Approach

- Initially inspected the data using describe,info and shape methods
- Derived target variable by analysing the LastWorkingDate
- Performed univariate analysis on all the features available
- Performed feature engineering to derived new features and trainable data as follows :-
    1. Derived tenure/duration worked as date difference between dateofjoining and lastworkingdate
    2. Grouped data based on Emp_ID,City,Education level and Joining Date
    3. Derived average age, average rating, average business value
    4. Derived promotion_flag which tells employee was promoted or not based on his last and present designations
    5. Derived rating_flag based on employees last and present ratings which indicate if ratings improved or not
    6. Derived salary_hike_flag which tells if employee received salary hike or not based on last and present salaries of employee
    7. Derived age and duration/tenure groups (categorical binning) based on average age and duration features
- The following observations were derived from EDA on these new features:
    1. Male employees had more attrition rate than female employees
    2. C20 city had highest attrition rate
    3. Employees with Master's Education had comparetively more attrition rate
    4. Age groups between 30-40 had highest attrition
    5. Employees lower designations had higher attrition
    6. Employees with no salary hike had highest attrition
    7. Employees lesser rating (lower performance) i.e no rating imporvement had highest attrition
    8. Employees with no promotion also had highest attrition
    9. Employees with less than 1 year tenure had highest attrition
- Filtered only employee ID that are not there in test data for training the model
- The data was highly imbalanced hence performed SMOTE to balance the data
- Performed one-hot encoding of categorical features and standard scaling on numerical features
- Initially performed logistic regression with RFE with 15 features
- The model was evaluated based on accuracy, precision, recall, auc_roc curve and f1-score
- The model gave around f1-score of around 0.77
- The model was test on test data with f1-score around 0.80
- The above procedures were carried out on Decision Tree Classifier, Random Forest Classifier and XGBoost Classifier but those returned unsatisfactory results