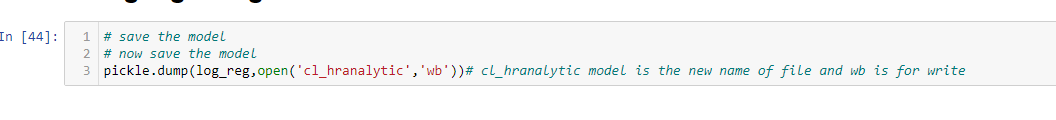
as we can see that decision tree has not good results ,so till now logistic reggresion is good but i will try another algoritm

**knn classifier**

****

so, all of three algo, only logistic regression gave satisfactory result so i am using logistic regression**.**

**SAVE THE MODEL**

****

1. **Concluding Remarks**

Let me go back to each step I have taken in this project. from beginning we have

Understanding the problem definition to go through EDA Processes. We went to some certain pre-processing steps and finally build Machine learning Model.

In this entire project I have given my best and put all possible potential that I have. Most of time I used my own code but when I found on internet some better code I opted that code.

I don’t take other good work on ego rather I do admire. Obviously, I do believe in learning buy other project but yes it does not mean that I copied the code.

For any of machine learning project my suggestion is first you have to understand the problem on ground level .if you don’t allow yourself to work with diligence .if you don’ t work harder anything that you are doing or will do , not only in case of machine learning but also in life cycle would be futile. Maybe, my endeavour assist you whenever you will get stuck

**Disclaimer**

I am new comer here in data science domain with some knowledge of 9 month. I have shared my effort in this blog to someone who is stepping in this field and can take some advantage from it. but to be honest it’s definitely inspired by others I saw many blog of this project on internet who have worked on this projects before me I just went through each of project and concluded my best way to make this blog.

THANK YOU

# Submitted By –

# SHUBHAM SHARMA