**Predicting Absenteeism at Workplace**

# Introduction

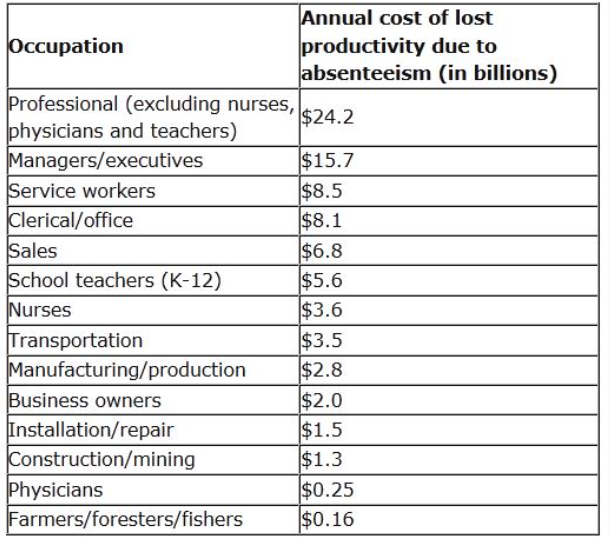
Competitiveness, market share, professional development and personal support to community action, health, culture, education and sport, are linked to a promising new market. Coupled with the development of organizations, the pressure to achieve goals more audacious, employees increasingly overwhelmed, they end up buying some disturbance in the health-related type of labor activity. The objective of this project is to apply some machine learning algorithms in the prediction of absenteeism at work.

Companies, in today's digital world, are collecting data about their workforce. This data enables us to apply statistical methods in order to analyze and come up with recommendations that can be implemented by companies in order to reduce absenteeism at work.

Majority of the work that's been done in the recent past has been to qualitatively address the issues of absenteeism and also identify and solve the root causes behind it. Analytics (with ML techniques) will help us address this issue quantitatively with pin-pointed recommendations around the factors that influence this issue

# Related work

The Gallup-Healthways Well-Being Index surveyed 94,000 workers across 14 major occupations in the U.S. Of the 77% of workers who fit the survey's definition of having a chronic health condition (asthma, cancer, depression, diabetes, heart attack, high blood pressure, high cholesterol or obesity), the total annual costs related to lost productivity totaled $84 billion. According to the survey, the annual costs associated with absenteeism vary by industry, with the greatest loss occurring in professional occupations (excluding nurses, physicians and teachers); the 14 occupations and corresponding costs of lost productivity are shown in Figure 1.



Absenteeism is an especially difficult problem to tackle, because there are both legitimate and poor excuses for missing work - and it can be challenging for employers to effectively monitor, control and reduce absenteeism. Unless a company requires a written excuse from a doctor, for example, it can be difficult to determine if an employee is actually sick when missing work. At the same time, however, it is important for employers to consider the added costs associated with a sick employee who spreads an illness that gets the whole division - or a lot of customers - sick.

To address problems like this, some companies, cities and states have moved toward a mandatory paid sick leave policy, where each employee receives a specified number of days each year to use when sick.

In an effort to reduce absenteeism, some companies offer incentives for going to work, such as earned time off or lotteries for workers who do not have any unexcused absences within a certain period. Other firms might try a more proactive approach, putting policies in place to focus on responses to employee health concerns, including:

a) Physical health

b) Psychological health

c) Work-home balance

d) Environmental health

e) Economic health

Absenteeism costs U.S. companies billions of dollars each year in lost productivity, wages, poor quality of goods/services and excess management time. In addition, the employees who do show up to work are often burdened with extra duties and responsibilities to fill in for absent employees, which can lead to feelings of frustration and a decline in morale.[[1]](#endnote-1)

# Data Set

This study was carried out on a dataset containing information about the employees of a courier company in Brazil and the reasons relating to the absenteeism from work between July 2007 to July 2010. The dataset has 666 rows of data and 21 distinctive features. These features range from reason for absence, age, Body Mass Index (BMI) to service time, number of children and number of pets among others. The final output is given as the number of hours an employee was absent for each instance of their absence.

**Attribute Information:**

1. Individual identification (ID)

2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioral disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

ANLY 530 Late Summer 2018

Course Project Instructions

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services.

. And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

17. Pet (number of pet)

18. Weight

19. Height

20. Body mass index

21. Absenteeism time in hours (target)

**Sample Data Values**

$ ID : int 11 36 3 7 11 3 10 20 14 1 ...

$ Reason.for.absence : int 26 0 23 7 23 23 22 23 19 22 ...

$ Month.of.absence : int 7 7 7 7 7 7 7 7 7 7 ...

$ Day.of.the.week : int 3 3 4 5 5 6 6 6 2 2 ...

$ Seasons : int 1 1 1 1 1 1 1 1 1 1 ...

$ Transportation.expense : int 289 118 179 279 289 179 361 260 155 235 ...

$ Distance.from.Residence.to.Work: int 36 13 51 5 36 51 52 50 12 11 ...

$ Service.time : int 13 18 18 14 13 18 3 11 14 14 ...

$ Age : int 33 50 38 39 33 38 28 36 34 37 ...

$ Work.load.Average.day : Factor w/ 36 levels "0","12","205,917",..: 8 8 8 8 8 8 8 8 8 8 ...

$ Hit.target : int 97 97 97 97 97 97 97 97 97 97 ...

$ Disciplinary.failure : int 0 1 0 0 0 0 0 0 0 0 ...

$ Education : int 1 1 1 1 1 1 1 1 1 3 ...

$ Son : int 2 1 0 2 2 0 1 4 2 1 ...

$ Social.drinker : int 1 1 1 1 1 1 1 1 1 0 ...

$ Social.smoker : int 0 0 0 1 0 0 0 0 0 0 ...

$ Pet : int 1 0 0 0 1 0 4 0 0 1 ...

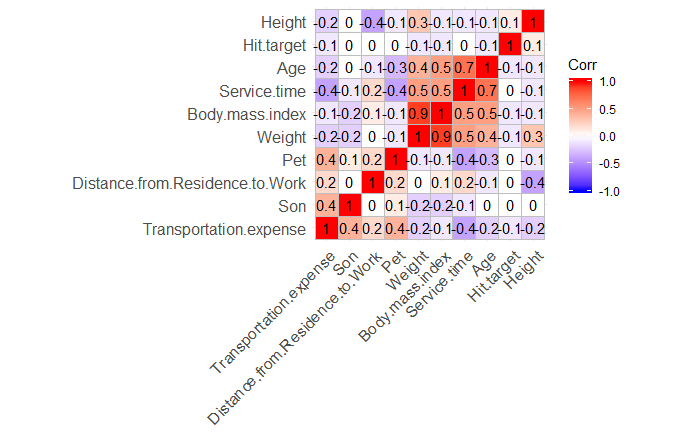
$ Weight : int 90 98 89 68 90 89 80 65 95 88 ...

$ Height : int 172 178 170 168 172 170 172 168 196 172 ...

$ Body.mass.index : int 30 31 31 24 30 31 27 23 25 29 ...

$ Absenteeism.time.in.hours : int 4 0 2 4 2 2 8 4 40 8 ...

1. Data pre-processing:
   1. Cleanup string value in Age
   2. Omit records with NA (3 of the 666)
   3. Created a new variable for grouping “Absenteeism Time in Hours” called “Absenteeism Group”
   4. Make the following variables as factors:
      1. Reason for absence
      2. Month of absence
      3. Day of the week
      4. Seasons
      5. Disciplinary Failure
      6. Education
      7. Social Drinker
      8. Social Smoker
      9. Absenteeism Group
   5. Correlation Matrix for Numerical Variables:



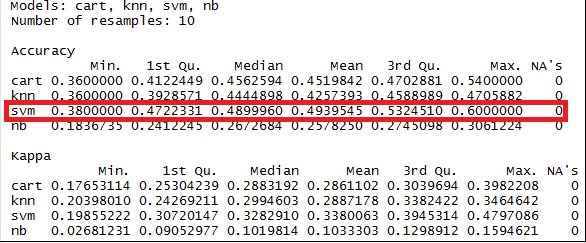
* 1. Eliminated the following variables:
     1. ID
     2. Work Load Average Day
     3. Absenteeism Time in Hours
     4. Body Mass Index

1. Creating training (500 rows;75%) and test (163 rows;25%) data set

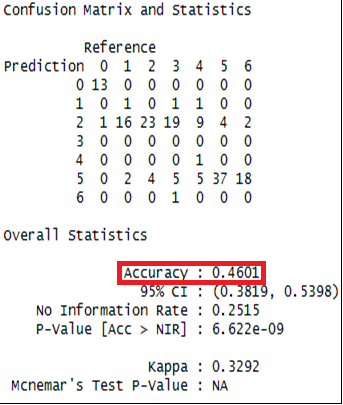
# Technical Approach

Following approach was used to arrive at the final model:

1. Run algorithms with 10-fold cross validation
2. Pre-process the data by centering it and scaling it
3. Run four different models on the training data set:
   1. Classification and Regression Trees
   2. K-Nearest Neighbors
   3. Support Vector Machines
   4. Naïve Bayes
4. Choose the most accurate model among the four



1. Predict the test data set based on the model selected



# Test and Evaluation

Based on the tests that we have ran, we have seen that SVM has the highest accuracy and the test data performs with 46% accuracy using the SVM model. Based on the exploratory data analysis, we have seen that “Reason for Absence” has two high repeating values: “Medical Consultation” and “Dental Consultation”.

The tests that we have run can easily be replicated by following the steps mentioned in this document.

# 6. Summary

The objective of the project was to identify trends in absenteeism and use the information to improve productivity and plan efficient utilization of Human Resources. To do this we build a Machine Learning model & use exploratory data analysis to predict trends in workplace absenteeism. A total of four different algorithms were built to select the best model with the highest prediction accuracy. Support vector machine (SVM) with an accuracy of 46% was the best when compared to k-nearest neighbor (KNN), Naïve Bayes (NB) and Classification & Regression Tree (CART). Dental consultation and medical consultation were identified as the two most important contributing reasons for absenteeism.

There we conclude that the organizations need to focus more on wellness programs, workshops which focus on improving overall employee fitness and health levels. Healthier employees are happier and tend to be more engaged at workplace resulting in better productivity.

7. References

Investopedia, “The Causes and Costs of Absenteeism In Workplace,” Forbes, 10 July 2013. [Online]. Available: <https://www.forbes.com/sites/investopedia/2013/07/10/the-causes-and-costs-of-absenteeism-in-the-workplace/#266c38d73eb6>

1. Investopedia, “The Causes And Costs of Absenteeism In Workplace,” Forbes, 10 July 2013. [Online]. Available: https://www.forbes.com/sites/investopedia/2013/07/10/the-causes-and-costs-of-absenteeism-in-the-workplace/#266c38d73eb6 [↑](#endnote-ref-1)