**EMPLOYEE ABSENTEEISM PREDICTION**

**PRIYANK KUMAR**

# Contents

## 1. Introduction

1.1. Problem Statement

1.2. Data

## 2. Methodology

2.1. Pre Processing

2.1.1. Missing Value Analysis

2.1.2. Feature Selection

2.1.3. Feature Scaling

2.2. Modelling

2.2.1. Model Selection

2.2.2. k-Nearest Neighbours

2.2.3. Decision Trees

2.2.4. Random Forest

2.2.5 Linear Regression

## 3. Conclusion

3.1. Model Evaluation

3.2 Model Selection

3.3 Summary

Chapter 1

Introduction

### 1.1 Problem Statement

Employee absenteeism is a major problem faced by every employer which eventually lead to the backlogs, piling of the work, delay in deploying the project and can have a major effect on company finances. The aim of this project is to find an issue which eventually leads toward the absence of an employee and provide a proper solution to reduce the absenteeism in the company.

### 1.2 Data

The task is to build a regression model which will predict the number of absent hours in relation to the employee depend upon the various factor.

Given below is the sample of the data set which is used to predict the number of absent hours related to the employee.

Table 1.1: Employee Absenteeism Sample Data (Columns: 1-8)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Reason for absence | Month of absence | Day of the week | Seasons | Transportation expense | Distance from  Residence to  Work | Service time |
| 11 | 26 | 7 | 3 | 1 | 289 | 36 | 13 |
| 36 | 0 | 7 | 3 | 1 | 118 | 13 | 18 |
| 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 |
| 7 | 7 | 7 | 5 | 1 | 279 | 5 | 14 |
| 11 | 23 | 7 | 5 | 1 | 289 | 36 | 13 |
| 3 | 23 | 7 | 6 | 1 | 179 | 51 | 18 |
| 10 | 22 | 7 | 6 | 1 |  | 52 | 3 |
| 20 | 23 | 7 | 6 | 1 | 260 | 50 | 11 |
| 14 | 19 | 7 | 2 | 1 | 155 | 12 | 14 |

Table 1.2: Employee Absenteeism Sample Data (Columns:9-18)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Workload Average/day | Hit target | Disciplinary failure | Education | Son | Social drinker | Social smoker | Pet | Weight |
| 33 | 2,39,554 | 97 | 0 | 1 | 2 | 1 | 0 | 1 | 90 |
| 50 | 2,39,554 | 97 | 1 | 1 | 1 | 1 | 0 | 0 | 98 |
| 38 | 2,39,554 | 97 | 0 | 1 | 0 | 1 | 0 | 0 | 89 |
| 39 | 2,39,554 | 97 | 0 | 1 | 2 | 1 | 1 | 0 | 68 |
| 33 | 2,39,554 | 97 | 0 | 1 | 2 | 1 | 0 | 1 | 90 |
| 38 | 2,39,554 | 97 | 0 | 1 | 0 | 1 | 0 | 0 | 89 |
| 28 | 2,39,554 | 97 | 0 | 1 | 1 | 1 | 0 | 4 | 80 |
| 36 | 2,39,554 | 97 | 0 | 1 | 4 | 1 | 0 | 0 | 65 |
| 34 | 2,39,554 | 97 | 0 | 1 | 2 | 1 | 0 | 0 | 95 |

Table 1.3: Employee Absenteeism Sample Data (Columns: 19-21)

|  |  |  |
| --- | --- | --- |
| Height | Body mass index | Absenteeism time in hours |
| 172 | 30 | 4 |
| 178 | 31 | 0 |
| 170 | 31 | 2 |
| 168 | 24 | 4 |
| 172 | 30 | 2 |
| 170 | 31 |  |
| 172 | 27 | 8 |
| 168 | 23 | 4 |
| 196 | 25 | 40 |

As per the above table following are the variable using which the regression model has to predict the number of hours the employee going to absent.

1. Individual identification (ID)
2. Reason for absence (ICD).

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

1. CERTAIN INFECTIOUS AND PARASITIC DISEASES
2. NEOPLASMS
3. DISEASES OF THE BLOOD AND BLOOD-FORMING ORGANS AND CERTAIN DISORDERS INVOLVING THE IMMUNE MECHANISM
4. ENDOCRINE NUTRITIONAL AND METABOLIC DISEASES
5. MENTAL AND BEHAVIOURAL DISORDERS
6. DISEASES OF THE NERVOUS SYSTEM
7. DISEASES OF THE EYE AND ADNEXA
8. DISEASES OF THE EAR AND MASTOID PROCESS
9. DISEASES OF THE CIRCULATORY SYSTEM
10. DISEASES OF THE RESPIRATORY SYSTEM
11. DISEASES OF THE DIGESTIVE SYSTEM
12. DISEASES OF THE SKIN AND SUBCUTANEOUS TISSUE
13. DISEASES OF THE MUSCULOSKELETAL SYSTEM AND CONNECTIVE TISSUE
14. DISEASES OF THE GENITOURINARY SYSTEM
15. PREGNANCY, CHILDBIRTH AND THE PUERPERIUM
16. CERTAIN CONDITIONS ORIGINATING IN THE PERINATAL PERIOD
17. CONGENITAL MALFORMATIONS, DEFORMATIONS AND CHROMOSOMAL ABNORMALITIES
18. SYMPTOMS, SIGNS AND ABNORMAL CLINICAL AND LABORATORY FINDINGS, NOT ELSEWHERE CLASSIFIED
19. INJURY, POISONING AND CERTAIN OTHER CONSEQUENCES OF EXTERNAL CAUSES
20. EXTERNAL CAUSES OF MORBIDITY AND MORTALITY
21. FACTORS INFLUENCING HEALTH STATUS AND CONTACT WITH HEALTH SERVICES.
22. PATIENT FOLLOW-UP
23. MEDICAL CONSULTATION
24. BLOOD DONATION
25. LABORATORY EXAMINATION
26. UNJUSTIFIED ABSENCE
27. PHYSIOTHERAPY

28 DENTAL CONSULTATION.

1. Month of absence
2. Day of the week (2. MONDAY 3. TUESDAY 4. WEDNESDAY 5. THURSDAY 6. FRIDAY)
3. Seasons (1. SUMMER 2. AUTUMN 3. WINTER 4. SPRING)
4. Transportation expense
5. Distance from Residence to Work (kilometres)
6. Service time
7. Age
8. Workload Average/day
9. Hit target
10. Disciplinary failure (yes=1; no=0)
11. Education (HIGH SCHOOL (1), GRADUATE (2), POSTGRADUATE (3), MASTER AND DOCTOR (4))
12. Son (number of children)
13. Social drinker (yes=1; no=0)
14. Social smoker (yes=1; no=0)
15. Pet (number of pet)
16. Weight
17. Height
18. Body mass index
19. Absenteeism time in hours

Chapter 2

Methodology

### 2.1 Pre-processing

The output of the Machine Learning data is fully depended upon the data feed in the algorithm, ML algorithm does not generate the proper outcome on raw data. So it’s important to transform the data so that the model accuracy of the model can be increased. In other words, we must apply some cleaning and processing for a better outcome with respect to the Machine learning algorithm. Some Machine learning algorithm required wellprocessed data like Decision Tree and random forest does not take null value. So it is important to process the data before feeding it in the respective machine learning algorithm.

Following are the pre-processing operation performed on the data to reduce the error rate and produce the optimal output.

#### 2.1.1 Missing Value Analysis

There are several options to handle missing value, but it mainly depends upon the nature of the data set, which missing analytics produce the optimum solution. Missing of data can occur due to nonresponse occur for example privacy concern. Moreover, if the value is missing completely at random, the data sample still represents the population but if the value is missing in some pattern, analysis produces bias output. There is two form of random missing values: MCAR and MAR. MCAR exists when the missing values are randomly distributed across all observations. This form can be confirmed by partitioning the data into two parts: one set containing the missing values, and the other containing the non-missing values. The second form is missing at random (MAR). In MAR, the missing values are not randomly distributed across observations but are distributed within one or more subsamples. Dropping the missing value is the good option if the data set has a low percentage of missing value and remaining data set can be used for further processing but it has its cons, dropping the missing value observation reduce the quantity of data. If our data set has a large number of missing value with respect to observation or feature, dropping is a good option. As fig 2.1 shows all the feature in the dataset is below 5% of missing value. Moreover, 13.65% of the data set has a missing value, so replacing the missing value with the statically analysis or clustering is a better solution. Multiple missing value approach can be implemented to replace the missing value with the predictive value generated by the approach. There are total 740 observations and 21 feature and by dropping all the missing value from the dataset to dimension reduce to 639 observations. This substantial reduction in the dimension can reduce to the accuracy of the machine learning algorithm.

Table 2.1: Missing Value Table

##### Variables Missing\_Val Missing\_per

|  |  |  |
| --- | --- | --- |
| Body mass index | 31 | 4.19% |
| Absenteeism time in hours | 22 | 2.97% |
| Height | 14 | 1.89% |
| Work load Average/day | 10 | 1.35% |
| Education | 10 | 1.35% |
| Transportation expense | 7 | 0.95% |
| Son | 6 | 0.81% |
| Disciplinary failure | 6 | 0.81% |
| Hit target | 6 | 0.81% |
| Social smoker | 4 | 0.54% |
| Age | 3 | 0.41% |
| Reason for absence | 3 | 0.41% |
| Service time | 3 | 0.41% |
| Distance from Residence to Work | 3 | 0.41% |
| Social drinker | 3 | 0.41% |
| Pet | 2 | 0.27% |
| Weight | 1 | 0.14% |
| Month of absence | 1 | 0.14% |
| Seasons | 0 | 0% |
| Day of the week | 0 | 0% |
| ID | 0 | 0% |

Table 2.2 represent all the approach parameter applied to check which missing value analysis produce the lowest variance with respect to the actual data and predicted data.

The first column represents the feature on which the missing value analysis applied. The second column represents the actual value in the cell. Sample array is randomly generated by the random generator and the value is replaced with NAN after fetching the actual value.

The third column shows the value when mean imputation is applied to the data. In mean imputation technique, the missing data is replaced with the mean of the respective feature. This technique produces a better result for a continuous variable. Mean imputation method does not produce the optimal output.

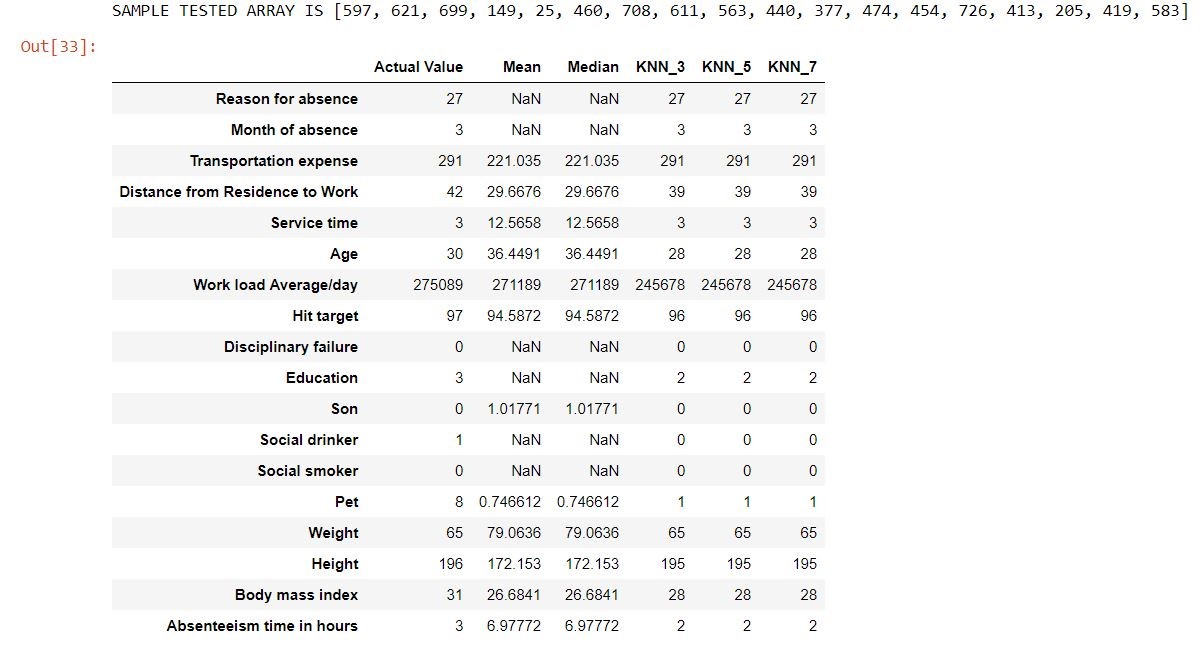
The fourth column represents the value with respect to the feature when the median imputation is applied to the data. Median imputation is similar to the mean imputation but rather than replacing the NAN value with the mean of the feature, this imputation method replaces it with the median of the feature. Similar to the mean imputation, median imputation should be applied on continues variable rather than applying on a categorical variable.

In the respective table, the mean and median column has NAN value because some feature in the data is of categorical nature.

The last third column represents the KNN imputation method. KNN imputation method checks all the nearest variable with respect to the missing value. KNN algorithm used for the matching point with the closest k neighbour in the multi-dimension. It can be used for data that are categorical, discrete and continuous which makes it more efficient with all kind of missing data. In the last three columns (KNN\_n) where n represents the kth unit.

K is the number of neighbour voting on the for the missing value and that the main reason the kth value is an odd number because even kth value will generate conflict in the voting process. KNN\_3 represent the value which was generated when applying KNN where k value = 3. Same for the KNN\_5 and KNN\_7 where k value was 5 and 7 respectively.

Table 2.2: Multiple Missing Value Analysis Test Table



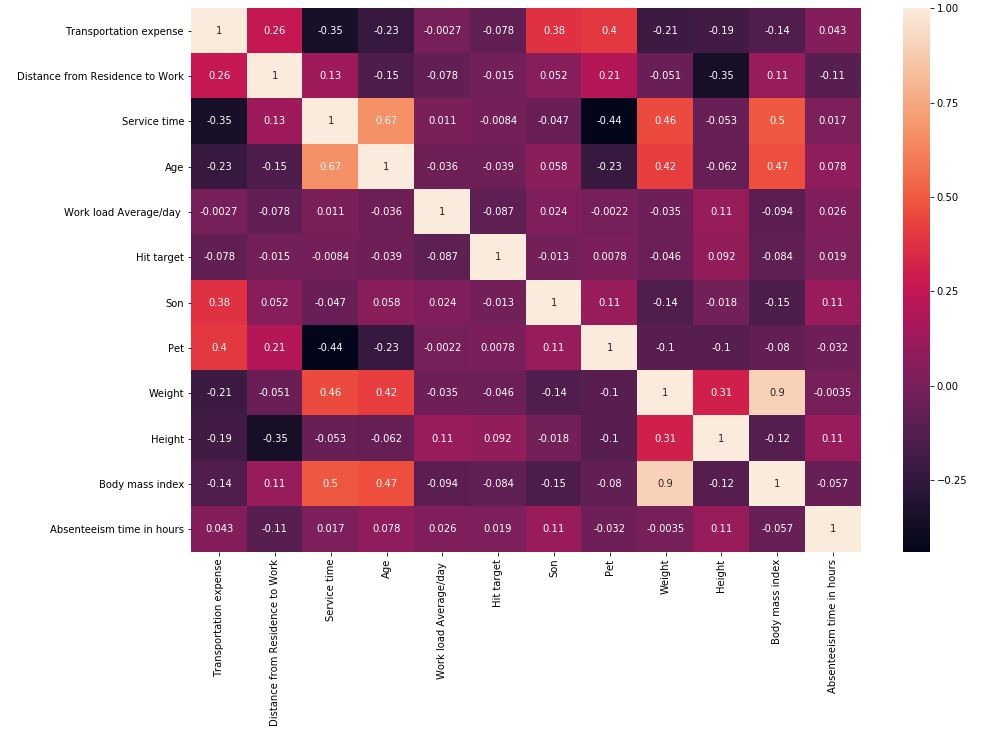
As per data represent in table 2.2, KNN imputation provides the optimum solution for absenteeism data with minimum variation with respect to actual data.

#### 2.1.2 Feature Selection

Feature selection is another pre-processing technique which decreases the load over machine learning algorithm checking the correlation between other feature and check which feature is highly correlated to another feature. If two feature carries the same data, this will create a bias output. Feature selection is a process where you can select those features which contribute most to the prediction output. Moreover, having a relevant feature in the data set can decrease the accuracy of the model.

Fig 2.3 represents the correlation heatmap with respect to the data. Darker colour represents the positive correlation and lighter colour scheme represent the negative correlation with respect to other features. Body mass index is highly positively related towards weight feature in the dataset which is 0.9. Moreover, no other feature is correlated with respect to other feature. Moreover, body mass index is the ratio of weight with respect to the square of height, so by dropping the two feature from the data set will reduce the time complexity with the same efficiency.

Fig 2.3: Correlation Heat map



#### 2.1.3 Feature Scaling

Feature scaling is a method to standardize the variable or the feature of the data set, It also knows as standardization of a data set. Some feature has magnitudes, unit or range and normalization should be performed when the scaling of the feature is irrelevant or have a high magnitude which will lead to the bases in the machine learning algorithm which will lean towards the higher magnitude dataset. Z-scoring/standardization is one technique in which the value of a feature is subtracted with the mean of the feature and their subtraction is divided by the standard deviation of the feature which will also set the feature in normalized bell shape form whereas another method involves the value of a feature is subtracted by the minimum value of the respective feature which is divided by the subtraction of the minimum value and maximum value of the respective feature. In our dataset, some feature has irrelevant scaling which will affect the machine learning algorithm. A feature like 'Transportation expense', 'Distance from Residence to Work','Service time','Workload Average/day ', 'Hit target','Height' and 'Weight' is scaled by applying the min-max scaling technique.

Min-Max scaling:

Yi = [Xi - min(X)]/[max(X) - min(X)]

### 2.2 Modelling

#### 2.1.1 Model Selection

In our earlier stage of analysis, we came to know that our dependent variable in our case is Absenteeism time in hours is numerical, so we are using a Regression analysis over our data set.

As we observed in missing value analysis KNN imputation generated the best result with respect our data set. So we are going to apply K-Nearest Neighbours regression first and will check all the error metrics and after Decision tree and random forest machine learning algorithm will be applied and evaluate which machine learning algorithm performs best and produce best predict for our data set.

#### 2.1.2 K-Nearest Neighbours

K-nearest Neighbour predicts the value by checking the distance with from other feature with respect to the Kth value.

Applying the K-nearest Neighbours on our data set with k =3. Checking which kth value fit best for the respective data set.

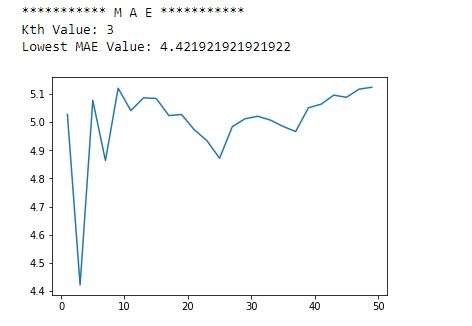
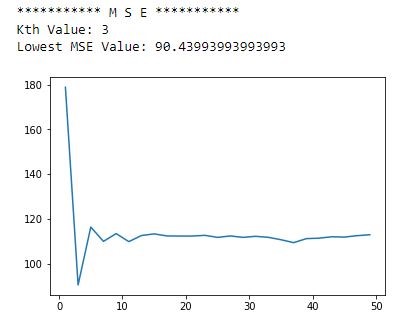
Python:

from sklearn.neighbors import KNeighborsRegressor

Pro\_KNN\_model = KNeighborsRegressor(n\_neighbors=3).fit(X\_TRAIN,Y\_TRAIN)

Final\_KNN\_predict = Pro\_KNN\_model.predict(X\_TEST) reg\_acc(Y\_TEST,Final\_KNN\_predict)

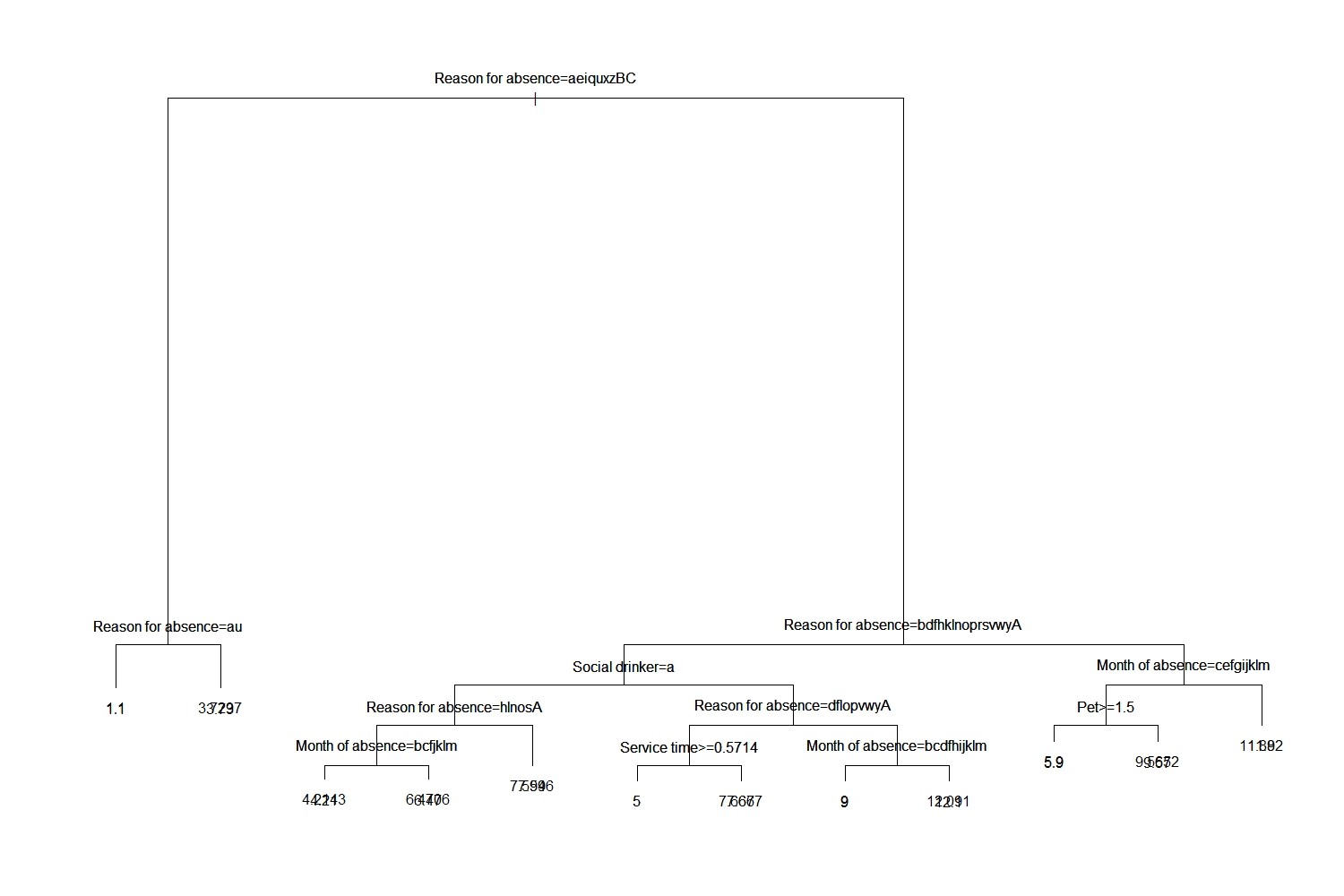
Following graph shows which kth value fits the data set and provide the optimal outcome.



As seen in the above fig the k=3 is the optimum value for the K-nearest Neighbour machine learning algorithm to predict the value with the lowest error rate.

#### 2.2.3. Decision Trees

Decision Tree algorithm creates a flow chat like tree structure where end nodes denote the outcome and the sub-nodes denote the decision flow of the tree. Following are the code and the decision tree with respect to the data set.



Python:

|  |
| --- |
| from sklearn import tree Deci\_tree\_pre =  tree.DecisionTreeRegressor().fit(X\_TRAIN,Y\_TRAIN).predict(X\_TEST) print('Decision Tree ML') reg\_acc(Deci\_tree\_pre,Y\_TEST) |

#### 2.2.4. Random Forest

Similar to the decision tree, we also applied a random forest algorithm to check the prediction rate of a data set. Random forest works the same as decision tree but rather than creating a single tree, random forest creates multiple trees corresponding to the data set.

Python:

from sklearn.ensemble import RandomForestRegressor

Ran\_for\_pre =

RandomForestRegressor(n\_estimators=50).fit(X\_TRAIN,Y\_TRAIN).predict(X\_TEST) print('Random Forest ML') reg\_acc(Ran\_for\_pre,Y\_TEST)

#### 2.2.5 Linear Regression

As the dependent variable is a numerical data, we are using linear regression model to ‘Absenteeism time in hours’. Following are the coefficient metrics with respect to all the feature of the data set to predict the value of Y (Absenteeism time in hours’).

## 3. Conclusion

### 3.1. Model Evaluation

As this is a regression model prediction, we are using MSE/RMSE and MAE to predict the accuracy and how well the value is predicted with respect to the machine learning algorithm.

Python Model Evaluation:

As the fig 3.1 show, python machine learning algorithm predicted the accuracy parameter.

MSE/RMSE parameter

MSE/RMSE of KNN : 130.57/11.42

MSE/RMSE of Decision Tree : 337.55/18.37 MSE/RMSE of Random Forest : 185.97/13.63 MSE/RMSE of Linear Regression : 136.75/11.69

MAE parameter

|  |  |
| --- | --- |
| MAE of KNN | : 4.94 |
| MAE of Decision Tree | : 6.73 |
| MAE of Random Forest | : 5.27 |
| MAE of Linear Regression | : 5.66 |

With respect to the above parameters, KNN and Random forest provided a better accurate result in comparison with the decision tree.

Fig 3.1: Error Metrics in python



R Model Evaluation:

As the fig 3.2 show, R machine learning algorithm predicted the accuracy parameter.

MSE/RMSE parameter

MSE/RMSE of KNN : 6.53/2.55

MSE/RMSE of Decision Tree : 8.54/2.92 MSE/RMSE of Random Forest : 4.21/2.05 MSE/RMSE of Linear Regression : 6.09/2.46

MAE parameter

|  |  |
| --- | --- |
| MAE of KNN | : 1.76 |
| MAE of Decision Tree | : 1.93 |
| MAE of Random Forest | : 1.49 |
| MAE of Linear Regression | : 1.67 |

With respect to the above parameters, KNN and Random forest provided better accurate result in comparison with the decision tree.

Fig 3.2: Error metric in R



### 3.2 Model Selection

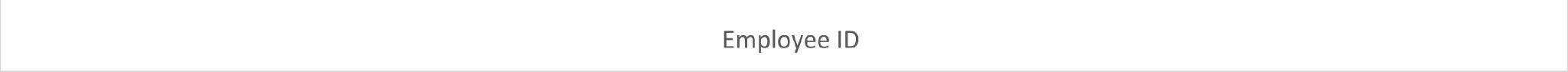
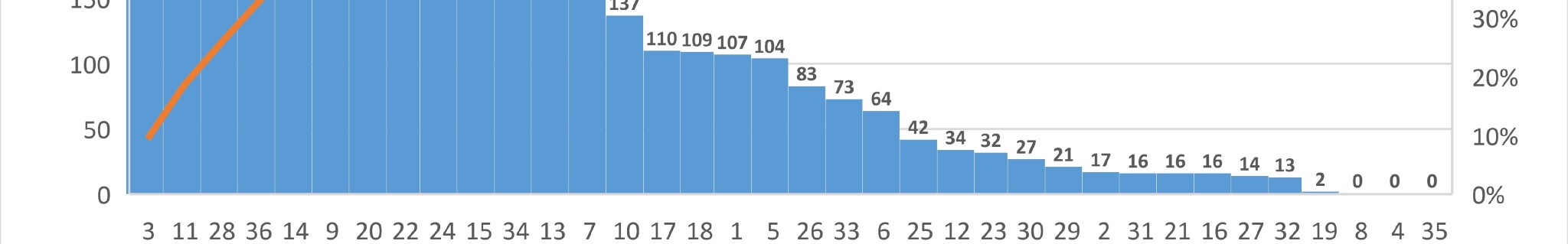
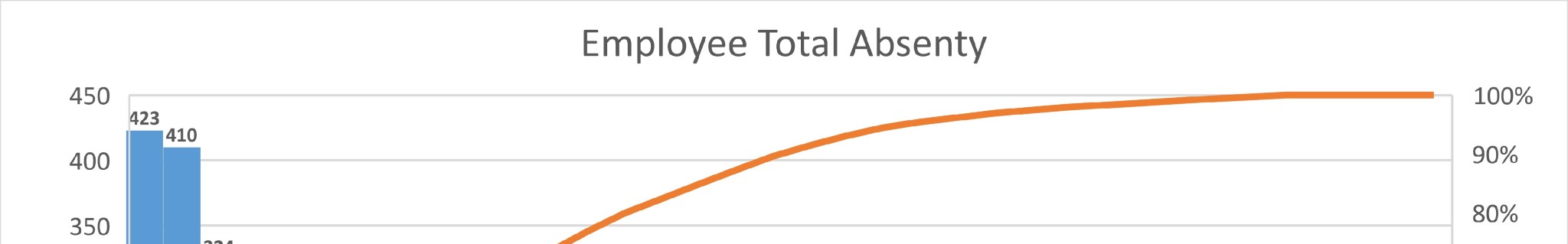
KNN machine learning algorithm produces the best outcome with respect to the data set and the prediction of the 'Absenteeism at work' was predicted quite accurately. Moreover, the other algorithm also has a reasonable accuracy in predicting the 'Absenteeism at work'.

### 3.3 Summary

With respect to the absenteeism data set, we have the following conclusion regarding the company absenteeism and the employees with respect to the employees.

The fig 3.3.1, show the employee ID 3 and 11 has the highest of 423 and 410 respectively. Moreover, these employees have a serious health condition which is leading towards the highest absenteeism.

Fig 3.3.1: Absenteeism with respect to Employee’s



The fig 3.3.2 represents, the employees with the age group of 28,33 and 38 has the highest absenteeism in the company. Employee ID 11 has an age of 33, who has serious health issue.

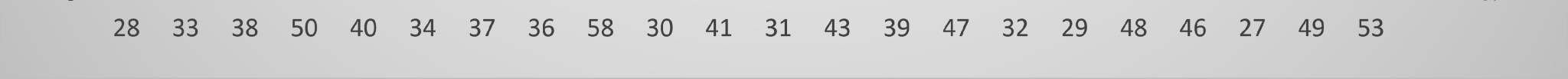
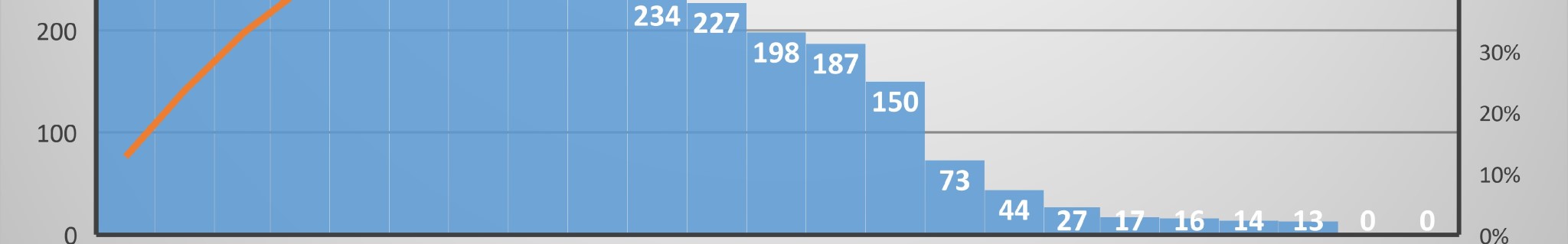
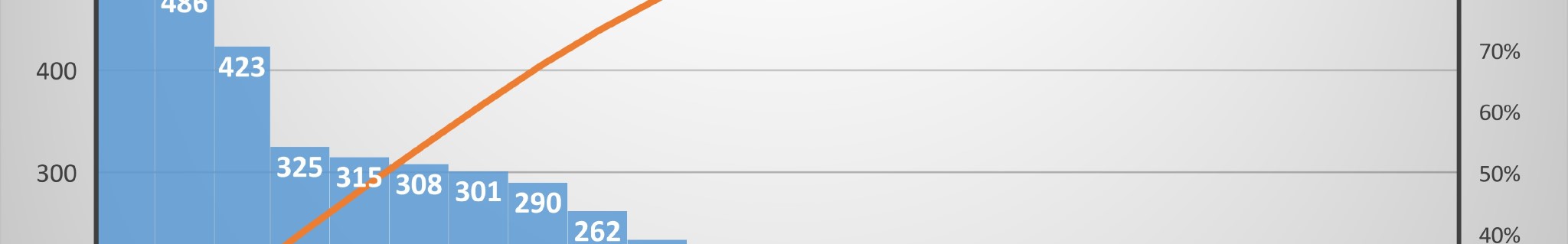
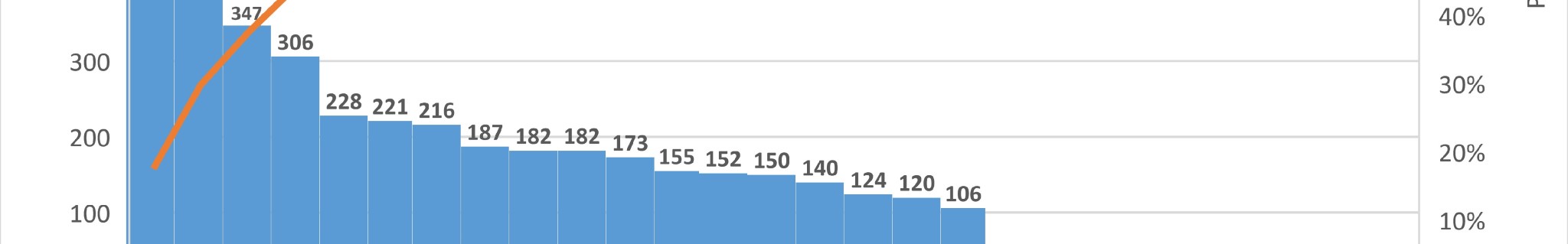
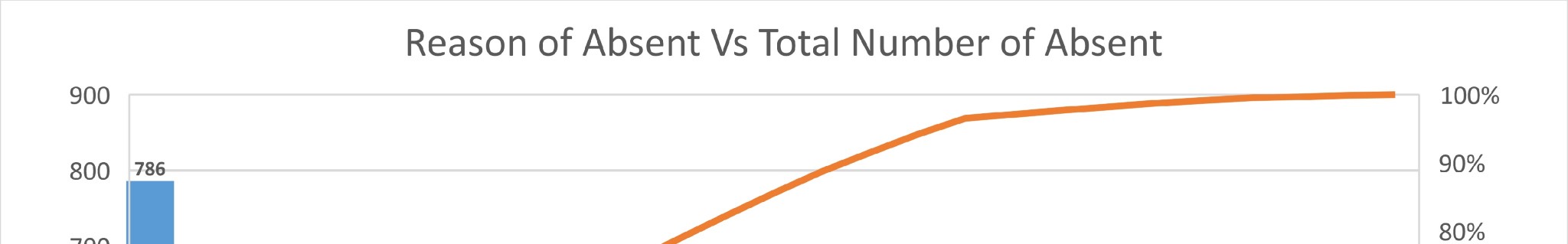


Fig 3.3.2: Age with respect to Absenteeism

Around half of the employee of the organisation is suffering from “diseases of the musculoskeletal system and connective tissue” (Reason of Absent: 13), which leads towards the highest absenteeism rate. Moreover, should also look after the employees how has “injury, poisoning and certain other consequences of external causes” (Reason of Absent 19) which is affecting 16 employees of the organisation.

Fig 3.3.3: Reason of Absent with respect to Absenteeism



Employees tend to take more absent leaves in the third season, there might be some employees has a season health issue which is affecting the employees. Moreover, as per the data set employees are taking more absently leave in the month of March and July.

Fig 3.3.4: Seasonal and monthly representation with respect to absenteeism

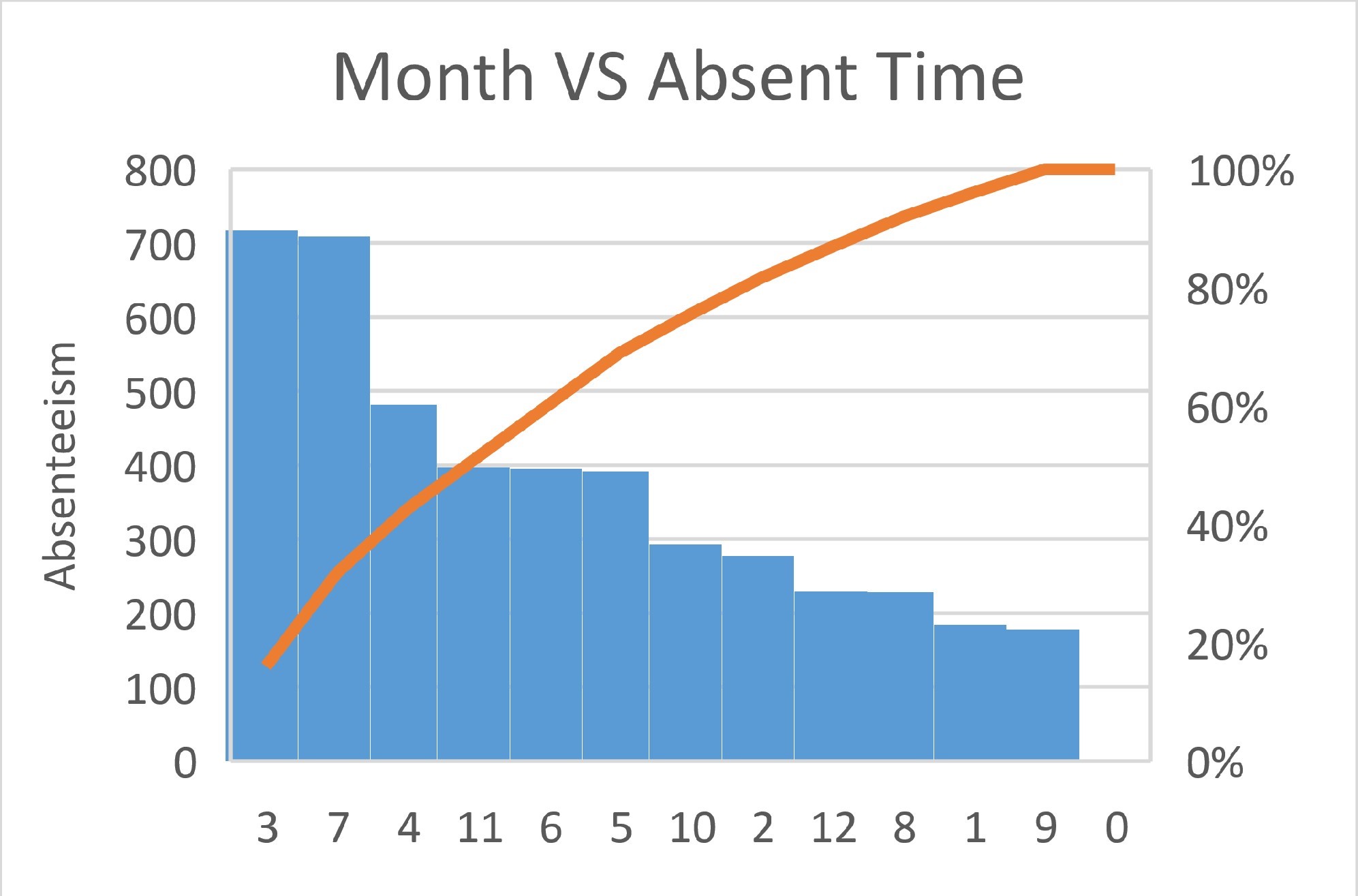
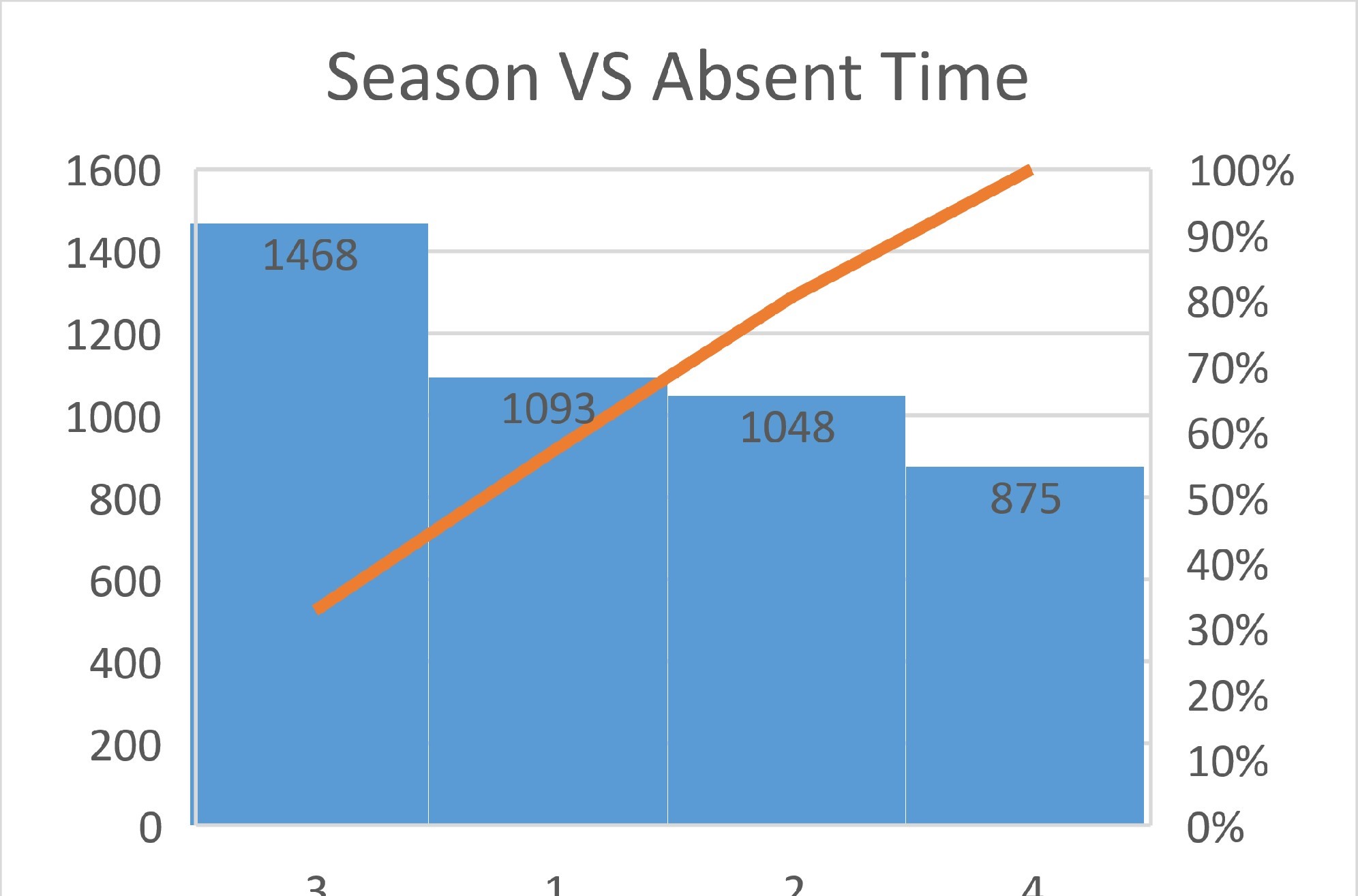
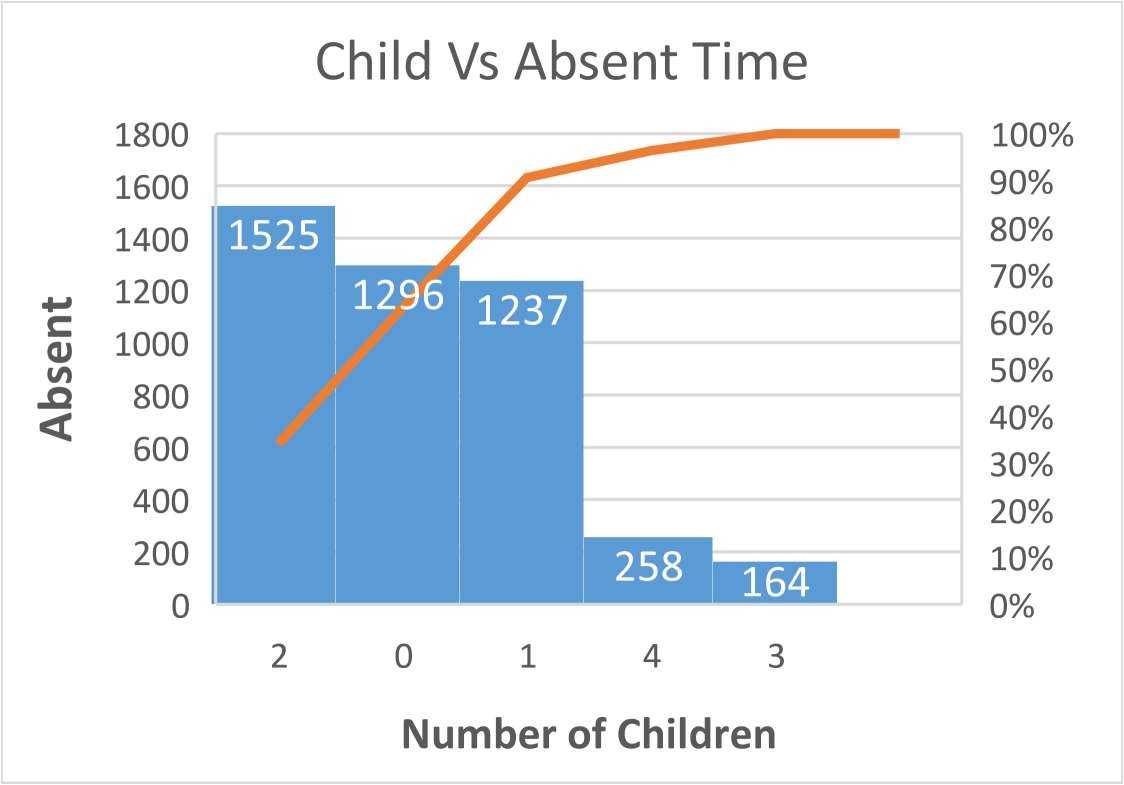


Fig 3.3.5: Absenteeism with respect to Children

 The company should emphasis on the employees who has 1 or 2 children, there might be a chance that the increase in absent is due to the employee children.

1. What changes the company should bring to reduce the number of absenteeism?

Some employees on the company have a series health condition which should be taken care off. Company management should focus on the employee with no child, as there might be chance of high health risk. Moreover, employee ID 3 and 11 should be taken into consideration as these employees have taken the maximum absentee and also having some serious health issue which leads to the high absenteeism.

“Diseases of the musculoskeletal system and connective tissue” and “Injury, poisoning and certain other consequences of external causes” are not a common disease, there might be an environmental issue. There might a chance that due to some environmental factor, maximum employees are affecting due to this disease. These are the serious issue which should be taken into consideration as this can also affect other employees of an organisation. Around 60% of the workforce is affected by this disease.

1. How much losses every month can we project in 2011 if the same trend of absenteeism continues? Fig 3.3.6 is projecting the forecasting with respect to the season, if this trend of absenteeism continued, there is a chance that the company will suffer the average of 6 (216 hours per season) per season. Fig 3.3.7 also shows the monthly forecasting with 90% confidential interval.

