BT2101 Decision Making Methods and Tools

SEMESTER II 2019-2020

Assessment of Machine Learning models on predicting Absenteeism from work

[redacted]

Andy Low Wei Liang [redacted]

[redacted]

TABLE OF CONTENTS

01 Background information and data modeling problem

Absenteeism at work is described as a habitual and frequent absence from work. Absenteeism at work is a serious issue that impacts the profit of companies (Grobler, Warnich, Carrell, Elbert, and Hatfield, 2006). By analysing the variables correlated with absenteeism, insights can be drawn about the characteristics of absenteeism. Models will help companies deploy manpower more efficiently and effectively, creating better workflow and/or layoff employees at a high risk of being absent for work. It can also help companies during recruitment, providing a better understanding of employees who are more likely to be absent from work.

1.1 Hypothesis

Before analysing the dataset, we came up with the following hypotheses:

- All 19 of the characteristics used will have predictive power for our model in finding out how they affect absenteeism in employees.

- The main and preferred evaluation model that we would be using for this dataset would be linear regression, due to the large number of characteristics and the feasibility of linear regression on such data types.

- Looking at disciplinary failure with respect to Absenteeism time in hours, it has the highest negative correlation. Hence, we hypothesize that disciplinary failure will be the strongest predictor of Absenteeism time in hours.

To evaluate our dataset, the following models were considered, and we would be doing a deeper analysis on selected models we deem fit.

1.2 Possible models

Model	Pros	Cons
Support Vector Machine (SVM)	• Effective for high-dimensional space • Kernel selection for non-linear correlation • Robust even with bias	• Black Box • Long and inefficient • Features may be dependent or highly correlated
Decision Tree	• Simple and easy to interpret	• Not very accurate
Neural Network	• Flexible model, able to use it with datasets that are large	• Due to it being a blackbox, it's explanatory is low
Naive Bayes	• Easy to comprehend • No distribution required	• Assumes features independence, almost impossible in the real world

02 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is an approach to analyse datasets, so as to summarise their main characteristics with the help of graphical methods. This maximises our data insights, testing our underlying assumptions and detects outliers and anomalies.

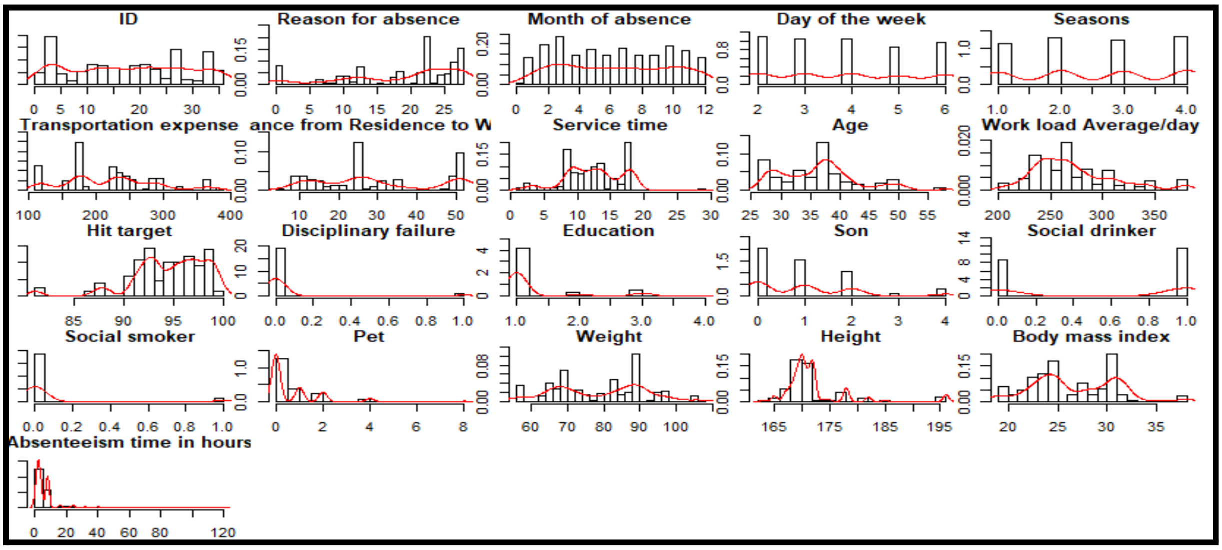
2.1 Data Overview

The Absenteeism at work dataset consists of 740 observations and 20 characteristics.Out of which Absenteeism time in hours is the dependent variable with 19 independent variables. Using sapply, we found that the data types of our variables are all integers.

2.2 Inconsistent values within the dataset

Legend for variable Month of absence: 1 = January; 2 = February; 3 = March; 4 = April; 5 = May; 6 = June; 7 = July; 8 = August; 9 = September; 10 = October; 11 = November; 12 = December. However, the range of values found were ranging from [0.12]. This means that there is an additional value of 0 that is unaccounted for.

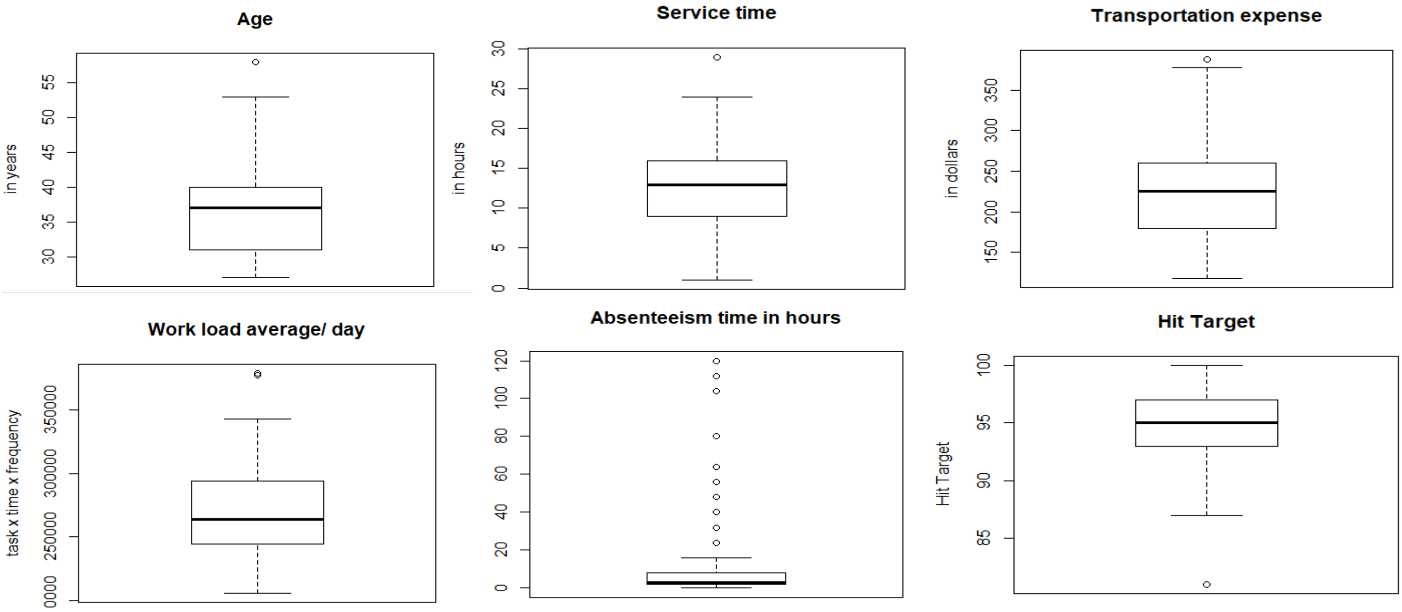
2.3 Checking for categorical data



The density plot for each attribute, based on continuity demonstrates if it is categorical. An example of which is the top right hand corner, “Seasons” which is categorical as observed by the discontinuity. The categorical data hence include, Month of Absence, Day of absence, Seasons, Disciplinary failure, Education, Social drinker, Social smoker and Reason for absence.

2.4 Distribution of continuous variables

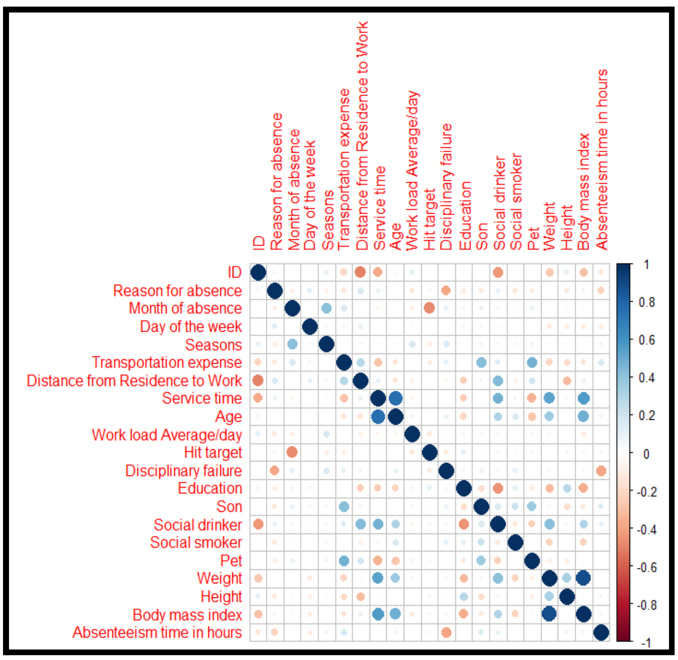
The box plots show the distribution of the continuous variables based on the maximum and minimum values. It shows the median, first quartile and the third quartile.



Looking at the boxplot for “Age”, “Service time”, “Transportation expense”, “Work load average/day”, and “Hit target”, we can see that these dependent variables have outliers. However, all these outliers are not necessarily wrong . We will have to run analysis before determining whether to keep or remove the outliers.

2.5 Correlation of variables

Looking at the correlation graph, we can see that Weight and Body mass index are highly correlated with each other. This is a cause of concern because it could lead to multicollinearity which would in turn affect the accuracy of our model



2.6 Dataset Evaluation

2.6.1. Skewed classes

As we can see, only approximately 6.08% of workers have never been absent as compared to 93.92% of workers with more than 0 hours of absenteeism.
Number of workers with 0 absenteeism hours: 45 (minority class)
Number of workers with >0 absenteeism hours: 695 (majority class)
Splitting the data as training and test could result in overpopulation by the majority class as compared to the minority. This would, in turn, affect the accuracy calculated.

2.6.2. Lack of data

There is a lack of examples in the dataset, there are a total of 36 workers and 740 observations. For algorithms which require more data, this will present implications. Hence, we should keep to simpler algorithms.

2.6.3. Too many features

For such a small dataset , we have a staggeringly high number of features, 20. This presents a Curse of Dimensionality as we are required to increase the number of examples we have exponentially for each feature added. Hence, we will conduct feature selection to select the strongest predictors of our dependent variable, Absenteeism hours.

03 Data Pre-Processing

3.1 Filtering Entries not consistent with data source

In the Month of absence column, there are values “0” which do not correspond to the legend for the table. Hence we chose to remove it as the Month of absence could skew the correlation between the month of absence and Absenteeism hours.

3.2 Removing variables with high/perfect multicollinearity

Since there is a high collinearity between Body Mass Index(BMI) and Weight, we decide to remove BMI from our dataset.

3.3 Removing the outliers

Since all attributes for the data points do not follow a normal distribution and some variables display covariance with one another. The Mahalanobis Distance can be used to identify the outliers. With the Mahalanobis Distance designs with Gaussian distribution, it is not necessary to have a joint multivariate normal distribution and it will still improve the objective functions to a greater extent in its variables/ attributes.

3.4 Conclusion

The dataset now consists of 692 observations with 19 characteristics.

04 Feature Selection

4.1 Definition

Feature selection is a necessary process in machine learning, modeling and statistics where selecting a subset of the most important features to the dependent variable is done, be it automatically or manually. This also means that the irrelevant features that have no predictive power would be taken out from the model which increases the accuracy of the results.

Feature selection increases accuracy, reduces overfitting, and speeds up the time needed for the algorithm to run our model. All these are beneficial to us and hence it is important to select the right features. Occam's Razor states "the simplest solution is always the best".

4.2 Feature Selection using Filter Methods

4.2.1 Correlation:

Correlation measures the degree of association between two numeric variables. Features with a high correlation with the dependent variable will be selected and included in our model.

With reference to the correlation matrix in section 2.5, other than the variable "Disciplinary Failure", the other independent variables have a relatively low correlation coefficient with the dependent variable (Absenteeism in hours). Thus, we are not able to conduct feature selection easily with the low correlation coefficients. Instead, a non-linear model may be more suitable for this dataset.

4.2.2 Hypothesis Testing (t-test and Chi-square Test):

To determine if the independent variables are statistically significant, hypothesis testing is carried out. This will be done using two kinds of tests: t-test and Chi-square test.
The t-test measures the degree of association between continuous independent variables and the dependent variable while the Chi-square test measures the association between two categorical variables and it will be used to test for association between categorical independent variables and the dependent variable. The following describes the null and alternate hypothesis:
a. Null Hypothesis: The independent variable is statistically insignificant
b. Alternate Hypothesis: The independent variable is statistically significant.

The p-values obtained will be used to determine whether we reject the null hypothesis. If it is less than the 5% level of significance, we reject the null hypothesis and conclude that the variable is statistically significant. The following are the p-values we obtained for each independent variable based on the t-test and Chi-square test:

t-test for continuous variables		Chi-square test for categorical variables	
Variable	**p-value**	**Variable**	**p-value**
ID	5.112549e-62	Reason for absence	1.241894e-95
Transportation.expense	5.646325e-298	Month.of.absence	0.0001253705
Distance.from.Residence.to.work	2.076261e-139	~~Day.of.the.week~~	0.113472
Service.time	4.291669e-30	Disciplinary.failure	3.363212e-108
Age	4.979349e-282	~~Education~~	0.8766481
Work.load.Average.day	1.137165e-60	~~Social.smoker~~	0.06750786
Hit.target	0	Social.drinker	0.006704489
Son	2.262669e-26	Seasons	1.200595e-09
Pet	9.314382e-29		
Weight	0		
Height	0		

For the continuous variables, we can observe that all of them have a p-value less than 0.05 so we can conclude that they are all statistically significant. As for the categorical variables, the variables, “Day of the week”, “Education” and “Social smoker” have a p-value greater than 0.05 so we can conclude that they are statistically insignificant. Thus, these variables can be excluded from our model as they do not contribute greatly to the prediction of our dependent variable. On the other hand, the remaining variables which have a p-value less than 0.05 should be included in our model as they are statistically significant.

In conclusion, the variables to be included in our model is as shown in the table.

4.2.3 Information Gain:

Information gain tells us how much information is given by the independent variable on the dependent variable.

Features are selected based on their information gain score and features with a non-zero information gain score are selected to be included in the model.

```
                                 attr_importance
ID                                    0.00000000
Reason.for.absence                    0.25352752
Month.of.absence                      0.00000000
Day.of.the.week                       0.00000000
Seasons                               0.00000000
Transportation.expense                0.03643298
Distance.from.Residence.to.Work       0.00000000
Service.time                          0.00000000
Age                                   0.00000000
Work.load.Average.day                 0.00000000
Hit.target                            0.00000000
Disciplinary.failure                  0.08700688
Education                             0.00000000
Son                                   0.00000000
Social.drinker                        0.00000000
Social.smoker                         0.00000000
Pet                                   0.00000000
Weight                                0.00000000
Height                                0.00000000
```

4.3 Feature Selection using Wrapper Methods

4.3.1 Stepwise Forward and Backward Selection:

This feature selection method helps us build a model by adding and removing certain characteristics. The following are different methods of stepwise regression:

a. Stepwise selection - A mixture of both forward and backward selection. At each iteration, the algorithm decides whether a variable is added or removed from the model.
b. Forward selection - The model starts off empty and then variables are progressively added to it.
c. Backward selection - The model starts off with all of the variables and then the least significant ones are removed from the model.

Output:
The following variables were selected from the stepwise regression selection.

```
> print(vars_step)
[1] "(Intercept)"          "Height"                "Reason.for.absence"
[4] "Disciplinary.failure" "Son"                   "Day.of.the.week"
[7] "Social.drinker"       "Seasons"
> print(vars_forward)
[1] "(Intercept)"          "Height"                "Reason.for.absence"
[4] "Disciplinary.failure" "Son"                   "Day.of.the.week"
[7] "Social.drinker"       "Seasons"
> print(vars_backward)
[1] "(Intercept)"          "Reason.for.absence"    "Day.of.the.week"
[4] "Disciplinary.failure" "Son"                   "Social.drinker"
[7] "Height"
```

4.3.2 Recursive Feature Elimination (RFE) Method:

Progressively, a model consisting of all variables drops the least significant feature, leaving behind the specified number of features. The optimal number of features in the model can be identified using cross-validation.

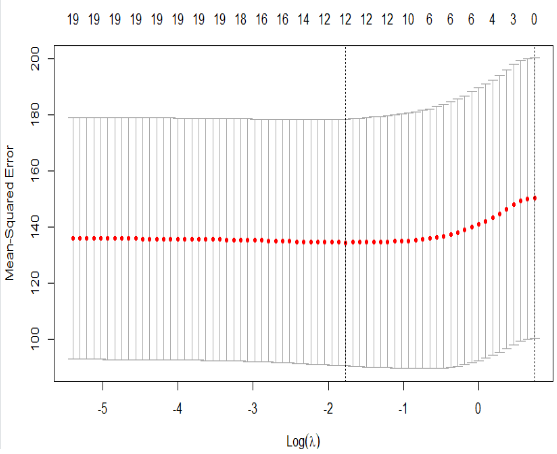
```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables  RMSE Rsquared   MAE RMSESD RsquaredSD  MAESD Selected
         1 9.985   0.1484 4.563  4.371     0.1104 1.1493
        19 9.672   0.2470 4.495  3.554     0.1297 0.9761        *

The top 5 variables (out of 19):
   Reason.for.absence, Disciplinary.failure, Height, Service.time, Seasons
```

The results of the RFE shows that the highest accuracy rate consists of the following 5 variables: Disciplinary.failure, Reason.for.absence,Height, Service.time and Seasons.

4.4 Feature Selection using Embedded Methods

4.4.1 Least Absolute Shrinkage and Selection Operator (Lasso):

```
(Intercept)                    -45.34349
ID                              -0.04293
Reason.for.absence              -0.39949
Month.of.absence                 0.11082
Day.of.the.week                 -0.70615
Seasons                          0.30524
Transportation.expense           0.00238
Distance.from.Residence.to.Work -0.02320
Service.time                     0.00000
Age                             -0.00737
Work.load.Average.day            0.03297
Hit.target                       0.15513
Disciplinary.failure           -15.48772
Education                       -0.42150
Son                              1.01658
Social.drinker                   1.63961
Social.smoker                    0.00000
Pet                             -0.35765
Weight                           0.00000
Height                           0.26929
```



This feature selection technique conducts regularisation whereby it shrinks the coefficients of the regression model as part of the penalisation. For feature selection, the variables which remain after the shrinkage process are included in the model.

We are unable to make inferences about the importance of the coefficients as the data has only been scaled individually and not scaled to have a common mean and standard deviation. Since our variables have different means and standard deviation, variables with larger averages will tend to have larger absolute coefficients.

Any variable with a coefficient of zero would be dropped from the model, because it shows that it has no predictive power. The following variables have a coefficient of zero and would hence be dropped for our model.

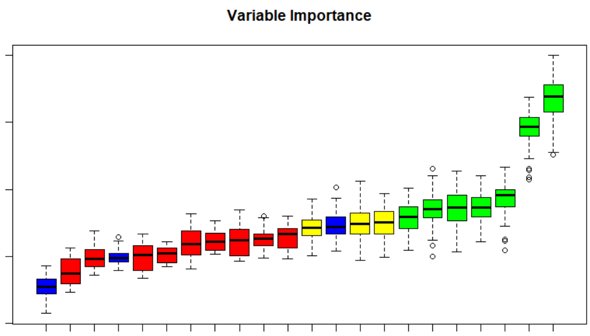
1. Service.time
2. Social.smoker
3. Weight

The remaining variables would then be considered in our model.

4.4.2 Boruta:

Boruta algorithm is another feature selection algorithm. Boruta is a wrapper built around the random forest classification algorithm.

At every iteration, Boruta runs and compares between a real feature and its shadow feature, whether or not the real feature has a higher importance. (i.e. comparing the Z score between the 2, whether the Z score of the real feature > max Z score of its shadow). The model also removes features which are deemed not significant. The algorithm completes when the various features are either confirmed or rejected.




```
                        meanImp  decision
Reason.for.absence     11.714432 Confirmed
Disciplinary.failure    9.566563 Confirmed
Service.time            4.352393 Confirmed
Height                  3.609344 Confirmed
Age                     3.582582 Confirmed
Transportation.expense  3.572832 Confirmed
```

From the results as shown in the figure above , we can see that the Boruta model confirmed the following 6 variables: Reasons for absence, Disciplinary Failure, Height , Age and Transportation.expense.

4.4.3 Random Forest:

This feature selection technique builds a random forest model and then provides a list of significant variables.

```
Random Forest

553 samples
 19 predictor

Pre-processing: scaled (19)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 497, 497, 498, 499, 497, 498, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    9.340947  0.2678349  4.536926
  10    9.524241  0.2964232  4.621465
  19    9.641721  0.3004031  4.692377
```

```
rf variable importance

                     Overall
Reason.for.absence    100.00
Disciplinary.failure   86.65
```

As seen from the results from the figure above,, we can see that the randomForest achieved an optimal model with the following variables: Reason for Absence and Disciplinary Failure.

4.5 Comparison Between Methods

Type	Method	No. of features selected	Features Selected
Filter	**Correlation**	N.A.	N.A.
	Hypothesis Testing	16	ID, Reason.for.absence, Month.of.absence, Seasons, Transportation.expense, Distance.from.Residence.to.Work, Service.time, Age, Work.load.Average.day, Hit.target, Disciplinary.failure, Son, Social.drinker, Pet, Weight, Height
	Information Gain	3	Transportation.expense, Disciplinary.failure, Reason.for.absence
Wrapper	**Stepwise Regression**	7 7 6	**<u>Both:</u>** Height, Reason.for.absence, Disciplinary.failure, Son, Social.drinker, Day.of.the.week, Seasons **<u>Forward:</u>** Height, Reason.for.absence, Disciplinary.failure, Son, Social.drinker, Day.of.the.week, Seasons **<u>Backward:</u>** Reason.for.absence, Disciplinary.failure, Son, Social.drinker, Day.of.the.week, Height
	Recursive Feature Elimination	5	Reason.for.absence, Disciplinary.failure, Height, Service.Time, Seasons
Embedded	**LASSO**	16	ID, Reason.for.absence, Month.of.absence, Day.of.the.week, Seasons, Transportation.expense, Distance.from.Residence.to.Work, Age, Work.load.Average.day, Hit.target, Disciplinary.failure, Education, Son, Social.drinker, Pet, Height
	Boruta	6	Reason.for.absence, Disciplinary.failure, Service.time, Height, Age, Transportation.expense
	Random Forest	2	Reason.for.absence, Disciplinary.failure

A different set of features can be obtained from each method allowing certain features to be filtered out for consideration. The table above shows the features we have identified from each feature selection method, for future deduction on the most accurate model.

5 Model Selection

5.1 Linear Regression

```
Multiple R-squared:  0.1066,    Adjusted R-squared:  0.07993
```

With the adjusted R-squared of 0.07993 being low, the linear regression model is not recommended as a predictor of Absenteeism.

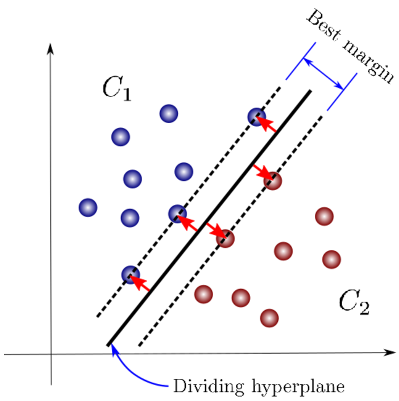
5.1.1 Logit Regression

```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.107e+01  7.993e+05   0.000    1.000
Month.of.absence                2.563e+00  1.042e+04   0.000    1.000
Day.of.the.week                 1.806e+01  6.191e+03   0.003    0.998
Seasons                        -1.118e+00  3.370e+04   0.000    1.000
Transportation.expense          1.009e-01  7.435e+02   0.000    1.000
Distance.from.Residence.to.Work -3.888e-01  3.083e+03   0.000    1.000
Service.time                    1.171e+00  1.180e+04   0.000    1.000
Age                            -1.844e+00  7.144e+03   0.000    1.000
Work.load.Average.day          -6.510e-01  3.077e+03   0.000    1.000
Hit.target                      1.346e-01  8.605e+03   0.000    1.000
Disciplinary.failure           -1.441e+02  1.267e+05  -0.001    0.999
Education                       9.002e+00  3.917e+04   0.000    1.000
Son                             2.546e+00  2.213e+04   0.000    1.000
Social.drinker                  1.023e+01  8.934e+04   0.000    1.000
Social.smoker                   5.249e+00  1.273e+05   0.000    1.000
Pet                            -1.620e-01  2.101e+04   0.000    1.000
Body.mass.index                 8.084e-01  5.918e+03   0.000    1.000
```

Due to the large spread of data points and poor feature selection, resulting in our logit regression falsely showing that all variables are insignificant in predicting Absenteeism.

5.2 Support Vector Machine (SVM)

For this model, we would plot each data item as a point in a n-dimension space (where n = 20 as it is the number of characteristics for our dataset) with the value of every characteristic being a coordinate. We would then conduct classification by finding out what would be the optimal hyper-plane for the selected characteristics.
This would be optimal for our dataset because it is effective in high dimensional spaces (high number of features).



5.2.1 Testing of the Kernels:

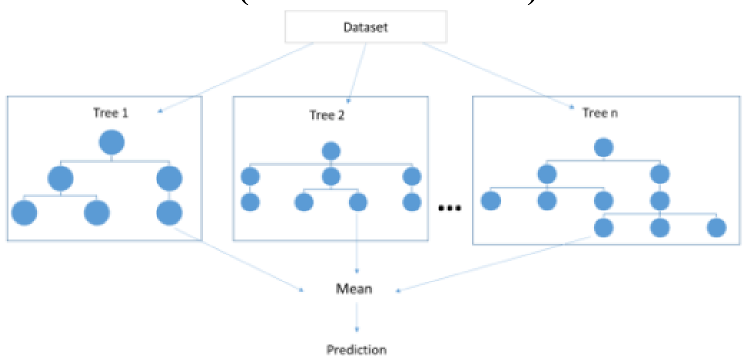
Kernel	Area Under Curve	Test Accuracy
Linear	29.10305%	43.16547%
Polynomial	44.92537%	41.72662%
Radial	40.8874%	44.60432%
Sigmoid	35.40026%	41.00719%

5.2.2 Model Evaluation:

From the results above, we can see the 3 models- Polynomial, Radial and Sigmoid returns a higher AUC. Hence, we will test the accuracy of the 3 models using the variables selected by the different feature selections.

	Train Data Accuracy			**Test Data Accuracy**		
	Polynomial / %	Radial / %	Sigmoid / %	Polynomial / %	Radial / %	Sigmoid / %
Hypothesis Testing	36.57407	40.8874	46.09375	43.16547	44.60432	37.41007
Information Gain	NaN	NaN	77.89855	36.69065	38.1295	33.09353
Step-wise(Forward)	53.10219	36.15288	NaN	38.1295	41.00719	39.56835
Step-wise(Backwards)	NaN	NaN	NaN	41.00719	40.28777	38.84892
RFE	44.89051	11.95652	75.47009	40.28777	41.00719	24.46043
LASSO	44.92537	41.97995	42.75194	42.44604	43.88489	40.28777
Boruta	11.95652	11.95652	48.98148	41.00719	40.28777	30.21583
Random Forest	NaN	NaN	NaN	33.09353	33.09353	38.1295

5.3 Random Forest (Decision Tree Model)

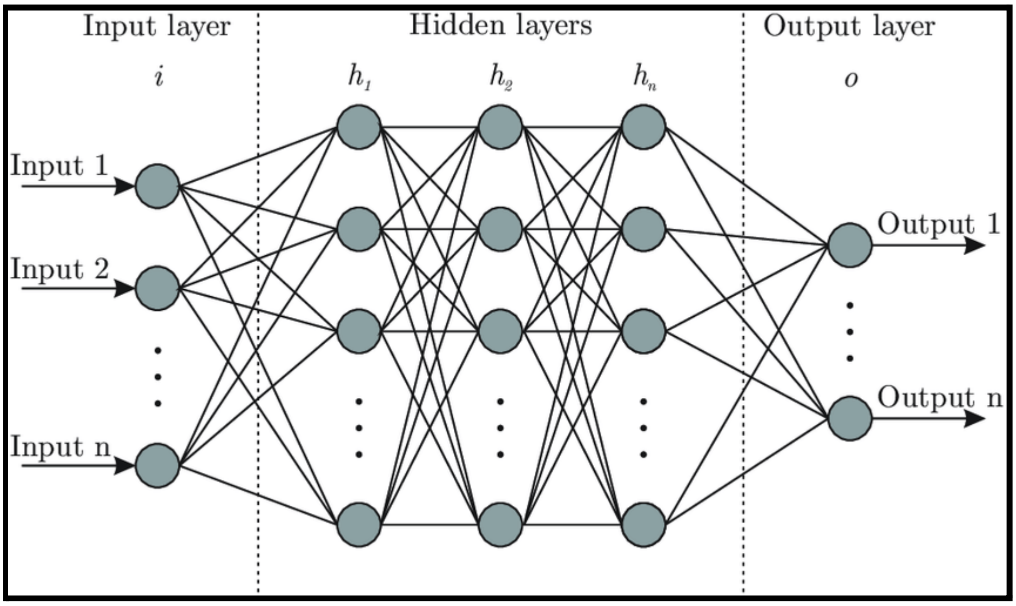


Random Forest consists of many individual decision trees that operate together. Each decision tree represents an independent variable and generates a class prediction, which contributes to a vote in the final prediction. Since decision trees are highly sensitive to the data they are trained on, small changes to the training set can lead to significantly different tree structures. Thus, random forest builds on this by allowing each individual tree to randomly sample from the dataset with replacement (bagging/bootstrap aggregation), resulting in different trees. With reference to section 4.1.1, the low correlation between the independent variables plays a key role in ensuring the accuracy of the random forest classifier. With the low correlation between trees, they are able to protect each other from potential errors that might occur.

5.3.1 Model Evaluation:

	Train Data Accuracy / %	Test Data Accuracy / %
Without Feature Selection	16.72	12.98
Hypothesis Testing	14.01	9.41
Information Gain	16.5	12.9
Step-wise	14.72	21.69
RFE	16.76	12.33
LASSO	14.74	13.46
Boruta	14.59	14.43
Random Forest	12.59	11.21

5.4 Neural Network



Neural networks are the workhorses of deep learning. They are black boxes trying to achieve good predictions. A neural network consists of both input and output neurons which are weighted. The weights will affect the degree of forward propagation that goes through the algorithm. When the back propagation happens, the weights are flexible enough to change and this is when the neural network learns.

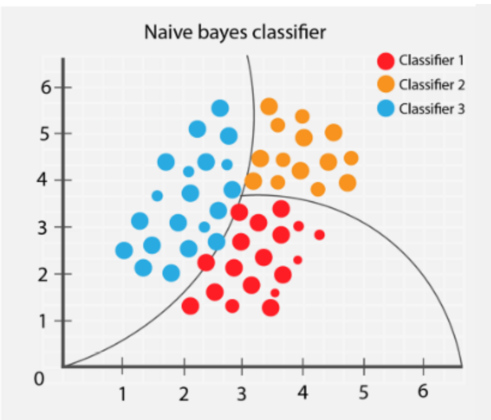
The constant process of forward and backward propagation is conducted iteratively for all data in the training set. The larger the dataset, the more the neural network will learn, and therefore the more accurate the algorithm will be at forecasting outputs.

5.4.1 Model Evaluation:

	Hidden =1		**Hidden =2**		**Hidden =3**	
	Root Mean Square Error(RSME)					
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Without Any Feature Selections	10.33613	15.03089	5.896172	24.79042	6.884369	38.79239
Hypothesis Testing	11.04617	15.88779	9.197887	11.48549	9.898163	15.29223
Information Gain	10.61523	15.08744	10.58664	15.03115	10.41181	16.48617
Step-wise (Backwards)	10.32608	14.68426	9.969812	14.46625	10.22883	14.71477
Step-wise (Forward)	10.29014	14.57731	9.830488	13.91881	10.11034	14.3922
RFE	10.50311	14.82188	10.48698	14.787	10.46755	14.76412
LASSO	10.2098	14.62024	10.05685	14.69654	7.845128	16.9419
Boruta	10.50148	14.86542	10.10994	14.79347	10.04438	14.33146
Random Forest	10.60692	15.06263	10.59927	15.04746	10.60638	15.05652

From the table above, we can see that the Neural Network model (with hidden layer=2) with the variables selected by Hypothesis Testing returns the lowest RMSE with the lowest deviation between the test and training dataset.

5.5 Naive Bayes Classifier



$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Mathematically, the Bayes theorem is represented as P(A|B). The Naive Bayes Classifier belongs to the family of probability classifier, using Bayesian theorem. This method solves classification problems using a probabilistic approach. However, it has a strong assumption that all the variables are independent of one another. This might not be the case for real-life examples. This is also why the model requires much less training data. Even if the assumption does not hold, this method could still prove to be an effective one.

5.5.1 Model Evaluation:

```
Accuracy
     trainAccuracy testAccuracy
[1,]         0.051         0.05
```

Without any feature selections, our Naive Bayes Classifier model returns an accuracy of 5.1% for the train dataset and an accuracy of 5% for the testing dataset. Since the accuracy for both datasets are low, we will not pursue the Naive Bayes Classifier model as one of the possible models for our dataset.

06 Conclusion

After evaluating the different models with variables from various feature selections, we concluded two best models, Neural Network and SVM sigmoid to determine the characteristics of Absenteeism from work. The variation of the accuracies are evaluated with a stratified K-fold cross validation to prevent overpopulation of the test sets with the majority class of data specified in 2.6.2.

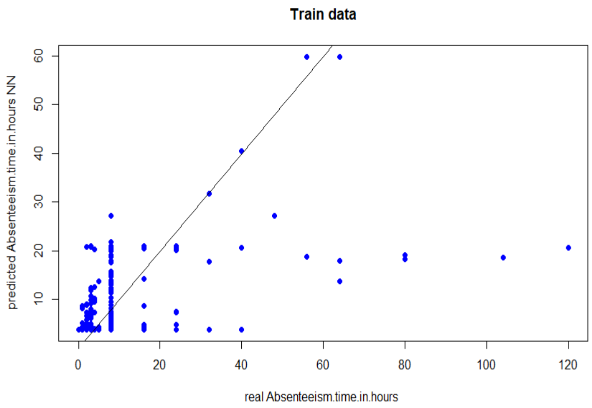
This table shows the RMSE/ Accuracy of the data with the stratified K-fold cross validations:

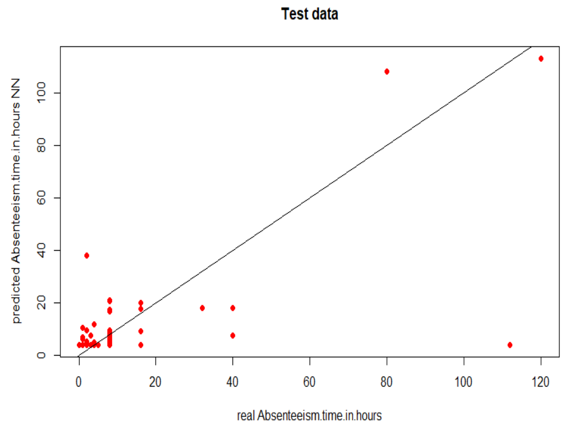
Model	**RSME(For Neural Network)/Accuracy(For SVM)**			
	2-fold	**5-fold**	**10-fold**	**15-fold**
Neural Network	Test- 20.07973 Train- 6.061027	Test-11.48549 Train-9.197887	Test-20.39662 Train-9.485357	Test-24.51548 Train-9.372849
SVM (Sigmoid)	Test- 39.88439% Train- 47.97101%	Test- 33.09353% Train-77.89855%	Test- 37.14286% Train-NaN	Test-28.57143% Train-73.93617%

As there is a significant variation with differing folds, predictions on test data with trained models may differ significantly based on the train-test split proportion. Hence, by following the 5-fold, our model results in the highest and most reliable accuracy.

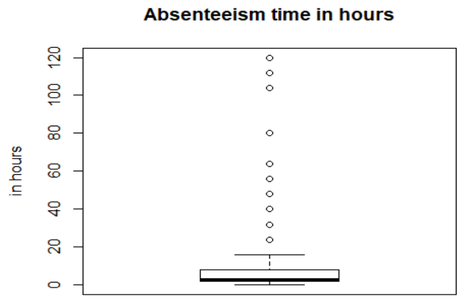
Firstly, the Neural Network Model (with 2 hidden layers with 6 nodes) with features selected by Hypothesis Testing. The model returned a low RMSE of 11.48549 in the training set and 9.197887 in the test set. The low RMSE and RMSE difference between train and test signifies a high predictive power of the model. However, the drawback of utilising this model is that since the neural network is a black box process, little insights can be drawn regarding the features.

Secondly, the Support Vector Machine Model (Sigmoid Kernel) with features selected by Information Gain. The model displayed a high 77.8955% accuracy in the training set and 33.09353% in the test set. Considering that our data has many features to begin with, SVM would be highly effective in the high dimensional space. However, while the training accuracy is undoubtedly high for our model, the test accuracy is conversely low, translating to a low predictive ability of the model.





The high disparity in accuracy for SVM can be attributed to the small sample size leading to overfitting and the high number of outliers in our dependent variable. In order to address the small sample size and reduce overfitting, we conducted feature selection to select only the most deterministic of features and evaluated it through various kernels, using the optimal train-test split ratio through cross validation. Nonetheless, the small sample size of 36 individuals provided in the data allowed little mitigation against overfitting.



In addition, the high number of outliers in the dependent variable impacted the accuracy of our models greatly and may be another causal factor for overfitting in our SVM model. With the outliers being included in the training set, the trained model will tend to overfit to accommodate the outliers resulting in lower test accuracy and predictive power. Hence, the evaluation of other models such as a decision tree (Random Forest) as well as the Naive Bayes Classifier faced the same problem. The Naive Bayes Classifier also relies on the assumption that all variables are independent of each other but in reality, such independence is close to an impossibility, hence returning inaccurate results when implemented onto this dataset.

In conclusion, as the dataset available only consists of 36 individuals, a larger dataset is required for a more conclusive conclusion. More features can also be considered based on the geographical demographics as only sons are included in this dataset instead of the general classification, children which require the same level of time commitment, in order to prevent omitted variable bias. There could have been incorrect inputs in Absenteeism hours negatively influencing the accuracy of models as well. In addition, more models can be considered and tested in future research with the various features selected which has proven to improve results.

07 Room for Improvement

7.1 Data Collection

The dataset that we have used consists of 740 tuples which is considered quite a small sample. To make matters worse, the experiment was done only on 36 different workers which is an extremely small sample size. Hence, the accuracy of our model may be easily affected by overfitting. This also provided little basis for the model to be generalised to the larger population.

The provided dataset has already been preprocessed. The original dataset stated by the author contained 38 attributes and 2243 records. This would have been useful to us in establishing a stronger model. However, the author has already filtered out several attributes and records. Hence, it will be hard for us to derive any predictions with such limited samples. The accuracy of our models could be improved had the original dataset been provided.

7.2 Outliers and potentially incorrect inputs

From the outliers identification in section 2.4, there are a significant number of outliers within the dependent variable-Absenteeism.time.in.hours. The large number of outliers will inadvertently affect both our test and train accuracy, making it prone to overfitting. The small size of the dataset aggravates this issue as inaccurate inputs in the dependent variable will be impossible to identify and rectify.

7.3 Class Imbalance

From section 2.6.2, a strong class imbalance can be observed where only 6.08% of workers have never been absent. This results in a stronger basis for error when predicting the absenteeism hours = 0 as compared to those >0. Hence, a more accurate model can be obtained should more data be collected with absenteeism hours = 0.

08 References

- Ferreira, R. P., & Martiniano, A., & Napolitano, D. & Prado Farias, E. B. & Sassi, R. J. (2018), *International Journal of Recent Scientific Research Vol. 9, Issue, 1(G), pp. 23332-23334, January, 2018*. doi: 10.24327/ijrsr.2018.0901.1447
- https://towardsdatascience.com/support-vector-machines-for-classification-fc7c1565e3
- https://medium.com/machinevision/overview-of-neural-networks-b86ce02ea3d1
- https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf
- https://towardsdatascience.com/random-forests-and-decision-trees-from-scratch-in-python-3e4fa5ae4249
- https://researchleap.com/critical-risk-analysis-absenteeism-work-place/
- http://recentscientific.com/artificial-neural-network-and-their-application-prediction-absenteeism-work

Kernels:

```
> auc(pred_linear,testset$Absenteeism.time.in.hours)
[1] 0.2910305
> mean(testset$Absenteeism.time.in.hours==pred_linear)
[1] 0.4316547
```

```
> auc(pred_polynomial,testset$Absenteeism.time.in.hours)
[1] 0.4492537
> mean(testset$Absenteeism.time.in.hours==pred_polynomial)
[1] 0.4172662
```

```
> auc(pred_radial,testset$Absenteeism.time.in.hours)
[1] 0.408874
> mean(testset$Absenteeism.time.in.hours==pred_radial)
[1] 0.4460432
```

```
> auc(pred_sigmoid,testset$Absenteeism.time.in.hours)
[1] 0.3540026
> mean(testset$Absenteeism.time.in.hours==pred_sigmoid)
[1] 0.4100719
```

With a low Area Under Curve (AUC) of 0.29, it shows that SVM is unable to clearly find a separating hyperplane due to overfitting. This resulted in a low accuracy of 0.43 of the model for the testing data.

With an AUC of 0.45, it shows that with the help of the polynomial kernel function, SVM is able to find a clearer hyperplane as compared to linear. This resulted in a higher accuracy of 0.42 of the model for the testing data.

With an AUC of 0.41, it shows that with the help of the radial kernel function, SVM is less able to form a defined hyperplane as compared to radial. It did however, manage to obtain a higher accuracy of 0.45 in the testing data.

With an AUC of 0.35, with the help of sigmoid kernel function, it is better than radial but worse than polynomial. It has the same accuracy as a polynomial of 0.41 for the testing data.

4.2 Filter Methods to conduct Feature Selection

4.2.1 Correlation:

Correlation measures the degree of association between two numeric variables. Features with a high correlation with the dependent variable will be selected and included in our model.

With reference to the correlation matrix in section 2.5, other than the variable "Disciplinary Failure", the other independent variables have a relatively low correlation coefficient with the dependent variable (Absenteeism in hours). Thus, we are not able to conduct feature selection easily with the low correlation coefficients. Instead, a non-linear model may be more suitable for this dataset.

4.2.2 Hypothesis Testing (t-test and Chi-square Test):

To determine if the independent variables are statistically significant, hypothesis testing is carried out. This will be done using two kinds of test: t-test and Chi-square test.
The t-test measures the degree of association between continuous independent variables and the dependent variable while the Chi-square test measures the association between two categorical variables and it will be used to test for association between categorical independent variables and the dependent variable. The following describes the null and alternate hypothesis:
a. Null Hypothesis: The independent variable is statistically insignificant
b. Alternate Hypothesis: The independent variable is statistically significant.

The p-values obtained will be used to determine whether we reject the null hypothesis. If it is less than the 5% level of significance, we reject the null hypothesis and conclude that the variable is statistically significant. The following are the p-values we obtained for each independent variable based on the t-test and Chi-square test:

T-test for continuous variables		**Chi-square test for categorical variables**	
Variable	**p-value**	**Variable**	**p-value**
Transport expenses	1.135922e-304	Reason for absence	4.006719e-90
Distance from work	1.142153e-130	Day of the week	2.38968e-05
Service Time	1.907112e-23	Seasons	0.1859655
Age	1.62962e-250	Disciplinary failure	2.89192e-08
Work load average	6.061992e-57	Education	0.9709759
Hit target	0	Social smoker	0.06856764

Weight	0	Social drinker	0.03080072
Height	0		
Son	1.105422e-23		
Pet	5.0099e-26		

For the continuous variables, we can observe that all of them have a p-value less than 0.05 so we can conclude that they are all statistically significant. As for the categorical variables, the variables, “Day of the week”, “Education”, “Social drinker”, “Social smoker” and “Pet” have a p-value greater than 0.05 so we can conclude that they are statistically insignificant. Thus, these variables can be excluded from our model as they do not contribute greatly to the prediction of our dependent variable. On the other hand, the remaining variables which have a p-value less than 0.05 should be included in our model as they are statistically significant.

In conclusion, the following variables will be included in our model:
Transport expenses, Distance from work, Service Time, Age, Work load average, Hit target, Weight, Height, Reason.for.absence, Month.of.absence, Day.of.the.week, Seasons, Disciplinary.failure, Education, Son, Social.drinker , Social.smoker, Pet.

4.2.3 Information Gain:

```
                                 attr_importance
Reason.for.absence                    0.28459327
Month.of.absence                      0.00000000
Day.of.the.week                       0.00000000
Seasons                               0.00000000
Transportation.expense                0.04068501
Distance.from.Residence.to.Work       0.00000000
Service.time                          0.00000000
Age                                   0.00000000
Work.load.Average.day                 0.00000000
Hit.target                            0.00000000
Disciplinary.failure                  0.08368046
Education                             0.00000000
Son                                   0.00000000
Social.drinker                        0.00000000
Social.smoker                         0.00000000
Pet                                   0.00000000
Weight                                0.00000000
Height                                0.00000000
```

4.3 Feature Selection using Wrapper Methods

4.3.1 Stepwise Forward and Backward Selection:

This feature selection method helps us build a model by adding and removing certain characteristics. The following are different methods of stepwise regression:

a. Forward selection - The model starts off empty and then variables are progressively added to it.
b. Backward selection - The model starts off with all of the variables and then the least significant ones are removed from the model.
c. Stepwise selection - A mixture of both forward and backward selection. At each iteration, the algorithm decides whether a variable is added or removed from the model.

Output:
The following variables were selected from the stepwise regression selection.

```
> print(vars_step)
 [1] "(Intercept)"                     "Height"
 [3] "Reason.for.absence"              "Disciplinary.failure"
 [5] "Son"                             "Social.drinker"
 [7] "Day.of.the.week"                 "Seasons"
 [9] "Distance.from.Residence.to.Work" "Hit.target"
> print(vars_forward)
 [1] "(Intercept)"                     "Height"
 [3] "Reason.for.absence"              "Disciplinary.failure"
 [5] "Son"                             "Social.drinker"
 [7] "Day.of.the.week"                 "Seasons"
 [9] "Distance.from.Residence.to.Work" "Hit.target"
> print(vars_backward)
 [1] "(Intercept)"                     "Reason.for.absence"
 [3] "Day.of.the.week"                 "Seasons"
 [5] "Distance.from.Residence.to.Work" "Hit.target"
 [7] "Disciplinary.failure"            "Son"
 [9] "Social.drinker"                  "Height"
> 
```

4.3.2 Recursive Feature Elimination (RFE) Method:

Progressively, a model consisting of all variables drops the least significant feature, leaving behind the specified number of features. The optimal number of features in the model can be identified using cross-validation.

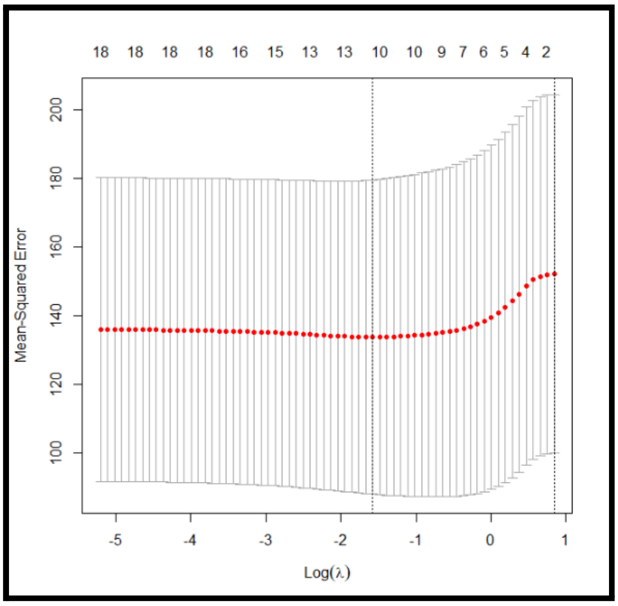
```
The top 2 variables (out of 2):
   Disciplinary.failure, Reason.for.absence
```

The results of the RFE shows that the highest accuracy rate consists of the following 2 variables: Disciplinary failure, Reason for absence.

4.4 Feature Selection using Embedded Methods

4.4.1 Least Absolute Shrinkage and Selection Operator (Lasso)

(Intercept)	-34.30200
Reason.for.absence	-0.40798
Month.of.absence	0.00000
Day.of.the.week	-0.57546
Seasons	0.57183
Transportation.expense	0.00000
Distance.from.Residence.to.Work	-0.03907
Service.time	0.00000
Age	0.00000
Work.load.Average.day	0.00000
Hit.target	0.16489
Disciplinary.failure	-15.62591
Education	0.00000
Son	1.16382
Social.drinker	3.04612
Social.smoker	0.00000
Pet	-0.10185
Weight	0.00000
Height	0.19152



This feature selection technique conducts regularisation whereby it shrinks the coefficients of the regression model as part of the penalisation. For feature selection, the variables which remain after the shrinkage process are included in the model.

We are unable to make inferences about the importance of the coefficients as the data has only been scaled individually and not scaled to have a common mean and standard deviation. Since our variables have different means and standard deviation, variables with larger averages will tend to have larger absolute coefficients.

Any variable with a coefficient of zero would be dropped from the model, because it shows that it has no predictive power. The following variables have a coefficient of zero and would hence be dropped for our model.

4. Month of Absence
5. Transportation Expense
6. Service Time
7. Age
8. Average Workload/ Day
9. Education
10. Social Drinker
11. Weight

The remaining variables would still be considered in our model.

4.4.2 Boruta

Boruta algorithm is another feature selection algorithm. Boruta is a wrapper built around the random forest classification algorithm.

```
                       meanImp  decision
Reason.for.absence   10.964907 Confirmed
Disciplinary.failure  9.628133 Confirmed
Service.time          4.765935 Confirmed
Age                   3.546094 Confirmed
Pet                   2.687534 Confirmed
Height                2.267262 Confirmed
```

At every iteration, Boruta runs and compares between a real feature and its shadow feature, whether or not the real feature has a higher importance. (i.e. comparing the Z score between the 2, whether the Z score of the real feature > max Z score of its shadow). The model also removes features which are deemed not significant. The algorithm completes when the various features are either confirmed or rejected.

From the results as shown in the figure above , we can see that the Boruta model confirmed the following 6 variables: Reasons for absence, Disciplinary Failure, Height , Age, Pet.

4.4.3 Random Forest

```
Random Forest

553 samples
 18 predictor

Pre-processing: scaled (18)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 497, 497, 498, 497, 497, 499, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    10.89517  0.1528534  4.954569
  10    11.24620  0.1696434  5.111034
  18    11.52299  0.1684981  5.261658

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 2.
```

```
> rfImp
rf variable importance

                              Overall
Reason.for.absence            100.000
Disciplinary.failure           97.093
Service.time                   47.519
Social.drinker                 42.482
Seasons                        40.510
Distance.from.Residence.to.Work 39.472
Month.of.absence               39.185
Height                         38.317
Pet                            34.355
Age                            30.055
Hit.target                     19.580
Son                            17.704
Day.of.the.week                14.446
Work.load.Average.day          12.852
Social.smoker                  11.308
Education                       6.333
Transportation.expense          5.067
Weight                          0.000
```

This feature selection technique builds a random forest model and then provides a list of significant variables.
As seen from the results from the figure above,, we can see that the randomForest achieved an optimal model with the following variables: Reason for Absence and Disciplinary Failure.

4.5 Comparison Between Methods

Type	Method	No. of features selected	Features Selected
Filter	**Correlation**	N.A.	N.A.
	Hypothesis Testing	15	Transportation.expense Distance.from.Residence.to.Work Service.time Age Work.load.Average.day Hit.target Weight Height Reason.for.absence Month.of.absence Seasons Disciplinary.failure Son Social Drinker Pet
	Information Gain	3	Transportation.expense Disciplinary.failure Reason.for.absence
Wrapper	**Stepwise Regression**	9 9	Both: Height Reason.for.absence Disciplinary.failure Son Social.drinker Day.of.the.week Seasons Distance.from.Residence.to.Work Hit.target Forward: Height Reason.for.absence Disciplinary.failure Son Social.drinker Day.of.the.week Seasons Distance.from.Residence.to.work Hit.target

		9	<u>Backward:</u> Reason.for.absence Disciplinary.failure Son Social.drinker Day.of.the.week Seasons Distance.from.Residence.to.work Hit.target Height
	Recursive Feature Elimination	2	Reason.for.absence Disciplinary.failure
Embedded	**LASSO**	10	Reason.for.absence Day.of.the.week Seasons Distance.from.Residence.to.Work Hit.target Disciplinary.failure Son Social.drinker Pet Height
	Boruta	6	Reason.for.absence Disciplinary.failure Service.time Age Pet Height
	Random Forest	2	Reason.for.absence Disciplinary.failure

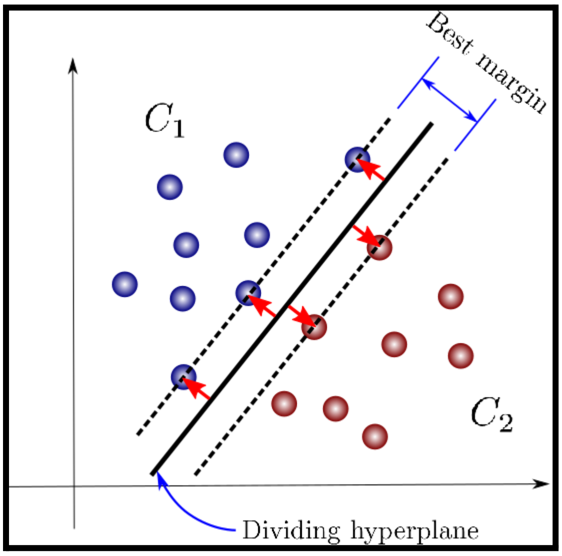
A different set of features can be obtained from each method and while some may not provide a definite set of features to be included, certain features can be filtered out for consideration. The table above shows the features we have identified from each model.
Hence, we conducted several tests with the different sets of features (namely from ________) and achieved the highest predictive accuracy when using _feature set_. Additionally, it can be noted that the features have been consistently selected as the top few variables across models. As it is beneficial for employers to monitor less features in predicting absenteeism, we chose _feature set_ as our inputs.mydata2

05 Model Selection

5.1 Support Vector Machine(SVM)

For this model, we would plot each data item as a point in a n-dimension space (where n = 20 as it is the number of characteristics for our dataset) with the value of every characteristic being a coordinate. We would then conduct classification by finding out what would be the optimal hyper-plane for the selected characteristics.

This would be optimal for our dataset because it is effective in high dimensional spaces (high number of features).



5.1.1 Testing of the Kernels:

```
> auc(pred_linear,test$Absenteeism.time.in.hours)
[1] 0.3526903
> mean(test$Absenteeism.time.in.hours==pred_linear)
[1] 0.4100719
```

With a low Area Under Curve (AUC) of 0.35, it shows that SVM is unable to clearly find a separating hyperplane due to overfitting. This resulted in a low accuracy of 0.41 of the model for the testing data.

```
> auc(pred_polynomial,test$Absenteeism.time.in.hours)
[1] 0.4594286
> mean(test$Absenteeism.time.in.hours==pred_polynomial)
[1] 0.4244604
```

With an AUC of 0.45, it shows that with the help of the polynomial kernel function, SVM is able to find a clearer hyperplane as compared to linear. This resulted in a higher accuracy of 0.42 of the model for the testing data.

```
> auc(pred_radial,test$Absenteeism.time.in.hours)
[1] 0.3539055
> mean(test$Absenteeism.time.in.hours==pred_radial)
[1] 0.4460432
```

With an AUC of 0.35, it shows that with the help of the radial kernel function, SVM is less able to form a defined hyperplane as compared to radial. It did however, manage to obtain a higher accuracy of 0.45 in the testing data.

```
> auc(pred_sigmoid,test$Absenteeism.time.in.hours)
[1] 0.4318182
> mean(test$Absenteeism.time.in.hours==pred_sigmoid)
[1] 0.4244604
```

With an AUC of 0.43, with the help of sigmoid kernel function, it is better than radial but worse than polynomial. It has the same accuracy as a polynomial of 0.42 for the testing data.

5.1.2 Model Evaluation:

	Train Data Accuracy		**Test Data Accuracy**	
	Polynomial	Radial	Polynomial	Radial
Hypothesis Testing	35.13431%	43.16547%	36.98925%	43.88489%
Information Gain	NaN	NaN	38.84892%	41.72662%
Step-wise	39.10853%	38.2622%	38.84892%	40.28777%
RFE/RF	NaN	NaN	36.69065%	35.97122%
LASSO	42.72017%	35.53763%	38.84892%	39.56835%
Boruta	27.73723%	27.73723%	43.88489%	38.1295%

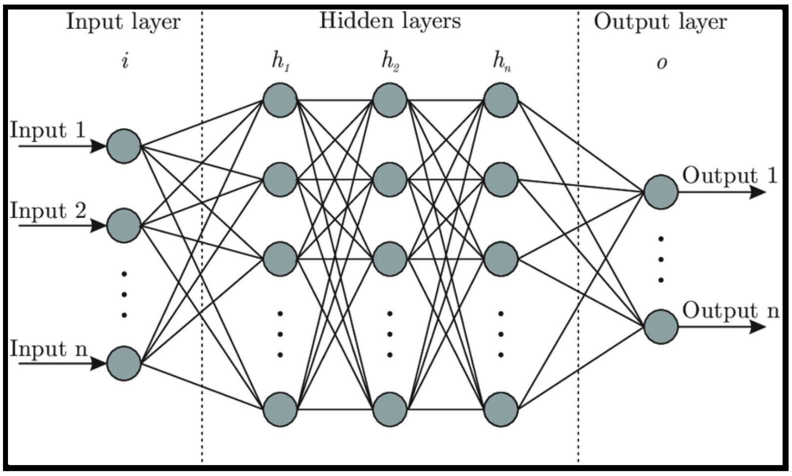
5.2 Random Forest (Decision Tree Model)

Random Forest consists of many individual decision trees that operate together. Each decision tree represents an independent variable and generates a class prediction, which contributes to a vote in the final prediction. Since decision trees are highly sensitive to the data they are trained on, small changes to the training set can lead to significantly different tree structures. Thus, random forest builds on this by allowing each individual tree to randomly sample from the dataset with replacement (bagging/bootstrap aggregation), resulting in different trees. With reference to section 4.1.1, the low correlation between the independent variables plays a key role in ensuring the accuracy of the random forest classifier. With the low correlation between trees, they are able to protect each other from potential errors that might occur.

5.2.1 Model Evaluation

	Train Data Accuracy	Test Data Accuracy
Hypothesis Testing	6.55%	16.73%
Information Gain	15.5%	15.06%
Step-wise	12.83%	16.42%
RFE/RF	12.88%	13.13%
LASSO	13.83%	16.85%
Boruta	11.61%	14.29%

5.3 Neural Network



Neural networks are the workhorses of deep learning. They are black boxes trying to achieve good predictions.

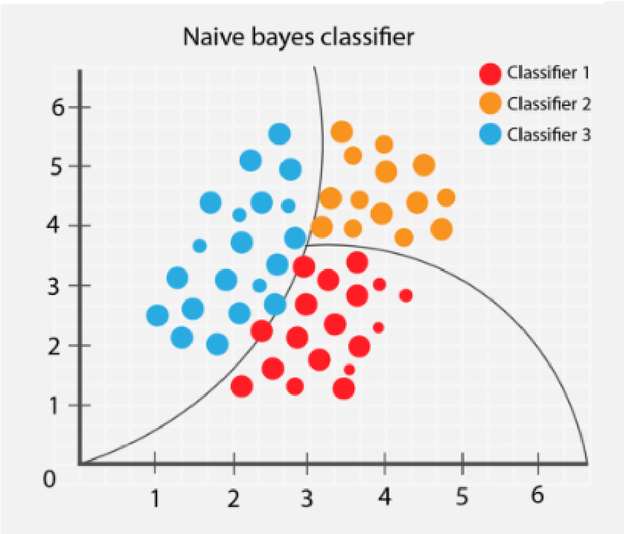
"A neural network has input and output neurons, which are connected by weighted synapses. The weights affect how much of the forward propagation goes through the neural network. The weights can then be changed during the back propagation — this is the part where the neural network is now learning.
This process of forward propagation and backward propagation is conducted iteratively on every piece of data in a training data set. The greater the size of the data set and the greater the variety of data set that there is, the more that the neural network will learn, and the better that the neural network will get at predicting outputs."

- *Overview of Neural Networks from Medium.com*
- *https://medium.com/machinevision/overview-of-neural-networks-b86ce02ea3d1*

5.3.1 Model Evaluation

	Train Data Accuracy	**Test Data Accuracy**
Hypothesis Testing		
Information Gain		
Step-wise		
RFE/RF		
LASSO		
Boruta		

5.4 Naive Bayes Classifier



$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Mathematically, the Bayes theorem is represented as P(A|B). The Naive Bayes Classifier belongs to the family of probability classifier, using Bayesian theorem. This method solves classification problems using a probabilistic approach. However, it has a strong assumption that all the variables are independent of one another. This might not be the case for real-life examples. This is also why the model requires much less training data. Even if the assumption does not hold, this method could still prove to be an effective one.

5.4.1 Model Evaluation

	Train Data Accuracy	**Test Data Accuracy**
Hypothesis Testing		
Information Gain		

Step-wise		
RFE/RF		
LASSO		
Boruta		

06 Conclusion

07 Room for Improvement

1. Other than the perfect multicollinearity that we stated in Homework 5, we might need to explore possible variables which are not perfectly correlated but have high correlation. This can be done in Feature Selection and help us narrow down important variables.

2. We only received one dataset which is insufficient to validate our proposed model in our study. Larger dataset with a higher number of employees would be desirable in order to come to a stronger conclusion.

08 References

09 Annex

4.2 Filter Methods to conduct Feature Selection

4.2.1 Correlation:

Correlation measures the degree of association between two numeric variables. Features with a high correlation with the dependent variable will be selected and included in our model.

With reference to the correlation matrix in section 2.5, other than the variable "Disciplinary Failure", the other independent variables have a relatively low correlation coefficient with the dependent variable (Absenteeism in hours). Thus, we are not able to conduct feature selection easily with the low correlation coefficients. Instead, a non-linear model may be more suitable for this dataset.

4.2.2 Hypothesis Testing (t-test and Chi-square Test):

To determine if the independent variables are statistically significant, hypothesis testing is carried out. This will be done using two kinds of test: t-test and Chi-square test.
The t-test measures the degree of association between continuous independent variables and the dependent variable while the Chi-square test measures the association between two categorical variables and it will be used to test for association between categorical independent variables and the dependent variable. The following describes the null and alternate hypothesis:
a. Null Hypothesis: The independent variable is statistically insignificant
b. Alternate Hypothesis: The independent variable is statistically significant.

The p-values obtained will be used to determine whether we reject the null hypothesis. If it is less than the 5% level of significance, we reject the null hypothesis and conclude that the variable is statistically significant. The following are the p-values we obtained for each independent variable based on the t-test and Chi-square test:

T-test for continuous variables		**Chi-square test for categorical variables**	
Variable	**p-value**	**Variable**	**p-value**
Transport expenses	1.135922e-304	Reason for absence	4.006719e-90
Distance from work	1.142153e-130	Day of the week	2.38968e-05
Service Time	1.907112e-23	Seasons	0.1859655
Age	1.62962e-250	Disciplinary failure	2.89192e-08
Work load average	6.061992e-57	Education	0.9709759
Hit target	0	Social smoker	0.06856764

Weight	0	Social drinker	0.03080072
Height	0		
Son	1.105422e-23		
Pet	5.0099e-26		

For the continuous variables, we can observe that all of them have a p-value less than 0.05 so we can conclude that they are all statistically significant. As for the categorical variables, the variables, “Day of the week”, “Education”, “Social drinker”, “Social smoker” and “Pet” have a p-value greater than 0.05 so we can conclude that they are statistically insignificant. Thus, these variables can be excluded from our model as they do not contribute greatly to the prediction of our dependent variable. On the other hand, the remaining variables which have a p-value less than 0.05 should be included in our model as they are statistically significant.

In conclusion, the following variables will be included in our model:
Transport expenses, Distance from work, Service Time, Age, Work load average, Hit target, Weight, Height, Reason.for.absence, Month.of.absence, Day.of.the.week, Seasons, Disciplinary.failure, Education, Son, Social.drinker , Social.smoker, Pet.

4.2.3 Information Gain:

```
                                  attr_importance
Reason.for.absence                     0.28459327
Month.of.absence                       0.00000000
Day.of.the.week                        0.00000000
Seasons                                0.00000000
Transportation.expense                 0.04068501
Distance.from.Residence.to.Work        0.00000000
Service.time                           0.00000000
Age                                    0.00000000
Work.load.Average.day                  0.00000000
Hit.target                             0.00000000
Disciplinary.failure                   0.08368046
Education                              0.00000000
Son                                    0.00000000
Social.drinker                         0.00000000
Social.smoker                          0.00000000
Pet                                    0.00000000
Weight                                 0.00000000
Height                                 0.00000000
```

4.3 Feature Selection using Wrapper Methods

4.3.1 Stepwise Forward and Backward Selection:

This feature selection method helps us build a model by adding and removing certain characteristics. The following are different methods of stepwise regression:

a. Forward selection - The model starts off empty and then variables are progressively added to it.
b. Backward selection - The model starts off with all of the variables and then the least significant ones are removed from the model.
c. Stepwise selection - A mixture of both forward and backward selection. At each iteration, the algorithm decides whether a variable is added or removed from the model.

Output:
The following variables were selected from the stepwise regression selection.

```
> print(vars_step)
 [1] "(Intercept)"                     "Height"
 [3] "Reason.for.absence"              "Disciplinary.failure"
 [5] "Son"                             "Social.drinker"
 [7] "Day.of.the.week"                 "Seasons"
 [9] "Distance.from.Residence.to.Work" "Hit.target"
> print(vars_forward)
 [1] "(Intercept)"                     "Height"
 [3] "Reason.for.absence"              "Disciplinary.failure"
 [5] "Son"                             "Social.drinker"
 [7] "Day.of.the.week"                 "Seasons"
 [9] "Distance.from.Residence.to.Work" "Hit.target"
> print(vars_backward)
 [1] "(Intercept)"                     "Reason.for.absence"
 [3] "Day.of.the.week"                 "Seasons"
 [5] "Distance.from.Residence.to.Work" "Hit.target"
 [7] "Disciplinary.failure"            "Son"
 [9] "Social.drinker"                  "Height"
> |
```

4.3.2 Recursive Feature Elimination (RFE) Method:

Progressively, a model consisting of all variables drops the least significant feature, leaving behind the specified number of features. The optimal number of features in the model can be identified using cross-validation.

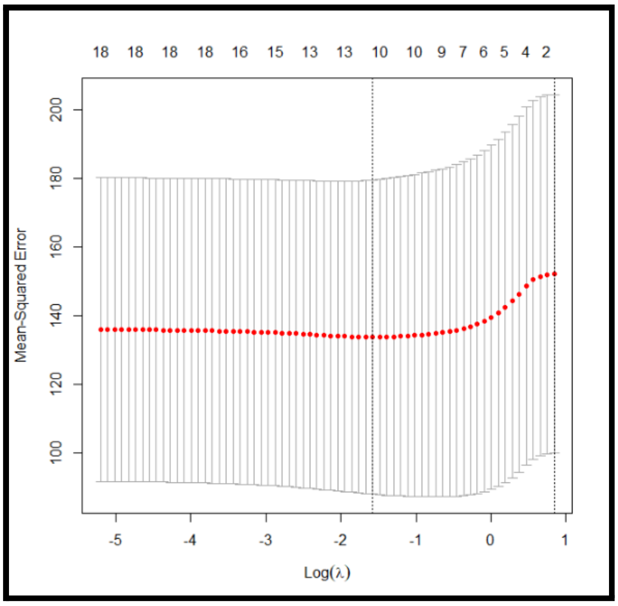
```
The top 2 variables (out of 2):
   Disciplinary.failure, Reason.for.absence
```

The results of the RFE shows that the highest accuracy rate consists of the following 2 variables: Disciplinary failure, Reason for absence.

4.4 Feature Selection using Embedded Methods

4.4.1 Least Absolute Shrinkage and Selection Operator (Lasso)

```
(Intercept)                         -34.30200
Reason.for.absence                   -0.40798
Month.of.absence                      0.00000
Day.of.the.week                      -0.57546
Seasons                               0.57183
Transportation.expense                0.00000
Distance.from.Residence.to.Work      -0.03907
Service.time                          0.00000
Age                                   0.00000
Work.load.Average.day                 0.00000
Hit.target                            0.16489
Disciplinary.failure                -15.62591
Education                             0.00000
Son                                   1.16382
Social.drinker                        3.04612
Social.smoker                         0.00000
Pet                                  -0.10185
Weight                                0.00000
Height                                0.19152
```



This feature selection technique conducts regularisation whereby it shrinks the coefficients of the regression model as part of the penalisation. For feature selection, the variables which remain after the shrinkage process are included in the model.

We are unable to make inferences about the importance of the coefficients as the data has only been scaled individually and not scaled to have a common mean and standard deviation. Since our variables have different means and standard deviation, variables with larger averages will tend to have larger absolute coefficients.

Any variable with a coefficient of zero would be dropped from the model, because it shows that it has no predictive power. The following variables have a coefficient of zero and would hence be dropped for our model.

12. Month of Absence
13. Transportation Expense
14. Service Time
15. Age
16. Average Workload/ Day
17. Education
18. Social Drinker
19. Weight

The remaining variables would still be considered in our model.

4.4.2 Boruta

Boruta algorithm is another feature selection algorithm. Boruta is a wrapper built around the random forest classification algorithm.

```
                       meanImp  decision
Reason.for.absence   10.964907 Confirmed
Disciplinary.failure  9.628133 Confirmed
Service.time          4.765935 Confirmed
Age                   3.546094 Confirmed
Pet                   2.687534 Confirmed
Height                2.267262 Confirmed
```

At every iteration, Boruta runs and compares between a real feature and its shadow feature, whether or not the real feature has a higher importance. (i.e. comparing the Z score between the 2, whether the Z score of the real feature > max Z score of its shadow). The model also removes features which are deemed not significant. The algorithm completes when the various features are either confirmed or rejected.

From the results as shown in the figure above , we can see that the Boruta model confirmed the following 6 variables: Reasons for absence, Disciplinary Failure, Height , Age, Pet.

4.4.3 Random Forest

```
Random Forest

553 samples
 18 predictor

Pre-processing: scaled (18)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 497, 497, 498, 497, 497, 499, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    10.89517  0.1528534  4.954569
  10    11.24620  0.1696434  5.111034
  18    11.52299  0.1684981  5.261658

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 2.
```

```
> rfImp
rf variable importance

                             Overall
Reason.for.absence           100.000
Disciplinary.failure          97.093
Service.time                  47.519
Social.drinker                42.482
Seasons                       40.510
Distance.from.Residence.to.Work 39.472
Month.of.absence              39.185
Height                        38.317
Pet                           34.355
Age                           30.055
Hit.target                    19.580
Son                           17.704
Day.of.the.week               14.446
Work.load.Average.day         12.852
Social.smoker                 11.308
Education                      6.333
Transportation.expense         5.067
Weight                         0.000
```

This feature selection technique builds a random forest model and then provides a list of significant variables.
As seen from the results from the figure above,, we can see that the randomForest achieved an optimal model with the following variables: Reason for Absence and Disciplinary Failure.

4.5 Comparison Between Methods

Type	Method	No. of features selected	Features Selected
Filter	**Correlation**	N.A.	N.A.
	Hypothesis Testing	15	Transportation.expense Distance.from.Residence.to.Work Service.time Age Work.load.Average.day Hit.target Weight Height Reason.for.absence Month.of.absence Seasons Disciplinary.failure Son Social Drinker Pet
	Information Gain	3	Transportation.expense Disciplinary.failure Reason.for.absence
Wrapper	**Stepwise Regression**	9 9	<u>Both:</u> Height Reason.for.absence Disciplinary.failure Son Social.drinker Day.of.the.week Seasons Distance.from.Residence.to.Work Hit.target <u>Forward:</u> Height Reason.for.absence Disciplinary.failure Son Social.drinker Day.of.the.week Seasons Distance.from.Residence.to.work Hit.target

		9	<u>Backward:</u> Reason.for.absence Disciplinary.failure Son Social.drinker Day.of.the.week Seasons Distance.from.Residence.to.work Hit.target Height
	Recursive Feature Elimination	2	Reason.for.absence Disciplinary.failure
Embedded	**LASSO**	10	Reason.for.absence Day.of.the.week Seasons Distance.from.Residence.to.Work Hit.target Disciplinary.failure Son Social.drinker Pet Height
	Boruta	6	Reason.for.absence Disciplinary.failure Service.time Age Pet Height
	Random Forest	2	Reason.for.absence Disciplinary.failure

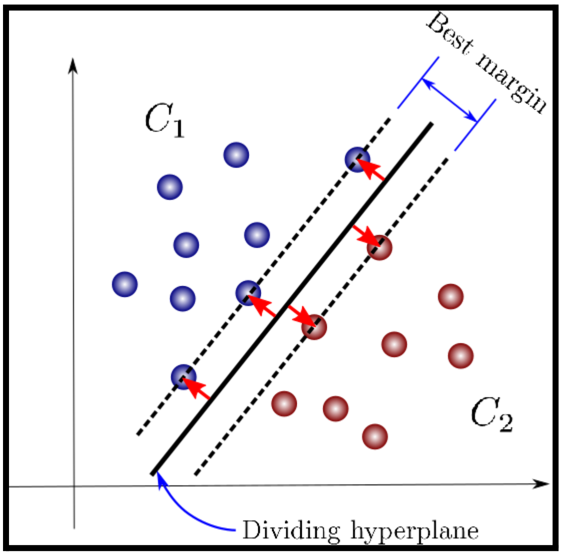
A different set of features can be obtained from each method and while some may not provide a definite set of features to be included, certain features can be filtered out for consideration. The table above shows the features we have identified from each model.
Hence, we conducted several tests with the different sets of features (namely from ________) and achieved the highest predictive accuracy when using _feature set_. Additionally, it can be noted that the features have been consistently selected as the top few variables across models. As it is beneficial for employers to monitor less features in predicting absenteeism, we chose _feature set_ as our inputs.mydata2

05 Model Selection

5.1 Support Vector Machine(SVM)

For this model, we would plot each data item as a point in a n-dimension space (where n = 20 as it is the number of characteristics for our dataset) with the value of every characteristic being a coordinate. We would then conduct classification by finding out what would be the optimal hyper-plane for the selected characteristics.

This would be optimal for our dataset because it is effective in high dimensional spaces (high number of features).



5.1.1 Testing of the Kernels:

```
> auc(pred_linear,test$Absenteeism.time.in.hours)
[1] 0.3526903
> mean(test$Absenteeism.time.in.hours==pred_linear)
[1] 0.4100719
```

With a low Area Under Curve (AUC) of 0.35, it shows that SVM is unable to clearly find a separating hyperplane due to overfitting. This resulted in a low accuracy of 0.41 of the model for the testing data.

```
> auc(pred_polynomial,test$Absenteeism.time.in.hours)
[1] 0.4594286
> mean(test$Absenteeism.time.in.hours==pred_polynomial)
[1] 0.4244604
```

With an AUC of 0.45, it shows that with the help of the polynomial kernel function, SVM is able to find a clearer hyperplane as compared to linear. This resulted in a higher accuracy of 0.42 of the model for the testing data.

```
> auc(pred_radial,test$Absenteeism.time.in.hours)
[1] 0.3539055
> mean(test$Absenteeism.time.in.hours==pred_radial)
[1] 0.4460432
```

With an AUC of 0.35, it shows that with the help of the radial kernel function, SVM is less able to form a defined hyperplane as compared to radial. It did however, manage to obtain a higher accuracy of 0.45 in the testing data.

```
> auc(pred_sigmoid,test$Absenteeism.time.in.hours)
[1] 0.4318182
> mean(test$Absenteeism.time.in.hours==pred_sigmoid)
[1] 0.4244604
```

With an AUC of 0.43, with the help of sigmoid kernel function, it is better than radial but worse than polynomial. It has the same accuracy as a polynomial of 0.42 for the testing data.

5.1.2 Model Evaluation:

	Train Data Accuracy		**Test Data Accuracy**	
	Polynomial	Radial	Polynomial	Radial
Hypothesis Testing	35.13431%	43.16547%	36.98925%	43.88489%
Information Gain	NaN	NaN	38.84892%	41.72662%
Step-wise	39.10853%	38.2622%	38.84892%	40.28777%
RFE/RF	NaN	NaN	36.69065%	35.97122%
LASSO	42.72017%	35.53763%	38.84892%	39.56835%
Boruta	27.73723%	27.73723%	43.88489%	38.1295%

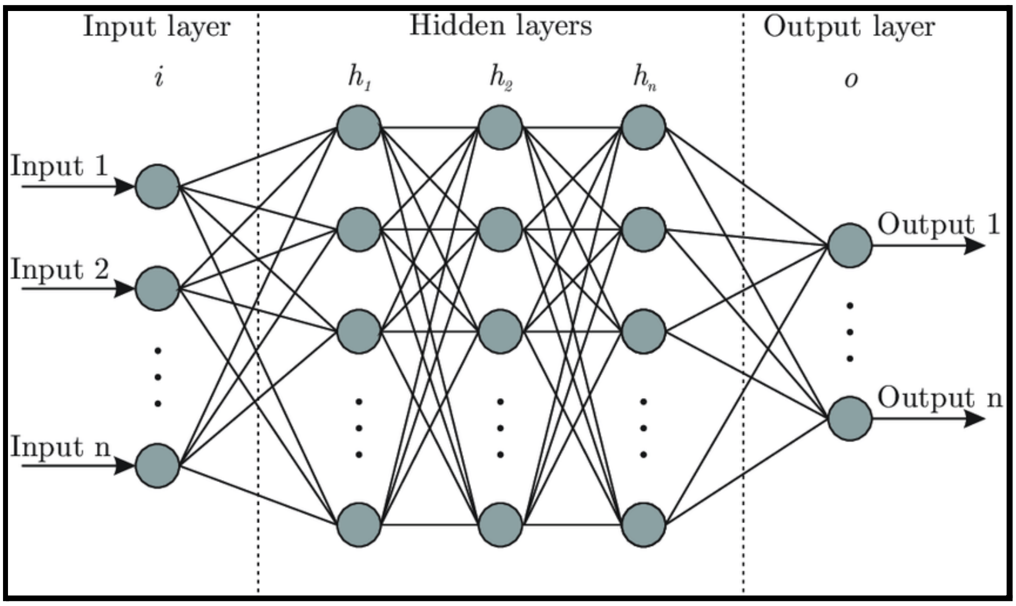
5.2 Random Forest (Decision Tree Model)

Random Forest consists of many individual decision trees that operate together. Each decision tree represents an independent variable and generates a class prediction, which contributes to a vote in the final prediction. Since decision trees are highly sensitive to the data they are trained on, small changes to the training set can lead to significantly different tree structures. Thus, random forest builds on this by allowing each individual tree to randomly sample from the dataset with replacement (bagging/bootstrap aggregation), resulting in different trees. With reference to section 4.1.1, the low correlation between the independent variables plays a key role in ensuring the accuracy of the random forest classifier. With the low correlation between trees, they are able to protect each other from potential errors that might occur.

5.2.1 Model Evaluation

	Train Data Accuracy	**Test Data Accuracy**
Hypothesis Testing	6.55%	16.73%
Information Gain	15.5%	15.06%
Step-wise	12.83%	16.42%
RFE/RF	12.88%	13.13%
LASSO	13.83%	16.85%
Boruta	11.61%	14.29%

5.3 Neural Network



Neural networks are the workhorses of deep learning. They are black boxes trying to achieve good predictions.

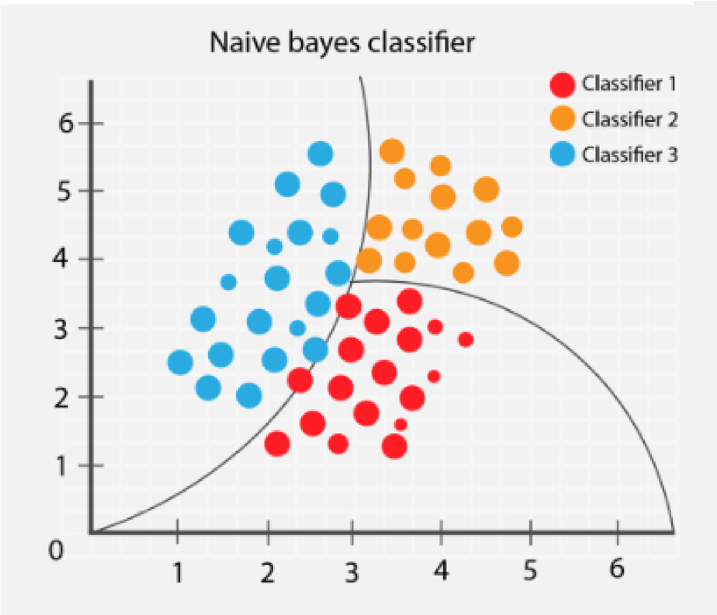
"A neural network has input and output neurons, which are connected by weighted synapses. The weights affect how much of the forward propagation goes through the neural network. The weights can then be changed during the back propagation — this is the part where the neural network is now learning.
This process of forward propagation and backward propagation is conducted iteratively on every piece of data in a training data set. The greater the size of the data set and the greater the variety of data set that there is, the more that the neural network will learn, and the better that the neural network will get at predicting outputs."

- *Overview of Neural Networks from Medium.com*
- *https://medium.com/machinevision/overview-of-neural-networks-b86ce02ea3d1*

5.3.1 Model Evaluation

	Train Data Accuracy	**Test Data Accuracy**
Hypothesis Testing		
Information Gain		
Step-wise		
RFE/RF		
LASSO		
Boruta		

5.4 Naive Bayes Classifier



$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Mathematically, the Bayes theorem is represented as P(A|B). The Naive Bayes Classifier belongs to the family of probability classifier, using Bayesian theorem. This method solves classification problems using a probabilistic approach. However, it has a strong assumption that all the variables are independent of one another. This might not be the case for real-life examples. This is also why the model requires much less training data. Even if the assumption does not hold, this method could still prove to be an effective one.

5.4.1 Model Evaluation

	Train Data Accuracy	**Test Data Accuracy**
Hypothesis Testing		
Information Gain		
Step-wise		
RFE/RF		
LASSO		
Boruta		

06 Conclusion

07 Room for Improvement

3. Other than the perfect multicollinearity that we stated in Homework 5, we might need to explore possible variables which are not perfectly correlated but have high correlation. This can be done in Feature Selection and help us narrow down important variables.

4. We only received one dataset which is insufficient to validate our proposed model in our study. Larger dataset with a higher number of employees would be desirable in order to come to a stronger conclusion.

08 References

Q1. Model selection:

Data Overview:

The Absenteeism at work dataset consists of 740 observations and 20 characteristics.Out of which Absenteeism time in hours is the dependent variable with 19 independent variables. By analysing the data, we could draw insights about the company's employees. Such models would help the company deploy manpower more efficiently and effectively, creating better workflow and/or layoff employees at a high risk of being absent for work.

Data Pre-Processing

- Removing null values(Month = 0)
- Removing outliers
- Removing Weight and Height due to Multicollinearity

The dataset now consists of 692 observations with 18 characteristics.

1. Skewed classes:

As we can see, only approximately 5.5% of workers have never been absent as compared to 94.5% of workers with more than 0 hours of absenteeism.
Number of workers with 0 absenteeism hours: 36 (minority class)
Number of workers with >0 absenteeism hours: 656 (majority class)
This presents an issue when we attempt to split the data as the training test could be overly populated by the majority class as compared to the minority. This would, in turn, affect the accuracy calculated.
Hence, greater emphasis is placed on the splitting of data and chosen metrics.

- Splitting the data: 80% training, 20% testing
  Training set will have 553 samples.
  Testing set will have 139 samples.

2. Lack of data:

There is a lack of examples in the dataset, 692 workers(after removing 3 workers with Month=0 and 45 outliers).
For algorithms that require more data, this will present implications.
Hence, we decided to keep to simpler algorithms.

3. Too many features:

For such a small dataset with 692 students, we have a staggeringly high number of features, 17. This presents a Curse of Dimensionality as we need to increase the number of examples we have exponentially for each feature added. Hence, to deal with such a high number of features, we will need to conduct feature selection to select the specific features which are the strongest predictors of our dependent variable, Absenteeism hours.

Q2 : Model Evaluation

Possible models

Model	Pros	Cons
Support Vector Machine (SVM)	• Effective for high-dimensional space • Selection of kernels for non-linear correlation • Even with bias, it remains robust	• Black Box • Long and inefficient • Features may be dependent or highly correlated
Logistic Regression/Linear Regression	• Easy to interpret • Probability output: Possible to rank instead of classifying • Possible to regularize the model to take into account errors and over-fitting	• Overfitting • Assumed independent observations • Outperformed by complex models

1) Linear Regression

```
Multiple R-squared:  0.1066,	Adjusted R-squared:  0.07993
```

With the adjusted R-squared of 0.07993 being low, the linear regression model is not recommended as a predictor of Absenteeism.

2) Linear SVM

```
> auc(pred_linear, test.data$`Absenteeism.time.in.hours`)
[1] 0.4924541
> mean(test.data$`Absenteeism.time.in.hours` == pred_linear)
[1] 0.4100719
```

With a low Area Under Curve of 0.49, it shows that SVM is unable to clearly find a separating hyperplane due to overfitting. This resulted in a low accuracy of 0.41 of the model for the testing data.

3) Polynomial SVM

```
> auc(pred_polynomial, test.data$`Absenteeism.time.in.hours`)
[1] 0.519086
> mean(test.data$`Absenteeism.time.in.hours` == pred_polynomial)
[1] 0.4172662
```

With a low Area Under Curve of 0.52, it shows that with the help of the kernel function, SVM is able to find a clear hyperplane as compared to linear. This resulted in a higher accuracy of 0.42 of the model for the testing data.

4) Logit Regression

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.107e+01  7.993e+05   0.000    1.000
Month.of.absence             2.563e+00  1.042e+04   0.000    1.000
Day.of.the.week              1.806e+01  6.191e+03   0.003    0.998
Seasons                     -1.118e+00  3.370e+04   0.000    1.000
Transportation.expense       1.009e-01  7.435e+02   0.000    1.000
Distance.from.Residence.to.Work -3.888e-01  3.083e+03   0.000    1.000
Service.time                 1.171e+00  1.180e+04   0.000    1.000
Age                         -1.844e+00  7.144e+03   0.000    1.000
Work.load.Average.day       -6.510e-01  3.077e+03   0.000    1.000
Hit.target                   1.346e-01  8.605e+03   0.000    1.000
Disciplinary.failure        -1.441e+02  1.267e+05  -0.001    0.999
Education                    9.002e+00  3.917e+04   0.000    1.000
Son                          2.546e+00  2.213e+04   0.000    1.000
Social.drinker               1.023e+01  8.934e+04   0.000    1.000
Social.smoker                5.249e+00  1.273e+05   0.000    1.000
Pet                         -1.620e-01  2.101e+04   0.000    1.000
Body.mass.index              8.084e-01  5.918e+03   0.000    1.000
```

Due to the large spread of data points and poor feature selection, resulting in our logit regression falsely showing that all variables are insignificant in predicting Absenteeism.

Conclusion

After evaluating the 4 possible models, it seems that a Polynomial SVM model is the best model to evaluate our data set. We believe that the low AUC (52%) is due to the poor feature selection that we did previously. We believe that removing insignificant variables will help in increasing the AUC, obtaining a more accurate model. After doing another round of proper feature selection, we will utilise stratified K fold testing to determine the number of folds which provides the highest level of accuracy for our Polynomial SVM model. We might also look into using Neural Network which was taught recently as a possible model for our project.

Q3: Rooms for Improvements

5. If we utilise logit regression models for prediction, we will have a heavy class imbalance with only 5.5% of workers with 0 absenteeism hours as compared to 95.5% with more than 0. As such, percentage error when predicting absenteeism hours equal 0 will be more than absenteeism hours more than 0. If more data is available with absenteeism hour equal 0, we will be able to obtain a more accurate model with less bias.

6. Other than the perfect multicollinearity that we stated in Homework 5, we might need to explore possible variables which are not perfectly correlated but have high correlation. This can be done in Feature Selection and help us narrow down important variables.

7. We should conduct feature selection before model selection to raise the accuracy of our model prediction. Such methods of feature selection include filter methods: Correlation, Hypothesis Testing, Information Gain, wrapper methods: Stepwise Forward and Backward Selection, Recursive Feature Elimination and Embedded Methods: Random Forest, Boruta.

8. We only received one dataset which is insufficient to validate our proposed model in our study. Larger dataset with a higher number of employees would be desirable in order to come to a stronger conclusion.

Logit Regression

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.107e+01  7.993e+05   0.000    1.000
Month.of.absence               2.563e+00  1.042e+04   0.000    1.000
Day.of.the.week                1.806e+01  6.191e+03   0.003    0.998
Seasons                       -1.118e+00  3.370e+04   0.000    1.000
Transportation.expense         1.009e-01  7.435e+02   0.000    1.000
Distance.from.Residence.to.Work -3.888e-01  3.083e+03   0.000    1.000
Service.time                   1.171e+00  1.180e+04   0.000    1.000
Age                           -1.844e+00  7.144e+03   0.000    1.000
Work.load.Average.day         -6.510e-01  3.077e+03   0.000    1.000
Hit.target                     1.346e-01  8.605e+03   0.000    1.000
Disciplinary.failure          -1.441e+02  1.267e+05  -0.001    0.999
Education                      9.002e+00  3.917e+04   0.000    1.000
Son                            2.546e+00  2.213e+04   0.000    1.000
Social.drinker                 1.023e+01  8.934e+04   0.000    1.000
Social.smoker                  5.249e+00  1.273e+05   0.000    1.000
Pet                           -1.620e-01  2.101e+04   0.000    1.000
Body.mass.index                8.084e-01  5.918e+03   0.000    1.000
```

Due to the large spread of data points and poor feature selection, resulting in our logit regression falsely showing that all variables are insignificant in predicting Absenteeism.

```
> postResample(predict(rfMod, newdata= test[,1:18]),test[,19])
      RMSE   Rsquared        MAE
11.1679226  0.1330206  4.5380213
```

	Training Data Accuracy	**Test Data Accuracy**
Linear Regression	R-squared of 0.07993	
Logistic Regression	-	-
Linear SVM		AUC of 0.35
Polynomial SVM		AUC of 0.45
Radial SVM		AUC of 0.35
Sigmoid SVM		AUC of 0.43
Random Forest		R-squared of 0.133

1

```
> print(boruta.train)
Boruta performed 818 iterations in 2.471952 mins.
 9 attributes confirmed important: Age, Disciplinary.failur
e,
Distance.from.Residence.to.Work, Month.of.absence, Seasons a
nd 4 more;
 7 attributes confirmed unimportant: Body.mass.index, Day.o
f.the.week, Education,
Hit.target, Pet and 2 more;
```

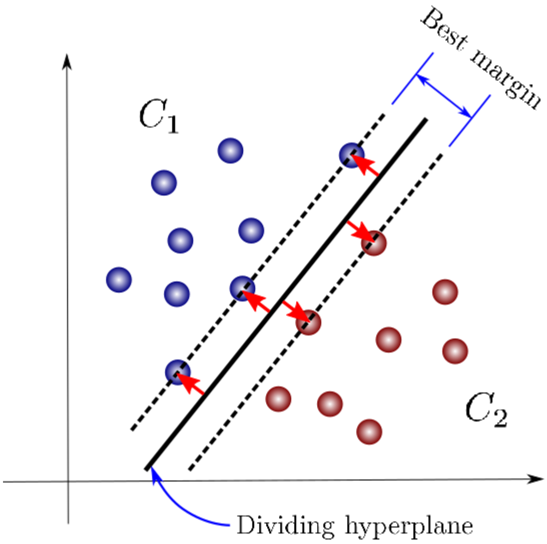
```
Call:
lm(formula = mydata$Absenteeism.time.in.hours ~ mydata$Day.of.the.week +
    mydata$Month.of.absence + mydata$Seasons + mydata$Transportation.expense +
    mydata$Distance.from.Residence.to.Work + mydata$Service.time +
    mydata$Age + mydata$Work.load.Average.day + mydata$Hit.target +
    mydata$Disciplinary.failure + mydata$Education + mydata$Son +
    mydata$Social.drinker + mydata$Social.smoker + mydata$Pet +
    mydata$Body.mass.index)

Residuals:
    Min      1Q  Median      3Q     Max
-15.236  -4.406  -1.824   1.233 108.975

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -23.42165   19.78144  -1.184 0.236929
mydata$Day.of.the.week                  -1.09035    0.36173  -3.014 0.002698 **
mydata$Month.of.absence                  0.12777    0.19558   0.653 0.513827
mydata$Seasons                           0.84225    0.54556   1.544 0.123217
mydata$Transportation.expense            0.00114    0.01055   0.108 0.914051
mydata$Distance.from.Residence.to.Work  -0.15603    0.04563  -3.420 0.000675 ***
mydata$Service.time                      0.09347    0.22526   0.415 0.678343
mydata$Age                              -0.11024    0.14158  -0.779 0.436538
mydata$Work.load.Average.day             0.03293    0.05113   0.644 0.519887
mydata$Hit.target                        0.37731    0.19318   1.953 0.051317 .
mydata$Disciplinary.failure             -9.23514    2.44140  -3.783 0.000173 ***
mydata$Education                         0.23242    0.89447   0.260 0.795084
mydata$Son                               1.39218    0.53397   2.607 0.009383 **
mydata$Social.drinker                    5.78621    1.43507   4.032 6.33e-05 ***
mydata$Social.smoker                     0.91634    2.11568   0.433 0.665102
mydata$Pet                               0.01191    0.50770   0.023 0.981288
mydata$Body.mass.index                  -0.09971    0.18005  -0.554 0.579956
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.8 on 536 degrees of freedom
Multiple R-squared:  0.1066,	Adjusted R-squared:  0.07993
F-statistic: 3.997 on 16 and 536 DF,  p-value: 3.287e-07
```

```
glm(formula = Absent ~ ., family = binomial(link = "logit"),
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6651   0.0000   0.0000   0.0000   0.7585

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.107e+01  7.993e+05   0.000    1.000
Month.of.absence             2.563e+00  1.042e+04   0.000    1.000
Day.of.the.week              1.806e+01  6.191e+03   0.003    0.998
Seasons                     -1.118e+00  3.370e+04   0.000    1.000
Transportation.expense       1.009e-01  7.435e+02   0.000    1.000
Distance.from.Residence.to.Work -3.888e-01  3.083e+03   0.000    1.000
Service.time                 1.171e+00  1.180e+04   0.000    1.000
Age                         -1.844e+00  7.144e+03   0.000    1.000
Work.load.Average.day       -6.510e-01  3.077e+03   0.000    1.000
Hit.target                   1.346e-01  8.605e+03   0.000    1.000
Disciplinary.failure        -1.441e+02  1.267e+05  -0.001    0.999
Education                    9.002e+00  3.917e+04   0.000    1.000
Son                          2.546e+00  2.213e+04   0.000    1.000
Social.drinker               1.023e+01  8.934e+04   0.000    1.000
Social.smoker                5.249e+00  1.273e+05   0.000    1.000
Pet                         -1.620e-01  2.101e+04   0.000    1.000
Body.mass.index              8.084e-01  5.918e+03   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221.6143  on 552  degrees of freedom
Residual deviance:   4.4987  on 536  degrees of freedom
AIC: 38.499

Number of Fisher Scoring iterations: 25
```

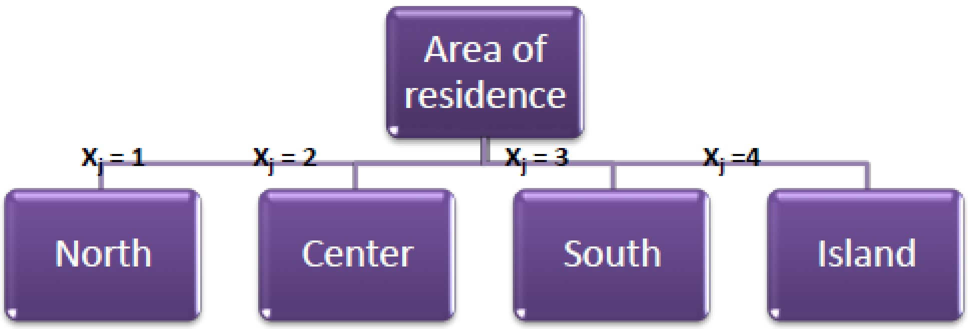
```
> print(boruta.train)
Boruta performed 818 iterations in 2.471952 mins.
 9 attributes confirmed important: Age, Disciplinary.failur
e,
Distance.from.Residence.to.Work, Month.of.absence, Seasons a
nd 4 more;
 7 attributes confirmed unimportant: Body.mass.index, Day.o
f.the.week, Education,
Hit.target, Pet and 2 more;
```

Support Vector Machine (SVM): For this model, we would plot each data item as a point in a n-dimension space (where n = 20 as it is the number of characteristics for our dataset) with the value of every characteristic being a coordinate. We would then conduct classification by finding out what would be the optimal hyper-plane for the selected characteristics.

This would be optimal for our dataset because it is effective in high dimensional spaces (high number of features).

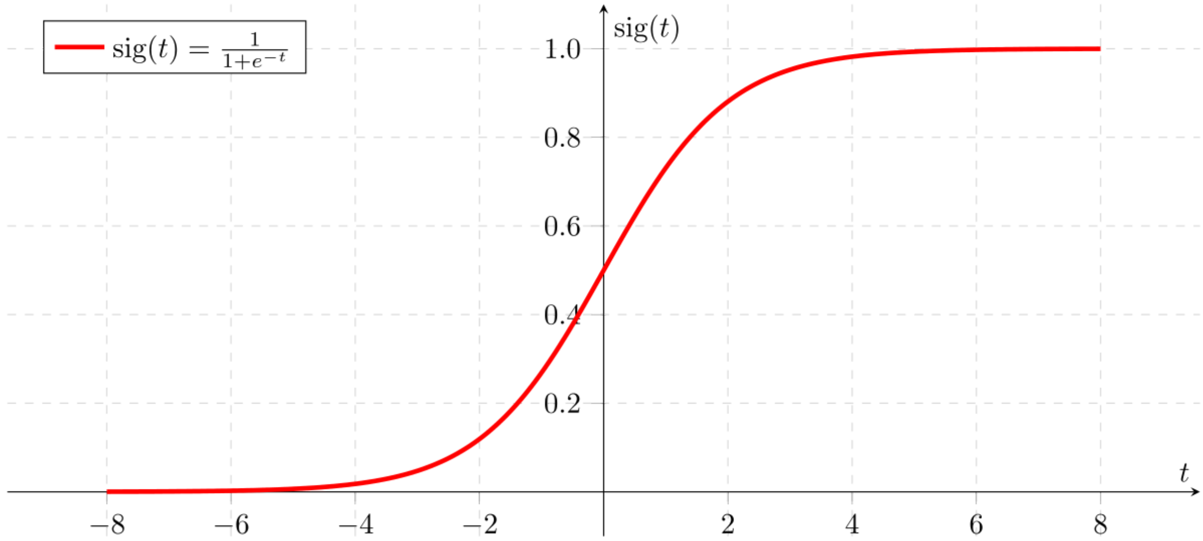


Multi-Split Classification Tree: For this model, one possible method would be to calculate the Entropy index or the Gini Index to decide which characteristics we should omit, and which characteristics we should focus on for our analysis. However, this process would be manually taxing and inefficient. Hence, we have decided not to use this modelling method.



Univariate split for a nominal attribute

Logistic Regression: A multiple variable logistic regression model would take into consideration all 19 of the characteristics of the dataset, with the dependent variable being the absenteeism. As our problem would be a binary classification problem (absent or not absent), a logistic regression model would be applicable.

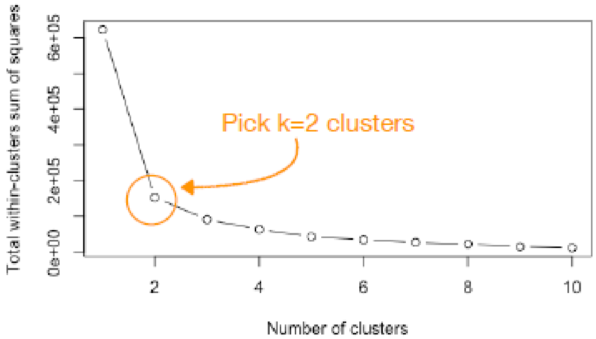


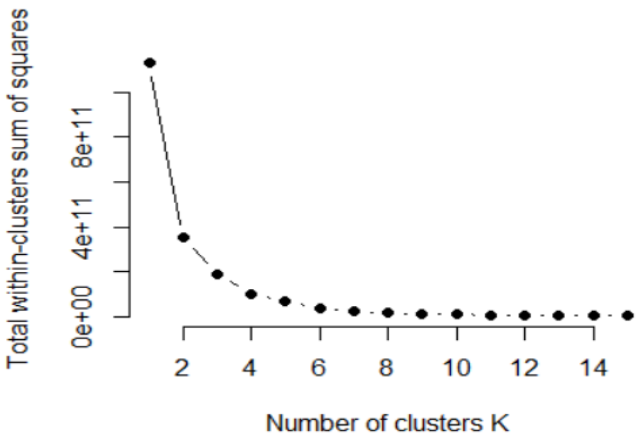
<u>Logit Regression of our training dataset</u>

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -23.42165   19.78144  -1.184 0.236929    
mydata$Day.of.the.week                -1.09035    0.36173  -3.014 0.002698 ** 
mydata$Month.of.absence                0.12777    0.19558   0.653 0.513827    
mydata$Seasons                         0.84225    0.54556   1.544 0.123217    
mydata$Transportation.expense          0.00114    0.01055   0.108 0.914051    
mydata$Distance.from.Residence.to.Work -0.15603    0.04563  -3.420 0.000675 ***
mydata$Service.time                    0.09347    0.22526   0.415 0.678343    
mydata$Age                            -0.11024    0.14158  -0.779 0.436538    
mydata$Work.load.Average.day           0.03293    0.05113   0.644 0.519887    
mydata$Hit.target                      0.37731    0.19318   1.953 0.051317 .  
mydata$Disciplinary.failure           -9.23514    2.44140  -3.783 0.000173 ***
mydata$Education                       0.23242    0.89447   0.260 0.795084    
mydata$Son                             1.39218    0.53397   2.607 0.009383 ** 
mydata$Social.drinker                  5.78621    1.43507   4.032 6.33e-05 ***
mydata$Social.smoker                   0.91634    2.11568   0.433 0.665102    
mydata$Pet                             0.01191    0.50770   0.023 0.981288    
mydata$Body.mass.index                -0.09971    0.18005  -0.554 0.579956    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 139.3007)

    Null deviance: 83574  on 552  degrees of freedom
Residual deviance: 74665  on 536  degrees of freedom
AIC: 4318

Number of Fisher Scoring iterations: 2
```

K-Means Clustering: Using the elbow method as shown below, we would decide on the optimal k number of clusters. Given k, identify the center of clusters ("centroid") that minimize within-cluster sum of squares (WCSS) and maximize between-cluster sum of squares (BCSS) based on the chosen distance metric.

Our dataset would be split into the optimal clusters, according to all the input characteristics. From there, we could categorize employees into clusters and gain absenteeism insights from there.






```
set.seed(123)
# Compute and plot wss for k = 2 to k = 15.
k.max <- 15
data <- mydata
wss <- sapply(1:k.max,
              function(k){kmeans(data, k, nstart=50,iter.max = 15 )$tot.withinss})
wss
plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```

As seen from the R plot shown above, it shows the optimal number of clusters should be 2. Possible splitting of clusters may include …

CONCLUSION

References

https://towardsdatascience.com/support-vector-machines-for-classification-fc7c1565e3

https://sefiks.com/2017/11/19/how-random-forests-can-keep-you-from-decision-tree/

https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

https://pythonprogramminglanguage.com/kmeans-elbow-method/

Q2.Model evaluation

Model evaluation procedures

1. **Training and testing on the same data**
    - **Rewards overly complex models that "overfit" the training data and won't necessarily generalize**
2. **Train/test split**
    - **Split the dataset into two pieces, so that the model can be trained and tested on different data**
    - **Better estimate of out-of-sample performance, but still a "high variance" estimate**
    - **Useful due to its speed, simplicity, and flexibility**
3. **K-fold cross-validation**
    - **Systematically create "K" train/test splits and average the results together**
    - **Even better estimate of out-of-sample performance**
    - **Runs "K" times slower than train/test split**

K-Means Clustering	• Easy to implement • Training would be done more efficiently	• Testing is slow • Have to correctly explain the reason for the cluster.

Model evaluation metrics

- **Regression problems: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error**
- **Classification problems: Classification accuracy**
    - **There are many more metrics, and we will discuss them today**

Q3. Discussion on whether there is any room for improvement.

```
No pre-processing
Resampling: Cross-Validated (2 fold, repeated 3 times)
Summary of sample sizes: 75, 75, 75, 75, 75, 75, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.9577778  0.9366667
   TRUE      0.9577778  0.9366667
```

Mean(training set) = 6.6075956.94
Mean(testing set) = 6.568345

<u>Data Exploration</u>

Total number of workers: 692
Number of features: 20
Number of workers with absenteeism hours > 0 : 656
Number of students with absenteeism hours = 0: 36
Absenteeism rate of the class: 94.51% (2 d.p.)

1. We predict that the disciplinary actions will be positively correlated to the hours of absenteeism as the actions proved to be deterrent against absenteeism.
2. We predict that distance will be positively related to the hours of absenteeism. For those who stay further away from their workplace, they will have to travel a greater distance which will lead to greater inconvenience and also higher transportation costs. Thus, they might be less inclined to travel to work, resulting in their absence.

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is an approach to analyse datasets, so as to summarise their main characteristics with the help of graphical methods. This maximises our data insights, testing our underlying assumptions and detects outliers and anomalies.

2.1 Data Overview

The Absenteeism at work dataset consists of 740 observations and 20 characteristics, of which Absenteeism time in hours is the dependent variable with 19 independent variables.

Absenteeism time in hours, Transportation Expense, Distance from Residence to work, Service Time, Age, Workload Average/day, Hit target, Son, Pet, Weight, Height, Body Mass Index consist of integer values.

Reason for absence, Month of absence, Day of the week, Seasons, Education consist of discrete categorical variables.

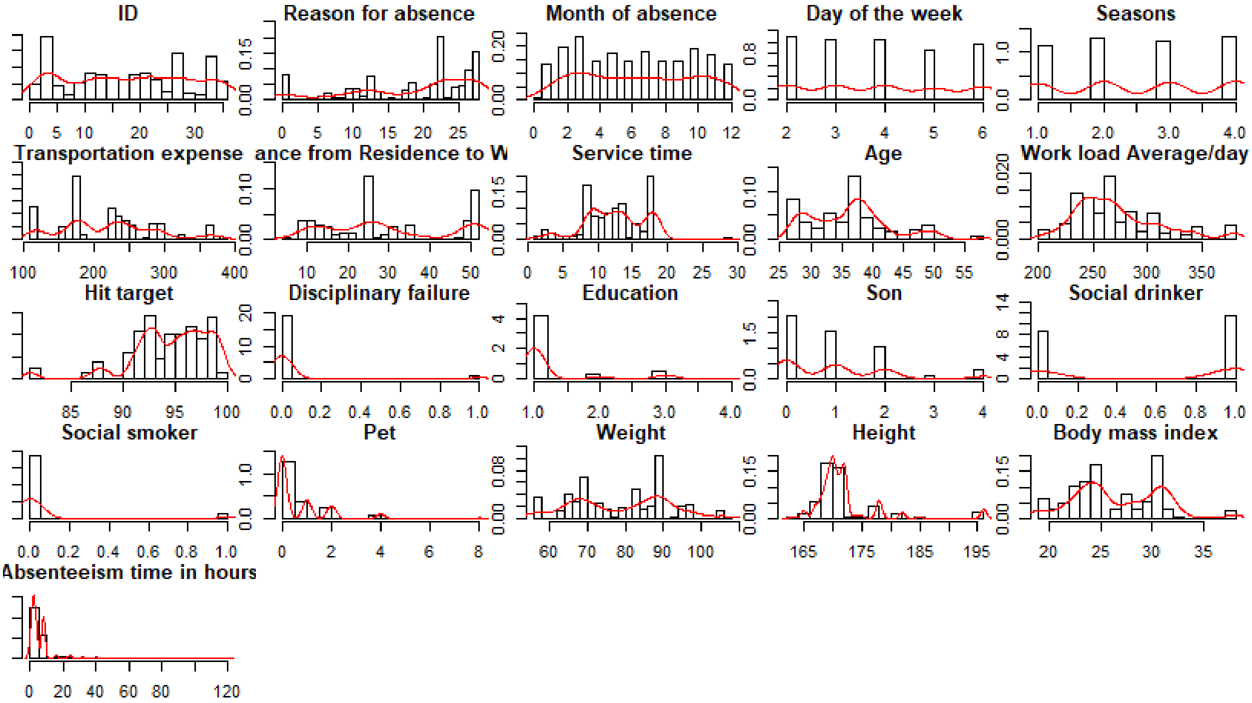
Disciplinary failure, Social drinker, Social smoker are dummy variables.

None of the variable columns has null/missing values.

2.2 Inconsistent values within dataset

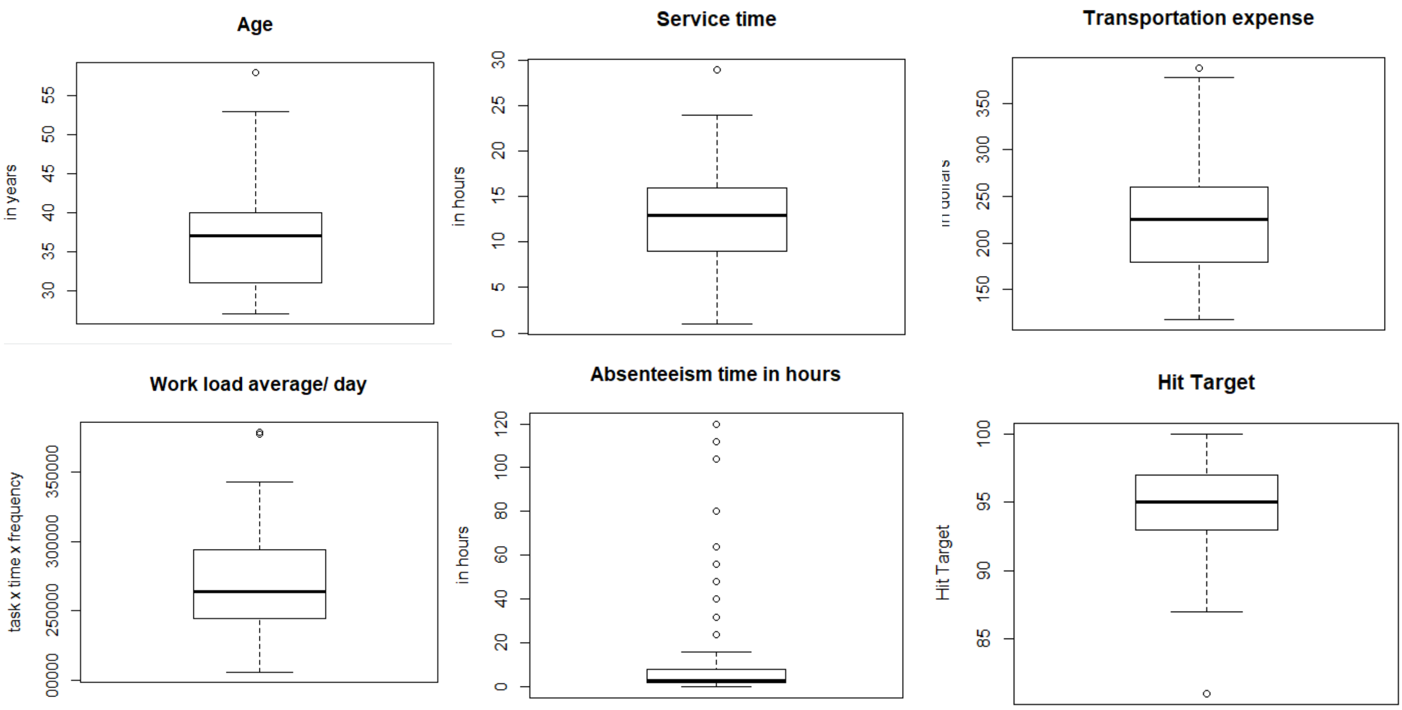
Legend for variable Month of absence: 1 = January; 2 = February; 3 = March; 4 = April; 5 = May; 6 = June; 7 = July; 8 = August; 9 = September; 10 = October; 11 = November; 12 = December. However, the range of values found were ranged from [0.12]. This means that there is an additional value of 0 that is unaccounted for.

2.3 Checking for categorical data



The density plot for each attribute, based on continuity demonstrates if it is categorical. An example of which is the top right hand corner, “Seasons” which is categorical as observed by the discontinuity.

2.4 Distribution of continuous variables



2.5 Spotting anomalies and outliers

- Looking at the boxplot for “Age”, “Service time”, “Transportation expense”, “Work load average/day”, and “Hit target”, we can see that these dependent variables have outliers. We would remove the outliers before moving on to analyse the dataset and check our hypothesis.
- For the aforementioned dependent variables, we would check the R^2 before and after removing the outlier, and if the R^2 increases, the removal of the outlier is validated.

3 Data Pre-Processing

3.1 Filtering Entries not consistent with data source

In the Month of absence column, there are values “0” which do not correspond to the legend for the table. Hence we chose to remove it as the Month of absence could skew the correlation between the month of absence and Absenteeism hours.

3.2 Removing outliers

Since all attributes for the data points do not follow a normal distribution and some variables display covariance with one another. The Mahalanobis Distance can be used to identify the outliers. With the Mahalanobis Distance designs with Gaussian distribution, it is not necessary to have a joint multivariate normal distribution and it will still improve the objective functions to a greater extent in its variables/ attributes.

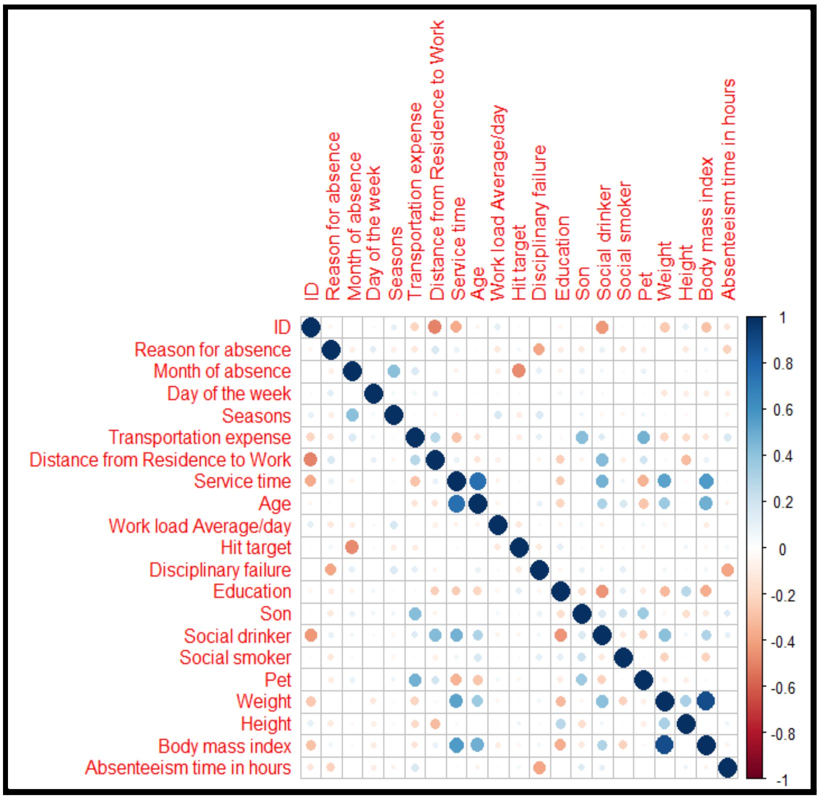
3.3 Multicollinearity between variables

We would be using multiple linear regression to find the effect of each dependent variable on the independent variable (absenteeism in hours). Using the p-value, we can determine which variables are statistically significant.

We would also cluster some of the dependent variables (reasons for absenteeism) into 4 clusters:

1. Employees with pre-existing medical conditions
2. Employees going for medical examination
3. Employees that went for a blood donation
4. Employees with unjustified reasonings for absenteeism

The dataset gives us a holistic and comprehensive overview and hence there are not any omitted variable biases that are prominent and important to include.



Assumptions

- We have to check for possible multicollinearity between the variables (Possible ones include Months and Seasons, Height & Weight and BMI, transport expenses and distance from work).
- If multicollinearity exists, we have to remove some of the variables.

  Analysing the correlation matrix, we can see that “Body Mass Index '' and “Weight” has a highly positive correlation coefficient. We would then have to take this into consideration and only use one of the variables.

  BMI is calculated using height and weight, hence we have decided to use BMI only instead of all 3 in our model.

Further analysis by running a linear regression model on BMI, height and weight, the high adjusted R^2 value = 0.9927 shows multicollinearity between BMI and height and weight. This further justifies our decision to remove height and weight.

4 Feature Selection

4.1 Definition

<u>…………</u>

4.2 Feature Selection through Filter Methods

4.2.1 Correlation

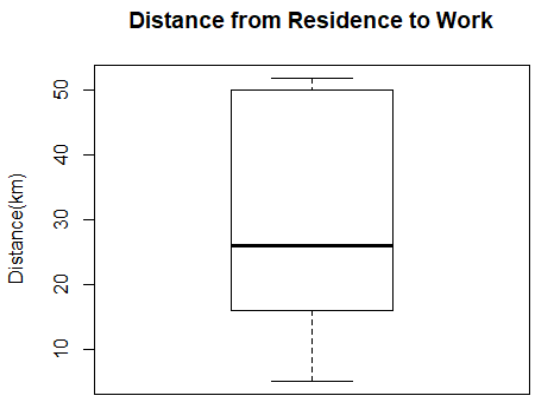
<u>Make correlation matrix</u>

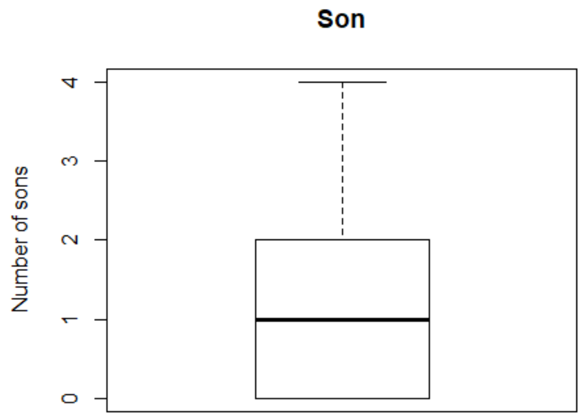
4.2.2 Hypothesis Testing

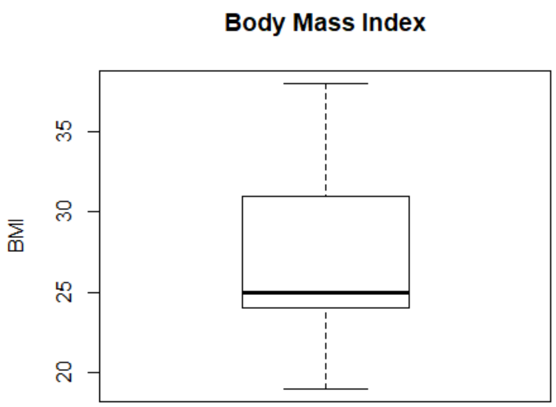
Q3: Possible hypothesis

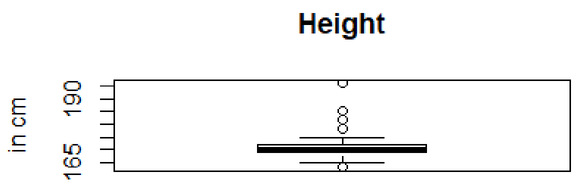
3. We predict that the cluster of employees with pre-existing medical conditions would be the biggest group. In other words, we foresee that employees within this group would have the highest number of hours of absenteeism due to health reasons. For this group, health-related variables (BMI, smoker, drinker, age) will have a stronger correlation with the hours of absenteeism compared to the other clusters.
4. We predict that the number of sons will be positively related to the hours of absenteeism. This is because these workers will most likely have to spend more time taking care of their children, thereby resulting in their absence from work.
5. We predict that the disciplinary actions will be positively correlated to the hours of absenteeism as the actions proved to be deterrent against absenteeism.
6. We predict that distance will be positively related to the hours of absenteeism. For those who stay further away from their workplace, they will have to travel a greater distance which will lead to greater inconvenience and also higher transportation costs. Thus, they might be less inclined to travel to work, resulting in their absence.

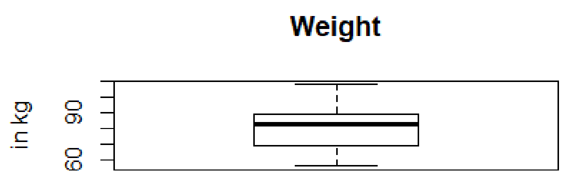
APPENDIX

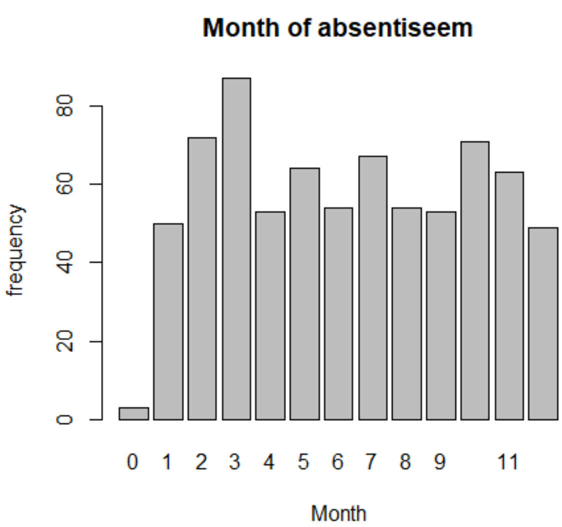

Distance from Residence to Work
Distance(km)
10 20 30 40 50

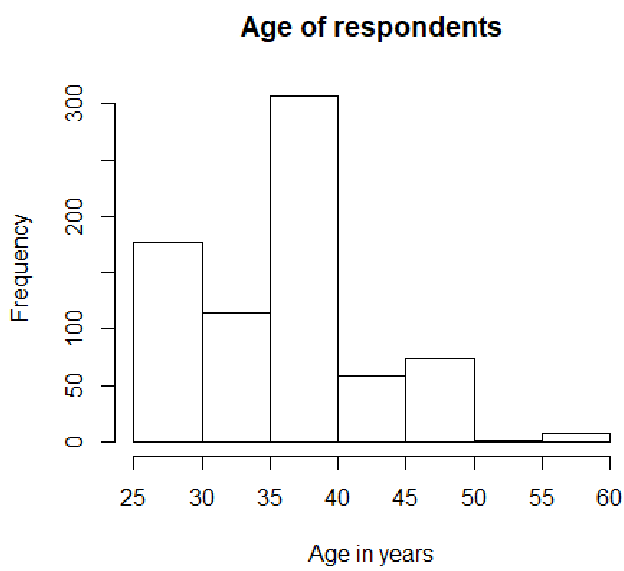

Son
Number of sons
0 1 2 3 4

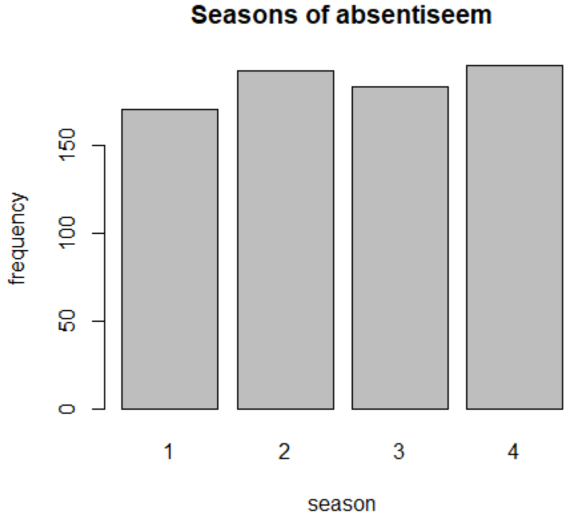

Body Mass Index
BMI
20 25 30 35

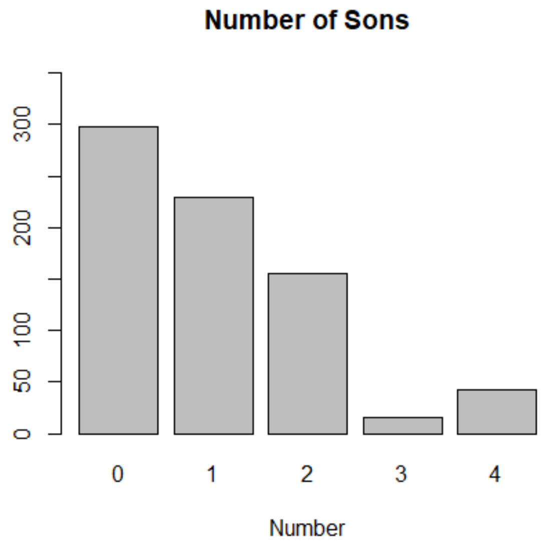

Height
in cm
165 190


Weight
in kg
60 90


Month of absentiseem
frequency
0 20 40 60 80
0 1 2 3 4 5 6 7 8 9 11
Month


Age of respondents
Frequency
0 50 100 200 300
25 30 35 40 45 50 55 60
Age in years






```
Call:
lm(formula = ds$Body.mass.index ~ ds$Height + ds$Weight)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.81144 -0.18143 -0.06357  0.21828  1.55572 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 53.116026   0.387409   137.1   <2e-16 ***
ds$Height   -0.312326   0.002353  -132.8   <2e-16 ***
ds$Weight    0.345631   0.001102   313.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3673 on 737 degrees of freedom
Multiple R-squared:  0.9927,    Adjusted R-squared:  0.9927 
F-statistic: 4.992e+04 on 2 and 737 DF,  p-value: < 2.2e-16
```

As the adjusted R^2 value = 0.9927 is high, there exists multicollinearity between BMI and Height and Weight. We decided to remove Height and Weight.

```
> cor(ds$Month.of.absence,ds$Seasons)
[1] 0.4075598
```

As the R^2 value = 0.4075598 is low, there is no multicollinearity between Months and Seasons.

```
> summary(ds)
       ID        Reason for absence Month of absence Day of the week    Seasons      Transportation expense Distance from Residence to Work  Service time
 Min.   : 1.00   Min.   : 0.00      Min.   : 0.000   Min.   :2.000   Min.   :1.000   Min.   :118.0          Min.   : 5.00                    Min.   : 1.00
 1st Qu.: 9.00   1st Qu.:13.00      1st Qu.: 3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:179.0          1st Qu.:16.00                    1st Qu.: 9.00
 Median :18.00   Median :23.00      Median : 6.000   Median :4.000   Median :3.000   Median :225.0          Median :26.00                    Median :13.00
 Mean   :18.02   Mean   :19.22      Mean   : 6.324   Mean   :3.915   Mean   :2.545   Mean   :221.3          Mean   :29.63                    Mean   :12.55
 3rd Qu.:28.00   3rd Qu.:26.00      3rd Qu.: 9.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:260.0          3rd Qu.:50.00                    3rd Qu.:16.00
 Max.   :36.00   Max.   :28.00      Max.   :12.000   Max.   :6.000   Max.   :4.000   Max.   :388.0          Max.   :52.00                    Max.   :29.00
      Age        Work load Average/day   Hit target     Disciplinary failure   Education          Son        Social drinker    Social smoker          Pet
 Min.   :27.00   Min.   :205917         Min.   : 81.00   Min.   :0.00000      Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:31.00   1st Qu.:244387         1st Qu.: 93.00   1st Qu.:0.00000      1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :37.00   Median :264249         Median : 95.00   Median :0.00000      Median :1.000   Median :1.000   Median :1.0000   Median :0.00000   Median :0.0000
 Mean   :36.45   Mean   :271490         Mean   : 94.59   Mean   :0.05405      Mean   :1.292   Mean   :1.019   Mean   :0.5676   Mean   :0.07297   Mean   :0.7459
 3rd Qu.:40.00   3rd Qu.:294217         3rd Qu.: 97.00   3rd Qu.:0.00000      3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
 Max.   :58.00   Max.   :378884         Max.   :100.00   Max.   :1.00000      Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.00000   Max.   :8.0000
     Weight           Height      Body mass index Absenteeism time in hours
 Min.   : 56.00   Min.   :163.0   Min.   :19.00   Min.   :  0.000
 1st Qu.: 69.00   1st Qu.:169.0   1st Qu.:24.00   1st Qu.:  2.000
 Median : 83.00   Median :170.0   Median :25.00   Median :  3.000
 Mean   : 79.04   Mean   :172.1   Mean   :26.68   Mean   :  6.924
 3rd Qu.: 89.00   3rd Qu.:172.0   3rd Qu.:31.00   3rd Qu.:  8.000
 Max.   :108.00   Max.   :196.0   Max.   :38.00   Max.   :120.000
```

```
> analysis <- lm( ds$`Absenteeism time in hours` ~ ds$`Day of the week` + ds$`Distance from Residence to Work`+
+                   ds$`Service time` + ds$Age + ds$`Work load Average/day` + ds$`Hit target` + ds$`Disciplinary failure` +
+                   ds$Education + ds$Son + ds$`Social drinker` + ds$`Social smoker`+ ds$Pet + ds$`Body mass index`)
> summary(analysis)

Call:
lm(formula = ds$`Absenteeism time in hours` ~ ds$`Day of the week` +
    ds$`Distance from Residence to Work` + ds$`Service time` +
    ds$Age + ds$`Work load Average/day` + ds$`Hit target` + ds$`Disciplinary failure` +
    ds$Education + ds$Son + ds$`Social drinker` + ds$`Social smoker` +
    ds$Pet + ds$`Body mass index`)

Residuals:
    Min      1Q  Median      3Q     Max
-14.262  -5.053  -2.090   0.866 108.243

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          8.552e+00  1.420e+01   0.602 0.547261
ds$`Day of the week`                -1.284e+00  3.429e-01  -3.745 0.000195 ***
ds$`Distance from Residence to Work` -1.155e-01  4.207e-02  -2.746 0.006174 **
ds$`Service time`                    2.165e-02  1.887e-01   0.115 0.908692
ds$Age                               1.798e-01  1.130e-01   1.591 0.112024
ds$`Work load Average/day`           5.545e-06  1.249e-05   0.444 0.657112
ds$`Hit target`                      7.534e-02  1.289e-01   0.584 0.559076
ds$`Disciplinary failure`           -8.309e+00  2.173e+00  -3.825 0.000142 ***
ds$Education                        -6.808e-01  8.439e-01  -0.807 0.420104
ds$Son                               1.143e+00  4.773e-01   2.395 0.016894 *
ds$`Social drinker`                  3.281e+00  1.266e+00   2.591 0.009758 **
ds$`Social smoker`                  -2.011e+00  2.015e+00  -0.998 0.318382
ds$Pet                               1.774e-01  4.540e-01   0.391 0.696178
ds$`Body mass index`                -3.867e-01  1.530e-01  -2.528 0.011698 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.91 on 726 degrees of freedom
Multiple R-squared:  0.07905,	Adjusted R-squared:  0.06255
F-statistic: 4.793 on 13 and 726 DF,  p-value: 4.776e-08
```

```
> analaysis.1 <- lm( ds$`Absenteeism time in hours` ~ ds$`Day of the week` + ds$`Distance from Residence to Work`+
+                    ds$`Disciplinary failure` + ds$Son + ds$`Social drinker` + ds$`Body mass index`)
> summary(analaysis.1)

Call:
lm(formula = ds$`Absenteeism time in hours` ~ ds$`Day of the week` +
    ds$`Distance from Residence to Work` + ds$`Disciplinary failure` +
    ds$Son + ds$`Social drinker` + ds$`Body mass index`)

Residuals:
    Min      1Q  Median      3Q     Max
-14.311  -5.098  -2.191   0.694 109.839

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          18.12771    3.65811   4.955 8.97e-07 ***
ds$`Day of the week`                 -1.24577    0.33941  -3.670 0.000260 ***
ds$`Distance from Residence to Work` -0.12896    0.03628  -3.555 0.000402 ***
ds$`Disciplinary failure`            -8.49000    2.11820  -4.008 6.75e-05 ***
ds$Son                                1.28504    0.45649   2.815 0.005008 **
ds$`Social drinker`                   3.83847    1.17091   3.278 0.001094 **
ds$`Body mass index`                 -0.20744    0.12132  -1.710 0.087703 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.9 on 733 degrees of freedom
Multiple R-squared:  0.07176,	Adjusted R-squared:  0.06416
F-statistic: 9.444 on 6 and 733 DF,  p-value: 5.256e-10
```

Regression with single variable
Absenteeism time in hours against Distance from Residence to Work

Absenteeism time in hours against Age

Absenteeism time in hours against Son