Final Report

1. Problem Statement

The objective of this analysis is to develop predictive models to estimate base salary for employees based on various features, such as 'Grade,' 'Overtime_Pay,' 'Longevity_Pay,' and categorical variables like 'Gender,' 'Department,' and 'Division.' The goal is to identify the most accurate model that can be used for forecasting base salary and supporting compensation-related decision-making.

2. Objective

The primary objectives are to:

Compare the performance of different regression models, including Linear Regression, Decision Tree, XGBoost, and Random Forest, in predicting base salary.

Optimize these models using techniques like Grid Search CV to enhance their predictive accuracy.

Implement data preprocessing techniques to improve the quality and interpretability of the models.

3. Dataset

The dataset includes the following features:

Numerical Features: 'Overtime_Pay,' 'Longevity_Pay,' 'Base_Salary.'

Categorical Features: 'Gender,' 'Department,' 'Division,' 'Grade.'

Target Variable: 'Base_Salary.'

4. Methodology

Data Preprocessing:

Log Transformation: Applied to 'Overtime_Pay' and 'Longevity_Pay' to reduce skewness and stabilize variance.

Label Encoding: Categorical columns such as 'Gender,' 'Department,' and 'Division' were encoded for model compatibility.

Ordinal Encoding: The 'Grade' column was encoded using a custom order based on the average base salary, ensuring that the ordinal nature of grades is preserved.

Modeling:

Developed and compared the following regression models: Linear Regression, Decision Tree, XGBoost, and Random Forest.

Conducted hyperparameter tuning using Grid Search CV for Decision Tree and Random Forest models to improve their performance.

Evaluated the models using performance metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared.

Evaluation Metrics:

MSE: Measures the average squared difference between predicted and actual base salaries.

RMSE: The square root of MSE, indicating the average error magnitude in the predictions.

MAE: The average absolute difference between predicted and actual base salaries.

R-squared: Represents the proportion of variance in the base salary explained by the independent variables.

5. Inference

Linear Regression provided a baseline with moderate performance in predicting base salary.

Decision Tree models, particularly after hyperparameter tuning, demonstrated improved accuracy, highlighting the importance of model tuning.

XGBoost emerged as a strong candidate, offering a good balance of accuracy and interpretability.

Random Forest (with Grid Search CV) performed the best, achieving the lowest error metrics and highest R-squared value, indicating a strong fit to the data.

MSE: 113,334,741.13

RMSE: 10,645.88

MAE: 6,628.64

R-squared: 0.8742

6. Conclusion

The analysis concluded that the Random Forest model with Grid Search CV is the most accurate model for predicting base salary, followed closely by XGBoost. The thorough preprocessing steps, including log transformations and ordinal encoding, contributed significantly to the accuracy of the predictions.

These models are recommended for deployment in predicting base salary, subject to further validation and testing to ensure their reliability in practical applications. The findings underscore the necessity of proper data preprocessing and model tuning in developing accurate predictive models. Further refinement, including additional feature engineering and data collection, may enhance the models' predictive power and applicability.