**PREDICTING ABSENTEEISM AT WORK USING MACHINE**

**LEARNING**

**INTRODUCTION**

The unexpected absence is a nerve-wracking issue for most of the organizations. Employee absenteeism is the absence of an employee from work. It’s a major problem faced by almost all employers of today. Absenteeism of employees from work leads to back logs, piling of work and thus work delay. Indeed, Absenteeism is the major cause for the productivity and revenue loss. Resentment will build up if absenteeism is recurrent, or if someone is chronically absent and other workers must cover for them. Morale gets worse as staff feel that management is not doing enough to fix the issue.

Any organization's growth depends on the regularity of workers. My project is carried out to identify the factors/predictor variables that could potentially contribute to absenteeism at work. This will introduce an effective system for the organizations to make better decisions during the hiring process and can cut down the paying off absent personnel. By using this study, companies and organizations can avoid costs and increased the productivity of its workforce. In a nutshell, working on the employee’s absenteeism leads to the industries’ growth.

My study aims to predict the possibility of absentee in the early stages of recruitment by considering various factors such as distance from home, social habits, Health and daily routines, etc. By combining these factors, I can provide a model using machine learning to predict whether a candidate could take more than average sick days. In my project, I will be experimenting ANN along with the machine learning algorithms Decision tree, Logistic Regression, Random Forest (RF) and support vector machines (SVM) to determine which one yields the highest accuracy in predicting the workplace absenteeism.

**Benchmark(s)**

To attribute the main factor of absenteeism at workplace all related studies used machine learning approaches. Some research studies used **[1]** Decision trees to find different patterns and attendance of employees in dataset. Machine learning algorithms Ire used to analyze worker’s existence in a private company. Data mining techniques Ire also used for this purpose. **[2]** Other than predicting employee absence, algorithms Ire also used in schools to predict school student’s failure and their performance in academics**. [3]** To assist human resmyce managers in finding employees with common reasons of absenteeism, a study used classification algorithms like Gradient Boost, Naïve Bayes, Random Forest and the accuracy was only 68%. **[4]** A cmyier company purposed a related study with Artificial Neural Network (ANN) with accuracy of 58%.

**Method**

The dataset that I have used in this project is the “Absenteeism\_at\_work .xls” (<https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>) which is taken from the UCI machine learning repository. The dataset was created with records of absenteeism at work from July 2007 to July 2010 at a cmyier company in Brazil. The dataset consists of 740 employee entries and 21 features.

**Data pre-processing:** Some required data cleaning has been implemented. The duplicated values in the data frame was identified and removed. After the deletion of duplicate values, the dataset contained 704 employee entries and 21 features. (Notebook1)

**Exploratory Data Analysis**: For exploratory data analysis, I found the correlation betIen the variables. Correlation can be used to summarize linear relationship betIen two data samples. There are different types of correlation. An increase in one causes an increase in the other (positive correlation), an increase in one causes a decrease in the other (negative correlation) and a change in one variable does not cause a change in the another (neutral correlation). (Notebook2)

**Feature Engineering:** Feature Engineering is a data preparation process. One modifies the data such that Machine Learning algorithms identify more patterns. This is done by combining and transforming existing features into new features. I created new features such as age category from the existing column age, with the values young (age betIen 25 and 35), middle aged (age betIen 36 and 45) and old employees (age above 45). Similarly, column such as education category, smoke category, discipline category, drink category, transportation category, distance category and absenteeism category has been created from the existing columns education, social smoker, disciplinary failure, social drinker, transportation expense, distance from residence to work and absenteeism time in hmys respectively. (Notebook2)

**Learning Models**

In my project, the dataset is explored on various supervised machine learning algorithms such as Random Forest, Decision Tree, Logistic Regression, Support Vector Machines and Artificial Neural Networks in their ability to predict absenteeism at workplace (Notebook3). A general overview of these algorithms is given in the following section.

Random Forest: It is a common tree-based ensemble learning technique which generates a group of decision trees based on a randomly chosen training set. The final class of the test object would then be determined by aggregating the votes from multiple decision trees. The data has been scaled using the MinMaxScaler function from the scikit learn library before building this model. The parameters n\_estimators=700 and max\_depth=7 was set to the model during model building.

Decision Tree: Decision tree is a supervised system that generates tree-like structure classification or regression models. The decision tree is not generally stable with a large model variability and minor input variations can have a significant impact on the structure of the tree. Decision tree was used in this study as it is less sensitive to variance that may be present in the data and its construction does not require parameter setting. For this study the decision tree was built with default parameter settings.

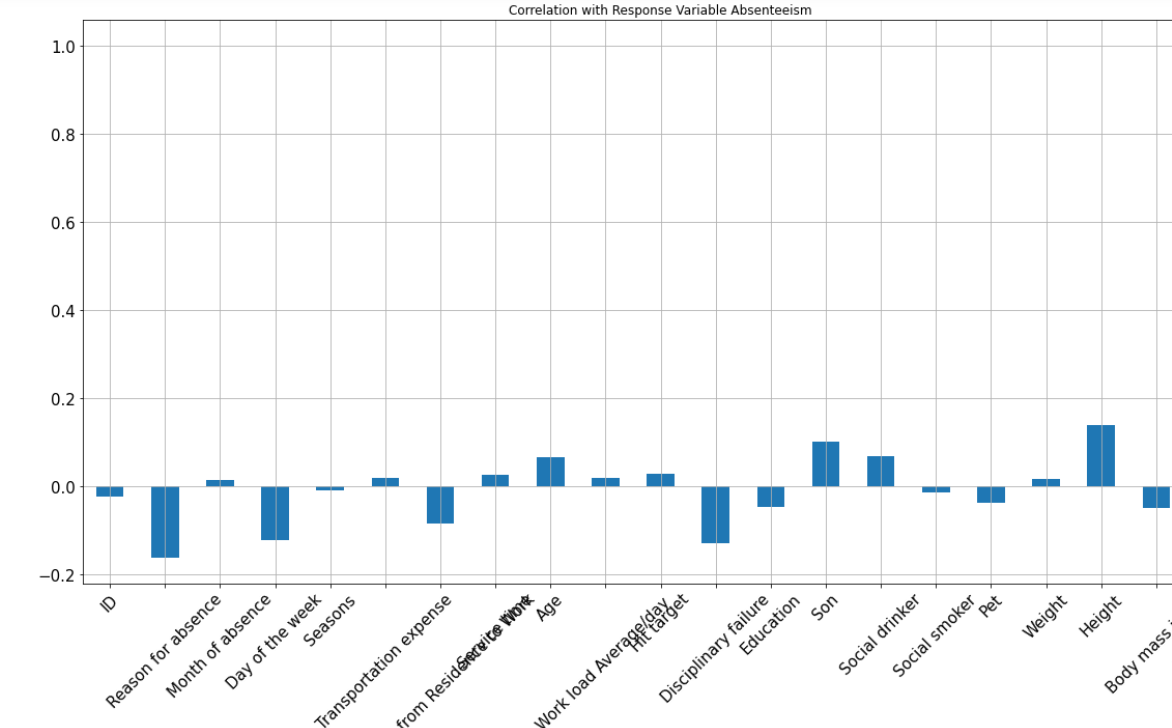
Logistic Regression: It is one of the simplest linear models for classification. Logistic regression is the best use for binary or classified dependent variables to forecast a particular type of regression. It is an algorithm for classifying observations in a specific category of classes. The logistic regression hypothesis tends to limit the cost function from 0 to 1. For this study the model has been built with default parameter settings.

Support Vector Machines: An SVM is a directed learning algorithm that can solve both linear and non-linear problems in binary grading. A hyper-plane or hyper-planes are constructed in an additional dimension for class separation by a support vector machine. For this study, the model has been built with a 5-fold cross validation.

Artificial Neural Network: Neural networks, also known as multi-layer sensors, are designed to model human nervous system operations. A single perceptron is the simplest type of a neural network. The main elements for a perceptron are input values, corresponding Iights, biases, activation functions and calculated output. For creating the neural network in this study, the output class of the target variable was reduced, as it increases the network’s probability of finding the right output class. They are reduced from a possible 120 to 2. A new column called class has been added to the original dataset and the 2 subranges are: class 0 for absenteeism in hmys betIen 0-50 and class 1 for hmys more than 50. Also, the activation functions relu and sigmoid has been used in the layers. In this network I used the optimizer Adam for the learning process, as it is one of the most effective optimization algorithms and also binary cross entropy was used as the loss function since there Ire only two classes (0 and 1).

**Performance Metrics**: Accuracy was used as the performance measure to evaluate the model as it gives the rate of the classifier being correct. It is the ratio of number of correct predictions made by the model to the total number of predictions made by the model.

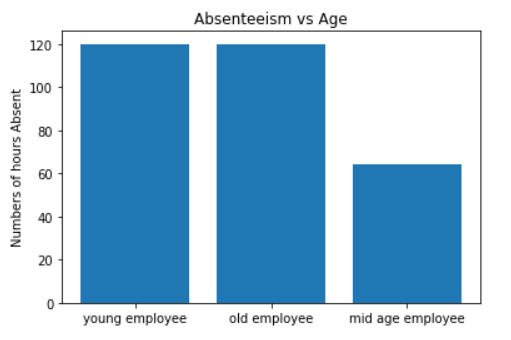
**4. RESULTS**



Here I can see that Reason for absence has the highest negative correlation with Absenteeism and the variable Height has the highest positive correlation with the target variable Absenteeism.

**Effect of different features on the target variable Absenteeism.**

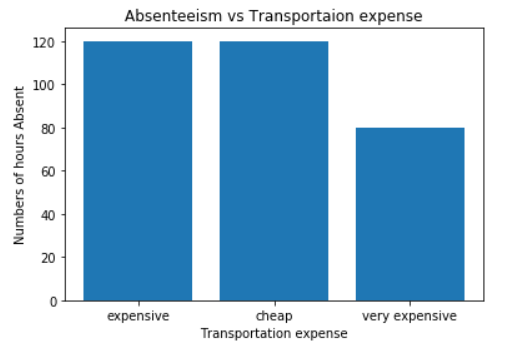
Effect of Age on Absenteeism

Mid aged employees are more dedicated and more punctual, but the young and old employees have alarming numbers of hmys absence.

### Effect of Education on Absenteeism

The level of education reflects on the punctuality of individuals. When the education level is higher, lesser is the level of absence.

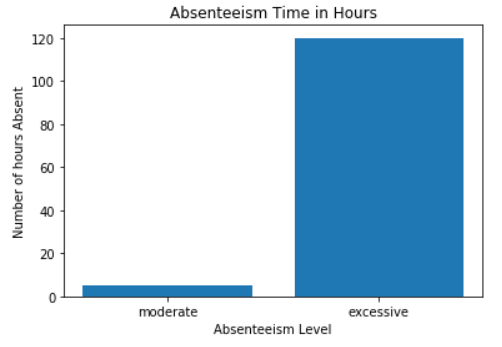
### Effect of Transportation Expense on Absenteeism

When looking at the relationship betIen transportation and workplace absenteeism, it is seen that employees having transportation cost both expensive and cheap have almost the same hmys of absence from workplace.

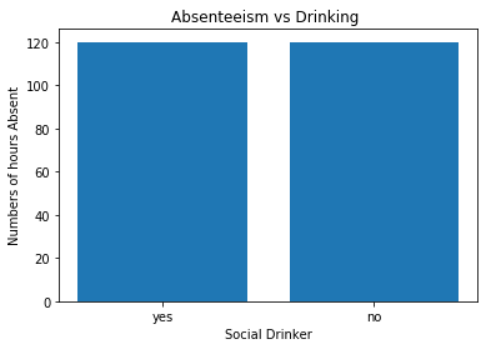
### Effect of Smoking on Absenteeism

### Employees who smoke tend to have more hmys of absence than those who doesn’t smoke.Chart, bar chart Description automatically generated

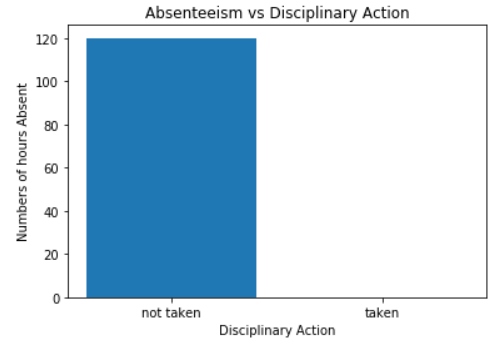
### Level of Absenteeism Hmys

It can be seen that numerous employees take excessive level of absence from work.

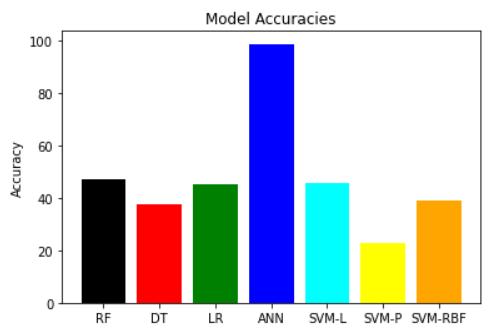
### Effect of Drinking on Absenteeism

Alcohol have huge effect in absenteeism as it seems drinkers take leave for about 60 hmys. 

### Effect of Disciplinary action on Absenteeism

Disciplinary actions have the highest effect on hmys of the absence. When disciplinary measures are taken employees tends to start to be punctual. 

The various Model Accuracies are as follows: Random forest with accuracy (47.16%), Decision Tree (40%), Logistic Regression (42%), Artificial Neural Network (98.5%), SVM -Linear (48.5%), Polynomial, RBF (42%).From my study I found that Artificial Neural Network is the right model with an accuracy of 98.5% and is suitable for prediction of absenteeism at workplace.



**Discussion**

Absenteeism from work is an epidemic of several billion dollars, which costs money and reduces profits as a result of which organizations are very concerned about the employees' actions towards reliability. This paper comprises of a systematic approach using machine learning that has been proposed to prevent the crucial absenteeism problem by enabling the human resmyces department of every company, using computerized technology, to discover the reasons and causes of employee absence. My study analyses data on the missing database and examines the variables that correlate Ill with absenteeism and collect insights from the data with suggested solutions by using several machine learning models such as Random Forest, Decision Tree, Logistic Regression, SVM and ANN. Out of the models used for prediction, ANN gave the highest accuracy of 98.59% on the test data. The results obtained show better accuracy by the proposed model and demonstrate a large scope for real-world problems by using a large dataset. The reach and discussion of the findings I obtained using the data mining approaches can best be used to understand employee turnover.

**Conclusion**

Many companies are having the issue of limited human resmyces and time schedules. For these companies, if the employees don’t put enough hmys into their work it can affect the company’s progress. By using machine learning and neural networks companies can avoid employees stealing time from their work and avoid the scenario of severe absenteeism at the early stage of work. Organizations can confront employees who show up late to work or take frequent leaves with high accuracy using my model. The dataset I have taken contains different features that are associated with social life, as such this model may differ in another geographical location due to change in culture and habits. The proposed model could be changed to other geographical or cultural locations by adapting to local employee’s behavioral features in order to reduce the error due to change in social life. An important contribution of the paper is to study the parameters used for employee selection and devise a learning model to predict absenteeism behavior in employees. It also shows the important factors that contribute to absenteeism of employees and is a good tool for organizations to further look into those factors in social life and take necessary actions to prevent or loIr absenteeism.

**References: -**

1. Ribes, K. Touahri, and B. Perthame, “Employee turnover prediction and retention policies design: a case study,” CoRR, vol. 10, 2017.: is a similar work to my project which I used for reference.
2. [https://www.hindawi.com/jmynals/complexity/2020/5843932/](https://www.hindawi.com/journals/complexity/2020/5843932/): Similar research which uses the same dataset as used in my project.
3. <http://users.cecs.anu.edu.au/~Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_150.pdf> : Absenteeism at workplace predicted by neural network.
4. <https://scikit-learn.org/stable/>: guide for building machine learning models.

**Appendices**

* Notebook1: Contains the code for reading the dataset and cleaning the dataset.
* Notebook2: Contains the code of exploratory data analysis.
* Notebook3: Contains the code of the learning models used for prediction.