# Employee Absenteeism

Eeshan Gupta

28 *September* 2018

# Contents

# Chapter 1

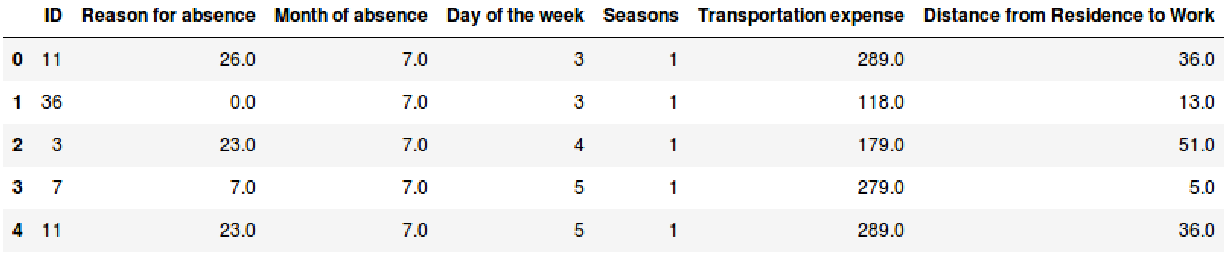
# Introduction

## 1.1 Problem Statement

The human capital plays an important role in collection, transportation and delivery for a company. The issue absenteeism is an important concern for the company. The aim of the project is to find the changes in the company human resource policy to reduce the number of absenteeism. The company also wants to calculate the losses if the current trend continues. The company want a model built which suggests the strategy that company should follow to reduce the absenteeism numbers.

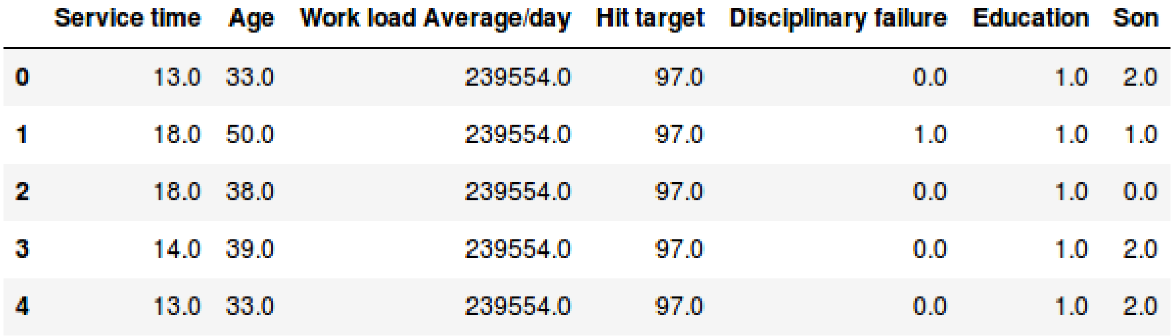
## 1.2 Data

Our Task is to build a model which help us identify the main reasons behind the absent employees and the duration of their absence. Given below is the data that is provided to us for the purpose of our analysis:



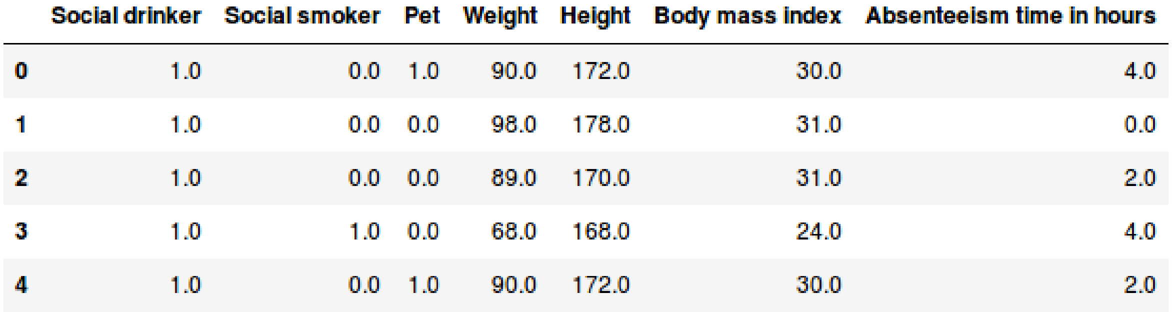
| | ID | Reason for absence | Month of absence | Day of the week | Seasons | Transportation expense | Distance from Residence to Work |
|---|---|---|---|---|---|---|---|
| **0** | 11 | 26.0 | 7.0 | 3 | 1 | 289.0 | 36.0 |
| **1** | 36 | 0.0 | 7.0 | 3 | 1 | 118.0 | 13.0 |
| **2** | 3 | 23.0 | 7.0 | 4 | 1 | 179.0 | 51.0 |
| **3** | 7 | 7.0 | 7.0 | 5 | 1 | 279.0 | 5.0 |
| **4** | 11 | 23.0 | 7.0 | 5 | 1 | 289.0 | 36.0 |

Figure 1.1: Columns 1 to 7



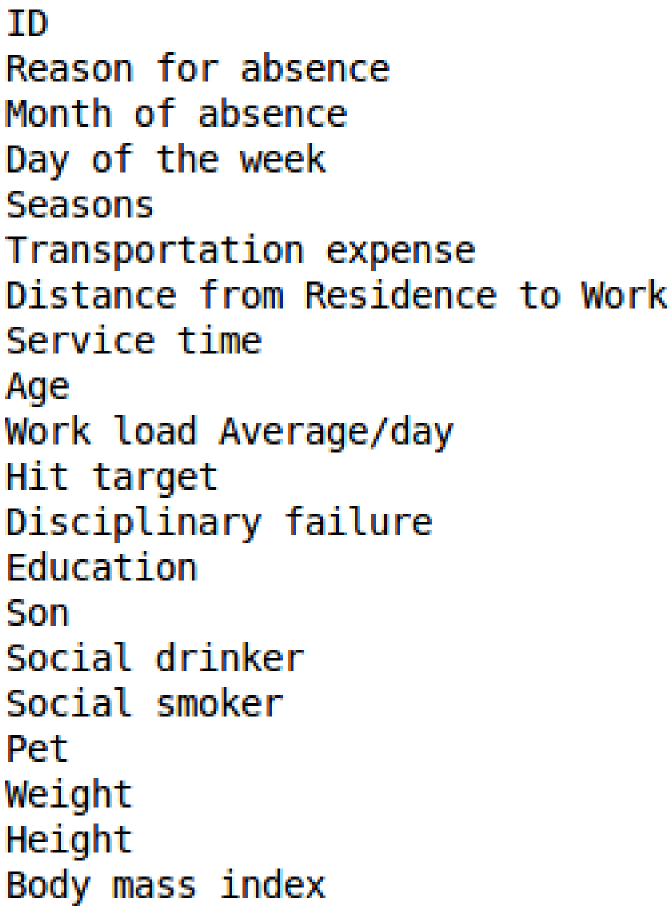
| | Service time | Age | Work load Average/day | Hit target | Disciplinary failure | Education | Son |
|---|---|---|---|---|---|---|---|
| **0** | 13.0 | 33.0 | 239554.0 | 97.0 | 0.0 | 1.0 | 2.0 |
| **1** | 18.0 | 50.0 | 239554.0 | 97.0 | 1.0 | 1.0 | 1.0 |
| **2** | 18.0 | 38.0 | 239554.0 | 97.0 | 0.0 | 1.0 | 0.0 |
| **3** | 14.0 | 39.0 | 239554.0 | 97.0 | 0.0 | 1.0 | 2.0 |
| **4** | 13.0 | 33.0 | 239554.0 | 97.0 | 0.0 | 1.0 | 2.0 |

Figure 1.2: Columns 8 to 14



| | Social drinker | Social smoker | Pet | Weight | Height | Body mass index | Absenteeism time in hours |
|---|---|---|---|---|---|---|---|
| **0** | 1.0 | 0.0 | 1.0 | 90.0 | 172.0 | 30.0 | 4.0 |
| **1** | 1.0 | 0.0 | 0.0 | 98.0 | 178.0 | 31.0 | 0.0 |
| **2** | 1.0 | 0.0 | 0.0 | 89.0 | 170.0 | 31.0 | 2.0 |
| **3** | 1.0 | 1.0 | 0.0 | 68.0 | 168.0 | 24.0 | 4.0 |
| **4** | 1.0 | 0.0 | 1.0 | 90.0 | 172.0 | 30.0 | 2.0 |

Figure 1.3: Columns 15 to 21



```
ID
Reason for absence
Month of absence
Day of the week
Seasons
Transportation expense
Distance from Residence to Work
Service time
Age
Work load Average/day
Hit target
Disciplinary failure
Education
Son
Social drinker
Social smoker
Pet
Weight
Height
Body mass index
```

Figure 1.4: Predicting Variables

It is clear from the data that there are 20 variables, which can be used to predict the total number of hours of absenteeism.

# Chapter 2

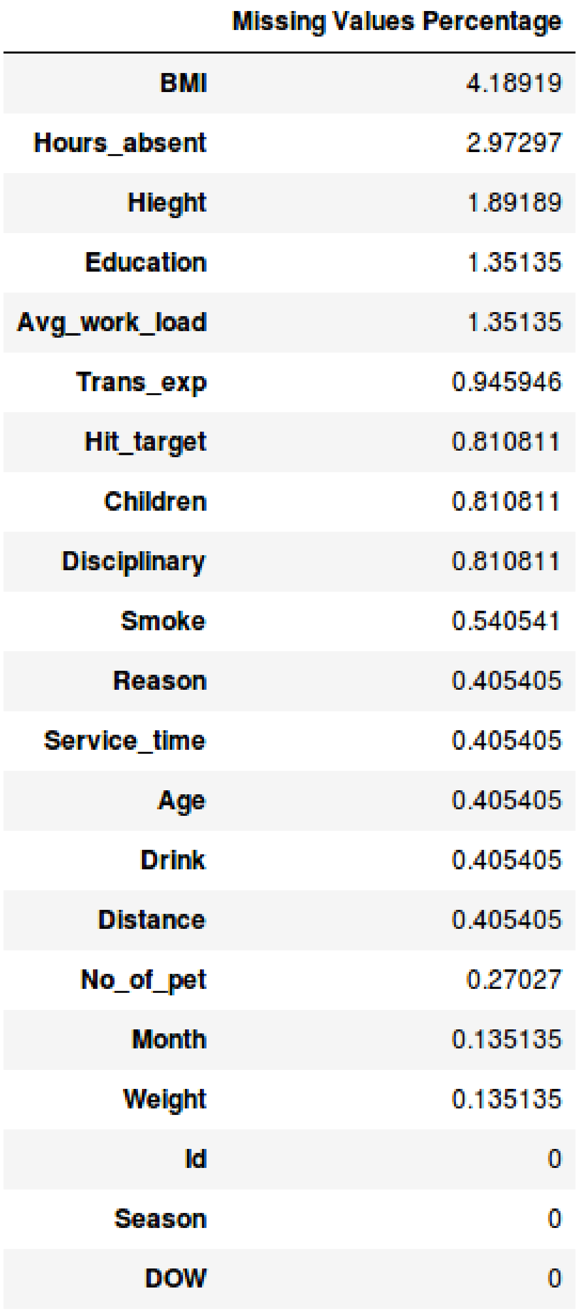
# Methodology

## 2.1 Pre-processing

The data we obtained for our analysis, is not clean i.e. not yet ready to use the data as an input to any algorithm. If such data is used, it may lead to bad results.
For the purposes of ease, the column names are changed.

### 2.1.1 Missing Value Analysis

Missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.
When missing value analysis is applied on our data the results were as in figure 2.1. The missing values can either be imputed i.e. replaced or the data points can be removed. The missing values in our analysis are imputed by the mean of rest of the values in the variable which are continuous and median of the values in the variable which are categorical.
In our analysis the missing values for each variable are denoted by NA. These values are then imputed by either mean or median of the rest of the values in the feature.



| | Missing Values Percentage |
|---|---|
| BMI | 4.18919 |
| Hours_absent | 2.97297 |
| Hieght | 1.89189 |
| Education | 1.35135 |
| Avg_work_load | 1.35135 |
| Trans_exp | 0.945946 |
| Hit_target | 0.810811 |
| Children | 0.810811 |
| Disciplinary | 0.810811 |
| Smoke | 0.540541 |
| Reason | 0.405405 |
| Service_time | 0.405405 |
| Age | 0.405405 |
| Drink | 0.405405 |
| Distance | 0.405405 |
| No_of_pet | 0.27027 |
| Month | 0.135135 |
| Weight | 0.135135 |
| Id | 0 |
| Season | 0 |
| DOW | 0 |

Figure 2.1: Missing Percentage

### 2.1.2 Outlier Analysis

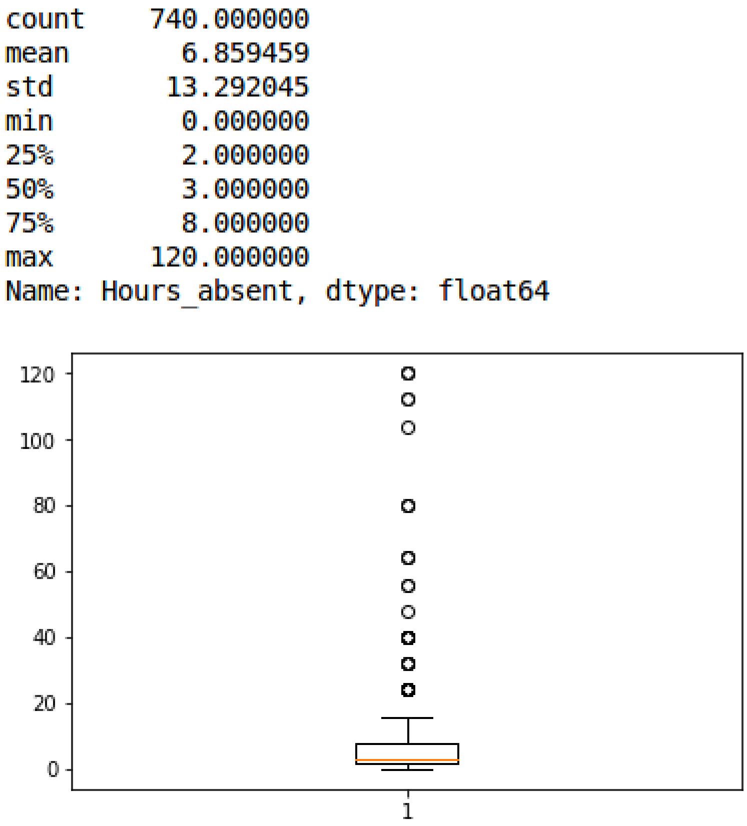
An outlier is an observation that appears to deviate markedly from other observations in the sample.
Identification of potential outliers is important for the following reasons.

1. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

2. In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation. However, if the data contains significant outliers, we may need to consider the use of robust statistical techniques.

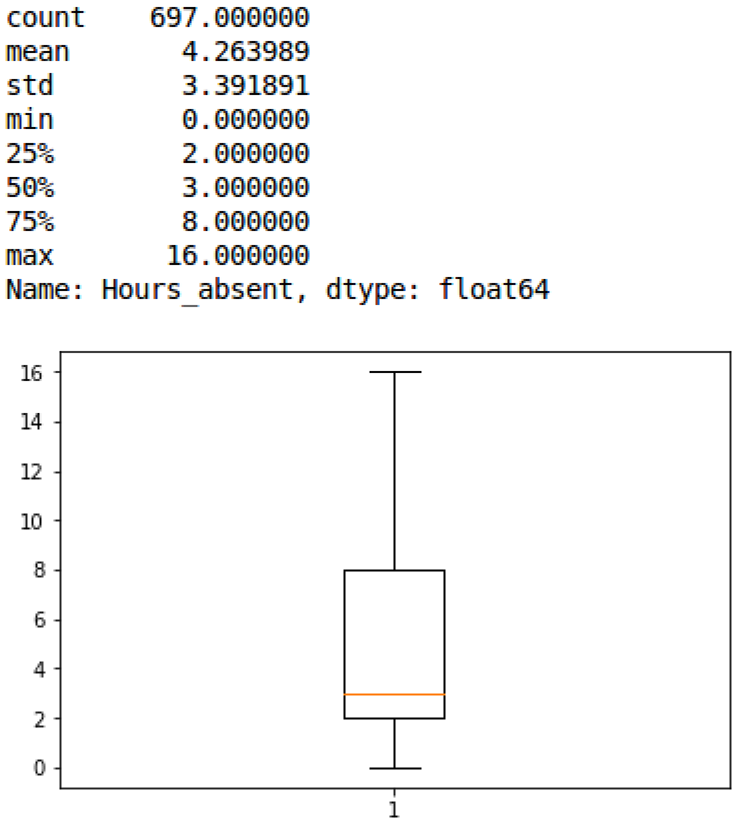
The data for our analysis consists of outliers in not many features. The most important variable i.e. the target variable is one with most outliers.
Technically, the outliers are separated from the other observations by the means of the 5 point summary. The minimum, the maximum, the median, the first element of the first quarter and the last element of the third quarter are calculated. The points lying above the maximum and the point lying below the minimum are termed as outliers.

In our analysis, the outliers for the target variable are removed from the dataset. This removal lead to better results.



```
count    740.000000
mean       6.859459
std       13.292045
min        0.000000
25%        2.000000
50%        3.000000
75%        8.000000
max      120.000000
Name: Hours_absent, dtype: float64
```

Figure 2.2: Boxplot of the target variable using the 5 point summary (Before removing the points)



```
count    697.000000
mean       4.263989
std        3.391891
min        0.000000
25%        2.000000
50%        3.000000
75%        8.000000
max       16.000000
Name: Hours_absent, dtype: float64
```

Figure 2.3: Boxplot of the target variable using the 5 point summary (After removing the points)

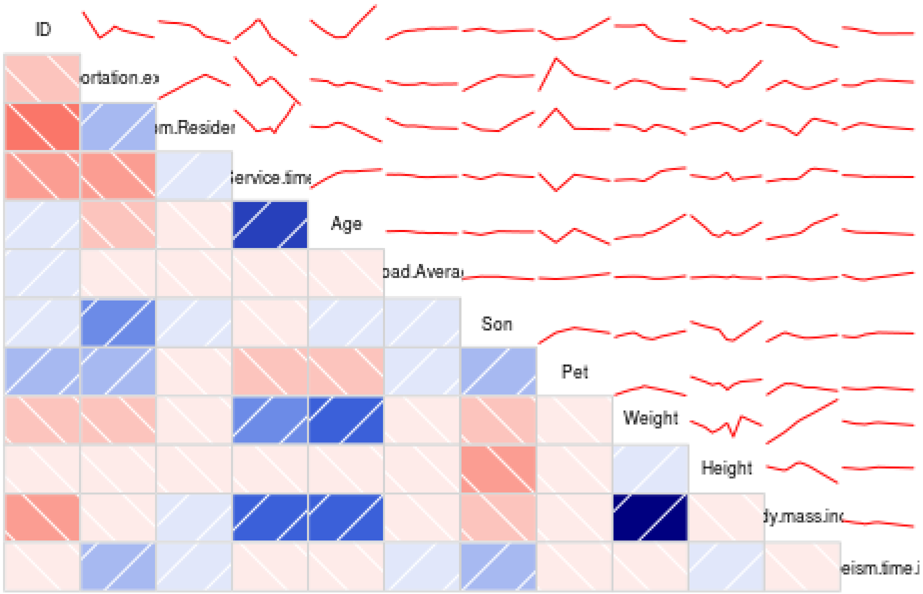


Figure 2.4: Heatmap of the correlation in between the features

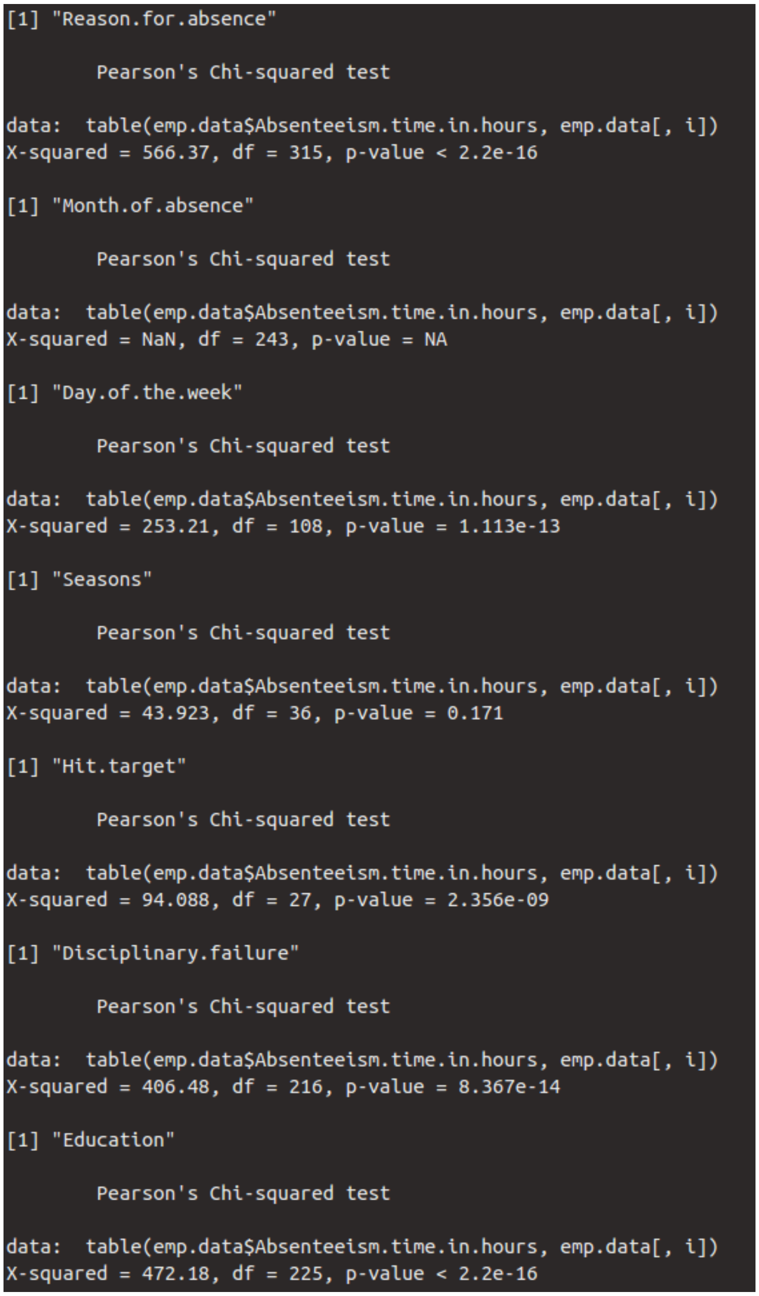
## 2.2 Feature Selection and Scaling

All the features in our data do not affect the target variable in a same fashion. When the features which do not affect the target variables are removed from the analysis, the results of the analysis gets better and more precise. This is called Feature Selection.

We, in this analysis have used 2 methods to find which feature affect the target variable. For the continuous features we find the correlation of the each variable. This shows whether the variable is linearly related to the another variable. Figure 2.4 is a heat-map showing the correlation values and the type of relation in between the features. The blue region in the heat-map represents a positive correlation and the red region represents a negative one. The intensity of the colours represent the value of the correlation i.e. darker the shade of the colour, more correlated the features are.

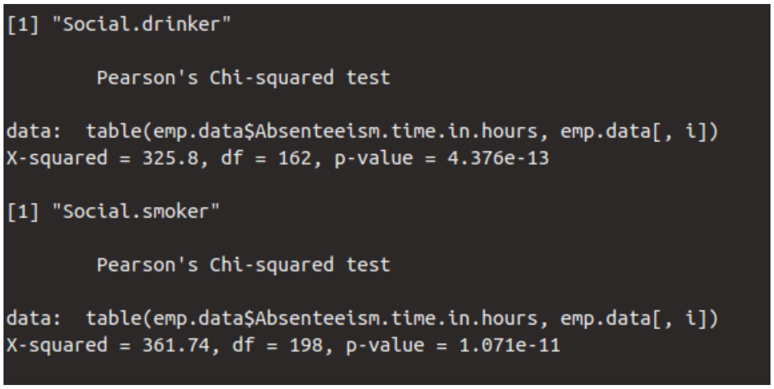
The categorical features are not chosen with the correlation values. There is a special statistical technique to find the usability of a categorical variable, known as chi-squared test. The chi-square test is always testing the null hypothesis, which states that there is no significant difference between the expected and observed result. Chi-square results are interpreted using p-value. P-value is calculated using the Chi-square value and degree of freedom. Figure 2.5 and 2.6 shows the calculated p values for the categorical features of the data.

The features like the distance, no of pets, weight, height, body mass index, hit target, and ID are removed from our analysis because the correlation values are near 0 and the p values are very low.



```
[1] "Reason.for.absence"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 566.37, df = 315, p-value < 2.2e-16

[1] "Month.of.absence"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = NaN, df = 243, p-value = NA

[1] "Day.of.the.week"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 253.21, df = 108, p-value = 1.113e-13

[1] "Seasons"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 43.923, df = 36, p-value = 0.171

[1] "Hit.target"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 94.088, df = 27, p-value = 2.356e-09

[1] "Disciplinary.failure"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 406.48, df = 216, p-value = 8.367e-14

[1] "Education"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 472.18, df = 225, p-value < 2.2e-16
```

Figure 2.5: Chi-squared Test



```
[1] "Social.drinker"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 325.8, df = 162, p-value = 4.376e-13

[1] "Social.smoker"

	Pearson's Chi-squared test

data:  table(emp.data$Absenteeism.time.in.hours, emp.data[, i])
X-squared = 361.74, df = 198, p-value = 1.071e-11
```

Figure 2.6: Chi-squared Test

### 2.2.1 Feature Scaling

The continuous features may have a very different range. For example, the age ranges from lets suppose, 18 - 56 years, but the hours of absenteeism ranges from 0 - 16 hours. If all the continuous features are brought in the same range the model will provide with better results. So all features are brought in the range of 0-1 by normalizing them using equation 2.1.

$$x' = \frac{x - min(X)}{max(X) - min(X)} \tag{2.1}$$

where $X$ is a feature, $max(.)$ and $min(.)$ are the function which return the maximum and the minimum values of the feature. $x$ represent the current data point value for the $X$ feature and $x'$ is the normalized value.

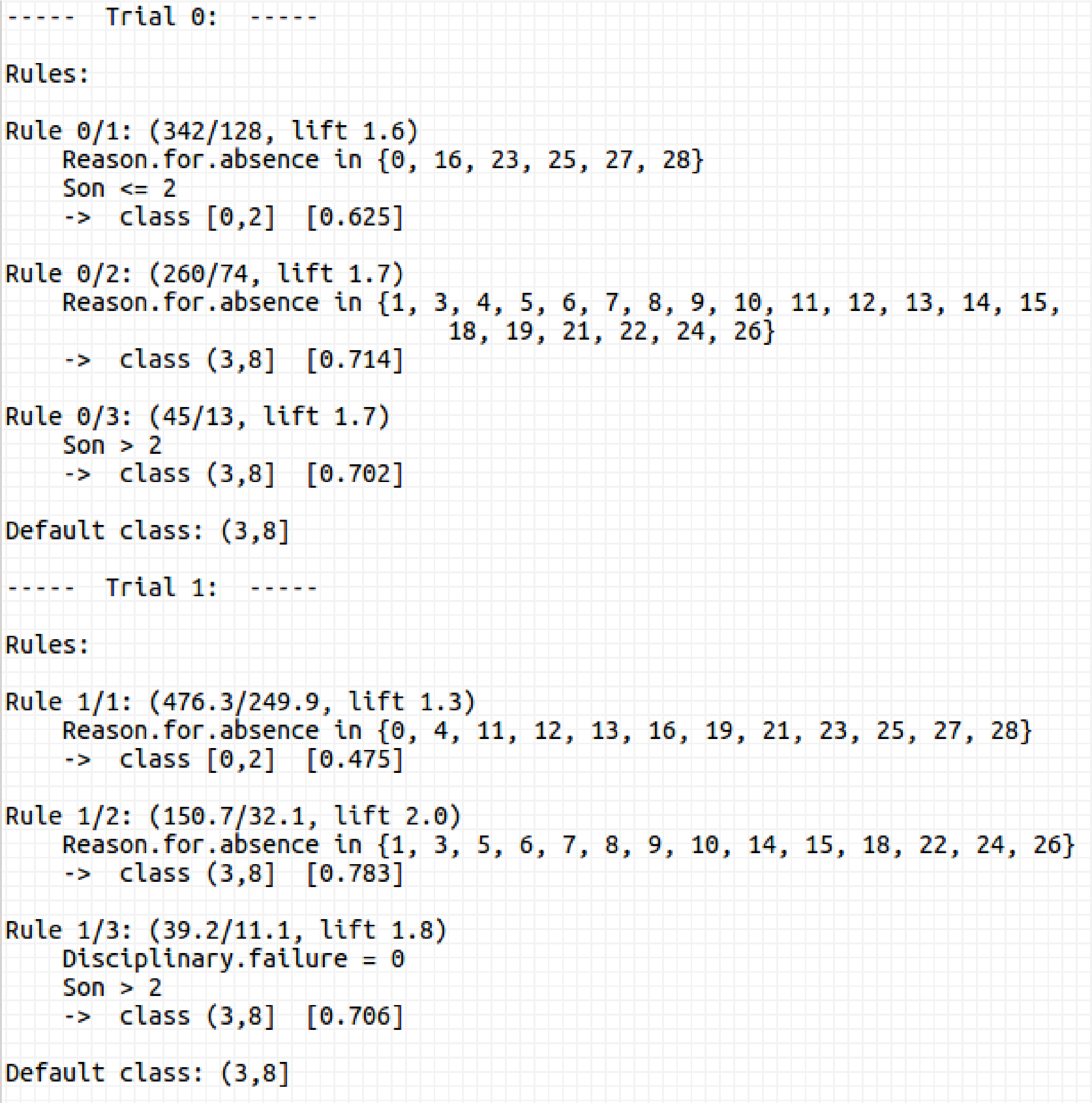
## 2.3 Model

The next and crucial step in our analysis is the development of the model. The model refers to the algorithm that we apply on the data, which is cleaned and pre-processed. The output of the algorithm is a model that has learned the rules of the data.

The data is first split into training data and testing data. The training data is used to train the model and get the rules of the data. The testing data is used for checking the precision of the rules. It is a necessary step as the trained model needs a check on its learning.
In our model, first step is to covert our target variable, which is continuous variable, to a categorical variable. This helps us to apply classification algorithm and rule mining algorithm on the data. The algorithm to train our model in C5.0.

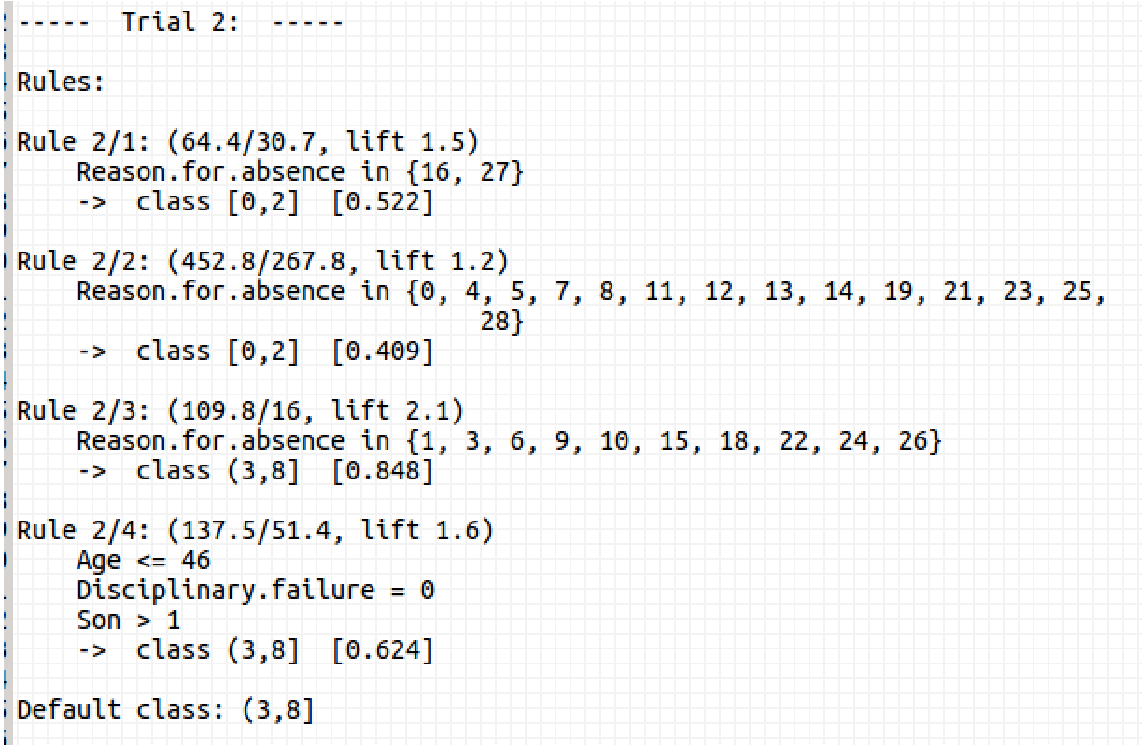
C5.0 learns the rule by building decision trees from the training data and then verifying the rules hence formed by the testing data.The decision tree calculates the feasibility of a classification rule.

### 2.3.1 Results



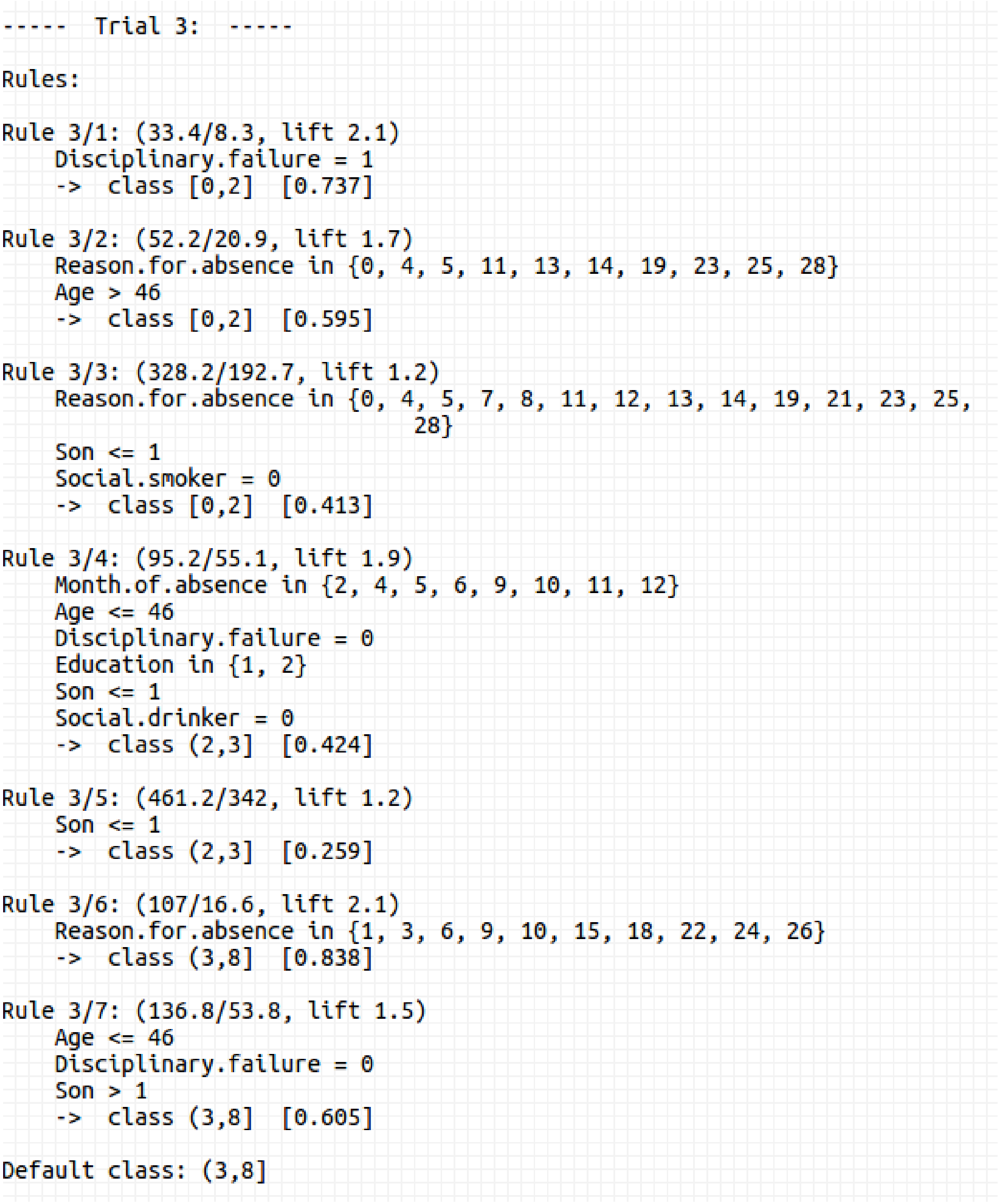
```
-----  Trial 0:  -----

Rules:

Rule 0/1: (342/128, lift 1.6)
    Reason.for.absence in {0, 16, 23, 25, 27, 28}
    Son <= 2
    ->  class [0,2]  [0.625]

Rule 0/2: (260/74, lift 1.7)
    Reason.for.absence in {1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
                           18, 19, 21, 22, 24, 26}
    ->  class (3,8]  [0.714]

Rule 0/3: (45/13, lift 1.7)
    Son > 2
    ->  class (3,8]  [0.702]

Default class: (3,8]

-----  Trial 1:  -----

Rules:

Rule 1/1: (476.3/249.9, lift 1.3)
    Reason.for.absence in {0, 4, 11, 12, 13, 16, 19, 21, 23, 25, 27, 28}
    ->  class [0,2]  [0.475]

Rule 1/2: (150.7/32.1, lift 2.0)
    Reason.for.absence in {1, 3, 5, 6, 7, 8, 9, 10, 14, 15, 18, 22, 24, 26}
    ->  class (3,8]  [0.783]

Rule 1/3: (39.2/11.1, lift 1.8)
    Disciplinary.failure = 0
    Son > 2
    ->  class (3,8]  [0.706]

Default class: (3,8]
```

Figure 2.7: Rules



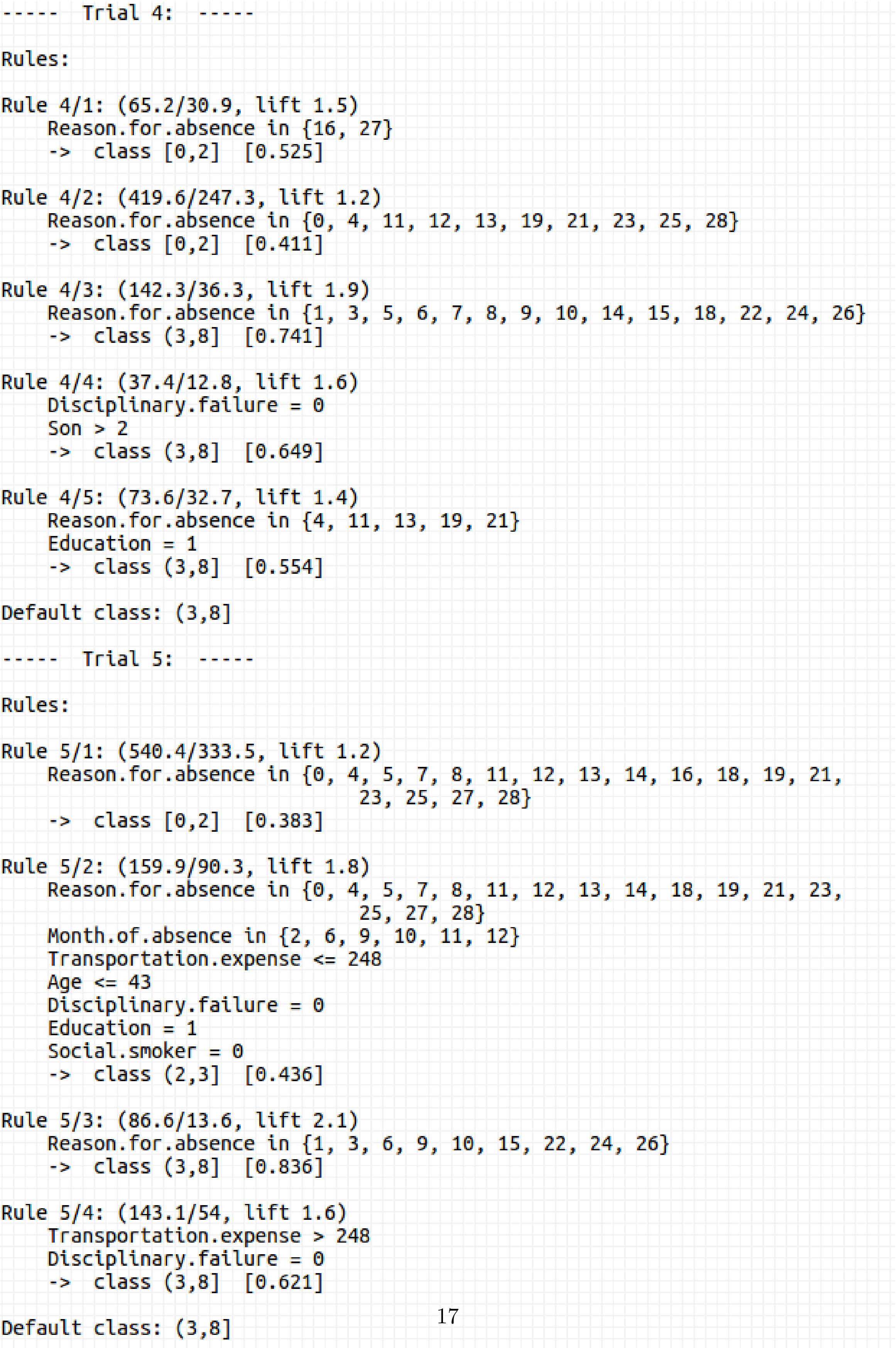
```
-----  Trial 2:  -----

Rules:

Rule 2/1: (64.4/30.7, lift 1.5)
	Reason.for.absence in {16, 27}
	->  class [0,2]  [0.522]

Rule 2/2: (452.8/267.8, lift 1.2)
	Reason.for.absence in {0, 4, 5, 7, 8, 11, 12, 13, 14, 19, 21, 23, 25,
	                       28}
	->  class [0,2]  [0.409]

Rule 2/3: (109.8/16, lift 2.1)
	Reason.for.absence in {1, 3, 6, 9, 10, 15, 18, 22, 24, 26}
	->  class (3,8]  [0.848]

Rule 2/4: (137.5/51.4, lift 1.6)
	Age <= 46
	Disciplinary.failure = 0
	Son > 1
	->  class (3,8]  [0.624]

Default class: (3,8]
```

Figure 2.8: Rules



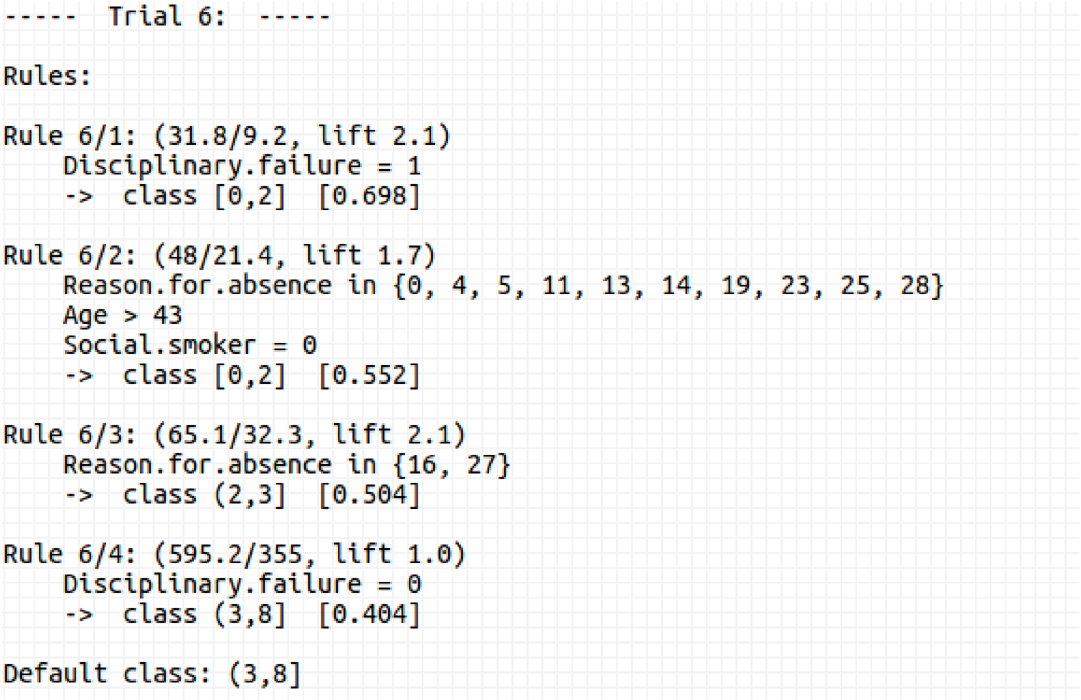
```
-----  Trial 3:  -----

Rules:

Rule 3/1: (33.4/8.3, lift 2.1)
    Disciplinary.failure = 1
    ->  class [0,2]  [0.737]

Rule 3/2: (52.2/20.9, lift 1.7)
    Reason.for.absence in {0, 4, 5, 11, 13, 14, 19, 23, 25, 28}
    Age > 46
    ->  class [0,2]  [0.595]

Rule 3/3: (328.2/192.7, lift 1.2)
    Reason.for.absence in {0, 4, 5, 7, 8, 11, 12, 13, 14, 19, 21, 23, 25,
                               28}
    Son <= 1
    Social.smoker = 0
    ->  class [0,2]  [0.413]

Rule 3/4: (95.2/55.1, lift 1.9)
    Month.of.absence in {2, 4, 5, 6, 9, 10, 11, 12}
    Age <= 46
    Disciplinary.failure = 0
    Education in {1, 2}
    Son <= 1
    Social.drinker = 0
    ->  class (2,3]  [0.424]

Rule 3/5: (461.2/342, lift 1.2)
    Son <= 1
    ->  class (2,3]  [0.259]

Rule 3/6: (107/16.6, lift 2.1)
    Reason.for.absence in {1, 3, 6, 9, 10, 15, 18, 22, 24, 26}
    ->  class (3,8]  [0.838]

Rule 3/7: (136.8/53.8, lift 1.5)
    Age <= 46
    Disciplinary.failure = 0
    Son > 1
    ->  class (3,8]  [0.605]

Default class: (3,8]
```

Figure 2.9: Rules



```
-----  Trial 4:  -----

Rules:

Rule 4/1: (65.2/30.9, lift 1.5)
    Reason.for.absence in {16, 27}
    ->  class [0,2]  [0.525]

Rule 4/2: (419.6/247.3, lift 1.2)
    Reason.for.absence in {0, 4, 11, 12, 13, 19, 21, 23, 25, 28}
    ->  class [0,2]  [0.411]

Rule 4/3: (142.3/36.3, lift 1.9)
    Reason.for.absence in {1, 3, 5, 6, 7, 8, 9, 10, 14, 15, 18, 22, 24, 26}
    ->  class (3,8]  [0.741]

Rule 4/4: (37.4/12.8, lift 1.6)
    Disciplinary.failure = 0
    Son > 2
    ->  class (3,8]  [0.649]

Rule 4/5: (73.6/32.7, lift 1.4)
    Reason.for.absence in {4, 11, 13, 19, 21}
    Education = 1
    ->  class (3,8]  [0.554]

Default class: (3,8]

-----  Trial 5:  -----

Rules:

Rule 5/1: (540.4/333.5, lift 1.2)
    Reason.for.absence in {0, 4, 5, 7, 8, 11, 12, 13, 14, 16, 18, 19, 21,
                           23, 25, 27, 28}
    ->  class [0,2]  [0.383]

Rule 5/2: (159.9/90.3, lift 1.8)
    Reason.for.absence in {0, 4, 5, 7, 8, 11, 12, 13, 14, 18, 19, 21, 23,
                           25, 27, 28}
    Month.of.absence in {2, 6, 9, 10, 11, 12}
    Transportation.expense <= 248
    Age <= 43
    Disciplinary.failure = 0
    Education = 1
    Social.smoker = 0
    ->  class (2,3]  [0.436]

Rule 5/3: (86.6/13.6, lift 2.1)
    Reason.for.absence in {1, 3, 6, 9, 10, 15, 22, 24, 26}
    ->  class (3,8]  [0.836]

Rule 5/4: (143.1/54, lift 1.6)
    Transportation.expense > 248
    Disciplinary.failure = 0
    ->  class (3,8]  [0.621]

Default class: (3,8]
```

Figure 2.10: Rules



```
-----  Trial 6:  -----

Rules:

Rule 6/1: (31.8/9.2, lift 2.1)
    Disciplinary.failure = 1
    ->  class [0,2]  [0.698]

Rule 6/2: (48/21.4, lift 1.7)
    Reason.for.absence in {0, 4, 5, 11, 13, 14, 19, 23, 25, 28}
    Age > 43
    Social.smoker = 0
    ->  class [0,2]  [0.552]

Rule 6/3: (65.1/32.3, lift 2.1)
    Reason.for.absence in {16, 27}
    ->  class (2,3]  [0.504]

Rule 6/4: (595.2/355, lift 1.0)
    Disciplinary.failure = 0
    ->  class (3,8]  [0.404]

Default class: (3,8]
```

Figure 2.11: Rules