Análise de Dados com Soluções Inteligentes

Discente: Glauber Borges Lindolfo
Inteligência Computacional
14 de Agosto de 2023

Visão Geral do Problema

Trata-se de uma análise do mercado de vídeo games.

Revisão dos dados adquiridos.

Pré Processamento dos Dados.

Análise Manual dos Dados.

Treino com os modelos.

Testes desses modelos.

Repetição com situações diferentes.

Conclusão

Dados Adquiridos

Dados de vendas de jogos ao redor do mundo.

Separação por vendas locais, e vendas globais.

Datas desde 1980 a 2020.

Dados de gênero, editora e plataforma estão inclusos.

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37

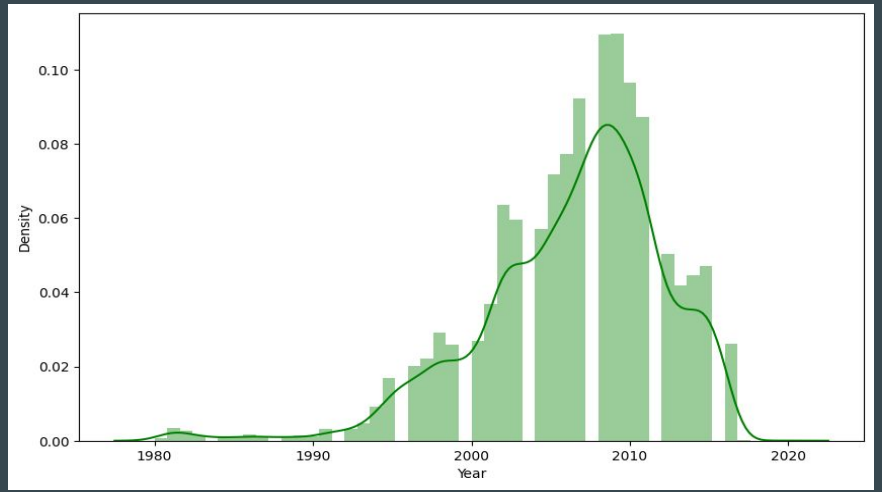
Análise de Tendências de Lançamento e Vendas

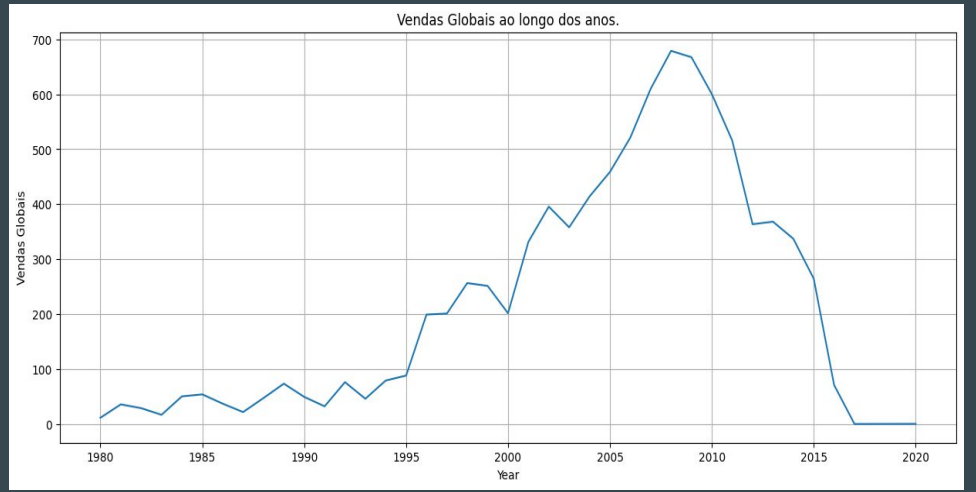
Número de Lançamento por Ano:

- Pico de Lançamento em 2009
- Editora com maior número de lançamentos: Electronic Arts.

Número de Vendas por Ano:

- Pico de Vendas em 2008
- Plataforma de maiores vendas: Playstation 2





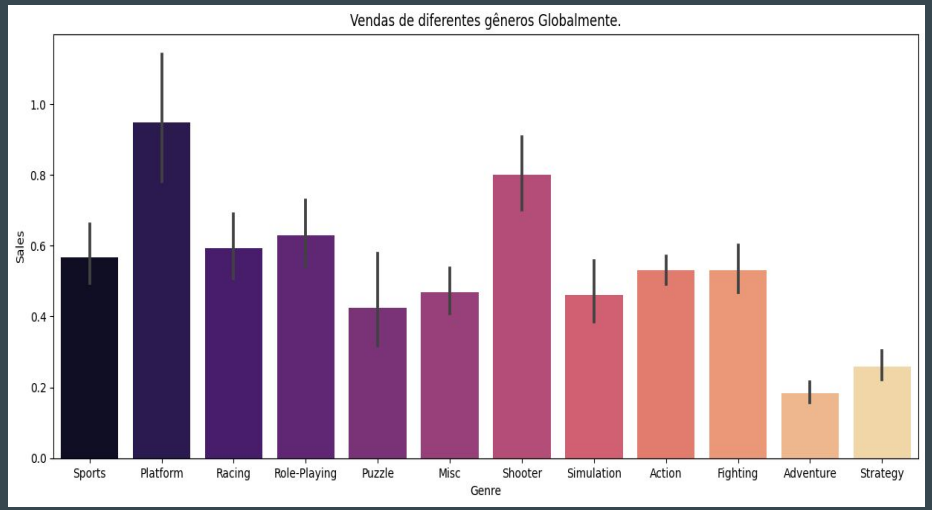
Análise Vendas por Gênero

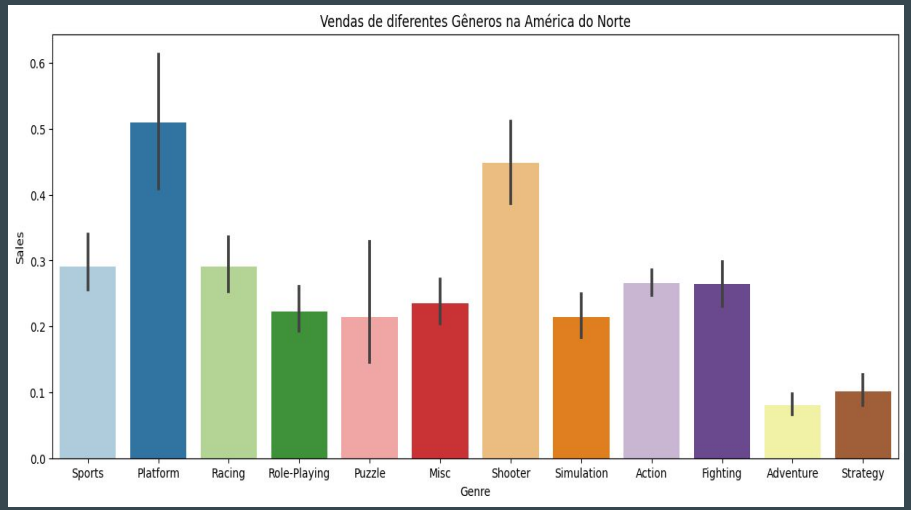
Número de Vendas por Gênero Globais

- Jogos de Plataforma
- Tendência de Shooters preencherem o mercado

Número de Vendas por Gênero Local:

- Semelhança com a tendência Global
- Vendas na América do Norte são a maior parte do mercado.

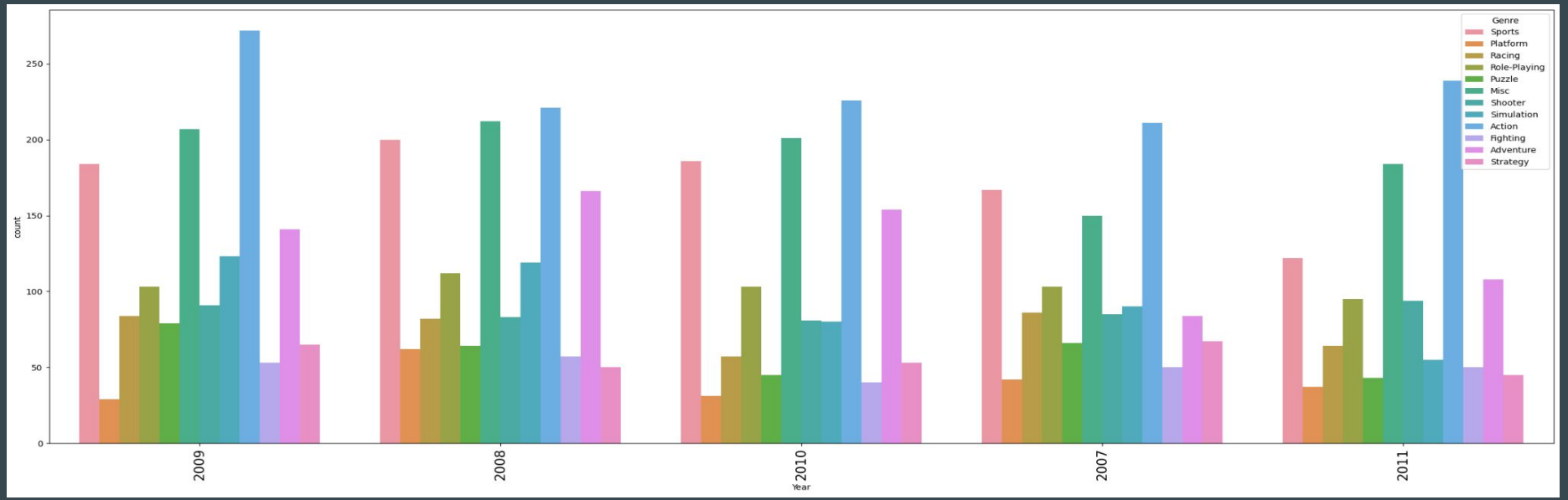




Análise Vendas por Gênero

Performance dos Gêneros ao longo dos 5 anos de maiores vendas

- De 2007 a 2011
- Tendências de alteração



Metodologia Utilizada

Modelos de Aprendizagem

Análise de Erros

Decision Tree

Uma Árvore de Decisão é um modelo de algoritmo de aprendizado de máquina usado para resolver problemas de classificação e regressão.

Etapas:

- Divisões
- Criação dos Ramos
- Critério de Parada
- Folhas e Previsões

Linear Regression

É uma técnica que visa encontrar a relação linear entre uma variável dependente e uma ou mais variáveis independentes.

O modelo de Regressão Linear assume que a relação entre as variáveis pode ser aproximada por uma linha reta.

Ridge

Ele é uma variação do modelo de regressão linear padrão que inclui um termo de regularização para controlar o tamanho dos coeficientes das características.

A regressão Ridge segue a mesma base da regressão linear, com a adição de um termo de regularização na função de perda.

Random Forest

Esses modelos são baseados em conjuntos de árvores de decisão e são conhecidos por sua capacidade de lidar com uma variedade de dados e problemas complexos.

O Random Forest Regressor é especificamente usado para problemas de regressão, ou seja, quando a tarefa é prever um valor numérico (contínuo) em vez de uma classe.

KNN

É um dos métodos mais simples e intuitivos de aprendizado de máquina, pois se baseia na ideia de que amostras semelhantes estão próximas umas das outras no espaço de características.

No caso da regressão, em vez de votos, os valores dos K vizinhos são usados para calcular uma média (ou outra medida de tendência central) que é atribuída ao novo ponto.

Coeficiente de Determinação

O Coeficiente de Determinação, frequentemente denotado como R ao quadrado , é uma métrica estatística utilizada para avaliar a qualidade de ajuste de um modelo de regressão em relação aos dados observados.

Ele fornece uma medida da proporção da variabilidade total da variável dependente que é explicada pelo modelo

Erro Médio Absoluto

É uma métrica comum de avaliação utilizada para medir o quão bem um modelo de regressão está prevendo os valores em relação aos valores reais.

Essa métrica fornece uma idéia direta da magnitude média dos erros de previsão em termos absolutos, sem levar em consideração sua direção.

Pré Processamento dos Dados

- Remoção de Duplicatas.
- Retirada de dados nulos.
- Remoção dos Outliers
- Mudança de Escala
- Label Encoding

(Feito com a função RobustScaler)

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	2	Super Mario Bros.	11	1985	4	359	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	26	2008	6	359	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	26	2009	10	359	15.75	11.01	3.28	2.96	33
4	5	Pokemon Red/Pokemon Blue	5	1996	7	359	11.27	8.89	10.22	1	31.37
5	6	Tetris	5	1989	5	359	23.2	2.26	4.22	0.58	30.26

Análise - Predição Global

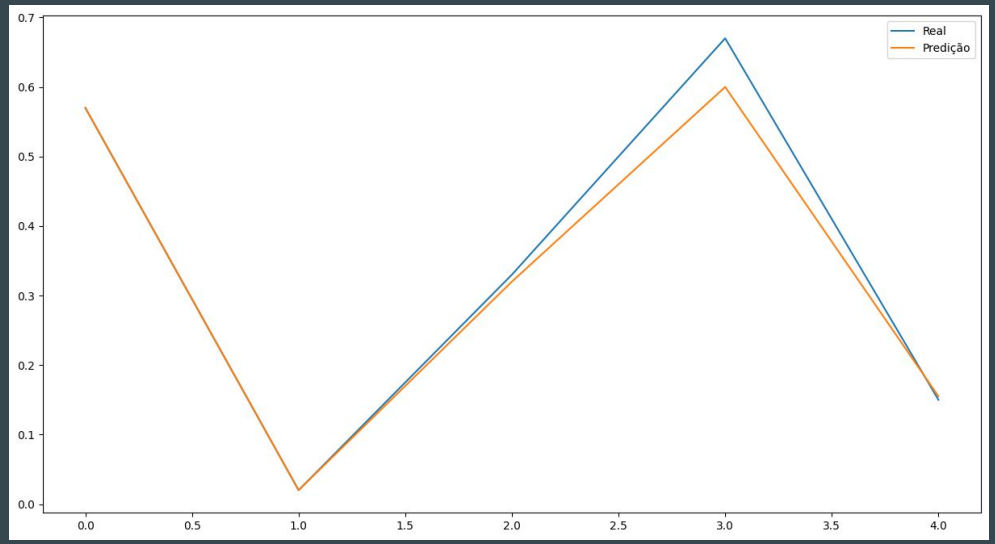
- Utilização dos dados de vendas Locais: NA, EU, JP e Others
- Cálculos lineares: Soma das vendas locais.
- Não utilização do Ano nem da Editora.
- Utilização de regressão linear.
- Identificação da Participação de mercado.

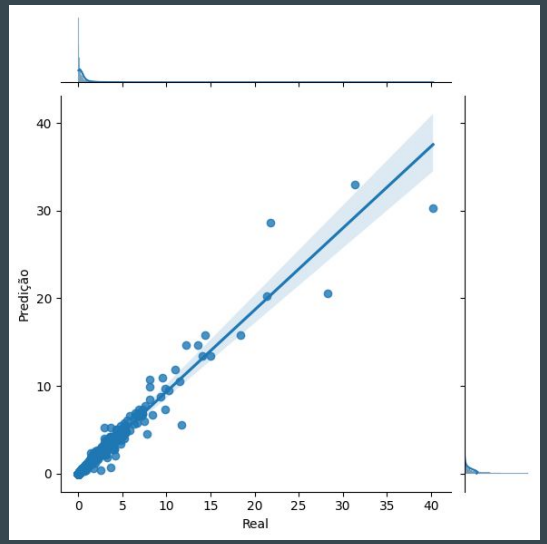
	Rank	Name	Platform	Genre	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	2	Super Mario Bros.	11	4	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	26	6	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	26	10	15.75	11.01	3.28	2.96	33
4	5	Pokemon Red/Pokemon Blue	5	7	11.27	8.89	10.22	1	31.37
5	6	Tetris	5	5	23.2	2.26	4.22	0.58	30.26

Resultados Predição Global

Decision Tree

Erro Médio Absoluto 0.04714600949645273
Coeficiente de Determinação 0.9600132190481512

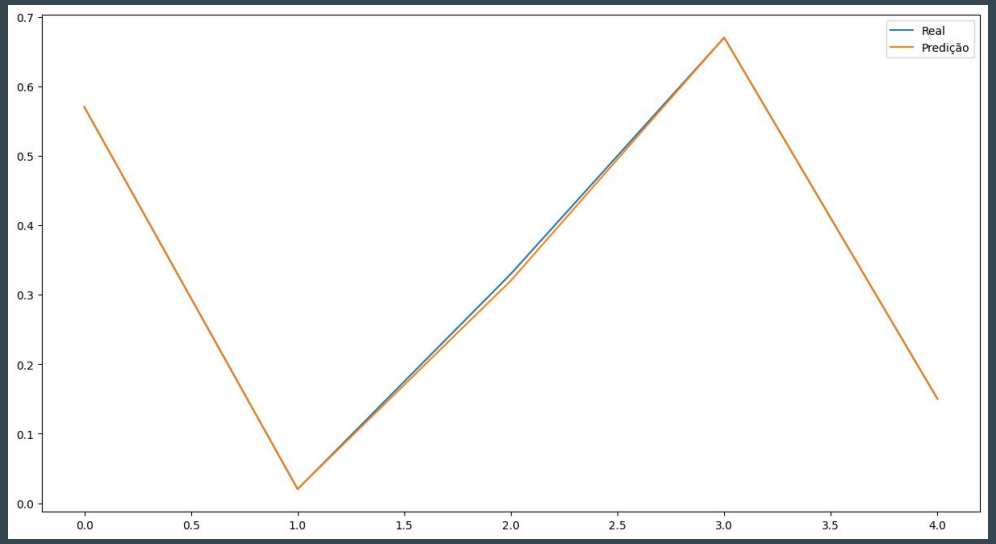


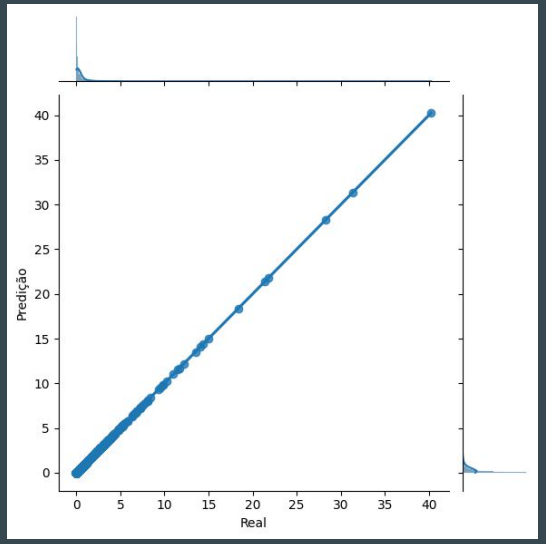


Resultados Predição Global

Ridge

Erro Médio Absoluto 0.002955137510258357
Coeficiente de Determinação 0.999987986050178

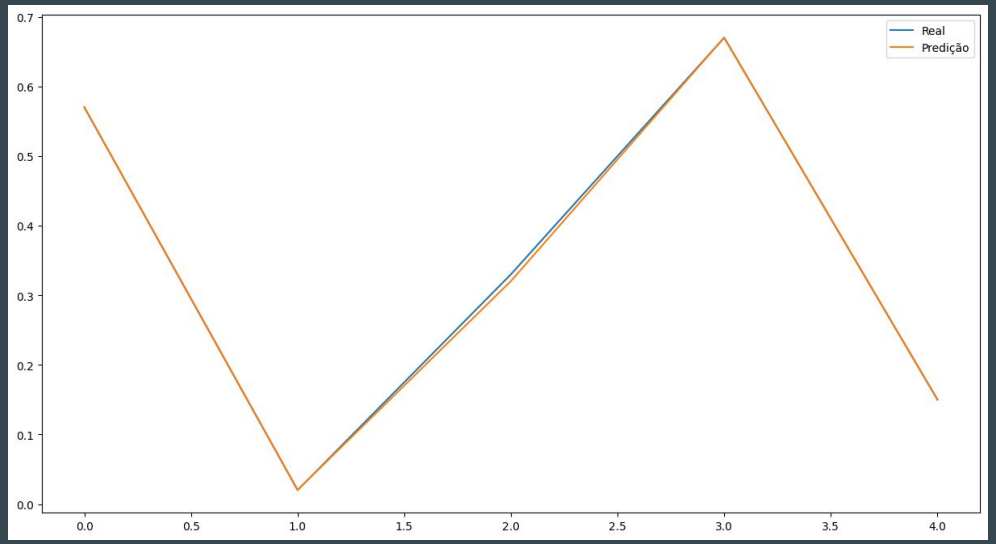


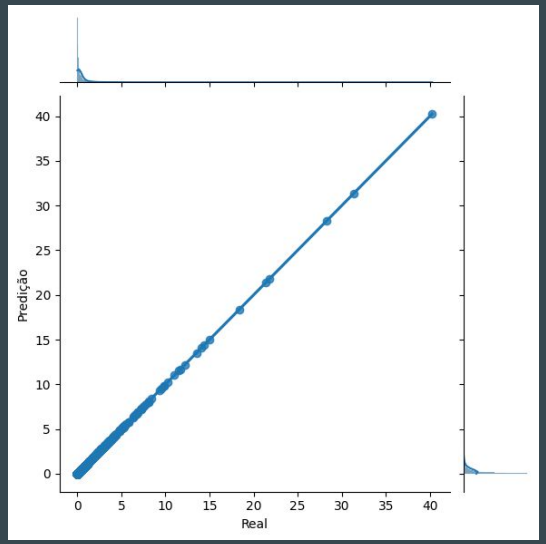


Resultados Predição Global

Linear Regression

Erro Médio Absoluto 0.0029547715225954007
Coeficiente de Determinação 0.9999879850329042

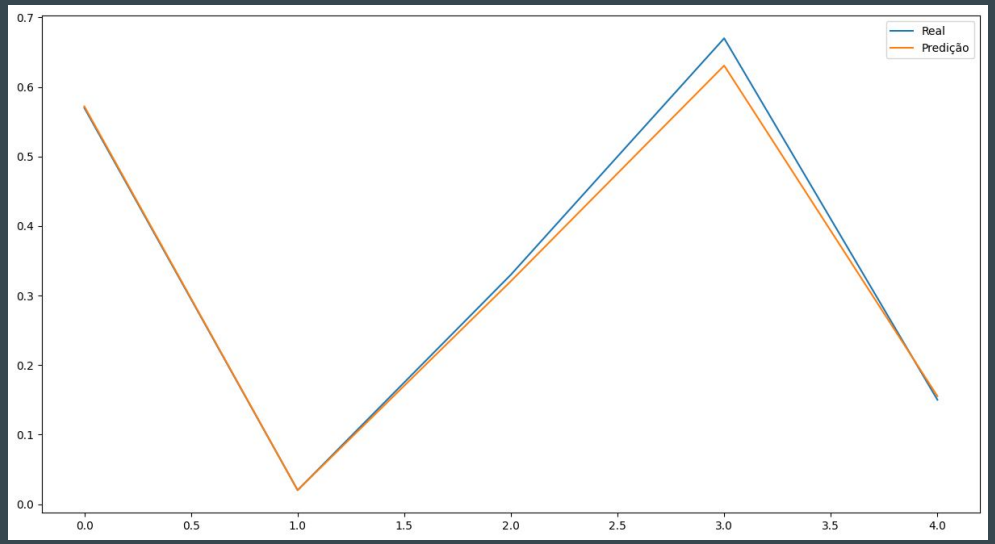


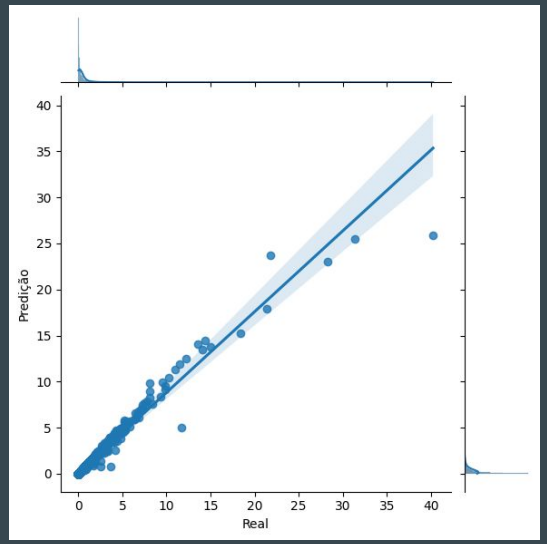


Resultados Predição Global

Random Forest Regressor

Erro Médio Absoluto 0.030947812566712654
Coeficiente de Determinação 0.9591450012506179

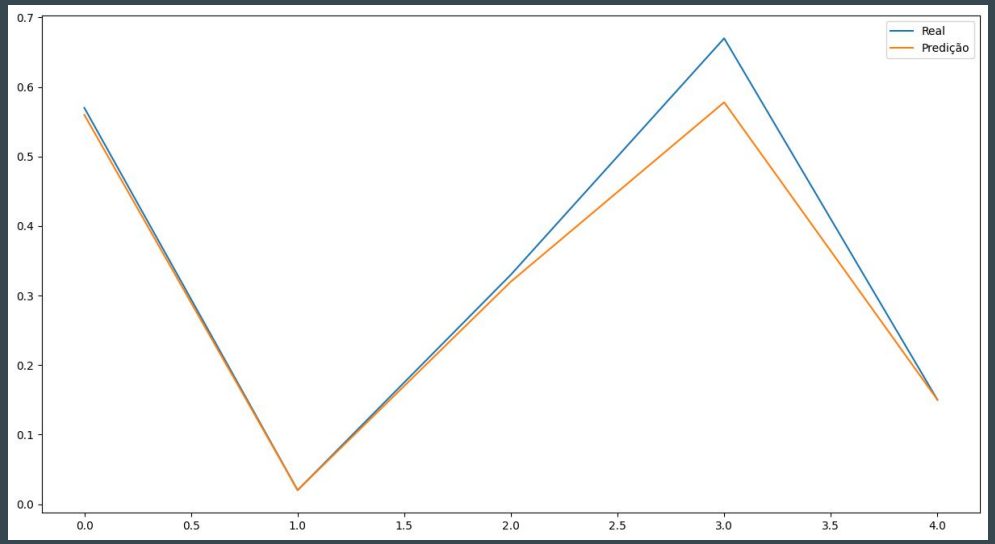


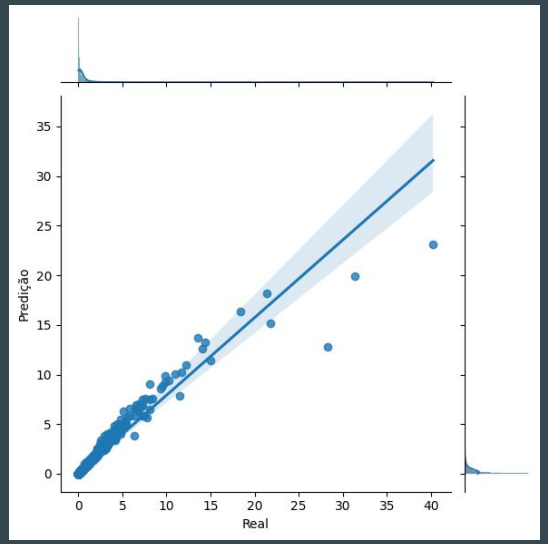


Resultados Predição Global

KNN

Erro Médio Absoluto 0.046915295850724285
Coeficiente de Determinação 0.9134505144593599





Conclusão

Melhor Modelo

Os melhores modelos foram:

1. Ridge
2. Regressão Linear

Implicações:

- Os dados iniciais obedecem uma regressão linear
- O modelo numérico

Coeficientes

Coeficientes do modelo numérico calculado

Plataform: -1.08639110e-04
Genre: -4.71152381e-05
NA_Sales: 2.39991637e-01
EU_Sales: 1.10008370e-01
JP_Sales: 3.99908258e-02
Other_Sales: 3.99685165e-02

Nova Análise - Predição Global com um Local

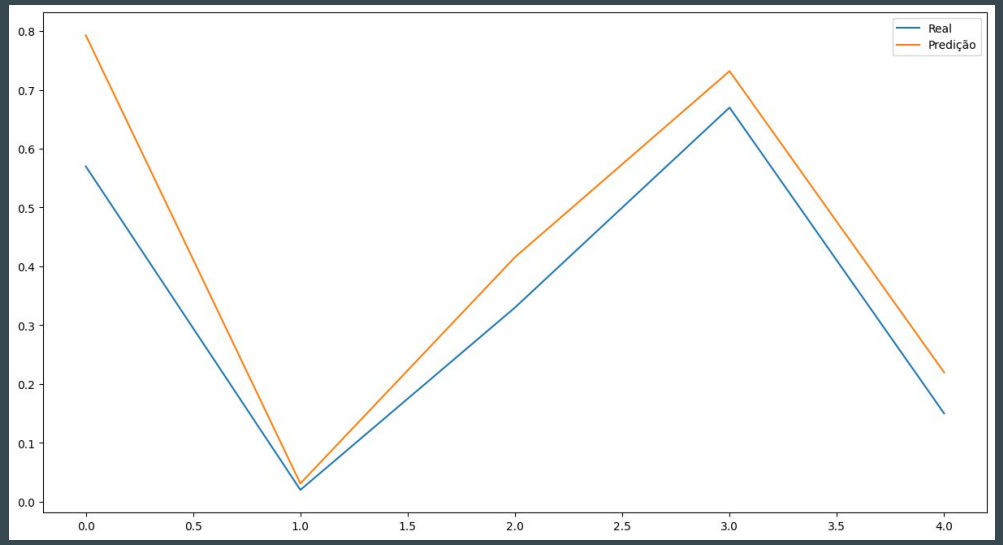
- Utilização dos dados de vendas Locais: NA, EU, JP e Others
- Predição dos Valores de venda global.
- Utilização de valores da editora, ano, plataforma e gênero.

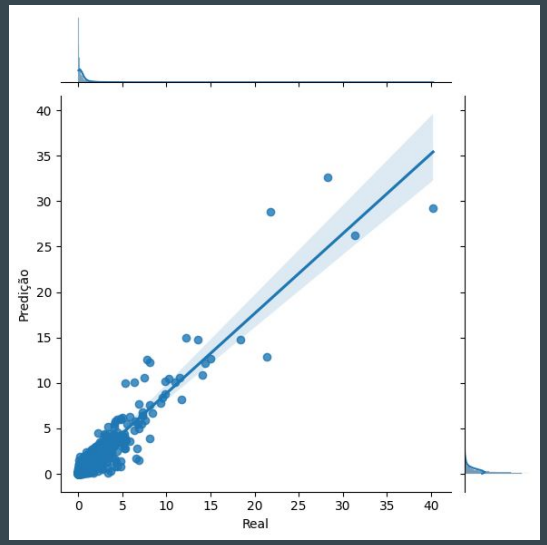
	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	Global_Sales
1	2	Super Mario Bros.	11	1985	4	359	29.08	40.24
2	3	Mario Kart Wii	26	2008	6	359	15.85	35.82
3	4	Wii Sports Resort	26	2009	10	359	15.75	33
4	5	Pokemon Red/Pokemon Blue	5	1996	7	359	11.27	31.37
5	6	Tetris	5	1989	5	359	23.2	30.26

Resultados NA - GLOBAL

Random Forest Regressor

Erro Médio Absoluto 0.17662931719771582
Coeficiente de Determinação 0.8956546552739884

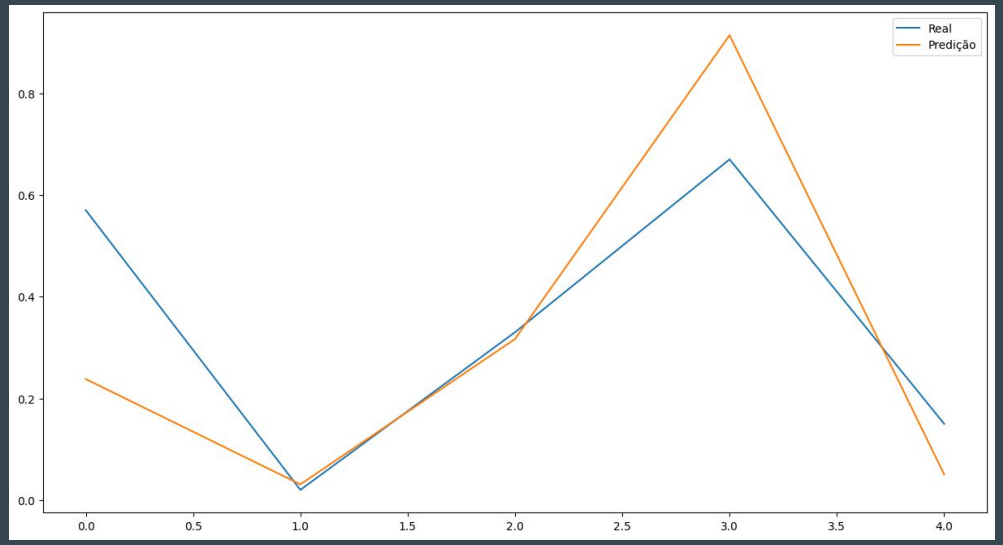


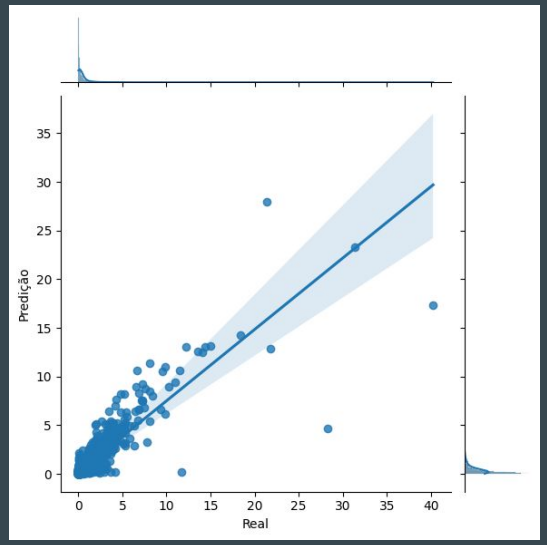


Resultados EU - GLOBAL

Random Forest Regressor

Erro Médio Absoluto 0.2160596997440449
Coeficiente de Determinação 0.7742814949430679

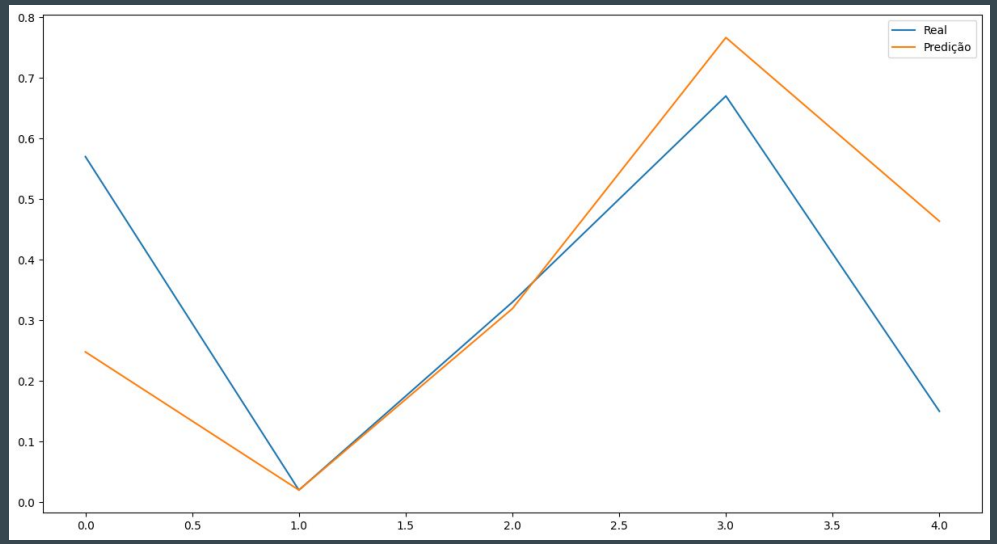


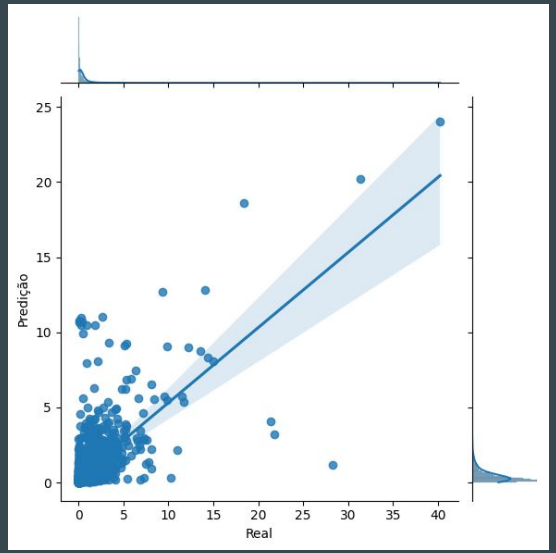


Resultados JP - GLOBAL

Random Forest Regressor

Erro Médio Absoluto 0.4953125911413111
Coeficiente de Determinação 0.4186670486736699

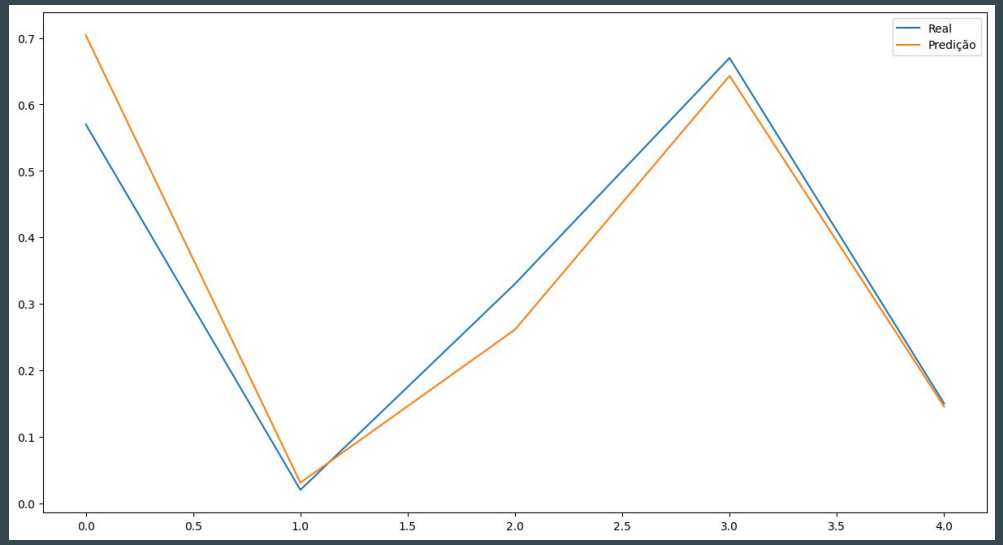


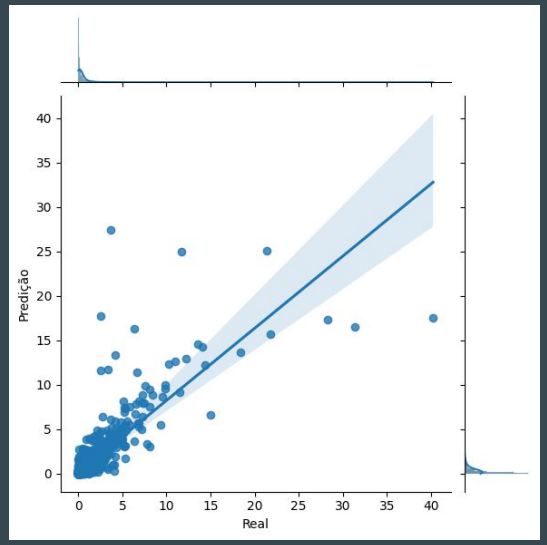


Resultados OU - GLOBAL

Random Forest Regressor

Erro Médio Absoluto 0.19301605685919718
Coeficiente de Determinação 0.6893291620452453





Conclusão

Melhor Modelo

Os melhores modelos foram:

1. América do Norte
2. Europa

Implicações:

- A participação de mercado é a mais influente.
- O Japão é o mercado mais afastado das tendências globais.
- O modelo não linear Random Forest foi o superior.

Coeficientes

Coeficientes do modelo numérico calculado para América do Norte

Platform: 0.02199668
Year: 0.02315024
Genre: 0.01497826
Publisher: 0.01729276
NA_Sales: 0.92258205

Nova Análise - Predição de Vendas Locais e Global

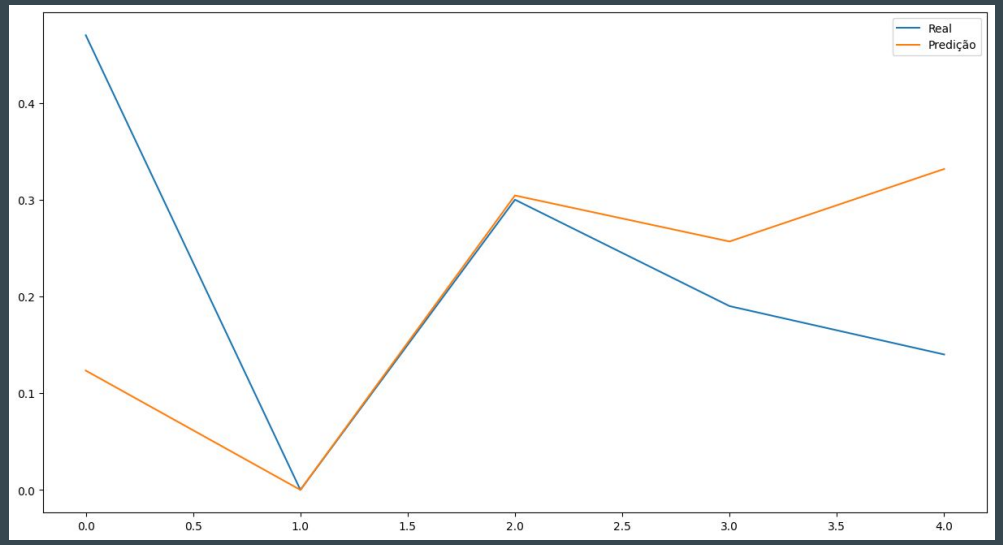
- Não utilização dos dados de vendas.
- Predição dos Valores de venda locais e global
- Utilização de valores da editora, ano, plataforma e gênero.
- Tentativa de identificação de características de mercado.

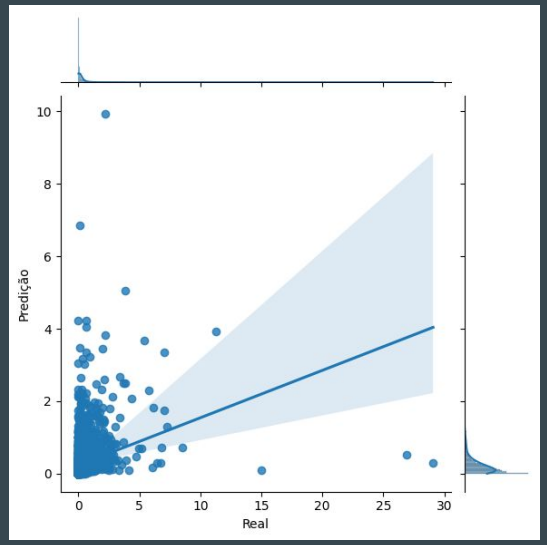
	Rank	Name	Platform	Year	Genre	Publisher	Global_Sales
0	1	Wii Sports	Wii	2006	Sports	Nintendo	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	33
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	31.37

Resultados DADOS - NA

Random Forest Regressor

Erro Médio Absoluto 0.26513793365877253
Coeficiente de Determinação 0.028587976364768952

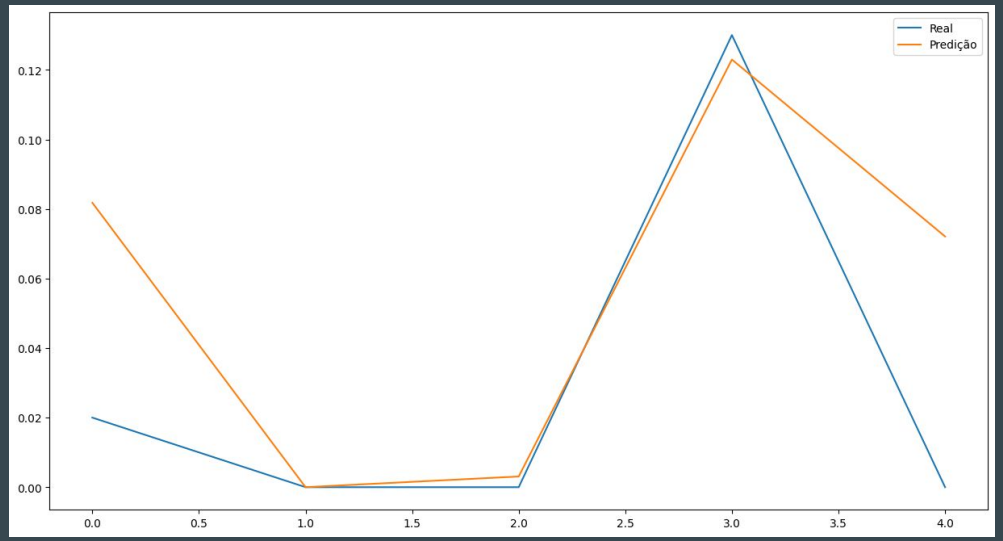


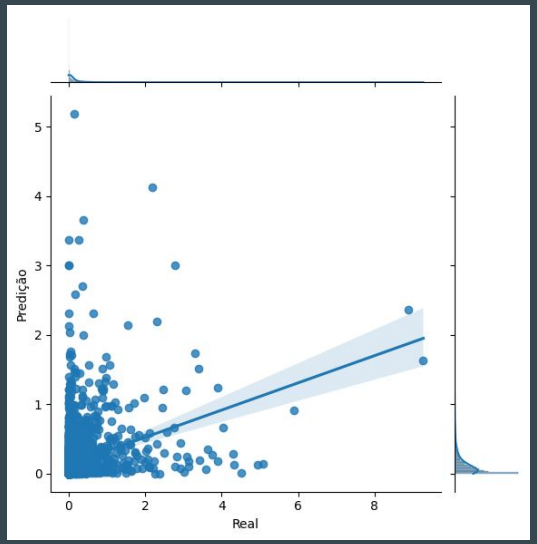


Resultados DADOS - EU

Random Forest Regressor

Erro Médio Absoluto 0.17077723744251078
Coeficiente de Determinação 0.03150984702923498

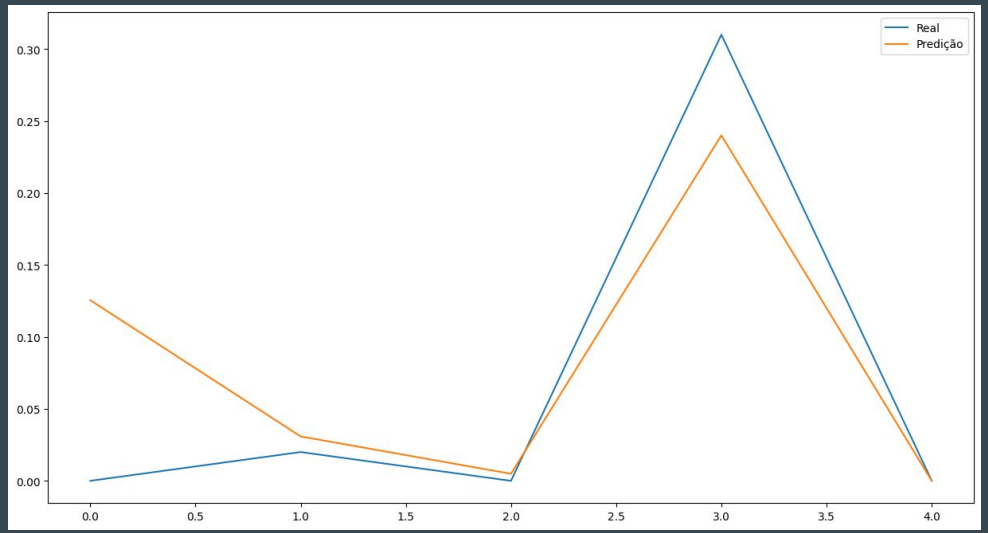


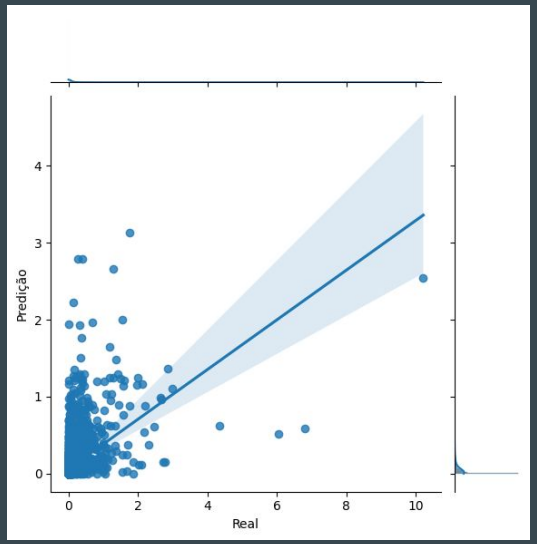


Resultados DADOS - JP

Random Forest Regressor

Erro Médio Absoluto 0.08388433793880226
Coeficiente de Determinação 0.21302576872646417

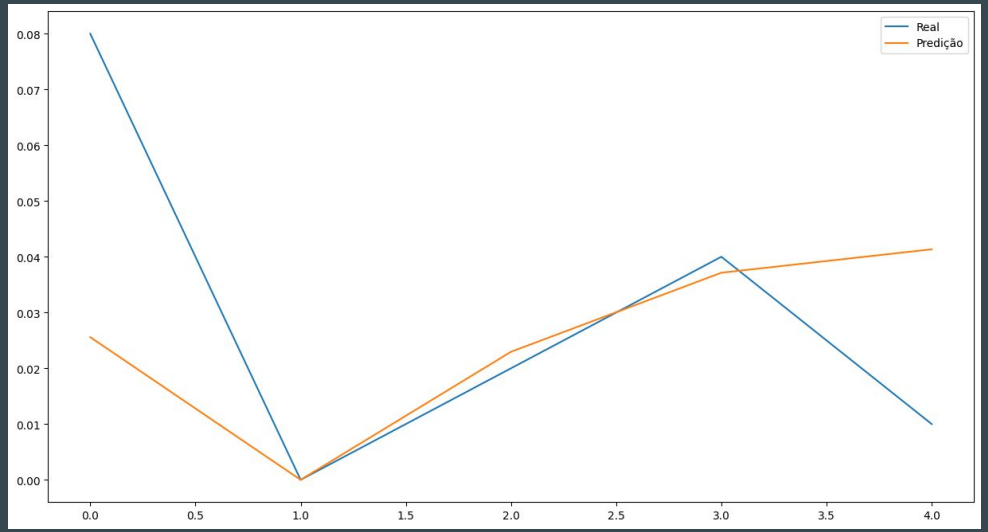


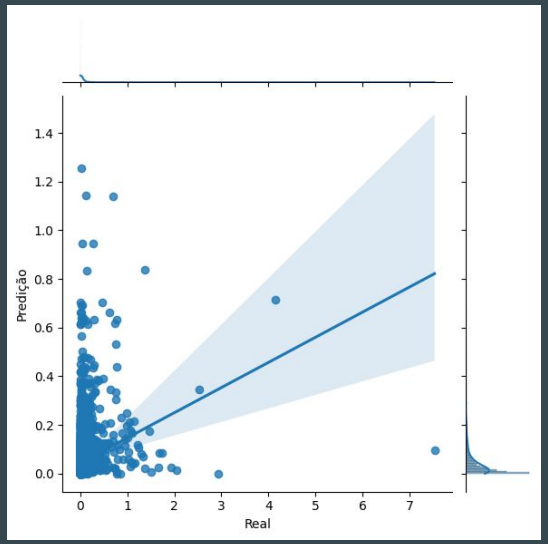


Resultados DADOS - OU

Random Forest Regressor

Erro Médio Absoluto 0.056234044563234865
Coeficiente de Determinação 0.0105511564524169999

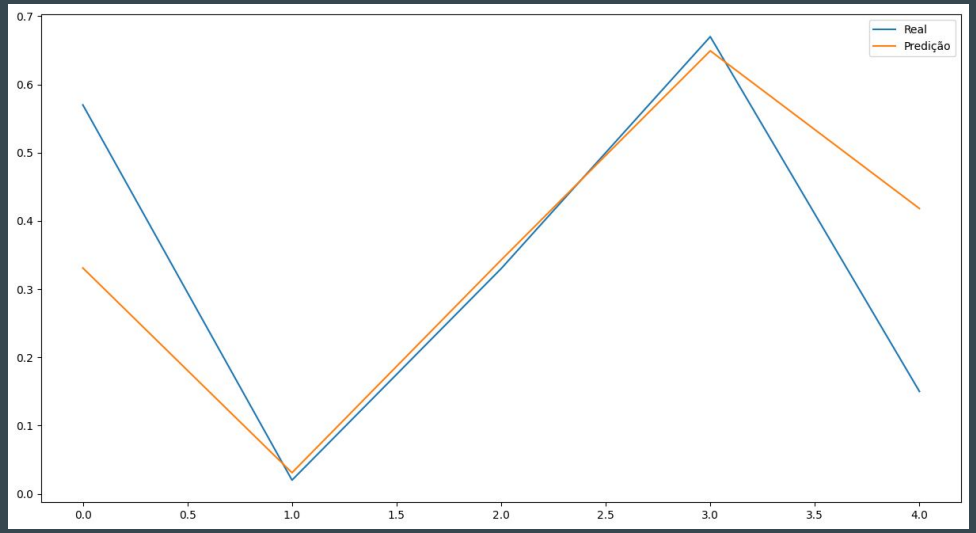


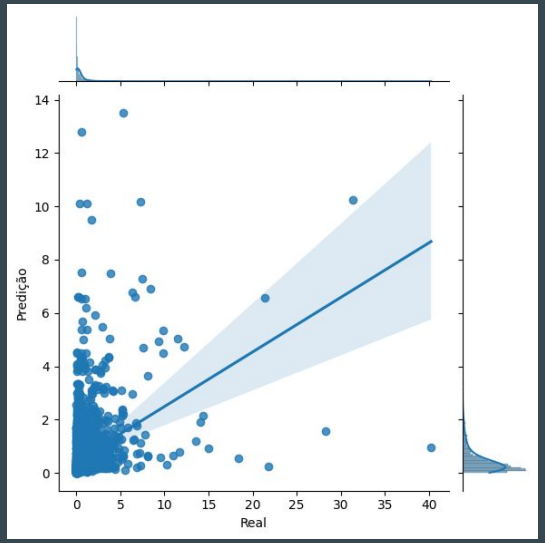


Resultados DADOS - GLOBAL

Random Forest Regressor

Erro Médio Absoluto 0.5151743266860834
Coeficiente de Determinação 0.0824564109887953





Conclusão

Melhor Modelo

Os melhores modelos foram:

1. Europa
2. Japão

Implicações:

- O método de medição utilizado foi o erro médio quadrático.
- Europa e Japão são os mercados mais influenciados pelo gênero do jogo.
- É possível achar uma tendência de mercado não linear.

Coeficientes

Coeficientes do modelo numérico calculado para Europa

Platform: 0.12964
Year: 0.37433817
Genre: 0.30163109
Publisher: 0.19439074

Análise de Influência - Predição de Vendas

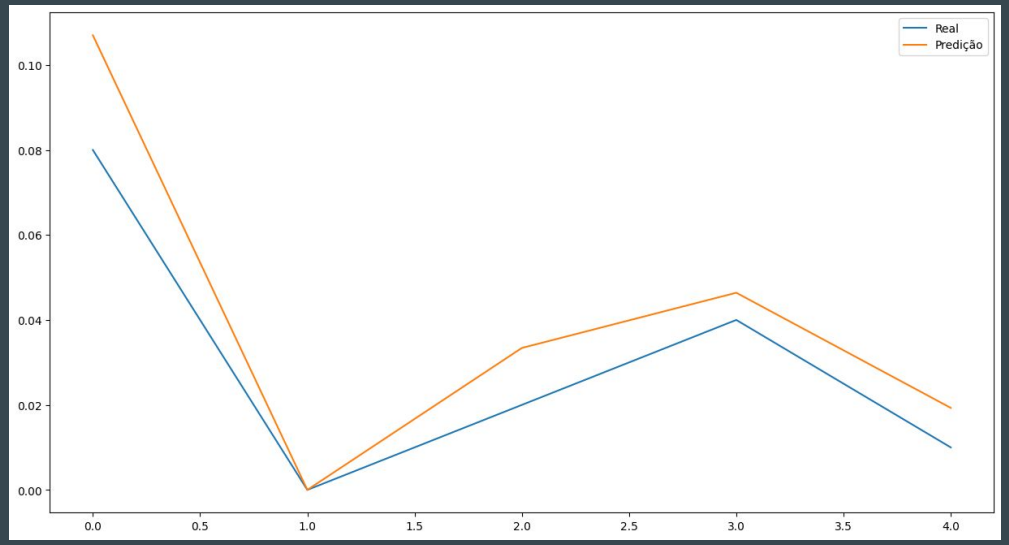
- Predição das vendas do outros(Fora dos pólos)
- Predição dos Valores de venda da América do Norte
- Utilização de valores da editora, ano, plataforma e gênero.
- Tentativa de identificação de características compartilhada entre o mercado NA e o resto do mundo.

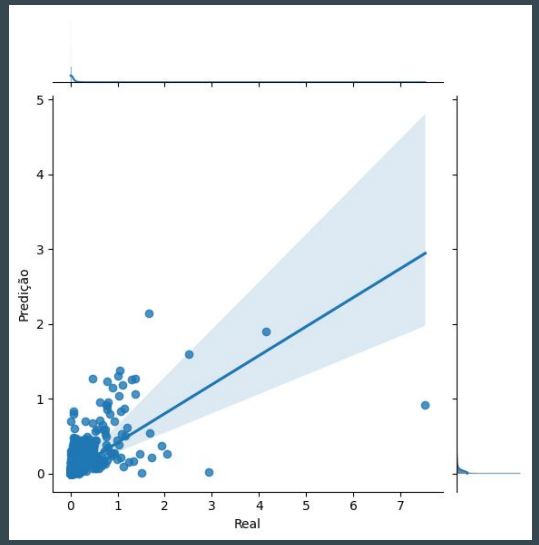
	Rank	Name	Platform	Year	Genre	Publisher	Global_Sales
0	1	Wii Sports	Wii	2006	Sports	Nintendo	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	33
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	31.37

Resultados NA - OU

Random Forest Regressor

Erro Médio Absoluto 0.02809158674801751
Coeficiente de Determinação 0.43071397290179736





Conclusão

Melhor Modelo

O melhor modelo foi novamente a Regressão por Floresta Randômica

Implicações:

- Existe uma relação não numérica entre o mercado Norte Americano e o resto do mercado mundial(Fora os grandes pólos)

Coeficientes

Coeficientes do modelo numérico calculado para Random Forest Regressor

Platform: 0.09420622
Year: 0.14278131
Genre: 0.07125195
Publisher: 0.05141113
NA_Sales: 0.64034938