Data Preprocessing and Cleaning :-

In this section, we describe the data preprocessing and cleaning steps taken to prepare the HR dataset for analysis. The dataset contains various features relevant to employee attrition, including demographic information, job roles, satisfaction levels, and performance metrics. Ensuring the dataset is clean and consistent is crucial for deriving accurate insights and building reliable predictive models.

Data Overview :-

The original dataset comprises the following columns: Age, Attrition, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Num Companies Worked, Over18, Over Time, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, and Years With Curr Manager. Each column represents a different aspect of the employees' profiles and working conditions.

Data Cleaning and Preprocessing Steps :-

1.Dropping Irrelevant Columns

- Employee Count, Employee Number, Over18, Standard Hours, and Over Time were identified as either redundant or non-informative for the analysis. These columns were removed from the dataset.

2. Handling Categorical Variables

- Consistent labelling was ensured for categorical variables. For instance, the Attrition column was converted from Yes/No to binary 1/0.

- Business Travel values were standardized to Rarely and Frequently.

- Gender was converted to binary 1 for Male and 0 for Female.

3. Handling Numerical Outliers

- For numerical columns, we applied the Interquartile Range method to identify and handle outliers. This ensures that extreme values do not skew the analysis.

4. Handling Missing Values

- We addressed missing values by forward filling them, which propagates the last valid observation forward to the next invalid observation.

5.Encoding Categorical Variables

- For compatibility with machine learning models, categorical variables were encoded using one-hot encoding. This process converts categorical variables into a format that can be provided to ML algorithms to do a better job in prediction.

6.Feature Engineering

- We created a new feature, Years In Role Proportion, representing the proportion of time an employee has spent in their current role relative to their total working years. This feature can provide insights into career progression and stability.

Exploratory Data Analysis ;-

After the data cleaning and preprocessing steps, we performed exploratory data analysis to understand the distributions and relationships within the dataset.

Conclusion :-

The data preprocessing and cleaning steps ensured that the HR dataset was in a suitable form for analysis. By removing irrelevant columns, standardizing categorical values, handling outliers and missing values, encoding categorical variables, and creating meaningful features, we prepared the dataset for subsequent analysis and model building. These steps are critical for ensuring the validity and reliability of the insights derived from the data.