Assignment Aidwise:

About Me:

Hi, I'm Parth Behl, an AI/ML enthusiast with a passion for leveraging technology to solve real-world problems. Currently pursuing my Bachelor's degree in Computer Science from Jaypee Institute of Information Technology, Noida, U.P., I am deeply fascinated by the potential of artificial intelligence and machine learning to revolutionize various industries.

My journey into the world of AI/ML began with a curiosity to understand how machines can be taught to learn and adapt, leading me to explore various concepts and algorithms in this field. I have engaged in hands-on projects, participated in online courses, and collaborated with peers to deepen my understanding and hone my skills.

I am excited about the opportunity to contribute my skills and expertise to meaningful projects and eager to continue learning and growing in the field of artificial intelligence and machine learning.

Feel free to connect with me on
LinkedIn-https://www.linkedin.com/in/parth-behl-080048227/
GitHub-https://github.com/ParthBehl1
to explore potential collaborations or discuss anything related to AI/ML!

Problem Statement: Predicting Employee Attrition [github link]

The objective of this project is to develop machine learning models to predict employee attrition within an organization. Employee attrition, or the rate at which employees leave a company, can have significant implications for organizational performance, productivity, and morale. By accurately predicting which employees are likely to leave, organizations can take proactive measures to retain talent and mitigate the negative impact of attrition.

Key Tasks:

1. **Data Wrangling:** Handle missing values, encode categorical variables, and standardize numerical features to prepare the dataset for modeling.
2. **Exploratory Data Analysis (EDA)**: Analyze the dataset to identify patterns, trends, and factors influencing employee attrition.
3. **Model Development:** Utilize various machine learning algorithms, such as Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Naive Bayes, to build predictive models for employee attrition.
4. **Model Evaluation:** Evaluate the performance of each model using appropriate metrics such as accuracy, precision, recall, and F1 score.
5. **Prediction and Recommendations:** Deploy the best-performing model to predict employee attrition and provide actionable recommendations for retention strategies based on model insights.

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

Introduction:

This submission aims to address the task of analyzing employee attrition using machine learning techniques. The dataset employed for this analysis is the "Employee Attrition Dataset," which contains various features related to employees' demographics, job roles, satisfaction levels, and other factors, along with a binary target variable indicating whether an employee has left the organization or not.

The primary objective of this assignment is to explore the dataset, identify factors contributing to employee attrition, preprocess the data, develop machine learning models to predict attrition, and evaluate the performance of these models. Key tasks performed include data preprocessing to handle missing values and encode categorical variables, exploratory data analysis (EDA) to uncover insights from the data, and model development using various algorithms such as Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and Naive Bayes.

Through this analysis, we aim to gain a deeper understanding of the factors influencing employee attrition and to build predictive models that can assist organizations in identifying at-risk employees and implementing targeted retention strategies.

Steps used to approach the problem:

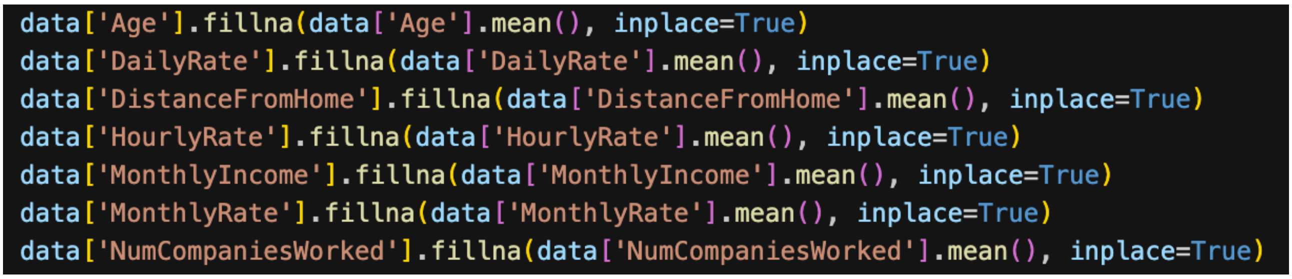
Data Wrangling:

1. Handling Missing Values:

- Missing values are identified using the isnull().sum() function, which calculates the sum of missing values for each column.

- Numerical missing values are replaced with the mean of their respective columns, while categorical missing values are filled with the mode.

- This ensures that the dataset is complete and ready for further analysis and modeling.



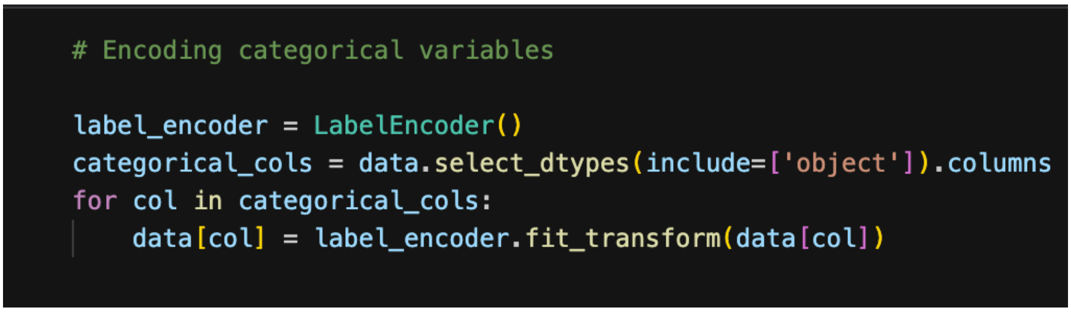
```
data['Age'].fillna(data['Age'].mean(), inplace=True)
data['DailyRate'].fillna(data['DailyRate'].mean(), inplace=True)
data['DistanceFromHome'].fillna(data['DistanceFromHome'].mean(), inplace=True)
data['HourlyRate'].fillna(data['HourlyRate'].mean(), inplace=True)
data['MonthlyIncome'].fillna(data['MonthlyIncome'].mean(), inplace=True)
data['MonthlyRate'].fillna(data['MonthlyRate'].mean(), inplace=True)
data['NumCompaniesWorked'].fillna(data['NumCompaniesWorked'].mean(), inplace=True)
```

Fig:1 Example of handling missing values from the code

2. Encoding Categorical Variables:

- Categorical variables are encoded using label encoding, which assigns a unique integer to each category within a column.

- This transformation is necessary for machine learning algorithms to interpret categorical data as numerical input.



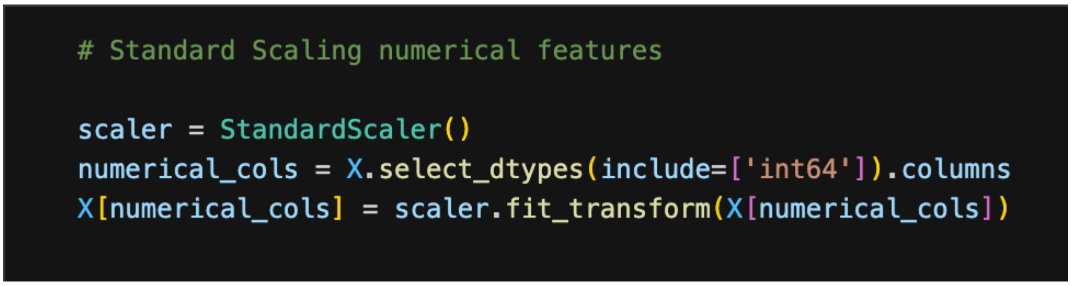
```
# Encoding categorical variables

label_encoder = LabelEncoder()
categorical_cols = data.select_dtypes(include=['object']).columns
for col in categorical_cols:
    data[col] = label_encoder.fit_transform(data[col])
```

Fig:2 Encoding categorical data part of the code

3. Standard Scaling Numerical Features:

- Numerical features are standardized using standard scaling, also known as z-score normalization.

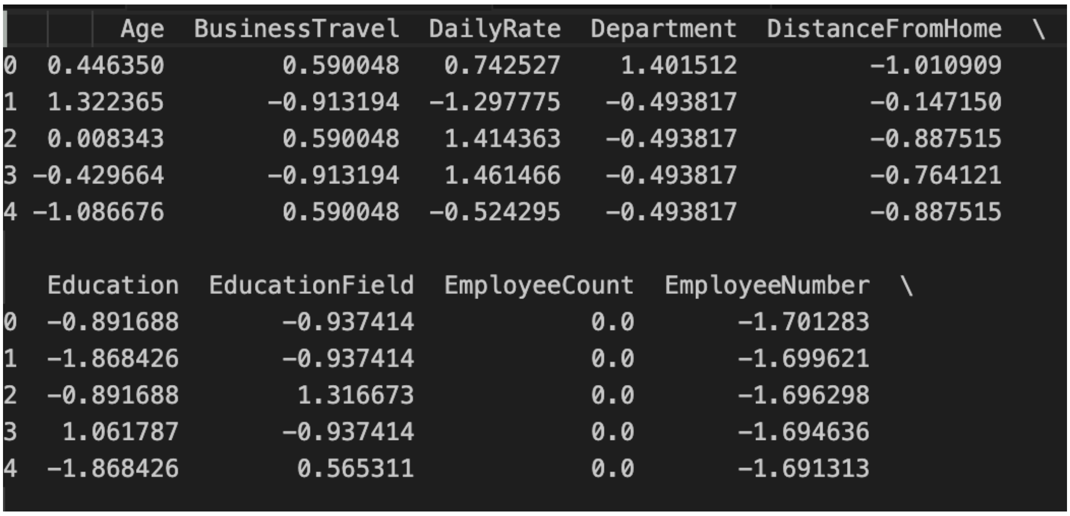
- This process transforms the data such that it has a mean of 0 and a standard deviation of 1, which helps in improving the performance of certain machine learning algorithms.



```
# Standard Scaling numerical features

scaler = StandardScaler()
numerical_cols = X.select_dtypes(include=['int64']).columns
X[numerical_cols] = scaler.fit_transform(X[numerical_cols])
```

Fig:3 Standard scaling numerical features part of the code

Here, is an example of how data looks after data preprocessing:



```
        Age  BusinessTravel  DailyRate  Department  DistanceFromHome  \
0  0.446350        0.590048   0.742527    1.401512         -1.010909
1  1.322365       -0.913194  -1.297775   -0.493817         -0.147150
2  0.008343        0.590048   1.414363   -0.493817         -0.887515
3 -0.429664       -0.913194   1.461466   -0.493817         -0.764121
4 -1.086676        0.590048  -0.524295   -0.493817         -0.887515

   Education  EducationField  EmployeeCount  EmployeeNumber  \
0  -0.891688       -0.937414            0.0       -1.701283
1  -1.868426       -0.937414            0.0       -1.699621
2  -0.891688        1.316673            0.0       -1.696298
3   1.061787       -0.937414            0.0       -1.694636
4  -1.868426        0.565311            0.0       -1.691313
```

Fig:4 Preprocessed data example

Exploratory Data Analysis (EDA):

The exploratory data analysis (EDA) performed on the dataset aims to uncover insights and patterns that may help understand the factors contributing to employee attrition. Various plots generated using libraries like Plotly, Matplotlib, and Seaborn provide visual representations of relationships between different variables and attrition.

1. Monthly Income vs. Attrition:

- **Plot**: A line plot using Plotly was generated to visualize the relationship between monthly income and attrition.

- **Insight**: The plot reveals that the attrition rate is high at lower income levels. As monthly income increases, the likelihood of an employee leaving the organization decreases. This suggests that compensation may play a significant role in employee retention.[Fig:5]

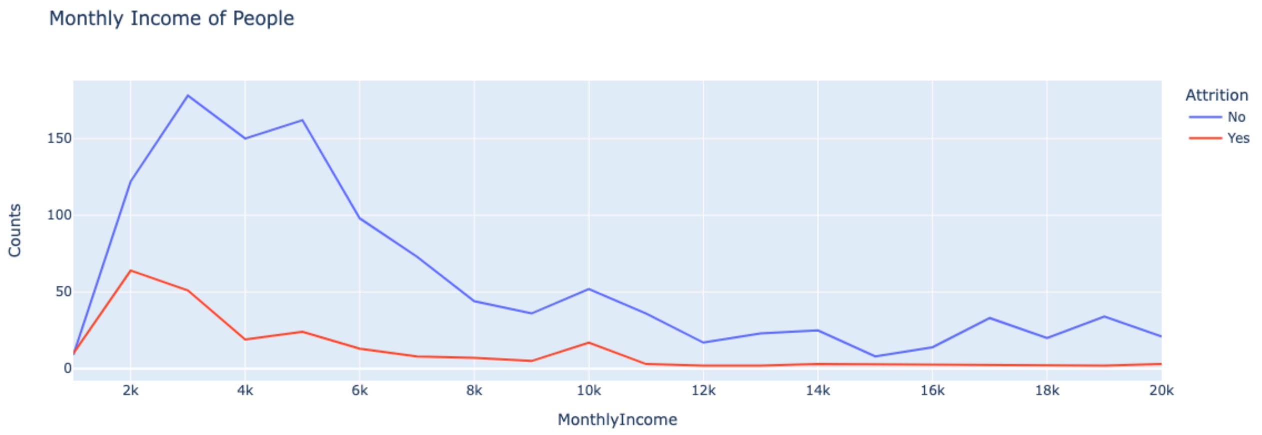


Fig:5 Monthly income and attrition plot

2. Age vs. Attrition:

- **Plot**: A line plot using Plotly was created to explore the relationship between age and attrition.
- **Insight**: The plot indicates that attrition is highest within certain age groups, particularly between the ages of 28-31. Additionally, younger employees (18-20 years) show a higher tendency to leave the organization compared to other age groups. [Fig:6]

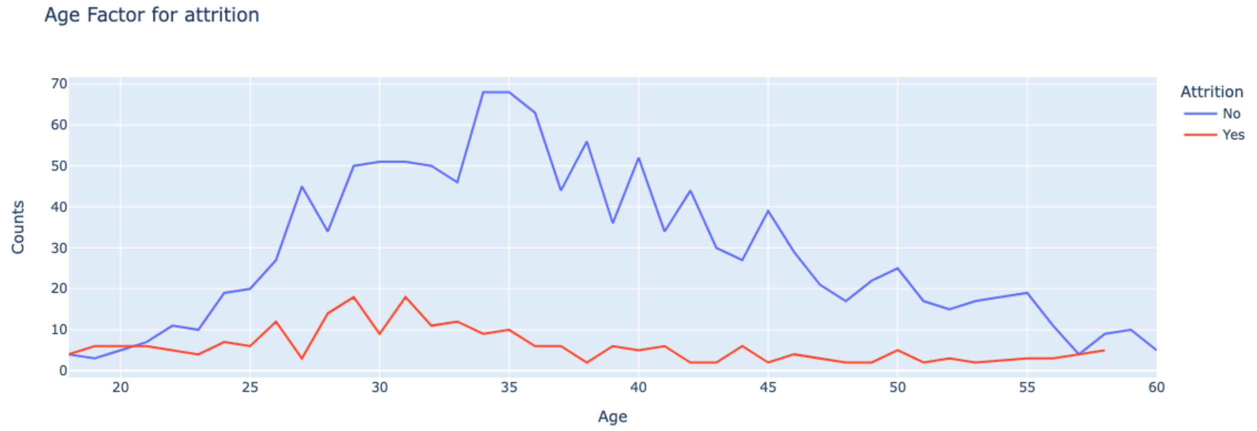


Fig:6 Employee Age vs Attrition plot

3. Years in Current Role vs. Attrition:

- **Plot**: Another line plot using Plotly was used to visualize the relationship between years in the current role and attrition.

- **Insight**: The plot suggests that employees are more likely to leave the organization within the initial years of their current role. However, as employees spend more time in their roles, the attrition rate tends to decrease, indicating that job stability may influence attrition.[Fig.7]

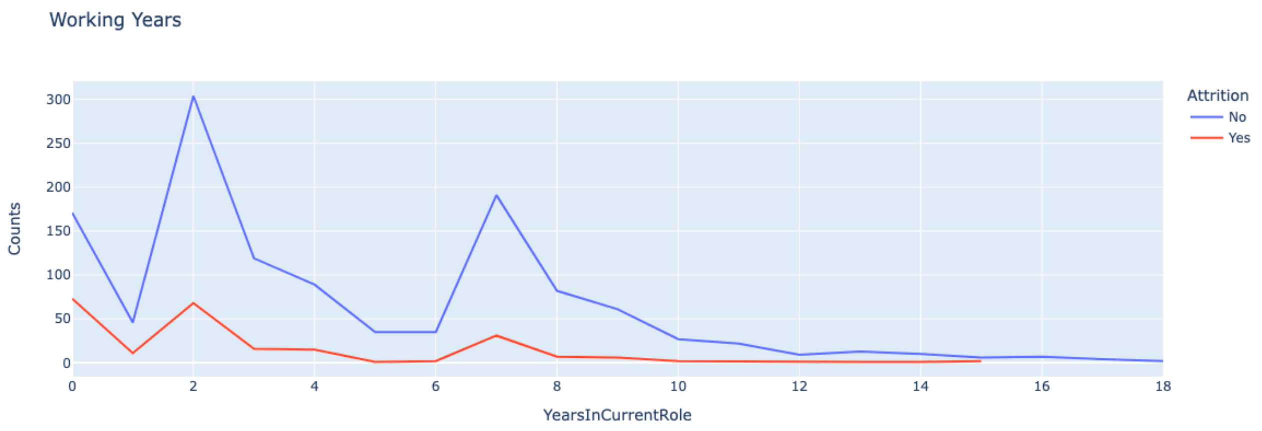


Fig:7 Employee years in a company vs attrition

4. Percent Salary Hike vs. Attrition:

- **Plot**: A line plot using Plotly illustrates the relationship between percent salary hike and attrition.
- **Insight**: The plot indicates that employees are more inclined to leave the organization when the percent salary hike is lower. Conversely, companies offering higher salary hikes tend to have lower attrition rates, suggesting that competitive compensation packages contribute to employee retention.[Fig. 7]

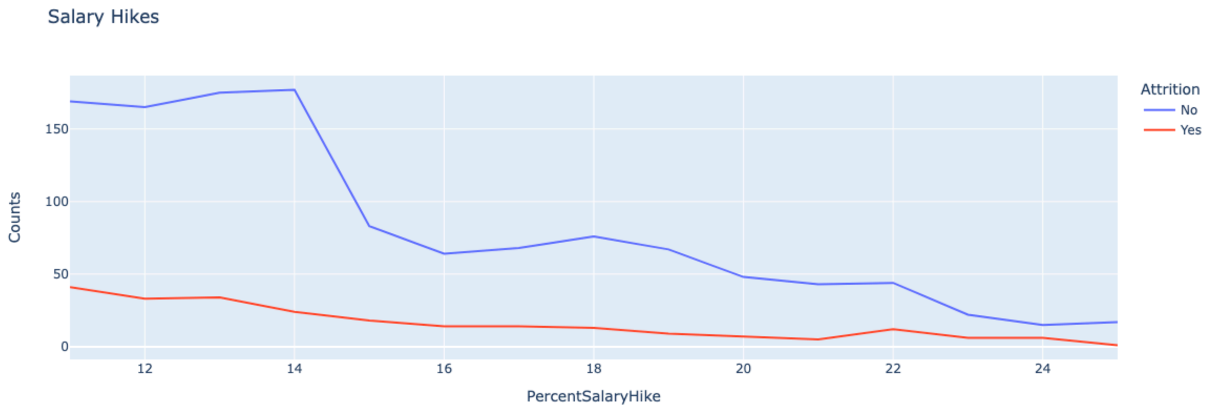


Fig:7 Salary Hikes of employee vs at the attrition rate

5. Department-wise Attrition:

- **Plot**: Pie charts were generated using Plotly to illustrate the distribution of attrition ('Yes' and 'No') within each department.[Fig.8]

- Percentage values were included on the pie charts to indicate the proportion of employees within each department who have left or stayed.

- **Insight**: 1) human resources and sales have almost same attrition rate,these departments show a relatively higher attrition rate.

  2) R&D department has least attrition rate among all.

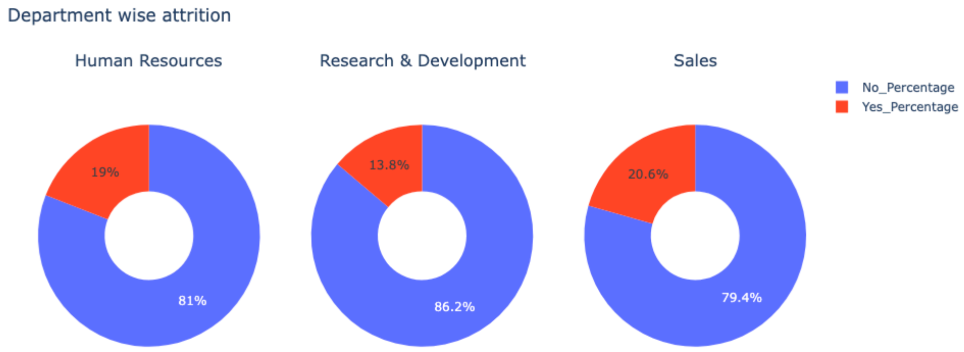


Fig:8 Pie chart representing the department vs attrition of top 3 departments with most number of employees

6. Correlation Heatmap:

- **Path**: The heatmap is a grid of squares, where each square represents the correlation coefficient between two features[Fig.10]. The color intensity of each square indicates the strength and direction of the correlation: darker colors represent stronger correlations (positive or negative), while lighter colors represent weaker or no correlations.

- **Insight**: **1) Positive Correlations:** Positive correlations (correlation coefficient close to 1) between features indicate that they tend to increase or decrease together. For example, we may observe positive correlations between job level and monthly income, indicating that higher job levels are associated with higher incomes.

2) Negative Correlations: Negative correlations (correlation coefficient close to -1) between features suggest that as one feature increases, the other tends to decrease, and vice versa.

For instance, there may be negative correlations between distance from home and job satisfaction, implying that longer commuting distances are associated with lower job satisfaction.

3) No Correlation: Features with correlation coefficients close to 0 indicate no linear relationship between them.

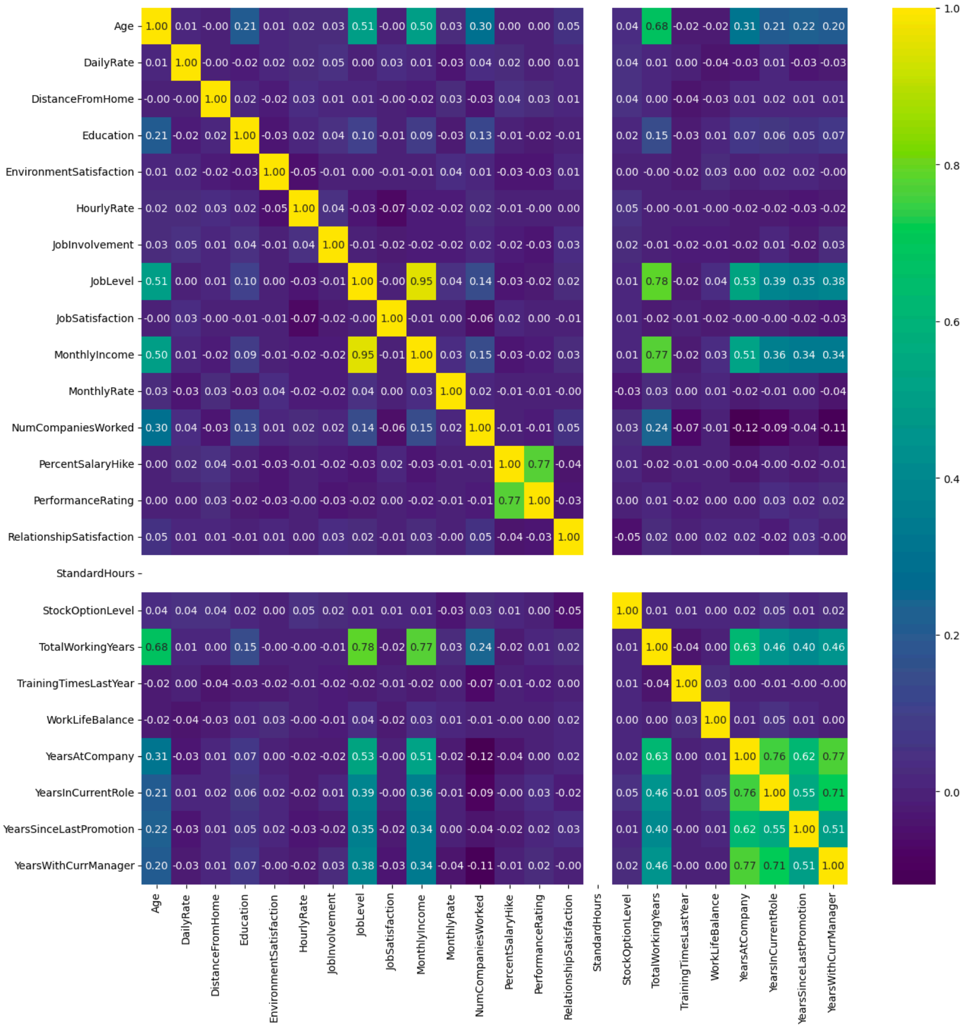


Fig:10 Correlation Heatmap using seaborn

Model Development :

The model development process involves several key steps, including data splitting, hyperparameter tuning using GridSearchCV, and evaluating various machine learning algorithms to identify the best model for predicting employee attrition.

1. **Data Splitting:**

- The dataset is divided into training and testing sets to train and evaluate the models, respectively.

- Typically, around 80% of the data is allocated for training and the remaining 20% for testing.

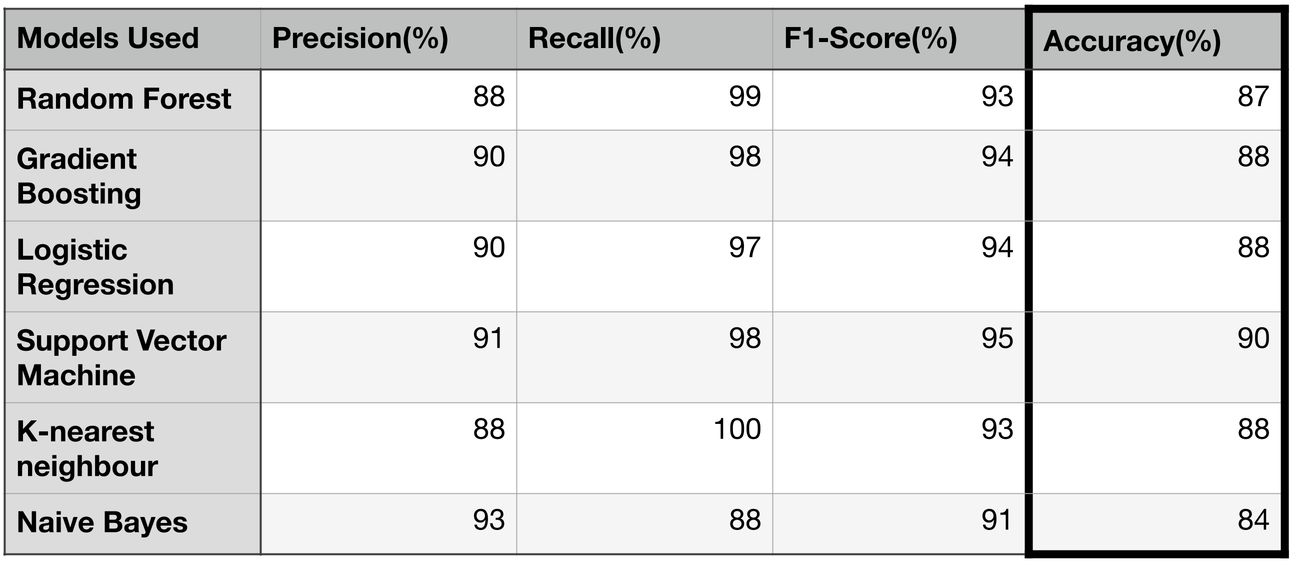
2. **Hyperparameter Tuning using GridSearchCV:**

- GridSearchCV is employed to systematically search for the best combination of hyperparameters for each machine learning algorithm.

- Cross-validation is used to evaluate the performance of each hyperparameter combination.

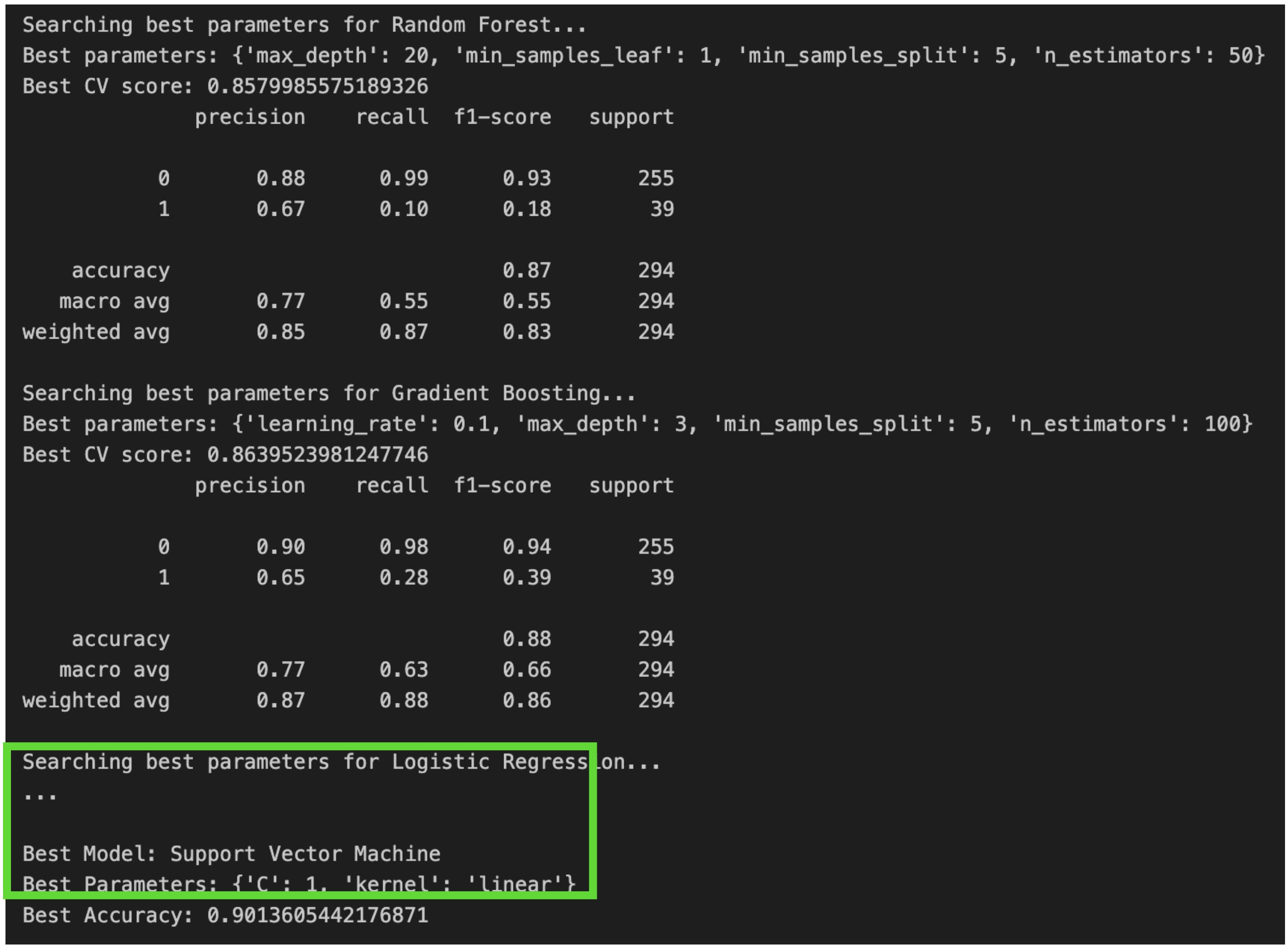
Hyperparameters: Hyperparameters are parameters that are set prior to the training process and determine the behavior of a machine learning algorithm. Unlike model parameters, which are learned during training, hyperparameters are not directly learned from the data but rather set externally by the user.

GridSearchCV (Grid Search Cross-Validation): It is a technique used to systematically search for the best combination of hyperparameters for a given machine learning algorithm.
(Refer Fig.11)



Models Used	Precision(%)	Recall(%)	F1-Score(%)	Accuracy(%)
Random Forest	88	99	93	87
Gradient Boosting	90	98	94	88
Logistic Regression	90	97	94	88
Support Vector Machine	91	98	95	90
K-nearest neighbour	88	100	93	88
Naive Bayes	93	88	91	84

Model Evaluation of different models used



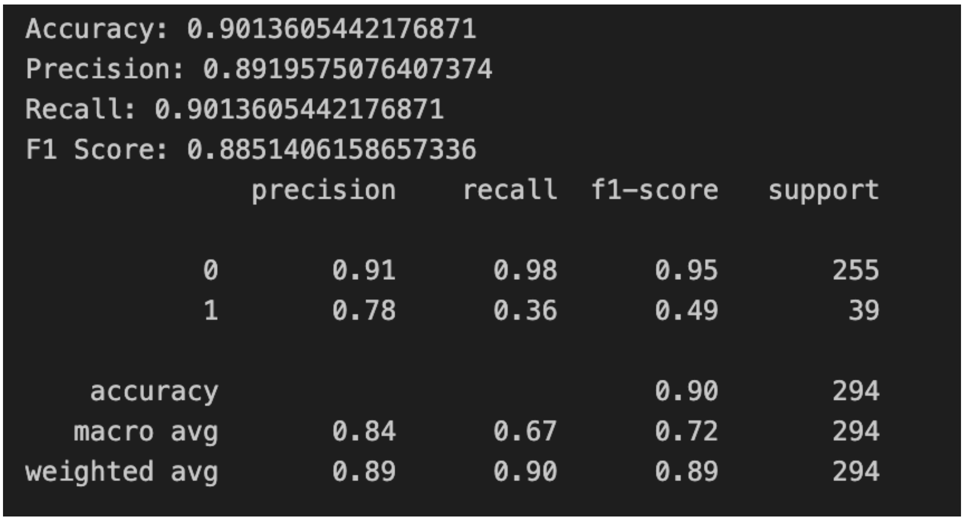
```
Searching best parameters for Random Forest...
Best parameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 50}
Best CV score: 0.8579985575189326
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       255
           1       0.67      0.10      0.18        39

    accuracy                           0.87       294
   macro avg       0.77      0.55      0.55       294
weighted avg       0.85      0.87      0.83       294

Searching best parameters for Gradient Boosting...
Best parameters: {'learning_rate': 0.1, 'max_depth': 3, 'min_samples_split': 5, 'n_estimators': 100}
Best CV score: 0.8639523981247746
              precision    recall  f1-score   support

           0       0.90      0.98      0.94       255
           1       0.65      0.28      0.39        39

    accuracy                           0.88       294
   macro avg       0.77      0.63      0.66       294
weighted avg       0.87      0.88      0.86       294

Searching best parameters for Logistic Regression...
...

Best Model: Support Vector Machine
Best Parameters: {'C': 1, 'kernel': 'linear'}
Best Accuracy: 0.9013605442176871
```

Fig:11 Result of GridSearchCV to find the best model and its parameters

Conclusion:

Best Model	Best parameters	Accuracy
Support Vector Machine(SVM)	C’ :1 ‘kernel’ : linear	90.01%

The SVM model achieved an accuracy of 90.14% on the testing data, indicating its effectiveness in distinguishing between employees who are likely to leave the organization (attrition = 'Yes') and those who are not (attrition = 'No'). The use of a linear kernel suggests that the decision boundary between the two classes is linear in feature space.
(Refer Fig:12 to see the code result)



```
Accuracy: 0.9013605442176871
Precision: 0.8919575076407374
Recall: 0.9013605442176871
F1 Score: 0.8851406158657336
              precision    recall  f1-score   support

           0       0.91      0.98      0.95       255
           1       0.78      0.36      0.49        39

    accuracy                           0.90       294
   macro avg       0.84      0.67      0.72       294
weighted avg       0.89      0.90      0.89       294
```

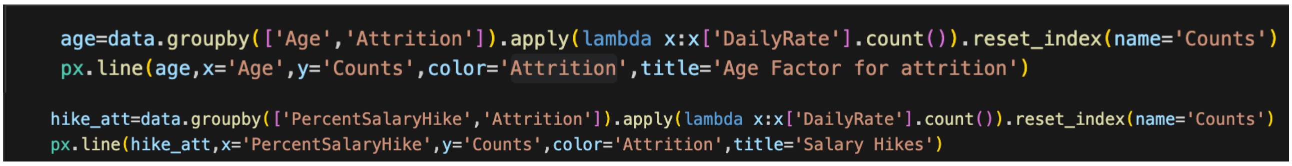
Fig:12 Code output of the Best model (SVM) and its parameters

3. Complexities in Visuals

Creating effective visualizations for the "Predicting Employee Attrition" project involves handling various complexities:

Functions:

- **Data Aggregation**: Summarizing large datasets into meaningful aggregates, such as average age or average years in the current role, to identify trends and patterns in employee attrition. For eg.(Refer Fig 13.1)



```
 age=data.groupby(['Age','Attrition']).apply(lambda x:x['DailyRate'].count()).reset_index(name='Counts')
 px.line(age,x='Age',y='Counts',color='Attrition',title='Age Factor for attrition')

hike_att=data.groupby(['PercentSalaryHike','Attrition']).apply(lambda x:x['DailyRate'].count()).reset_index(name='Counts')
px.line(hike_att,x='PercentSalaryHike',y='Counts',color='Attrition',title='Salary Hikes')
```

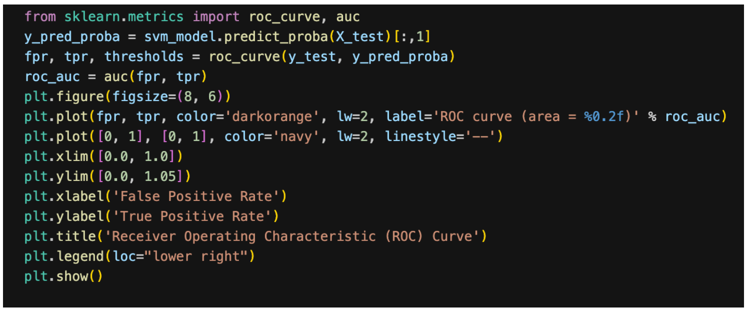
Fig:13.1 Project code reference to implementing Data Aggregation

- **Custom Functions**: Writing custom functions to calculate metrics like attrition rate across different departments or age groups.

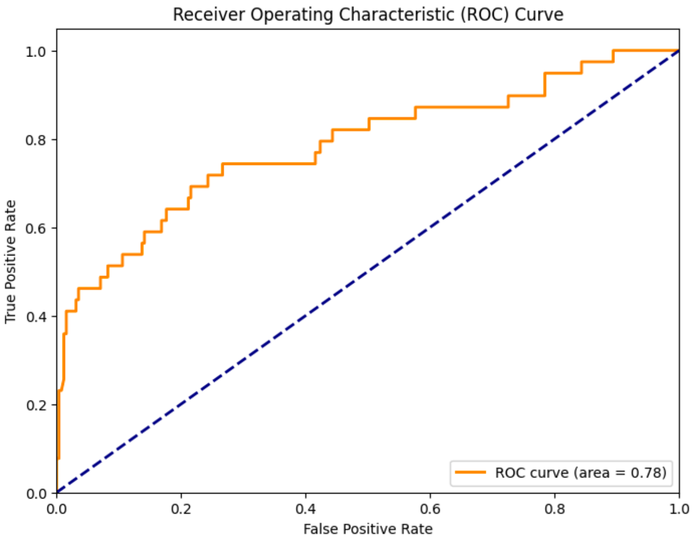
Measures:

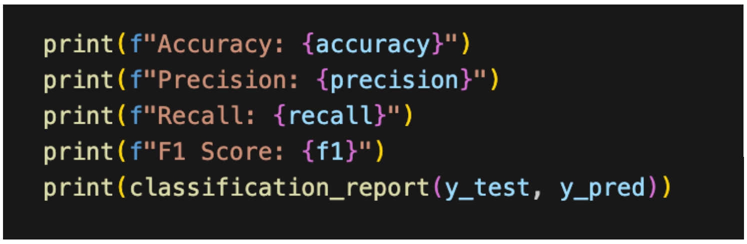
- **Choosing Appropriate Metrics**: Selecting relevant metrics like accuracy, precision, recall, and F1 score to evaluate model performance and displaying these metrics in a comprehensible manner.

- **Multiple Metrics**: Displaying multiple metrics simultaneously, such as confusion matrices and ROC curves, to provide a comprehensive view of model performance.(Refer Fig:13.2)



```
from sklearn.metrics import roc_curve, auc
y_pred_proba = svm_model.predict_proba(X_test)[:,1]
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
roc_auc = auc(fpr, tpr)
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()
```





```
print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1}")
print(classification_report(y_test, y_pred))
```

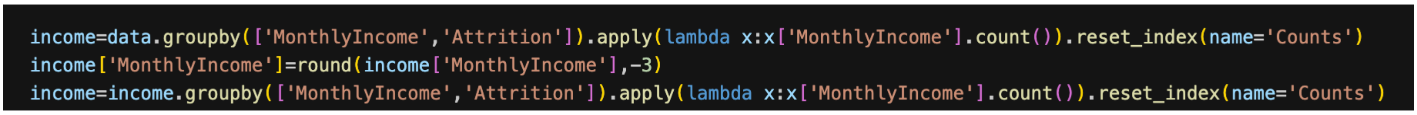
Fig:13.2 Example of code and results of appropriate measures used in the code

Data Wrangling:

- **Cleaning Data**: Handling missing values, removing duplicates, and correcting data inconsistencies are crucial steps before visualization.
- **Encoding Categorical Variables**: Converting categorical variables (e.g., job role, department) into numerical values using techniques like one-hot encoding to make them suitable for visualization.
- **Standardizing Data**: Scaling numerical features (e.g., income, years at company) to a common range, which is important for certain visualizations and machine learning models.

Connections:

- **Data Integration**: Combining data from multiple sources, such as employee performance records and HR databases, and ensuring consistency across datasets. (Refer Fig:13.3)



```
income=data.groupby(['MonthlyIncome','Attrition']).apply(lambda x:x['MonthlyIncome'].count()).reset_index(name='Counts')
income['MonthlyIncome']=round(income['MonthlyIncome'],-3)
income=income.groupby(['MonthlyIncome','Attrition']).apply(lambda x:x['MonthlyIncome'].count()).reset_index(name='Counts')
```

Fig:13.3 Implementation of Data Integration

Challenges Encountered:

1. **Choosing the Best Model:**

- The initial challenge was to select the most suitable machine learning model for predicting employee attrition from a wide range of options.

- Experimentation with different algorithms, such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine, was necessary to identify the model with the best performance.

2. **Studying the Wide Dataset of 30 Columns:**

- Analyzing a dataset with 30 columns required thorough exploration and understanding of the relationships between various features and their impact on employee attrition.

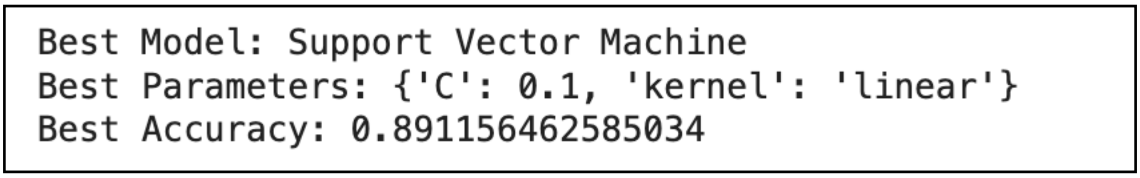
- Performing exploratory data analysis (EDA) and correlation analysis helped in gaining insights into the dataset's characteristics and identifying relevant features for modeling.

3. **Improving Accuracy:**

- Initially, achieving a satisfactory level of accuracy posed a challenge, as the logistic regression model yielded an accuracy of 86%.

- Hyperparameter tuning using techniques like GridSearchCV was crucial in fine-tuning the models and improving accuracy incrementally.[Fig.13] shows the accuracy of the previous best model.

- Further adjustments to hyperparameters and model selection led to significant improvements in accuracy, eventually reaching 90.1% with the Support Vector Machine (SVM) model by fine tuning and adding extra parameters and removing the dropped columns line of code taking in consideration the all the features. [Fig.11] shows the accuracy of the improved model.



```
Best Model: Support Vector Machine
Best Parameters: {'C': 0.1, 'kernel': 'linear'}
Best Accuracy: 0.891156462585034
```

Fig.13 Accuracy of the previous best model before improvement

Recommendations for Reducing Employee Attrition:

1. Implement Employee Engagement Initiatives:

   - Foster a positive work environment by promoting open communication, recognition programs, and opportunities for professional development.
   - Conduct regular employee satisfaction surveys to identify areas of improvement and address concerns proactively.

2. Offer Competitive Compensation and Benefits:

   - Ensure that salaries and benefits packages are competitive within the industry to attract and retain top talent.
   - Provide opportunities for performance-based bonuses, incentives, and career advancement.

3. Promote Work-Life Balance:

   - Encourage flexible work arrangements, such as remote work options and flexible hours, to accommodate employees' personal commitments and promote work-life balance.
   - Offer wellness programs and initiatives to support employees' physical and mental well-being.

4. Provide Opportunities for Growth and Development:

   - Offer training programs, workshops, and mentorship opportunities to enhance employees' skills and capabilities.
   - Provide clear pathways for career advancement and opportunities for employees to take on new challenges and responsibilities.

5. Foster a Positive Organizational Culture:

   - Promote values such as diversity, inclusion, and respect in the workplace to create a supportive and inclusive culture.
   - Encourage teamwork, collaboration, and camaraderie among employees to build strong relationships and a sense of belonging.

Dashboards improvements for future:

1. **Interactivity and User Control**

   - **Filters and Selectors**: Implemented dropdown filters for departments, job roles, and age groups to allow users to slice and dice the data.

   - **Dynamic Tooltips**: Added tooltips with detailed information on hover for each chart element, providing exact numbers and additional context.

2. **Comparative and Detailed Visuals**

   - **Time Series Plots**: Introduced time series plots to visualize attrition trends over different periods, identifying patterns.

   - **Heatmaps**: Added heatmaps to display correlations between various factors and attrition rates.

   - **Side-by-Side Comparisons**: Included side-by-side bar charts to compare attrition rates across different categories, such as departments and job roles.

3. **Predictive Analytics Integration**

   - **Risk Scores**: Integrated a predictive model to assign attrition risk scores to employees, highlighting those at high risk.

   - **What-If Scenarios**: Enabled what-if scenario analysis to assess the impact of potential changes on attrition rates.

5. **User Experience and Customization**

   - **Responsive Design**: Ensured the dashboard is responsive and accessible on various devices.

   - **Customization Options**: Allowed users to customize their dashboard views and save their preferences.

   - **Consistent UI/UX**: Maintained a consistent and clean design throughout the dashboard.

6. **Additional Metrics**

   - **Employee Satisfaction**: Included metrics related to employee satisfaction and engagement.

   - **Financial Impact**: Provided analysis on the financial impact of employee attrition, illustrating costs related to hiring and training.

Summary:

After going thorough the data, these are the following observations:

1. People are tending to switch to a different jobs at the start of their careers, or at the earlier parts of it. Once they have settled with a family or have found stability in their jobs, they tend to stay long in the same organizatin.

2. Salary and stock options have a great motivation on the employees and people tend to leave the organization much lesser. Higher pay and more stock options have seen more employees remain loyal to their company.

3. Work life balance is a great motivation factor for the employees. However, people with a good work-life balance, tend to switch in search of better opportunities and a better standard of living.

4. Departments where target meeting performance is very much crucial (for e.g. Sales) tend to have a greater chances of leaving the organization as compared to departments with more administration perspective (For e.g. Research and development)

5. People with a good Job Satisfaction and Environment satisfaction are loyal to the organization- and this speaks loud for any Organization. However, people who are not much satisfied with their current project- tend to leave the organization far more.

References:

Research papers:

1. https://core.ac.uk/download/pdf/234697248.pdf
2. https://www.mdpi.com/2073-431X/9/4/86
3. https://ieeexplore.ieee.org/abstract/document/8605976