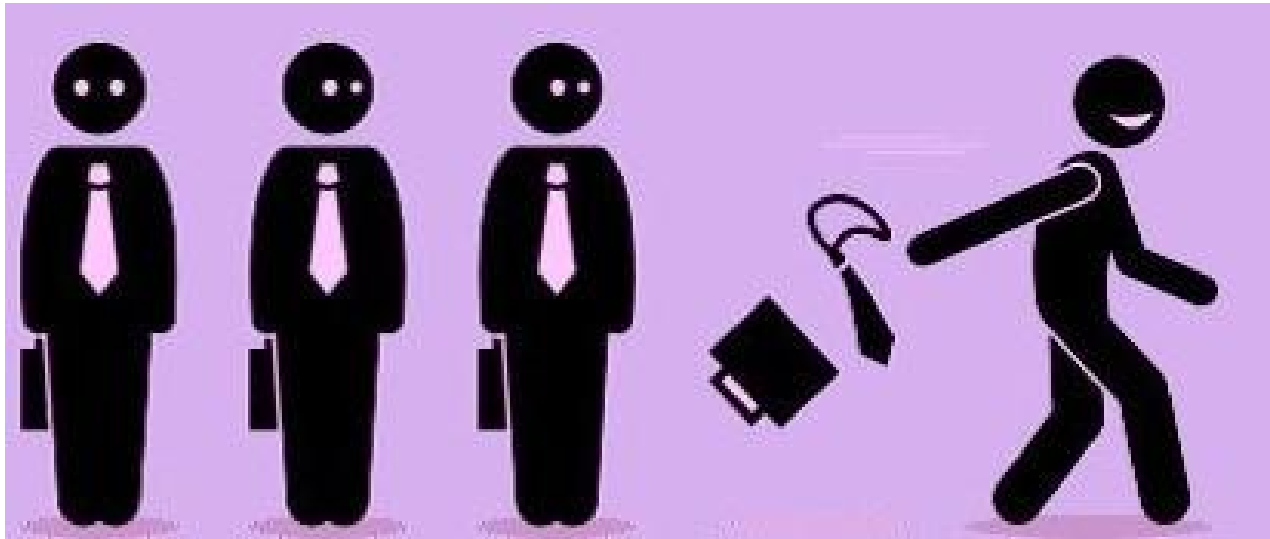


Wayne State University
CSC5800- Intelligent Systems: Algorithms and Tools
Project Title: EMPLOYEE ATTRITION ANALYSIS
Date of submission: 12/17/2021



Project by:

Akhila Dandu(hj4522)

Snigdha Pendhota(hj5002)

Problem Statement

The data is for a company which is trying to control employee attrition. There are two kinds of datasets: "Existing employees" and "Employees who have already left". Analysis of what kind of employees are leaving and determining which employees are about to leave next using appropriate machine learning models.

Dataset

Dataset used is from the Taken Mind dataset

<http://139.59.22.214/wp-content/uploads/2018/05/TakenMind-Python-Analytics-Problem-case-study-1-1.xlsx>

It consists of approx 15000 employee details in total with 10 attributes.

Attributes used are:

- **Satisfaction_level:** It explains the employee satisfaction point, which ranges from 0-1.
- **last_evaluation:** It is an evaluated performance of the employer, which also ranges from 0-1.
- **Number_projects:** It explains how many numbers of projects are assigned to an employee.
- **Average_monthly_hours:** It states how many average numbers of hours worked by an employee in a month.
- **Time_spent_company:** time spent company means employee experience. The number of years an employee spent in the company.
- **Work_accident:** Discusses whether an employee has had a work accident or not.
- **Promotion_last_5years:** Discusses whether an employee has had a promotion in the last 5 years or not.
- **Departments:** Employee's working department/division.
- **Salary:** Salary range of the employee such as low, medium and high.
- **left:** Describes whether the employee has left the company or not.

Language

In this project Python has been used to program the models and data analysis.

Literature Review

In one of the research papers, the author wanted to use the kNearest Neighbors algorithm to predict whether an employee would leave the company. Features which were included are employee performance evaluation, average monthly working hours, and years of service at the company. The dataset was split into 70% which is used for algorithm training and 30% which is used for testing.

In another article, the author suggested using deep learning techniques with some preprocessing steps to improve employee attrition predictions. They analyzed some of the factors that lead to employee attrition, reveal their interdependencies, and identify the dominant factors. Their work has been tested with an unbalanced dataset from IBM Analytics that includes 35 features for 1470 employees. Finally, cross evaluation is done and their work is evaluated.

To build an analysis model that uses a decision tree to analyze and predict what employees are prone to leave next. To do this we have concatenated the two datasets into a single dataset and applied a decision learning algorithm to predict employee attrition. The previous research papers have focused mainly on the analysis and prediction of employee attrition using k- Nearest Neighbors and deep learning techniques.

Implementation

Our current datasets consist of 10 features and more than 14000 records. All of the features are related to the employees' attrition problems. The left dataset represents the records of the employees who left the company. The existing dataset represents the records of employees who are still currently working in the company.

```
[ ] exist.head()
```

	Emp ID	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	promotion_last_5years	dept	salary
0	2001	0.58	0.74	4	215	3	0	0	sales	low
1	2002	0.82	0.67	2	202	3	0	0	sales	low
2	2003	0.45	0.69	5	193	3	0	0	sales	low
3	2004	0.78	0.82	5	247	3	0	0	sales	low
4	2005	0.49	0.60	3	214	2	0	0	sales	low

```
[ ] left.head()
```

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

Our dataset consists of 15,000 records approx out of which 3,571 were left, and 11,428 stayed. The remaining number of employees is 23% of total employment.

Data Preprocessing

This stage involves preparing and analyzing the data.

To check for missing values

```
✓ [8] print(exist.isnull().sum()) # To check the number of missing values in the data set
```

```
Emp ID          0
satisfaction_level 0
last_evaluation   0
number_project   0
average_monthly_hours 0
time_spend_company 0
Work_accident    0
promotion_last_5years 0
dept             0
salary           0
dtype: int64
```

```
✓ ▶ print(left.isnull().sum())
```

```
[ ] map1 = {1 : 'Yes', 0 : 'No'}
exist['promotion_last_5years'] = exist['promotion_last_5years'].map(map1)
left['promotion_last_5years'] = left['promotion_last_5years'].map(map1)
exist['Work_accident'] = exist['Work_accident'].map(map1)
left['Work_accident'] = left['Work_accident'].map(map1)
```

No missing values were found in our dataset

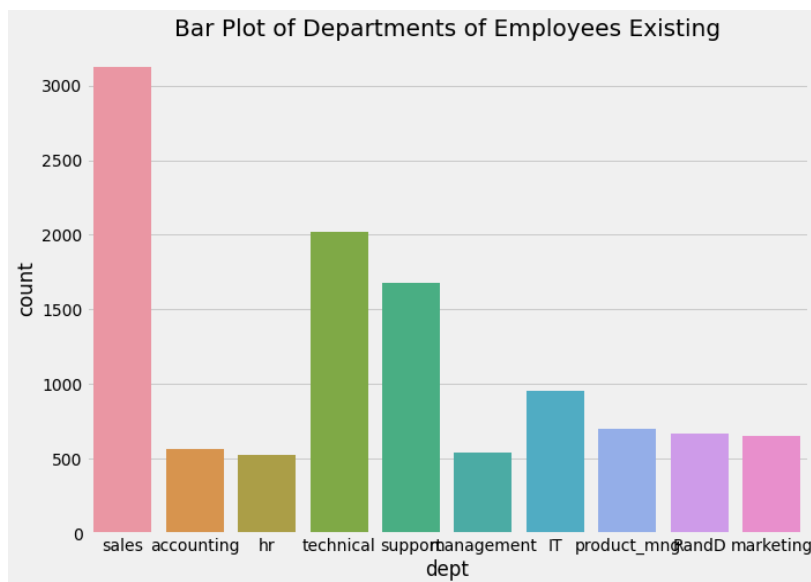
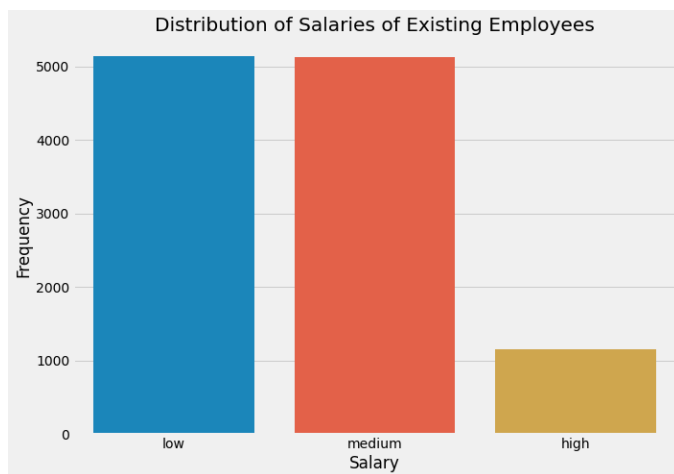
```
[ ] exist.describe()
```

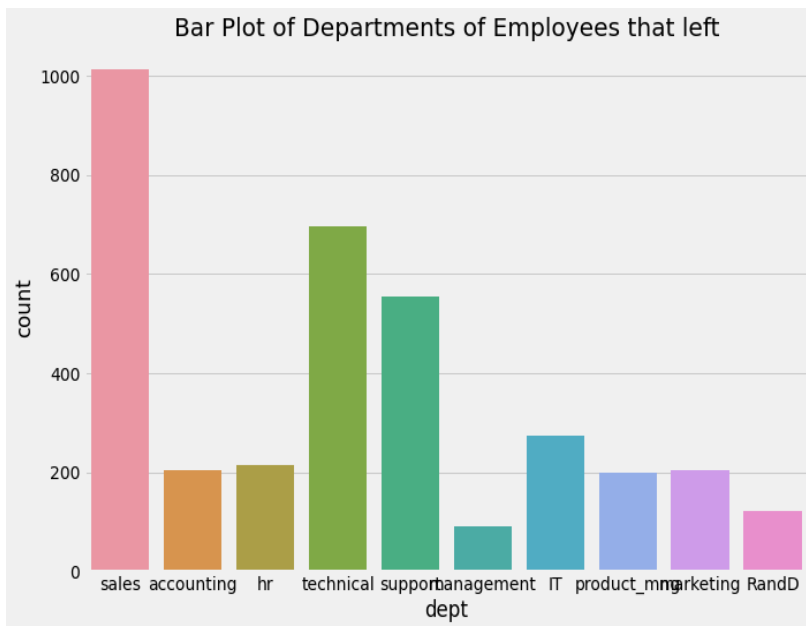
	Emp ID	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company
count	11428.000000	11428.000000	11428.000000	11428.000000	11428.000000	11428.000000
mean	7812.340742	0.666810	0.715473	3.786664	199.060203	3.380032
std	3453.947461	0.217104	0.162005	0.979884	45.682731	1.562348
min	2001.000000	0.120000	0.360000	2.000000	96.000000	2.000000
25%	4857.750000	0.540000	0.580000	3.000000	162.000000	2.000000
50%	7714.500000	0.690000	0.710000	4.000000	198.000000	3.000000
75%	10571.250000	0.840000	0.850000	4.000000	238.000000	4.000000
max	14211.000000	1.000000	1.000000	6.000000	287.000000	10.000000

```
▶ left.describe()
```

Describe() function in pandas is convenient in getting many summary statistics. The function returns the count, mean, standard deviation, minimum and maximum values and quantiles of the data.

Analysis based on Salary and Department

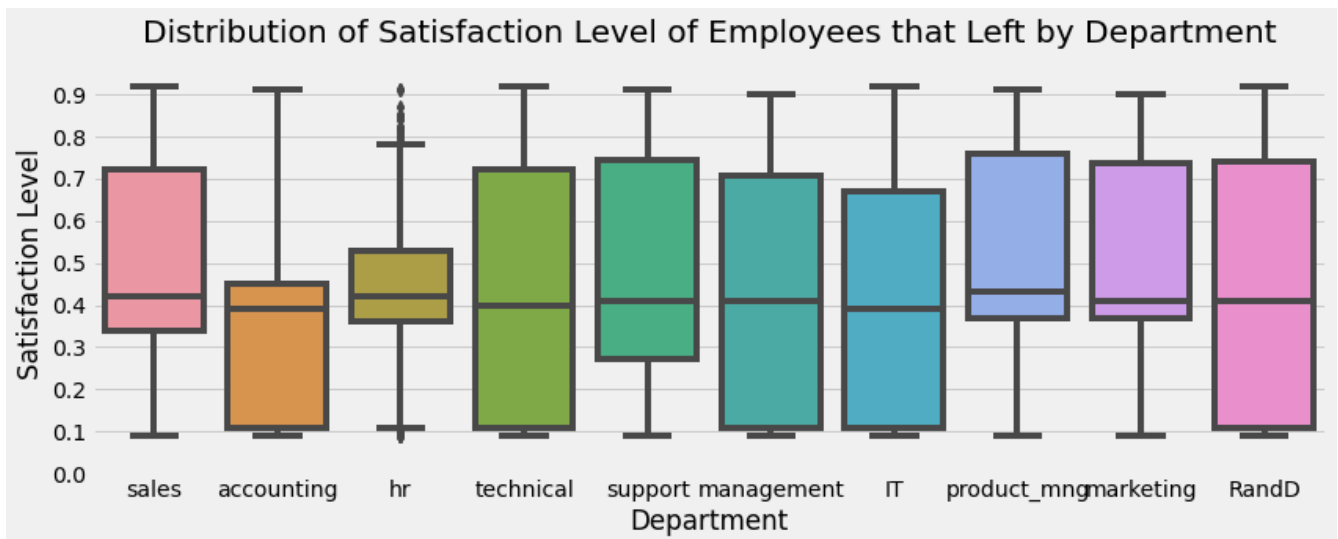
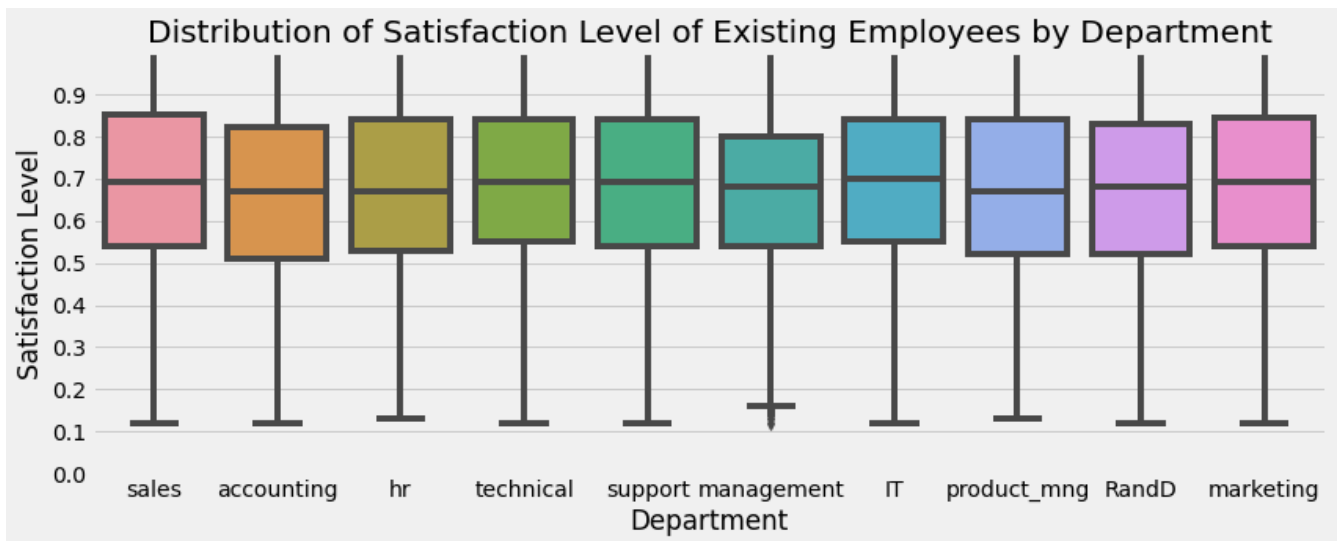




Inferences

- Most employees have medium or low salaries.
- Those left are in the low and medium salary range, whereas employees with a high salary range are less likely to leave than those of low and medium.
- The first figure shows the salary distribution of employees who left the company. From this figure, we conclude that the majority of left employees have low or medium salaries.
- The last figure represents the distribution of departments versus employee's left. The company shows that the technical, sales and support departments are one of the top three with the highest number of employees.

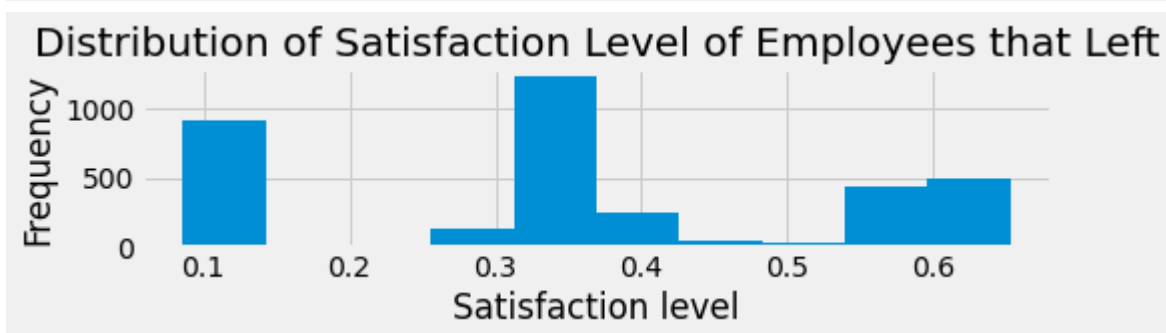
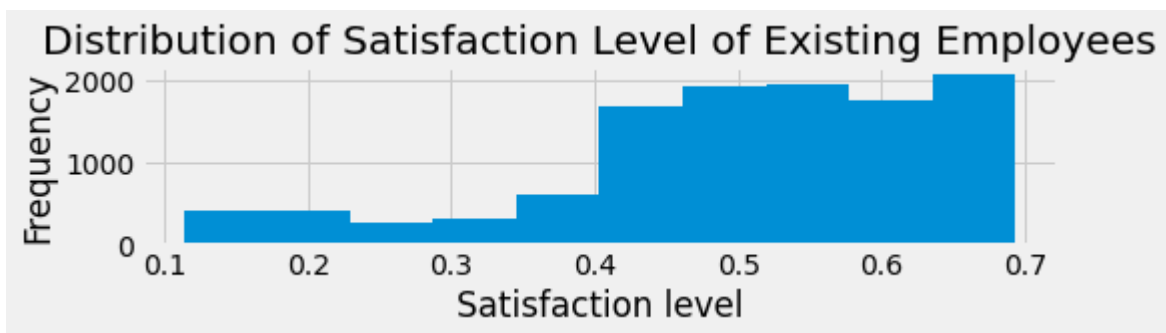
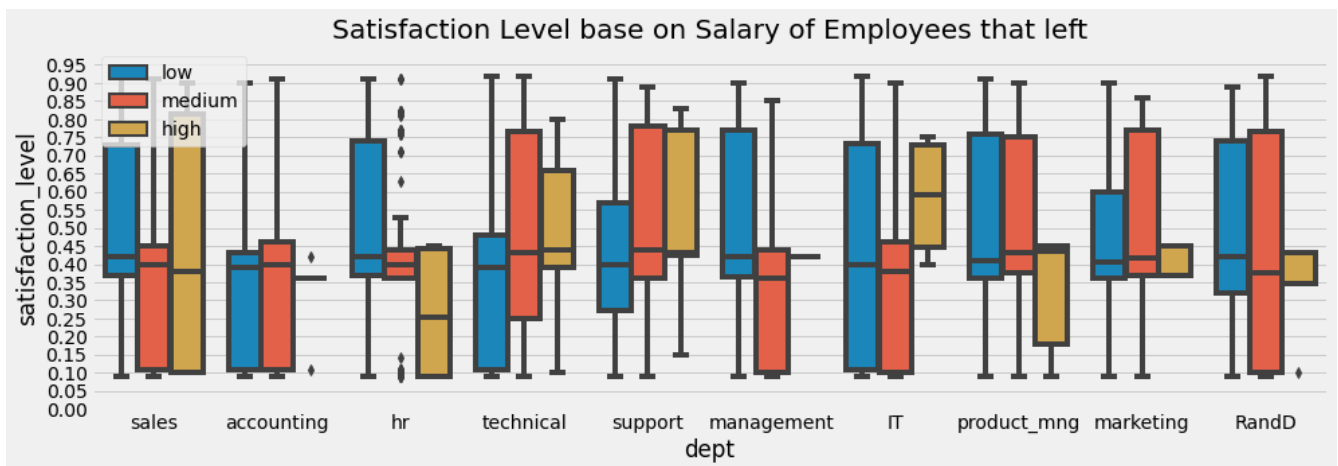
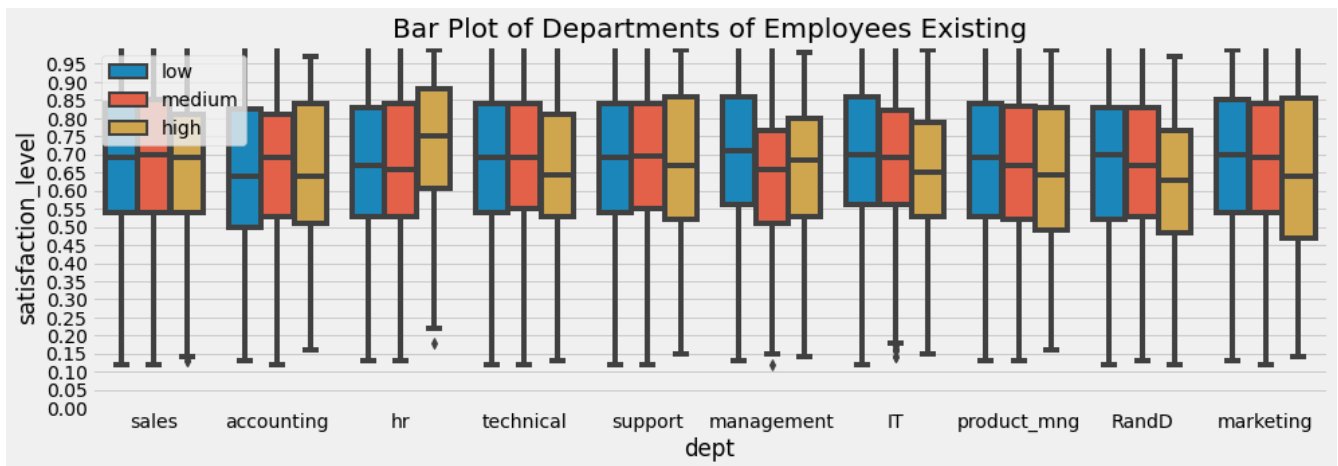
Analysis based on Satisfaction level



From the Box plot of both Existing Employees and those that left we can discover the following

- The median is 40 - 50% for the Satisfaction level of employees that left .
- The median is 65-70% for the Satisfaction level of employees existing in the company.

In order to strengthen the discovery further, analysis on satisfaction level has been performed.



- A thorough analysis of salary groups for both existing and left employees showed that left employees in different salary ranges were less than 45% satisfied, while existing employees were 60% satisfied across different salary ranges.
- This gives us a strong conviction that satisfaction level is a key factor and determinant in explaining the type of employees that left and also employees that are prone to leave using this metric.

After the satisfaction of employees who left the company has been determined, it is important to have a valid statistical information on the percentage of employees that are left in this category.

```
[ ] #checking for the job satisfaction greater then 60%
satjob=exist[exist["satisfaction_level"]>0.6]
print(((len(exist)-len(satjob))/len(exist))*100,"% People are less than 60% satisfied by job")
print(satjob["satisfaction_level"].mean())
```

```
36.13055652782639 % People are less than 60% satisfied by job
0.7987532538703787
```

```
[ ] #checking for the people who have left the job and their satisfaction
satjob1=left[left["satisfaction_level"]>0.6]
print(len(satjob1))
print(len(left))
print(((len(left)-len(satjob1))/len(left))*100,"% People have left job when less than 60% satisfied by job")
```

```
971
3571
72.80873704844582 % People have left job when less than 60% satisfied by job
```

```
✓ [30] [left[left['satisfaction_level']<0.45][ 'promotion_last_5years'].value_counts(normalize=True)]*100
```

```
No      99.566349
Yes      0.433651
Name: promotion_last_5years, dtype: float64
```

The results show that 99% of the employees that have left and have less than 45% satisfaction have not been promoted in the last five years, which is considered the main reason for attrition.


Another factor that identifies these employees in this category area is salary range.

```
#To determine the Salary proportion of Employees with less than 45% Satisfaction level
(left[left['satisfaction_level']<0.45][ 'salary'].value_counts(normalize=True))*100
```

```
low      60.754553
medium    37.120555
high      2.124892
Name: salary, dtype: float64
```

The results of the analysis show that about 61% of them had low salaries.

It makes sense to check and also verify the status of current employees based on the findings of left employees.

```
✓  #To determine the mean satisfaction level of employees that exist  
exist['satisfaction_level'].mean()*100  
  
66.6809590479516
```

From the output, we can see that the average level of existing satisfaction is about 66%.

Therefore, use the left employee metric as an insight to checkmate existing employees who tend to leave considering their status.

```
[ ] # To determine the percentage of employees that left had a satisfaction level < 45  
(len(left[left['satisfaction_level']<0.45])/len(left)) * 100  
  
64.57574908989079
```

The output shows that about 65% of those who leave the company have less than 45% satisfaction which needs to be questioned.

These findings increase the curiosity to identify factors that may have influenced low satisfaction. Promotion was an intuitive factor, prompting a deeper analysis to determine the percentage of people who left that had less than 45% and the status of their promotion.

```
[ ] print (str((len(exist[exist['satisfaction_level']<0.45])/len(exist)) * 100) + '%' + ' of employees that exist have satisfaction level less than 45%')  
  
13.816940847042353% of employees that exist have satisfaction level less than 45%
```

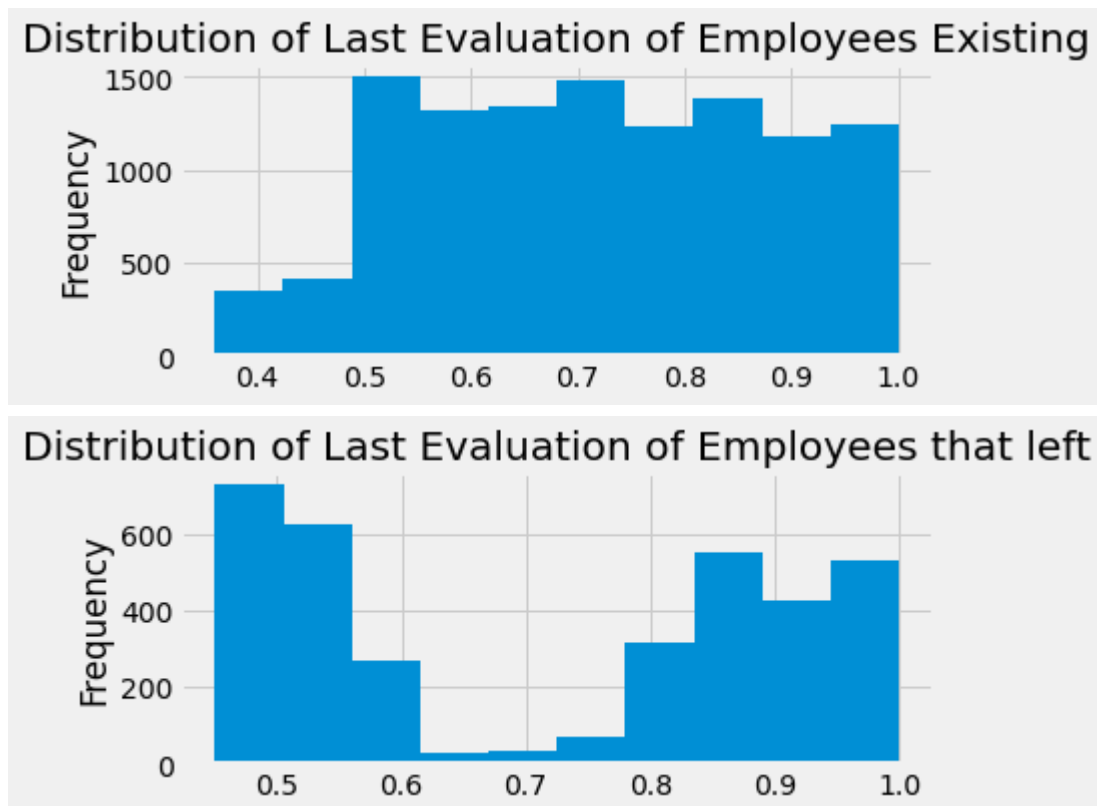
It can be seen that of the 13% of employees with less than 45% satisfaction, about 98% have not experienced a promotion in the last 5 years.

Inferences

- When it comes to salary satisfaction, employees in all salary classes share a higher level of satisfaction with existing employees in each department. In contrast, the satisfaction of those who stayed is very low.
- The satisfaction level ranges between (0.6-0.70) for the existing employees and for the employees left, it ranges from (0.25-0.4)

Based on Last Evaluation

The figures show the employee's count versus the last evaluation. This also represents the bio-modal distribution of employees who have left the company.



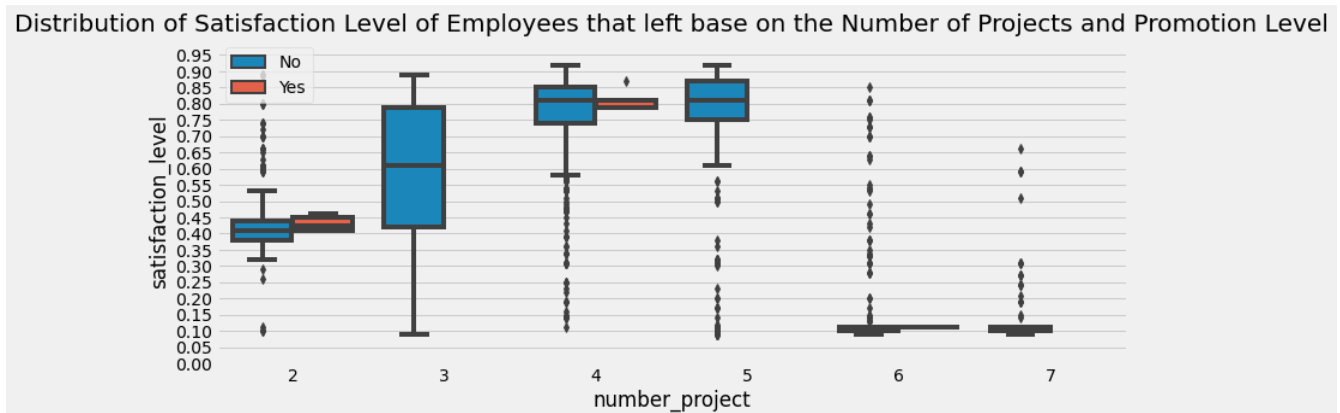
```
[ ] exist['last_evaluation'].median(),exist['last_evaluation'].mean()
(0.71, 0.7154733986699274)
```

```
[ ] left['last_evaluation'].median(), left['last_evaluation'].mean()
(0.79, 0.7181125735088183)
```

Inference

- From the above result for the last evaluation, there are no specific trends or patterns that encourage further analysis to determine if there are hidden insights.
- Poor performance and high performance are two key criteria for employees who tend to leave the company. The last evaluation is 0.6 to 0.8, which is a good range for employees staying at the company.

Bi/Multivariate Analysis



Over the last five years, significantly fewer employees have been promoted.

```
print (str((len(exist[exist['satisfaction_level']<0.45])/len(exist)) * 100) + '%' + ' of employees that exist have satisfaction level less than 45%')
exist[exist['satisfaction_level']<0.45]['promotion_last_5years'].value_counts(normalize=True)*100
```

```
13.816940847042353% of employees that exist have satisfaction level less than 45%
No      97.720076
Yes      2.279924
Name: promotion_last_5years, dtype: float64
```

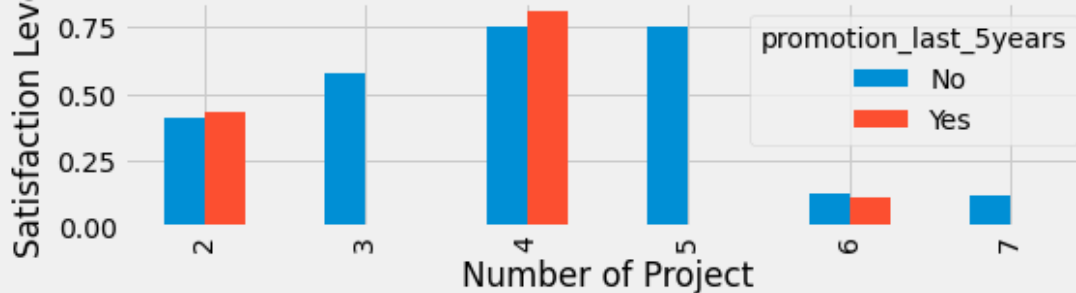
```
[ ] [(left[left['satisfaction_level']<0.45]['promotion_last_5years'].value_counts(normalize=True))*100]
No      99.566349
Yes      0.433651
Name: promotion_last_5years, dtype: float64
```

The figure above supports the finding that a significant percentage of employees who left the company are unsatisfied and most of them have not been promoted in the last five years.

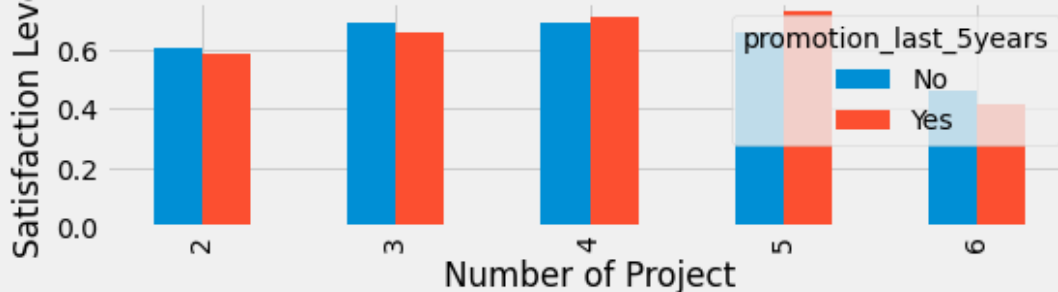
Satisfaction Level of employees that left base on Salary and Work accident



Satisfaction Level Vs Number of Projects with respect to Promotion



Satisfaction Level Vs Number of Projects with respect to Promotion



Inference

- The results displayed show that those with low satisfaction who left employment have rarely been promoted in the last five years, which is considered to be the main reason for attrition.
- Most of the employees are doing the project from 3-5.
- Employees who had five or more projects were among those who left the company.

Inference

- The band which has lighter color has high correlation. Boxes in the center have high correlation.
- The number of projects is also strongly correlated with the average monthly working hours, and as the number of years worked by employees increases, so does the number of projects.
- There is also a high correlation between last evaluation and time spent at company which depicts that employees have greater last evaluation owing to their time with the company.

Decision tree

Added a new feature 'Attrition' to both the datasets with the value Yes and No for existing dataset and left dataset respectively. Created a new dataset called employee_df by concatng both the existing and left dataset to create a new dataset to be able to apply Machine learning algorithm.

```
[54] exist['Attrition'] = 'No'
      exist.head()
```

	Emp ID	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	promotion_last_5years	dept	salary	Attrition
0	2001	0.58	0.74	4	215	3	No	No	7	1	No
1	2002	0.82	0.67	2	202	3	No	No	7	1	No
2	2003	0.45	0.69	5	193	3	No	No	7	1	No
3	2004	0.78	0.82	5	247	3	No	No	7	1	No
4	2005	0.49	0.60	3	214	2	No	No	7	1	No

```
[55] left['Attrition'] = 'Yes'
      left.head()
```

```
[56] all_list = [exist, left]
      employee_df = pd.concat(all_list)
      employee_df.to_excel("AllEmployees.xlsx", index=False)
```

```
[57] employee_df
```

- At first the employee_df dataset is shuffled and the emp_id is dropped as it doesn't contribute for analysis in the future. Next we separated the numerical columns from categorical columns and numerically encoded our categorical values from the new dataset.
- After identifying which features contained the categorical data, we started digitally encoding the data. To do this, we used Pandas' get_dummies method in

Python. This method creates a dummy variable encoded from a categorical variable.

```
✓ [58] employee_df = employee_df.sample(frac=1).reset_index(drop=True)
      employee_df

✓ [59] employee_df=employee_df.drop(columns=['Emp ID'])

✓ [60] categorical = []
      for col, value in employee_df.iteritems():
          if value.dtype == 'object':
              categorical.append(col)

      # Store the numerical columns in a list numerical
      numeric = employee_df.columns.difference(categorical)
      numeric

Index(['average_monthly_hours', 'dept', 'last_evaluation', 'number_project',
      'salary', 'satisfaction_level', 'time_spend_company'],
      dtype='object')

✓ [61] at_cat = employee_df[categorical]
      at_cat = at_cat.drop(['Attrition'], axis=1)

✓ [62] at_cat = pd.get_dummies(at_cat)
      at_cat.head(3)
```

The attrition column(target) contains categorical values which have been encoded to numeric values i.e., we digitally encoded by creating a dictionary that mapped the values of Yes to 1 and No to 0.

```
✓ [63] at_num = employee_df[numeric]
```

```
✓ ▶ at_final = pd.concat([at_num, at_cat], axis=1)  
at_final.head()
```

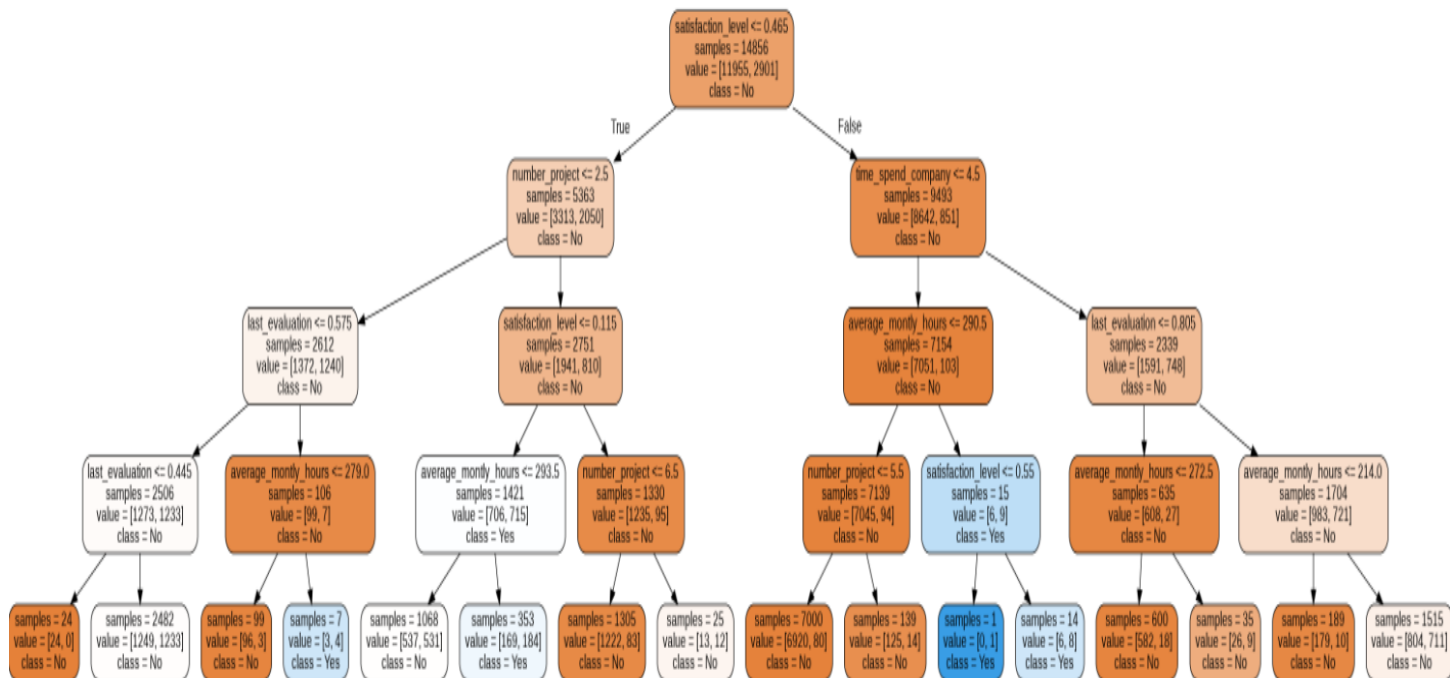
```
✓ [65] target_map = {'Yes':1, 'No':0}  
# Use the pandas apply method to numerically encode our attrition target variable  
target = employee_df["Attrition"].map(target_map)  
target.head(5)
```

```
0    0  
1    1  
2    0  
3    0  
4    0  
Name: Attrition, dtype: int64
```

We split the data into both training and testing sets as well as for validation and testing using the function `train_test_split()`.

```
✓ ▶ from sklearn.model_selection import train_test_split  
  
# Split data into train and test sets as well as for validation and testing  
train, test, target_train, target_val = train_test_split(at_final,  
                                                         target,  
                                                         train_size= 0.80,  
                                                         random_state=0);
```

A decision tree has been constructed to traverse through all the features of our dataset. The `DecisionTreeClassifier` object which is obtained from `sklearn` has been used. `sklearn` has an `export_graphviz` method to export the tree diagram to `.png` format which can be viewed from the output of this kernel.



Performances Evaluation

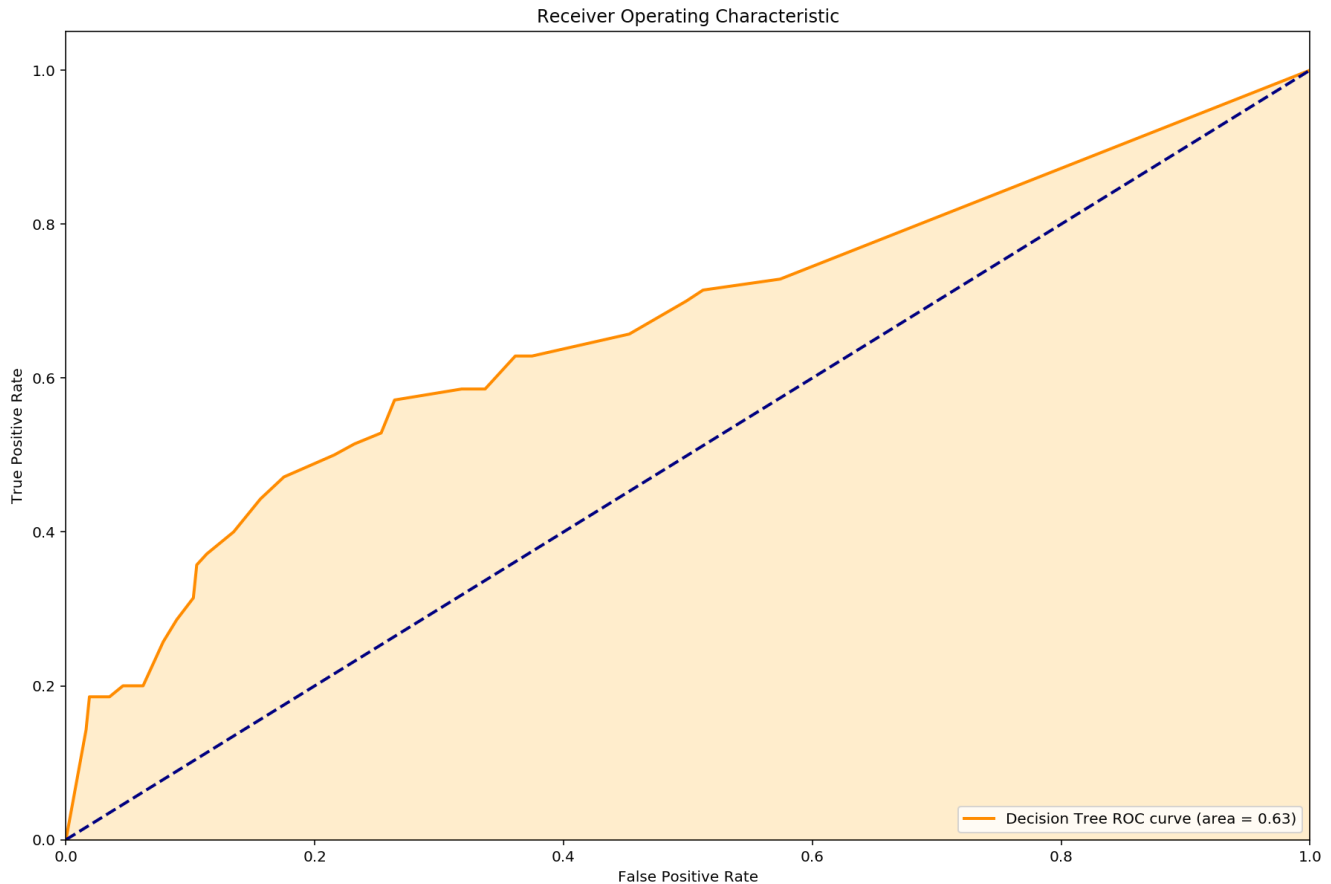
✓
0s



```
from sklearn import metrics
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
print("Accuracy score: {}".format(accuracy_score(target_val, y_pred)))
```

➤ Accuracy score: 0.7945611200861604

As observed, our decision tree provides about 80% accuracy for its predictions.



Result

- Satisfaction is starting to decline as more than four projects have been taken over by left employees, which can also be seen as a trend for current employees. Therefore, as a suggestion to further maintain employee satisfaction, the company needs to ensure that employees do not have too many projects assigned to them and thus lose the excitement that could lead to employees leaving.
- The salary system of people who have suffered a work accident is in the mid-low range, and the satisfaction level was analyzed to be about 50%. However, checking this information with people working at the company, It was found that their salaries were in the low, medium, and high range, and that their satisfaction was even higher. This shows that the company maintains a good salary system against work accidents.

Project Contribution

Akhila Dandu - Data Preprocessing and Data visualization

Snigdha Pendhota - Implementation of Decision tree and Performances Evaluation