MEHRAN UNIVERSITY OF ENGINEERING AND TECHNOLOGY

JAMSHORO

DATA SCIENCE AND ANALYTICS PROJECT

TITLE:

ANALYSIS OF HR DATASET

SUBMITTED TO

DR SANAM NAREJO

SUBMITTED BY

BILAWAL ABBASI(17CS45)

SANOBER ASGHAR(17CS69)

UME SAMANA (17CS63)

DEPARTMENT OF COMPUTER SYSYTEM ENGINEERING MUET JAMSHORO

TABLE OF CONTENTS

1. INTRODUCTION:

Human resources (HR) is the department of a company responsible for locating, screening, hiring, and training job applicants and administering employee benefit programs. In the twenty-first century, HR plays a critical role in assisting organizations in dealing with a rapidly changing business environment and a higher need for quality employees. Large corporations use it mostly to find innovative ways to increase revenues and cut costs. Data mining examines data and aids in discovering hidden aspects, allowing for the creation of meaningful patterns and information. These kinds of findings can assist any company in making future product decisions. There are several classification techniques in data mining such as the decision tree, neural network, rough set theory, baisian theory and fuzzy, see Phyu [1].

2. LITERATURE REVIEW:

In today's corporate world, numbers are considered the language of business. Organizational decision-makers make decisions based on figures derived from descriptive, predictive, and prescriptive studies. As a result, firms use data analytics to increase decision accuracy while also increasing their efficacy and efficiency. To make appropriate judgments on employees' difficulties, data relating to every aspect of the organization should be carefully reviewed, appraised, and analyzed (Lochan et al., 2018). "Workforce analytics," "human capital analytics," or "HR analytics" are terms used to describe the use of data in HR. HR analytics is a powerful instrument that can offer positive value to the HR department's functions while also boosting the efficacy and efficiency of all linked aspects through logical and quantifiable explanations. HR professionals use HR analytics to make decisions to attract, retain, and improve employee performance. An organization can only continue its success in the long run if it stays current with the newest trends in HR analytics (Reena et al., 2019). HR analytics has several advantages, one of which is evidence-based research that assists HR professionals in making rational decisions while increasing strategic impact. On three feature selection approaches, A. Chaudhary, and his colleagues [2] present a Bayes of simple performance and a custom mobile device evaluation. The practical approaches, such as the gain ratio method and the information gain method, are applied in this work. According to Ahmad and his colleagues, HRM (Human Resource Management) is the critical coalition of administration that deals with the firm's most precious resource, which is its human resource [3].

3. REPORT OBJECTIVE:

- **a)** Assess what are the relationship between the 10 variables and what are the significant variables to describe the dataset.
- **b)** Understand who are the employees that have left.
- **c)** Focus the analysis on the most valuable employees who have left.
- **d)** Develop a predictive model to assess the likelihood of an employee leaving.

4. TOOLS AND TECHNIQUES:

4.1 ANACONDA

Anaconda was designed with data scientists in mind. Over 20 million people use our technology to tackle the world's most complex challenges. Anaconda solutions are serious data science and machine learning technology.

4.2 JUPYTER NOTEBOOK

Jupyter is an interactive development environment for Jupyter notebooks, code, and data accessible via the web. Jupyter is adaptable: you may customise and arrange the user interface to support a variety of data science, scientific computing, and machine learning workflows. JupyterLab is modular and expandable, allowing you to create plugins that add new features and integrate with current ones.

4.3 PYTHON:

Numpy is a Python library for manipulating arrays. It also provides functions for working with matrices, the Fourier transform, and the domain of algebra. It's used to collect a variety of colors.

4.4 NUMPY:

Python is a high-level, general-purpose programming language that is interpreted. The use of considerable indentation in its design philosophy emphasizes code readability. Its language elements and object-oriented approach aim to assist programmers in writing clear, logical code for both small and large-scale projects.

4.5 PANDAS:

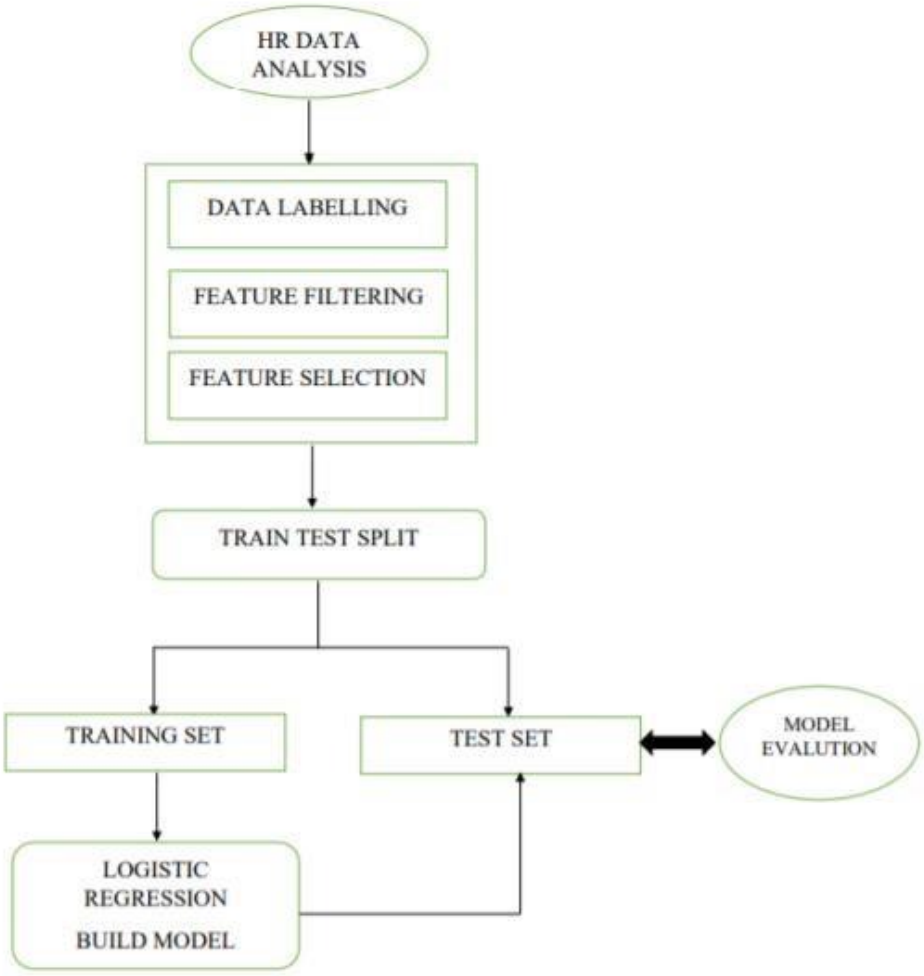
Pandas is a data manipulation Data Frame object with built-in indexing that is quick and easy to use. Its tools read and write data between in-memory data structures and several formats, including CSV and text files, Microsoft Excel, SQL databases, and the speedy HDF5 format.

4.6 SEABORN:

Seaborn is a matplotlib-based Python data visualization package. It has a high-level interface for creating visually appealing and instructive statistics visuals.

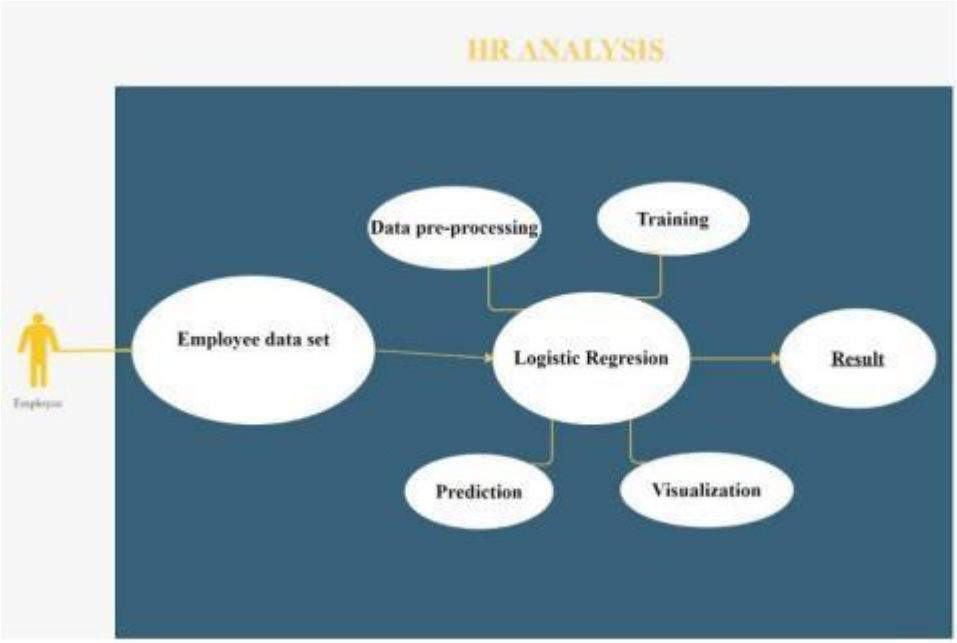
5. DESGN

5.1 WORKFLOW DIAGRAM:



5.2 FLOW CHART DIAGRAM:

5.3 SOFTWARE ENGINEERING DIAGRAM:



6. IMPLEMENTATION

6.1 DATASET LINK

https://www.kaggle.com/giripujar/hr-analytics

6.2 PREPROCESS DATA.

6.2.1 Data Quality Check

First, we will perform basic statistical analysis and understand the type of factors.

```
In [2]: df = pd.read_csv("HR_comma_sep.csv")
        df.head()
```

Out[2]:

satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Department	salary
0.38	0.53	2	157	3	0	1	0	sales	low
0.80	0.86	5	262	6	0	1	0	sales	medium
0.11	0.88	7	272	4	0	1	0	sales	medium
0.72	0.87	5	223	5	0	1	0	sales	low
0.37	0.52	2	159	3	0	1	0	sales	low

Looking at data we can make the following assumptions related the nature of the 10 variables:

1. **Satisfaction level**: A numeric indicator filled out by the employee ranging from 0 to 1
2. **Last Evaluation:** A numeric indicator filled in by the employee's manager ranging from 0 to 1
3. **Number Project**: An integer that indicates the number of projects the employee has worked on ranging from 2-7
4. **Average Monthly Hours:** The number of hours employees work in the month
5. **Time Spend Company:** An integer value indicated the years of service
6. **Work Accident:** A dummy variable assessing whether (1) or not (0) they had an accident
7. **Left**: A dummy variable, leave (1), not leave (0)
8. **Promoted Last 5years:** A dummy variable, promoted (1), not promoted (0)
9. **Department**: A categorical variable assessing the department in which employee is working (sales, technical, support, IT, product, marketing, other)
10. **Salary**: A 3-level categorical variable (low, medium, high)

6.2.2 Data Quality Report

First of all we assess that there are no missing data with function is.na(myData) that we would not report in the code and and we perform basic summary statistic of the dataset.

```
In [3]: df.isna().sum()
```

```
Out[3]: satisfaction_level      0
        last_evaluation         0
        number_project          0
        average_montly_hours    0
        time_spend_company      0
        Work_accident           0
        left                    0
        promotion_last_5years   0
        Department              0
        salary                  0
        dtype: int64
```

```
In [4]: df.min(axis = 0)
```

```
Out[4]: satisfaction_level      0.09
        last_evaluation         0.36
        number_project             2
        average_montly_hours      96
        time_spend_company         2
        Work_accident              0
        left                       0
        promotion_last_5years      0
        Department                IT
        salary                  high
        dtype: object
```

```
In [5]: df.max(axis = 0)
```

```
Out[5]: satisfaction_level              1
        last_evaluation                 1
        number_project                  7
        average_montly_hours          310
        time_spend_company             10
        Work_accident                   1
        left                            1
        promotion_last_5years           1
        Department                sales
        salary                   technical
        dtype: object
```

```
In [6]: df.mean(axis = 0)
```

```
Out[6]: satisfaction_level      0.612834
        last_evaluation         0.716102
        number_project          3.803054
        average_montly_hours  201.050337
        time_spend_company      3.498233
        Work_accident           0.144610
        left                    0.238083
        promotion_last_5years   0.021268
        dtype: float64
```

```
In [7]: df.median(axis = 0)
```

```
Out[7]: satisfaction_level      0.64
        last_evaluation         0.72
        number_project          4.00
        average_montly_hours  200.00
        time_spend_company      3.00
        Work_accident           0.00
        left                    0.00
        promotion_last_5years   0.00
        dtype: float64
```

6.3 EDA TECHNIQUE AND VISUALIZATION.

Data exploration and visualization

In [7]:
```
left = df[df.left==1]
left.shape
```

Out[7]: (3571, 10)

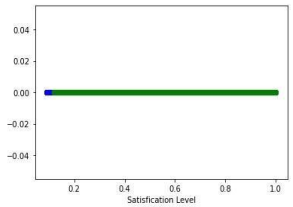
In [8]:
```
retained = df[df.left==0]
retained.shape
```

Out[8]: (11428, 10)

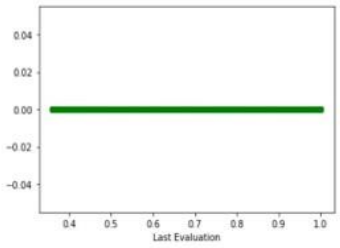
1. Univariate Analysis EDA

In [9]:
```
df_left = df.loc[df['left']==1]
df_retain = df.loc[df['left']==0]
```

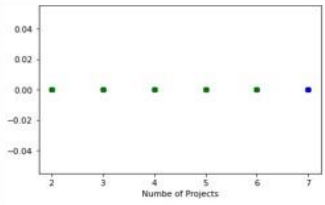
In [10]:
```
plt.plot(df_left['satisfaction_level'],np.zeros_like(df_left['satisfaction_level']),'o',color='blue')
plt.plot(df_retain['satisfaction_level'],np.zeros_like(df_retain['satisfaction_level']),'o',color='green')
plt.xlabel('Satisfication Level')
plt.show()
```



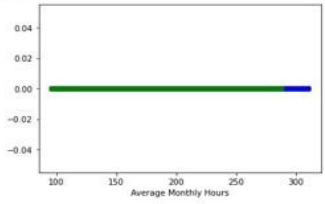
In [11]:
```
plt.plot(df_left['last_evaluation'],np.zeros_like(df_left['last_evaluation']),'o',color='blue')
plt.plot(df_retain['last_evaluation'],np.zeros_like(df_retain['last_evaluation']),'o',color='green')
plt.xlabel('Last Evaluation')
plt.show()
```



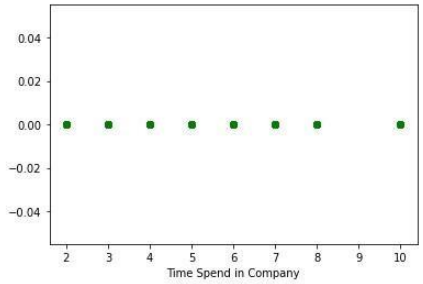
In [12]:
```
plt.plot(df_left['number_project'],np.zeros_like(df_left['number_project']),'o',color='blue')
plt.plot(df_retain['number_project'],np.zeros_like(df_retain['number_project']),'o',color='green')
plt.xlabel('Numbe of Projects')
plt.show()
```



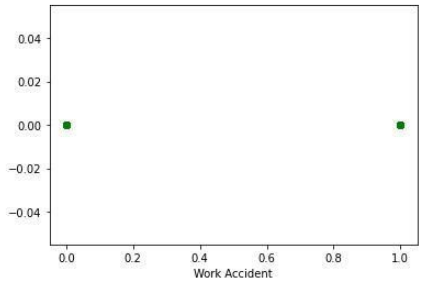
In [13]:
```
plt.plot(df_left['average_montly_hours'],np.zeros_like(df_left['average_montly_hours']),'o',color='blue')
plt.plot(df_retain['average_montly_hours'],np.zeros_like(df_retain['average_montly_hours']),'o',color='green')
plt.xlabel('Average Monthly Hours')
plt.show()
```



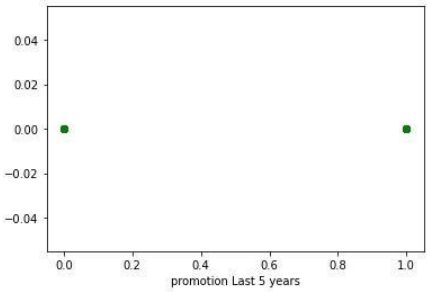
In [14]:
```
plt.plot(df_left['time_spend_company'],np.zeros_like(df_left['time_spend_company']),'o',color='blue')
plt.plot(df_retain['time_spend_company'],np.zeros_like(df_retain['time_spend_company']),'o',color='green')
plt.xlabel('Time Spend in Company')
plt.show()
```



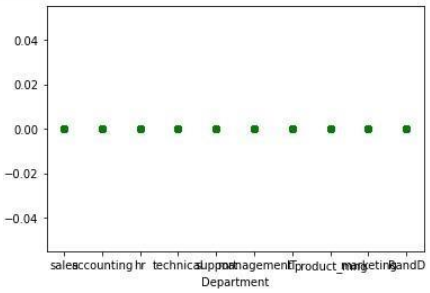
In [15]:
```
plt.plot(df_left['Work_accident'],np.zeros_like(df_left['Work_accident']),'o',color='blue')
plt.plot(df_retain['Work_accident'],np.zeros_like(df_retain['Work_accident']),'o',color='green')
plt.xlabel('Work Accident')
plt.show()
```



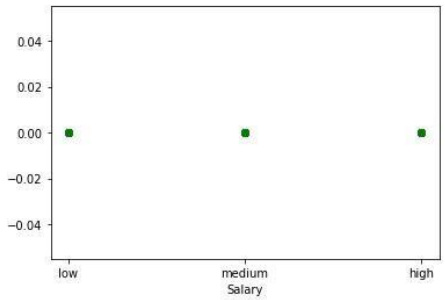
In [16]:
```
plt.plot(df_left['promotion_last_5years'],np.zeros_like(df_left['promotion_last_5years']),'o',color='blue')
plt.plot(df_retain['promotion_last_5years'],np.zeros_like(df_retain['promotion_last_5years']),'o',color='green')
plt.xlabel('promotion Last 5 years')
plt.show()
```



In [17]:
```
plt.plot(df_left['Department'],np.zeros_like(df_left['Department']),'o',color='blue')
plt.plot(df_retain['Department'],np.zeros_like(df_retain['Department']),'o',color='green')
plt.xlabel('Department')
plt.show()
```



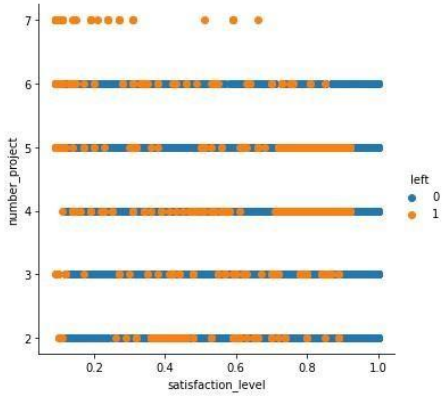
In [18]:
```
plt.plot(df_left['salary'],np.zeros_like(df_left['salary']),'o',color='blue')
plt.plot(df_retain['salary'],np.zeros_like(df_retain['salary']),'o',color='green')
plt.xlabel('Salary')
plt.show()
```



2. Multivariate Analysis

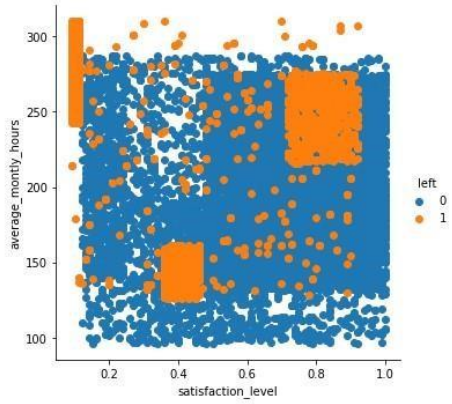
In [19]:
```
sns.FacetGrid(df,hue="left",size=5).map(plt.scatter,"satisfaction_level","number_project").add_legend();
plt.show()
```

```
C:\Users\sufi sahab\.conda\envs\tensorflow\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has bee
n renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



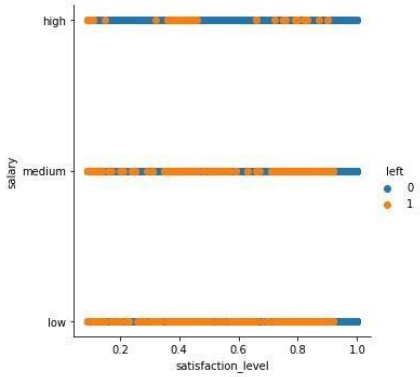
In [20]:
```
sns.FacetGrid(df,hue="left",size=5).map(plt.scatter,"satisfaction_level","average_montly_hours").add_legend();
plt.show()
```

```
C:\Users\sufi sahab\.conda\envs\tensorflow\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has bee
n renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



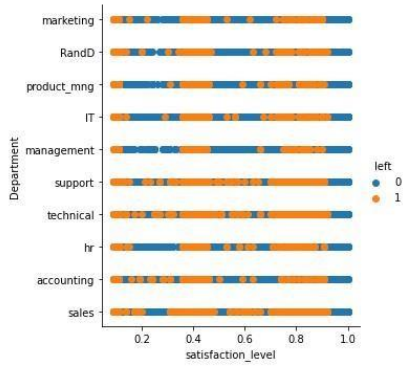
In [21]:
```
sns.FacetGrid(df,hue="left",size=5).map(plt.scatter,"satisfaction_level","salary").add_legend();
plt.show()
```

```
C:\Users\sufi sahab\.conda\envs\tensorflow\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has bee
n renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [22]:
```
sns.FacetGrid(df,hue="left",size=5).map(plt.scatter,"satisfaction_level","Department").add_legend();
plt.show()
```

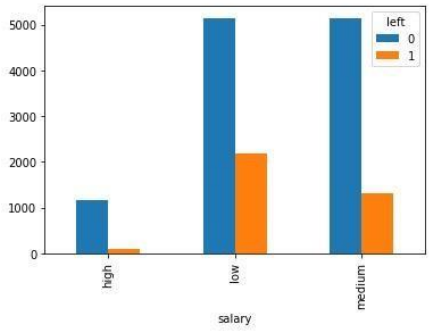
```
C:\Users\sufi sahab\.conda\envs\tensorflow\lib\site-packages\seaborn\axisgrid.py:337: UserWarning: The `size` parameter has bee
n renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



Histogram Impact of salary on employee retention

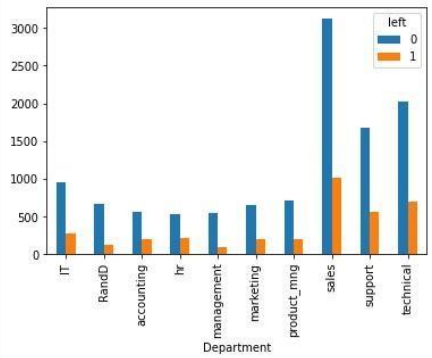
In [23]:
```
pd.crosstab(df.salary,df.left).plot(kind='bar')
```

Out[23]: <AxesSubplot:xlabel='salary'>



Department wise employee retention rate

```
In [24]: pd.crosstab(df.Department,df.left).plot(kind='bar')
```



From above chart there seem to be some impact of department on employee retention but it is not major hence we will ignore department in our analysis

6.4 PRINCIPAL COMPONENT ANALYSIS.

Average numbers for all columns

```
In [25]: df.groupby('left').mean()
```

Out[25]:

left	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years
0	0.666810	0.715473	3.786664	199.060203	3.380032	0.175009	0.026251
1	0.440098	0.718113	3.855503	207.419210	3.876505	0.047326	0.005321

From above table we can draw following conclusions,

1. **Satisfaction Level**: Satisfaction level seems to be relatively low (0.44) in employees leaving the firm vs the retained ones (0.66)
2. **Average Monthly Hours**: Average monthly hours are higher in employees leaving the firm (199 vs 207)
3. **Promotion Last 5 Years**: Employees who are given promotion are likely to be retained at firm
4. **Salary bar chart shows employees with high salaries are likely to not leave the company**

From above table and EDA exploration we can draw following conclusions,

- Satisfaction level seems to be relatively low (0.44) in employees leaving the firm vs the retained ones (0.66).
- Average monthly hours are higher in employees leaving the firm (199 vs 207).
- Employees who are given promotion are likely to be retained at firm.
- Salary bar chart shows employees with high salaries are likely to not leave the company.

6.5 CREATE MACHINE LEARNING MODEL.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,train_size=0.3)
```

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
```

```
model.fit(X_train, y_train)
```

```
LogisticRegression()
```

6.6 TEST AND EVALUATE ACCURACY OF MODEL.

```
model.predict(X_test)
```

```
array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

Accuracy of the model

```
model.score(X_test,y_test)
```

```
0.7812380952380953
```

7. CONCLUSION

Managing employee satisfaction, according to our findings, is critical to retaining staff. This is particularly true for individuals who have worked with the company for more than three years. Aside from that, the number of projects and staff evaluations should be kept track of. To establish a successful talent retention policy, this firm's HR Director should create systems to keep track of these indicators.

8. REFRENCES:

[1] T. N. Phyu, "Survey of classification techniques in data mining," in Proceedings of the International Multiconference of Engineers and Computer Scientists, vol. 1, 2009, pp. 18–20.

[2] N. A. A. Shashoa, N. A. Salem, I. N. Jleta, and O. Abusaeeda, "Classification depend on linear discriminant analysis using desired outputs," in Sciences and Techniques of Automatic Control and Computer Engineering (STA), 2016 17th International Conference on. IEEE, 2016, pp. 328– 332.

[3] S. Ahmad, "Green human resource management: Policies and practices," Cogent Business & Management, vol. 2, no. 1, p. 1030817, 2015