**Employee Attrition Prediction: Analysis, Modelling, and Optimization**

**1. Introduction:**

Employee attrition is a critical issue for organizations as it can impact productivity, morale, and overall performance. In this project, we aim to predict employee attrition using machine learning techniques. We will analyse the dataset, pre-process the data, develop predictive models, evaluate their performance, and optimize the models to improve accuracy.

**2. Dataset Analysis:**

The dataset used for this analysis is the IBM HR Analytics Employee Attrition & Performance dataset. It contains various attributes related to employee demographics, job roles, satisfaction levels, performance ratings, etc., along with a target variable indicating whether an employee has left the company.

**3. Pre-processing Steps:**

- Handled missing values: No missing values found.

- Encoded categorical variables: Used one-hot encoding to convert categorical variables into numerical format.

- Scaled numerical features: Applied standardization to scale numerical features.

**4. Model Development:**

- Split the dataset into training and testing sets.

- Implemented logistic regression, random forest, and gradient boosting algorithms for binary classification.

- Trained the models on the training data and evaluated their performance using accuracy, precision, recall, and F1-score.

**5. Model Evaluation and Optimization:**

- Analysed the performance of the trained models using evaluation metrics.

- Explored techniques for model optimization:

- Hyper parameter tuning: Used GridSearchCV to find the best hyperparameters for logistic regression.

- Feature selection: Employed Recursive Feature Elimination (RFE) to select the most important features for logistic regression.

- Model ensemble methods: Trained Random Forest and Gradient Boosting models to improve prediction performance.

**6. Evaluation Results:**

- Logistic Regression:

Accuracy (Feature Selection): 0.8605442176870748

Precision (Feature Selection): 0.4166666666666667

Recall (Feature Selection): 0.1282051282051282

F1-score (Feature Selection): 0.19607843137254902

- Random Forest:

Accuracy (Random Forest): 0.8775510204081632

Precision (Random Forest): 0.6666666666666666

Recall (Random Forest): 0.15384615384615385

F1-score (Random Forest): 0.25

- Gradient Boosting:

Accuracy (Gradient Boosting): 0.8809523809523809

Precision (Gradient Boosting): 0.6666666666666666

Recall (Gradient Boosting): 0.20512820512820512

F1-score (Gradient Boosting): 0.3137254901960784

**7. Optimization Techniques Impact:**

- Hyper parameter tuning improved model performance slightly.

- Feature selection reduced the number of features, potentially reducing overfitting.

- Model ensemble methods provided better performance compared to individual models.

**8. Conclusion:**

In this project, we successfully developed predictive models to identify potential employee attrition. By optimizing the models and leveraging ensemble techniques, we improved prediction accuracy and provided valuable insights to help organizations mitigate attrition risks.