

Organizational Employee Attrition Classification

Rahul Kotecha

Employee Attrition

- Employee Attrition is a term that describes the situation of an employee leaving a company
- Employee attrition can be for any given reason- voluntary, involuntary, internal, retirement, death, etc.
- HR analyze company's attrition rate to know how many employees left the company over a specific period of time
- Employee Attrition analysis helps understand what factors company needs to improve to preserve good working talent
- Top- level management uses Employee Attrition analysis to develop retention strategies
- As per a research by People Analytics- salary is not the only major reason for employee attrition, there are several internal and external factors
- 75% of employee attrition factors can be prevented with right changes in the company and team policies



Dataset

- The dataset used for the purpose of the project: IBM HR Analytics Employee Attrition & Performance
- The dataset was created by Data Scientists at IBM (Fictional dataset)
- Dataset provider: Pavan Subhash
- Dataset source: Kaggle
- Dataset year of publication: 2016
- Kaggle Usability score: 8.8/10
- Dataset link: [IBM HR Analytics Employee Attrition & Performance](#)



About the project

- The dataset has 35 total columns/features/ dimensions- 34 independent features and 1 dependent feature
- The dataset has a total of 1470 records
- The dataset has 9 Categorical features and the rest are Numerical features
- We are using Python programming language and its important libraries- Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, Imbalanced- learn for the purpose of the project
- Project requires to analyze which features are actually important for the binary- classification task
- The dependent feature is imbalanced in the dataset- if developed a model with the same, may lead to more likelihood of underfitting or overfitting



Project Question

- What are the major factors that lead to employee attrition?
- Based on 34 different factors classify will the employee leave the company or not?

Data Dictionary

- Age: The age represents the age of the employees. The mean age of all the employees in the company is 37. The lowest age is 18 years and the highest age is 60 years for an employee
- Daily Rate: The Daily Rate represents the per-day pay of an employee. The average daily rate is 802 with a minimum value of 102 and maximum of 1499 for an employee
- Distance from Home: Distance from Home explains how far the employee has to travel daily to reach to office and also explains how much time and money he has to spend to reach to the work location. The average distance from home is 2.91 units, with a Minimum value of 1 unit and a Maximum value of 29 units
- Education: Education has 5 major categories- Below college, college, bachelors, master, doctor. The average population has a Bachelor's degree, very small group of people have a Doctorate education
- Employee Number: Employee number refers to the Employee internal- organizational unique identification number



Data Dictionary

- **Environment Satisfaction:** Environment Satisfaction refers to the level of comfort and ease an employee feels in a work environment. Environment Satisfaction has 4 categories- low, medium, high and very high
- **Hourly rate:** The Daily Rate represents the hourly pay of an employee. The average daily rate is 65.89 with a minimum value of 30 and maximum of 100 for an employee
- **Job Involvement:** Job involvement described the level of devotion that employees show towards their work in an organization. Job Involvement has 4 categories- low, medium, high and very high
- **Job level:** Job level refers to the level of seniority in an organization and also relates to the amount of responsibility and rewards
- **Gender:** gender refers to the sex of the employee- Male, female, etc
- **Marital Status:** Marital status provides information regarding the marriage of an employee, this can include- Single, Married, Divorced, etc



Data Dictionary

- Job Satisfaction: Job satisfaction refers to the feeling of enjoyment and fulfilment from the work an individual undertakes. Job satisfaction has 4 categories- low, medium, high and very high
- Performance Rating: Performance rating refers to the level of work undertaken which in terms is relative to the expectations of the manager. Performance rating has 4 main categories- low, good, excellent and outstanding
- Relationship satisfaction: Relationship satisfaction refers to all relationships that an employee has built over time be it with the organization itself, other employees, other stakeholders, etc. Relationship satisfaction has 4 major categories- low, medium, high and very high
- Work-Life Balance: Work- life balance refers to how well are the employees of an organization able to prioritize their work and their personal life outside the organization. For the study we have 4 major categories in which we define the work- life balance- bad, good, better and best
- Standard Hours: Standard Hours refers to the number of hours an employee works per week



Data Dictionary

- Years at Company: This feature describes how long an employee has been associated with a specific organization
- Training Times Last Year: This feature describes how long an employee had undergone training in the last year
- Department: Department describes which organizational team the employee works in
- Years With Current Manager: This feature describes how long an employee has been working with his current manager
- Business Travel: This feature describes how often an employee has to travel because of business need-rarely, frequent, non-travel



Data Dictionary

Category No	Education	Environment Satisfaction	Job Involvement	Job Satisfaction	Performance Rating	Relationship Satisfaction	Work Life Balance
1	Below College	Low	Low	Low	Low	Low	Bad
2	College	Medium	Medium	Medium	Good	Medium	Good
3	Bachelor	High	High	High	Excellent	High	Better
4	Master	Very High	Very High	Very High	Outstanding	Very High	Best
5	Doctor	-	-	-	-	-	-



What is Machine Learning

- Machine Learning is a branch of Computer Science that allows the computer to learn without explicitly being programmed
- Machine learning can also be described as the process of extracting patterns or structures from data and using it to make predictions in regards to the data.
- To summarize it all we can say that Machine Learning is the semi-automated extraction of knowledge from data. In Machine Learning we use/implement programming code, algorithms on data in order to derive useful insights from the data.
- Machine Learning has a very different process compared to traditional learning.
- Machine learning can be described as a combination of Computer Science, Business understanding, and Statistics.
- The branch of Machine Learning can broadly be classified into- Supervised Machine Learning, Unsupervised Machine Learning, and finally Reinforcement Machine Learning



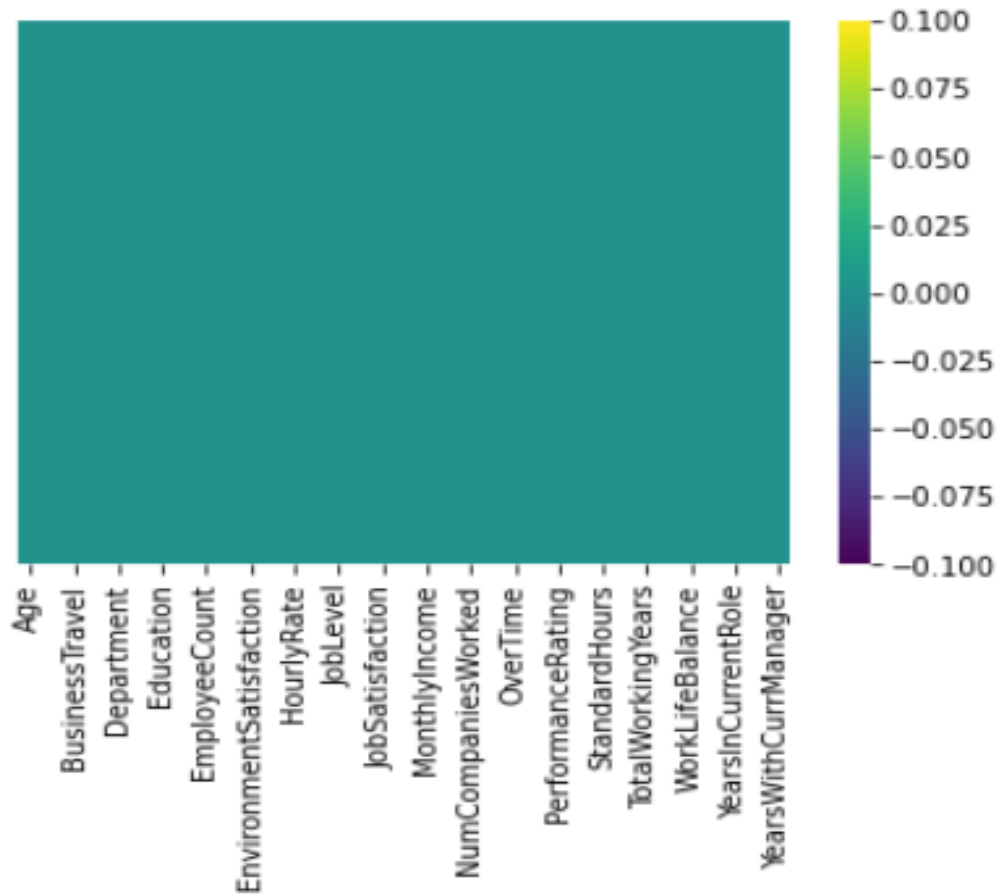
Exploratory Data Analysis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                                Non-Null Count  Dtype
---  -
 0   Age                                  1470 non-null   int64
 1   Attrition                           1470 non-null   object
 2   BusinessTravel                       1470 non-null   object
 3   DailyRate                           1470 non-null   int64
 4   Department                           1470 non-null   object
 5   DistanceFromHome                     1470 non-null   int64
 6   Education                             1470 non-null   int64
 7   EducationField                       1470 non-null   object
 8   EmployeeCount                        1470 non-null   int64
 9   EmployeeNumber                       1470 non-null   int64
10   EnvironmentSatisfaction               1470 non-null   int64
11   Gender                               1470 non-null   object
12   HourlyRate                           1470 non-null   int64
13   JobInvolvement                       1470 non-null   int64
14   JobLevel                             1470 non-null   int64
15   JobRole                              1470 non-null   object
16   Jobsatisfaction                       1470 non-null   int64
17   MaritalStatus                        1470 non-null   object
18   MonthlyIncome                        1470 non-null   int64
19   MonthlyRate                          1470 non-null   int64
20   NumCompaniesWorked                   1470 non-null   int64
21   Over18                               1470 non-null   object
22   OverTime                             1470 non-null   object
23   PercentSalaryHike                    1470 non-null   int64
24   PerformanceRating                    1470 non-null   int64
25   RelationshipSatisfaction              1470 non-null   int64
26   StandardHours                        1470 non-null   int64
27   StockOptionLevel                     1470 non-null   int64
28   TotalWorkingYears                    1470 non-null   int64
29   TrainingTimesLastYear                1470 non-null   int64
30   WorkLifeBalance                      1470 non-null   int64
31   YearsAtCompany                       1470 non-null   int64
32   YearsInCurrentRole                   1470 non-null   int64
33   YearsSinceLastPromotion              1470 non-null   int64
34   YearsWithCurrManager                 1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

- We are using Python's Pandas library to draw useful insights out of the data
- There are a total of 35 columns in the dataset
- There are a total of 1470 records/ rows
- 26 columns have Integer datatype and 9 columns have object datatype
- The entire dataset consumes 402.1+ KB of memory
- The names of different columns can be seen in the chart
- We can see if a specific column has a null value or not



Checking for Null values



- We have created a Heatmap to depict the presence of Null values in the columns of the dataset

Code:

```
sns.heatmap(df.isna(),yticklabels=False,cmap="viridis")  
plt.show()
```

- On looking at the Heatmap we can see that there is no Null value in any of the columns in the dataset

Descriptive statistics for Numerical Columns

Column	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel
count	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00	1470.00
mean	36.92	802.49	9.19	2.91	1.00	1024.87	2.72	65.89	2.73	2.06
std	9.14	403.51	8.11	1.02	0.00	602.02	1.09	20.33	0.71	1.11
min	18.00	102.00	1.00	1.00	1.00	1.00	1.00	30.00	1.00	1.00
25%	30.00	465.00	2.00	2.00	1.00	491.25	2.00	48.00	2.00	1.00
50%	36.00	802.00	7.00	3.00	1.00	1020.50	3.00	66.00	3.00	2.00
75%	43.00	1157.00	14.00	4.00	1.00	1555.75	4.00	83.75	3.00	3.00
max	60.00	1499.00	29.00	5.00	1.00	2068.00	4.00	100.00	4.00	5.00






Descriptive statistics (Continued)






Column1	Relationship Satisfaction	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear
count	1470.00	1470.00	1470.00	1470.00	1470.00
mean	2.71	80.00	0.79	11.28	2.80
std	1.08	0.00	0.85	7.78	1.29
min	1.00	80.00	0.00	0.00	0.00
25%	2.00	80.00	0.00	6.00	2.00
50%	3.00	80.00	1.00	10.00	3.00
75%	4.00	80.00	1.00	15.00	3.00
max	4.00	80.00	3.00	40.00	6.00

Descriptive statistics (Continued)

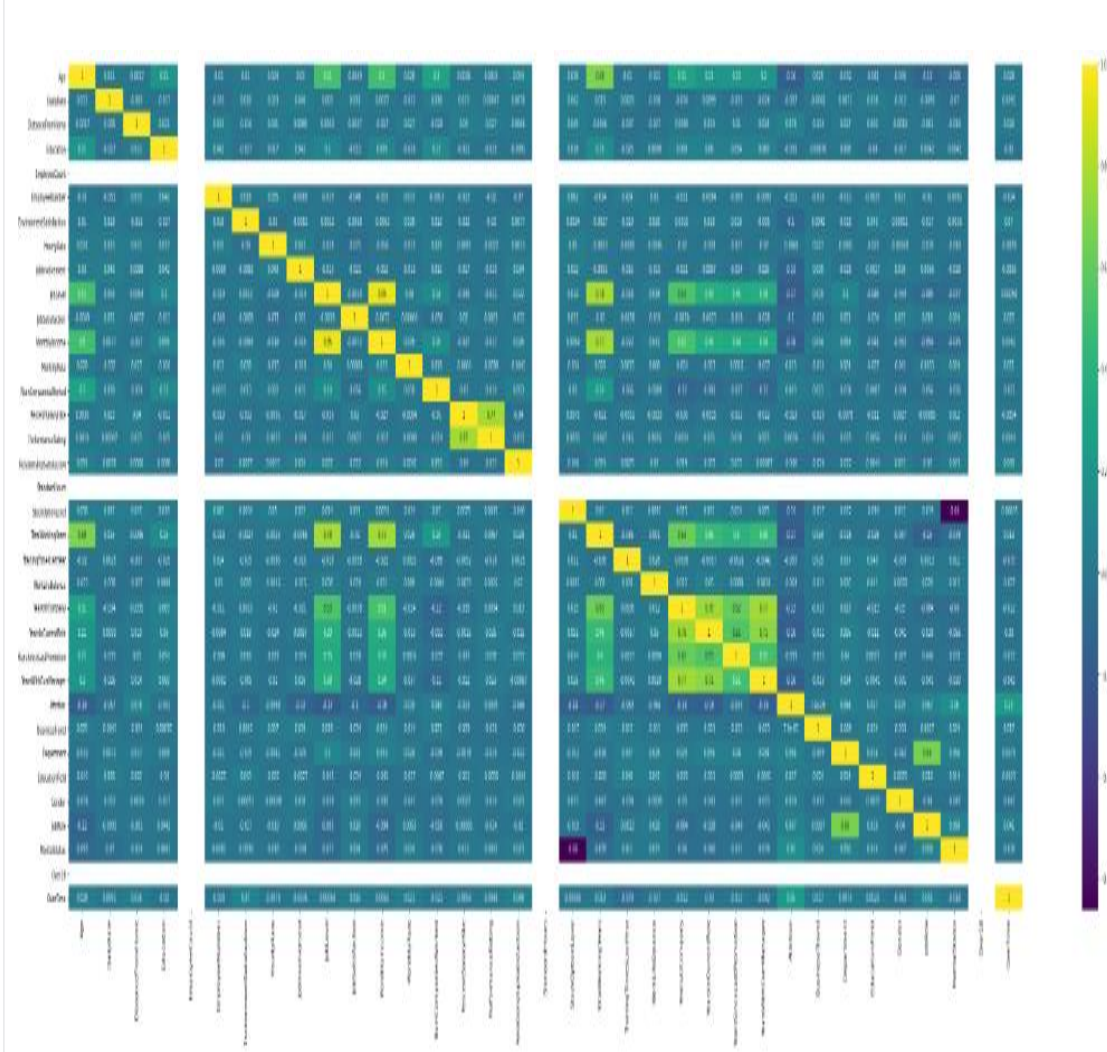
Column1	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
count	1470.00	1470.00	1470.00	1470.00	1470.00
mean	2.76	7.01	4.23	2.19	4.12
std	0.71	6.13	3.62	3.22	3.57
min	1.00	0.00	0.00	0.00	0.00
25%	2.00	3.00	2.00	0.00	2.00
50%	3.00	5.00	3.00	1.00	3.00
75%	3.00	9.00	7.00	3.00	7.00
max	4.00	40.00	18.00	15.00	17.00

Descriptive statistics for Categorical Columns

Column 	Attrition 	BusinessTravel 	Department 	EducationField 
count	1470	1470	1470	1470
unique	2	3	3	6
top	No	Travel_Rarely	Research & Development	Life Sciences
freq	1233	1043	961	606

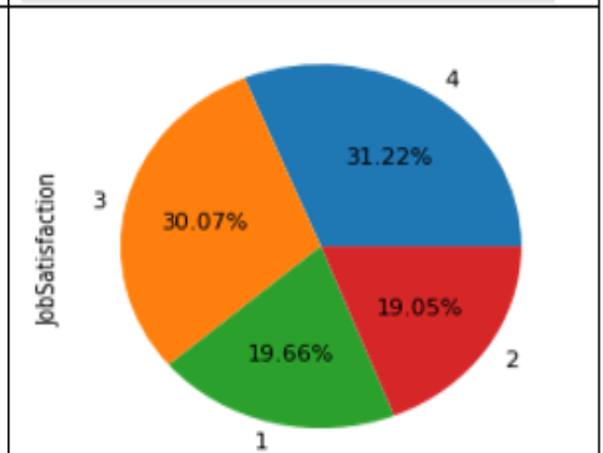
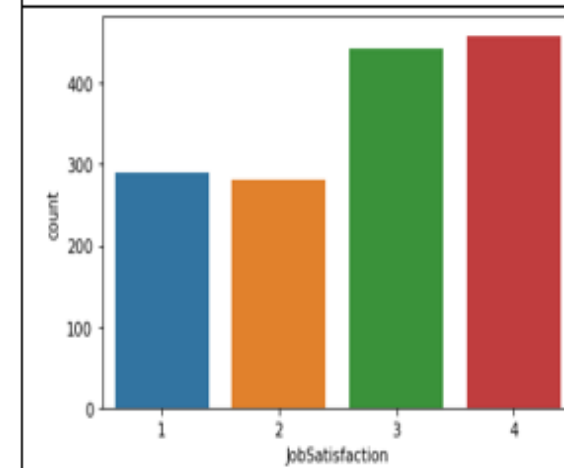
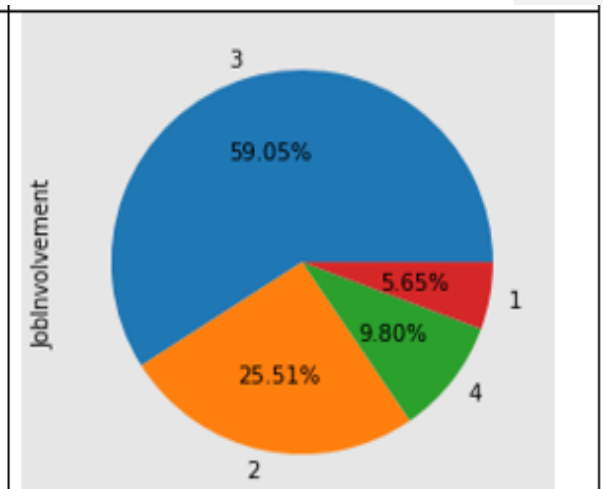
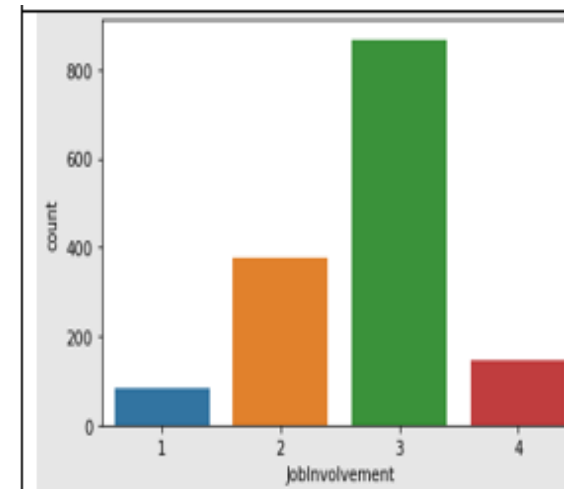
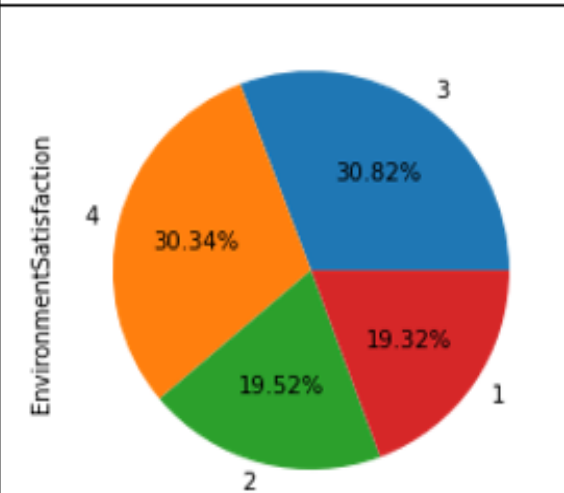
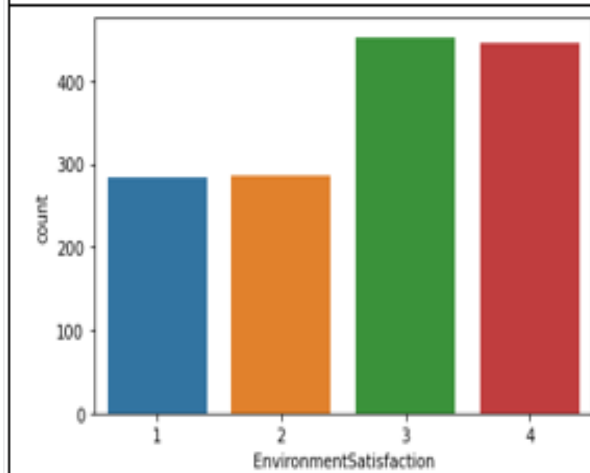
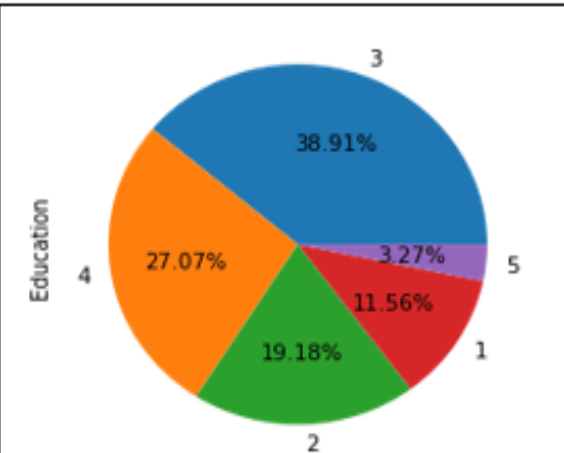
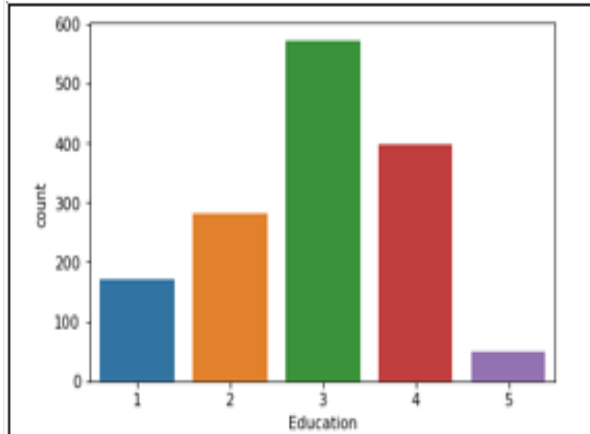
Column1 	JobRole 	MaritalStatus 	Over18 	OverTime 
count	1470	1470	1470	1470
unique	9	3	1	2
top	Sales Executive	Married	Y	No
freq	326	673	1470	1054

Correlation of Independent variables with the Dependent variable

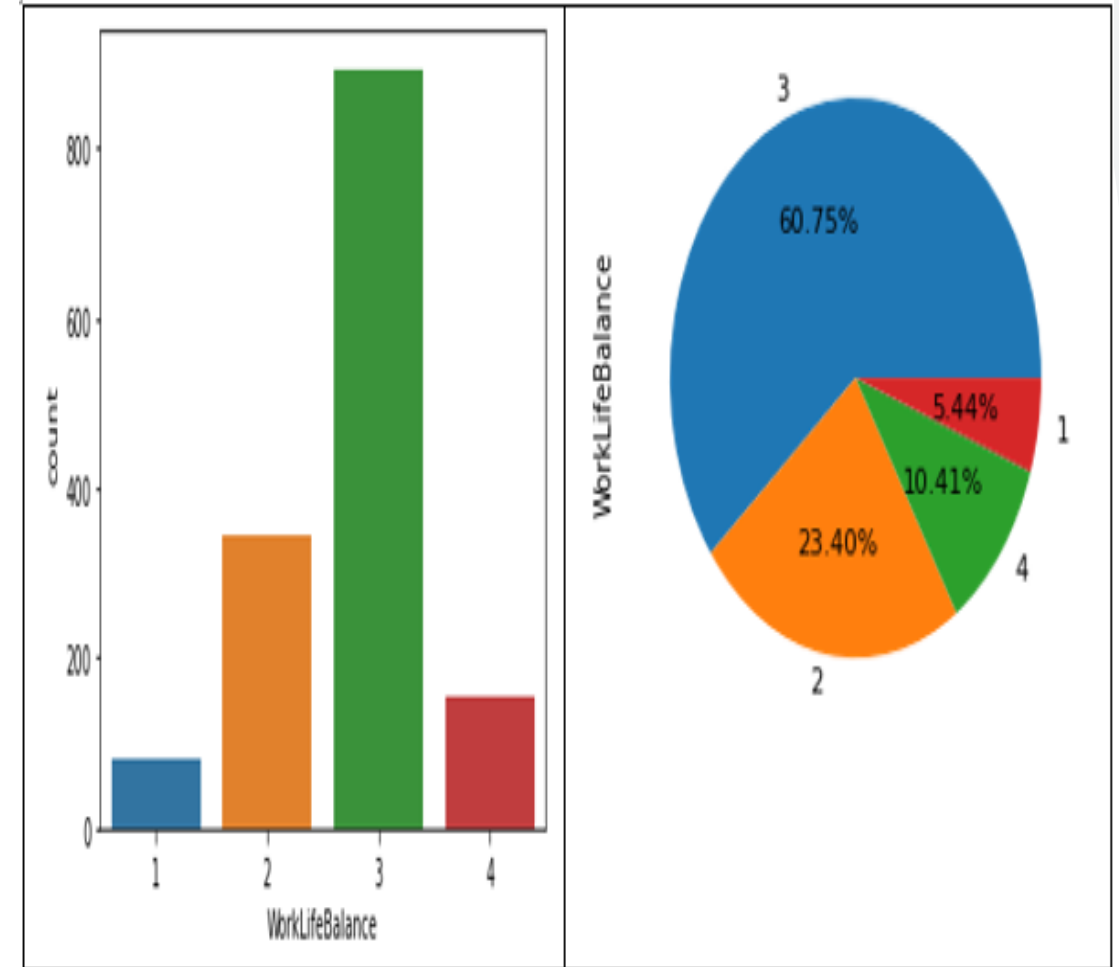
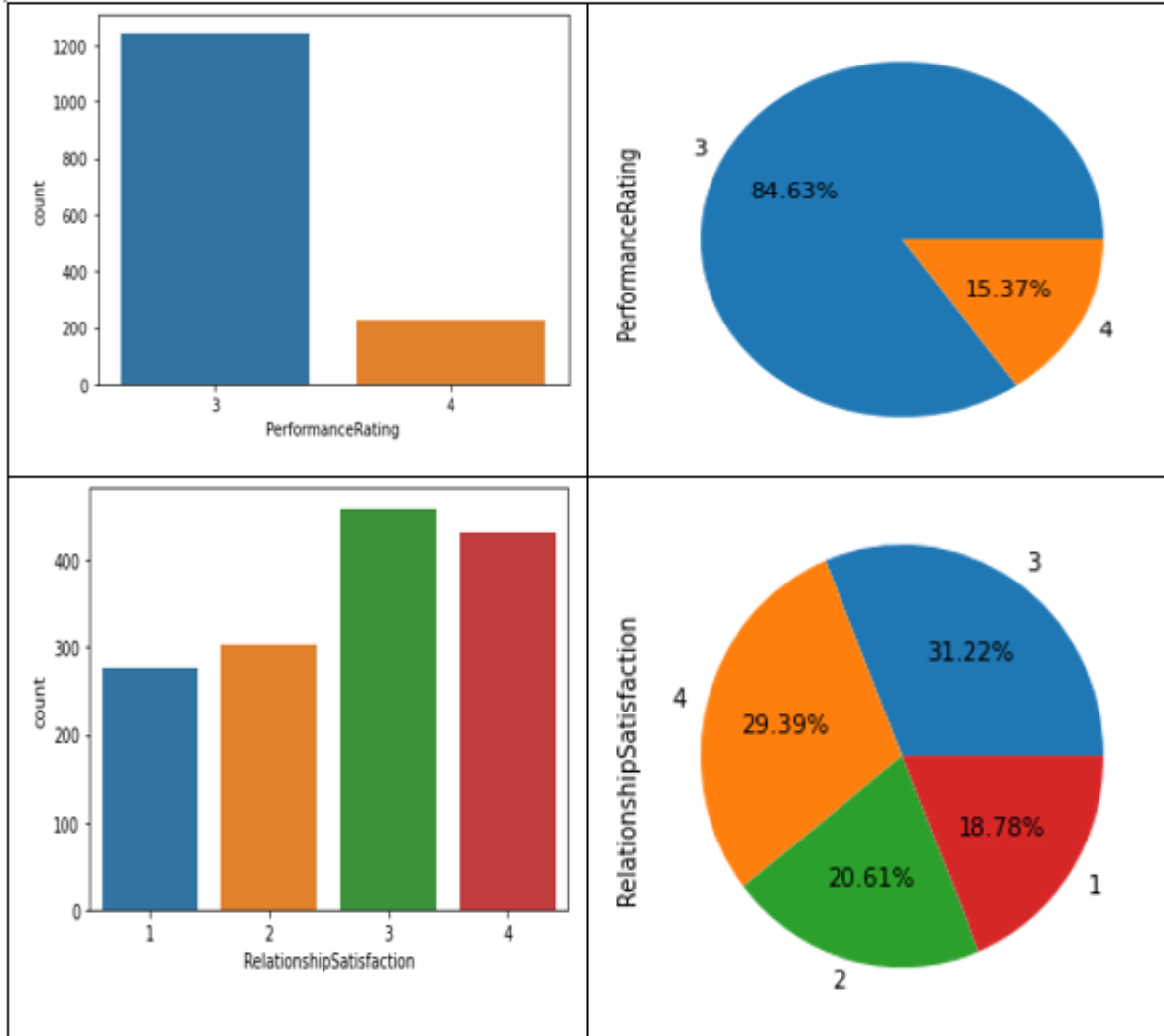


TotalWorkingYears	-0.171063
JobLevel	-0.169105
YearsInCurrentRole	-0.160545
MonthlyIncome	-0.159840
Age	-0.159205
YearswithCurrManager	-0.156199
StockOptionLevel	-0.137145
YearsAtCompany	-0.134392
JobInvolvement	-0.130016
JobSatisfaction	-0.103481
EnvironmentSatisfaction	-0.103369
WorkLifeBalance	-0.063939
TrainingTimesLastYear	-0.059478
DailyRate	-0.056652
RelationshipSatisfaction	-0.045872
YearsSinceLastPromotion	-0.033019
Education	-0.031373
PercentSalaryHike	-0.013478
EmployeeNumber	-0.010577
HourlyRate	-0.006846
BusinessTravel	0.000074
PerformanceRating	0.002889
MonthlyRate	0.015170
EducationField	0.026846
Gender	0.029453
NumCompaniesWorked	0.043494
Department	0.063991
JobRole	0.067151
DistanceFromHome	0.077924
MaritalStatus	0.162070
OverTime	0.246118
Attrition	1.000000
EmployeeCount	NaN
StandardHours	NaN
Over18	NaN

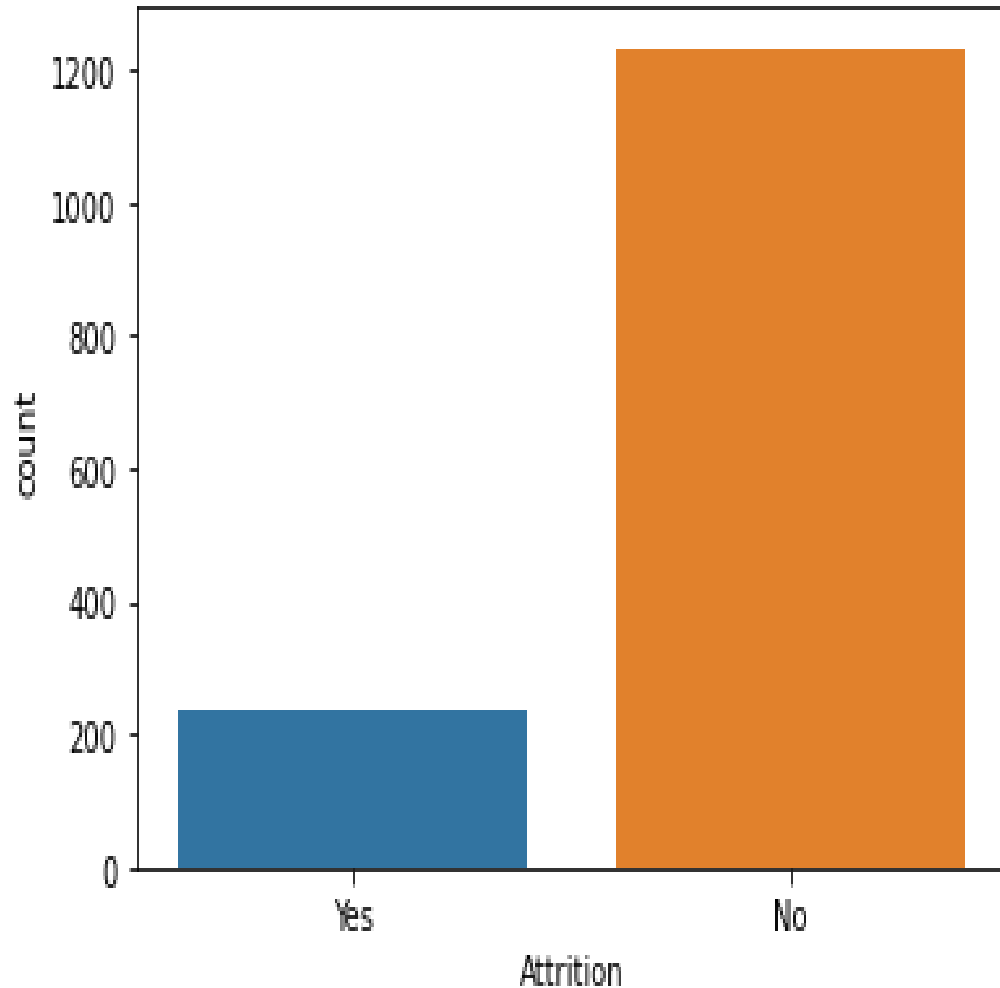
Data Visualization



Data Visualization (Continued)



Imbalanced Dataset

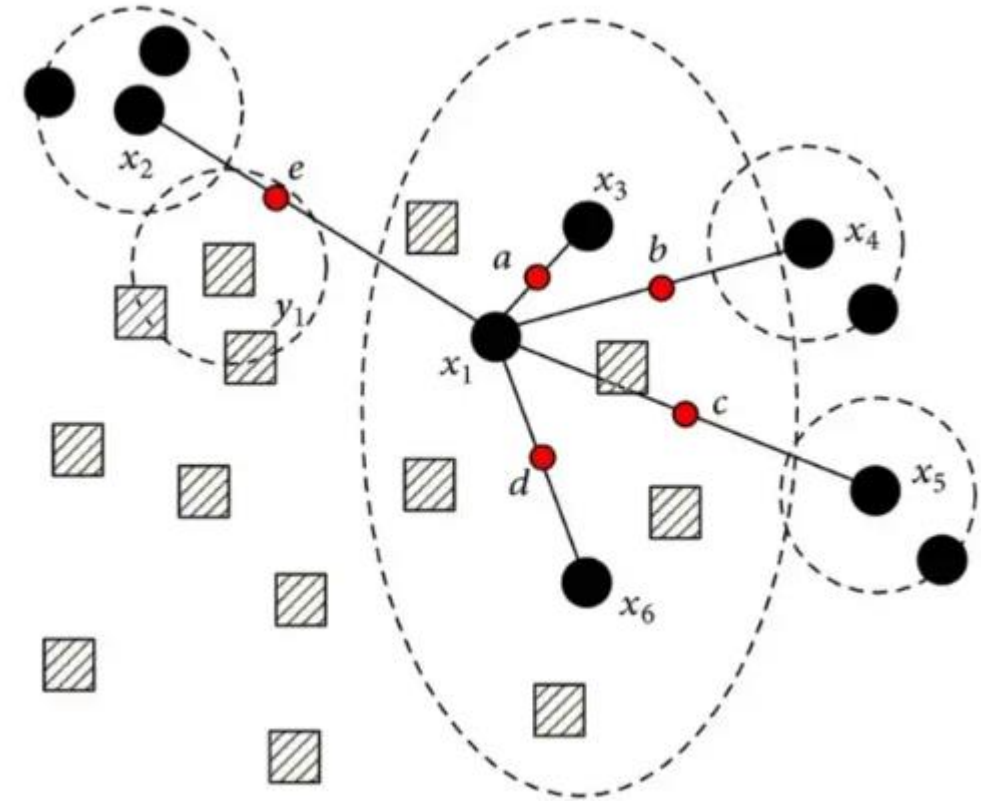


- By means of undertaking this project, our goal is to be able to predict based on multiple factors, will an employee leave the organization or not
- Based on the data set, we try to undertake Machine Learning, i.e, teaching the Machine to be able to predict if the employee will leave the organization or not
- However from the current dataset, the machine will learn more examples of employees not leaving the company and less examples of the employees leaving the company
- YES(1) to attrition is a total of $237 / 1470 = 16.1\%$ of total and NO(0) to attrition is a total of $1233 / 1470 = 83.9\%$ of total

Techniques to deal with Imbalanced Dataset

SMOTE- Synthetic Minority Oversampling Technique:

- A statistical method for evenly expanding the number of instances in your dataset
- For each of the samples in the class, SMOTE determines the n -nearest neighbors in the minority class
- Following that, it creates random spots on the lines between the neighbors, thus generating new data points
- These data points will have consideration of the data and will also help computer learn more about the minority class



(Source: [ResearchGate](#))

Feature Selection

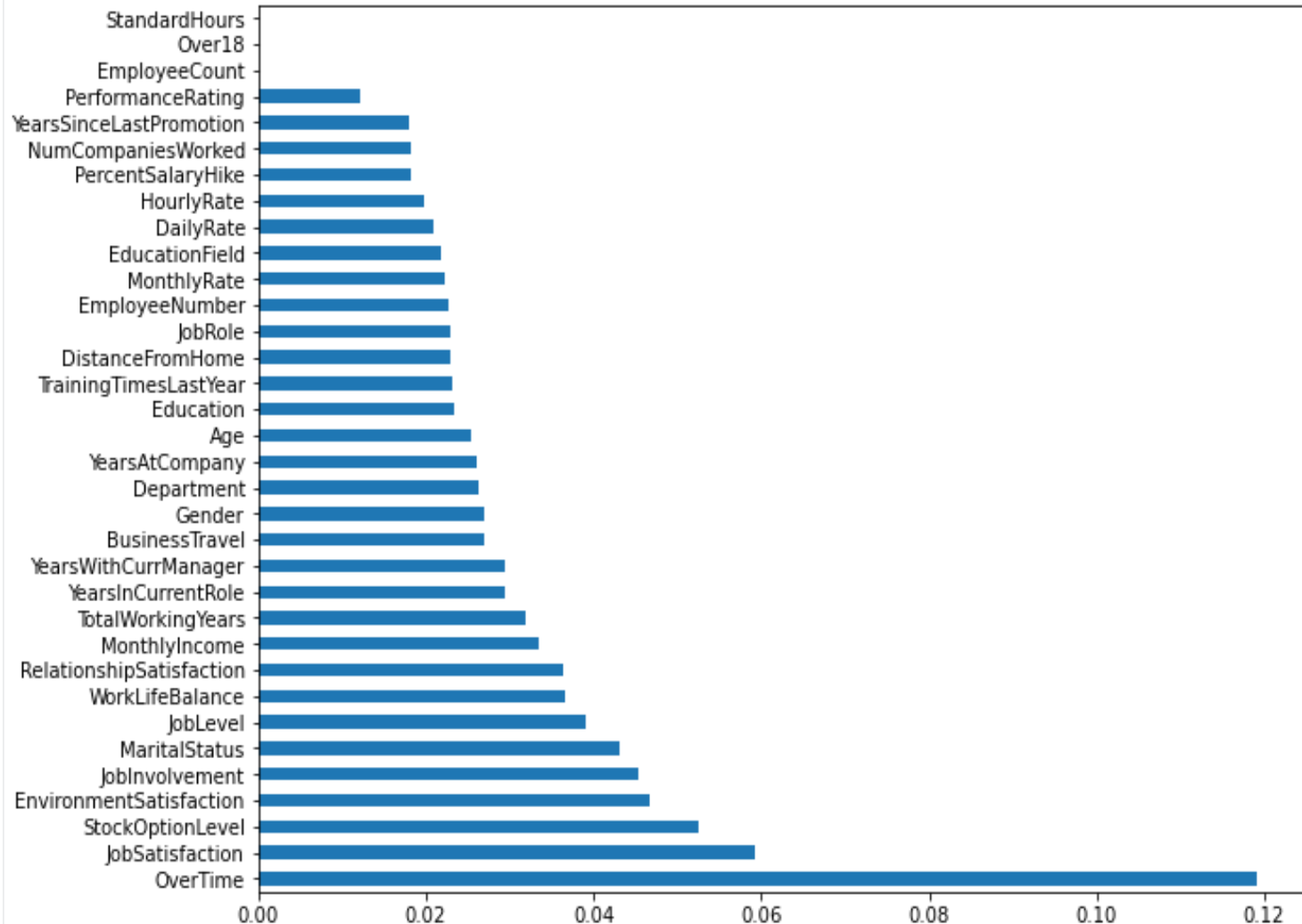
- Feature selection is undertaken using the Extremely Randomized Trees Classifier, also known as Extra Trees Classifier
- Extra Trees Classifier is an Ensemble learning technique
- In Extra Trees Classifier- for the construction of the forest, the normalized total reduction in Gini Index is used
- In some ways it is similar to Random Forest Classifier
- Each feature is sorted in descending order of Gini Index/ Gini Importance
- For calculation of Gini Index, we need to calculate Entropy

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2(p_i)$$

(Source: [Entropy](#))

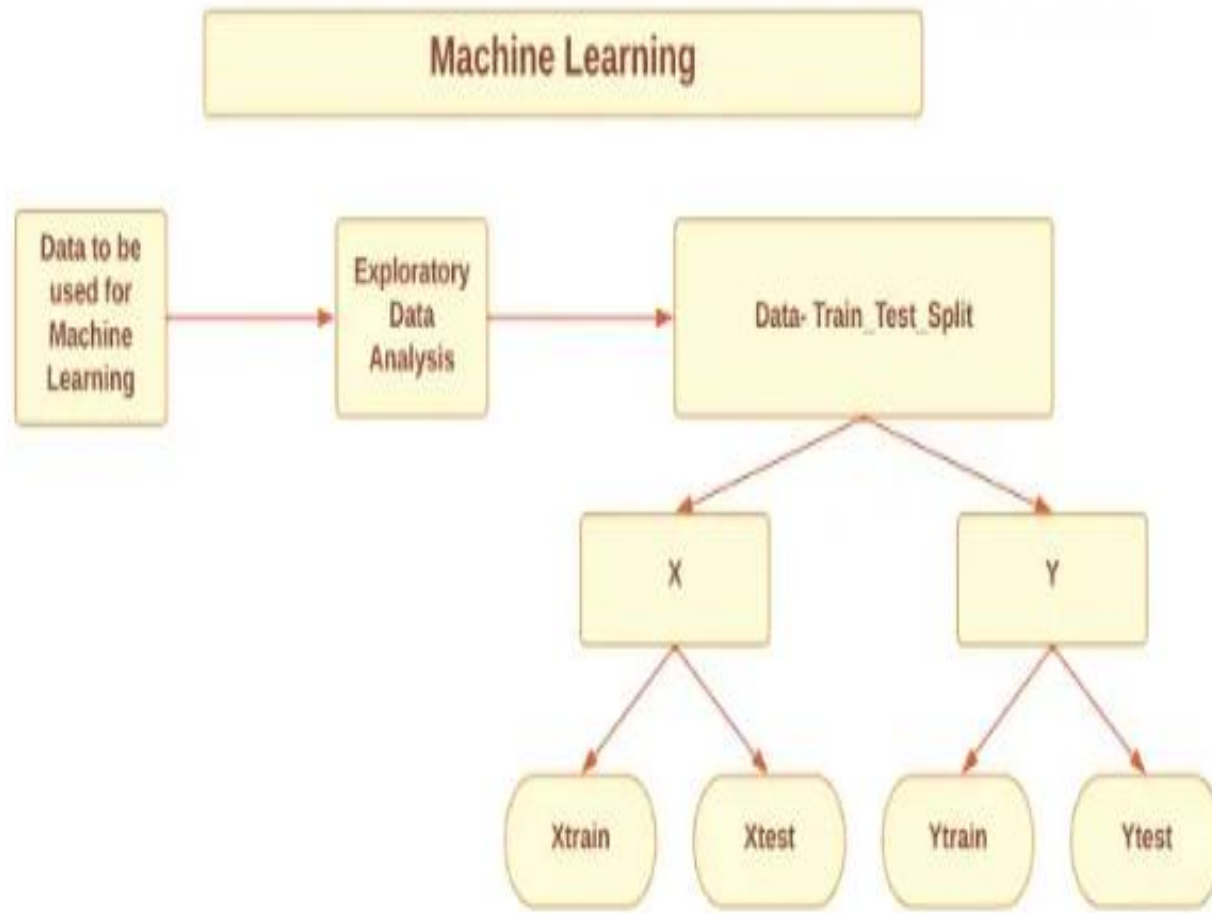
$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

(Source: [Information Gain](#))



- Based on Extra Trees Classifier, the most important feature that has an impact on the feature- 'Attrition' is 'OverTime'
- The features with the least importance in terms of making a prediction regarding Attrition are- 'StandardHours', 'Over18' and 'EmployeeCount'
- Based on analysis, we remove the 3 least important features because their individual importance is less than 0.01

Train- Test Split



- Since we have the independent variables- age, dailyrate, education, etc and we have the dependent variable- attrition, we will split the dataframe- independent variables as “x” and the dependent variable as “y”. Post the x and y split we will undertake the train_test_split
- In order to undertake the creation of a model what we do in Machine Learning is that we divide the existing data into suppose 70% and 30% wherein 70% of data we will use for Training and 30% for testing. What this means is that at random we will split the data into 70% and 30% portions and these x and y will further be split into xtrain and ytrain portion and the xtest and ytest portion.

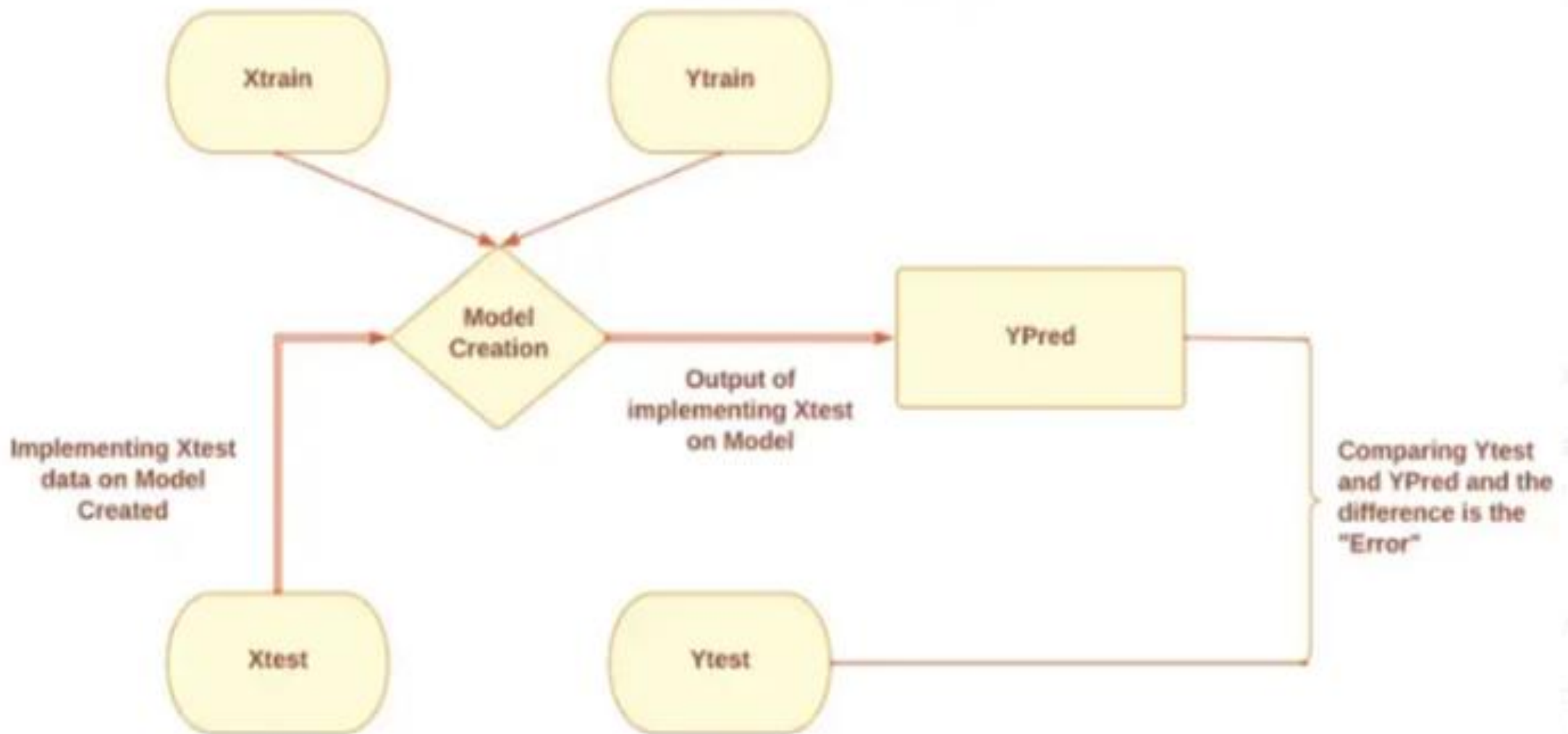
Feature Scaling

- When dealing with a real- world dataset, there often are different types of variables that have different scales of measurement.
- Example: weight of a human, salary of a human, etc
- To study these features together, it is important to reduce these independent variables/ features to a common ground or an even field where they can be compared/ analysed and used to predict something
- This process is known as Feature Scaling
- Scaling can be undertaken by different ways- Standardization and Normalization
- Usually StandardScaler is used when the data has a normal distribution and otherwise we use the MinMaxScaler or Normalization

Standard Scaler	$\frac{x_i - \text{mean}(x)}{\text{stdev}(x)}$
MinMax Scaler	$\frac{x_i - \min(x)}{\max(x) - \min(x)}$

(Source: [Medium.com](#))





Classification Models

- Classification can be described as the process of categorizing/ grouping items into groups
- Classification is a type of Supervised Machine Learning task
- These groups can be based on variable factors depending upon the type and item of classification being undertaken
- If there are only 2 groups- Binary classification and if there are more than 2 groups- Multi- class classification
- The major Machine Learning Classifiers are as follows:
 - KNeighborsClassifier
 - Logistic Regression
 - GaussianNB
 - XGBClassifier
 - SVC- Support Vector Classifier
 - DecisionTreeClassifier
 - RandomForestClassifier



Model Accuracy Analysis

Model Name	Accuracy	Cross Validation	F1 (Attrition)	F1 Score (Non-Attrition)
KNeighborsClassifier	0.83	0.71	0.84	0.84
LogisticRegression	0.81	0.61	0.78	0.83
GaussianNB	0.75	0.72	0.74	0.76
XGBClassifier	0.90	0.67	0.88	0.92
SVC	0.85	0.59	0.83	0.87
DecisionTreeClassifier	0.83	0.59	0.81	0.84
RandomForestClassifier	0.91	0.87	0.89	0.92



Conclusion

- The dataset is too small to make very perfect decision in regards to 35 dimensional data
- With an increase in data, we can actually let the machine learn better and also improve accuracy
- SMOTE is a better technique to deal with Imbalanced data as compared to RandomUnderSampling and RandomOverSampling as they lead to extreme Overfitting and Underfitting situations
- Based on all model considerations and analysis, we find that the best machine learning model for the given dataset was the Random Forest Classifier
- The Type 1 and Type 2 errors are low with Random Forest Classifier
- The F1 score for both the categories is good which means we have reduced error to the maximum





THANK YOU

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030