Employee Attrition Analysis Summary

Data Exploration:

- **Data Source:** CSV file
- **Dataset Size:** 1470 rows (employees) x 35 columns (features)
- **Missing Values:** None
- **Duplicate Values:** None
- **Class Imbalance:** Significant (1233 non-attrition vs. 237 attrition)

Analysis Steps:

1. **Univariate & Bivariate Analysis:**
    - Performed using histograms, countplots, and kdeplots for numerical data.
    - Used chi-square tests and correlation heatmaps for categorical data.
2. **Feature Selection:**
    - Identified key features influencing attrition using the above analyses.
    - Created a reduced dataset (final_df) containing these features.

Preprocessing and Modeling:

1. **Categorical Encoding:** Converted categorical features into numerical representations suitable for modeling.
2. **Standardization:** Scaled numerical features for improved model performance.
3. **Classification Models:**
    - Evaluated various models: Logistic Regression, Random Forest, XGBoost, Gradient Boosting, Decision Tree, and ANN (Deep Learning).
    - Addressed class imbalance using data duplication, SMOTE, and undersampling techniques.

Results:

- Initial models achieved high accuracy (~85%) but suffered from misclassification of the minority class (attrition).
- Applying techniques for class imbalance led to significant improvement, with Random Forest achieving over 97% accuracy.