UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

VINICIUS NASCIMENTO TARGA
Orientador: Rodrigo César Pedrosa Silva

CUT THE TAILS: UMA ABORDAGEM PARA MODELOS DE REGRESSÃO COM CAUDAS PESADAS

Ouro Preto, MG
2023

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

VINICIUS NASCIMENTO TARGA

CUT THE TAILS: UMA ABORDAGEM PARA MODELOS DE REGRESSÃO COM CAUDAS PESADAS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Rodrigo César Pedrosa Silva

Ouro Preto, MG
2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Vinicius Nascimento Targa

Cut the Tails: Uma abordagem para modelos de regressão com caudas pesada

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 21 de Agosto de 2023.

Membros da banca

Rodrigo César Pedrosa Silva (Orientador) - Doutor - Universidade Federal de Ouro Preto
Josemar Coelho Felix (Coorientador) - Mestre - Universidade Federal de Ouro Preto
Lauro Angelio Gonçalves de Morais (Examinador) - Mestre - Universidade Federal de Ouro Preto
Gabriel Bicalho Ferreira (Examinador) - Bacharel - Universidade Federal de Ouro Preto

Rodrigo César Pedrosa Silva, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 21/08/2023.

Documento assinado eletronicamente por **Rodrigo Cesar Pedrosa Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 25/08/2023, às 11:42, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0577011** e o código CRC **2CBA10CE**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho a meu Pai e mentor, Claudio Norbiato Targa, por sempre estar do meu lado a vida inteira.

Agradecimentos

À instituição de ensino Universidade Federal de Ouro Preto, essencial no meu processo de formação profissional, pela dedicação, e por tudo o que aprendi ao longo dos anos do curso. Ao professor Rodrigo César Pedrosa Silva, por ter sido meu orientador e ter desempenhado tal função com dedicação. A todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho.

O sucesso nada mais é que ir de fracasso em fracasso sem que se perca o entusiasmo.
(Winston Churchill)

Resumo

A presença de dados que apresentam uma distribuição de caudas pesadas é mais comum do que se espera em determinadas áreas, pois um número considerável de casos no mundo real apresenta esse tipo de comportamento. O uso de técnicas de aprendizado de máquina tradicionais nesses tipos de distribuição pode não ser satisfatório o suficiente para gerar um modelo com acurácia considerável. O objetivo deste trabalho é criar uma abordagem que possa ser usada em distribuições de caudas pesadas, e que possa melhorar a acurácia dos modelos de aprendizado de máquina. Portanto, este trabalho propõe um método chamado "Cut the Tails", que, em sua metodologia, separa a distribuição da variável alvo de suas caudas pesadas e produz modelos de aprendizado de máquina separadamente para cada parte da distribuição. Ao comparar essa estratégia com as abordagens tradicionais de aprendizado de máquina, observou-se uma melhoria no erro absoluto médio percentual nos casos particulares que apresentaram curtose e assimetria elevados, sendo o melhor caso uma redução de erro de aproximadamente 48% ao usar a estratégia proposta.

Palavras-chave: Distribuição de Caudas Pesadas. Aprendizado de Máquina. Estatística.

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

Em muitos modelos estatísticos, assume-se que dados seguem uma distribuição normal. No entanto, estes dados podem apresentar um comportamento diferente da normalidade, exibindo assimetria em sua distribuição com valores distantes aparecendo mais frequentemente. Este comportamento da distribuição é denominado "Cauda Pesada"(SUN; FREES; ROSENBERG, 2008).

Contudo, uma distribuição de cauda pesada pode apresentar outras características, a mais importante sendo que esse tipo de distribuição não é exponencialmente limitada, isto é, ao ser comparada com uma distribuição da família exponencial (por exemplo, a distribuição normal ou exponencial), as suas caudas apresentam uma probabilidade de ocorrência maior do que as caudas da família exponencial, sendo possível observar visualmente caudas com valores mais elevados.

Existe uma necessidade de tratamento deste tipo de distribuição, já que ela ocorre em diversos problemas no mundo real, quando a normalidade não se aplica devido a quantidades significativas de valores extremos. Um exemplo são em dados de assistência médica, onde um pequeno número de pacientes requerem um custo mais elevado de recursos para tratamentos (MANNING; BASU; MULLAHY, 2005).

Neste trabalho, examinaremos abordagens para tratar dados com distribuição de cauda pesada em problemas reais. Este trabalho tem como objetivo testar uma adaptação da abordagem de Felix (2022) em outros conjuntos de dados com presença de caudas pesadas e verificar se pode haver melhoras na acurácia nos casos estudados. A versão adaptada é denominada *Cut the Tails*, onde, em lugar de um método de separação de *inliers* e *outliers* utilizando os quartis propostos por Felix (2022), são utilizados quantis para realizar um corte inferior e superior na distribuição, separando o pico da distribuição e a(s) cauda(s) da distribuição. Após o corte das caudas, são feitos os modelos de aprendizado de máquina para as caudas e o pico, para predizer os valores da variável alvo. Além disso, será observado como este algoritmo se comporta ao ser aplicado em diferentes tipos de distribuições com cauda pesada, comparando o desempenho com técnicas tradicionais de aprendizado de máquina, como Regressão Linear, Redes Neurais e *Random Forests*.

Os objetivos específicos deste trabalho são:

- Implementar um algoritmo de corte de caudas pesadas, com o uso de algoritmos de otimização para a seleção dos percentis.
- Elaborar uma metodologia de aprendizado de máquina para a predição das variáveis objetivo;

- Fazer um levantamento de bases de dados com diferentes índices de curtose e assimetria para observar o comportamento do algoritmo;
- Comparar os resultados com abordagens de aprendizado de máquina tradicionais, para verificar se houve um aumento na acurácia da predição.

Este trabalho está organizado da seguinte maneira:
O segundo capítulo se trata da revisão bibliográfica e trabalhos relacionados com este. O capítulo 3 se define o referencial teórico, onde será descrito os conceitos e métodos utilizados no trabalho. O capítulo 4 remete a metodologia/desenvolvimento do trabalho, onde será descrito os métodos e materiais utilizados para reproduzir os resultados. O capítulo 5 se trata da metodologia experimental do trabalho, descrevendo as bases de dados que foram utilizadas. O capítulo 6 se refere aos resultados obtidos e as discussões que foram levantadas com os resultados do artigo. O capítulo 7 contém a conclusão do artigo e trabalhos futuros.

2 Revisão Bibliográfica

2.1 Trabalhos Relacionados

Neste capítulo são discutidos alguns métodos que são utilizados para tratar problemas que apresentam caudas pesadas.

A distribuição de cauda pesada pode ser tratada com diversas estratégias para a modelagem de problemas de regressão. Uma dessas estratégias (FELIX, 2022) é elaborada para a predição do tempo de manutenção de vagões de trens utilizando uma abordagem de separação de *inliers* e *outliers*, gerando dois modelos de regressão independentes. Utilizando esses dois modelos e um algoritmo de detecção de *outliers*, foi notada uma melhora na acurácia na predição nos tempos de manutenção dos vagões em comparação com a técnica padrão na indústria, a cronoanálise.

Uma abordagem utilizando cópulas para modelar dados longitudinais com cauda pesada foi criada em (SUN; FREES; ROSENBERG, 2008). Uma cópula é uma distribuição multivariada com distribuição marginal uniforme no intervalo (0, 1). Essa abordagem utiliza as cópulas para modelar as dependências de dados e a distribuição de cauda pesada para modelar as margens, possibilitando a modelagem de dados que tenham caudas pesadas positivas ou negativas.

Bourguignon, Santos-Neto e Castro (2021) desenvolveram um modelo de regressão onde é possível modelar caudas pesadas e assimétricas, mas com a desvantagem de que a variável de interesse tem de estar restringida a um intervalo de valores reais positivos, e se relacionar com as outras variáveis por meio de parâmetros de média e precisão. Uma vantagem deste modelo é sua flexibilidade ao ser utilizado com dados com alta assimetria em sua distribuição, sendo uma boa alternativa a ser considerada em comparação com os modelos já existentes.

Takeuchi, Bengio e Kanamori (2002) propõem um método de regressão linear e não linear robusto que pode ser aplicado em distribuições com caudas pesadas, o *Robust Regression for Asymmetric Tails* (RRAT), onde ela pode ser aplicada em diversos problemas desse tipo de distribuição, como na área de seguros. O método é comparado com o *Least Squares Regression* (LS) com uma análise assintótica e com bases de dados artificiais. Na análise assintótica, o RRAT se provou melhor do que o LS quando o grau de assimetria da distribuição é elevado, e testes com dados artificiais e dados de seguros apontam que o método proposto tem maior desempenho que o *Least Squares Regression*.

3 Referencial Teórico

Neste capítulo são apresentados os conceitos por trás dos métodos e algoritmos selecionados para este trabalho.

3.1 Estatística e Probabilidade

Esta seção apresenta conceitos de estatística e probabilidade que serão utilizados no trabalho.

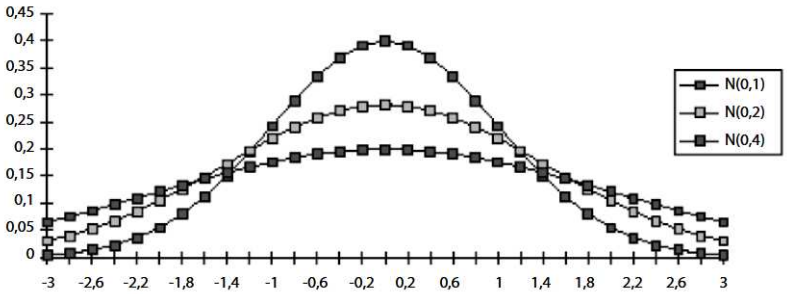
3.1.1 Distribuição Normal

Segundo o livro de Silva, Fernandes e Almeida (2015), uma distribuição normal com média u, desvio padrão σ e variância σ^2 tem as propriedades a seguir:

- $f(x)$ integra a 1;
- Os limites de $f(x)$ quando x tende a $+\infty$ e $-\infty$, são iguais a zero;
- $f(x)^3 \geq 0$ sempre;
- A densidade $N(u, \sigma^2)$ é simétrica em torno de u, ou seja, $f(u + x) = f(u - x)$;
- O valor máximo de $f(x)$ ocorre em $x = u$.
- Os pontos de inflexão de $f(x)$ são $x = u + \sigma$ e $x = u - \sigma$.

É possível observar na Figura 3.1 as características visuais de três distribuições normais, com sua variância igual a 1, 2 e 4 respectivamente e com média zero. O eixo vertical representa a probabilidade de ocorrência e o eixo horizontal representa os valores que a função pode assumir.

Figura 3.1 – Distribuições normais, com variância 1, 2 e 4.

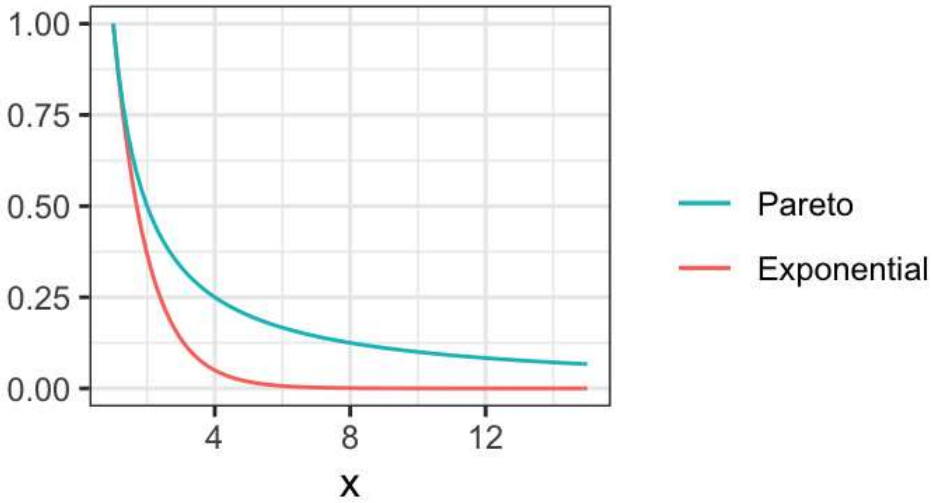


Fonte: (SILVA; FERNANDES; ALMEIDA, 2015).

3.1.2 Distribuição com Cauda Pesada

Segundo os artigos de Sigman (1999) e Goldie e Klüppelberg (1998), uma distribuição de cauda pesada pertence à um grupo de distribuições denominado distribuições sub-exponenciais. Este grupo de distribuições é caracterizado por possuírem caudas que demonstram uma probabilidade de ocorrência significantemente maiores que qualquer distribuição exponencial. As distribuições de cauda pesada tem como sua maior representante a distribuição de Pareto, demonstrada na Figura 3.2 em azul, onde é possível observar a diferença de frequência de valores nas caudas de ambas as distribuições, e como a distribuição de Pareto apresenta uma cauda mais pesada em comparação com uma distribuição exponencial, em vermelho.

Figura 3.2 – Comparação entre distribuições de Pareto e Exponencial.



Fonte: Vincenzo Coia and Michael Gelbart.

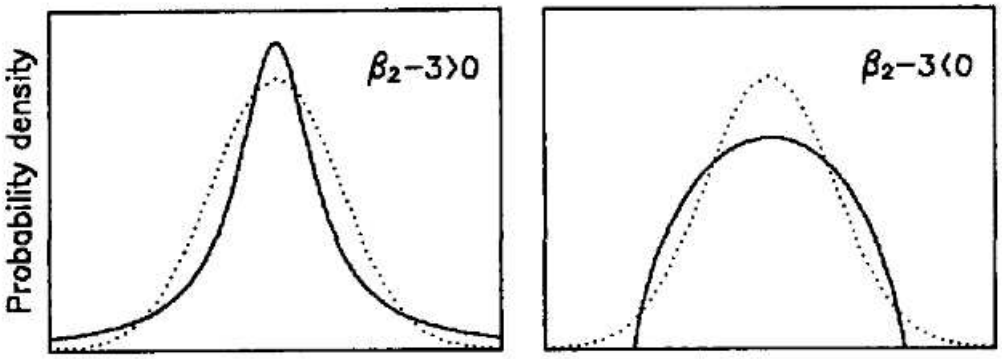
3.1.3 Curtose

DeCarlo (1997) define a curtose como um índice de 'peso' das caudas de uma distribuição em comparação com a distribuição normal. Distribuições com curtose positiva, indicam caudas mais pesadas em comparação com uma distribuição normal e um ápice mais 'pontudo' (*peakedness*), mas distribuições com curtose negativa, indicam caudas mais leves e uma ápice mais 'plano' (*flatness*). Um exemplo prático pode ser visto na Figura 3.3.

Existem classificações da curtose, estas sendo:

- Mesocúrtica (curtose = 0): A distribuição tem uma curtose próxima de zero, o que indica que suas caudas e a região central se assemelham à curva normal.
- Leptocúrtica (curtose > 0): Uma distribuição com curtose positiva tem uma concentração maior de valores na região central, o que significa que as caudas da distribuição são mais pesadas do que a curva normal.

Figura 3.3 – Ilustração de curtose, representada pela expressão $\beta_2 - 3$. As linhas pontilhadas denotam uma distribuição normal, e as linhas contínuas denotam uma distribuição com curtose positiva (esquerda) e negativa (direita).



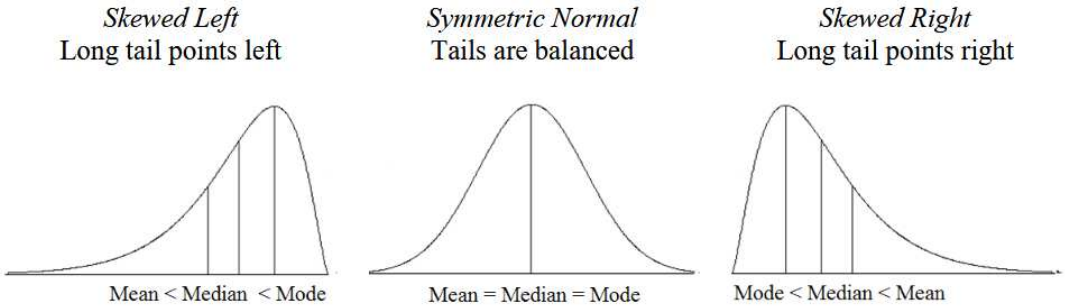
Fonte: (DECARLO, 1997).

- Platicúrtica (curtose < 0): Uma distribuição com curtose negativa tem uma dispersão maior de valores em relação à média, resultando em caudas mais leves e uma região central menos concentrada.

3.1.4 Assimetria

Doane e Seward (2011) discute sobre a assimetria de uma distribuição, definindo assimetria como a falta de simetria em uma distribuição qualquer. Enquanto uma distribuição simétrica tem uma média, mediana e moda alinhadas, uma distribuição assimétrica é tem a tendência de se inclinar para um lado, dependendo do valor do índice de assimetria. A Figura 3.4 demonstra as características visuais de uma distribuição assimétrica.

Figura 3.4 – Ilustração demonstrando a posição da média, mediana e moda de uma distribuição com assimetria negativa (esquerda), nula (meio) e positiva (direita).



Fonte: (DOANE; SEWARD, 2011).

Existem classificações da assimetria, estas sendo:

- Assimetria positiva (à direita, maior que 1): A cauda da distribuição se estende mais para

a direita (valores maiores) em relação à média. A mediana e a moda geralmente estão à esquerda da média;

- Assimetria próxima de zero (entre 1 e -1): A distribuição tem uma simetria razoável, mas não necessariamente perfeita;
- Assimetria negativa (à esquerda, menor que -1): A cauda da distribuição se estende mais para a esquerda (valores menores) em relação à média. A mediana e a moda geralmente estão à direita da média.

3.1.5 Percentis

De acordo com o livro de Silva, Fernandes e Almeida (2015), os percentis são medidas estatísticas que dividem um conjunto de dados ordenados em 100 partes iguais. Eles são usados para avaliar a posição relativa de um valor dentro de um conjunto de dados, indicando a porcentagem de valores que estão abaixo desse valor específico.

3.1.6 Erro Percentual Absoluto Médio

O Erro Percentual Absoluto Médio (em inglês, *Mean Absolute Percentage Error* - MAPE) é uma medida estatística utilizada para avaliar a eficácia de previsões. Segundo o artigo de Kim e Kim (2016) O MAPE é expresso como uma porcentagem e fornece uma indicação da magnitude média dos erros percentuais entre as previsões e os valores reais, Um MAPE baixo significa que as previsões estão próximas dos valores observados, enquanto um valor alto indica que as previsões estão longe dos valores observados. O cálculo do MAPE pode ser expressado da seguinte equação:

$$MAPE = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{A_t - F_t}{A_t} \right|$$

Onde N é o número de dados, A_t é o valor atual do dado e F_t é o valor da predição do modelo de aprendizado de máquina.

3.2 Aprendizado de Máquina Supervisionado

O aprendizado de máquina supervisionado tem como objetivo capacitar sistemas computacionais a aprender a partir de dados de entrada, a fim de realizar tarefas específicas com acurácia e autonomia. Este método de aprendizado é definido por sua abordagem guiada por exemplos previamente rotulados, onde um algoritmo é treinado com um conjunto de dados de treinamento contendo pares de entrada e saída desejada. A partir desse treinamento, a máquina pode generalizar seu conhecimento para fazer previsões ou tomar decisões em dados não vistos anteriormente.

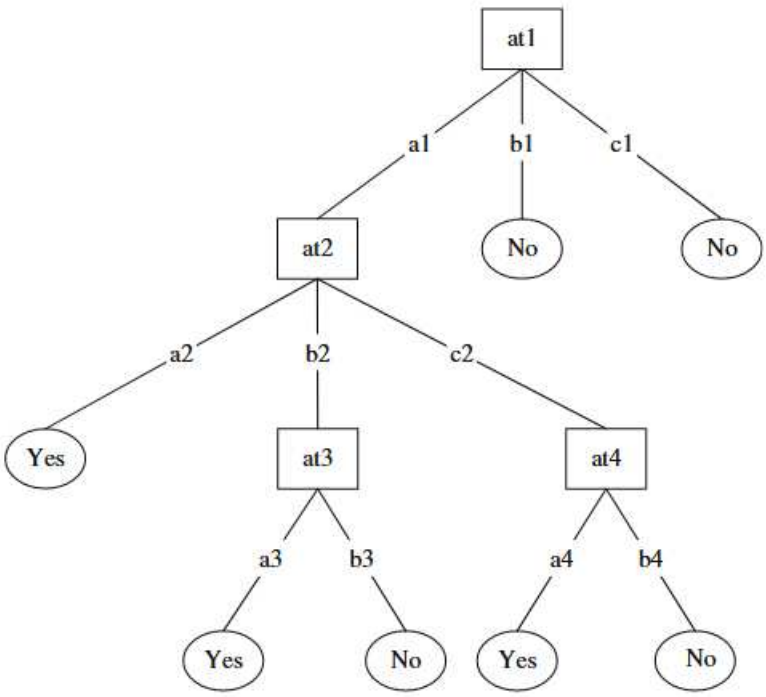
3.2.1 Decision Tree

Segundo Kotsiantis (2013), uma *decision tree*, ou árvore de decisão, é um tipo de modelo sequencial que utiliza uma série de testes simples de forma lógica. Cada teste envolve a comparação de um atributo numérico com um valor limite ou de um atributo nominal com um conjunto de valores possíveis. O algoritmo tenta encontrar padrões nos dados de entrada e tenta dividir as instâncias em classes, com base nos testes que o algoritmo formula, dando a forma de árvore. Diferentemente de modelos complexos como redes neurais, as *decision trees* são classificadores simbólicos e se destacam por sua compreensibilidade.

Além disso, elas podem ser utilizadas para identificar as características mais importantes para a previsão, o que é útil para a seleção de recursos. No entanto, as *decision trees* tendem a ser instáveis e sensíveis a pequenas variações nos dados, o que pode levar a *overfitting*.

Um exemplo de *decision tree* pode ser observado na Figura 3.5:

Figura 3.5 – Uma árvore de decisão.



Fonte: (KOTSIANTIS, 2013).

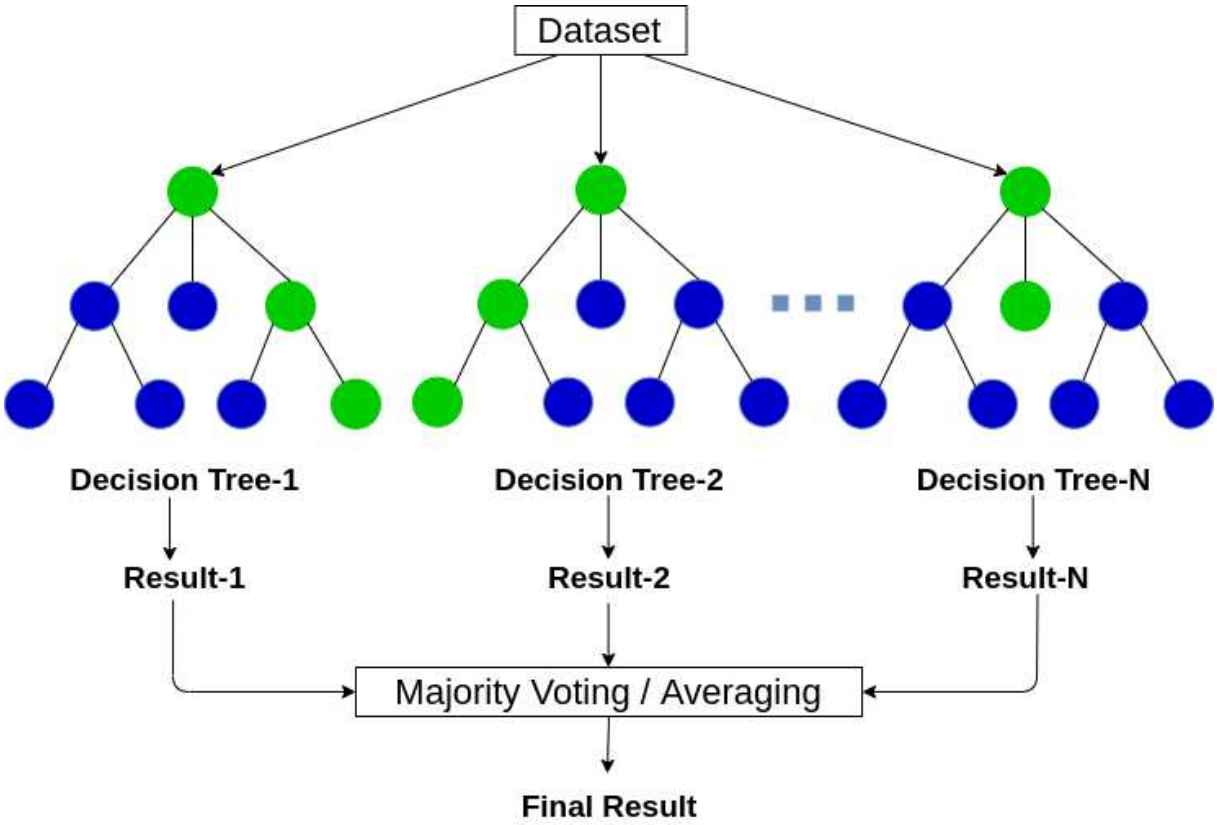
Na *decision tree* da Figura 3.5, é possível observar os nós $at1$, $at2$, $at3$ e $at4$ representados

pelas caixas na Figura 3.5 que representam os atributos dos dados que serão testados, e os valores nos nós folhas representados por elipses, representam os testes feitos nos valores dos atributos. Ao se realizar todos os testes e um nó folha for encontrado, a instância será classificada com o valor que o nó contém.

3.2.2 Random Forest

Random Forest é um algoritmo de aprendizado de máquina que combina várias *decision trees* em um modelo preditivo robusto e preciso. Cada árvore depende de valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores na floresta (BREIMAN, 2001). Uma das principais vantagens da *Random Forest* é que ela é capaz de lidar com dados de alta dimensionalidade e com uma grande quantidade de amostras de dados, enquanto mantém uma boa precisão preditiva. Ela também é resistente a *overfitting*, pois utiliza a média ou votação majoritária de várias árvores de decisão em vez de depender de uma única árvore.

Figura 3.6 – Esquemática de uma *Random Forest*.



Fonte: Analytics Vihdya.

A Figura 3.6 demonstra como um algoritmo de *random forest* funciona. São criadas diversas *decision trees*, cujo número é definido pelo usuário, e cada uma dessas árvores gera o seu próprio resultado. Os resultados então são computados para gerar o resultado final da *random*

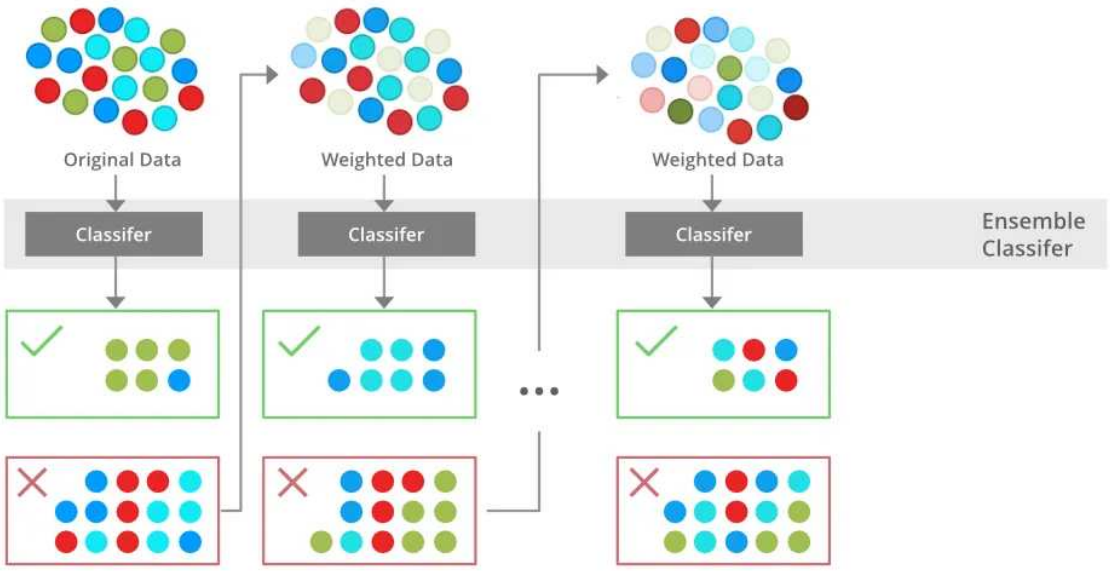
forest, que pode ser decidido em ser a média dos resultados das *decision trees* ou por voto da maioria, onde o resultado mais frequente nas *decision trees* é escolhido para ser o resultado final.

3.2.3 XGBoost

O algoritmo XGBoost (*Extreme Gradient Boosting*) criado por Chen e Guestrin (2016), é uma biblioteca de código aberto de *gradient boosting* otimizada, usada para desenvolver modelos de aprendizado de máquina em tarefas como classificação, regressão e ranking, sendo muito utilizado por conta de sua escalabilidade em diversos cenários diferentes. Este algoritmo usa um técnica denominada '*gradient boosting*', o que significa que no momento de combinar os modelos gerados, é utilizado um algoritmo de descida de gradiente para minimizar as perdas.

A Figura 3.7 ilustra o funcionamento do algoritmo:

Figura 3.7 – Funcionamento do algoritmo XGBoost.



Fonte: Medium.

O XGBoost funciona treinando uma quantidade de *decision trees*, onde cada árvore é treinada num subconjunto de dados da base de dados original, gerando vários modelos 'fracos' e as predições de cada árvore são combinadas para gerar um modelo 'forte' e a previsão final mais precisa.

3.2.4 *Overfitting*

O overfitting em aprendizado de máquina, é um fenômeno crítico que ocorre quando um modelo se adapta excessivamente aos dados de treinamento, resultando em um modelo que se

torna específico para predizer o conjunto de treinamento e que não consegue realizar predições com precisão para novos dados.

3.3 Otimização

Esta seção descreve conceitos e algoritmos de otimização.

3.3.1 Conceitos

3.3.1.1 Função Objetivo

A função objetivo é uma expressão matemática que quantifica a medida de desempenho ou qualidade de uma solução candidata em um problema de otimização. Ela representa o critério do problema a ser maximizado ou minimizado.

3.3.1.2 Solução Otimizada

Uma solução otimizada é um conjunto de valores que satisfazem todas as restrições do problema e atendem ao critério definido pela função objetivo.

3.3.1.3 Ótimo Global

O ótimo global é o valor mais alto ou mais baixo que a função objetivo pode atingir em todo o espaço de busca, que inclui todas as possíveis soluções. O ótimo global é o valor mais alto se o problema é de maximização e é o valor mais baixo se é um problema de minimização.

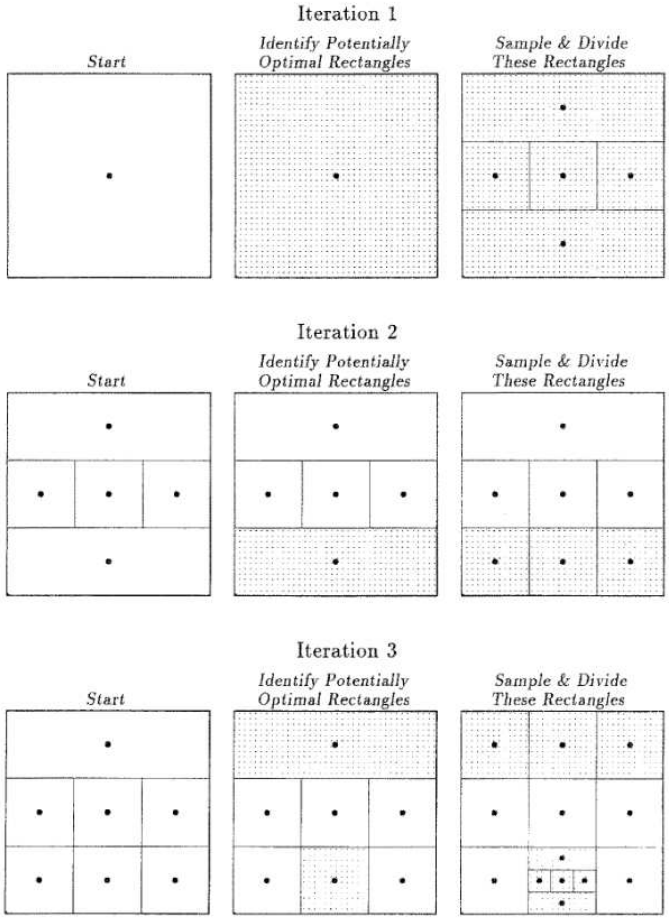
3.3.2 Força Bruta

O algoritmo de otimização *Brute Force*, ou força bruta, é uma abordagem simples e direta para resolver problemas de otimização. Ele consiste em testar todas as possíveis soluções em um espaço de busca para encontrar a solução ótima. Esse método é conhecido por sua abordagem exaustiva, pois avalia todas as combinações possíveis. Embora seja eficaz para problemas de tamanho pequeno, o tempo de execução do algoritmo cresce exponencialmente à medida que o tamanho do problema aumenta.

3.3.3 *Direct*

De acordo com o artigo de Jones, Perttunen e Stuckman (1993) O algoritmo de *Dividing Rectangles*, também conhecido como *Direct*, é um método iterativo para resolver problemas de otimização. Este método foi criado como uma alternativa à otimização Lipschitziana sem expressar a constante de *Lipschitz*, uma constante que descreve a taxa máxima de variação de uma função em relação à distância entre os pontos no domínio da função.

Figura 3.8 – Funcionamento do algoritmo *Direct* utilizando a função de Branin em suas três primeiras iterações.



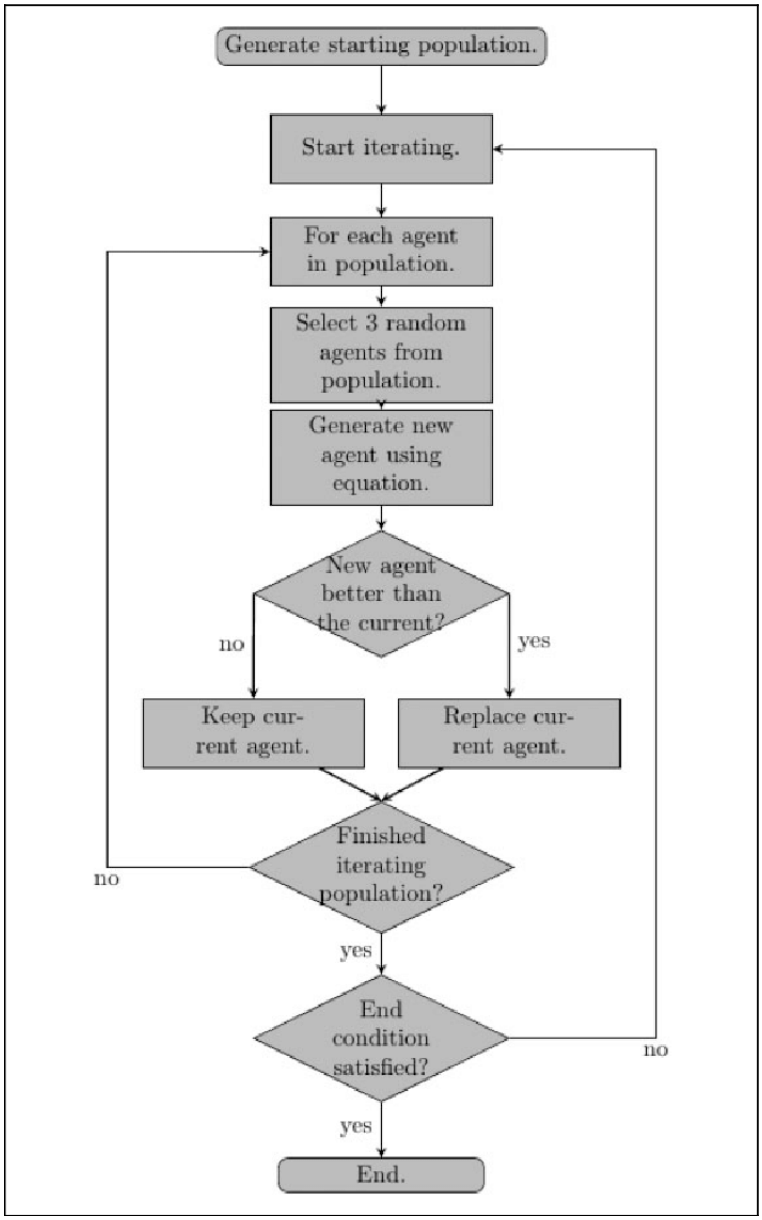
Fonte: (JONES; PERTTUNEN; STUCKMAN, 1993).

Com base na Figura 3.8, a primeira coluna representa o espaço de busca no início de cada iteração. O algoritmo *Direct* divide o espaço de busca em retângulos menores e avalia a função objetivo em cada um desses retângulos, de acordo com a segunda coluna da Figura 3.8, na área destacada. Com base nos resultados, o algoritmo identifica a área que possui maior potencial de conter a solução ótima e continua a dividir essa área em retângulos ainda menores, como visto na terceira coluna. Esse processo é repetido até que uma solução otimizada seja encontrada ou um número máximo de iterações é atingido.

3.3.4 *Differential Evolution*

O algoritmo *Differential Evolution*, ou Evolução Diferencial, é uma técnica de otimização baseada em populações criada por Storn e Price (1997). Ele simula o processo de evolução natural para buscar soluções otimizadas em um espaço de busca multidimensional. O algoritmo utiliza uma população de indivíduos que representam soluções candidatas. Cada indivíduo é avaliado de acordo com uma função objetivo e, em cada iteração, são realizadas operações de mutação, recombinação e seleção para gerar novas soluções.

Figura 3.9 – Fluxograma demonstrando o funcionamento do algoritmo *Differential Evolution*.



Fonte: ResearchGate.

O *Differential Evolution* é executado de acordo com a Figura 3.9, onde é gerado uma população inicial aleatoriamente e são gerados novos vetores de parâmetros ao adicionar um vetor de diferença ponderada entre dois indivíduos da população a um terceiro membro. Se o vetor resultante produzir um valor de função objetivo mais baixo do que o de um membro predefinido da população, o vetor recém-gerado substituirá o vetor com o qual foi comparado na geração seguinte. O algoritmo continua a ser executado até que se encontre o mínimo global da função objetivo.

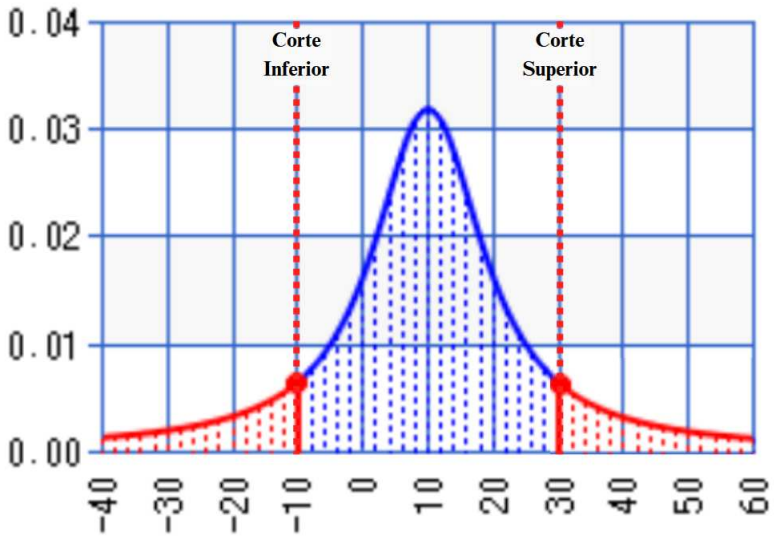
4 *Cut the Tails*

4.1 Os Cortes

No artigo de Felix (2022), é utilizado um método de separação simples utilizando quartis, que dividem o conjunto ordenado de dados em quatro partes iguais de 25% (vinte e cinco porcento) cada. Para realizar a separação de *inliers* e *outliers* são calculados os quartis e intervalo interquartil da variável objetivo para separar os *outliers*.

Neste trabalho, em vez de quartis, serão selecionados percentis, que dividem o conjunto de dados em 100 (cem) partes. Serão selecionados uma dupla de percentis referentes à distribuição da variável alvo. Os percentis selecionados serão denominados 'Corte Inferior' e 'Corte Superior' e tem finalidade de separar os dados do conjunto em diferentes classes. A Figura 4.1 exemplifica o uso dos cortes em uma distribuição:

Figura 4.1 – Corte das caudas da distribuição de Cauchy



Fonte: De autoria própria.

Na Figura 4.1 é possível observar a separação da distribuição ao se aplicar os cortes inferior e superior. Todos os dados que se situam abaixo do corte inferior são classificados como pertencentes à 'cauda inferior', e todos os dados que se situam acima do corte superior são classificados como pertencentes à 'cauda superior'. Os dados que se situam entre os cortes são classificados como pertencentes ao 'pico' da distribuição.

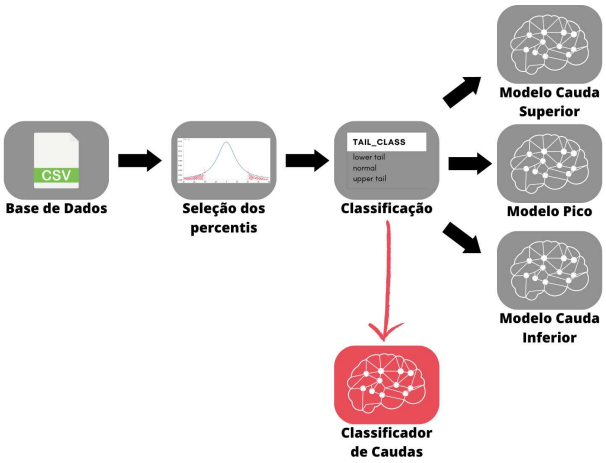
4.2 Aprendizado de máquina com o *Cut the Tails*

Para se fazer um modelo de aprendizado de máquina com a técnica *Cut the Tails*, este procedimento requer que o conjunto de dados de treinamento sejam previamente classificados de acordo com os cortes. Este conjunto de dados será dividido para o treinamento do respectivo modelo, gerando até três modelos independentes.

Para construir os modelos para a predição, é necessário seguir os seguintes passos, que também são ilustrados na Figura 4.2:

1. Dada uma base de dados de treinamento t, classificada de acordo com os cortes;
2. Separar a base t em uma base de treinamento da cauda superior t_s, da cauda inferior t_i e do pico t_p, seguindo a classificação realizada na fase de corte das caudas;
3. Construir o modelo do pico M_p, a partir de t_m.
4. Construir o modelo da cauda superior M_s, a partir de t_s.
5. Construir o modelo da cauda inferior M_i, a partir de t_i.

Figura 4.2 – Metodologia do *Cut the Tails*

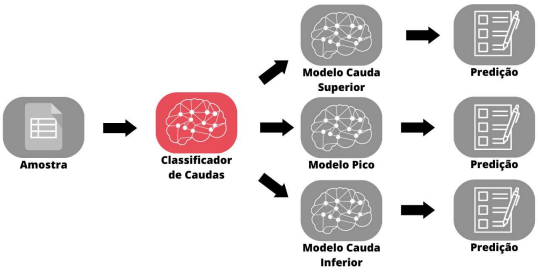


Fonte: De autoria própria.

4.3 Classificador de caudas

Um classificador é treinado com o objetivo de identificar a classificação de cada amostra no conjunto de teste do experimento. Segundo a Figura 4.3, o classificador pode denominar a amostra como pertencente à cauda inferior, pico, ou cauda superior, para que no momento do teste dos modelos treinados, sejam utilizados as amostras correspondentes ao tipo de cauda que ela pertence.

Figura 4.3 – Funcionamento do classificador de caudas



Fonte: De autoria própria.

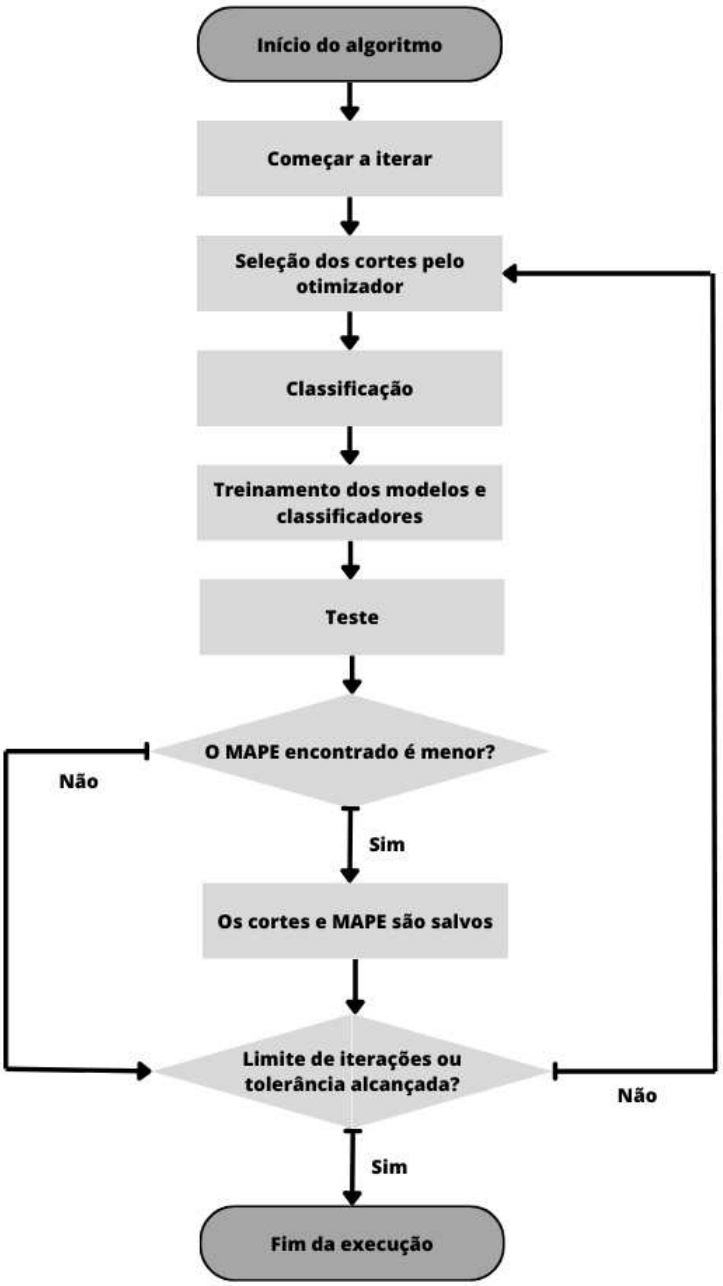
4.4 Otimização dos Cortes

Para selecionar os percentis referentes aos cortes superior e inferior das caudas de uma maneira não arbitrária, foram utilizados algoritmos de otimização global para encontrar os cortes da cauda que remetem ao menor erro possível no teste dos modelos treinados. Os algoritmos de otimização global utilizados para este trabalho tem a função de explorar as todas as possibilidades de cortes na distribuição da variável-alvo do base de dados, retornando os percentis respectivos para o corte inferior e superior que tiveram menor MAPE.

A função objetivo dos otimizadores pode ser expressa na seguinte equação:

$$Z = minimize(MAPE)$$

Figura 4.4 – Fluxograma do funcionamento dos otimizadores.



Fonte: De autoria própria.

Observando a Figura 4.4, é possível ver os passos tomados pelo otimizador em sua execução para medir o MAPE de cada corte selecionado, onde apenas os cortes que gerarem o menor mape no processo serão retornados pelo algoritmo. O processo tem um limite de iterações e tolerância, onde se o algoritmo ultrapassar o máximo de iterações ou a diferença dos MAPE's coletados se mostrarem inferiores a uma certa precisão, o algoritmo será finalizado.

Foram testados três otimizadores globais da biblioteca *SciPy* para selecionar os cortes e analisar o MAPE de cada base de dados com seus respectivos parâmetros:

- Força Bruta;
  - $fulloutput = True$
  - $finish = None$
- Direct;
  - $tol = 1e - 3$
- Evolução Diferencial.
  - $eps = 1e - 3$

Os otimizadores foram implementados com base nas seguintes restrições:

- Intervalo da solução de corte inferior/superior (*Bounds*): Entre 0.0 e 1.0;
- Tolerância relativa: $1e - 3$

Para definir qual otimizador mais se adequa aos requisitos deste trabalho, um *script* foi configurado para executar os respectivos otimizadores com as dez bases de dados, com o classificador e regressor definidos como *Random Forest*, visto que este algoritmo gerou os melhores resultados entre os outros. Para verificar se os cortes e erros são consistentes, cada base foi processada com um mesmo otimizador cinco vezes usando cada otimizador e o método *baseline*. Os resultados deste experimento são ilustrados na Figura 6.2, onde cada caixa do *boxplot* representa a distribuição do resultado das cinco execuções de cada otimizador (e *baseline*) em sua respectiva base.

4.5 O Algoritmo *Cut the Tails*

O algoritmo do método proposto nesta monografia será executado da seguinte maneira:

Os experimentos são realizados particionando as bases de dados em conjuntos de treinamento e teste, onde 80% dos dados são dedicados para o treinamento do modelo, e os 20% restantes são dedicados ao teste do modelo treinado.

Considere:

- b representa uma base de dados;
- d é um dado/instância de b;
- x é o percentil referente ao valor da variável alvo de d na distribuição do histograma da variável alvo;
- C_i e C_s são os cortes inferior e superior, respectivamente.

Algoritmo *Cut the Tails*:

1. Dada uma base de dados b;
2. Selecionar os cortes C_i e C_s da distribuição da variável alvo de b com um algoritmo de otimização global.
3. Classificar as instâncias da base b:
   a) Se $x \leq C_i$, d será classificado como pertencente à cauda inferior;
   b) Se $x > C_i$ e $x < C_s$, d será classificado como pertencente ao pico;
   c) Se $x \geq C_s$, d será classificado como pertencente à cauda superior.
4. A base b é partilhada para o treinamento e teste dos modelos e classificador;
5. Os modelos de aprendizado de máquina das caudas M_i e M_s, do pico M_p e do classificador são treinados.
6. Os dados destinados a teste são classificados pelo classificador.
7. Os dados de teste classificados são direcionados ao seu respectivo modelo e suas variáveis alvo são preditas.

5 Metodologia Experimental

Para que o método *Cut the Tails* seja testado, serão utilizados os regressores e classificadores, com seus respectivos parâmetros para este experimento:

- *Random Forest*:
  - $maxdepth = 5$.
- *Decision Tree*:
  - $maxdepth = 5$.
- *XGBoost*:
  - $nestimators = 100$;
  - $maxleaves = 0$.

A comparação será feita utilizando um modelo $baseline$, que será construído sem as estratégias do método *Cut the Tails*, e um modelo que utiliza a estratégia supracitada. Para esta parte do trabalho, a linguagem utilizada para os testes é *Python* no *Jupyter Notebook*, utilizando a biblioteca *Scikit-learn* para usar os regressores e classificadores. A métrica utilizada para comparar os modelos e abordagens é o Erro Absoluto Médio Percentual (MAPE), onde serão medidos o Erro Absoluto Médio Percentual de cada execução de algoritmo, e serão comparados os erros para cada base de dados descrita.

Após esta comparação de modelos, o modelo que apresentar melhor resultado será utilizado com os algoritmos de otimização global para encontrar os cortes ótimos de cada base. Após os cortes, será comparado o MAPE do método *baseline* com cada um dos otimizadores para determinar se houve uma melhoria significante entre eles.

5.1 Bases de dados

Neste capítulo serão descritas e analisadas as bases de dados que serão utilizadas neste trabalho.

As bases apresentam diversos comportamentos em suas distribuições, que podem ser vistos no histograma da variável alvo de cada base. Dez bases foram escolhidas com diversas características em suas distribuições e índices para que sejam executadas pelo algoritmo do *Cut the Tails* e analisadas.

As métricas de curtose e assimetria serão utilizadas para identificar as bases com caudas pesadas e exibidas na Tabela 5.1.

Tabela 5.1 – Tabela de índices de Curtose e Assimetria das bases de dados

Base	Curtose	Assimetria
Blueberry Yield	-0.382	-0.321
Employee Performance	0.413	-0.828
Financial Distress	1449.229	30.860
Car Price	20.620	4.156
Real Estate	2.138	0.597
Wind Speed	0.211	0.645
Bike Sharing	1.416	1.277
Solar Radiation	0.510	1.369
House Rent	840.220	21.403
Medical Cost	1.595	1.514

Fonte: De autoria própria.

5.2 *Blueberry Yield* - Base de Predição de Produção de Mirtilos

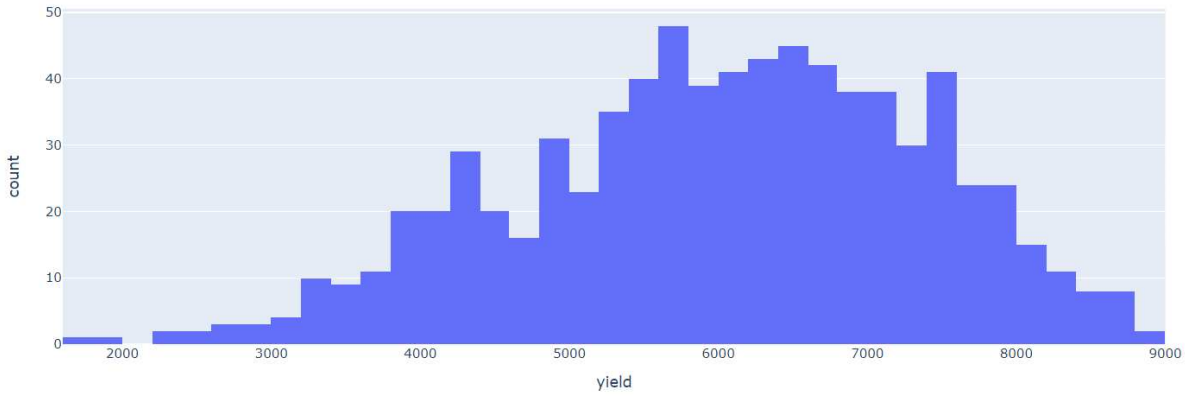
5.2.1 Descrição da Base de Dados

O setor agricultural tem investido em aprendizado de máquina para predizer de maneira mais precisa a produção das colheitas, e quais fatores mais influenciam em maior produção. Esta base de dados é gerado pelo *Wild Blueberry Pollination Model*, uma simulação construída a partir de dados coletados do estado de Maine nos Estados Unidos, dos últimos 30 anos (OBSIE; QU; DRUMMOND, 2020). Os atributos da base relevantes para este trabalho são:

1. *Clonesize*: Tamanho médio do arbusto de mirtilo, em metros quadrados;
2. *Honeybee*: Densidade de abelhas no campo, em abelhas por metro quadrado por minuto;
3. *Bumblebee*: Densidade de abelhões no campo, em abelhões por metro quadrado por minuto;
4. *Andrena*: Densidade de abelhas da famíla Andrena no campo, em abelhas por metro quadrado por minuto;
5. *Osmia*: Densidade de abelhas da famíla Osmia no campo, em abelhas por metro quadrado por minuto;
6. *RainingDays*: Número total de dias que ocorreu precipitação, na temporada de floração;
7. *AverageRainingDays*: Média dos dias chuvosos na temporada inteira;
8. *Yield* (Alvo): Quantidade de produção de mirtilos da temporada;

5.2.2 Comportamento da Variável *Yield*

Figura 5.1 – Histograma ilustrando a distribuição dos valores da variável *Yield*.



Fonte: De autoria própria.

Visualizando o comportamento da variável *Yield* na Figura 5.1 e na Tabela 5.1, observa-se a presença de uma cauda à esquerda, e a presença de assimetria com o valor de -0.321 e curtose com -0.382, ambos os valores negativos próximos de zero.

5.3 *Employee Performance* - Base de Predição de Performance de Empregados

5.3.1 Descrição da Base de Dados

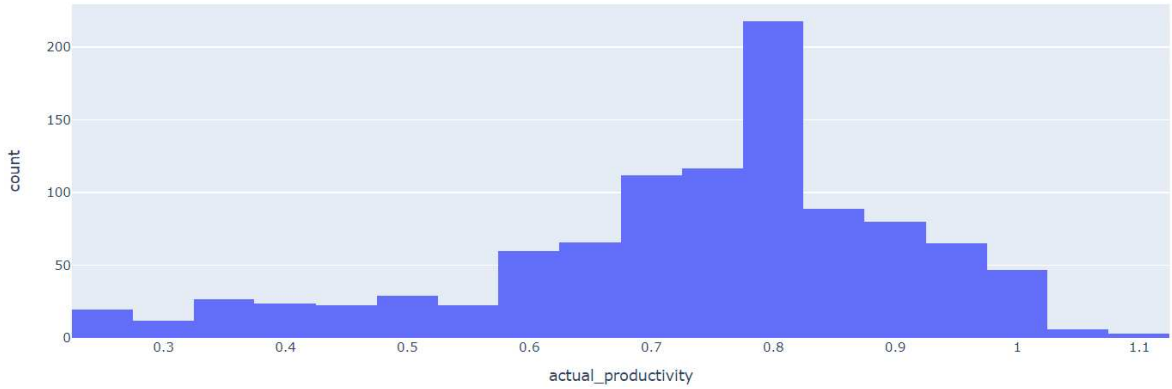
As indústria têxtil necessita de muita mão de obra para saciar a demanda de roupas e tecidos no mundo. Um problema enfrentado por essa indústria é a produtividade dos empregados. Por ser uma atividade laboriosa, muitos empregados não alcançam as metas definidas, resultando em uma perda de produção. Esta base apresenta dados de empregados no processo de manufatura de uma empresa de vestuário que foram validados por especialistas da indústria, com o objetivo de prever a produtividade dos empregados (DUTTA, 2022). Os atributos da base relevantes para este trabalho são:

1. *Day*: Dia da semana;
2. *Quarter*: Uma porção do mês, um mês é dividido em quatro partes;
3. *Teamno*: Número do time associado com a instância;
4. *Noofworkers*: Número de empregados em cada time;
5. *Noofstylechange*: Número de mudanças de estilo de um produto;

6. *smv*: Tempo alocado para uma tarefa;
7. *overtime*: Número de horas extra de um time, em minutos;
8. *Targetedproductivity*: Produtividade esperada do time;
9. *Actualproductivity* (Alvo): Produtividade real do time, dada em um percentual da variável *Targetedproductivity*.

5.3.2 Comportamento da Variável *Actualproductivity*

Figura 5.2 – Histograma ilustrando a distribuição dos valores da variável *Actualproductivity*.



Fonte: De autoria própria.

Analisando o histograma na Figura 5.2 e a Tabela 5.1, existe uma cauda prolongada à esquerda, justificada pelo índice de assimetria em -0.828, sendo negativa e próximo a -1.0 e curtose em 0.413.

5.4 *Financial Distress* - Base de Predição de Crise Financeira

5.4.1 Descrição da Base de Dados

Grandes e pequenas empresas podem ser afetados por uma crise financeira, devido a diversos fatores econômicos, a predição de uma crise em uma empresa pode ajudar na prevenção do pior caso. Esta base de dados tem o objetivo de predizer a crise financeira de uma amostra de empresas com base em períodos de tempo e características financeiras e não-financeiras (EBRAHIMI, 2018). Os atributos relevantes para esta base de dados são:

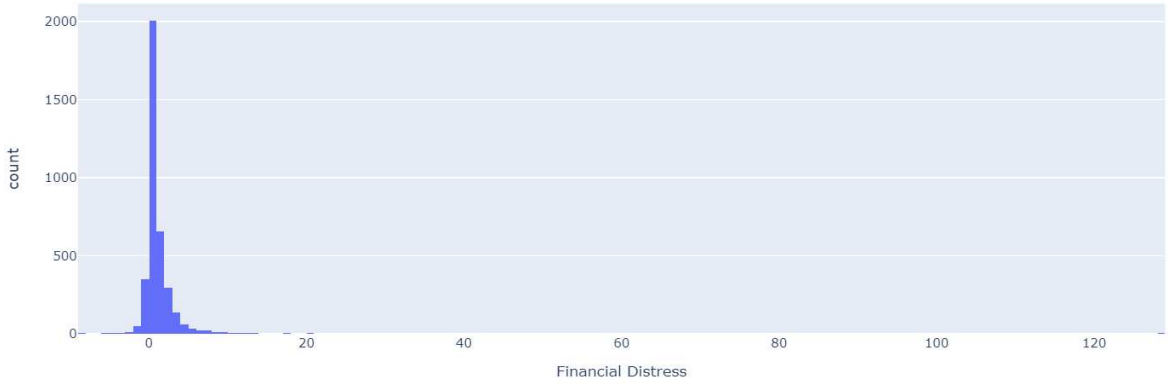
1. *Time*: Período de tempo em que os dados foram retirados;
2. *Financial Distress* (Alvo): Variável que denomina a situação de crise financeira da empresa, se maior que -0,5. a empresa é saudável, ao contrário, a empresa está financeiramente em crise;

3. x1 ... x83: Atributos que denominam as características financeiras e não-financeiras tomadas pela empresa.

5.4.2 Comportamento da Variável *Financial Distress*

Analisando a distribuição da variável *Financial Distress* em um histograma, é possível chegar na seguinte conclusão:

Figura 5.3 – Histograma ilustrando a distribuição dos valores da variável *Financial Distress*.



Fonte: De autoria própria.

Na Figura 5.3 e a Tabela 5.1, a distribuição apresenta valores extremos à direita, gerando índices de assimetria e curtose elevados (1449.229 e 30.860) e positivos e possibilitando a categorização desta distribuição como uma cauda pesada.

5.5 *Car Price* - Base de Predição de Preço de Automóveis

5.5.1 Descrição da Base de Dados

O valor de um automóvel usado pode variar consideravelmente, dependendo de vários fatores, incluindo o número de proprietários anteriores, a quilometragem percorrida, o ano de fabricação e outros. Isso pode tornar desafiador para os proprietários determinar o preço justo do veículo sem a ajuda de um profissional do setor ou de uma corretora (BIRLA; VERMA; KUSHWAHA, 2023). Nesse contexto, a base de dados contém informações sobre carros usados pode ser utilizada para prever o valor de um carro. Os atributos relevantes para este trabalho incluem.

1. *name*: Nome do automóvel;
2. *year*: Ano em que o automóvel foi comprado;
3. *km driven*: Número de quilômetros que o automóvel percorreu;

4. *fuel*: Tipo de gasolina;
5. *seller type*: Tipo de vendedor do automóvel;
6. *transmission*: Tipo de marcha do automóvel;
7. *owner*: Número de donos anteriores do automóvel;
8. *selling price* (Alvo): Preço de venda do automóvel, em rupias indianas;

5.5.2 Comportamento da Variável *Selling Price*

Figura 5.4 – Histograma ilustrando a distribuição dos valores da variável *Selling Price*.

Fonte: De autoria própria.

O histograma da Figura 5.4 e a Tabela 5.1 definem uma distribuição que tem uma cauda definida á direita e ausência de uma cauda à esquerda. A presença de valores extremos dão a esta distribuição índices de curtose e assimetria altos, assumindo valores respectivos de 20.620 e 4.156, classificando esta distribuição como uma cauda pesada.

5.6 *Real Estate* - Base de Predição de Preço de Imobiliária

5.6.1 Descrição da Base de Dados

Esta base de dados apresenta a avaliação do valor de casas de uma imobiliária tailandesa, onde são considerados os fatores externos e localidade da casa, como coordenadas e proximidade a estação de metrô mais próxima (YEH, 2018). Os atributos relevantes para este trabalho são:

1. *House Age*: Idade da casa, em anos;
2. *Distance to the Nearest MRT Station*: Distância da estação de metrô mais próxima, em metros;

3. *Number of Convenience Stores*: Número de lojas de conveniência próximas da casa;
4. *Latitude*: Latitude da localização da casa, em graus;
5. *Longitude*: Longitude da localização da casa, em graus;
6. *House Price* (Alvo): Preço da casa por Ping (onde 1 ping = 3,3 metros quadrados), em milhares dólares tailandeses.

5.6.2 Comportamento da Variável *House Price*

Figura 5.5 – Histograma ilustrando a distribuição dos valores da variável *House Price*.

Fonte: De autoria própria.

O histograma da Figura 5.5 e a Tabela 5.1 definem uma distribuição que tem uma cauda pouco definida à direita, e seus índices de curtose e assimetria assumem os valores 2.138 e 0.597 respectivamente, mostrando uma curtose elevada.

5.7 *Wind Speed* - Base de Predição da Velocidade do Vento

5.7.1 Descrição da Base de Dados

Eventos climáticos extremos como tornados, ventos fortes e tempestades podem ser disruptivos para a vida diária de muitas pessoas e resultar em danos consideráveis. Mesmo que a predição da velocidade do vento ainda é um desafio para meteorologistas do mundo inteiro, ela é vital para que seja possível se preparar ou prevenir futuros danos causados por condições extremas do clima no mundo (FEDESORIANO, 2022).

Esta base de dados apresenta 6574 instâncias de variáveis do clima coletadas de sensores de uma estação meteorológica com o objetivo de prever a velocidade do vento. As principais variáveis de interesse para este trabalho são:

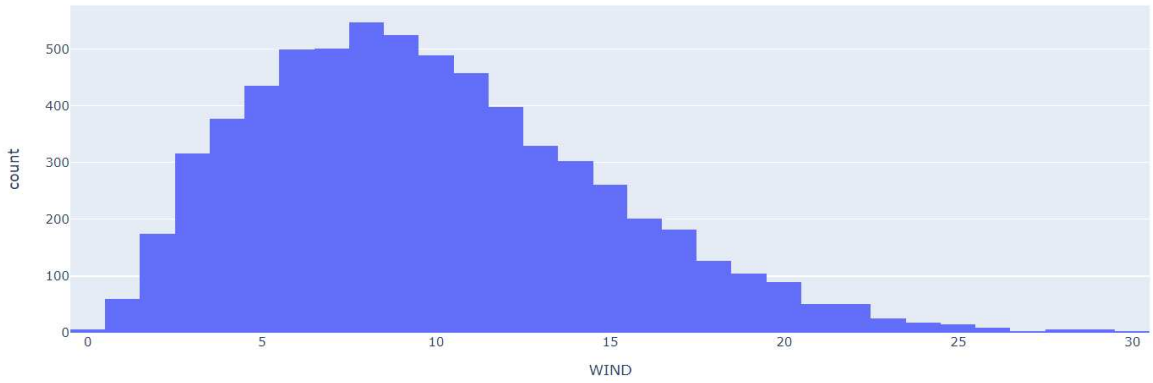
1. *Date*: Data da coleta da informação;

2. *Wind* (Alvo): Velocidade média do vento em nós;
3. *Rain*: Quantidade de precipitação em milímetros;
4. *T.Max*: Temperatura máxima em Celsius;
5. *T.Min*: Temperatura mínima em Celsius.

5.7.2 Comportamento da Variável *Wind*

Ao se analisar a variável *Wind* na base e transferir os respectivos valores para um histograma, é observado o seguinte comportamento.

Figura 5.6 – Histograma ilustrando a distribuição dos valores da variável *Wind*.



Fonte: De autoria própria.

Com a Figura 5.6 e a Tabela 5.1, é possível observar uma pequena assimetria de 0.645 e curtose baixa de 0.211, mesmo tendo uma cauda bem definida à direita.

5.8 *Bike Sharing* - Base de Predição de Sistema de Bicicletas Públicas

5.8.1 Descrição da Base de Dados

O aluguel de bicicletas é um processo que é praticado em diversos países do mundo, onde se é possível alugar uma bicicleta e devolvê-la em lugares diferentes automaticamente. O objetivo da base em análise é prever a quantidade de aluguéis de bicicletas públicas com base em atributos como clima, estação do ano, hora do dia, etc. Para entender em quais momentos existe uma circulação maior de alugueis de bicicleta (PATEL, 2022). Os atributos relevantes para esta base de dados são:

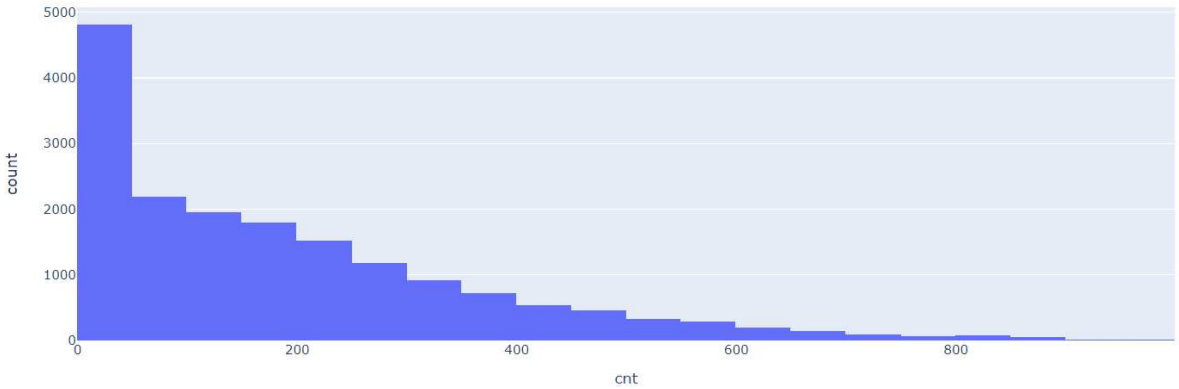
1. *Dteday* : Data de quando os dados foram coletados;

2. *Season*: Estação do ano;
3. *Holiday*: Determina se o dia em questão é ou não um feriado;
4. *Workday*: Determina se o dia é um dia útil;
5. *Weathersit*: Situação do clima;
6. *Temp*: Temperatura normalizada em Celsius;
7. *Hum*: Percentual da umidade normalizada;
8. *Windspeed*: Velocidade do vento, em nós;
9. *Casual*: Quantidade de usuários casuais;
10. *Registered*: Quantidade de usuários registrados;
11. *Cnt* (Alvo): Quantidade de aluguéis totais de bicicletas.

5.8.2 Comportamento da Variável *Cnt*

Analisando a distribuição da variável *Cnt* em um histograma, é possível chegar na seguinte conclusão:

Figura 5.7 – Histograma ilustrando a distribuição dos valores da variável *Cnt*.



Fonte: De autoria própria.

A distribuição apresentada pela variável *Cnt* na Figura 5.7 e a Tabela 5.1 tem a presença de um índices de curtose e assimetria assumindo valores razoáveis de 1.416 e 1.277 respectivamente, mostrando uma única cauda elongada á direita.

5.9 *Solar Radiation* - Base de Predição de Radiação Solar

5.9.1 Descrição da Base de Dados

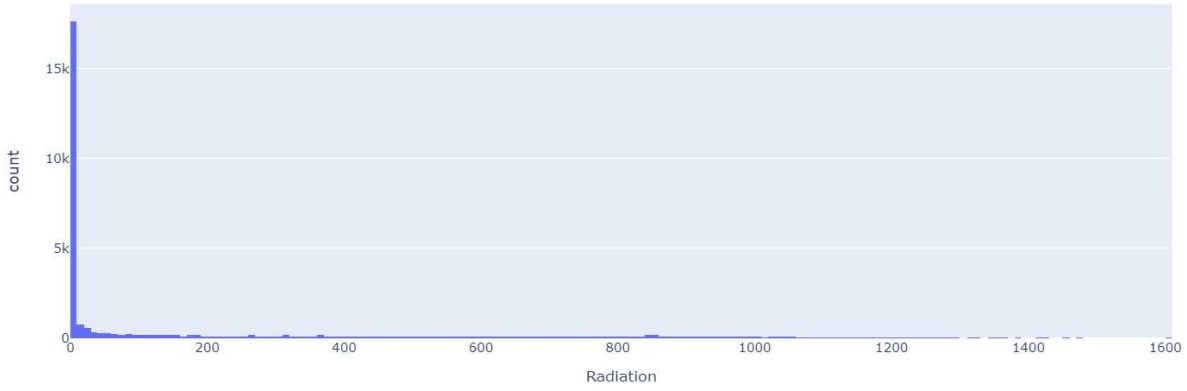
Esta base de dados é um conjunto de dados meteorológicos da estação climática *HI-SEAS* por 4 (quatro) meses, com o objetivo de criar um modelo de predição para prever o nível de radiação solar em função dos dados do clima (ANDREY, 2017). Os atributos relevantes para este trabalho são:

1. *Time*: Horário em que o dado foi recolhido;
2. *Radiation* (Alvo): Quantidade de radiação solar captada pela estação, em watts por metro quadrado;
3. *Temperature*: Temperatura do ambiente, em Fahrenheit;
4. *Pressure*: Pressão atmosférica do ambiente, em polegadas de mercúrio;
5. *Humidity*: Percentual de umidade do ambiente;
6. *Wind Direction*: direção do vento em graus;
7. *Speed*: Velocidade do vento em milhas por hora.

5.9.2 Comportamento da Variável *Radiation*

Ao se analisar a variável *Radiation* na base e transferir os respectivos valores para um histograma, é observado o seguinte comportamento:

Figura 5.8 – Histograma ilustrando a distribuição dos valores da variável *Radiation*.



Fonte: De autoria própria.

Mesmo que a uma quantidade considerável de valores extremos à direita do histograma, na Figura 5.8, esta base apresenta curtose baixa de 0.510 e assimetria alta de 1.369.

5.10 *House Rent* - Base de Predição de Valor de Aluguel de Casas

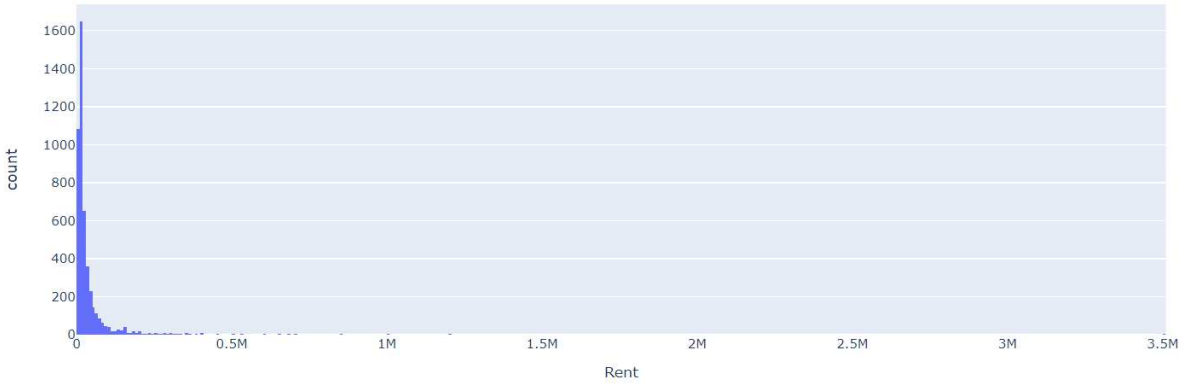
5.10.1 Descrição da Base de Dados

Esta base de dados contém informação de mais de 4700 casas e apartamentos para a predição do valor de aluguel na Índia (BANERJEE, 2022). Os atributos relevantes para a predição desta base de dados são:

1. *BHK*: Número de Quartos, salões e cozinhas do imóvel;
2. *Rent* (Alvo): Valor do aluguel da casa/apartamento, em Rupias Indianas;
3. *Size*: Tamanho do imóvel. em metros quadrados;
4. *Area Type*: Tipo de área que o imóvel se situa;
5. *Location*: Localização do imóvel;
6. *Furnishing status*: Situação da presença de mobília no imóvel;
7. *Bathroom*: Quantidade de Banheiros do imóvel.

5.10.2 Comportamento da Variável *Rent*

Figura 5.9 – Histograma ilustrando a distribuição dos valores da variável *Rent*.



Fonte: De autoria própria.

O comportamento da variável *Rent* ao ser distribuída em um histograma pode ser observada de acordo com a Figura 5.9 e Tabela 5.1. São observados apresenta valores muito distantes de uma distribuição normal, apresentando curtose e assimetria elevadas, com valores 840.220 e 21.403 respectivamente, portanto esta distribuição pode ser considerada uma cauda pesada.

5.11 *Medical Cost* - Base de Predição de Seguro Médico

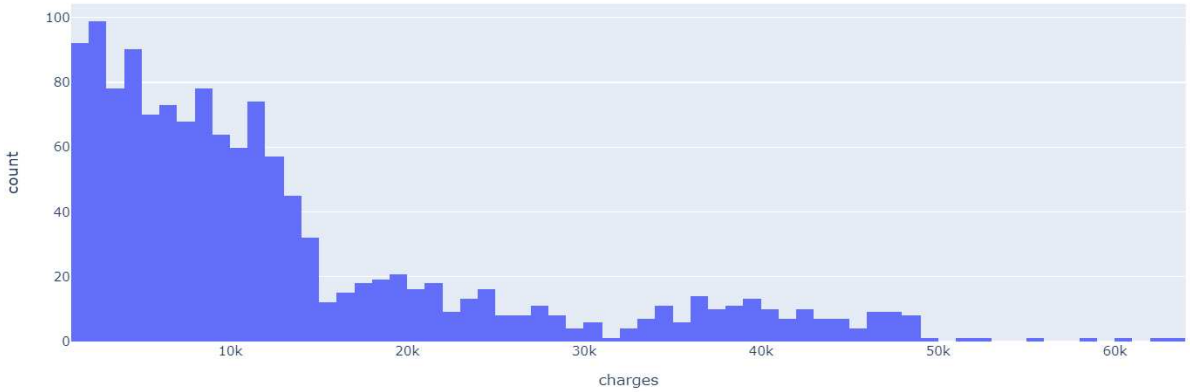
5.11.1 Descrição da Base de Dados

A área de seguros de saúde costuma apresentar distribuições com caudas pesadas devido a uma pequena porcentagem de pacientes que demandam mais recursos. A base de dados fornecida contém informações sobre os beneficiários de uma empresa de seguros, o que pode ajudar a estimar o custo médio a ser cobrado pelo seguro de saúde. Os atributos relevantes para este trabalho são:

1. *age*: Idade do beneficiário principal;
2. *sex*: Sexo do contratante do seguro;
3. *bmi*: Índice de massa corporal, em quilogramas por metro quadrado;
4. *children*: Número de crianças que seguro de saúde cobre;
5. *smoker*: Índice se o beneficiário principal fuma, ou não;
6. *region*: Área residencial do beneficiário nos Estados Unidos;
7. *charges* (Alvo): Custo médico individual cobrados pelo seguro de saúde, em dólares.

5.11.2 Comportamento da Variável *Charges*

Figura 5.10 – Histograma ilustrando a distribuição dos valores da variável *charges*.



Fonte: De autoria própria.

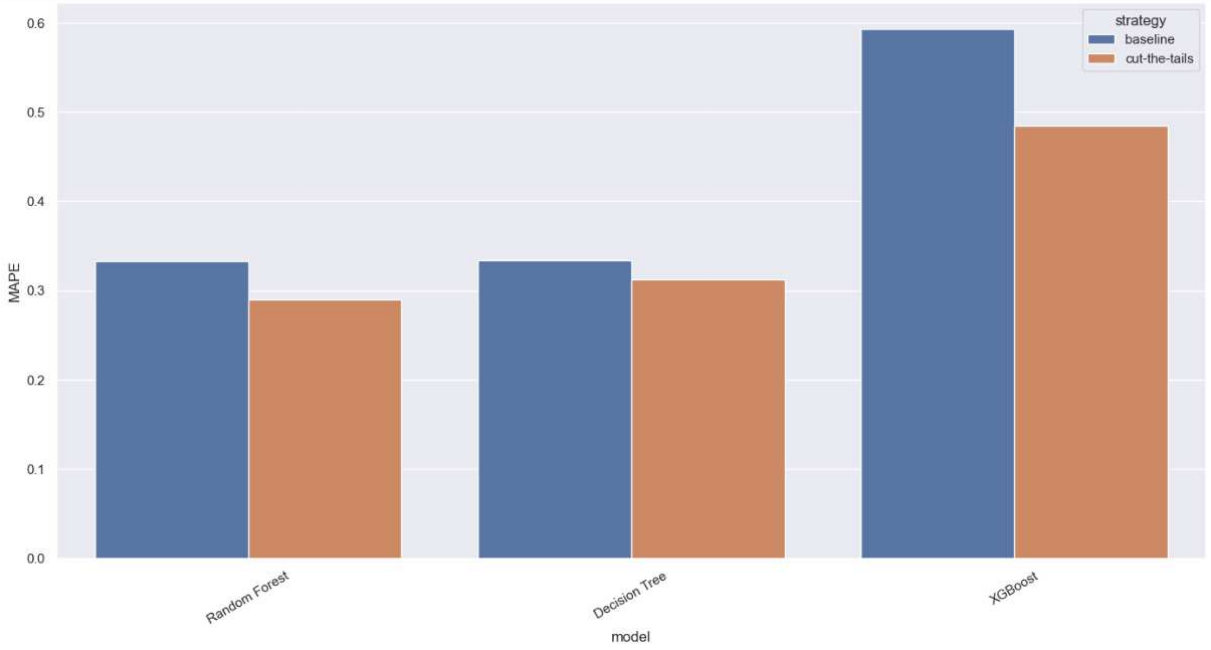
Ao observar o histograma mostrado na Figura 5.10 e a Tabela 5.1, é perceptível que a distribuição apresenta uma cauda definida à direita, justificada pelo seus índices razoáveis de curtose e assimetria (1.595 e 1.514).

6 Resultados

A Figura 6.1 representa o resultado da média arimética entre o MAPE de todas as bases, separando os resultados pelos algoritmos de regressão escolhidos. Cada regressor foi executado e testado em cada base uma vez.

6.1 Comparação de Desempenho Geral Por Modelo

Figura 6.1 – Barplot ilustrando o MAPE de dos modelos selecionados.



Fonte: De autoria própria.

Observando a Figura 6.1, é possível observar uma diferença do desempenho da abordagem *Cut the Tails*, produzindo o MAPE em todos os modelos usados para o experimento. É possível notar a diferença entre o erro dos modelos, onde o *XGBoost* se destaca exibindo uma maior média de MAPE em comparação com os outros modelos, e o modelo Random Forest exibe a menor média de MAPE entre todos os outros modelos.

6.2 Resultados dos Algoritmos de Otimização

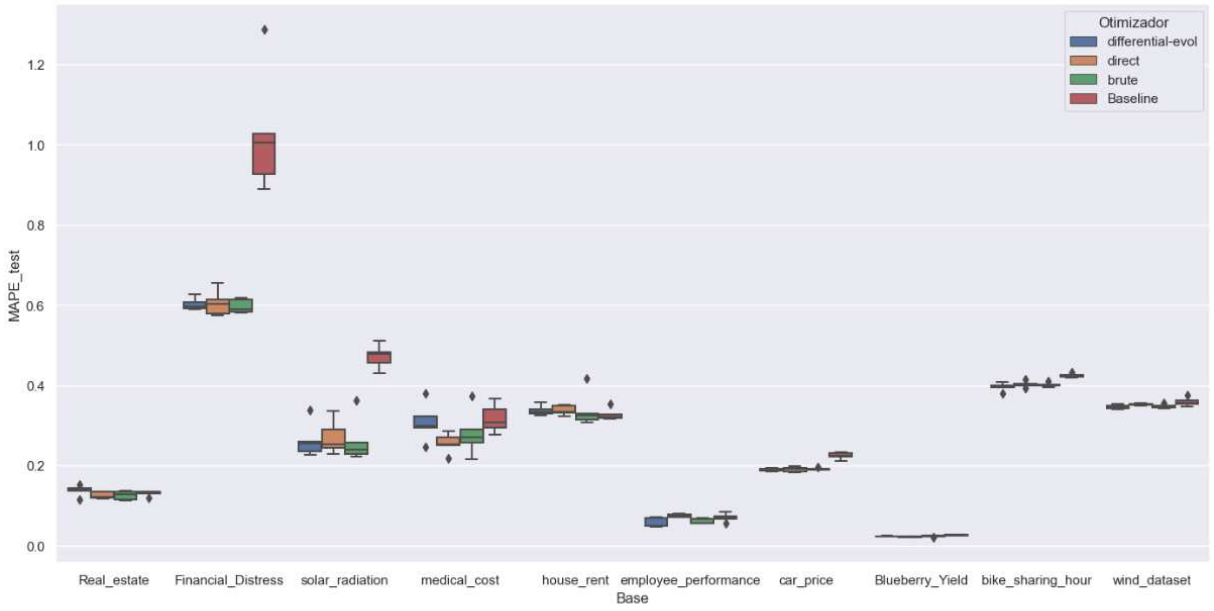
A Tabela 6.1 ilustra as diferença de MAPE de cada base de dados e otimizador em comparação com o *baseline*. Para obter este resultado, foi realizada a média aritmética entre todos os resultados das réplicas de cada otimizador, e subtraído do MAPE do método *baseline*. Uma melhora positiva demonstra que a abordagem *Cut the Tails* se prova como superior em comparação ao método *baseline*, caso contrário a melhora será negativa.

Tabela 6.1 – Tabela de melhoras dos otimizadores.

Base	Otimizador	Melhora de MAPE
Blueberry Yield	Força Bruta	0,0028
Blueberry Yield	*Direct*	0,0042
Blueberry Yield	*Diff. Evol.*	0,0030
Employee Performance	Força Bruta	0,0073
Employee Performance	*Direct*	-0,0053
Employee Performance	*Diff. Evol.*	0,0128
Financial Distress	Força Bruta	0,4294
Financial Distress	*Direct*	0,4213
Financial Distress	*Diff. Evol.*	0,4238
Car Price	Força Bruta	0,0329
Car Price	*Direct*	0,0338
Car Price	*Diff. Evol.*	0,0345
Real Estate	Força Bruta	0.0044
Real Estate	*Direct*	0.0048
Real Estate	*Diff. Evol.*	-0.0076
Wind Speed	Força Bruta	0,0102
Wind Speed	*Direct*	0,0062
Wind Speed	*Diff. Evol.*	0,0127
Bike Sharing	Força Bruta	0,0231
Bike Sharing	*Direct*	0,0219
Bike Sharing	*Diff. Evol.*	0,0280
Solar Radiation	Força Bruta	0,2100
Solar Radiation	*Direct*	0,2017
Solar Radiation	*Diff. Evol.*	0,2083
House Rent	Força Bruta	-0,0114
House Rent	*Direct*	-0,0107
House Rent	*Diff. Evol.*	-0,0103
Medical Cost	Força Bruta	0,0360
Medical Cost	*Direct*	0,0620
Medical Cost	*Diff. Evol.*	0,0089

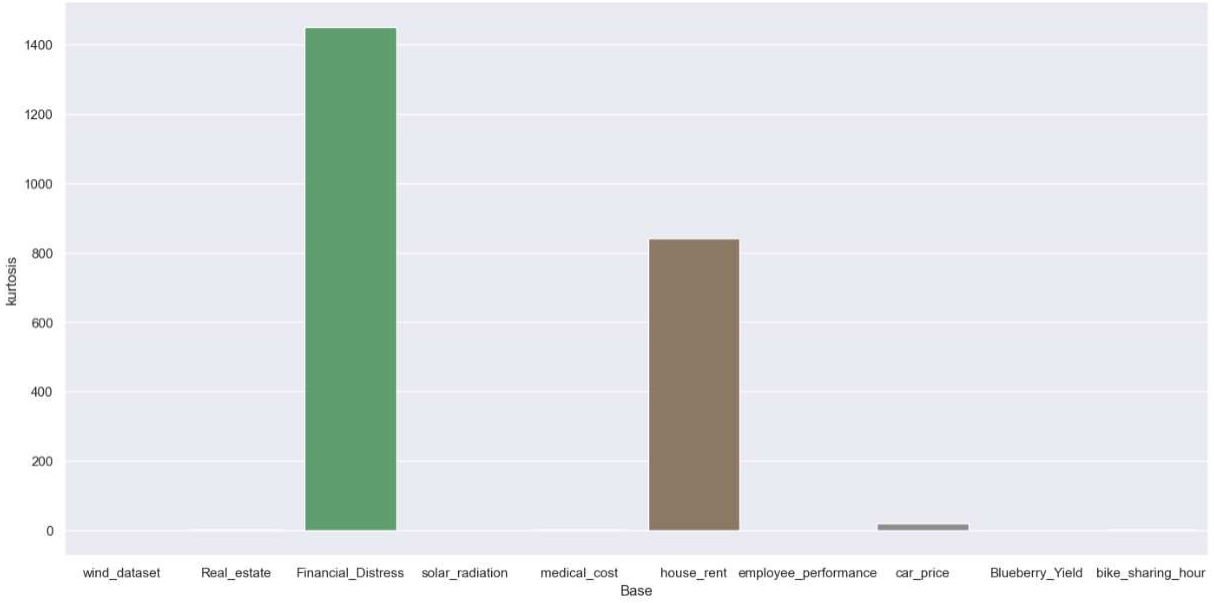
Fonte: De autoria própria.

Figura 6.2 – Histograma ilustrando o MAPE de todas as bases de dados.



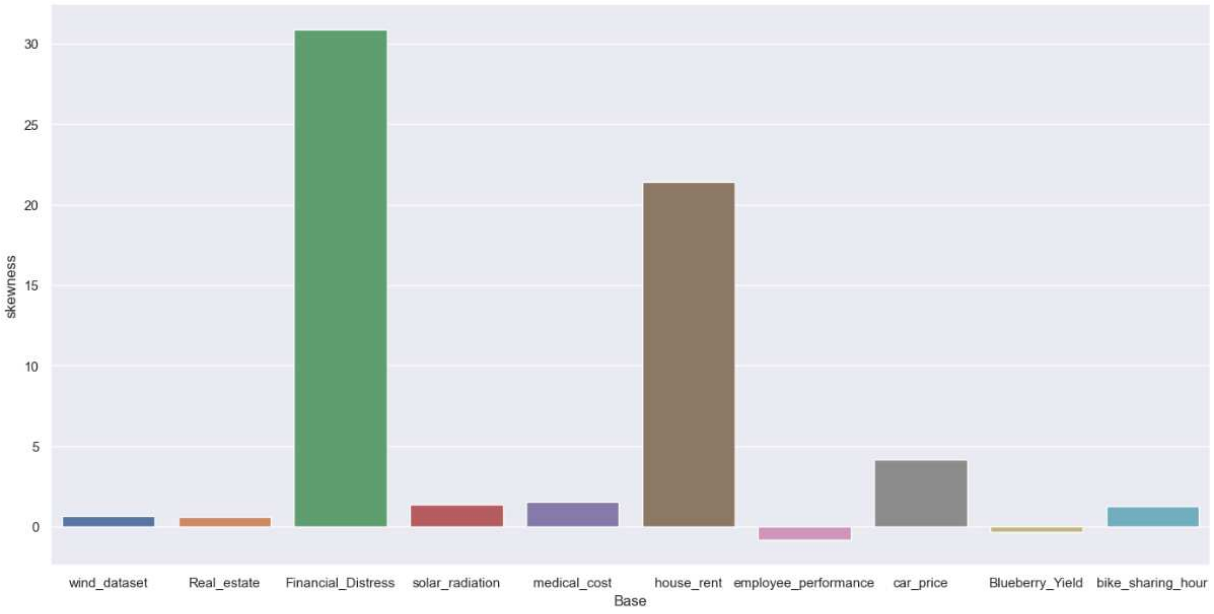
Fonte: De autoria própria.

Figura 6.3 – Histograma ilustrando a Curtose de cada Base.



Fonte: De autoria própria.

Figura 6.4 – Histograma ilustrando a Assimetria de cada Base.



Fonte: De autoria própria.

6.3 Análise dos Modelos

Ao analisar os resultados da Tabela 6.1, é possível extrair as seguintes informações em destaque:

Pode ser observado nos resultados dos modelos que o modelo *Random Forest* mostra o menor MAPE em comparação com os outros modelos de maneira consistente, mas não significa que são os melhores resultados dentre eles. Aparentemente o modelo *Random Forest* conseguiu se adaptar às bases com mais êxito do que os outros métodos, sendo possível concluir que, no geral, ele é o melhor modelo para estes conjunto de bases de dados.

6.4 Análise dos Algoritmos de Otimização

Comparando os otimizadores da Figura 6.2, é possível perceber que todos tem um desempenho equiparável em diversas bases de dados, portanto é seguro concluir que nenhum otimizador tem vantagem clara sobre os outros e a escolha entre estes otimizadores fica à critério do experimentador.

6.5 Análise Final

Ao analisar a Figura 6.2 e a Tabela 6.1, é possível enumerar os cinco melhores resultados:

1. *Financial Distress*, com uma melhora de 42%;
2. *Solar Radiation*, com uma melhora de 20%;
3. *Medical Cost*, com uma melhora de 6%;
4. *Car Price*, com uma melhora de 3%;
5. *Bike Sharing*, com uma melhora de 2%.

Para se analisar mais profundamente a razão pelo qual os modelos se adaptaram melhor a estas bases, foi necessário utilizar técnicas de estatística denominadas de curtose e assimetria. As Figuras 6.3 e 6.4 apresentam o valor de curtose e assimetria de cada base, respectivamente.

Ao observar a Figura 6.3, é possível extrair a informação de que as bases *Financial Distress*, *House Rent* e *Car Price* tem índice de curtose elevado em comparação com as outras bases.

Analisando a Figura 6.4, é possível entender que mesmo que todas as bases possuam uma assimetria elevada (com ressalva das bases *Employee Performance* e *Blueberry Yield*) as bases que se destacaram com assimetria elevada foram: *Financial Distress*, *Solar Radiation*, *Medical Cost*, *House Rent*, *Car Price* e *Bike Sharing*.

Mesmo que a base *House Rent* tenha se destacado nas Figuras 6.3 e 6.4, apresentando ambas curtose e assimetria elevadas, na Figura 6.2 ela não demonstrou nenhuma melhora significativa em relação ao método *baseline*. Este caso foi considerado uma anomalia nos resultados, pois na fase de otimização, os cortes não convergiram para um intervalo semelhante de percentis, mas convergiram para diversos cortes diferentes com o MAPE semelhante que não condiziam com a forma da distribuição.

7 Considerações Finais

7.1 Conclusão

Neste trabalho é proposto uma abordagem de aprendizado de máquina para distribuições de cauda pesada utilizando um método de seleção de cortes para separar a distribuição da variável alvo. A abordagem foi implementada com o uso de algoritmos de otimização para selecionar os percentis para os cortes. A abordagem proposta foi testada comparando seus resultados com outros modelos de aprendizado tradicional sem nenhuma abordagem especial denominada *baseline*. Quando implementado e testado, o algoritmo *Cut the Tails* resultou em melhoras em comparação a sua versão *baseline* em casos particulares, onde as distribuições apresentaram índices de curtose e assimetria elevados. Como um exemplo, a base de dados *Financial Distress* apresentou uma melhora de 42% em relação a abordagem *baseline* com todos os algoritmos de otimização selecionados. Em contrapartida, os resultados que não obtiveram melhoras apresentam uma melhora de aproximadamente 1% à favor da abordagem *baseline* .

Considerando os resultados, a técnica *Cut the Tails* se prova superior à técnica *baseline* em casos de distribuições com curtose e assimetria elevadas.

7.2 Trabalhos Futuros

Após o término deste trabalho, foram feitas algumas considerações para os trabalhos futuros neste tema:

- Realizar validação cruzada nos modelos das bases de dados;
- Classificar objetivamente as bases como distribuições de cauda pesada.

Referências

ANDREY. *Solar Radiation Prediction*. 2017. Disponível em: <https://www.kaggle.com/datasets/dronio/SolarEnergy>.

BANERJEE, S. *House Rent Prediction Dataset*. 2022. Disponível em: <https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset?select=House_Rent_Dataset.csv>.

BIRLA, N.; VERMA, N.; KUSHWAHA, N. *Vehicle dataset*. 2023. Disponível em: <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv>.

BOURGUIGNON, M.; SANTOS-NETO, M.; CASTRO, M. de. A new regression model for positive random variables with skewed and long tail. *METRON*, 2021. Disponível em: <https://link.springer.com/article/10.1007/s40300-021-00203-y#article-info>.

BREIMAN, L. Random forests. *Machine Learning*, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. Association for Computing Machinery, 2016. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

DECARLO, L. T. On the meaning and use of kurtosis. *Psychological methods*, American Psychological Association, v. 2, n. 3, p. 292, 1997.

DOANE, D. P.; SEWARD, L. E. Measuring skewness: a forgotten statistic? *Journal of statistics education*, Taylor & Francis, v. 19, n. 2, 2011.

DUTTA, G. *Employee Performance Prediction*. 2022. Disponível em: <https://www.kaggle.com/datasets/gauravduttakiit/employee-performance-prediction?select=train_dataset.csv>.

EBRAHIMI. *Financial Distress Prediction*. 2018. Disponível em: <https://www.kaggle.com/datasets/shebrahimi/financial-distress>.

FEDESORIANO. *Wind Speed Prediction Dataset*. 2022. Disponível em: <https://www.kaggle.com/datasets/fedesoriano/wind-speed-prediction-dataset>.

FELIX, J. C. *Algoritmos de aprendizado de máquina para previsão do tempo da manutenção de vagões*. [S.l.]: Dissertação de Mestrado, 2022.

GOLDIE, C. M.; KLÜPPELBERG, C. Subexponential distributions. *A practical guide to heavy tails: statistical techniques and applications*, Birkhäuser Boston, p. 435–459, 1998.

JONES, D.; PERTTUNEN, C.; STUCKMAN, B. Lipschitzian optimisation without the lipschitz constant. *Journal of Optimization Theory and Applications*, v. 79, p. 157–181, 01 1993.

KIM, S.; KIM, H. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, v. 32, n. 3, p. 669–679, 2016. ISSN 0169-2070. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207016000121>.

KOTSIANTIS, S. B. Decision trees: a recent overview. *Artificial Intelligence Review*, 2013. Disponível em: <https://doi.org/10.1007/s10462-011-9272-4>.

MANNING, W. G.; BASU, A.; MULLAHY, J. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, v. 24, n. 3, p. 465–488, 2005. ISSN 0167-6296. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167629605000056>.

OBSIE, E. Y.; QU, H.; DRUMMOND, F. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Computers and Electronics in Agriculture*, v. 178, p. 105778, 2020. ISSN 0168-1699. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016816992031156X>.

PATEL, A. *Rental Bike Sharing Dataset*. 2022. Disponível em: <https://www.kaggle.com/datasets/imakash3011/rental-bike-sharing>.

SIGMAN, K. Appendix: A primer on heavy-tailed distributions. *Queueing Systems*, 1999. Disponível em: <https://doi.org/10.1023/A:1019180230133>.

SILVA, J. L. d. C.; FERNANDES, M. W.; ALMEIDA, R. L. F. d. Estatística e probabilidade. *3ª edição, Ceará. Editora ABEU*, 2015.

STORN, R.; PRICE, K. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, v. 11, p. 341–359, 01 1997.

SUN, J.; FREES, E. W.; ROSENBERG, M. A. Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics*, v. 42, n. 2, p. 817–830, 2008. ISSN 0167-6687. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167668707000972>.

TAKEUCHI, I.; BENGIO, Y.; KANAMORI, T. Robust Regression with Asymmetric Heavy-Tail Noise Distributions. *Neural Computation*, v. 14, n. 10, p. 2469–2496, 10 2002. ISSN 0899-7667. Disponível em: <https://doi.org/10.1162/08997660260293300>.

YEH, I.-C. *Real estate valuation Data Set*. 2018. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set>.