Projeto Final - Data Science 2024.2

Integrantes: Anderson Santos Alves da Silva (asas@cesar.school)
Carlo Romero Lira (crls2@cesar.school)

Dataset : https://www.kaggle.com/datasets/yashpaloswal/ann-car-sales-price-prediction

O dataset escolhido tem como objetivo prever o valor de compra do carro com base nas informações dos consumidores. A imagem abaixo ilustra as primeiras linhas do CSV:

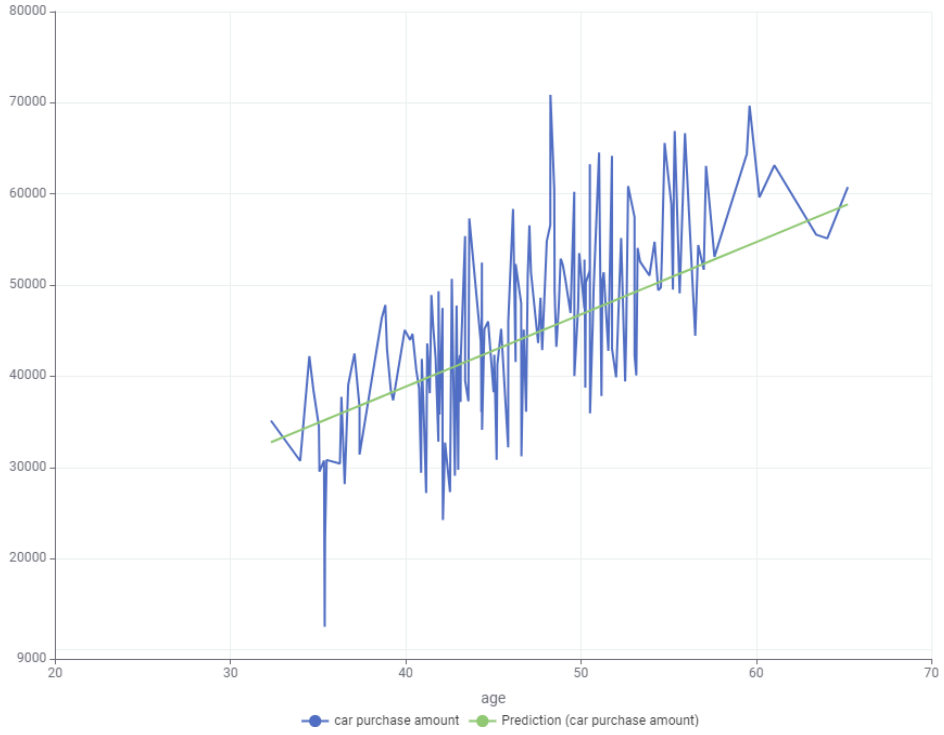
Rows: 500 | Columns: 9 — Table | Statistics

#	RowID	customer ... String	customer ... String	country String	gender Number (integ...	age Number (doub...	annual Sa... Number (doub...	credit car... Number (doub...	net worth Number (doub...	car purch... Number (doub...
1	Row0	Martina Avila	cubilia.Curae.Ph	Bulgaria	0	41.852	62,812.093	11,609.381	238,961.25	35,321.459
2	Row1	Harlan Barnes	eu.dolor@diam.	Belize	0	40.871	66,646.893	9,572.957	530,973.908	45,115.526
3	Row2	Naomi Rodrique	vulputate.mauri	Algeria	1	43.153	53,798.551	11,160.355	638,467.177	42,925.709
4	Row3	Jade Cunningha	malesuada@dig	Cook Islands	1	58.271	79,370.038	14,426.165	548,599.052	67,422.363
5	Row4	Cedric Leach	felis.ullamcorpe	Brazil	1	57.314	59,729.151	5,358.712	560,304.067	55,915.462
6	Row5	Carla Hester	mi@Aliquamera	Liberia	1	56.825	68,499.852	14,179.472	428,485.36	56,611.998
7	Row6	Griffin Rivera	vehicula@at.co.	Syria	1	46.607	39,814.522	5,958.46	326,373.181	28,925.705
8	Row7	Orli Casey	nunc.est.mollis(	Czech Republic	1	50.193	51,752.234	10,985.697	629,312.404	47,434.983
9	Row8	Marny Obrien	Phasellus@sed	Armenia	0	46.585	58,139.259	3,440.824	630,059.027	48,013.614
10	Row9	Rhonda Chavez	nec@nuncest.c	Somalia	1	43.324	53,457.101	12,884.079	476,643.354	38,189.506

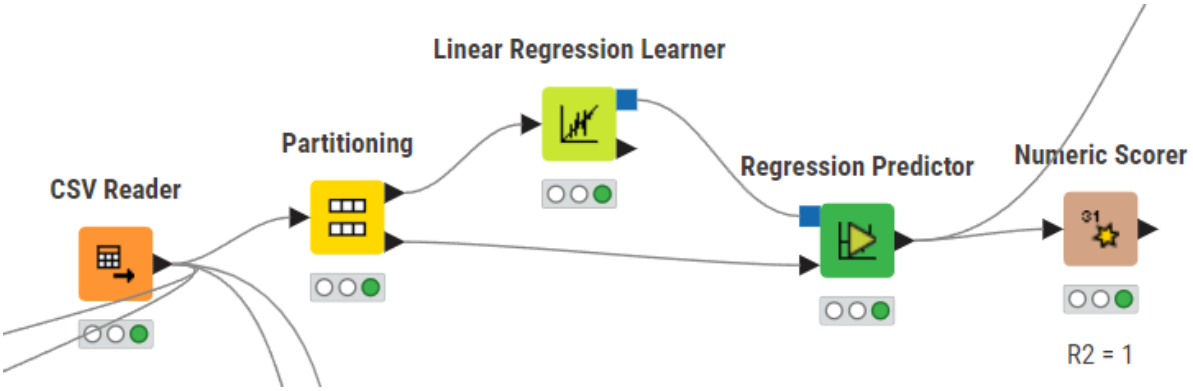
O primeiro passo foi analisar a correlação entre as colunas, utilizando o *node* ***Linear Correlation***. A imagem abaixo apresenta as colunas com maior relação: *Age*, *Annual Salary* e *Net Worth*.

#	RowID	customer ... Number (doub...	customer ... Number (doub...	country Number (doub...	gender Number (doub...	age Number (doub...	annual Sal... Number (doub...	credit car... Number (doub...	net worth Number (doub...	car purch... Number (doub...
9	car pur	?	?	?	-0.066	0.633	0.618	0.029	0.489	1

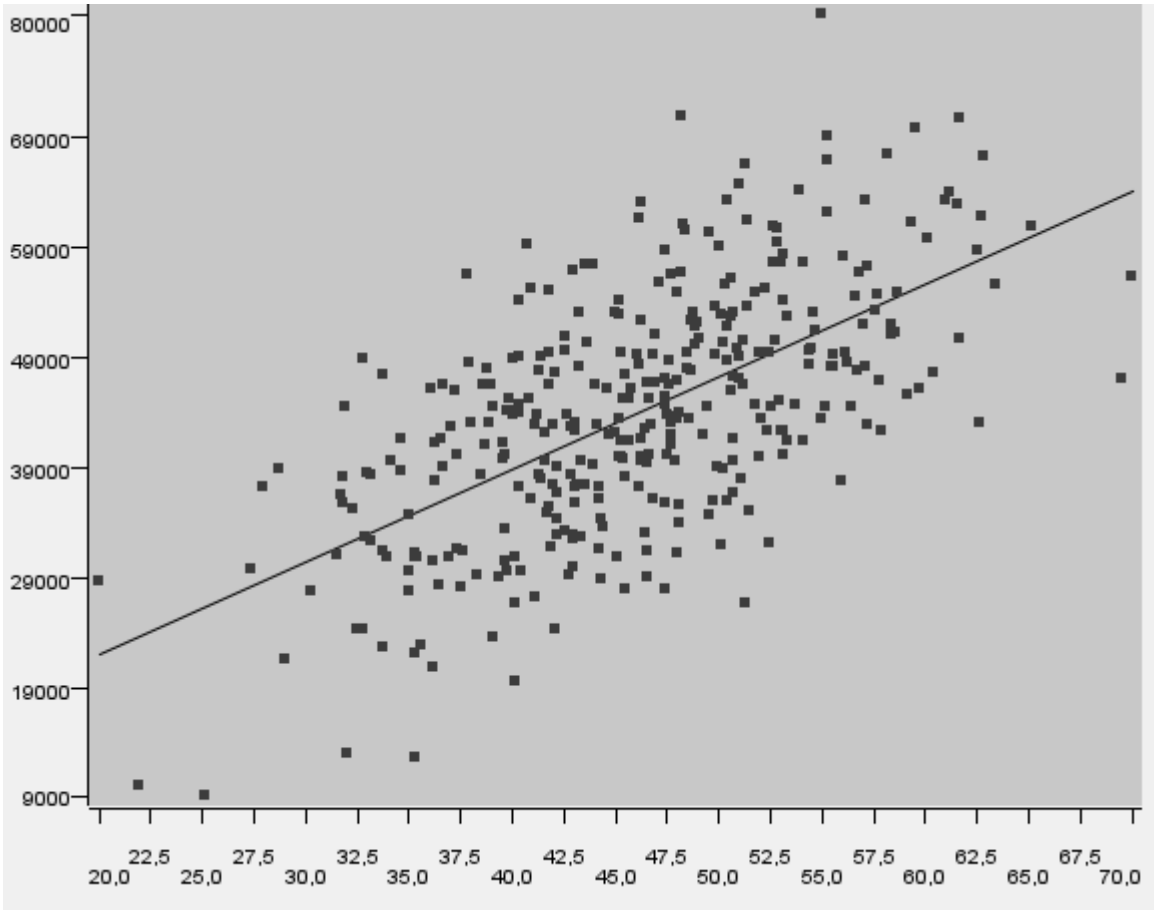
Treinando o *Learner* com input da coluna com maior correlação (*Age*), sem normalizar e particionando 70% para treinamento, obtive um **R^2** de apenas **0.422.** A previsão baseada apenas na idade não ficou precisa, teve *Underfitting*. A imagem abaixo mostra a diferença entre a previsão e o valor alvo.



O próximo passo foi treinar o *Learner* com os 3 parâmetros correlatos: *Age*, *Annual Salary* e *Net Worth*. Com as novas colunas, foi obtido o **R²** de ±**1**.



Abaixo está o Scatterplot view do LinearRegressionLearner, assim como as estatísticas.



Variable	Coeff.	Std. Err.	t-value	P>|t|
age	841,5743	0,01	83.772,1019	0.0
annual Salary	0,5623	6,70E-6	83.871,5746	0.0
net worth	0,029	4,68E-7	61.945,7277	0.0
Intercept	-42.147,9226	0,655	-64.343,2615	0.0

Utilizar todas as colunas no Learner também traz uma boa acurácia mas resulta em um Erro Quadrático Médio (MSE) e Erro Médio Absoluto (MAE) maiores comparados a utilizar apenas as 3 colunas mais correlatas.

AGE, ANNUAL SALARY E NET WORTH

RowID	Prediction (car purchase amount) *Number (double)*
R^2	1
mean absolute error	1.116
mean squared error	1.947

TODAS AS COLUNAS

RowID	Prediction (car purchase amount) *Number (double)*
R^2	1
mean absolute error	1.226
mean squared error	2.218

Abaixo segue a visualização completa do workflow: