A MINOR PROJECT REPORT

ON

**EMPLOYEE LAYOFF PREDICTION**

SUBMITTED IN PARTIAL FULFILLMENT FOR THE AWARD OF DEGREE OF

**BACHELOR OF TECHNOLOGY**

**IN**

**ELECTRONICS AND COMMUNICATION ENGINEERING**



**Submitted by:** **Under the Guidance of**

**PRAGY UPADHYAY(9922102016) DR. BAJRANG BANSAL**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**

**JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA (U.P.)**

**November, 2024**

**DECLARATION**

We hereby declare that this written submission represents our own ideas in our own words and where others' ideas or words have been included, have been adequately cited and referenced the original sources.We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission.

Place: Noida, Uttar Pradesh

Date: 18/11/2024

Name:Pragy Upadhyay

Enrollment: 9922102016

**CERTIFICATE**

This is to certify that the minor project report entitled, “**Employee Layoff Prediction**” submitted by **Shukla, Pragy Upadhyay, in** partial fulfillment of the requirements for the award of Bachelor of Technology Degree in **Electronics and Communication Engineering** of the Jaypee Institute of InformationTechnology, Noida is an authentic work carried out by them under my supervision and guidance. The matter embodied in this report is original and has not been submitted for the award of any other degree.

**Signature of Supervisor:**

**Name of the Supervisor: Dr. Bajrang Bansal**

**ECE Department,**

**JIIT, Sec-128,**

**Noida-201304**

**Dated:18/11/2024**

**ABSTRACT**

In the modern business environment, efficient workforce management is essential to maintaining organizational stability and competitiveness. This project, Employee Layoff Prediction System, aims to address the need for predictive insights into employee layoffs using machine learning models. By analyzing historical employee data, this system predicts the likelihood of individual layoffs, allowing organizations to make informed HR decisions that can minimize turnover and support proactive workforce planning.

The system is developed using Python and focuses on implementing machine learning algorithms to analyze key employee features, such as tenure, performance metrics, role, and department. Data preprocessing techniques, including data cleaning and normalization, prepare the dataset for accurate model training, enhancing prediction reliability. Various machine learning models were tested to identify the best approach, balancing predictive accuracy with interpretability.

To present the results clearly, data visualization techniques are employed, enabling HR professionals to understand layoff probabilities and the underlying factors driving these predictions. Graphs, probability distributions, and key feature impacts provide transparency into the model’s functioning and foster data-driven insights into potential layoff risks within the organization.

Key features of this project include its high predictive accuracy, achieved through model selection and tuning, and its scalability, making it adaptable to diverse organizational requirements and larger datasets. Preliminary results indicate that this system can significantly enhance HR departments' decision-making capabilities, enabling them to manage workforce transitions more effectively and with reduced unexpected turnover. This project ultimately aims to contribute to organizational resilience and strengthen support for employee welfare by providing actionable insights into potential layoffs before they occur.

**ACKNOWLEDGEMENTS**

I would like to express my heartfelt gratitude to all those who provided invaluable support and guidance throughout the course of this project. I am especially grateful to my supervisor and mentor, Dr. Bajrang Bansal, whose expertise, encouragement, and insightful feedback were instrumental in guiding this work to completion.

I extend my sincere thanks to the members of my panel coordinators, Dr. Priyanka Kwatra and Mr. Abhay Pratap Singh, for their valuable insights, constructive criticism, and support during the evaluation process. Their feedback has been crucial in refining the quality of this project.

I am also grateful to Jaypee Institute of Information and Technology, Noida Sector-128, for fostering an academic environment that encouraged learning and exploration and for providing the resources necessary for the success of this project.

Finally, I would like to acknowledge the assistance of the technical staff and my peers, whose support and feedback were invaluable at various stages of this project.

**TABLE OF CONTENTS**

1. Introduction 7
2. Literature Survey 8
3. Problem Formulation 21
4. Dataset Review 22
5. Data Preprocessing 24
   1. Model Training 24
   2. Evaluation 25
   3. Hyperparameter Tuning 25
   4. Performance Evaluation 26
6. Methodology 27
   1. Logistic Regression 27
   2. AdaBoost 28
   3. Random Forest 28
   4. Support Vector Machines (SVM) 29
   5. Decision Trees 29
   6. K-Nearest Neighbor (KNN) 30
   7. Naïve Bayes 31
   8. System Architecture 32
7. Result & Discussion 32
8. Conclusion & Future Scope 35
   1. Conclusion 35
   2. Future Scope 35
      1. Integration of Advanced Machine Learning Techniques 35
      2. Real-Time Attrition Monitoring 35
      3. Expanded Data Collection 36
      4. Improved Handling of Imbalanced Data 36
      5. Ethical and Fair AI 36
9. References 37
10. Appendices 38
    1. Dataset Summary 38
    2. Machine Learning Models Overview 38
    3. Performance Metrics Formulas 39
    4. Figures and Vizualizations 39
    5. References 39
11. **INTRODUCTION**

Employee attrition, which occurs when individuals leave a firm for various reasons, presents significant challenges for both organizations and employees. High attrition rates impose a substantial financial burden on companies, impact service quality, and can lead to a loss of clients. In recent years, attrition rates have surged across the industry; studies indicate an increase to 19% from 12.9% in 2021 among leading IT firms in India, with current average rates hovering around 25% in the sector. Although some types of attrition, such as retirements or relocations, are inevitable, excessive attrition can severely harm a company’s financial health and corporate culture. Retaining skilled employees has thus become a top priority, pushing organizations to foster employee-centric cultures that address workforce needs and improve overall job satisfaction[1].

The evolution of artificial intelligence (AI) and machine learning (ML) has revolutionized how organizations address complex issues like employee attrition. Machine learning’s practical applications now enable us to analyze vast datasets, identify patterns, and predict outcomes, supporting proactive decision-making. Within the context of employee attrition, ML algorithms can help uncover root causes and identify factors that increase the risk of employee turnover. By understanding these factors, companies can take targeted steps to create supportive work environments, enhancing employee retention. However, limited access to quality data remains a key obstacle, as companies often guard sensitive internal information. Additionally, quantifying emotional and behavioural influences poses a significant challenge, leaving gaps in current attrition models.

Researchers have applied a variety of ML algorithms to predict attrition, including logistic regression, decision trees, and support vector machines (SVM), as well as ensemble methods such asAdaBoost, and Random Forest. Ensemble methods combine multiple algorithms to improve prediction accuracy, with Random Forest being a prominent example that aggregates the outputs of several decision trees. In this project, we focus on understanding these techniques and exploring how ML can be applied to effectively predict and address employee attrition[2].

The ability to predict employee attrition with high accuracy is critical for organizations looking to mitigate its negative impacts. By leveraging machine learning, companies can develop predictive models that analyze factors such as job satisfaction, compensation, career growth opportunities, work-life balance, and even leadership style, which are commonly linked to employee turnover. These models not only allow companies to identify at-risk employees but also provide insights into which interventions—such as changes in job roles, management styles, or benefits packages—could improve retention rates. Furthermore, data-driven approaches to employee engagement can help organizations implement personalized strategies that foster long-term commitment, ensuring that employees feel valued and motivated[3].

1. **LITERATURE SURVEY**
   1. **PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING TECHNIQUES**

The research paper titled "Predicting Employee Attrition Using Machine Learning Techniques"investigates how machine learning can be leveraged to predict employee turnover, providing actionable insights for human resources (HR) departments. Employee attrition, defined as voluntary or involuntary separation from a company, poses significant challenges to organizations, as it leads to the loss of valuable talent and resources invested in recruiting and training. The study aims to identify the primary factors influencing an employee's decision to leave and develop predictive models that can forecast attrition. By utilizing data-driven methodologies, the paper emphasizes how artificial intelligence (AI) can replace subjective decision-making in HR with objective analysis. The research uses a dataset provided by IBM, comprising 1,500 samples and 35 features, including variables related to employee demographics, job roles, levels of satisfaction, compensation, and professional experiences. Preprocessing steps, such as data cleaning, encoding categorical variables, and exploratory analysis, were carried out to ensure data quality and readiness for machine learning. Features like monthly income, distance from home, overtime, job satisfaction, and work-life balance were found to have strong correlations with employee attrition. To build predictive models, the authors tested a range of machine learning algorithms, including Gaussian Naïve Bayes, Logistic Regression, Decision Trees, and Support Vector Machines (SVM). The dataset was split into training and testing sets, ensuring balanced distributions of the target variable—attrition status. Cross-validation techniques were also used to enhance the reliability of the results. Among the algorithms, the Gaussian Naïve Bayes classifier emerged as the most effective model, achieving a recall rate of 54% and a false negative rate of 4.5%. This indicates that the model was successful in identifying the majority of employees who were at risk of leaving, a critical factor in mitigating turnover. The study's analysis revealed several key predictors of attrition. Employees with lower monthly incomes exhibited the highest rates of turnover, highlighting the importance of competitive compensation. Similarly, younger employees and those early in their careers were more likely to leave, potentially due to unmet career growth expectations. Workplace factors such as long commutes and excessive overtime also played significant roles, as they contribute to dissatisfaction. Employees reporting lower levels of job satisfaction, environment satisfaction, or job involvement were more likely to consider leaving, emphasizing the need for supportive and engaging work environments[4].

* + 1. **RESEARCH GAP**

The research concludes that machine learning can be a powerful tool for HR departments, enabling them to adopt a proactive approach to employee management. By identifying at-risk employees early, organizations can implement strategies to address dissatisfaction, such as improving compensation, providing career development opportunities, or fostering a better work-life balance. The study also serves as a foundation for further exploration into predictive HR analytics. Future research could expand on these findings by incorporating larger and more diverse datasets, applying advanced feature engineering techniques, and considering external factors such as labor market conditions or adverse workplace scenarios. Overall, the paper highlights the transformative potential of AI in HR management, shifting the focus from reactive to preventative measures. By replacing intuition and guesswork with data-driven models, companies can better understand and address the underlying causes of attrition, ultimately fostering a more satisfied and stable workforce[4].

* 1. **ANALYZING EMPLOYEE ATTRITION USING MACHINE LEARNING**

The journal article "Analysing Employee Attrition Using Machine Learning" explores the application of machine learning techniques, particularly classification and clustering, to predict employee attrition and inform HR management strategies. Employee turnover is a critical issue for organizations, especially those relying on skilled knowledge workers, as it adversely affects productivity and competitiveness. Using the IBM Employee Attrition dataset with 1,470 records and 35 features, the study employs the Weka tool to preprocess the data and evaluate the performance of various algorithms in identifying patterns associated with employee attrition. Classification techniques such as J48 (a decision tree based on the C4.5 algorithm) and Naïve Bayes were implemented to predict attrition outcomes. J48 demonstrated superior accuracy, achieving 82.4% with tenfold cross-validation and 82.8% with a 70% data split, though it required more processing time compared to Naïve Bayes. The Naïve Bayes algorithm, while slightly less accurate, provided faster computations, with accuracy rates of 78.8% and 81.0% under similar conditions. Additionally, clustering methods, including K-Means and Expectation Maximization (EM), were used to group employees based on shared characteristics related to attrition or retention. Among these, K-Means showed better performance with a correct classification rate of 57.3% compared to EM's 55.1%. However, clustering methods were found to be less precise than supervised classification models for individual predictions. The study concludes that J48 is the most effective algorithm for predicting employee attrition due to its higher accuracy and ability to uncover actionable patterns for HR decision-making. Clustering methods, while helpful in identifying group-level trends, are less reliable for pinpointing individual attrition risks. Machine learning tools like Weka provide HR departments with powerful predictive capabilities, enabling them to identify at-risk employees and implement targeted strategies to reduce turnover. The study highlights the role of data-driven approaches in mitigating attrition and fostering organizational stability[5].

* + 1. **RESEARCH GAP**

Despite these contributions, the research has certain gaps. Firstly, the dataset used is limited to fictional employee data provided by IBM, which may not fully capture the complexities of real-world scenarios. Moreover, the study primarily focuses on commonly used algorithms like J48 and Naïve Bayes, overlooking advanced machine learning techniques such as ensemble methods, deep learning, or hybrid models that might yield improved accuracy and insights. The research also lacks an exploration of external factors, such as industry trends, labor market conditions, or macroeconomic variables, which could significantly impact attrition rates. Future studies could address these limitations by incorporating real-world datasets, exploring more sophisticated algorithms, and analysing external variables to provide a more comprehensive understanding of employee attrition[5].

* 1. **EXPLAINING AND PREDICTING EMPLOYEES’ ATTRITION: A MACHINE LEARNING APPROACH**

The journal article "Explaining and Predicting Employees’ Attrition: A Machine Learning Approach" focuses on using machine learning techniques to address the issue of employee attrition. Employee turnover presents significant challenges for organizations, as it not only results in the loss of valuable skills but also incurs costs related to recruitment and training. The study seeks to develop a predictive model capable of identifying employees at risk of leaving and to explore the factors influencing attrition. By leveraging machine learning, the researchers aim to enhance retention strategies and provide actionable insights for HR management. The study uses a human resources dataset containing over 14,000 records and ten key features such as employee satisfaction, last evaluation, number of projects, salary, and promotions. The data was pre-processed to address missing values and categorized into predictors and target variables. Analytical methods included univariate and bivariate analyses to identify correlations and visualize relationships between variables. For instance, satisfaction levels and salary were found to be critical factors in predicting attrition, with lower satisfaction and salaries correlating strongly with higher turnover. The authors implemented several machines learning algorithms, including Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF), to train and validate predictive models. Cross-validation and confusion matrices were employed to evaluate model performance, focusing on metrics such as precision, recall, and F1-score. Among the models, Random Forest emerged as the most effective, achieving 99% accuracy in identifying employees likely to leave, outperforming SVM and Decision Tree models. The study concludes that machine learning can significantly enhance HR decision-making by enabling targeted interventions for at-risk employees, thereby reducing turnover and associated costs[6].

* + 1. **RESEARCH GAP**

While the research provides valuable insights, certain gaps remain. The dataset used is limited to ten features, potentially omitting critical variables such as work-life balance, leadership quality, or external labour market conditions that may influence attrition. Additionally, while the study demonstrates high accuracy, it does not address the potential biases in the dataset that could affect model generalizability across different industries or regions. Future research could explore the inclusion of more comprehensive datasets, advanced machine learning techniques such as ensemble methods or deep learning, and the impact of external factors on attrition. Addressing these gaps would provide a more holistic understanding of employee turnover and further refine predictive capabilities[6].

* 1. **MACHINE LEARNING APPROACH FOR EMPLOYEE ATTRITION ANALYSIS**

This research paper explores the use of machine learning techniques to analyze and predict employee attrition, a critical challenge faced by organizations aiming to retain their workforce. Employee attrition not only disrupts organizational operations but also incurs significant costs in terms of hiring and training new talent. The study aims to identify key factors driving attrition and evaluate the accuracy of various machine learning algorithms in predicting it. By understanding these dynamics, organizations can proactively address attrition risks and develop effective retention strategies. The researchers utilized a simulated dataset from Kaggle comprising 15,000 employee records, with attributes such as satisfaction level, performance evaluations, number of projects, monthly work hours, years of tenure, history of accidents, recent promotions, departmental roles, and salary levels. These factors were chosen for their potential influence on employee decisions to stay or leave. The dataset was divided into training, testing, and validation subsets in proportions of 70%, 15%, and 15% respectively, ensuring a robust evaluation of model performance. Four supervised learning algorithms were employed: Decision Trees, Random Forest, Support Vector Machines (SVM), and Linear Regression. Decision Trees were used as a baseline to map the relationships between input attributes and attrition outcomes, with satisfaction level emerging as the primary predictor. Random Forest extended this analysis by constructing multiple decision trees on random data subsets, achieving the highest predictive accuracy. SVM attempted to classify attrition data using hyperplanes, while Linear Regression provided a traditional statistical baseline but was significantly outperformed by the other models. The evaluation relied on confusion matrices and Pseudo R-Square values, revealing that Random Forest, with a Pseudo R-Square value of 0.9773, was the most effective model, outperforming Decision Trees (0.8473), SVM (0.8315), and Linear Regression (0.2299). The study concludes that employee satisfaction level is the most critical determinant of attrition, surpassing factors like salary, promotion history, and departmental role. The findings suggest that organizations focusing on enhancing employee satisfaction could significantly reduce attrition rates. The Random Forest model's high accuracy underscores its utility for HR analytics, offering a reliable tool for predicting attrition and aiding strategic workforce management[7].

* + 1. **RESEARCH GAP**

While the research provides valuable insights, it highlights certain gaps that limit its applicability and open avenues for future exploration. A significant limitation lies in the use of a simulated dataset from Kaggle. Although comprehensive, such datasets may not fully reflect the complexities of real-world organizational environments. Future studies could enhance the generalizability of these findings by applying similar methodologies to real-world datasets across various industries and geographies. Additionally, the study primarily focuses on quantitative variables, such as satisfaction levels and salary, but excludes qualitative factors like workplace culture, employee engagement, and individual career aspirations. Incorporating these factors could provide a more holistic understanding of the causes of attrition. Another notable gap is the exclusion of advanced and emerging machine learning techniques. While the study evaluates prominent models like Random Forest and SVM, it does not explore deep learning approaches or ensemble methods beyond Random Forest. These newer techniques might offer improved predictive accuracy and better adaptability to complex datasets. Moreover, the dataset used is cross-sectional, representing a single time point. Employing longitudinal data that tracks employees over time could yield deeper insights into attrition trends and help identify early warning signs. The research also lacks a discussion on the practical deployment of these models. While the algorithms were evaluated in a controlled setting, integrating them into real-time HR decision-making processes remains unexplored. Issues of scalability, model retraining, and alignment with organizational workflows are critical for operationalizing these insights. Addressing these gaps will not only refine the application of machine learning to HR challenges but also expand its potential for strategic workforce management[7].

* 1. **PREDICTING EMPLOYEE ATTRITION USING MACHINE LEARNING APPROACHES**

This research paper presents a comprehensive exploration of using machine learning (ML) techniques to predict employee attrition, addressing a persistent challenge in human resource management. Attrition, defined as the reduction in workforce due to employee departures, can result in substantial organizational costs. With reports showing a 57.3% attrition rate in 2021 and average hiring costs exceeding $4,000 per employee, the study underscores the necessity for predictive tools that enable proactive retention strategies. The research leverages the IBM HR Employee Attrition dataset, which contains 1,470 employee records and 35 features, encompassing attributes like age, income, job level, and years of service. The study applies Employee Exploratory Data Analysis (EEDA) to uncover trends and relationships within the dataset, revealing that factors such as lower monthly income, younger age, and fewer years of service are significant predictors of attrition. Given the dataset’s inherent imbalance, the authors employed the Synthetic Minority Oversampling Technique (SMOTE) to ensure equitable representation of target classes during model training, thereby improving model performance. Four machine learning algorithms were evaluated: Extra Trees Classifier (ETC), Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree Classifier (DTC). The ETC approach, an advanced ensemble learning technique, emerged as the most effective, achieving a prediction accuracy of 93%. This outperformed the other models, with SVM achieving 87%, DTC at 83%, and LR at 72%. The study employed rigorous methods, including k-fold cross-validation, hyperparameter tuning, and standard evaluation metrics such as precision, recall, F1 score, and ROC curve analysis, to validate model performance. Metrics such as precision and recall for the ETC model reached 93%, underscoring its reliability for predicting employee attrition. Key findings highlighted that monthly income, job level, hourly rate, and age were the most influential factors affecting employee turnover. The analysis also demonstrated that attrition rates were higher among younger employees with limited tenure and lower income levels. The proposed ETC model was benchmarked against prior studies utilizing methods such as Random Forest and Gradient Boosting, demonstrating superior performance and offering a robust predictive tool for HR practitioners. These insights equip organizations with the ability to target at-risk employees with tailored retention strategies, ultimately improving workforce stability[8].

* + 1. **RESEARCH GAP**

Despite the study's robust methodological framework and significant findings, it leaves room for further exploration. The research relies on a relatively small and structured dataset of 1,470 records from a single source, which may limit the generalizability of its results across diverse organizational settings. Additionally, the analysis emphasizes structured data while excluding unstructured inputs, such as textual employee feedback or performance evaluations, which could provide a richer understanding of attrition dynamics. Future studies could address these limitations by using larger, multi-industry datasets and incorporating Natural Language Processing (NLP) to analyse unstructured data. Moreover, the absence of deep learning models leaves unexplored potential for uncovering complex relationships between predictors, especially in larger datasets. Operationalizing these predictive models into real-time HR systems and studying their direct impact on employee retention policies would further enhance the practical utility of this research[8].

* 1. **EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING**

This research paper explores the application of machine learning techniques to predict employee attrition, a significant issue faced by organizations globally. Employee attrition leads to substantial costs due to recruitment and training, in addition to the loss of expertise and experience within an organization. The paper emphasizes the importance of predicting attrition rates using data-driven models to help organizations adopt proactive measures for employee retention The study begins with data collection from the IBM HR Analytics dataset, which contains 35 attributes such as age, monthly income, years with the current manager, job role, and distance from home. This dataset was subjected to pre-processing to remove irrelevant features and handle data imbalances. The authors used the Synthetic Minority Oversampling Technique (SMOTE) to address the imbalance, as the majority of records indicated non-attrition, which could skew predictive accuracy. Exploratory Data Analysis (EDA) revealed patterns such as higher attrition rates among younger employees with lower incomes and shorter tenures. Features like monthly income, job role, and overtime were identified as significant predictors of attrition. The authors conducted comparative analyses of five machine learning algorithms: Random Forest, K-Nearest Neighbours (KNN), Decision Tree, Logistic Regression, and Stochastic Gradient Descent (SGD). Random Forest outperformed other models with the highest accuracy, demonstrating its reliability in handling complex datasets with multiple features. This model works by constructing numerous decision trees during training and voting on the most probable outcomes, ensuring robust predictions. Performance metrics such as accuracy, precision, recall, and F1-score validated the effectiveness of the Random Forest algorithm. Other models like KNN and Decision Tree also showed promise but lagged in accuracy compared to Random Forest. The study underscores the importance of feature selection in improving model performance. Features like employee count, employee number, and irrelevant categorical variables were removed to streamline the model and focus on significant attributes. Feature importance analysis revealed that attributes such as monthly income, years at the company, and overtime had a substantial impact on attrition predictions. The authors propose that by identifying atrisk employees early, organizations can implement strategies to improve job satisfaction and reduce turnover[9].

* + 1. **RESEARCH GAP**

Despite the study’s insightful findings, it has several limitations that suggest areas for further research. The authors relied on the IBM HR Analytics dataset, which, although widely used, represents a controlled dataset and may not generalize across industries or cultural contexts. Expanding the study to include diverse datasets from various organizational settings could enhance the model’s applicability. Additionally, the study focuses on structured data but does not integrate unstructured data sources such as employee feedback, performance reviews, or exit interviews, which could provide deeper insights. Future research could also explore advanced techniques like deep learning models or ensemble methods beyond Random Forest to uncover complex relationships within the data. Lastly, while the models show high predictive accuracy, the operationalization of these predictions in real-world HR systems remains unexplored, including how such insights can be seamlessly integrated into decision- making processes. Addressing these gaps could elevate the practical utility and scope of attrition prediction models[9].

* 1. **MACHINE LEARNING FOR PREDICTING EMPLOYEE ATTRITION**

The paper titled "Machine Learning for Predicting Employee Attrition" by Mansor, Sani, and Aliff delves into the application of machine learning to address the critical issue of employee attrition in organizations. Employee attrition, whether due to voluntary resignation or external factors, poses a significant challenge for companies, leading to the loss of valuable knowledge, increased hiring costs, and disruptions in operations. In this context, the paper aims to enhance the predictive capabilities of machine learning models to identify potential attrition and assist human resource departments in proactive decision-making. The study evaluates three popular machine learning algorithms—Decision Tree (DT), Support Vector Machines (SVM), and Artificial Neural Networks (ANN)—to identify the most effective model for this predictive task. The authors used the IBM Human Resource Analytic Employee Attrition and Performance dataset, a synthetic dataset featuring 1,470 employee records with 35 attributes, including factors such as age, salary, job role, and job satisfaction. The dataset, characterized by an imbalanced class distribution (16% attrition), underwent extensive preprocessing to ensure data quality and relevance. This involved cleaning, reducing, and transforming the dataset by removing irrelevant features and standardizing numerical attributes. A critical step in this process was addressing the class imbalance using the Synthetic Minority Oversampling Technique (SMOTE), which created synthetic examples of the minority class to improve model accuracy. The research methodology followed a structured approach, beginning with data preprocessing to enhance the dataset’s suitability for machine learning. Feature selection was performed using techniques such as correlation analysis and gain ratio to identify the 15 most predictive attributes from the original 35. These attributes included critical factors such as "Overtime," "Monthly Income," and "Years with Current Manager." Following this, the dataset was split into training and testing subsets, and three classification algorithms were implemented. Each model was trained using k-fold cross-validation to ensure robust validation. The study also incorporated parameter tuning and regularization techniques to optimize the performance of the models and mitigate overfitting. The comparative analysis of the algorithms revealed distinct strengths and limitations. Decision Tree (DT) demonstrated simplicity and quick training times but lagged behind the other models in terms of predictive accuracy. On the training dataset, DTachieved 84.40% accuracy, but its performance dropped slightly on the test dataset to 80.95%, indicating limited generalization capability. Support Vector Machines (SVM), on the other hand, emerged as the best-performing model. With an optimized Pearson VII Kernel Function (PUK) and a regularization parameter of 10, SVM achieved a training accuracy of 88.87% and a test accuracy of 87.76%. This superior performance was attributed to its robustness in handling high-dimensional data and class imbalances. The ANN model showed promising results, with a training accuracy of 87.98%, but was computationally expensive and slightly less accurate on the test dataset compared to SVM, rendering it less practical for real-world applications[10].

* + 1. **RESEARCH GAP**

While the study successfully demonstrated the effectiveness of machine learning in predicting employee attrition, it also pointed to several limitations and potential avenues for future research. The use of a synthetic dataset limits the generalizability of the findings to real-world scenarios, as synthetic data may not capture the nuanced patterns present in actual employee behaviour. Furthermore, the study did not delve deeply into understanding the causal relationships between the features and employee attrition. Analysing these relationships could yield actionable insights for organizational policies and interventions. Although the authors conducted parameter tuning, the use of more advanced hyperparameter optimization techniques, such as grid search or Bayesian optimization, could further enhance model performance. Additionally, the study focused on only three machine learning algorithms. Incorporating ensemble methods like Random Forest or Gradient Boosting could provide a more comprehensive evaluation of predictive techniques. Another notable limitation was the lack of consideration for temporal trends, which are often pivotal in understanding and predicting employee behaviour. Incorporating time-series analysis could significantly improve the models’ predictive power. By addressing these research gaps, future studies can develop more robust, scalable, and actionable solutions to help organizations mitigate the adverse impacts of employee attrition[10].

* 1. **EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING MODELS: A REVIEW PAPER**

The paper titled "Employee Attrition Prediction Using Machine Learning Models: AReview Paper" by Haya Alqahtani, Hana Almagrabi, and Amal Alharbi provides a comprehensive review of research in the field of employee attrition prediction using machine learning (ML) techniques. Employee attrition, a significant concern for organizations, refers to the reduction in workforce due to voluntary or involuntary departures. Attrition negatively impacts organizational productivity, incurs high replacement costs, and disrupts overall operational continuity. This review paper explores the various machine learning methods employed in predicting employee attrition, identifying effective approaches and highlighting recent trends in this domain. The authors emphasize the growing importance of predictive analytics in human resources to forecast attrition rates and enable informed decision-making. The paper surveys literature from 2019 to February 2024, categorizing studies based on datasets, models, and feature selection methods. Notably, the IBM HR Analytics Employee Attrition dataset is identified as the most frequently used dataset. This dataset provides valuable employee-related attributes but is often criticized for being synthetic and highly imbalanced. To address these issues, researchers have applied oversampling techniques like SMOTE to enhance model performance. A significant portion of the paper is devoted to comparing the performance of various machine learning algorithms. Random Forest (RF) and XGBoost (XGB) ensemble methods frequently emerge as top performers across multiple studies due to their robustness, high accuracy, and ability to handle imbalanced datasets. Deep learning techniques, including Artificial Neural Networks (ANN) and Deep Neural Networks (DNN), have also shown promise, particularly in scenarios requiring complex pattern recognition. However, their computational demands and susceptibility to overfitting are notable limitations. The review highlights the critical role of feature selection in improving model performance. Many studies employ algorithms such as Recursive Feature Elimination (RFE), mutual information methods, and Chi-square tests to identify key predictors of attrition. Commonly cited influential features include monthly income, years of experience, job role, and environmental satisfaction. These features enable organizations to develop targeted retention strategies by addressing specific employee concerns. The authors also explore ensemble learning techniques, which combine multiple models to achieve superior accuracy and reliability. Models such as AdaBoost, Gradient Boosting, and voting classifiers demonstrate the potential for outperforming single-model approaches by leveraging the strengths of multiple algorithms. The paper concludes by discussing the application of these methods in real-world scenarios. Predictive analytics enables HR managers to anticipate employee behaviour, design interventions to improve satisfaction, and reduce turnover rates. This proactive approach contributes to resource optimization and enhances overall organizational performance. However, the authors acknowledge the limitations of current research and the need for further innovation[11].

* + 1. **RESEARCH GAP**

The final section of the paper identifies several research gaps that warrant attention. First, the reliance on synthetic datasets like the IBM dataset limits the generalizability of findings, as these datasets fail to capture the nuances of real-world employee behaviour. Future studies should incorporate real, domain-specific datasets to provide more accurate predictions. Second, while many studies focus on identifying critical features, the causal relationships between these features and attrition remain underexplored. Understanding these relationships can provide actionable insights for retention strategies. Additionally, most existing models are static and fail to account for temporal trends that influence attrition rates. Incorporating time-series analysis could enhance the models' predictive power. Finally, the paper highlights the need for explainable AI techniques, as HR professionals require interpretable models to justify decisions and build trust in predictive systems. Addressing these gaps will lead to more robust, practical, and actionable solutions for tackling employee attrition[11].

* 1. **EMPLOYEE ATTRITION PREDICTION USING STACKING AND ITS EVALUATION**

The research paper titled \*"Employee Attrition Prediction Using Stacking and Its Evaluation"\* explores the application of machine learning (ML) techniques to predict employee attrition and identify effective strategies to retain valuable employees. The paper highlights that attrition, often referred to as the wastage or turnover rate, poses a significant challenge for organizations by increasing recruitment and training costs and disrupting workplace efficiency. To mitigate these impacts, the study focuses on leveraging machine learning to analyse historical and current employee data, thereby providing actionable insights to reduce attrition rates. The authors used the IBM HR Analytics Employee Attrition dataset, which contains approximately 14,000 rows and 10 features. This dataset provides a comprehensive view of various employee attributes, such as job satisfaction, project count, evaluation scores, and time spent at the company. The study emphasizes the importance of data preprocessing, including cleaning and feature selection, to ensure the data is structured and relevant for predictive modelling. Following preprocessing, the dataset was divided into training and testing sets, comprising 75% and 25% of the data, respectively. The authors then applied several classification algorithms, including Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), to analyse the data. Additionally, they implemented an ensemble stacking model to improve predictive accuracy by combining the strengths of these base algorithms. The study finds that the stacked hybrid model, which uses a Decision Tree as the metalearner, outperformed individual machine learning models in predicting employee attrition. By combining the predictions of base models, the stacking method achieved an accuracy of 98%, making it the most effective approach among the tested methods. The analysis also identified key features influencing employee attrition, such as job satisfaction, years at the company, and evaluation scores. Employees with low job satisfaction or extreme workloads—either underworked with fewer than 150 hours per month or overworked with more than 250 hours—were found to be at higher risk of leaving. Similarly, mid-level employees with 4 to 5 years of tenure were observed to have high turnover rates, likely due to unmet career progression expectations. These findings underscore the critical role of human resource management in addressing specific employee concerns and implementing retention strategies based on predictive insights[12].

* + 1. **RESEARCH GAP**

Despite its contributions, the study acknowledges several limitations and research gaps. The reliance on the IBM dataset, a widely used synthetic dataset, limits the study's applicability to real-world scenarios as it does not fully capture the complexity of organizational contexts. Future research should prioritize using diverse, real-world datasets that reflect industry-specific challenges. Moreover, while the study highlights the effectiveness of stacking, it does not explore the comparative performance of alternative ensemble methods like boosting or bagging in detail. Another limitation is the lack of focus on temporal trends, which could provide a deeper understanding of how employee behavior and attrition risks evolve over time. Additionally, the study primarily evaluates model performance based on accuracy, without considering metrics such as interpretability or fairness, which are crucial for HR applications. Addressing these gaps could lead to more robust and practical solutions for managing employee attrition[12].

* 1. **PREDICTION OF EMPLOYEE ATTRITION USING MACHINE LEARNING**

The paper titled "Prediction of Employee Attrition Using Machine Learning" presents a comprehensive study aimed at addressing one of the critical challenges faced by organizations: employee attrition. Employee attrition refers to the phenomenon where employees leave an organization due to resignation, retirement, or other reasons, leaving a vacuum that is challenging to fill. The paper emphasizes that understanding the factors leading to attrition and predicting it accurately is essential for organizations to implement effective retention strategies, reduce hiring costs, and improve organizational stability. The study explores the application of machine learning techniques to predict employee attrition using a dataset containing employee details, such as age, salary, department, job satisfaction, and work-life balance. The dataset used in the research is a fictional dataset developed by IBM data scientists, and the authors focus on preparing the data through preprocessing techniques like cleaning, sampling, and formatting. Data preprocessing ensures that the dataset is in a usable format, devoid of missing or irrelevant values, thus facilitating the effective application of machine learning models. The research highlights the application of two machine learning algorithms: Random Forest and Support Vector Machine (SVM). Random Forest, an ensemble learning method, builds multiple decision trees and aggregates their results to produce more accurate and robust predictions. It is lauded for its stability, reduced bias, and ability to handle both categorical and numerical features. SVM, on the other hand, classifies data by finding the optimal hyperplane that separates different classes. The study compares the performance of these algorithms, with Random Forest achieving an accuracy of 77% and SVM attaining 70%. The results indicate that factors such as job dissatisfaction, low salaries, limited promotions, and high workload are significant predictors of employee attrition. The study concludes that organizations must address these factors by fostering a positive work environment, offering competitive salaries, and providing growth opportunities to retain talent. The authors stress that understanding the root causes of attrition can help organizations implement targeted interventions, enhance productivity, and achieve longterm business goals[13].

* + 1. **RESEARCH GAP**

While the study provides valuable insights into predicting employee attrition and highlights key contributing factors, it is limited by its reliance on a fictional dataset rather than real-world data. This restricts the generalizability and applicability of the findings to actual organizational scenarios. Additionally, the study primarily focuses on Random Forest and SVM algorithms, overlooking other advanced machine learning techniques such as deep learning or gradient boosting that could potentially yield better accuracy and more nuanced insights. Furthermore, the research does not account for external factors like industry trends, economic conditions, or social influences, which could significantly impact attrition rates. These gaps provide opportunities for future research to incorporate real-world data, explore advanced algorithms, and integrate external contextual factors for a more comprehensive understanding of employee attrition dynamics[13].

1. **PROBLEM FORMULATION**

**1. High Employee Turnover:**

* Employee attrition leads to increased costs, reduced productivity, and disruption in team dynamics.
* High turnover negatively impacts organizational performance and efficiency.

**2. Complex Causes:**

* Factors like employee dissatisfaction, lack of career growth, compensation issues, and work-life balance contribute significantly to attrition.
* These issues require accurate identification and effective solutions.

**3. Limitations of Traditional Methods:**

* Manual analysis and subjective assessments are insufficient to address the complexities of modern workplaces.
* Predicting and preventing attrition through traditional means is time-consuming and unreliable.

**4. Emergence of Machine Learning:**

* Machine learning offers data-driven predictions to identify employees at risk of attrition and provides actionable insights.
* Despite its potential, challenges like inaccessible datasets and lack of tailored models hinder its widespread adoption.

**5. Need for a Robust Solution:**

* Developing accurate predictive systems can help organizations proactively address attrition and retain talent.
* This can ensure stability, improve productivity, and provide a competitive edge in the industry.

1. **DATASET REVIEW**

For this project on predicting employee attrition, we have gathered a dataset comprising 1,470 employees. This dataset includes essential attributes related to employee satisfaction, income, seniority, and various demographic details. Each attribute provides valuable insights for understanding and predicting employee behaviour, specifically regarding their likelihood of leaving the organization. Below is an overview of the key features in the dataset and their descriptions:

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| **Age** | Numerical value representing the age of the employee. |
| **Attrition** | Binary indicator of employee leaving the company (0 = No, 1 = Yes). |
| **Business Travel** | Categorical value indicating travel frequency (1 = No Travel, 2 = Travel Frequently, 3 = Travel Rarely). |
| **Daily Rate** | Numerical value representing the daily salary level of the employee. |
| **Department** | Categorical value representing the department (1 = HR, 2 = R&D, 3 = Sales). |
| **Distance From Home** | Numerical value indicating the distance from home to the workplace. |
| **Education** | Numerical value representing the education level. |
| **Education Field** | Categorical value indicating the field of education (1 = HR, 2 = Life Sciences, 3 = Marketing, 4 = Medical Sciences, 5 = Others, 6 = Technical). |
| **Employee Count** | Numerical count representing the total number of employees (typically constant in the dataset). |
| **Employee Number** | Unique identifier for each employee (Employee ID). |
| **Environment Satisfaction** | Numerical value representing the employee's satisfaction with their work environment. |
| **Gender** | Categorical value for gender (1 = Female, 2 = Male). |
| **Hourly Rate** | Numerical value indicating the employee's hourly wage. |
| **Job Involvement** | Numerical value representing the level of job involvement. |
| **Job Level** | Numerical level indicating the seniority or level of the job position. |
| **Job Role** | Categorical value representing the job role (e.g., 1 = HC Rep, 2 = HR, 3 = Lab Technician, etc.). |
| **Job Satisfaction** | Numerical value indicating the level of job satisfaction. |
| **Marital Status** | Categorical value indicating marital status (1 = Divorced, 2 = Married, 3 = Single). |
| **Monthly Income** | Numerical value representing the monthly salary. |
| **Monthly Rate** | Numerical rate for the employee on a monthly basis. |
| **NumCompanies Worked** | Numerical count of companies the employee has previously worked for. |
| **Over 18** | Binary indicator if the employee is over 18 (1 = Yes, 2 = No). |
| **Overtime** | Binary indicator for overtime status (1 = No, 2 = Yes). |
| **Percent Salary Hike** | Numerical value indicating the percentage increase in salary over time. |
| **Performance Rating** | Numerical rating based on the employee’s performance. |
| **Relation Satisfaction** | Numerical value indicating satisfaction with workplace relationships. |
| **Standard Hours** | Standard working hours, typically a constant. |
| **Stock Option Level** | Numerical level of stock options available to the employee. |
| **Total Working Years** | Total years the employee has worked in their career. |
| **Training Times Last Year** | Numerical value representing hours spent in training over the past year. |
| **Work-Life Balance** | Numerical rating of the balance between work and personal life. |
| **Years at Company** | Total number of years the employee has been with the company. |
| **Years in Current Role** | Total years spent in the current role. |
| **Years Since Last Promotion** | Years since the employee's last promotion. |
| **Years with Current Manager** | Total years spent working under the current manager. |

Table 4.1 Dataset Matrix

This dataset provides a comprehensive view of various factors affecting employee satisfaction, performance, and decisions to stay or leave the organization. By analyzing these attributes through machine learning models, we can identify significant patterns and factors that contribute to employee attrition. This information is crucial for HR departments to develop effective retention strategies and improve employee satisfaction.

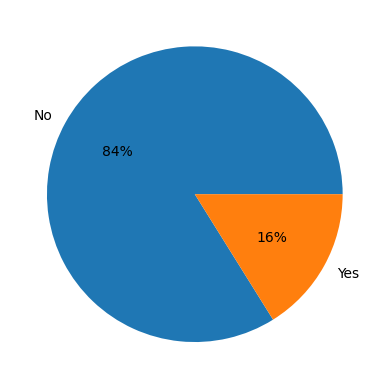


Fig 4.2 Employee Layoff Distribution

1. **DATA PREPROCESSING**

* **Data Cleaning**: Removed rows with missing categorical data; filled missing numerical values with zeroes. Assigned zero values for fields like promotions to ensure completeness.
* **Categorical Encoding**: Applied one-hot encoding to convert categorical features (e.g., *Business Travel*, *Department*) into binary fields for compatibility with the model.
* **Class Balancing**: Used **SMOTE** to address class imbalance in the target variable, generating synthetic samples for underrepresented classes.
* **Feature Scaling**: Standardized features like Daily Rate, Hourly Rate, and Distance from Home for consistent feature contribution across models.
* **Feature Selection**: Selected relevant features based on correlation analysis to reduce noise and improve model efficiency[2].



Fig 5.1 Data Preprocessing Workflow

* 1. **MODEL TRAINING**
* **Train-Test Split**: Split dataset in an 80:20 ratio (80% for training, 20% for testing).
* **Regularization and Hyperparameter Tuning**: Applied regularization techniques and optimized penalty parameters to prevent overfitting and improve model performance.
* **Training Process**: Trained multiple models using tuned hyperparameters, identifying optimal configurations to enhance predictive accuracy[2].
  1. **EVALUATION**
* **Performance Metrics**: Evaluated models based on accuracy, precision, recall, and F1-score using the testing dataset.
* **Confusion Matrix**: Generated confusion matrices for each model to analyze true/false positives and negatives, enabling a comprehensive model comparison[2].
  1. **HYPERPARAMETER TUNING**

Hyperparameter tuning optimizes model parameters to achieve the best possible performance. For this project, we used **Grid Search** and **Random Search** methods to fine-tune the parameters of each model.

* **Logistic Regression**: Tuned the regularization strength (C) to manage overfitting and optimized the penalty type (L1 or L2) for the best trade-off between accuracy and complexity.
* **AdaBoost**: Adjusted the n\_estimators (number of boosting rounds) and learning\_rate to balance the contributions of each weak learner, refining the ensemble’s predictive accuracy.
* **Random Forest**: Tuned the n\_estimators (number of trees), max\_depth (depth of trees), and min\_samples\_split to optimize the diversity and depth of each tree without causing overfitting.
* **Support Vector Machine (SVM)**: Focused on optimizing the C parameter for regularization, kernel type (e.g., linear or RBF), and gamma for the RBF kernel to enhance boundary separation and accuracy.
* **Decision Tree**: Tuned max\_depth, min\_samples\_split, and min\_samples\_leaf to prevent overfitting by controlling tree complexity and depth.
* **K-Nearest Neighbors (KNN)**: Optimized the n\_neighbors parameter (number of neighbors) and distance metric (e.g., Euclidean or Manhattan) to improve local region accuracy.
* **Naive Bayes**: Adjusted the var\_smoothing parameter in Gaussian Naive Bayes to control variance in numerical features, enhancing stability on small datasets.

Each model’s hyperparameters were fine-tuned using **cross-validation** to validate performance across different data subsets, selecting the optimal parameters based on validation accuracy and overall stability. This process led to an improvement in model performance, achieving reliable and robust predictions on employee attrition.

* 1. **PERFORMANCE EVALUATION**

The performance assessment of ML model are done by using some performance metrics. 70% of the data set is used for training purpose and remaining 20% is used as test purpose. We have used weighted average of precision, recall, F1-score and accuracy as performance metrics. All the metrics are measured on following four parameters: True positive (Tp): It is the result of the model successfully predicting the positive class. True negative (T): It is the result of the model successfully predicting the negative class. False Positive (Fp): When the class is negative but the model predicts it as positive. False negative (F): When the class is positive but the model predicts it as negative.

Precision: It determines how much of the affirmative identification is correct. It is represented as:

(5.1)

Accuracy: Accuracy is the percentage of correct predictions made our model. It is measured as:

(5.2)

Recall: Recall is a measure of how well our model detects true positives. It is measured as:

(5.3)

F1 score: The F1 score is the harmonic mean of precision and recall, with 1.0 being the highest and 0.0 being the poorest.

(5.4)

1. **METHODOLOGY**

This project utilizes various machine learning models to predict employee attrition effectively. The methodology consists of data preprocessing, model selection, training, hyperparameter tuning,and evaluation, using a diverse set of algorithms for comparative analysis.

* 1. **LOGISTIC REGRESSION**

Logistic Regression is a fundamental model in classification tasks, particularly suited to binary outcomes like employee attrition (whether an employee leaves or not). Unlike simple linear regression, which predicts a continuous outcome, logistic regression uses a sigmoid function to map any real-valued number into a probability between 0 and 1. This probability indicates the likelihood of an employee leaving, based on various input features (e.g., job satisfaction, years at the company,income level). Logistic regression is interpretable and efficient, making it a solid baseline model forunder standing key drivers of attrition.

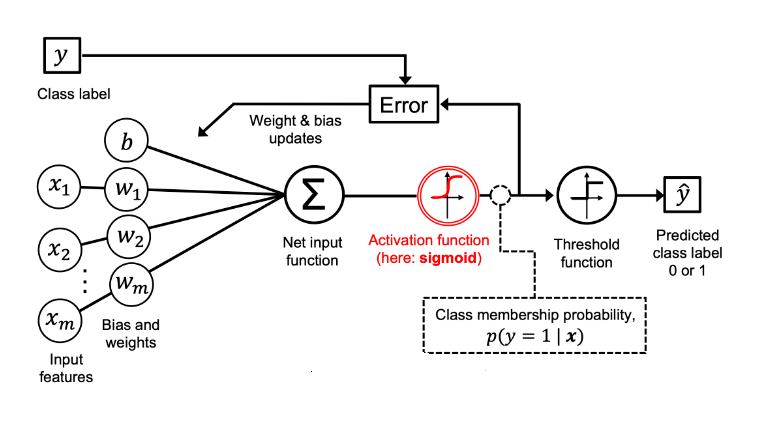


Fig 6.1 Logistic Regression Model

* 1. **ADABOOST(ADAPTIVE BOOSTING)**

AdaBoost is an ensemble technique that combines multiple “weak learners” to create a stronger predictive model. Each weak learner is a simple model that may struggle to predict accurately on its own, but AdaBoost enhances performance by training each learner sequentially. It gives more weight to misclassified data points after each iteration, directing the model’s attention toward difficult cases. This “boosting” method can be highly effective for employee attrition, as it accounts for nuanced patterns that might be missed by a single model. AdaBoost is particularly useful for handling imbalanced datasets, as it can adjust focus dynamically to handle rare cases.



Fig 6.2 AdaBoost Working

* 1. **RANDOM FOREST**

Random Forest is another ensemble method, but it takes a different approach by building multiple decision trees, each trained on different subsets of data. Unlike AdaBoost, which adjusts weights, Random Forest creates a variety of trees and aggregates their predictions, reducing the risk of overfitting and capturing complex patterns. Random Forest is highly resilient to noise and outliers, making it valuable for real-world data like employee records, where there may be inconsistencies or missing information. Its flexibility allows it to capture interactions between features, such as how income level and job satisfaction together affect the likelihood of attrition.

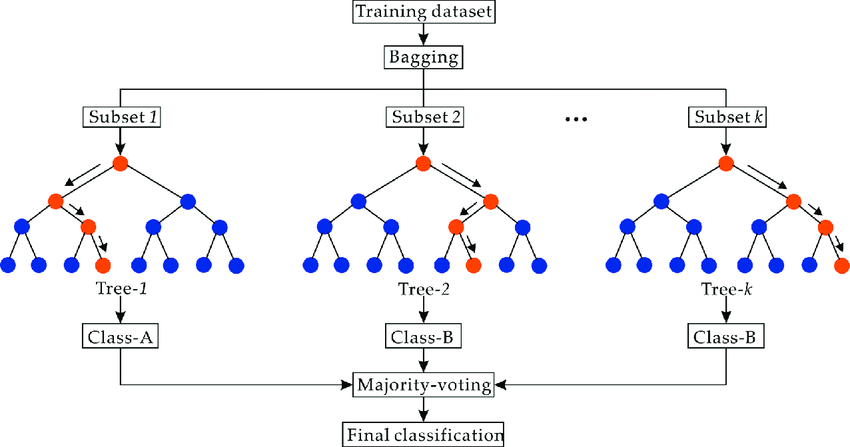


Fig 6.3 Random ForestWorking Flowchart

* 1. **SUPPORT VECTOR MACHINES (SVM)**

SVM is a powerful classification model that seeks to find the best boundary, or hyperplane, to separate different classes. It works by maximizing the margin between classes, making it effective at handling high-dimensional datasets. SVM can use different “kernels” to handle non-linear data patterns, which may arise from complex employee behaviors and organizational influences on attrition. In the context of predicting employee attrition, SVM helps in situations where there is a clear but non-linear boundary between employees who are likely to stay and those who may leave.

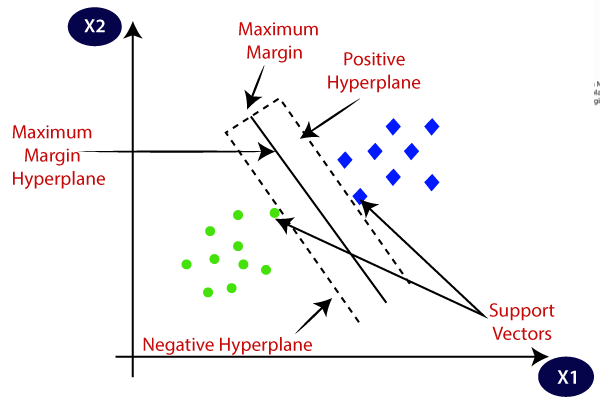


Fig 6.4 Support Vector MachinesRepresentation

* 1. **DECISION TREES**

Decision Trees are intuitive models that split data into branches based on feature values, making decisions at each “node.” They allow easy visualization of the paths leading to different outcomes, which is helpful in understanding the specific factors influencing attrition. For instance, a decision tree might first split data by employee tenure, then by job satisfaction level, and finally by income level to predict attrition. Although decision trees can sometimes overfit, they are simple to interpret and provide transparency about which features play key roles in employee decisions to stay or leave.

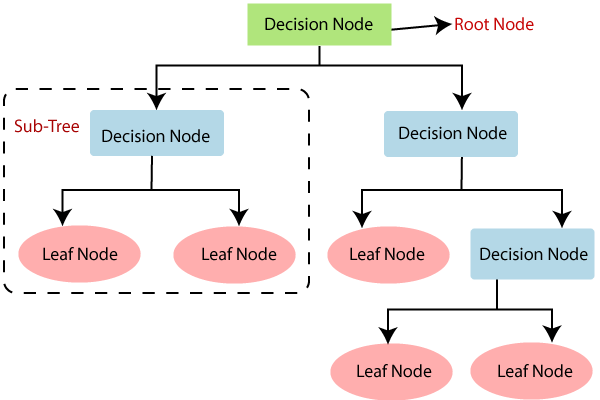


Fig 6.5 Decision Trees Flowchart

* 1. **K-NEAREST NEIGHBOR (KNN)**

KNN is a “lazy” learning algorithm that makes predictions based on the proximity of a data point to its “k” nearest neighbors. For employee attrition prediction, KNN could group employees by shared attributes like department, age, or satisfaction level, using these “neighbors” to determine the likelihood of attrition. KNN’s performance depends on the distance metric and number of neighbors chosen, and it can capture local patterns effectively, particularly for similar types of employees. KNN is sensitive to feature scaling, so preprocessing (such as standardization) is essential for optimal performance.

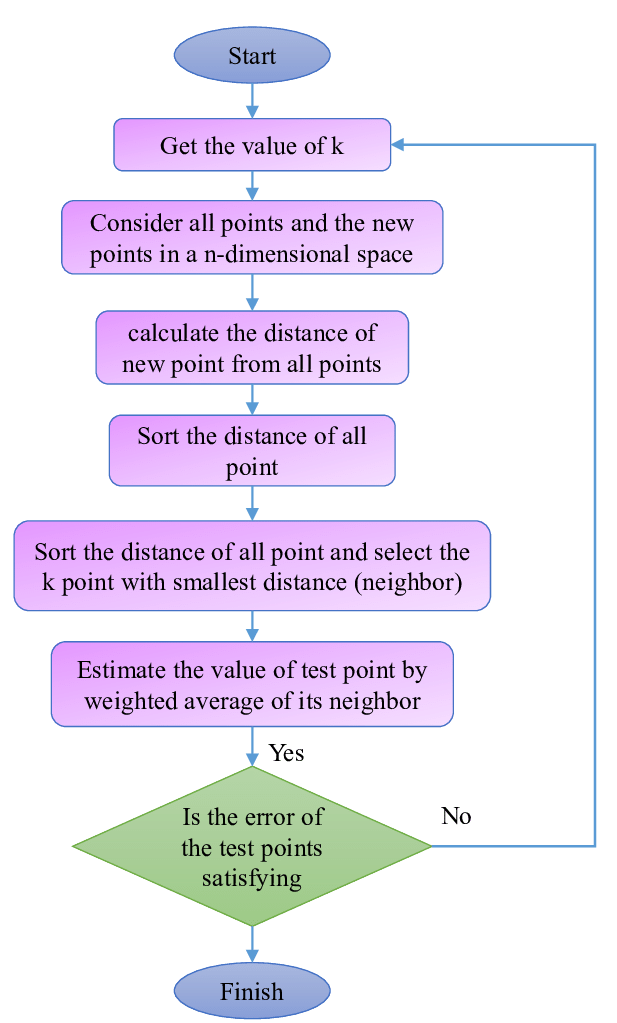


Fig 6.6 K-Nearest Neighbor Flowchart

* 1. **NAÏVE BAYES**

Naive Bayes is a probabilistic classifier grounded in Bayes’ theorem. It assumes independence between features, meaning it treats each factor (like income level, job satisfaction, and seniority) as contributing individually to the outcome. While this “naive” assumption does not always hold in complex data like employee attrition, Naive Bayes is effective when there are strong individual correlations between features and the target variable. It is particularly useful for high-dimensional data with numerous categorical variables. Naive Bayes is computationally efficient and can serve as a good benchmark model.

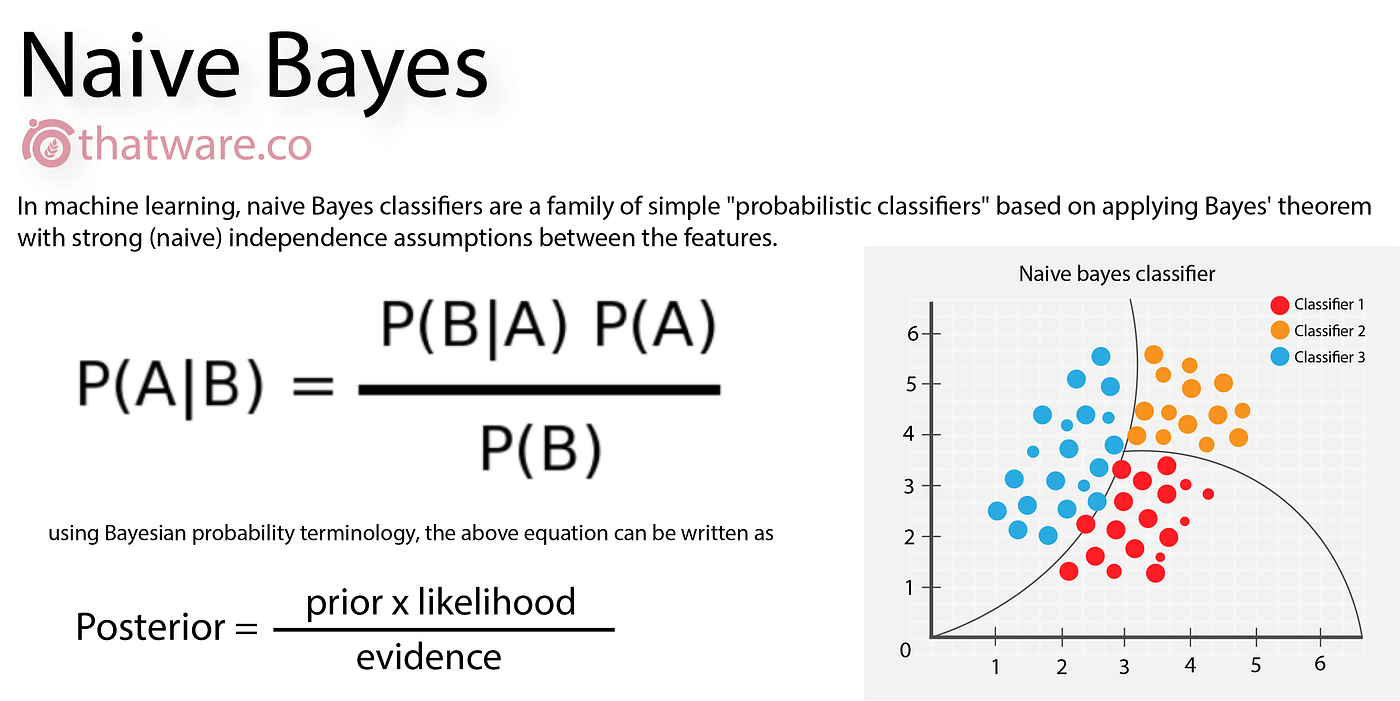


Fig 6.7 Naïve Bayes Working

* 1. **SYSTEM ARCHITECTURE**

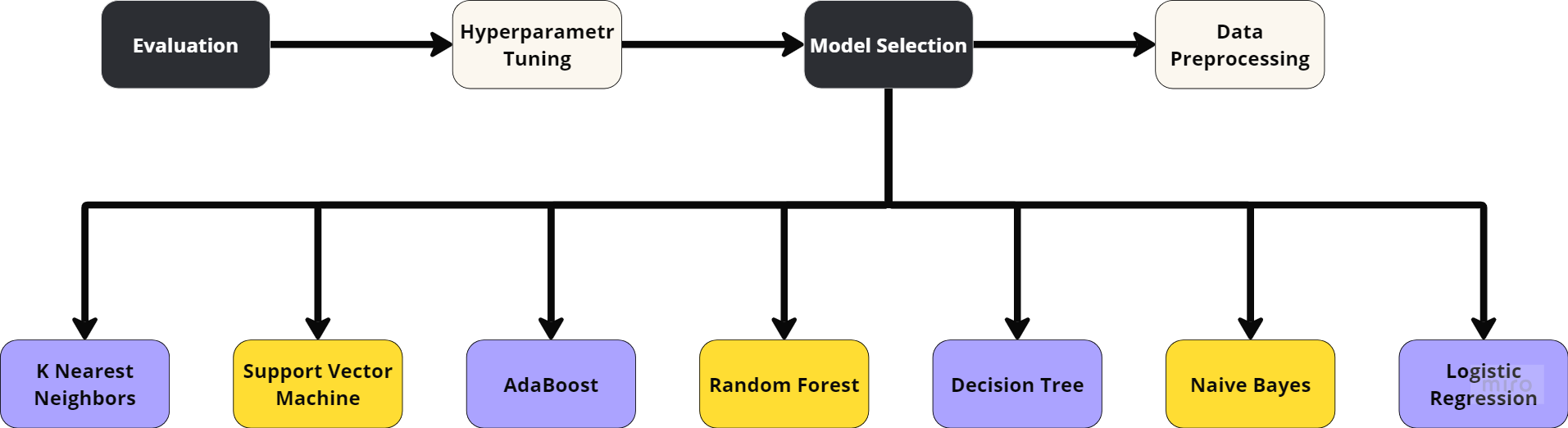
****

Fig 6.8 Flowchart showcasing project Methodology

1. **RESULTS & DISCUSSION**

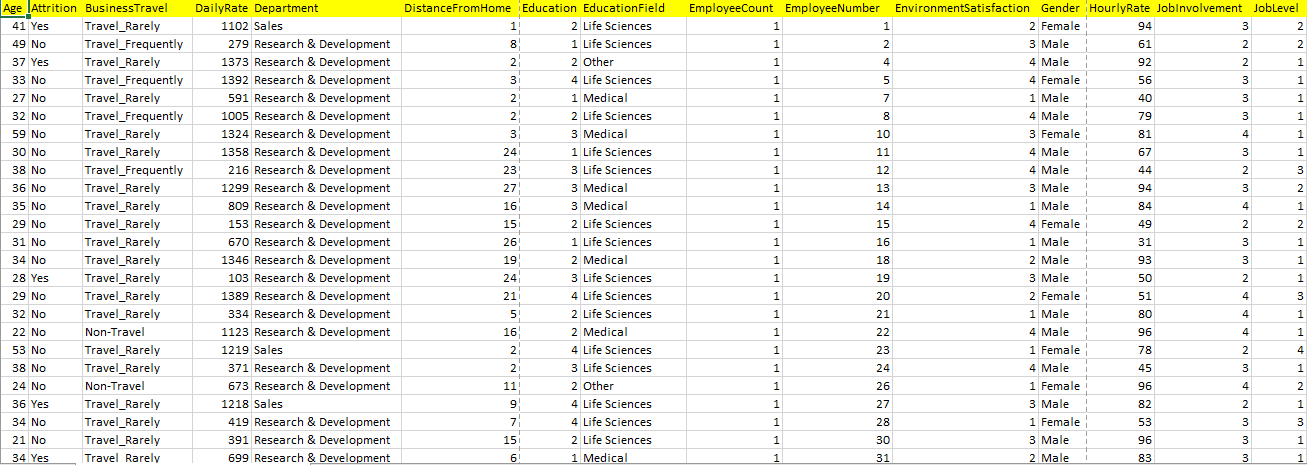


Fig 7.1 Sample Employee Details

The Best performance was obtained in Random Forest Model with Under sampling with an accuracy of 74.6%, the precision of 74.5%, recall of 75.1% and F1 Score of 74.7% shown in Fig 7.2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** |
| Logistic Regression | 0.734 | 0.739 | 0.734 | 0.734 |
| Naïve Bayes | 0.653 | 0.624 | **0.776** | 0.691 |
| Decision Trees | 0.639 | 0.650 | 0.615 | 0.629 |
| **Random Forest** | **0.746** | **0.745** | **0.751** | **0.747** |
| AdaBoost | 0.736 | 0.739 | 0.734 | 0.735 |
| SVM | 0.729 | 0.734 | 0.730 | 0.729 |
| KNN | 0.736 | **0.792** | 0.641 | 0.707 |

Fig 7.2 Undersampling Model

The Best performance was obtained in Random Forest Model with Oversampling with an accuracy of 99.1%, the precision of 98.5%, recall of 99.7% and F1 Score of 99.1% shown in Fig 7.3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** |
| Logistic Regression | 0.753 | 0.755 | 0.751 | 0.752 |
| Naïve Bayes | 0.680 | 0.649 | 0.782 | 0.709 |
| Decision Trees | 0.872 | 0.826 | 0.944 | 0.880 |
| **Random Forest** | **0.991** | **0.985** | **0.997** | **0.991** |
| AdaBoost | 0.987 | 0.977 | **0.997** | 0.987 |
| SVM | 0.604 | 0.590 | 0.680 | 0.632 |
| KNN | 0.824 | 0.741 | **0.997** | 0.850 |

Fig 7.3 Oversampling Model

The Best performance was obtained in Random Forest Model with PCA and Oversampling with an accuracy of 99.2%, the precision of 98.6%, recall of 99.8% and F1 Score of 99.2% shown in Fig 7.4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** |
| Logistic Regression | 0.758 | 0.751 | 0.771 | 0.761 |
| Naïve Bayes | 0.680 | 0.649 | 0.782 | 0.709 |
| Decision Trees | 0.841 | 0.796 | 0.917 | 0.852 |
| **Random Forest** | **0.992** | **0.986** | **0.998** | **0.992** |
| AdaBoost | **0.989** | 0.982 | **0.998** | 0.990 |
| SVM | 0.951 | 0.912 | **0.998** | 0.953 |
| KNN | 0.894 | 0.826 | **0.998** | 0.904 |

Fig. 7.4 Oversampling with PCA Model

We observe that the most important features were MonthlyIncome followed by OverTime and Age, while the least important features were Performance Rating, Gender and BusinessTravel shown in Fig 7.5.

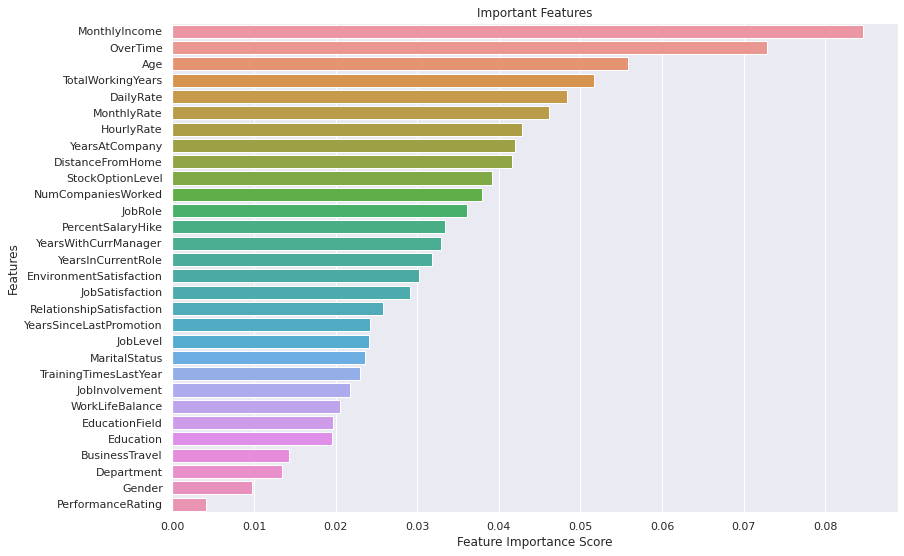


Fig 7.5 Feature Importance w.r.t Random Forest with Oversampling

1. **CONCLUSION & FUTURE SCOPE**
   1. **CONCLUSION**

We applied multiple supervised classification models, including Logistic Regression, Naive Bayes, Decision Trees, Random Forest, AdaBoost, Support Vector Machines and K-Nearest Neighbors, and summarized their results in this project. The best results were achieved with the Random Forest model combined with PCA and oversampling, yielding an accuracy of 99.2%, a precision of 98.6%, a recall of 99.8%, and an F1 score of 99.2%. Other model, such as SVM , also demonstrated high performance, consistently achieving accuracy and F1 scores above 90%. PCA with oversampling improved most models, except for Logistic Regression and Naive Bayes, while tree-based models had the highest overall metrics. According to the EDA, factors like Monthly Income, Age, Overtime, and Total Working Years significantly influenced attrition, while Gender did not have an impact.

* 1. **FUTURE SCOPE**
     1. **INTEGRATION OF ADVANCED MACHINE LEARNING TECHNIQUES**
* Deep Learning Models: Employ neural networks, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, to analyze sequential and time-series data related to employee behavior.
* Explainable AI (XAI): Incorporate tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to enhance the interpretability of predictions, enabling actionable insights for HR teams.
  + 1. **REAL-TIME ATTRITION MONITORING**
* Dashboard Development: Build a real-time employee attrition dashboard that updates continuously based on inputs such as employee feedback, performance ratings, and training participation.
* Automated Alerts: Introduce alerts for HR managers to flag high-risk employees, enabling timely interventions.
  + 1. **EXPANDED DATA COLLECTION**
* Comprehensive Metrics: Integrate datasets that include psychological, emotional, and behavioral metrics derived from employee surveys, social media activity (adhering to ethical boundaries), and engagement levels.
* IoT and Wearable Technology: Leverage data from IoT devices (e.g., stress levels, activity patterns) to gain insights into employee well-being.
  + 1. **IMPROVED HANDLING OF IMBALANCED DATA**
* Advanced Oversampling: Utilize methods like SMOTE-ENN or Generative Adversarial Networks (GANs) to address data imbalance issues effectively.
* Cost-Sensitive Algorithms: Design custom algorithms to mitigate underrepresentation in attrition datasets.
  + 1. **ETHICAL AND FAIR AI**
* Fairness-Aware Algorithms: Ensure that predictions do not discriminate based on attributes like gender or age, implementing fairness-aware techniques.
* Data Privacy Policies: Establish transparent privacy policies to build employee trust and ensure compliance with regulations like GDPR.

1. **REFERENCES**

[1] S. George, K. A. Lakshmi, and K. T. Thomas, "Predicting Employee Attrition Using Machine Learning Algorithms," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 700-705, doi: 10.1109/ICAC3N56670.2022.10074131.

[2] D. Mamtani and B. Malukani, "Predicting Employee Attrition Using Machine Learning," \*European Economic Letters (EEL)\*, vol. 13, no. 4, pp. 937–941, 2023.

[3] K.Konar, S. Das, and S. Das, "Employee attrition prediction for imbalanced data using genetic algorithm-based parameter optimization of XGB Classifier," \*2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)\*, Kolkata, India, 2023, pp. 1-6, doi: 10.1109/ICCECE51049.2023.10085402.

[4] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting Employee Attrition Using Machine Learning Techniques," \*Computers\*, vol. 9, no. 4, p. 86, 2020.

[5] Usha P. M. and N. V. Balaji, "Analyzing Employee Attrition Using Machine Learning," Karpagam Journal of Computer Science, vol. 13, pp. 277-282, Sep.-Oct. 2019.

[6] P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees’ attrition: a machine learning approach," SN Applied Sciences, vol. 2, no. 757, pp. 1-12, Mar. 2020, doi: 10.1007/s42452-020-2519-4

[7] R. S. Kamath, S. S. Jamsandekar, and P. G. Naik, "Machine Learning Approach for Employee Attrition Analysis," International Journal of Trend in Scientific Research and Development (IJTSRD), Conference Issue, pp. 62-67, March 2019.

[8] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," \*Applied Sciences\*, vol. 12, no. 13, p. 6424, Jun. 2022. doi: [10.3390/app12136424]

[9] P. N. Gandhi, A. Jangale, G. S. Mhaske, and A. Kadlag, "Employee attrition prediction using machine learning," \*Journal of Emerging Technologies and Innovative Research (JETIR)\*, vol. 7, no. 9, pp. 1086–1094, Sep. 2020

[10] N. Mansor, N. S. Sani, and M. Aliff, "Machine Learning for Predicting Employee Attrition," \*International Journal of Advanced Computer Science and Applications (IJACSA)\*, vol. 12, no. 11, pp. 435–445, 2021.

[11] H. Alqahtani, H. Almagrabi, and A. Alharbi, "Employee Attrition Prediction Using Machine Learning Models: A Review Paper," \*International Journal of Artificial Intelligence and Applications (IJAIA)\*, vol. 15, no. 2, pp. 1–25, Mar. 2024. DOI: 10.5121/ijaia.2024.1520223.

[12] Rema, Ruchitha V., Mythri K., and Hariprasad V., "Employee Attrition Prediction Using Stacking and Its Evaluation," \*International Research Journal of Engineering and Technology (IRJET)\*, vol. 8, no. 8, pp. 764–771, Aug. 2021.

[13] S. Kaushik, R. Anusha, B. Kavitha, and N. Krishna Priya, "Prediction of Employee Attrition Using Machine

Learning," Journal of Engineering Sciences, vol. 11, no. 4, pp. 25-31, Apr. 2020

**10. APPENDICES**

**10.1 APPENDIX: DATASET SUMMARY**

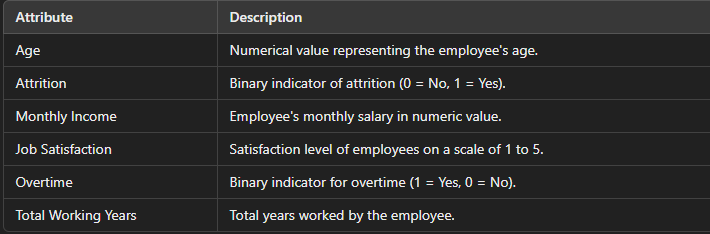


Fig 10.1 Key Features of the Dataset

**10.2 APPENDIX: MACHINE LEARNING MODELS OVERVIEW**  
**MODELS USED:**

• Logistic Regression

• Random Forest

• AdaBoost

• Decision Trees

• Support Vector Machines (SVM)

• K-Nearest Neighbors (KNN)

• Naïve Bayes

**10.2.1 HYPERPARAMETER TUNING DETAILS:**

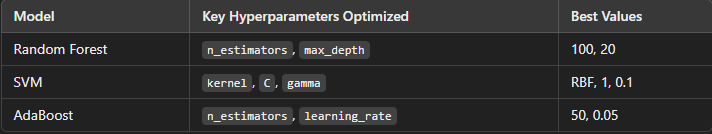


Fig 10.2 Hyperparameter Tuning

**10.3 APPENDIX: PERFORMANCE METRICS FORMULAS**

**10.4 APPENDIX: FIGURES AND VISUALIZATIONS**

* Feature Importance from Random Forest Model
* Comparison of Model Performance Graph showing accuracy, precision, recall, and F1-score for all models.

**10.5 APPENDIX: REFERENCES**

* S. George et al., Predicting Employee Attrition Using Machine Learning Algorithms. (2022)
* D. Mamtani and B. Malukani, Predicting Employee Attrition Using Machine Learning. (2023)
* R. S. Kamath et al., Machine Learning Approach for Employee Attrition Analysis. (2019)