

# Previsão de Vendas de Serviços de Coleta de Resíduos Utilizando Modelos de Aprendizado de Máquina e ARIMA

Antônio Maurício Brito Junior e Daniel Moraes de Lima

Universidade de Fortaleza, Fortaleza, Ceará, Brasil, 2021

**Resumo** Modelos de previsões confiáveis são necessários para subsidiar o planejamento e a tomada de decisão nos negócios. Este artigo realiza a previsão de vendas de serviços de coleta de resíduos sólidos, usando modelos de aprendizado de máquina, tendo como referência comparativa (*baseline*) o popular modelo estatístico autoregressivo integrado de médias móveis (ARIMA). As previsões foram realizadas utilizando uma série temporal univariada, obtida a partir de uma série histórica multivariada de vendas que compõem os diversos serviços da empresa.

**Keywords:** Resíduos Sólidos · Previsão em Séries Temporais · Aprendizado de Máquina · ARIMA.

## 1 Introdução

O crescimento populacional em 2020 no Brasil foi da ordem de 0,77% em relação a 2019, superando a marca de 211 milhões de brasileiros, segundo dados do IBGE. O crescimento da população vem seguido do aumento dos padrões de consumo e, por consequência, da maior geração de lixo por parte das pessoas, empresas e outras organizações. Todavia, nem sempre o descarte dos resíduos é feito de forma correta, trazendo consequências para o meio ambiente, como a contaminação do lençol freático e impactando na saúde humana. Com o intuito de reduzir o impacto do descarte dos resíduos sólidos no meio ambiente, foi criada, em 2010, a Política Nacional de Resíduos Sólidos (PNRS), tal política foi instituída pela Lei nº 12.305/10 e regulamentada pelo Decreto 7.404/10. Está política se propõe a organizar a forma com que o país lida com o lixo e exige de empresas públicas e privadas o correto descarte dos resíduos. Entre os tipos de resíduos citados na lei estão os domiciliares, industriais, resíduos de saneamento público, da saúde, da construção civil e resíduos tóxicos. Nesse cenário, a prestação de serviços de coleta de resíduos sólidos é uma atividade promissora, demandada em todos os lugares e por diversos setores da economia e contribui para que pessoas, empresas e governos gerenciem e realizem o correto descarte de seus resíduos.

Disponer de previsões confiáveis é fundamental para a tomada de decisão. Entretanto, dependências do ambiente macroeconômico diversificado, demandas de clientes que mudam constantemente e fatores como sazonalidade e eventos

emergenciais, são grandes desafios para que muitas empresas realizem boas previsões. A previsão de vendas para o horizonte de 30 dias ou mais, é crucial para o planejamento dos processos operacionais de uma empresa de gestão de resíduos. Uma suposição comum, é que se consegue melhores previsões de uma série temporal multivariada do que com uma série temporal escalar. No entanto, as séries multivariadas contêm vários tipos de informações e pode ser difícil manipulá-las para prever os passos seguintes. Por isso, a previsão univariada pode ocasionalmente ser superior à previsão multivariada [1].

Neste artigo, realizamos previsões para 3 períodos, usando uma série temporal de vendas como entrada para os modelos de aprendizado de máquina e autoregressivo integrado de médias móveis (ARIMA). Ao longo do estudo, nosso objetivo foi esclarecer o processo de análise e preparação da série temporal e as etapas necessárias para selecionar os melhores modelos de previsão que pudessem realizar previsões confiáveis a partir das informações disponíveis e colaborar para o melhor planejamento das atividades empresariais.

O artigo está estruturado da seguinte forma: Na seção 2, apresentamos os trabalhos relacionados (“estado da arte”), na seção 3, descrevemos a metodologia e técnicas usadas. Na seção 4, explicamos as métricas de avaliação e apresentamos os resultados dos experimentos, E, por fim, as conclusões e recomendações para trabalhos futuros.

## 2 Trabalhos Relacionados

No estudo de [2], os autores propõem dois modelos para prever demandas de curto e longo prazo para a compra de veículos elétricos. O primeiro modelo usado foi o análise de espectro singular (SSA, na sigla do inglês) que é um modelo não paramétrico e baseado em dados [3]. Um segundo modelo foi construído fundamentado em vetores autoregressivos (VAR, na sigla em inglês) para fazer previsões em séries multivariadas [4], uma vez que se considerou as influências externas de seis indicadores econômicos na previsão. Eles concluíram que os dois modelos foram adequados para realizar as previsões de demanda de veículos, ressaltando que o modelo SSA é satisfatório com menos requisitos de dados e o modelo VAR foi útil ao considerar fatores econômicos e a preferência dos consumidores. No artigo de [5], eles relatam o uso de uma rede neural perceptron de multicamadas (MLP) para fazer previsões de vendas para uma empresa. Segundo os autores, a seleção do modelo com rede neural foi devido a pequena quantidade de dados disponíveis na série temporal usada. Os resultados apresentados, indicam que o modelo MLP apresentou flexibilidade para lidar com a previsão de vendas usando uma série temporal curta. As previsões para um mês apresentaram erro inferior a 5%. No estudo de [6], é feito uso de modelos multiníveis com abordagem de empilhamento [7] para criar os modelos de previsão. Nesta abordagem, os resultados das previsões no conjunto de validação são tratados como regressores de entrada para os modelos de próximo nível. Foram utilizados nos níveis seguintes, modelos como a floresta aleatória, rede neural e o ARIMA. Eles concluíram que, usando a abordagem de empilhamento

para modelos de aprendizado de máquina, alcançaram melhor desempenho dos modelos preditivos.

### 3 Metodologia

#### 3.1 Conjunto de Dados e Séries Temporais

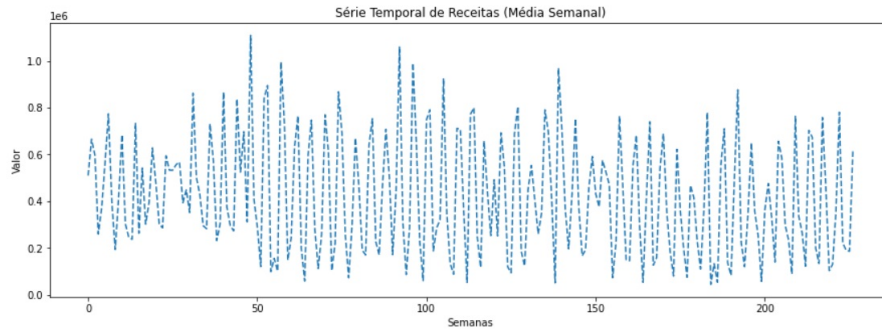
Para as previsões realizadas neste trabalho, foi utilizada uma série temporal, não categórica, mais concretamente de valores numéricos reais, que compreendem as vendas de serviços realizados de janeiro de 2017 a maio de 2021. Uma série temporal pode ser definida como uma coleção de variáveis aleatórias indexadas de acordo com a ordem em que são obtidas no tempo [8]. A série original é composta de 479 serviços distintos, caracterizando uma série multivariada. Como cada serviço que compunham a série original possui características distintas (unidade de medidas, precificação e até clientes), optamos por transformá-la em uma série univariada. Segundo o estudo de [9], as séries multivariadas são mais complexas de se obter as informações úteis para se fazer boas previsões, também, afirma que uma série escalar pode propiciar melhor desempenho que série multivariada quando os passos que se quer prever são pequenos. Para fazer as previsões foi realizada uma etapa de pré-processamento no conjunto de dados, retirando colunas que não eram relevantes para o modelo. Para gerar a série univariada as vendas foram totalizadas por dia e, em seguida, calculada a média semanal. Em resumo, a série usada é composta por 227 observações da média semanal das vendas dos serviços prestados, entre janeiro de 2017 a maio de 2021. Estes dados foram obtidos a partir de registros dos sistemas gerenciais da empresa. A figura 1 apresenta uma amostra dos dados utilizados e a figura 2 apresenta o gráfico da série temporal.

	Data	Produto	Qtd	Valor
0	2021-05-06	10209571015003	1.0	24596.64
1	2021-05-06	10209571015003	1.0	11108.16
2	2021-05-06	10209571015003	1.0	10116.36
3	2021-05-06	10209571015003	1.0	13290.12
4	2021-05-06	10910761015001	1.0	1850.67

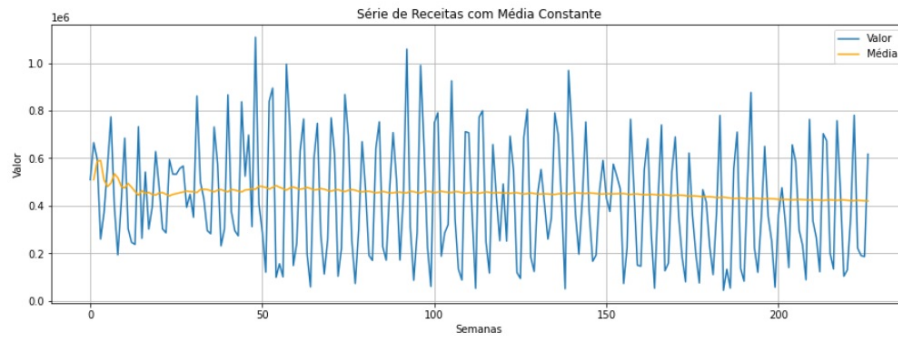
**Figura 1.** Amostra do Conjunto de Dados.

#### 3.2 Previsão de Vendas

Objetivando selecionar os modelos de previsão e tirar o máximo desempenho deles, foi realizada uma etapa de análise da série. Para [8] o primeiro passo em



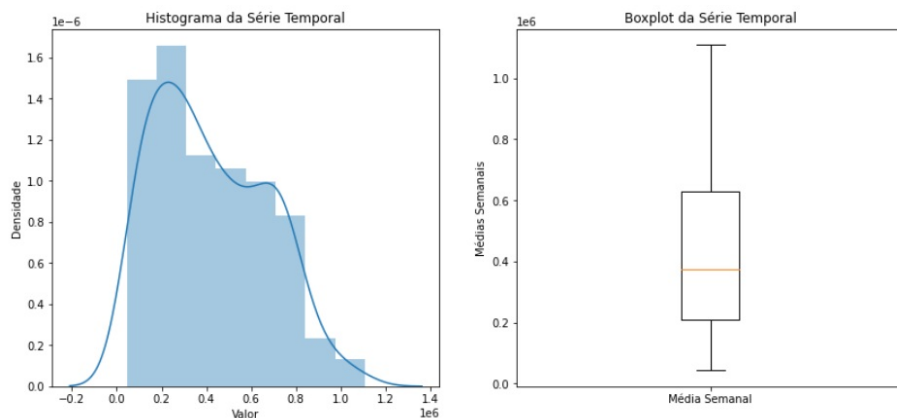
**Figura 2.** Série de Vendas (Médias Semanais)



**Figura 3.** Média da Série ao Longo do Período

qualquer investigação de série temporal sempre envolve um exame cuidadoso dos dados registrados plotados ao longo do tempo (ver figura 2). Inicialmente, foram calculadas as medidas de estatística descritiva. Na série histórica, temos o valor mínimo de R\$ 44.510,91, a média de R\$ 420.951,49, o valor máximo de R\$ 1.108.648,15 e o desvio padrão de R\$ 251.421,57. Em seguida, a visualização da série por meio dos gráficos nos permite maior aprofundamento nas análises. Analisando o Boxplot, verifica-se que a série não apresenta *outliers*. O gráfico das médias da série (figura 3), sugere uma leve tendência decrescente ao longo do período observado. Através do histograma da figura 4, é possível observar que os dados não seguem uma distribuição normal e apresentam uma simetria positiva, a não normalidade, também, foi checada pelo teste de Shapiro-Wilk [10,11] que apresenta valor de p igual a 0 (zero) e o teste estatístico 0,9542, indicando que os mesmos precisarão sofrer uma transformação (como uma diferenciação) para que possam ser utilizados modelos auto-regressivos, por exemplo. Também, fica claro que a maior quantidade de amostras são de valores abaixo da média verificada.

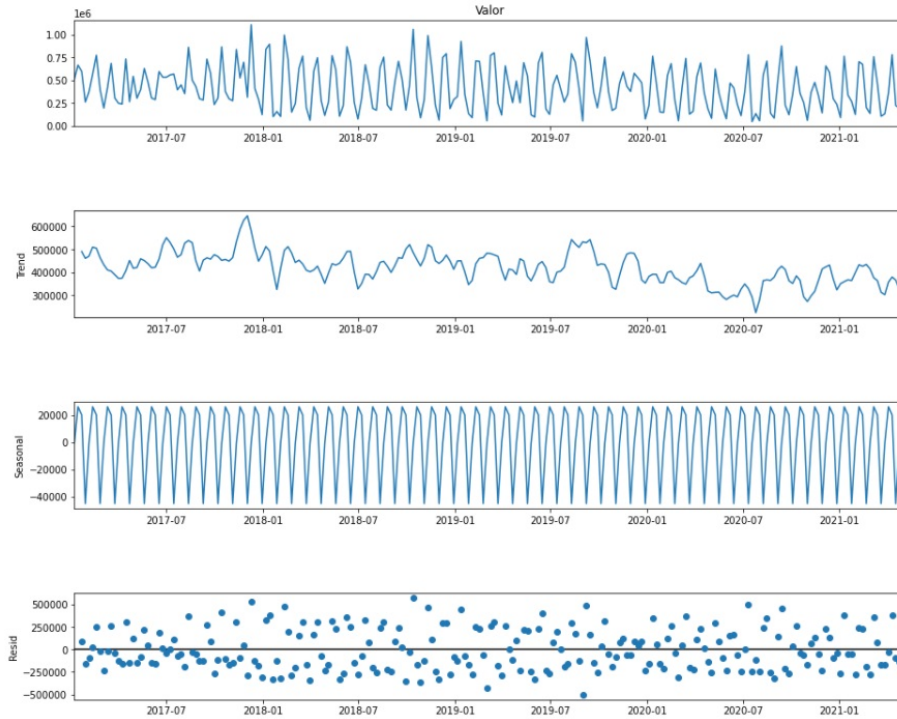
Na sequência, foi realizada a decomposição da série temporal para melhor iden-



**Figura 4.** Histograma e Boxplot da Série Temporal

tificar o comportamento dos componentes de tendência, de sazonalidade e de resíduos da amostra, e que também ajudará para a escolha dos métodos de previsão [9]. Na figura 5, percebe-se que há uma leve inclinação negativa e que os dados possuem uma variabilidade constante, com ciclos trimestrais, em torno da tendência. A tendência negativa, foi causada pela redução das vendas, provavelmente, a redução foi influenciada pelo período recessivo da atividade econômica no país. A sazonalidade se apresenta em ciclos mensais (4 semanas), apresentando picos no meio do mês e, do ponto de vista anual, há um incremento gradual com ápice nos finais de ano. No gráfico de componentes residuais, percebe-se uma grande quantidade de resíduos que se apresentam em ciclos mensais, apresentando picos aleatórios na receita dos serviços. Conforme o estudo de [12] devemos considerar o comportamento de todos estes padrões na análise dos riscos da previsão.

No passo seguinte de análise, realizou-se os testes de Dickey-Fuller e KPSS, ambos para checar a estacionariedade da série. Estacionariedade é um termo introduzido para descrever uma regularidade assumida em uma série de dados [8]. Nas séries temporais estacionárias, suas propriedades não dependem do tempo em que a série é observada [13]. No teste de Dickey-Fuller, obtivemos 0,2878 para o valor de  $p$ . Como a hipótese nula deste teste é que a série é não-estacionária caso o valor de  $p$  seja maior que o valor de  $p$  limite definido (no caso, adotou-se 5% ou 0,05) e, o valor do teste estatístico de -1,9971 é negativo e maior que o valor crítico estabelecido de 5% (-2,8751). Portanto, para o teste de Dickey-Fuller, conclui-se que a série é não-estacionária. Se chega nessa mesma conclusão através do teste KPSS (teste criado por Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt e Yongcheol Shin). Nesse teste, a hipótese nula é que a série temporal é estacionária. Portanto, caso o valor de  $p$ , seja maior que 0,05, a série é considerada estacionária. Caso o valor de  $p$  seja menor ou igual a limite, a série será considerada não-estacionária. No caso, o valor de  $p$  obtido foi 0,0100,

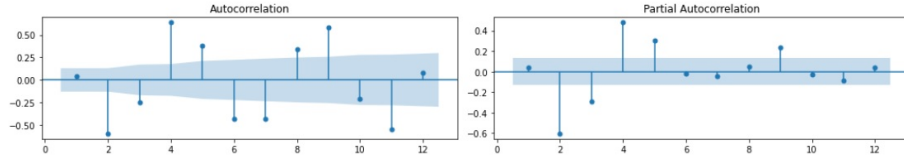


**Figura 5.** Decomposição da Série Temporal

que é inferior ao valor de  $p$  limite e, o valor do teste estatístico de 0,8937 é maior que o valor crítico estabelecido de 5% (0,4630), indicando que a série é não-estacionária.

Na etapa final investigativa da série temporal, foi analisada a autocorrelação, que significa comparar o valor de um período presente com valores de períodos passados (conhecidos como lags) da mesma série. A autocorrelação parcial leva em conta apenas a correlação direta com o período anterior, enquanto a autocorrelação, considera a correlação direta e indireta. No gráfico da figura 6, observa-se que o há autocorrelação negativa com o segundo lag e positiva com o quarto lag, ainda que indiquem correlação baixa, estes lags podem ser usados para criar previsões para eventos futuros.

**Modelo de Previsão SARIMA.** Neste trabalho, foi adotado o modelo de previsão SARIMA (Sazonal ARIMA) como referência comparativa para os modelos baseados em aprendizado de máquina. A popular família de modelos ARIMA [8], da qual o SARIMA faz parte, são baseados nas autocorrelações existentes nos dados e são aplicados para fazer previsões em séries que mostram evidências de não estacionariedade e que uma diferenciação inicial pode ser realizada, uma ou mais vezes, para eliminar a não estacionariedade [13]. Modelos ARIMA não



**Figura 6.** Gráfico de Autocorrelação e Autocorrelação Parcial

sazonais são geralmente denotados como  $ARIMA(p, d, q)$ , em que os parâmetros  $p$ ,  $d$  e  $q$  são números inteiros não negativos,  $p$  é a ordem (número de defasagens) do modelo auto-regressivo,  $d$  é o grau de diferenciação (o número de vezes em que os dados tiveram valores passados subtraídos) e  $q$  é a ordem do modelo de média móvel. Modelos  $ARIMA$  sazonais são denotados como  $ARIMA(p, d, q)(P, D, Q)_m$ , em que  $m$  se refere ao número de períodos em cada temporada (no caso, 52 semanas) e  $P$ ,  $D$  e  $Q$  se referem aos termos de auto-regressão, diferenciação e média móvel para a parte sazonal do modelo  $ARIMA$  [14,13]. O algoritmo utilizado neste artigo está disponível na biblioteca statsmodels para Python, hospedado em statsmodels.org [15]. Para a definição dos parâmetros do modelo, além da análise dos gráficos de autocorrelação e autocorrelação parcial, foi avaliada a métrica AIC (Akaike's Information Criterion) [16], uma vez que o modelo misto Sazonal  $ARIMA$  é difícil de identificar apenas por métodos visuais, a métrica AIC forneceu uma estimativa aproximada dos valores possíveis  $p$ ,  $d$ ,  $q$ ,  $P$ ,  $D$ , e  $Q$ . Por esta métrica, o melhor modelo é aquele com o valor AIC mais baixo [17]. Para este artigo, foi desenvolvida uma rotina para avaliar as possibilidades e encontrar o menor valor de AIC, obtendo os seguintes parâmetros para o modelo final:  $SARIMA(1, 1, 1)(1, 1, 0)_{52}$ .

**Algoritmos de Aprendizado de Máquina.** Segundo [12] a previsão de vendas é mais um problema de regressão do que um problema de série temporal. A prática mostra que o uso de abordagens de regressão muitas vezes pode nos dar melhores resultados em comparação com métodos mais utilizados com séries temporais. Os algoritmos de aprendizado de máquina permitem encontrar padrões nas séries temporais. Podemos encontrar padrões complicados na dinâmica de vendas, usando métodos de aprendizado de máquina supervisionados. Esses modelos são construídos de forma supervisionada, o que significa que os dados com as variáveis alvo desejadas devem ser preparados com antecedência [18], dentre estes, os mais populares são algoritmos de aprendizado de máquina baseados em árvore [19], como por exemplo, *Random Forest* [20], *Gradient Boosting Machine* [18,21,22], *Linear Regression* [23] e redes neurais artificiais. Neste estudo, não há rigor na definição dos hiperparâmetros dos algoritmos, pois o propósito não foi buscar o melhor desempenho de cada algoritmo, mas apenas, avaliar de maneira geral, quais algoritmos realizaram boas previsões.

### 1 - Regressão Linear

A regressão linear realiza a tarefa de prever um valor de variável dependente ( $y$ ) com base em uma determinada variável independente ( $x$ ) [23]. Portanto, esta técnica de regressão descobre uma relação linear entre  $x$  (entrada ou atributos) e  $y$  (saída ou alvo). A regressão linear pode ser expressa pela seguinte equação:

$$y = \theta_1 + \theta_2 x \quad (1)$$

O objetivo do algoritmo de regressão linear é encontrar os melhores valores para  $\theta_1$  e  $\theta_2$ . Ele consegue isso, através de uma função de custo e da descida em gradiente.

### 2 - Árvore de Decisão (*Decision tree*)

Árvores de classificação e regressão são métodos de aprendizado de máquina para construir modelos de previsão a partir de dados. Os modelos são obtidos particionando recursivamente o espaço de dados e ajustando um modelo de predição simples em cada partição. Como resultado, o particionamento pode ser representado graficamente como uma árvore de decisão. As árvores de regressão são para variáveis dependentes que assumem valores discretos contínuos ou ordenados, com erro de predição tipicamente medido pela diferença quadrática entre os valores observados e preditos [24].

### 3 - Florestas Aleatórias (*Random forests*)

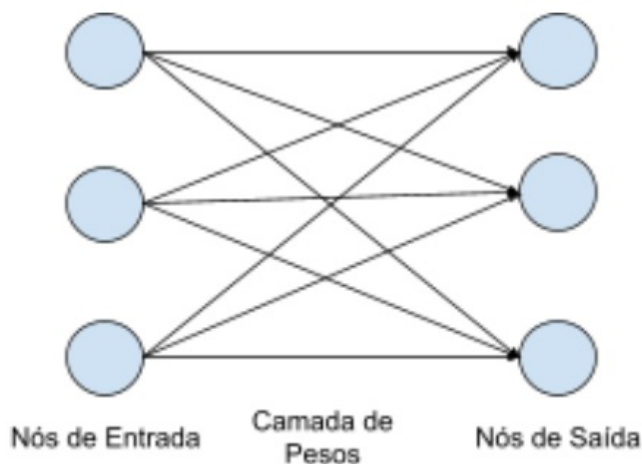
Florestas aleatórias é um algoritmo de aprendizado supervisionado criado por Leo Breiman [20]. É um algoritmo muito flexível e fácil de usar. Uma floresta é feita de árvores. Diz-se que quanto mais árvores você tem, mais robusta é a floresta, isso também reduz a possibilidade de sobreajuste (*overfitting*). Ele funciona da seguinte maneira: o algoritmo seleciona amostras aleatórias de um determinado conjunto de dados e, em seguida, constrói uma árvore de decisão para cada amostra e obtém um resultado de previsão de cada árvore de decisão. Para a regressão, a média das previsões de cada árvore é o resultado final do algoritmo.

### 4 - Máquina de Aumento de Gradiente (*Gradient Boosting Machine*)

As técnicas comuns de *ensemble* [25], como florestas aleatórias, dependem de uma média simples de modelos de *ensemble*. A família de métodos de *boosting* é baseada em uma estratégia diferente e construtiva de formação de *ensembles*. A ideia principal do *boosting* é adicionar novos modelos de *ensemble* sequencialmente. A cada iteração particular, um novo modelo de previsão fracos, geralmente árvores de decisão, é treinado com relação ao erro de todos os *ensembles* aprendido até então [18]. Em Máquinas de Aumento de Gradiente, ou simplesmente da sigla em inglês, GBMs, o procedimento de aprendizagem se ajusta consecutivamente a novos modelos para fornecer uma estimativa mais precisa da variável de resposta. A ideia principal por trás desse algoritmo é construir os novos aprendizes básicos para serem correlacionados ao máximo com o gradiente negativo da função de perda, associado a todos os *ensembles*. Os GBMs mostraram um sucesso considerável não apenas em aplicações práticas, mas também em vários desafios de



aprendizado de máquina e mineração de dados [26,27].



**Figura 7.** Ilustração de um Perceptron Simples

#### 5 - Perceptron de Multi-camadas (*Multi-layer Perceptron*)

Redes Neurais Artificiais são um paradigma de programação que busca emular a microestrutura do cérebro e são amplamente utilizadas em problemas de inteligência artificial. O *Multi-layer Perceptron* (MLP) é um exemplo de rede neural artificial amplamente usada para a solução de vários problemas diferentes, incluindo reconhecimento de padrões, desde tarefas simples de reconhecimento de padrões até interpolação [28]. Elas geralmente superam o desempenho de modelos de aprendizado de máquina tradicionais porque têm as vantagens de não linearidade, a possibilidade de interações entre variáveis e personalização de sua arquitetura e parâmetros. Uma rede neural Perceptron mais simples, consiste em uma camada de neurônios de entrada, acoplada a uma camada de neurônios de saída, e uma única camada de pesos entre eles, como mostrado na Figura 7. O processo de aprendizado, de uma forma simples, o perceptron recebe entradas, multiplica-as por alguns pesos e, em seguida, as passa para uma função de ativação (como logística, linear retificada - relu, tangente hiperbólica, identidade) para produzir uma saída. A representação esquemática dada na Figura 7 é frequentemente como as redes neurais são descritas na literatura, embora matematicamente seja útil pensar nas camadas de entrada e saída como vetores de valores (I e O respectivamente) e os pesos como uma matriz [28].

Para o desenvolvimento dos experimentos, utilizamos o ambiente Jupyter Notebook, com a linguagem Python e as bibliotecas Pandas, Scikit Learn, Numpy, Scipy, Matplotlib e Seaborn.

## 4 Avaliação e Resultados

Nesta seção, descreveremos as técnicas e métricas usadas para a avaliação e apresentamos os resultados obtidos.

**Métricas.** Foram selecionadas métricas populares na avaliação de séries temporais e que fossem fáceis de as interpretar. Por estes critérios, escolhemos MAE, RMSE, MAPE e  $R^2$ . A seguir detalhamos cada uma delas.

### 1 - MAE (*Mean Absolute Error*).

O erro médio absoluto é a média dos erros da série avaliada, porém, antes os erros de valores negativos são transformados em positivos. Essa métrica é muito usada em séries temporais, pois há casos que o erro negativo pode zerar o positivo ou dar uma ideia que o modelo é preciso, no caso do MAE isso não acontece, pois essa métrica mostra o quanto a previsão está longe dos valores reais, independente se a maior ou a menor. O MAE é popular porque é fácil de entender e calcular ao comparar métodos de previsão aplicados a uma única série temporal ou a várias séries temporais com as mesmas unidades. Um método de previsão que minimiza o MAE levará a previsões da mediana [13]. Seja  $Y_t$  a observação no tempo  $t$  e  $P_t$  a previsão de  $Y_t$ . Então, define-se o erro de previsão por:

$$e_t = Y_t - P_t \quad (2)$$

Logo, temos que:

$$MAE = média(|e_t|) \quad (3)$$

### 2 - RMSE (*Root Mean Square Error*)

Para entender o RMSE, primeiro se deve entender o MSE ou erro quadrático médio, Essa métrica coloca mais peso nos erros maiores, pois cada valor individual do erro é elevado ao quadrado e posteriormente a média desses valores é calculada. Assim, essa métrica é muito sensível à *outliers* e coloca bastante peso nas previsões com erros mais significativos. O RMSE é simplesmente a raiz quadrada do MSE, onde o erro volta a ter a unidade de medida dos dados, ela é muito usada em séries temporais porque é mais sensível a erros maiores devido ao processo de elevação ao quadrado que a originou [29]. É expressa pela formula:

$$MSE = \sqrt{(média(e_t^2))} \quad (4)$$

$$RMSE = \sqrt{MSE} \quad (5)$$

### 3 - MAPE (*Mean Absolute Percentage Error*).

O erro percentual médio absoluto é dado por  $p_t = 100e_t/y_t$ . Os erros de porcentagem têm a vantagem de não possuírem unidades e, portanto, são freqüentemente

usados para comparar desempenhos de previsão entre conjuntos de dados [29]. É expressa pela equação:

$$MAPE = média(|p_t|) \quad (6)$$

Medidas baseadas em erros percentuais têm a desvantagem de serem infinitas ou indefinidas se  $y_t = 0$  para qualquer  $t$  no período de interesse, e ter valores extremos se qualquer  $y_t$  for próximo a zero. Eles também têm a desvantagem de colocar uma penalidade mais pesada nos erros negativos do que nos positivos [29,13].

4 - Coeficiente de Determinação. O coeficiente de determinação, também conhecido como  $R^2$  avalia a qualidade da previsão do modelo utilizado (equação 7). Seus valores são apresentados entre 0 e 1 e expressa a quantidade da variância dos dados que é explicada pelo modelo linear. Assim, quanto maior o  $R^2$ , mais explicativo é o modelo linear, ou seja, melhor ele se ajusta à amostra [30]. Usamos o  $R^2$  como referência nas comparações de acurácia dos modelos avaliados neste estudo.

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

Onde  $n$  é o numero de observações,  $y_i$  é o valor observado,  $\bar{y}$  é a média das observações e  $\tilde{y}_i$  é o valor previsto de  $y_i$ .

#### 4.1 Resultados Dos Experimentos.

Para realizar as previsões utilizamos a série histórica de vendas de janeiro de 2017 a maio de 2021. A amostra foi dividida em conjunto de dados de treino e de teste. É importante mencionar que, no caso de previsão de série temporal, não podemos usar uma abordagem convencional de separação. Temos que dividir a série histórica em conjuntos de dados de treinamento e de teste, de modo que os dados de treinamento fiquem no primeiro período de tempo e os de teste fiquem no intervalo seguinte [12]. Foram feitas previsões para os períodos de 4, 12 e 52 semanas com os modelos selecionados. Importante destacar que, a menos que a série em análise seja estacionária, conforme o horizonte de previsão cresce, os erros de previsão aumentam sem limites [8].

#### 4.2 Previsão de vendas para 4 semanas

O modelo LGBM obteve o melhor desempenho, com  $R^2$  de 0,9784, o Random Forest obteve 0,9189, ficando com o segundo melhor coeficiente e o SARIMA, em terceiro, obteve 0,8470. Para o período analisado, percebe-se uma forte tendência de aumento das vendas. Os algoritmos apresentaram boa estabilidade e excelente desempenho para a previsão de 4 semanas. Porém, o DT e o MLP se mostraram mais instáveis [31], por esse motivo, os experimentos foram repetidos 100 vezes e calculadas as médias para as métricas apresentadas. Outra métrica que merece

**Tabela 1.** Métricas para a previsão de 4 semanas.

Modelo	MAE	RMSE	MAPE(%)	$R^2$
SARIMA	60.243,36	70.898,18	30,06	0,8470
LR	88.046,07	95.445,45	32,91	0,7226
DT	102.715,37	123.389,24	49,82	0,5364
RF	40.702,96	53.038,93	16,34	0,9143
LGBM	22.689,53	26.634,62	9,12	0,9784
MLP	68.580,36	69.787,01	28,55	0,8517

destaque é o MAE, pois ela mostra maior "espaçamento" entre os algoritmos em relação ao  $R^2$ . Para exemplificar, o RF fica a 6,55% de alcançar o LGBM, porém quando considera-se o MAE, a defasagem supera os 79%. Isso pode indicar que as previsão do RF para os picos da série são piores que o indicado pelo  $R^2$ . A tabela 1 apresenta as métricas e a tabela 2 os valores previstos pelos modelos para 4 semanas.

**Tabela 2.** Valores Previstos Para 4 Semanas (R\$).

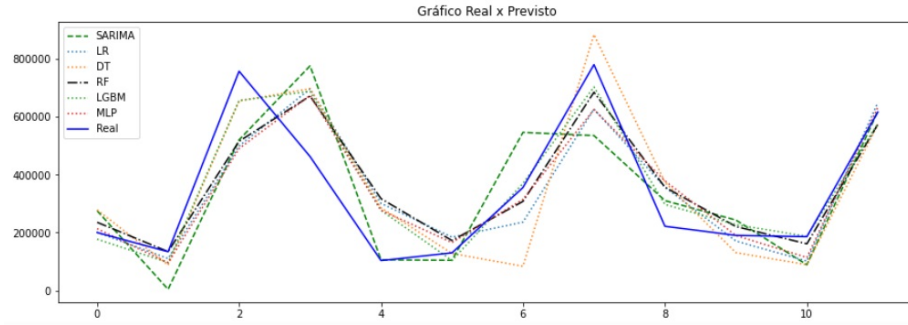
Semana	REAL	SARIMA	LR	DT	RF	LGBM	MLP
1	222.107,59	310.772,14	262.012,47	362.279,17	319.696,88	229.163,62	310.445,24
2	190.551,25	243.486,55	92.610,40	137.471,56	202.762,43	232.510,95	138.438,94
3	186.117,31	88.969,44	112.741,67	380.820,59	203.278,80	174.077,72	118.515,97
4	616.894,22	619.119,95	757.857,13	593.987,30	581.044,35	646.597,02	683.164,37

#### 4.3 Previsão de vendas para 12 semanas

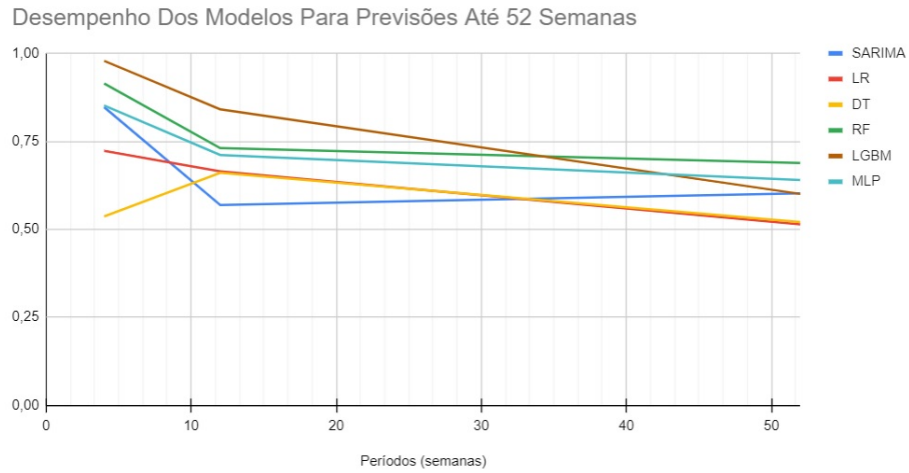
A tendência indica crescimento moderado. Novamente, o modelo LGBM apresenta os melhores indicadores, ficando o  $R^2$  em 0,8407 e o MAE de 69.341,63. Como era esperado, o erro da previsão aumentou com o aumento do período. Essa degradação foi da ordem de 11% na média dos modelos avaliados na comparação com a previsão de 4 semanas. O gráfico da figura 8 mostra as séries previstas e real para 12 semanas.

#### 4.4 Previsão de vendas para 52 semanas

A tendência se mostra estável. Porém, o modelo que apresentou melhor desempenho foi o RF, seu  $R^2$  foi de 0,6880 e em segundo lugar o MLP com 0,6396. Novamente, verifica-se um aumento dos erros das previsões, alcançando o percentual médio de 24,84%. O gráfico da figura 9 apresenta a queda de desempenho dos modelos para previsões até 52 semanas.



**Figura 8.** Série Temporal Real e Previstas por Cada Modelo.



**Figura 9.** Desempenho dos Modelos para Previsões ate 52 Semanas ( $R^2$ )

## 5 Conclusão

Realizar boas previsões de vendas e com o menor erro possível é fundamental para o planejamento e a competitividade nos negócios [32]. Desenvolver um modelo confiável de previsão de vendas recebe especial atenção das empresas. Neste artigo, analisamos modelos de previsão em uma série temporal de vendas de resíduos sólidos, com algoritmos de aprendizado de máquina e o popular modelo auto-regressivo ARIMA para os períodos de 4, 12, e 52 semanas. Através dos experimentos realizados, comprovamos que os modelos de aprendizado de máquina superaram o modelo estatístico de auto-regressão. O modelo LGBM obteve o melhor desempenho para a previsão de 4 e 12 semanas, com  $R^2$  de 0,9784 e 0,8407, respectivamente. Na previsão de 4 semanas, o LGBM superou o ARIMA em 15,51% e para 12 semanas em 47,78%. Na previsão de 52 sema-

nas, o modelo floresta aleatória (RF) obteve o melhor desempenho, com  $R^2$  de 0,6880, superando o ARIMA em 14,25%. Em trabalhos futuros, experimentos com modelos híbridos poderão ser avaliados visando melhorar as previsões de curto e longo prazo. Além disso, previsões com redes neurais profundas poderá colaborar para realizar previsões com séries temporais mais complexas, como as multivariadas e considerar atributos exógenos para enriquecimento do modelo e fornecer previsões mais confiáveis.

## Referências

1. M. Chayama and Y. Hirata, “When univariate model-free time series prediction is better than multivariate,” *Physics Letters A*, vol. 380, no. 31-32, pp. 2359–2365, 2016.
2. Y. Zhang, M. Zhong, N. Geng, and Y. Jiang, “Forecasting electric vehicles sales with univariate and multivariate time series models: The case of china,” *PloS one*, vol. 12, no. 5, p. e0176729, 2017.
3. N. Golyandina, V. Nekrutkin, and A. A. Zhigljavsky, *Analysis of time series structure: SSA and related techniques*. CRC press, 2001.
4. M. G. Dekimpe and D. M. Hanssens, “Time-series models in marketing:: Past, present and future,” *International journal of research in marketing*, vol. 17, no. 2-3, pp. 183–193, 2000.
5. R. M. C. Croda, D. E. G. Romero, and S.-O. C. Morales, “Sales prediction through neural networks for a small dataset,” *IJIMAI*, vol. 5, no. 4, pp. 35–41, 2019.
6. B. Pavlyshenko, “Using stacking approaches for machine learning models,” in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 255–258.
7. D. H. Wolpert, “Stacked generalization,” *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.
8. R. H. Shumway, D. S. Stoffer, and D. S. Stoffer, *Time series analysis and its applications*. Springer, 2000, vol. 3.
9. G. P. Souza, R. W. Samohyl, and R. Meurer, “Previsão do consumo de energia elétrica do setor industrial em santa catarina—um estudo comparativo entre diferentes métodos de previsão através de suas discrepâncias,” *Anais do Simpósio Brasileiro de Pesquisa Operacional*, 2004.
10. S. Shaphiro and M. Wilk, “An analysis of variance test for normality,” *Biometrika*, vol. 52, no. 3, pp. 591–611, 1965.
11. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, 2020.
12. B. M. Pavlyshenko, “Machine-learning models for sales time series forecasting,” *Data*, vol. 4, no. 1, p. 15, 2019.
13. R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

14. S. Inc, "Csas/stat® 9.3 user's guide. care," NC: SAS Institute Inc, 2011.
15. S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *Proceedings of the 9th Python in Science Conference*, vol. 57. Austin, TX, 2010, p. 61.
16. H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.
17. D. Eni *et al.*, "Seasonal arima modeling and forecasting of rainfall in warri town, nigeria," *Journal of Geoscience and Environment Protection*, vol. 3, no. 06, p. 91, 2015.
18. A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.
19. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.
20. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
21. J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.
22. —, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.
23. J. T. Pohlman and D. W. Leitner, "A comparison of ordinary least squares and logistic regression," 2003.
24. W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.
25. D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.
26. A. Bissacco, M.-H. Yang, and S. Soatto, "Fast human pose estimation using appearance and motion via multi-dimensional boosting regression," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.
27. R. Hutchinson, L.-P. Liu, and T. Dietterich, "Incorporating boosted regression trees into ecological latent variable models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011.
28. L. Noriega, "Multilayer perceptron tutorial," *School of Computing. Staffordshire University*, 2005.
29. R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.
30. C. Catal, E. Kaan, B. Arslan, and A. Akbulut, "Benchmarking of regression algorithms and time series analysis techniques for sales forecasting," *Balkan Journal of Electrical and Computer Engineering*, vol. 7, no. 1, pp. 20–26, 2019.
31. G. Fan and J. B. Gray, "Regression tree analysis using target," *Journal of Computational and Graphical Statistics*, vol. 14, no. 1, pp. 206–218, 2005.
32. M. Bohanec, M. K. Borštnar, and M. Robnik-Šikonja, "Explaining machine learning models in sales predictions," *Expert Systems with Applications*, vol. 71, pp. 416–428, 2017.