NUS

National University of Singapore

BT2101 Decision Making Methods and Tools

SEMESTER II 2019-2020

Assessment of Machine Learning models on predicting Absenteeism from work

BT2101 Group 4:

Andy Low Wei Liang	A0187919H	(e0323503@u.nus.edu)
Aster Ooi	A0187855J	(aster.ooi@u.nus.edu)
Byron Yip Pak Lun	A0200090A	(e0407071@u.nus.edu)
Chee Zhong Quan	A0183873R	(e0310668@u.nus.edu)
Edwin Lim Ze Xin	A0203808J	(e0421014@u.nus.edu)

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.

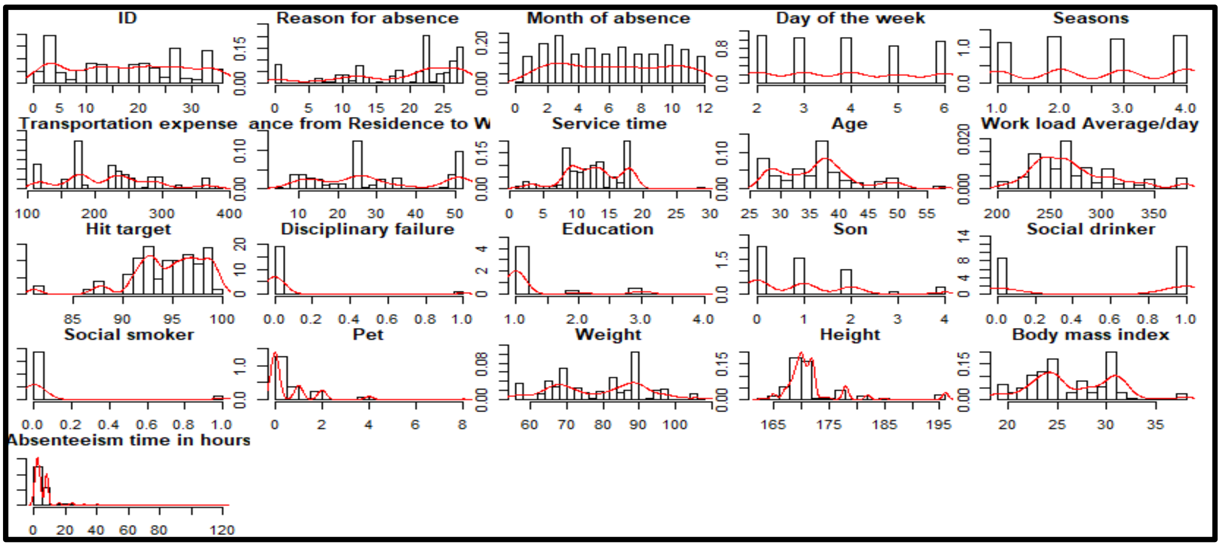
2.1 Data Overview

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.

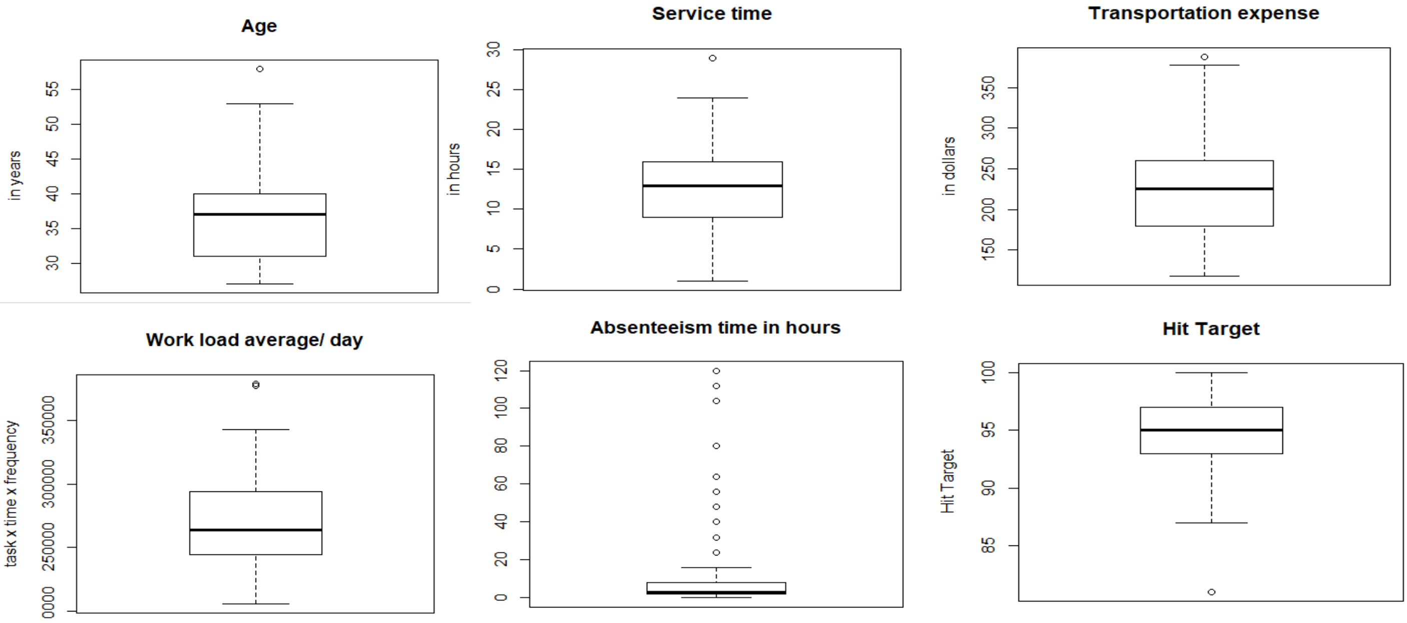
2.3 Checking for categorical data



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

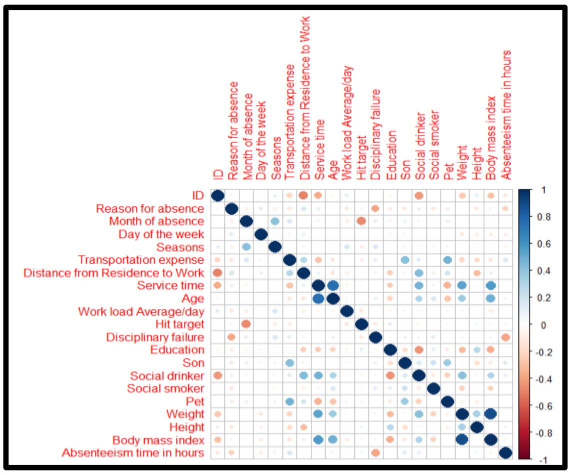
The box plots show the distribution of the continuous variables based on the maximum and minimum values. It shows the median, first quartile and the third quartile.



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		

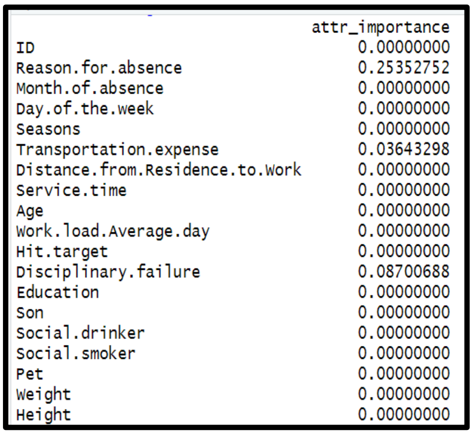
For the continuous variables, we can observe that all of them have a p-value less than 0.05 so we can conclude that they are all statistically significant. As for the categorical variables, the variables, “Day of the week”, “Education” and “Social smoker” have a p-value greater than 0.05 so we can conclude 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
```

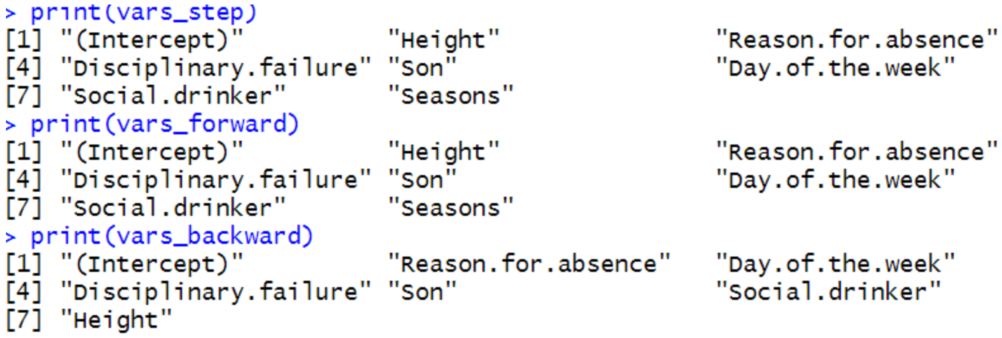
4.3 Feature Selection using Wrapper Methods

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.

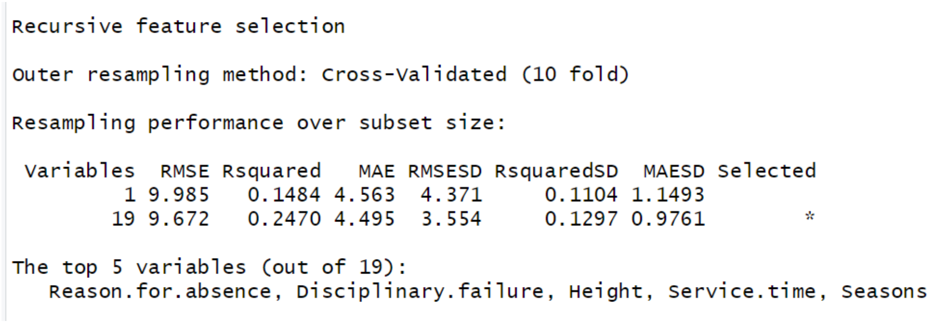
Output:
The following variables were selected from the stepwise regression selection.



```
> 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.

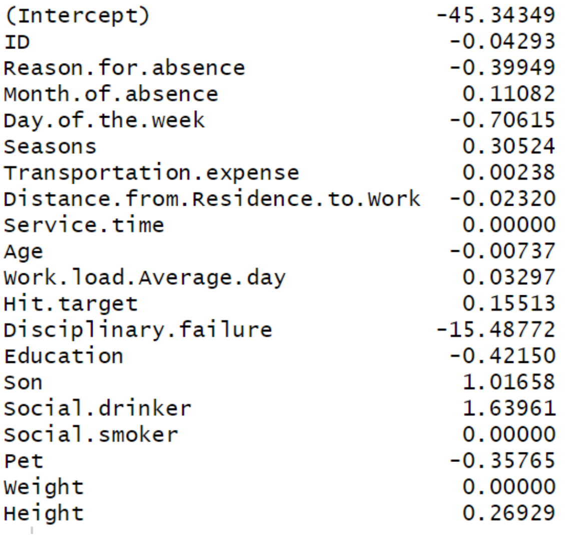


```
Recursive feature selection

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
```

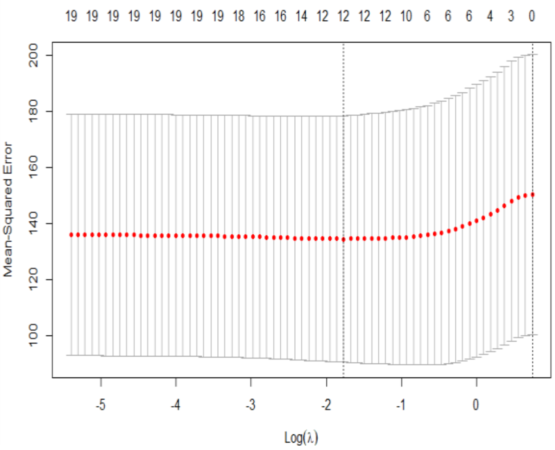
The results of the RFE shows that the highest accuracy rate consists of the following 5 variables: Disciplinary.failure, Reason.for.absence,Height, Service.time and Seasons.

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.

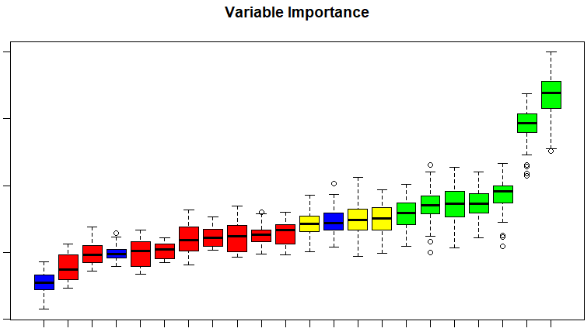
1. Service.time
2. Social.smoker
3. Weight

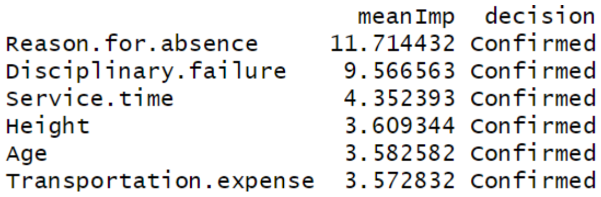
The remaining variables would then be considered in our model.

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.



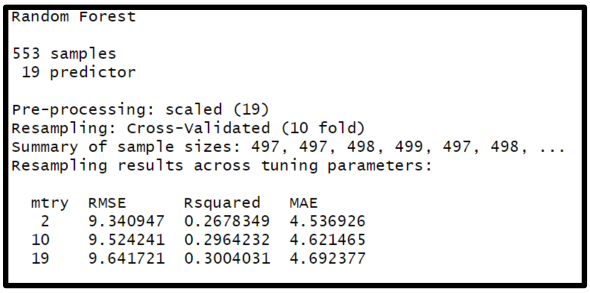


```
                         meanImp  decision
Reason.for.absence     11.714432 Confirmed
Disciplinary.failure    9.566563 Confirmed
Service.time            4.352393 Confirmed
Height                  3.609344 Confirmed
Age                     3.582582 Confirmed
Transportation.expense  3.572832 Confirmed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

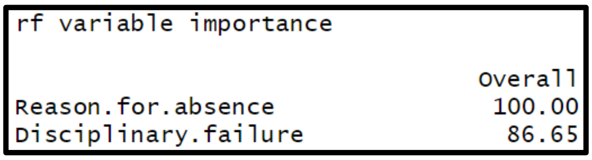
From the results as shown in the figure above , we can see that the Boruta model confirmed the following 6 variables: Reasons for absence, Disciplinary Failure, Height , Age and Transportation.expense.

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 Random Forest achieved an optimal model with the following variables: Reason for Absence and Disciplinary Failure.

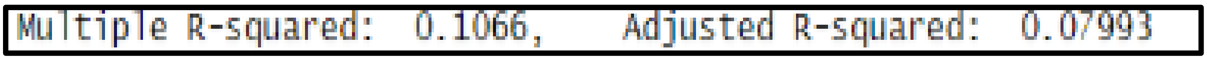
4.5 Comparison Between Methods

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	**Both:** Height, Reason.for.absence, Disciplinary.failure, Son, Social.drinker, Day.of.the.week, Seasons **Forward:** Height, Reason.for.absence, Disciplinary.failure, Son, Social.drinker, Day.of.the.week, Seasons **Backward:** 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

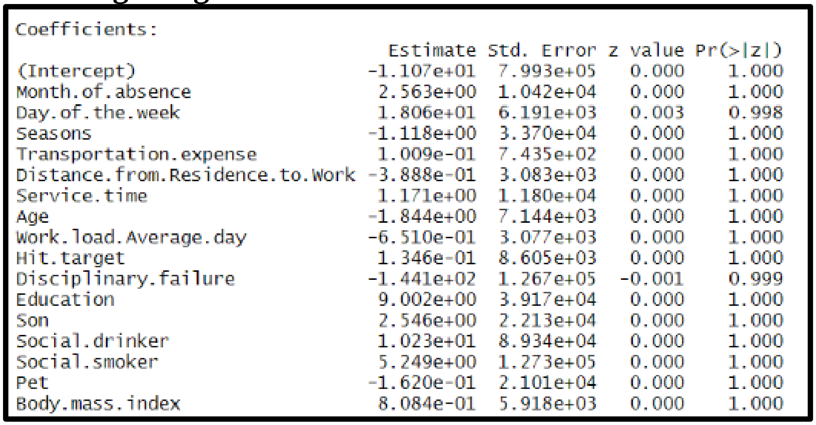
5.1 Linear Regression



```
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

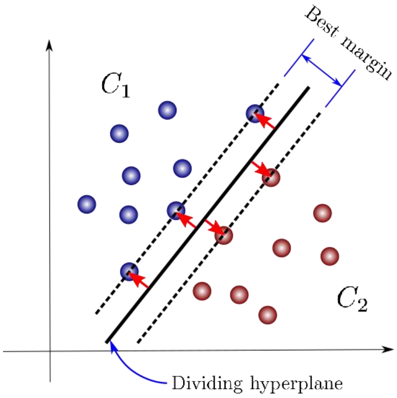


```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.107e+01  7.993e+05   0.000    1.000
Month.of.absence            2.563e+00  1.042e+04   0.000    1.000
Day.of.the.week             1.806e+01  6.191e+03   0.003    0.998
Seasons                    -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:

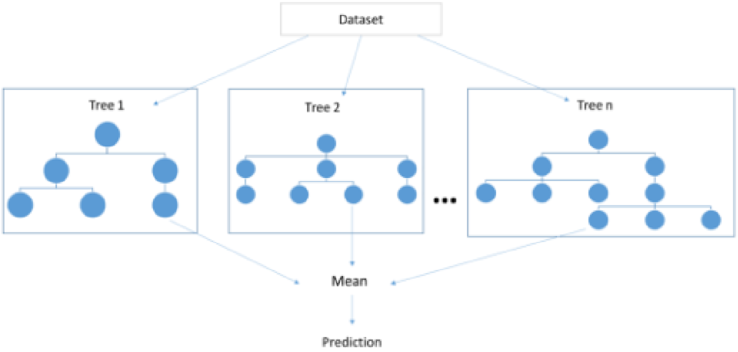
Kernel	Area Under Curve	Test Accuracy
Linear	29.10305%	43.16547%
Polynomial	44.92537%	41.72662%
Radial	40.8874%	44.60432%
Sigmoid	35.40026%	41.00719%

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)



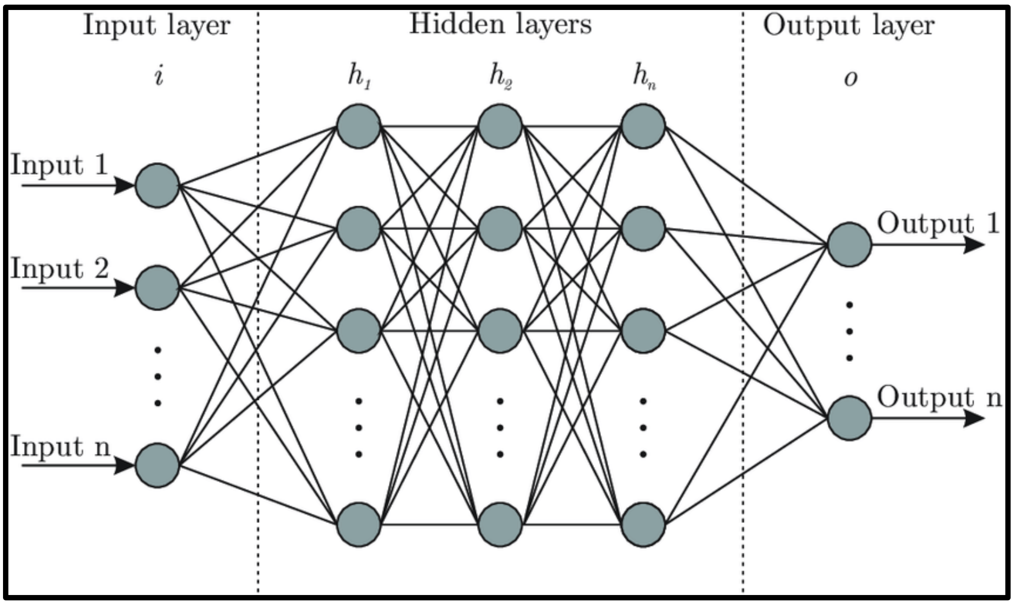
Random Forest consists of many individual decision trees that operate together. Each decision tree represents an independent variable and generates a class prediction, which contributes to a vote in the final prediction. Since decision trees are highly sensitive to the data they are trained on, small 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.

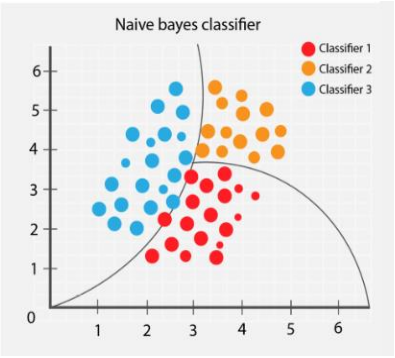
The constant process of forward and backward propagation is conducted iteratively for all data in the training set. The larger the dataset, the more the neural network will learn, and therefore the more accurate the algorithm will be at forecasting outputs.

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.

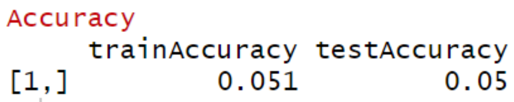
5.5 Naive Bayes Classifier



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

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.

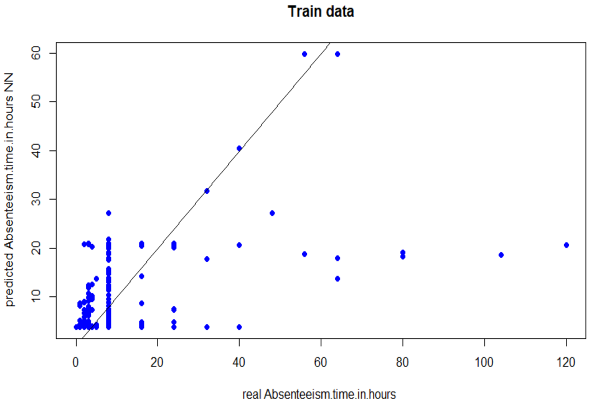
This table shows the RMSE/ Accuracy of the data with the stratified K-fold cross validations:

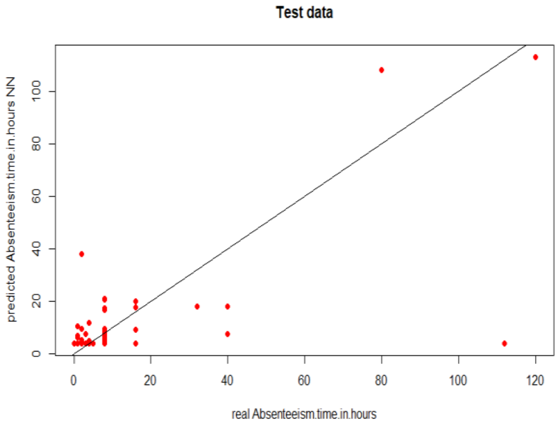
Model	**RSME (For Neural Network)/Accuracy (For SVM)**			
	2-fold	**5-fold**	**10-fold**	**15-fold**
Neural Network	Test- 20.07973 Train- 6.061027	Test-11.48549 Train-9.197887	Test-20.39662 Train-9.485357	Test-24.51548 Train-9.372849
SVM (Sigmoid)	Test- 39.88439% Train- 47.97101%	Test- 33.09353% Train-77.89855%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

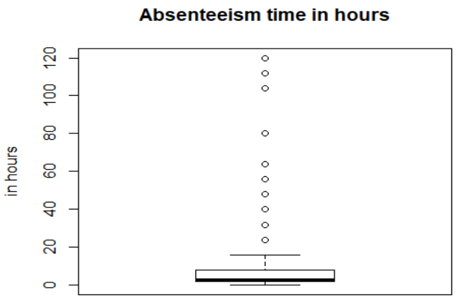
Firstly, the Neural Network Model (with 2 hidden layers with 6 nodes) with features selected by Hypothesis Testing. The model returned a low RMSE of 11.48549 in the training set and 9.197887 in the test set. The low RMSE and RMSE difference between train and test signifies a high predictive power of 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 pre-processed. 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