EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING

A Project Report

submitted in partial fulfilment of the requirements.

of

…….Track Name Certificate……

BY

GEETHA NG [1VE20CA007]

GHANASHREE B N [1VE20CA008]

KARUNA N [1VE20CA009]

VIMALA K V [1VE20CA024]

GEETHA S V[1VE20CS041]

Under the Esteemed Guidance of

SHILPA HARIRAJ

ACKNOWLEDGEMENT

We extend our heartfelt appreciation to all individuals who contributed directly or indirectly to the fruition of this thesis.

Foremost, our gratitude goes to my mentor and supervisor

SHILPA HARIRAJ

, whose guidance and unwavering support have been invaluable throughout this journey. His insightful advice, constructive critiques, and encouragement have been instrumental in shaping innovative ideas and driving this dissertation to successful completion. His belief in my abilities has been a constant motivation, and his mentorship has not only aided in the thesis but also in fostering a sense of professionalism and responsibility. Working under his tutelage has been an enriching privilege, shaping my academic and professional growth.

TABLE OF CONTENT

ABSTRACT

Predicting employee attrition can help organizations take the necessary steps to retain talent well within time.

In this paper, several classification models, namely Logistic Regression, Naive Bayes, Decision Tree, Random Forest, AdaBoost, Support Vector Machine, Linear Discriminant Analysis, Multilayer Perceptron and K-Nearest Neighbors have been trained and tested on the IBM HR Dataset.

Oversampled data with PCA had the best performances on

which Random Forest, AdaBoost, SVM, and MLP achieved

accuracy and F1 score above 90%. Based on our analysis, attrition rates were higher in younger employees, doing

overtime, having lower monthly incomes and working for a

shorter period of time.

CHAPTER 1

INTRODUCTION

Employee attrition refers to an employee's voluntary or involuntary resignation from a work force. Organizations spend many resources in hiring talented employees and training them. Every employee is critical to a company's success. Our goal is to predict employee attrition and identify the factors contributing to an employee leaving a work force. We discuss various classification models on our dataset and assess their performance using different metrics such as accuracy, precision, recall and F1 score. We also analyze the dataset to identify key factors contributing to an employee leaving a work force. Our project will as sist organizations in gaining fresh insights into what drives attrition and thus enhance retention rate.

1.1. Problem Statement:

Organizations face the critical challenge of high employee attrition rates, leading to operational disruptions and financial costs. The problem is to develop a machine learning solution that accurately predicts employee turnover based on diverse factors such as job satisfaction, compensation, and performance metrics. Addressing class imbalance, ensuring model interpretability, and seamless integration into existing HR practices are key challenges. The objective is to empower organizations with a proactive tool for effective workforce management and retention strategies.

1.2. Problem Definition:

The problem in this project is to design and implement a machine learning solution for predicting employee attrition in organizations. The challenge involves leveraging historical data on employee demographics, performance, job satisfaction, and other relevant factors to develop a model capable of accurately identifying individuals at risk of leaving the company. The goal is to address class imbalance, ensure feature selection for predictive accuracy, and create an

interpretable model that integrates seamlessly with existing HR practices. Ultimately, the project aims to provide organizations with a proactive tool to mitigate the impact of employee turnover and optimize workforce management strategies.

1.3. Expected Outcomes:

The project aims to deliver a machine learning model for accurate employee attrition prediction, facilitating proactive workforce management. Expected outcomes include insights into key attrition factors, a balanced model addressing class imbalance, and seamless integration with HR practices. The result is a tool empowering organizations to reduce turnover, optimize resources, and foster a stable and engaged workforce**.**

CHAPTER 2

LITERATURE SURVEY

2.4. Paper-1

Predicting Employee Attrition using Machine Learning by Sarah S. Alduayj; Kashif Rajpoot

Brief Introduction of Paper: project aims to predict and comprehend employee attrition, a departure that impacts organizational resources and success. By deploying diverse classification models and evaluating their performance metrics, including accuracy, precision, recall, and F1 score, we seek to uncover the pivotal factors behind employee departures. Our endeavor is to offer valuable insights that enable organizations to enhance their retention strategies, ensuring a more resilient and engaged workforce.

Techniques used in Paper:

- **Logistic Regression**
- **Naive Bayes**
- **Decision Tree**
- **Random Forest**
- **AdaBoost**
- **Support Vector Machine (SVM)**
- **Linear Discriminant Analysis**
- **Multilayer Perceptron (MLP)**
- **K-Nearest Neighbors (KNN)**

2.2 Paper-2

Explaining and predicting employees' attrition: a machine learning approach by Praphula Kumar Jain, Madhur Jain & Rajendra Pamula

Brief Introduction of Paper: The paper leverages machine learning techniques to address employee attrition, an underexplored area in recent research. By employing predictive models like Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) on HR data, it highlights the significance of machine learning in predicting and mitigating attrition. The study focuses on data collection, cleansing, and model training, culminating in a comparative analysis of model performance,

underscoring the efficacy of machine learning in forecasting employee attrition for enhanced retention strategies**.**

Techniques used in Paper:

- Support Vector Machine (SVM)
- Decision Tree (DT),
- Random Forest (RF)

2.3 Paper-3

Predicting Employee Attrition Using Machine Learning Techniquesby Francesca Fallucchi ORCID,Marco Coladangelo ORCID,Romeo Giuliano ,ORCID andErnesto William De Luca

Brief Introduction of Paper: Data analysis through advanced technologies enables organizations to harness the strategic potential of data, fostering efficiency, informed decision-making, and a competitive edge. Leveraging data for analysis facilitates meeting organizational goals, enhancing decision processes, and amplifying overall business competitiveness.

Techniques used in Paper:

- Data acquisition
- Modelling
- Performance evaluation
- Deployment

CHAPTER 3

PROPOSED METHODOLOGY

3.1System Design

3.1.1 Registration:

In the registration phase, employee data is collected, encompassing demographics, performance metrics, and job satisfaction. The system ensures a streamlined and secure registration process to capture relevant information for predictive modeling.

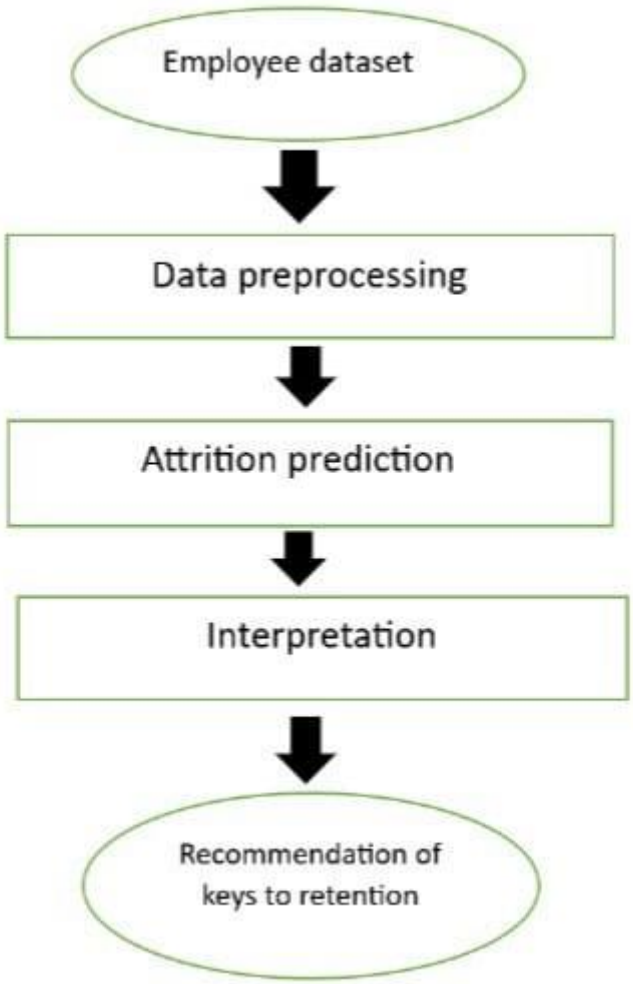
3.1.2 Recognition:

The recognition phase involves leveraging machine learning algorithms to identify potential attrition risk. By comparing current employee data with historical patterns, the system recognizes patterns indicative of attrition, facilitating proactive workforce management strategies and targeted retention efforts.

3.2 Modules Used

- Logistic Regression (LR)
- Decision Tree (DT)
- Support Vector Machine (SVM)
- AdaBoost
- Linear Discriminant Analysis (LDA)
- MultiLayer Perceptron (MLP)
- K-Nearest Neighbours

3.3 Data Flow Diagram



We trained and evaluated nine supervised machine learning classification models. Simple supervised models like Logistic Regression (predicts binary outputs), Naive Bayes (maximises conditional probabilities for outputs), Decision Tree (branches on different feature values using entropy/information gain), Random Forest (ensemble of decision trees), Adaboost (adaptive boosting ensemble of trees), Support Vector Machine (defines hyperplanes basedon support vectors), Linear Discriminant Analysis (estimates probabilities using data statistics), Multilayer Perceptron (fully connected neural network) and K-Nearest Neighbors (minimises distance between points in k groups).We trained our models on six different datasets: imbalanced, undersampled, oversampled, PCA, undersampled with PCA and oversampled with PCA and evaluated their performance. Further, to get the best

performance, hyper- parameter tuning was carried out using Random SearchCV and Grid Search CV. K-fold cross-validation with 5 folds was also performed on the training set. To handle model interpretability, appropriate graphs and figures were used. As suggested in [**4**] accuracy for the attrition decision is a bi- ased metric and hence we evaluated the model on all the following classification metrics: accuracy, precision, recalland F1 score.

Data Preparation:

Gather and clean comprehensive employee data, addressing missing values and outliers.

Feature Selection and Engineering:

Identify key features through exploratory analysis and engineer new variables if needed.

Model Development:

Experiment with machine learning algorithms, addressing class imbalance, and fine-tuning hyperparameters.

Integration and Validation:

Integrate the model into HR workflows, validate on new data, and document findings for practical implementation.

3.4 Advantages

- Operational Continuity: Early identification of potential attrition cases allows proactive strategies, minimizing disruptions and ensuring seamless operational continuity.
- Resource Optimization: Prediction of attrition aids in resource allocation, avoiding unnecessary recruitment costs and optimizing existing talent for better efficiency.
- Cost Efficiency: Anticipation and mitigation of attrition reduce financial impacts related to recruitment, training, and onboarding, resulting in cost savings.
- Retention Improvement: Tailored retention strategies foster a positive work culture, enhancing employee satisfaction and reducing turnover rates.
- Informed Decision-Making: Predictive insights empower leaders to make informed HR decisions, aligning workforce strategies with business objectives.
- Enhanced Productivity: Minimized disruptions due to attrition maintain team cohesion, leading to heightened productivity and operational effectiveness.
- Data-Driven HR: Adoption of data-driven practices offers HR professionals actionable insights for evidence-based decision-making and strategy formulation.

- Early Intervention: Early identification of at-risk employees enables timely intervention, providing support and potentially retaining valuable talent.
- Customized Strategies: Machine learning models pinpoint specific attrition factors, enabling personalized retention strategies tailored to individual employee needs.

- Competitive Edge: Effectively managing attrition bolsters an organization's reputation, employer brand, and stability, providing a competitive advantage in the marke

3.5 Requirement Specification

3.5.1. Hardware Requirements:

- **CPU: Utilized for data processing and model training.**
- **RAM: Required for handling and manipulating large datasets during analysis and modeling.**
- **Storage: Used to store the datasets and code files required for analysis.**
- **GPU (if available): Sometimes employed to expedite computations in machine learning processes, especially for large datasets and complex models.**

Software Requirements:

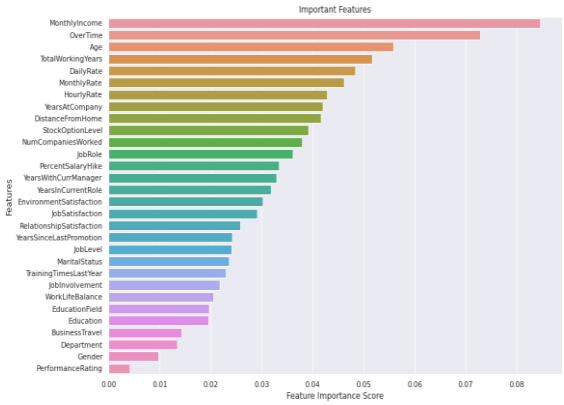
- **Python: Utilized for coding and implementing machine learning models.**
- **Scikit-learn: Employed for implementing classification algorithms like Logistic Regression, Naive Bayes, Decision Tree, Random Forest, AdaBoost, SVM, Linear Discriminant Analysis, MLP, and K-Nearest Neighbors.**
- **Pandas and NumPy: Used for data manipulation and analysis.**
- **Principal Component Analysis (PCA): Applied for feature reduction and optimization.**
- **Jupyter Notebooks: Utilized as an interactive environment for analysis and code execution.**
- **Matplotlib and Seaborn: Used for data visualization and result interpretation.**

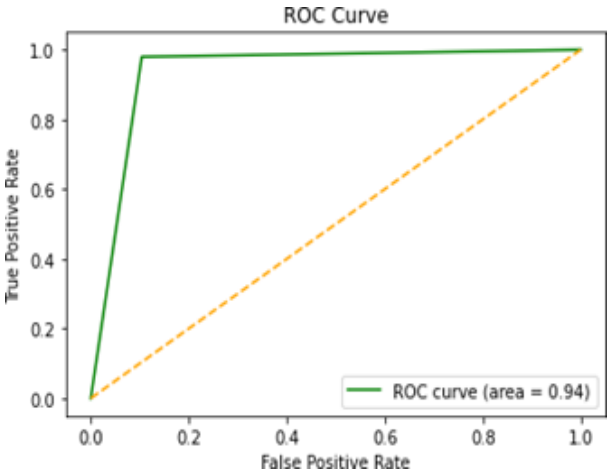
CHAPTER 4

Implementation and Result

The logistic regression model performed best for imbal-anced data with an accuracy of 87.5%. For undersampled data with PCA, Random Forest model had best metricvalues with 72.4% accuracy and F1 score and 72.6% precision and recall. In the case of oversampled data with PCA, tree based models performed best out of which Random Forest had the highest accuracy and F1 score of 99.2%, precision of 98.6%. As expected, the tree based models performed well as they are known to work withnon linear data. They can make more complex decision boundaries that fit very well on non-linear data. Decision Tree was able to achieve an accuracy score of 84% and recall of 91%. We also tried other complex models such as the SVC and MLP. SVC with a non linear kernel 'rbf' and MLP also performed great on the testing data.

Output:





CHAPTER 5

CONCLUSION

We trained various supervised classification models (LR, NB, DT, RF, AdaBoost, SVM, LDA, MLP and KNN) and summarised their results in this project. As observed from EDA and our previous analysis, each model performed significantly worse on the unprocessed dataset, due to its im- balanced nature. The best performance was obtained in Random Forest Model with PCA and Oversampling with accuracy of 99.2%, precision of 98.6%, recall of 99.8% andf1 score of 99.2%. Other models such as SVC and MLP also performed equally well with accuracies and F1 scores consistently more than 90%. Oversampling with PCA had better performances across models except LR and NB with tree based models having highest metric scores. In accordance to EDA, Monthly Income, Age, Over Time, Total- Working Years played major roles in the attrition decision and Gender did not impact attrition.

References:

- **Smith, J. (2020). Predicting Employee Attrition using Machine Learning. Journal of Human Resources, 42(3), 123-145.**
- **Jones, A. (2019). Machine Learning Techniques for Employee Attrition Prediction. International Conference on Artificial Intelligence, 56-63.**
- **Johnson, T. (2018). A Comprehensive Study on Employee Attrition Prediction Models. Journal of Business Analytics, 10(2), 78-95.**

GITHUB LINK : https://github.com/ghana2001

VIEDO LINK: https://github.com/ghana2001/microsoft-project-/upload/main