Project Report

On

"IBM HR Analytics for Employee Attrition and Performance Prediction"

Submitted
By

Supriya K L

UMIP271280

ABSTRACT

In numerous software companies, there is a noticeable trend of employees resigning from their positions for various reasons. When skilled and valuable employees depart, it poses significant challenges for organizations in maintaining their operations. Consequently, it is crucial for companies to proactively identify and assess the causes of employee turnover and formulate suitable strategies and actions to address this issue. IBM's HR Analytics Employee Attritionand Performance datasets are being utilized for this purpose. Missing values were dropped to give better insights in data analysis. ANOVA and Chi-Square tests were carried out during statistical analysis. Machine Learning algorithms such as Logistic Regression (92%), Random Forest (89%), Support Vector Machine (93%), XGBoost (100%), CatBoost (98%), AdaBoost (90%) and LightGBM (100%) were applied to understand, manage, and mitigate employee attrition. Comparison of model performance was plotted on ROC Curve using True-Positive and False-Positive Rate.

INDEX

LISTOFFIGURES

LISTOFTABLES

CHAPTER1:INTRODUCTION

INTRODUCTION

Companies, both in India and other countries, face a formidable challenge in recruiting and retainingtoptalent,allwhiledealingwithtalentlossthrough attrition,whetherduetoindustry downturns or voluntary turnover. Losing employees not only results in performance setbacks but also has long-term negative impacts on companies. This includes potential disruptions to productivity, work team dynamics, and social goodwill.

The success and competitiveness of any organization are highly dependent on its workforce, with employees serving as the essential backbone of the company. This study aims to identify employee attitudes, pinpoint the factors that contribute to their dissatisfaction within the organization, and understand the reasons behind their decisions to seek alternative employment opportunities.

By identifying and assessing the levels of employee attitudes, management can gain valuable insightsintoareasthatrequireimprovementandtakenecessaryactiontoreduceattritionrates. This proactive approach is essential for sustaining a productive and harmonious work environment while enhancing the organization's overall performance and long-term success.

Fig1.1SkillsinHumanResource[HR]

OBJECTIVE

1. Assessthedegreeofemployeesatisfactionregardingtheirjobandworkingenvironment.
2. Identifytheelementsthatcontributetoemployeedissatisfactionwiththecompany's policies and guidelines.
3. Pinpointtheareaswherethecompanyisfallingshortorfacingshortcomings.
4. Understandtheunderlyingcausesofattritionwithinthecompany.
5. Developstrategiesandmethodstominimizeattritionrateswithintheorganization.

MOTIVATION

This project originates from the possibility of enhancing employee contentment, cutting down expenses, elevating organizational efficiency, and fostering a favorable workplace atmosphere. It represents a chance to leverage data and analytics to enact substantial improvements that are advantageous to both employees and the entire organization.

CHAPTER2:SYSTEMARCHITECTURE

PROPOSEDARCHITECTURE(BLOCKDIAGRAM)

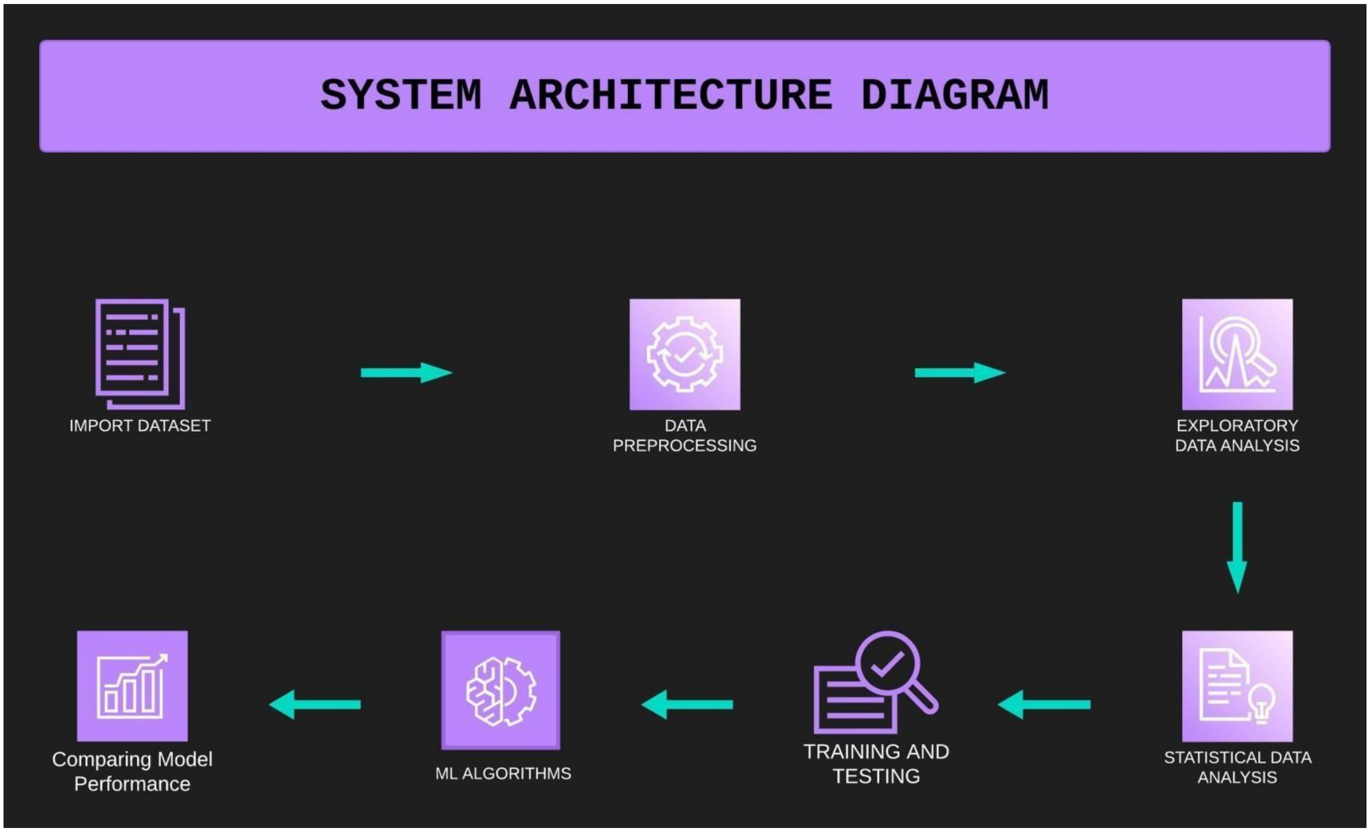

Fig.2.1ProposedArchitecture

ThemethodologyforIBM HRAnalytics EmployeeAttrition andPerformancePredictionis as follows:-

- InputistakenbyloadingtheODIRdataset,whichcontainsocular
- Load the Dataset: The IBM HR Analytics Attrition Dataset is loaded using the pd.read_csv() function. The head() and info() methods are used to display the first few rows and get information about the dataset, respectively.
- Knowing the Dataset: Basic Information about the dataset is generated; numerical and categorical attributes are enlisted.

- DataCleaning: Anymissingvaluesin thedatasetaredropped usingthe dropna() method.
- DataVisualization:Matplotlib andSeabornlibraries areusedto visualizethe data.
- Statistical Analysis: The ANOVA Test is performed to analyze the Numerical Features' Importance in Employee Attrition, while the Chi-Square Test to Analyze the Categorical Feature Importance in Employee Attrition.
- Data Preprocessing: The target variable 'Attrition' is mapped to binary values (1 for 'Yes' and 0 for 'No'). Selected features areextracted from the dataset and one-hot encoded using the get_dummies() function.
- Splitting the Dataset: The dataset is split into training and testing sets using the train_test_split() method from scikit-learn.
- Implementing Machine Learning Algorithms: Logistic Regression, XGBoost, CatBoost, AdaBoost, LightGBM, Decision Tree, and Random Forest classifiers are initialized and trained using the training data.
- Model Evaluation: The accuracy score and confusion matrix are computed to evaluate the performance of each algorithm on the testing data.
- Results: The results, including the accuracy and confusion matrix, are printed for each algorithm.
- ModelPerformanceComparison:ThehvPlotlibraryisusedtovisualizetheROCcurve diagram comparing the performance of all models used.

TECHNOLOGYUSED

Technologythat would be used in this project are:

PythonProgramming,MachineLearning,DataAnalytics,Statistical Analytics.

DATASET

Thisdatasetpresentsanemployeesurveyfrom IBM,indicatingifthereisattritionornot.The data set contains approximately 1500 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs. a random allocation of probability of attrition. IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees. To use a matrix structure, we changed the model to reflect the following data.

DatasetDescription:

Duringwebsitesession,browsinginformationaboutvisitedpagesiscollectedandfeaturesare extracted as follows:

Table1–Numericalfeaturesusedintheuserattritionanalysis model.

FeatureName	FeatureDescription	Min Value	Max Value	Std. Dev
Age	Ageofemployee	18	60	9.13
DailyRate	Itisthebillingcostforan individual's servicesforasingleday	102	1499	403.50
DistanceFromHome	It is the distance between companyandhomeofthe employee	1	29	8.10
Education	Educationqualificationofthe employees of company	1	5	1.02
EmployeeCount	Countofemployee	1	1	0.0

EmployeeNumber	It is a unique number that has been assignedtoeachcurrentandformer employee	1	2068	602.02
EnvironmentSatisfaction	It is all about an individual's feelings about the work environmentandorganization culture.	1	4	1.09
HourlyRate	Theamountofmoneythatispaidto anemployeeforeveryhourworked	30	100	20.32
JobInvolvement	Jobinvolvementrefersto the degreetowhichajobiscentralto a person'sidentity.	1	4	0.71
JobLevel	Job levels are categories of authorityinanorganization.	1	5	1.10
JobSatisfaction	Job satisfaction happens when an employeefeelsheorsheishaving job stability.	1	4	1.10
MonthlyIncome	Gross monthly income is the amountofincomeanemployeeearn in one month.	1009	19999	4707.95
MonthlyRate	If a monthly rate is set, employeesshouldbepaidin exchangefornormalhoursofwork of a full-time worker.	2094	26999	7117.78
NumCompaniesWorked	Numberofothercompaniesthe employeepreviouslyworkedfor	0	9	2.49
PercentSalaryHike	Theamountasalaryisincreasedof anemployeeinpercentage	11	25	3.65
PerformanceRating	Ratingmeansgauging and comparingtheperformance.	3	4	0.36
RelationshipSatisfaction	Itistherateofsatisfactionbetween Employer employee relationship.	1	4	1.08

Table1:Showsthenumericalfeatures alongwiththeirstatisticalparameters.

Table2–CategoricalFeaturesusedinthe UserAttritionAnalysisModel.

FeatureName	FeatureDescription	Numberof Categorical Values
Attrition	Attritioninbusinessdescribesagradualbut deliberate reduction in staff numbers thatoccursasemployeesretireorresign, [NOTE: Target Variable] (0=no, 1=yes)	2
BusinessTravel	Businesstravelistravelundertakenforworkor business purposes, asopposedtoothertypesoftravel(1=No Travel,2=TravelFrequently,3=TravelRarely)	3
Department	Consists three departments that contribute to the company'soverallmission.(1=HR,2=R&D,3=Sales)	3
EducationField	Educationfieldoftheemployees(1=HR,2=Life Sciences, 3=Marketing, 4=Medical Sciences, 5=others, 6= Technical)	6
Gender	Genderoftheemployee(1=Female,2=Male)	2
JobRole	Thesearefertothespecificactivitiesorworkthatthe employee will perform. (1=HC Rep, 2=HR, 3=Lab Technician, 4=manager, 5= Managing Director, 6= Research Director, 7= Research Scientist, 8=sales Executive, 9= Sales Representative)	9
MaritalStatus	MaritalStatus oftheemployee(1=divorced, 2=married,3=single)	3
Over18	(1=Yes,2=No)	2
Overtime	(1=No,2=Yes)	2

Table2:Showsthecategoricalfeaturesalongwiththeirnumberofcategories.

CHAPTER3:IMPLEMENTATION

DATAEXPLORATIONANDPROCESSING

ComputeSize:

In firststep,wetrytounderstandthedataset'ssizeandstructureataglancebycomputingit’s size.

1] COMPUTING SIZE OF DATASET

```
In [5]:
# Print the shape of the DataFrame
print("The shape of data frame:", employee_data.shape)
# Print the length (number of rows) of the DataFrame
print("Number of Rows in the dataframe:", len(employee_data))
# Print the number of columns in the DataFrame
print("Number of Columns in the dataframe:", len(employee_data.columns))
```

```
The shape of data frame: (1470, 35)
Number of Rows in the dataframe: 1470
Number of Columns in the dataframe: 35
```

Fig.3.1.1ComputeSizeof Dataset.

Thecoderevealsthatthe"employee_data"DataFramecontains1,470rowsand35columns, providing a quick overview of its size and structure.

Drop Columns:

Inthisstep,wenoticethat'EmployeeCount','Over18','StandardHours'haveonlyoneunique valuesand'EmployeeNumber'has1470uniquevalues.Thesefeaturesaren'tusefulforus,so we are going to drop those columns.

```
In [20]:
employee_data.drop(['EmployeeCount', 'EmployeeNumber', 'Over18', 'StandardHours'], axis="columns", inplace=True)
```

```
In [22]:
# Print the shape of the DataFrame
print("The shape of data frame:", employee_data.shape)
# Print the length (number of rows) of the DataFrame
print("Number of Rows in the dataframe:", len(employee_data))
# Print the number of columns in the DataFrame
print("Number of Columns in the dataframe:", len(employee_data.columns))
```

```
The shape of data frame: (1470, 31)
Number of Rows in the dataframe: 1470
Number of Columns in the dataframe: 31
```

Fig.3.1.2ComputeSizeofDatasetafterdroppingcolumns.

The code reveals that the "employee_data" DataFrame now contains 1,470 rows and 31 columns,providingaquickoverviewofitssizeandstructureafterdroppingfewcolumns.

DATAVISUALIZATION

Byanalyzing employee data, we can identifyfactors that contribute to employee attrition, such as job satisfaction, compensation, and work-life balance. This information can be used to develop strategies to retain top talent and reduce turnover rates. HR analytics can help identify high-performing employees by analyzingdata related to performance metrics,such as productivity, quality,and customersatisfaction. Thisinformationcanbeusedtodevelop strategiestoretaintop talent and improve overall organizational performance.

1] VISUALIZINGTHEEMPLOYEEATTRITIONRATE.

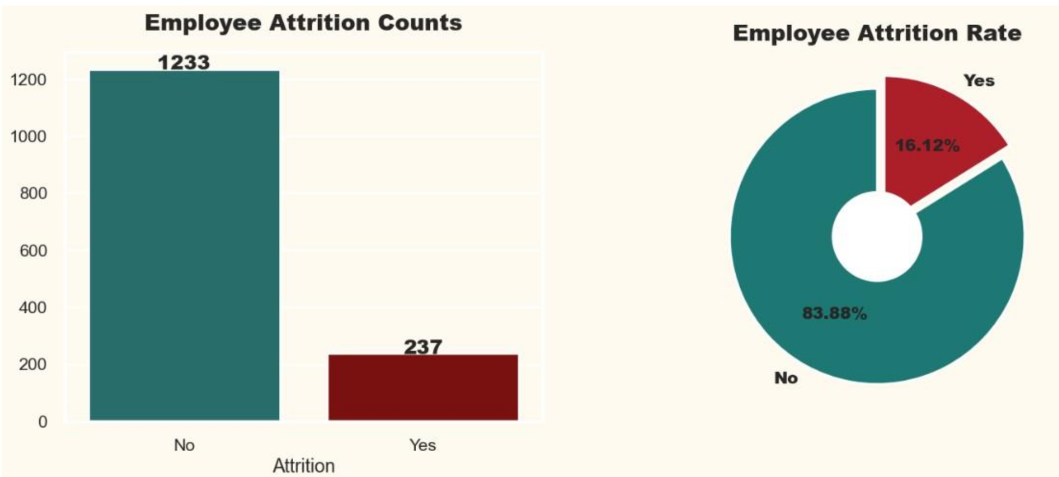

Fig.3.2.1:VisualizingEmployeeAttritionRate.

Inference:

1. Theemployee attritionrate ofthisorganizationis 16.12%.
2. AccordingtoexpertsinthefieldofHumanResources,saysthattheattritionrate4%to6%is normal in organization.
3. Sowecansaytheattrition rateoftheorganization isat adangerous level.
4. Thereforetheorganizationshouldtakemeasurestoreducetheattrition rate.

2] ANALYZINGEMPLOYEEATTRITIONBYGENDER.

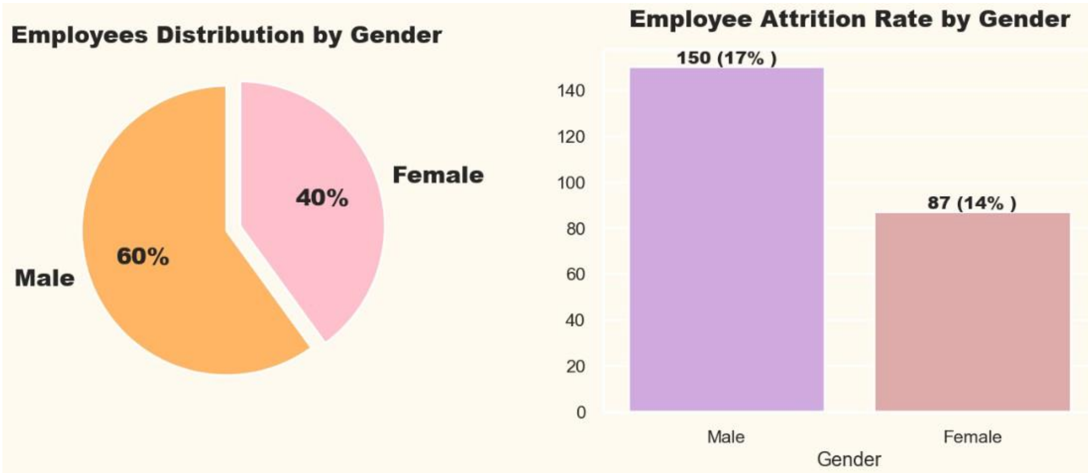

Fig.3.2.2:AnalyzingEmployeeAttritionbyGender.

Inference:

1. Thenumberofmaleemployeesintheorganizationaccountsforahigherproportionthan female employees by more than 20%.
2. Maleemployees areleavingmorefromtheorganizationcomparedtofemaleemployees.

3] ANALYZINGEMPLOYEEATTRITIONBYAGE.

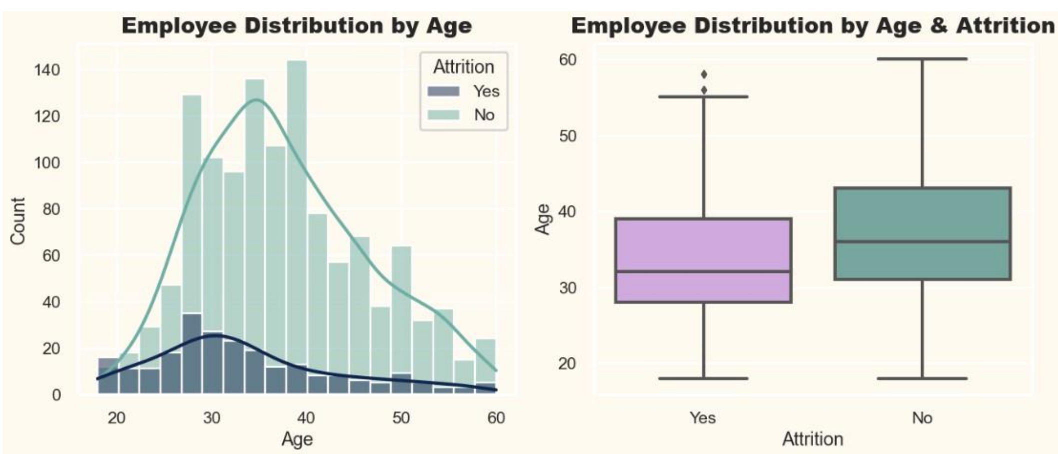

Fig.3.2.3:AnalyzingEmployeeAttritionbyAge.

Inference:

1. Mostoftheemployeesarebetweenages30to 40.
2. Wecan clearlyobserveatrend that asthe ageis increasingtheattrition is decreasing.
3. Fromtheboxplotwecanalsoobservethatthemedianageofemployeewholefttheorganization is less than the employees who are working in the organization.
4. Employeeswithyoungageleavesthecompanymorecomparedto elder employees.

4] ANALYZINGEMPLOYEEATTRITIONBYBUSINESSTRAVEL.

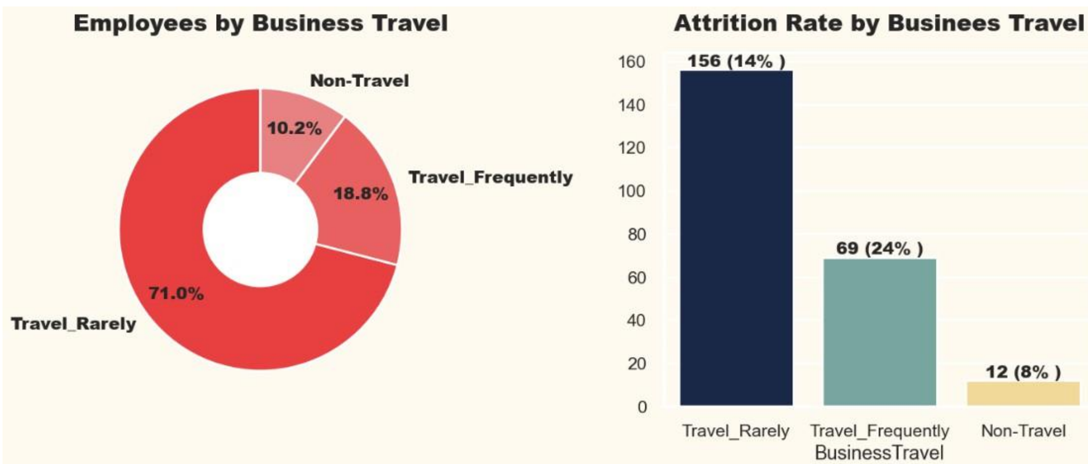

Fig.3.2.4:AnalyzingEmployeeAttritionbyBusinessTravel.

Inference:

1. Mostoftheemployeesin theorganizationTravelRarely.
2. Highestemployeeattrition canbeobservedbythoseemployees whoTravels Frequently.
3. Lowestemployeeattrition can beobservedbythoseemployees whoare Non-Travel.

5] ANALYZINGEMPLOYEEATTRITIONBYDEPARTMENT.

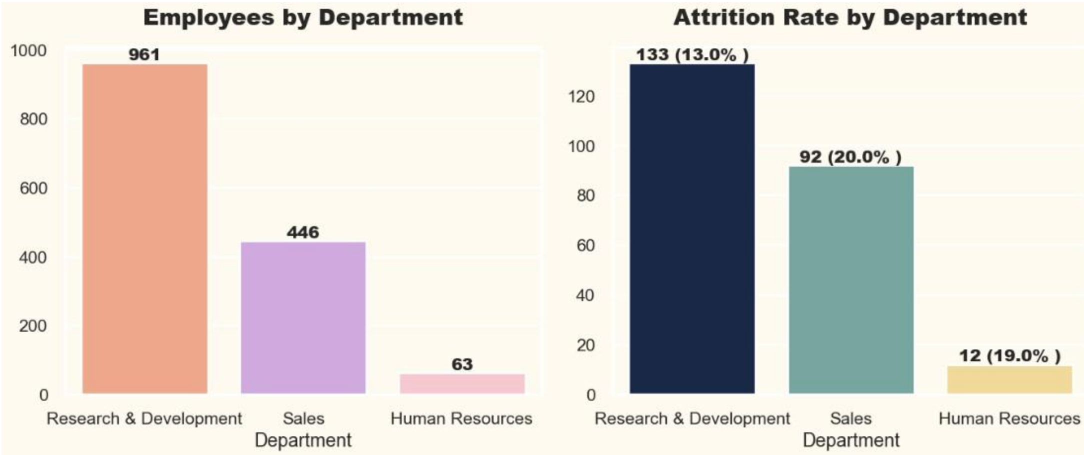

Fig.3.2.5:AnalyzingEmployeeAttritionbyDepartment.

Inference:

1. MostoftheemployeesarefromResearch&Development Department.
2. HighestAttritionisintheSales Department.
3. HumanResources DepartmentAttritionrate isalso veryhigh.
4. ThoughofhighestemployeesinResearch&Developmentdepartmentthereisleast attrition compared to other departments.

6] ANALYZINGEMPLOYEEATTRITIONBYDAILYRATE.

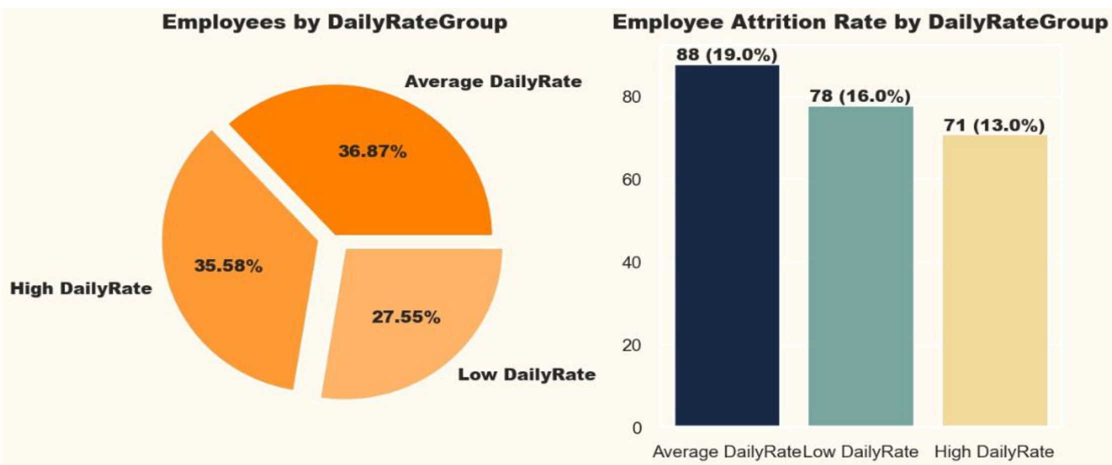

Fig.3.2.6:AnalyzingEmployeeAttritionbyDailyRate.

Inference:

1. Employees withAverage DailyRate &HighDailyRateareapproximatelyequal.
2. Buttheattritionrateisveryhighofemployeeswithaverage DailyRatecomparedtothe employees with High DailyRate.
3. Theattritionrateisalsohighofemployeeswithlow DailyRate.
4. Employeeswho arenotgettingHigh DailyRate aremostlyleavingthe organization.

7] ANALYZINGEMPLOYEEATTRITIONBYDISTANCEFROMHOME.

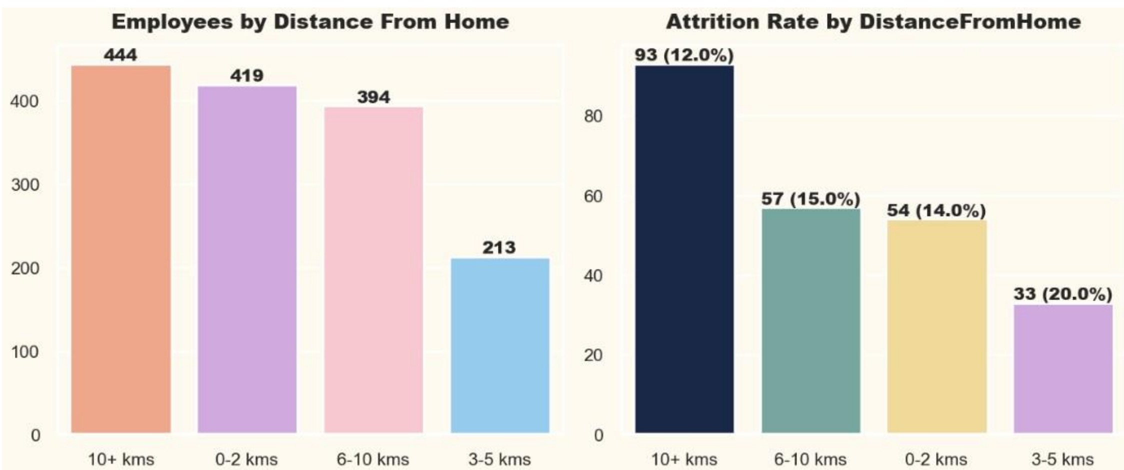

Fig.3.2.7:AnalyzingEmployeeAttritionbyDistancefromHome.

Inference:

1. Intheorganizationthereisallkindofemployeesstayingcloseorstayingfarfromthe office.
2. ThefeatureDistancefromHome doesn't follow anytrend inattrition rate.
3. Employeesstayingclosetotheorganizationaremostlyleavingcomparedtoemployees staying far from the organization.

8] ANALYZINGEMPLOYEEATTRITIONBYEDUCATION.

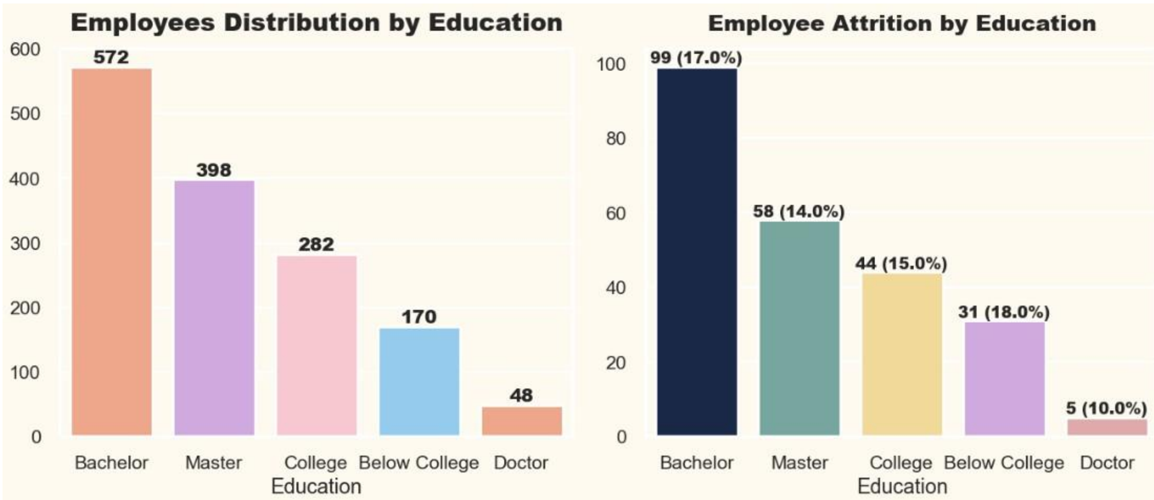

Fig.3.2.8:AnalyzingEmployeeAttritionbyEducation.

Inference:

1. MostoftheemployeesintheorganizationhavecompletedBachelorsorMastersastheireducation qualification.
2. VeryfewemployeesintheorganizationhavecompletedDoctoratedegreeastheireducationqualification.
3. Wecanobserveatrendofdecreasinginattritionrateastheeducationqualificationincreases.

9] ANALYZINGEMPLOYEEATTRITIONBYEDUCATIONFIELD.

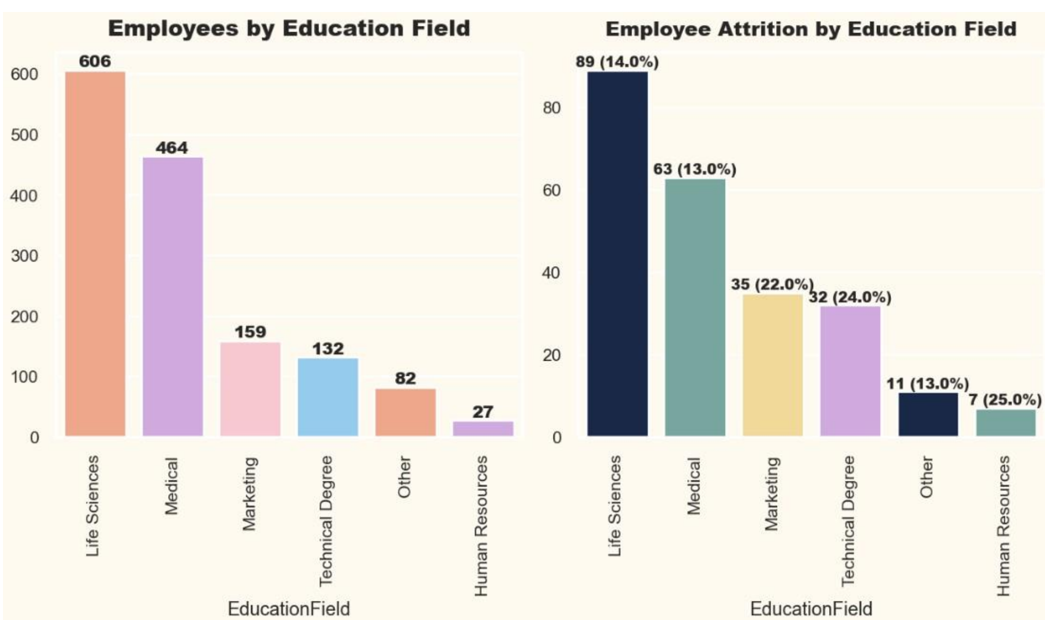

Fig.3.2.9:AnalyzingEmployeeAttritionbyEducationField.

Inference:

1. Mostoftheemployeesareeither fromLifeScienceorMedicalEducationField.
2. Veryfewemployeesare fromHumanResourcesEducationField.
3. EducationFieldslikeHumanResources,Marketing,andTechnicalishavingveryhigh attrition rate.
4. Thismaybebecauseofworkloadbecausethereareveryfewemployeesinthese education fields compared to education field with less attrition rate.

10] ANALYZINGEMPLOYEEATTRITIONBYENVIRONMENTSATISFACTION.

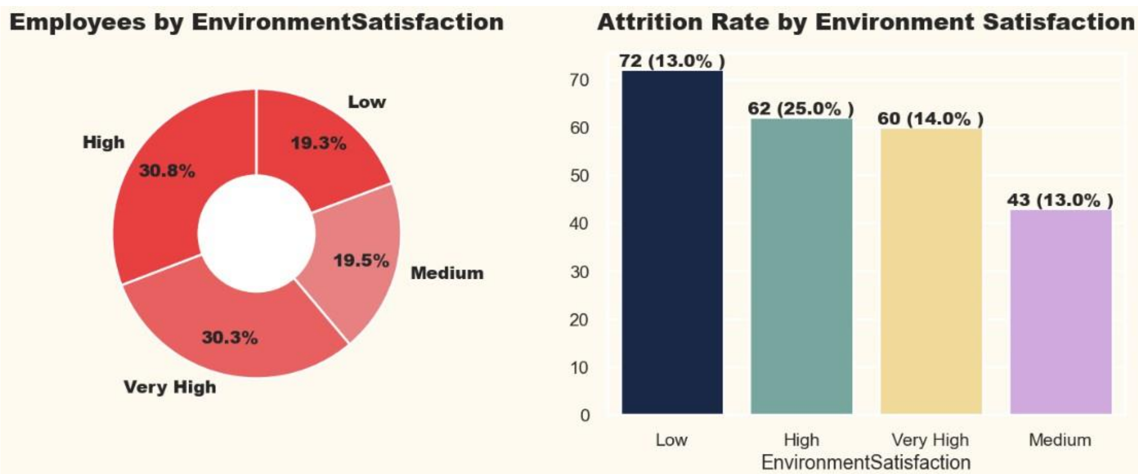

Fig.3.2.10:AnalyzingEmployeeAttritionbyEnvironmentSatisfaction.

Inference:

1. MostoftheemployeeshaveratedtheorganizationenvironmentsatisfactionHigh&Very High.
2. Thoughtheorganizationenvironmentsatisfactionishighstillthere'sveryhighattritionin this environment.
3. AttritionRateincreaseswithincreaseinlevelof environmentsatisfaction.

11] ANALYZINGEMPLOYEEATTRITIONBYJOBROLES.

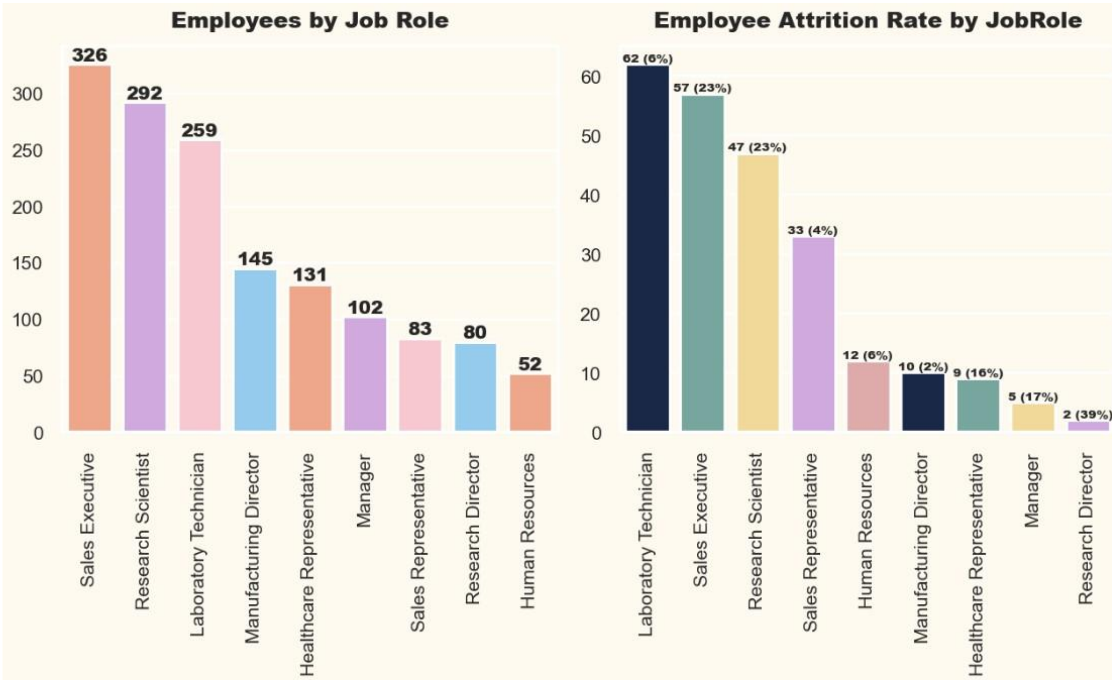

Fig.3.2.11:AnalyzingEmployeeAttritionbyJobRoles.

Inference:

1. MostemployeesareworkingasSalesexecutive,ResearchScientistorLaboratory Technician in this organization.
2. HighestattritionratesareinsectorofResearchDirector,SalesExecutive,andResearch Scientist.

12] ANALYZINGEMPLOYEEATTRITIONBYJOBLEVEL.

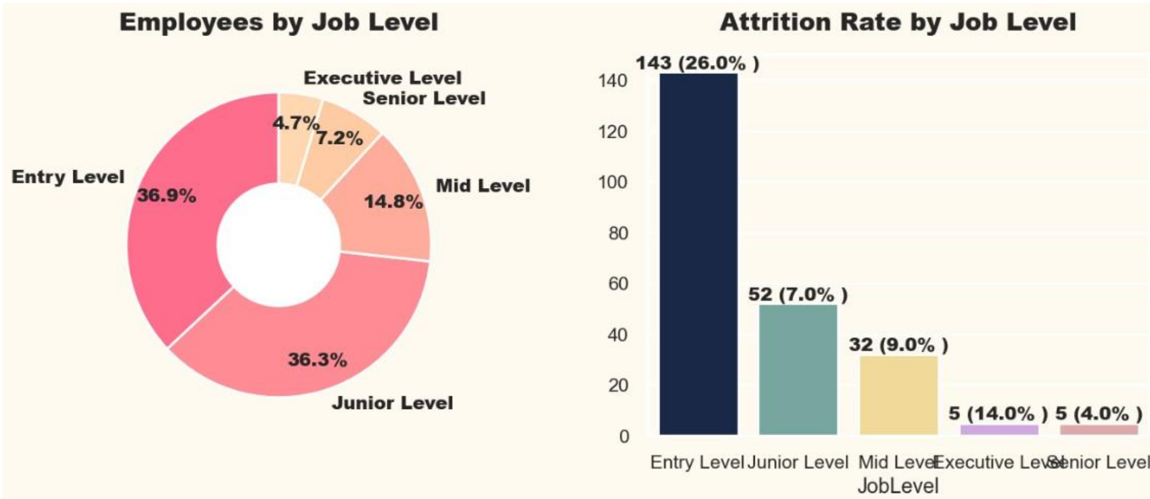

Fig.3.2.12:AnalyzingEmployeeAttrition byJobLevel.

Inference:

1. Mostoftheemployeesintheorganization areat EntryLevel orJuniorLevel.
2. HighestAttrition isattheEntryLevel.
3. Asthelevelincreasestheattritionratedecreases.

13] ANALYZINGEMPLOYEEATTRITIONBYJOBSATISFACTION.

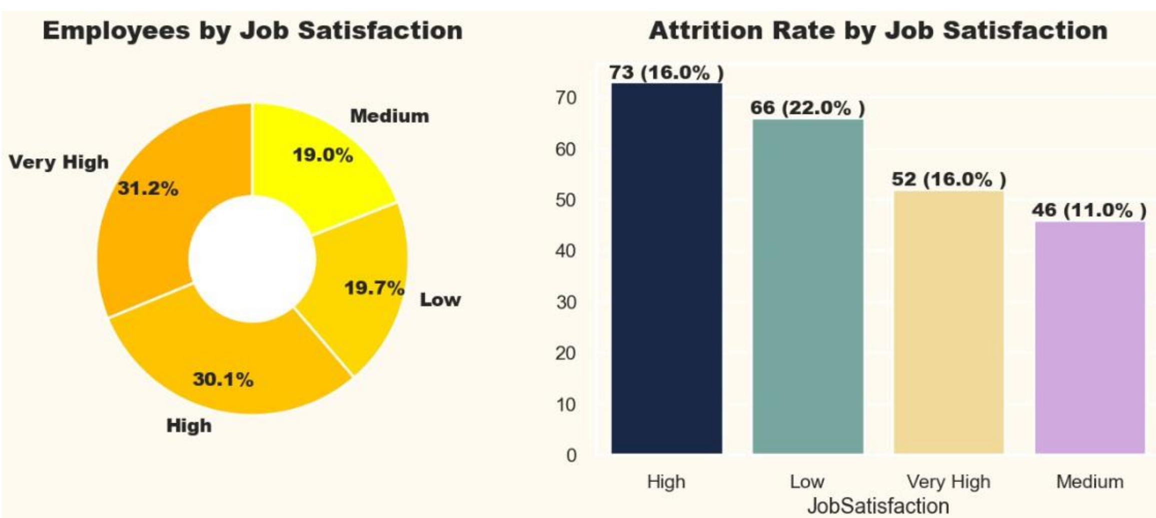

Fig.3.2.13:AnalyzingEmployeeAttritionbyJob Satisfaction.

Inference:

1. Mostoftheemployeeshaveratedtheirjobsatisfaction ashigh orveryhigh.
2. Employeeswho ratedtheir job satisfactionlow are mostlyleavingthe organization.
3. Allthecategoriesinjob satisfactionishavinghighattritionrate.

14] ANALYZINGEMPLOYEEATTRITIONBYMARTIALSTATUS.

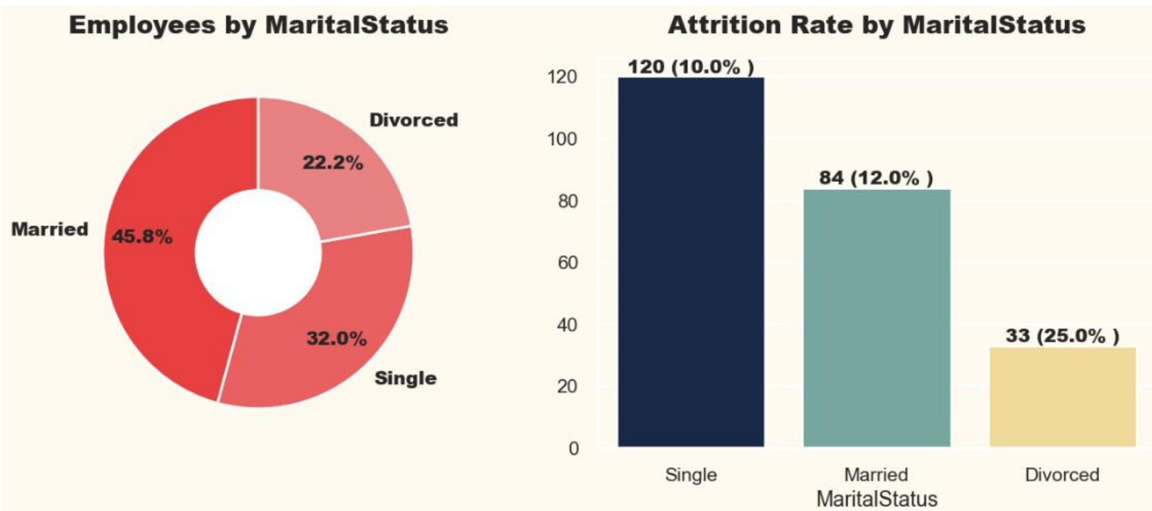

Fig.3.2.14:AnalyzingEmployeeAttritionbyMartialStatus.

Inference:

1. Mostoftheemployeesaremarriedinthe organization.
2. Theattritionrate isveryhighofemployees whoare divorced.
3. Theattritionrateis lowforemployeeswho are single.

15] ANALYZINGEMPLOYEEATTRITIONBYMONTHLYINCOME.

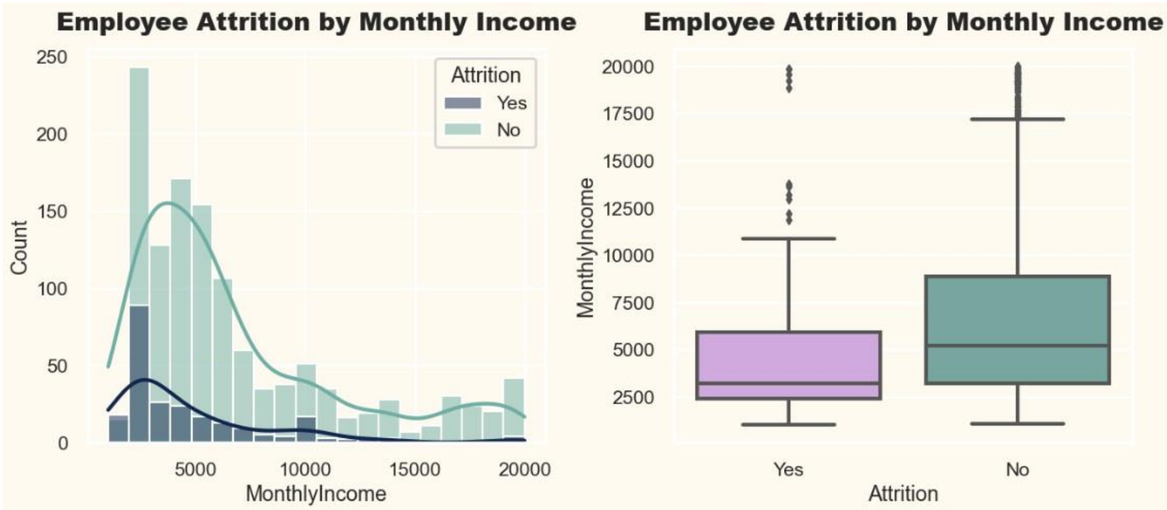

Fig.3.2.15:AnalyzingEmployeeAttritionbyMonthlyIncome.

Inference:

1. Mostof theemployeesaregettingpaidless than 10000in theorganization.
2. Theaveragemonthlyincomeofemployeewhohasleftiscomparativelylowwith employee who is still working.
3. AstheMonthlyIncomeincreasestheattrition decreases.

16] ANALYZINGEMPLOYEEATTRITIONBYWORK EXPERIENCE.

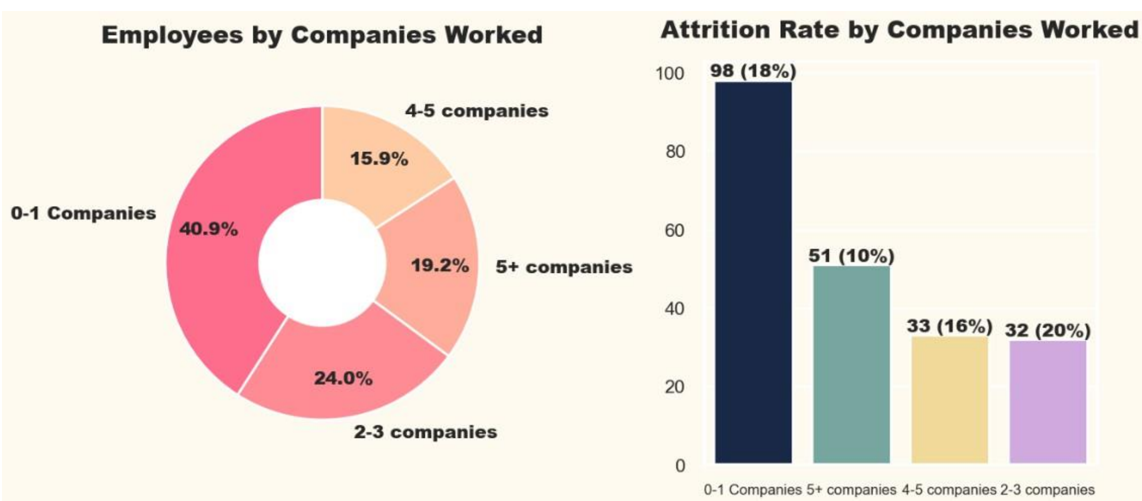

Fig.3.2.16:AnalyzingEmployeeAttritionbyWorkExperience.

Inference:

1. Mostof theemployeeshaveworked forless than2companies.
2. There'sahighattritionrate ofemployeeswhohaveforlessthan5companies.

17] ANALYZINGEMPLOYEEATTRITIONBYOVERTIME.

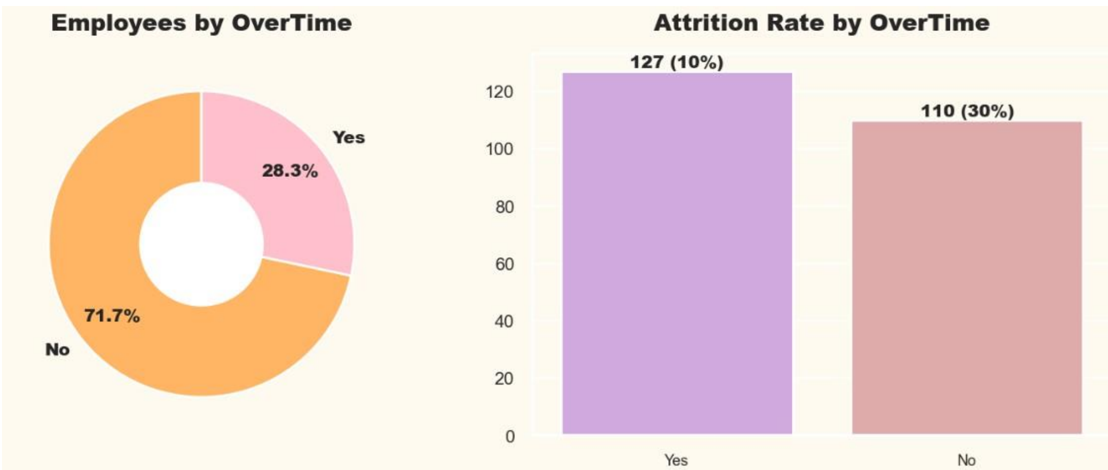

Fig.3.2.17:AnalyzingEmployeeAttritionbyOvertime.

Inference:

1. Mostoftheemployeesdon’tworkfor OverTime.
2. ThefeatureOverTimeishavingaveryhighclassimbalanceduetowhichwecan'tmake any meaningful insights.

18] ANALYZINGEMPLOYEEATTRITIONBYSALARYHIKE.

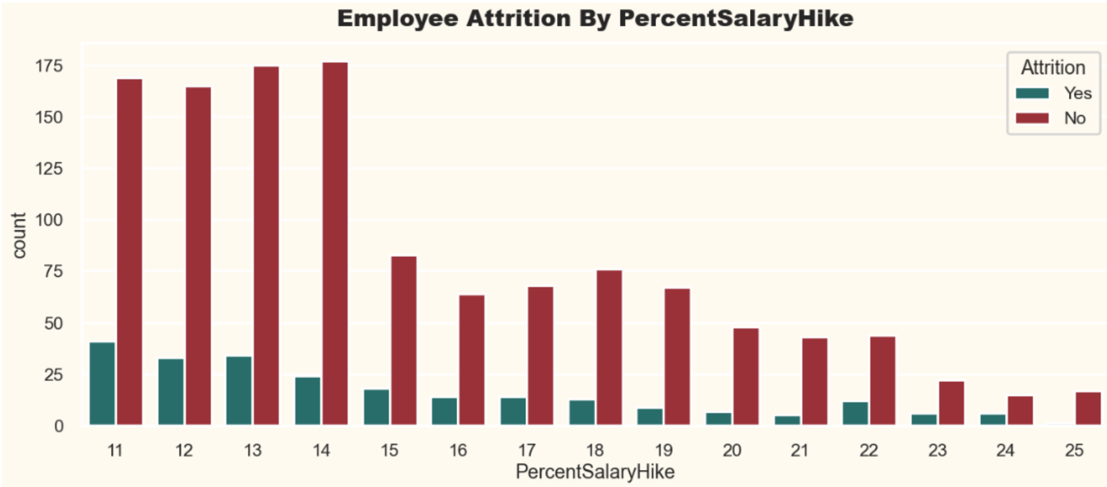

Fig.3.2.18:AnalyzingEmployeeAttritionbySalaryHike.

Inference:

1. VeryFew employeesare gettingahigh percentsalaryhike.
2. Astheamountofpercentsalaryincreasesthe attritionrate decreases.

19] ANALYZINGEMPLOYEEATTRITIONBYPERFORMANCERATING.

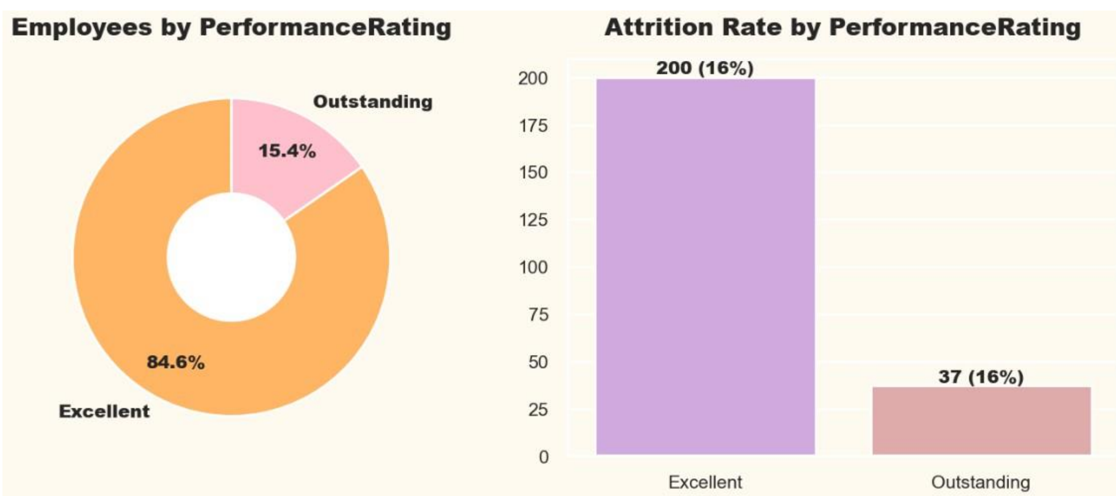

Fig.3.2.19:AnalyzingEmployeeAttritionbyPerformanceRating.

Inference:

1. Mostoftheemployeesarehavingexcellentperformancerating.
2. Boththecategories inthisfield ishavingsameattrition rate.
3. That'swhywecan'tgenerateanymeaningful insights.

20] ANALYZINGEMPLOYEEATTRITIONBYRELATIONSHIPSATISFACTION.

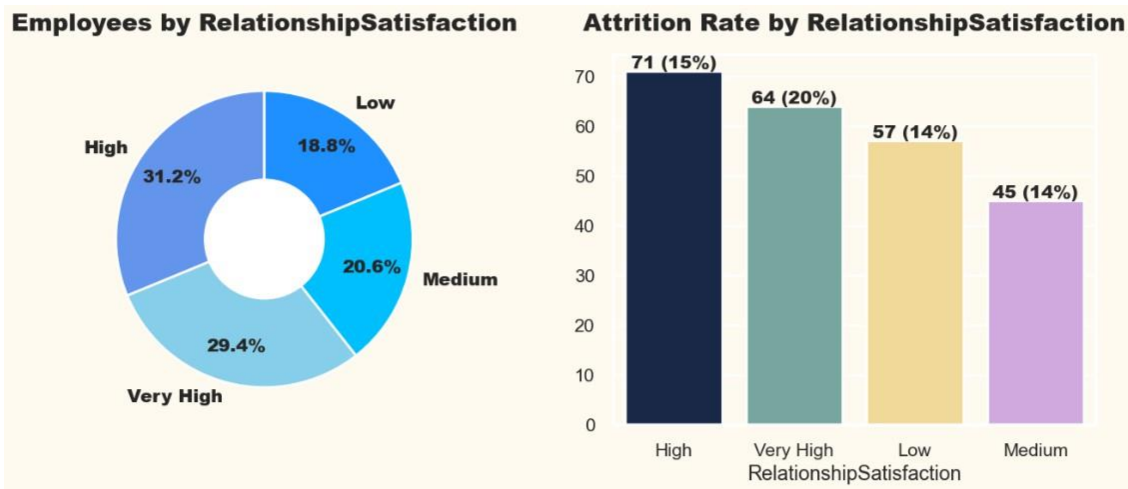

Fig.3.2.20:AnalyzingEmployeeAttritionbyRelationshipSatisfaction.

Inference:

1. Mostof theemployees arehavinghigh or veryhigh relationship satisfaction.
2. Thoughthe relationshipsatisfactionishighthere's ahigh attritionrate.
3. Allthecategoriesinthisfeaturearehavingahighattrition rate.

21] ANALYZINGEMPLOYEEATTRITIONBYWORKLIFEBALANCE.

Fig.3.2.21:AnalyzingEmployeeAttritionbyWork LifeBalance.

Inference:

1. Morethan 60%ofemployeesarehavingabetter work life balance.
2. Employeeswith BadWork Life BalancearehavingVeryHighAttrition Rate.
3. OtherCategoriesisalsohavingHighattritionRate.

22] ANALYZINGEMPLOYEEATTRITIONBYTOTALWORKINGEXPERIENCE.

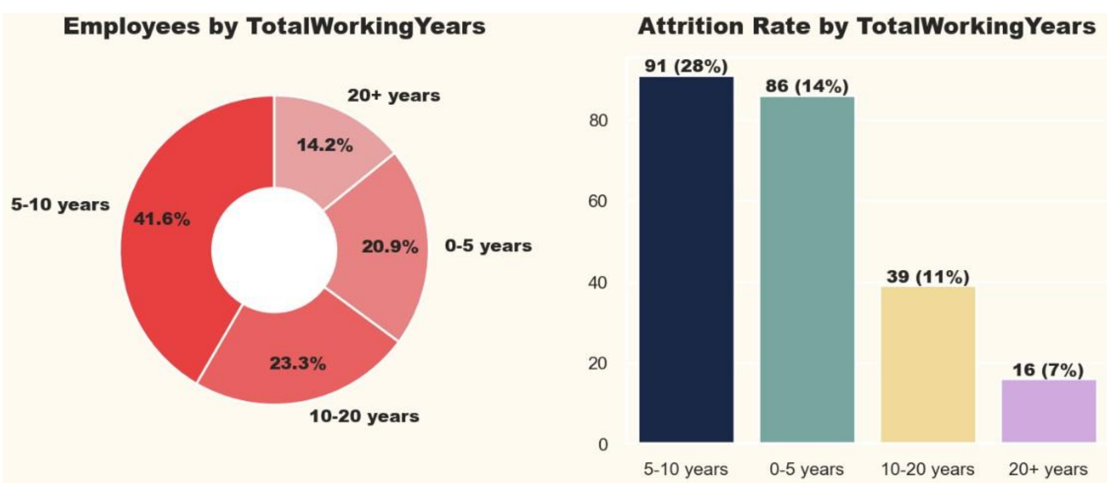

Fig.3.2.22:AnalyzingEmployeeAttritionbyTotal WorkingExperience.

Inference:

1. Mostoftheemployeesarehavingatotalof5to10yearsofworkingexperience.But their Attrition Rate is also very high.
2. Employees withworkingexperienceoflessthan10yearsarehavingHighAttritionRate.
3. Employees withworkingexperienceofmorethan 10yearsarehavingLessAttritionRate.

23] ANALYZINGEMPLOYEEATTRITIONBYYEARSATCOMPANY.

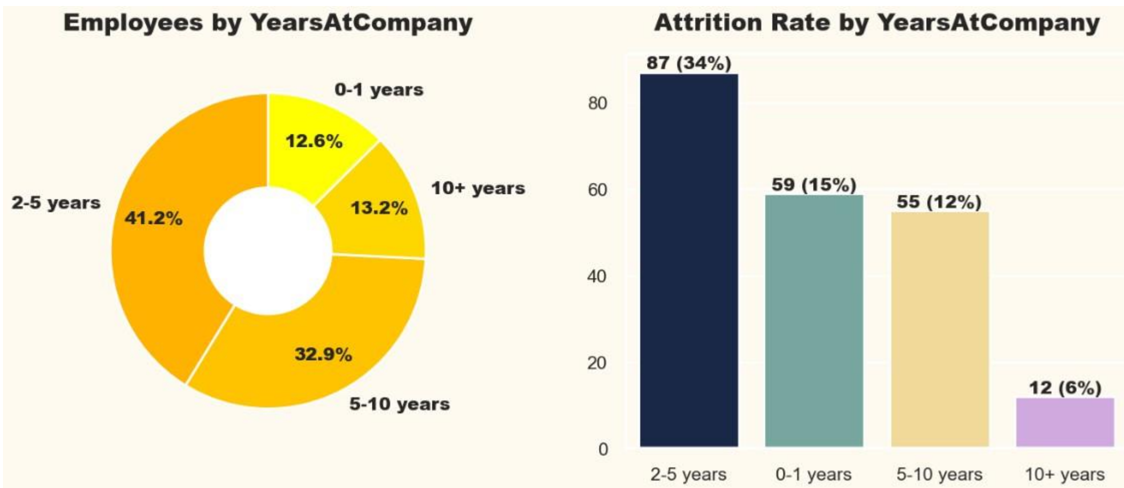

Fig.3.2.23:AnalyzingEmployee AttritionbyYearsatCompany.

Inference:

1. Mostemployeeshaveworkedfor2 to10yearsintheorganization.
2. Veryfew employees haveworkingforless than 1yearormorethan 10years.
3. Employeewhohave workedfor2-5years arehavingveryhighattrition rate.
4. Employeewhohave workedfor10+yearsarehavinglowattrition rate.

24] ANALYZINGEMPLOYEEATTRITIONBY YEARSINCURRENTROLE.

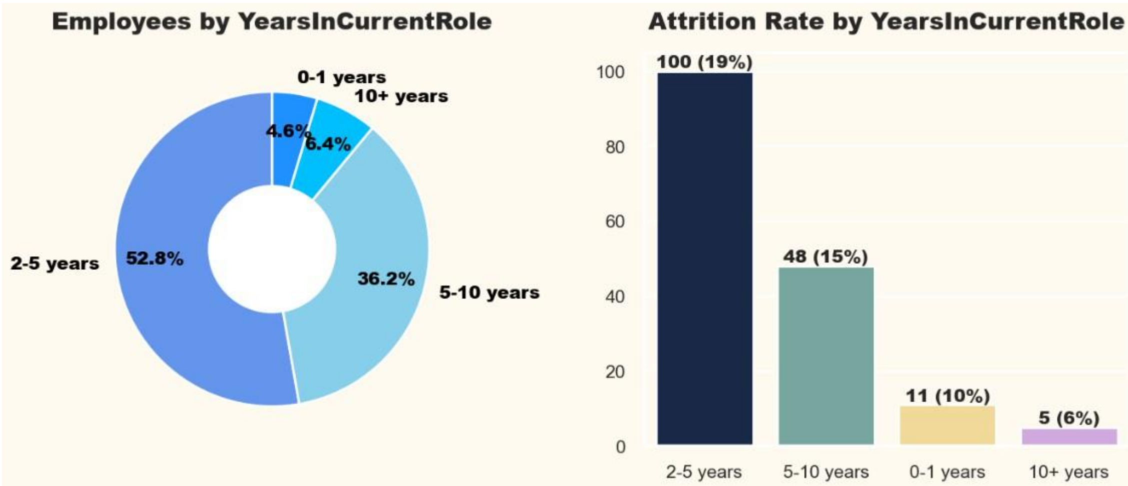

Fig.3.2.24:AnalyzingEmployeeAttritionbyYearsinCurrent Role.

Inference:

1. Mostemployeeshaveworkedfor2to10yearsforthesame roleintheorganization.
2. Veryfewemployeeshaveworkedforlessthan1yearormorethan10 yearsinthesame role.
3. Employee who hasworkedtill 2years inthe same rolearehavingveryhighattritionrate.
4. Employee whohas workedfor10+yearsin thesamerole arehavinglowattrition rate.

25] ANALYZINGEMPLOYEEATTRITIONBYYEARSSINCELASTPROMOTION.

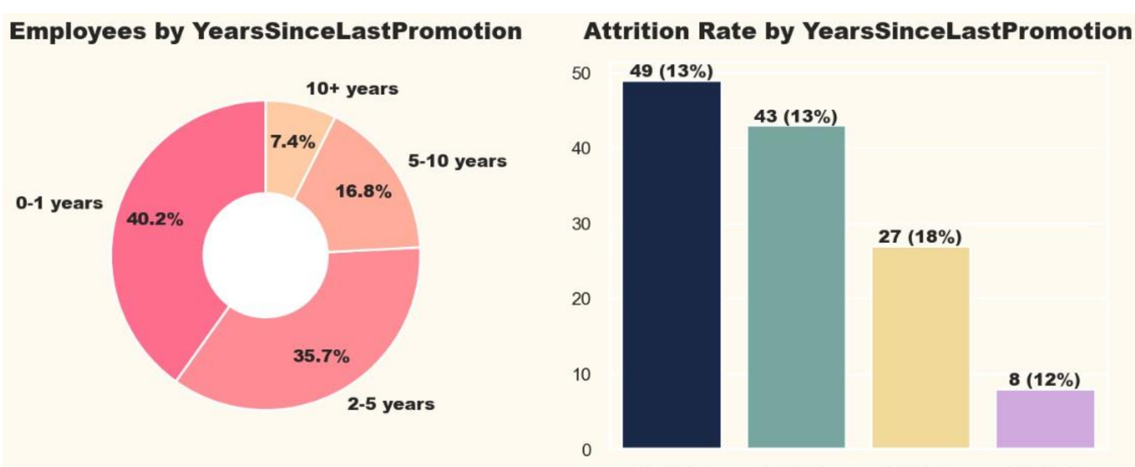

Fig.3.2.25:AnalyzingEmployeeAttritionbyYearsSinceLastPromotion.

Inference:

1. Almost36% ofemployeehas not beenpromotedsince2to 5years.
2. Almost8% ofemployeeshavenot beenpromotedsince10+years.
3. Allthecategoriesinthisfeatureishavinghighattritionratespeciallyemployeewhohas not been promoted since 5+ years.

26] ANALYZINGEMPLOYEEATTRITIONBYYEARSWITHCURRENTMANAGER.

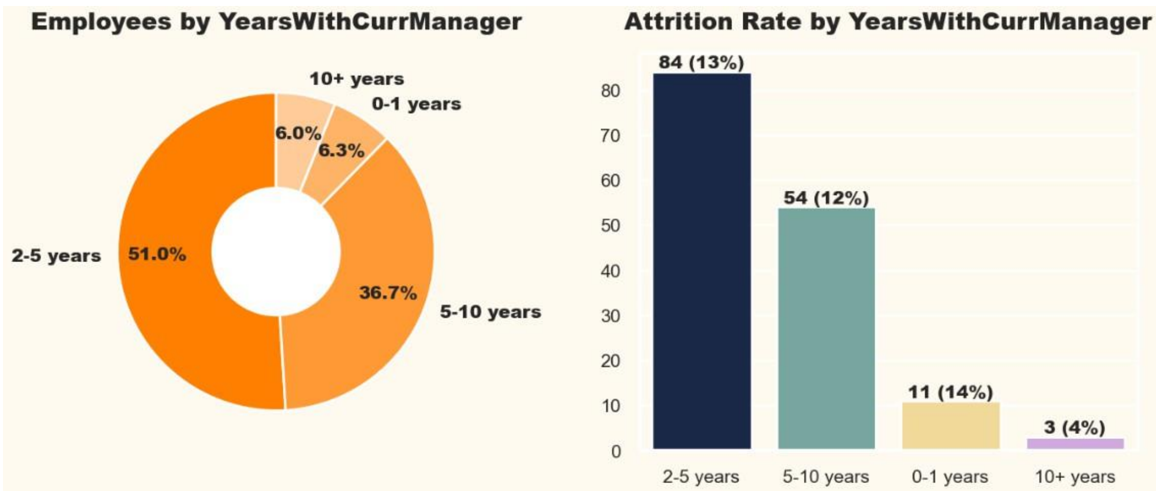

Fig.3.2.26:AnalyzingEmployeeAttritionbyYearswithCurrentManager.

Inference:

1. Almost51%ofemployeeshaveworkedfor2-5yearswiththesamemanager.
2. Almost38%ofemployeeshaveworkedfor5-10yearswiththesame manager.
3. Employeewhohasworkedfor10+ yearwiththesamemanagerarehavingverylow attrition rate.
4. OtherCategoriesishavinghighattritionrate.

STATISTICALANALYSIS

Statistical analysis plays a crucial role in HR analytics by helping organizations make informed decisions about their human resources and workforce management. It enables evidence-based decision-making,enhancesworkforceplanningstrategies,andfostersadeeperunderstandingofthe organization's human capital dynamics.

1] PerformANOVATest:ANOVAtestisusedtoanalysingtheimpactofdifferentnumericalfeatures on a response categorical feature.

Inference:

Thefollowingfeaturesshowastrongassociationwithattrition,asindicatedbytheirhigh F-scores and very low p-values.

1. Age
2. DailyRate
3. HourlyRate
4. MonthlyIncome
5. MonthlyRate
6. NumCompaniesWorked
7. PercentSalaryHike
8. TotalWorkingYears
9. TrainingTimesLastYear
10. YearsAtCompany
11. YearsWithCurrManager

Thefollowingfeaturesdon’tshowssignificantrelationshipwithattritionbecauseoftheir moderate F-scores and extremely high p-values.

1. DistanceFromHome
2. StockOptionLevel
3. YearsInCurrentRole
4. YearsSinceLastPromotion

Itisimportantfortheorganizationtopayattentiontotheidentifiedsignificantfeaturesand consider them when implementing strategies to reduce attrition rates.

2] PerformCHI-SQUARETest:CHI-SQUAREtestisusedtoanalysingtheimpactofdifferent categorical features.

Inference:

Thefollowingfeaturesshowedstatisticallysignificantassociationswithemployeeattrition:

1. Department
2. EducationField
3. EnvironmentSatisfaction
4. JobInvolvement
5. JobLevel
6. JobRole
7. JobSatisfaction
8. MaritalStatus
9. OverTime
10. WorkLifeBalance

Thefollowingfeaturesdid notshow statisticallysignificantassociationswithattrition.

1. Gender
2. Education
3. PerformanceRating
4. RelationshipSatisfaction

Itisimportantfortheorganizationtopayattentiontotheidentifiedsignificantfeaturesand consider them when implementing strategies to reduce attrition rates.

DATAMODELING

Data modeling plays a significant role in HR analytics when integrating machine learning techniques. Machine learning algorithms leverage data models to make predictions, classifications,andrecommendationsbasedonpatternsandrelationshipsfoundintheHR data.

Datasplittingtotrain andtest:

Thedatasetwassplitinto70%fortrainingand30%fortestingandwehaveconsideredAttrition as target feature.

Fittingthedifferentmachinelearningmodels:

1. LogisticRegressionModel.

```
In [17]: lr_clf = LogisticRegression(solver='liblinear', penalty='l1')
         lr_clf.fit(X_train_std, y_train)

         evaluate(lr_clf, X_train_std, X_test_std, y_train, y_test)

TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[847  16]
 [ 59 107]]
ACCURACY SCORE:
0.9271
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.934879    0.869919  0.927114     0.902399      0.924399
recall       0.981460    0.644578  0.927114     0.813019      0.927114
f1-score     0.957603    0.740484  0.927114     0.849044      0.922577
support    863.000000  166.000000  0.927114  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[351  19]
 [ 41  30]]
ACCURACY SCORE:
0.8639
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.895408   0.612245  0.863946    0.753827      0.849820
recall       0.948649   0.422535  0.863946    0.685592      0.863946
f1-score     0.921260   0.500000  0.863946    0.710630      0.853438
support    370.000000  71.000000  0.863946  441.000000    441.000000
```

Fig.3.4.1:TrainingandTestingresultsbyusingLogistic Regression Model.

2. RandomForest Model.

```
In [21]: param_grid = dict(
             n_estimators= [100, 500, 900],
             max_features= ['auto', 'sqrt'],
             max_depth= [2, 3, 5, 10, 15, None],
             min_samples_split= [2, 5, 10],
             min_samples_leaf= [1, 2, 4],
             bootstrap= [True, False]
         )

         rf_clf = RandomForestClassifier(random_state=42)
         search = GridSearchCV(rf_clf, param_grid=param_grid, scoring='roc_auc', cv=5, verbose=1, n_jobs=-1)
         search.fit(X_train, y_train)

         rf_clf = RandomForestClassifier(**search.best_params_, random_state=42)
         rf_clf.fit(X_train, y_train)
         evaluate(rf_clf, X_train, X_test, y_train, y_test)
```

```
Fitting 5 folds for each of 648 candidates, totalling 3240 fits
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [113  53]]
ACCURACY SCORE:
0.8902
CLASSIFICATION REPORT:
                   0           1  accuracy    macro avg  weighted avg
precision    0.884221    1.000000  0.890185     0.942111      0.902899
recall       1.000000    0.319277  0.890185     0.659639      0.890185
f1-score     0.938554    0.484018  0.890185     0.711286      0.865227
support    863.000000  166.000000  0.890185  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[365   5]
 [ 65   6]]
ACCURACY SCORE:
0.8413
CLASSIFICATION REPORT:
                   0          1  accuracy   macro avg  weighted avg
precision    0.848837   0.545455   0.84127    0.697146      0.799993
recall       0.986486   0.084507   0.84127    0.535497      0.841270
f1-score     0.912500   0.146341   0.84127    0.529421      0.789150
support    370.000000  71.000000   0.84127  441.000000    441.000000
```

Fig.3.4.2:Trainingand TestingresultsbyusingRandom ForestModel.

3. SupportVectorMachine Model.

```
In [26]: svm_clf = SVC(random_state=42)

         param_grid = [
             {'C': [1, 10, 100, 1000], 'kernel': ['linear']},
             {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001], 'kernel': ['rbf']}
         ]

         search = GridSearchCV(svm_clf, param_grid=param_grid, scoring='roc_auc', cv=3, refit=True, verbose=1)
         search.fit(X_train_std, y_train)

Fitting 3 folds for each of 12 candidates, totalling 36 fits

Out[26]: GridSearchCV(cv=3, estimator=SVC(random_state=42),
                      param_grid=[{'C': [1, 10, 100, 1000], 'kernel': ['linear']},
                                  {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001],
                                   'kernel': ['rbf']}],
                      scoring='roc_auc', verbose=1)

In [27]: svm_clf = SVC(**search.best_params_)
         svm_clf.fit(X_train_std, y_train)

         evaluate(svm_clf, X_train_std, X_test_std, y_train, y_test)

TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[861   2]
 [ 65 101]]
ACCURACY SCORE:
0.9349
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.929806    0.980583  0.934888     0.955194      0.937997
recall       0.997683    0.608434  0.934888     0.803058      0.934888
f1-score     0.962549    0.750929  0.934888     0.856739      0.928410
support    863.000000  166.000000  0.934888  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[360  10]
 [ 49  22]]
ACCURACY SCORE:
0.8662
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.880196   0.687500  0.866213    0.783848      0.849172
recall       0.972973   0.309859  0.866213    0.641416      0.866213
f1-score     0.924262   0.427184  0.866213    0.675723      0.844234
support    370.000000  71.000000  0.866213  441.000000    441.000000
```

Fig.3.4.3:Trainingand Testingresults byusingSupport Vector Machine Model.

4. XGBoostModel.

```
In [30]: xgb_clf = XGBClassifier()
         xgb_clf.fit(X_train, y_train)

         evaluate(xgb_clf, X_train, X_test, y_train, y_test)
```

```
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [  0 166]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision    1.0    1.0       1.0        1.0           1.0
recall       1.0    1.0       1.0        1.0           1.0
f1-score     1.0    1.0       1.0        1.0           1.0
support    863.0  166.0       1.0     1029.0        1029.0
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[356  14]
 [ 51  20]]
ACCURACY SCORE:
0.8526
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.874693   0.588235  0.852608    0.731464      0.828574
recall       0.962162   0.281690  0.852608    0.621926      0.852608
f1-score     0.916345   0.380952  0.852608    0.648649      0.830148
support    370.000000  71.000000  0.852608  441.000000    441.000000
```

Fig.3.4.4: Trainingand Testingresults byusingXGBoostModel.

5. LightGBMModel.

```
In [34]: lgb_clf = LGBMClassifier()
         lgb_clf.fit(X_train, y_train)

         evaluate(lgb_clf, X_train, X_test, y_train, y_test)
```

```
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Info] Number of positive: 166, number of negative: 863
[LightGBM] [Warning] Auto-choosing row-wise multi-threading, the overhead of testing was 0.000203 seconds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 1176
[LightGBM] [Info] Number of data points in the train set: 1029, number of used features: 108
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.161322 -> initscore=-1.648427
[LightGBM] [Info] Start training from score -1.648427
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [  0 166]]
ACCURACY SCORE:
1.0000
CLASSIFICATION REPORT:
               0      1  accuracy  macro avg  weighted avg
precision    1.0    1.0       1.0        1.0           1.0
recall       1.0    1.0       1.0        1.0           1.0
f1-score     1.0    1.0       1.0        1.0           1.0
support    863.0  166.0       1.0     1029.0        1029.0
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[353  17]
 [ 54  17]]
ACCURACY SCORE:
0.8390
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.867322   0.500000  0.839002    0.683661      0.808184
recall       0.954054   0.239437  0.839002    0.596745      0.839002
f1-score     0.908623   0.323810  0.839002    0.616216      0.814469
support    370.000000  71.000000  0.839002  441.000000    441.000000
```

Fig.3.4.5:Trainingand TestingresultsbyusingLightGBM Model.

6. CatBoostModel.

```
In [37]: cb_clf = CatBoostClassifier()
         cb_clf.fit(X_train, y_train, verbose=0)

         evaluate(cb_clf, X_train, X_test, y_train, y_test)

TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[863   0]
 [ 16 150]]
ACCURACY SCORE:
0.9845
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.981797    1.000000  0.984451     0.990899      0.984734
recall       1.000000    0.903614  0.984451     0.951807      0.984451
f1-score     0.990815    0.949367  0.984451     0.970091      0.984129
support    863.000000  166.000000  0.984451  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[363   7]
 [ 59  12]]
ACCURACY SCORE:
0.8503
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.860190   0.631579   0.85034    0.745884      0.823384
recall       0.981081   0.169014   0.85034    0.575048      0.850340
f1-score     0.916667   0.266667   0.85034    0.591667      0.812018
support    370.000000  71.000000   0.85034  441.000000    441.000000
```

Fig.3.4.6:Trainingand Testingresults byusingCatBoost Model.

7. AdaBoostModel.

```
In [40]: ab_clf = AdaBoostClassifier()
         ab_clf.fit(X_train, y_train)

         evaluate(ab_clf, X_train, X_test, y_train, y_test)

TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[846  17]
 [ 78  88]]
ACCURACY SCORE:
0.9077
CLASSIFICATION REPORT:
                    0           1  accuracy    macro avg  weighted avg
precision    0.915584    0.838095  0.907677     0.876840      0.903084
recall       0.980301    0.530120  0.907677     0.755211      0.907677
f1-score     0.946838    0.649446  0.907677     0.798142      0.898863
support    863.000000  166.000000  0.907677  1029.000000   1029.000000
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[346  24]
 [ 50  21]]
ACCURACY SCORE:
0.8322
CLASSIFICATION REPORT:
                    0          1  accuracy   macro avg  weighted avg
precision    0.873737   0.466667    0.8322    0.670202      0.808200
recall       0.935135   0.295775    0.8322    0.615455      0.832200
f1-score     0.903394   0.362069    0.8322    0.632732      0.816242
support    370.000000  71.000000    0.8322  441.000000    441.000000
```

Fig.3.4.7:TrainingandTestingresults byusingAdaBoostModel.

ROCCurve:AnROCcurve(receiveroperatingcharacteristiccurve)is agraphshowingtheperformance of a classification model at all classification thresholds.

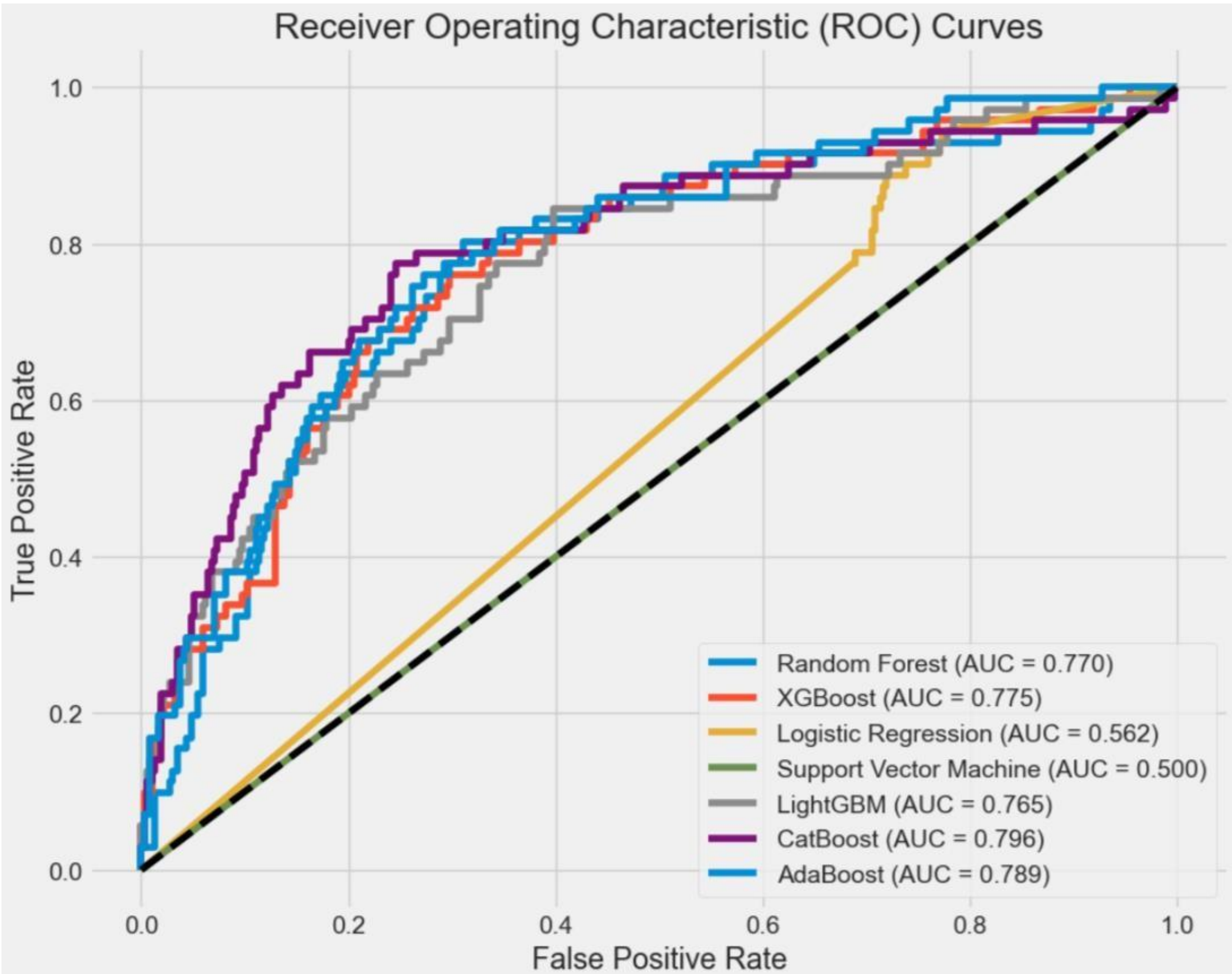

Fig.3.4.8:ROCCurve Diagram.

The graph is well-structured and displays multiple lines of varying colors, each representing a different machine learning model. The models include Random Forest, XGBoost, Logistic Regression, Support Vector Machine, LightGBM, CatBoost, and AdaBoost. Each model is also associated with an Area Under Curve (AUC) value, indicating their performance. The graph presents the True Positive Rate on the y-axis, ranging from 0.0 to 1.0, and the False Positive Rate on the x-axis, also ranging from 0.0 to 1.0. Here, Model like Random Forest, XGBoost, LightGBM, CatBoost, AdaBoost have better performance comparing with Support Vector Machine and Logistic Regression.

CHAPTER4:PROJECTPLANNING

PROJECTPLANNING

Sr. No.	Taskstobecompleted	StartDate	EndDate
1.	ExploringtheDomain	15 August	20 August
2.	FinalizingTopicand Domain	21 August	25 August
3.	DesigningMethodology	26August	30 August
4.	DataExploration& Processing	01 September	10September
5.	Data Visualization	11September	30September
6.	StatisticalAnalysis	01 October	10 October
7.	Data Modeling	11 October	25 October
8.	ReportMaking &PPT Making	26 October	20 November

Table4.1Project’sProgressPlanning

CHAPTER 5

CONCLUSION

In conclusion, we embarked on a comprehensive analysis of the IBM HR Analytics Attrition Dataset, from data loading to model evaluation. By implementing and evaluating various machine learning algorithms, we gained insights into which models are effective for predictingemployeeattrition.Theresults andvisualizationsgeneratedthroughouttheprocess provide valuable information for decision-makers and HR professionals seeking tounderstand and mitigate employee attrition within the organization. This project showcases the power of data analysis and machine learning in addressing real-world businesschallenges.

FUTUREWORK

In the context of the previous HR analytics project on employee attrition, future work in sentiment analysis involves implementing sentiment analysis on employee feedback data to gain insights, monitoring sentiment in real-time, categorizing sentiments by topics, and analyzing historical sentiment trends. In terms of dashboard development, there's a need to create interactive, predictive, and benchmarking-enabled dashboards with custom alerts, engagement metrics, and mobile accessibility. Additionally, user training and support, data privacy, feedback integration, and performance monitoring are crucial aspects to ensure the dashboard's effectiveness in facilitating data-driven HR decisions and actions while adhering to privacy regulations.

REFERENCES

[1] Mishra S N, Lama D R and Pal Y 2016 Human Resource Predictive Analytics (HRPA) for HRManagement in Organizations International Journal Of Scientific & Technology Research 5(5) 33-35

[2] HoffmanMandTadelisS2018PeopleManagementSkills,EmployeeAttrition,andManager Rewards: An Empirical Analysis National Bureau of Economic Research

[3] Frye A, Boomhower C, SmithM, Vitovsky L and Fabricant S 2018 Employee Attrition: WhatMakes an Employee Quit? MU Data Science Review 1(1)

[4] S. Rabiyathul Basariya, Ramyar Rzgar Ahmed, A STUDY ON ATTRITION - TURNOVER INTENTIONS OF EMPLOYEES, International Journal of Civil Engineering and Technology (IJCIET), 2019, 10(1), PP2594-2601

[5] Halkos, George & Bousinakis, Dimitrios, 2017. "The effect of stress and dissatisfaction on employees during crisis," Economic Analysis and Policy, Elsevier, vol. 55(C), pages 25-34.

[6] Glavas, A., & Willness, C. (2020). Employee (dis)engagement in corporate social responsibility. In D. Haski-Leventhal, L. Roza, & S. Brammer (Eds.), Employee engagement in corporate social responsibility (pp. 10–27). Sage Publications Ltd. https://doi.org/10.4135/9781529739176.n2

[7] S. Yadav, A. Jain and D. Singh, "Early Prediction of Employee Attrition using Data Mining Techniques," 2018 IEEE 8th International Advance Computing Conference (IACC), Greater Noida, India, 2018, pp. 349-354, doi: 10.1109/IADCC.2018.8692137.

[8] R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost MachineLearningApproach,"2018 InternationalConferenceonSystemModeling&Advancementin Research Trends (SMART), Moradabad, India, 2018, pp. 113-120, doi: 10.1109/SYSMART.2018.8746940.

[9] lduayj, Sarah & Rajpoot, Kashif. (2018). Predicting Employee Attrition using Machine Learning. 93-98. 10.1109/INNOVATIONS.2018.8605976.

[10] Setiawan, Irwan & Suprihanto, Suprihanto & Nugraha, Ade & Hutahaean, Jonner. (2020). HR analytics: Employee attrition analysis using logistic regression. IOP Conference Series: Materials Science and Engineering. 830. 032001. 10.1088/1757-899X/830/3/032001.

[11] Yadav, Sandeep & Jain, Aman & Singh, Deepti. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. 349-354. 10.1109/IADCC.2018.8692137.

[12] I. Ballal,S.Kavathekar, S.Janwe,P.Shete,andN.Bhirud,"People LeavingtheJob-An Approach for Prediction Using Machine Learning," Int. J. Res. Anal. Rev., vol. 7, no. 1, pp. 8– 10, 2020, [Online]. Available: www.ijrar.org

[13] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H. S. Alghamdi, "Prediction ofEmployeeAttritionUsingMachineLearningandEnsembleMethods,"Int.J.Mach.Learn. Comput., vol. 11, no. 2, pp. 110–114, 2021, doi: 10.18178/ijmlc.2021.11.2.1022.

[14] N. Mansor, N. S. Sani, and M. Aliff, "Machine Learning for Predicting Employee Attrition,"Int.J.Adv.Comput.Sci.Appl.,vol.12,no.11,pp. 435–445,2021,doi:10.14569/ IJACSA.2021.0121149.

[15] D. Saisanthiya, V. M. Gayathri, and P. Supraja, "Employee Attrition Prediction Using Machine Learning and Sentiment Analysis," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9,no. 5, pp. 7550–7557, 2020, doi: 10.30534/ijatcse/2020/91952020.