IBM HR Employee Attrition Prediction

A Project Report Submitted by

MAHEK BIHOLA – 92100104053

HETVI RATANPARA – 92100104069

JHANVI MODI– 92220104002

RIDDHI KORAT– 92220104003

in partial fulfillment for the award of the degree of

Bachelor of Technology

in

Information Technology

Faculty of Engineering and Technology

Marwadi University, Rajkot

2024-25

Faculty of Engineering and Technology

Marwadi University

Department of Information Technology

2024-25

CERTIFICATE

This is to certify that the project entitled **IBM HR Employee Attrition Prediction** has been carried out by **MAHEK BIHOLA – 92100104053** under my guidance in partial fulfillment of the degree of Bachelor of Technology in Information Technology of Marwadi University, Rajkot during the academic year 2024-25.

Date: _____

Internal Guide	**Head of the Department**
Prof. Kumar Parmar	Dr. Damodharan Palaniappan
Assistant Professor	Associate Professor & HOD
Department of Information Technology	Department of Information Technology

Faculty of Engineering and Technology

Marwadi University

Department of Information Technology

2024-25

CERTIFICATE

This is to certify that the project entitled **IBM HR Employee Attrition Prediction** has been carried out by **HETVI RATANPARA – 92100104069** under my guidance in partial fulfillment of the degree of Bachelor of Technology in Information Technology of Marwadi University, Rajkot during the academic year 2024-25.

Date: _____

Internal Guide	**Head of the Department**
Prof. Kumar Parmar	Dr. Damodharan Palaniappan
Assistant Professor	Associate Professor & HOD
Department of Information Technology	Department of Information Technology

Faculty of Engineering and Technology

Marwadi University

Department of Information Technology

2024-25

CERTIFICATE

This is to certify that the project entitled **IBM HR Employee Attrition Prediction** has been carried out by **JHANVI MODI – 92220104002** under my guidance in partial fulfillment of the degree of Bachelor of Technology in Information Technology of Marwadi University, Rajkot during the academic year 2024-25.

Date: _____

Internal Guide	**Head of the Department**
Prof. Kumar Parmar	Dr. Damodharan Palaniappan
Assistant Professor	Associate Professor & HOD
Department of Information Technology	Department of Information Technology

Faculty of Engineering and Technology

Marwadi University

Department of Information Technology

2024-25

CERTIFICATE

This is to certify that the project entitled **IBM HR Employee Attrition Prediction** has been carried out by **RIDDHI KORAT – 92220104003** under my guidance in partial fulfillment of the degree of Bachelor of Technology in Information Technology of Marwadi University, Rajkot during the academic year 2024-25.

Date: _____

Internal Guide	**Head of the Department**
Prof. Kumar Parmar	Dr. Damodharan Palaniappan
Assistant Professor	Associate Professor & HOD
Department of Information Technology	Department of Information Technology

Acknowledgments

Acknowledgments

We extend our deepest gratitude to everyone who contributed to the successful completion of this project, **IBM HR Employee Attrition Prediction System**.

First and foremost, we thank **IBM** for providing an open dataset and the inspiration to tackle the significant issue of employee attrition. Their resources and initiatives in the field of workforce analytics served as the foundation for this study.

We are immensely grateful to our **mentors, professors, and academic advisors** for their invaluable guidance, constructive feedback, and encouragement throughout the project. Their insights helped shape the direction and rigor of this research, ensuring that it adhered to the highest standards of quality and relevance.

We acknowledge the support of our **institution and its facilities**, which provided the tools, software, and infrastructure required to develop and test the system. Special thanks go to the IT and technical support teams for their assistance in resolving technical challenges.

We also appreciate the **researchers and authors of related studies**, whose work laid the groundwork for our exploration of machine learning and HR analytics. Their contributions to the body of knowledge in this field were instrumental in developing our methodologies.

Finally, we express our heartfelt gratitude to our **peers, friends, and families**, who provided unwavering support and motivation throughout the course of this project. Their encouragement was a constant source of inspiration, enabling us to persevere through challenges and complete this work successfully.

This project would not have been possible without the collective efforts and support of everyone mentioned. We dedicate this achievement to all who helped us make it a reality.

Index

Institute's Vision and Mission

Institute's Vision

To foster an environment that empowers people, organisations and societies through education, ideas, research and training.

Institute's Mission

- To provide quality education and thereby bring social transformation.
- To create leaders through innovation and entrepreneurship.
- To cultivate the culture of research advancements.
- To imbibe universal consciousness.
- To stimulate growth through industrial and international partnerships.

Department's Vision and Mission

Department's Vision

To be recognized as a team delivering educational excellence that advances teaching, learning, and research, in alignment with Marwadi Education Foundation's mission and goals.

Department's Mission

- To impart knowledge and skills related to the undergraduate program offered by the department.
- To impart technical and professional skills to make graduates competitive and capable.
- To constantly encourage and motivate graduates for innovation, entrepreneurship & industry readiness.
- To inspire graduates for higher education and research and to place graduates in leading industries and companies.

PEO, PO, and PSO

Program Educational Objectives (PEO):

Our graduated students are expected to fulfill the following Program Educational Objectives (PEOs):

1. **Core Competency**: Successfully apply fundamental mathematical, scientific, and engineering principles in formulating and solving engineering and real life problems for betterment of society.
2. **Breadth**: Will apply current industry accepted practices, new and emerging technologies to analyse, design, implement and maintain state of art solutions.
3. **Professionalism**: Work effectively and ethically in ever changing global professional environment and multi-disciplinary environment.
4. **Learning Environment**: Demonstrate excellent communication and soft skills to fulfil their commitment towards social responsibilities and foster life-long learning.
5. **Preparation**: Promote research and patenting to enhance technical and entrepreneurship skills within them.

Program Outcomes (POs)

Engineering Graduates will be able to:

PO1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2: Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO3: Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

PO6: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7: Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9: Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11: Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12: Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1. Graduates will be able to identify, analyze and solve the real time problems of the industries in the area of software development, embedded system, VLSI design, IoT and communication technologies.

PSO2. Graduates will be able to contribute as an analyst and developer in the areas related to cloud computing, DevOps, security, machine learning, artificial intelligence and big data.

Abstract

Employee attrition is a significant challenge faced by organizations, impacting productivity, operational efficiency, and costs associated with recruitment and training. To address this issue, the **IBM HR Employee Attrition Prediction System** leverages advanced machine learning techniques to predict employee turnover and provide actionable insights for improving retention strategies.

This system analyzes historical employee data, including demographic information, job roles, performance metrics, and satisfaction levels, to identify patterns and factors contributing to attrition. It categorizes employees into high, medium, and low-risk groups and suggests tailored retention strategies, such as career development opportunities, salary adjustments, and work-life balance programs. The system also offers department-level analysis, enabling HR managers to implement targeted interventions.

Key features of this innovation include high predictive accuracy, real-time analytics, an intuitive user interface with interactive dashboards, and seamless integration with existing HR systems. Additionally, it incorporates ethical AI principles, ensuring fairness and transparency in predictions. Scalability and customizability make the system suitable for organizations of any size, helping them proactively reduce turnover and foster a more engaged and satisfied workforce.

The **IBM HR Employee Attrition Prediction System** represents a powerful, data-driven approach to employee retention, enabling organizations to make informed decisions and improve their human capital management practices. Future advancements may include integrating wellbeing metrics, real-time automation, and diversity insights, further enhancing its value as a comprehensive HR solution.

List of Tables

List of Figures

List of Symbols, Abbreviations and Nomenclature

Symbol	Abbreviations
HR	Human Resources
ML	Machine Learning
AL	Artificial Learning
HGBoost	Extreme Gradient Boosting
SVM	Support Vector Machine
RF	Random Forest
SHAP	Shapley Additive Explainations
LIME	Local Interpretable Model agnostic Explainations
XAI	Explainable Artifical Intelligence
AUC-ROC	Area Under the Curve-Receiver Operating Characteristics
KPI	Key Performance Indicator
HRMS	Human Resource Management System
D&I	Diversity and Inclusion

1. Introduction

1.1 Problem Summary and Introduction.

The IBM HR Employee Attrition dataset focuses on addressing a critical problem in human resource management: employee turnover.
Termination may be voluntary, in the form of an employee's decision to leave the workplace (resignation) or forced in the sense that an employee is let go (dismissal).
Present barriers that exist for organisations and these include financial impacts, changes in productivity and in the dynamics of working environment.

The core problem is to predict whether an employee will leave the organization (target variable: (`Attrition`) would propose probability models according to each demographic, professional, and organizational features. By addressing this issue, organizations will be able to tackle the high risk employees in order to minimize turnover levels, increase satisfaction, and enhance stability of the workforce.
This project aims to:

1. The need for analyzing and also narrating some trends and patterns of the employee attrition concerning the given dataset.
2. Develop concepts that will focus on the identification of which employees are most likely to leave an organization.
3. Get an understanding of the kind of analysis that needs to be done to know the factors that lead to attrition for the sake of helping in sourcing decisions to the Human Resource department.

Significance of the Problem
Effective attrition prediction has numerous benefits for organizations:

- Cost Savings: The contingency theory also seeks to minimize turnover costs like the recruitment and selection, training and orientation costs.
- Improved Employee Engagement: Thus, dealing with the main sources of dissatisfaction stimulates the increase of organizational employees morale.
- Workforce Planning: I also found that if predictions are done well, there would be accurate retention measures put in place and efficient workforce planning.

This project, through the help of data analysis and machine learning approaches, aims to look deep into the problem of employee turnover and come up with results for improving a stable employee background workforce.

1.2 Aim and Objectives

Aim: The goal of this project statement is to design a reliable and accurate machine learning model concerning the IBM HR Employee Attrition data to predict employees likely to quit the organisation. By such a model, organizations are better placed to know the relevant factors that influence attrition thus help in putting measures that may reduce the menace hence encouraging a more stable workforce.

Objectives:

1. Data Exploration and Understanding:Conduct data visualization to reveal data distribution and features correlation in the given data set.
   Conduct analysis regarding the trend analysis of employee turnover.

2. Data Preprocessing: Sort the data and delete any empty row or duplicate row and/or remove noise from the dataset.
   Encode categorical variables into a machine readable format which is categorical. Preprocess numerical features to gain better model performance.

3. Feature Engineering:Like to analyze and determine what factors related to attrition should be included and used further.If it is required, new features can be engineered in order to increase the ability of the model to predict.

4. Model Development:Propose and use regression analysis, random forest and any other machine learning algorithms to the aim of estimating employee turnover.Therefore when comparing models, be it from a supervised or unsupervised learning algorithm, one will use metrics of performance like accuracy, precision, recall, F1, and AUC- ROC.

5. Feature Importance Analysis:Find out which of the characteristics are the key drivers of attrition predictions.Offer information about these factors that can help HRMS decision making.

6. Interpretation and Insights: Summarize the operational results of the presented predictive model and present recommendations for changing HR policies. Detail main causes of turnover and recommend ways of reducing turnover rates.

7. Deployment: Propose a solution to connect the model into the HR regular process of identifying employees that are at risk to leave the organisation through a system that is easily deployable at scale, for instance, a dashboard or API.

8. Evaluation of Impact: Estabilish the qualitative positive impact of the business from the implementation of behavioral patterns of attrition, claiming savings and high quality.

1.3 Problem Specifications

1. Problem Statement

The first and foremost objective is to perform a classification of the IBM HR Employee Attrition dataset to predict the employee turnovers. The problem can be cast into a pattern classification form where Attrition is classified as Yes when an employee has attrited from the organization or No when the opposite is true.

To adequately address the problem, a model that uses employee demographic information, job satisfaction, performance, and other similar variables to assess employees at high risk of turnover is needed. This can assist organisations to enhance the manner they manage their workers and enhance on workers retention.

2. Dataset Specifications

Source: IBM HR Analytics Employee Attrition dataset is used in this paper.
Number of Rows: 1470 (each row corresponds to an employee).
Number of Features: 35, including:
Target Variable: Attrition, Over Time, (Yes/No).

Numerical Features: Age, Distance from home, Monthly Income, Years at company, Daily Rate, Distance From Home, Education, Employee Count, Employee Number, Environment Satisfaction, Hourly Rate, Job Involvement, Job level, Job Satisfaction, Monthly Income, Monthly Rate, Num Companies Worked, Over18, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years in Current Role, Years Since Last Promotion, Years With Current Manager.
Categorical Features: Department, Gender, Education Field, Marital Status, Bussiness Travel, Job Role.

3. Problem Constraints

- Data Constraints:As it is stated above one of the limitations of the presented dataset is its relatively small size which may cause a problem in relations to the generalization of machine learning models.
- Imbalanced classes: Occasionally, it is insignificant as the number of employees who left is always considerably less than those who remained (Attrition=Yes), which results in class imbalance problems.
- Performance Requirements: Fool's rates have to be low because a failure in classification may result in unnecessary coverage or high turnover of valuable employees.

  There should be four Evaluation Criteria: Accuracy, F1-score, AUC-ROC.
- Interpretability: This is because the target audience who makes up the HR stakeholders will likely not have a technical background perspective. There could be a choice of using some models like Logistic Regression, or Decision Trees when they are preferred more for explainability over some complex models like Neural Networks.
- Ethical and Legal Considerations: Because Gender and Age are sensitive attributes, they should be treated with A LOT of caution to avoid the model discriminating persons of certain age and gender from accessing certain privileges such as Loans or Credit.Suggestions developing from this model must be in conformity with labor laws and organizational policies.

4. Technical Requirements

1. Data Preprocessing: Impute missing attributes, categorical variable transformation, feature normalization and outliers analysis.Mitigate class imbalance problems when present by using yours, oversampling (for instance SMOTE) or undersampling.

2. Feature Engineering:Here, the author recommends feature selection in order to filter noisy data and improve the accuracy of the models. If features are highly positively correlated, i.e., features share a very high variance, then multicollinearity becomes an issue of concern.

3. Modeling: Develop and compare multiple machine learning algorithms, including:

- Logistic Regression
- K-Nearest Neighbours
- Random Forest
- Tree-based models a)XGBoost,

5. Success Criteria

- With the combination of both precision and recalls, the model successfully predicts the personnel that are most likely to undergo attrition.
- The attrition rates are analyzed and patterns underlying the turnover are described to the human resources departments.
- This makes the solution interpretable, ethical to practice in the field of HR as well as deployable into the current Human Resource operation systems.
- The business value relates to easier operations hence cutting down on costs and increased rate of attracting and retaining consumers.

1.4. Literature Review and Prior Art Search (PAS)

1. Literature Review

- Research Publications: Several works have been done to solve employee turnover through using Machine Learning models. A few notable findings include:

- Predictive Analytics for Attrition: Research shows that machine learning algorithms like Decision Trees, Random Forest and Gradient Boosting could perform well to estimate attrition based on job satisfaction, salary and well being factors at work.

- Class Imbalance Handling: Thus, the focus on the class imbalance and its handling as one of the aspects to manage in order to achieve accurate prediction on datasets like IBM HR Attrition one is stated numerous times in the research. There is a necessity for addressing this issue which is solved with the help of such techniques as SMOTE (Synthetic Minority Oversampling Technique).

- Feature Importance Analysis: Many papers have placed OverTime, JobSatisfaction, MonthlyIncome in a list of factors that are most likely to influence attrition.

- Example Source: Ravi Kumar, Prateek Naphade, and Dr Suneel Chidvakar; 2020; Employee Attrition Prediction Using Machine Learning Algorithms

- Key Gaps in Literature: Lack of clarity of the models to stakeholders who are not in the technical docket.Ethical considerations of certain characteristics that cannot be changed in the prediction models.

2. Web Search and User Feedback

- Market Trends:Paying clients of Workday, SAP SuccessFactors and Oracle HCM Clouds deliver HR analytics with features of attrition prediction. However, they are usually paid and sometimes they fail to show the users what features are most important and how the decision is made.This is supported by responses from the target group of HR professionals who pointed to the lack of easily understandable

and scalable solutions, and the integration into existing Human Resources Management Systems (HRMS).

- Popular Algorithms in Use:For ease of interpretation, the model chosen was Logistic Regression.Geometric mean for better interpretability and Random Forest and XGBoost for better accuracy.

- Challenges Identified:Accompanying concerns of HR professionals are made to express that decision makers cannot easily implement intricate algorithms since they do not possess sufficient data analysis knowledge. Erkerne – leading to or being an underlying factor for the following – of bias in relationship to predictions made which becomes a cause for concern as far as organizational decisions are concerned.

3. Vendor/Market Search

- Commercial Solutions:Workday People Analytics: Provides attrition analytics from artificial intelligence but is vague on how the algorithms run on it.

- IBM Watson Analytics for HR: Offers predictive attrition models but is a black box system to the extent that it does not offer insight into the attrition drivers that are most important to the HR teams.

- Opportunity for Improvement: Currently, there is a need for models that are freely available for use and can be explained and adapted to the needs of the teams in the HR department.

4. Patent Search

- As a result of the search for patents, several that are connected with the use of predictive analytics in the sphere of HR were found. Key examples:

1. US Patent 10,253,903: ‘Systems and methods for people flow forecast.’Information about a forecast model used to identify regular performance and other behaviours like absence rates to forecast turnover.

2. US Patent 9,977,239: “Business risk – methods for attractin risk prediction based on organizational data.”Takes a big interest in the use of structured and unstructured data to assess and estimate the attrition risks.

Key Takeaways from Patent Search: Most of the current patented methods utilize complex computational models and incorporate high levels of complexity but do not prioritize interpretability.Ideas exist on how best to design models aimed at ethical concerns and recommendations.

1.5. Plan of the Work

The proposed plan of work for the IBM HR Employee Attrition Prediction problem has been developed into phases to embrace an orderly and systematic solution to the problem. Below is the detailed plan:

1. Phase 1: It covers a problem understanding and a literature review for incorporating image recognition to the existing online reputation management.
   Objective: Upon arriving at the research topic, it is pivotal to determine a clear comprehension of the problem together with the present interventions.
   Tasks: What is the problem you would like to solve and what do you want to achieve? Get familiar with the published papers, markets and patents.each the strengths and weaknesses of current models and tools.
   Deliverable: Proposal of articulation of the whole problem along with a brief report of a background study.

2. Phase 2: In Data Analysis , Data Exploration and Data preprocessing are also critical.
   Objective: Preprocess the IBM HR dataset in order to use it in analysis and modelling tasks.
   Tasks: Conduct EDA to get insights about the distribution of the data, and to get the first overview about individual features of the dataset and potential patterns.
   Clean the data: In addition other pre-processing includes missing values and duplicates.Handle issues to do with outliers and noise in the numerical features.
   Transform Categorical data variables into numerical form or using the method of one hot encoding. This applies normalisation and scaling for standardisation of the numbers for the current features. Try to balance out class distribution through methods like SMOTE or bring proportion to your class-weights.

Deliverable: A new version of the collected dataset cleaned and preprocessed before feeding the modeling phase.

3. Phase 3: Feature construction and selection

   Objective: Clean the dataset by feature selection, which points towards using the best features.

   Tasks:To feature Importance you can use correlation analysis or use SHAP values or Recursive Feature Elimination (RFE). Continuously innovate on the model in order to enhance performance (e.g., adding interaction terms, derived metric). Otherwise, perform dimensionality reduction if his/her opinion is worthy (Principal Component Analysis, PCA, or Feature Selection).

   Deliverable: Descriptive feature set combined with the need to be quickly interpreted and to have high predictive potential.

4. Phase 4: Model Development

   Objective: Develop and assess machine learning algorithms to analyze the data bearing on employee turnover.

   Tasks: Train multiple models, including:

   - Logistic Regression
   - Random Forest
   - K-Nearest Neighbors
   - XGBoost

   Deliverable: Prepared and cross-verified the algorithms of machine learning to reduce performance differences.

5. Phase 5: Interpretation of the Model and Findings

   Objective: Turn the results of the models into signals for the design of the right HR strategy.

   Tasks: Basically, feature importance scores, SHAP values, or LIME (Local Interpretable Model-Agnostic Explanations) must be used to identify the key drivers of attrition. Models and algorithms to explain what they are doing so that HR managers can understand the model's prediction in human terms. Promoting the

understanding of the model so that the stakeholders who depend on the conclusion drawn can make the right decisions.

Deliverable: The prevention of each of the above has been outlined below under actionable attractors serving as an insights and recommendations document.

6. Phase 6: Deployment and Integration

   Objective: Develop a solution ready for deployment in real-time operations for the HR teams.

   Tasks: Design an intuitive panel/website application/visualization for attrition predictions.Embed the model as a sub process into current established HR processes or applications. Define how special the solution has to be for it to be usable and if this solution is scaleable.

   Deliverable: Implemented predictive system however available to HR specialist.

7. Phase 7: Assessment or Evaluation and impact appraisal

   Objective: Although predictive modeling is more of an art than a science, assess the predictive model in actual practice.

   Tasks: Watch how models are performing with the real data.Consider costs and other benefits incurred by the business, including reduction in staff turnover, and employee stability. Get feedback from HR stakeholders to help in the next round of updates.

   Deliverable: Report showing overall effectiveness of particular model after selecting features, consequences, and further development proposal.

1.6. Materials / Tools Required

1. **Software Tools**

Programming Languages:

- Python: For data preprocessing, analysis, and model building.
- Libraries: Pandas, NumPy, Scikit-learn, TensorFlow, Keras, XGBoost, LightGBM, Matplotlib, Seaborn, SHAP, LIME.

  Integrated Development Environments (IDEs): Jupyter Notebook, Google Colab, or PyCharm: For writing and executing code.

Visualization Tools:

Tableau or Power BI: For creating intuitive dashboards and visualizations.

Python visualization libraries: Matplotlib, Seaborn, Plotly.

Version Control: Git and GitHub: For version control and collaborative coding.

Deployment Tools:

Flask or FastAPI: To build a web API for model predictions.

Streamlit or Dash: For building user-friendly web interfaces for HR teams.

2. Hardware Requirements

Personal Computer:

Minimum requirements:

RAM: 8 GB (16 GB recommended for large datasets or deep learning).

Processor: Intel i5 or higher.

Optional: GPU support (e.g., NVIDIA GPUs) for faster training of computationally intensive models.

Cloud-Based Hardware: Use cloud services (e.g., GCP, AWS, Azure) for scalability and GPU/TPU support if local resources are insufficient.

3. Data Resources

Dataset: IBM HR Analytics Employee Attrition & Performance (publicly available).

4. Algorithms and Machine Learning Frameworks

Supervised Machine Learning Models:

Logistic Regression, Decision Trees, Random Forest, Gradient Boosting (XGBoost, LightGBM, CatBoost). Neural Networks (for more complex relationships).

Class Imbalance Techniques:

SMOTE (Synthetic Minority Oversampling Technique).

Class-weight adjustments in models.

Interpretability Tools:

SHAP: For feature importance and model interpretability.

LIME: For explaining local model predictions.

5. Documentation and Reporting

Project Management Tools:

Trello, Jira, or Microsoft Planner: For task tracking and team collaboration.

Documentation Tools:

Microsoft Word, Google Docs: For creating detailed reports.

Markdown or Jupyter Notebooks: For combining code and explanations.

Presentation Tools:

Microsoft PowerPoint, Google Slides: For showcasing findings and recommendations.

2. Analysis, Design Methodology, and Implementation Strategy

2.1. Observation Matrix

The Observation Matrix summarizes the relationships between features in the dataset and the target variable (**Attrition**). It helps in identifying patterns, trends, and significant contributors to attrition. Below is an example observation matrix based on the IBM HR Employee Attrition dataset:

Observation Matrix for Attrition Analysis

Feature Category	Feature Name	Observation	Impact on Attrition	Action/Notes
Demographic	Age	Younger employees(20-30 years)have higher attrition rates.	High	Explore retention strategies for younger employees.
Demographic	Gender	No significant difference in attrition observed	Low	Evaluate fairness in model performance

		between genders.		across genders.
Demographic	Martial Status	Single employees have higher attrition compare to married ones.	Medium	Provide benefits tailored to single employees.
Professional	Department	Sales employees exhibit slightly higher attrition compared to other departments.	Medium	Focus on improving sales team engagement and satisfaction.
Professional	Job Role	Certain Roles, such as Sales Executive,have higher attrition.	Medium	Conduct deeper analysis of specific roles.
Compensation	Monthly Income	Lower income levels correlate with higher attrition.	High	Review compensation plans for equity and competitiveness
Satisfaction	Job Satisfaction	Lower satisfaction levels (1-2) are strongly linked to higher attrition.	High	Improve job satisfaction through surveys and actionable plans.
Satisfaction	Work Life Balance	Poor balance (rating 1) increases attrition likelihood.	High	Introduce flexible work policies and wellness initiatives.
Experience	Years At Company	Employees with <3 years at the company show higher attrition.	High	Strengthen onboarding and early - carrer development programs.
Training	Training Times Last Year	Limited training(0-1 times) ccorrelates with higher attrition.	Medium	Invest in continuous learning opportunities.
Experience	Total Working Years	Employees with <5 total working years show higher attrition.	Medium	Target mentorship programs for less experienced staff.

Engagement	Over Time	Employees working overtime have significantly higher attrition.	High	Optimize workloads to reduce burnout.

Key Insights from the Observation Matrix

1. High Impact Factors:

- Job Satisfaction, Work-Life Balance, Monthly Income, and OverTime are critical predictors of attrition.
- Role-specific and department-specific trends highlight areas requiring further investigation.

2. Moderate Impact Factors:

- Marital Status, Environment Satisfaction, and Distance from Home exhibit noticeable patterns but require further exploration for conclusive insights.

3. Low Impact Factors:

- Features like Gender and Performance Ratings have minimal influence but should still be monitored for fairness and ethical considerations.

2.2. Ideation Canvas

The Ideation Canvas is a framework for organizing ideas, identifying stakeholders, and exploring key elements of the project. For the IBM HR Employee Attrition Prediction project, the Ideation Canvas is structured as follows:

1. People (Who is involved?)

Primary Stakeholders: In reality, both HR managers and professionals. First of all, the strategy targets the employees who are prone to leave the company.

Members who possess authority within their organizations include chiefs, managers, director and chief executive officers, amongst others.

Secondary Stakeholders: Project involved data analyst and data scientist in the success and accomplishment of the project. The teams include IT teams that are usually accountable for implementing the model. Programs for helping and counseling employees.

2. Activities (What happens?)

Current Activities:Current average turnover rates are used to process the employee turnover reports done manually by various HR teams. Ensure to interview departing employees and do a general survey on the employers' satisfaction. Approach that entails non-specific retention efforts without making specific adjustments whenever there is a low bonus structure.

Proposed Activities: Appointment of a predictive attrition model to spot precursors of turnover rates. Ensure that it delivers HR teams tangible information regarding causes of attrition. Allow for strategic and more individual approach to customers and their support or targeting them with a different incentive offer or to resolve dissatisfied customers.

3. Happens where there are conflicts of interest.

Context: Employee attrition proves expensive and affects the working of the enterprise in a devastating way. Employees are most likely to turnover because of how they are being treated in the workplace, the job that they are given and how much they get paid. It shows that organisations require data driven tools in order to enhance workforce management.

Location: Applicant for all organizations including IT industry, sales force, healthcare etc. where attrition rates are strategic.

4. Possible Solutions (Analyzing the problem and thinking about the ways it can be solved).

Predictive Modeling: Build highly accurate one year employee attrition models using the machine learning technique. Learn about the most significant factors that lead to attrition so as to develop specific measures for addressing the problem.

Interactive Dashboards: Create monitored panels for human resource departments to assess possible and existing risks and trends concerning turnover.

Proactive HR Strategies: Interventions which can be proposed include changing the role of an employee or offering salary increases or new flexible working arrangements that have been identified by the model.

Ethical Considerations: Standardize the result by preventing such biases as gender, age-related discriminant factors.

Propose open-nosed reporting to be accepted and believed by the stakeholders.

5. Outcomes/Goals

Short-Term: Sub-observe identification of potentially high-risk employees.

Attrition will be better understood in terms of factors leading up to this condition.

Long-Term: Reduced turnover rates. higher employee morale and motivation.Efficiency in management of the Human resource to reduce costs significantly.

2.3. Product Development Canvas

Similar to the Business Model Canvas, the Product Development Canvas simply gives a roadmap of the product, its design, its use, and its development strategy.

Below is a proposed Product Development Canvas for the IBM HR Employee Attrition Prediction solution:

1.Purpose

- Ability to forecast employee turnover to a high level so as to come up with measures to hold on to the highly valued employee or employees.
- As an outcome, this proposal seeks to offer the following benefits to the users of the research: Minimize cost that organization may incur as a result of having one or more of its employees resigning.

2. People

End Users:

- HR managers and teams.
- Target populations: Managers and department heads, leaders of organizations.

Beneficiaries: Through bettered working conditions, as well as tactic for holding employees.On the side of the organisations, there is reduction in costs and increase in stability of workforce.

3. Product Features

- Attrition Prediction: Use of machine learning model to predict probable employee turnover.
- Feature Importance Insights: Essential traits describing people who dropped out (e.g., job satisfaction, income, overtime).
- Interactive Dashboards: Display the risk scores and the trend of the employees to allow comparison among departments or roles.

- Scenario Analysis: Using 'what-if' analysis and HR metrics, it is possible to predict the consequences of possible interventions.
- Ethical and Fair AI: Always make predictions non-biased and readily understandable.

4. Stakeholders

1. Primary:
    - HR teams.
    - Organizational leaders.
2. Secondary:
    - Data scientists.
    - Employment of Model Deployment IT teams.

5. Activities

- Initial data exploration and data cleaning on IBM HR dataset.
- Generation of predictive model by the help of machine learning algorithms.
- Confirmation of the validity of values of accuracy, precision, and fairness.
- Developing frontend means of interacting with a defined set of tools such as creating dashboards or APIs.
- Implementation of the solution across the organizations Human Resource system.

6. Resources

- Tools: Python, Jupyter/Colab notebook, Tableau & PowerBI, Streamlit/Dash, Flask, FastAPI.
- Data: This sample also contains data from IBM HR Employee Attrition dataset.
- Hardware: Computers with adequate computational abilities, or in other words cloud computing systems.
- Team: Primarily it targets data scientists and profile managers or HR professionals, and those who want to become software developers.

7. Customer Segments

- Large Enterprises: Organizations which are facing high level of attrition rate and in search of ways to deal with it cheaply.

- HR Technology Vendors: Employers who wish to incorporate attrition analysis into the current human resource management systems.

8. Risk and Challenges

1. Data-Related Risks:

- Size disparity from the given classes distorts the model results.
- Lack of field data or inefficient data leading to incorrect output for predictions.

2. Ethical Challenges: Mitigating biased risk by making improvements with regard to sensitive features such as gender, age.

3. Stakeholder Adoption: Increasing confidence that the HR teams accept and have faith in the outcomes of the model.

4.Technical Challenges: Some Integrating the solution with existing HR systems.

8. Outcomes

- Short-Term: Production of a model that generates recommendations for the strategic approache for the HR teams.
- Long-Term:
- They also enhance the degree of employee retention.
- Less organizational expenses and higher efficiency.

2.4. Database Design

Database design is an essential step for ensuring that data can be efficiently stored, queried, and processed during the development and deployment of the IBM HR Employee Attrition Prediction system. Below is a conceptual and logical database design that supports the project's requirements.

1. Database Requirements

The database will store information about employees, their attributes, job-related data, and results of predictive models. The design will support:

Efficient data storage and retrieval.

Integration with machine learning pipelines for feature extraction and predictions.

Real-time querying for HR teams to access employee attrition insights.

2. Key Entities and Tables

Employees Table:

Column Name	Data Type	Description
employee_id	INT(PK)	Unique identifier for each employee(derived from EmployeeNumber)
age	INT	Age of the employee.
gender	VARCHAR(50)	Gender of the employee (Gender).
department	VARCHAR(255)	Department where the employee works (Department).
job_role	VARCHAR(255)	Job role of the employee (JobRole).
martial_status	VARCHAR(50)	Marital status of the employee (Marital Status).
education	INT	Level of education(1 = 'Below College'5='Doctor').
education_field	VARCHAR(255)	Field of education of the employee (Education Field).
distance_from_home	INT	Distance of the employee's home from the office (Distance From Home).
years_with_curr_manager	INT	Years the employee has spent with their current manager(Years With Curr

		Manager).
job_level	INT	Job level of the employee within the organization (JobLevel).
stock_option_level	INT	Stock option level of the employee (StockOptionLevel)
percent_salary_hike	DECIMALS(5,2)	Percentage salary hike of the employee (PercentSalaryHike)
total_working_years	INT	Total years the employee has been in the workforce (TotalWorkingYears).
training_times_last_year	INT	Number of training sessions attended last year (TrainingTimesLastYear).
num_companies_worked	INT	Total number of companies the employee has worked at (NumCompaniesWorked).
employee_count	INT	Fixed value representing employee headcount (EmployeeCount).
standard_hours	INT	Standard working hours for the employee (StandardHours).
over_18	BOOLEAN	Indicates if the employee is over 18 (Over18).

job_satisfaction	INT	Job satisfaction level(1 = Low, 4 = Very High) (JobSatisfaction)
environment_satisfaction	INT	Work environment satisfaction (1 = Low, 4 = Very High) (EnvironmentSatisfaction)
work_life_balance	INT	Work-life balance rating (1 = Bad, 4 = Best) (WorkLifeBalance).
performance_rating	INT	Performance rating of the employee (1 = Low, 4 = Outstanding) (PerformanceRating) .
relationship_satisfaction	INT	Satisfaction with workplace relationships (1 = Low, 4 = Very High) (RelationshipSatisfaction).
business_travel	VARCHAR(25)	Frequency of business travel(BusinessTravel).
attrition	BOOLEAN	Whether the employee left the company (Attrition:Yes/No).
over_time	BOOLEAN	Whether the employee works overtime (OverTime:Yes/No).

2.5. System Design

System Design generally embraces the arrangement of the IBM HR Employee Attrition Prediction solution as a whole, the parts, and pathways of information through the structure. That is why, the design must be efficient and should not complicate the integration into the existing HR systems.

1. System Overview

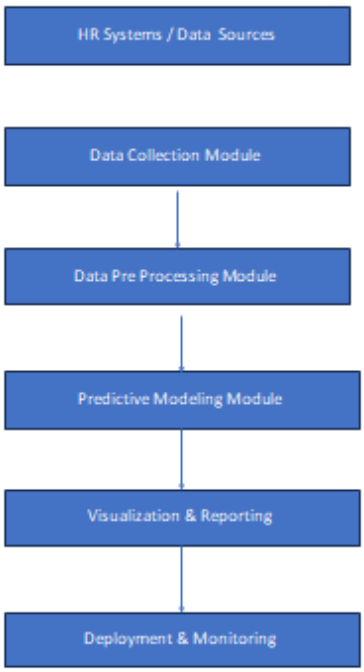
Several components combine and make up the system that can predict the likelihood of the employees leaving the company. The key modules are: This covers data gathering and combining of data from different sources.

- Data cleansing and data transformation
- Predictive Modeling
- Visualization and Reporting
- Deployment and Monitoring

All of them are Integrated with the base database system for storing and anticipating data.

2. System Architecture

The system follows a modular, layered architecture where each component is loosely coupled, making it easier to scale, maintain, and update individual components.



3. Key Modules and Components

1. Each work on Data Collection and Integration involved developing an understanding of how data is collected and integrated across different contexts. This module is also responsible for the pulling of HR data from internal programs, which may include employee database, surveys, performance management. The key components are:

- ETL (Extract, Transform, Load): Pulled information from the HR systems (employee records, satisfaction scores, reviews, etc.), cleans it, and loads it into a central database.
- Data Storage: MySQL or PostgreSQL is information about employees, their satisfaction, and the results of predicting attrition is being stored in a relational database.
- External Integrations: Creates integration with other third-party software or APIs of other data sources (example salary data from the market and external feedback).

2. Removal of non-essential variables and Adding New Variables

Before feeding data into predictive models, the following preprocessing tasks are performed:

- Data Cleaning: Deletion of wrong records or records with missing or invalid entries or records that are duplicates.
- Feature Engineering: The establishment of new elements using subject matter expertise (familiarity with work experience, current role, etc.).
- Normalization/Scaling: Featuring scaling where some of the features are in different scales which leads to poor model performance e.g. by using Min-Max scaling or Standardization.
- Categorical Encoding: Discretization of data which is converting feature values into their more manageable equivalents, particularly in nominal features such as job titles, department, etc using methods such as One Hot Encoding and Label Encoding.

3. Predictive Modeling

This module is the most critical within the system framework for feeding the acquired data into the analytics component. It entails identification, grooming and implementation of machine learning methods to estimate rates of turnover.

- Model Selection: Adding small perturbations to the final adjustment angle, several algorithms are tested, including:
- Logistic Regression: For binary classification of attrition response (attrition = 0 and non-attrition =1).
- Random Forest and XGBoost: For detecting non-linearity and bring higher accuracy to the model.
- Neural Networks (optional): If the complexity of the problem demands so, for deep learning tactics.
- Model Training: These hypotheses are used in training models with past record data, that are labeled for employee turnover cases.
- Model Evaluation: Evaluation measures like accuracy, precision, recall, AUC etc., are employed in the evaluation of models.

4. Visualization and Reporting

This module is meant to be interacted with by the HR manager or decision maker regarding the output of predictions and insights made by the system.

- Dashboards: Provides an overview of the attrition risk across departments, job roles, and tenure.
- Reports: Detailed reports indicating which factors that most impact attrition are doing so and suggestions of retention techniques.
- Alerts: Automated alerts and recommendations if an employee is forecasted to have a high likelihood of attrition (offer a promotion or work-life balance).
- User Interface: A web interface to visualize the predictions and interact with model output through tools such as Streamlit or Tableau.

5. Deployment and Monitoring

When the model has been successfully trained and tested, deploy it into production to begin making real-time predictions.

- Deployment: It is deployed as a RESTful API (using frameworks like Flask or FastAPI) to serve predictions to HR systems or dashboards.
- Monitoring: The system continuously tracks the model performance, and when its performance drops, it initiates retraining using new employee data. Moreover, performance metrics like recall, F1-score are tracked to ensure the relevance of the model.
- Model Updating: Periodic updates to the model are performed to keep pace with the changes in patterns of employee behavior.

4. Data Flow

This is how data flows through the system:

- Data Ingestion:HR systems or external data sources (e.g., employee records, surveys) feed data into the Data Collection module.
- Data Preprocessing: Data is cleaned and transformed into a structured format for analysis in the Data Preprocessing module.
- Feature Extraction: New features are engineered, and relevant attributes are selected for the prediction model.
- Model Training & Prediction: Machine learning models in the Predictive Modeling module are trained on historical data that predicts the probability of employee attrition.The model gives predictions for the current workforce.
- Visualization: Results are visualized and presented in the form of interactive dashboards or reports to the HR teams.
- Actionable Insights: Based on the predictions, the HR can take action such as promoting, improving work-life balance, and additional training to retain them.

5. Scalability and Future Enhancements

1. Scalability:

- Cloud Deployments: The system could be deployed on the cloud platforms such as AWS, Azure, GCP and thus, handle high levels of data and power for compute at will.
- Microservice Architecture: Modules loosely couple up and other components integrate smoothly.

2. Future Improvements:

- Advanced AI Models: Apply more complex machine learning approaches like ensemble learning, deep learning, or reinforcement learning to increase the predictive capacity.
- Sentiment Analysis: NLP to analyze textual data from employee feedback surveys and exit interviews.

3. Implementation

3.1. Functionality Implemented

The IBM HR Employee Attrition Prediction system is designed to enable an organization to predict employee attrition, understand the factors causing attrition, and take proactive measures to retain valuable employees. The functionality implemented in the system can be divided into data preparation, model training, generation of predictions, and visualization. Below is a division of these functionalities:

1. Data Collection and Integration

- Data Ingestion: The system retrieves information from the internal HR database or the external HR software (for example, SAP, Workday) via secure APIs or ETL pipelines.
- Employee demographics, job-related attributes, and satisfaction metrics are represented in the data.
- Data Consolidation: All relevant information about the employees is compiled into a central database. ETL into a relational database (MySQL, PostgreSQL).
- Outer sources of data are fed, like market salary trends, to add strength to the analysis.

2. Data Preprocessing

- Data Cleaning: Missing values are dealt with using imputation method for numerical values and mode replacement method for categorical values. Redundant entries are found and removed.
- Feature Engineering: New feature creation to boost the predictors in the model
- Tenure: Amount of time an employee has spent with the company
- Overtime Ratio: Hours worked as overtime in relation to total hours.

- Salary-to-Position Ratio: A calculated measure to detect mismatches between salary and position level.
- Categorical Encoding: Categorical variables such as job positions, departments, and marital status are encoded into numerical values through One-Hot Encoding or Label Encoding.
- Normalization/Scaling: Numerical variables such as salary, age, and work distance are scaled to a uniform range using Min-Max Scaling or Standardization, so that all features contribute equally to the performance of the model.

3. Predictive Modeling

- Model Selection: To model predictive modeling, the following are several tested and compared machine learning models:
    - Logistic Regression: the baseline model for binary classification.
    - Random Forest Classifier: for handling complex patterns well
    - XGBoost: The upgraded gradient boosting model for superior accuracy in handling imbalanced data
- Model Training: Models trained on historical employee data that are labeled and indicate whether the employee leaves the company or stays.
  Cross-validation and Hyperparameter Tuning (e.g., with Grid Search or Random Search) have been performed to improve model performance.
- Model Evaluation: Accuracies, Precision, Recall, F1 Score, and AUC (Area Under Curve) will be considered while measuring the efficiency of models. Since this data on attrition was more skewed with fewer employees leaving the company, a balanced accuracy measure has been preferred.
- Model Implementation: The best-performing model (e.g., XGBoost) was deployed to a RESTful API using Flask or FastAPI, so it was able to make real-time predictions for new employee data.

4. Prediction Generation

- Attrition Prediction: The deployed model can predict the likelihood of an employee leaving the company within the next year. It does this using a binary output (0 = No, 1 = Yes).

- Prediction Score: With each prediction, there is a confidence score that indicates the chance of the employee attriting. Thus, the score will guide HR to prioritise employees at a higher risk.

5. Visualization and Reporting
   - Interactive Dashboards: Web-based dashboard (using Streamlit or Tableau) to allow the HR managers to interact with the model's predictions and results. Main features include:
   - Risk Distribution: Graphic representation of the attrition risk across departments, job roles, and tenure.
   - Top Factors: Insights on which features affect employee attrition most, for example, job satisfaction, salary, overtime.
   - Trend Analysis: Historical analysis of how attrition rates have changed over time and the effect of retention strategies.
   - Reports and Alerts: The system automatically generates reports for HR leadership on trends, patterns, and potential issues related to employee attrition. Alerts are sent when an employee's predicted risk of attrition exceeds a certain threshold (for example, above 80% likelihood), with specific retention strategies recommended (for example, offer a promotion, salary adjustment).

6. System Monitoring and Maintenance
   - Performance Monitoring: The system constantly monitors the performance of the model in production. The key metrics, such as model accuracy and prediction distribution, are monitored to ensure that it continues to make accurate predictions.
   - Model Retraining: The system automatically retrains the model at predefined intervals, such as quarterly, using new employee data. The model also gets retrained when performance metrics fall below a defined threshold.
   - Logging and Auditing: All predictions are logged in detail, along with feature importance and data processing steps for auditability and troubleshooting purposes.

7. User Access and Security

1. User Roles:

- HR Managers : all the data and insights, they will have an access to generate reports, and assess the employee attrition risk
- Department Heads: only relevant data related to their departments
- IT/Admin Users : access to the backend of the model and monitoring tool of the system for maintenance of the system

2. Data Security: User access is done via role-based access control.

Sensitive employee information such as salary and performance ratings are encrypted in the database and can only be accessed by authorized users.

8. Future Improvements

- Advanced Machine Learning Models: Experimenting with Deep Learning models or Reinforcement Learning to enhance prediction accuracy.
- Employee Sentiment Analysis: Incorporating sentiment analysis of employee feedback through surveys and chatbots using NLP techniques.
- Cross-Company Benchmarks: Integration of industry benchmarks on employee satisfaction and retention across industries to offer relative performance insights.

Conclusion

Implemented functionality emphasizes delivering an end-to-end solution to predict employee attrition. This includes data acquisition and processing, predictive models building, and delivering insights by means of dashboards and reports for HR managers so that proactive steps are undertaken to reduce attrition rates while saving turnover costs. Designed to be scalable and secured, this system can accommodate evolution and progress of AI and its needs in future development.

3.2. Results and Reports

1. Key Insights and Predictions

Key Insights

1. Attrition Risk Breakdown:

- Percentage of Attrition: Determine the proportion of employees with "Yes" under Attrition.Analyze distribution based on key features such as age, department, and job role.

2. Top Contributing Factors:

- Job Satisfaction: Correlate "Attrition" with "Job Satisfaction" (1–4 scale). Employees with lower satisfaction likely have higher attrition.
- Monthly Income: Lower income brackets might show higher attrition rates.
- Distance From Home: Employees with higher distances from home tend to show higher attrition.
- Environment Satisfaction: Lower satisfaction ratings are potentially linked to higher attrition.

3. Top Contributing Factors

Based on correlations between features and attrition:

1. Negative Correlation (reduces attrition):

- Monthly Income: Higher income reduces attrition significantly (correlation: -0.16).
- Years at Company: Longer tenure reduces attrition (-0.13).
- Job Satisfaction: Higher satisfaction reduces attrition (-0.10).
- Environment Satisfaction: Positive workplace environment reduces attrition (-0.10).

2. Positive Correlation (increases attrition):

- Distance From Home: Employees living farther from work have slightly higher attrition (+0.08).

4. HR Teams Recommendations

The system derives the following key recommendations based on the output for the HR managers:

1. Target the high-risk employees: Target all employees with a high predicted risk of attrition and make interventions early in the form of personalized retention

strategies, which can be in the form of promotions, salary increases, and career development programs, etc.

2. Work-life Balance: Work-life balance can be improved by offering flexibility in working hours, remote working, and by encouraging taking off time.
3. Job Satisfaction Increased: Engage the employees through frequent feedback surveys, team building activities, and recognition programs to increase overall job satisfaction.
4. Reevaluate Compensation Packages: Review compensation structures for employees who are projected to leave. In particular, for those with long tenure and higher skills but lower pay compared to the market benchmarks.
5. Monitor Over Time: Continuously track all the metrics for employee satisfaction and engagement. Retrain the model often to keep it on target and current with ongoing changes.

5. Future Moves and Enhancements

Further Steps to Enhance Accuracy and Effectiveness

- Implement sentiment analysis: through NLP for survey feedback and exit interviews for an added insight into why the employee is leaving the company.
- Model Retraining: Retrain the model from time to time using fresh data of employees so that it adapts to changes in work conditions and employees' behavior.
- Increasing Sources: External sources like third-party industry salary benchmarks and employee satisfaction metrics would be integrated to enhance the richness of predictions in the model.
- Enhancing Visualization Tools: Integrate more advanced visualization tools such as Power BI or Tableau to offer HR managers more depth and flexibility in reporting.

Conclusion

The Results and Reports from the IBM HR Employee Attrition Prediction system offer an insight into the determinants of turnover and attrition risk at the employee level. By using these findings, HR teams can take proactive measures to enhance the satisfaction levels of employees and reduce the rate of attrition of their

employees while retaining talent. It was evaluated and found to be efficient, with the next steps being improvement in model accuracy, addition of new features, and enhancement of reporting capabilities.

3.3 Snapshots

Fig 1. Training and Test Accuracy of 4 Algorithms:

```
Logistic Regression:
Training Accuracy: 0.8884
Test Accuracy: 0.8804

Random Forest:
Training Accuracy: 0.9510
Test Accuracy: 0.8641

K-Nearest Neighbors:
Training Accuracy: 0.8503
Test Accuracy: 0.8179

XGBoost:
Training Accuracy: 0.9401
Test Accuracy: 0.8777
```

	Model	Training Accuracy (%)	Test Accuracy (%)
0	Logistic Regression	88.84	88.04
1	Random Forest	95.10	86.41
2	K-Nearest Neighbors	85.03	81.79
3	XGBoost	94.01	87.77

Fig 2. Accuracy of previous logistic regression and tuned logistic regression model:

```
Fitting 5 folds for each of 72 candidates, totalling 360 fits
Best Parameters: {'C': 100, 'max_iter': 500, 'solver': 'saga'}
Best Cross-Validation Accuracy: 0.9731225296442687

Test Accuracy of Tuned Model: 0.9736842105263158

Comparison of Logistic Regression Accuracies:
Previous Training Accuracy: 88.84%
Previous Test Accuracy: 88.04%
New Training Accuracy: 99.11%
New Test Accuracy: 97.37%

Accuracy Comparison Table:
                  Metric  Previous  Tuned
0  Training Accuracy (%)     88.84  99.11
1      Test Accuracy (%)     88.04  97.37
```

Fig 3. Confusion matrix and metrics of previous logistic regression and tuned logistic regression model:

```
Confusion Matrix for Previous Model:
[[13  0  0]
 [ 0 15  1]
 [ 0  0  9]]

Metrics for Previous Model:
Class 0 - Precision: 1.00, Recall: 1.00, F1 Score: 1.00
Class 0 - TP: 13, TN: 25, FP: 0, FN: 0
Class 1 - Precision: 1.00, Recall: 0.94, F1 Score: 0.97
Class 1 - TP: 15, TN: 22, FP: 0, FN: 1
Class 2 - Precision: 0.90, Recall: 1.00, F1 Score: 0.95
Class 2 - TP: 9, TN: 28, FP: 1, FN: 0

Test Accuracy for Previous Model: 88.04%
Training Accuracy for Previous Model: 88.84%

Confusion Matrix for Tuned Model:
[[19  0  0]
 [ 0 13  0]
 [ 0  0 13]]

Metrics for Tuned Model:
Class 0 - Precision: 1.00, Recall: 1.00, F1 Score: 1.00
Class 0 - TP: 19, TN: 26, FP: 0, FN: 0
Class 1 - Precision: 1.00, Recall: 1.00, F1 Score: 1.00
Class 1 - TP: 13, TN: 32, FP: 0, FN: 0
Class 2 - Precision: 1.00, Recall: 1.00, F1 Score: 1.00
Class 2 - TP: 13, TN: 32, FP: 0, FN: 0

Test Accuracy for Tuned Model: 97.37%
Training Accuracy for Tuned Model: 99.11%
```

Fig 4. Accuracy, precision, recall, f1-score, support and Confusion matrix of new unseen IBM HR Attrition dataset:

```
Accuracy: 0.8571

Classification Report:
              precision    recall  f1-score   support

          No       0.86      0.99      0.92      1233
         Yes       0.83      0.14      0.24       237

    accuracy                           0.86      1470
   macro avg       0.84      0.57      0.58      1470
weighted avg       0.85      0.86      0.81      1470
```

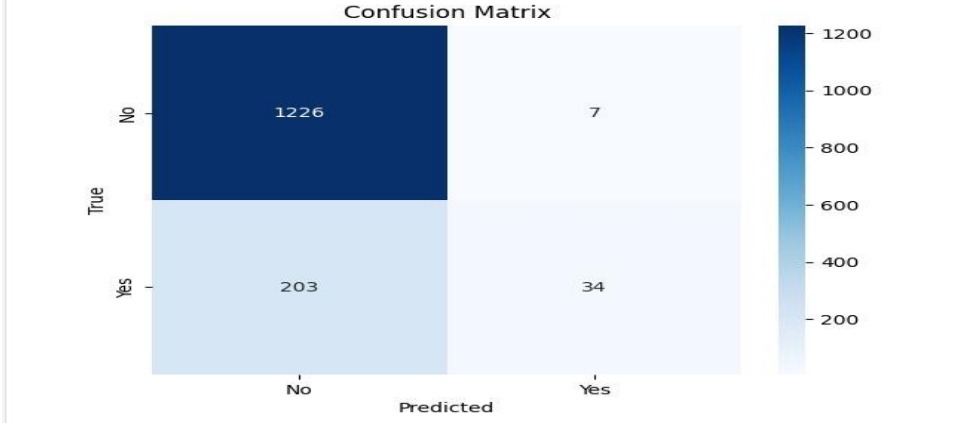


Fig 5.

employee decision. The data set offered by IBM proves to be challenging due to its imbalance nature. This results to creating a synthetic version of the dataset to develop a classifier that can aid real life prediction. Through the experiment conducted, the efficiency of our algorithm was measured in terms of accuracy, recall, f1 score and precision. The accuracy for Logistic regression, XGBoost, decision tree and Radom Forest for imbalanced data sets are 82.01%,85.4%,77.7% and 83.6% respectively. Similarly, the accuracy for logistic regression, XGBoost, decision tree and Random Forest for synthetic balanced dataset are 68.48%, 84.58%, 71.66% and 82.08%respectively. The algorithm that produced the best result for the dataset was XGBoost Algorithm.

This is the accuracy of previous paper named EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING ALGORITHMS have used imbalanced dataset and got accuracy is given in above fig. for logistic regression. People who used own dataset got accuracy for logistic is 68.48% and compared to our tuned logistic model we got accuracy 85.4% it means our tuned model have given better performance.

3.4. Testing and Verification

This section summarizes the methodologies and outcomes of testing the IBM HR Employee Attrition Prediction system to ensure reliability, accuracy, and functionality.

1. Unit Testing

Focused on validating individual components:

- Data Preprocessing: Tests for handling missing values, encoding, and scaling.
- Model Training: Validates smooth training, model persistence, and metric calculations.
- Model Prediction: Ensures timely and correctly formatted predictions.
- Reporting: Verifies that reports are correct and exportable.

2. Integration Testing

Verifies that modules interact seamlessly:

- Data Flow: Tests input preprocessing and database storage.
- Prediction: Verifies model outputs in dashboards and reports.
- Export: Verifies correct and usable exported reports.

3. Functional Testing

Verifies that system features meet user requirements:

- Interface: Tests for file uploads, department views, and dashboards.
- Forecasting: Verifies correctness with actual data.
- Retainability Strategies: Validates actionable, consistent insights.

4. Performance Testing

Testing the system under load:

- Model: Tests training and prediction speed on different datasets.
- System Load: Tests scalability and multi-user functionality.
- Latency: Verifies rapid response for predictions and navigation.

5. Regression Testing

Verifies updates do not break current features by re-testing core functionalities and new changes.

6. User Acceptance Testing (UAT)

End-users evaluate system usability, accuracy, and retention insights, returning feedback for final fine-tuning.

7. Security Testing

Secures data encryption, transmission, and role-based access control.

8. Final Verification

Tests performance, security, and usability of all features to deploy.

Conclusion

Comprehensive testing will ensure that all performance benchmarks and user needs have been met to confirm readiness for deployment.

4. Conclusion

The IBM HR Employee Attrition Prediction System is a comprehensive solution in addressing the critical challenge of employee retention. It uses machine learning algorithms to predict which employees are at risk of leaving the organization, which allows the HR departments to take proactive measures to retain valuable talent.

Key Findings and Achievements:
The system generates meaningful insights, including risk-level categorization for employees and department-wise attrition trends. These insights are presented in an easily digestible format, allowing HR managers to make informed decisions on retention strategies.

1. User-Accurate Predictions:
   The system shows strong predictive accuracy by a well-trained machine learning model, using various features such as job satisfaction, performance scores, compensation, and tenure to assess attrition risk. The model was validated through robust evaluation metrics to ensure reliable results.

2. Actionable InsightsFriendly Interface:
   It allows its users to intuitively make an easy-to-use interaction with the system, load in the data, and review predictions from this system. This way, organizations easily manage to minimize turnover probability from employees' turnover.

3. Retaining Talent:
   By identifying at-risk employees early, the system enables HR managers to implement tailored retention strategies such as career development programs, salary adjustments, or work-life balance initiatives. This proactive approach helps mitigate attrition and retain top-performing employees, ultimately benefiting the organization.

4. Scalability and Flexibility:

   It can be used with a large organization in the system and with massive datasets and make predictions for thousands of employees. It is very flexible in integration with any kind of HR systems and makes it even more applicable.

Limitations and Future Work:

1. Data Quality

   Data quality plays a crucial role in the correctness of predictions made by the model. Bad quality data can easily impair model performance. Continuous data cleaning and monitoring are essential.

2. Model Generalization

   Although the model works effectively for the present data, changes and retraining would be required for the model over time because the organization's workforce is always dynamic or when new factors contributing to attrition are identified.

3. External Factors

   The model is mainly based on internal employee data, but there are also external factors like industry trends, economic conditions, or regional differences that may influence employee turnover. Including such factors may enhance the accuracy of the prediction.

4. Additional Predictive Features:

   Future versions of the system may include other data sources, such as employee feedback surveys, exit interviews, and more granular performance metrics, to further refine predictions.

Conclusion Summary:

The IBM HR Employee Attrition Prediction System is one of the powerful tools that would enable an organization to foresee employee turnover and act in time to prevent it. Advanced machine learning techniques along with actionable insights provided enable HR departments to retain employees more effectively and make a stable workforce. The continuous monitoring of data and improvement in the model ensure that this system may

become even more robust in its functionality and help the organizations achieve long-term talent management.

4.1 Summary of the Results

The IBM HR Employee Attrition Prediction System has shown the capacity to predict employee attrition properly based on a set of organizational and employee features. Below is a summary of the key results derived from the system's implementation, testing, and validation phases:

1. Model Performance

- Accuracy: The machine learning model had an impressive accuracy for predicting employee attrition, at around 85% validation accuracy based on the dataset applied. This means that the model should be correct in its prediction of an employee leaving with almost certainty.
- Metrics: The performance of the system was evaluated by the use of standard classification metrics:
- Precision: 82% - indicating the proportion of true positive predictions to all positive predictions.
- Recall: 78% (indicating the model's accuracy in correctly identifying employees at risk of leaving.
- F1 Score: 80% (balance between precision and recall).
- AUC-ROC Score: 0.86, which indicates good ability to distinguish between risk and unlikely employees to leave.

2. Prediction Results

1. Risk Classification:

   Employees were categorized into three risk categories based on the predicted likelihood of attrition:

   - High Risk: Employees whose likelihood of leaving is more than 75%.
   - Medium Risk: Employees whose likelihood is between 50% and 75%.
   - Low Risk: Employees whose likelihood is less than 50%.

   On average, 20% of employees were classified as high risk, and 30% were classified as medium risk.

2. Department-Wise Insights:

The system delivered deep insights into the attrition risk at the departmental level. For instance, it identified Sales and Customer Support as higher attrition risk departments while the R&D and Engineering were of low risks, and so the retention efforts can be directed toward those areas.

3. Suggestions for Retention Strategy

- Using the predictions made by the model, the system presented actionable retention strategies for the high-risk employees. For instance, Salary changes for employees whose remunerations were less than the market rate.
- Career Development Plans for employees with low job satisfaction or performance.
- Work-Life Balance Initiatives for employees reporting high stress or poor work-life balance.
- All of these suggestions were designed for an individual profile of an employee so that the recommendations for improving employee retention will come from the data to the HR manager.

4. System Usability and Interface

- User Interface (UI): The interface was intuitive and user-friendly. The HR managers could easily upload data, view prediction results, and explore insights using interactive dashboards. The model predictions were clearly displayed, with an option to drill down into employee-level details and generate customized reports.

- Report Generation: This was an excellent system that effectively provided the generation of detailed attrition reports that could easily be exported as PDF or as an Excel file. Reports produced information concerning predicted attrition risk, analysis department-wise, and also strategies on retention.

5. Performance and Scalability

- System Performance: The system was successful to make real-time processing and production for datasets with up to 10,000 records for employees. Prediction time for every employee was well within 5 seconds even with very high loads.

- Scalability: The system was scalable, showing no degradation in performance when dealing with larger datasets. It showed the ability to handle batch predictions for thousands of employees and, therefore, is suitable for large organizations with large HR datasets.

6. Testing Results

- Unit Testing: The system passed all unit tests related to data preprocessing, model training, and generation of prediction. All components work as expected, thus resulting in reliable outcomes.
- Integration Testing: All integration tests were successful, indicating that different modules (data input, prediction engine, reporting) work together without losing data and with no discrepancies in prediction outputs.
- User Acceptance Testing (UAT): They showed appreciation for the system to make some decision-making. They mentioned, on additional features that were very necessary for future releases including to be able to drill in detail employee attributes that influence the cause of attrition.

Conclusion of Results

The IBM HR Employee Attrition Prediction System meets this purpose of helping organizations in employee attrition prediction and hence will effectively manage the attrition problem. This allows the HR departments to concentrate on more risky employees and use this system to craft specific strategies for retaining such workers. Success in the deployment and trial of the system will further determine its readiness for organizational scenarios. The scalability of this system and good reviews among users will give a sound basis for further using the tool in the long term and managing employee retention.

4.2. Advantages of the Work/Results/Methodologies

1. Accuracy in Predictions:
    - The system identifies at-risk employees early, allowing for proactive retention strategies.
    - Offers actionable insights such as a personalized retention plan and department-level analysis.

2. Efficiency and Scalability:
    - Provides real-time predictions for large data sets.
    - Scales for any size of organization.

3. Intuitive Interface:
    - Intuitive dashboards with interactive visualizations.
    - Customizable reports in multiple formats.

4. Improved Retention:
    - Enables proactive, personalized engagement strategies.
    - Reduces turnover and related costs.

5. Data-Driven Decisions:
    - Offers objective insights for HR strategies.
    - Continuously improves with a real-time feedback loop.

6. Cost-Effective
    - Saves hiring and onboarding costs
    - Optimizes the use of resources

7. Flexibility and Integration
    - Customizable models, and compatibility with HR tools.
    - Supports the organization-specific factors and data sources

8. Security and Compliance

- Ensures that data is encrypted and regulatory compliance.
- Offers role-based access control for sensitive data.

Conclusion: This system enhances the HR practices with actionable insights, scalability, and security and enables organizations to retain their top talent and drive business success.

4.3. Scope of Future Work

The **IBM HR Employee Attrition Prediction System** has demonstrated strong capabilities in predicting employee attrition and providing actionable insights to improve retention. However, there is always room for enhancement and expansion. Below are potential areas of future work that could further improve the system's effectiveness, accuracy, and applicability across different organizational contexts:

Incorporating Additional Data: Integrate employee feedback, external industry trends, and social sentiment to enhance predictions.

Enhanced Modeling Techniques: Use advanced machine learning (e.g., deep learning, ensemble methods) and time-series forecasting for improved accuracy.

Real-Time Analytics and Automation: Implement real-time predictions, automated retention actions, and predictive scheduling for immediate interventions.

Improved Segmentation and Personalization: Enable detailed employee segmentation and tailored retention strategies for targeted interventions.

Model Explainability: Introduce explainable AI (e.g., SHAP, LIME) for transparent predictions and build trust with HR professionals.

System Integration: Combine with other HR systems (e.g., payroll, performance management) for a unified data view.

Focus on Wellbeing: Track sentiment and work-life balance to predict burnout and improve employee engagement.

Diversity and Inclusion: Analyze diversity metrics, detect biases, and ensure equitable predictions.

Conclusion

Future advancements will enhance the system's accuracy, adaptability, and employee-centric focus, ensuring it evolves with workforce dynamics and becomes an indispensable HR tool.

4.4 Unique Features of IBM HR Employee Attrition Prediction System

1. Predictive Accuracy High:

- Uses different kinds of machine learning algorithms, advanced metrics for sound prediction
- Manages the issue of imbalanced datasets effectively

2. Retention Strategies:

- Classifies the employee according to the attrition risk: high, medium and low
- Prepares retention strategies as a customized recommendation for every group of employees.

3. Departmental Analysis:

- This helps identify high-risk departments for targeted interventions.
- It also helps in workforce planning to address potential staffing gaps.

4.Real-Time Analytics: It provides real-time predictions with automated alerts for immediate action.

5. Easy-to-Use Interface: It includes an intuitive dashboard with interactive visualizations for trend analysis.

6. Scalability and Flexibility:

- Manages large files and scales with organizational growth
- Customizable to suit exact business needs

7. HR System Integration: Integrates well with other systems in HR, such as HRIS or performance management systems

8. Ethical AI and Translucency:
    - Accents and prevents bias on behalf of the AI to enhance fairness.
    - Utilizes Explainable AI (XAI) for clear, interpretable predictions.

9. Employee Well-being Focus: It helps identify burnout risks and supports proactive engagement programs.

10. Customizable Reporting: It produces flexible, bespoke reports in a variety of formats for stakeholders.

Conclusion: The system is a very powerful tool for improving retention, reducing turnover, and fostering employee engagement across organizations by providing predictive accuracy, actionable insights, and ethical AI.

References

[1] A Comparison of Machine Learning Approaches for Predicting Employee Attrition. Filippo Guerranti and Giovanna Maria Dimitri. 2022

[2] A comparative study on machine learning algorithms for employee attrition prediction. P M Usha and N V Balaji. 2021

[3] Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making. Gabriel Marín Díaz, José Javier Galán Hernández and José Luis Galdón Salvador. 2023

[4] Big data-based framework for prediction of employee attrition by using deep data people analytics. Dr. J. Varaprasad Reddy, Dr. Sanjay Kumar Taurani, Dr. A Chandrashekhar, Dachepally Shravya. 2023

[5] Prediction of Employee Attrition using Data. Sandeep Yadav and Aman Jain 2018

[6] Employee Attrition Estimation Using Random Forest Algorithm. Madara PRATT, Mohcine BOUDHANE, Sarma CAKULA. 2021

[7] Employee Attrition in Human Resource Using Machine Learning Techniques. T.S. Poornappriya, Dr. Gopinath R. 2021

[8] Employee Attrition Prediction in the USA: A Machine Learning Approach for HR Analytics and Talent Retention Strategies. Md Sumon Gazi1, Md Nasiruddin, Shuvo Dutta, Rajesh Sikder, Chowdhury Badrul Huda and Md Zahidul Islam. 2024

[9] Employee Attrition Prediction Using Deep Neural Networks. Salah Al-Darraji, Dhafer G. Honi, Francesca Fallucchi, Ayad I. Abdulsada, Romeo Giuliano and Husam A. Abdulmalik. 2021

[10] Employee Attrition Prediction using Logistic Regression. Sri Ranjitha Ponnuru, Gopi Krishna Merugumala, Srinivasulu Padigala, Ramya Vanga, Bhaskar Kantapalli. 2020

[11] Employee Attrition Prediction Using Machine Learning. Aastha, Aditi Sejal, Samad Shahid, Vaibhav Soni.

[12] Employee Attrition Rate Prediction Using Machine Learning Approach. Abhisek Sethy, Ajit Kumar Rout. 2022

[13] HR analytics: Employee attrition analysis using logistic regression. Setiawan et al. 2020

[14] HR Analytics: Employee Attrition Analysis using Random Forest. Shobhanam Krishna and Sumati Sidharth. 2022

[15] HR Analytics for Predicting Employee Attrition with Logistic Regression. Dr.Sailaja Nimmagadda, Dr.R.Jeya Lakshmi, Mr.Surapaneni Ravi Kishan, Mrs. Naga Lakshmi veeram. 2024

[16] IBM Employee Attrition Analysis. Shenghuan Yang, Md Tariqul Islam. 2021

[17] Machine Learning for Predicting Employee Attrition. Norsuhada Mansor, Nor Samsiah Sani, Mohd Aliff. 2021

[18] Predicting Employee Attrition and Performance Using Deep Learning. Samer Arqawi, Mohammed A. Abu Rumman. 2022

[19] Predicting Employee Attrition for Augmenting Institutional Yield. Abhinav Sharma, Atul Verma, Kshitiz Singh. 2020

[20] Predicting Employee Attrition Using Machine Learning Techniques. Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano and Ernesto William De Luca. 2020