Understanding Absenteeism Patterns in the Workplace

Mohsen Selseleh

Assignment for course: **Data 6100**, Introduction to data science (2021-22, Fall),

University of Guelph

Introduction:

Workplace absenteeism has an enormous impact on performance, employee morale, and productivity, which worries businesses all over the world. Developing methods that effectively manage absenteeism requires an understanding of the elements that contribute to this problem. This research proposal uses a comprehensive dataset that includes individual characteristics, reasons for absence, and other relevant information to investigate patterns of workplace absenteeism. Through the analysis of this dataset, we hope to determine the main causes of absence and create models that can be employed to predict absenteeism patterns in the future.

Objectives:

1. To examine the relationship between individual characteristics and absence trends.
2. To investigate into how absenteeism rates are affected by the reasons given for absences.
3. To create forecasting models for absence patterns using the factors that have been found.

Methodology:

Data Collection:

- Utilize the provided dataset containing information on individual characteristics, reasons for absence, month of absence, day of the week, seasons, transportation expenses, distance from residence to work, service time, age, workload, disciplinary history, education level, family size, social habits (drinking and smoking), pet ownership, weight, height, and body mass index (BMI). Sample size is 740, no missing value.

Goals:

1. Descriptive Statistics: Examine the distribution of variables and identify any trends or patterns.

2. Correlation Analysis: Investigate the relationships between different variables, such as the correlation between reasons for absence and absenteeism rates.

3. Regression Analysis: Build regression models to explore the impact of individual characteristics, reasons for absence, and contextual factors on absenteeism.

4. Machine Learning Model: Develop predictive model using machine learning algorithms such as logistic regression to forecast absenteeism trends.

Evaluation:

To evaluate the predictive model's performance, determines including accuracy, precision, recall, and F1 score are assessed. This thorough analysis helps to determine the best method for absenteeism forecasting.

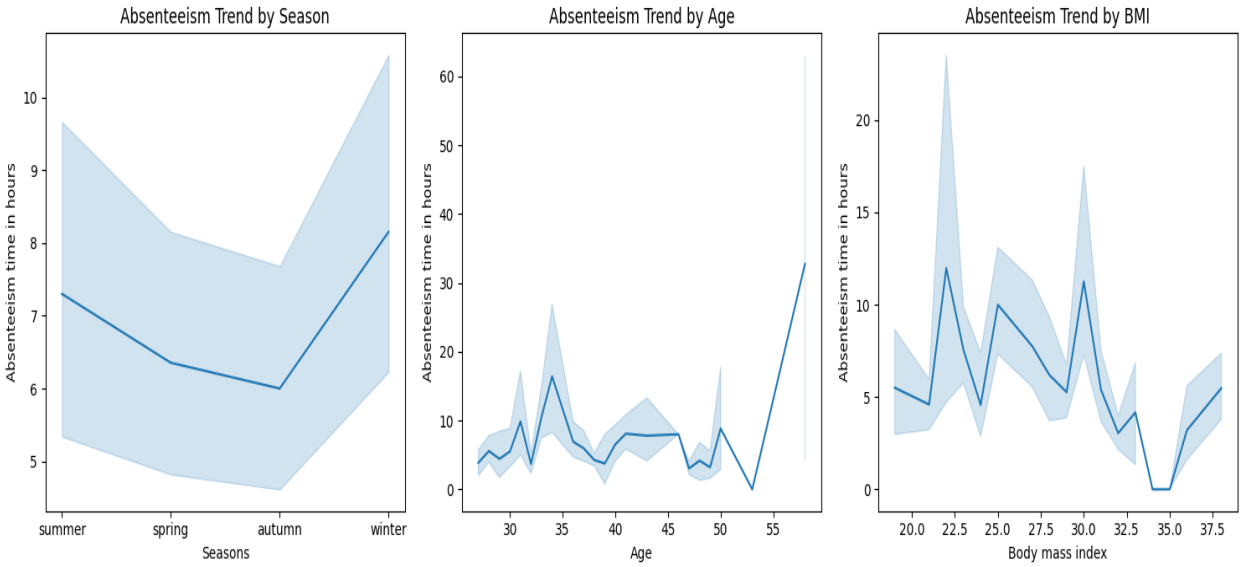
Expected Outcomes

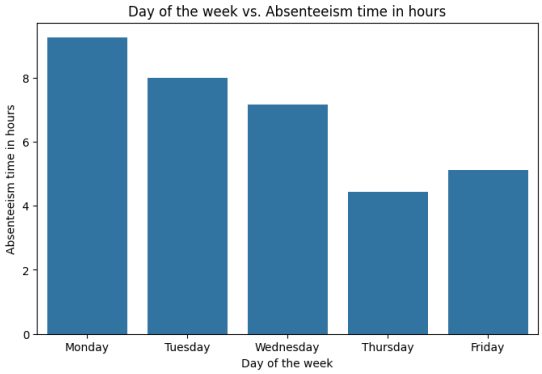
1. Identification of key factors contributing to absenteeism in the workplace.

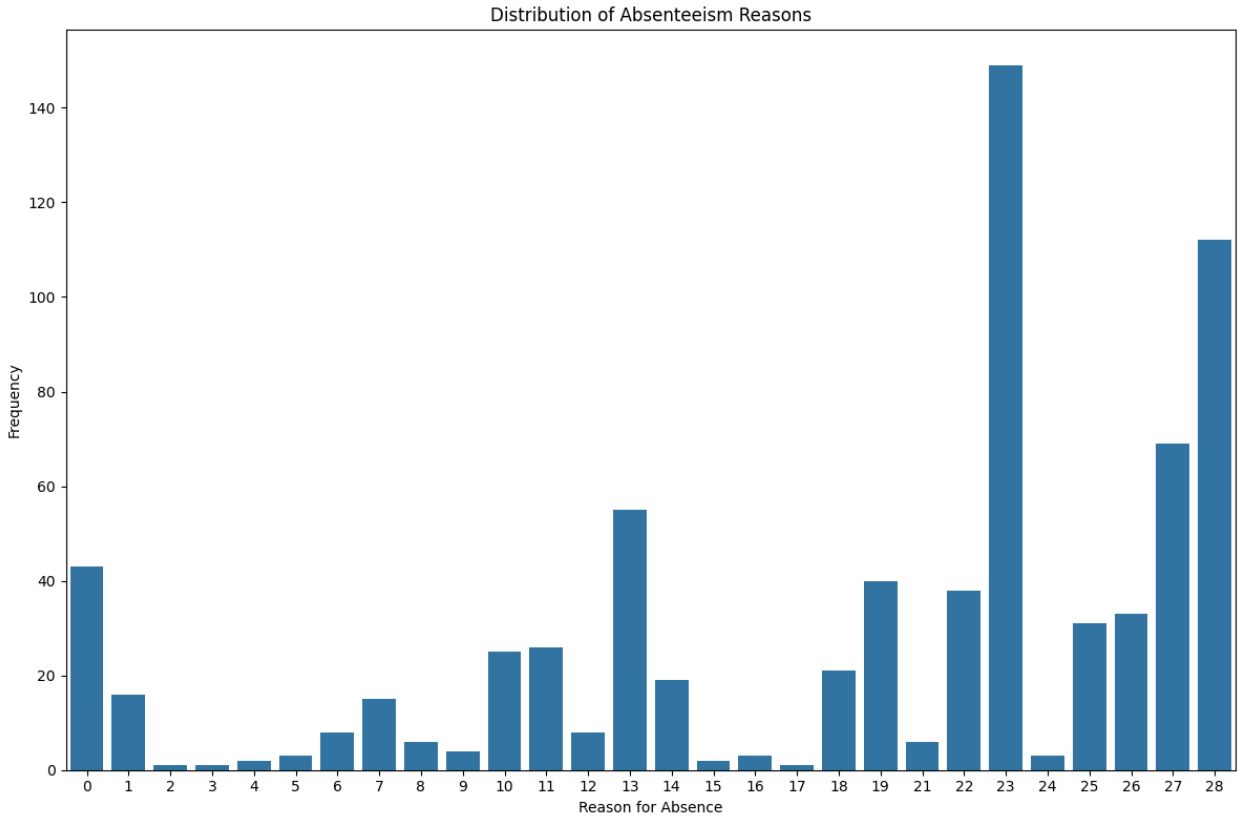
2. Development of predictive models to forecast absenteeism trends with high accuracy.

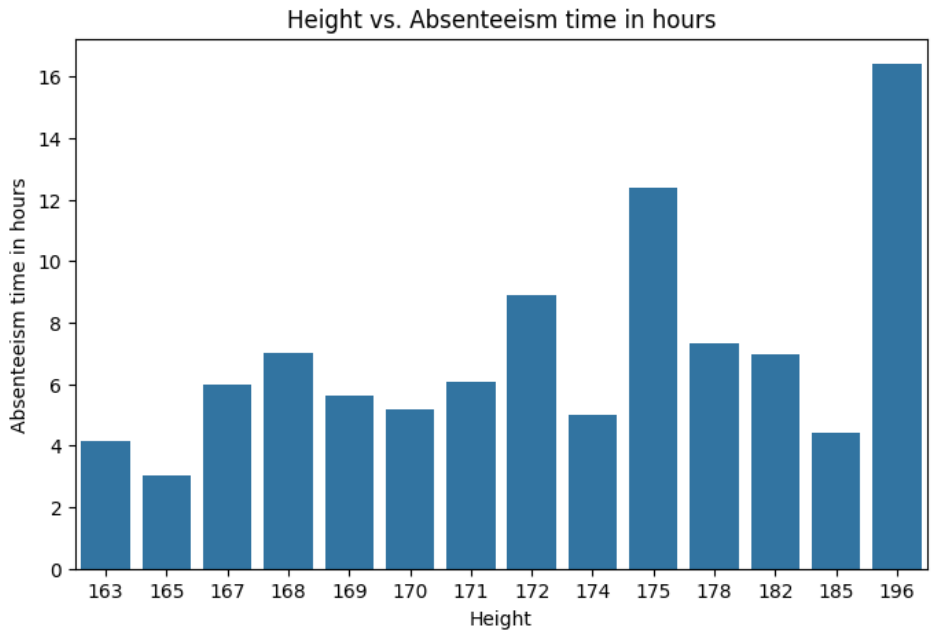
3. Insights into strategies for mitigating absenteeism based on the identified factors.

Result:


Absenteeism Trend by Season
Absenteeism time in hours
summer
spring
autumn
winter
Seasons
Absenteeism Trend by Age
Absenteeism time in hours
Age
Absenteeism Trend by BMI
Absenteeism time in hours
Body mass index


Day of the week vs. Absenteeism time in hours
Absenteeism time in hours
Monday
Tuesday
Wednesday
Thursday
Friday
Day of the week


Distribution of Absenteeism Reasons
Frequency
Reason for Absence
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21 22 23 24 25 26 27 28
0 20 40 60 80 100 120 140


Height vs. Absenteeism time in hours
Absenteeism time in hours
Height
163 165 167 168 169 170 171 172 174 175 178 182 185 196
0 2 4 6 8 10 12 14 16

The analysis of absenteeism reasons reveals that medical consultation had the highest frequency, followed by dental consultation and physiotherapy. Conversely, neoplasms, blood diseases, and congenital malformations showed the lowest frequencies, indicating less occurrence as reasons for absenteeism.

Season-wise analysis indicates that winter and summer recorded the highest absenteeism rates, while autumn had the lowest.

Concerning age, individuals aged 35 exhibited the highest absenteeism, while those aged 54 showed the lowest. Interestingly, absenteeism tends to increase with age after 54. Additionally, individuals with a higher body mass index demonstrated lower absenteeism rates.

Analysis by day of the week reveals that Monday, followed by Tuesday and Wednesday, had the highest absenteeism rates.

Factors such as distance from residence to work, disciplinary failure, and height also showed meaningful correlations. Taller individuals tended to have higher rates of absenteeism.

Regression:

The regression model displayed significance at a 5% level, with an adjusted R-squared of 0.321 and a notable F-statistic of 18.46 (p-value < 0.05). Among the predictors—age, reason for absenteeism, day of the week, education, and number of children—only age, reason for absenteeism, day of the week, education, distance from residence to work, disciplinary failure, Number of children, height, were statistically significant predictors of absenteeism.

To enhance the adjusted R-squared value, consider including additional variables, transforming existing ones, addressing multicollinearity, refining model specification, detecting and treating outliers, exploring nonlinear modeling techniques, and incorporating interaction terms. These approaches may improve the model's explanatory power.

Logistic Regression:

In summary, the logistic regression model performs well with high accuracy, precision, and recall scores of approximately 0.94, indicating effective classification. Although it achieves a balanced F1-score of around 0.97, the ROC-AUC score of approximately 0.61 suggests moderate discriminative ability. While the model demonstrates strong predictive capabilities overall, there's room for improvement to enhance its ability to differentiate between classes.

Conclusion:

We have provided a thorough framework for examining workplace absence patterns and creating predictive models for predicting future trends through this study proposal. Organizations can strategically engage to decrease absenteeism and improve overall performance by analyzing the factors that influence it. This proactive strategy promotes a healthier and more productive work environment in addition to reducing the disruptions caused by absenteeism.

Ethical Considerations

- Ensure data privacy and confidentiality.

- Avoid discriminatory practices in absenteeism management.

- Use findings to support employee well-being and organizational effectiveness.

Reference

UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/445/absenteeism+at+work

Please keep in mind that codes and dataset are available at:

https://github.com/Mohsenselseleh/Absenteeism