IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

Career Catalyst: Anticipating staff attrition through predtive modelling

T. Shalini[1], Sk. Sheema[2], G. Lahari[3], M. Baby Jyothsna[4], P. Pranathi[5]

[1] Assistant Professor, Dept. of Computer Science & Engineering, Vijaya Institute of Technology for Women, Enikepadu, Vijayawada-521108

[2,3,4,5] Students, Dept. of Computer Science & Engineering, Vijaya Institute of Technology for Women, Enikepadu, Vijayawada-521108

Email id: shalinitammana09@gmail.com[1], shaiksheema444@gmail.com[2], giritipallilahari@gmail.com[3], maddalibabyjyothsna@gmail.com [4], ppranathi1001@gmail.com [5]

Abstract:

Employee attrition prediction is a critical task for organizations seeking to maintain workforce stability and productivity. This study investigates the effectiveness of machine learning methods, specifically logistic regression and random forest algorithms, in predicting employee attrition. Utilizing a comprehensive dataset comprising various employee attributes such as demographics, job satisfaction, and performance metrics, logistic regression is initially employed to model the probability of attrition. This method allows for the identification of significant predictors contributing to attrition risk, providing valuable insights for HR decision-makers. Subsequently, the study integrates random forest, an ensemble learning technique renowned for its ability to capture complex interactions and nonlinear relationships within the data. By combining the strengths of logistic regression and random forest, the predictive model achieves enhanced accuracy in identifying employees at risk of attrition. Through the utilization of this machine learning-powered approach, organizations gain a proactive tool for identifying potential turnover, enabling them to implement targeted retention strategies and mitigate the adverse effects of employee attrition on business continuity and performance. The results demonstrate the efficacy of the proposed methodology in accurately predicting employee attrition, thereby empowering organizations to make informed decisions regarding workforce management and retention efforts. By leveraging machine learning techniques, such as logistic regression and random forest, organizations can proactively address attrition challenges, fostering a more stable and engaged workforce while optimizing operational efficiency and productivity. This research underscores the importance of utilizing advanced analytics in HR practices to anticipate and mitigate employee turnover, ultimately contributing to organizational success and competitiveness in the dynamic business landscape.

Keywords: Career Catalyst, machine learning methods, HR practices.

Introduction

The project utilizes logistic regression and random forest algorithms to predict employee attrition, using a dataset with employee demographics and job satisfaction data. Logistic regression estimates attrition probabilities, while random forest enhances predictive accuracy by capturing complex data patterns. The aim is to help organizations proactively manage attrition and implement targeted retention strategies for a stable workforce.

In this project, logistic regression and random forest algorithms were employed to address the pressing issue of employee attrition within organizations. Logistic regression provided a foundational understanding by modeling the probability of turnover based on a variety of factors including demographics, job satisfaction, and performance metrics. This approach offered clear insights into the significance of individual predictors in influencing attrition risk, allowing for targeted intervention strategies.

Furthermore, the integration of random forest brought added depth to the analysis by capturing intricate patterns and nonlinear relationships present in the data. By leveraging the strengths of both logistic regression and random forest, the predictive model aimed to accurately identify employees at risk of attrition, thereby enabling organizations to proactively implement retention measures and foster a more stable workforce. Through the utilization of advanced machine learning techniques, this project sought to empower organizations with actionable insights to mitigate the detrimental effects of employee turnover and optimize workforce management strategies.

The purpose of the project was to develop a predictive model using logistic regression and random forest algorithms to forecast employee attrition within organizations. By anticipating turnover before it occurs, businesses can implement proactive retention strategies tailored to individual employee needs, thereby fostering a more stable and engaged workforce while mitigating the costs associated with recruitment and loss of productivity.

2.LITERATURE REVIEW

Mohd Aliff [1] The study aims to compare the performance of three machine learning classifiers - Decision Tree (DT), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) - in predicting employee attrition. It underscores the significance of employing machine learning techniques to anticipate employee turnover, facilitating timely interventions by HR departments to mitigate attrition's adverse effects.

Rohit Punnoose and Pankaj Ajit [2] in their research published in the International Journal of Advanced Research in Artificial Intelligence (IJARAI) in 2016, Rohit Punnoose and Pankaj Ajit address the challenge of accurately predicting employee turnover, which is often under-funded compared to other domains within organizations

Qasem A, A.Radaideh, and Eman A Nagi. [3] The study applies data mining techniques to develop a classification model for predicting employee performance. The authors utilize the CRISP-DM data mining methodology in their work, focusing on the Decision Tree algorithm as the primary tool for building the classification model.

EXISTING SYSTEM:

Present existing systems for addressing employee attrition typically rely on traditional methods such as surveys, exit interviews, and HR analytics tools. These systems often involve manual data collection and analysis, leading to delays in identifying attrition risks and implementing retention strategies. While some organizations may utilize statistical methods like regression analysis, they often lack the ability to capture complex relationships and nonlinear patterns present in employee data. As a result, there's a growing recognition of the need to integrate advanced machine learning techniques into existing systems to enhance predictive accuracy and enable proactive attrition management. These modern systems leverage algorithms like logistic regression and random forest to analyze diverse employee attributes and provide organizations with actionable insights for optimizing workforce retention efforts.

PROPOSED SYSTEM:

The impetus behind the Employee Attrition Prediction Project stems from the pressing challenges organizations face in retaining valuable talent amidst high turnover rates. With turnover rates soaring, the imperative to develop proactive solutions becomes paramount. Studies indicate that traditional methods of attrition prediction often fall short, lacking the sophistication to capture intricate patterns and predictors effectively. Thus, the project seeks to address this gap by harnessing advanced machine learning

techniques, namely logistic regression and random forest algorithms, to predict employee attrition accurately.

SYSTEM ARCHITECTURE

Creating a system architecture for predicting staff attrition involves several components, from data collection and preprocessing to modeling and deployment. Below, I outline a typical architecture for a predictive system focused on staff attrition:

Data Collection

- Human Resources (HR) Data: Collect data from HR systems, including employee demographics, tenure, job roles, salary, performance evaluations, training records, and promotion history.
- Employee Surveys: Periodic surveys to gather data on employee satisfaction, engagement, and workplace environment.
- Operational Data: Information from other systems that might indirectly indicate employee satisfaction or dissatisfaction, such as attendance records, frequency of late arrivals, or overtimes.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

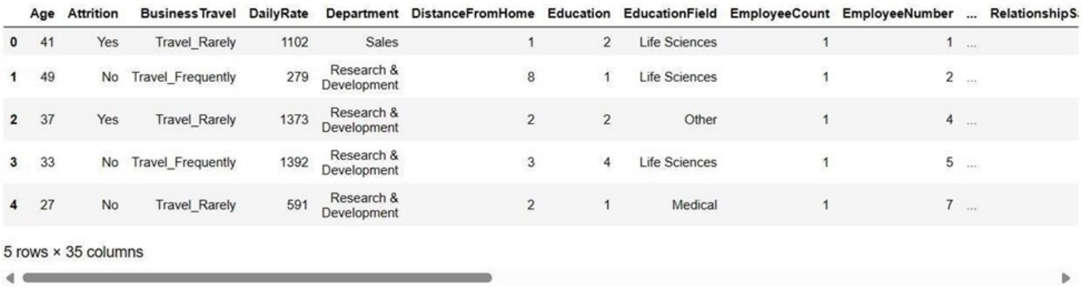
INPUT DESIGN

The goal of the coding or programming phase is to translate the design of the system produced during the design phase into code in a given programming language, which can be executed by a computer and that performs the computation specified by the design. The coding phase affects both testing and maintenance. The goal of coding is not to reduce the implementation cost but the goal should be to reduce the cost of later phases. In other words the goal is not to simplify the job of programmer. Rather the goal should be to simplify the job of the tester and maintainer. Designing the input for a staff attrition project involves collecting relevant data and information to analyze the factors contributing to employee turnover. Here's a breakdown of the input design.

OUTPUT DESIGN

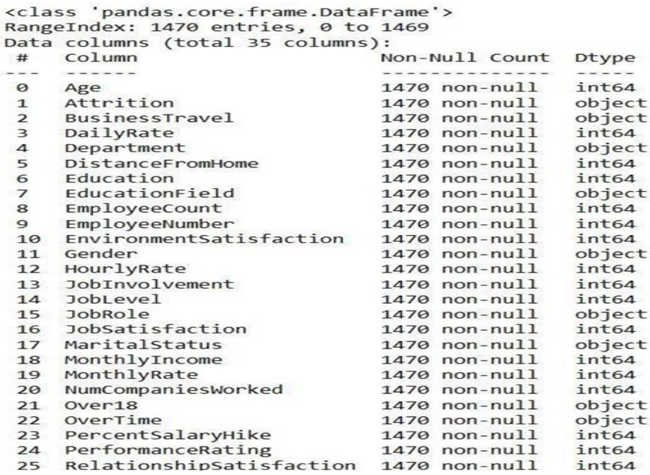
A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipS
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	

5 rows × 35 columns



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
```

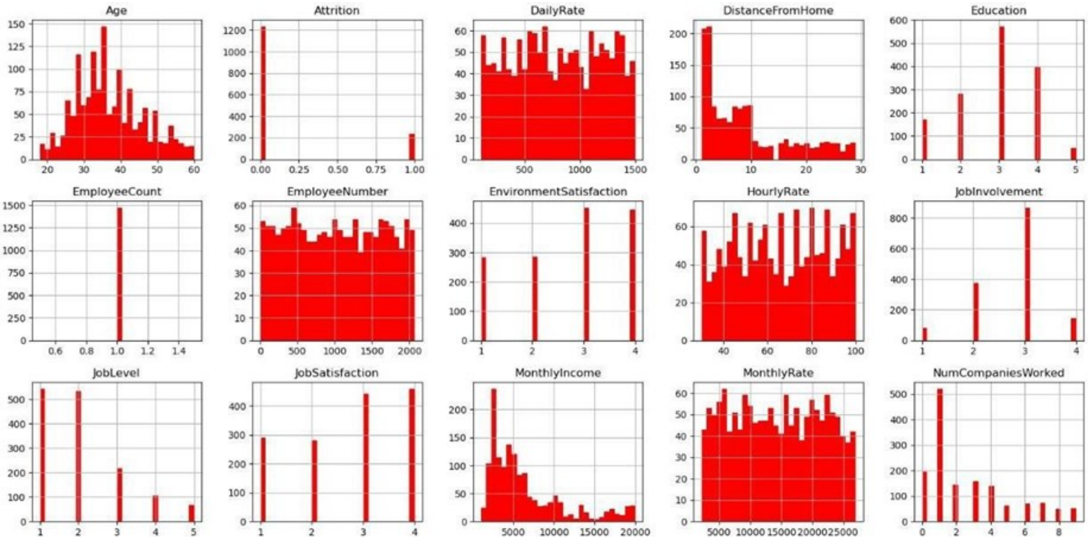
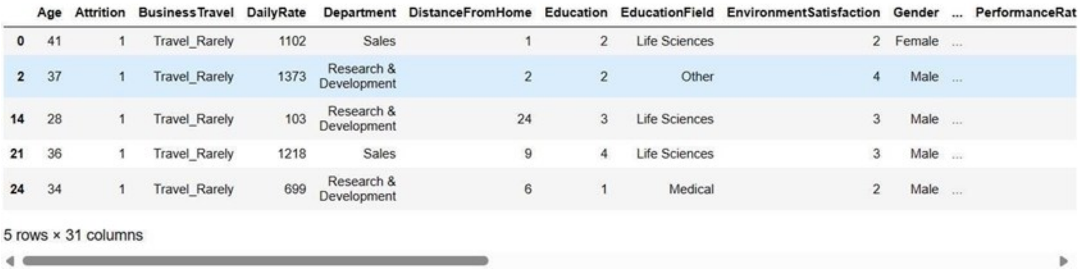
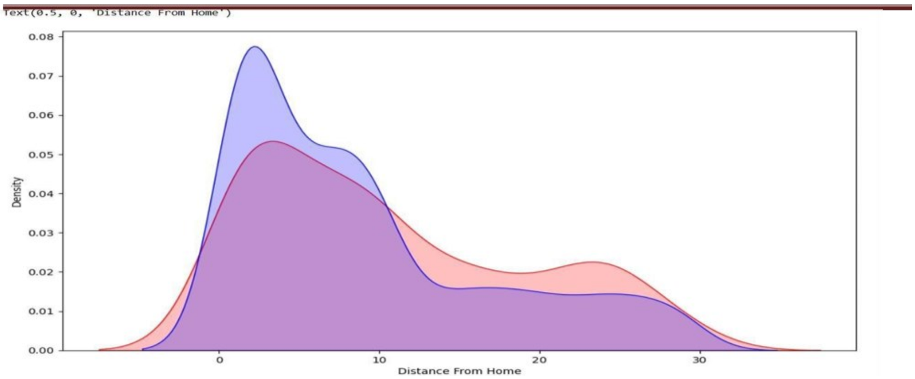


Figure: Employee_ df .history



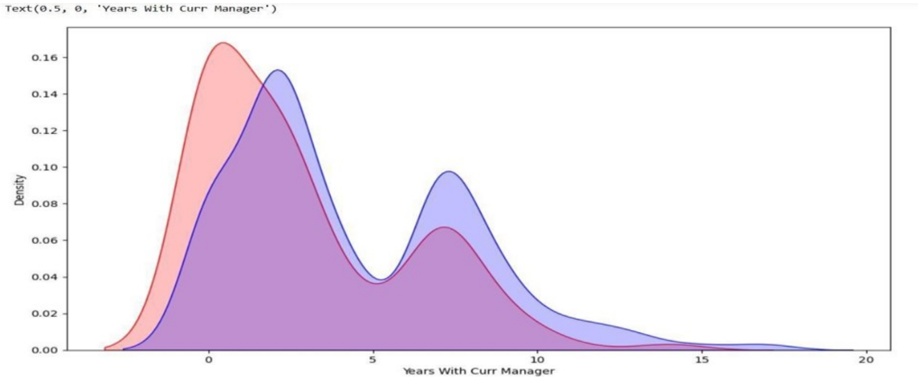
	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	...	PerformanceRat
0	41	1	Travel_Rarely	1102	Sales	1	2	Life Sciences	2	Female	...	
2	37	1	Travel_Rarely	1373	Research & Development	2	2	Other	4	Male	...	
14	28	1	Travel_Rarely	103	Research & Development	24	3	Life Sciences	3	Male	...	
21	36	1	Travel_Rarely	1218	Sales	9	4	Life Sciences	3	Male	...	
24	34	1	Travel_Rarely	699	Research & Development	6	1	Medical	2	Male	...	

5 rows × 31 columns



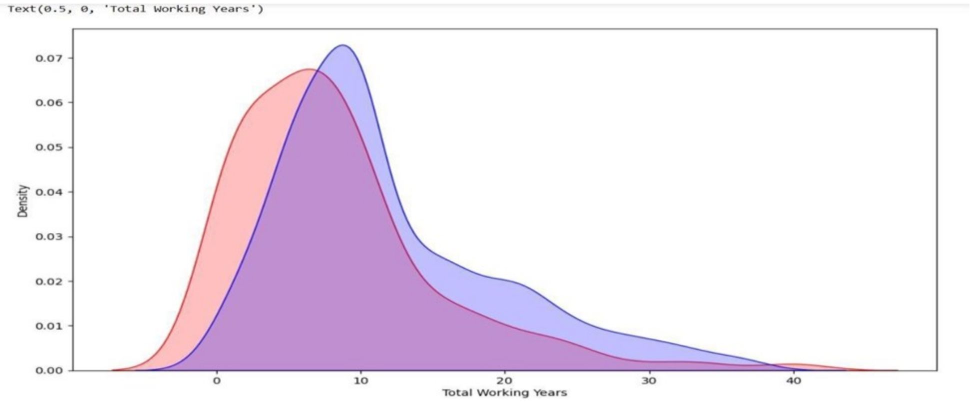
```
Text(0.5, 0, 'Distance From Home')
```

Figure: KDE Plot-1



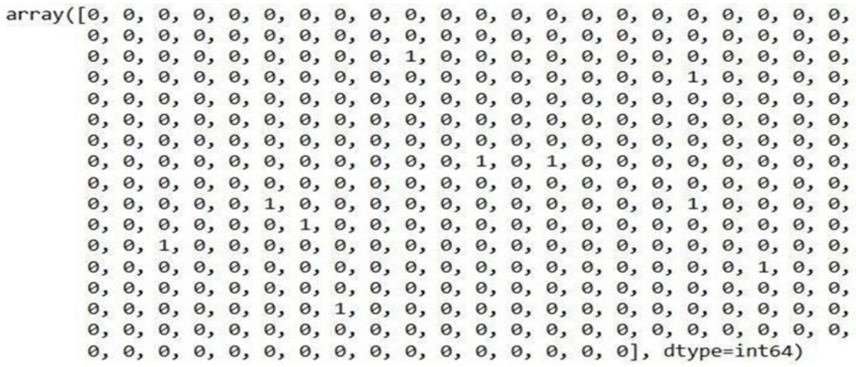
```
Text(0.5, 0, 'Years With Curr Manager')
```

Figure: KDE Plot-2



```
Text(0.5, 0, 'Total Working Years')
```

Figure: KDE Plot-3



```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int64)
```

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

<AxesSubplot:>

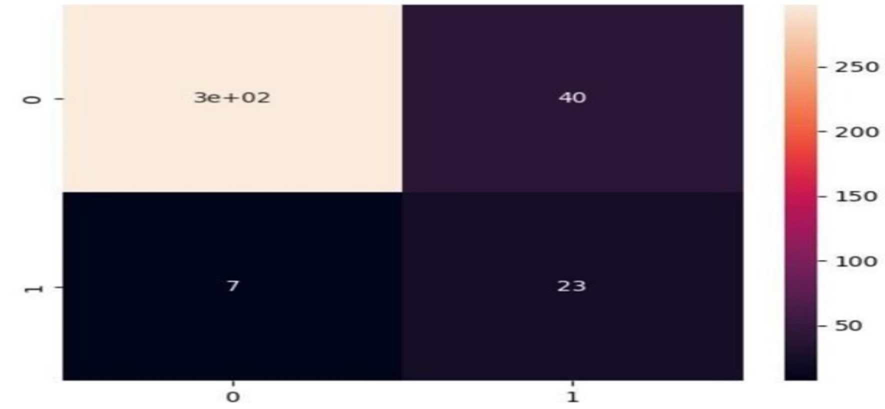


Figure: Confusion _matrix

CONCLUSION:

The purpose of the project was to develop a predictive model using logistic regression and random forest algorithms to forecast employee attrition within organizations. By anticipating turnover before it occurs, businesses can implement proactive retention strategies tailored to individual employee needs, thereby fostering a more stable and engaged workforce while mitigating the costs associated with recruitment and loss of productivity. Through the analysis of diverse employee attributes such as demographics, job satisfaction, and performance metrics, the project aimed to provide organizations with actionable insights to optimize workforce management practices, including recruitment, talent development, and employee engagement initiatives. Ultimately, the goal was to empower organizations to make informed decisions that promote long-term organizational success by reducing turnover and enhancing employee retention. Predicting the factors for Employee attrition based on machine learning algorithms for taking steps to overcome the employee attrition percentage in the company with help of ml model performance and accuracy, and finding Top 10 Factors based on above ML algorithms.

Future scope:

Controlling the Employee attrition based on predictions performed by us and taking steps for overcome the current situation and make changes in employment and the attrition percentage in the organizations and companies after the prediction

References:

1. Peng, B. Statistical analysis of employee retention. In Proceedings of the International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2021), Nanjing, China, 26–28 November 2021; Volume 12163, pp. 7–15.
2. Here's What Your Turnover and Retention Rates Should Look Like. Available online: https://www.ceridian.com/blog/turnover-and-retention-rates-benchmark (accessed on 6 May 2022).
3. SHRM Survey: Average Cost Per Hire Is $4129. Available online: https://www.businessmanagementdaily.com/46997/shrm-survey-average-cost-per-hire-is-4129/ (accessed on 6 May 2022).
4. Gandomi, A.H.; Chen, F.; Abualigah, L. Machine Learning Technologies for Big Data Analytics. Electronics 2022, 11, 421.
5. Jia, X.; Cao, Y.; O'Connor, D.; Zhu, J.; Tsang, D.C.W.; Zou, B.; Hou, D. Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. Environ. Pollut. 2021, 270, 116281.