**Project Definition**

In this project we used a dataset of cereals along with their different percentage of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. Besides, we did a hierarchical clustering for finding the reasons of absents.

**Literature Survey**

Dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/00445/

The meaning of each column:

1. Individual identification (ID)

2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (1to 21) 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 behavioral 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.

And 7 categories without (CID)

Patient follow-up (22),

Medical consultation (23),

Blood donation (24),

Laboratory examination (25),

Unjustified absence (26),

Physiotherapy (27),

Dental consultation (28).

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

17. Pet (number of pet)

18. Weight

19. Height

20. Body mass index

21. Absenteeism time in hours (target)

**Methods**

The dataset has been clustered by the hierarchical clustering technique. The Data set has been divided 12 clusters. Then similarities between instances of individual clusters and dissimilarities between instances of different clusters have been analyzed.

**Hierarchical cluster tree with cutting point:**

=== Run information ===

**Scheme:** weka.clusterers.HierarchicalClusterer -N 12 -L COMPLETE -P -A "weka.core.EuclideanDistance -R first-last"

**Relation**: Absenteeism\_at\_work\_A-weka.filters.unsupervised.attribute.AddCluster-Wweka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -**t**2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

**Instances:** 740

**Attributes**: 22

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

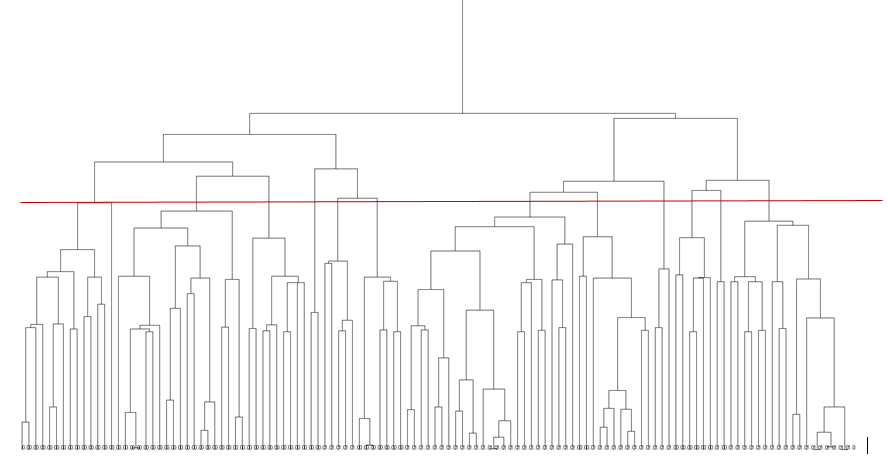
Weight

Height

Body\_mass\_index

Absenteeism\_time\_in\_hours

cluster



**Cluster Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster No.** | **No. of Instances** | **Analysis** | **Findings** |

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 76 | * Social Smoker: all (except 2) * Hit Target:96% * Education: High School (all) | * Most High school people are smokers with hit hits. |
| 2 | 53 | * Distance :Average( Mostly 50) * Education: high School(most) * Hit Target: High * Son: Most has 4 sons * Social Drinker: 99% * BMI: 90% underweight | * Has 4 son and drinker are underweight |
| 3 | 66 | * Season: Autumn * Hit Rate:96% * Disiplinary Failure: All 0 * Social Smoker :98% | * Has high hits and no disciplinary failure but most are social smoker |
| 4 | 58 | * Education:97% high school * Social Drinker:98% * Social Smoker: 2% * BMI:Class1 Obesity | * Most are drinker but does’t smoke and are not healthy people. |
| 5 | 47 | * Transportation Expense: high * Social Drinker:100% | * All are Drinker |
| 6 | 50 | * Age: 98% Middle Aged * Hit Target :High | * Middle aged having high hits |
| 7 | 98 | * Season: Spring 80% * Social Drinker:1 | * Hos only one drinker |
| 8 | 36 | * Hit Target: High * Son:2(75%) * Disiplinary Failure:15% * BMI: Normal Weight | * Most has 2 sons will very less disciplinary failure. |
| 9 | 60 | * Distance : High * Service time : Medium(Avg18) * Age: Middle Age(avg28.2) * Education:High School (100%) * Social Smoker:0% * BMI: Class 1 Obesity | * All are high school grads and sober despite none are smokers. |
| 10 | 72 | * Age: Middle Age(avg 37.8) * Disiplinary failure:0%. * Social Drinker:0% * Social Drinker:0% | * Non-smoker and no drinker with no disciplinary failure. |
| 11 | 44 | * Season: spring * Distance: High * Transportation: High * Social Smoker:0% * BMI: Class 1 obesity | * Distance is high so the transportation expense is high and no smoker but has bad health |
| 12 | 89 | * Service Time: Low * Hit target: High * Edutacion:80% post grad * BMI: normal Weight | * Service time is low despite high hits, most are post-graduates. |

Using the analysis and the unique findings from the tables we can get many co-relation between the attributes which may be directly or indirectly related to the findings.