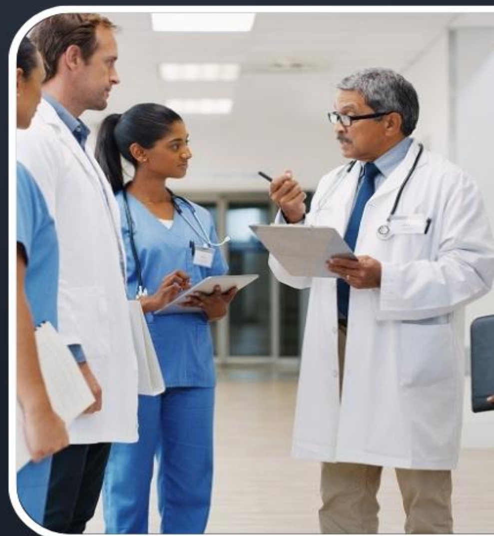
ANALYSIS OF EMPLOYEE ATTRITION FOR HEALTHCARE

Prepared for :
IS 4007



Prepared by :
Tharindu Darshana
s16341

Abstract

Employee attrition is a major concern in the healthcare sector due to its direct impact on service quality and operational efficiency. This study aims to develop a predictive model to identify employees at risk of leaving the organization based on various demographic and job-related features. Using an HR dataset from a healthcare context, exploratory data analysis (EDA) was performed to understand key patterns, followed by data preprocessing steps such as encoding, scaling, and handling class imbalance. Several classification models including Logistic Regression, Support Vector Machine, and XGBoost were applied and evaluated. Logistic Regression emerged as the best-performing model in predicting attrition, particularly in identifying the minority class, with the highest AUC-PR score of 0.7382 and accuracy of 90%. The analysis revealed that working overtime and younger age were the strongest predictors of attrition. Based on these findings, targeted recommendations were proposed to reduce turnover. The results of this study can help healthcare organizations implement proactive retention strategies and improve workforce stability.

Content

List of Figures and Tables

Introduction

Employee attrition—also known as employee turnover—is a critical issue faced by organizations. It refers to the scenario of reduction of the workforce in an organization due to employees leaving, typically without being immediately replaced. This scenario becomes particularly crucial in the healthcare sector. High attrition rates in healthcare sector can be a critical issue as it leads to patient safety and quality of care, loss of experienced professionals. Therefore, it's essential to understand the key factors behind attrition in order to develop effective retention strategies and maintain workforce stability.

The objectives of this study are,

- Perform Exploratory Data Analysis to uncover patterns associated with employee turnover
- To build a predictive model that accurately forecasts the likelihood of an employee leaving the organization. Hence identify key demographic and job-related factors that contribute to employee attrition in the healthcare sector.
- To provide actionable insights for HR and management teams to design targeted employee retention strategies.

Understanding and predicting employee attrition is very important in the healthcare sector, where workforce shortages can directly impact to patient safety and organizational performance. This study aims to provide valuable insights that can guide strategic decision-making in human resource management by identifying the key drivers. By acting as an early warning system, the predictive model can help healthcare organizations reduce operational disruptions, implement targeted retention initiatives, and retain a stable, experienced workforce

Literature Review

Employee attrition remains a significant issue for organizations due to high costs associated with hiring, training, and productivity loss. To mitigate this, recent research has focused on predicting attrition using machine learning techniques. Jain et al. (2020) applied decision tree (DT), support vector machine (SVM), and random forest (RF) models on a 14,999-record HR dataset, identifying key predictors such as satisfaction level, number of projects, and salary. Among these, the random forest model delivered the highest predictive performance, though specific accuracy metrics were not reported. Qutub et al. (2021) evaluated five algorithms—Logistic Regression (LR), Random Forest, Decision Tree, Adaboost, and Gradient Boosting—using the IBM HR Analytics dataset of 1,470 records. Their results showed that LR achieved the highest accuracy at 88.43%, with a recall of 0.46 and AUC of 0.8593. Ensemble models like DT+LR achieved slightly lower accuracy at 86.39% but offered better generalizability. In a healthcare-specific context, Egwom et al. (2024) used SMOTETomek to handle data imbalance in a modified IBM dataset with 1,676 samples, where only 199 indicated attrition. They tested RF, SVM, KNN, and XGBoost, with the SMOTETomek-enhanced Random Forest model achieving 98.0% accuracy, outperforming all others. This underscores the importance of data resampling and domain-specific adaptation in improving model reliability.

Across these studies, ensemble methods and logistic regression models consistently performed well, especially when paired with thoughtful preprocessing techniques. These findings guide the current study's approach, which will integrate various models with resampling strategies like SMOTE to ensure high accuracy and practical relevance in predicting employee attrition in healthcare.

Theory and Methodology

Theoretical Framework

Employee attrition can be treated as a binary classification problem, where the target variable indicates whether an employee has left the organization (Yes) or not (No). There are several statistical and machine learning methods suitable for analyzing this type of problem, including logistic regression, decision trees, and ensemble models. Before modeling, exploratory data analysis (EDA) and data preprocessing steps are necessary to understand the structure and relationships in the dataset.

Exploratory Data Analysis (EDA)

This is conducted to understand the data distributions, detect outliers and uncover relationships between features. Statistical summaries, correlation matrices, and visualization tools such as box plots, bar charts, and heatmaps are commonly used to perform this.

Logistic Regression

Logistic regression is a widely used statistical method for binary classification. It is preferred because of its simplicity and interpretability. It estimates the probability that a given input belongs to a particular category using a linear function.

Decision Tree and Random Forest

They are tree based non-parametric models that partition the data based on feature values to predict outcomes. Decision tree uses a single tree so it is interpretable and capable of handling non-linear relationships. However, they are prone to overfitting. On the other hand, Random forest is an ensemble learning method that uses multiple decision trees and combines their result to get the outcome.

XGBoost (Extreme Gradient Boosting)

This is an advanced ensemble learning algorithm based on gradient boosting. It builds a series of decision trees sequentially, where each new tree aims to correct the errors made by the previous ones minimizing a specified loss function using gradient descent techniques.

Methodology

The methodology of this study consists of several steps. They are Data preprocessing, Exploratory Data Analysis, Model Building and Evaluation and Feature Importance Analysis.

Initial Data Exploration and Cleaning

This step involves checking the overview of the dataset, i.e. checking the number of observations, number of columns, looking at the variable types and performing basic data type conversions. It also includes checking for the presence of missing values to assess data completeness and determine appropriate handling strategies.

Exploratory Data Analysis (EDA)

EDA was conducted to examine distributions and identify potential patterns. This is done by plotting suitable charts according to the variable type such as bar charts, histograms, scatter plots. This is carried out in two parts, univariate analysis and bivariate analysis. Univariate analysis involves analyzing a single variable at a time. Done to identify the data distribution. Bivariate analysis is done by analyzing two variables at a time, examining the relationship between the two variables

Feature Engineering and Final Preprocessing

In this stage new features were created using the existing features, encoding categorical features using one-hot encoding to convert them into a numerical format suitable for machine learning algorithms and numeric features were scaled to ensure that they are on a comparable scale, which is important for distance-based algorithms

Feature Importance Analysis

For the best performing model, feature importance was examined to identify the most influential factors that lead to employee turnover.

Data

This dataset contains details about employees in the healthcare sector, obtained from the Kaggle platform, including information on their demographics as well as job related factors. Containing details about 1676 employees across 35 features. The following is the description of the variables that are most related to this study.

- Age - Age of the employee (in years)
- Attrition - Whether the employee has left the company – the target variable (Yes/No)
- BusinessTravel - Frequency of business travel (e.g., Rarely, Frequently)
- Department - Department to which the employee belongs (Cardiology, Maternity, Neurology)
- DistanceFromHome - Distance from the employee’s home to the workplace (in km)
- Education - Education level
- EducationField - Field of study (Life Sciences, Medical, Marketing etc.)
- EnvironmentSatisfaction - Satisfaction with the work environment (1 = Low, 4 = Very High)
- Gender - Gender of the employee (Male/Female)
- JobInvolvement - Level of involvement in the job (1 = Low, 4 = Very High)
- JobRole - Job title or role (e.g., Nurse, Manager, Lab Technician)
- JobSatisfaction - Level of job satisfaction (1 = Low, 4 = Very High)
- MaritalStatus - Marital status (e.g., Single, Married, Divorced)
- MonthlyIncome - Monthly income of the employee
- OverTime - Whether the employee works overtime (Yes/No)
- PercentSalaryHike - Percentage increase in salary compared to previous year
- PerformanceRating - Performance score (1 = Low, 4 = Excellent)
- Shift - Employee’s work shift
- TotalWorkingYears - Total number of years of professional experience
- YearsAtCompany - Number of years the employee has been with the company
- YearsInCurrentRole - Number of years in the current job role
- YearsSinceLastPromotion - Number of years since the last promotion

Quantitative variables

Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, StandardHours, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager

Qualitative variables

Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18, OverTime, Education, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobSatisfaction, RelationshipSatisfaction, WorkLifeBalance, PerformanceRating, Shift

The dataset does not have any missing values, and there are no issues with the data types of the variables. Therefore, the initial data preprocessing was not carried out.

Exploratory Data Analysis

This section aims to gain a deeper understanding of the distribution of the variables in the dataset, detect anomalies and uncover relationships between variables. The analysis was carried out in two main parts: Univariate analysis and bivariate analysis. Univariate analysis focused on examining the distribution and characteristics of individual variables and bivariate analysis investigated the relationship between the target variable (Attrition) and other features.

Univariate Analysis

<u>Distribution of Age</u>

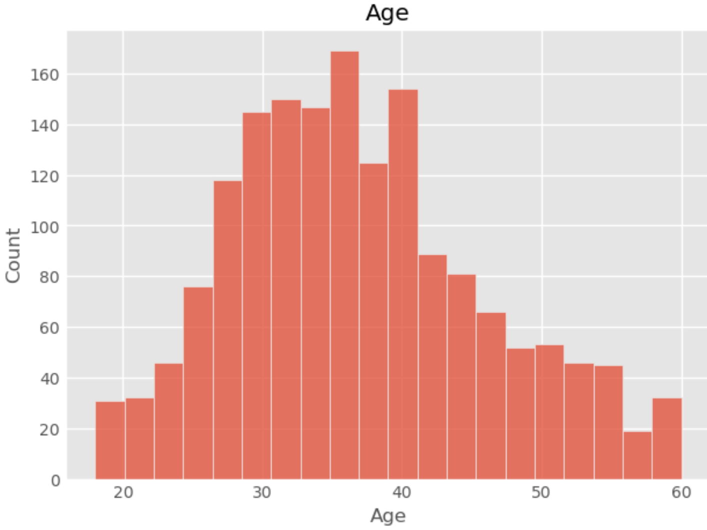


Figure 1 - Distribution of Age

<u>Distribution of Monthly Income</u>

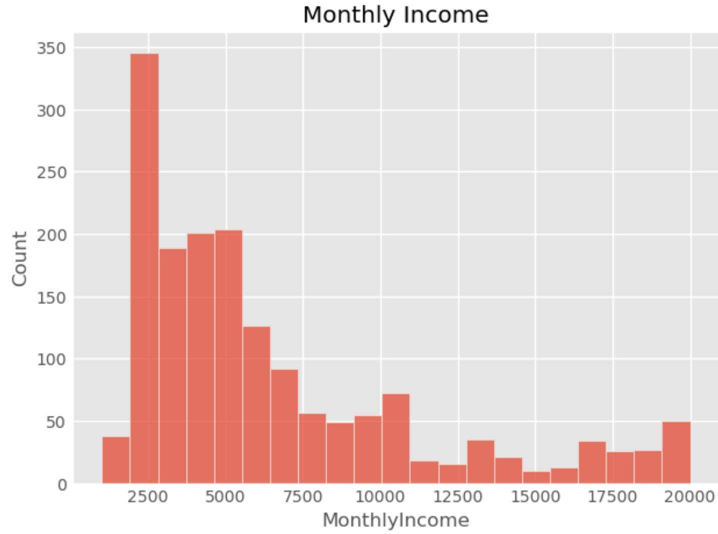


Figure 2 - Distribution of Monthly income

The distribution of the age approximately follows a normal distribution, with the majority falling within the mid-age range of 30 to 40 years.

Monthly income distribution is right skewed, meaning that the majority of employees receive a low income while few employees receive a high income

<u>Distribution of Distance from Home</u>

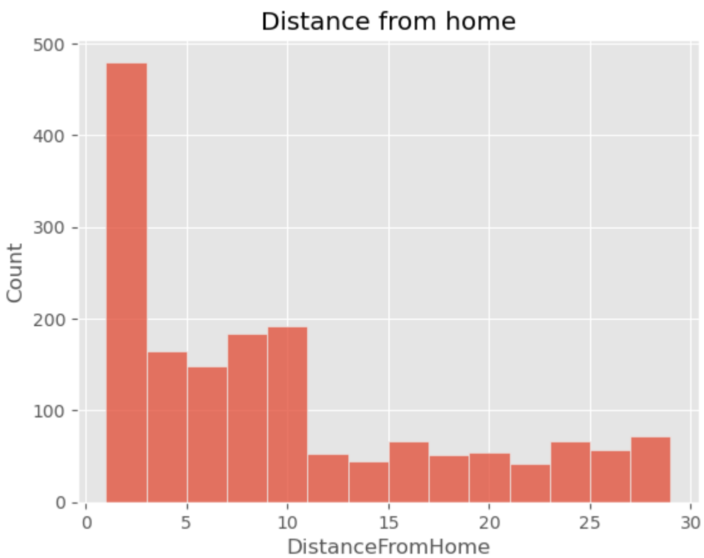


Figure 3 - Distribution of Distance from home

<u>Distribution of Years in current role</u>

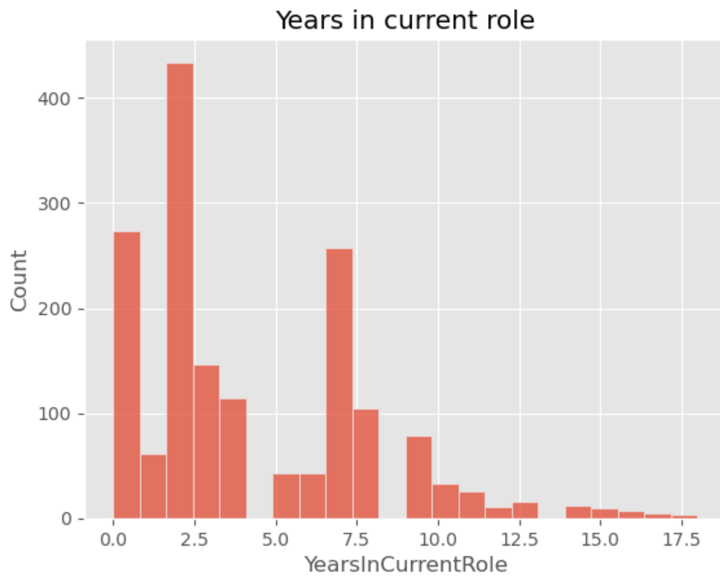


Figure 4 - Distribution of years in current role

Distribution is right skewed, indicating that most employees live relatively close to the workplace, while few employees have high distance

Distribution is right skewed, meaning that while few employees stays within the same role for long time, the majority stayed short time within the same role

The following are some distributions of categorical variables

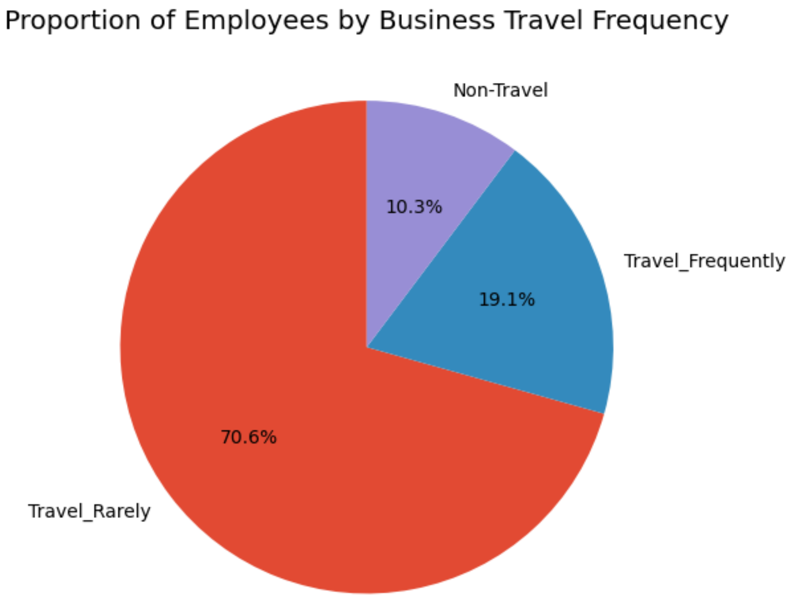


Figure 5 - Distribution of employees by business travel

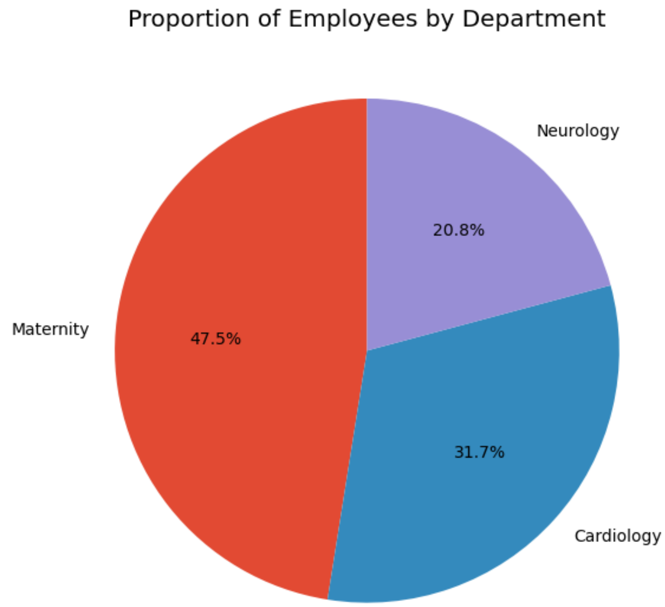


Figure 6 - Distribution of employees by Department

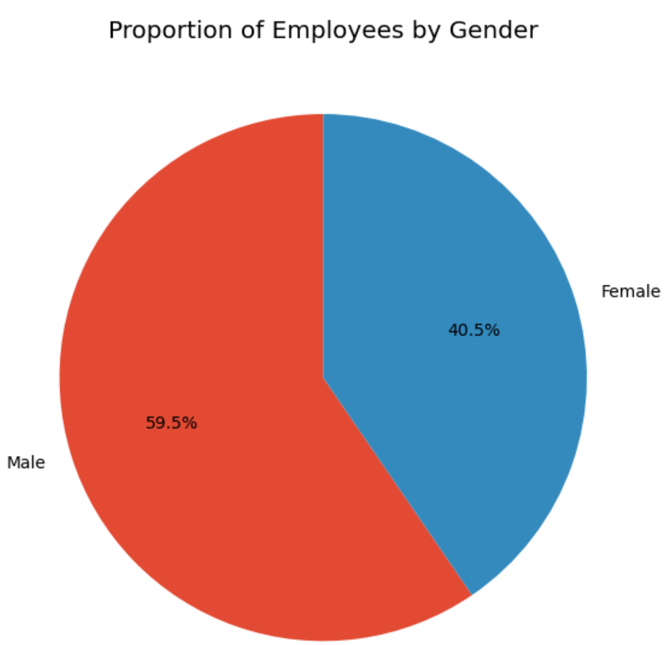


Figure 7 - Distribution of employees by gender

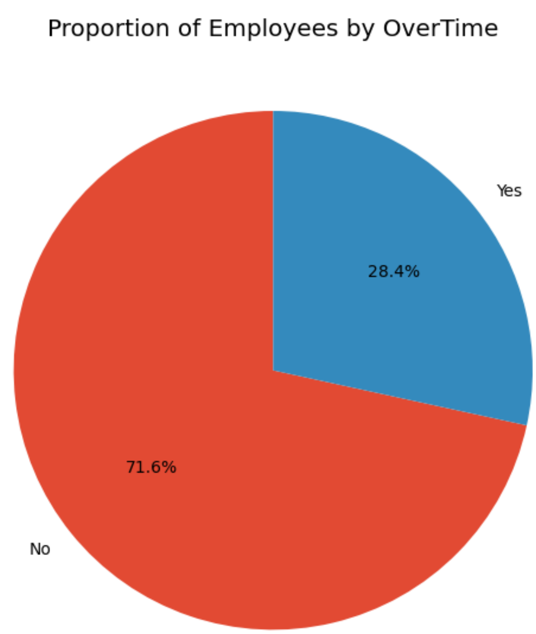


Figure 8 - Distribution of employees by Overtime

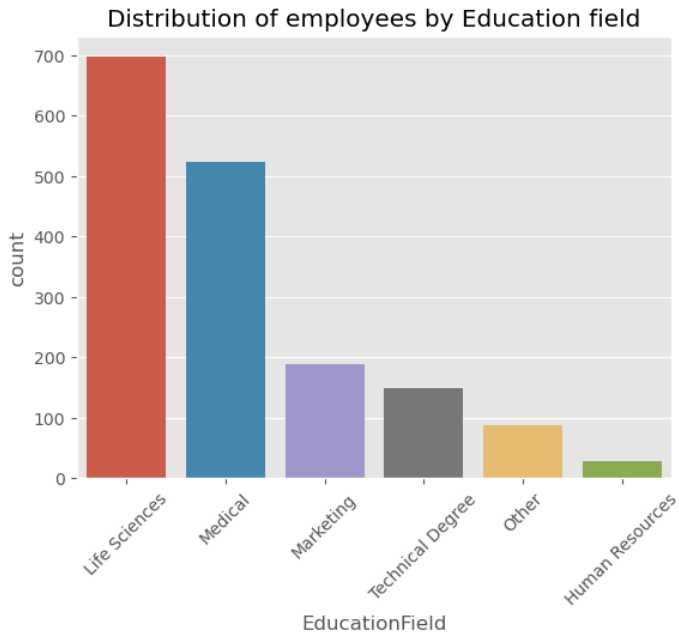


Figure 10 - Distribution of employees by Education field

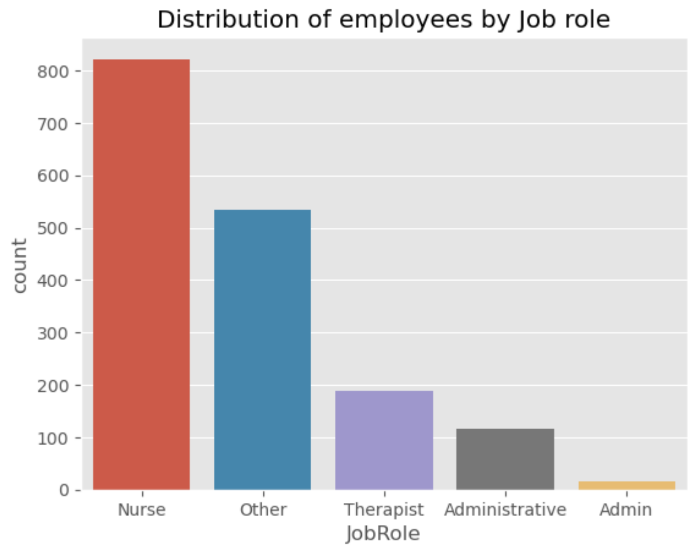


Figure 9 - Distribution of employees by Job role

Bivariate Analysis

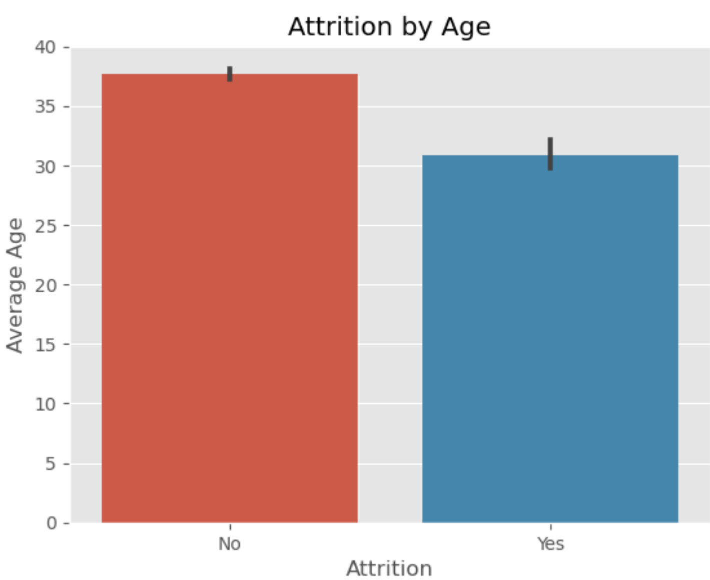


Figure 11 - Attrition by Age

It is observed that employees who left the organization tend to be younger on average compared to those who stayed.

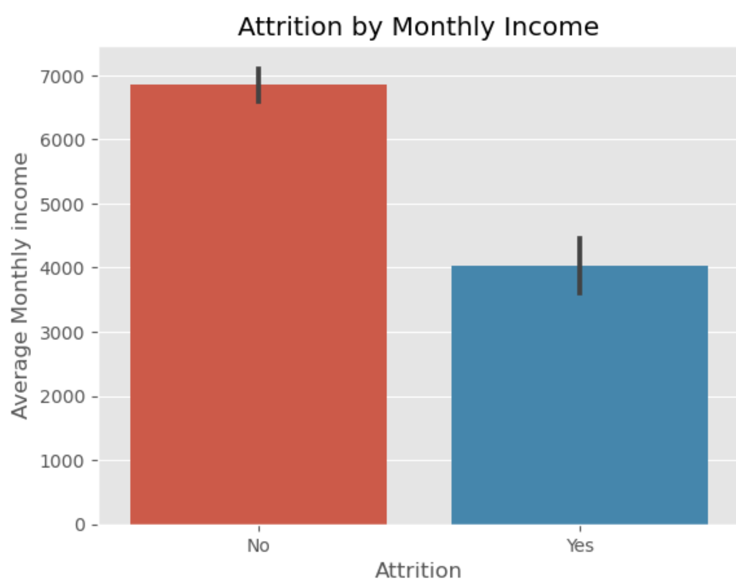


Figure 12 - Attrition by monthly income

It can be seen that employees who left the organization had less monthly income than the employees those who stayed

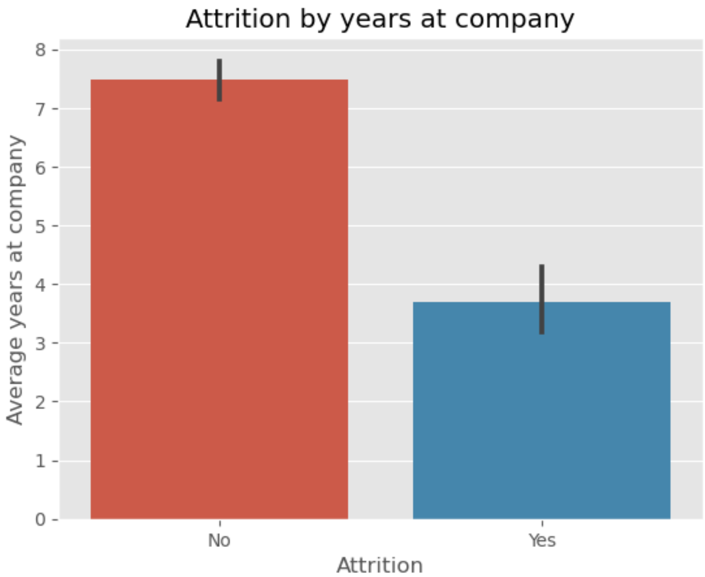


Figure 13 - Attrition by years at company

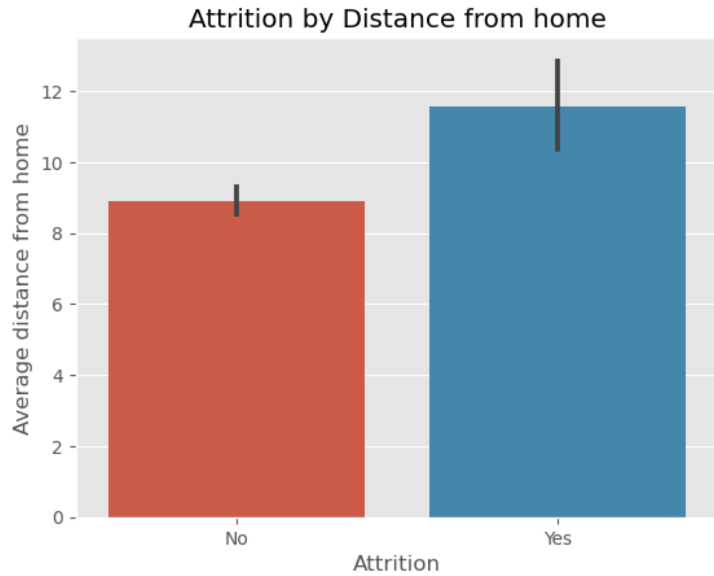


Figure 14 - Attrition by Distance from home

It is observed that the average number of years at the company is less for those who left the organization compared to those who remained. Indicating seniors in the organization tend to stay.

The average distance from home is higher in those who leave the company. Raising the question that a high distance from home might be a factor in employee turnover

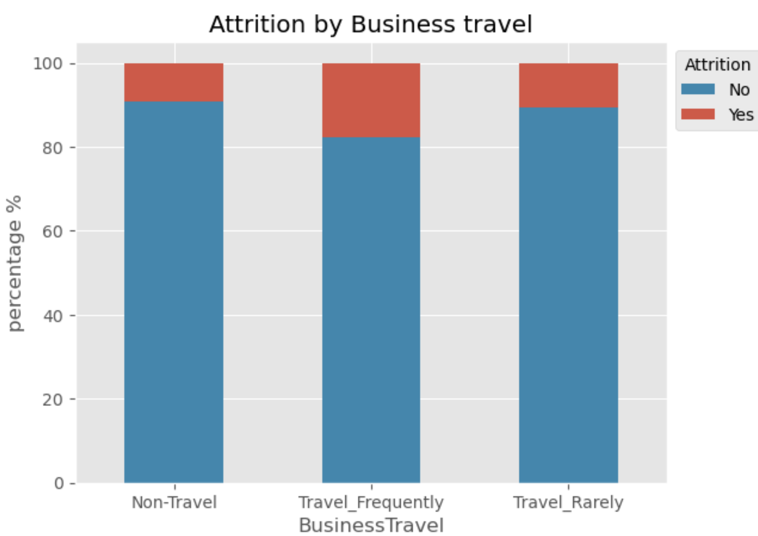


Figure 15 - Attrition by Business level

It is observed that the attrition percentage is higher among the employees who travel frequently than among employees who travel rarely and non-travel employees

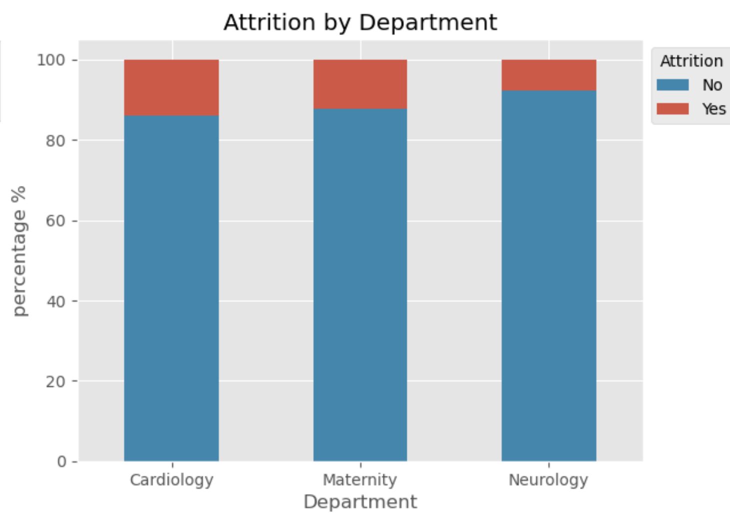


Figure 16 - Attrition by Department

There is not much difference in the attrition percentage among the departments. However it is little bit less in Neurology department

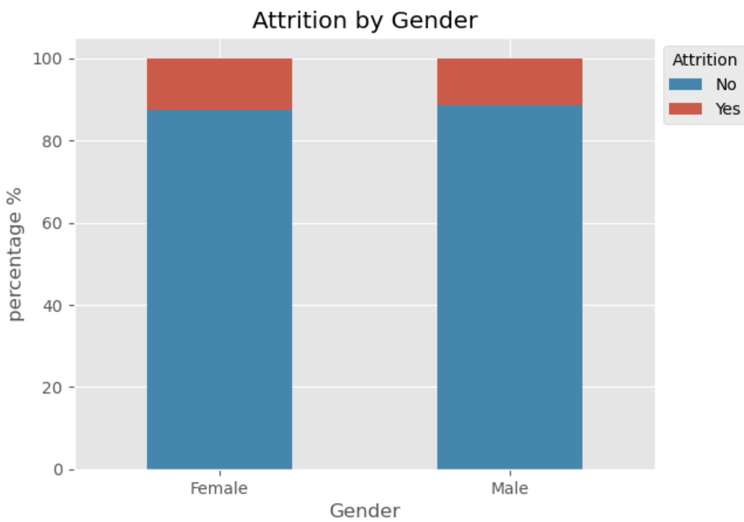


Figure 17 - Attrition by Gender

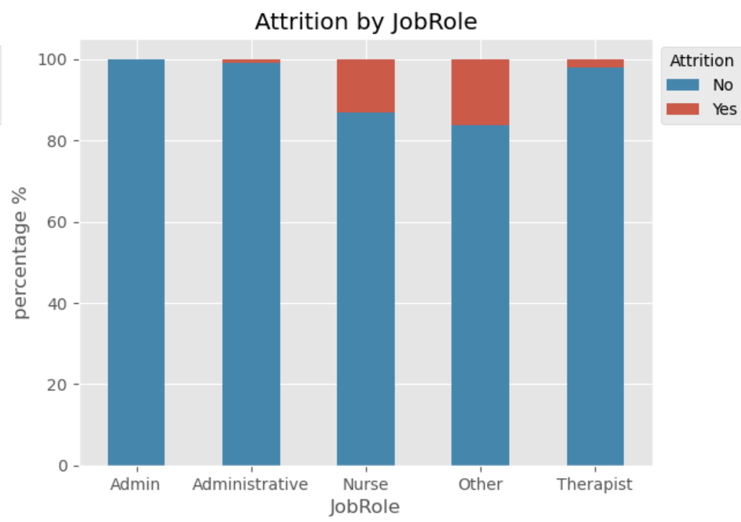


Figure 18 - Attrition by Job role

There is little to no difference in attrition rates between the two genders.

It is observed that Nurses have a high attrition rate compared to other job roles. Apart from the defined job roles, employees who have other job roles experience high attrition rate

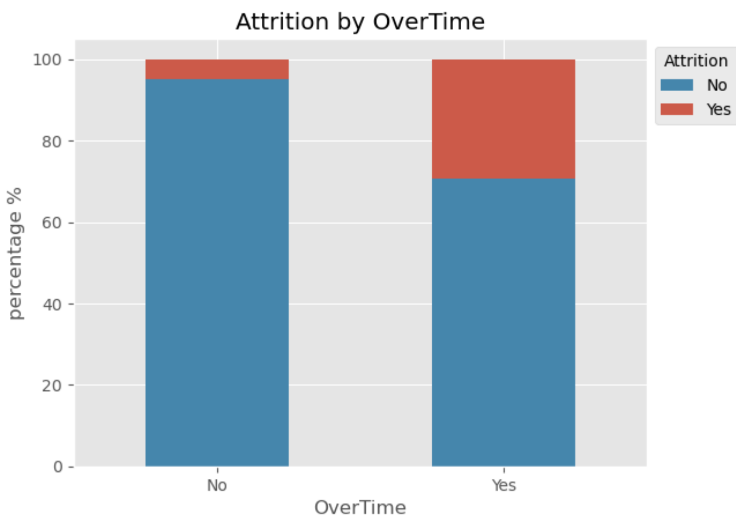


Figure 20 - Attrition by Overtime

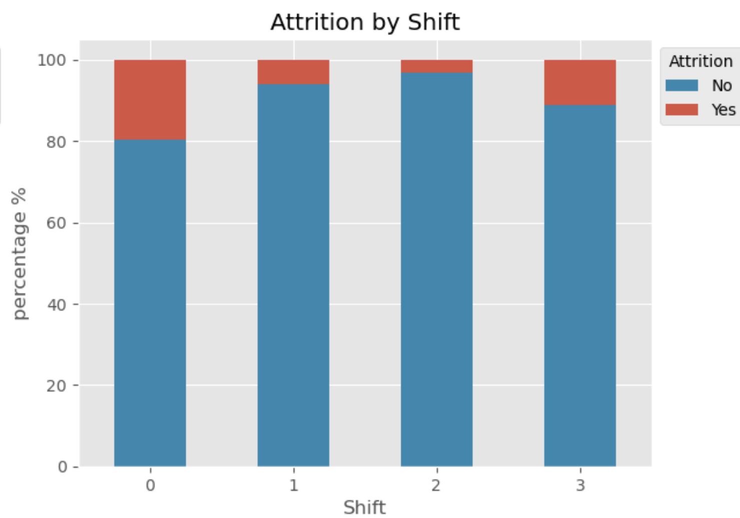


Figure 19 - Attrition by Shift

It is clear that employees who do overtime have a higher attrition rate than those who do not do overtime

Among the employees who are doing shift no.0, have a higher attrition rate compared to other shifts. The percentage is approximately 20% of those who are doing shift 0. The second highest attrition rate was observed among the employees doing shift 3.

Advanced Analysis

In this phase, various statistical and machine learning models were developed to predict employee attrition based on the given features. The primary objective was to build a reliable and interpretable model that can help the organization proactively identify employees at risk of leaving. Several classification algorithms were explored, including logistic regression, decision trees, random forest, and XGBoost. Model performance was evaluated using appropriate metrics such as accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), with special attention given to handling class imbalance through techniques like resampling and class weighting.

Model Building

First, the dataset was split into parts, the training set and the testing set such that 80% of the observations are for the training set and 20% of the observations are for the testing set. After that, to handle the imbalance, SMOTE (Synthetic Minority Oversampling Technique) was applied only to the training set. Categorical variables were encoded and numerical variables were scaled before being fed into the machine learning models. After that, various statistical and machine learning models were trained on the training set and evaluated their performance on the testing set.

Since the **Logistic Regression, Support Vector Machine, and XGBoost** models demonstrated high predictive accuracy, these three were selected for further improvement through hyperparameter tuning to optimize their performance. After performing hyperparameter tuning on these three models, classification reports were to examine the performance of the three models.

model	score
LogisticRegression()	0.902
DecisionTreeClassifier()	0.830
(DecisionTreeClassifier(max_features='sqrt', r...	0.878
SVC()	0.890
KNeighborsClassifier()	0.768
XGBClassifier(base_score=None, booster=None, c...	0.893

Table 1 - Performance of models

XGBoost (Extreme Gradient Boosting)

	precision	recall	f1-score	support
0	0.93	0.95	0.94	289
1	0.66	0.57	0.61	47
accuracy			0.90	336
macro avg	0.80	0.76	0.78	336
weighted avg	0.89	0.90	0.90	336

Table 2 - Classification report of XGBoost model

Logistic regression

	precision	recall	f1-score	support
0	0.95	0.93	0.94	289
1	0.61	0.70	0.65	47
accuracy			0.90	336
macro avg	0.78	0.81	0.80	336
weighted avg	0.90	0.90	0.90	336

Table 3 - Classification report of Logistic regression

Support Vector Machine

	precision	recall	f1-score	support
0	0.94	0.94	0.94	289
1	0.63	0.66	0.65	47
accuracy			0.90	336
macro avg	0.79	0.80	0.79	336
weighted avg	0.90	0.90	0.90	336

Table 4 - Classification report of Support Vector Machine

It is observed that the accuracy of the three models is almost equal. But the logistic regression does a better job of predicting the positive class than the other two models. To examine this better, Precision-Recall curves were generated for the top three models to evaluate their performance in handling class imbalance.

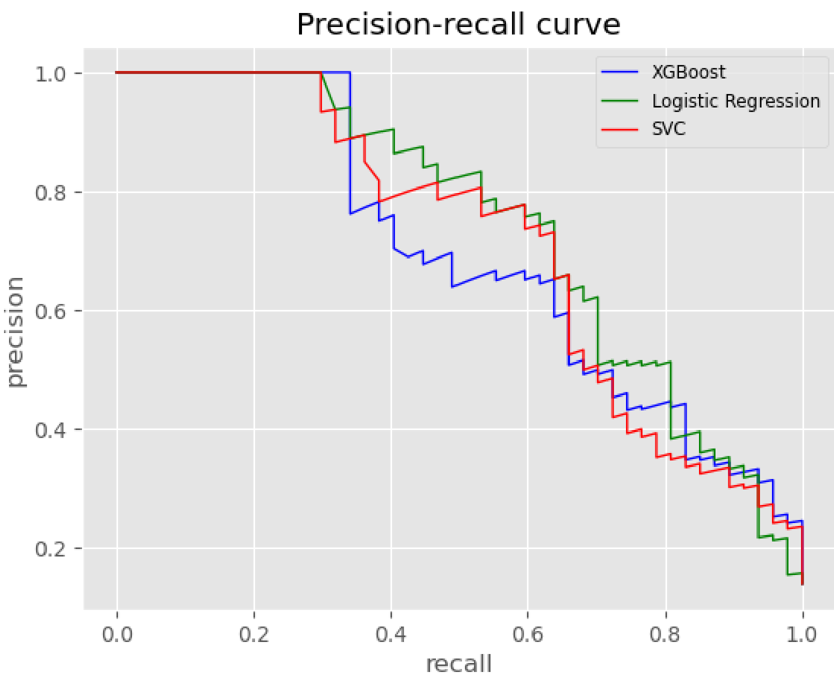


Figure 21 - Precision-recall curve

```
AUC-PR of XGBoost:  0.6908
AUC-PR of Logistic Regression:  0.7382
AUC-PR of SVC:  0.7096
```

Table 5 - AUC – PR of models

Among the three models evaluated, **Logistic Regression** achieved the highest Area Under the Precision-Recall Curve (AUC-PR) with a score of **0.7382**, indicating its high ability to correctly identify employees at risk of attrition. This suggests that Logistic Regression is more effective in balancing precision and recall, making it the most reliable model for detecting the positive class in an imbalanced setting. Therefore, **Logistic Regression** was chosen as the best performing model.

After selecting the best model, it is of interest that find a better threshold that balances the precision and recall of the positive class. Therefore, the new threshold was found, and it is 0.7. That is, if the probability of an employee leaving the organization is higher than 70%, the model classifies it as a 'Yes' otherwise 'No'. The default threshold of the model was 50%. That is if the probability of an employee leaving the organization is higher than 50%, the model classifies it as a 'Yes' otherwise 'No'. To compare these two situations, a classification report was obtained for the predicted results obtained from the new threshold value.

	precision	recall	f1-score	support
0	0.93	0.98	0.95	289
1	0.79	0.55	0.65	47
accuracy			0.92	336
macro avg	0.86	0.76	0.80	336
weighted avg	0.91	0.92	0.91	336

Table 6 - classification report of logistic regression for the new threshold

It can be seen that although the accuracy was higher than the default threshold, there is a significant decrease in the recall of the positive class, meaning that a considerable number of actual attrition cases were not correctly identified by the model. Since we are interested in predicting employee attrition as much as possible, the default threshold was preferred, where there was a 70% recall for the positive class.

The obtained model formula can be written as follows,

$\ln\left(\frac{p}{1-p}\right)$ = 2.73 + 1.91(BusinessTravel_Travel_Frequently) + 0.99(BusinessTravel_Travel_Rarely)

-0.78(Department_Maternity) – 0.85(Department_Neurology) + 0.82(EducationField_Life Sciences) + 0.98(EducationField_Marketing) + 0.11(EducationField_Medical) - 0.67(EducationField_Other) + 0.22(EducationField_Technical Degree) – 0.06(Gender_Male)

-0.36(JobRole_Administrative) + 2.08(JobRole_Nurse) + 1.65(JobRole_Other) – 0.63(JobRole_Therapist) + 0.27(MaritalStatus_Married) + 1.56(MaritalStatus_Single) + 4.27(OverTime_Yes) – 3.14(Age) – 0.58(DailyRate) + 2.23(DistanceFromHome) – 0.38(HourlyRate) + 0.23(MonthlyIncome) + 0.37(MonthlyRate) + 1.41(NumCompaniesWorked)

+ 0.15(PercentSalaryHike) – 2.98(TotalWorkingYears) – 1.35(TrainingTimesLastYear) – 1.76(YearsAtCompany) – 2(YearsInCurrentRole) + 0.9(YearsSinceLastPromotion) – 1.32(YearsWithCurrManager) + 0.05(Education) – 0.43(EnvironmentSatisfaction) – 0.73(JobInvolvement) - 1.15(JobLevel) – 0.27(JobSatisfaction) – 0.12(PerformanceRating) – 0.1(RelationshipSatisfaction) – 0.88(Shift) – 0.22(WorkLifeBalance)

; where p is the probability of an employee leaving the organization (Attrition = ‘Yes’)

Feature Importance

Since Logistic Regression was identified as the best-performing model, feature importance was determined based on the magnitude of its model coefficients. These coefficients reflect the impact of each feature on the likelihood of employee attrition. Features with larger absolute coefficient values were considered more influential.

The following table depicts the most important features and corresponding coefficients and their absolute values.

Feature	coefficient	|coefficient|
OverTime_Yes	4.271549	4.271549
Age	-3.144287	3.144287
TotalWorkingYears	-2.981486	2.981486
DistanceFromHome	2.233094	2.233094
JobRole_Nurse	2.080806	2.080806
YearsInCurrentRole	-1.997434	1.997434
BusinessTravel_Travel_Frequently	1.910274	1.910274
YearsAtCompany	-1.760616	1.760616
JobRole_Other	1.653409	1.653409
MaritalStatus_Single	1.562822	1.562822
NumCompaniesWorked	1.413706	1.413706
TrainingTimesLastYear	-1.349605	1.349605
YearsWithCurrManager	-1.322894	1.322894
JobLevel	-1.147641	1.147641
BusinessTravel_Travel_Rarely	0.993517	0.993517
EducationField_Marketing	0.984094	0.984094
YearsSinceLastPromotion	0.901604	0.901604
Shift	-0.884693	0.884693
Department_Neurology	-0.853201	0.853201
EducationField_Life Sciences	0.823457	0.823457
Department_Maternity	-0.783098	0.783098
JobInvolvement	-0.734975	0.734975
EducationField_Other	-0.668026	0.668026
JobRole_Therapist	-0.630650	0.630650
DailyRate	-0.581271	0.581271
EnvironmentSatisfaction	-0.430909	0.430909
HourlyRate	-0.384878	0.384878

Table 7 - Model coefficients

The most influential factor for employee attrition is doing overtime. Doing overtime increases the log-odds of attrition by 4.27. The second most influential factor is the age of the employee. The year increase in employee age reduces the log-odds of attrition by 3.14. And the other most influential factors are Total working years, distance from home, being a nurse, years in current role and frequent business travel.

General Discussion and Conclusion

- The project aimed to predict employee attrition in the healthcare sector using various machine learning models.
- Exploratory data analysis (EDA) helped uncover patterns and relationships, such as higher attrition among younger employees and those working overtime.
- The dataset was imbalanced, with fewer cases of attrition compared to non-attrition. This was addressed using the SMOTE technique before providing them to the machine learning models.
- After training various models, best performing models were selected to proceed to do hyperparameter tuning on them. After tuning hyperparameters, models were evaluated using evaluation metrics and Logistic Regression showed the best performance in identifying the minority class, with the highest AUC-PR (0.7382).
- Feature importance analysis based on logistic regression coefficients revealed that OverTime, Age, Total working years, and Job role were significant predictors of attrition.
- Since 'OverTime' was a key predictor of attrition, steps should be taken to reduce excessive workloads and promote better work-life balance.
- Younger employees and those with fewer years at the company showed higher attrition rates. Mentorship programs and growth opportunities can improve their engagement and retention.
- Those with fewer total working years and shorter tenure in current roles may feel less stable. Provide growth opportunities and recognize early contributions.
- Since distance from home appears as a major factor for attrition, organization can arrange flexible work arrangements, or provide transportation support to those employees.
- Nurses were identified as a high-risk group for attrition. Management should pay attention to enhancing their job satisfaction by reducing their stress and workload. And increase the staff to reduce the workload if needed.
- Since frequent travel is associated with higher attrition, providing logistical support and accommodations can also help reduce the burden on frequently travelling staff.

References

Jessica, N. E. O. ., Emmanuel, N. L., Hambali, N. M. A., & Galadima, N. K. R. (2024). An intelligent analysis and prediction of employee attrition rate in healthcare using machine learning techniques. *International Journal of Emerging Multidisciplinaries Computer Science & Artificial Intelligence*, *3*(1). https://doi.org/10.54938/ijemdcsai.2024.03.1.352

Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 86. https://doi.org/10.3390/computers9040086

Predicting Employee Attrition using Machine Learning. (2018, November 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8605976/

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of employee attrition using machine learning and ensemble methods. *International Journal of Machine Learning and Computing*, *11*(2), 110–114. https://doi.org/10.18178/ijmlc.2021.11.2.1022

Jain, P. K., Jain, M., & Pamula, R. (2020). Explaining and predicting employees' attrition: a machine learning approach. *SN Applied Sciences*, *2*(4). https://doi.org/10.1007/s42452-020-2519-4

Dataset Source: Employee Attrition for Healthcare

https://www.kaggle.com/datasets/jpmiller/employee-attrition-for-healthcare?select=watson_healthcare_modified.csv