Case Study

Absenteeism Analysis for a Renowned Company

Background:

In today's competitive corporate environment, employee productivity is one of the most significant factors determining a company's success. A renowned multinational corporation (MNC), known for its commitment to operational efficiency, has noticed a concerning trend in its workforce—an increase in absenteeism. Employees are frequently absent, and management is unsure whether these absences are due to legitimate reasons, personal issues, or other factors. The company leadership is concerned that absenteeism is affecting overall productivity and morale, and they want to understand the underlying causes so they can implement effective interventions.

The company has provided a dataset with detailed information about the factors influencing employee absenteeism. The dataset captures employee demographics, transportation expenses, workload, social habits, and absenteeism hours.

Scenario:

A mid-sized manufacturing company, XYZ Corp, has been experiencing a significant increase in employee absenteeism over the past year. This rise in absenteeism has led to decreased productivity, missed deadlines, and increased operational costs. The HR department has collected detailed absenteeism data, including reasons for absence, transportation expenses, distance from residence to work, and other relevant factors. The company seeks to understand the key factors contributing to absenteeism and develop a predictive model to identify employees who are at risk of high absenteeism. This will help the company implement targeted interventions to reduce absenteeism and improve overall productivity.

Problem Statement

XYZ Corp is facing a challenge with increasing employee absenteeism, which is negatively impacting productivity and operational efficiency. The company needs to identify the key factors contributing to absenteeism and develop a predictive model to forecast absenteeism hours. This will enable the HR department to implement targeted interventions and policies to reduce absenteeism and improve employee attendance.

Objectives

1. Analyze the absenteeism data to identify key factors contributing to employee absenteeism.

2. Develop a machine learning model to predict absenteeism hours based on the identified factors.
3. Provide actionable recommendations to the HR department to reduce absenteeism and improve productivity.

Methodology

1. **Data Collection:** Use the absenteeism data provided by the HR department.
2. **Data Preprocessing:** Clean and prepare the data for analysis, including handling missing values and encoding categorical variables.
3. **Exploratory Data Analysis (EDA):** Perform statistical analysis and visualizations to identify patterns and correlations in the data.
4. **Machine Learning Model:** Develop and evaluate Linear Regression and Decision Tree Regressor models to predict absenteeism hours.
5. **Recommendations:** Based on the insights from the analysis and model predictions, provide recommendations to the HR department.

Expected Outcomes

1. Identification of key factors contributing to employee absenteeism.
2. A predictive model that accurately forecasts absenteeism hours.
3. Actionable recommendations for the HR department to implement targeted interventions and policies to reduce absenteeism.

Your analysis will help the company understand what factors are contributing to higher absenteeism rates and how these factors vary across different employee groups.

Dataset Overview:

The dataset includes the following key columns:

1. **ID**: Employee identification number.
2. **Reason for absence**: A categorical variable indicating the reason for the employee's absence (e.g., medical, personal issues).
3. **Month of absence**: The month in which the absence occurred.

4. **Day of the week**: The day of the week on which the absence occurred.
5. **Transportation expense**: The transportation cost incurred by the employee.
6. **Distance from Residence to Work**: The distance (in kilometers) between the employee's residence and workplace.
7. **Service time**: The number of years the employee has worked at the company.
8. **Age**: The employee's age.
9. **Workload Average/day**: The average daily workload of the employee.
10. **Disciplinary failure**: A binary variable indicating whether the employee has faced any disciplinary actions.
11. **Education**: The education level of the employee.
12. **Social drinker**: Whether the employee consumes alcohol socially (Yes/No).
13. **Social smoker**: Whether the employee smokes socially (Yes/No).
14. **Absenteeism time in hours**: The total number of hours the employee was absent.

Questions to Consider:

1. **Absenteeism Trends**:
    - Are there specific **days of the week** or **months** where absenteeism is more common? This can help the company understand if absences are related to personal scheduling or external factors.
2. **Employee Characteristics and Absenteeism**:
    - Do factors such as **age**, **distance from work**, or **transportation expenses** correlate with higher absenteeism rates? Are older employees or those living farther away more likely to be absent?
3. **Social Habits and Absenteeism**:
    - Are **social habits** (e.g., smoking or drinking) linked to higher absenteeism hours? Are social drinkers or smokers more likely to miss work compared to their peers?
