

ML - Machine Learning

Lidia Fernandes Magalhães, Gisele Ferreira Araujo, Ruam Coimbra

Instituto de Ciências Exatas e Informática - Pontifícia Universidade Católica de Minas
Gerais

Instituto de Educação Continuada - Pós Graduação Lato Sensu em Ciência de Dados e
Big Data

1. Machine Learning

Machine learning ou aprendizado da máquina é uma área da ciência da computação que permite automatizar respostas ao usuário a partir de inteligência artificial e big data. Ela permite que seja desenvolvidos modelos onde o computador entende e aprende por conta própria aprimorando cada vez mais seu entendimento sobre os fatos.

A história do ML começa na década de 1950, quando o pai da computação, Alan Turing, fez a seguinte pergunta: “As máquinas podem pensar?”. Nessa época, ele desenvolveu o famoso Teste de Turing, que testava a capacidade das máquinas de raciocinarem como seres humanos.

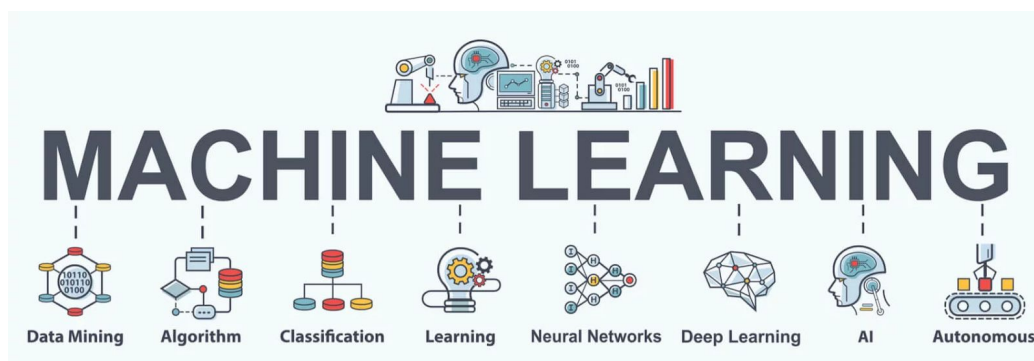


Figura 1 - ML

2. Motivação do projeto

Identificamos uma necessidade de melhorar a forma que é realizado o cálculo de tendência de algumas métricas (faturamento e volume) de uma empresa. Hoje o calculo é realizado baseado em uma série de cálculos onde é levado em conta o peso de acordo com o dia que está sendo calculado, exemplo:

Dia 05/03/2020

Peso 5

Valor faturado até o momento R\$ 10.000,00

1	5 30	10.000,00 x
2	300.000,00	5x
3	x	300.000,00 5
4	x	60.000,00

Figura 2 - Regra de Cálculo

Para prever os valores dos demais dias o cálculo realizado era baseado na troca do denominador pela quantidade de dias corridos até o momento, sem desconsiderar feriados e fins de semanas.

3. Hipóteses

Ao realizar uma análise sobre o histórico de dados de faturamento da empresa espera-se obter um modelo que retorne uma previsão de faturamento para o ano de 2020.

4. Metodologia

Para realização deste trabalho foram utilizados os dados de faturamento de uma unidade da empresa e as seguintes etapas foram realizadas:

4.1 Problema

Devido ao crescimento da empresa viu-se a necessidade de criação de novas metodologias para entendimento do resultado como um todo, pois foi identificado que algumas técnicas utilizadas para calcular resultados financeiros eram obsoletos ou desatualizados. Desse modo, entendemos que é de grande importância o desenvolvimento de um modelo que realize a predição de alguns resultados, baseados em uma base histórica rica em informações, e que garantisse a performance dos resultados de planejamento com mais fidelidade do que os cálculos realizados atualmente.

4.2 Coleta de dados

Foi realizada extração dos resultados de faturamento e volume do ano de 2019 de algumas unidades, sorteadas aleatoriamente e com seus dados anonimizados. A extração de resultados foi realizada utilizando a ferramenta de ETL QlikView, realizando assim o primeiro tratamento das informações, isto é, separando apenas as colunas necessárias para realização dos cálculos:

- Data da competencia;

- Unidade;
- Município;
- Exame;
- Volume;
- Valor unitário.

	%IDData	%IDUnidade	%IDMunicipic	%IDExame	Qtde. Venda	Vlr. Vendas	Vlr. Unit. Vendas
	01/01/2019	51	3106705	11	1	32,9	32,9
	01/01/2019	54	3106200	13	1	7,46	7,46

Tabela 1 - Exemplo dos dados extraídos

4.3 Pré-Processamento

Utilizamos o Knime para desenvolver o workflow de machine learning, de modo a obter um modelo permita a predição dos valores de faturamento da empresa.

Dos dados extraídos foram utilizados apenas os dados do município 3144805. Por meio do excel um filtro foi aplicado de modo que permanecem apenas os dados da empresa supracitada, conforme tabela abaixo:

	A	B	C	D	E	F	G
1	%IDData	%IDUnid	%IDMunic	%IDExame	Qtde. Ven	Vlr. Vendas	Vlr. Unit. Ven
10	01/01/2019	54	3144805	13	1	20.48	20.48
21	01/01/2019	54	3144805	13	14	150.08	10.72
27	01/01/2019	54	3144805	15	1	13.87	13.87
62	01/01/2019	54	3144805	25	74	888	12.
65	01/01/2019	54	3144805	25	112	1008	9.
67	01/01/2019	54	3144805	25	124	3622.04	29.21

Tabela 2 - Exemplo dos dados extraídos município 3144805.

4.3.1 Correlação entre as variáveis

Foram identificadas a correlação entre as variáveis numéricas IDData, Qtde.Vendas, Vlr.Vendas e Vlr.Unit.vendas, conforme tabela abaixo.



Figura 3 - Workflow Correlação entre variáveis

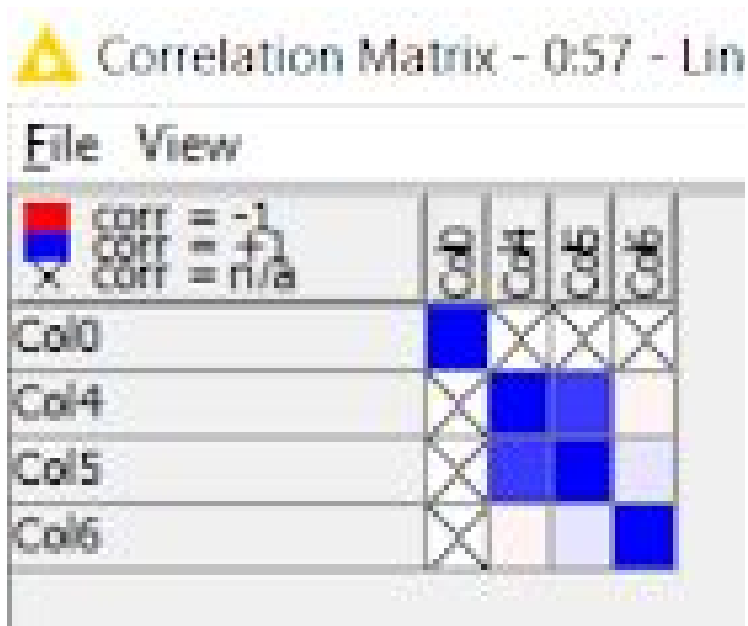


Figura 4 - Matrix de Correlação

4.3.2 Leitura dos dados

File Table - 0:7 - File Reader (Leitura dos dados)

File Hilite Navigation View

Table "3144805.csv" - Rows: 4897 Spec - Columns: 7 Properties Flow Variables

Row ID	S Col0	I Col1	I Col2	I Col3	I Col4	D Col5	D Col6
Row0	01/01/2019	54	3144805	13	1	20.48	20.48
Row1	01/01/2019	54	3144805	13	14	150.08	10.72
Row2	01/01/2019	54	3144805	15	1	13.87	13.87
Row3	01/01/2019	54	3144805	25	74	888	12
Row4	01/01/2019	54	3144805	25	112	1,008	9
Row5	01/01/2019	54	3144805	25	124	3,622.04	29.21
Row6	01/01/2019	54	3144805	46	1	477.15	477.15
Row7	01/01/2019	54	3144805	50	3	33.72	11.24
Row8	01/01/2019	54	3144805	55	1	8.41	8.41
Row9	01/01/2019	54	3144805	55	35	374.5	10.7
Row10	01/01/2019	54	3144805	56	1	8.24	8.24
Row11	01/01/2019	54	3144805	56	3	14.82	4.94
Row12	01/01/2019	54	3144805	58	5	175	35
Row13	01/01/2019	54	3144805	59	6	96	16
Row14	01/01/2019	54	3144805	61	1	15.14	15.14
Row15	01/01/2019	54	3144805	61	6	64.32	10.72
Row16	01/01/2019	54	3144805	62	1	17.21	17.21
Row17	01/01/2019	54	3144805	62	5	53.6	10.72
Row18	01/01/2019	54	3144805	64	1	10.07	10.07
Row19	01/01/2019	54	3144805	64	7	27.3	3.9
Row20	01/01/2019	54	3144805	68	2	23.36	11.68
Row21	01/01/2019	54	3144805	68	67	522.6	7.8
Row22	01/01/2019	54	3144805	72	7	183.54	26.22
Row23	01/01/2019	54	3144805	73	3	32.16	10.72
Row24	01/01/2019	54	3144805	74	2	21.44	10.72
Row25	01/01/2019	54	3144805	76	1	17.91	17.91
Row26	01/01/2019	54	3144805	93	2	307.88	153.94
Row27	01/01/2019	54	3144805	95	1	439.84	439.84

Figura 5 - Leitura dos dados

4.3.3 Tratar dados ausentes

Quando não houver valor substituir pelos valores fixos 0 e 0.0.

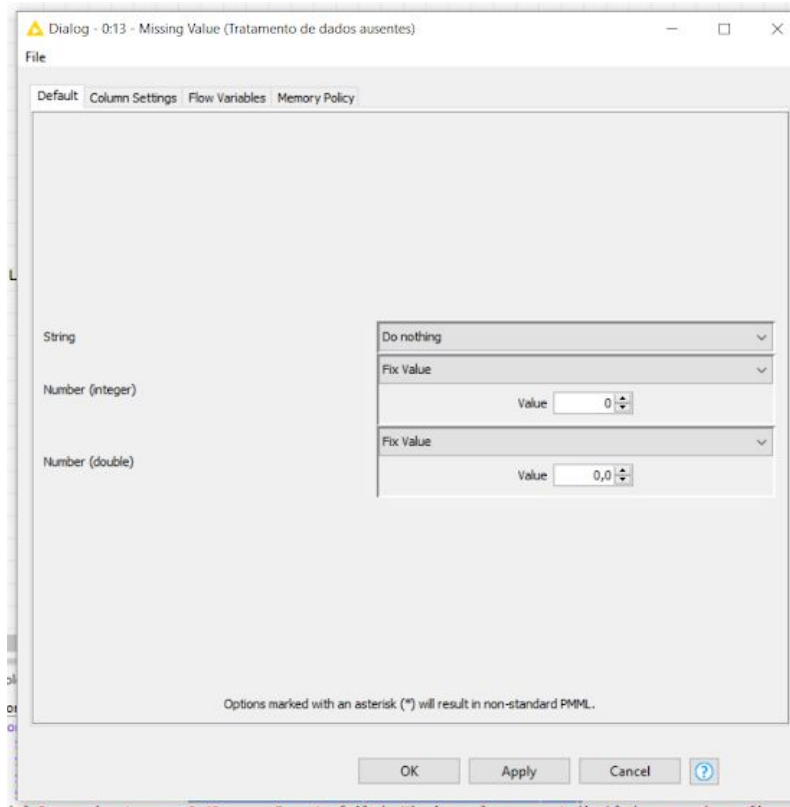


Figura 6 - Tratamento de Missing Value

4.3.4 Seleção de colunas

Foram selecionadas as colunas: iddata, qtde.vendas e vlr.vendas.

Filtered table - 0:8 - Column Filter (Seleccionar columnas)

File Hilite Navigation View

Table "default" - Rows: 4897 Spec - Columns: 3 Properties Flow Vi

Row ID	S Col0	I Col4	D Col5
Row0	01/01/2019	1	20.48
Row1	01/01/2019	14	150.08
Row2	01/01/2019	1	13.87
Row3	01/01/2019	74	888
Row4	01/01/2019	112	1,008
Row5	01/01/2019	124	3,622.04
Row6	01/01/2019	1	477.15
Row7	01/01/2019	3	33.72
Row8	01/01/2019	1	8.41
Row9	01/01/2019	35	374.5
Row10	01/01/2019	1	8.24
Row11	01/01/2019	3	14.82
Row12	01/01/2019	5	175
Row13	01/01/2019	6	96
Row14	01/01/2019	1	15.14
Row15	01/01/2019	6	64.32
Row16	01/01/2019	1	17.21
Row17	01/01/2019	5	53.6
Row18	01/01/2019	1	10.07
Row19	01/01/2019	7	27.3
Row20	01/01/2019	2	23.36
Row21	01/01/2019	67	522.6
Row22	01/01/2019	7	183.54
Row23	01/01/2019	3	32.16
Row24	01/01/2019	2	21.44
Row25	01/01/2019	1	17.91
Row26	01/01/2019	2	307.88
Row27	01/01/2019	1	439.84
Row28	01/01/2019	2	879.68
Row29	01/01/2019	47	266.96
Row30	01/01/2019	1	72.99
Row31	01/01/2019	1	471
Row32	01/01/2019	2	24.8
Row33	01/01/2019	54	176.04
Row34	01/01/2019	5	101.75
Row35	01/01/2019	4	24.4
Row36	01/01/2019	43	94.6
Row37	01/01/2019	14	312.76

Figura 7 - Filtrando columnas

5. Modelo De Regressão Linear

A regressão linear verifica a existencia de relacao entre duas variáveis, isto é dado x e y, quanto que x explica y.

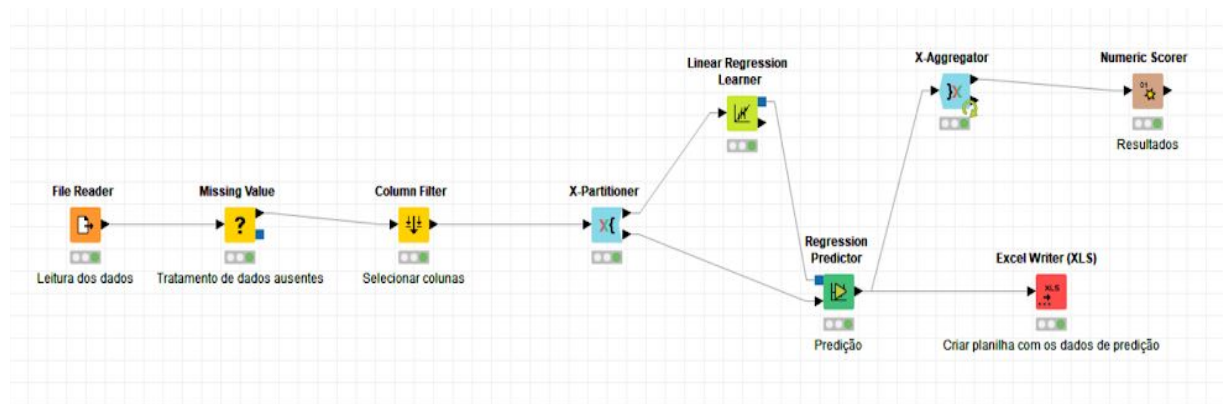


Figura 8 - Workflow de Regressão Linear

5.1 Resultados Da Análise De Regressao Linear

▲ Predicted data - 0:15 - Regression Predictor (Predição)

File Hilite Navigation View

Table "default" - Rows: 489 Spec - Columns: 4 Properties Flow Variables

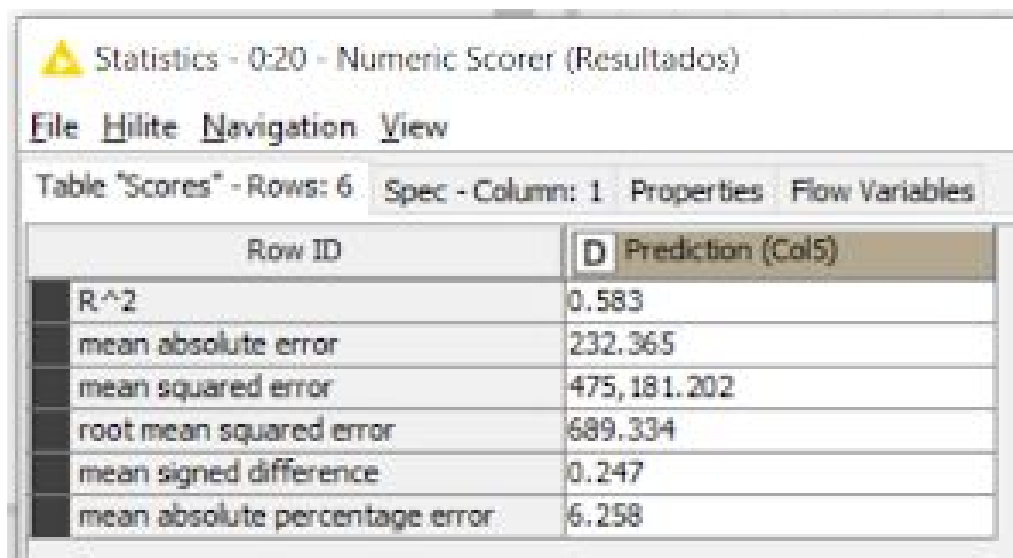
Row ID	S Col0	I Col4	D Col5	D Predict...
Row 11	01/01/2019	3	14.82	172.132
Row 15	01/01/2019	6	64.32	178.551
Row 21	01/01/2019	67	522.6	309.072
Row 61	01/01/2019	15	94.5	197.808
Row 71	01/01/2019	2	8.2	169.992
Row 81	01/01/2019	1	19.29	167.852
Row 84	01/01/2019	22	204.6	212.786
Row 107	01/01/2019	56	285.6	285.535
Row 128	01/01/2019	1	7.77	167.852
Row 134	01/01/2019	9	261	184.97
Row 138	01/01/2019	2	26	169.992
Row 144	01/01/2019	9	74.25	184.97
Row 150	01/01/2019	3	13.14	172.132
Row 158	01/01/2019	76	319.2	328.329
Row 161	01/01/2019	1010	3,939	2,326.799
Row 169	01/01/2019	1	33.68	167.852
Row 185	01/01/2019	9	77.22	184.97
Row 192	01/01/2019	356	1,068	927.442
Row 227	01/01/2019	83	614.2	343.307
Row 237	01/01/2019	548	1,644	1,338.263
Row 247	01/01/2019	691	4,215.1	1,644.238
Row 253	01/01/2019	6	25.8	178.551
Row 275	01/01/2019	22	350.02	212.786
Row 299	01/01/2019	40	93.2	251.3
Row 307	01/01/2019	5	31.65	176.411
Row 341	01/01/2019	1	6.62	167.852
Row 377	01/01/2019	20	229.4	208.506
Row 382	01/01/2019	3	20.55	172.132
Row 399	01/01/2019	19	77.9	206.367
Row 402	01/01/2019	8	133.36	182.83
Row 414	01/01/2019	1	9.35	167.852
Row 425	01/01/2019	1	8.56	167.852
Row 428	01/01/2019	18	124.2	204.227
Row 438	01/01/2019	1	5.03	167.852
Row 453	01/01/2019	1	40.89	167.852
Row 472	01/01/2019	1	38	167.852
Row 517	01/02/2019	2	24.8	157.968
Row 527	01/02/2019	1	12.15	155.829
Row 530	01/02/2019	1	45.83	155.829
Row 532	01/02/2019	7	66.64	168.667
Row 533	01/02/2019	5	40.5	164.387
Row 538	01/02/2019	2	94	157.968
Row 547	01/02/2019	2	107.2	157.968
Row 574	01/02/2019	2	30.96	157.968
Row 590	01/02/2019	2	15.5	157.968
Row 620	01/02/2019	9	123.3	172.946
Row 629	01/02/2019	3	54	160.108
Row 631	01/02/2019	2	60	157.968
Row 638	01/02/2019	1	9.9	155.829
Row 641	01/02/2019	1	9.8	155.829
Row 643	01/02/2019	1	30.17	155.829
Row 668	01/02/2019	1	9.9	155.829
Row 704	01/02/2019	4	39.6	162.248
Row 705	01/02/2019	121	375.1	412.592
Row 714	01/02/2019	63	407.61	288.49
Row 745	01/02/2019	12	126.12	179.365
Row 760	01/02/2019	1	32	155.829
Row 767	01/02/2019	1	28.43	155.829
Row 777	01/02/2019	1	23	155.829
Row 804	01/02/2019	2	12.66	157.968
Row 818	01/02/2019	1	29	155.829
Row 847	01/02/2019	3	318.45	160.108
Row 855	01/02/2019	1	10.85	155.829
Row 862	01/02/2019	2	9	157.968

Figura 9 - Resultado da Regressao Linear

5.2 Extração de Conhecimento

Na análise de Regressão Linear um coeficiente de determinação obtido foi de R^2 0,583.

O coeficiente de determinação, também chamado de R^2 , é uma medida de ajuste de um modelo estatístico linear generalizado que varia entre 0 e 1. Quanto maior o R^2 mais explicativo é o modelo linear, ou seja, melhor ele se ajusta à amostra.



The screenshot shows a software window titled "Statistics - 0:20 - Numeric Scorer (Resultados)". It has a menu bar with "File", "Hilite", "Navigation", and "View". Below the menu bar, there are tabs: "Table 'Scores' - Rows: 6", "Spec - Column: 1", "Properties", and "Flow Variables". The "Table 'Scores' - Rows: 6" tab is active, displaying a table with two columns: "Row ID" and "Prediction (Col5)". The table contains six rows of statistical data.

Row ID	Prediction (Col5)
R^2	0.583
mean absolute error	232.365
mean squared error	475,181.202
root mean squared error	689.334
mean signed difference	0.247
mean absolute percentage error	6.258

Figura 10 - Resultados Estatísticos do modelo de Regressão Linear

6. Modelo De Regressao Polinomial

A regressão polinomial pode ser utilizado para problemas de relações não lineares.

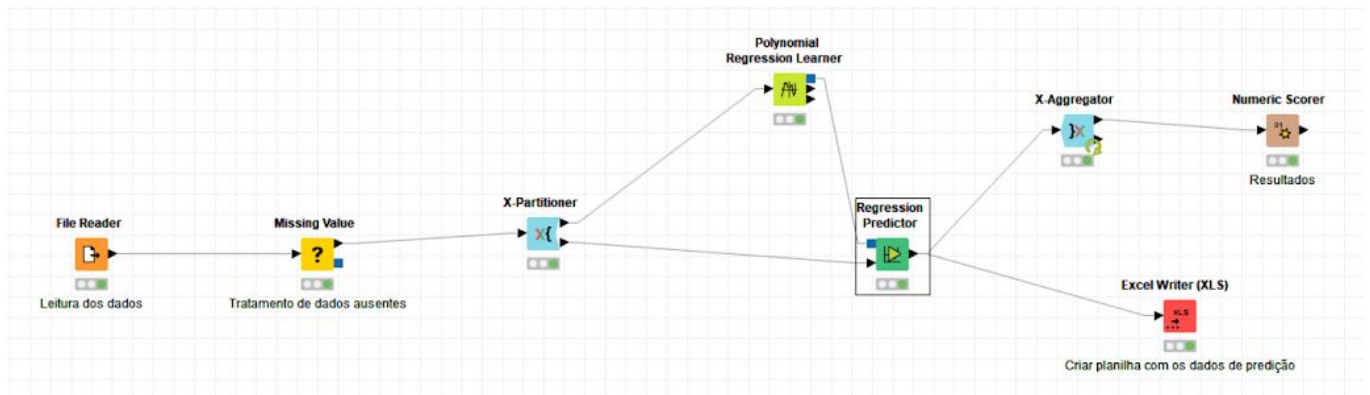


Figura 11- Workflow Regressao Polinomial

Foram selecionados as colunas apenas na configuração da Regressão Polinomial: Idunidade, Idexame, Qtde.vendas.

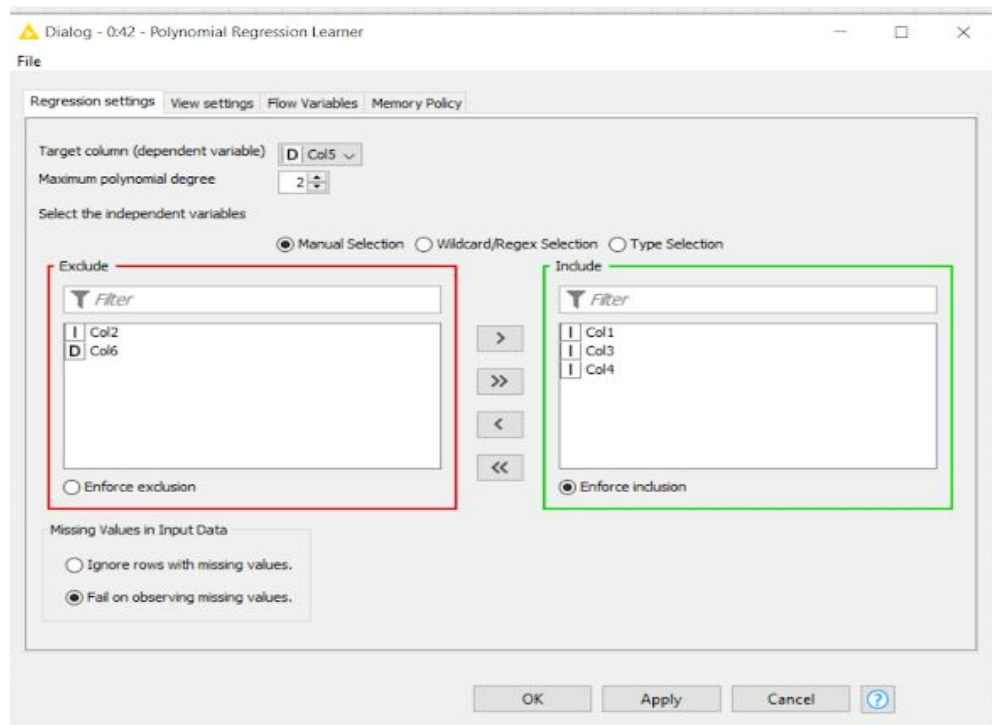


Figura 12 - Configuração do modelo de Regressão Polinomial

6.1 Resultados Da Análise De Regressao Polinomial

▲ Predicted data - 0.43 - Regression Predictor

File Hilite Navigation View

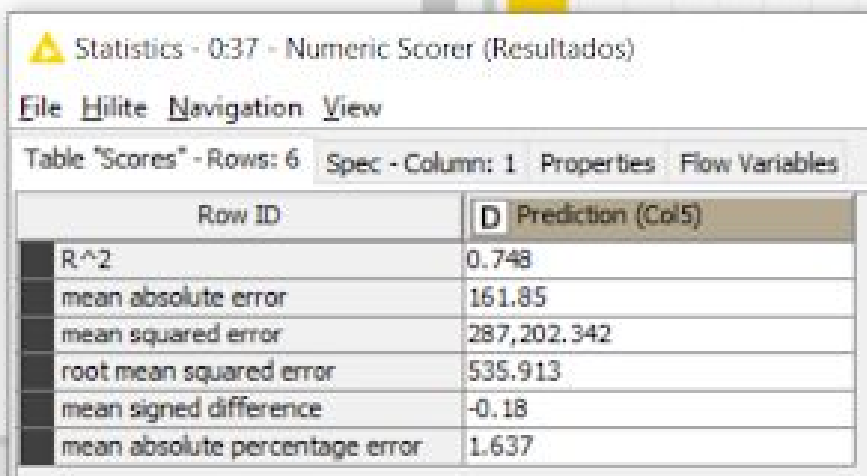
Table "default" - Rows: 489 Spec - Columns: 8 Properties Flow Variables

Row ID	S Col0	I Col1	I Col2	I Col3	I Col4	D Col5	D Col6	D Predict...
Row1546	01/04/2019	54	3144805	633	111	954.6	8.6	532.99
Row1547	01/04/2019	54	3144805	640	1	134	134	61.166
Row1560	01/04/2019	54	3144805	751	9	261	29	98.98
Row1575	01/04/2019	54	3144805	848	2	9.6	4.8	71.215
Row1579	01/04/2019	54	3144805	897	97	397.7	4.1	480.864
Row1588	01/04/2019	54	3144805	997	1	30.17	30.17	70.92
Row1592	01/04/2019	54	3144805	1000	3	18.9	6.3	79.699
Row1594	01/04/2019	54	3144805	1001	3	15.78	5.26	79.727
Row1602	01/04/2019	54	3144805	1033	5	168.4	33.68	89.288
Row1619	01/04/2019	54	3144805	1185	3	29.7	9.9	84.701
Row1625	01/04/2019	54	3144805	1212	124	622.48	5.02	603.671
Row1641	01/04/2019	54	3144805	1285	225	472.5	2.1	1,026.18
Row1654	01/04/2019	54	3144805	1404	72	462.96	6.43	388.028
Row1657	01/04/2019	54	3144805	1413	1	66.69	66.69	82.117
Row1682	01/04/2019	54	3144805	1486	2	15	7.5	88.412
Row1693	01/04/2019	54	3144805	1523	289	809.2	2.8	1,293.354
Row1725	01/04/2019	54	3144805	1666	2	151.2	75.6	93.187
Row1729	01/04/2019	54	3144805	1701	17	101.32	5.96	159.226
Row1733	01/04/2019	54	3144805	1718	4	133.36	33.34	103.256
Row1734	01/04/2019	54	3144805	1724	1	28.67	28.67	90.369
Row1743	01/04/2019	54	3144805	1882	1	7.36	7.36	94.522
Row1747	01/04/2019	54	3144805	1942	2	130.76	65.38	100.442
Row1754	01/04/2019	54	3144805	2016	7	66.64	9.52	124.105
Row1774	01/04/2019	53	3144805	2294	1	6.62	6.62	128.611
Row1778	01/04/2019	54	3144805	2315	1	3.42	3.42	105.77
Row1793	01/04/2019	54	3144805	2521	8	251.2	31.4	141.476
Row1818	01/04/2019	56	3144805	2782	1	9	9	68.566
Row1828	01/04/2019	56	3144805	2882	3	27	9	79.784
Row1836	01/04/2019	54	3144805	2957	1	24.58	24.58	122.084
Row1848	01/04/2019	54	3144805	3025	1	4.67	4.67	123.786
Row1862	01/04/2019	54	3144805	3144	40	66	1.65	295.592
Row1869	01/04/2019	54	3144805	3193	1	5.03	5.03	127.972
Row1872	01/04/2019	54	3144805	3201	2	3.3	1.65	132.52
Row1880	01/04/2019	54	3144805	3270	80	248	3.1	470.188
Row1883	01/04/2019	54	3144805	3801	1	57.16	57.16	142.872
Row1889	01/04/2019	54	3144805	3894	3	18	6	153.815
Row1913	01/04/2019	54	3144805	35614	1	15.7	15.7	380.409
Row1924	01/05/2019	54	3144805	59	5	80	16	62.398
Row1940	01/05/2019	54	3144805	94	1	330	330	45.989
Row1941	01/05/2019	54	3144805	96	1	439.84	439.84	46.045
Row1948	01/05/2019	54	3144805	128	1	33.92	33.92	46.944
Row1952	01/05/2019	54	3144805	137	47	103.4	2.2	246.165
Row1963	01/05/2019	54	3144805	214	5	47.6	9.52	66.744
Row1966	01/05/2019	54	3144805	270	6	47.94	7.99	72.653
Row1978	01/05/2019	54	3144805	420	1	5.26	5.26	55.089
Row1984	01/05/2019	54	3144805	460	295	619.5	2.1	1,288.731
Row1988	01/05/2019	54	3144805	463	1	19.29	19.29	56.28
Row2003	01/05/2019	54	3144805	517	1	10.72	10.72	57.774
Row2015	01/05/2019	54	3144805	551	10	44.3	4.43	97.821
Row2017	01/05/2019	54	3144805	570	1	24.12	24.12	59.238
Row2026	01/05/2019	54	3144805	633	2	12.86	6.43	65.323
Row2037	01/05/2019	54	3144805	683	1	11.89	11.89	62.348
Row2040	01/05/2019	54	3144805	685	1	8.36	8.36	62.403
Row2058	01/05/2019	54	3144805	848	1	4.8	4.8	66.865
Row2062	01/05/2019	54	3144805	897	131	537.1	4.1	624.643
Row2073	01/05/2019	54	3144805	998	4	23.36	5.84	83.993
Row2079	01/05/2019	54	3144805	1001	2	12.4	6.2	75.378
Row2087	01/05/2019	54	3144805	1033	1	33.68	33.68	71.896
Row2091	01/05/2019	54	3144805	1083	4	39.6	9.9	86.295
Row2092	01/05/2019	54	3144805	1084	1	9.9	9.9	73.277
Row2101	01/05/2019	54	3144805	1185	5	49.5	9.9	93.394
Row2103	01/05/2019	54	3144805	1203	15	29.85	1.99	137.283
Row2109	01/05/2019	54	3144805	1218	1	8.74	8.74	76.891
Row2110	01/05/2019	54	3144805	1235	1	59	59	77.348

Figura 13 - Resultados do modelo de Regressão Polinomial

6.2 Extração De Conhecimento

Na análise de Regressão Polinomial coeficiente de determinação obtido foi de R^2 0,749.



The screenshot shows a software window titled "Statistics - 0:37 - Numeric Scorer (Resultados)". It has a menu bar with "File", "Hilite", "Navigation", and "View". Below the menu bar, there are tabs: "Table 'Scores' - Rows: 6", "Spec - Column: 1", "Properties", and "Flow Variables". The "Table 'Scores' - Rows: 6" tab is active, displaying a table with two columns: "Row ID" and "Prediction (Col5)". The table contains six rows of data.

Row ID	Prediction (Col5)
R^2	0.748
mean absolute error	161.85
mean squared error	287,202.342
root mean squared error	535.913
mean signed difference	-0.18
mean absolute percentage error	1.637

Figura 14 - Resultados Estatísticos modelo de Regressão Linear

7. Experimentos e Conclusões

Algoritmo	Valor de R^2
Regressão Linear	0.583
Regressão Polinomial	0.748

Tabela 3 - Comparação de resultados

Conforme apresentado na tabela de resultados o algoritmo que apresentou melhor precisão na predição de faturamento foi a Regressão Polinomial, no entanto, outros modelos e variáveis precisam ser testadas, bem como dados dos anos anteriores para que a previsão seja mais assertiva.

REFERÊNCIAS

Bassani, Hansenclever F. O impacto da aprendizagem profunda na sociedade e academia. Revista da Sociedade Brasileira de Computação, Rio Grande do Sul, ano 2019, Ed.1. Disponível em: http://www.sbc.org.br/images/flippingbook/computacaobrasil/computa_39/pdf/CompBrasil_39_180.pdf. Acesso em: 10 fev .2020.

Deep Learning Book. Disponível em: <http://datascienceacademy.com.br/blog/17-casos-de-uso-de-machine-learning/>. Acesso em: 11 fev. 2020.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2015). Big Data, bigger dilemmas: a critical review. Journal of the Association for Information Science and Technology, paginas 1523-1545.

inference, and prediction. New York: Springer; 2008. Disponível em [The Elements of Statistical Learning](#)

Russell, Stuart J. (Stuart Jonathan), 1962 Inteligência artificial / Stuart Russell, Peter Norvig; tradução Regina Célia Simille. – Rio de Janeiro: Elsevier, 2013.

17 CASOS DE USO DE MACHINE LEARNING. Disponível em: <http://datascienceacademy.com.br/blog/17-casos-de-uso-de-machine-learning/>. Acesso em: 10 fev. 2020