Previsão do nível de educação do ausente.

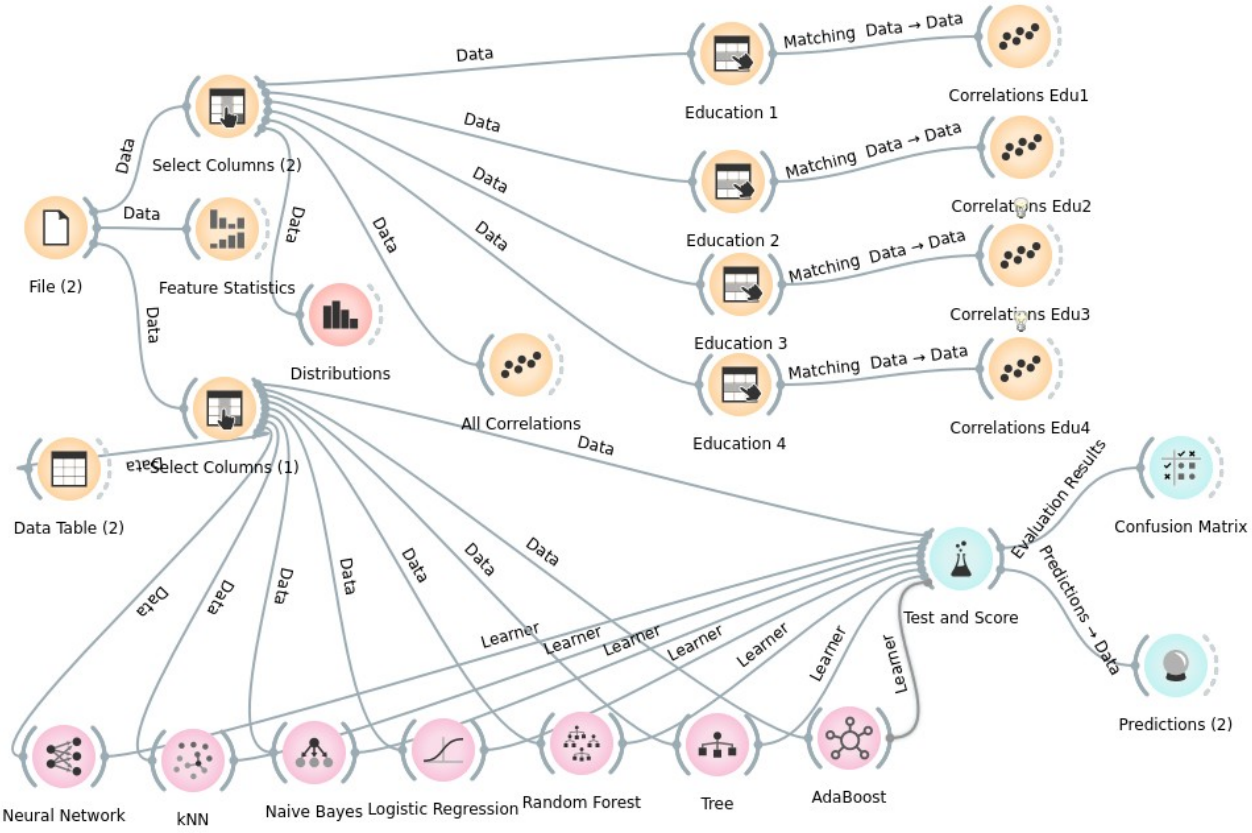
Autor: Raisler Voigt

Dataset escolhido: Absenteeism_at_work

Objetivo e metodologia da análise:

O nível de educação tem uma correlação com os outros atributos do dataset? O objetivo é responder essa pergunta de algumas formas diferentes, utilizando algoritmos de classificação para classificar utilizando o grau de educação(1 a 4) como a classe a ser prevista e ver a acurácia que podemos atingir, e usando condicionais para ver correlações mais fortes do dataset condicionando o dataset ser apenas com as linhas que possuem o atributo Educação igual ao nível escolhido.

Pipeline



Pré-Processamento

Todos os atributos, exceto Educação que é do tipo categórico que será utilizado para classificação, para este projeto será do tipo numérico. E intuitivamente, foram escolhidos apenas alguns atributos para uso no treinamento dos algoritmos. Os atributos escolhidos para olhar as correlações foram: 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)

Todas as correlações

Pearson correlation

(All combinations)

Filter ...

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

Todas as correlações

Usando o coeficiente de Pearson tentei ver quais eram as correlações mais fortes do dataset, sem nenhuma condicional, porém não houve resultados expressivo, a correlação mais forte que é de 90% foi da indíce de massa corporal com o peso, o que já era um pouco esperado, idade e o tempo de serviço fazem sentidos, mas houve apenas 60%, esperava mais, o resto não considerei significante, olhando assim é possível concluir que graficamente os data points estão bem dispersos.

Aplicando condicionais no dataset baseado no grau de Educação

Correlações entre os atributos apenas de quem possui educação de nível 1

Pearson correlation

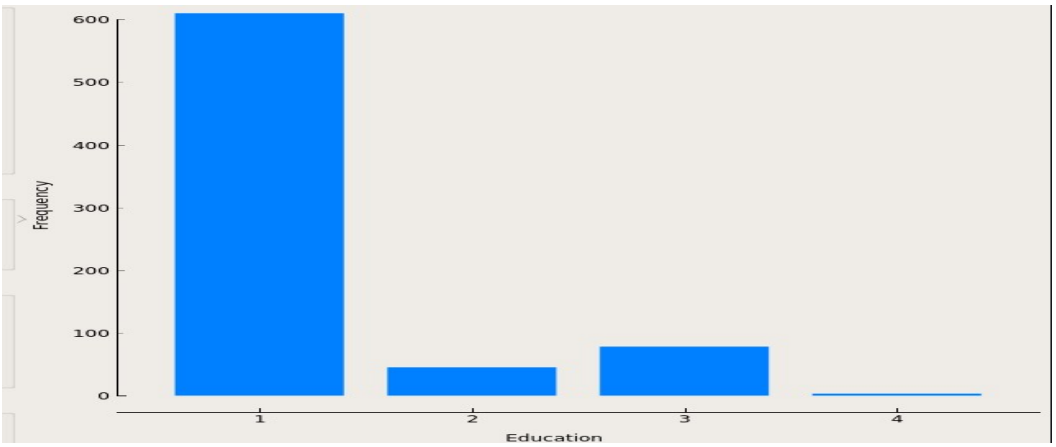
(All combinations)

Filter ...

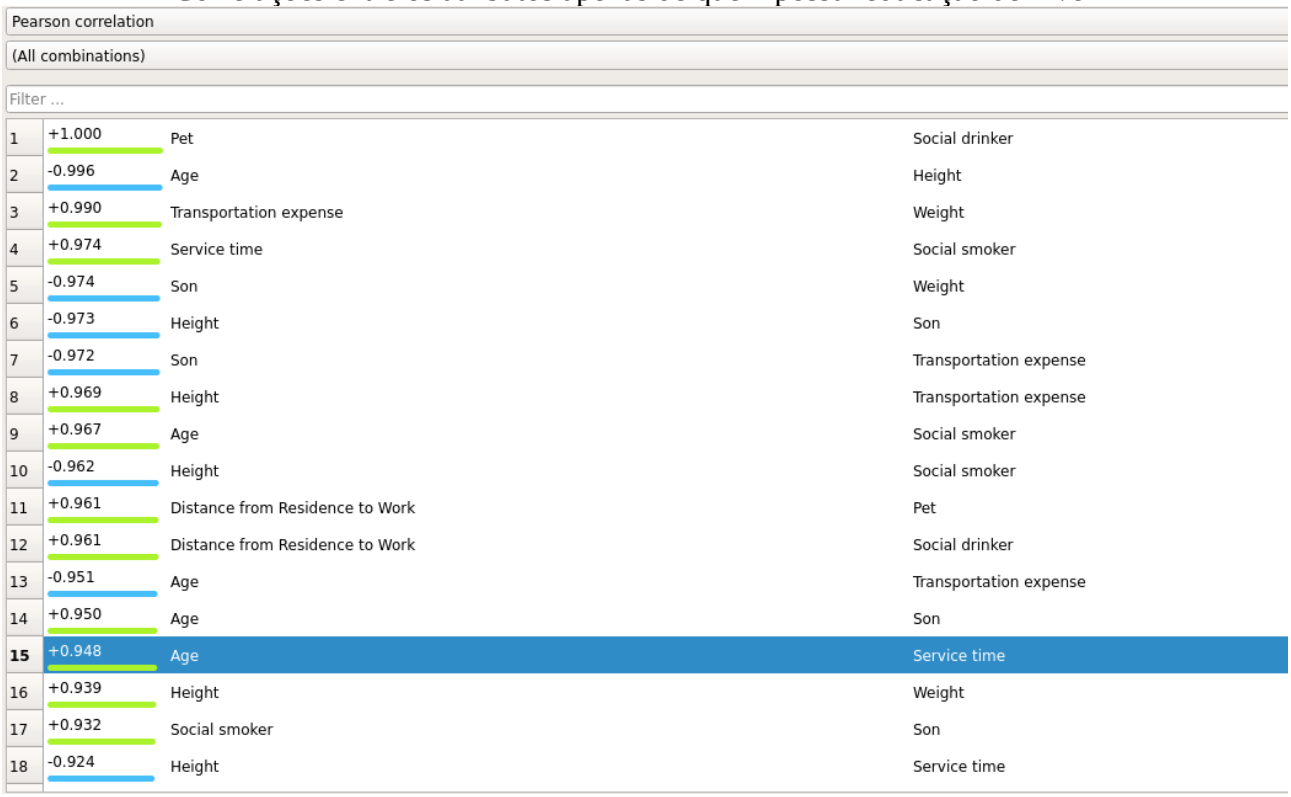
1	+0.880	Body mass index	Weight
2	+0.627	Age	Service time
3	-0.505	Pet	Service time
4	+0.503	Service time	Weight

Ainda continua disperso, isso porque segundo o gráfico de distribuição abaixo, nível 1 é a maioria, e já sabendo que os dados não possuem um padrão que facilite a análise, condicionando o dataset apenas para o nível, seguiria mais ou menos o que todas as correlações anteriores mostraram.

Distribuição das classes (níveis de Educação)



Correlações entre os atributos apenas de quem possui educação de nível 2



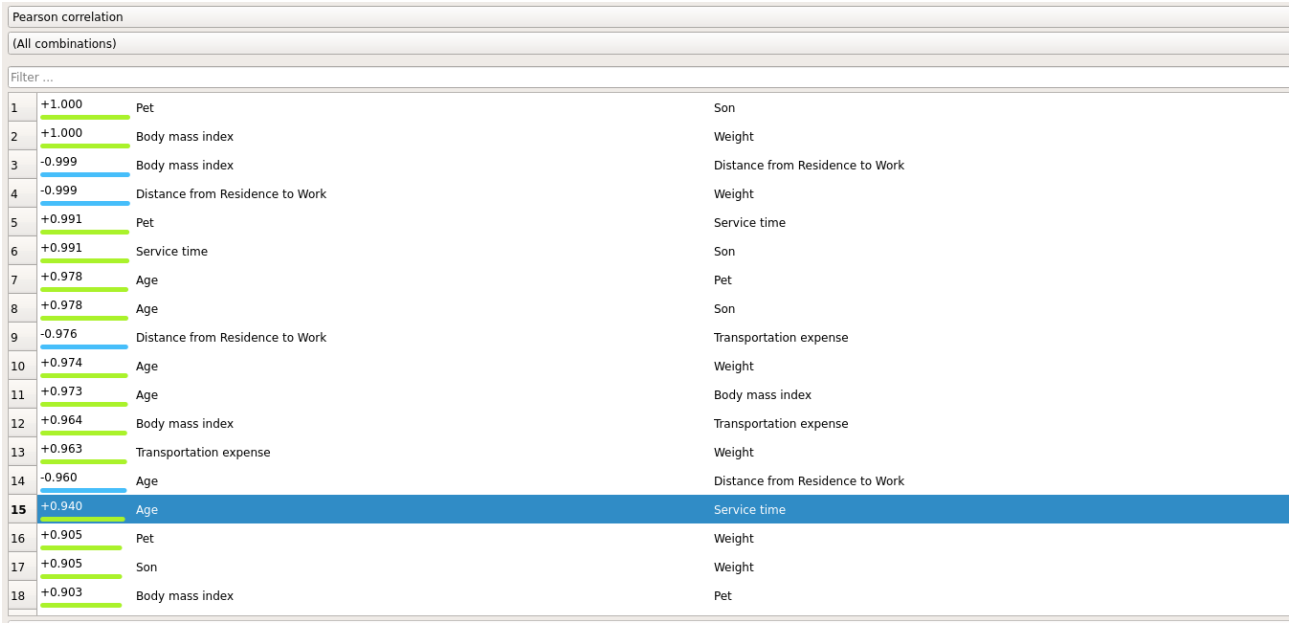
Pearson correlation

(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

Correlações entre os atributos apenas de quem possui educação de nível 3



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

Não olharemos de nível 4, já que existe apenas 4 e não vai ser de grande relevância no momento, mas é possível ver que que pelo menos em nosso dataset é possível ver que entre pessoas com gradução e pós-graduação há uma padronização melhor do que acontece com elas, e fica mais fácil correlacionar as causas e efeitos, há mais de 18 correlações fortes (mais de 90%), inclusive é possível analisar existem correlações negativas e positivas, o mais legal no nível 3 é olhar que a peso da pessoa tem uma correlação negativa com a distância da casa ao trabalho.

Classificação

A tentativa será ver se algum algoritmo de classicação consegue prever qual o nível de educação da pessoa baseado nos valores de alguns atributos escolhidos com base nas correlações observadas anteriormente. Mesmo nossas classes não sendo bem distruibuídas e podendo ter enviesamento, pode ser que no grau 2 e 3 ele consiga acertar pelas correlações fortes e o resto ele coloque como 1, será testado qual algoritmo performa melhor, e para isso utilizaremos quase todos os algoritmos presentes na ferramenta Orange.

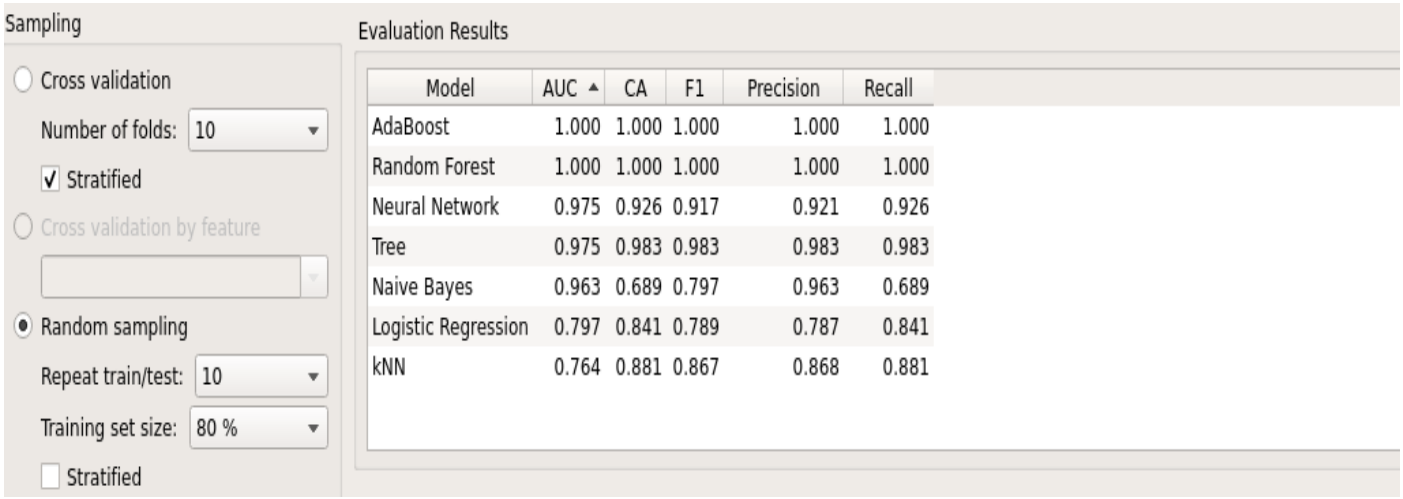
Nossas features: Reason for absence, Distance from Residence to Work, Service Time, Age, Transportation expense, Work load Average/day, Son, Social drinker, Social smoker, Weight, Pet, Height, Body mass index, Absenteeism time in hours.

Lista de algoritmos utilizados (apenas os presentes na ferramenta)

- 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)

Os testes foram feitos com 20% dos dados, e o treino com 80%.

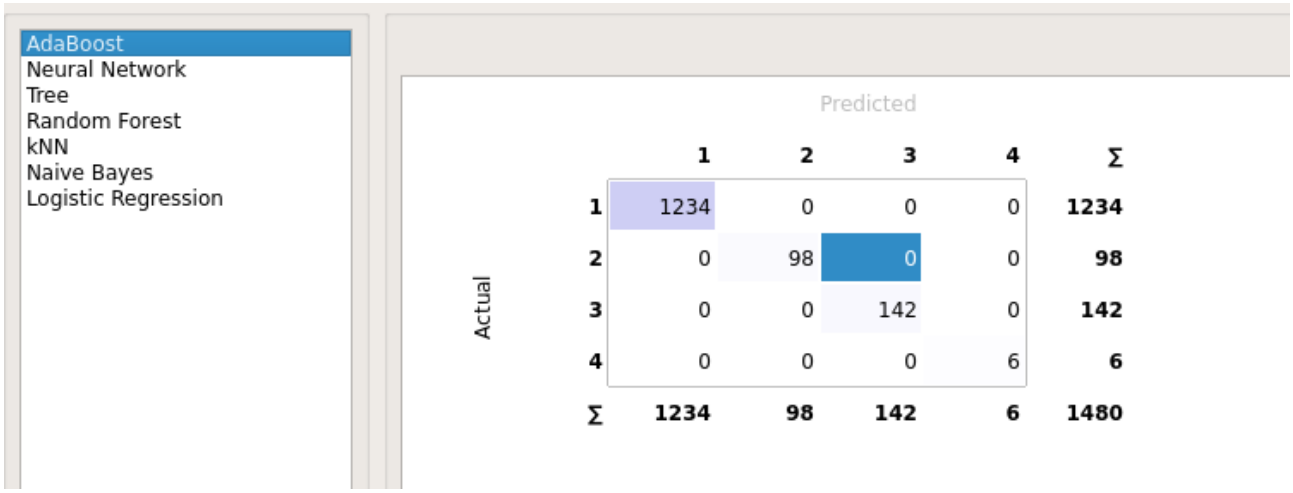
Evaluation



Model	AUC	CA	F1	Precision	Recall
AdaBoost	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	1.000	1.000	1.000	1.000
Neural Network	0.975	0.926	0.917	0.921	0.926
Tree	0.975	0.983	0.983	0.983	0.983
Naive Bayes	0.963	0.689	0.797	0.963	0.689
Logistic Regression	0.797	0.841	0.789	0.787	0.841
kNN	0.764	0.881	0.867	0.868	0.881

Apesar do enviesamento para acertar o nível de educação já que a maioria é de nível 1, o algoritmo AdaBoost nesse teste e em vários outros foi o que acertou com 100%, como pode-se ver na matriz de confusão.

Matriz de Confusão



Conclusão

Mesmo com uma distribuição desigual de classes, foi possível prever qual o nível de educação do indivíduo, e tendo resultados de 100% em dois algoritmos, e bons resultados em outros algoritmos.

Isso se deve ao ter uma correlação muito forte nos níveis 2 e 3, e uma dispersão no nível 1, aparentemente quando as features estão muito bem correlacionadas ele tende para as classes de educação 2 e 3.