**Data Visualization Project Writing Part   
  
  
  
Abstract**

The project aims to explore the analysis of an employee performance dataset sourced from Kaggle. The study will use Exploratory Data Analysis (EDA) and visualization techniques to infer useful insights, especially on attrition. It has become pertinent to understand and evaluate the performance of employees in a remote and hybrid work culture due to the thrust of Covid-19. Traditional methods of performance assessment have become either inefficient or ineffective; therefore, newer ideas are required.

The dataset will first be pre-processed with Python to clean and make it ready for analysis. Next, Exploratory Data Analysis (EDA) will be carried out to study the patterns and relationships among data features by making use of visualization libraries like ggplot, seaborn, and matplotlib. It aims to communicate the findings clearly. Important features identified while performing EDA will be incorporated into the predictive model. For the employee attrition prediction model, we’ll be making using of the XGBoost classifier algorithm. We will also be using several model evaluation techniques to check the performance of our model.

The expected outcomes of the project include a comprehensive understanding of the factors explaining employee performance, informative visualizations with clear depictions of the insights, and a predictive model capable of accurately assessing whether or not an employee will leave the company. In this direction, the current study attempts to contribute valuable information to human resource management in the provision of analytic insights and tools to facilitate improved processes of employee appraisal. Eventually, the output will help organizations upgrade their work processes to adapt to a transformed environment while being equally productive.

**Introduction**

The Covid-19 pandemic has turned the work world upside down. In just a year, we've seen a massive shift to remote and hybrid work. This big change has made keeping talented employees a real challenge. The old ways of keeping staff happy and engaged, which relied a lot on face-to-face interactions, just don't cut it anymore in this new landscape.

There's a pressing need for fresh, data-smart ways to understand why people leave their jobs and how to keep them around in this new remote work reality.

Employee turnover has always been a major headache for companies. When good people leave, it hits productivity, team morale, and the company's bottom line. But figuring out why people quit has always been tricky, especially in today's diverse and fast-changing workplaces. With so many people working from home now, it's even harder to spot the warning signs that someone might be thinking of leaving

In this project, we're going to dig deep into employee data to uncover hidden patterns about why people leave their jobs. We'll use some clever data science techniques to look at the information in new ways. It's like being a detective, but instead of solving crimes, we're solving the mystery of employee attrition.

We'll use some smart computer programs to predict who might be at risk of leaving. These tools are great at spotting complex patterns that humans might miss. We'll try out different methods, like Logistic Regression and Random Forest, to see which one is best at predicting who might quit based on various factors we uncover.

To make all this data easy to understand, we'll create lots of eye-catching charts and graphs. These visuals will help bosses and HR teams quickly grasp what's going on and take action to keep their best people.

The main goal here is to really understand what makes people want to stay or leave their jobs, especially when they're working from home. By doing this, we hope to help companies keep their talented employees happy and productive, even in these new and challenging times. This project isn't just about numbers and charts – it's about helping businesses adapt to the new world of work and keep their most valuable asset: their people.

**Previous Work**

Human interest has increased in applying machine learning algorithms to predict employee performance over the past years. A series of studies explores diverse techniques and approaches to understand and improve the accuracy of predictive models.

A major study by Thomas Paul in 2024 also noted the difficulties in applying machine learning models to assess employee performance. This emphasizes the fact that comprehensive and interpretable models need to be developed, which are robust in handling the complexities and nuances embedded in employee performance data. Paul's work underlines the importance of generalization from different datasets and features into organizational contexts.

In a similar study, Siahaan (2021) conducted an in-depth analysis of the performance of contract employees using five different machine learning models: Decision Tree, Naive Bayes, K-Nearest Neighbors, Support Vector Machine, and Random Forest. It shows that the Random Forest algorithm obtained the highest performance, with superior accuracy and robustness compared to other models. Siahaan's study proves the effectiveness of ensemble methods particularlly Random Forest in handling high-dimensional and complex datasets, making it a preferred method in predictive modeling within HR analytics.

In yet another crucial research, Jide Kehinde Adeniyi (2024) compared the prediction of employee performance by three important algorithms: Logistic Regression, Decision Tree, and Artificial Neural Network. The study found that ANN provided better classification accuracy than other models for predicting employee performance. Adeniyi's work reflects the capability of deep learning techniques in identifying complex patterns present in the data, which traditional methods often miss out on.

Laura Gabriela Tanasescu extended this research by using six different machine learning algorithms to optimize and predict employee performance. These was Logistic Regression, Decision Tree, Random Forest, Gradiant Boosting Machine (GBM), XGBost and the Support Vector Machine (SVM). Her study also showed that Random Forest consistently outperformed all the other models, confirming the earlier work of Siahaan (2021). The research emphasized feature selection and engineering to enhance the performance of the models. By selecting and transforming features carefully, Tanasescu managed to upgrade the predictive power of the models.

Together, these studies emphasize the predictive potential of machine learning with regard to employee performance but also point out the challenges and important considerations required for model building. Importantly, the selected algorithm, chosen features, and steps of data pre-processing are critical factors that largely determine the performances of the predictive models.

On this note, the present project is focused on accomplishing a detailed Exploratory Data Analysis to identify significant features that influence employee performance. We will use different visualization techniques to uncover the patterns and relationships present in the dataset. The insights gained from EDA will guide the selection of features for constructing our predictive model.

For our predictive model, we will explore different machine learning algorithms that will including the Logistic Regression, decision Tree, Random Forest and also the possibly deep learning method, for example, artificial neural networks. The models will evaluated by using metrics such as the accuracy, precision, recall, and the F1 score. By comparing all the results, we aim for to identify the best model for to predicting employees performance in this specific context.

Our approach will also incorporate techniques for model interpretability, ensuring that the results are not only accurate but also understandable to stakeholders, which is essential for acceptance within the organization. In this regard, prior work in the field has demonstrated the potential of using machine learning for predicting employee performance but also highlighted several challenges. Building on that prior work and including new techniques of data analysis and modeling, our project aims to provide valuable insights and tools toward improving employee performance evaluation in the dynamic work setup.

METHODOLOGY

This study employed a comprehensive approach to analyze employee attrition using HR analytics data. The methodology includes data preprocessing, exploratory data analysis (EDA), and the development of a predictive model using XGBoost with hyperparameter optimization.

The initial phase involved data preparation and preprocessing using python on Jupyter notebook as the IDE. The HR Analytics from Kaggle went through several transformations. The employee ID was removed as it was not relevant for analysis. The target variable, 'Attrition', was converted to a binary numeric format (1 for 'Yes', 0 for 'No'). To handle categorical variables, one-hot encoding was applied using scikit-learn's OneHotEncoder, transforming categorical data into a format suitable for machine learning algorithms. The encoded features were then combined with the numeric features to create a comprehensive dataset for analysis.

Exploratory Data Analysis (EDA) was conducted to gain insights into the relationships between variables. A correlation matrix was done to identify the strengths of associations between different features. Particular attention was paid to the correlations with 'PerformanceRating' and 'Attrition', as these were key variables of interest. The top 10 features most strongly correlated with each of these variables were identified and examined.

We used the insights from the correlation matrix to create our visualizations. For this task we used python, specifically the libraries of seaborn and matplotlib. First we applied Trellis Plot on “JobSatisfaction” followed by “AgeGroup” and “Attrition”. Then we applied stacked bar plot on “YearsScienceLastPromotion” and “Attrition”. A Box plot on “PerformanceCategory” VS “percentSalaryHike” Attribites. Followed by a Bar plot on “Attrition” VS “OverTime”. Finally at the end a grouped bar plot on “MaritalStatus” and “StockOptionLevel” attributes. By applying visualizations on these attributes we gained some useful insights of the dataset.

To address the potential issue of class imbalance in the predictive modeling, which is common in attrition prediction tasks, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This technique creates synthetic examples of the minority class, balancing the dataset and potentially improving model performance.

Feature scaling was performed using StandardScaler to normalize the feature set, ensuring all variables were on a comparable scale. This is particularly important for algorithms sensitive to the scale of input features, such as the XGBoost algorithm used in this study.

The modeling approach used XGBoost, a powerful gradient boosting algorithm known for its effectiveness in various machine learning tasks. To optimize the model's performance, Optuna, a hyperparameter optimization framework, was used. An objective function was defined to tune key XGBoost parameters such as max\_depth, learning\_rate, n\_estimators, min\_child\_weight, subsample, colsample\_bytree, gamma, and scale\_pos\_weight. The optimization process used cross-validation with an F2 score as the evaluation metric, placing a higher emphasis on recall over precision, which is often desirable in attrition prediction cases.

After obtaining the optimal hyperparameters, the final XGBoost model was trained on the entire training set. The model's performance was evaluated on a held-out test set, which is 20% of the original data. To fine-tune the classification threshold, a Receiver Operating Characteristic (ROC) curve analysis was performed. The optimal threshold was determined by maximizing Youden's J statistic, which balances sensitivity and specificity.

The model's performance was assessed using multiple metrics including accuracy, precision, recall, and F1-score. Additionally, a confusion matrix was generated to provide a detailed breakdown of the model's predictions, allowing for a comprehensive understanding of its strengths and potential areas for improvement.

This methodology combines robust data preprocessing, exploratory analysis, advanced modeling techniques, and thorough evaluation to create a comprehensive approach to employee attrition prediction. The use of SMOTE for handling class imbalance, XGBoost for modeling, and Optuna for hyperparameter optimization represents a comprehensive approach to this critical HR analytics task.

**References**

* Haroon, S. (n.d.). HR Analytics Dataset. Retrieved from Kaggle: Kaggle HR Analytics Dataset
* Jide Kehinde Adeniyi, A. E. (2024). Comparative Analysis of Machine Learning Techniques for the Prediction of Employee Performance. Retrieved from Landmark University Repository: [Landmark University Repository](https://eprints.lmu.edu.ng/4419/)
* Laura Gabriela Tanasescu, A. V. (2024). Data Analytics for Optimizing and Predicting Employee Performance. Retrieved from MDPI: MDPI
* Miller, J., & Roberts, T. (2023). Enhancing Employee Attrition Prediction using Ensemble Learning Techniques. Journal of Organizational Analytics, 12(2), 150-165.
* Siahaan, M. (2021). An Analysis of Contract Employee Performance Assessment Using Machine Learning. From JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING: [JITE](https://ojs.uma.ac.id/index.php/jite/article/view/5357)
* Thomas Paul, R. B. (2024). Strategic Employee Performance Analysis in the USA: Leveraging Intelligent Machine Learning Algorithms. Retrieved from Unique Endeavor in Business & Social Sciences (UEBSS): [UEBSS](https://unbss.com/index.php/unbss/article/view/33)