Prediction of the absentee's level of education

Dataset: Absenteeism_at_work
Author: Raisler Voigt | GitHub , Linkedin

A Brief Explanation

Absenteeism is the habitual non-presence of an employee at his or her job. Habitual non-presence extends beyond what is expected as a normal amount of time away for reasons such as scheduled vacation or occasional illness.

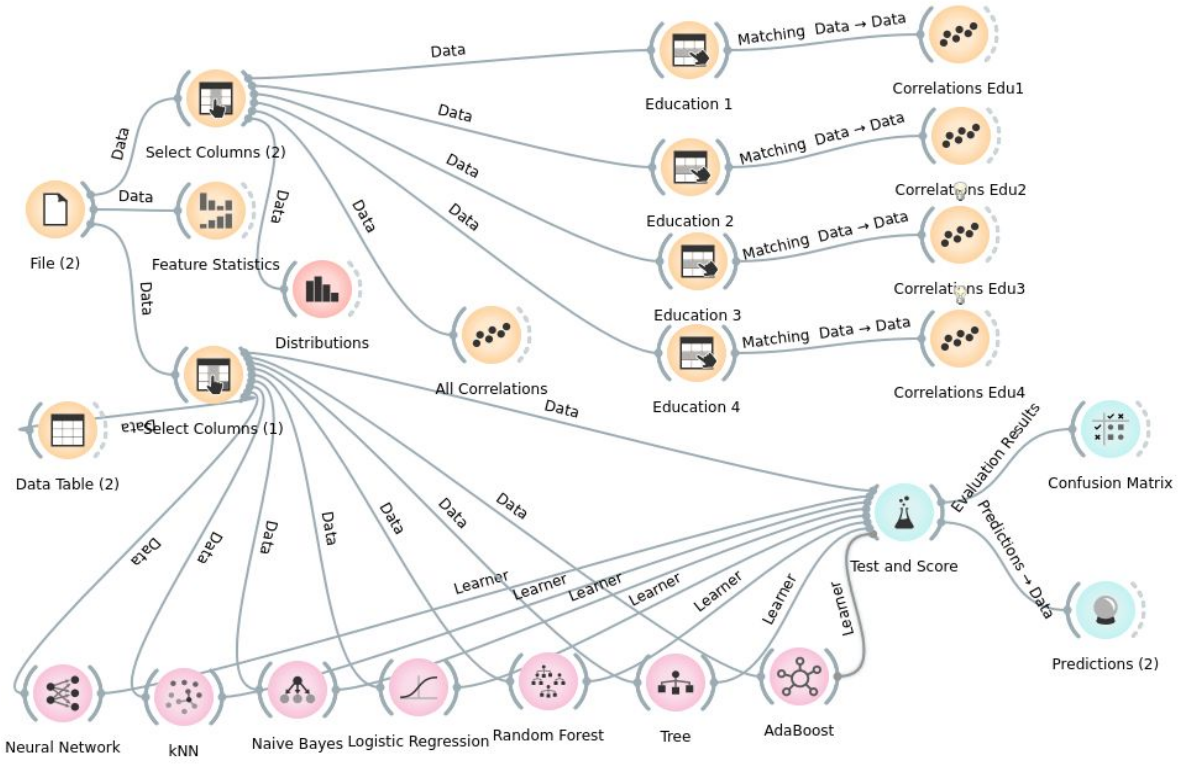
The goal and analysis methodology

Does the level of education correlation with the other attributes of the dataset? The objective is to answer this question in a few different ways, and see if it is possible to predict the education level.

We can see whether it makes sense for the education level in the absenteeism of the people, using classification algorithms to classify using the level of education(1 to 4) as a class to predict and to see the accuracy we can achieve.

I will not use python, not directly, there is a great tool to data mining called Orange, with widgets to make some actions, including widgets with machine learning models.

Pipeline



Data

You can see all attributes in the link as the information about each one.

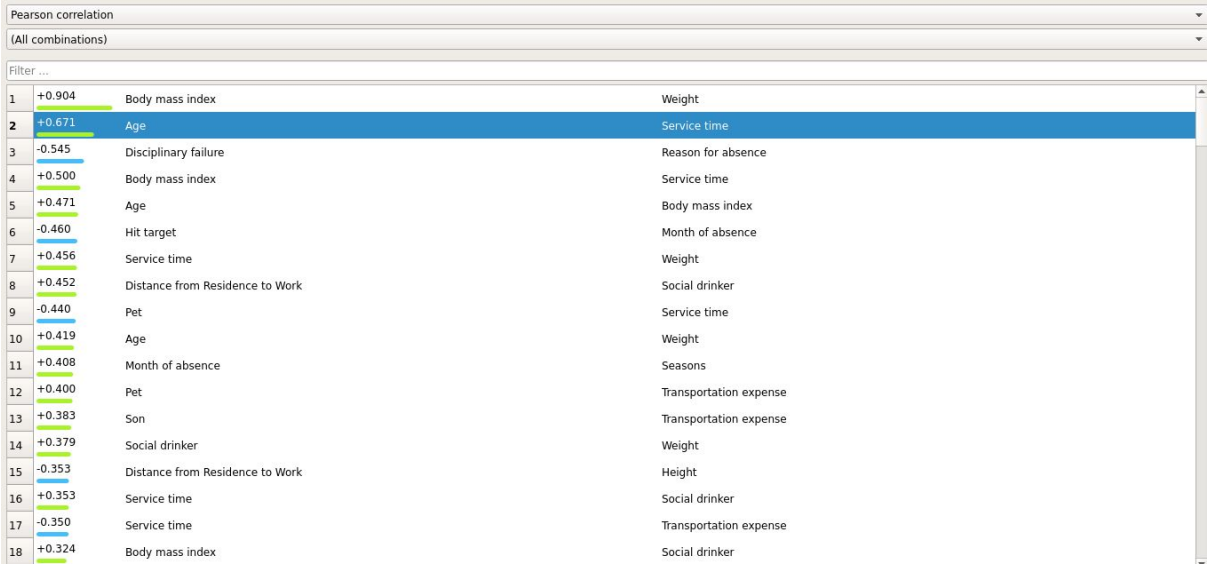
Preprocessing

All attributes, except education, for this project will be the numeric type. Intuitively, only a few attributes were chosen for use in training the algorithms.

The chosen ones: Reason for absence (ICD), Month of absence, Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6)), Seasons, Transportation expense, Distance from Residence to Work (kilometers), Service time, Age, Work load Average/day, Disciplinary failure (yes=1; no=0), Education (high school (1), graduate (2), postgraduate (3), master and doctor (4)), Son (number of children), Social drinker (yes=1; no=0), Social smoker (yes=1; no=0).

There is no outliers, missing values or anomalies, based on features statistics.

All Correlations



1	+0.904	Body mass index	Weight
2	+0.671	Age	Service time
3	-0.545	Disciplinary failure	Reason for absence
4	+0.500	Body mass index	Service time
5	+0.471	Age	Body mass index
6	-0.460	Hit target	Month of absence
7	+0.456	Service time	Weight
8	+0.452	Distance from Residence to Work	Social drinker
9	-0.440	Pet	Service time
10	+0.419	Age	Weight
11	+0.408	Month of absence	Seasons
12	+0.400	Pet	Transportation expense
13	+0.383	Son	Transportation expense
14	+0.379	Social drinker	Weight
15	-0.353	Distance from Residence to Work	Height
16	+0.353	Service time	Social drinker
17	-0.350	Service time	Transportation expense
18	+0.324	Body mass index	Social drinker

All Correlations

With the Pearson correlation coefficient, I tried to see what correlations was stronger, no condition, just with the naked data, but had no meaningful results, the stronger correlation was about body mass index with weight (an expected correlation), age and time service makes sense, but I expected more. It's possible to conclude the data points are well dispersed.

Applying conditions in the dataset based in the level education

The idea is to see the differences if we put the conditional data only for one level of education.

Correlations between attributes of only those with level 1 in education

Pearson correlation

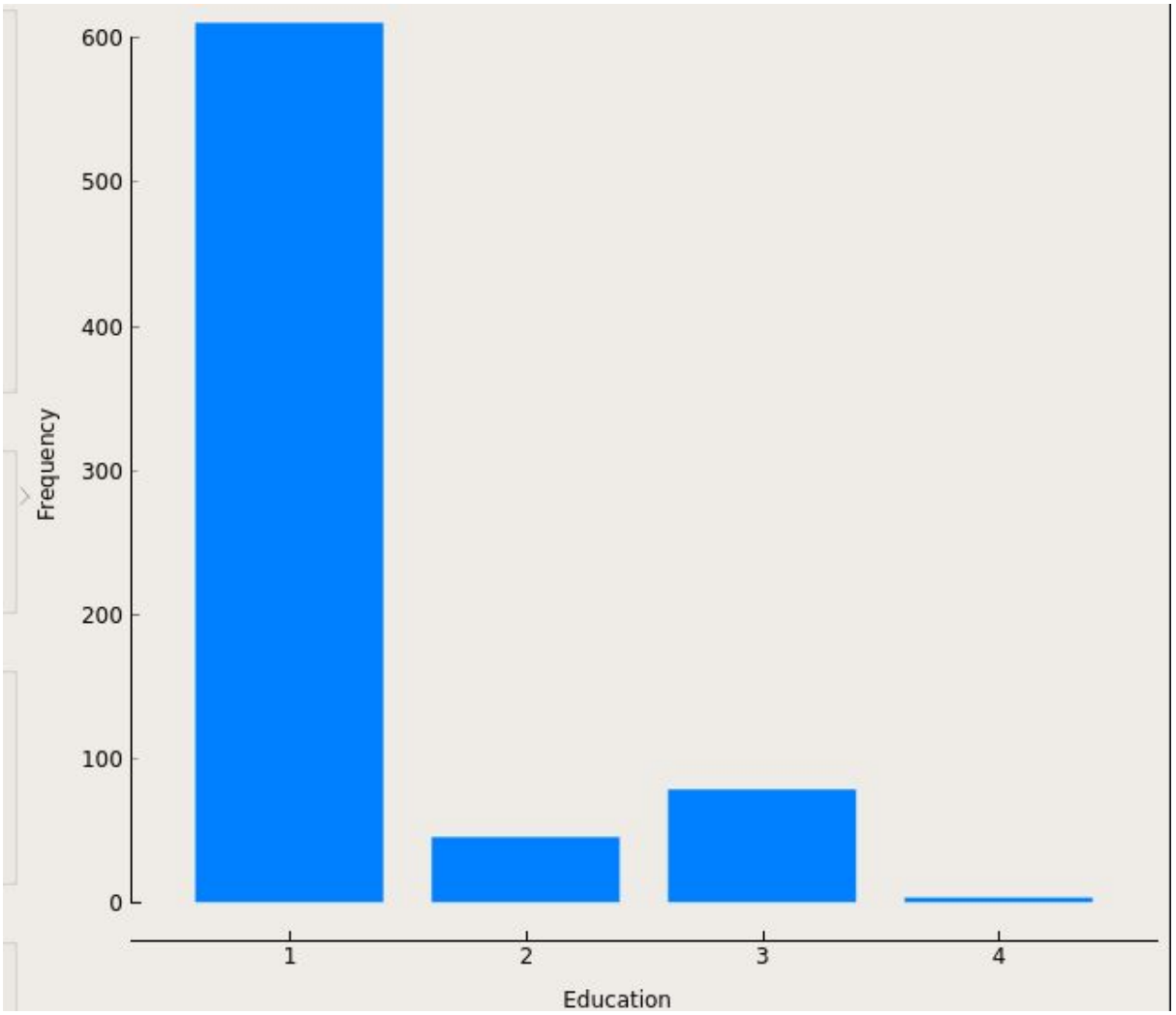
(All combinations)

Filter ...

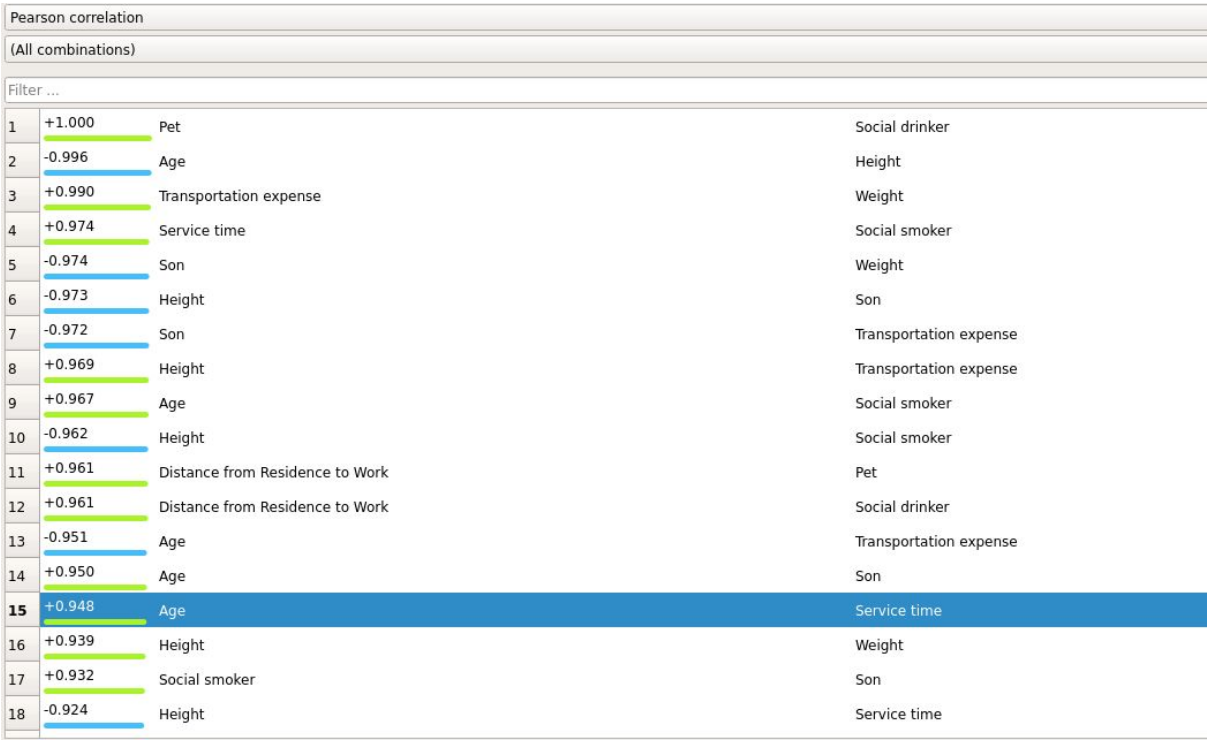
1	+0.880	Body mass index	Weight
2	+0.627	Age	Service time
3	-0.505	Pet	Service time
4	+0.503	Service time	Weight

It is still dispersed, because according to the distribution chart below, level 1 is the majority, and already knowing that the data do not have a pattern that facilitates the analysis, conditioning the dataset only to the level, would follow more or less what all previous correlations showed.

Classes Distribution (Education Level)



Correlations between attributes of only those with level 2 in education



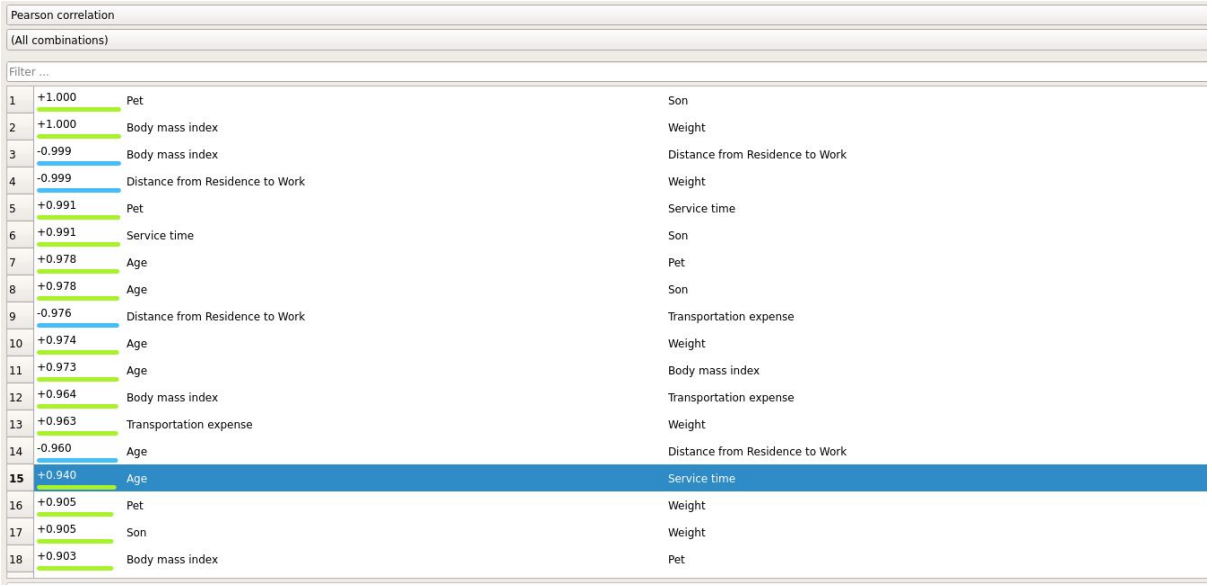
Pearson correlation

(All combinations)

Filter ...

1	+1.000	Pet	Social drinker
2	-0.996	Age	Height
3	+0.990	Transportation expense	Weight
4	+0.974	Service time	Social smoker
5	-0.974	Son	Weight
6	-0.973	Height	Son
7	-0.972	Son	Transportation expense
8	+0.969	Height	Transportation expense
9	+0.967	Age	Social smoker
10	-0.962	Height	Social smoker
11	+0.961	Distance from Residence to Work	Pet
12	+0.961	Distance from Residence to Work	Social drinker
13	-0.951	Age	Transportation expense
14	+0.950	Age	Son
15	+0.948	Age	Service time
16	+0.939	Height	Weight
17	+0.932	Social smoker	Son
18	-0.924	Height	Service time

Correlations between attributes of only those with level 3 in education



Pearson correlation

(All combinations)

Filter ...

1	+1.000	Pet	Son
2	+1.000	Body mass index	Weight
3	-0.999	Body mass index	Distance from Residence to Work
4	-0.999	Distance from Residence to Work	Weight
5	+0.991	Pet	Service time
6	+0.991	Service time	Son
7	+0.978	Age	Pet
8	+0.978	Age	Son
9	-0.976	Distance from Residence to Work	Transportation expense
10	+0.974	Age	Weight
11	+0.973	Age	Body mass index
12	+0.964	Body mass index	Transportation expense
13	+0.963	Transportation expense	Weight
14	-0.960	Age	Distance from Residence to Work
15	+0.940	Age	Service time
16	+0.905	Pet	Weight
17	+0.905	Son	Weight
18	+0.903	Body mass index	Pet

We will not look at level 4, since there are only 4 and it will not be of great relevance at the moment, but it is possible to see that at least in our dataset it is possible to see that among people with undergraduate and graduate there is a better standardization than what happens with them, and it is easier to correlate causes and effects, there are more than 18 strong correlations (more than 90%), it is even possible to analyze there are negative and positive correlations, the coolest thing in level 3 is to look that the weight of the person has a negative correlation with the distance from home to work.

Classification

The attempt will be to see if any classing algorithm can predict the level of education of the person based on the values of some attributes chosen based on the correlations observed previously. Even if our classes are not well distributed and can have bias, it can be that in the degree 2 and 3 he can get right by the strong correlations and the rest he put as 1, what algorithm performs better? and for that we will use almost all the algorithms present in the Orange tool.

Features: Reason for absence, Distance from Residence to Work, Service Time, Age, Transportation expense, Work load Average/day, Son, Social drinker, Social smoker, Weight, Pet, Height, Body mass index, Absenteeism time in hours.

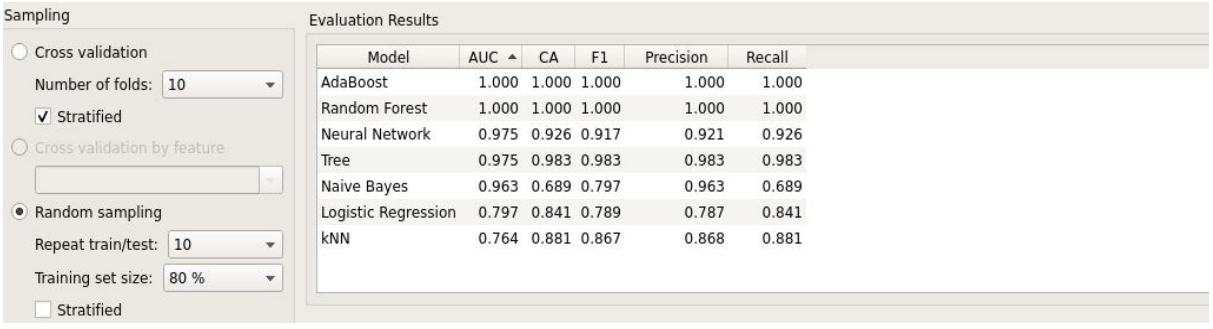
Algorithms list (only those present in the tool):

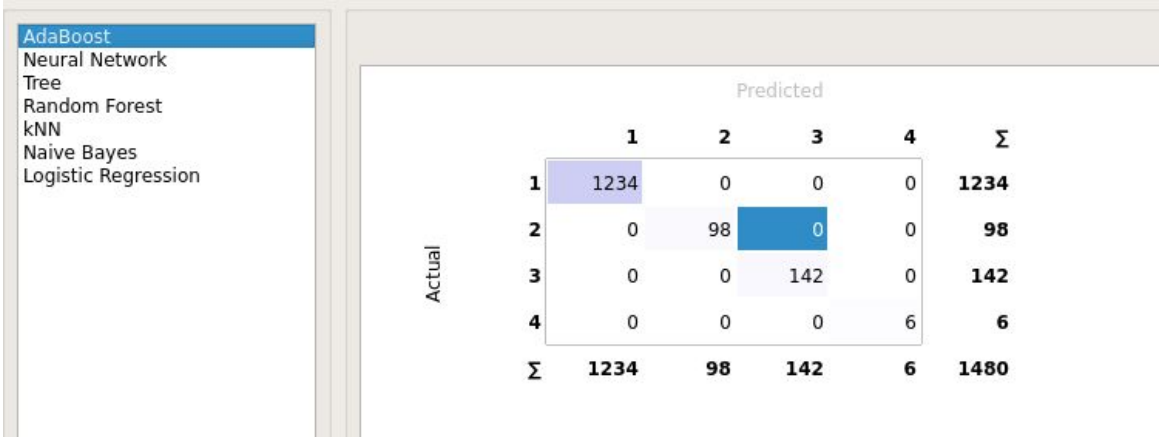
- Naive Bayes
- Decision Tree
- Logistic Regression (Multi-Class)
- Rede Neural (200 iterations, solver: SGD, Ativação: Relu, 100 neurônios)
- kNN (10 neighbors, metric: Euclidean, weight: Distance)
- Random Forest (20 trees)
- AdaBoost (50 estimators, Learning Rate = 1, classifier: SAMME.R, regression loss function: Linear, base estimator: tree)

Metrics

The tests were done with 20% of the data, and the training with 80%.

Evaluation and Confusion Matrix





Conclusion

Even with an unequal distribution of classes, it was possible to predict the level of education of the individuals with 100% in all metrics with the algorithms AdaBoost and Random Forest, a great indicator.

It is possible to know why showing a very strong correlation at levels 2 and 3, and a dispersion at level 1, apparently when the features are very well correlated it tends to the education classes 2 and 3.

Could have more data, especially to balance the education level in this case, but there is a pattern about the people with higher education, and it’s clear, even with the small sample and the objective based on a single attribute.