**HR Attrition Machine Learning Project**

This repository hosts a Machine Learning project focused on predicting HR attrition using a comprehensive dataset. The project aims to identify key factors contributing to employee turnover and build a predictive model to help organizations proactively address attrition challenges.

**Table of Contents**

* [Project Overview](https://www.google.com/search?q=%23project-overview)
* [Dataset](https://www.google.com/search?q=%23dataset)
* [Methodology](https://www.google.com/search?q=%23methodology)
* [Key Findings](https://www.google.com/search?q=%23key-findings)
* [Files in this Repository](https://www.google.com/search?q=%23files-in-this-repository)
* [How to Run the Project](https://www.google.com/search?q=%23how-to-run-the-project)
* [Libraries Used](https://www.google.com/search?q=%23libraries-used)

**Project Overview**

This project tackles the critical business problem of HR attrition. By analyzing various employee-related features, a machine learning model is developed to predict whether an employee is likely to leave the company. This predictive capability can enable HR departments to implement targeted retention strategies, ultimately saving costs and improving employee satisfaction.

**Dataset**

The project utilizes two datasets:

Train\_Dataset.csv and Test\_Dataset.csv1.

The Train\_Dataset.csv contains 7810 entries and 22 columns, including:

* EmployeeID 2
* Attrition 3
* Age 4
* TravelProfile 5
* Department 6
* HomeToWork 7
* EducationField 8
* Gender 9
* HourlnWeek 10
* Involvement 11
* WorkLifeBalance 12
* Designation 13
* JobSatisfaction 14
* ESOPs 15
* NumCompaniesWorked 16
* OverTime 17
* SalaryHikelastYear 18
* WorkExperience 19
* LastPromotion 20
* CurrentProfile 21
* MaritalStatus 22
* MonthlyIncome 23

The

Test\_Dataset.csv has 2630 entries and 21 columns, with similar features to the training dataset but without the 'Attrition' column, as it is meant for prediction24.

**Methodology**

The project involves several key steps:

1. **Data Loading**: Train\_Dataset.csv and Test\_Dataset.csv are loaded into pandas DataFrames25.
2. **Data Preprocessing**:

* Duplicate rows are dropped from the training dataset26.
* EmployeeID column is dropped from both training and testing datasets as it is not relevant for model training27.
* Missing values in the  
  Attrition column of the training set are filled with '1.0'28.
* A 'data' column is added to differentiate between training and testing sets before concatenation29.
* The training and testing datasets are concatenated for unified preprocessing30.

1. **Exploratory Data Analysis (EDA)**: (Based on typical ML project flow, though specific EDA steps weren't detailed in the provided snippets, it's a standard practice)
2. **Feature Engineering**: (Likely performed, but specific details not available in snippets)
3. **Model Training**: (Specific model not identified, but the project builds a machine learning model)
4. **Prediction**: The trained model makes predictions on the test dataset.
5. **Output Generation**: The final predictions are saved to a CSV file named Final\_output\_for\_Attrition.csv31.

**Key Findings**

* The training dataset initially contains 7810 entries, with 5180 non-null values for  
  Attrition, indicating some missing values that were handled during preprocessing32.
* The  
  Attrition column in the training data shows a distribution of 3735 instances of 0.0 (no attrition) and 1445 instances of 1.0 (attrition) before imputation of null values33.

**Files in this Repository**

* HR Attrition - Capstone project.ipynb: The main Jupyter Notebook containing the Python code for data loading, preprocessing, model building, and prediction. 34
* HR attrtition - ML capstone project.pdf: A PDF version of the Jupyter Notebook, possibly for presentation or easy viewing. 35
* Train\_Dataset.csv: The dataset used for training the attrition prediction model. (Assumed to be in the repository as it's read by the notebook)
* Test\_Dataset.csv: The dataset used for making predictions. (Assumed to be in the repository as it's read by the notebook)
* Final\_output\_for\_Attrition.csv: The output file generated by the notebook, containing the attrition predictions for the test dataset. 36

**How to Run the Project**

To run this project, follow these steps:

1. **Clone the repository:**  
   Bash  
   git clone <repository-url>
2. **Navigate to the project directory:**  
   Bash  
   cd <project-directory>
3. Install the required libraries:  
   The project uses standard data science libraries. You can install them using pip:  
   Bash  
   pip install numpy pandas matplotlib seaborn warnings
4. **Open the Jupyter Notebook:**  
   Bash  
   jupyter notebook "HR Attrition - Capstone project.ipynb"
5. **Run all cells:** Execute all cells in the notebook to perform data loading, preprocessing, model training, and generate the Final\_output\_for\_Attrition.csv file.

**Libraries Used**

* os 37
* numpy 38
* pandas 39
* matplotlib.pyplot 40
* seaborn 41
* warnings 42