**A**

**Project Report On**

“**HUMAN RESOURCEANALYTICS”**

Submitted in partial fulfilment of the requirement

As internship in

**MACHINE LEARNING WITH PYTHON**

**Submitted by**

1. Neha Reddy - 18311A05H7
2. Pranavi Chitti - 19315A0518
3. Nandini Pola - 18H61A05N3
4. Balduri Vamshi - 18311A05Q7

**Under the esteemed guidance of**

**MR. C.NAVEEN KUMAR**

**MR. R.RAKESH**

**ACKNOWLEDGEMENTS**

We would like to express our immense gratitude and sincere thanks to Mr.Rakesh Sir and Mr.Naveen Sir for thier guidance ,valuable suggestions and encouragement in completing the internship project work with in the stipulated time.

We would like to express the sincere thanks to GoalStreet ,Mr.Sheshu Kandregula sir, director of Goal Street, Mr.Durga Naveen Kandregula sir, co-founder of Goalstreet for permitting us to do our Internship project work at GoalStreet pvt ltd.Finally,we would also like thank the people who have directly or indirectly helped us and other faculty,friends and goalstreet team for their cooperation in complete in the project work.

**ABSTRACT**

Employee turn-over is a very costly problem for companies. The cost of replacing an employee if often larger than 100K USD, taking into account the time spent to interview and find a replacement, placement fees, sign-on bonuses and the loss of productivity for several months. It is only natural then that data science has started being applied to this area. Understanding why and when employees are most likely to leave can lead to actions to improve employee retention as well as planning new hiring in advance. This application of DS is sometimes called people analytics or people data science (if you see a job title: people data scientist, this is your job).

In this project, we attempt to predict when employees are going to quit by understanding the main drivers of employee churn. The goal is to predict employee retention and understand its main drivers.

**CONTENTS**

Title page…. 1

Acknowledgements… 2

Abstract……………………………….3

1.Introduction………………………...5

2.Data Collection……………………..6

3.Data Representation……………….8

3.1. Feature Engineering……….12

3.2 Data Visualisation…………..18

4.Models

4.1. Logistic regression…………….19

4.2. Random Forest………………..20

4.3. Decision Trees………………...21

5.Conclusion………………………….22

6.References…………………………..23

**1.INTRODUCTION**

**Employee retention refers to the various policies and practices which let the employees stick to an organization for a longer period of time**. Every organization invests time and money to groom a new joinee, make him a corporate ready material and bring him at par with the existing employees. The organization is completely at loss when the employees leave their job once they are fully trained. Employee retention takes into account the various measures taken so that an individual stays in an organization for the maximum period of time.

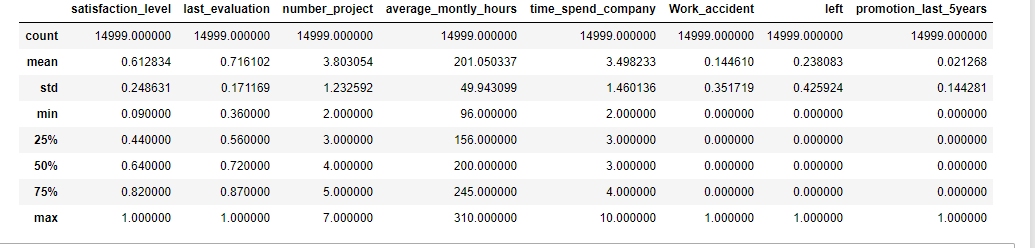
**Research says that most of the employees leave an organization out of frustration and constant friction with their superiors or other team members**. In some cases low salary, lack of growth prospects and motivation compel an employee to look for a change. The management must try its level best to retain those employees who are really important for the system and are known to be effective contributors.

It is the responsibility of the line managers as well as the management to ensure that the employees are satisfied with their roles and responsibilities and the job is offering them a new challenge and learning every day.

We aim to build other classification models like logistic regression,random forest and others to tune in to the parameters and make it work by training and testing the model.

**2.DATA COLLECTION**

The HR dataset consists of around 593 records.These records show the information related to employees of a company.it contains around 10 attributes like satisfaction level of employee, average monthly hours spent by an employee,last evaluation, salary,whether an employee lest etc..Based on this information, we need to predict whether an employee is likely to quit or stay.



This figure describes our datas

**3.DATA REPRESENTATION**

The Dataset we are using is HRA(Human Resource Analytics).It consists of various details like employee satisfaction level,last evaluation,salary,number of projects,average time spent etc. of around 593 employees.The following are the steps to be performed on the dataset to solve the problem.

● Read the Input dataset.

● Perform all necessary Data Normalization, Standardization processing to prepare the transformed format of given input dataset.

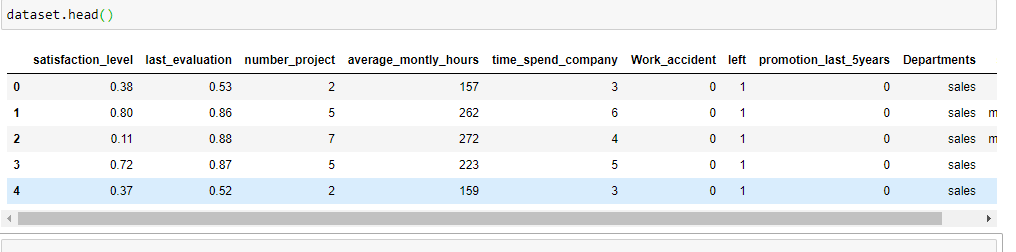
● Handle Missing values

● Perform Exploratory Data Analysis/Visualization and bring insights of the predictor variables.

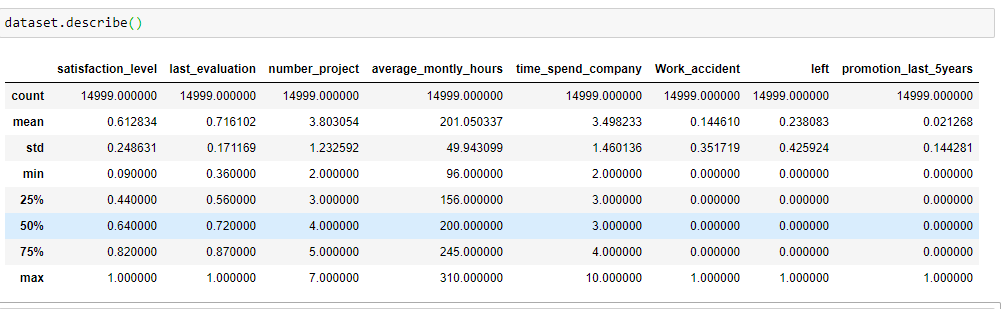
● Apply Logistic Regression Classifier,Random Forest,Decision Tree Classifier algorithms by splitting the data into train and test sets

● Measure and compare the performance of the models using confusion matrix and metrics like Precision and Recall

First five records of the dataset are as follows



Description of the dataset is as follows:



To determine if any null values are present,we should check info of the dataset which gives clear picture of dataset as follows

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 14999 entries, 0 to 14998

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 satisfaction\_level 14999 non-null float64

1 last\_evaluation 14999 non-null float64

2 number\_project 14999 non-null int64

3 average\_montly\_hours 14999 non-null int64

4 time\_spend\_company 14999 non-null int64

5 Work\_accident 14999 non-null int64

6 left 14999 non-null int64

7 promotion\_last\_5years 14999 non-null int64

8 Departments 14999 non-null object

9 salary 14999 non-null object

dtypes: float64(2), int64(6), object(2)

memory usage: 1.1+ MB

In our dataset target column is ‘left’’,it has two categorical values encoded into 1 and 0.Which indicates whether the patient has disease as 1 and not as 0.

**3.1 Feature Engineering**

Algorithms require features with some specific characteristic to work properly. Here, the need for **feature engineering** arises. I think feature engineering efforts mainly have two goals:

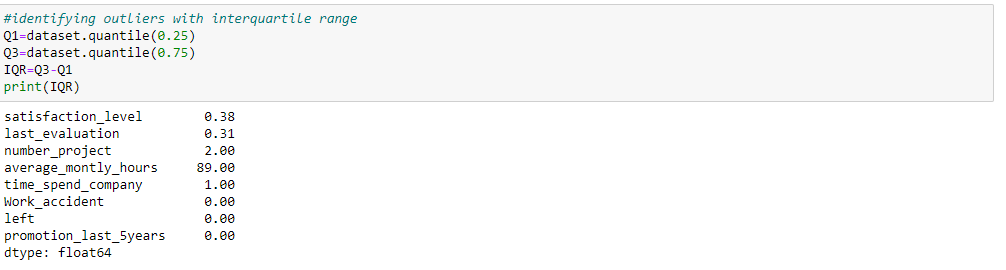
● Preparing the proper input dataset, compatible with the machine learning algorithm requirements.

● Improving the performance of machine learning models.

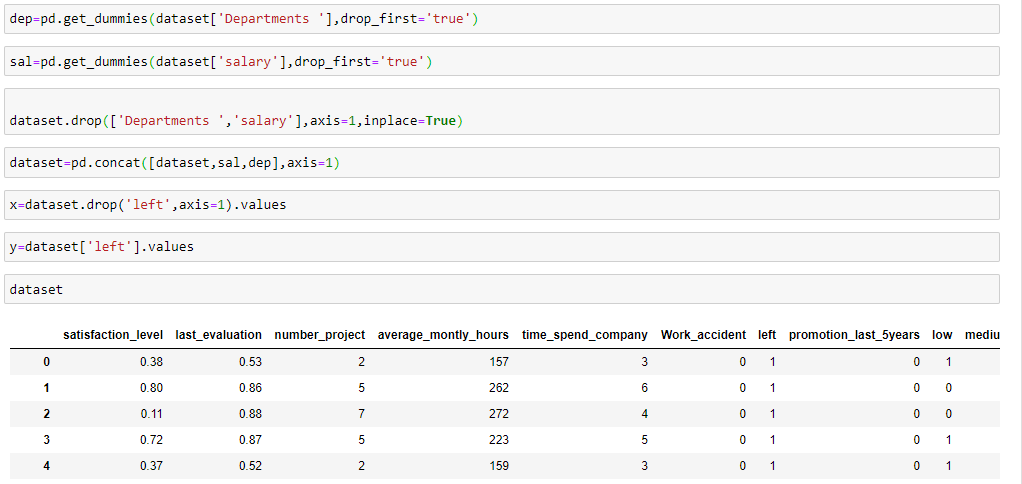
**Handling outliers**

Using Inter Quantile range to remove outliers and to make a lower and upper boundary for each and every feature in the dataset.

In descriptive statistics, the interquartile range, also called the midspread, middle 50%, or H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q₃ − Q₁.



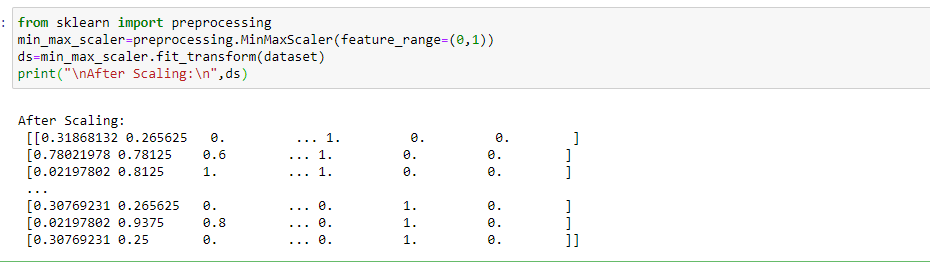
We need to convert all the attributes of dataset into numericals before scaling.For our dataset it is as follows



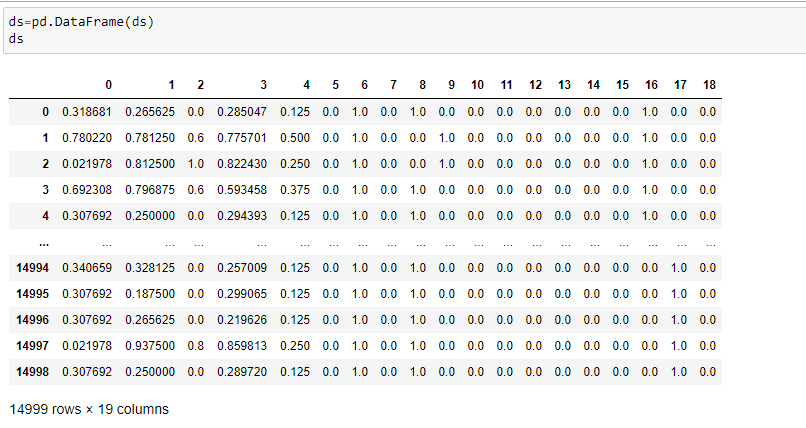
In case of Logistic Regression,Random Forest and Decision Trees,they internally do scaling and therefore we don’t need to do any scaling explicitly.If feature scaling should be done in any other case,it can be done as follows.

**Min Max Feature Scaling**

**Min**-**max normalization** is one of the most common ways to **normalize** data. For every feature, the minimum value of that feature gets transformed into a 0, the **maximum** value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1.



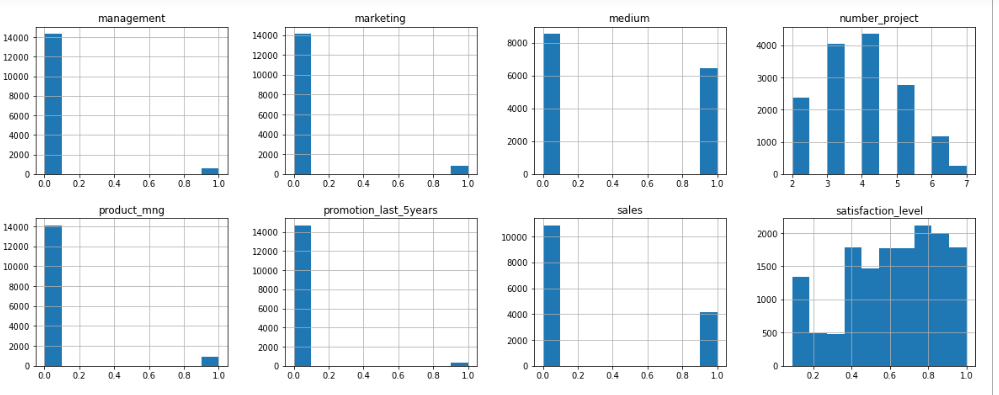
After feature scaling the ‘ds’ ndarray needs to be converted into the dataframe.

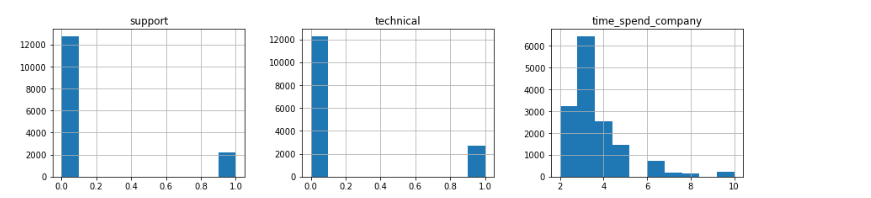


**3.2. Data Visualisation**



Contd…



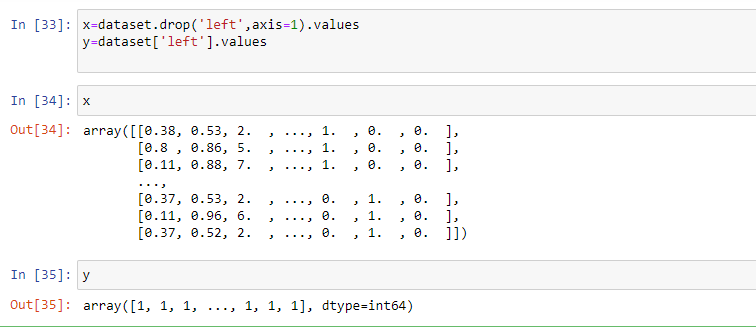
Contd..

Here the purpose of Visualization is to observe whether the data is distributed normally or not and to observe and remove the outliers. If data in each attribute is not normally distributed we have to apply some normalization techniques.But we have data distributed normally after removing outliers and normalization.And we can gain some meaningful insights from the observations above histograms.

Now, the imbalance in the final class is:



Now, we are seperating the features and label values into two different dataframes



**Feature Selection**

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

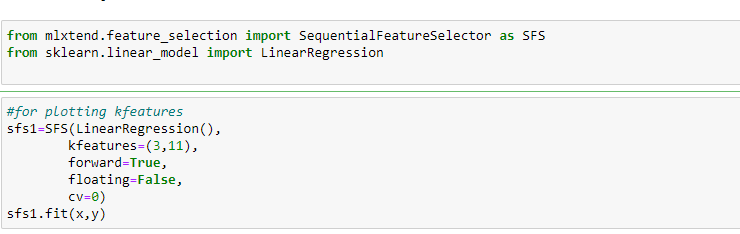
Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

**i) Backward Elimination:** In backward elimination, we start with the full model (including all the independent variables) and then remove the insignificant feature with highest ***p-value*(> *significance level*).** This process repeats again and again until we have the final set of ***significant*** features.

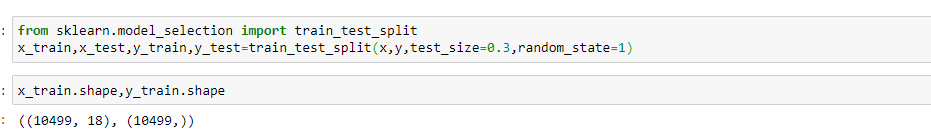


The above Backward elimination method will result in the important features which will contribute much to the ‘left’ Label.

For performing this operation we have to import ‘*SequentialFeatureSelector*’ from *mlxtend.featureselection* and we are applying the prediction plot based on *Linear regression* classifier with *kfeatures*.



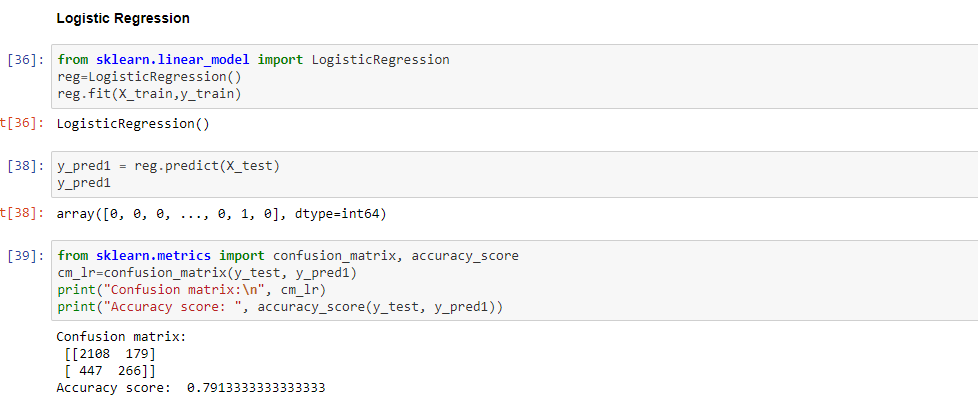
**Separated features and target label for training purpose.**



**4.APPLYING CLASSIFICATION MODELS**

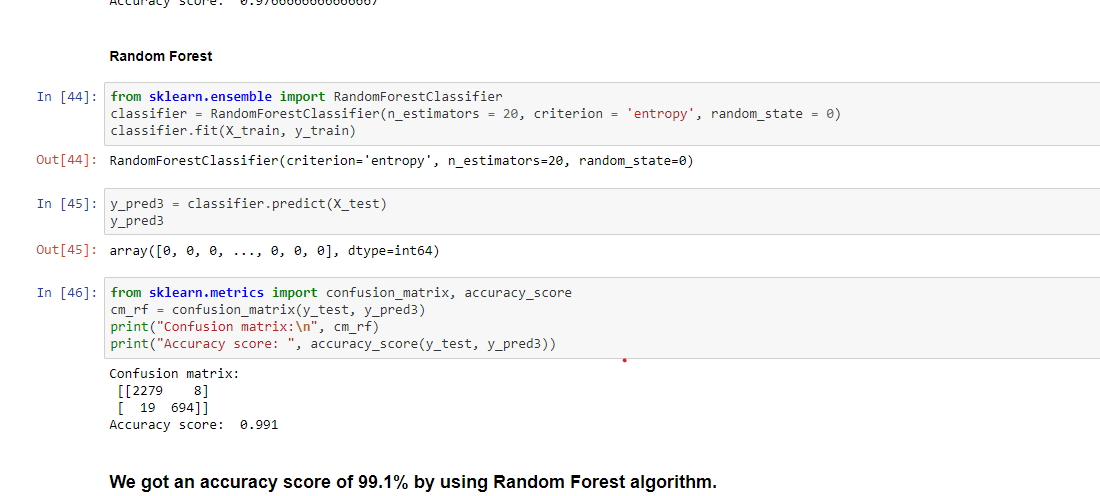
**4.1 LogisticRegressionClassifie**r

**Logistic regression** is a **classification** algorithm used to assign observations to a discrete set of classes. ... **Logistic regression** transforms its output using the **logistic** sigmoid function to return a probability value.



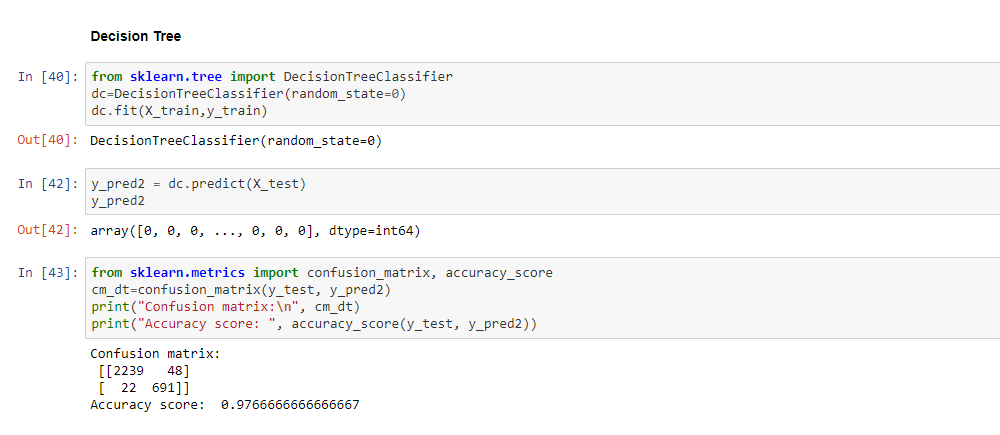
**4.2 Random Forest Classifier**

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.



**4.3 Decision Tree Classifier**

Decision Tree Classification in Python 1 Information Gain. Shannon invented the concept of entropy, which measures the impurity ... 2 Gain Ratio. Information gain is biased for the attribute with many outcomes. 3 Gini index. Another decision tree algorithm CART (Classification and Regression Tree) ... 4 Importing Required Libraries.



**5.CONCLUSION**

* Our project can help by processing the huge amounts of HR data in most of the companies, and help them understand why certain employees are leaving, what is the cause factor, and help them improve themselves in that particular department.
* Human Resources Analytics can play a huge role in companies by giving them the right tools to analyze their data. Being able to flag which employees are at risk enables organizations to step in with preventative measures and avoid the cost of losing productivity and the cost of re-hiring.
* We can use the algorithms given previously and perform analysis on the data to reduce the overall cost of hiring again

**6.REFERENCES**

<https://github.com/SkillsSquad/Machine-Learning>

<https://in.search.yahoo.com/search?fr=mcafee&type=E211IN714G0&p=skills+squad>

<https://in.search.yahoo.com/search?fr=mcafee&type=E211IN714G0&p=decision+tree+classifier>

<https://in.search.yahoo.com/search?fr=mcafee&type=E211IN714G0&p=random+forest+classifier>

<https://in.search.yahoo.com/search?fr=mcafee&type=E211IN714G0&p=logistic+regression>