**Report: Employee Turnover Prediction**

**1. Introduction**

This report outlines the approach taken to develop a machine learning model for predicting employee turnover using a provided dataset. The project involved data preprocessing, feature engineering, and model development to identify the key factors influencing employee attrition and to select the most effective predictive model.

**2. Approach**

**Data Preprocessing**

* **Data Cleaning:** Missing values were handled, and inconsistent entries were corrected.
* **Feature Engineering:** New features were created, including the *satisfaction\_change* metric calculated as *satisfaction\_level - last\_evaluation*.
* **Normalization/Scaling:** Numerical features such as *number\_project*, *average\_montly\_hours* and *time\_spend\_company* were normalized to ensure consistent scaling and improve model performance.

**Feature Engineering**

* **Satisfaction Change:** Introduced as a new feature to capture changes in employee satisfaction relative to their last evaluation.

**Model Development**

Five machine learning algorithms were trained and evaluated on the dataset:

1. **Logistic Regression:** A simple, interpretable model suitable for binary classification.
2. **Random Forest Classifier:** An ensemble method that combines multiple decision trees for improved accuracy.
3. **Naive Bayes (GaussianNB):** A probabilistic model based on feature independence assumptions.
4. **Support Vector Machine (SVC):** A powerful model that maximizes the margin between classes.
5. **Decision Tree Classifier:** A model that provides clear interpretability and handles both categorical and numerical features.

**3. Results and Findings**

**Model Performance Metrics**

* **Logistic Regression:** The lowest performance with an accuracy of 76%, precision of 0.79 for class 0, and 0.47 for class 1.
* **Random Forest Classifier, Naive Bayes (GaussianNB), Support Vector Machine (SVC), Decision Tree Classifier:** All exhibited similar high performance with an accuracy of 98%, precision of 0.99 for class 0, and 0.97 for class 1. Among these, GaussianNB was the fastest in terms of training and prediction time.

**Visualization Insights**

* **Satisfaction Level by Salary:** Higher salaries are associated with somewhat higher satisfaction levels, though the relationship is not strong.
* **Histograms of Metrics:** Provided insights into employee satisfaction, work hours, project numbers, and promotions, revealing inconsistencies and workload imbalances.
* **Employee Turnover by Department:** Higher turnover is observed in departments with more employees.
* **Impact of Work Accidents:** Minimal impact on turnover, suggesting that work accidents are not a significant factor in employee attrition.
* **Turnover by Salary:** Higher turnover rates in low salary brackets, indicating a potential risk of underpaid employees leaving.
* **Turnover by Promotion:** Low promotion rates may contribute to higher turnover.

**4. Conclusion**

Based on the performance metrics and insights from the visualizations, the Random Forest Classifier, Naive Bayes (GaussianNB), Support Vector Machine (SVC), and Decision Tree Classifier are all strong candidates for predicting employee turnover, with GaussianNB being the most efficient. The analysis revealed that salary and promotion rates are significant factors in employee retention, with turnover being higher among employees with lower salaries and fewer promotions.

**Final Model Selection:** Given its high accuracy and speed, GaussianNB is recommended for deployment, balancing performance and efficiency effectively.