# Predicting Employee Attrition - Report

1. Dataset used : IBM HR Analytics Employee Attrition & Performance from kaggle.

2. Libraries: The libraries that were required for the analysis and building of the model such as Pandas, Numpy, Matplotlib, seaborn and scikit-learn.

3. Dataset Analysis:
    - The data is loaded as ‘df’ using the Pandas library. Then the structure of the dataset was checked such as info, describe, columns, shape, etc.
    - The dataset was checked for null values in order to clean the data. The fields were then encoded using the Label Encoder.
    - Then using the seaborn library various histogram representations were made to visualize the data for further analysis.
    - The correlation heatmap also helped to gain insights into the data distribution and relationships between variables.

4. Feature Engineering and Pre-Processing steps:
    - Based on the correlation heatmap produced, the irrelevant fields that do not have any impact on the result are removed such as 'EmployeeCount', 'StandardHours', 'EmployeeNumber', 'Over18' from the dataset.
    - Some features that were producing high correlation were removed and a few of the new features were introduced from existing ones by applying logarithmic transformations and to capture meaningful relationships.
    - And a better correlation heatmap is produced so that data can be used for model development and evaluation.

5. Model Development:
    - This firstly includes the separation of the target variable and then splitting of train and text data from the dataset that is processed.
    - Three machine learning models were developed for the process of prediction: Logistic Regression, Random Forest Classifier, and Decision Tree Classifier.
    - The models were trained on a portion of the dataset and evaluated on a separate test set to assess their performance.

6. Evaluation Results:
    - Model performance was evaluated using accuracy as the primary metric. Other metrics such as precision, recall, F1-score, and ROC-AUC could provide additional insights into model performance.
    - The initial accuracy that was observed without any optimisation or processing is 87.755% but later with the further improvement the accuracy was observed to be 90.136% which is the highest (Logistic Regression).
    - Logistic Regression demonstrated the highest accuracy among the models evaluated, indicating its effectiveness in predicting employee attrition in this context.
    - The other models accuracy are as follows: 88.435% for Random Forest Classifier and 80.272% for Decision Tree Classifier.

7. Optimisations used:
    - Scaling features using StandardScaler was applied to logistic regression to improve convergence and model stability.
    - The removal of redundant features helped simplify the model and reduce the risk of overfitting.
    - Feature engineering was a crucial optimization technique used to create new features that capture relevant information and potentially improve model performance.

Conclusion:

- The analysis, preprocessing, model development, and evaluation conducted on the IBM HR Analytics Attrition Dataset provided valuable insights into employee attrition prediction.
- Further optimization and experimentation, such as fine-tuning model hyperparameters or exploring advanced ensemble techniques, could be pursued to enhance predictive accuracy and robustness.