VIT

Vellore Institute of Technology

Oecmed io be University urvder section 3 ni' üGC Act 1956)

# EDA on Datasets of IBM Watson Employees (SDG – Goal 8)

Priyanka S | Kavya N | Hema B | Dr. Asnath Victy Phamila Y | SCOPE

## Motivation/ Introduction

This project is based on SDG Goal 8 – "Decent Work and Economic Growth," focusing on employee attrition and workplace satisfaction. Using the IBM Watson Employee dataset, we aim to uncover patterns and drivers behind why employees leave organizations. Exploratory Data Analysis (EDA) was applied to visualize data distributions, identify key variables, and detect patterns that influence attrition. The purpose of this study is to help businesses retain employees by understanding underlying data-driven insights.

## SCOPE of the Project

This study performs an in-depth EDA on a real-world dataset containing 1676 records and 35 attributes. The project explores missing value treatments using mean, KNN, and MICE imputation methods, identifies outliers via Z-score and IQR, and uses visual tools like histograms, boxplots, and heatmaps. Additionally, feature selection techniques such as ANOVA, Mutual Information, and RFE were used to isolate the most important attributes. Pattern mining was done using Apriori to discover rules in employee behavior.

## Methodology

The dataset was imported using Google Colab and examined for missing values, which were found to be zero. Imputation techniques (mean, KNN, MICE) were demonstrated for robustness, though not required. Outlier detection was performed using Z-score and IQR methods, with further visualization using KMeans clustering. Histograms showed distribution of numerical features; bar charts were used for categorical ones. Feature correlations were assessed using a heatmap. ANOVA and Mutual Information helped identify top features influencing attrition such as Age, Total Working Years, Job Level, and Years at Company. Dimensionality reduction techniques PCA and LDA provided visual separability of classes. A simple linear regression model was used to analyze the relation between Age and Monthly Income. Frequent pattern mining with the Apriori algorithm uncovered associations among departments, roles, and marital status.

The equation Yi=β0+β1XiY_i = \beta_0 + \beta_1 X_iYi=β0+β1Xi represents a **Simple Linear Regression model** used to explore the relationship between an employee's **Age** and their **Monthly Income**. Here, YiY_iYi denotes the monthly income of the *i-th* employee, and XiX_iXi represents the age of that employee. The term β0\beta_0β0 is the intercept, indicating the expected income when the employee's age is zero, which is not meaningful in a practical sense but is mathematically necessary for the model. The slope β1\beta_1β1 shows how much the income is expected to increase with each additional year of age.

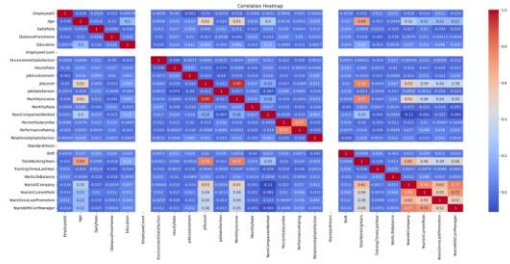
This model was applied in the project to understand whether age could be used as a predictor for income. A regression line was fitted to the scatter plot of Age vs. Monthly Income to visualize this trend. The model returned an R2R^2R2 value of approximately **0.31**, meaning that only about 31% of the variation in monthly income can be explained by age alone. Although the positive slope indicated that older employees tend to earn more, the wide spread of data points around the regression line suggested that age is **not a strong predictor** on its own.

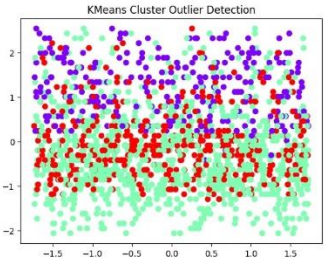
This analysis highlighted the complexity of salary structures in organizations. Factors like job level, total working years, department, and performance rating likely play a larger role in determining income. The regression equation provided a useful but limited view into this relationship, reinforcing the need for more complex models or multiple regression to capture the full picture.

$$Y_i = \beta_0 + \beta_1 X_i$$

## Results

No missing values or duplicates were found in the dataset. Most employees (88%) had no attrition. Key numerical features such as Total Working Years and Years With Current Manager showed strong influence on attrition. Z-score and IQR revealed outliers in MonthlyIncome, YearsSinceLastPromotion, and TrainingTimesLastYear. The correlation matrix revealed that YearsAtCompany correlated highly with YearsInCurrentRole and YearsWithCurrManager. Feature selection revealed Age, JobLevel, and MonthlyIncome as strong predictors. PCA and LDA confirmed class separability for both Attrition and Performance Rating. Regression analysis indicated a moderate correlation ($R^2 = 0.31$) between Age and Monthly Income. Association rules such as "If Department is Marketing, then Cardiology" were discovered with high lift (3.15), indicating a strong connection between departments.






## Conclusion/ Summary

This analysis found that younger employees, those with lower job levels, or fewer years in their roles were more likely to leave. These features can be used by HR teams to proactively address attrition. The data did not require much preprocessing, and visual tools helped in identifying patterns clearly. Association rules revealed key relationships between departments and roles. The study concludes that attrition is influenced by multiple overlapping factors and recommends deeper modeling for future predictions.

## Contact Details

priyanka.s2023c@vitstudent.ac.in

kavya.n2023@vitstudent.ac.in

hema.b2023@vitstudent.ac.in

## Acknowledgments/ References

IBM Watson Employee Dataset.
Tools used: Python, Pandas, Seaborn, Scikit-learn, MLxtend in Google Colab.
Poster aligns with SDG Goal 8 – Decent Work and Economic Growth.