Employee Salary Prediction Report

Objective:

The goal of this project is to develop a machine learning model capable of predicting the base salary of employees based on various features in the provided dataset. The project encompasses several key stages: data exploration, preprocessing, feature engineering, model building, and evaluation, all to be completed within a day.

Data Exploration:

1. **Understanding the Dataset**:
    - The initial step involves examining the dataset's structure, including the types and number of features, as well as identifying any anomalies or patterns. This process provides a foundational understanding of the data.
2. **Univariate Analysis**:
    - Conducting univariate analysis helps in understanding the distribution of individual variables. For example, histograms are used to visualize how each variable is distributed, aiding in identifying whether they follow a normal distribution or not.
3. **Bivariate and Multivariate Analysis**:
    - Heatmaps are employed to assess the correlation between different variables. This analysis reveals significant correlations, such as the strong relationship between the 'Division' and 'Department' features.
    - Boxplots are used to detect outliers in key variables like **'Base_Salary1,' 'Overtime_Pay_1,'** and **'Longevity_Pay_1,'** which could potentially skew the model's predictions.

Data Preprocessing:

1. **Handling Missing Values**:
    - The **fillna** method is used to address missing values, ensuring the dataset is complete and ready for modeling.
2. **Handling Outliers**:
    - Outliers in the dataset are treated using the capping method, where extreme values are replaced with more reasonable ones, based on statistical thresholds.
3. **Encoding Categorical Variables**:
    - Categorical variables are converted into numerical form using label encoding, which transforms categories into integers, making them suitable for machine learning algorithms.
4. **Splitting the Data**:
    - The dataset is divided into training and testing sets, typically using an 80/20 split, to allow for the evaluation of model performance on unseen data.

Feature Engineering:

Identifying and selecting the most relevant features is crucial for enhancing model performance. In this case, '**Base_Salary1'** is identified as a key feature that directly correlates with the prediction target.

Model Building:

Several machine learning algorithms are employed to build and compare models:

1. **Linear Regression**:
    - Achieved an MSE (Mean Squared Error) of 0.009511914910418967 and an R^2 score of **0.9999994362016347**, indicating a near-perfect fit. This model outperformed the others in terms of accuracy.
2. **Decision Tree**:
    - The Decision Tree model yielded an MSE of 6350.427786748267 and an R^2 score of **0.6235920065402036**. This indicates that while it captures some patterns, it does not generalize as well as other models.
3. **Random Forest**:
    - This ensemble method produced an MSE of 2621.793155098221 and an R^2 score of **0.8445988311470158**, showing improved performance over the Decision Tree model by reducing overfitting and increasing accuracy.
4. **Gradient Boosting**:
    - Gradient Boosting achieved an MSE of 1234.752982660588 and an R^2 score of **0.9268126639292507**. This model provided a good balance between bias and variance, resulting in better generalization on unseen data compared to the Decision Tree and Random Forest models.

Conclusion:

The Linear Regression model demonstrated the highest accuracy and best fit among the tested models, making it the most suitable choice for predicting employee base salaries in this dataset. Future work could involve refining the model further, exploring additional features, or employing advanced techniques like hyperparameter tuning to enhance predictive performance.