1. Introduction

Employee attrition, the departure of personnel from an organization, stands as a persistent challenge impacting operational continuity, team dynamics, and overall productivity. To address this multifaceted issue, this study harnesses the potential of supervised learning techniques—specifically, logistic regression, random forest, Support Vector Machines (SVM), and decision trees—to forecast and comprehend instances of employee turnover. The essence lies in leveraging historical data encompassing a diverse spectrum of employee attributes, ranging from basic demographics and performance metrics to nuanced factors like tenure, job role satisfaction, and engagement levels. By delving into these supervised learning methodologies, the research aims not only to construct predictive models capable of accurate attrition prediction but also to unravel the intricate interplay of influential factors precipitating workforce turnover. The predictive models are forged upon historical datasets that serve as a rich repository of insights into employee behavior, performance trends, and intrinsic job-related elements. The objective is not merely predictive accuracy but also the extraction of actionable insights crucial for proactive human resource management strategies. The study involves a meticulous evaluation of each model's performance, employing robust metrics such as accuracy, precision, recall, and F1-score to gauge their efficacy in predicting employee attrition. This evaluation phase stands as a cornerstone in discerning the strengths and limitations of each model, guiding the selection of the most reliable and precise predictive approach. Beyond the realm of prediction, the research endeavors to unravel the underlying factors contributing to attrition, thus empowering organizations with insights imperative for formulating proactive retention strategies. Such insights, derived from a detailed analysis of predictive models, serve as a compass for human resource decision-makers, guiding them in devising interventions and policies aimed at curbing attrition rates. Ultimately, this research aims not only to construct accurate predictive models but also to provide a comprehensive understanding of the dynamic landscape influencing employee turnover. By bridging the gap between data-driven insights and actionable strategies, it aspires to arm organizations with the tools necessary to foster a more stable, engaged, and retention-focused work environment.

2. Literature Review

The exploration of employee attrition utilizing supervised learning techniques has garnered substantial attention in contemporary organizational research. Various studies have investigated predictive models employing logistic regression, random forest, Support Vector Machines (SVM), and decision trees to anticipate and comprehend workforce turnover. Research by Johnson and Tandon (2018) delved into logistic regression models, highlighting their effectiveness in identifying key predictors of attrition, such as job satisfaction, performance metrics, and tenure. Similarly, Chen et al. (2019) extended this exploration, employing random forest algorithms to predict attrition. Their findings underscored the model's capability in capturing nonlinear relationships among diverse employee attributes, offering enhanced predictive accuracy. Moreover, SVM algorithms have attracted attention in mitigating employee attrition risks.Thereby enabling precise identification of factors contributing to attrition, especially in heterogeneous workforce settings. Additionally, decision tree-based models have emerged as insightful tools in comprehending employee turnover. Research conducted by Li et al. (2017) revealed decision tree models' capacity to elucidate hierarchical relationships among predictors, aiding in identifying critical attributes influencing attrition among different employee cohorts. Collectively, these studies underscore the significance of supervised learning techniques in anticipating and understanding employee attrition. While logistic regression provides interpretability, random forest and SVM offer robustness in handling complex relationships within the data. Decision trees, on the other hand, unveil intricate attribute hierarchies. However, there remains a continuous pursuit to enhance model accuracy, interpretability, and generalizability across diverse organizational contexts. Future research directions may entail the fusion of these techniques or the integration of advanced methodologies to provide holistic insights for effective attrition management strategies.

3. Statement of Problem

The high rate of employee attrition can have significant implications for organizations, leading to increased recruitment and training costs, reduced productivity, and decreased employee morale. It is crucial for companies to understand the underlying factors contributing to attrition and take proactive measures to address them.

4. Scope of Study

The study focuses on predicting employee attrition using supervised learning techniques like logistic regression, random forest, SVM, and decision trees employing historical employee data encompassing demographics, performance metrics, tenure, and satisfaction. This involves data preprocessing, feature selection, model implementation, and evaluation using various metrics to compare model performance. The objective is to determine the most effective algorithm for predicting attrition of employees.

5. Objective of Study

The objective of this project is to perform an Exploratory Data Analysis (EDA) on the IBM HR Analytics Employee Attrition & Performance dataset using Python. The goal is to gain insights into the factors influencing employee attrition within the organization. By examining various features and conducting statistical analysis, we aim to identify patterns, trends, and potential areas for improvement that can help the company better understand and manage employee attrition.

6. Methodology

a. Data Sources

The IBM HR employee dataset was utilized for learning model building and model evaluation process. The comparison of state-of-the-art machine learning methods was applied to predict employee attrition in IBM and the dataset link is given below:

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

b. Data Pre-processing

The three-stage system based on preprocessing, processing, and post-processing techniques was proposed to predict employee attrition. The IBM HR employee dataset was utilized for framework training and testing. The max-out feature selection technique was utilized for the dimension reduction stage. The techniques like Logistic regression, Random Forest, SVM and Decision tree was utilized for employee attrition prediction.

c. Description of the tools used

i. **Google Collab (Collaboratory):** Google Collab is a cloud-based platform provided by Google that allows users to write and execute Python code in a Jupyter Notebook environment.

ii. **Jupyter Notebook:** Jupyter Notebook is an open-source web application that enables interactive computing. It allows users to create and share documents containing live code, equations, visualizations, and explanatory text.

iii. **Power BI (Business Intelligence):** Power BI is a business analytics tool developed by Microsoft. It allows users to visualize and share insights from their data through interactive reports and dashboards.

iv. **Dash:** Dash is a Python framework built on top of Flask

v. **Flask:** Flask is a lightweight and flexible web application framework for Python. It is designed to make getting started with web development quick and easy, with the ability to scale up to complex applications.

vi. **Plotly:** Plotly is an open-source, interactive data visualization library for creating a wide variety of graphical plots and charts.

7. Analysis

a. Algorithms used

1. Logistic Regression
2. Random Forest
3. SVM
4. Decision Tree

b. ML models and techniques

i. Logistic regression (LR) estimates the parameters of a logistic (or logit) binomial regression model. Logistic regression is commonly used when the target variable is categorical and with problems of two class values such as pass/fail, win/lose or leave/stay as in the IBM attrition dataset

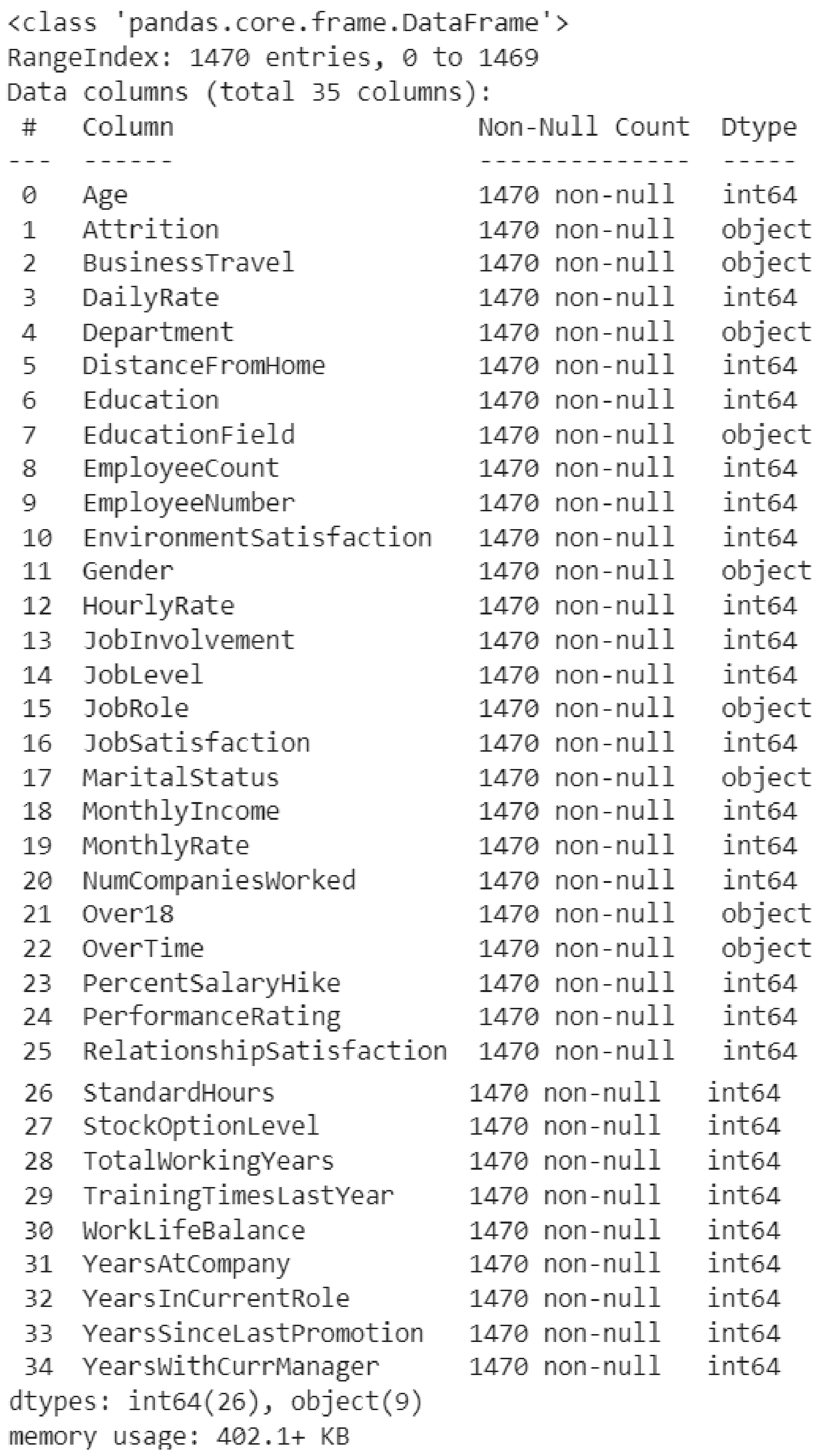
ii. Random forest (RF) is an ensemble learning method for classification and regression that combines many decision trees (weak learners) to form a stronger learner and get a more accurate and stable prediction. Those decision trees vote on how to classify a given instance of input data and outputting the class that is the mode of the classes in case of classification tasks or the mean of predictions in the case of regression tasks. Thus, random forest reduces the problem of overfitting. The more trees in the forest, the better the result will be produced. Random forest learning algorithm is flexible and widely used. It is able to produce good results even without hyper-parameter tuning.

iii. Support Vector Machine (SVM) is used for classification and regression tasks. Its primary objective in classification is to find a hyperplane in an N-dimensional space (where N is the number of features) that distinctly separates data points into different classes. The hyperplane chosen is the one that maximizes the margin, which is the distance between the hyperplane and the nearest data points (called support vectors) from each class. SVM can handle both linear and non-linear classification tasks by using different kernel functions to transform the input space into a higher-dimensional space, where a linear separation can be achieved.

iv. Decision trees (DT) are very powerful algorithms, capable of fitting complex datasets and have been applied to wide range of tasks such as medical diagnosis and credit risk of loan application. Decision tree learning approximates a target function which is represented as a tree of "if-then" rules to improve human readability. Thus, it breaks down a dataset into smaller and smaller subsets starting from the topmost node called "root", and then an associated decision tree is incrementally developed. The final decision tree has two nodes; decision nodes and leaf nodes, which can handle both categorical and numerical data

c. Exploratory data Analysis

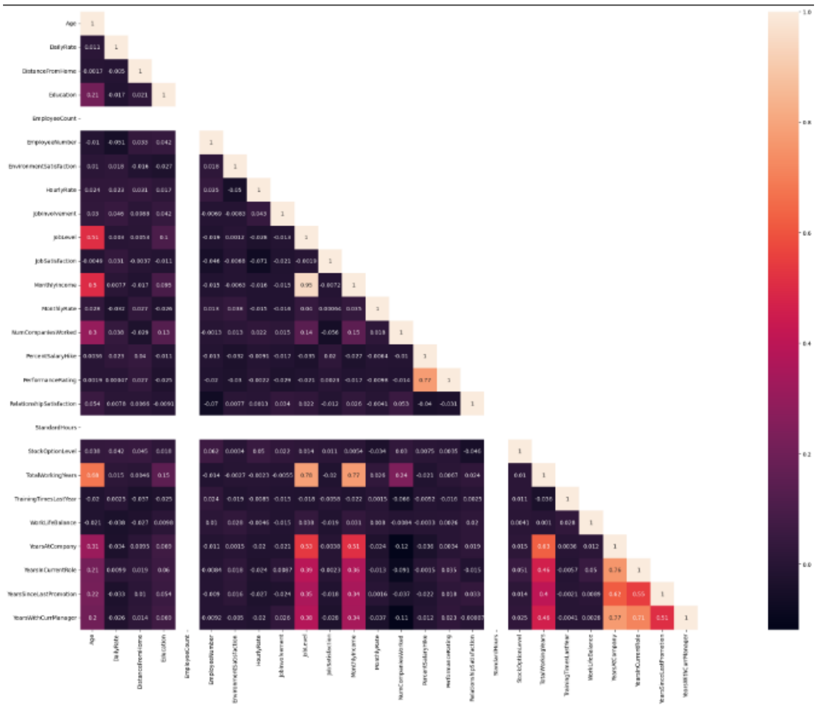
Dataset



```
<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
```

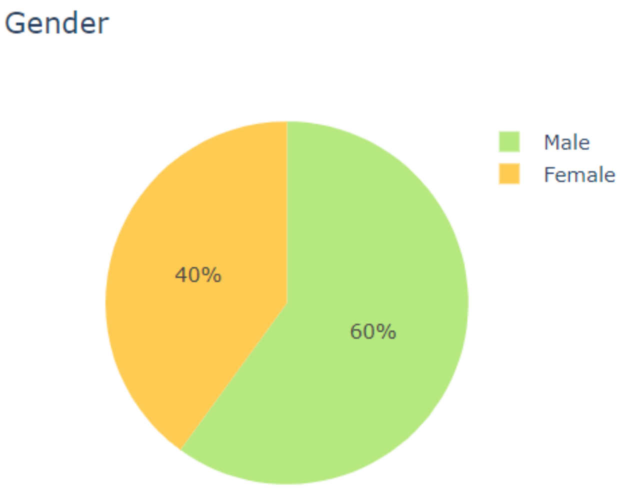
There is total 1470 Records or Rows in the dataset. There is total 35 Features or columns in the dataset. The primary key factor is attrition. There are 26 Numerical Attributes and 9 Categorical Attributes in the dataset. The 9 categorical columns consist of 1470 recordings and with the highest frequency of 1233 for attrition column and lowest frequency 606 for life sciences column. There are no missing values in each of the column. The average age of the employee is 36, the average distance from home is 9km, the average hourly rate is 65 hours, the average monthly income is 6500, the average percentage of hike is 15%, the average of total working years is 11 years, the average years at company is 7 years and the average years in current role is 4 years. The minimum age of the employee is 18, the minimum distance from home is 1km, the minimum hourly rate is 30 hours, the minimum monthly income is rs1000, the minimum percentage of hike is 11%, the minimum years at company is 0 years and the minimum years in current role is 0 years. The maximum age of the employee is 60, the maximum distance from home is 29km, the maximum hourly rate is 100 hours, the maximum monthly income is rs20000, the maximum percentage of hike is 25%, the maximum of total working years is 40 years, the maximum years at company is 40 years and the maximum years in current role is 18 years. 25%, 50%, 75% values are also noted. There are no any duplicate values in the dataset. The unique vales in the object datatype are checked.

Correlation heatmap



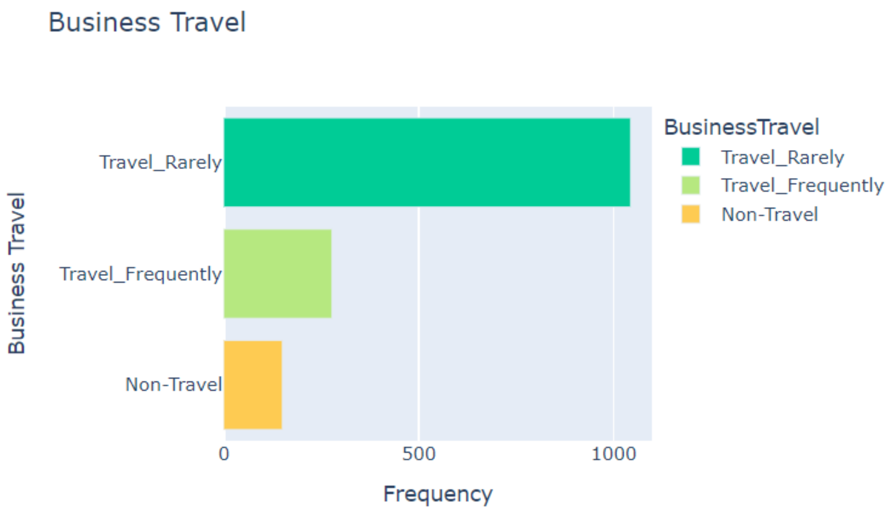
The correlation coefficient value, ranging from -1 to 1, indicates the strength and direction of the linear relationship between variables. From the graph we say that there is highly positive correlation between monthly income and job level. From this we know that monthly income and job level plays a major role in the attrition process. Next comes the total working hours and job level and monthly income. This says that job position and salary depending on total working hours also have a slight high correlation.

Distribution by Gender



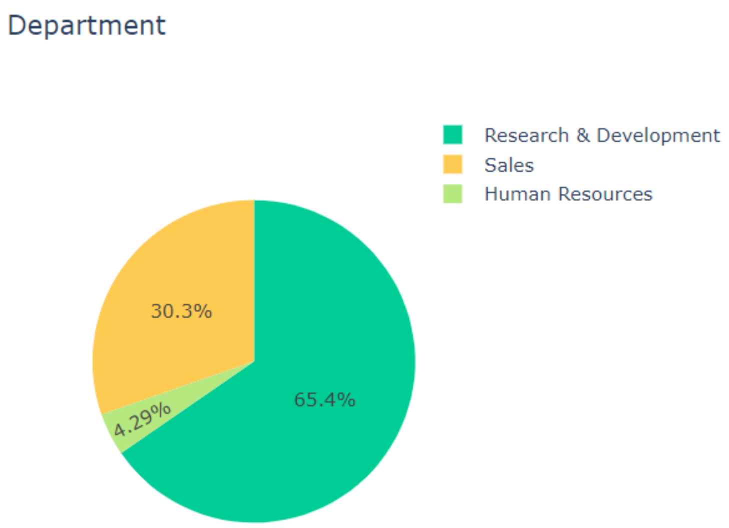
The graph represents that IBM have 60% that is 882 male employees and 40% that is 588 female employees work in the organization.

Distribution by business travel



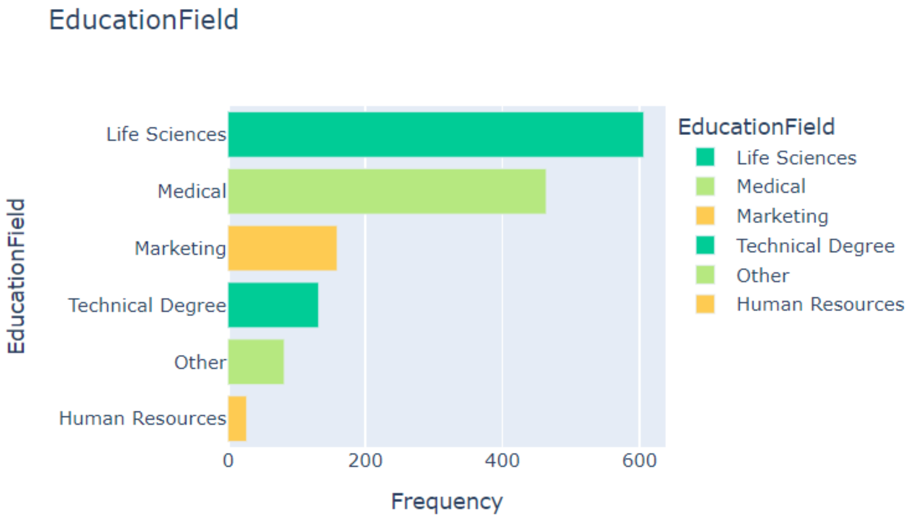
The graph represents that the employees travel rarely as the frequency rate is of 71% that is 1043 and some employees travel frequently with the frequency rate of 18% that is 277 and 10% that is 150 frequency rates of non travel.

Distribution by department



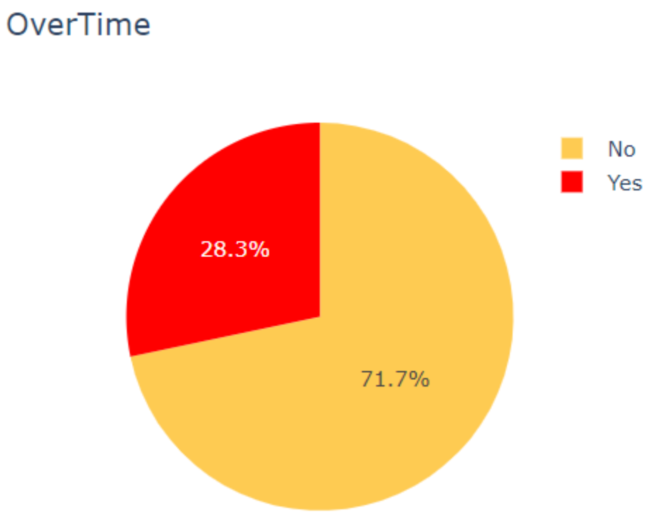
The graph represents that 961 that is 65.4% employees work in R&D Department and 446 that is 30.3% employees work in sales department and 63 that is 4.29% employees work in human resource department.

Distribution by educational field



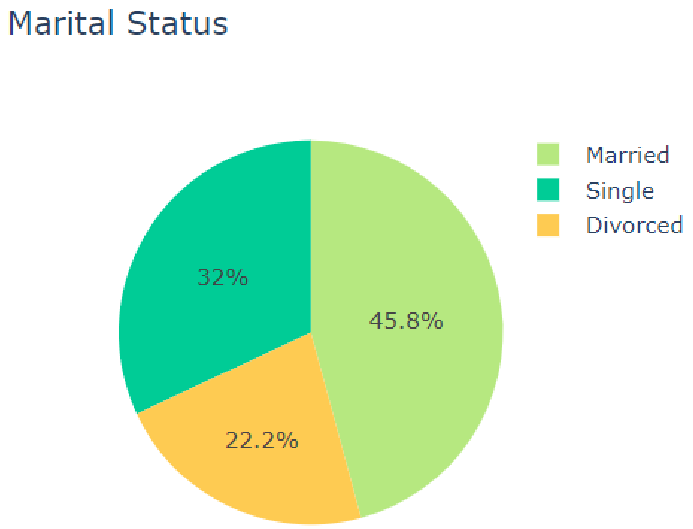
The graph represents that the 606 that is 41% employees in educational field of life sciences and 464 that is 31% employees in educational field of medical and 159 that is 10% employees in educational field of marketing and 132 that is 8% employees in educational field of technical degree and 82 that is 5% employees in educational field of others and 27 that is 2% employees in the field of human resource are working in the organization.

Distribution by over time



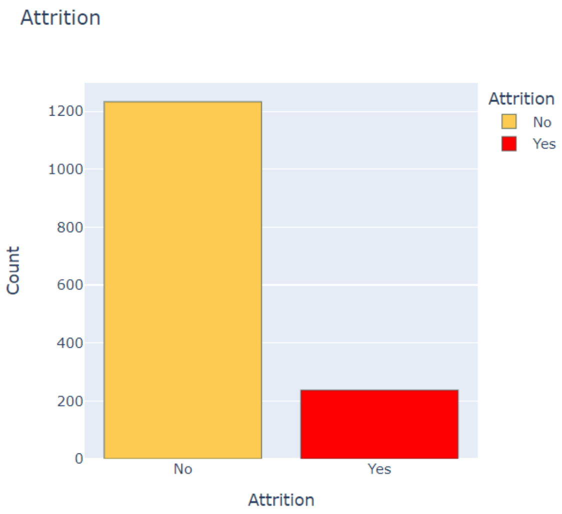
The graph represents that 1054 that is 71.7% employees work overtime and 416 that is 28.3% employees does not work overtime.

Distribution by marital status



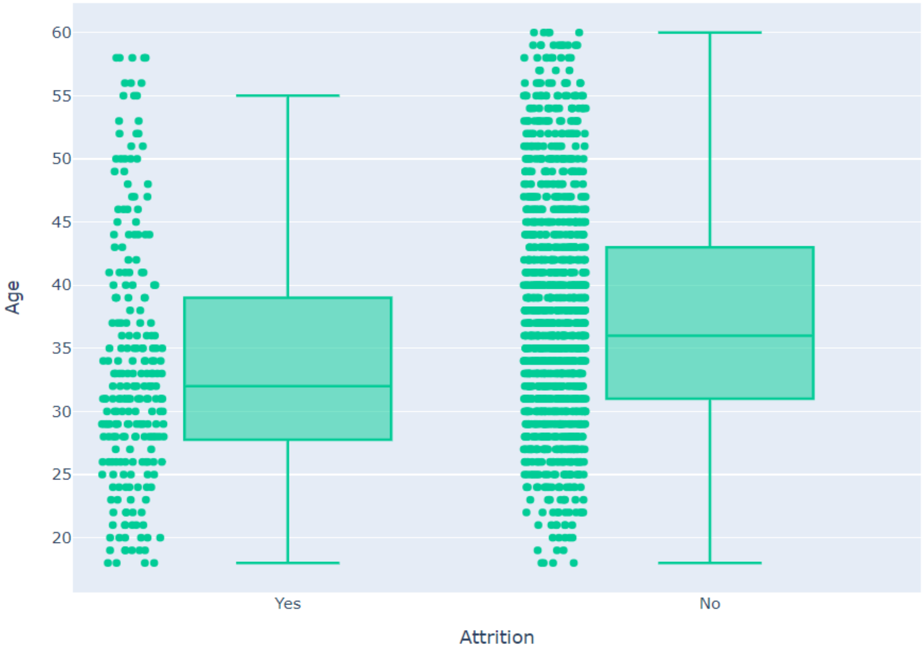
The graph represents that 45.8% that is 673 employees are married and 32% that is 470 employees are single and 22.2% that is 327 employees are divorced in the organization.

Distribution by attrition



The graph represents attrition count of employees and it is found that 1233 employees that is 84% does not need attrition and 237 employees that is 16% need attrition.

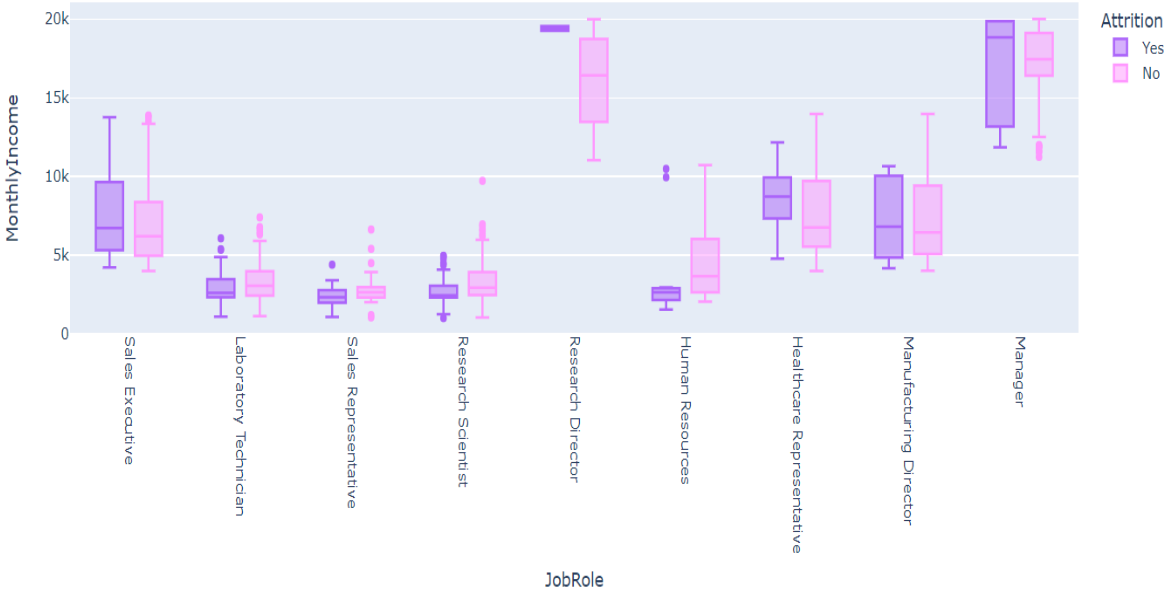
Attrition by Age



Most of the employees are between age 30 to 40. We can clearly observe a trend that as the age is increasing the attrition is decreasing. From the boxplot we can also observe that the median age of employee who left the organization is 32 is less than the median age of employees of 36 who are working in the organization. From the box plot we can observe that the minimum age of employee who left the organization is

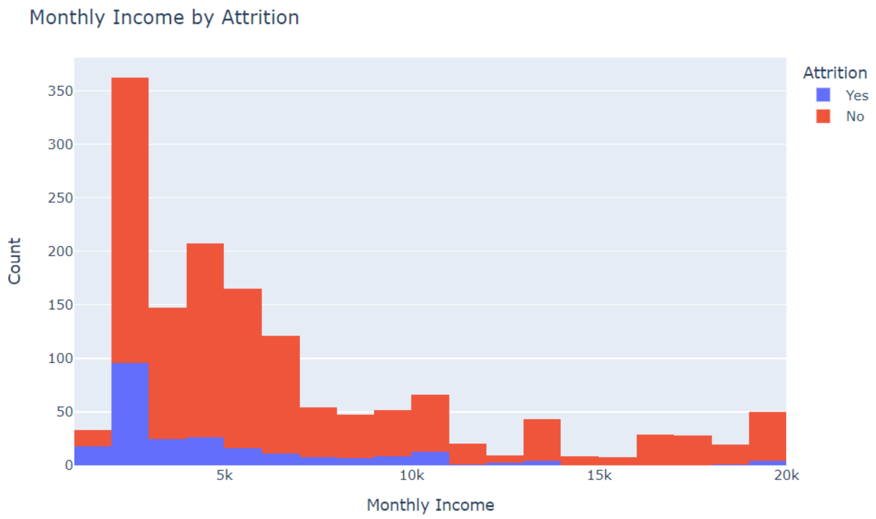
28 is less than the minimum age of employees of 31 who are working in the organization. From the box plot we can observe that the maximum age of employee who left the organization is 39 is less than the maximum age of employees of 43 who are working in the organization. Employees with young age leaves the company more compared to elder employees.

Attrition by Job role and Monthly income



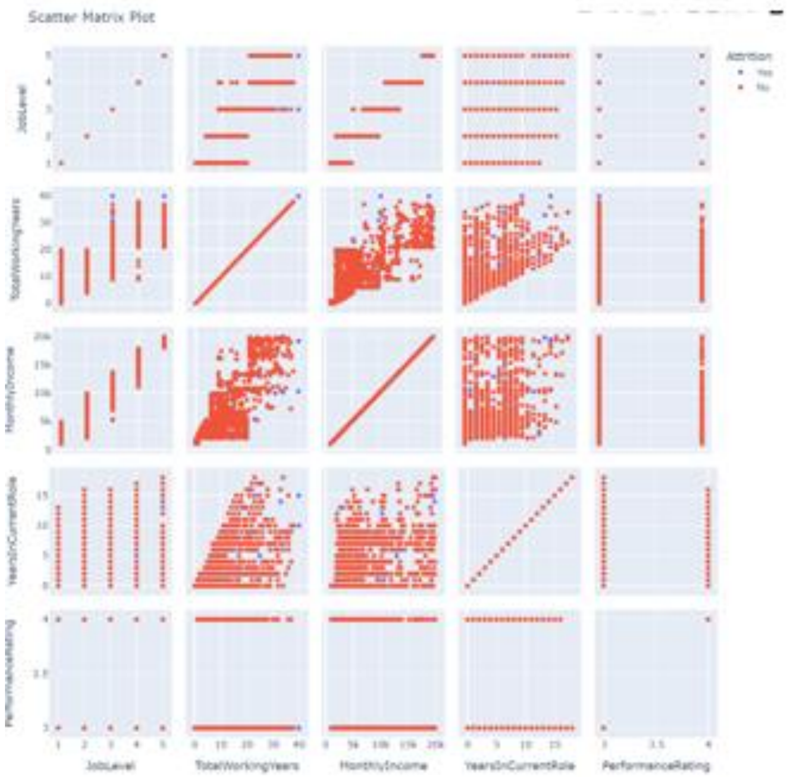
The plot represents the attrition by job role and monthly income it shows that there is less attrition of 2 employee of 78 in Research director and it is found that employees in that department does not quit job and the laboratory technician have the highest of 62 employees of 197 in attrition rate. The sales executive has 57 of 269, sales representative has 33 of 50, the research scientist have 47 of 245, the human resource have 12 of 40, the healthcare representative have 9 of 122, the manufacturing director have 10 of 135 and the manager have 5 of 97 of attrition rate.

Monthly income by attrition



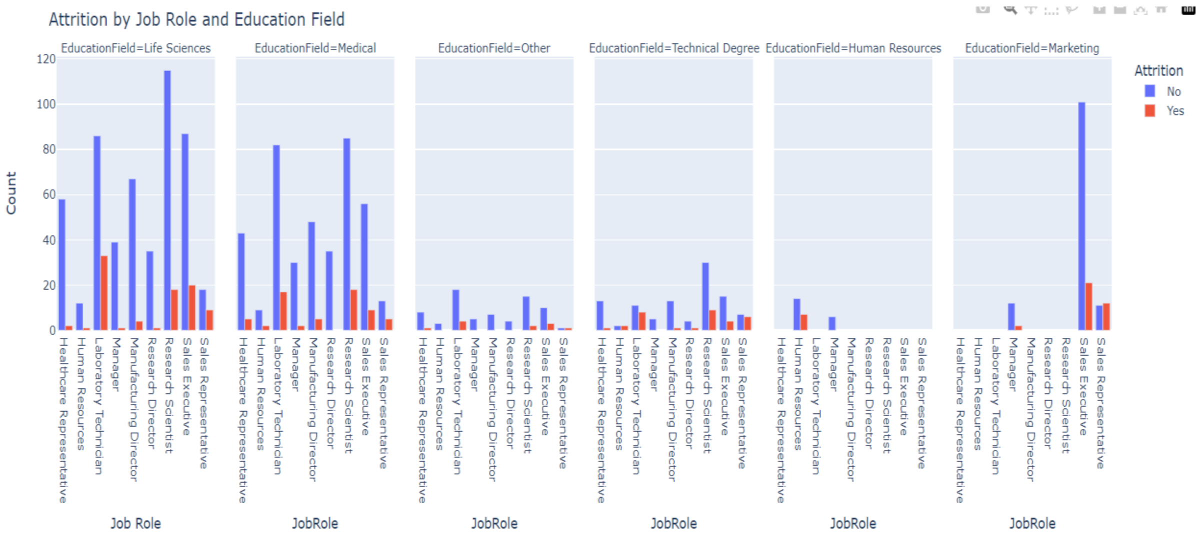
The graph represents the attrition by monthly income it is clear that 267 employees have salary of 2000-3000 but even though they have less salary the attrition rate is less in the organization. The employees who get salary 20000 have a attrition rate of 10% due to many factors. The employees with 10000 seeks better jobs so they leave the organization by 30%.

Scatter matrix plot for comparison of variables



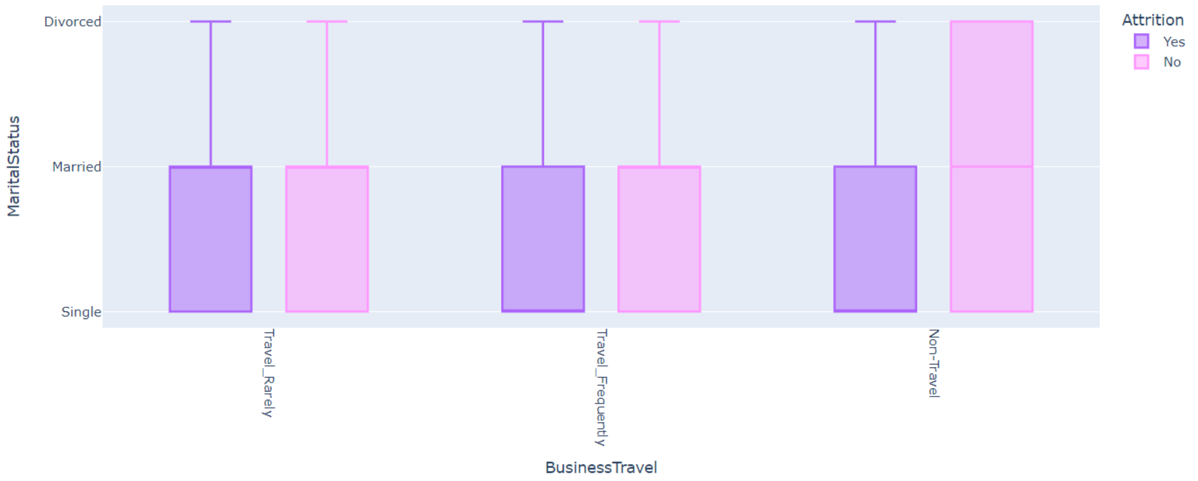
The above graph displays the relationship between the variable like job level, total working hour, monthly income, years in current role and performance rating by taking attrition as the key factor. It is found that the attrition rate is less as we compared with the variables.

Attrition by Job role and Educational Field



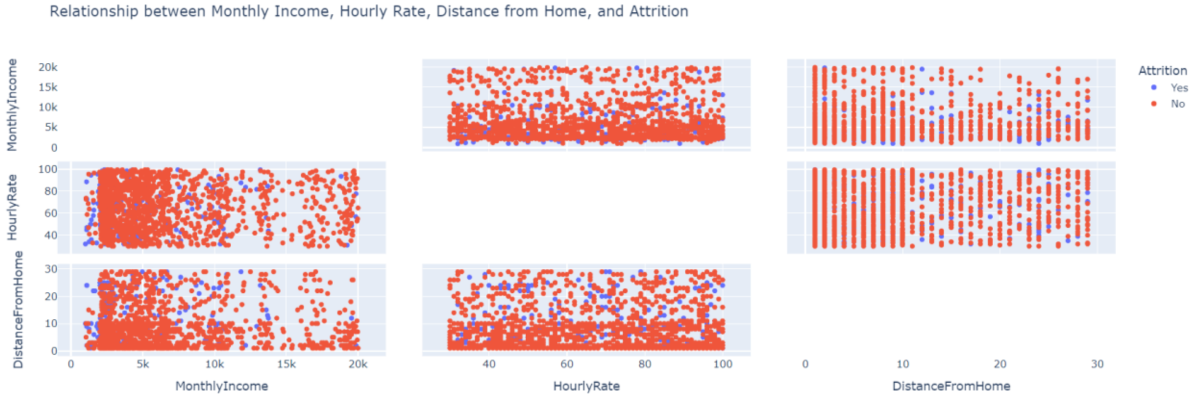
The graph represents that employees in the educational field of life sciences do the jobs what they have studied that is Research scientist of 115 employees so there is less attrition even though employees working rate is high. The employees in the educational field of medical work in the jobs as research scientist and laboratory technician of 80-85 employees as the employees in the studies field the attrition rate is low. The employees in the field of technical degree have the attrition little equal to count as the have more job vacancy in their field so attrition is high. The employees in the field of Human resource have attrition rate half to count in human resource as they have many job offers in many organizations. The employees in the field of marketing the employees work in sales department of 100 employees and the attrition rate is low by 20%.

Attrition by Business travel and Marital status



The graph shows attrition by business travel and marital status and we know the married employees are 673 in that 59 are non travel employees and 118 are travel frequently and 496 are travel rarely. so the married employees are non travel persons so the attrition rate is very low. Single employees are 470 in that most people travel frequently so the attrition rate is very high for single employees.

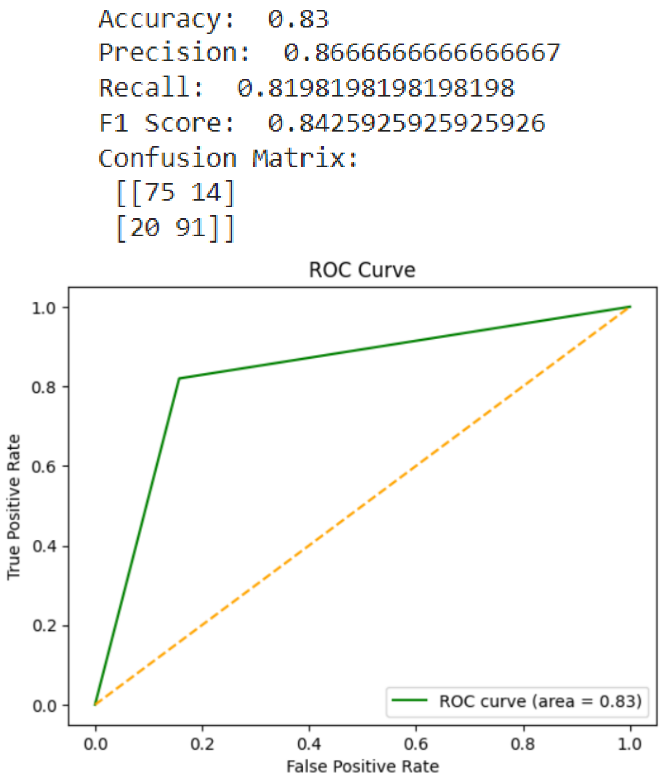
Relationship between Monthly Income, Hourly Rate, Distance from Home and Attirtion



As we can see the employees leaving their jobs even in high salary with the rate of 237. Employees with higher monthly income may be leaving the company because they have better job opportunities elsewhere. Employees with a higher hourly rate may be more satisfied with their pay and therefore less likely to leave the job. Employees who live further away from work may find it difficult to commute and therefore be more likely to leave the even they have good salary and less workinh hour.

Machine Learning models

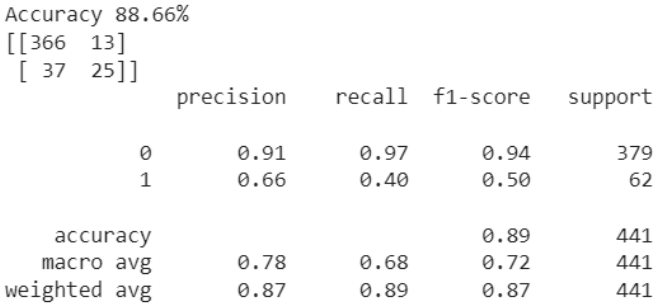
ROC Curve



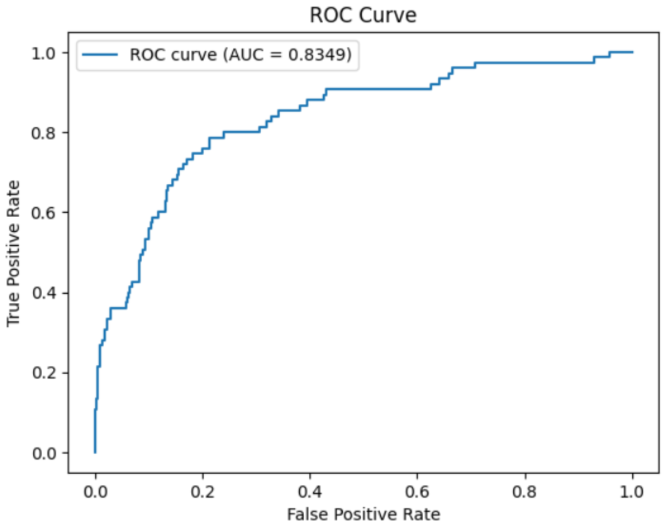
```
Accuracy:  0.83
Precision:  0.8666666666666667
Recall:  0.8198198198198198
F1 Score:  0.8425925925925926
Confusion Matrix:
 [[75 14]
 [20 91]]
```

The ROC (Receiver Operating Characteristic) curve is a graphical representation used to evaluate the performance of a classification model at various threshold settings. It plots the true positive rate (sensitivity/recall) against the false positive rate (1 - specificity) at different classification thresholds. Essentially, it illustrates the trade-off between sensitivity and specificity for different threshold values. By using the ROC curve in predicting employee attrition, HR teams can assess, select, and compare models effectively, aiding in the identification of the most suitable model for accurate predictions and supporting informed decision-making in retention strategies.

Logistic Regression

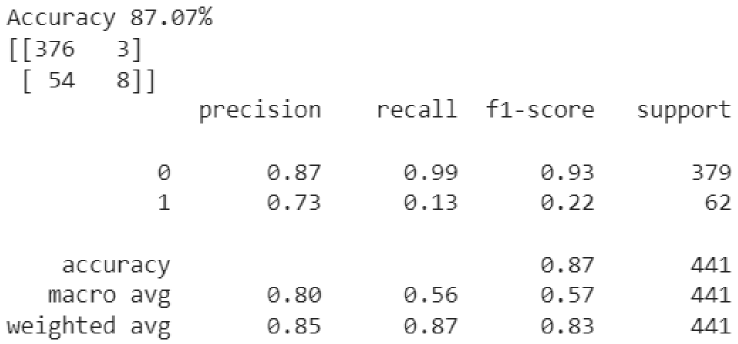


```
Accuracy 88.66%
[[366  13]
 [ 37  25]]
              precision    recall  f1-score   support

           0       0.91      0.97      0.94       379
           1       0.66      0.40      0.50        62

    accuracy                           0.89       441
   macro avg       0.78      0.68      0.72       441
weighted avg       0.87      0.89      0.87       441
```

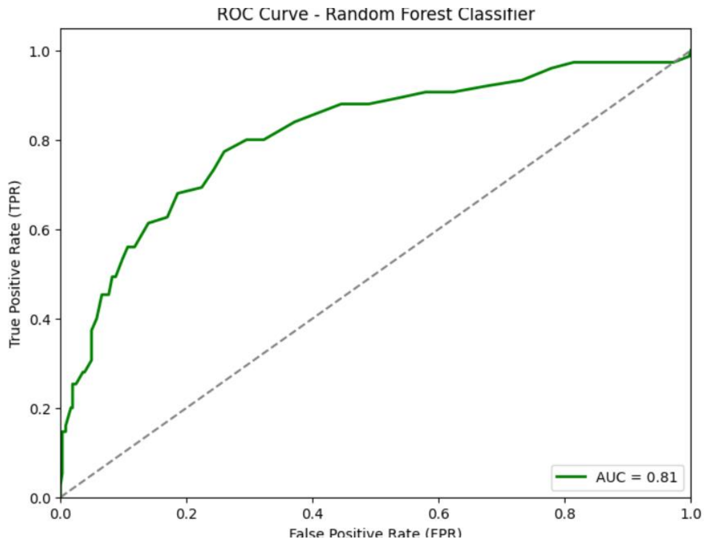


Logistic regression is a statistical method used for binary classification tasks, where the outcome or dependent variable is categorical and has two possible classes. It's particularly useful when predicting the probability of a categorical outcome based on one or more predictor variables. Despite its name, logistic regression is used for classification, not regression tasks. Logistic regression is a statistical method utilized in employee attrition analysis to predict whether an employee will stay or leave an organization. Leveraging various predictors like age, tenure, job satisfaction, and performance metrics, it estimates the probability of attrition. The model exhibits strong accuracy in identifying employees who are likely to remain within the organization but demonstrates notable limitations in correctly predicting those likely to leave. While it effectively identifies employees staying (class 0) with high precision and recall, its performance in detecting employees leaving (class 1) is considerably weaker, reflected in lower precision and recall scores.

Random Forest Classifier



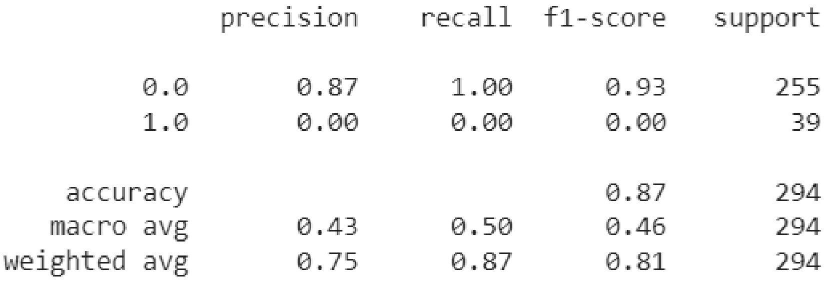
```
Accuracy 87.07%
[[376   3]
 [ 54   8]]
              precision    recall  f1-score   support

           0       0.87      0.99      0.93       379
           1       0.73      0.13      0.22        62

    accuracy                           0.87       441
   macro avg       0.80      0.56      0.57       441
weighted avg       0.85      0.87      0.83       441
```



Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of individual trees. Each tree in the forest is trained on a random subset of the data and uses a random subset of features, resulting in a diverse set of trees that collectively make predictions. Random Forest is an ensemble learning technique utilized in employee attrition analysis for its robust predictive accuracy and ability to handle complex data relationships. By aggregating multiple decision trees trained on subsets of data and features, it offers higher accuracy, identifies influential predictors, and effectively handles nonlinearities and interactions among various employee attributes. The model displays exceptional accuracy and precision in identifying employees likely to stay, demonstrated by perfect recall and high precision scores for non-attrition cases. However, its ability to detect employees likely to leave is significantly limited, with low recall and F1-score for attrition cases.

While it excels in retaining predictions for employees staying, it struggles to effectively capture instances of actual employee turnover, emphasizing the need for improvements in identifying and addressing factors associated with attrition.

Support Vector Machine (SVM)



```
              precision    recall  f1-score   support

         0.0       0.87      1.00      0.93       255
         1.0       0.00      0.00      0.00        39

    accuracy                           0.87       294
   macro avg       0.43      0.50      0.46       294
weighted avg       0.75      0.87      0.81       294
```

Support Vector Machines (SVM) are supervised learning models used for classification and regression tasks. SVM aims to find the optimal hyperplane that best separates data points into different classes while maximizing the margin between the classes. Support Vector Machines (SVM) are utilized in employee attrition analysis due to their capability to discern complex patterns and nonlinear relationships within employee datasets. These models excel in high-dimensional spaces, effectively separating employees likely to stay or leave by maximizing the margin between classes. The model demonstrates strong accuracy and precision in identifying employees likely to stay within the organization, achieving perfect recall and high precision for non-attrition cases. However, its performance in detecting employees likely to leave is notably deficient, indicated by zero recall, precision, and F1-score for attrition cases. While proficient in retaining predictions for employees staying, the model severely lacks the capability to identify instances of actual employee turnover, emphasizing the necessity for substantial improvements in recognizing and addressing factors associated with attrition.

Decision Tree

Accuracy = 0.8605442176870748

F1 Score = 0.36923076923076914

```
Precision = 0.5454545454545454

Recall = 0.27906976744186046

Confusion Matrix:

 [[241 ,10]

 [[ 31 ,12]]
```

The decision tree with an accuracy of approximately 86.05%, the model demonstrates a commendable overall correctness in predicting outcomes. However, a deeper examination through the F1 score of around 0.369 suggests a moderate balance between precision and recall. Precision, measuring around 54.55%, signifies the proportion of correctly identified positive cases among all the predicted positives. Meanwhile, recall, standing at approximately 27.91%, indicates the model's ability to identify the actual positives out of all true positives in the dataset.

The root node begins with the feature 'OverTime', splitting the data into two branches: 'OverTime' less than or equal to 0.50 and 'OverTime' greater than 0.50.

OverTime <= 0.50:

If 'TotalWorkingYears' is less than or equal to 2.50, it further divides based on 'HourlyRate' and 'EnvironmentSatisfaction'. However, these splits predominantly result in class 0 predictions.

When 'TotalWorkingYears' is greater than 2.50, the tree focuses on 'NumCompaniesWorked', 'WorkLifeBalance', and 'Age' features, but most splits lead to class 0 predictions.

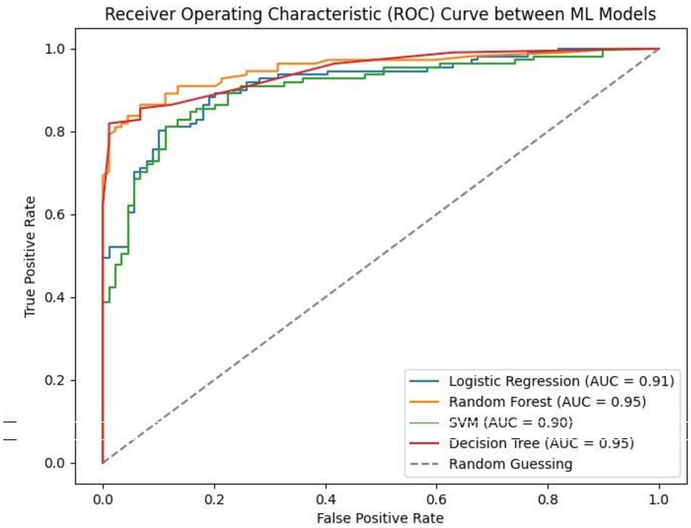
OverTime > 0.50:

For 'MonthlyIncome' less than or equal to 2475.00, the tree considers 'DailyRate', 'Age', and 'YearsInCurrentRole'. It generally predicts class 1.

If 'MonthlyIncome' is above 2475.00, it further segregates based on 'StockOptionLevel', 'JobRole', and 'YearsInCurrentRole', yet tending towards class 0 predictions.

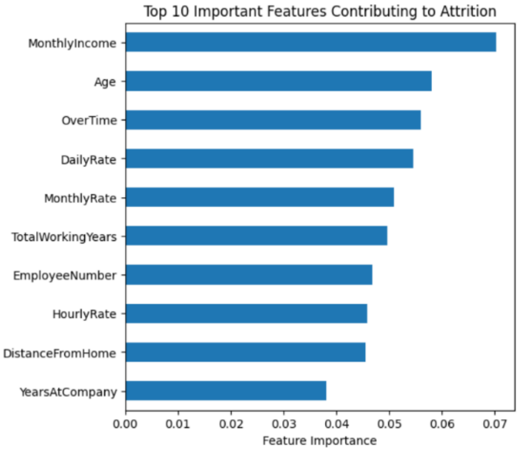
8. Findings for each Model

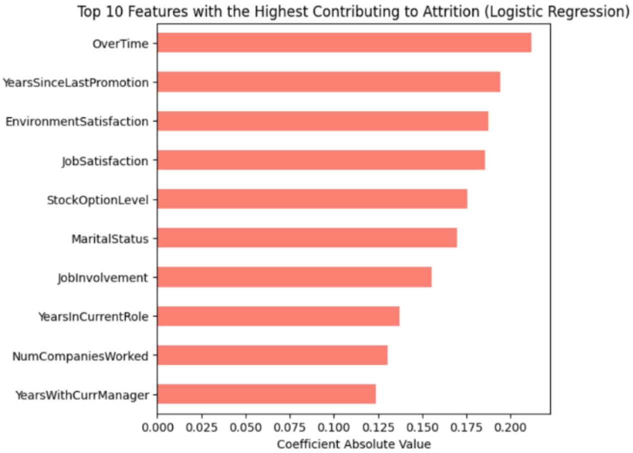
ROC Curve Comparison:



In contrast, Random Forest achieved an accuracy of 87.75%, with significantly higher precision (83.94%), recall (54.93%), and an F1 score of 55.78%. It showed better predictive ability, correctly classifying 254 instances of one class and 4 of the other while misclassifying 1 and 35, respectively mirrored Logistic Regression's performance with an accuracy of 86.73% and identical precision, recall, and F1 score values. It, too, succeeded in predicting one class but struggled with the other, resulting in 255 correct and 39 incorrect predictions.

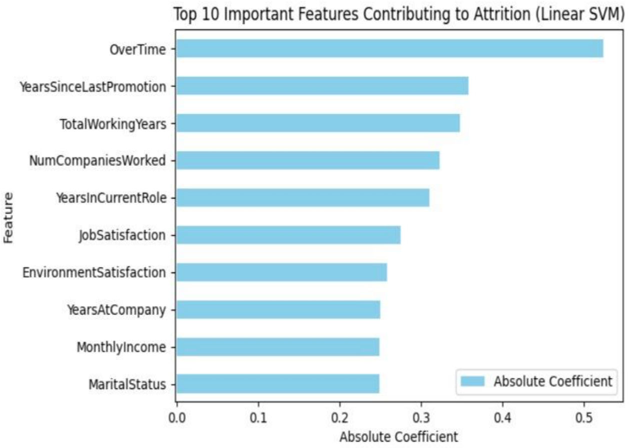
Top 10 important features contributing to attrition

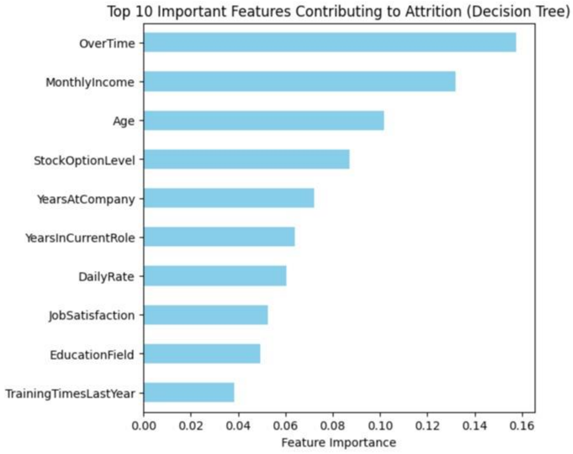




Random Forest

Logistic Regression



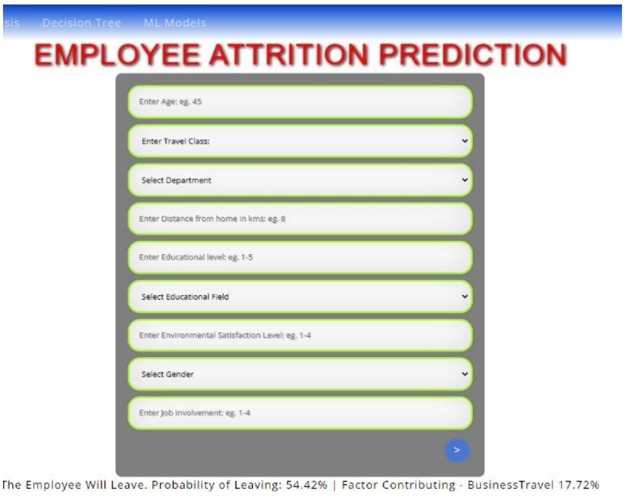


SVM

Decision Tree

The top 10 features contributing to attrition on the basis of random forest model are Monthly income, Age, Over time, Daily rate, Monthly rate, Total working years, Employee number, Hourly rate, Distance from home, Years at company. The top 10 features contributing to attrition on the basis of logistic regression are Over time, Years since last promotion, Environment satisfaction, Job satisfaction, Stock option level, Marital status, Job involvement, Years in current role, Number of companies worked, Years in current manager. Over time plays a major role attrition by 0.200 coefficient absolute value. In all the above graph Overtime play an important features of attrition process in order to reduce the attrition the company should be focuses on the overtime in order to reduce attrition.

Prediction Model



EMPLOYEE ATTRITION PREDICTION

Enter Age: eg. 45

Enter Travel Class:

Select Department

Enter Distance from home in kms: eg. 8

Enter Educational level: eg. 1-5

Select Educational Field

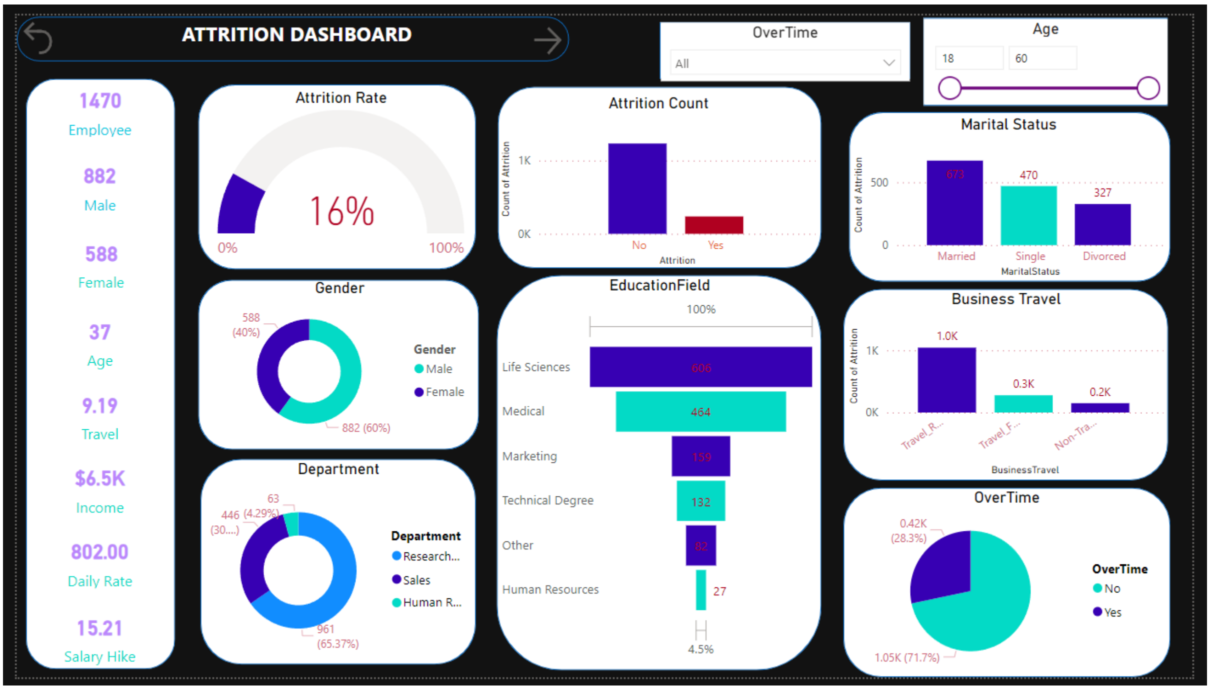
Enter Environmental Satisfaction Level: eg. 1-4

Select Gender

Enter Job involvement: eg. 1-4

The Employee Will Leave. Probability of Leaving: 54.42% | Factor Contributing - BusinessTravel 17.72%

Power Bi



ATTRITION DASHBOARD

OverTime: All

Age: 18 – 60

1470 Employee

882 Male

588 Female

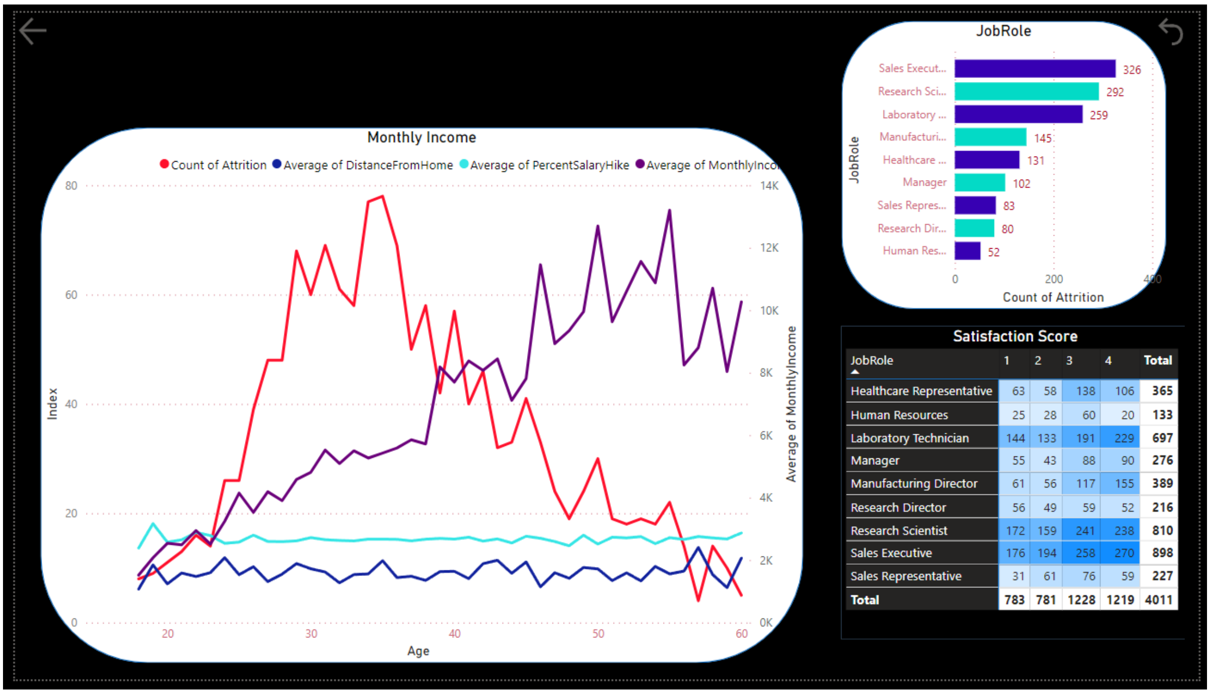
37 Age

9.19 Travel

$6.5K Income

802.00 Daily Rate

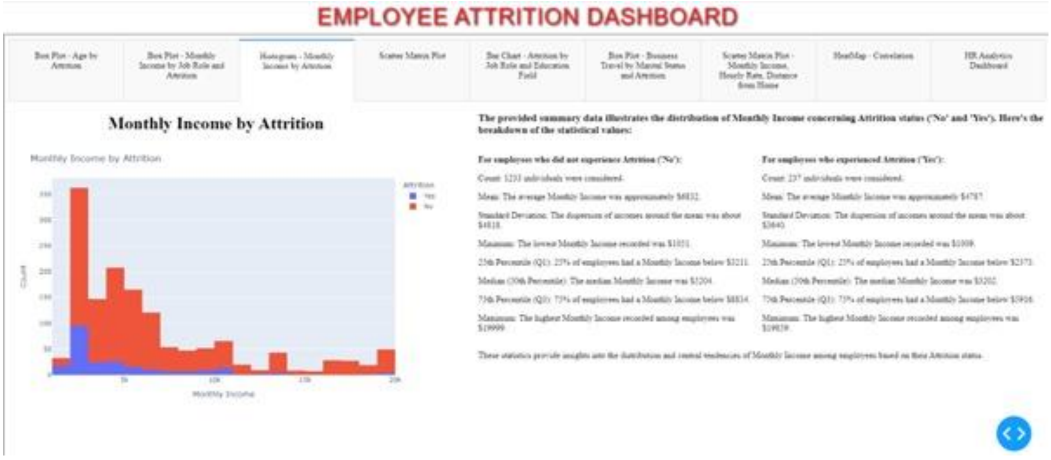
15.21 Salary Hike



Satisfaction Score

JobRole	1	2	3	4	Total
Healthcare Representative	63	58	138	106	365
Human Resources	25	28	60	20	133
Laboratory Technician	144	133	191	229	697
Manager	55	43	88	90	276
Manufacturing Director	61	56	117	155	389
Research Director	56	49	59	52	216
Research Scientist	172	159	241	238	810
Sales Executive	176	194	258	270	898
Sales Representative	31	61	76	59	227
Total	783	781	1228	1219	4011

Output:



EMPLOYEE ATTRITION DASHBOARD

Box Plot - Age by Attrition
Box Plot - Monthly Income by Job Role and Attrition
Histogram - Monthly Income by Attrition
Scatter Matrix Plot
Bar Chart - Attrition by Job Role and Education Field
Box Plot - Business Travel by Marital Status and Attrition
Scatter Matrix Plot - Monthly Income, Hourly Rate, Distance from Home
HeatMap - Correlation
HR Analytics Dashboard
Monthly Income by Attrition
Attrition
Yes
No
Count
Monthly Income
5k
10k
15k
20k

Monthly Income by Attrition

The provided summary data illustrates the distribution of Monthly Income concerning Attrition status ('No' and 'Yes'). Here's the breakdown of the statistical values:

For employees who did not experience Attrition ('No'):

Count: 1233 individuals were considered.

Mean: The average Monthly Income was approximately $6832.

Standard Deviation: The dispersion of incomes around the mean was about $4818.

Minimum: The lowest Monthly Income recorded was $1051.

25th Percentile (Q1): 25% of employees had a Monthly Income below $3211.

Median (50th Percentile): The median Monthly Income was $5204.

75th Percentile (Q3): 75% of employees had a Monthly Income below $8834.

Maximum: The highest Monthly Income recorded among employees was $19999.

For employees who experienced Attrition ('Yes'):

Count: 237 individuals were considered.

Mean: The average Monthly Income was approximately $4787.

Standard Deviation: The dispersion of incomes around the mean was about $3640.

Minimum: The lowest Monthly Income recorded was $1009.

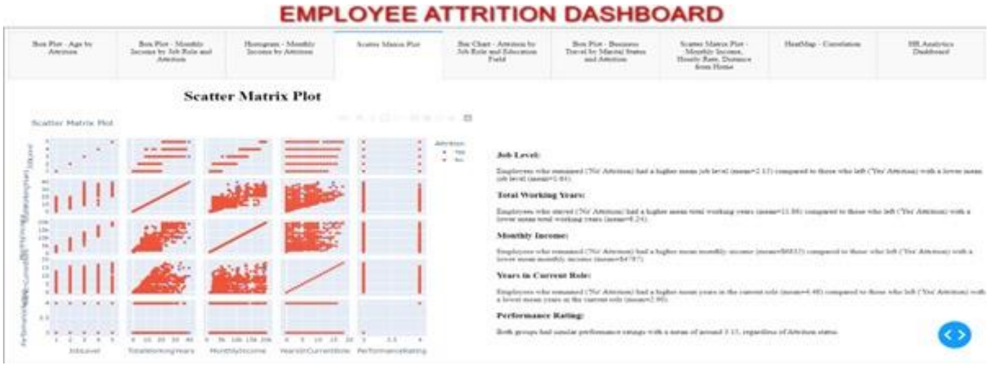
25th Percentile (Q1): 25% of employees had a Monthly Income below $2373.

Median (50th Percentile): The median Monthly Income was $3202.

75th Percentile (Q3): 75% of employees had a Monthly Income below $5916.

Maximum: The highest Monthly Income recorded among employees was $19859.

These statistics provide insights into the distribution and central tendencies of Monthly Income among employees based on their Attrition status.



EMPLOYEE ATTRITION DASHBOARD

Box Plot - Age by Attrition
Box Plot - Monthly Income by Job Role and Attrition
Histogram - Monthly Income by Attrition
Scatter Matrix Plot
Bar Chart - Attrition by Job Role and Education Field
Box Plot - Business Travel by Marital Status and Attrition
Scatter Matrix Plot - Monthly Income, Hourly Rate, Distance from Home
HeatMap - Correlation
HR Analytics Dashboard
Scatter Matrix Plot
Attrition
Yes
No
JobLevel
TotalWorkingYears
MonthlyIncome
YearsInCurrentRole
PerformanceRating

Scatter Matrix Plot

Job Level:

Employees who remained ('No' Attrition) had a higher mean job level (mean=2.15) compared to those who left ('Yes' Attrition) with a lower mean job level (mean=1.64).

Total Working Years:

Employees who stayed ('No' Attrition) had a higher mean total working years (mean=11.86) compared to those who left ('Yes' Attrition) with a lower mean total working years (mean=8.24).

Monthly Income:

Employees who remained ('No' Attrition) had a higher mean monthly income (mean=$6833) compared to those who left ('Yes' Attrition) with a lower mean monthly income (mean=$4787).

Years in Current Role:

Employees who remained ('No' Attrition) had a higher mean years in the current role (mean=4.48) compared to those who left ('Yes' Attrition) with a lower mean years in the current role (mean=2.90).

Performance Rating:

Both groups had similar performance ratings with a mean of around 3.15, regardless of Attrition status.

EMPLOYEE ATTRITION DECISION TREE
True
False

9. Conclusion

The employee attrition, the natural departure of employees from the workforce due to various factors, has been analyzed using machine learning models—Logistic Regression, Random Forest, and Support Vector Machines (SVM). The organization experiences an overall attrition rate of 16.12%, with notable gender disparities and age-related trends. Despite male employees constituting a larger segment, their attrition rate surpasses that of females. Employees within the 30 to 40 age range exhibit a decreasing attrition trend as age increases. Additionally, frequent travelers show higher attrition rates, while non-travelers exhibit the lowest rates. Surprisingly, the Research & Development department, despite having the highest employee count, displays the lowest attrition rate among all departments.

The predictive performance of the machine learning models varied. Logistic Regression achieved an 86.73% accuracy but showed moderate Precision (43.37%) and Recall (50%), failing to predict positive instances. Random Forest performed slightly better with 87.75% accuracy, demonstrating higher Precision (83.94%) and Recall (54.93%). However, it misclassified one instance and missed 35 positive instances in its confusion matrix. SVM mirrored Logistic Regression's metrics, also achieving 86.73% accuracy with identical Precision, Recall, and F1 Score.

The top contributing features to attrition were identified by the ML models Monthly Income, Age, Over Time, among others, Years since Last Promotion, and factors related to job satisfaction and involvement. Over Time emerged as a pivotal factor, indicating a significant impact on attrition. Thus, addressing overtime concerns could potentially reduce attrition rates within the company.

Recommendation:

Age-specific Approaches: Tailoring development programs, mentorship initiatives, and flexible work arrangements to meet the diverse career aspirations and needs of employees across different age groups.

Compensation Review: Ensuring competitive compensation packages and introducing performance-based incentives to motivate employees and acknowledge their contributions.

Career Growth: Providing opportunities for advancement, skill enhancement, and cross-functional training. Clear career paths, regular feedback, and evaluations support employee growth and engagement, customized based on job roles and responsibilities.

Job Satisfaction: Enhancing engagement and satisfaction by fostering a supportive work environment, offering professional development opportunities, and ensuring transparent processes for promotions and career development.

Work-life Balance: Focusing on factors like environment satisfaction, job involvement, satisfaction, and managing overtime demands. Regular surveys to comprehend employee concerns and taking proactive steps to address identified areas of improvement.

Organizational Culture: Creating a positive culture that values employee well-being, work-life balance, and growth. Encouraging open communication, feedback, and continuous evaluation of retention strategies to adapt to evolving employee needs.

Overtime: This is one the important feature among all the above feature contributes the attrition. Reduce the overtime in the company that would reduce the attrition among the company.

10. Reference

1. https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/
2. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
3. https://medium.com/dschunks/ibm-hr-dataset-analysis-58af52641cac
4. https://towardsdatascience.com/using-ml-to-predict-if-an-employee-will-leave-829df149d4f8
5. https://www.semanticscholar.org/paper/IBM-Employee-Attrition-Analysis-Yang-Islam/e47d6b3274a3ca7fbc94940cd491409cc001c0be