Project Title	**Employee Attrition Analysis and Prediction**
Skills take away From This Project	**Data Preprocessing and Cleaning, Exploratory Data Analysis (EDA), Feature Engineering, Machine Learning Model Development, Model Evaluation, Streamlit Application Development**
Domain	**HR Analytics: Predicting and Preventing Employee Attrition**

Problem Statement:

Employee turnover poses a significant challenge for organizations, resulting in increased costs, reduced productivity, and team disruptions. Understanding the factors driving attrition and predicting at-risk employees is critical for effective retention strategies. This project aims to analyze employee data, identify key drivers of attrition, and build predictive models to support proactive decision-making in workforce management.

Business Use Cases:

- **Employee Retention:** Identify at-risk employees and implement targeted strategies to reduce turnover.
- **Cost Optimization:** Minimize recruitment, onboarding, and training costs associated with high attrition rates.
- **Workforce Planning:** Use predictive insights to align retention strategies with organizational goals and improve employee satisfaction.

Approach:

- **Data Collection & Preprocessing:** Gather relevant employee data, including demographics, job roles, performance, tenure, and exit interviews. Clean and preprocess the data to handle missing values, outliers, and categorical variables.
- **Exploratory Data Analysis (EDA):** Conduct in-depth analysis to identify patterns, correlations, and key factors influencing attrition, such as salary, job satisfaction, and department.
- **Feature Engineering:** Create new features like tenure categories, performance metrics, and engagement scores to enhance model accuracy.
- **Model Building:** Implement predictive models such as logistic regression, decision trees, or random forests to predict employee attrition. Evaluate model performance using metrics like accuracy, precision, and recall.
- **Actionable Insights & Visualization:** Create STREAMLIT dashboards and reports that provide HR teams with insights on attrition trends and at-risk employees, enabling data-driven retention strategies.

Results:

- **Predictive Model Accuracy:**
  A highly accurate predictive model identifying employees at risk of attrition, with an optimal performance metric (e.g., accuracy, AUC-ROC) above a certain threshold (e.g., 85%).
- **Key Drivers of Attrition:**
  Identified significant factors influencing employee turnover, such as low job satisfaction, inadequate compensation, lack of career growth, or poor work-life balance.
- **At-Risk Employees:**
  A ranked list of employees with a high probability of leaving, enabling HR teams to take proactive measures such as offering retention bonuses, career development opportunities, or addressing specific concerns.
- **Impact of Retention Strategies:**
  Insights from the model suggesting potential retention strategies, leading to a reduction in the overall attrition rate, improved employee engagement, and reduced hiring and training costs.
- **Visual Dashboards & Reports:**
  Interactive dashboards visualizing attrition trends, key factors, and predictive insights, helping HR departments make informed decisions.

Project Evaluation metrics:

- **Accuracy:**
  Measures the overall performance of the predictive model, indicating how often the model correctly predicts whether an employee will leave or stay.
- **Precision and Recall:**
  - Precision: The percentage of correctly predicted "attrition" cases out of all predicted "attrition" cases.
  - Recall: The percentage of correctly predicted "attrition" cases out of all actual "attrition" cases.
    These metrics help assess the balance between false positives and false negatives in predicting attrition.
- **F1-Score:**
  The harmonic mean of precision and recall, providing a single metric to evaluate the model’s ability to correctly identify at-risk employees.
- **AUC-ROC** (Area Under the Curve - Receiver Operating Characteristic):
  A metric to assess the model's ability to discriminate between employees who will leave and those who will stay. A higher AUC indicates better model performance.
- **Confusion Matrix:**
  A matrix showing the true positives, true negatives, false positives, and false negatives, helping visualize how well the model is performing across different categories.
- **Model Training Time and Computational Efficiency:**
  Evaluating how long the model takes to train and its computational efficiency, ensuring that it can scale effectively for large datasets.
- **Business Impact Metrics:**
  - Attrition Rate Reduction: Improvement in retention rates following the implementation of the model’s recommendations.
  - Cost Savings: Reduction in recruitment, onboarding, and training costs due to better retention strategies informed by the model.

Technical Tags:

- Data Analytics
- Machine Learning
- Classification Algorithms
- Feature Engineering
- Data Preprocessing
- Exploratory Data Analysis (EDA)

- Scikit-learn
- Matplotlib
- Model Evaluation Metrics
- AUC-ROC
- F1-Score
- Streamlit (for Dashboards)

Data Set:

Data_set link: Employee-Attrition

Dataset Explanation:

1. Age

- **Description**: The age of the employee.
- **Data Type**: Integer
- **Purpose**: Used to analyze the age distribution and its correlation with other factors like job satisfaction, attrition, or performance.

2. Attrition

- **Description**: Whether the employee left the company (1) or stayed (0).
- **Data Type**: Categorical (Binary: 0 = Stayed, 1 = Left)
- **Purpose**: Used to track employee turnover and analyze factors influencing attrition.

3. BusinessTravel

- **Description**: The frequency of business travel for the employee.
- **Data Type**: Categorical (e.g., "Travel_Rarely", "Travel_Frequently", "Non-Travel")
- **Purpose**: Used to understand travel patterns and their impact on work-life balance, job satisfaction, and performance.

4. DailyRate

- **Description**: The employee's daily rate of pay.
- **Data Type**: Numeric (Currency)
- **Purpose**: Provides insight into the pay scale and its relationship with other factors like job satisfaction, performance, or attrition.

5. Department

- **Description**: The department the employee works in (e.g., "Sales", "Research & Development", "HR").
- **Data Type**: Categorical
- **Purpose**: Used for departmental analysis to see if attrition or job satisfaction differs across departments.

6. DistanceFromHome

- **Description**: The distance in miles from the employee's home to the workplace.
- **Data Type**: Numeric (Distance)
- **Purpose**: Can be used to analyze the impact of commute distance on attrition, satisfaction, and work-life balance.

7. Education

- **Description**: The highest education level attained by the employee (e.g., "1" = Below College, "2" = College, etc.).
- **Data Type**: Integer
- **Purpose**: Helps to analyze the correlation between education level and job satisfaction, performance, or attrition.

8. EducationField

- **Description**: The field of study in which the employee obtained their degree (e.g., "Life Sciences", "Medical", "Marketing").
- **Data Type**: Categorical
- **Purpose**: Analyzes the relationship between the employee's education field and performance or job satisfaction.

9. EmployeeCount

- **Description**: A constant value, usually the total number of employees.
- **Data Type**: Integer (Constant: always the same for all records)
- **Purpose**: Often not useful for analysis but provides consistency in the dataset.

10. EmployeeNumber

- **Description**: Unique identifier for each employee.
- **Data Type**: Integer
- **Purpose**: Identifies each employee in the dataset.

11. EnvironmentSatisfaction

- **Description**: Satisfaction with the work environment (e.g., "1" = Low, "2" = Medium, "3" = High, "4" = Very High).
- **Data Type**: Integer (Ordinal)

- **Purpose**: Used to assess the impact of work environment satisfaction on employee retention and job satisfaction.

12. Gender

- **Description**: The gender of the employee (e.g., "Male", "Female").
- **Data Type**: Categorical
- **Purpose**: Used to study gender-based differences in satisfaction, pay, and attrition.

13. HourlyRate

- **Description**: The employee's hourly rate of pay.
- **Data Type**: Numeric (Currency)
- **Purpose**: Analyzes the relationship between pay and employee satisfaction, retention, or performance.

14. JobInvolvement

- **Description**: The level of involvement the employee has in their job (e.g., "1" = Low, "2" = Medium, "3" = High).
- **Data Type**: Integer (Ordinal)
- **Purpose**: Used to study how job involvement correlates with performance, satisfaction, and attrition.

15. JobLevel

- **Description**: The job level of the employee within the company (e.g., "1" = Entry Level, "2" = Mid-Level, etc.).
- **Data Type**: Integer
- **Purpose**: Can help analyze the relationship between job level and performance, satisfaction, or attrition.

16. JobRole

- **Description**: The specific role or position of the employee (e.g., "Sales Executive", "Research Scientist").
- **Data Type**: Categorical
- **Purpose**: Used to analyze how different job roles impact satisfaction, performance, and attrition.

17. JobSatisfaction

- **Description**: Satisfaction with the job (e.g., "1" = Low, "2" = Medium, "3" = High, "4" = Very High).
- **Data Type**: Integer (Ordinal)

- **Purpose**: Analyzes the correlation between job satisfaction and attrition, performance, or other factors.

18. MaritalStatus

- **Description**: The marital status of the employee (e.g., "Single", "Married", "Divorced").
- **Data Type**: Categorical
- **Purpose**: Used to analyze how marital status affects employee satisfaction, performance, or retention.

19. MonthlyIncome

- **Description**: The monthly salary of the employee.
- **Data Type**: Numeric (Currency)
- **Purpose**: Helps analyze the relationship between income and other factors like job satisfaction, performance, or attrition.

20. MonthlyRate

- **Description**: The monthly rate for the employee's work, typically indicating the total work output value.
- **Data Type**: Numeric (Currency)
- **Purpose**: Used to understand the correlation between monthly work rates and performance, satisfaction, or attrition.

21. NumCompaniesWorked

- **Description**: The number of companies the employee has worked for in the past.
- **Data Type**: Integer
- **Purpose**: Analyzes work history and its relationship with performance, satisfaction, and attrition.

22. Over18

- **Description**: A binary indicator that shows if the employee is over 18 years old (always "Y").
- **Data Type**: Categorical (Constant)
- **Purpose**: Usually not useful for analysis, as it is a constant value for all employees.

23. OverTime

- **Description**: Whether the employee works overtime (e.g., "Yes", "No").
- **Data Type**: Categorical
- **Purpose**: Used to study the relationship between overtime work and job satisfaction, performance, or attrition.

24. PercentSalaryHike

- **Description**: The percentage of the employee's salary increase over the last year.
- **Data Type**: Numeric (Percentage)
- **Purpose**: Analyzes the impact of salary increases on employee satisfaction, performance, and retention.

25. PerformanceRating

- **Description**: The employee’s performance rating (e.g., "1" = Low, "2" = Medium, "3" = High).
- **Data Type**: Integer (Ordinal)
- **Purpose**: Helps assess the relationship between performance and other factors like attrition, job satisfaction, or compensation.

26. RelationshipSatisfaction

- **Description**: Satisfaction with relationships at work (e.g., "1" = Low, "2" = Medium, "3" = High, "4" = Very High).
- **Data Type**: Integer (Ordinal)
- **Purpose**: Analyzes the impact of interpersonal relationships at work on job satisfaction and attrition.

27. StandardHours

- **Description**: The standard hours of work per week (typically constant, e.g., 40 hours).
- **Data Type**: Integer (Constant)
- **Purpose**: Usually not useful for analysis as it is a constant value.

28. StockOptionLevel

- **Description**: The level of stock options granted to the employee (e.g., "0" = No options, "1" = Low, "2" = Medium, "3" = High).
- **Data Type**: Integer
- **Purpose**: Analyzes the relationship between stock options and job satisfaction, performance, or retention.

29. TotalWorkingYears

- **Description**: The total number of years the employee has been working.
- **Data Type**: Integer
- **Purpose**: Helps in analyzing experience-related trends, such as the impact of work experience on performance or attrition.

30. TrainingTimesLastYear

- **Description**: The number of training sessions the employee attended in the past year.
- **Data Type**: Integer
- **Purpose**: Used to analyze the impact of training on performance, satisfaction, or attrition.

31. WorkLifeBalance

- **Description**: Employee’s work-life balance satisfaction (e.g., "1" = Low, "2" = Medium, "3" = High, "4" = Very High).
- **Data Type**: Integer (Ordinal)
- **Purpose**: Analyzes the relationship between work-life balance and job satisfaction, performance, or attrition.

32. YearsAtCompany

- **Description**: The number of years the employee has worked at the company.
- **Data Type**: Integer
- **Purpose**: Helps analyze the relationship between tenure and job satisfaction, performance, or attrition.

33. YearsInCurrentRole

- **Description**: The number of years the employee has been in their current role.
- **Data Type**: Integer
- **Purpose**: Used to understand how long employees stay in their roles and how this affects satisfaction and retention.

34. YearsSinceLastPromotion

- **Description**: The number of years since the employee was last promoted.
- **Data Type**: Integer
- **Purpose**: Helps assess the impact of promotions on employee satisfaction and attrition.

35. YearsWithCurrManager

- **Description**: The number of years the employee has worked with their current manager.
- **Data Type**: Integer
- **Purpose**: Analyzes the impact of long-term managerial relationships on job satisfaction, performance, and attrition.

Project Deliverables:

- **Source Code:** Python scripts for data preprocessing, model training, and Streamlit app deployment.
- **Model Files:** Trained models in Pickle/Joblib format along with hyperparameter tuning configurations and evaluation results.
- **Data:** Cleaned and preprocessed dataset in CSV/Excel format.
- **Documentation:** Brief report covering the approach, model evaluation, and deployment instructions.
- **Performance Metrics:** Key model evaluation metrics like accuracy, precision, recall, F1-score, and AUC-ROC.

Potential Predictions and Examples:(Predict any Two)

1. **Predicting Employee Attrition (Turnover Prediction)**:
   - **Goal**: Predict whether an employee will leave the company (attrition).
   - **Target Variable**: `Attrition`
   - **Features to Use**: Age, Department, Monthly Income, Job Satisfaction, Years at Company, Marital Status, Overtime, etc.
   - **Example**:
     - An employee who is older, works in a department with high turnover, has low job satisfaction, and works overtime frequently might be more likely to leave the company.
     - You can build a machine learning model (e.g., logistic regression or decision tree) to predict the probability of attrition based on these features.
2. **Predicting Performance Rating**:
   - **Goal**: Predict the employee's performance rating.
   - **Target Variable**: `PerformanceRating`
   - **Features to Use**: Education, Job Involvement, Job Level, Monthly Income, Years at Company, Years in Current Role, etc.
   - **Example**:
     - Employees with more years at the company, higher job involvement, and higher job levels are likely to receive higher performance ratings. A model can use these features to predict performance ratings.
3. **Predicting Employee Promotion Likelihood**:
   - **Goal**: Predict the likelihood of an employee getting promoted.
   - **Target Variable**: **YearsSinceLastPromotion** (i.e., predict when the employee will be promoted).
   - **Features to Use**: Job Level, Total Working Years, Years in Current Role, Performance Rating, Education, etc.

  - **Example**:
    - An employee with a high performance rating, long tenure, and experience in a relevant role might be more likely to receive a promotion soon.

Example of How You Can Approach It:

- **Attrition Prediction**:
  - **Problem**: Predict if an employee will leave the company.
  - **Approach**: Train a **classification model** (e.g., Random Forest, Logistic Regression) to predict whether an employee will leave (`Attrition = 1`) or stay (`Attrition = 0`) based on features like `Age`, `Gender`, `MonthlyIncome`, `JobSatisfaction`, `OverTime`, etc.
  - **Result**: The model will give the probability of attrition for each employee, and you can identify employees who are at high risk of leaving the company, helping to take preventative actions.
- **Job Satisfaction Prediction**:
  - **Problem**: Predict the job satisfaction level (on a scale of 1 to 4).
  - **Approach**: Train a **regression model** (e.g., Linear Regression, Decision Trees) using features like `JobInvolvement`, `WorkLifeBalance`, `JobRole`, and `YearsAtCompany` to predict the job satisfaction score.
  - **Result**: Predict how satisfied each employee is with their job and focus on improving job satisfaction for employees with lower satisfaction scores.

Project Guidelines:

- **Coding Standards:**
  Write clean, readable code with descriptive variable and function names. Organize code into functions and modules for better maintainability.
- **Version Control:**
  Use Git for version control with clear and concise commit messages. Maintain separate branches for features and bug fixes to ensure a clean development flow.

- **Testing:**
  Test data preprocessing, model training, evaluation, and the Streamlit app thoroughly.
  Write unit tests for critical functions to ensure the reliability of the code.
- **Documentation:**
  Include a README.md file with setup, installation, and usage instructions.
  Provide a project report detailing the approach, methodology, results, and key insights.
- **Deployment:**
  Ensure the Streamlit app is deployable and provide clear deployment instructions.
  Document any dependencies or environment setup needed for deployment.

Timeline:

The project must be completed and submitted **within 14 days from the assigned date**.

References:

Streamlit recording (Tamil)	**STREAMLIT_SESSION.mp4**
Reference Document	**https://docs.streamlit.io/develop/api-reference**
Streamlit recording (English)	**Special session for STREAMLIT(11/08/2…**
Project Live Evaluation	**Project Live Evaluation**
EDA Guide	**Exploratory Data Analysis (EDA) Guide**

Capstone Explanation Guideline	**Capstone Explanation Guideline**
GitHub Reference	**How to Use GitHub.pptx**
Project Orientation (Tamil)	Project Orientation Session : Employee …
Reference PPT(How to do the project)	ppt.pptx

PROJECT DOUBT CLARIFICATION SESSION (PROJECT AND CLASS DOUBTS)

About Session: The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.
Note: Book the slot at least before 12:00 Pm on the same day

Timing: Monday-Saturday (4:00PM to 5:00PM)

Booking link :https://forms.gle/XC553oSbMJ2Gcfug9

LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)

About Session: The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.
Note: This form will Open only on Saturday (after 2 PM) and Sunday on Every Week

Timing: Monday-Saturday (05:30PM to 07:00PM)

Booking link : https://forms.gle/1m2Gsro41fLtZurRA

Approval Workflow

Created By:	Verified By:	Approved By:
Gomathi A	Shadiya P P	Nehlath Harmain