Predicting Employee Attrition in an Organization Through Advanced Data Mining Technique

Yasmin P. Zacarias
San Carlos College
Talang, San Carlos City,
Pangasinan, Philippines

Millbert S. Secretario
San Carlos College
Bacnar, San Carlos City,
Pangasinan, Philippines

Dexter C. Macaraeg
San Carlos College
Pallas, Binmaley,
Pangasinan, Philippines

Tracy Anne M. Agbuya
San Carlos College
Urbiztondo, Pangasinan,
Philippines

Jenniea A. Olalia
San Carlos College
San Carlos City,
Pangasinan, Philippines

Maynard Gel F. Carse
San Carlos College
Abanon, San Carlos City
Pangasinan, Philippines

ABSTRACT
Employees play a big role in an organization, and they greatly contribute to its success and functioning. Every level from operational tasks to strategic decision-making, their collective efforts contribute to achieving the organization's mission, vision, and objectives. If the attrition rate of employees continuously increases that will be a big problem for the company. Understanding and forecasting turnover at the firm and departmental levels is essential for reducing attrition as well as for effectively planning, budgeting, and recruiting in the human resource field [6].

Advanced data mining techniques help organizations predict attrition proactively to address workforce stability by leveraging insights derived from historical data. In this study, the proponents identified key predictors of employee attrition using feature selection methods, specifically Recursive Feature Elimination (RFE) and SelectKBest. After evaluating both methods with Random Forest and SVM models, the Random Forest model combined with RFE achieved the highest overall performance with an accuracy of 84.2% and a precision of 0.700. This combination offered the most reliable balance, making it a valuable tool for organizations to more accurately identify potential attrition risks.

Keywords
Employee Attrition, Recursive Feature Elimination, SelectKBest, Random Forest, Support Vector Machine

1. INTRODUCTION
Predicting employee's attrition manually basically rely on the subjective interpretation of human judgment and the analysis of qualitative data, frequently leading to a prolonged process characterized by a restricted scope and the possibility of inherent bias. Relying on human assessment can introduce variability and subjectivity, extending the time required for analysis and might constrain the breadth of insights that can be gleaned. Moreover, the qualitative nature of the data may not capture the full spectrum of factors influencing employee attrition. Limiting the effectiveness of manual approaches in predicting and addressing turnover risks comprehensively.

In this study the aim is to develop a predictive model for employee attrition using data mining techniques. The framework involves several key stages, starting with acquiring and preprocessing an employee dataset, ensuring it's comprehensive and clean for analysis. To understand the distribution of variables, outliers are identified, and explore potential patterns or trends related to attrition by conducting exploratory data analysis. Subsequently, feature selection methods are employed such as Recursive Feature Elimination (RFE) and Select K Best to identify the most relevant predictors of employee turnover. Some of the features in the dataset are categorical which in most machine learning algorithms cannot be used directly. To solve this problem, a data encoding technique will help convert categorical values into numerical ones. After preparing the data, it is split into training and testing sets to evaluate the performance of the predictive models. Following this, various advances machine learning based techniques such as Random Forest and Support Vector Machine (SVM) are trained and evaluated using cross-validation techniques to predict attrition.

2. RELATED LITERATURE
This study was guided by several related works that served as important references. Among these is the proposed project by R. Sihva Shankar entitled "Prediction Of Employee Attrition Using Datamining" examine the factors driving employee turnover, from workplace dynamics to personal issues. Utilizing different data classification methodologies on the IBM HR Employee Attrition dataset, the study aims to identify and predict attrition triggers. By analyzing 1470 records with 35 features, the study extracts relevant characteristics, segregates the dataset into training and test sets, and applies classification techniques to provide insights for minimizing turnover [8].

Another study proposed by Ali Raza entitled "Predicting Employee Attrition Using Machine Learning Approaches" applied machine learning techniques to predict employee attrition, address turnover issues and optimize retention strategies. The authors utilize IBM HR employee attrition dataset and employ Employee Exploratory Data Analysis (EEDA) a feature engineering technique to find the best-fit parameters through feature correlation for model building and prediction purposes. The machine learning model was trained on 85% of the data and tested on the remaining 15% of the data. By partitioning the dataset into separate training and testing subsets, researchers can accurately evaluate the model's performance [5].

To assist the human resources (HR) department by identifying

factors influencing employee attrition Sari and Lhaksmana in their study entitled "Employee Attrition Prediction Using Feature Selection with Information Gain and Random Forest Classification" employ Random Forest as the predictive model and compare 7 three feature selection methods namely Information Gain, Select K Best, and Recursive Feature Elimination. These methods are evaluated based on their performance metrics Recursive Feature Elimination yields 88.8% accuracy while Select K Best achieved 87.8% accuracy. The two feature selection methods mentioned Select K Best and Recursive Feature Elimination will be used in this study to identify the most important factors from the dataset, allowing the model to focus on key predictors and improve prediction accuracy [7].

In a proposed project by Madara Pratt entitled "Employee Attrition Estimation Using Random Forest Algorithm" addressed the challenges of employee retention in companies despite various factors such as culture and finance influencing attrition, many firms struggle to assess employee satisfaction, often resulting in unexpected resignations. Their study highlights Random Forest which achieves a high accuracy of 85.12%, signaling its potential for managers to anticipate and mitigate employee turnover. Given the insights from Pratt, et. al. (2021), the proponents aim to utilize the high predictive accuracy of the Random Forest algorithm in their study to delve deeper into the dynamics of employee attrition and provide managers with actionable strategies for improving retention rates [4].

To explore the effectiveness of machine learning techniques in predicting and mitigating attrition's impact a comparative study was conducted by Norsuhada Mansor, Nor Samsiah Sani and Mohd Aliff. Their study "Machine Learning for Predicting Employee Attrition" presents a comparative analysis of three machine learning algorithms, the study aimed to identify the most accurate predictor of employee attrition. Utilizing dataset from IBM the study found that the optimized Support Vector Machine achieved the highest accuracy of 88.87% followed by Artificial Neural Network and Decision Tree. These findings offer valuable insight for organizations seeking to develop proactive strategies to retain talent and minimize attrition-related costs [3].

3. METHODOLOGY

A. Employee Dataset

For this study, the proponents used the IBM Human Resource Attrition dataset, which was downloaded from the Kaggle Dataset Repository. The dataset consists of various employee attributes that help predict employee turnover and identify factors contributing to attrition. It includes 35 variables, such as demographic details such as age, gender, marital status, job-related information like department, job role, satisfaction, and performance metrics such as years at company, education level, distance from home. The dataset is 222 KB in size, making it a compact yet informative resource for analysis.

B. Preprocessing

To prepare the raw data for analysis, modeling and to help improve the quality of the data ensuring better performance of the machine learning algorithms data preprocessing is performed by the proponents. Data preprocessing involves several key steps which will be described in the following sections.

1) Employee Exploratory Data Analysis

In this study, the Employee Exploratory Data Analysis (EEDA) was applied to gain valuable insights from the HR employee attrition dataset. This analysis was conducted to thoroughly investigate the key features, check for null values, and examine the data types of the columns. The process involved identifying and removing unnecessary features, as well as gaining a deeper understanding of the factors contributing to employee attrition. This study explored these features using various tables, graphs, plots, and pie charts.

The dataset consisted of 35 features, categorized into 9 categorical or string columns and 26 numerical columns. Prior to the analysis, the proponents check for null values in the dataset and confirmed that there were no missing values, ensuring the integrity of the analysis.

Table 1. Dataset Summary

Dataset Summary	Value
Rows	1,470
Columns	35
Categorical Columns	9
Numerical Columns	26
Null Values	0

The data distribution pie chart in Figure 1 for employee attrition shows that out of the 1470 employees, 16% of the employees left their job for various reasons, whereas 84% of the employees chose to continue their employment with the company.

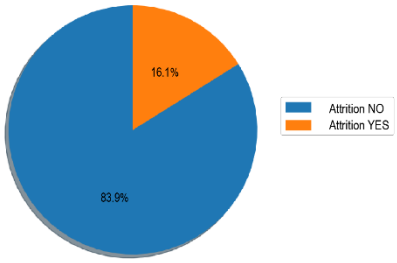

Fig 1: Distribution of Employee Attrition

The proponents explore the significance of feature names in the dataset, focusing on their role in model interpretability and data exploration. The analysis includes examining the distributions and relationships of the features with the target variable, ultimately demonstrating how this assessment impacts the overall quality and performance of the machine learning model. Notably, the proponents identify five features the StandardHours, EmployeeCount, Over18, EmployeeNumber, and StockOptionLevel that were removed from the dataset due to their limited relevance.

The bar chart from Figure 2 shows how employee age relates to attrition, comparing those who stayed and those who left. Most employees are between 29 and 36 years old, and within this age range, attrition is higher, especially for those aged 28 to 33. Younger employees tend to leave more often, while older employees (especially over 40) are more likely to stay, with attrition decreasing significantly after age 50. This suggests that younger employees are more likely to leave the company, while older employees tend to stay longer.

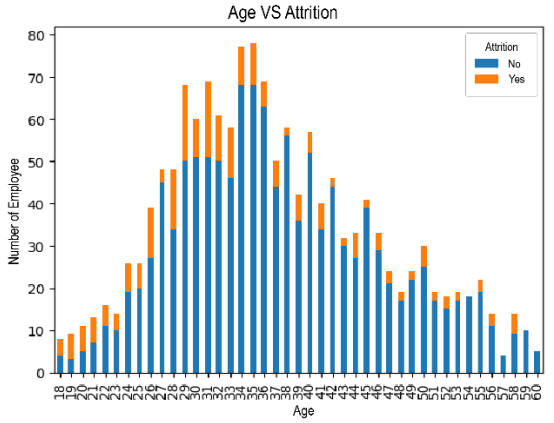

Fig 2: Age and Attrition Bar Chart

The data chart in figure 3 illustrates that employee attrition is most prevalent in the early years of tenure, particularly within the first year, where a significant number of employees leave the company. Attrition rates remain relatively high during the first five years but decline steadily thereafter, with fewer employees leaving as tenure increases. Beyond 10 years at the company, attrition becomes infrequent, suggesting that long-tenured employees are more likely to stay. This trend highlights the importance of focusing retention efforts on newer employees, as "YearsAtCompany" appears to be a key factor in predicting employee attrition.

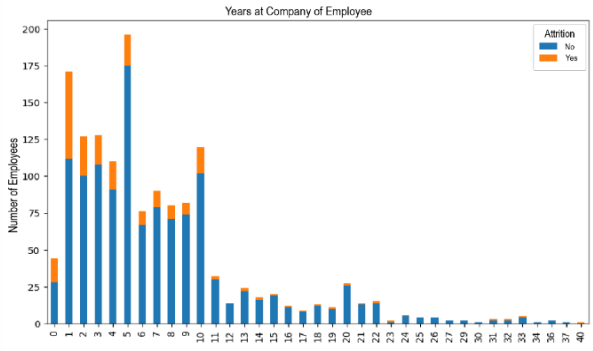

Fig 3: Years at Company and Attrition Bar Chart

The data distribution box plot for the analysis of monthly income shows that employees who left tend to have lower median incomes, with most incomes ranging between 2,500 and 5,000, and a few outliers above 12,500. On the other hand, employees who stayed have a higher median income, with a wider range of income levels, including outliers reaching up to 20,000. This suggests that higher-income employees are less likely to leave, indicating that income may be a factor in employee retention.

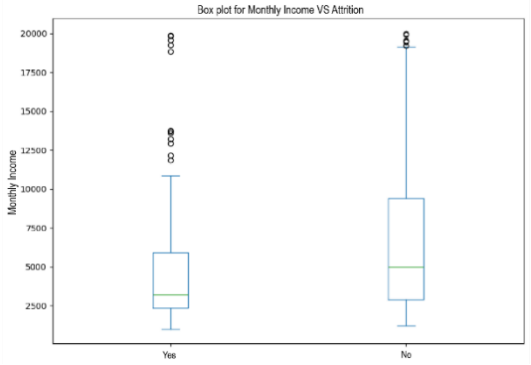

Fig 4: Monthly Income and Attrition Box Plot

The proponents compare the number of employees who work overtime with those who don't, highlighting their respective attrition rates. The bar chart for overtime and attrition shows that the employees who don't work overtime represent a much larger group, with most of them remaining with the company, as indicated by the smaller proportion of those who left. The group of employees who work overtime is smaller but exhibits a significantly higher employee turnover, possibly due to factors like burnout or job dissatisfaction.

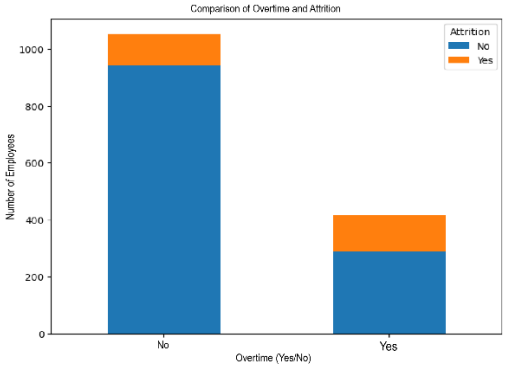

Fig 5: Overtime and Attrition Bar Chart

The donut chart in figure 6 visualizes the relationship between gender and attrition, with the outer ring representing gender and the inner ring showing attrition status. For both males and females, the majority of the inner ring is orange, indicating that most employees, regardless of gender, did not leave the company. However, there is a small blue section for each gender that represents employees who have experienced attrition. The blue section is slightly larger for males, suggesting that male employees have a marginally higher attrition rate compared to females. Overall, while attrition is relatively low for both genders, it appears to be slightly more prevalent among males.

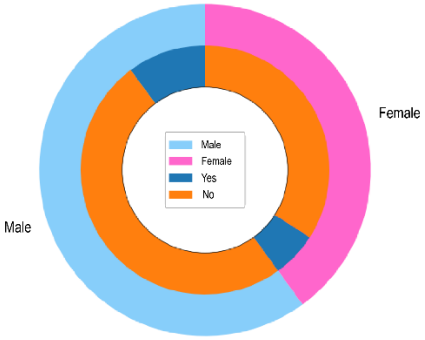

Fig 6: Gender and Attrition Donut Chart

2) Feature Selection

This study examines the impact of two feature selection methods namely Select K Best and Recursive Feature Elimination (RFE) and each feature selection is taken by 10 features. The Select K Best method utilizes scoring metrics to evaluate the features based on their relationship to the target variable. RFE, in contrast, ranks features by recursively eliminating the least important ones until the desired number of features is achieved. Select K Best.

a. Recursive Feature Elimination (RFE): The approach taken by the RFE method is to recursively select the optimal subset of features based on their importance to the prediction process. In each iteration, features will be ranked, and non-optimal or

irrelevant features will be removed [9]. The advantages of RFE are that it is easy to configure, and use, and can efficiently select features to predict target variables [7].

Table 2 presents the features ranked by the Recursive Feature Elimination (RFE) method from the IBM HR Employee Attrition dataset, indicating their importance in predicting employee attrition. RFE works by ranking features based on their importance, where features with a rank of 1 are the most significant. For this analysis, the objective is to select the top 10 feature.

Table 2. Key Features Identified by RFE

	Features	Ranking	Support
0	Age	1	True
1	BusinessTravel	11	False
2	DailyRate	1	True
3	Department	20	False
4	DistanceFromHome	1	True
5	Education	12	False
6	EducationField	15	False
7	EnvironmentSatisfaction	5	False
8	Gender	17	False
9	HourlyRate	1	True
10	JobInvolvement	19	False
11	JobLevel	18	False
12	JobRole	6	False
13	JobSatisfaction	3	False
14	MaritalStatus	2	False
15	MonthlyIncome	1	True
16	MonthlyRate	1	True
17	NumCompaniesWorked	4	False
18	OverTime	1	True
19	PercentSalaryHike	1	True
20	PerformanceRating	16	False
21	RelationshipSatisfaction	9	False
22	TotalWorkingYears	1	True
23	TrainingTimesLastYear	1	True
24	WorkLifeBalance	13	False
25	YearsAtCompany	10	False
26	YearsInCurrentRole	8	False
27	YearsSinceLastPromotion	7	False
28	YearsWithCurrentManager	14	False

Table 3 shows that feature such as Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, OverTime, PercentSalaryHike, TotalWorkingYears and TrainingTimesLastYear were identified as essential and selected during the RFE process. These features are considered highly relevant for predicting employee attrition.

Table 3. Top 10 Features from RFE

No.	Featured Features
1	Age
2	DailyRate
3	DistanceFromHome
4	HourlyRate
5	MonthlyIncome
6	MonthlyRate
7	Overtime
8	PercentSalaryHike
9	TotalWorkingYears
10	TrainingTimesLastYear

b. Select K Best (SKB): SKB is a module in the scikit learn library that selects the K Feature that has the highest score. The score is calculated based on univariate statistical analysis, which is an analysis of variables one by one [2].

Table 4 shows the result of applying the Select K Best feature selection method to the IBM employee attrition dataset. Select K Best is a technique used to identify the most important features based on their relationship with the target variable, employee attrition. The goal of this method was to select the top 10 features most relevant for predicting attrition. Each featured received a score, with higher scores indicating greater importance.

Table 4. Key Features Identified by Select K Best

	Features	Score	Status
0	Age	38.17588	Selected
1	BusinessTravel	7.99037	Not Selected
2	DailyRate	4.71885	Not Selected
3	Department	6.03587	Not Selected
4	DistanceFromHome	8.96827	Not Selected
5	Education	1.44630	Not Selected
6	EducationField	1.05872	Not Selected
7	EnvironmentSatisfaction	15.85520	Not Selected
8	Gender	1.27458	Not Selected
9	HourlyRate	0.06879	Not Selected
10	JobInvolvement	25.24198	Selected
11	JobLevel	43.21534	Selected
12	JobRole	6.64967	Not Selected

13	JobSatisfaction	15.89000	Not Selected
14	MaritalStatus	39.59976	Selected
15	MonthlyIncome	58.75056	Selected
16	MonthlyRate	0.34251	Not Selected
17	NumCompaniesWorked	2.78228	Not Selected
18	OverTime	94.65645	Selected
19	PercentSalaryHike	0.26672	Not Selected
20	PerformanceRating	0.01225	Not Selected
21	RelationshipSatisfaction	3.09557	Not Selected
22	TotalWorkingYears	44.52363	Selected
23	TrainingTimesLastYear	5.21164	Not Selected
24	WorkLifeBalance	6.02611	Not Selected
25	YearsAtCompany	28.05146	Selected
26	YearsInCurrentRole	38.83830	Selected
27	YearsSinceLastPromotion	1.60221	Not Selected
28	YearsWithCurrentManager	36.71231	Selected

From Table 5, it can be observed that the features with the highest scores and selected for the model are Age, JobInvolvement, JobLevel, MaritalStatus, MonthlyIncome, OverTime, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, and YearsWithCurrManager. These features are considered the most important for predicting employee attrition when using Select K Best method.

Table 5. Top 10 Features from Select K Best

No.	Featured Features
1	Age
2	JobInvolvement
3	JobLevel
4	MaritalStatus
5	MonthlyIncome
6	OverTime
7	TotalWorkingYears
8	YearsAtCompany
9	YearsInCurrentRole
10	YearsWithhCurrManager

3) Data Encoding

Machine Learning algorithms typically require numerical input. Therefore, categorical variables such as Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18 and Overtime were encoded into numerical values using LabelEncoder, transforming them into a format suitable for model input and processing. Each category is assigned a unique numerical value, converting categorical data into a format suitable for machine learning applications.

C) Dataset Splitting

To evaluate the model's performance effectively, the dataset was split into training and testing sets following an 85:15 ratio. With a total of 1,470 records, this means that approximately 1,249 records (85%) were used to train the machine learning models, enabling them to learn patterns and relationships within the data, while the remaining 221 records (15%) were reserved as a testing set to assess generalization capability on unseen data. The researchers applied this splitting approach to two machine learning methods, namely Random Forest and Support Vector Machine (SVM). This provides a balanced evaluation, helping to prevent overfitting and offering a reliable estimate of how well the models might perform on new employee data.

D) Machine Learning Approach

This study involves selecting two machine learning techniques to pinpoint the most suitable machine learning algorithm for predicting employee attrition. The considered machine learning algorithms are as follows:

a. Random Forest: Random forest, a tree-based algorithm, is well-known in machine learning problems. Random Forest (RF) is utilized for classification problems that works by producing multiple decision trees generates multiple random training subsets. Then it creates a tree with random training subsets [1].

b. Support Vector Machine (SVM): SVM is a supervised machine learning algorithm that can be used for either classification or regression challenges. However, it is mostly used in classification problems. It performs classification by finding the hyperplane that completely separates the vector into two non-overlapping classes. The vectors that define the hyperplane are the support vectors [8].

E) Deployment

The final stage of this project involves deploying or designing a web application prototype. At this stage, the developed and tested model is implemented as a web-based tool ready for user interaction. For this study on predicting employee attrition, the web application prototype is built using the Random Forest model and incorporates selected features, including Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, OverTime, PercentSalaryHike, TotalWorkingYears and TrainingTimesLastYear.

4. RESULTS AND DISCUSSION

All trained models in this project were evaluated using the following metrics: accuracy, precision, recall and F1 score, as described below:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$
$$Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN)$$
$$F1\ Score = 2 * (Precision * Recall) / (Precision + Recall)$$

The terms TP, TN, FP, and FN are used in classification models to evaluate their performance. TP (True Positives) refer to the number of correctly predicted positive cases, while TN (True Negatives) represent the correctly predicted negative cases. FP (False Positives) are instances where the model incorrectly predicted a positive outcome when it should have been negative, and FN (False Negatives) are instances where the

model incorrectly predicted a negative outcome when it should have been positive. Together, these values help calculate performance metrics, such as accuracy, precision, recall, and F1-score that assess how well the model is making predictions.

In table 6, the Random Forest model, using RFE for feature selection resulted in an accuracy of 0.842, meaning the model correctly predicted 84.2% of the cases. The precision score of 0.700 shows that, among all cases predicted as positive, 70% were correct. The recall score is 0.179, indicating that the model identified 17.9% of all actual positive cases. The F1-Score, which combines precision and recall, is 0.286, showing the overall balance between these two measures.

By applying Select K Best for feature selection, the Random Forest model attained an accuracy of 0.837, which translates to an 83.7% accuracy rate. The precision dropped to 0.615, meaning that 61.5% of positive predictions were correct. The recall also dropped to 0.205, capturing 20.5% of actual positives. The F1-Score, however, increased slightly to 0.308, indicating a better balance between precision and recall compared to the RFE method.

For the Support Vector Machine (SVM) model, the RFE-selected features gave an accuracy of 0.561, meaning 56.1% of predictions were correct. The precision was 0.241, so 24.1% of positive predictions were correct. The recall was higher at 0.692, meaning the model captured 69.2% of actual positives. The F1-Score was 0.358, showing a reasonable balance between precision and recall.

Using Select K Best with the SVM model improved accuracy to 0.670, achieving a 67% success rate. The precision increased to 0.250, meaning 25% of positive predictions were correct, but the recall dropped to 0.436. The F1-Score slightly decreased to 0.318, indicating a small drop in the balance between precision and recall.

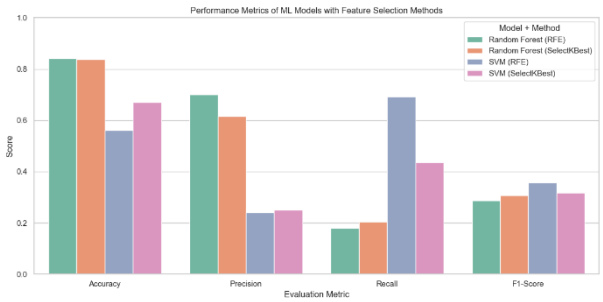

Figure 7: Performance Metrics Graph

Overall, the Random Forest model performed better in terms of accuracy and precision, especially with RFE for feature selection. However, the SVM model achieved a higher recall, particularly with the RFE method, showing its strength in identifying more actual positives. This suggests that Random Forest is more reliable for accurate predictions.

The proponents selected the Random Forest model with RFE because it achieved the highest accuracy of 84.2%, demonstrating the best overall performance. Despite the SVM model having higher recall in some cases, the Random Forest model's higher accuracy and better balance between precision and recall made it the more reliable and effective choice for predictions.

The Flask Framework is widely used for building web applications. This study on employee attrition, researchers utilized Flask to develop an application that predicts the chances of turnover. This framework made it simple for the proponents to integrate Machine Learning models and display the results using clear and interactive visualizations on the app's dashboard. Flask also enables the proponents to set up endpoints where users can input data and receive predictions instantly.

The homepage of the project, as depicted in Figure 8 serves as an introduction to the Employee Attrition Prediction system. It presents a visually appealing design and provides an overview of the system's purpose. The page explains how the system aids organizations in identifying employees at risk of leaving, facilitating early interventions to improve retention, reduce turnover costs, and enhance workforce planning. Additionally, it highlights the factors contributing to employee attrition to inform workplace policies and optimize talent management.

Table 6. Machine Learning Model Performance Evaluation

Method	**Random Forest**				**Support Vector Machine**			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
RFE	0.842	0.700	0.179	0.286	0.561	0.241	0.692	0.358
SelectKBest	0.837	0.615	0.205	0.308	0.670	0.250	0.436	0.318

Fig 8: Home Page

The Employee Attrition Prediction tool helps predict the chances of employees leaving the company. It uses 10 key factors that can influence an employee's decision to quit, including Age, Training Times Last Year, Daily Rate, 32 Total Working Years, Distance From Home, Percent Salary Hike, Overtime, Monthly Rate, Monthly Income, and Hourly Rate. After entering these details, the tool processes the data and gives a percentage that shows the likelihood of an employee leaving.

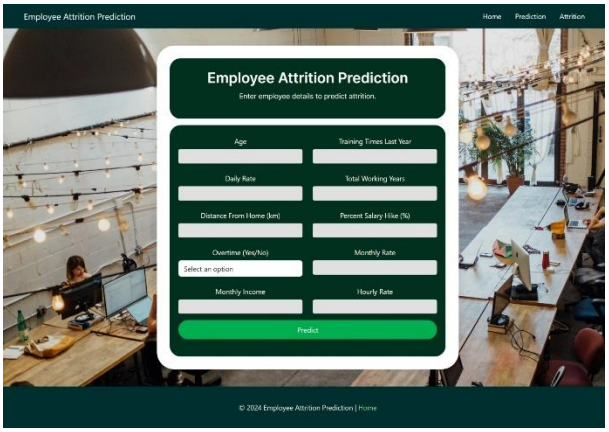

Fig 9: Prediction Page

The Attrition page in the web application displays the key factors that influence employee attrition, which are the inputs required for the prediction tool. The page features two buttons left and right, which when clicked, show the next factor along with a brief description explaining why it is an important contributor to attrition.

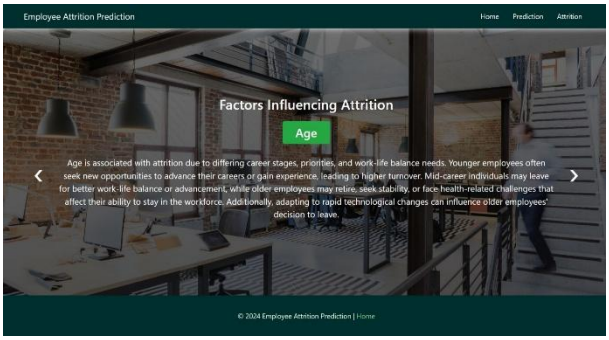

Fig 10: Attrition Page

5. CONCLUSION AND FUTURE WORK

The employee attrition prediction process analyzes employee data to predict turnover and identify the main reasons behind it. Through employee exploratory data analysis of the IBM Human Resource Attrition dataset, the researchers identified several significant trends regarding employee attrition. Younger employees, particularly those aged 28 to 33, were found to have higher turnover rates, with attrition decreasing those aged 28 to 33, were found to have higher turnover rates, with attrition decreasing significantly after the age of 40 and becoming minimal after 50.

The analysis also showed that employees with shorter tenures, particularly those in the first year, had the highest attrition rates, with turnover decreasing as tenure increasing. Income played a role in attrition, with lower-income employees more likely to leave, while those with higher incomes tended to stay. Overtime work was another contributing factor, as employees working overtime showed higher attrition rates, possibly due to burnout or job dissatisfaction. Additionally, the data revealed that male employees had slightly higher attrition rates than their female counterparts. These findings suggest that organizations could improve retention by focusing on younger employees on younger employees, offering competitive compensation, managing workloads, and addressing gender-specific factors.

In this study the proponents identified key predictors of employee attrition using feature selection methods, specifically Recursive Feature Elimination (RFE) and SelectKBest. According to the RFE method, the top 10 features for predicting attrition including Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, OverTime, PercentSalaryHike, TotalWorkingYears, and TrainingTimesLastYear. On the other hand, SelectKBest highlighted Age, JobInvolvement, JobLevel, MaritalStatus, MonthlyIncome, OverTime, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, and YearsWithCurrManager as the most relevant predictors. These findings underscore the importance of both personal and professional factors, such as age, compensation, work hours, and career progression, in understanding employee attrition.

Based on the performance comparison, the combination of Recursive Feature Elimination (RFE) and the Random Forest model proved to be the most effective in terms of accuracy and precision. Although its recall was slightly lower, its overall performance met the criteria for a strong method, making it the optimal choice for this case.

To reduce employee turnover, organizations should focus on key areas. Since younger employees (ages 28–33) are more likely to leave, companies should invest in mentorship, career development, and clear promotion paths. Competitive salaries are also essential, as lower-income employees tend to leave

more often. Regular salary reviews and performance-based incentives can help employees feel valued.

Workload is another factor—those working overtime showed higher attrition rates, likely due to stress. Employers should monitor workloads, offer flexible schedules, and support overworked staff. For new hires, especially in their first year, strong onboarding and regular check-ins are important to boost engagement and retention.

The study also found that male employees had slightly higher attrition rates. Organizations should adopt inclusive policies that support equal growth opportunities and work-life balance, while also addressing concerns specific to male employees.

To identify employees at risk of leaving, companies are encouraged to use the web application developed in this project. It analyzes key factors like age, income, and overtime status to support proactive HR strategies. Regular updates and adding features like engagement levels and career development data can further improve its accuracy.

Organizations are encouraged to regularly update the employee attrition prediction tool with new data to ensure its accuracy and effectiveness. Including additional relevant features, such as career development opportunities and employee engagement levels, could improve the tool's predictive capability.

6. REFERENCES

[1] Bada, O., and Lekan, A. J. 2022. Employee attrition prediction using machine learning algorithms. ResearchGate.

[2] Desyani, T., Saifudin, A., and Yulianti, Y. 2020. Feature selection based on naive Bayes for Caesarean section prediction. IOP Conference Series: Materials Science and Engineering, 879(1), 012091.

[3] Mansor, N. N. B. A., Sani, N. S., and Aliff, M. 2021. Machine learning for predicting employee attrition. International Journal of Advanced Computer Science and Applications, 12(11).

[4] Pratt, M., Boudhane, M., and Cakula, S. 2021. Employee attrition estimation using random forest algorithm. Baltic Journal of Modern Computing, 9(1).

[5] Raza, A., Munir, K., Almutairi, M., Younas, F., and Fareed, M. M. S. 2022. Predicting employee attrition using machine learning approaches. Applied Sciences, 12(13), 6424.

[6] Repaso, J. A. A., Capariño, E. T., Hermogenes, M. G. G., and Perez, J. G. 2022. Determining factors resulting to employee attrition using data mining techniques. International Journal of Education and Management Engineering, 12(3), 22–29.

[7] Sari, S. F., and Lhaksmana, K. M. 2022. Employee attrition prediction using feature selection with information gain and random forest classification. Journal of Computer System and Informatics, 3(4), 410–419.

[8] Shankar, R., Rajanikanth, J., Sivaramaraju, V. V., and Murthy, K. V. 2018. Prediction of employee attrition using datamining. In Proceedings of the International Conference on Smart Computing and Applications (ICSCAN).

[9] Wardhani, F. H., and Lhaksmana, K. M. 2022. Predicting employee attrition using logistic regression with feature selection. Sinkron, 7(4), 2214–2222