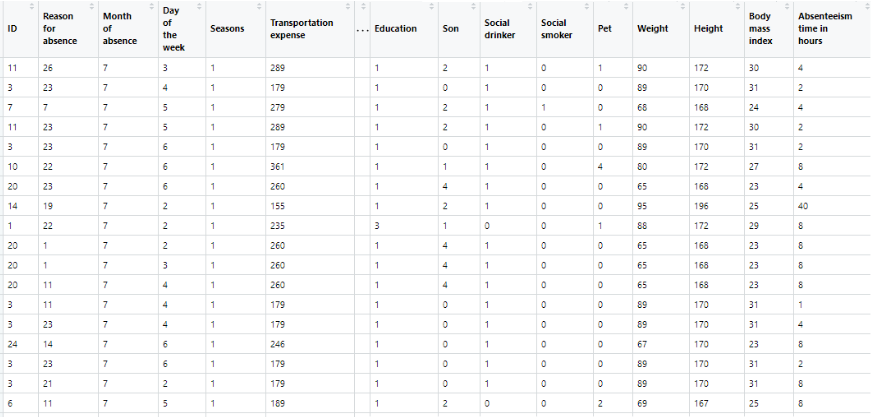
Multivariate Analysis

Catarina Costa nº86582
Inês Tavares nº86593
Catarina Oliveira nº86963
Gonçalo Mestre nº 87005

Instituto Superior Técnico, Lisboa

January 2020

Initial Dataset



ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	...	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
11	26	7	3	1	289		1	2	1	0	1	90	172	30	4
3	23	7	4	1	179		1	0	1	0	0	89	170	31	2
7	7	7	5	1	279		1	2	1	1	0	68	168	24	4
11	23	7	5	1	289		1	2	1	0	1	90	172	30	2
3	23	7	6	1	179		1	0	1	0	0	89	170	31	2
10	22	7	6	1	361		1	1	1	0	4	80	172	27	8
20	23	7	6	1	260		1	4	1	0	0	65	168	23	4
14	19	7	2	1	155		1	2	1	0	0	95	196	25	40
1	22	7	2	1	235		3	1	0	0	1	88	172	29	8
20	1	7	2	1	260		1	4	1	0	0	65	168	23	8
20	1	7	3	1	260		1	4	1	0	0	65	168	23	8
20	11	7	4	1	260		1	4	1	0	0	65	168	23	8
3	11	7	4	1	179		1	0	1	0	0	89	170	31	1
3	23	7	4	1	179		1	0	1	0	0	89	170	31	4
24	14	7	6	1	246		1	0	1	0	0	67	170	23	8
3	23	7	6	1	179		1	0	1	0	0	89	170	31	2
3	21	7	2	1	179		1	0	1	0	0	89	170	31	8
6	11	7	5	1	189		1	2	0	0	2	69	167	25	8

- 21 variables
- 740 rows

Data Preliminary Analysis

Division of the variable *Reason for Absence* in classes

- Disease **Reason for Absence = 1**
- Patient follow-up - **Reason for Absence = 2**
- Medical Consultation - **Reason for Absence = 3**
- Blood Donation - **Reason for Absence = 4**
- Laboratory Examination - **Reason for Absence = 5**
- Unjustified Absence - **Reason for Absence = 6**
- Physiotherapy - **Reason for Absence = 7**
- Dental Consultation - **Reason for Absence = 8**

Data Preliminary Analysis

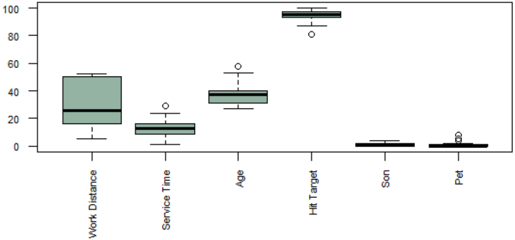
Missing values

The rows in which the variable *Absenteeism* took the value 0 were removed.

Division of the target variable *Absenteeism* in classes

- Number of hours < 8 - **Absenteeism = 1**
- Number of hours $= 8$ - **Absenteeism = 2**
- $40 \leq$ Number of hours < 8 - **Absenteeism = 3**
- Number of hours > 40 - **Absenteeism = 4**

Data Preliminary Analysis - Outliers



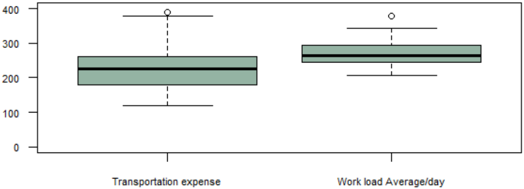
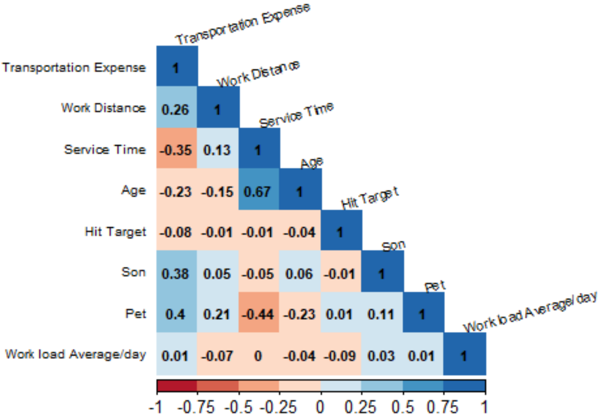


Figure: Boxplots of *Work Distance*, *Service Time*, *Age*, *Hit Target*, *Son*, *Pet*, *Work Load Average/day* and *Transportation Expense*

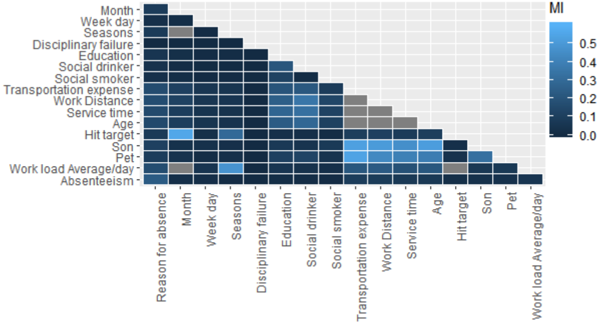
Data Preliminary Analysis - Pearson's Correlation



	Transportation Expense	Work Distance	Service Time	Age	Hit Target	Son	Pet	Work load Average/day
Transportation Expense	1							
Work Distance	0.26	1						
Service Time	-0.35	0.13	1					
Age	-0.23	-0.15	0.67	1				
Hit Target	-0.08	-0.01	-0.01	-0.04	1			
Son	0.38	0.05	-0.05	0.06	-0.01	1		
Pet	0.4	0.21	-0.44	-0.23	0.01	0.11	1	
Work load Average/day	0.01	-0.07	0	-0.04	-0.09	0.03	0.01	1

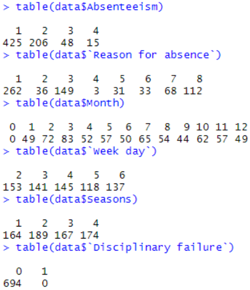
Figure: Correlation between the quantitative variables

Data Preliminary Analysis - Mutual Information



- *Transportation Expense* and *Work Distance* - 0.854;
- *Transportation Expense* and *Age* - 0.834;
- *Service Time* and *Work Distance* - 0.868;
- *Age* and *Work Distance* - 0.869;
- *Age* and *Service Time* - 0.829;
- *Transportation Expense* and *Service Time* - 0.789.

Data Preliminary Analysis - Categorical/Binary Variables



```
> table(data$Absenteeism)

  1   2   3   4 
425 206  48  15 
> table(data$`Reason for absence`)

  1   2   3   4   5   6   7   8 
262  36 149   3  31  33  68 112 
> table(data$Month)

 0  1  2  3  4  5  6  7  8  9 10 11 12 
 0 49 72 83 52 57 50 65 54 44 62 57 49 
> table(data$`Week day`)

  2   3   4   5   6 
153 141 145 118 137 
> table(data$Seasons)

  1   2   3   4 
164 189 167 174 
> table(data$`Disciplinary failure`)

  0   1 
694   0 
```

Figure: *Table* of *Absenteeism*, *Reason for Absence*, *Seasons*, *Week Day*, *Month* and *Disciplinary Failure*

- Most commonly used reasons for absence: Disease (38%), Medical Consultation (21%) and Dental Consultation (16%).
- The absenteeism time is mostly lower than 8 hours (61%), or equal to 8 hours (30%);
- 100% of the work absences are from people without disciplinary failures.

Data Preliminary Analysis - Categorical/Binary Variables

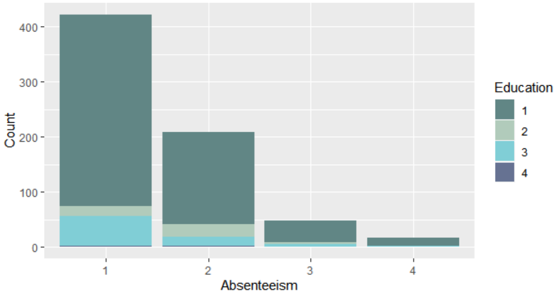


Figure: Proportion of each level of Education by Absenteeism's class

- 82% of the work absences are from people with only high school level of education;

Data Preliminary Analysis - Categorical/Binary Variables

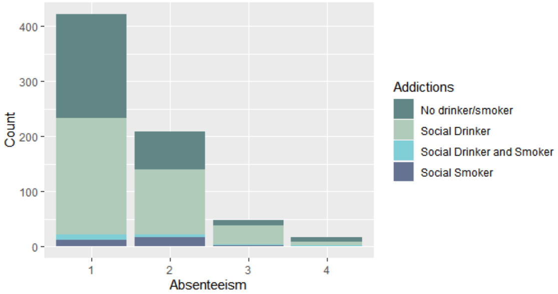
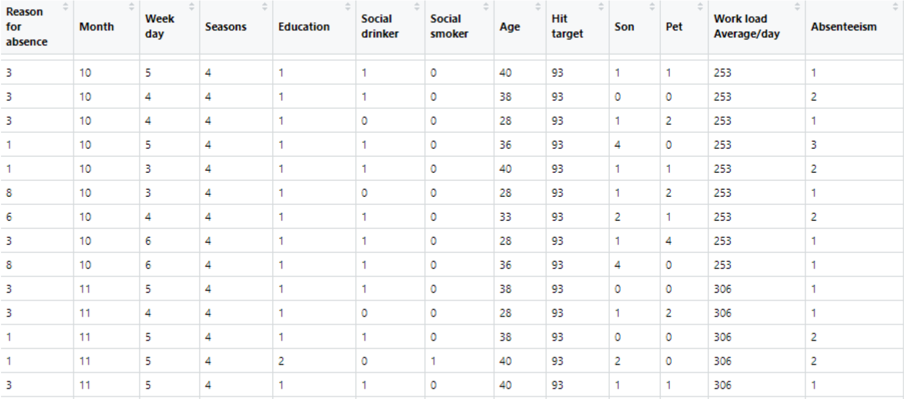


Figure: Proportion of social smokers/drinkers and non drinkers and smokers by Absenteeism's class

- 93% of the work absences are from people that do not smoke;
- 39% of absences are from non smokers and non drinkers;

Data Preliminary Analysis - Final dataset

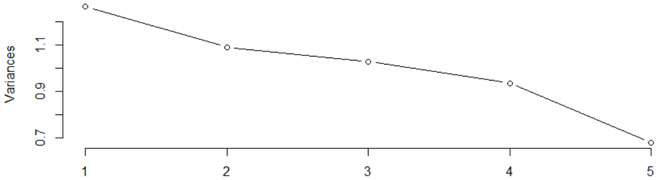


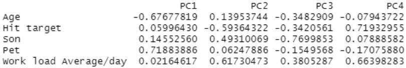
Reason for absence	Month	Week day	Seasons	Education	Social drinker	Social smoker	Age	Hit target	Son	Pet	Work load Average/day	Absenteeism
3	10	5	4	1	1	0	40	93	1	1	253	1
3	10	4	4	1	1	0	38	93	0	0	253	2
3	10	4	4	1	0	0	28	93	1	2	253	1
1	10	5	4	1	1	0	36	93	4	0	253	3
1	10	3	4	1	1	0	40	93	1	1	253	2
8	10	3	4	1	0	0	28	93	1	2	253	1
6	10	4	4	1	1	0	33	93	2	1	253	2
3	10	6	4	1	1	0	28	93	1	4	253	1
8	10	6	4	1	1	0	36	93	4	0	253	1
3	11	5	4	1	1	0	38	93	0	0	306	1
3	11	4	4	1	0	0	28	93	1	2	306	1
1	11	5	4	1	1	0	38	93	0	0	306	2
1	11	5	4	2	0	1	40	93	2	0	306	2
3	11	5	4	1	1	0	40	93	1	1	306	1

- 13 variables
- 694 rows

Dimensionality Reduction

PCA standardized - Quantitative variables (5)



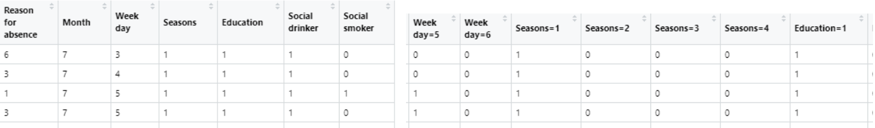


```
                                PC1         PC2        PC3         PC4
Age                     -0.67677819  0.13953744 -0.3482909 -0.07943722
Hit target               0.05996430 -0.59364322 -0.3420561  0.71932955
Son                      0.14552560  0.49310069 -0.7699853  0.07888582
Pet                      0.71883886  0.06247886 -0.1549568 -0.17075880
Work load Average/day    0.02164617  0.61730473  0.3805287  0.66398283
```

- The first 4 principal components explain **86.37%** of the variability

Dimensionality Reduction

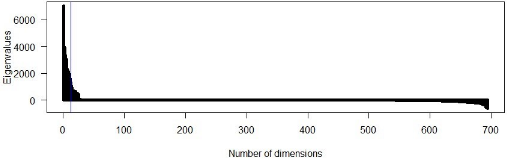
Multidimensional Scaling - Categorical variables (7)



Reason for absence	Month	Week day	Seasons	Education	Social drinker	Social smoker
6	7	3	1	1	1	0
3	7	4	1	1	1	0
1	7	5	1	1	1	1
3	7	5	1	1	1	0

Week day=5	Week day=6	Seasons=1	Seasons=2	Seasons=3	Seasons=4	Education=1
0	0	1	0	0	0	1
0	0	1	0	0	0	1
1	0	1	0	0	0	1
1	0	1	0	0	0	1

Figure: Turning all variables into binary



- *Hamming distance* is good for binary variables
- The *Goodness of fit* is **0.805** using 12 dimensions

Division of dataset

Training and testing data

The dataset after PCA and MDS was divided into training (**70%**) and testing (**30%**) data.

Cross Validation

- *k-fold* cross validation using training data, with $k = 10$
- Testing the performance of each model using testing data

Supervised Learning Methods for Classification

Linear Algorithms

- Linear Discriminant Analysis

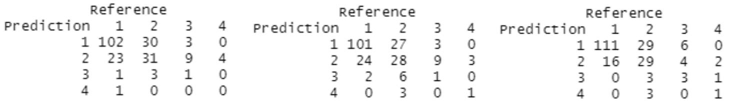
Nonlinear Algorithms

- Regularized Discriminant Analysis
- Support Vector Machine
- Naive Bayes Classifier
- Neural Networks
- K-Nearest Neighbour

Ensemble Algorithms

- Random Forest

Supervised Learning Methods for Classification



```
          Reference
Prediction   1   2   3   4
         1 102  30   3   0
         2  23  31   9   4
         3   1   3   1   0
         4   1   0   0   0
```

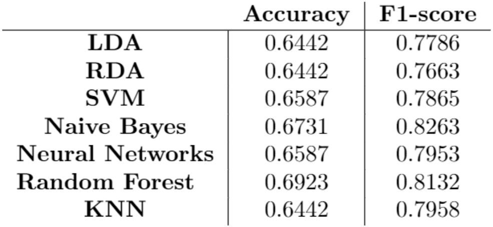
```
          Reference
Prediction   1   2   3   4
         1 101  27   3   0
         2  24  28   9   3
         3   2   6   1   0
         4   0   3   0   1
```

```
          Reference
Prediction   1   2   3   4
         1 111  29   6   0
         2  16  29   4   2
         3   0   3   3   1
         4   0   3   0   1
```

Figure: Confusion matrices using LDA, Neural Networks and Random Forest, respectively

- LDA (F1-Score = 0.779) does not classify any observation into class 4 - class with the lowest probability
- Neural Networks (F1-Score = 0.800) is a good model for this data
- Random Forest (F1-Score = 0.813) is the best algorithm for this dataset

Supervised Learning Methods for Classification



	Accuracy	F1-score
LDA	0.6442	0.7786
RDA	0.6442	0.7663
SVM	0.6587	0.7865
Naive Bayes	0.6731	0.8263
Neural Networks	0.6587	0.7953
Random Forest	0.6923	0.8132
KNN	0.6442	0.7958

Figure: Overall performance of each method

Cluster Analysis - First Approach

Distance Metric

Gower's Distance:

$$d_G(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{p} \frac{|x_k - y_k|}{r_k}$$

Cluster methods

- Hierarchical methods
- Partitioning method
- Density Based method

Cluster Analysis - Second Approach

Dimensionality Reduction

- Multidimensional Scaling and Principal Component Analysis

Distance Metric

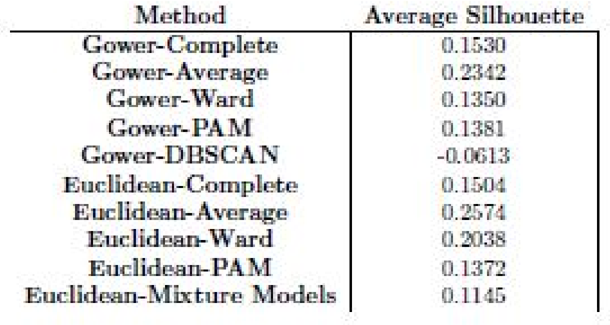
Euclidean Distance:

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{p} (x_k - y_k)^2}$$

Cluster methods

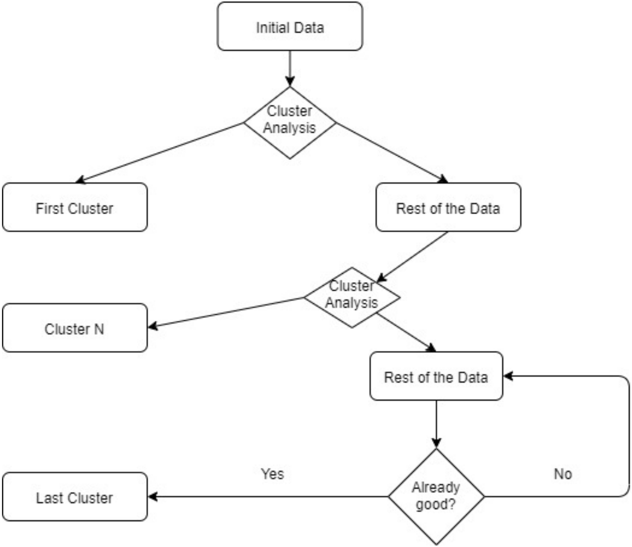
- Hierarchical methods
- Partitioning method
- Density Based method
- Mixture model method

Cluster Analysis - Results



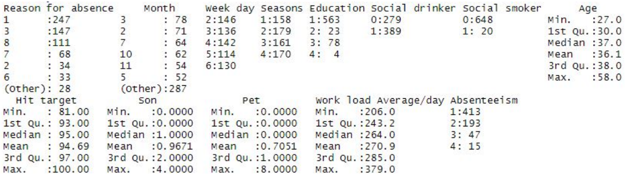
Method	Average Silhouette
Gower-Complete	0.1530
Gower-Average	0.2342
Gower-Ward	0.1350
Gower-PAM	0.1381
Gower-DBSCAN	-0.0613
Euclidean-Complete	0.1504
Euclidean-Average	0.2574
Euclidean-Ward	0.2038
Euclidean-PAM	0.1372
Euclidean-Mixture Models	0.1145

Cluster Analysis - Results(2)



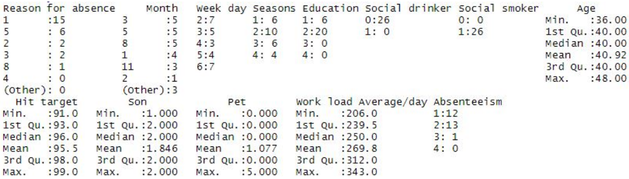
Cluster Analysis - Results(3)

First Cluster



```
Reason for absence      Month        Week day Seasons Education Social drinker Social smoker       Age
1        :247        3        : 78   2:146    1:158   1:563     0:279          0:648          Min.   :27.0
3        :147        2        : 71   3:136    2:179   2: 23     1:389          1: 20          1st Qu.:30.0
8        :111        7        : 64   4:142    3:161   3: 78                                   Median :37.0
7        : 68        10       : 62   5:114    4:170   4:  4                                   Mean   :36.1
2        : 34        11       : 54   6:130                                                    3rd Qu.:38.0
6        : 33        5        : 52                                                            Max.   :58.0
(Other): 28          (Other):287
  Hit target          Son              Pet            Work load Average/day Absenteeism
Min.   : 81.00   Min.   :0.0000   Min.   :0.0000   Min.   :206.0          1:413
1st Qu.: 93.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:243.2          2:193
Median : 95.00   Median :1.0000   Median :0.0000   Median :264.0          3: 47
Mean   : 94.69   Mean   :0.9671   Mean   :0.7051   Mean   :270.9          4: 15
3rd Qu.: 97.00   3rd Qu.:2.0000   3rd Qu.:1.0000   3rd Qu.:285.0
Max.   :100.00   Max.   :4.0000   Max.   :8.0000   Max.   :379.0
```

Second Cluster



```
Reason for absence      Month     Week day Seasons Education Social drinker Social smoker       Age
1        :15         3        :5  2:7       1: 6   1: 6      0:26           0: 0           Min.   :36.00
5        : 6         5        :5  3:5       2:10   2:20      1: 0           1:26           1st Qu.:40.00
2        : 2         8        :5  4:3       3: 6   3: 0                                    Median :40.00
3        : 2         1        :4  5:4       4: 4   4: 0                                    Mean   :40.92
8        : 1         11       :3  6:7                                                      3rd Qu.:40.00
4        : 0         2        :1                                                           Max.   :48.00
(Other): 0           (Other):3
  Hit target        Son             Pet           Work load Average/day Absenteeism
Min.   :91.0   Min.   :1.000   Min.   :0.000   Min.   :206.0          1:12
1st Qu.:93.0   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:239.5          2:13
Median :96.0   Median :2.000   Median :0.000   Median :250.0          3: 1
Mean   :95.5   Mean   :1.846   Mean   :1.077   Mean   :269.8          4: 0
3rd Qu.:98.0   3rd Qu.:2.000   3rd Qu.:0.000   3rd Qu.:312.0
Max.   :99.0   Max.   :2.000   Max.   :5.000   Max.   :343.0
```

Cluster Analysis - Problems and Recommendations

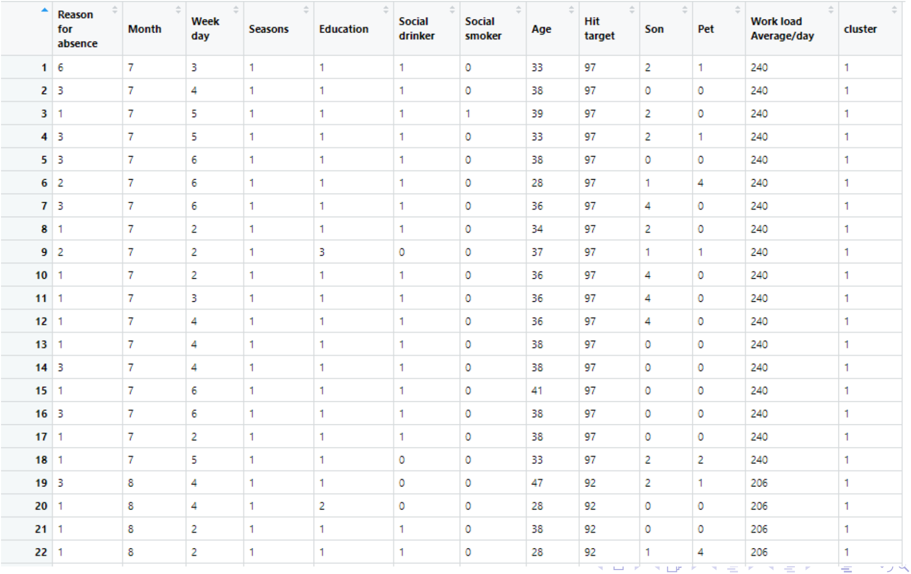
Problems

- High Dimensionality
- Imbalanced Dataset

Recommendations

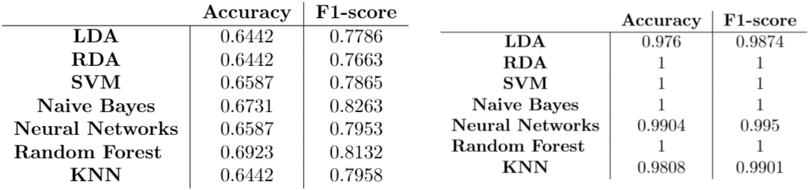
- Oversampling
- Fuzzy Clustering

Classification with new diagnosis variable



	Reason for absence	Month	Week day	Seasons	Education	Social drinker	Social smoker	Age	Hit target	Son	Pet	Work load Average/day	cluster
1	6	7	3	1	1	1	0	33	97	2	1	240	1
2	3	7	4	1	1	1	0	38	97	0	0	240	1
3	1	7	5	1	1	1	1	39	97	2	0	240	1
4	3	7	5	1	1	1	0	33	97	2	1	240	1
5	3	7	6	1	1	1	0	38	97	0	0	240	1
6	2	7	6	1	1	1	0	28	97	1	4	240	1
7	3	7	6	1	1	1	0	36	97	4	0	240	1
8	1	7	2	1	1	1	0	34	97	2	0	240	1
9	2	7	2	1	3	0	0	37	97	1	1	240	1
10	1	7	2	1	1	1	0	36	97	4	0	240	1
11	1	7	3	1	1	1	0	36	97	4	0	240	1
12	1	7	4	1	1	1	0	36	97	4	0	240	1
13	1	7	4	1	1	1	0	38	97	0	0	240	1
14	3	7	4	1	1	1	0	38	97	0	0	240	1
15	1	7	6	1	1	1	0	41	97	0	0	240	1
16	3	7	6	1	1	1	0	38	97	0	0	240	1
17	1	7	2	1	1	1	0	38	97	0	0	240	1
18	1	7	5	1	1	0	0	33	97	2	2	240	1
19	3	8	4	1	1	0	0	47	92	2	1	206	1
20	1	8	4	1	2	0	0	28	92	0	0	206	1
21	1	8	2	1	1	1	0	38	92	0	0	206	1
22	1	8	2	1	1	1	0	28	92	1	4	206	1

Supervised Learning Methods for Classification



	Accuracy	F1-score
LDA	0.6442	0.7786
RDA	0.6442	0.7663
SVM	0.6587	0.7865
Naive Bayes	0.6731	0.8263
Neural Networks	0.6587	0.7953
Random Forest	0.6923	0.8132
KNN	0.6442	0.7958

	Accuracy	F1-score
LDA	0.976	0.9874
RDA	1	1
SVM	1	1
Naive Bayes	1	1
Neural Networks	0.9904	0.995
Random Forest	1	1
KNN	0.9808	0.9901

Figure: Overall performance of each method with *Absenteeism* variable and clustering results

Compare Performance: K- Nearest Neighbors

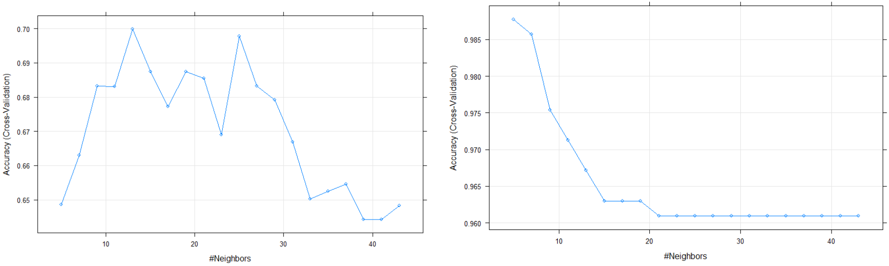
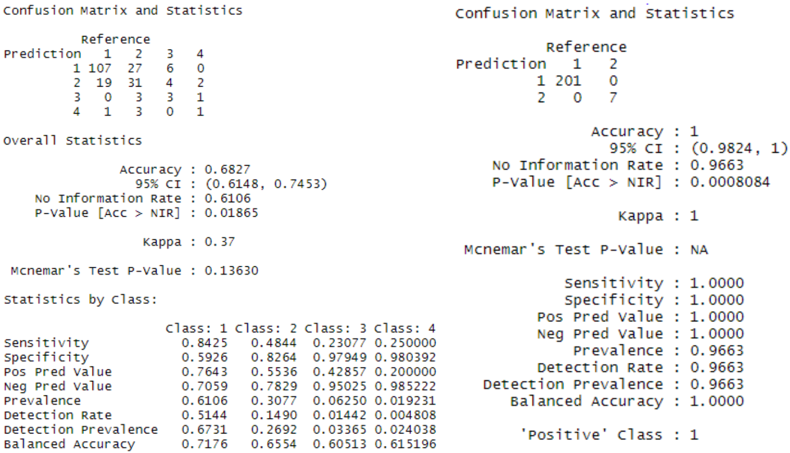


Figure: Relation between *k* values and accuracy of the respectively model, with *Absenteeism* variable on the left and clustering results on the right

Compare Performance: Random Forest



```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4
         1 107  27   6   0
         2  19  31   4   2
         3   0   3   3   1
         4   1   3   0   1

Overall Statistics

               Accuracy : 0.6827
                 95% CI : (0.6148, 0.7453)
    No Information Rate : 0.6106
    P-Value [Acc > NIR] : 0.01865

                  Kappa : 0.37

 Mcnemar's Test P-Value : 0.13630

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8425   0.4844  0.23077 0.250000
Specificity            0.5926   0.8264  0.97949 0.980392
Pos Pred Value         0.7643   0.5536  0.42857 0.200000
Neg Pred Value         0.7059   0.7829  0.95025 0.985222
Prevalence             0.6106   0.3077  0.06250 0.019231
Detection Rate         0.5144   0.1490  0.01442 0.004808
Detection Prevalence   0.6731   0.2692  0.03365 0.024038
Balanced Accuracy      0.7176   0.6554  0.60513 0.615196
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 201   0
         2   0   7

               Accuracy : 1
                 95% CI : (0.9824, 1)
    No Information Rate : 0.9663
    P-Value [Acc > NIR] : 0.0008084

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.9663
         Detection Rate : 0.9663
   Detection Prevalence : 0.9663
      Balanced Accuracy : 1.0000

       'Positive' Class : 1
```

Figure: RF's performance evaluators with *Absenteeism* variable and clustering results

Conclusion

- Strongly imbalanced data
- Best obtained classification method for *Absenteeism*: Random Forest
- Few relations that allow group into clusters
- Generalized clusters which favors prediction

Multivariate Analysis

Catarina Costa nº86582
Inês Tavares nº86593
Catarina Oliveira nº86963
Gonçalo Mestre nº 87005

Instituto Superior Técnico, Lisboa

January 2020