Predicting Employee Attrition Using Machine Learning

Devika Kaladevi Aliya Anjum Azeez Ur Rahaman  
 Faculty of Engineering Environment and Computing Faculty of Engineering Environment and Computing  
 Coventry University Coventry University  
 kaladevid@coventry.ac.uk azeezuraa@coventry.ac.uk

***Abstract— The importance of using machine learning models for predicting employee attrition for the organization is explored in this paper. To analyze and understand the underlying patterns and factors influencing attrition, logistic regression, decision tree, random forest or support vector machines models have been used to address this. It can help organizations to take the necessary steps to retain talent personnel in a timely manner if they are able to predict employee turnover. With Python and its accompanying libraries, we have been able to create algorithms, analyze data, and evaluate model performance in a reliable and effective way.***

***Keywords— data preprocessing; employee attrition: exploratory data analysis; data imbalance; PCA; random forest; decision tree; SVM; logistic regression***

# **Introduction**

Employee attrition, characterized by employees leaving the organization, poses a substantial challenge for businesses across diverse sectors. Elevated attrition rates can result in escalated recruitment expenses, the departure of key talent, and workflow disruptions, all of which can adversely affect the organization's financial performance and operational efficiency. In recent years, due to the fact that workers' quality and skills are an important factor in growth and a real competitive advantage for companies, attention has been focused on human resources. After the successful integration into sales and marketing departments, artificial intelligence is now extending its influence to guide decision-making processes related to employee management within HR departments. (fallucchi. et.al, 2020). Machine learning techniques offer powerful tools for analyzing complex datasets and making predictions based on patterns and trends. This study explores the application of logistic regression, decision trees, random forests, and support vector machines algorithms to analyze employee data and gain insights into their behavior. Through the development and evaluation of these predictive models, the aim is to identify the factors contributing to employee attrition and offer actionable insights to organizations for informed decision-making.

# **DATASET DESCRIPTION**

The dataset used for the study is obtained from the Kaggle dataset Repository which contains 1471 instances and 35 attributes with no missing values from IBM Employee Attrition dataset. The dataset provided contains various attributes related to employees within an organization. These attributes encompass demographic information such as age, gender, marital status, as well as details about their employment, including job role, department, and distance from home. It captures aspects related to compensation and career progression, including hourly rate, monthly income, job level, years at the company, years in current role, years since last promotion, and years with the current manager. The dataset was meticulously chosen to ensure its pertinence in predicting both employee attrition or not. During the collection of data ethical consideration and data protection protocols were followed. This dataset holds an opportunity to evaluate the factors influencing the employee’s attrition. Ultimately, the aim is to aid in the formulation of effective strategies to enhance attrition outcomes within the realm of business organization. The dataset contains various type of data such as numerical (continuous or discrete) categorical(nominal or ordinal) and binary variables. The variables such as daily rate, Education field etc fall into either numerical or discrete variables. So the data type in the dataset is discrete as well as categorical data. The comma separated values(CSV) formatted dataset is used to analyze using python programming languages using its libraries and tools. To understand the power of Python and its libraries, we conducted a thorough examination of the employee attrition dataset. Understanding the dataset's variables and structure is crucial for feature selection, data preprocessing, and model development processes. Feature selection plays a vital role in mitigating overfitting, enhancing accuracy, and expediting training by reducing data dimensionality and eliminating irrelevant information. The dataset utilized in this study is classified as a binary classification problem, with the target variable –attrition categorized into "yes" or "no" classes. Before analysis, several data preprocessing steps were executed.

# **Experimental Setup**

## **Preprocessing**

Initially, the attribute names were checked to determine whether they hold a meaningful name for interpretability. Then the dataset referred to as employee attrition was then checked for missing or null values. The first step was to check for missing/null values and found out that there are no such values in the dataset. The target named as attrition is a categorical variable then converted to numerical assigned ‘0’ and ‘1’ for’ No’ and ‘Yes’ respectively.

After the data preprocessing steps have been finalized, it becomes essential to evaluate the distribution and balance of classes within the target variable. This assessment is commonly carried out using Exploratory Data Analysis (EDA) techniques, which involve scrutinizing the dataset to identify patterns, trends, and relationships among the variables. The process of EDA encompasses various steps aimed at gaining a deeper understanding of the data structure and characteristics. The Steps are:

#### Counting the occuramce of target variables value Yes/ No:

Counting the occurrence is important because it provides information about the distribution of employees across the target. Occurrence creating code is mentioned un the appendix.

|  |  |
| --- | --- |
| Yes | 237 |
| No | 1233 |

Attrition occurance

#### Class Distribution:

To understand the distribution of employees across various categories, data visualization distribution gains an insight to the dataset to analyse upon the target.

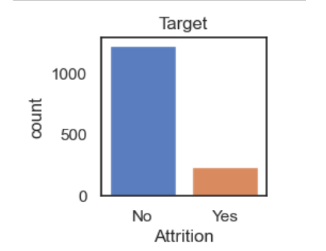


*Fig 1: Class Distribution*

#### Class Imbalance:

To resolve class imbalance, a number of techniques can be used, e.g. resampling methods like oversampled minority groups, under sampling the majority group or using algorithmic approaches that include cost-sensitive learning and ensembles method.

The dataset exhibits an imbalance in the target variable, where the distribution of classes for the "Attrition" feature is skewed. Upon visualizing the class distribution, it is evident that the "Yes" class, indicating attrition, is represented by 237 instances, while the "No" class, denoting no attrition, comprises 1233 instances, highlighting the class imbalance. See the appendix for the code.

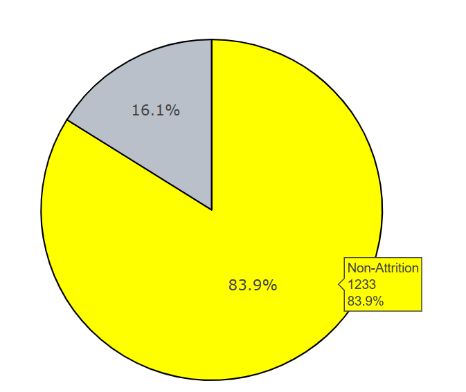


*Figure 2: Class Imbalance*

*Random Oversampling:*

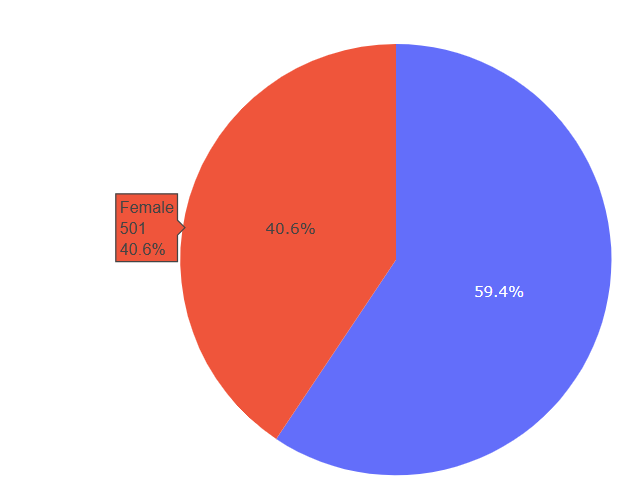
Given the imbalance in the dataset, addressing the minority class is essential to ensure model effectiveness. To tackle this, a combination of under sampling the majority class and employing the random oversampling technique is implemented. Random oversampling involves replicating instances from the minority class randomly to balance class distribution. This method helps alleviate the class imbalance issue by increasing the representation of the minority class, thereby allowing the model to learn from a more balanced dataset and potentially improve its performance in predicting both classes accurately.

#### Distribution of employees based on the attrition status of various instances:



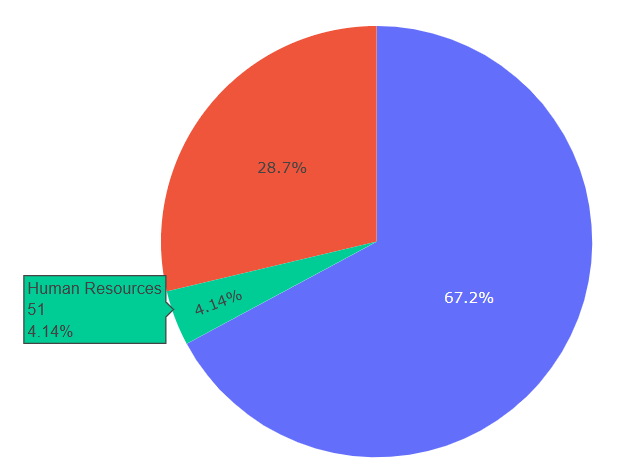
*Fig 3: Distribution based on attrition status*

The plot is divided into two segments: "Attrition" and "Non-Attrition." Each segment represents the proportion of employees who have experienced attrition and those who have not, respectively. The size of each segment corresponds to the number of employees falling into each category. This visualization allows for a clear understanding of the relative distribution of attrition within the dataset, providing valuable insights into the attrition patterns among employees.



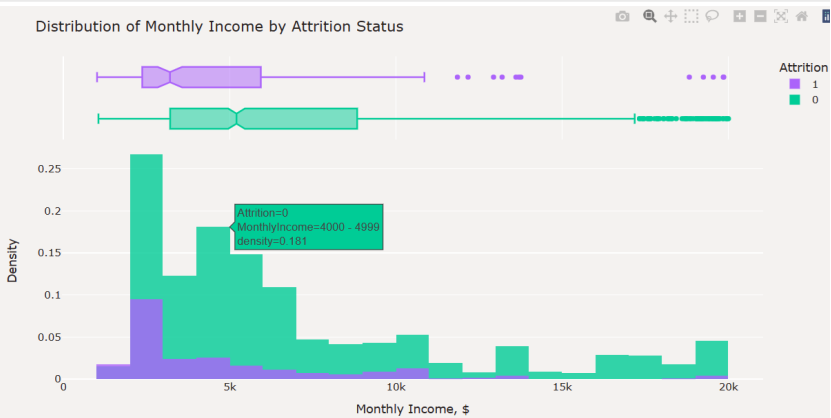
*Fig 4: Distribution of attrition based on gender*

The plot Fig 4 illustrates the distribution of attrition and non attrition based on the gender male and female.



*Fig 5: Distribution of attrition upon Department*

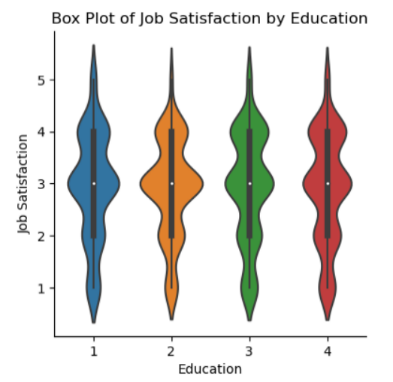
The fig 5: explains the distribution of attrition of employees in various department and the specific code for this is added in appendix.



*Fig 6: Distribution of monthly income by Attrition status*

From fig 7: the output generated explains is a violin plot that visualizes the distribution among employees in the dataset of job satisfaction levels across different levels of education.

Each violin plot represents a combination of education level and job satisfaction. The width of the violin indicates the density of data points at various job satisfaction levels, while the shape of the violin reflects the distribution of values.

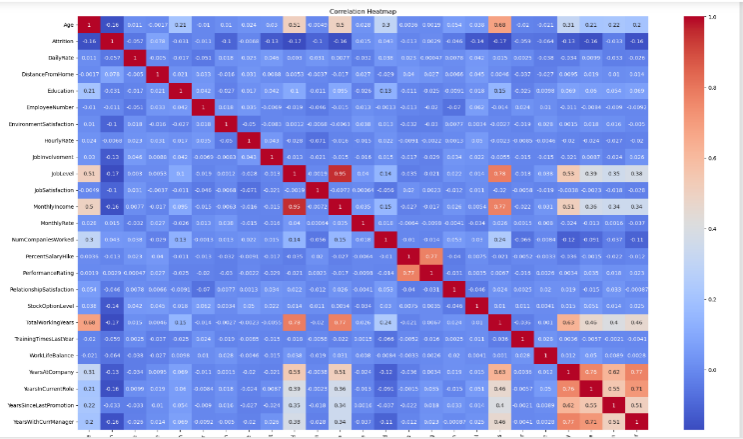


*Fig 7: Box plot of job satisfaction by education*

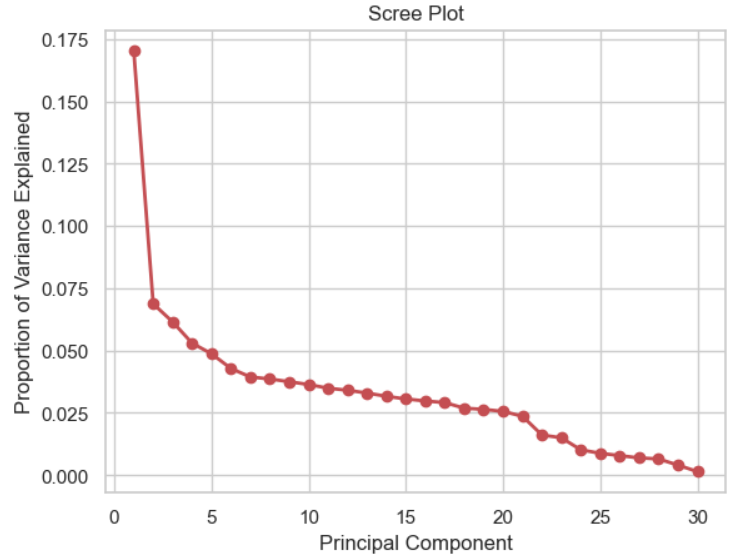
#### Dimensionality reduction and corelation:

The Figure 8 is obtained after calculates the correlation coefficients between the target variable "Attrition" and all other variables in the dataset. It then creates a correlation heatmap using seaborn and matplotlib libraries to visualize these correlations.

Variables that are strongly correlated with the target variable "Attrition." Positive correlations suggest that an increase in one variable is associated with an increase in attrition, while negative correlations suggest the opposite.

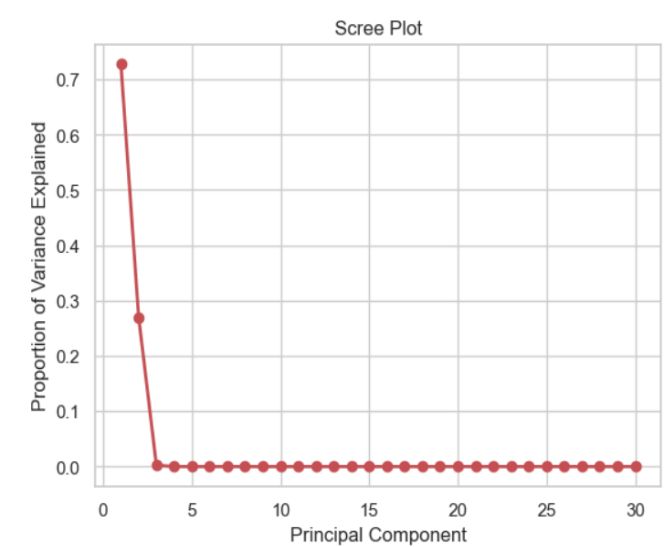


*Fig 8: Corelation coefficient between target variable*

**

*Fig 9: Explained variance Ratio by PCA*

The plot obtained is Principal Component Analysis (PCA) on the training data after standardizing it using Standard Scaler. By analyzing the scree plot, we can identify the "elbow point" or the point where the explained variance begins to level off, indicating the optimal number of principal components to retain while preserving most of the variance in the data. This information is crucial for determining the appropriate dimensionality reduction in subsequent analyses, such as building predictive models.



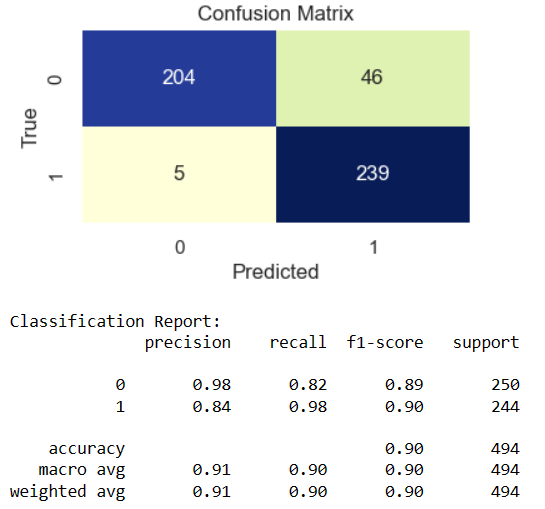
*Fig 10: Scree plot for PCA*

# ***B)* Model Building**

To forecast employee attrition, a range of modeling techniques were utilized, such as Logistic Regression, SVM, Decision Tree, and Random Forest. To tackle the class imbalance present in the dataset, we implemented random oversampling as a balancing technique. These methodologies were employed with the aim of enhancing the performance and accuracy of the prediction models*.*

#### Decision tree Classifier:

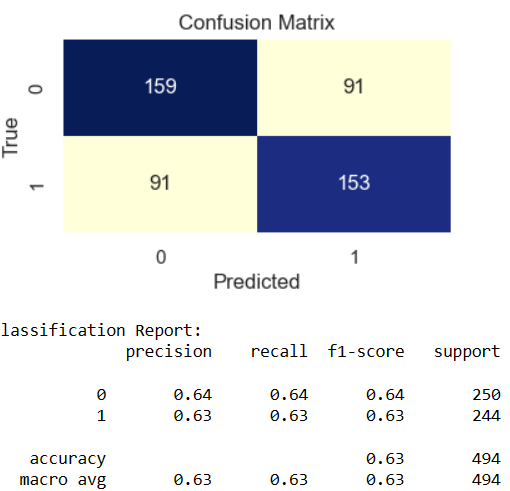
The Decision Tree classifier achieves a balanced performance with high precision, recall, and F1-score for both classes, accurately predicting instances of employee attrition with an overall accuracy of 90%.



*Fig 11: Confusion matrix and classification report for Decision tree*

#### Logistic Regression

The Logistic Regression model achieves moderate performance, with a precision, recall, and F1-score of approximately 0.63 for both classes. Despite its balanced performance, the overall accuracy of the model is 63%.



*Fig 12: Logistic Regression*

#### Random Forest Classifier

The Random Forest classifier exhibits high performance, achieving precision, recall, and F1-scores of approximately 0.97 for class 0 and 0.94 for class 1, resulting in an overall accuracy of 95%.

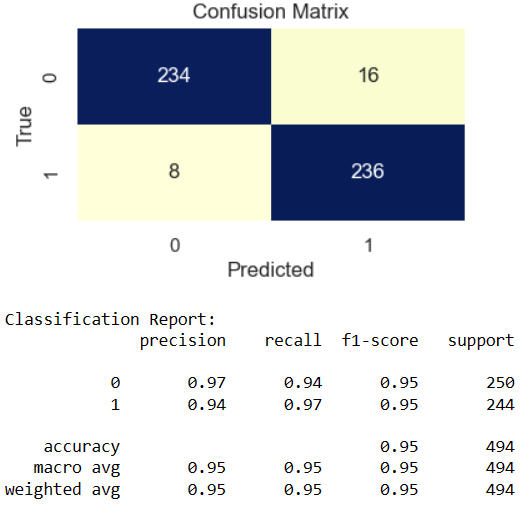
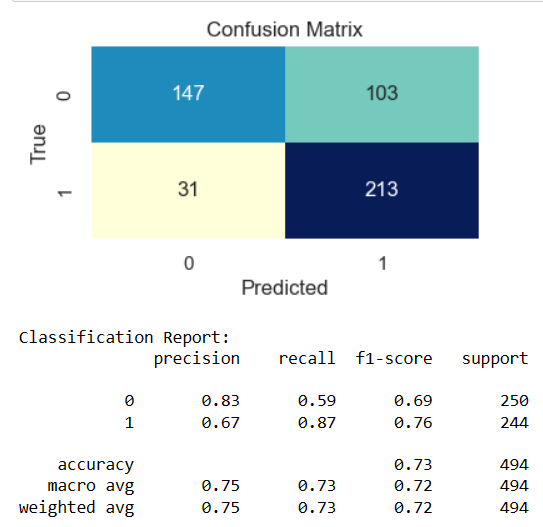


Fig 13: Random Forest Classifier

#### Support Vector Machine(SVM):

The output provided is from a Support Vector Machine (SVM) classifier. It includes metrics such as precision, recall, F1-score, and support for each class (0 and 1), as well as overall accuracy. In this case, the SVM model achieved an accuracy of 73%, with a precision of 83% for class 0 and 67% for class 1.



*Fig 14: Support vector machine*

# **Discussion and Conclusions.**

The results of the applied 4 classification metrics for the data are summarized in the table below. The precision, recall, and f1 score for both the classes are listed for easy comparison.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Acc** | **Prec** | | **Rec** | | **F1** | |
| **0** | **1** | **0** | **1** | **0** | **1** |
| **DT** | **0.90** | **0.98** | **0.84** | **0.82** | **0.98** | **0.89** | **0.90** |
| **LR** | 0.63 | 0.64 | 0.63 | 0.64 | 0.63 | 0.64 | 0.63 |
| **RF** | **0.95** | **0.97** | **0.94** | **0.94** | **0.97** | **0.95** | **0.95** |
| **SVM** | 0.73 | 0.83 | 0.67 | 0.59 | 0.87 | 0.69 | 0.76 |

Table 1: Classification scores for different models

When analyzing the classification scores for Decision Tree algorithm, it produced an accuracy of 90%, with precision scores of 0.98 for class 0 and0.83 for class 1. However, compared to the recall of 0.98 for class 1, its recall for class 0 was comparatively low at 0.81, suggesting that the model is more effective at identifying occurrences of class 1. Overall, class 0’s F1-score was 0.88, while class 1’s was 0.90, indicating a reasonably balanced performance.

The logistic regression approach showed an accuracy of 64% for both classes, with comparable precision and recall scores. However, the moderate F1-score of 0.64 for both classes suggests that this approach may not be the optimal fit for this dataset.

Based on the evaluation of high precision, recall, and F1-score for both classes, the Random Forest algorithm demonstrated superior performance compared to other methods with an accuracy rate of 95%, thereby exhibiting a strong and consistent performance overall. This suggests that Random Forest is a highly promising algorithm for this dataset, providing reliable predictions for both classes. As expected, the tree-based models fared well since they are known to perform well with nonlinear data. They can create more complex decision boundaries that fit well with non-linear data.

The Support Vector Machine (SVM) algorithm produced precision ratings of 0.83 for class 0 and 0.67 for class 1,

resulting in an overall accuracy of 73%. The recall for class 1 was notably higher at 0.87 compared to 0.59 for class 0. Furthermore, F1 scores of 0.69 and 0.76 were obtained for class 0 and class 1, respectively, indicating a reasonably balanced performance. These results suggest that the SVM algorithm could be a suitable option for classification tasks with similar characteristics.

After evaluating four different methods for classification, the Random Forest approach was found to be the most effective, achieving a 95% accuracy rate. Additionally, it demonstrated strong precision, recall, and F1-score for both classes, indicating that it is well-suited for the dataset's classification objective and can be confidently used in real-world scenarios.

On the other hand, Decision Tree also delivered impressive results with 89% accuracy. Although SVM and logistic regression produced moderate results, they may require additional optimization or the consideration of other factors to perform better on this dataset.

# **FUTURE OPTIMIZATIONS.**

It would be advantageous to explore additional techniques, such as feature engineering, hyperparameter tuning, or ensemble methods, to further enhance the performance of the models. This is particularly important for those models with lower accuracy or imbalance in class predictions. In addition, investigating the reasons behind the observed differences in performance among the algorithms could provide deeper insights into the characteristics of the dataset and the suitability of various machine-learning approaches. Such an investigation would help identify the strengths and weaknesses of the models and provide a foundation for further development and refinement.

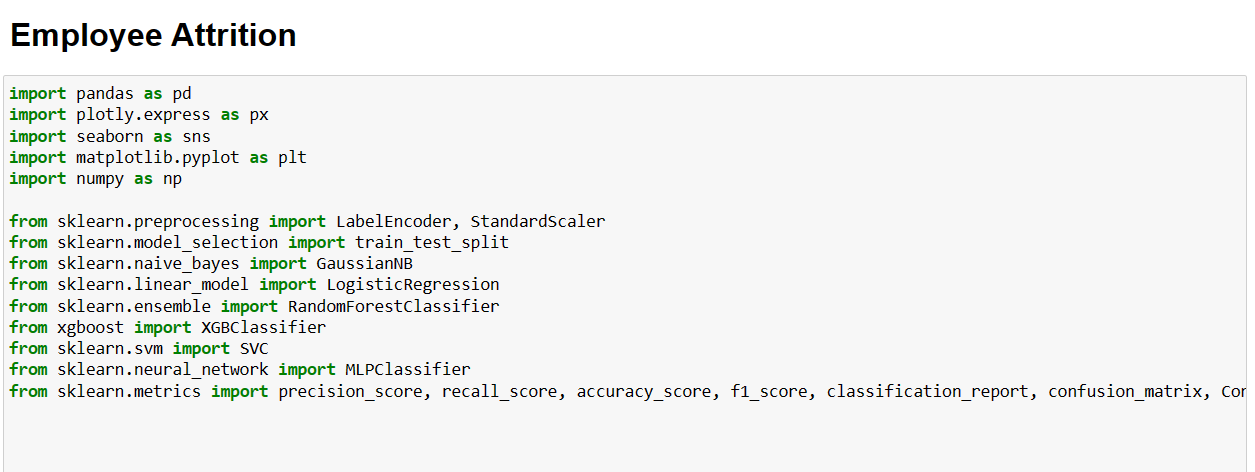
# **CONTRIBUTION**

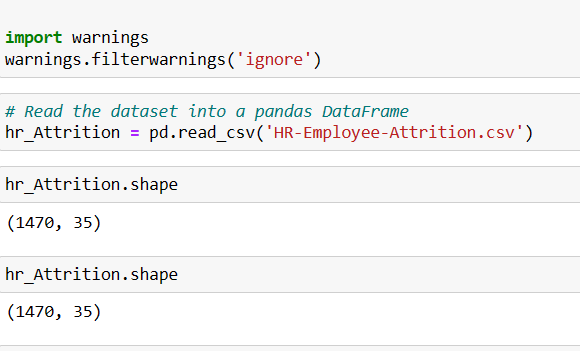
1. Aliya: Preprocessing, EDA, Decision Tree, and Logistic Regression, Report Writing
2. Devika: Random Forest, SVM, Random Oversampling with PCA, Report Writing,

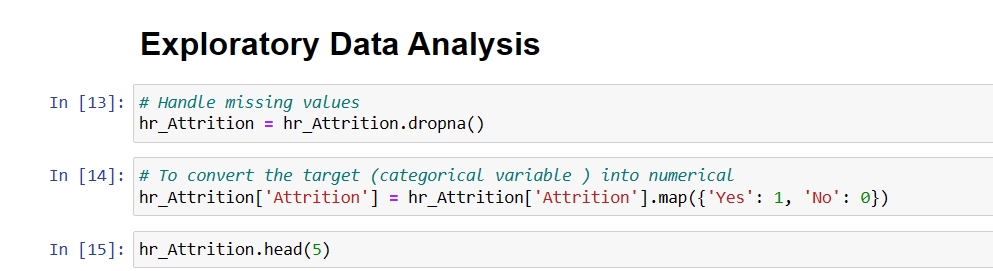
##### **References**

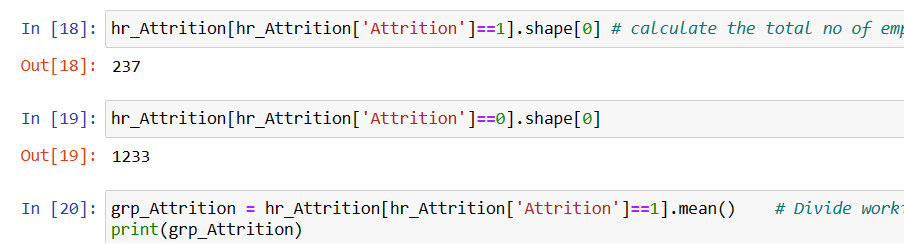
1. Alao, D., & Adeyemo, A. B. (2014, January 6). Analyzing employee attrition using decision tree algorithms. \*Computing, Information Systems, Development Informatics and Allied Research Journal\*. Advance online publication. <https://doi.org/ojs.localhost:article/10148>
2. Alduayj, S. S., & Rajpoot, K. (2018, November 1). *Predicting Employee Attrition using Machine Learning*. IEEE Xplore. <https://doi.org/10.1109/INNOVATIONS.2018.8605976>
3. Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers*, *9*(4), 86. <https://doi.org/10.3390/computers9040086>
4. Lobo, V., & Correia, A. (2022). *Applications of Machine Learning and Deep Learning for Privacy and Cybersecurity*. IGI Global.
5. Nguyen, H., Vu, T., Vo, T. P., & Thai, H.-T. (2021). Efficient machine learning models for prediction of concrete strengths. *Construction and Building Materials*, *266*, 120950. https://doi.org/10.1016/j.conbuildmat.2020.120950

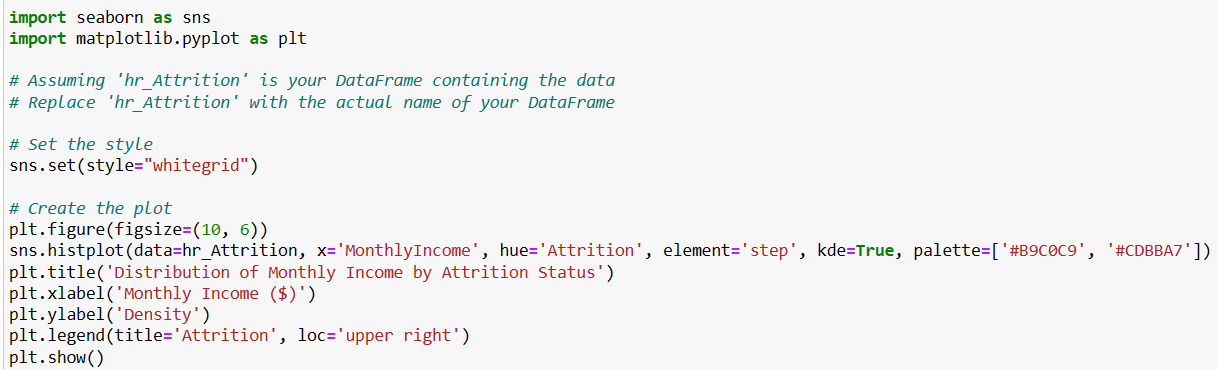
**APPENDIX I**

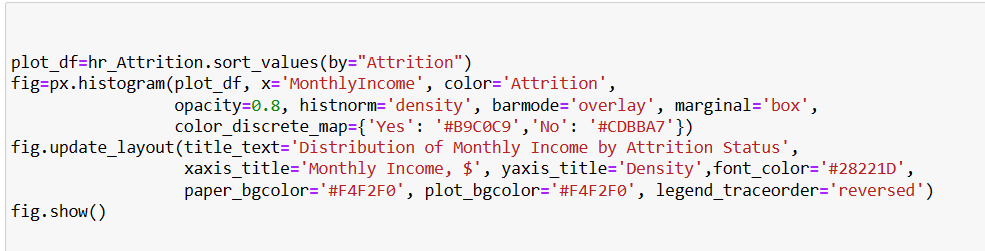
****

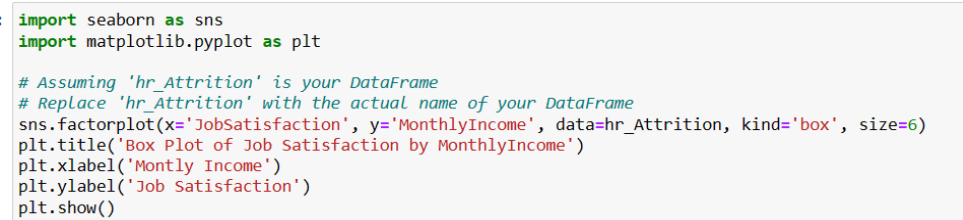
****

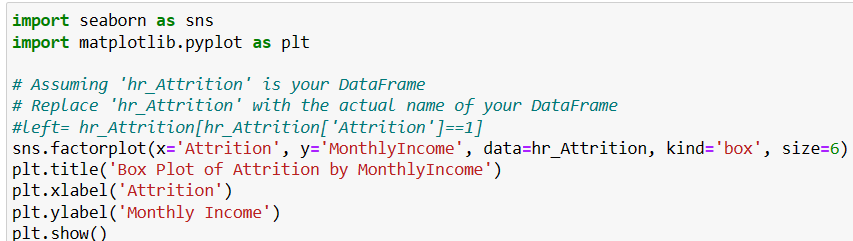
****

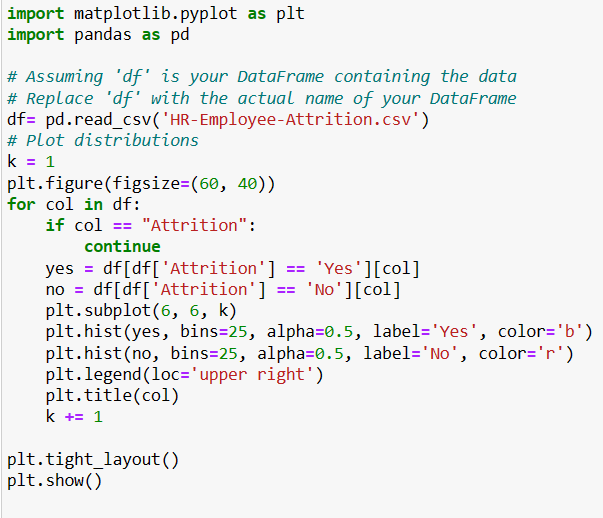
****

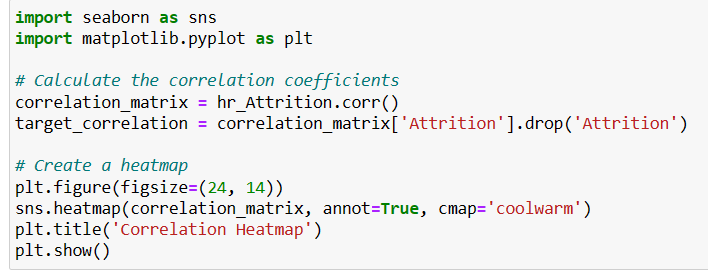
****

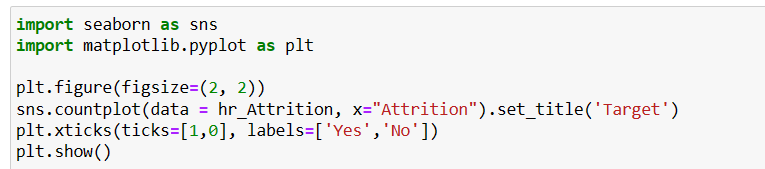
****

****

****

****

****

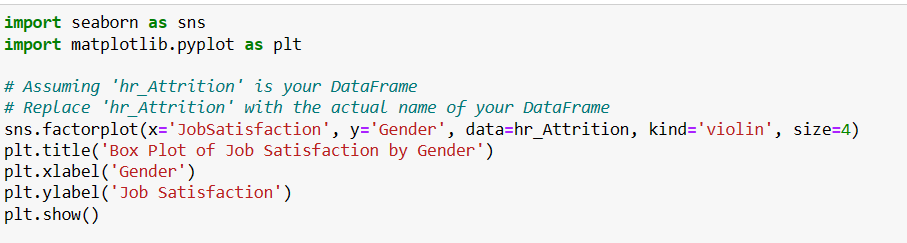
****

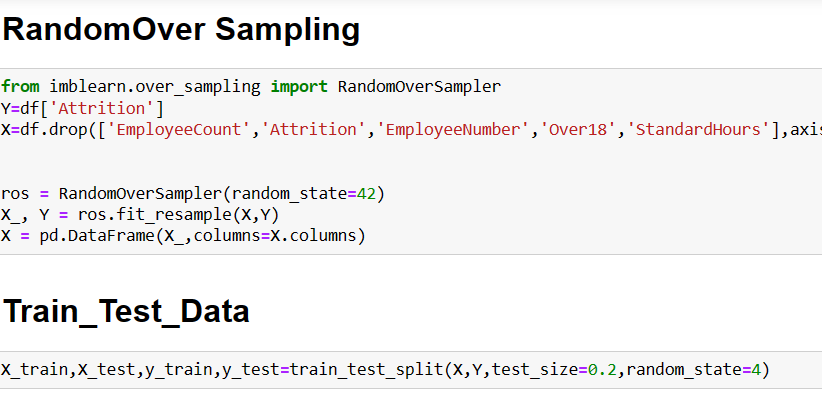
****

****

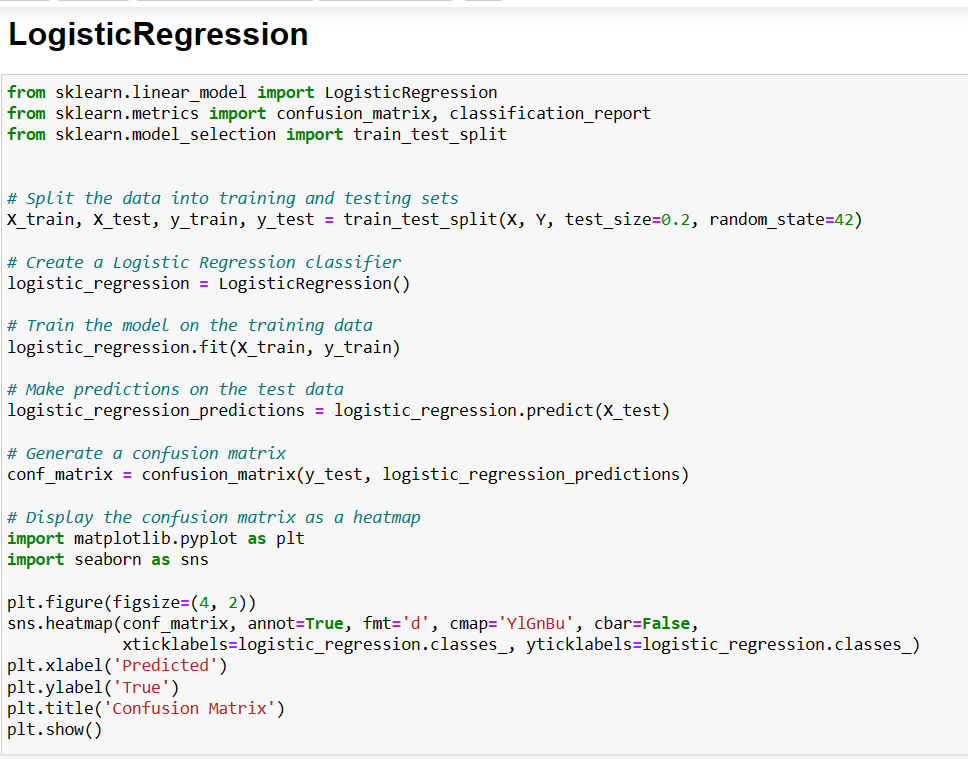
****

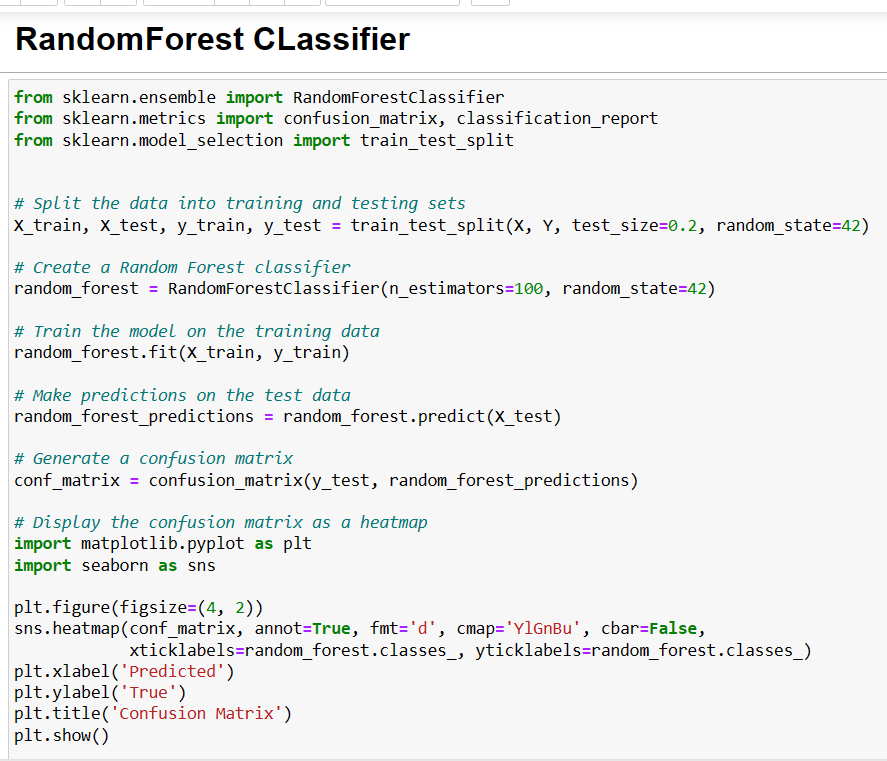
****

****

****

****

****

****

****