**Dataset Name:** Absenteeism at work **.**

**Project Definition:** In Absenteeism at work dataset our aim is to perform hierarchical clustering to finding new bases (knowledge) .This data set is recorded of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. This dataset contain unlabelled data so this is an unsupervised learning. Unsupervised machine learning technique is to find similarities in the data point and clustered similar data points together .For this project target attribute is absenteeism time in hours. There are 21 attribute in this dataset .

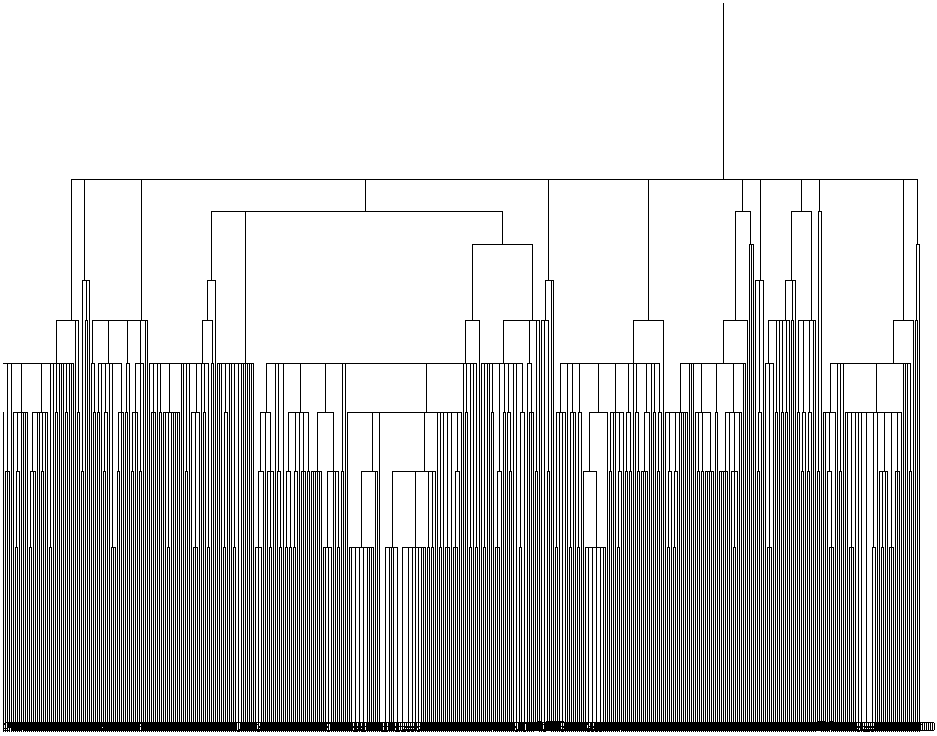
**Literature Survey:** Absenteeism at work dataset is prepared for machine learning purpose. The high competitiveness in the market, professional development combined with the development of organizations and the pressure to reach the goal, create increasingly overburdened employees and end up acquiring some disturbance in the state of health related to the type of work activity. Absenteeism is defined as absence to work as expected, represents for the company the loss of productivity and quality of work. The purpose of the paper [1] was to apply an artificial neural network to prediction of absenteeism at work. The objective of this paper is to apply a neuro fuzzy network in the prediction of absenteeism at work. A neuro fuzzy network was developed using an artificial neural network architecture multilayer perceptron with the error back propagation algorithm .The 21 attributes were used in the experiments to prediction of absenteeism. ANN are models consisting of simple processing units, called artificial neurons, these models are inspired by the structure of the brain and aim to simulate human behaviour, such as learning, association, generalization and abstraction when submitted to training. The experiments with the ANN presented the expected results in prediction of absenteeism at work.

**Methods:**

We use hierarchical clustering in absenteeism at work data set .We use 2 method :

* Agglomerative bottom up (consider Euclidean distance)
* Single link clustering

**Hierarchical cluster tree with cutting point:**



Clustered Instances

0 460 ( 62%)

1 6 ( 1%)

2 109 ( 15%)

3 23 ( 3%)

4 32 ( 4%)

5 7 ( 1%)

6 6 ( 1%)

7 37 ( 5%)

8 15 ( 2%)

9 20 ( 3%)

10 8 ( 1%)

11 2 ( 0%)

12 7 ( 1%)

13 4 ( 1%)

14 2 ( 0%)

15 1 ( 0%)

16 1 ( 0%)

**Knowledge discovery(Finding new bases) :** I got 17 cluster from the marked cutting point of the dendrogram (hierarchical clustering tree) . Then I decorated the data set instances by their belonging cluster .Then cluster wise knowledge easily can discovered .The data set have total 740 instances among them cluster-0 contains 460 instances .I was not considered Cluster (1,5,6,10,11,12,13,14,15,16) for discover knowledge .Because these clusters are contain only 1% of instances from the total dataset .So, if I discovered any knowledge from these cluster did not give enough support of the total data set.

Cluster-8 contain the highest average absenteeism time 12 hours. The most frequent reason for absenteeism is 22 (patient follow up) for cluster-8 and this absenteeism is happened most frequently in October(10) month .Average transport expense of cluster-8 is 369 ,which could be the reason for the highest absenteeism time in hours. This absenteeism is most frequently occurred in spring(4) season .

Cluster-4 contain average absenteeism time 10 hours .The most frequent reason for absenteeism is 23 (medical consultation) for cluster-4 and this absenteeism is happened most frequently in March (3) month . This absenteeism is most frequently occurred in autumn(2) season .

Cluster-0 contain average absenteeism time 7 hours .The most frequent reason for absenteeism is 28 (dental consultation) for cluster-4 and this absenteeism is happened most frequently in March (3) month .This absenteeism is most frequently occurred in autumn(2) season.

Cluster-7 contain average absenteeism time 7 hours .The most frequent reason for absenteeism is 23 (medical consultation) for cluster-7 and this absenteeism is happened most frequently in September (9) month . This absenteeism is most frequently occurred in spring(4) season .

Cluster-2 contain average absenteeism time 6 hours .The most frequent reason for absenteeism is 23 (medical consultation) for cluster-2 and this absenteeism is happened most frequently in June & December (6 & 12) month . This absenteeism is most frequently occurred in spring(4) season .

Cluster-9 contain average absenteeism time 6 hours .The most frequent reason for absenteeism is 25 (laboratory examination) for cluster-9and this absenteeism is happened most frequently in May (5) month . This absenteeism is most frequently occurred in autumn(2) season.

Cluster-3 contain average absenteeism time 5 hours .The most frequent reason for absenteeism is 23 (medical consultation) for cluster-3 and this absenteeism is happened most frequently in August (8) month . This absenteeism is most frequently occurred in summer(1) season .Average service time of cluster-3 is only 8, which could be reason for the lowest absenteeism time in hours.

**Reference:**

1. Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE