Data Science and Artificial Intelligence: Project

Final Project: By Vivek Kulthe

Why Do Employees Quit?

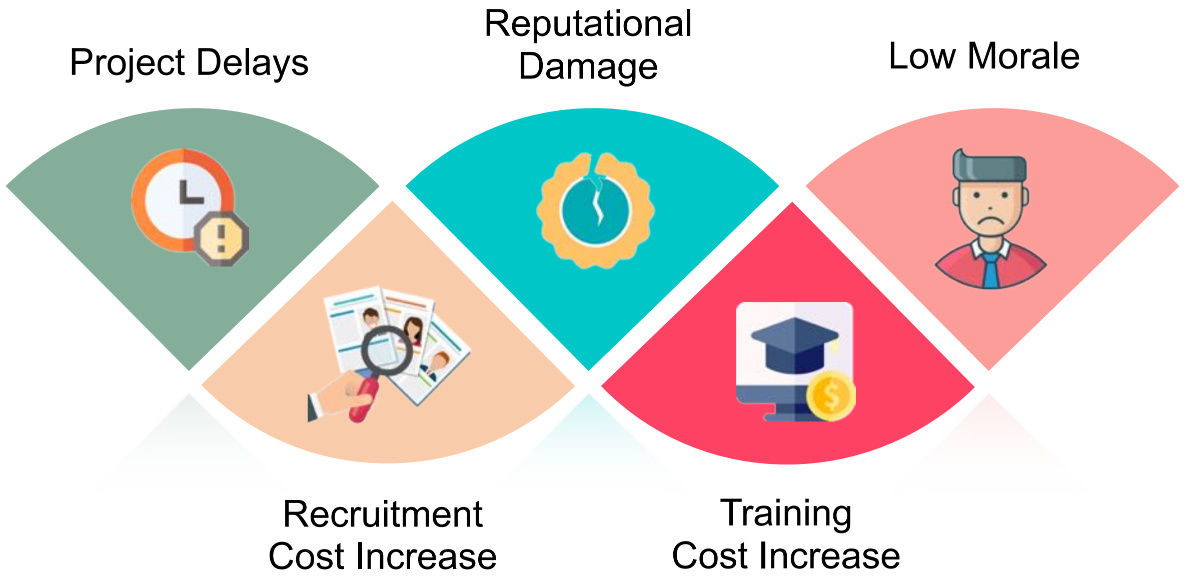
Employee Attrition Analysis & Prediction

Background

- **Organization Background:** ABC Technologies, is facing a concerning trend of high employee attrition ***even though they offer a competitive salary and benefits package.***

- **Problem Statement:** This suggests the ***root cause of attrition lies beyond financial compensation***. Company wants to *identify the underlying factors* driving employee departures to ***improve retention*** and maintain a strong talent pool.

- **Challenges Faced:**



Background (Team Background)

ABC Technologies has formed its **HR Analytics Team** with a combined skillset of HR expertise and data analytics

This team is expected to work with other teams such as Survey team, Communications team, HR Apps team and Data Engineering team to leverage existing employee ***data from HR applications*** and ***conduct targeted surveys*** to gather additional insights

This **data-driven approach** will allow for a deeper understanding of why employees leave, ultimately leading to a stronger talent pool for the company

Background (Skills & Resources Requirement)

Other Teams / Resources Needed:

Survey team to create targeted surveys in collaboration with HR experts

Internal Comms team to circulate and get survey responses from employees

HR Apps team to provide employee data that is already available through HR Applications

Data Engineering team to source & process raw data for further analytics and ML model building

Data Scientist to do Analytics and ML Model building

Required Skills / Tools / Libraries:

python + colab

pandas

NumPy

Microsoft Excel

matplotlib

seaborn

scikit learn

plotly

Objective

Identify factors influencing employee departures from ABC Technologies

GOAL

Develop a model to predict the likelihood of employee attrition

- Analyze employee data to uncover trends and patterns related to employee attrition.
- Identify key attributes associated with departing employees
- Visualize data to present findings clearly and concisely

EXPECTATIONS

- Build a logistic regression model to predict employee departure probability
- Use employee data as input features to train the model
- Analyze model performance to identify the most significant predictors of attrition

Objective (Expectations)

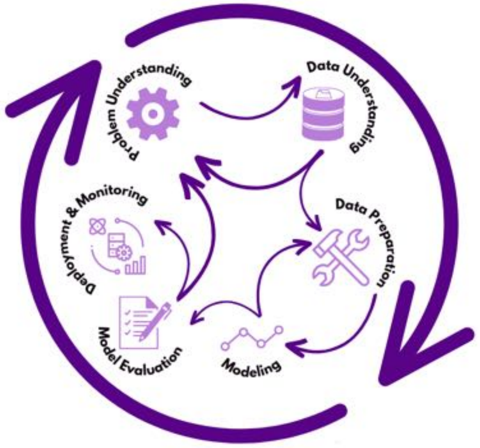
- **Analyze & Visualize the employee data** to uncover trends and patterns related to employee attrition
- Identify key attributes associated with departing employees, viz;
  - **Demographic Factors**
  - **Compensation and Benefits**
  - **Work Environment**
  - **Job Satisfaction**
  - **Managerial Influence**
  - **Career Progression**

- Build a **Classification Model** to predict if an employee is at a risk of attrition
- Analyze model performance to identify the most significant factors of attrition
- Provide recommendation to reduce the attrition

Objective (Benefits)

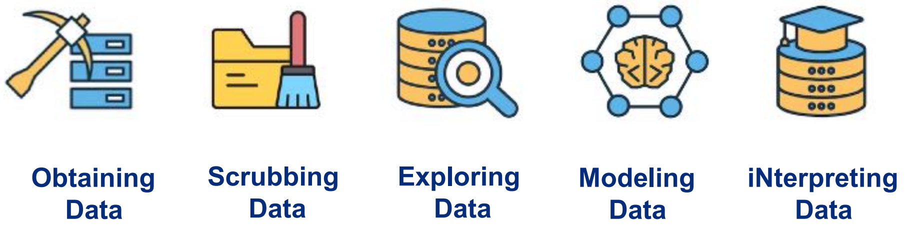
Approach / Methodology (EDA Framework)

CRISP-DM



- CRISP-DM is a traditional, iterative framework
- Better for traditional sectors and stakeholders
- CRISP-DM include business-focused phases like deployment, operations, and optimization
- Does not appeal where the data is not going to be iterative that required continuous learning and enhancement of the model

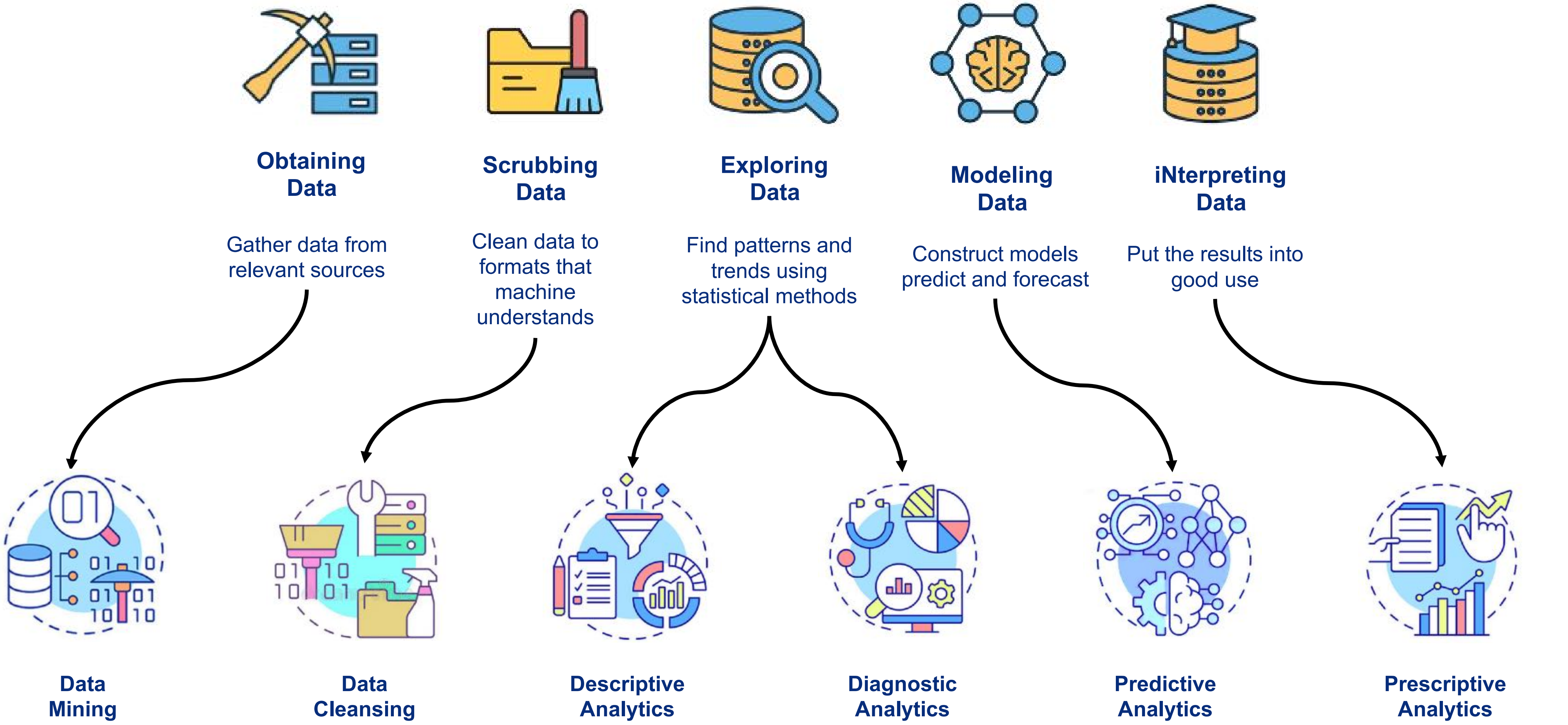
OSEMN



- A popular, non-iterative model
- Better for smaller and more focused projects, such as exploratory research
- It's a higher-level approach that doesn't include business-focused phases like deployment, operations, and optimization.
- OSEMN is appealing for startups or educational projects that want to iterate quickly

Analytical Technique (EDA Framework)

OSEMN



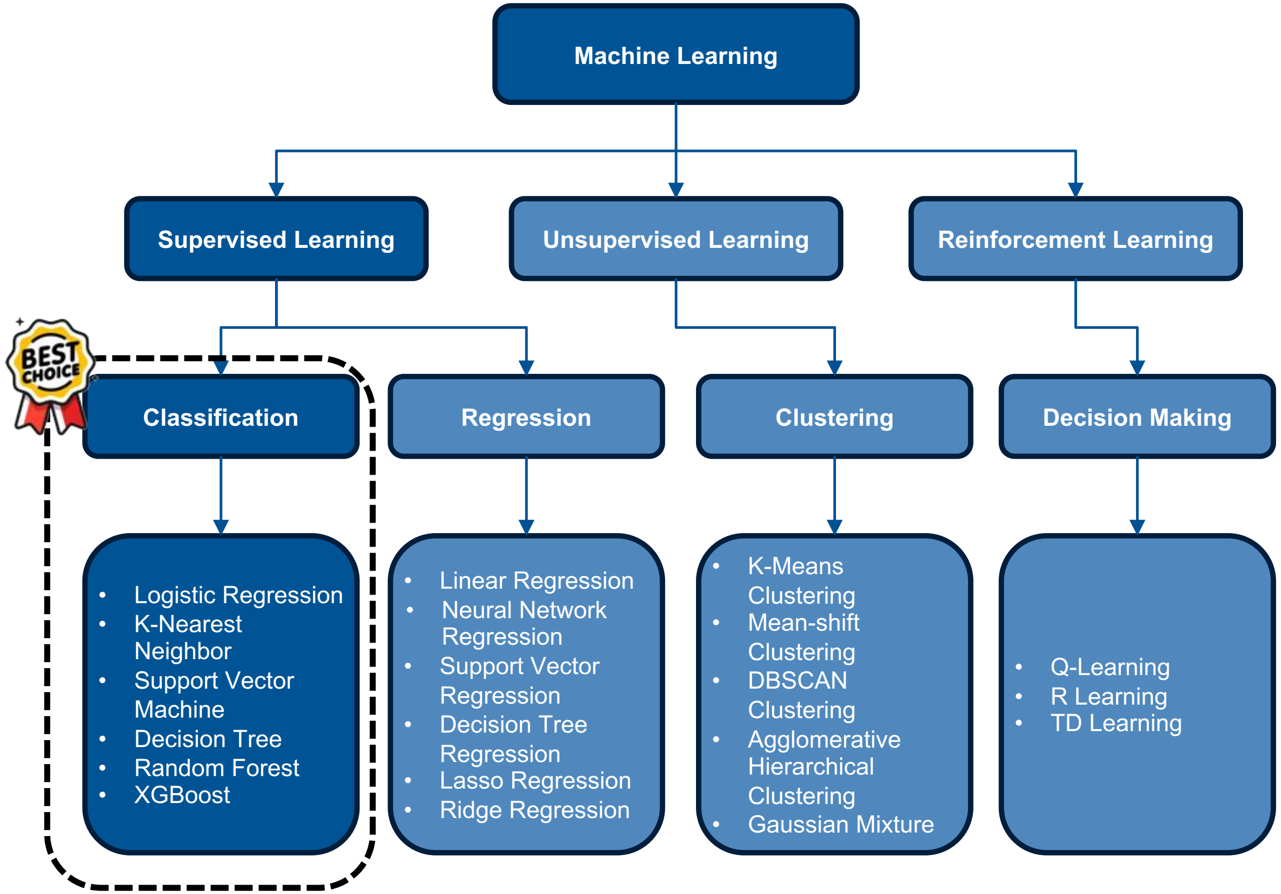
Analytical Technique (ML Model)

Key Factor To Decide Between Supervised and Unsupervised Learning:

- We have chosen **Supervised Learning** in this scenario as we have a labeled dataset and aim to make predictions.
- Supervised Learning is ideal for tasks like employee churn prediction, where our goal is to classify categories or predict continuous values based on past data.

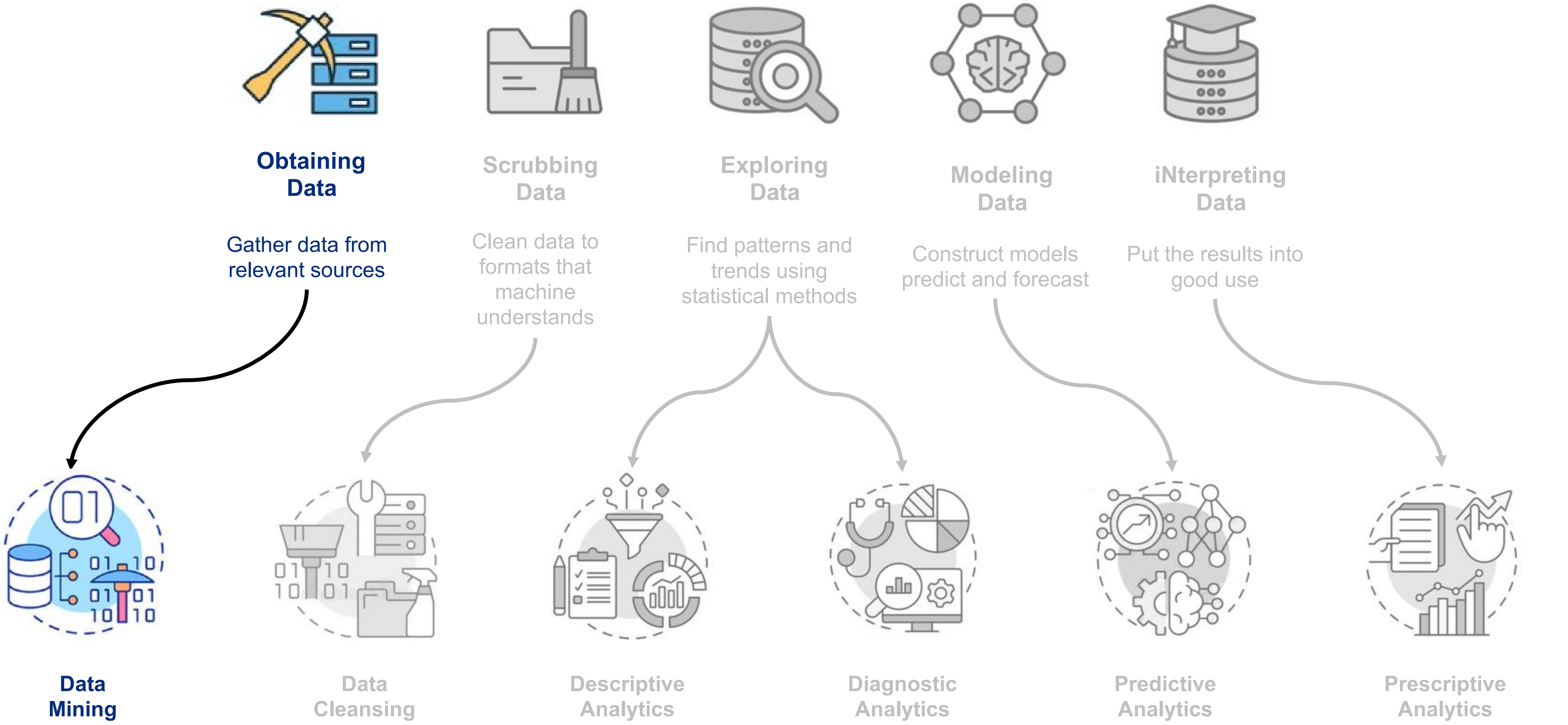
Key factor to deciding between Classification and Regression lies in target variable:

- **Classification:** Classification model is used when target variable is discrete and falls into distinct categories. These categories can be binary (like Attrition Yes/No) or have multiple classes (e.g., classifying handwritten digits into 0-9).
- **Regression:** Regression model is used when target variable is continuous. This means it can take on any numerical value within a range. Common examples include predicting house prices, weather forecasts (temperature), or customer lifetime value.



Obtaining Data

OSEMN



Obtaining Data

- The dataset used is taken from *kaggle* and contains ***HR analytics data of employees that stay and leave***

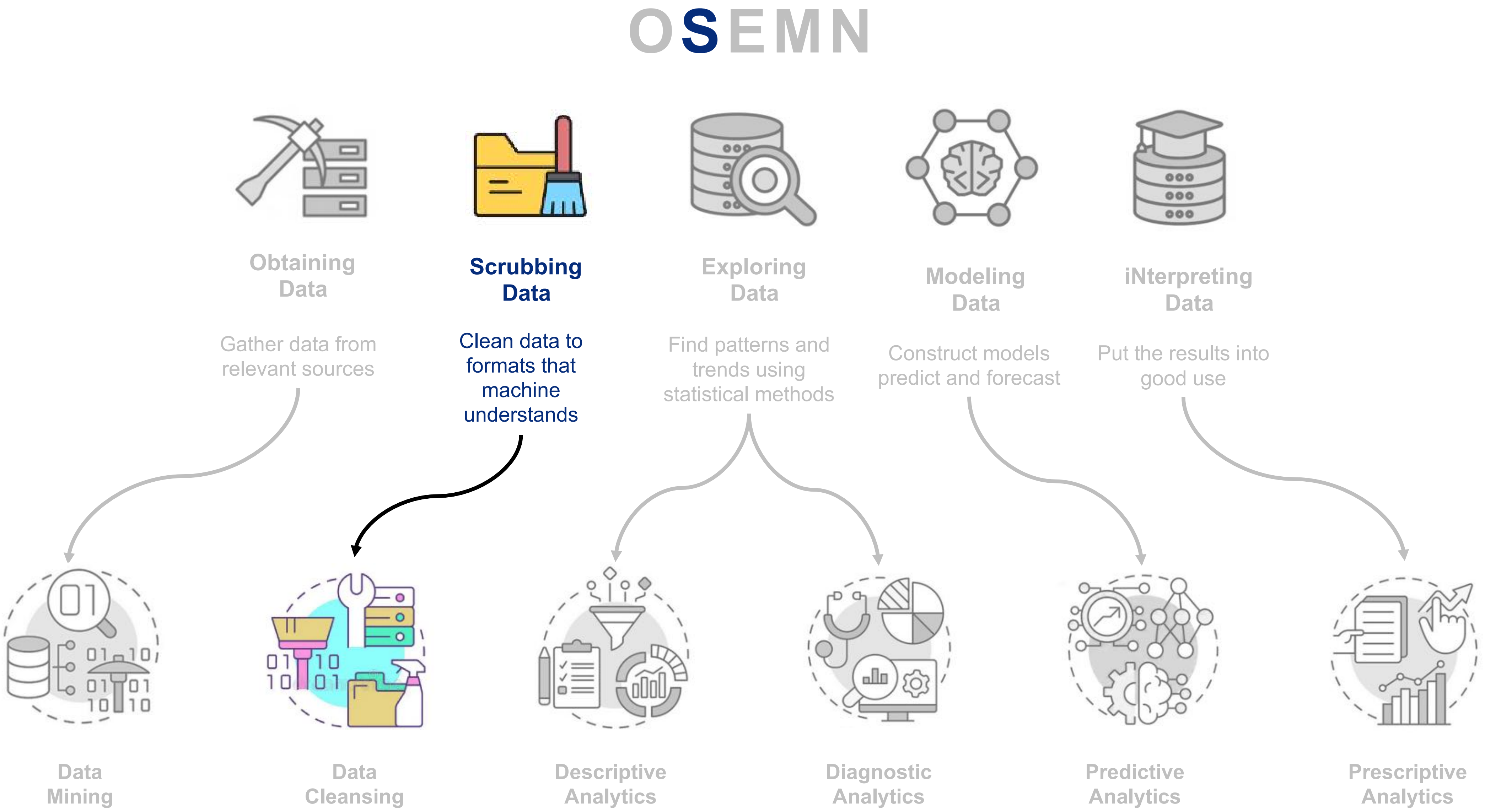
Asset	License	Source Link
Employee Attrition Data	Open Database License (ODbL) Database Content license (DbCL)	Kaggle

- **Factors & Attributes in the Dataset:**

Career Progression	Compensation & Benefits	Demographic Factors	Job Satisfaction	Work Environment
JobLevel	*DailyRate*	*Age*	*JobInvolvement*	*BusinessTravel*
PerformanceRating	*HourlyRate*	*Education*	*JobRole*	*Department*
TotalWorkingYears	*MonthlyIncome*	*EducationField*	*JobSatisfaction*	*DistanceFromHome*
TrainingTimesLastYear	*PercentSalaryHike*	*Gender*	*RelationshipSatisfaction*	*EmployeeCount*
YearsAtCompany	*StockOptionLevel*	*MaritalStatus*		*EnvironmentSatisfaction*
YearsInCurrentRole		*NumCompaniesWorked*		*OverTime*

- **Total Number of Records:** *1470*
- **Classification Target:** ***`Attrition`*** *(Yes or No)*

Scrubbing Data



Scrubbing Data

Handling Single & Unique Data:

- </> : `df_clean.nunique()`
- Single data: *`EmployeeCount, Over18, StandardHours`* are columns that single/same value across all records.
- Unique data: *`EmployeeNumber`* is a column whose entire row contains a unique value
- And hence these columns will not add any value to our analysis and ***"can be dropped"***

Handling Missing Data:

- </> : `df_clean.isnull().sum()`
- There is "***no empty (NULL) data***" in any column in the data frame

Handling Duplicated Data:

- </> : `df_clean.duplicated().sum()`
- There are "***no duplicates***" in any of the column

Scrubbing Data (continued)

Handling Outliers:

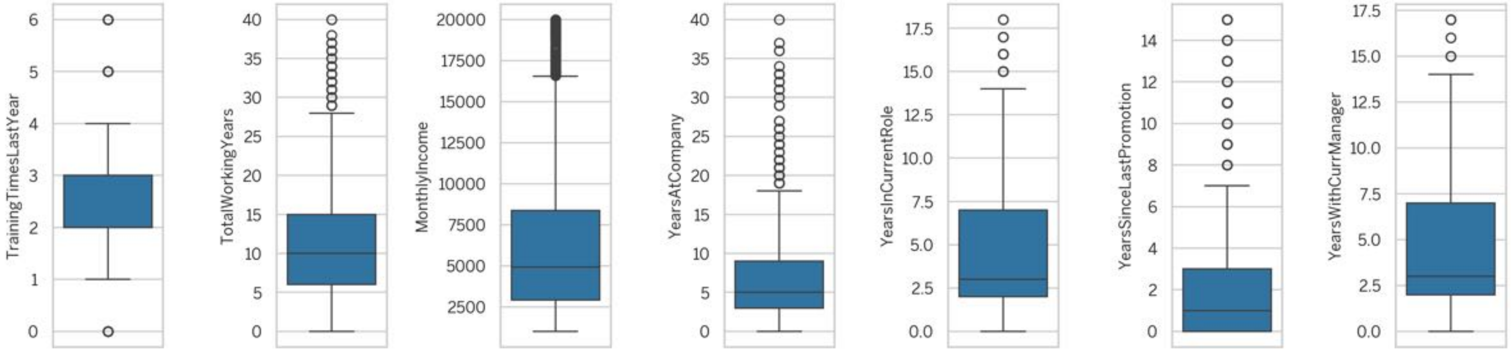
- To identify the outliers in the data, the box plots are used.

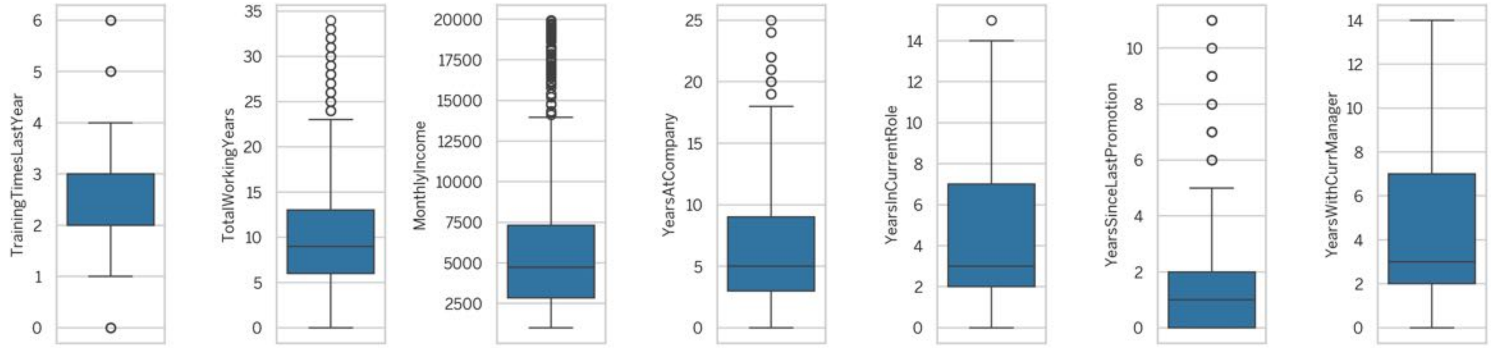
- Based on the **boxplots**, we can see that there are outliers in the following columns:
    - *TotalWorkingYears*
    - *TrainingTimesLastYear*
    - *YearsAtCompany*
    - *YearsInCurrentRole*
    - *YearsSinceLastPromotion*
    - *YearsWithCurrManager*
    - *MonthlyIncome*

- The outliers in the data are removed using **Z-Score method**:
    - Calculated the absolute z-score
    - Kept <3 absolute z-scores
    - Filtered out records whose Z-Scores are below 3

- Number of rows after removing outliers: **1387**

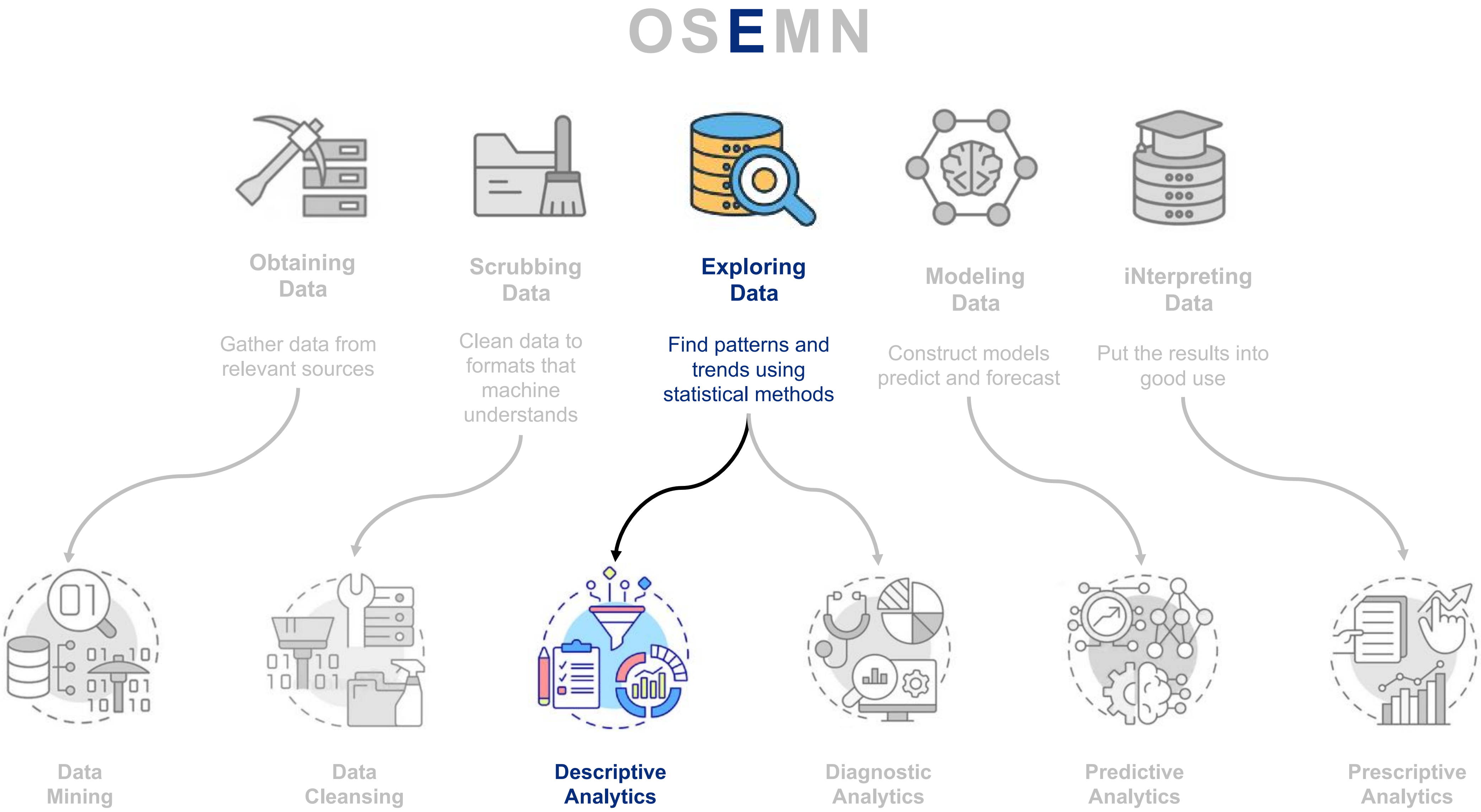
Before Removing Outliers ↓





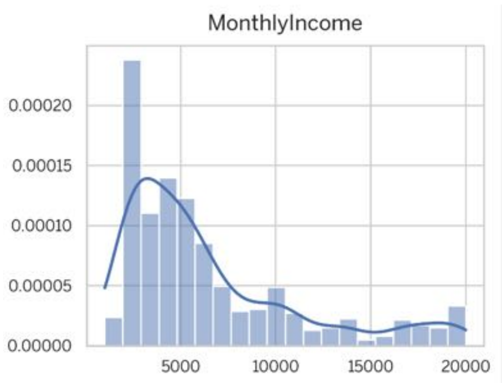
After Removing Outliers ↑

Exploring Data (Descriptive Analytics)



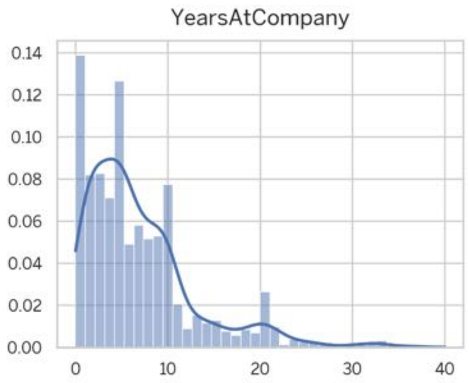
Exploring Data (Descriptive Analytics)

- Using **Quantitative** Feature Distribution ***(Histogram)***

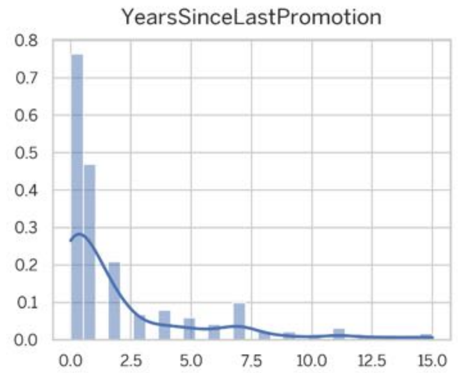


- The histograms shows large number of employees are having *low monthly income* (in the range of 2000-2500)

- The histograms reveals most of the employees have received *hike less than 15%*



- The histogram shows that a large number of employees are *new hires,* signifying the company has seen a *recent increase in employee turnover*



- A large number of employees are *recently promoted*

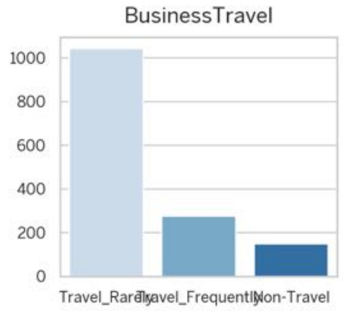
- Apart from those columns, *the distribution looks normal*

</> Link to the code and charts generated for quantitative other attributes.

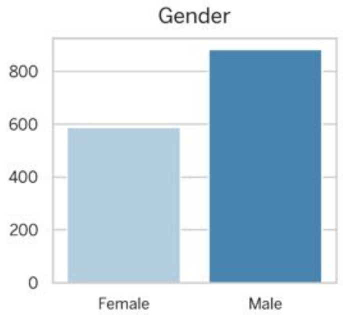
Exploring Data (Descriptive Analytics)

- Using **Qualitative** Feature Distribution ***(Bar Chart)***

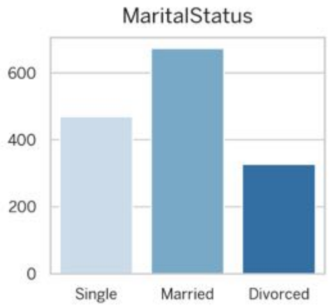
Research & Development is the department with the largest number of employees, around 900 employees



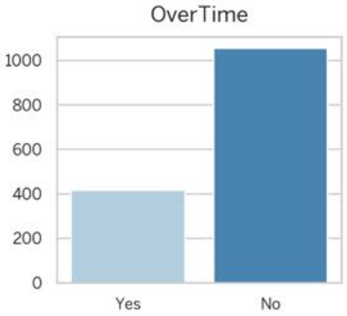
Many employees have a high frequency of business travel, around 1000+ employees



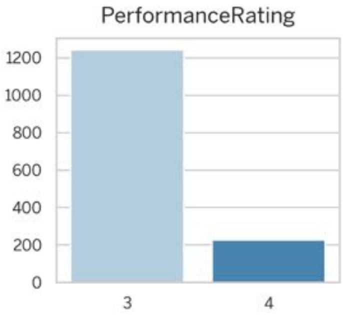
The company is dominated by 800 male employees.



The company is dominated by married employees



Relatively small number of employees work overtime

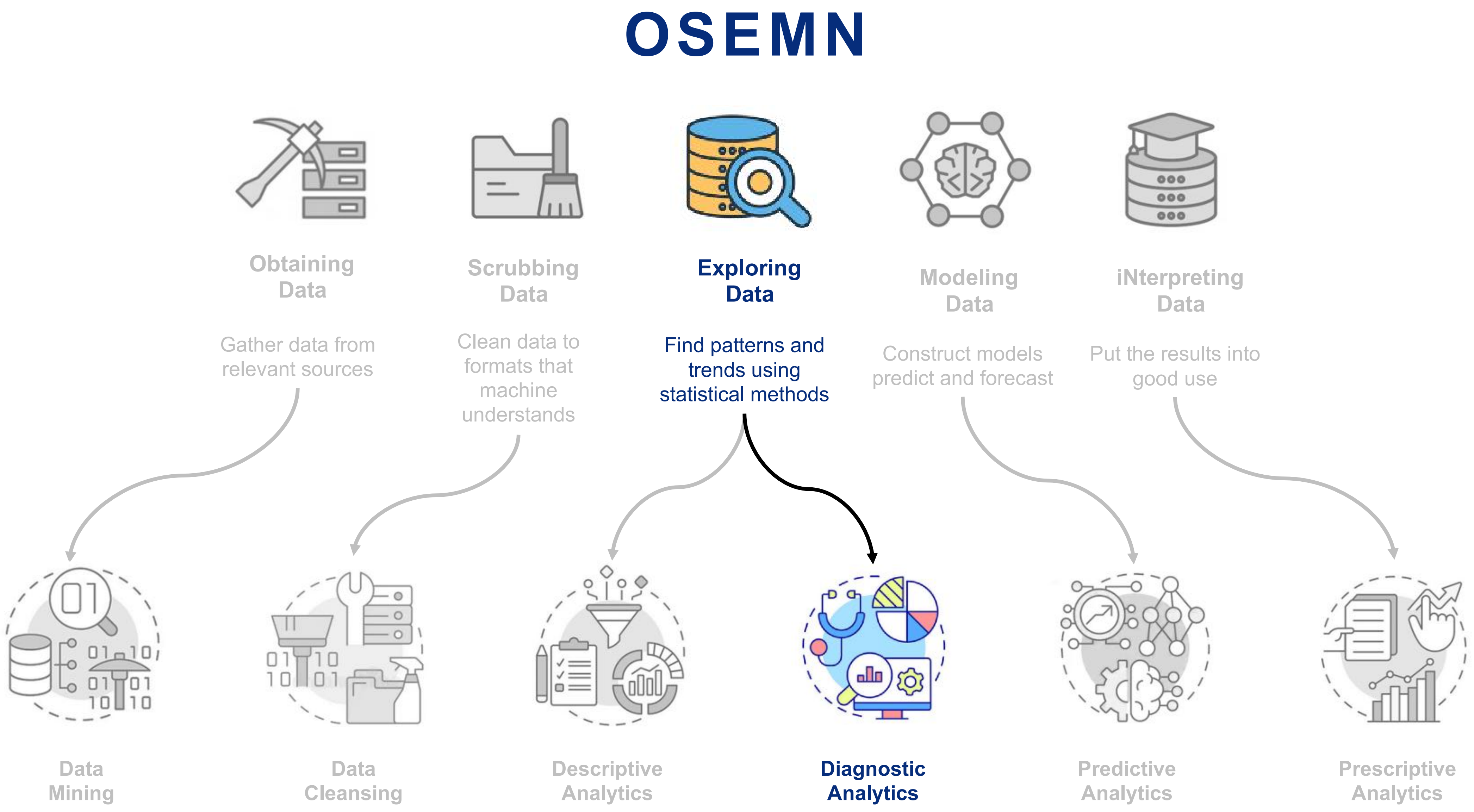


Only a small number of employees show extraordinary performance (4: *outstanding*)

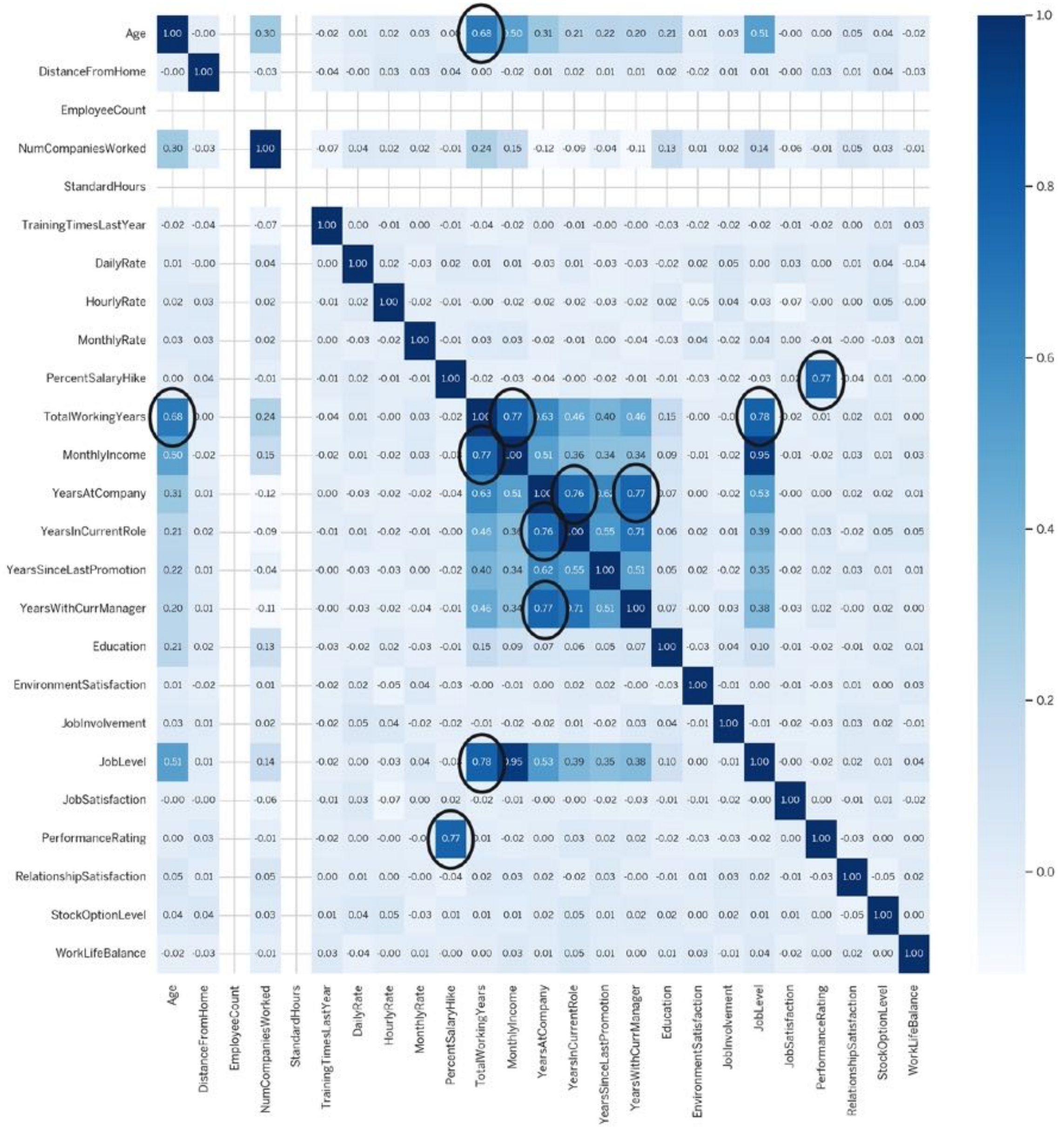
And there are no employees who have low performance (1: *low*)

[</> Link to the code and charts generated for qualitative other attributes.]()

Exploring Data (Diagnostic Analytics)



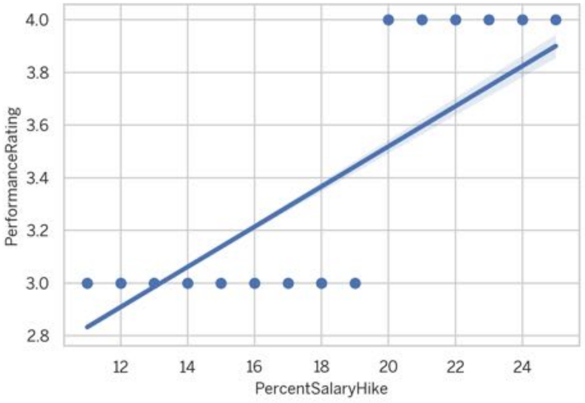
Exploring Data (Diagnostic Analytics)



Numerical & Categorical Ordinal Correlation (Heatmap)

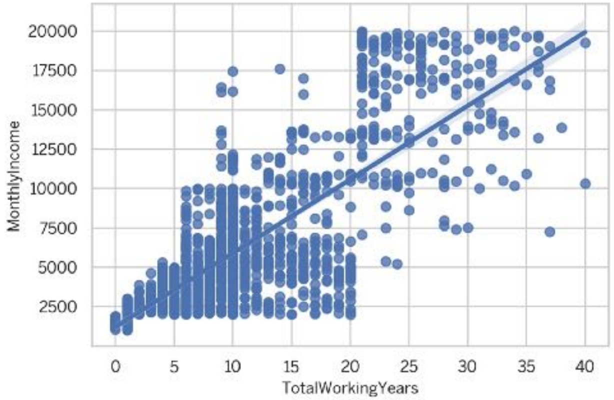
- `PercentSalaryHike` and `PerformanceRating` have a strong positive relationship
- `TotalWorkingYears` has a strong positive relationship with `Age`, `MonthlyIncome`, and `JobLevel`
- `YearsAtCompany` has a strong positive relationship with `YearsInCurrentRole` and `YearsWithCurrManager`

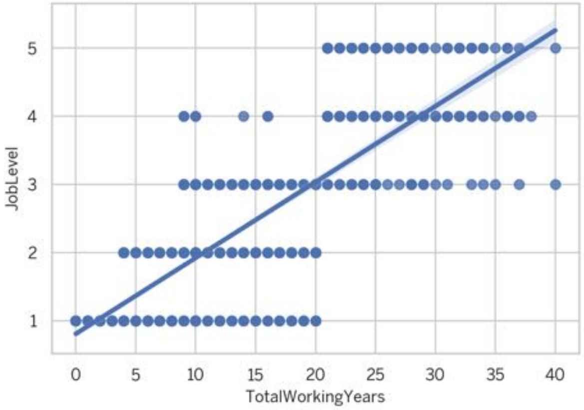
Multivariate Analysis (Correlation)



Salary has a great influence on the employee's performance

The longer employees work in the company, the more pay they will get





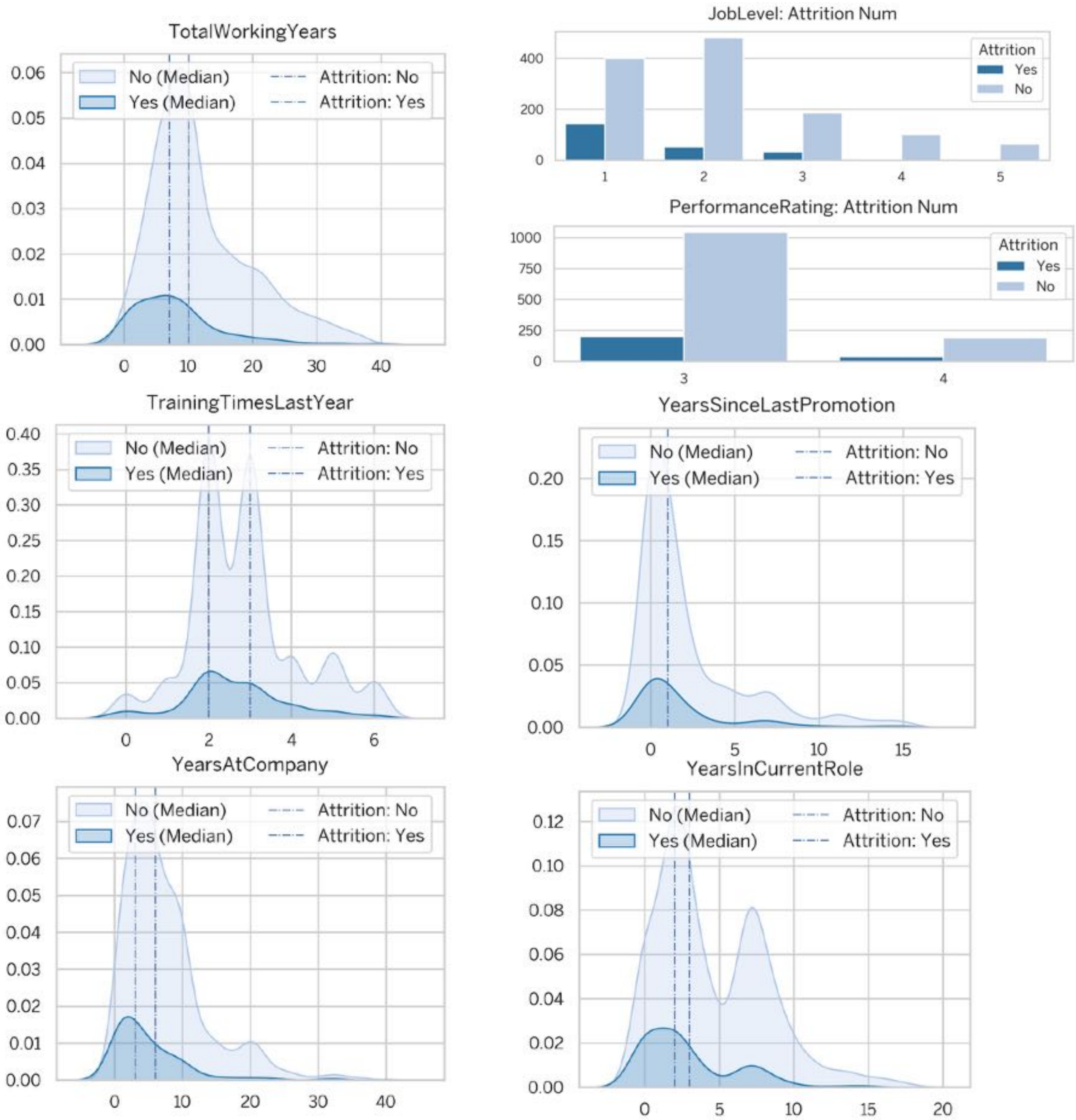
Most employees who are in the higher position have worked for the company for a long time

Exploring Data (Diagnostic Analytics)

- Distribution of attributes as Qualitative and Quantitative is done so that;
  - Graphical analysis for Qualitative data can be done using KDE Plots
  - And graphical analysis for Quantitative data can be done using Bar Chart

	Career Progression	Compensation & Benefits	Demographic Factors	Job Satisfaction	Work Environment
Analysis of Qualitative (Categorical) Data is done using KDE Plot	*JobLevel*	*StockOptionLevel*	*Education*	*JobInvolvement*	*BusinessTravel*
	PerformanceRating		*EducationField*	*JobRole*	*Department*
			Gender	*JobSatisfaction*	*EnvironmentSatisfaction*
			MaritalStatus	*RelationshipSatisfaction*	*OverTime*
Analysis of Quantitative (Numerical) Data is done using Bar Chart	*TotalWorkingYears*	*DailyRate*	*Age*		*DistanceFromHome*
	TrainingTimesLastYear	*HourlyRate*	*NumCompaniesWorked*		*EmployeeCount*
	YearsAtCompany	*MonthlyIncome*			
	YearsInCurrentRole	*PercentSalaryHike*			

Exploring Data (Diagnostic Analytics)



How *Career Progression* attributes impact the attrition?

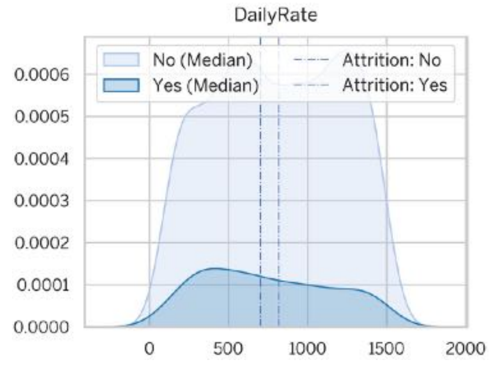
Attributes:

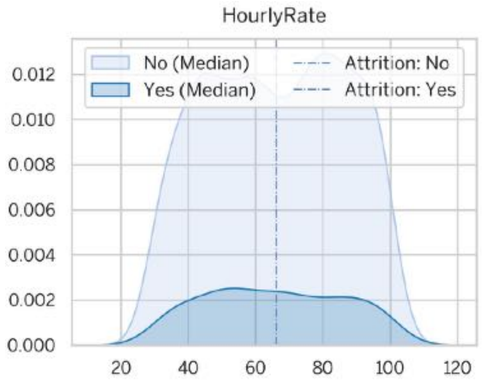
`JobLevel`, `PerformanceRating`, `TotalWorkingYears`, `TrainingTimesLastYear`, `YearsAtCompany`, `YearsInCurrentRole`, `YearsSinceLastPromotion`

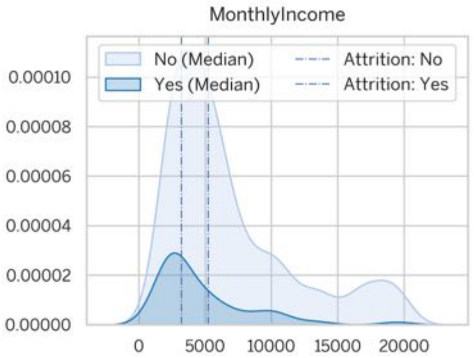
Observations:

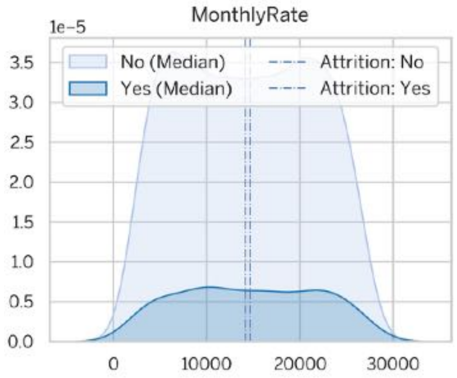
We see higher turnover among employees with fewer years of experience (0-5 years) compared to those who have been with the company for over a decade.

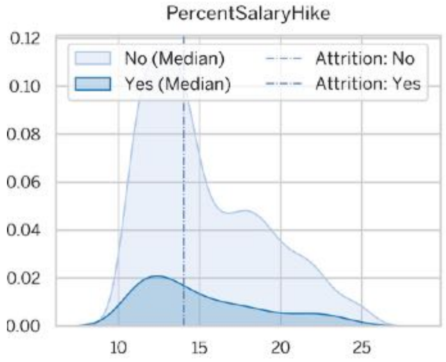
Exploring Data (Diagnostic Analytics)

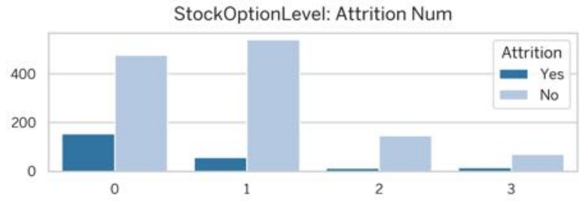












How *Compensation and Benefits* attributes impact the attrition?

Attributes:

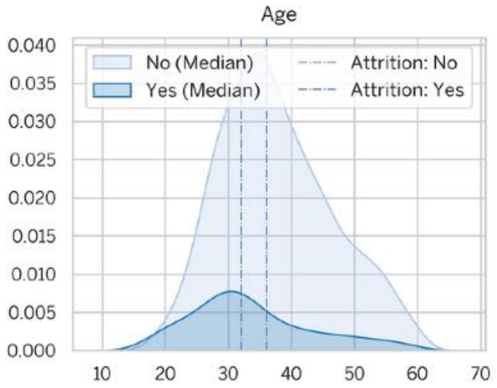
DailyRate, *HourlyRate*, *MonthlyIncome*, *MonthlyRate*, *PercentSalaryHike*, *StockOptionLevel*

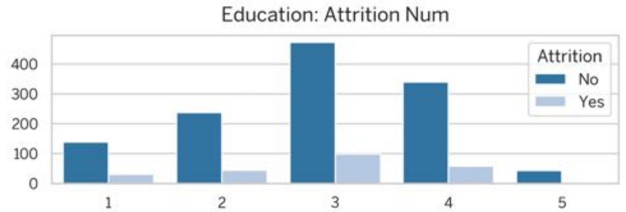
Observations:

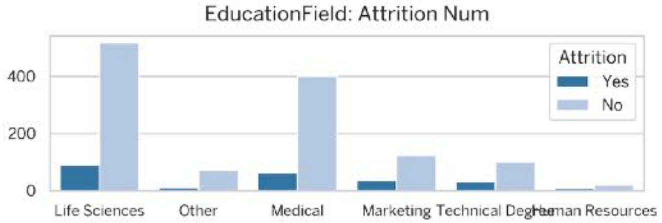
While salary may not be the primary driver of departures across the board, the data shows a higher turnover rate among employees earning lower monthly salaries (USD 0-5,000).

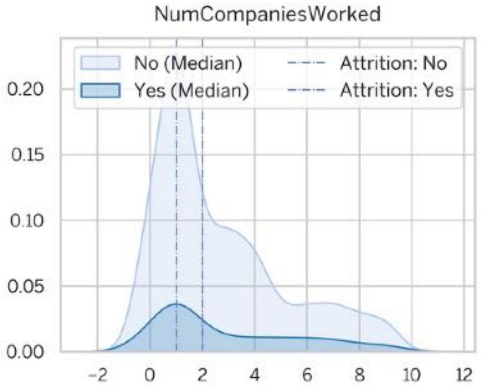
Further investigation into reasons for leaving specifically within this income bracket is recommended.

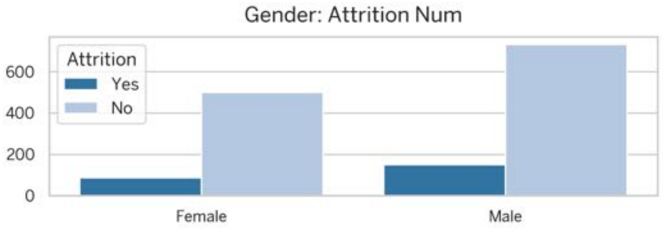
Exploring Data (Diagnostic Analytics)











How *Demographic Factors* attributes impact the attrition?

Attributes:

`Age`, `Education`, `EducationField`, `Gender`, `MaritalStatus`, `NumCompaniesWorked`

Observations:

Our analysis indicates that employees who leave tend to be younger (25-35 years old) compared to those who stay.

Additionally, we see a higher turnover rate among single employees, with a greater number of men leaving the company overall (approximately 150 people).

Exploring Data (Diagnostic Analytics)

How *[Job Satisfaction]* attributes impact the attrition?

Attributes:

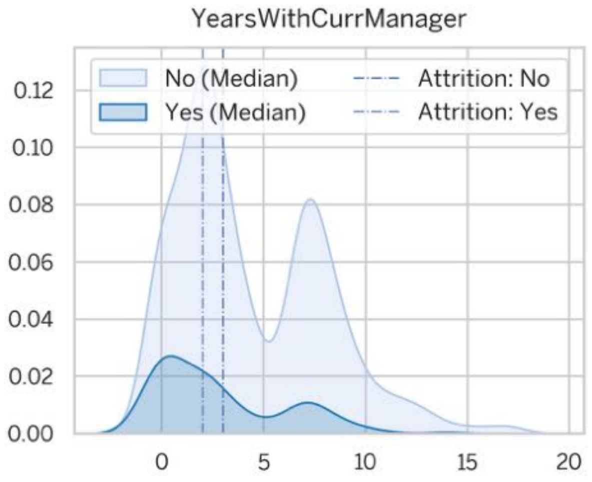
`JobInvolvement`, `JobRole`, `JobSatisfaction`, `RelationshipSatisfaction`

Observations:

We've identified Sales Representatives, followed by Laboratory Technicians and Human Resources personnel, as having the highest turnover rates within the company.

Conversely, employees with higher job involvement tend to stay with the company longer.

Exploring Data (Diagnostic Analytics)



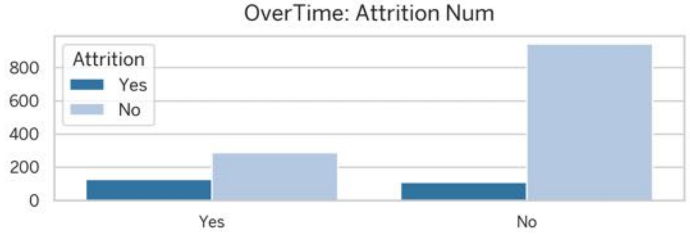
How *Managerial Influence* attributes impact the attrition?

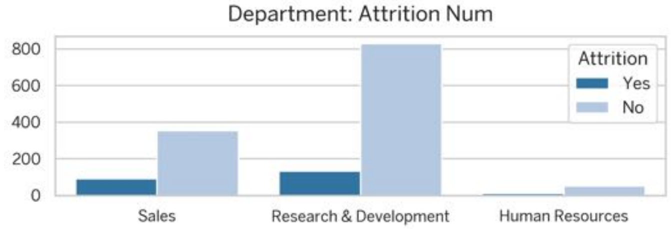
Attribute: *`YearsWithCurrManager`*

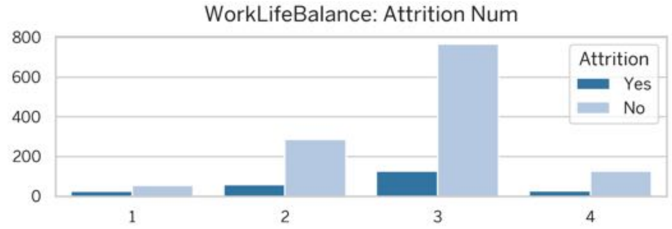
Observations:

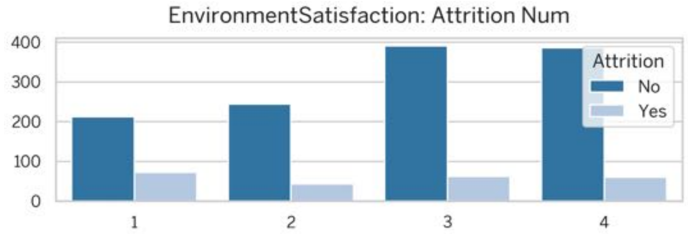
We could not observe any specific pattern with respect to this attribute.

Exploring Data (Diagnostic Analytics)









How *Work Environment* attributes impact the attrition?

Attributes: `BusinessTravel`, `Department`, `EnvironmentSatisfaction`, `OverTime`, `WorkLifeBalance`

Observations:

Travel: Frequent travel correlates with higher attrition

Sales (20% attrition) has the highest rate, but R&D incurs the most cost due to size

Overtime: Employees working overtime are 3x more likely to leave (30% attrition)

Work Environment: Lower Environment Satisfaction leads to higher attrition

Consolidated Findings (In Simple Words)

- **Age:** Younger employees (25-35 years old) are more likely to leave compared to those with longer tenures.

- **Travel:** Rare Business Travel correlates with attrition. While the Sales department has the highest Attrition Rate (21%), the R&D Department incurs the most cost due to its larger size (Attrition of 128 employees).

- **Compensation:** While salary distribution doesn't significantly differ between departing and remaining employees, lower monthly income (0 - 5,000) shows a higher turnover rate.

- **Overtime:** Employees working overtime are three times more likely to leave (30% attrition rate).

- **Job Satisfaction:** Lower satisfaction with the work environment (EnvironmentSatisfaction) is linked to higher turnover.

- **Department:** Sales representatives, followed by Laboratory Technicians and HR Personnel, has the highest turnover rates. Interestingly, HR has the lowest number of employees leaving, despite a high attrition rate (25%).

- **Demographics:** A higher number of men leave the company compared to women (approximately 150). Marital status also plays a role, with single employees showing a higher attrition rate (25%).

Feature Selection

Create new features for Machine Learning Analysis

- Creating Feature Groups from Age (`Group Age`)
- Finding Median Monthly Income by Job Level (`MedIncome`)
- Creating a Feature for Below Median Income (`BelowMedIncome`)
- Creating Interaction Features (`GroupAge_Overtime` , `JobLevel_Overtime` , `JobLevel_BelowMedIncome_Overtime`)

Feature Encoding:

- *Traditional Labelling :* for columns containing binary values (e.g. Yes/No, Male/Female for Attrition, OverTime etc.)
- *One Hot Encoding :* for columns with more than two unique values(1,2,3,4,5), (e.g.: Education)

Feature Scaling:

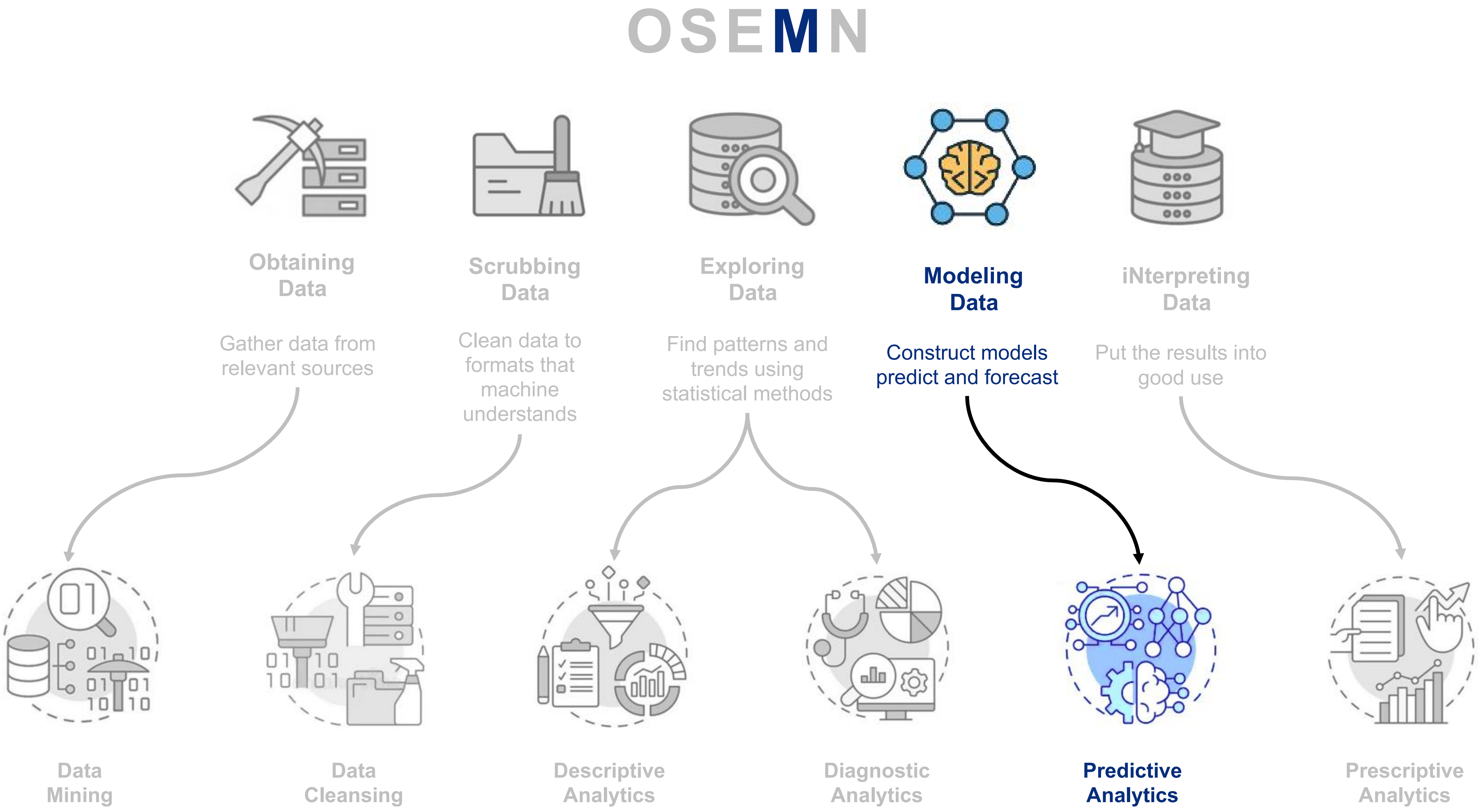
- Standardization *(StandardScaler + FitTransform) :* Age Column (normal distribution)
- Normalization *(MinMaxScalaer StandardScaler + FitTransform :* Numerical Columns except Age

```
***sklearn.preprocessing import MinMaxScaler, StandardScaler
```

Feature Selection:

- Perform feature importance analysis to identify the most influential factors for employee attrition.
- Select a subset of features based on importance and potentially other considerations

Modelling Data



Modelling Data (ML Model)

AS WE HAVE SEEN EARLIER…

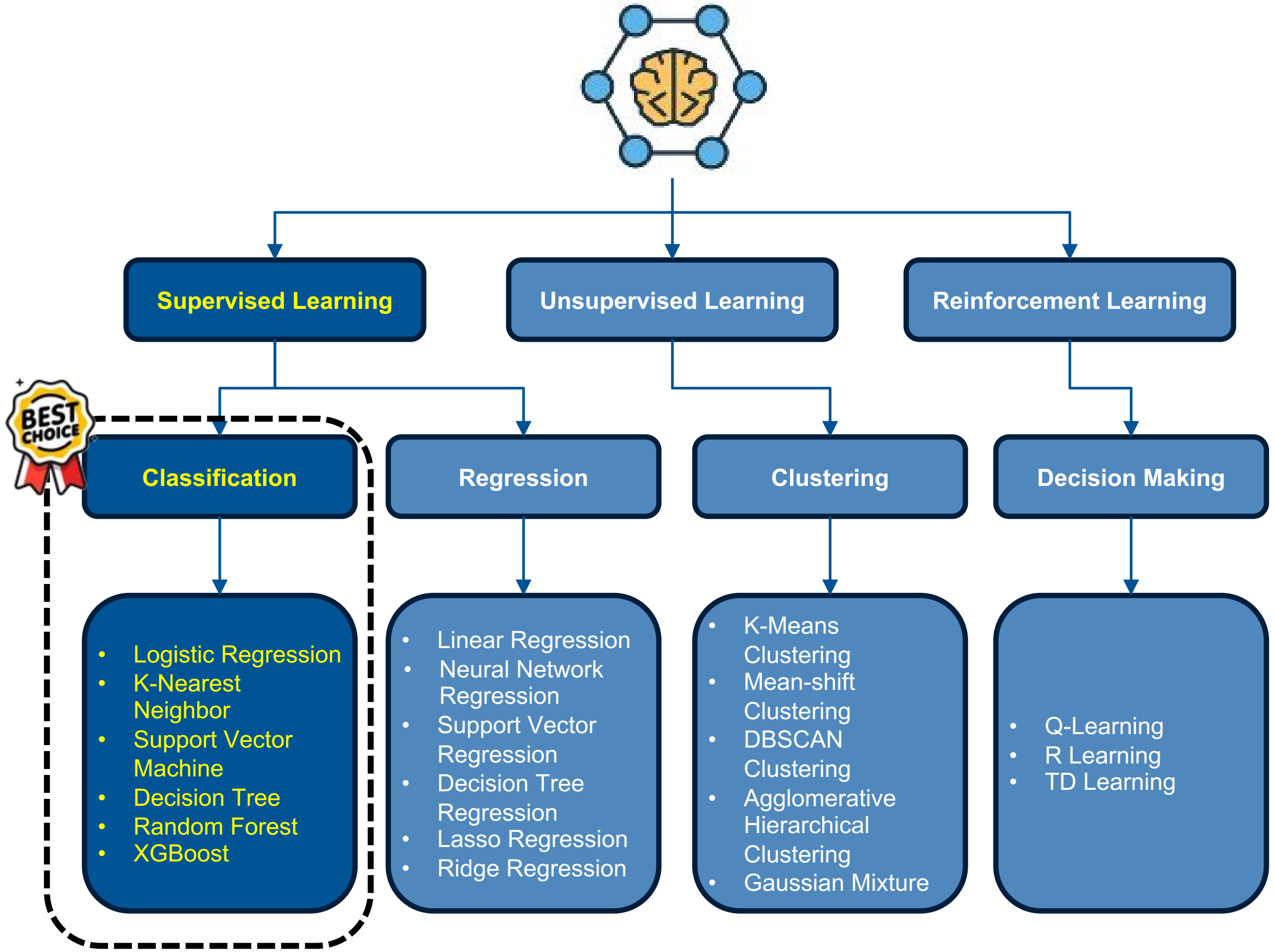
Key Factor To Decide Between Supervised and Unsupervised Learning:

- We have chosen Supervised Learning in this scenario as we have a labeled dataset and aim to make predictions.
- Supervised Learning is ideal for tasks like employee churn prediction, where our goal is to classify categories or predict continuous values based on past data.

Key factor to deciding between Classification and Regression lies in target variable:

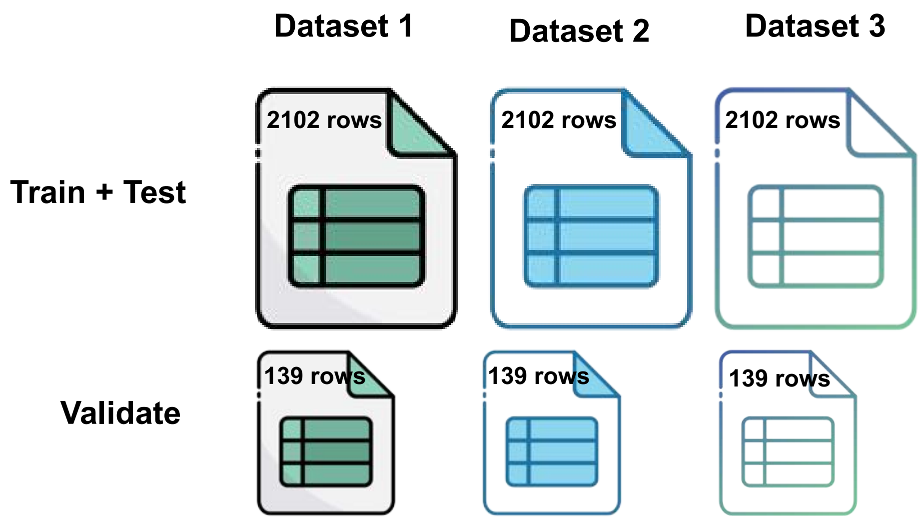
- Classification: Classification model is used when target variable is discrete and falls into distinct categories. These categories can be binary (like Attrition Yes/No) or have multiple classes (e.g., classifying handwritten digits into 0-9).

- **Regression:** Regression model is used when target variable is continuous. This means it can take on any numerical value within a range. Common examples include predicting house prices, weather forecasts (temperature), or customer lifetime value.



Data Splitting (Train + Test & Validate)

- The data is prepared for training by separating features (without Attrition column) and the target variable (Attrition).
- The data is split into training and testing sets using a 70:30 ratio.
- The training set is used to train the machine learning model, and the testing set is used to evaluate its performance on unseen data.
- It creates a validation set, which can be used for hyperparameter tuning (finding the best settings for the model).
- By splitting the data into these sets, we ensured that the model is not simply memorizing the training data and can generalize well to unseen data.



- The number of rows and columns of df_clean1 : (1248, 84)
- The number of rows and columns of df_clean2 : (1248, 20)
- The number of rows and columns of df_clean3 : (1248, 28)
- The number of rows and columns of df_valid1 : (139, 84)
- The number of rows and columns of df_valid2 : (139, 20)
- The number of rows and columns of df_valid3 : (139, 28)

Classification Models Used

- K-Nearest Neighbor
- Support Vector Machine
- Decision Tree
- Random Forest
- Logistic Regression
- XGBoost

Classification Report Metrics Used

- **Accuracy** — This metric measures the proportion of correct predictions made by the model across the entire dataset. It is calculated as the ratio of true positives (TP) and true negatives (TN) to the total number of samples.

- **Precision** — Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated as the ratio of TP to the sum of TP and false positives (FP).

- **Recall** — Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances. It is calculated as the ratio of TP to the sum of TP and false negatives (FN).

- **F1 Score** — F1 Score is a metric that balances precision and recall. It is calculated as the harmonic mean of precision and recall. F1 Score is useful when seeking a balance between high precision and high recall, as it penalizes extreme negative values of either component.

- **ROC AUC (Receiver Operating Characteristic - Area Under Curve)** — This metric is used to measure the performance of a classification model at various threshold settings. The ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

Classification Report

	Model	Accuracy	Precision	Recall	F1	AUC
Dataset 1	**Logistic Regression (1)**	0.848921	0.761905	0.5	0.603774	0.726636
	Support Vector Machine (1)	0.791367	0.548387	0.53125	0.539683	0.700204
	Random Forest (1)	0.769784	0.5	0.46875	0.483871	0.664282
	K-Nearest Neighbor (1)	0.676259	0.341463	0.4375	0.383562	0.592582
	Decision Tree (1)	0.705036	0.354839	0.34375	0.349206	0.578417
	XGBoost (1)	0.805755	0.619048	0.40625	0.490566	0.665742
Dataset 2	**Logistic Regression (2)**	0.676259	0.385965	0.6875	0.494382	0.680199
	Support Vector Machine (2)	0.654676	0.362069	0.65625	0.466667	0.655228
	Random Forest (2)	0.741007	0.428571	0.375	0.4	0.612734
	K-Nearest Neighbor (2)	0.633094	0.306122	0.46875	0.37037	0.575496
	Decision Tree (2)	0.719424	0.36	0.28125	0.315789	0.565859
	XGBoost (2)	0.748201	0.4	0.1875	0.255319	0.551694
Dataset 3	**Logistic Regression (3)**	0.726619	0.434783	0.625	0.512821	0.691005
	Support Vector Machine (3)	0.683453	0.384615	0.625	0.47619	0.662967
	Random Forest (3)	0.769784	0.5	0.40625	0.448276	0.642377
	K-Nearest Neighbor (3)	0.654676	0.34	0.53125	0.414634	0.611419
	Decision Tree (3)	0.697842	0.272727	0.1875	0.222222	0.518984
	XGBoost (3)	0.76259	0.454545	0.15625	0.232558	0.550088

From the comparison table above, it is found that;

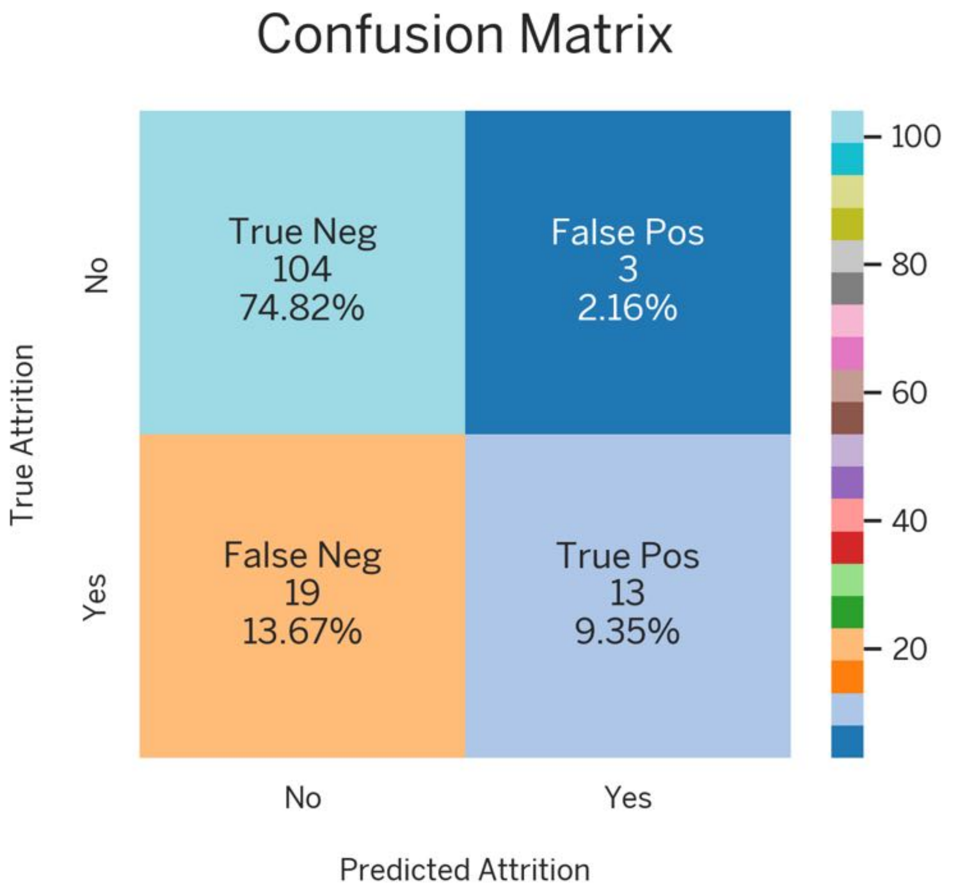
Logistic Regression (1) is the best algorithm with the highest

- Accuracy and
- F1 score

Final Training Model

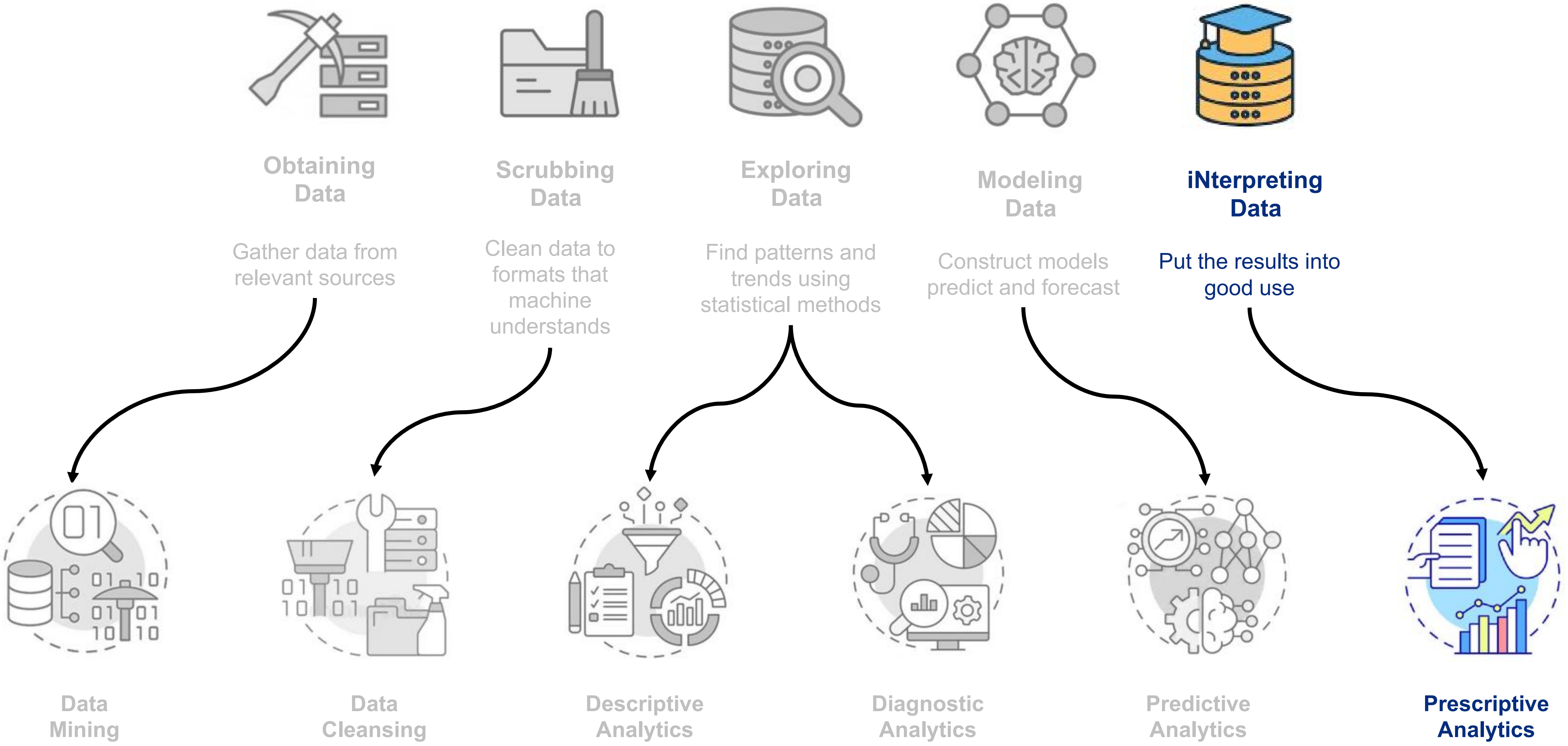
The confusion matrix summarizes how many data points were correctly or incorrectly classified into different categories.

Carrying out model training using the Logistic Regression algorithm with all features scenario.

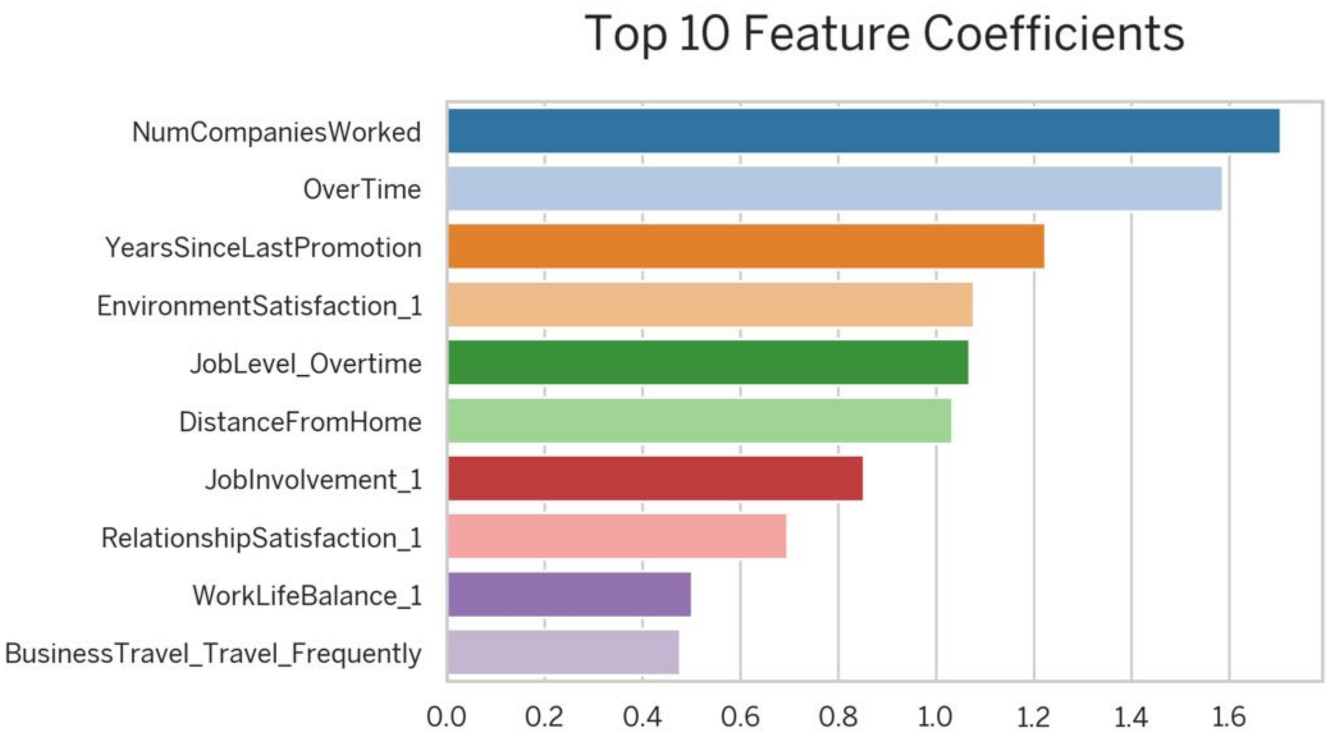


iNterpreting Data

OSEMN



Coefficients Feature



Analyze the coefficients of a trained Logistic Regression model to understand the relative importance of features in predicting the target variable.

Based on the top list of Coefficient Feature, we can predict that

- '`OverTime`' and
- '`YearsSinceLastPromotion`'

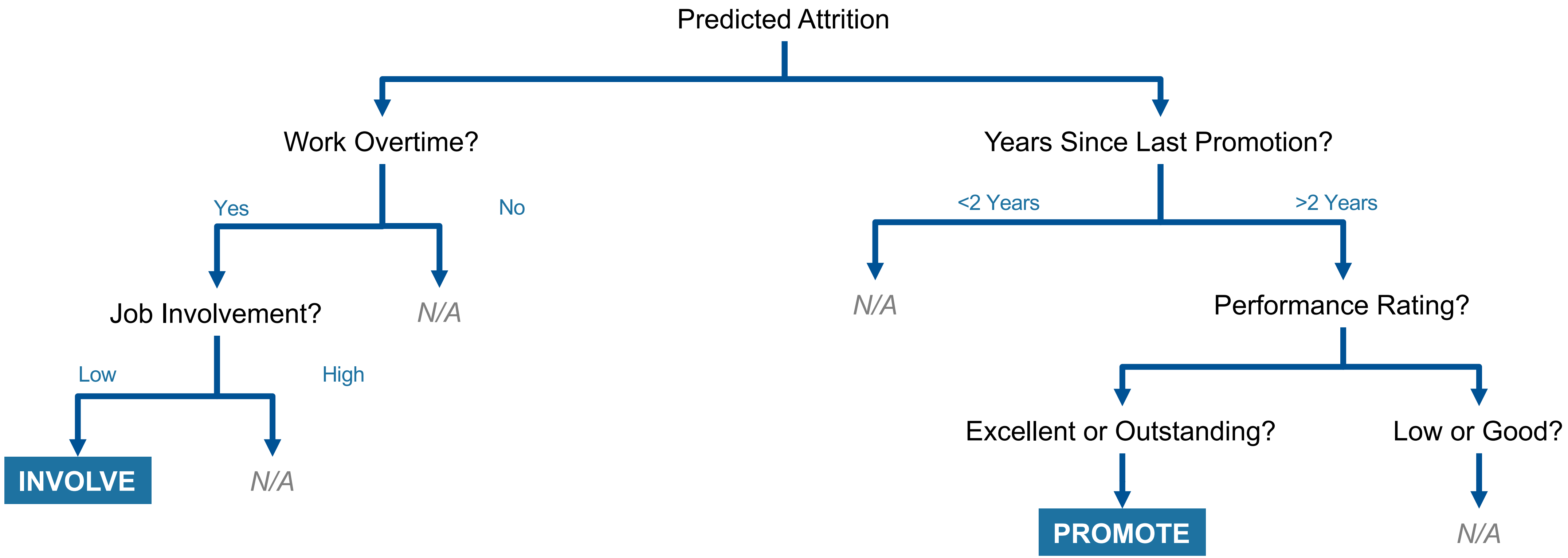
are the key factors affecting attrition.

'`NumCompaniesWorked`' feature cannot be controlled by ABC Technologies' HR Department to reduce the Attrition, hence not used.

Data Interpretation & Recommendations

- With the help of ML models, the probability of potential employees leaving the company (before the treatment) was 142 out of 1470 people (i.e. 9.66%).
- Following up on the prediction results, the HR department recommend ABC Technologies leadership to treat employees based on features with high coefficient values i.e. **Overtime** and **Years Since Last Promotion**.
- What To Look For?
  - Are there employees with a performance rating above average (excellent and outstanding) which means they deserve to be promoted, but have not received a promotion for years?
  - Are there employees who work overtime, despite they have low job involvement?

Recommendation Flow for Simulation



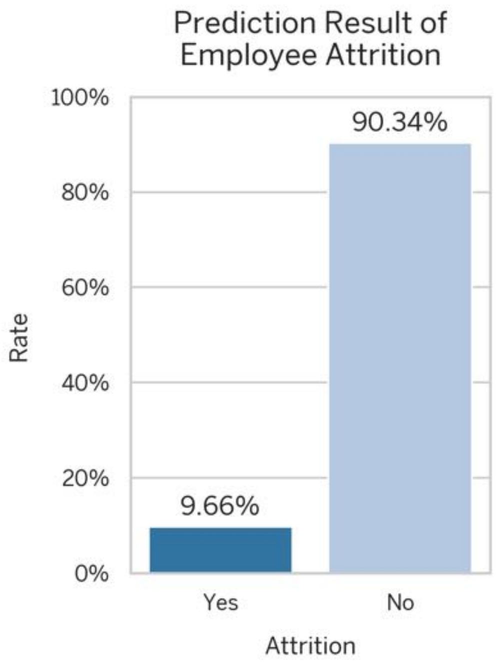
In this case we try to increase 1 Job Involvement level by filtering as follows:
- Predicted attrition = Yes
- Job involvement <= 3
- Overtime = Yes

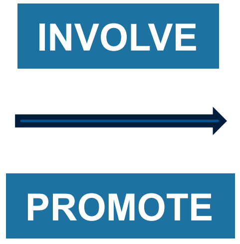
Data changes made:
- Increase a level of **Job Involvement**, for example: from 1 to 2, 2 to 3, etc.

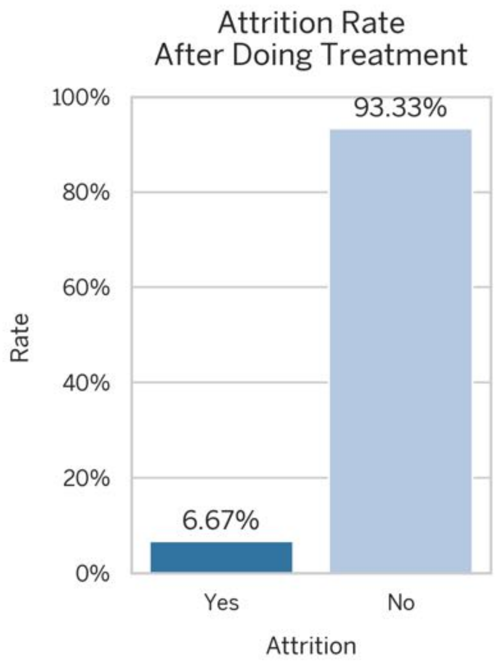
In this case, we try to give promotions to several employees who are worthy of promotion by filtering as follows:
- Predicted attrition = Yes
- Performance rating >= 3 (excellent and outstanding)
- Years since last promotion >= 2 (last promoted above equals 2 years)

Data changes made:
- Increase **Job Level** by 1, for example: from 1 to 2, 2 to 3, etc.
- Changed the **Years Since Last Promotion** value to 0

Results Pre & Post Simulation







The attrition rate reduced from 9.66% to 6.67% **(reduced 2.99%)**, leaving only 98 employees with the potential to leave the company.

Recommendation To Leadership

The HR Analytics Department, would recommend the leadership team to;

- **PROMOTE!**
  - Pay attention to the performance of your employees who are entitled to a promotion
- **INVOLVE!**
  - Improve The Job Involvement of Employees:
    - Coaching or mentoring, and give positive feedback
    - Employee involvement programs
    - Open-communication and suggestion boxes
  - Reduce Employee's Overtime:
    - Cross-train your employees
    - Try flexible work schedules

Bibliography

- Link to the Google Colab notebook:

  https://go.vivekkulthe.com/iimk-dsai

- This presentation has been designed using images from

  https://www.flaticon.com

  https://www.dreamstime.com

- Dataset used for this project is taken from

  https://www.kaggle.com

INDIAN INSTITUTE OF MANAGEMENT KOZHIKODE

योगः कर्मसु कौशलम्

भारतीय प्रबंध संस्थान कोषिक्कोड

Indian Institute of Management Kozhikode

भारतीय विचारधारा का वैश्वीकरण *Globalizing Indian Thought*