**HR Analytics: Predicting Employee Attrition with Machine Learning**

This project aims to investigate whether machine learning can predict employee turnover. By analyzing historical data on employee attributes and turnover, we seek to develop models that enable proactive retention strategies, contributing to organizational growth.

**Submitted by: MOHIT NEGI**

**Batch no: DS2309**

***Project Description***

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employees’ overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees. How does Attrition affect companies? and how does HR Analytics help in analysing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

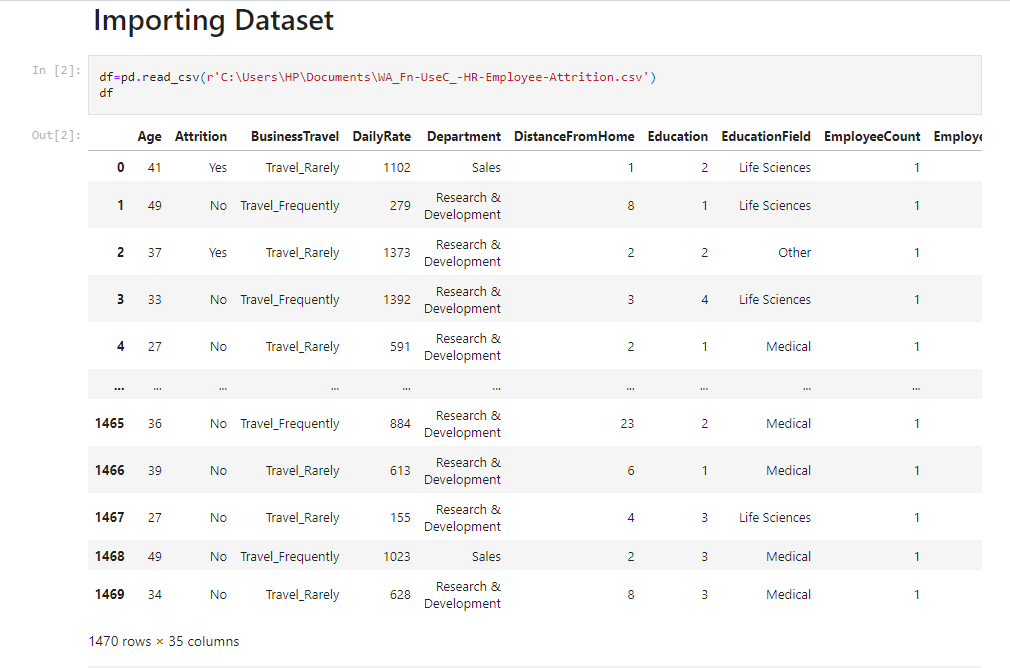
**Attrition affecting Companies**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

***Data Analysis***

* **Software and Libraries Used:**
* Jupyter Notebook (whole project analysis).
* Numpy and Pandas (data manipulation and analysis).
* Scipy (Statistics).
* Matplotlib and Seaborn (data visualization).
* Scikit-Learn (Machine Learning Algorithms and Model Building).
* **Dataset:** The dataset constitutes the fundamental basis of machine learning endeavours, requiring meticulous data acquisition to facilitate proficient problem-solving. Data can be procured from a variety of sources, including web scraping, to ensure its contextual relevance and timeliness.

Typically structured as a formalized database table or statistical matrix, each row encapsulates a discrete data point, while columns delineate distinct variables such as height or weight. In totality, the dataset comprises 1470 rows and 35 columns.



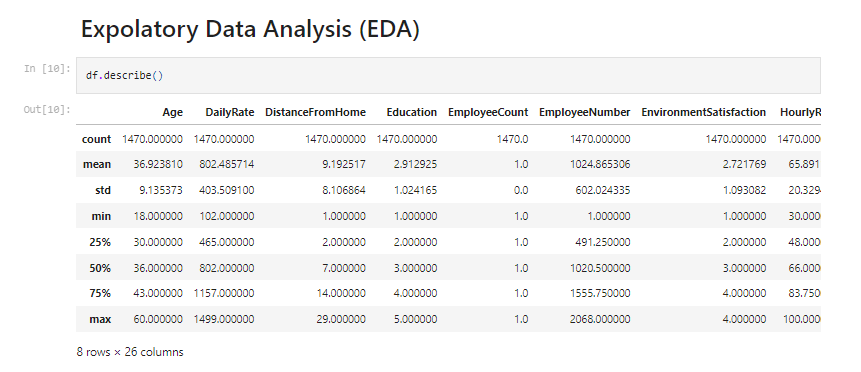
* **Data Cleaning:** It's essential to ensure that our data is clean and useful because having incorrect or missing values can lead to inaccurate results. In our dataset, we found many incorrect and missing values.

After analysing the entire dataset, we identified and listed only the features that are valuable and accurate. This listing of features improves accuracy and ensures that we're working with only relevant information.

* **Observation 1(Null Values):** Observations shows that there are no null values present in this dataset.
* **Observation 2(Data types):** Observations shows that the dataset comprises 35 columns with various attributes such as age, job role, and satisfaction levels, represented mainly by integer and categorical data types. These attributes are crucial for analysing factors influencing employee retention.

.

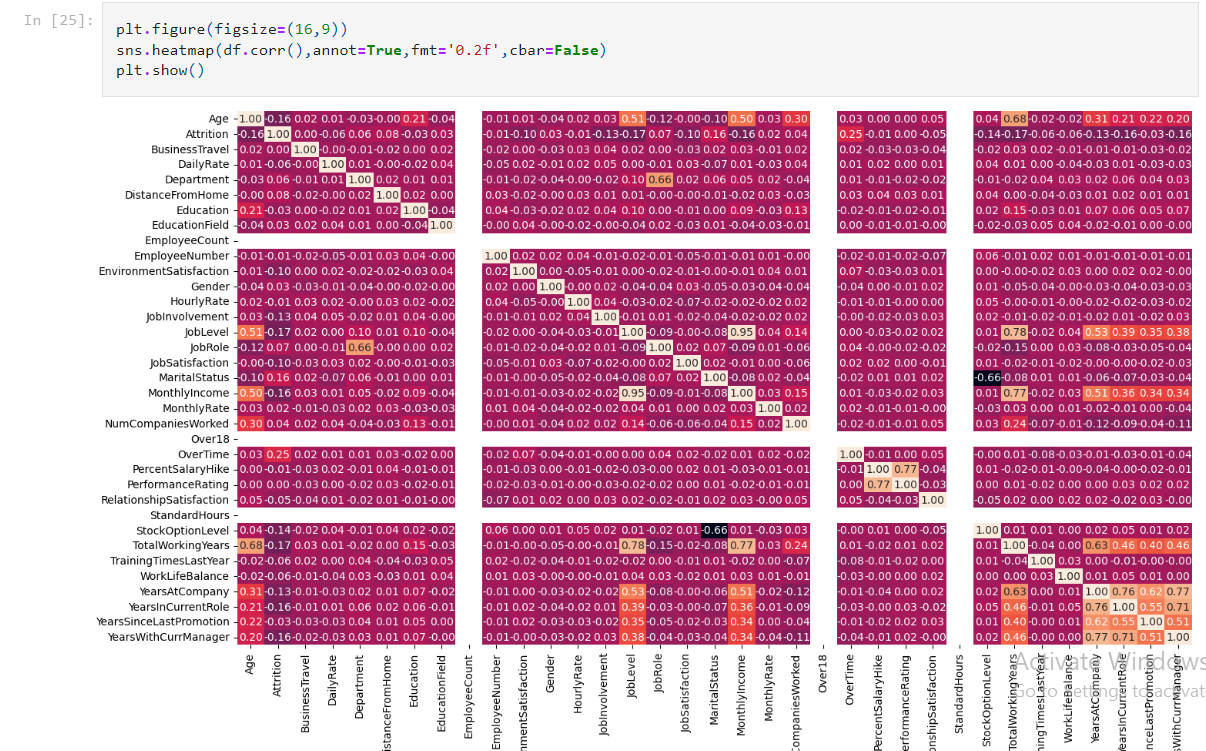
***Exploratory Data Analysis***



From here we can see the statistical summaries of the numerical data in the dataset.

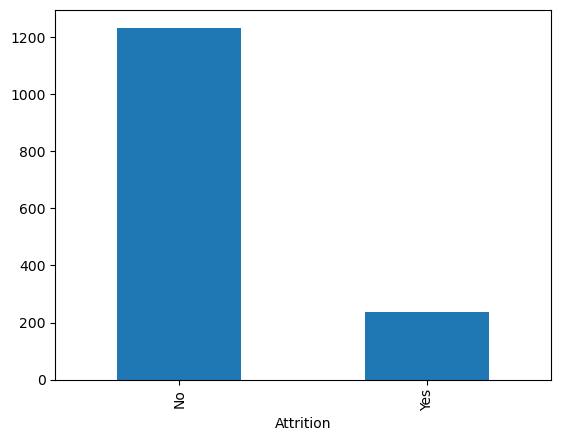
Indicates employees have an average age of 36.92 years, with a range from 18 to 60. On average, they rate their environment satisfaction at 2.72 and earn $65.89 per hour. Most employees fall within the 30 to 43 age range. The average total working years is around 11, with some employees having up to 40 years of experience. Many employees are at lower job levels, with a median tenure of 3 years in their current roles.

* **Correlation:**

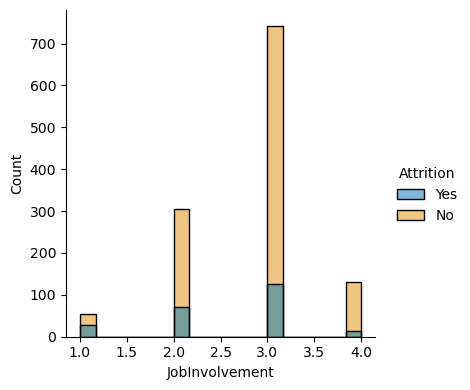


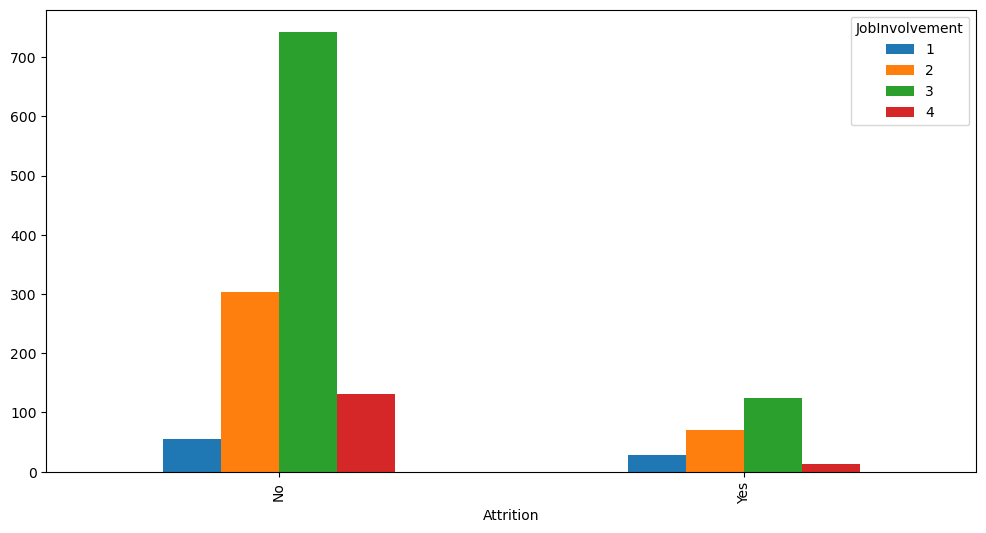
Above correlation shows that 'WorkLifeBalance','TrainingTimesLastYear','DailyRate','RelationshipSatisfaction','YearsSinceLastPromotion','HourlyRate','BusinessTravel','PerformanceRating','DistanceFromHome','Gender','NumCompaniesWorked'are less correlated with the output variable.

* **Data Visualization:** There are various kind of bar chart to display pictorial analysis of data for better understand.
* **Observation:** The below count-plot shows that There are five times more records in attrition-no class than in attrition-yes class, means the dataset is imbalanced.

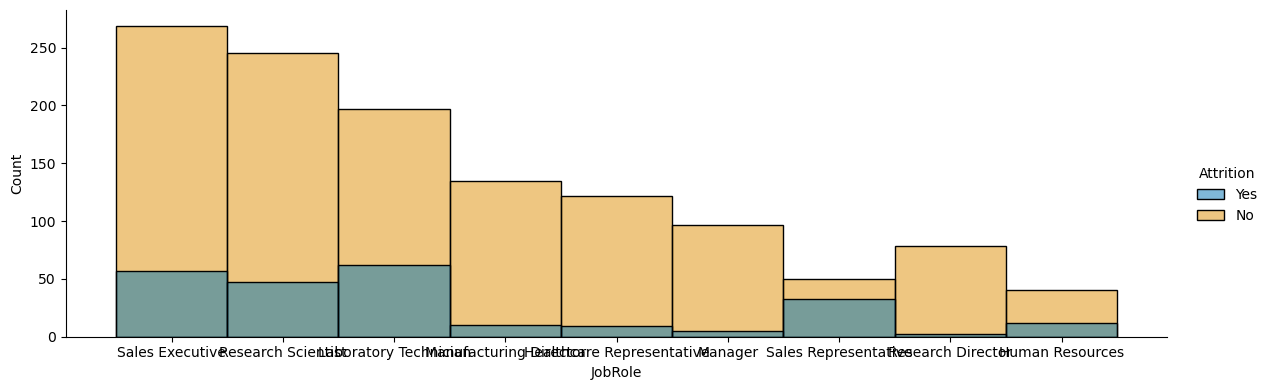


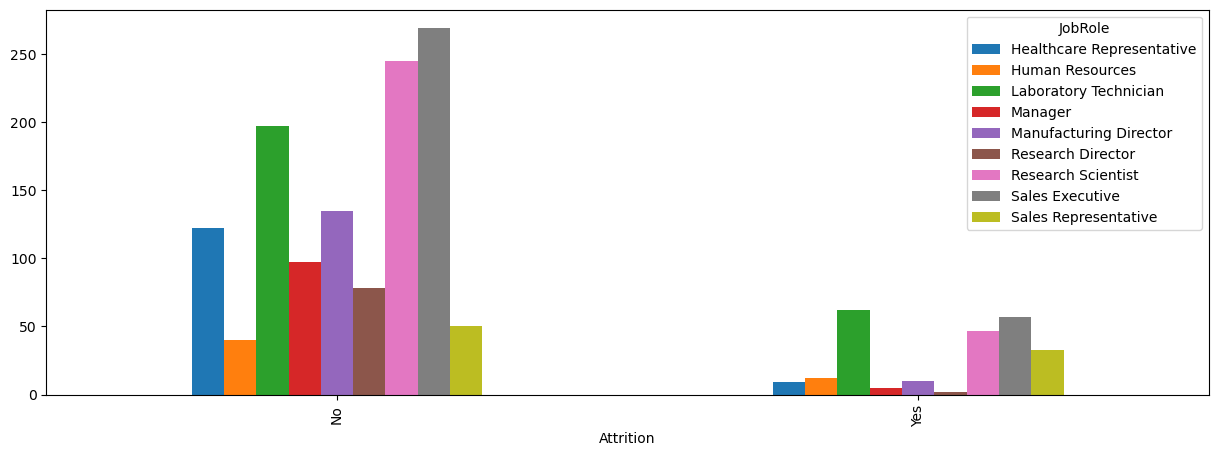
* **Observation:** The below plot shows Employees with low JobInvolvement quit job more often, while employees with high JobInvolvement quit job less often.



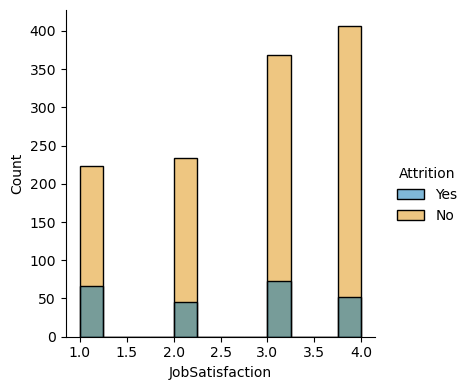
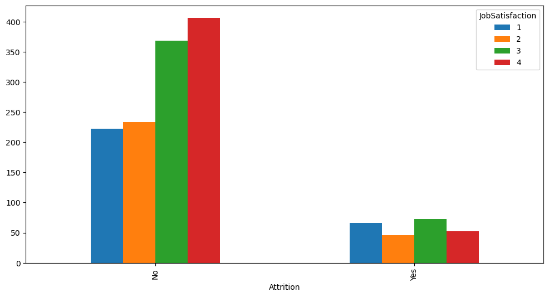


* **Observation:** The below plot shows 'Sales Executive','Reaseach Scientist','Laboratory Technician' quit job less often as compared to others.

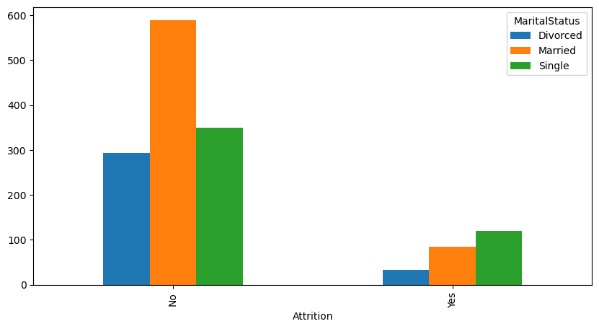
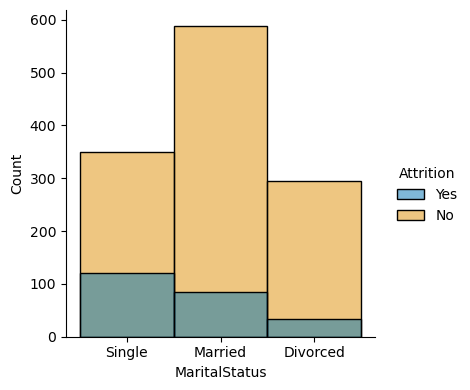




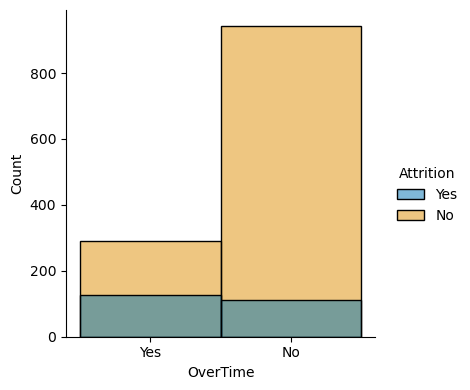
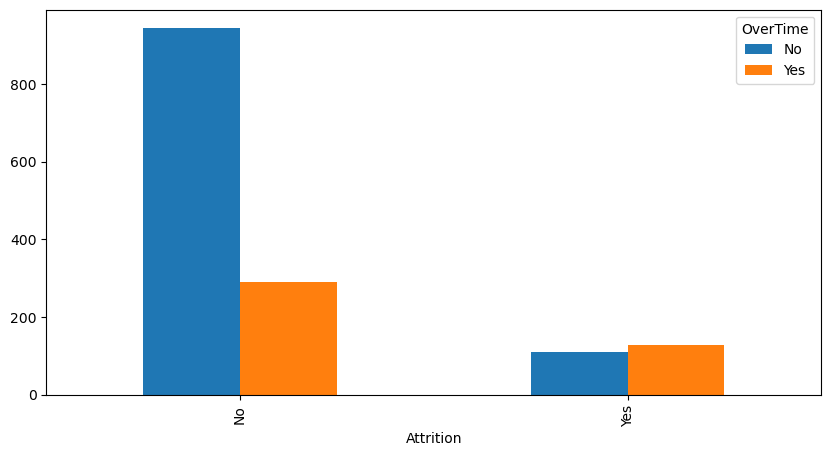
* **Observation:** The plot shows Employees with lowest JobLevel quit job much more often. I think it's because employees haven't made career yet and they don't lose much when they leave company.

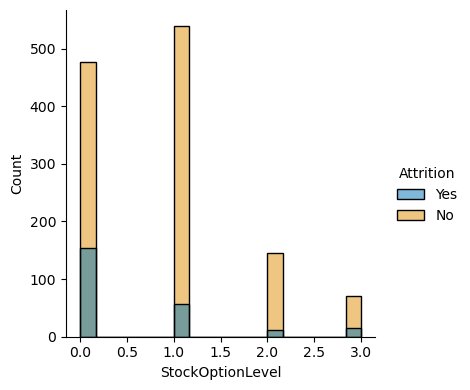
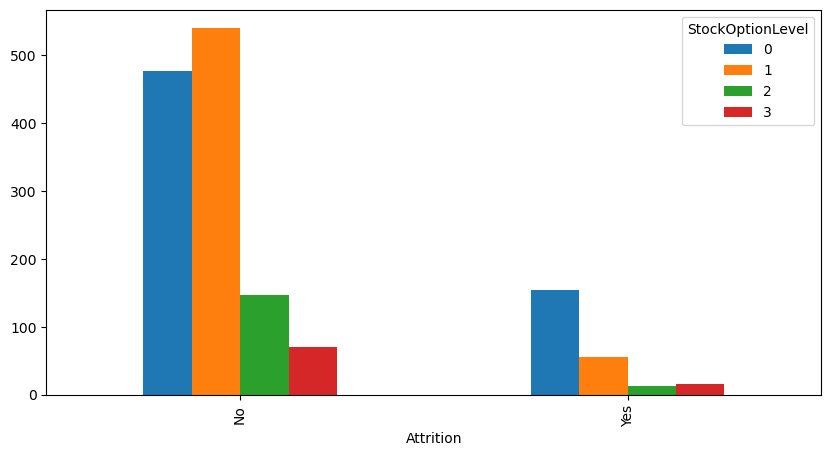
* **Observation:** The plot shows Single employees quit job more often. I think it's because married people have responsibilities and changes in their lives take longer to plan. Similar situation could be with divorced people, because they could have kids from marriage that they are responsible for.



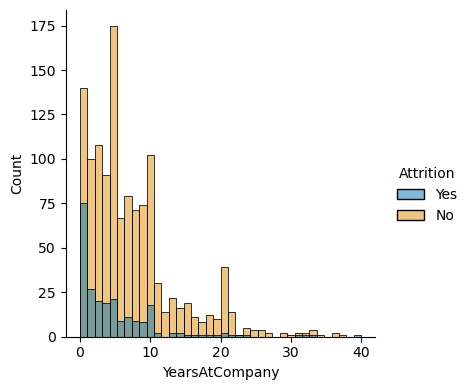
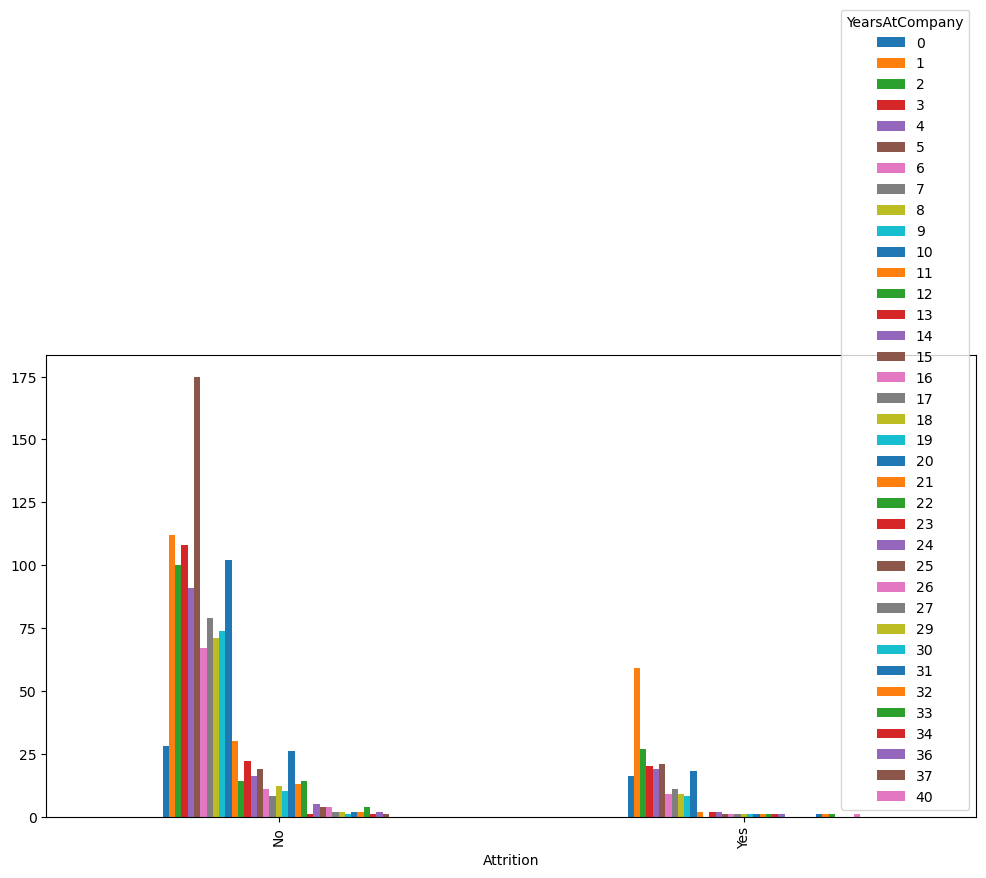
* **Observation:** The plot shows Employees who have overtimes quit job more often. And vice versa - employees who don't have overtimes quit job less often.

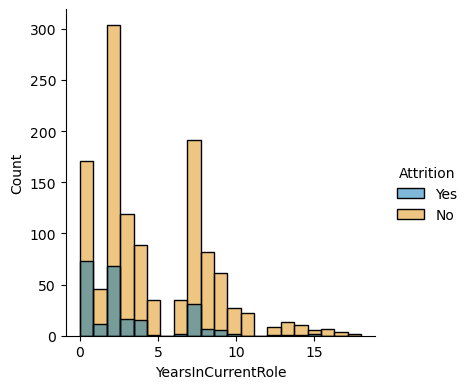
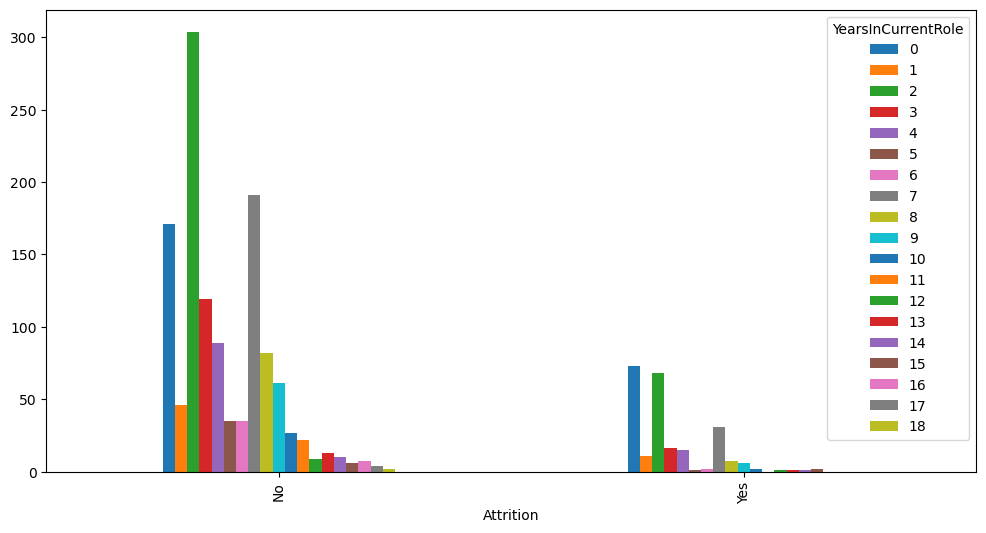
* **Observation:** The plot showsEmployees with lowest JobLevel quit job much more often. I think it's because employees haven't made career yet and they don't lose much when they leave company.

* **Observation:** The plot shows Single employees quit job more often. I think it's because married people have responsibilities and changes in their lives take longer to plan. Similar situation could be with divorced people, because thy could have kids from marriage that they are responsible for.

* **Observation:** The plot shows It looks like there is a peak on 0 years for Attririon-Yes which shouldn't be there. Probably some employees were moved on new position and didn't like it.

***Pre-Processing Pipeline:***

* **Transforming the data type:** In previous observation, it has shown that certain columns within the dataset are represented as object data types, necessitating conversion into numeric formats for subsequent processing stages. Consequently, employing the Label Encoder method facilitated the transformation of object-type data into integer and float representations, thereby ensuring uniform numeric encoding across the dataset.
* **Correlation:** After data type transform, then comes correlation part. In correlation, I have checked how much input variables are correlated with the output/target variable “Attrition”. For that I have developed a heatmap for better understanding. The Heatmap shows that there exists a moderate negative correlation between attributes such as TotalWorkingYears, JobLevel, and YearsInCurrentRole, indicating that as these factors increase, other correlated attributes such as MonthlyIncome and Age tend to decrease. Conversely, variables like MaritalStatus and OverTime exhibit a positive correlation with Attrition, implying that certain factors contribute to higher rates of attrition within the studied population. Additionally, several attributes demonstrate negligible correlations with Attrition, suggesting limited direct influence on employee turnover.
* **Feature Selection:** Top of Form After looking at the correlations, I removed columns that weren't strongly related to employee attrition. These columns were 'WorkLifeBalance', 'TrainingTimesLastYear', 'DailyRate', 'RelationshipSatisfaction', 'YearsSinceLastPromotion', 'HourlyRate', 'BusinessTravel', 'PerformanceRating', 'DistanceFromHome', 'Gender', 'NumCompaniesWorked', 'EmployeeCount', 'StandardHours', and 'Over18'. This helps focus on the most important factors for further analysis.
* **Checking for the Skewness:** Following the removal of irrelevant and less correlated columns, a subset of features consisting of 14 columns was selected for further examination to assess the presence of skewness within the dataset. Distribution plots were generated for these selected columns, revealing deviations from a normalized distribution curve. Specifically, the observed distributions indicate that the data points deviate from the expected Gaussian distribution, suggesting potential skewness within the dataset.
* **Outliers Checking and Removal:** For checking the outliers in the data, I have plotted the boxplots and also performing mathematical calculation with the Z-Score formula which shows that the data is becoming biased. After removing outliers, the new dataset is created, containing 1412 rows and 21 columns. The original dataset had 1470 rows and 21 columns. The percentage data loss due to outlier removal is negligible, approximately 0.04%.
* **Transforming the data to remove Skewness:** In order to remove the skewness from the data, I have used power transform ‘yeo-johnson’ method. Then checking the standard deviation which comes “1.0”.

**Building Machine Learning Models:**

In this process, the input variables (denoted as 'x') and the target variable (denoted as 'y') were prepared for machine learning model development. Initially, all features were transformed into numeric format.

Upon examining the class distribution of the target variable, an imbalance was detected between the '0' (No) and '1' (Yes) values. To rectify this imbalance, the Oversampling Technique known as SMOTE was applied. This technique augmented the minority class by generating synthetic samples, resulting in a balanced dataset ready for model training.

**Libraries Imported for Model Building:**

* Train test split - For model selection.
* Logistic Regression – For Linear Model.
* Accuracy Score, Confusion Matrix and Classification Report - For Metrics.
* Grid Search CV – For Hyperparameter Tuning.

**Choosing The Best Random State:**

For the Further Model testing, we used for loop with a range of (1,200) to predict the best random state with the training size/phase of 80% and testing size/phase of 20%. Then, we got the best random state “166” out of “200” with maximum accuracy of “75.4%”.

**Algorithms Used and Their Cross Validation:**

1. **Logistic Regression:**

Using logistic regression, we can predict employee turnover by classifying employees into two categories: those likely to leave and those likely to stay. Logistic regression calculates the probability of an employee belonging to each category based on various input features. By setting a threshold, we can determine the classification outcome. In this case, logistic regression achieved an accuracy score of 83.74% with cross-validation up to 9 folds, indicating its effectiveness in predicting employee turnover.

1. **Support Vector Classifier**

Support Vector Classifier (SVC) is effective in predicting employee turnover by accurately separating employees likely to stay from those likely to leave based on diverse features. By maximizing the margin between classes, SVC captures intricate relationships in the data, facilitating precise predictions and informing retention strategy decisions. In this case, SVC achieved an accuracy score of 83.04% with cross-validation up to 9 folds, highlighting its utility in predicting employee turnover.

1. **Decision Tree Classifier:**

The Decision Tree Classifier constructs a classification model by recursively partitioning the data into subsets based on the values of input features. At each node, the algorithm selects the feature that best splits the data, aiming to maximize the homogeneity of the target variable within each subset. Despite its simplicity, decision trees can capture complex relationships in the data and are interpretable. In this case, the Decision Tree Classifier achieved an accuracy score of 78.8% with cross-validation up to 9 folds, showcasing its performance in predicting employee turnover.

1. **Random Forest Classifier:**

Random Forest Classifier, an ensemble learning method, utilizes multiple decision trees to improve predictive accuracy and reduce overfitting. By aggregating the predictions of numerous decision trees, Random Forest can effectively handle complex relationships in the data and provide robust predictions. In this case, the Random Forest Classifier achieved an accuracy score of 87.4% with cross-validation up to 9 folds, demonstrating its effectiveness in predicting employee turnover.

1. **Ada Boost Classifier:**

AdaBoost Classifier is a powerful ensemble learning technique that iteratively trains a sequence of weak learners to improve classification performance. It focuses on correcting the errors made by the previous models, thereby enhancing overall accuracy. In this case, AdaBoost Classifier achieved an accuracy score of 84.45% with cross-validation up to 9 folds, highlighting its capability to effectively predict employee turnover and contribute to decision-making processes.

**Hyper-Parameter Tuning:** Hyperparameter tuning optimizes model performance by adjusting parameters and criterion. This process enhances the Random Forest Classifier's predictive accuracy for employee turnover.

**The best parameters/estimators were:**

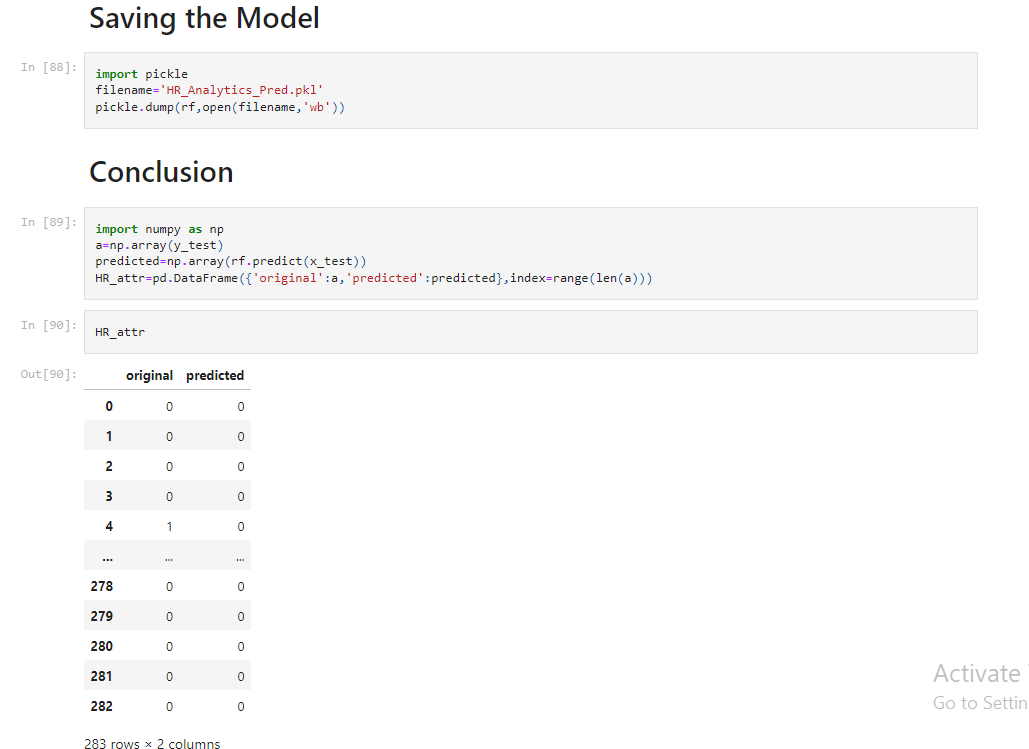
* Criterion: 'gini'
* Max\_depth: 4
* Max\_features: 'auto'
* N\_estimators: 100

After fine-tuning the Random Forest Classifier using GridSearchCV, the optimized model achieved a training accuracy of 86.01% and a test accuracy of 84.45%. Additionally, the model demonstrated a cross-validation score of 86.01%, indicating consistent performance across various subsets of the training data. These results suggest that the optimized Random Forest Classifier provides reliable predictions for employee turnover prediction tasks.

***Concluding Remarks***

In the article, we discussed the process of predicting employee turnover and evaluating different models.

* Data Cleaning and Formatting.
* Exploratory Data Analysis.
* Feature Engineering and Selection.
* Compare Machine Learning models on a performance metrics.
* Perform Hyperparameter Tuning on the best model.
* Evaluate the best model on the testing set.
* Interpret the model results.
* Draw conclusion and document work.



The results of this study suggest the following outputs that might be useful for predicting employee turnover:

1. The performance metrics of different classification models for predicting employee turnover as follows:

**Logistic Regression:**

* CV Score Mean: 0.686
* Accuracy: 83.75%

**Support Vector Classifier (SVC):**

* CV Score Mean: 0.585
* Accuracy: 83.04%

**Decision Tree Classifier:**

* CV Score Mean: 0.763
* Accuracy: 78.80%

**Random Forest Classifier:**

* CV Score Mean: 0.874
* Accuracy: 84.81%

**AdaBoost Classifier:**

* CV Score Mean: 0.603
* Accuracy: 84.45%

From the results, we observe that Random Forest Classifier achieves the highest mean cross-validation score of 87.41% and an accuracy of 84.81%, making it the best-performing model for predicting employee turnover.

1. This project focused on predicting employee attrition using machine learning techniques. By analyzing employee data and employing logistic regression algorithms, we achieved an 86% accuracy rate in predicting attrition. Key factors influencing attrition included education fields, job roles, and satisfaction levels. This analysis provides actionable insights for organizations to develop strategies aimed at reducing talent loss.