Boston University – Metropolitan College

CS677 - Data Science with Python

FINAL PROJECT

Exploring Employee Absenteeism

Department of Computer Science

Student ID - U42035592

Professor -
Eugene Pinsky

Submitted by -
Akanksha Ankam

Summary:

Understanding Employee Absenteeism -

In this data science project, my quest to grasp the complexities of employee absenteeism, I undertook a diverse journey of data analysis, visualization, and predictive modeling. This exciting adventure aimed to uncover deep-seated insights by diving into the intricacies of employee data. The main goal was to uncover the various factors that contribute to absenteeism in an organization and build a strong data-driven base to shape smarter HR strategies.

This process allowed me to explore data like a detective, piecing together information through visualizations and models. By navigating through missing data and creatively engineering new features, I not only overcame challenges but also equipped HR teams with valuable insights. This experience highlighted the power of data in driving informed decisions, making it clear that numbers can reveal fascinating stories about employee behavior.

Data Preprocessing:
The project started with data preprocessing, including handling missing values. Numerical columns were filled with the mean of the column, while categorical columns were filled with the mode. The categorical variables were one-hot encoded to facilitate analysis.

Feature Engineering:
Feature engineering was crucial for better understanding the data. New features like "TotalExperience" (sum of age and length of service) were created to capture complex interactions. These engineered features provided insights into non-linear relationships and potential factors affecting absenteeism.

Exploratory Data Analysis (EDA):
EDA revealed key insights such as the positive correlation between age and absentee hours. Gender-based differences in absenteeism were identified, with females showing higher absenteeism rates. The analysis also highlighted divisional variations and the impact of tenure and experience on absenteeism.

Machine Learning:
Machine learning models were used to predict absentee hours. “Linear Regression, Random Forest Regression”, and Hyperparameter Tuning using GridSearchCV were employed. The best-performing model after tuning was the Random Forest Regressor, which demonstrated the lowest Mean Absolute Error and Mean Squared Error.

Feature Importance Analysis:

Feature importance analysis was conducted to identify key drivers of absenteeism. It showed that features like age, total experience, and length of service were critical contributors. This analysis provided actionable insights for HR teams to develop targeted interventions.

Data Visualizations:

Visualizations breathed life into the analysis. Scatterplots vividly displayed the age-absenteeism relationship, exposing a trend of increased absentee hours with age. Bar plots eloquently communicated the disparity in absenteeism across gender and divisions, engaging stakeholders with compelling visuals. Heat maps elegantly summarized feature correlations, providing a holistic view of interdependencies.

Conclusions:

The project's main findings were:

- Age and absentee hours had a positive correlation, indicating that older employees tended to have higher absenteeism rates.
- Divisions exhibited varied absenteeism rates, with the "Finance and Accounting" division having the highest rates.
- Gender disparities were observed, with female employees experiencing higher absenteeism.
- Tenure and experience influenced absenteeism trends, suggesting the need for tailored strategies.

Significance:

This project underscores the power of data science in HR decision-making. By leveraging data-driven insights, organizations can optimize workforce management and enhance employee engagement. The findings guide HR strategies, ensuring employee well-being and overall organizational success.

Future Implications:

Moving forward, organizations can use these insights to design targeted interventions, optimize employee engagement, and foster a positive work environment. Integrating data-driven strategies into HR practices can lead to improved absenteeism management and overall workforce effectiveness.

Thank you!!!