**Employee Absenteeism Prediction**

Sonachi Mogbogu

Machine Learning Project

June 25, 2020

**Table of Contents**

[INTRODUCTION](#_p38m6jxm6z3m) 3

[RELATED WORKS](#_vqgjxptvt9gx) 4

[DATA](#_co6yej530qh0) 5

[TECHNICAL APPROACH](#_9sdzvfhd7vfs) 11

[TEST AND EVALUATION](#_l49iogeqjmd5) 13

[REFERENCES](#_x0kjozn6d0uq) 13

INTRODUCTION

XY&Z is a courier company that prides itself on its capability to transport goods and services on time. Although, as good and well serving the company has been to its customers, it is not without flaws. XY&Z has recently discovered that it is currently facing absenteeism. Absenteeism at work occurs when the employee does not show up during business hours, or the employees are mentally unavailable or are battling with difficulties at their positions. We have been presented with absenteeism records from 2007 to 2010, showing a notable decrease in productivity and revenue. We have sought ways to mitigate this issue, and one way to tackle it is by conducting a root cause analysis. Machine learning has proven to be beneficial in root-cause analysis, especially in the healthcare industry. Therefore, we would be employing machine learning algorithms and techniques to help us understand the current dilemma the company is facing and also predict the absenteeism hours in the future. The root-cause analysis process comprises the following steps: Data collection and Exploration, Prediction models, and Evaluation.

**Keywords**

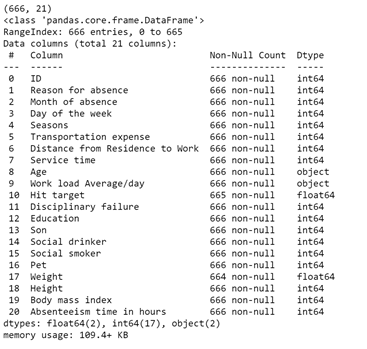
Absenteeism; Machine Learning; Supervised Learning; Classification Trees

# **RELATED WORKS**

There have been multiple studies conducted concerning absenteeism at work. Below are some of the related works that we found to be helpful for this project.

# **DATA**

As mentioned, we have been provided with the dataset titled: *Absenteeism\_at\_work\_train.csv*, to help better address the issue of absenteeism. This dataset runs from July 2007 to July 2010 of the company’s absenteeism records. The dataset has the shape of (666,21), i.e., 666 observations and 21 variables. Two variables (Weight and Hit Target) have less than 666 records, indicating the presence of missing values. Below is a display of the variables in the dataset.

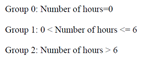


The variable “Reason for absence” uses the International Classification of Diseases (ICD) as a significant predictor in the analysis of absenteeism. The ICD gives guidelines in identifying and classifying diseases globally. In this dataset, the ICD is stratified into 21 categories, and seven other categories are not certified by the ICD. Together, they would be fed into the model along with other feature selections.

As per the following requirements, all these variables (Reason for absence, Month of absence, Day of the week, Seasons, Disciplinary failure, ‘Education, Social drinker, and Social smoker) were converted to string data types while the rest remained as int types. The target variable, “Absenteeism time in hours” was later converted to string type following the breakdown in groups. Since this is a supervised learning strategy, we implemented the techniques to arrive at our results. We used the Anaconda 3 distribution from Jupyter.org. Also, all codes were written in Python 3, and Jupyter will serve as the python shell.

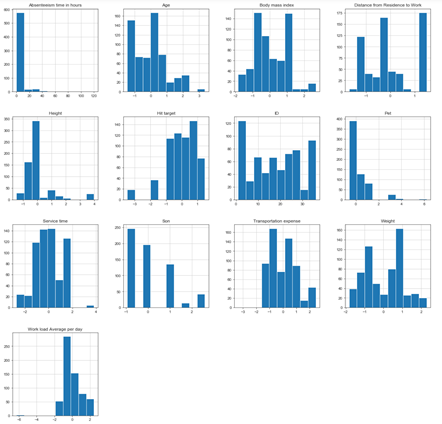
To fulfil task 1:

The target variable was grouped into the following category:

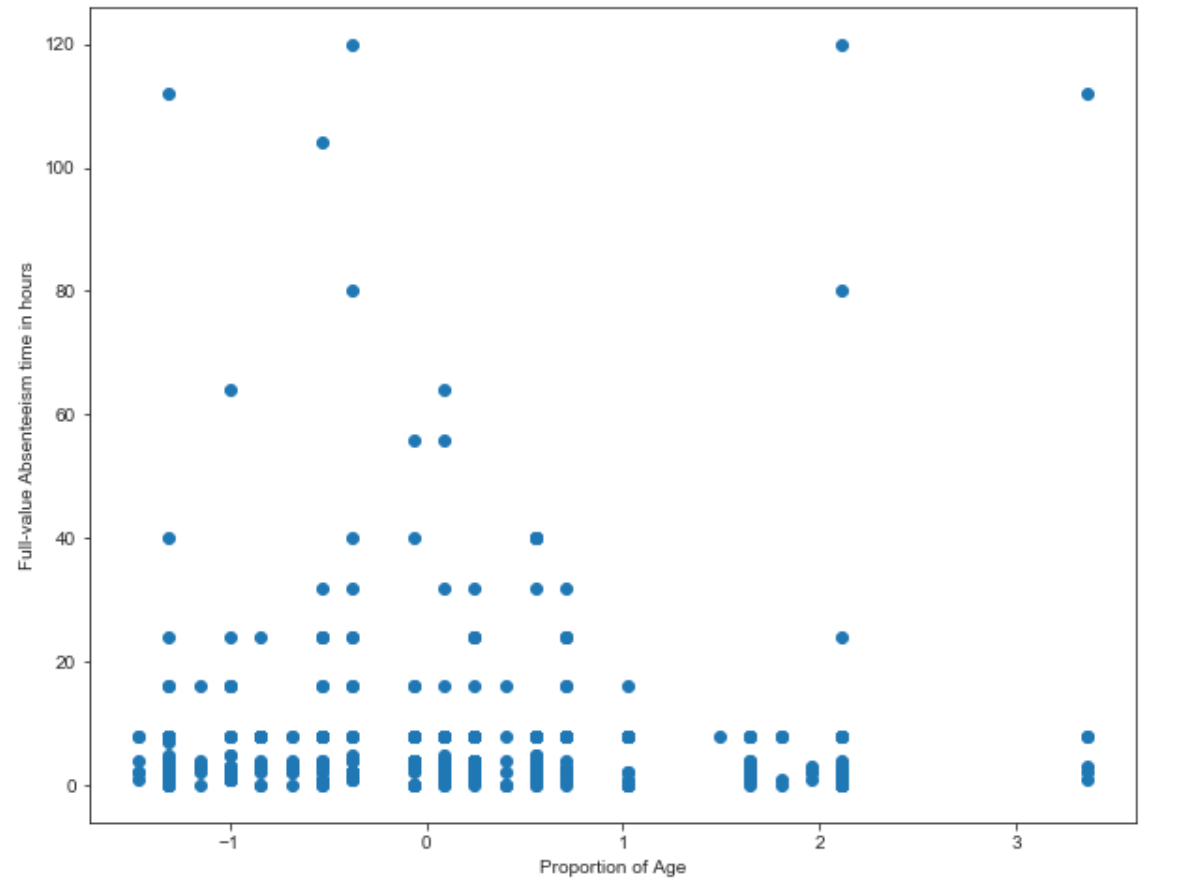


The **preprocessing** of the dataset involved removing the following: removing duplicates: we found 27 duplicates. We also noticed the variable “Reason for absence” had a minimum value of 0 when the minimum value should be 1 to represent Certain infectious and parasitic diseases, so we resolved the issue but resulted in missing values. For resolving missing values, we removed data values less than five percent. The “reason for absence” variable had 5.63 percent, so we used a mode substitution since it was a categorical variable and deleted the missing values in the remaining variables. Also, for the outlier analysis, we detected and dropped 71 data points. Last, we standardized features by removing mean and scaling to unit variance. We removed ID as it did not have a pattern to include as significant features after performing the necessary preprocessing steps.

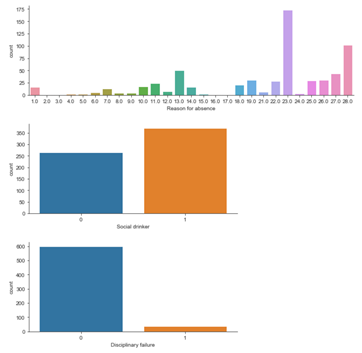
Exploratory Data Analysis:



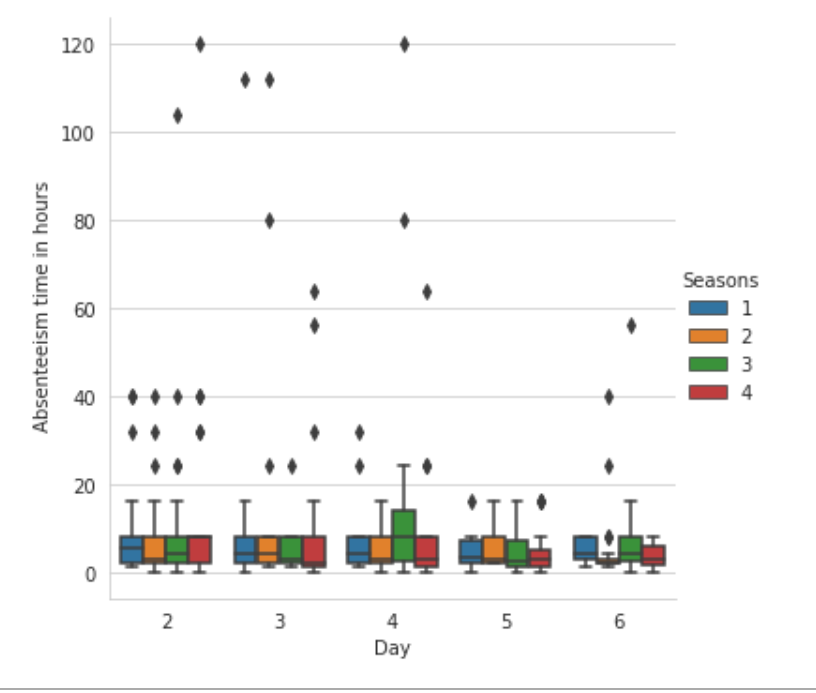
The above image shows a histogram distribution of all continuous variables in the dataset. We scaled the data values to see if we could achieve a gaussian or near Gaussian distribution, but that was not the outcome. The variable weight appears to show two peaks, resulting in a bimodal distribution.



The above image illustrates how an absent employee's time in hours might change depending on the older he or she is. We wanted to see if there is a strong relationship between these two variables. From the plot, we see the points line up, depicting a weak negative correlation. We also noticed some points on the upper right of the graph, which does not fit the pattern of the other data points, indicating some outliers in the data. The outlier in the plot shows that one or more employee(s) of age 30 and above, but the absent time in hours stayed at 120.



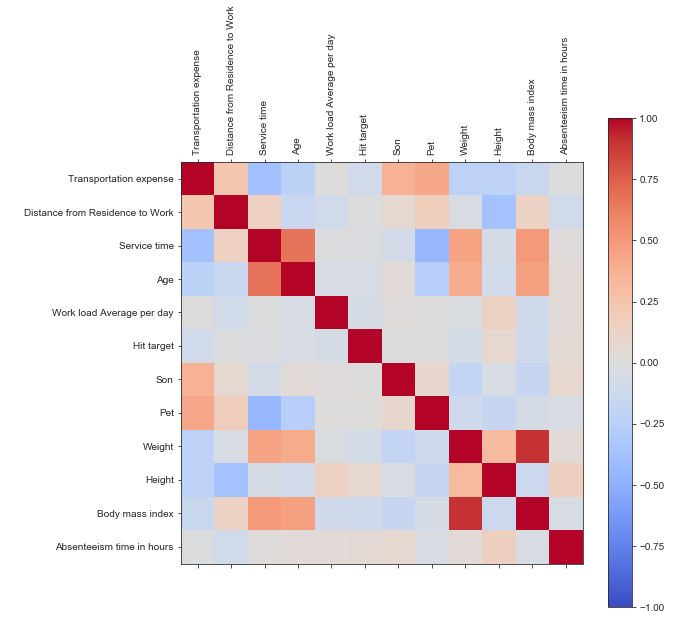
The above image shows a bar graph of the frequency of reasons for absence, disciplinary failure, and social drinker. Here you can see that most of the records are employees who are social drinkers. The most reason for being absent was medical consultation (23).



The above diagram shows a grouped boxplot of the target variable and two independent variables: Seasons and Day. We used this type of graph because we wanted to see the five summaries: min, max, IQR: first and third quartile, and median. From this plot, there is evidence of outliers characterized by points that lie outside the whiskers, indicating a non-symmetrical distribution. We detected and removed some data points found in outlier analysis during the preprocessing part of this project, but we can not entirely remove them, and that is still acceptable. The plot shows that most of the absent time in hours was from the day (4) and in season 3.

**Correlation Plot:**

Below is the correlation plot for the dataset. We used this plot to determine significant features to feed to our machine learning algorithm. We set the threshold cut off at 0.60 to find features index that was greater than the cutoff and dropped two variables (Age and Body mass index).



# **TECHNICAL APPROACH**

The processes involved in this study were data gathering and preprocessing, exploratory analysis, feature selection from correlation, partitioning/splitting, machine learning algorithm, and evaluation.

For splitting/partitioning, we used a 75-25% ratio split for training and testing datasets. The train-test split was essential so we could achieve accurate or near accurate predictions. The training set used 75% or 475 instances out of 634 total instances, whereas the remaining 25% or 159 instances were reserved for testing, with random seed set at 123.

The Decision Tree classification algorithm was used because the decision variable is categorical/discrete (since we transformed the target variable to categorical), excludes unimportant features, and a useful technique for supervised learning. Decision tree, as we know, takes the shape of recursive partitioning, which inserts the data into subsets and then splits even further on the subset nodes until the process ends. We used gini index, which comes default from scikit-learn to measure the quality of a split. We used a confusion matrix to measure our model's accuracy, as you will see in the test and evaluation section.

The Gaussian Naive Bayes algorithm was also used to find the probability of a response/target variable given a selected list of observed features. We attempted this algorithm to see if it will take on the gaussian shape after scaling the distribution. Since our distribution is not gaussian or near gaussian, we expect this algorithm to yield a lower accuracy score than the accuracy for the decision tree.

Last, the support vector machine (SVM) was used because it allows the learning of non-linear models. We used a polynomial and Radial basis function kernel (RBF) as the two types of SVM kennels. However, we decided to keep our data values for categorical features instead of one-hot encoding.

# **TEST AND EVALUATION**

# **REFERENCES**