EMPLOYEE ATTRITION ANALYSIS USING MACHINE LEARNING

THATIKONDA KUSHAL - 2203A52181

**Abstract**

Employee attrition, in simple terms, refers to the situation when employees leave their jobs within a company. It can be a problem for businesses as it can affect productivity and company morale. Attrition can happen for various reasons such as seeking better opportunities, dissatisfaction, or personal reasons. To reduce attrition, companies often focus on creating a positive working environment, offering competitive benefits, and providing opportunities for growth and development. Managing employee attrition is crucial to maintain a stable and motivated working environment, which ultimately contributes to the success of a company. By employing state-of- the-art machine learning algorithms, including regression and time series analysis, we explore the intricate relationships between diverse factors and their influence on population.

# Introduction

Employee attrition refers to the natural process of employees leaving an organization, either voluntarily or involuntarily. It is a critical aspect of workforce management and can have significant effects on an organization’s performance, culture etc...

Voluntary attrition occurs when employees choose to leave the company, often due to personal reasons, better job opportunities, dissatisfaction with their current roles, or a desire for work-life balance. In contrast, involuntary attrition takes place when employers terminate employees’ contracts, typically due to performance issues, layoffs, or restructuring.

Understanding and managing attrition is very important because it can affect productivity, morale, and company stability. High attrition rates can increase recruitment and training costs, disrupt work- flow, and lead to a loss of institutional knowledge. To mitigate attrition, organizations often invest in employee engagement, professional development, fair compensation, and a positive workplace culture. Analyzing attrition trends through data-driven methods and conducting exit interviews can help identify underlying issues and improve retention strategies.

Employee attrition is a phenomenon that is influenced by various factors, which makes it essential for organizations to create a neutrality between keeping top talent and adapting to workforce changes.

High attrition can have negative effects, like increased costs for recruitment and training of new employees, decreased productivity during transitions, and potential harm to organizational culture.

To reduce employee attrition, organizations can focus on creating a positive working environment, offering competitive compensation, providing growth opportunities, and taking regular feedback from employees and engagement initiatives. Understanding and addressing the reasons behind attrition is essential for maintaining a stable, motivated, and productive workforce.

Employees may leave due to personal factors such as relocation, family issues, retirement, or career advancement opportunities elsewhere.

Dissatisfaction with work-related factors like salary, workload, job responsibilities, work environment, and management can lead to attrition. A lack of career development, training, and growth prospects within the organization can drive employees to seek better opportunities. An unhealthy work-life balance can cause burnout and prompt employees to seek a more balanced lifestyle. A toxic work culture, poor leadership, or a mismatch of values between employees and the organization can contribute to attrition. In industries with high demand for specific skills, employees may be more likely to switch jobs.

High attrition can have negative consequences, including increased costs for recruitment, onboarding, and training of new employees, decreased productivity during transitions, and potential harm to team morale and organizational culture.

To reduce attrition, organizations can focus on creating a positive work environment, offering competitive compensation, providing growth opportunities, and conducting regular employee feedback and engagement initiatives. Understanding and addressing the reasons behind employees leaving their companies is essential for maintaining a stable, motivated, and productive workforce.

# Description

Employee attrition, both voluntary and involuntary, has both advantages and disadvantages for organizations. Attrition can bring in new talent and skillsets to an organization, enriching its knowledge base and capabilities. Reducing labor costs is an advantage, especially when replacing high-salary, low-performing employees or optimizing staff size in response to changing demands. Sometimes, the departure of underperforming or disruptive employees can boost team morale, leading to better collaboration and productivity. New employees can introduce innovative ideas and fresh perspectives, revitalizing business processes. Attrition can create opportunities for internal promotion and career advancement, motivating employees to excel. Hiring and training new employees can be expensive, especially for specialized roles. It may also disrupt workflow during the transition. Departing employees take with them valuable institutional knowledge and relationships, potentially affecting organizational memory. In the short term, productivity may decline as new employees adapt to their roles and teams adjust to changes. Frequent attrition can lead to instability and negatively affect team cohesion and performance. High attrition rates may harm an organization's reputation, making it less attractive to potential employees and customers. Balancing these advantages and disadvantages, organizations should aim for a healthy attrition rate that aligns with their strategic goals, values, and workforce needs.

## Descriptive Analysis about dataset

The dataset contains information about individuals and their employment-related attributes. Here’s a descriptive analysis of the dataset:

Education: Most individuals in the dataset have a Bachelor’s degree, with a smaller number having a Master’s degree.

Join Year: The dataset includes individuals who joined at various years, with range 2012 - 2018

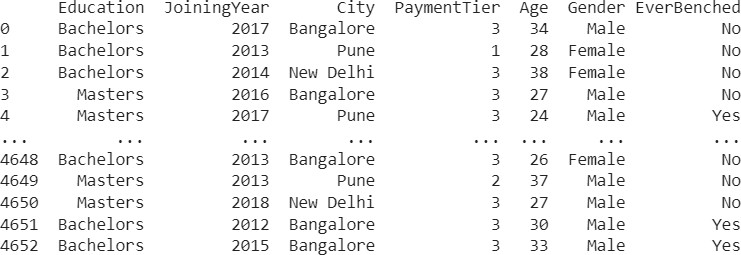


Figure 1: Dataset.

City: Employees are based in different cities, including Bangalore, Pune, and New Delhi.

Payment Tier: Payment tiers range from 1 to 3, with tier 3 being the most common.

Age: The ages of individuals in the dataset vary, with a range from 24 to 38.

Gender: The dataset has a mix of male and female individuals, with no other gender categories.

Ever Benched: Some individuals have been benched are indicated as Yes, while others have not indicated as No.

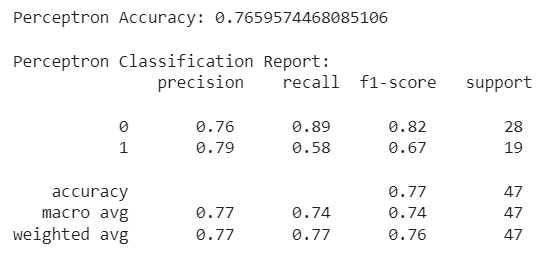
Experience in Current Domain: Experience in the current domain varies, with a range from 0 to 5 years.

Leave or Not: The dataset contains binary values (0 or 1) indicating whether an individual left their job (1) or not (0).

This dataset captures a diverse set of individuals’ employment-related information, making it suit- able for various analyses related to factors influencing employee attrition or retention.

## Perceptron Learning

Perceptron learning is a binary classification algorithm and one of the fundamental building blocks of neural networks. The perceptron is a simple model designed to classify data into two categories. It operates by learning a linear decision boundary that separates the classes in feature space. Start by initializing the weights and bias to small random values or zeros. For each training example, calculate the weighted sum of the input features. If the weighted sum is greater than a threshold (typically zero), the perceptron predicts one class; otherwise, it predicts the other class. If the prediction is incorrect, adjust the weights and bias to correct the error. The adjustment is proportional to the input and a learning rate, which controls the step size. Iterate through the training examples until the perceptron makes correct predictions for all data points or until a predefined number of iterations is reached. Perceptron learning is suitable for linearly separable data but may not converge for more complex, nonlinear problems. For such cases, multilayer neural networks (perceptron’s connected in layers) and backpropagation algorithms are used to address more intricate patterns and decision boundaries.

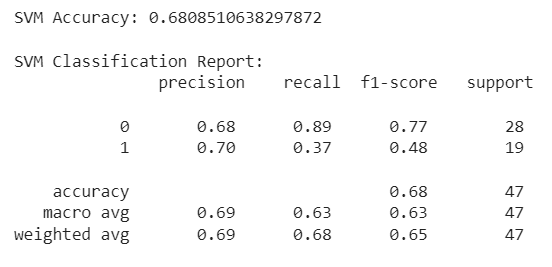


## Figure 2: perceptron learning.

## 2.3 Support Vector Machine

Support Vector Machine (SVM) is a powerful machine learning algorithm used for classification and regression tasks. SVM excels in finding optimal decision boundaries, particularly in scenarios with clear class separation. SVM aims to find the hyperplane that maximizes the margin between classes. This hyperplane is the one that maintains the maximum distance between the nearest data points of each class, called support vectors. SVM can handle linear and nonlinear data by using kernel functions that map the original feature space into a higher-dimensional space. This enables the discovery of complex decision boundaries. SVM's performance is influenced by hyperparameters and gamma. SVM focuses on generalization, aiming to make accurate predictions on new, unseen data, which is why it often provides robust results. SVM is primarily designed for binary classification, but it can be extended to handle multiclass problems.

SVM's ability to handle high-dimensional data, its robustness, and the versatility provided by kernel functions make it a popular choice for various machine learning tasks, from text categorization to image recognition.

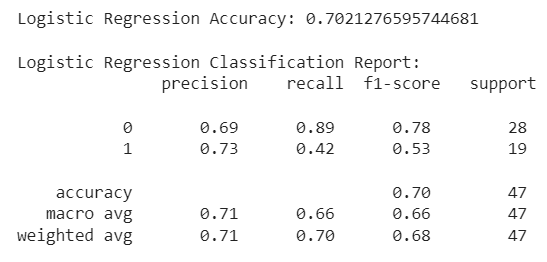


## 

## Figure 3: SVM

## Logistic Regression

Logistic Regression is a statistical model and machine learning algorithm used for binary classification tasks, where the goal is to predict one of two possible outcomes. It works by estimating the probability of the binary response variable based on a set of predictor variables. Logistic Regression uses the logistic function to transform a linear combination of input features into a value between 0 and 1, representing the probability of the positive class. The model is trained using a logistic loss function, and it outputs a decision boundary that separates the two classes. It is widely used in areas like medical diagnosis, marketing, and risk assessment due to its simplicity and interpretability.



## K - nearest neighbors

K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for classification and regression tasks. In KNN, data points are classified or predicted based on the majority class or average value of their 'k' nearest neighbors in a feature space. It assumes that similar data points are close to each other. To make predictions, KNN calculates distances (usually Euclidean) between data points and selects the 'k' nearest neighbors. In classification, it assigns the class most common among the neighbors, while in regression, it calculates the average of their values. The choice of 'k' influences the model's performance, with smaller 'k' values leading to more flexible, noise-sensitive models, and larger 'k' values producing smoother decision boundaries.

## Bootstrapping

Bootstrapping is a resampling technique in statistics and machine learning used to estimate the sampling distribution of a statistic. It involves repeatedly sampling a dataset with replacement to create multiple resampled datasets of the same size. Each resampled dataset is treated as a new sample from the population. This process allows for the assessment of the variability and uncertainty of a statistic without the need for additional data collection. Bootstrapping is particularly useful for estimating confidence intervals, performing hypothesis tests, and evaluating the robustness of statistical models. It is a powerful and versatile tool for making inferences and understanding the reliability of statistical estimates.

# Results:

## Perceptron Learning

## Randomly resampling the training data (X\_train and y\_train) using the `resample` function to create new training sets (X\_train\_boot and y\_train\_boot). Training a Perceptron classifier on each bootstrapped training set, with a maximum of 100 iterations. Making predictions on the test data (X\_test) using the trained Perceptron classifier. Calculating the accuracy of the Perceptron model's predictions on the test data.

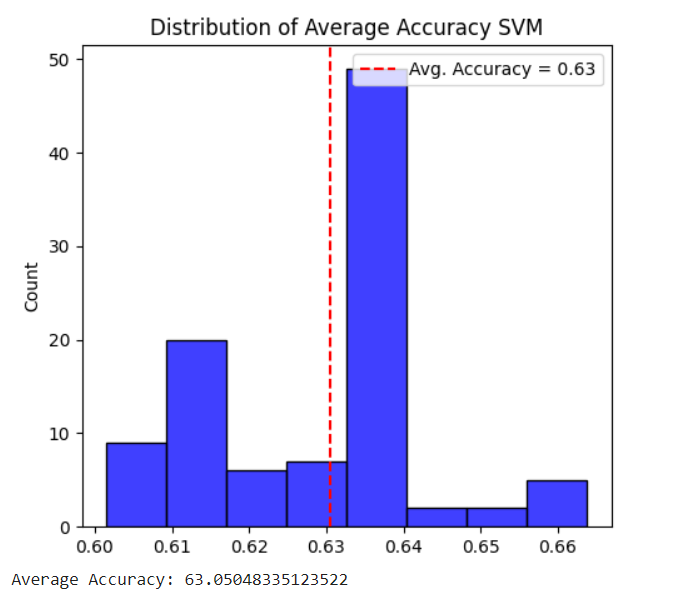
## Storing the accuracy score in the `results perceptron` list. This process is repeated 100 times to assess the variability in Perceptron model performance due to different bootstrapped training samples. The final `results\_perceptron` list contains accuracy scores for each iteration, allowing for the analysis of model performance across multiple resampled datasets.

## 

## 

* 1. **Support Vector Machine**

Random resampling of the training data (X\_train and y\_train) is done using the `resample` function with a fixed random state to create bootstrapped training sets (X\_train\_boot and y\_train\_boot).A Support Vector Machine (SVM) classifier with a linear kernel is trained on each bootstrapped training set, where SVM is a machine learning algorithm. Predictions are made on the test data (X\_test) using the SVM classifier. The accuracy of the SVM model's predictions on the test data is calculated. The accuracy score is stored in the `results\_svm` list. This process is repeated 100 times to assess the variability in SVM model performance due to different bootstrapped training samples. The final `results\_svm` list contains accuracy scores for each iteration, allowing for the analysis of model performance across multiple resampled datasets.

****

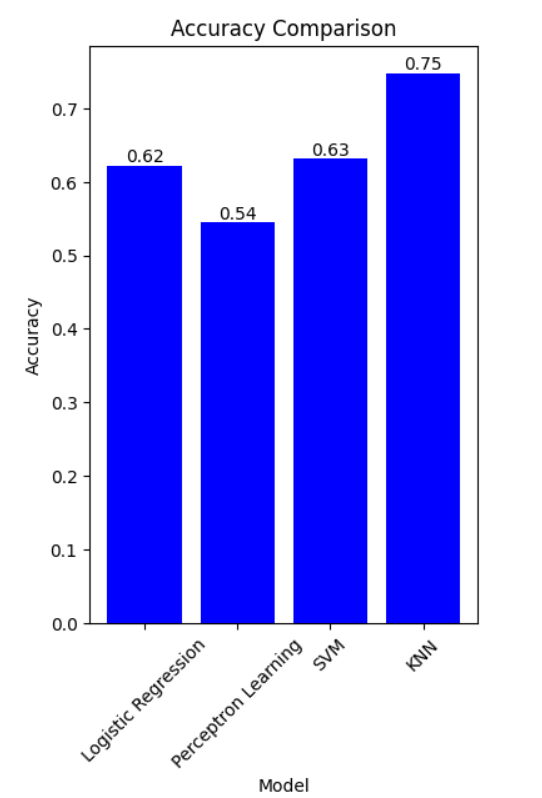
## 

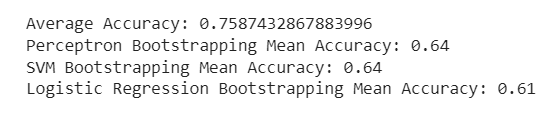
## Logistic Regression

## For each iteration, a new training set (X\_train\_boot and y\_train\_boot) is created by resampling the original training data (X\_train and y\_train) with a fixed random state (42). A Logistic Regression classifier is trained on the bootstrapped training data. Predictions are made on a test dataset (X\_test) using the trained Logistic Regression model. The accuracy of the Logistic Regression model's predictions on the test data is computed. The accuracy score is stored in the `results logistic` list, allowing for the assessment of model performance across different bootstrapped training datasets. This process is repeated 100 times to estimate the variability in the Logistic Regression model's accuracy due to different resampled training sets.

## 

* 1. **Bootstrapping Accuracies**

****



# 

# 4 CONCLUSION

# After running all the machine learning models like SVM, Perceptron Learning, Logistic Regression and KNN using bootstrapping the KNN

# Has highest average accuracy i.e. 0.75.

**5 REFERENCE**

**1. Ref :** [**refrence**](https://www.coursera.org/projects/employee-attrition-prediction?action=enroll)

**2. Ref :** [**capstone\_code**](https://github.com/KUSHAL113/statml.git)

THANK YOU