



Aplicação de Inteligência Artificial na Previsão e Otimização de Processos Químicos

Edgar da Silva Vieira dos Santos

Mestrado em Química

Departamento de Química e Bioquímica

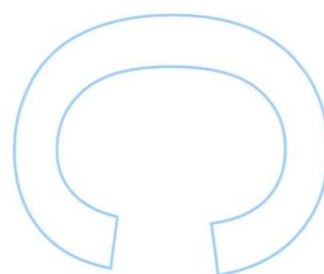
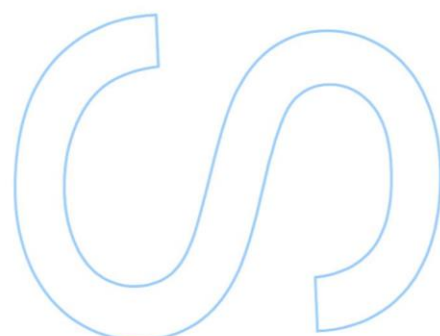
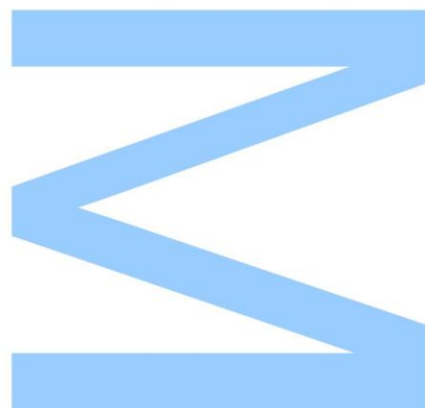
2021

Orientador

Maria Natália Dias Soeiro Cordeiro, Professora associada, Faculdade de Ciências

Coorientador

Filipe Carlos Teixeira Gil, Investigador, Faculdade de Ciências

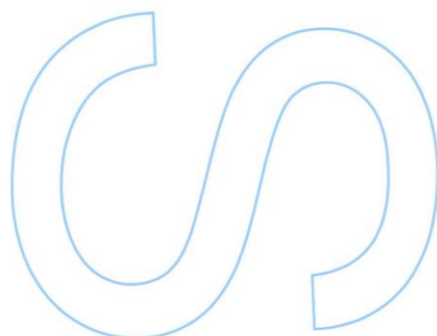
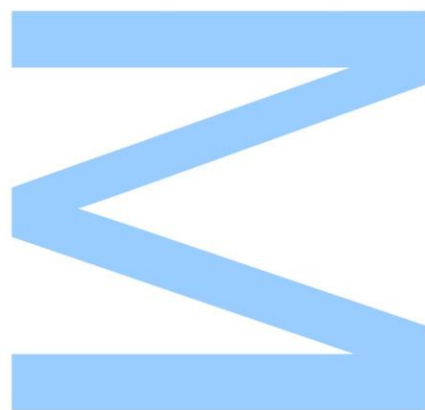




Todas as correções determinadas
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____



Agradecimentos

Aqui, ficam os meus mais sinceros agradecimentos a todos os que, de uma forma direta ou indireta, contribuíram para o desenvolvimento e estruturação do presente trabalho.

Aos meus orientadores, Doutora Natália Cordeiro e Doutor Filipe Teixeira, que trabalharam na construção dos alicerces em que esta dissertação assenta, permitindo assim o desenvolvimento do trabalho efetuado. A eles também, que incansavelmente fizeram todos os esforços para responder a todas as questões que surgiam e realizar os reparos necessários que o decurso do trabalho exigia.

À minha família, em particular, à minha mãe e irmã pela motivação contínua e apoio emocional que se pressupõem necessários na realização deste trabalho.

À minha namorada, Ana Bezerra, não apenas pela paciência e afeto demonstrados, mas por todo o conhecimento partilhado e auxílio prestado em diversos tópicos pertinentes à dissertação.

Finalmente, a todos os meus amigos que demonstraram uma presença invariável nos momentos mais importantes.

A todos, expresso aqui a minha profunda gratidão.

Resumo

Nitroarenos são um subproduto de combustão de combustíveis com elevada toxicidade, sendo por este motivo, a sua conversão mais sustentável do que a sua eliminação. De modo a assegurar a reutilização e reciclagem destes produtos secundários, é frequentemente empregue a sua redução a aminoarenos. Este grupo de compostos por sua vez, é passível de ser aplicado, como precursor, em várias indústrias de síntese química. São exemplos destas, a tintureira e a farmacêutica.

Da perspetiva da química sustentável, é essencial que tanto o dispêndio de energia para a execução da reação seja minimizado, e maximizado o aproveitamento dos reagentes. Para isso, este trabalho propõe-se a encontrar as melhores condições de utilização para diferentes catalisadores. Mais ainda, é também tido por objetivo encontrar as melhores características que podem ser exibidas por um catalisador desta reação.

Para a execução deste trabalho foi fornecida uma base de dados previamente construída, bem como, um modelo de *Random Forest* que prevê os valores de k_1 de reações de redução de nitroarenos, através de condições de reação, características de catalisador e substrato utilizado. Para a realização da otimização das condições de reação foram utilizados dois tipos de algoritmos de otimização: Monte Carlo, e algoritmos de enxame. Inicialmente, foi realizado um estudo dos dados e do modelo fornecidos, para que se pudesse aferir quanto à distribuição dos dados, possíveis *outliers* e principais variáveis representativas dos dados.

É realizada posteriormente, a otimização das condições para todos os catalisadores catalogados. Mediante os resultados obtidos e as características de cada catalisador, foram selecionados seis para prosseguir com a análise comparativa entre os otimizadores. São então comparados os resultados obtidos entre os otimizadores, e entre as suas diferentes versões. Para a otimização das características do catalisador, utiliza-se o que obteve melhor desempenho nos ensaios anteriores.

Para complementar todo o trabalho realizado, define-se o objetivo secundário de obter correlações entre as condições de utilização. Para tal efeito, sempre que pertinente, aplicam-se técnicas de correlação aos resultados obtidos. No final apresentam-se as ilações de todo o trabalho realizado.

Abstract

Nitroarenes are a highly toxic fuel combustion by-product, making it, for this reason, their conversion more sustainable than their disposal. To ensure the reutilization and recycling of these secondary products, their reduction to aminoarenes is frequently employed. On their behalf, this last group of compounds can be applied as a precursor in several industries, among the largest are the dye and pharmaceutical industries.

From the sustainable chemistry standpoint, it is essential that both the energy costs are minimized, as well as the product yield maximized, as to ensure the consumption of the maximum substrate possible. To achieve this, the present work sets out to find the best set of reaction conditions for a set of catalysts. Additionally, an attempt is also made to find the best properties in a catalyst for this reaction.

To carry out this work, a previously constructed database, as well, as a Random Forest model were supplied. This model aims to predict the k_1 values of a certain reaction from its conditions, catalyst properties and substrate used. To perform the optimization of the reaction conditions, two different optimization algorithm types were used: Monte Carlo based optimizer, and a Swarm intelligence optimizer. First, both the model and the data were studied, to make inferences about data distribution, possible outliers, and the most important variables.

Then, the optimization results for the reaction conditions of all the catalysts are presented. According to the obtained results and each catalysts' characteristics, six were selected to proceed to the comparative analysis between optimizers. These optimizations are then performed, and their result compared between optimizers and each of their versions. Proceeding to the catalysts' characteristics optimization, the algorithm that showed the best performance in the past trials was selected.

Complementing everything that was achieved, the secondary objective of finding correlations among reaction conditions. To achieve this, correlation techniques are applied to the optimized results whenever deemed suitable. In the end, the conclusions drawn from this work are summarized.

Índice

Agradecimentos.....	i
Resumo	ii
Abstract	iii
Índice.....	iv
Índice de figuras	vi
Índice de tabelas.....	viii
1. Introdução.....	1
1.1. Cinética química	1
1.2. Redução de nitroarenos	3
1.3. Métodos de obtenção e de tratamento de dados experimentais	5
1.3.1. Tratamento de bases de dados	5
1.3.2. Data mining	6
1.4. Aprendizagem automática	9
1.4.1. Árvores de decisão	11
1.4.2. Métodos de ensemble: Random Forest	12
1.5. Métodos de otimização	14
1.5.1. Enxame de abelhas artificiais	15
1.5.2. Algoritmo de alcateia de lobos cinzentos	16
2. Metodologia.....	17
2.1. Otimização de Monte Carlo	18
2.2. Algoritmos de enxame	19
3. Resultados e discussão	20
3.1. Caracterização dos dados	20
3.2. Caracterização do modelo de <i>Random Forest</i>	25
3.3. Otimização de Monte Carlo	33
3.3.1. Otimização de parâmetros.....	34
3.3.2. Otimização de condições de reação	36
3.4. Otimização de enxame	50
3.4.1. Otimização de parâmetros.....	50
3.4.2. Otimização de condições de reação – Abelhas artificiais.....	51
3.4.3. Otimização de condições de reação – Lobos cinzentos.....	57
3.5. Comparação das condições otimizadas.....	60
3.6. Comparação do processo de otimização	63

3.7.	Otimização de características de catalisador	65
3.7.1.	Lobos cinzentos – Início aleatório	65
3.7.2.	Lobos cinzentos – Covil	67
4.	Conclusões	68
	Bibliografia	71
	Anexos	79
A.	Caraterização de dados e modelo de <i>Random Forest</i>	79
B.	Otimização de Monte Carlo	82
C.	Otimização de enxame	83

Índice de figuras

Figura 1 - As cinco fases de KDD. ²⁵	6
Figura 2 - Exemplo de aplicação de técnicas de análise de frequências. ²⁸	8
Figura 3 - Sistematização do emprego da função kernel em SVM. ⁴⁴	11
Figura 4 - Árvore de decisão para jogar ténis.	12
Figura 5 - Esquematização de uma Random Forest. ⁶³	14
Figura 6 - Projeção dos catalisadores nas duas primeiras componentes principais.	22
Figura 7 - Projeção dos catalisadores, e das respetivas variáveis com exceção do XRD, nas primeiras duas componentes principais.	23
Figura 8 - Outliers (a azul) indicados pelo método de florestas de isolamento.	24
Figura 9 - Projeção das características dos catalisadores (sem outliers) nas primeiras duas componentes principais.....	25
Figura 10 - Dados originais de pk_1 e valores previstos pelo modelo.....	26
Figura 11 - Resposta parcial do modelo em função da percentagem de oxigénio no catalisador (%).	27
Figura 12 – Resposta parcial do modelo em função da massa de catalisador (g).	28
Figura 13 - Resposta parcial do modelo em função da média do diâmetro de partícula do catalisador (nm).	28
Figura 14 - Resposta parcial do modelo em função da quantidade de substrato (g). ..	29
Figura 15 - Resposta parcial do modelo em função do tempo de reação (s).	30
Figura 16 - Resposta parcial do modelo em função do volume de solução (mL).	31
Figura 17 - Resposta parcial do modelo em função da percentagem de azoto na composição do catalisador (%).	31
Figura 18 - Resposta parcial do modelo em função da percentagem de carbono na composição do catalisador (%).	32
Figura 19 - Resposta parcial do modelo em função da área de superfície do catalisador (m^2/g).	32
Figura 20 - Resposta parcial do modelo em função do diâmetro médio dos poros do catalisador (nm).	33
Figura 21 - 1ª abordagem de otimização, $P_{acpt} = 0.45$	34
Figura 22 - 2ª abordagem de otimização, $P_{acpt} = 0.45$	35
Figura 23 - 3ª abordagem do otimizador, $P_{acpt} = 0.45$	36
Figura 24 - 3ª abordagem do otimizador, $P_{acpt} = 0.80$	36

Figura 25 - Diferença entre o valor inicial e final de k_1 para os 10 catalisadores com maiores melhorias registadas após as corridas do otimizador de Monte Carlo.	37
Figura 26 - Matriz de correlação das condições otimizadas e do k_1	38
Figura 27 - Projeção dos catalisadores e condições otimizadas nas primeiras duas componentes principais.	39
Figura 28 - Método do cotovelo, distorção do espaço pelo número de clusters de k-médias.	40
Figura 29 - PC 1 a 4 e representação dos clusters obtidos.	41
Figura 30 - PC 5 a 8 e representação dos clusters obtidos.	41
Figura 31 - Projeção dos dados nas primeiras 4 componentes principais, subdivididos em 5 clusters.	42
Figura 32 - Melhorias de k_1 para os catalisadores selecionados.	49
Figura 33 - Teste de otimização para: (a) 15 agentes; (b) 20 agentes.	51
Figura 34 - Melhoria do k_1 obtido por corrida do algoritmo das abelhas artificiais para os catalisadores: (a) 0 ; (b) 18 ; (c) 65 ; (d) 73 ; (e) 120 ; (f) 181	52
Figura 35 - Melhorias absolutas por catalisador do algoritmo de abelhas artificiais de início aleatório na corrida 7.	55
Figura 36 - Melhor k_1 obtido por corrida do algoritmo de lobos cinzentos para os catalisadores (a) 0 ; (b) 18 ; (c) 65 ; (d) 73 ; (e) 120 ; (f) 181	58
Figura 37 - Gráfico de matriz de correlação das variáveis obtidas pelo otimizador de lobos cinzentos de início aleatório para o catalisador 0	61
Figura 38 - Gráfico de matriz de correlação das variáveis obtidas pelo otimizador de lobos cinzentos de início aleatório para o catalisador 18	63
Figura 39 - Melhor k_1 obtido por corrida, para ambos os substratos (a) 2-Nitrofenol; (b) 4-Nitrofenol.	65

Índice de tabelas

Tabela 1 - Valores fixados, e respetivas unidades, para a otimização das características do catalisador.	18
Tabela 2 - Lista de variáveis utilizadas para o treino do modelo.	21
Tabela 3 - Importância por permutação das 10 variáveis mais relevantes do modelo de Random Forest.	26
Tabela 4 - Variância explicada pelas componentes principais.	39
Tabela 5 - Centroides dos clusters obtidos em k-médias.	43
Tabela 6 - Características dos catalisadores selecionados.	46
Tabela 7 - Valores de condições de partida do otimizador.	47
Tabela 8 - Resultados da última corrida do otimizador de Monte Carlo.	48
Tabela 9 - Valores de k_1 estimados para as condições iniciais e otimizadas.	50
Tabela 10 - Condições experimentais obtidas após a 7ª corrida do algoritmo de abelhas artificiais, para os catalisadores selecionados.	53
Tabela 11 - Condições experimentais obtidas após a 8ª corrida do algoritmo de abelhas artificiais, para os catalisadores selecionados.	53
Tabela 12 - Condições experimentais obtidas após a 9ª corrida do algoritmo de abelhas artificiais, para os catalisadores selecionados.	54
Tabela 13 - Melhores condições obtidas pelo algoritmo de abelhas artificiais de início predefinido.	55
Tabela 14 - Condições ótimas obtidas pelo algoritmo de abelhas artificiais - colmeia dispersa.	56
Tabela 15 - Condições ótimas obtidas pelo algoritmo de lobos cinzentos – covil.	57
Tabela 16 - Médias e desvios-padrões de todos os catalisadores para as corridas 2, 3, 6, 7, 9 e 10 da versão de início aleatório de lobos cinzentos.	59
Tabela 17 - Valores de k_1 estimados para as condições utilizadas experimentalmente e para as melhores condições encontradas pelos algoritmos de enxame.	64
Tabela 18 - Tempo médio de corrida de cada tipo de algoritmo.	64
Tabela 19 - Características ótimas, para ambos os substratos, obtidas pelo otimizador de início aleatório.	66
Tabela 20 - Média e desvio-padrão de cada característica de catalisador para o 2-Nitrofenol, para todas as corridas	66
Tabela 21 - Média e desvio-padrão de cada característica de catalisador para o 4-Nitrofenol, para todas as corridas.	67

Tabela 22 - Características ótimas, para ambos os substratos, obtidas pelo otimizador de covil.	68
--	----

1. Introdução

No dia 27 de agosto de 1859 assistia-se ao nascimento da indústria petrolífera, tendo Edwin Drake, extraído petróleo do solo pela primeira vez.¹ Durante os anos seguintes, a humanidade explorou de forma exaustiva as potencialidades deste combustível fóssil, expandindo desta forma largamente, o conhecimento e a tecnologia. Contudo, em meados do século XX começaram a surgir as sequelas ambientais desta intensa exploração, aumentando a preocupação pela preservação do meio ambiente. Na sequência do início desta cadeia de pensamento, surgiram, no final do século XX os 12 princípios da química verde:²

1. Prevenção de resíduos é preferível à sua eliminação ou tratamento;
2. A síntese deve aproveitar ao máximo todas as matérias-primas;
3. A síntese deve utilizar e gerar substâncias inócuas;
4. Os produtos devem ser concebidos para serem inócuos;
5. Devem ser utilizados catalisadores em vez de reagentes;
6. O uso de aditivos deve ser minimizado;
7. Os requisitos de energia para a síntese devem ser minimizados;
8. Os reagentes devem ser progressivamente mais renováveis;
9. Os compostos derivados devem ser evitados;
10. Os produtos de reação devem decompor-se em produtos inócuos;
11. Obter melhor controlo dos processos químicos;
12. As substâncias utilizadas devem ter o mínimo potencial para acidentes;

Estes princípios têm sido aplicados até à data para tornar a síntese química mais sustentável e segura.

1.1. Cinética química

Como citado nos princípios da química verde, a síntese deve sempre aproveitar ao máximo as matérias-primas, e minimizar os gastos de energia. Uma vez que as condições da reação exigem, normalmente, dispêndio de energia (seja em temperatura, pressão ou controlo de outras condições), quanto mais tempo demorar uma reação a ocorrer, mais energia será dissipada. Para uma reação completa, a velocidade da reação pode depender, ou não, da concentração dos reagentes. Por outro lado, para uma reação em equilíbrio, a velocidade da reação em qualquer ponto dependerá

também da extensão da reação inversa. Neste aspeto, todas as reações podem ser classificadas em diversas ordens de cinética química. A ordem global de uma reação depende do mecanismo de conversão dos reagentes. Para uma reação elementar genérica: $aA + bB \rightarrow cC + dD$, a sua ordem global é dada pela Equação 1, na qual s e t são os expoentes que constam da sua lei de velocidade descrita pela Equação 2.³

$$\text{Ordem} = s + t,$$

Equação 1 - Ordem global da reação.

$$v = k[A]^s \cdot [B]^t$$

Equação 2 - Lei de velocidade da reação.

onde, k é a contante de velocidade (ou velocidade específica), $[A]$, $[B]$ são as concentrações dos reagentes A e B em mol/dm³, correspondendo s e t à ordem cinética de cada um dos reagentes.

Podem ser estudadas segundo estas equações reações de várias ordens, contudo, para medir a ordem de uma reação é necessário efetuar um procedimento experimental que permita obter esta informação. Normalmente, é seguido um procedimento, tal que permita obter informação quanto à concentração de um dos reagentes.⁴ Por este motivo, são mais desejáveis, do ponto de vista do estudo da cinética química, reações de ordem zero ou de primeira-ordem. As restantes podem ser influenciadas pela concentração de dois ou mais reagentes, algo que é impossível de medir, pelo procedimento comum. A solução para este problema passa muitas vezes por adicionar quase todos os intervenientes da reação em excesso. Um dos intervenientes, cuja ordem de reação se estime ser um, é adicionado numa porção previamente medida, de modo que a sua concentração seja conhecida ao longo da reação. Estas são denominadas reação de pseudo-primeira ordem, e a sua constante é doravante referida como k_1 . Este método é particularmente útil em reações, que não dependem apenas de um reagente, e das quais se pretende conhecer a constante de velocidade.⁵

Conhecer então o k_1 , de reações cuja velocidade dependa de mais que um reagente, é feito através do mesmo método que reações de primeira ordem. Ou seja, utilizando o mesmo exemplo em que $aA + bB \rightarrow cC + dD$, e assumindo que se trata de uma reação de primeira ordem em relação a A:

$$v = k \cdot [A] = -\frac{d[A]}{dt}$$

Equação 3 - Velocidade de uma reação de primeira ordem.

então, integrando a Equação 3 em ordem ao tempo t , retira-se, que:

$$\ln|A|_t = \ln|A|_0 - kt$$

Equação 4 - Lei integrada da velocidade para reações de primeira ordem.

Pela Equação 4, observa-se que a constante de velocidade de uma reação de primeira ordem, ou de pseudo-primeira ordem, é igual ao declive de $\ln|A|_t$ em função do tempo. Por este motivo, para conhecer o k_1 de uma reação, é traçada tal variação e calculado o declive. Este procedimento tem a evidente limitação, no caso de se tratar de uma reação que entre em equilíbrio, o k_1 tem de ser calculado no início da reação quando a concentração de produto, ainda é desprezível em relação à concentração de reagente.⁶

1.2. Redução de nitroarenos

Sendo o petróleo um dos principais componentes primários e transversal a várias indústrias, torna-se importante aproveitar ao máximo todos os compostos que podem ser dele extraídos ou produzidos após a sua utilização. Se os produtos gerados pela utilização do petróleo forem tóxicos, então a sua reutilização e transformação tem ainda maior importância. Se estes produtos não forem reutilizados terão de ser despejados, e para poderem ser despejados terão de ser previamente convertidos em produtos inócuos para o ambiente.²

São exemplos disto os nitroarenos, compostos aromáticos com um grupo nitro. Este grupo de compostos é um produto secundário da combustão de derivados do petróleo.⁷

Os nitroarenos são, normalmente, gerados como produtos secundários em reações de combustão de petróleo e derivados. Isto deve-se à oxidação de N_2 a NO_x e posterior adição a compostos aromáticos não consumidos no processo de combustão, ou formados como produtos da combustão incompleta do combustível.⁸ Como tal, de modo a obter um maior aproveitamento do petróleo, utilizando ao máximo os seus diversos subprodutos, torna-se interessante o estudo destes compostos. Dentro das transformações de nitroarenos, é particularmente interessante o estudo da sua redução para aminoarenos. Isto, porque este tipo de compostos pode ser facilmente empregue em diversas indústrias, como por exemplo, as indústrias agrícola, tintureira e farmacêutica.^{9,10} Mesmo sendo possível executar esta reação sem recurso a vias

catalíticas, uma vez que requer agentes redutores mais fortes, isto é bastante incomum, sendo normalmente utilizados catalisadores de modo a poupar reagentes e tornar a reação mais eficiente.¹¹

Podem ser utilizados diversos tipos de catalisadores para a execução da mesma reação, desde grafenos a nanopartículas de silicas,^{11–16} sendo desta forma importante avaliar as vantagens e desvantagens inerentes à utilização de cada um.

De forma a conhecer as propriedades de cada um destes catalisadores existem diversos métodos para os caracterizar. De entre os quais, o de rápida execução e interpretação geral simples, será a transformada de *Fourier* de Infravermelho (FTIR), que consiste na captação de radiação, cujo comprimento de onda se situa na zona de infravermelho (1 mm a 700 nm).¹⁷ Também com uma interpretação fácil, sendo, contudo, um pouco mais dispendioso, existe o método de espectroscopia de raios-X por energia dispersiva (EDS). Esta técnica confere informação acerca dos elementos, e respetiva percentagem, que compõem um ponto da amostra. Normalmente realizadas em simultâneo com esta técnica estão as técnicas de microscopia eletrónica de varrimento (SEM) e microscopia eletrónica de transmissão (TEM). Estas conferem informação quanto a aglomerações e tamanho de partículas.¹⁸ Conferindo também informação relativa às propriedades físicas do catalisador, encontra-se o método de adsorção de gás (normalmente N₂) para efetuar o cálculo da área superficial, conhecido como o método de Brunauer-Emmett-Teller (BET).¹⁹ Finalmente, uma técnica também muito utilizada, particularmente em catalisadores metálicos, é a difração de raios-X (XRD). Esta técnica confere informação quanto à cristalografia do material, sendo que quanto mais cristalino, maior o número de sinais encontrados no seu espectro e maior a sua intensidade. Este método, é mais frequentemente visto a ser empregue em catalisadores metálicos, uma vez que não é espetável que catalisadores de carbono demonstrem elevada cristalinidade.^{18–21}

Dadas as inúmeras composições possíveis para estes catalisadores, e a todas as formas de os caracterizar, aferir cada um quanto a vantagens e desvantagens do seu uso é um processo moroso. Para atingir este objetivo, será primeiro necessário conhecer bem a eficácia de cada um e, conseqüentemente, as melhores condições de utilização dos mesmos. O método mais imediato, será a experiência por tentativa/erro, ou seja, a experiência de diversas condições (teoricamente sãs), até ser encontrado um conjunto de condições tal, que possa ser considerado como condições ótimas. Naturalmente, isto requer um trabalho laboratorial muito extenso e uma quantidade

elevada de energia e de reagentes, de modo a cobrir todas, ou quase todas, as possibilidades.

1.3. Métodos de obtenção e de tratamento de dados experimentais

Um outro método que poderá ser aplicado na otimização de condições de uma dada reação, seria um *design of experiments* (DoE), considerando que a otimização das condições da reação seria realizada apenas para um dos catalisadores.²² Para atingir este objetivo, era necessário primeiro definir uma ou mais variáveis como resposta. Algumas das variáveis mais vulgarmente utilizadas para definir a extensão de uma reação são, a percentagem máxima de conversão e o tempo até ser atingido este ponto.^{23,24} Desta forma, pode ser avaliado o impacto de cada condição da reação, e a interação entre as condições reacionais. Assim, é possível minimizar o trabalho experimental a realizar, contudo, não se obteria informação transversal a outros catalisadores.

Para que a informação seja transversal a múltiplos catalisadores seria necessário obter informações sobre vários catalisadores em várias condições. Na literatura estão já documentadas diversas experiências com vários catalisadores em diferentes condições. Sendo possível (ainda que moroso), extrair a informação previamente documentada e efetuar o tratamento dos dados registados.

1.3.1. Tratamento de bases de dados

Os dados, depois de registados, serão normalmente convertidos em subconjuntos de dados consoante as variáveis que se entenderem ser mais importantes para o objetivo em causa. Isto é normalmente reconhecido como a primeira das cinco fases de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases - KDD*).²⁵

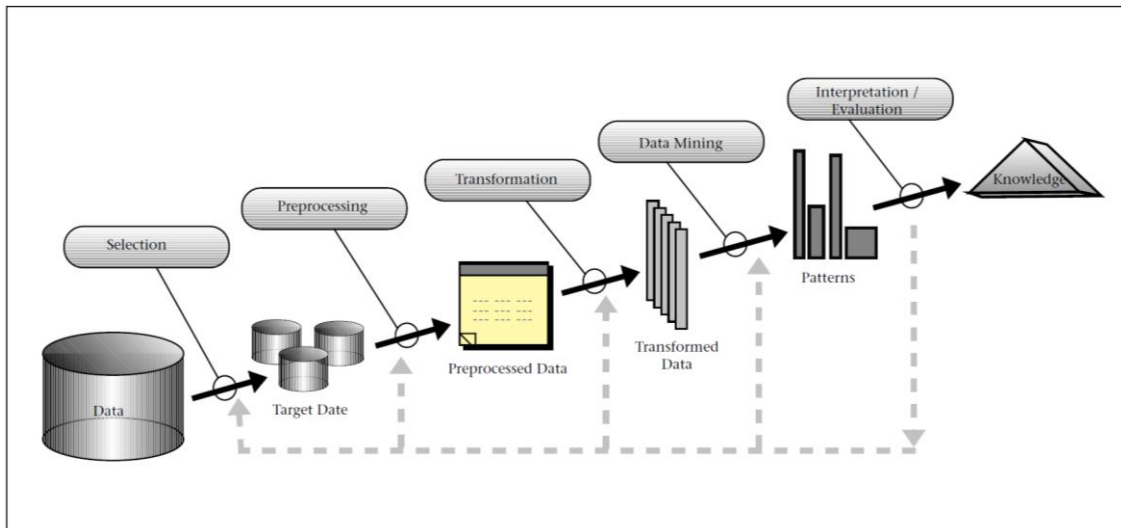


Figura 1 - As cinco fases de KDD.²⁵

Na Figura 1, estão traduzidas de uma forma sintética as etapas nas quais assenta a descoberta através de bases de dados. Existem também outras formas de subdividir o processo e denominação de cada etapa, como por exemplo o processo SEMMA (*Sample, Explore, Modify, Model, Access*) desenvolvido e cunhado pelo instituto SAS.²⁶ Tal como foi referido, a primeira etapa assenta na identificação do objetivo e seleção de variáveis em função do objetivo que se pretende atingir. A segunda fase é o pré-processamento dos dados, que consiste em identificar e tratar o ruído, e decidir como lidar com dados em falta. A terceira etapa, denominada de transformação, reside principalmente na redução de dados, sendo isto particularmente útil em bases de dados com um número de variáveis muito elevado.²⁷ A quarta fase é geralmente identificada como *data mining*. Esta é a fase principal, e é aqui que são observados os padrões nos dados e criados modelos que os repliquem.²⁵ Por fim, a quinta e última fase, é a avaliação e implementação do conhecimento adquirido.

1.3.2. Data mining

A definição de *data mining* (mineração de dados) utilizada com mais frequência é a de Fayyad e Piatetsky-Shapiro.²⁶ De acordo com esta definição, *data mining* define o emprego de algoritmos para executar a extração de padrões de uma série de dados. Assim, *data mining* refere-se a uma das etapas de KDD.

A extração de padrões subjacentes a múltiplas entradas numa base de dados é levada a cabo através de vários métodos existentes para este fim. Os diferentes

métodos disponíveis terão domínios de aplicabilidade diferentes. Alguns métodos mais simples baseiam-se na avaliação de frequência de um ou mais acontecimentos sob certas circunstâncias, enquanto outros, podem basear-se, por exemplo, em aglomeração ou classificação.²⁸

Um bom exemplo da aplicação do primeiro conjunto de técnicas é o caso de aplicação de probabilidades condicionais num supermercado. Esta aplicação pode ser utilizada para determinar que produtos são comprados ao mesmo tempo, e ajudar assim, a dispor os produtos de forma a que os clientes possam finalizar as compras no menor tempo possível.²⁹ Claro está que o objetivo final de qualquer negócio é gerar lucro. Ainda que seja de facto útil para o consumidor ter os produtos que pretende comprar dispostos lado-a-lado por conveniência, é também evidente que esta estratégia pretende induzir a compra de segundos produtos por consumidores que estariam apenas a considerar trazer os adjacentes.

Na Figura 2 está representada uma situação onde a aplicação da técnica de análise de frequência, de uma forma simplificada, permite aferir que existe, pelo menos um conjunto de produtos que é frequentemente comprado ao mesmo tempo. Aqui observa-se que, clientes que açúcar, geralmente também compram ovos. Isto, será provavelmente utilizado na confeção de bolos. Contudo, nota-se também que clientes que compram leite geralmente compram pão. Neste caso, esta análise visa também induzir a compra de leite nos clientes que já compram pão e vice-versa. Isto é muitas vezes feito com recurso a “promoções” destes produtos, utilizando assim os dados das compras de alguns dos clientes para induzir a compra em futuros clientes.

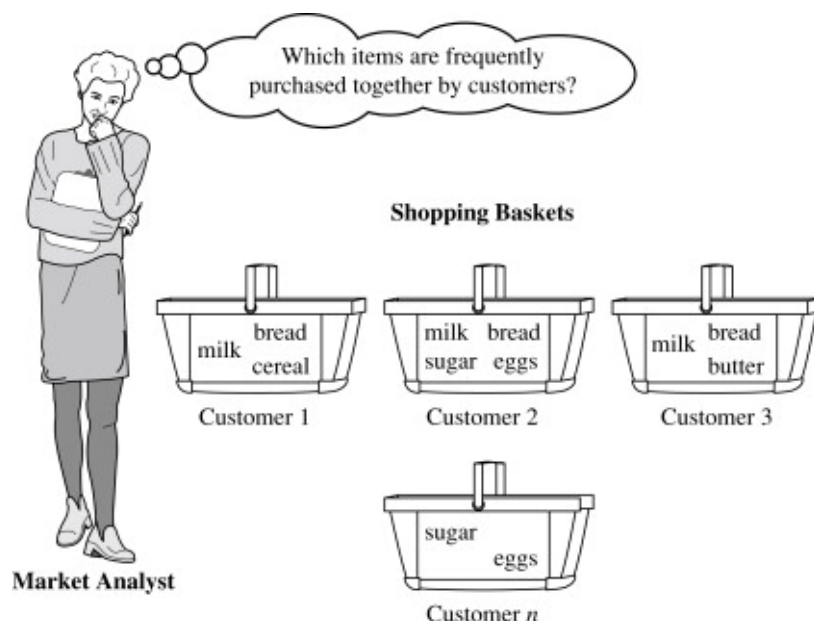


Figura 2 - Exemplo de aplicação de técnicas de análise de frequências.²⁸

Para que se possa chegar a estas conclusões, são realizadas todas as etapas de *KDD* descritas anteriormente. Destas, o pré-tratamento de dados é uma etapa crucial da análise para que possam ser tratados por métodos de *Machine Learning*. O número baixo de observações juntamente com a presença de várias variáveis explicativas conduz a uma dificuldade de análise elevada. Um conjunto de dados pode conter a presença de *outliers*, observações que se encontram bastante distante das restantes numa certa população amostral. A alimentação dos dados em “cru” a um modelo de *Machine Learning* pode conduzir a modelos mais instáveis e à atribuição de uma importância incorreta às variáveis explicativas.³⁰

A identificação de *outliers* pode ser uma tarefa árdua, principalmente em dados em que não existe uma variável que funciona como classificador. Sendo estes pontos pouco frequentes, é necessário conhecer bastante bem os dados a tratar, e se é possível ou não retirar as observações que recaiam sobre esta categoria. Existem várias técnicas de identificação de *outliers*, como florestas de isolamento.³¹ Nesta técnica é aplicada uma árvore de decisão ao conjunto de dados em que as decisões são com base nas distâncias entre as observações, assim pontos mais dispersos da nuvem de distribuição irão ser separados das restantes observações. No final, os pontos que se encontram mais perto da base da árvore de decisão poderão ser identificados como *outliers*.³²

1.4. Aprendizagem automática

Aprendizagem automática (*Machine Learning*) é o estudo e construção de algoritmos computacionais que visam modelar a capacidade humana de aprender, melhorando-se por experiência.³³ Integra-se como uma pequena parte, na área de inteligência artificial, sendo esta por sua vez, definida por toda a inteligência demonstrada por máquinas, não estando por isso, restrita à capacidade de aprendizagem.³⁴

Técnicas de *Machine Learning* (ML), podem ser divididos em técnicas supervisionadas e não-supervisionadas. Técnicas supervisionadas são assim descritas por requererem um conjunto de dados de treino e um conjunto de dados de teste. Os dados de treino são constituídos pelos respetivos valores de entrada e de resposta desejados, e têm por função a construção de um modelo de regressão ou classificação. Os dados de teste são utilizados para fazer uma apreciação da explicação do modelo em dados para os quais não foi treinado.³⁵ Índices de correlação muito altos em dados de teste e baixos em dados de treino, apontam para o *overfitting* do modelo. De outra forma, técnicas não-supervisionadas traduzem-se por técnicas que apenas requerem os dados de entrada existentes, sem que seja necessária a sua divisão em subconjuntos. São exemplos disto, técnicas de *clustering*. Este tipo de técnicas, avaliam a disposição geométrica dos dados e criam grupos nos quais dividem os dados pelos valores apresentados nas variáveis a correlacionar.

Por outro lado, técnicas de análise de classificação (ou de regressão) são técnicas de aprendizagem supervisionadas e pressupõem já, uma existência prévia de grupos (ou valores) aos quais vão ser atribuídos os diversos casos. Tanto técnicas de *clustering* como técnicas de classificação são frequentemente utilizadas em diferentes áreas desde processamento de imagem, a análise financeira.³⁶

Uma grande parte da área de *Machine Learning*, está intimamente relacionada com o estudo de estatística computacional, que se foca principalmente, na realização de previsões, sejam estas numéricas ou classificativas. Estas técnicas produzem posteriormente um modelo baseado nos dados de treino, que traduz a aprendizagem efetuada pelo computador.³⁷ Neste sentido, existem já vários algoritmos que permitem efetuar estas previsões, sendo alguns dos mais utilizados:³⁸

1. Análise de componentes principais (*Principal component analysis*);
2. Regressão linear;

3. Regressão logística;
4. Máquina de vetores de suporte (*Support Vector Machines*);
5. *K*-médias;
6. Árvores de decisão;
7. Métodos de *ensemble*.

O PCA (*Principal component analysis*) é um método frequentemente utilizado em pré-tratamento de dados geralmente em conjuntos que apresentam uma grande quantidade de variáveis explicativas.³⁹ Este método tem como intuito a redução da dimensionalidade em componentes principais, ortogonais entre si. Em implementações modernas, este método recorre normalmente à técnica de SVD (*Single Value Decomposition*). As componentes principais derivam da decomposição da matriz de dados em várias matrizes de onde se retiram os valores próprios e os respetivos vetores próprios. Isto permite selecionar previamente o número de componentes que se pretende extrair.⁴⁰ O conjunto de todas as componentes principais captura a variância total dos dados, sendo que as primeiras componentes capturam a maior quantidade da variância. Assim é possível reduzir um conjunto de dados com várias variáveis explicativas a um número pequeno de eixos ortogonais facilitando, tanto visualização dos dados, como a respetiva análise e conhecimento do tipo de dados a tratar.⁴⁰

A regressão linear será, provavelmente, um dos métodos mais intuitivos. Serve para representar uma variável dependente que varia linearmente em função de uma ou mais variáveis independentes.⁴¹

Apesar do nome, a regressão logística, é também um método de classificação binário. Serve para estimar os parâmetros de um modelo, que por sua vez irá determinar a probabilidade da variável dependente assumir uma dada classificação, com base nos valores das variáveis independentes.⁴² Por exemplo, dados os resultados de análises de um paciente, este modelo já treinado, poder-se-ia referir quanto à probabilidade do paciente padecer de uma tal condição 'A' previamente ponderada.

Support Vector Machines (SVM) trata-se de um método de classificação que utiliza vetores de classificação de suporte para separar os dados em dois grupos. Estes vetores de classificação são desenhados de forma a maximizar a margem entre os vetores e os pontos em questão. De forma a criar um hiperplano capaz de separar os dados no plano em que se encontram é utilizada a função cerne (ou *kernel*), que em algumas implementações é apenas uma função linear dos planos já existentes.^{43,44}

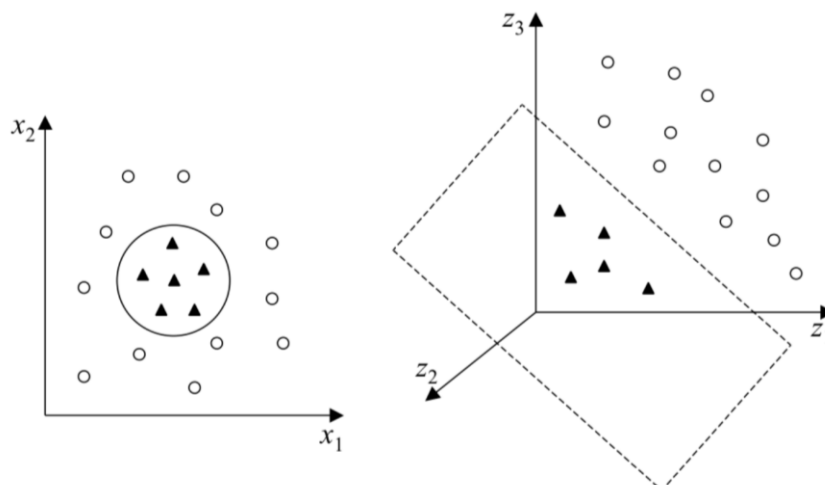


Figura 3 - Sistematização do emprego da função kernel em SVM.⁴⁴

K-Means é um método de *clustering*, considerado ser não-supervisionado ou semi-supervisionado. O método diz-se não-supervisionado, uma vez que não é necessário conhecimento prévio dos dados, sendo atribuído um número de clusters n iterativamente a k pontos, e associando os restantes casos a esses centroides n .³⁶ Podem, contudo, ser utilizados outros métodos, alguns dos quais podem ser mais eficientes, dependendo das situações.^{45,46} Adicionalmente, este método é conhecido por ser também semi-supervisionado, uma vez que podem ser aplicados os conhecimentos prévios sobre o conjunto de dados a trabalhar para definir o número de *clusters*. Isto é, se for conhecido k (o número de grupos presentes no conjunto de dados), isto permite passar à frente o processo de iniciação do algoritmo, alimentando o número de centroides já conhecido, ultrapassando assim, a maior desvantagem deste método.^{47,48}

1.4.1. Árvores de decisão

As árvores de decisão foram desenvolvidas como ferramentas de suporte de decisão onde são apresentados os diferentes eventos possíveis e as respetivas probabilidades, se conhecidas.⁴⁹ A Figura 4 representa uma esquematização do procedimento de decisão para jogar ténis. Neste exemplo simples, observa-se o processo de decisão automática mediante a observação de certas variáveis.

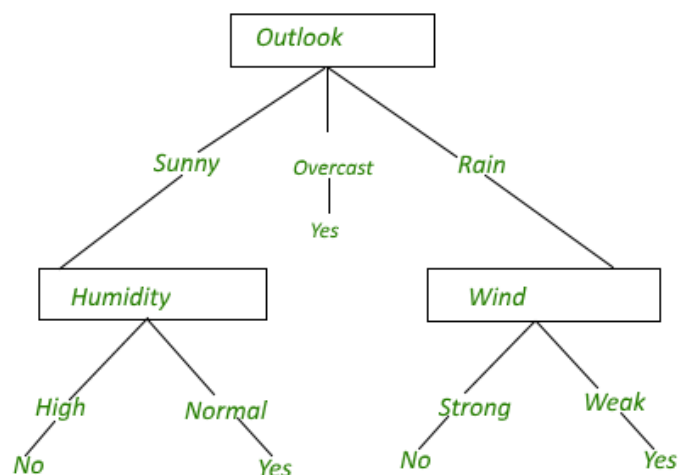


Figura 4 - Árvore de decisão para jogar ténis.

Uma vez que não existe um limite concreto no número de nodos a utilizar numa árvore, cada problema pode ser melhor explicado por um número diferente de nodos, contudo, é relevante referir que quantos mais forem descritos, mais complexa será a árvore, e consequentemente mais difícil será a sua execução.⁵⁰ Para mitigar este efeito, muitas vezes recorre-se ao método de *pruning* (poda), que consiste em retirar alguns dos nodos sem os quais o erro não aumenta consideravelmente.

Mesmo não sendo a forma mais intuitiva da sua utilização, é possível utilizar uma variante deste método para descrever um conjunto de variáveis independentes, e ainda, para prever não só variáveis dependentes discretas, ou qualitativas, mas também contínuas. No primeiro caso, as árvores serão denominadas de árvores de classificação, pois classificam cada caso, mediante as variáveis listadas.⁵¹ No segundo caso, são relatadas árvores de regressão, uma vez que estas almejam prever o valor exato de uma variável contínua. Aqui, um modelo de regressão é atribuído a cada nodo final da árvore, de modo que a variável possa seguir princípios diferentes para cada conjunto de variáveis independentes.⁵²

1.4.2. Métodos de ensemble: Random Forest

As árvores de decisão regressivas viram já a sua aplicação em alguns trabalhos na área da química,^{53,54} particularmente no estudo de catálises.^{55,56} Demonstram em certos casos uma boa capacidade de descrever um modelo capaz de prever uma grande maioria dos casos.^{52,55,56} Contudo, quando a base de dados a trabalhar é composta por

uma elevada quantidade de dados, por vezes, é difícil controlar a complexidade do modelo, mesmo com o recurso a *pruning*. Assim, em vez de tentar simplificar uma só árvore, utilizam-se várias, avaliando-se o resultado de cada uma e usando o contributo de todas para obter um resultado final. Estes tipos de métodos fazem parte de métodos de *ensemble*.

Existem vários métodos para executar este tipo de algoritmos, sendo que todos têm por base a construção de um modelo de regressão ou classificação, que combine os pontos fortes de modelos mais simples.⁵⁷ Entre as técnicas de *ensemble* de árvores mais populares estão *Adaboost* e as *Random forests*.^{51,58}

Random Forest é um método que tem por objetivo final criar um modelo que dependerá de vários modelos mais simples de árvores de classificação ou de regressão. O resultado do modelo final é a média aritmética de cada uma das previsões realizadas por cada árvore, no caso de regressões. Para classificações, é avaliada a classificação que obteve a maioria dos votos e é atribuída essa classificação ao caso em estudo.⁵⁹ O método tem este nome uma vez que cada árvore é plantada de forma aleatória, recorrendo a *bootstrap*. Esta técnica descreve o procedimento de retirar m amostras de uma população com tamanho n , de forma aleatória e com reposição ($m < n$).^{58,60}

Depois de feita a primeira amostragem é plantada uma árvore com base nas amostras recolhidas, na qual a ramificação da árvore é feita de maneira que cada nível, consiga separar o melhor possível as amostras em grupos de menor dimensão. Isto significa, que cada nodo vai ramificar pela variável que conseguir obter o menor grupo possível. O procedimento é repetido até ser atingido o número de árvores desejado.^{61,62}

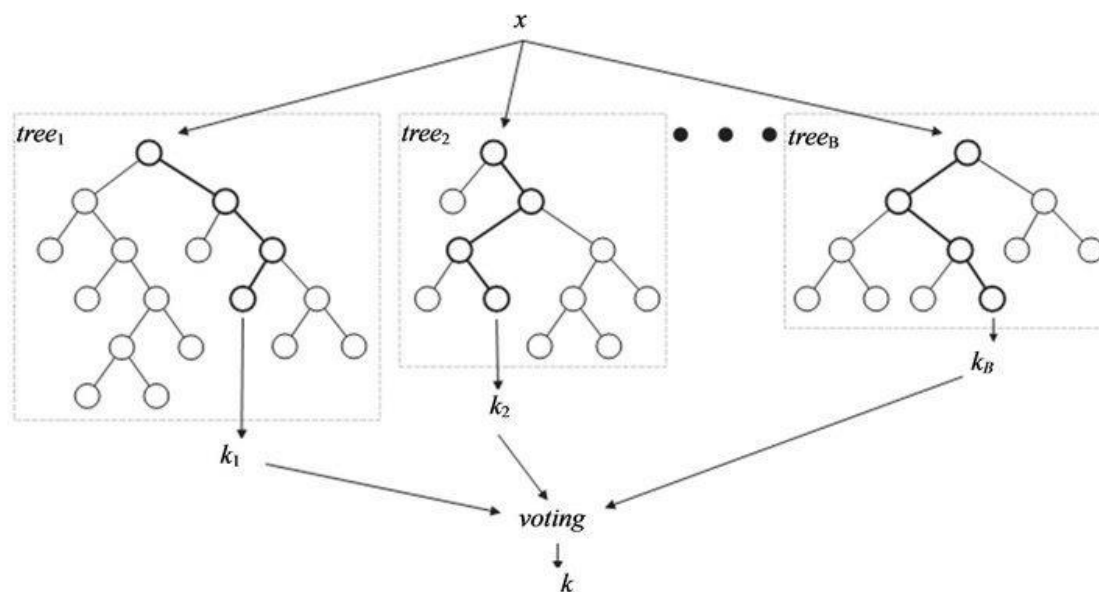


Figura 5 - Esquemática de uma Random Forest.⁶³

Finda a construção do modelo com base nos dados de treino executa-se, então, o teste do modelo. Nesta fase, técnicas de *ensemble* como *Random Forest* tem uma vantagem em relação a métodos que dependem de um só modelo, nomeadamente a capacidade de testar e minimizar o erro *out-of-bag (oob)*. Isto é, realizar previsões a partir das árvores em que algumas observações aleatórias não foram utilizadas, e calcular o desvio da previsão do modelo para o valor real destas observações.^{64,65}

Para o método descrito, observa-se que o resultado será sempre um modelo sem tradução por expressão analítica, tratando-se de um método estatístico não-paramétrico.⁶⁶ Muitas vezes é útil percorrer estes métodos para encontrar certos parâmetros que originaram soluções ótimas ou ideais.^{67,68} Para tal efeito, recorre-se a outra área da inteligência artificial, nomeadamente, a área de otimização.

1.5. Métodos de otimização

De entre os métodos de otimização mais antigos e comuns, estão os métodos de *Gradient descent* e de Monte-Carlo. O primeiro método, tenta encontrar o mínimo de uma função calculando alguns pontos e as respetivas derivadas até atingir um ponto em que a função é decrescente e a derivada suficientemente próxima de zero. Este método é menos eficaz em funções muito uniformes ou “planas” com pequenas regiões de

mínimos, uma vez que tende a convergir rapidamente para o primeiro mínimo local que encontra.⁶⁹

De forma a colmatar este problema, o método de otimização de Monte-Carlo, tenta iterativamente encontrar o mínimo de uma função efetuando saltos aleatórios, de forma a percorrer a maioria do domínio da função. Este método é um dos mais utilizados para otimizar funções muito ruidosas, uma vez que os valores atribuídos não dependem da topologia da função, contrastando assim, com o método de Newton.⁷⁰ Neste método pode ainda ser introduzido um parâmetro P_{acpt} , que determina a probabilidade do algoritmo aceitar valores que resultam num ponto mais elevado da função. Apesar de serem os métodos mais comuns, existem alguns que assentam noutros pressupostos para obter informação quanto aos mínimos ou máximos absolutos de uma função.

Inteligência de enxame (*Swarm Intelligence*) é a disciplina que visa o estudo do comportamento de sistemas naturais ou artificiais auto-organizados e coordenados pelo controlo descentralizado.⁷¹ Este princípio tem vindo a ser aplicado a problemas de otimização, dando origem a algoritmos de otimização mais eficientes, devido a esta descentralização e comunicação entre os agentes. Geralmente, os mais utilizados são os que derivam de sistemas naturais representando a interação de grupos de animais, como, por exemplo, enxames de abelhas⁷² e alcateias de lobos.⁷³

1.5.1. Enxame de abelhas artificiais

O primeiro algoritmo mencionado tenta retratar um enxame de abelhas que procura, ao longo de um campo, o local com maior densidade de pólen ou néctar para alimentar a colmeia. Para o fazer, criam-se três grupos de agentes: abelhas operárias, espectadoras e batedoras. As abelhas batedoras percorrem o campo em busca de fontes de alimentação, as operárias recolhem a comida encontrada e as espetadoras escolhem o ponto com base no movimento das abelhas operárias. Este movimento serve de comunicação entre as abelhas, para se informarem sobre a localização de pontos de alimentação.

As abelhas operárias mantem informação acerca de fontes de alimentação e partilham-na com as espectadoras. Já para o algoritmo, todos os agentes atuam tendo em vista a informação recolhida até então. Quando as abelhas operárias esgotam uma fonte de alimentação, ou seja, quando uma solução que estavam a explorar é melhorada ao máximo, tornam-se batedoras, e estas soluções são apagadas. Nesta analogia, o

número de fontes de alimentação representa o número de soluções da população, a posição representa as coordenadas desta solução, e a quantidade de néctar ou pólen, a qualidade da mesma.⁷⁴

1.5.2. Algoritmo de alcateia de lobos cinzentos

O último algoritmo mencionado tenta retratar a liderança hierárquica e o mecanismo de caça de lobos cinzentos na natureza. Estes animais tem uma hierarquia bem definida estratificada em quatro níveis, aqui denominados: alfa, beta, gama, ómega. Os lobos alfa lideram a alcateia e são os principais responsáveis pela tomada de decisões, podendo, contudo, ceder o poder de decisão a um dos outros elementos. Os lobos beta servem de ligação entre o alfa e a restante alcateia de modo que as ordens sejam comunicadas aos restantes elementos, e todos estejam cientes das ordens ditadas. Também comunicam ao alfa a forma de como a ordem foi recebida. Lobos delta estão abaixo dos supramencionados, mas acima dos ómegas que normalmente, tomam o papel de cuidadores das crias.

O objetivo deste algoritmo passa por retratar esta hierarquia, para o fazer, começa-se por atribuir a estratificação a cada um dos agentes iniciais, sendo o melhor ponto encontrado o alfa, o segundo melhor o beta, o terceiro o delta e os restantes ómegas. Os três melhores lideram a caça, decidindo para onde cada lobo na alcateia vai em seguida, contudo, não é atribuído um maior peso ao alfa nem ao beta, tendo a decisão dos três líderes o mesmo peso e sendo feita a média das três decisões.⁷⁴

2. Metodologia

O objetivo do presente trabalho incide na avaliação do uso de métodos de inteligência artificial na otimização da redução de nitroarenos. Para esse efeito, foi fornecida uma base de dados onde estavam catalogadas várias reações de redução de nitroarenos (de referir, que a base de dados à data de escrita, não era pública, não sendo por este motivo mencionada a sua referência). Foi também fornecido um modelo de *Machine Learning* (ML) descrevendo a eficiência catalítica de vários catalisadores (como pk_1 ; vide Equação 5 abaixo) sob diferentes condições experimentais.

$$pk_1 = -\log_{10} \left(\frac{k_1}{1s^{-1}} \right)$$

Equação 5 - Transformação de k_1 em pk_1

O trabalho foi dividido em duas etapas. Na primeira etapa, os métodos de otimização foram usados para otimizar as condições de utilização de/dos catalisadores já conhecidos. Seguidamente, os mesmos otimizadores foram usados para encontrar pontos ótimos da função de resposta do modelo de ML, procurando assim encontrar combinações ótimas de características de um catalisador para síntese futura.

A referir que todos os otimizadores além das versões originais dos otimizadores de enxame (“início aleatório”), foram implementadas pelo próprio autor, sendo sempre verificadas e testadas antes da sua execução. Além da implementação de raiz do otimizador de Monte-Carlo e da implementação das alterações para as versões acrescentadas dos algoritmos de enxame, foram também implementadas pelo autor as conexões entre os otimizadores e a base de dados.

Na primeira etapa, pretendia-se apenas otimizar as variáveis pertinentes às condições de utilização dos catalisadores. Como tal, as restantes variáveis eram fixadas com base nas características de cada catalisador que estivesse a ser otimizado. Na segunda etapa, como se pretendia obter o melhor catalisador para as condições de utilização mais acessíveis (as que envolvem o menor dispêndio de energia), as condições foram fixadas de acordo com a Tabela 1 abaixo. O significado das variáveis aqui mencionado será explicitado posteriormente (Tabela 2).

Tabela 1 - Valores fixados, e respetivas unidades, para a otimização das características do catalisador.

Variáveis	Valores fixados	Unidades
m_{cat}	0.07	g
m_{sub}	0.005	g
m_{red}	0.05	g
m_{adt}	0	g
T	298	K
T_{ind}	0	s
cdr	1	N/A
V	50	mL

Para a primeira etapa, o substrato fixado foi o 4-Nitrofenol para todos os otimizadores. No segundo caso foram utilizados dois substratos diferentes, o 2-Nitrofenol e o 4-Nitrofenol, e fixadas também as variáveis referentes ao espectro de XRD em 0, e assim otimizadas as características do catalisador em função destes substratos.

2.1. Otimização de Monte Carlo

Este método de otimização descreve um algoritmo que tenta iterativamente pontos diferentes da função de forma aleatória. Nesta implementação, foi tido por objetivo que o otimizador tivesse alguma tendência para chegar aos pontos mínimos da função, ao invés de apenas percorrer a função de forma aleatória. Para tal efeito, foi atribuído um parâmetro à função de otimização chamado P_{acpt} . De cada vez que um valor da função na iteração atual fosse superior ao valor anterior encontrado, ele compara o valor do parâmetro com um número gerado aleatoriamente, se o valor aleatório for inferior ou igual o algoritmo aceita o novo valor proposto, caso contrário rejeita e tenta novas coordenadas. Sendo este valor ajustável, permite que o otimizador percorra o domínio da função de forma focando-se mais nas zonas ótimas da função.

Cada corrida deste algoritmo era composta por 1 000 000 de iterações, ou seja, avaliações da função. Em particular otimizaram-se as condições de todos os catalisadores com este otimizador. Isto serve para que posteriormente fossem escolhidos 6, sendo estes depois comparados a todos os otimizadores utilizados. Definiu-se, também, como objetivo secundário, a análise de *clustering* a estes resultados, para verificar a criação de aglomerados pelas condições otimizadas e as características dos catalisadores.

2.2. Algoritmos de enxame

Os algoritmos de enxame utilizados provêm do repositório “*Swarm Package*”.⁷⁴ Em todos os algoritmos de enxame utilizados, foi acrescentada uma versão que permite registar os melhores pontos de uma corrida, e começar a partir desses mesmos pontos em corridas seguintes. A estas versões foram atribuídos nomes diferentes, alusivos ao tipo de algoritmo de que se trata, e foram testadas paralelamente aos algoritmos do código original.

De modo a ser possível comparar os algoritmos de enxame com o de Monte-Carlo, é necessário que o número de avaliações da função seja idêntico. Para tal efeito, e atentando que o número de iterações no algoritmo de Monte-Carlo foi 1 000 000, o número de iterações nos algoritmos de enxame será calculado de acordo com a fórmula descrita na Equação 6.

$$N^{\circ} \text{ iterações} = \frac{1\,000\,000}{N^{\circ} \text{ agentes}}$$

Equação 6 - Cálculo do nº de iterações para algoritmos de enxame.

3. Resultados e discussão

O presente capítulo apresenta os resultados e faz a discussão dos mesmos. Inicia-se pela descrição e caracterização dos dados, seguindo-se a caracterização do modelo de *Random Forerst*. Em seguida, são discutidos os resultados obtidos por tipo de otimização a realizar, em subcapítulos que estão divididos por: otimização de parâmetros do otimizador, caso haja parâmetros a avaliar; otimização das condições de reação, onde são apresentadas e discutidas as otimizações feitas às condições de utilização dos catalisadores por versão de otimizador; comparação das condições obtidas, onde é feita a representação adicional dos dados necessária e discutida em detalhe a comparação entre as condições dos catalisadores. Depois são comparados todos os otimizadores e os resultados obtidos para que só os mais eficientes sejam utilizados na fase seguinte. Finalmente, discute-se a otimização às características do catalisador.

3.1. Caracterização dos dados

A base de dados fornecida era composta por diversas reações de redução de nitroarenos a aminoarenos. Apresenta várias variáveis que é possível dividir em três grupos: condições da reação; características do catalisador utilizado; substrato utilizado. Em cada reação catalogada nesta base de dados, o redutor utilizado era NaBH_4 . A Tabela 2 representa a lista dos nomes das variáveis e o que cada uma representa no modelo.

Tabela 2 - Lista de variáveis utilizadas para o treino do modelo.

Tipo de variável	Variável	Significado
Condições de reação	m_{cat}	Massa de catalisador (g)
	m_{sub}	Massa de substrato (g)
	m_{red}	Massa de reductor (g)
	m_{adt}	Massa de carbonato de cálcio (g)
	T	Temperatura (K)
	t	Tempo total da reação (s)
	T_{ind}	Tempo de indução do catalisador (s)
	cdr	Ciclo de reciclagem do catalisador
	V	Volume de solução (mL)
Características de catalisador	D_{part}	Diâmetro médio das partículas do catalisador (nm)
	D_{pore}	Diâmetro médio dos poros do catalisador (nm)
	S_{BET}	Área superficial do catalisador medida através da adsorção de N_2 (m^2/g)
	$\%_{Ag}$	Percentagem mássica de prata no catalisador
	$\%_C$	Percentagem mássica de carbono no catalisador
	$\%_{Cu}$	Percentagem mássica de cobre no catalisador
	$\%_{Fe}$	Percentagem mássica de ferro no catalisador
	$\%_N$	Percentagem mássica de azoto no catalisador
	$\%_O$	Percentagem mássica de oxigénio no catalisador
	$\%_P$	Percentagem mássica de fósforo no catalisador
	$\%_{Pd}$	Percentagem mássica de paládio no catalisador
	$\%_{Ru}$	Percentagem mássica de ruténio no catalisador
	$\%_S$	Percentagem mássica de enxofre no catalisador
	$\%_{Si}$	Percentagem mássica de silício no catalisador
	XRD(x)*	Picos apresentados no espetro do XRD

Quase todas as variáveis são contínuas, à exceção do ciclo de catalisador e do substrato que descreve apenas que substrato foi utilizado. Tendo sido catalogados seis substratos diferentes, utilizou-se a técnica de *one-hot encoding* para referenciar qual

* Os picos de XRD utilizados para o treino do modelo vão de $2\theta = 0^\circ$ a $2\theta = 89^\circ$, constituindo assim 90 variáveis no total.

dos substratos está presente na reação a estudar. Assim, o substrato a utilizar corresponde a seis variáveis, uma para cada substrato catalogado:

1. 2,4,6-Trinitrofenol;
2. 2,4-Dinitrofenol;
3. 2-Nitrofenol;
4. 3-Nitrofenol;
5. 4-Nitroanilina;
6. 4-Nitrofenol.

Inicia-se a caracterização dos dados pela análise de componentes principais às características dos 97 catalisadores.

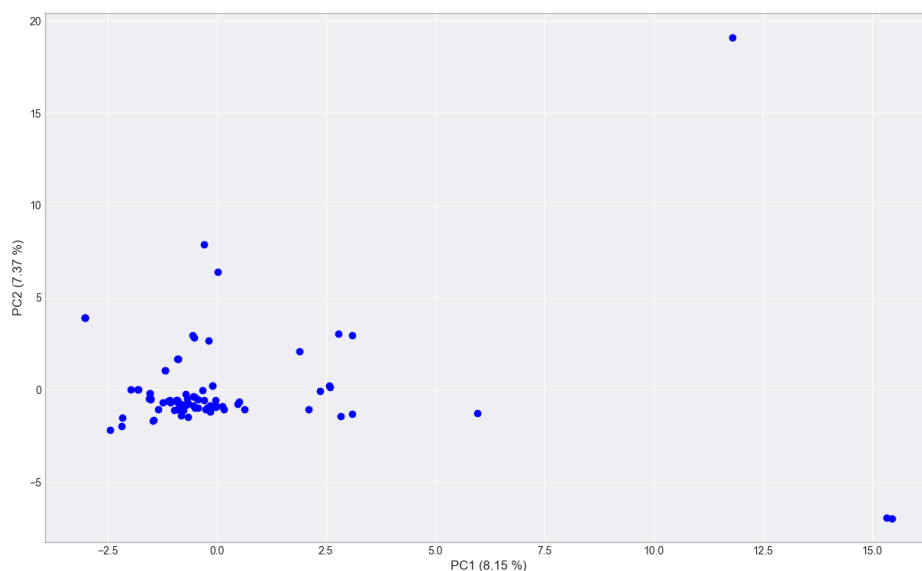


Figura 6 - Projeção dos catalisadores nas duas primeiras componentes principais.

Através de uma breve análise da Figura 6, observa-se que os dados estão muito dispersos, e que, para que seja mais fácil correlacionar informação acerca dos catalisadores com informações obtidas posteriormente através da otimização, será necessário um processamento adicional dos dados.

Pensando apenas acerca do tipo de informação que seria útil retirar, esta incidiria mais nas características do catalisador e na sua correlação com as condições otimizadas, da forma mais abrangente possível. Por este motivo, o passo seguinte consistiu em realizar a mesma projeção, excluindo desta vez a informação contida nos espectros de

XRD por dois motivos: a singularidade conferida por esta informação, e a falta da mesma em muitos dos catalisadores.

Sendo que XRD produz espectros muito singulares a cada composto, e não sinais em zonas características de alguns compostos/elementos, os espectros registados serão pertencentes e únicos aos respetivos catalisadores. Como o objetivo do trabalho e das análises subsequentes é extrapolar por informação pertinente e transversal a vários catalisadores que pertençam a determinados grupos, doravante, opta-se por realizar o tratamento dos dados sem recurso ao XRD. Adicionalmente, dado que apenas um determinado número de catalisadores possui esta informação, os resultados obtidos pela análise da otimização podiam ser enviesados pela falta e/ou existência dos mesmos.

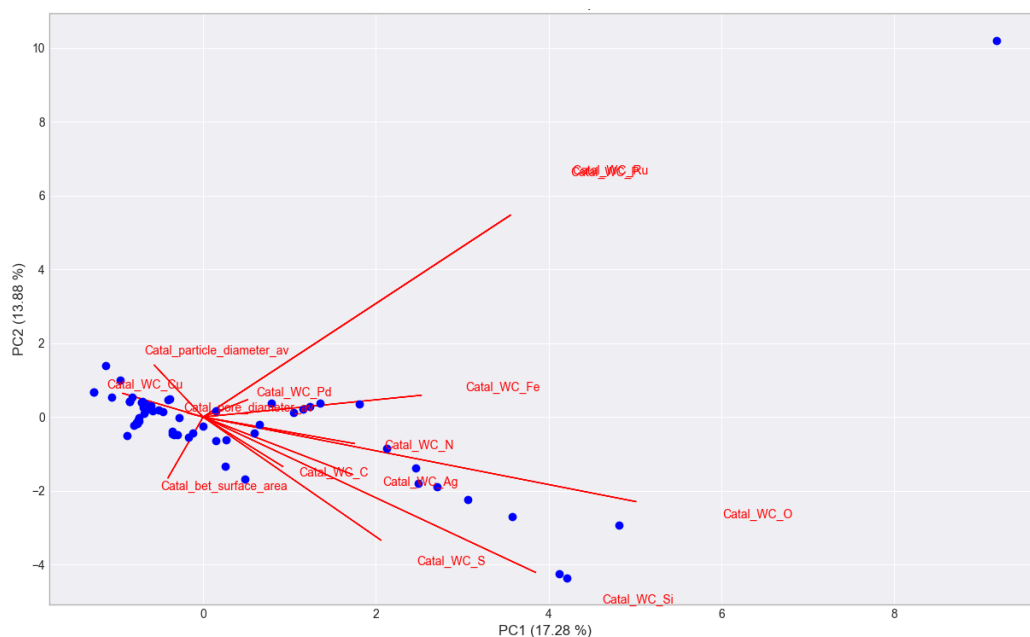


Figura 7 - Projeção dos catalisadores, e das respetivas variáveis com exceção do XRD, nas primeiras duas componentes principais.

A partir da Figura 7, observa-se que apesar de existir uma dispersão dos dados nestas duas componentes, existem alguns dados muito desfasados dos restantes. Como tal, foi realizado o teste de florestas de isolamento aos dados para aferir quanto à existência de *outliers*.

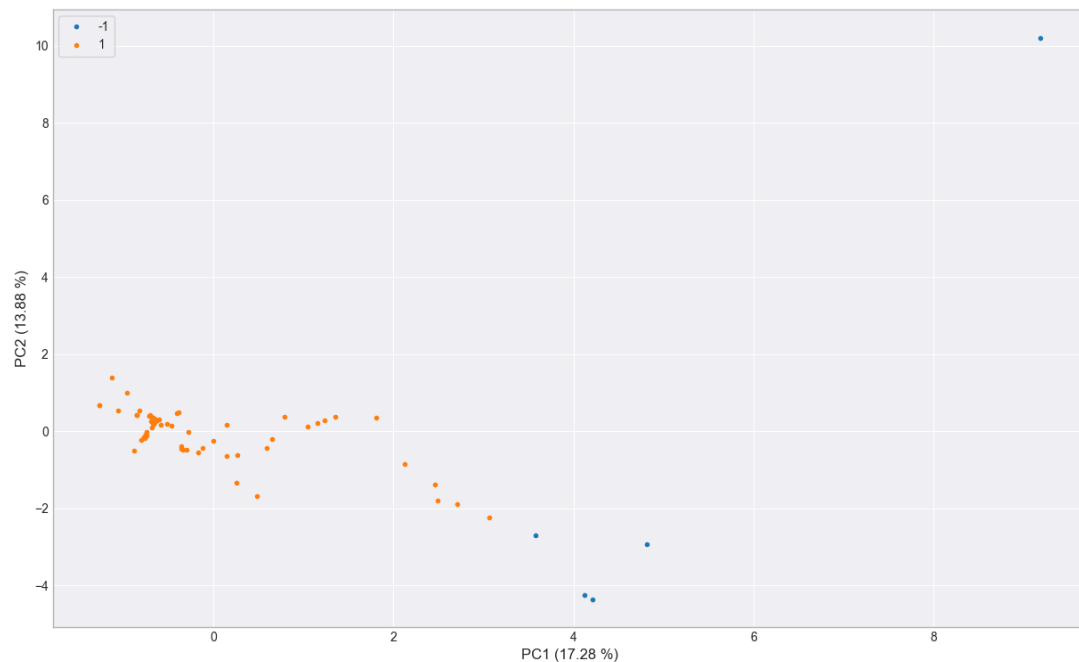


Figura 8 - Outliers (a azul) indicados pelo método de florestas de isolamento.

Pelo gráfico representado na Figura 8 obtido através do método referido, considera-se a existência de 5 *outliers*, que parecem coincidir com pontos que, graficamente, parecem também ser *outliers* no gráfico de projeção nas componentes principais. A Figura 9 representa a mesma projeção, sem os catalisadores que o método de florestas de isolamento, considerou serem *outliers*. Também na Figura 9, vê-se a mesma projeção com a adição da projeção das variáveis. Isto permite aferir quanto à representação das variáveis nas duas componentes principais representadas.

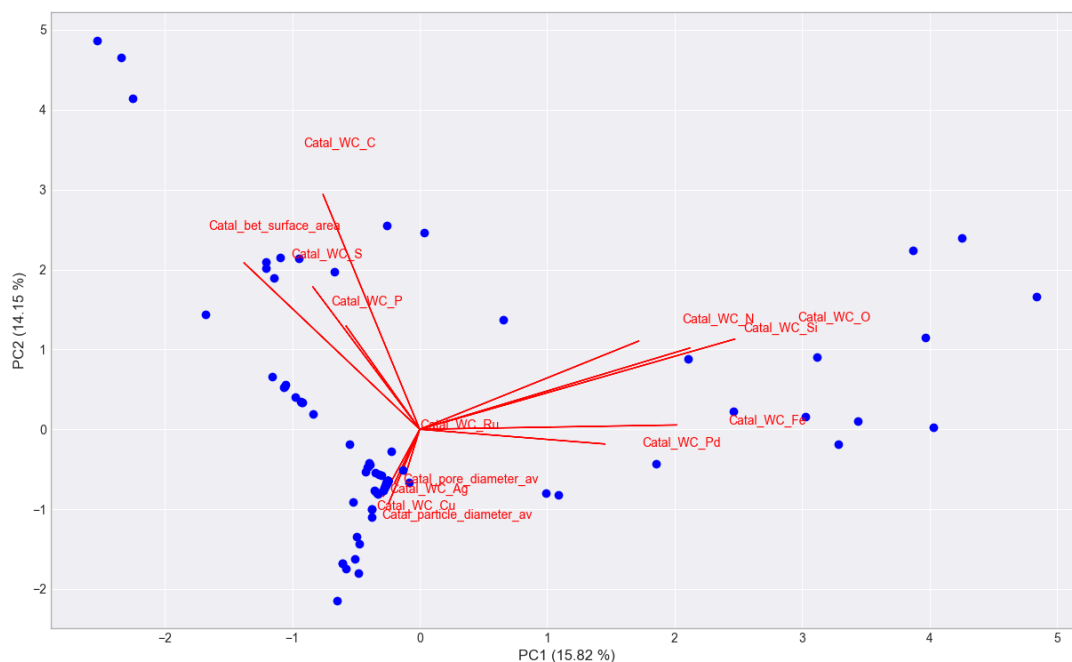


Figura 9 - Projeção das características dos catalisadores (sem outliers) nas primeiras duas componentes principais.

3.2. Caraterização do modelo de *Random Forest*

Uma vez concluída a caraterização dos dados iniciais, importa aferir também algumas conclusões retiradas pela análise do modelo construído. O modelo de *Random Forest* apresenta valores de coeficiente de determinação (R^2) para os conjuntos de treino e teste de 0.9242 e 0.8246, respetivamente. A Figura 10 apresenta os valores estimados pelo modelo para todos os casos e os valores de pk_1 reportados. Note-se que foi utilizado o valor de pk_1 como variável dependente, pois o modelo apresentava melhores resultados quando era aplicado o simétrico do logaritmo do k_1 .

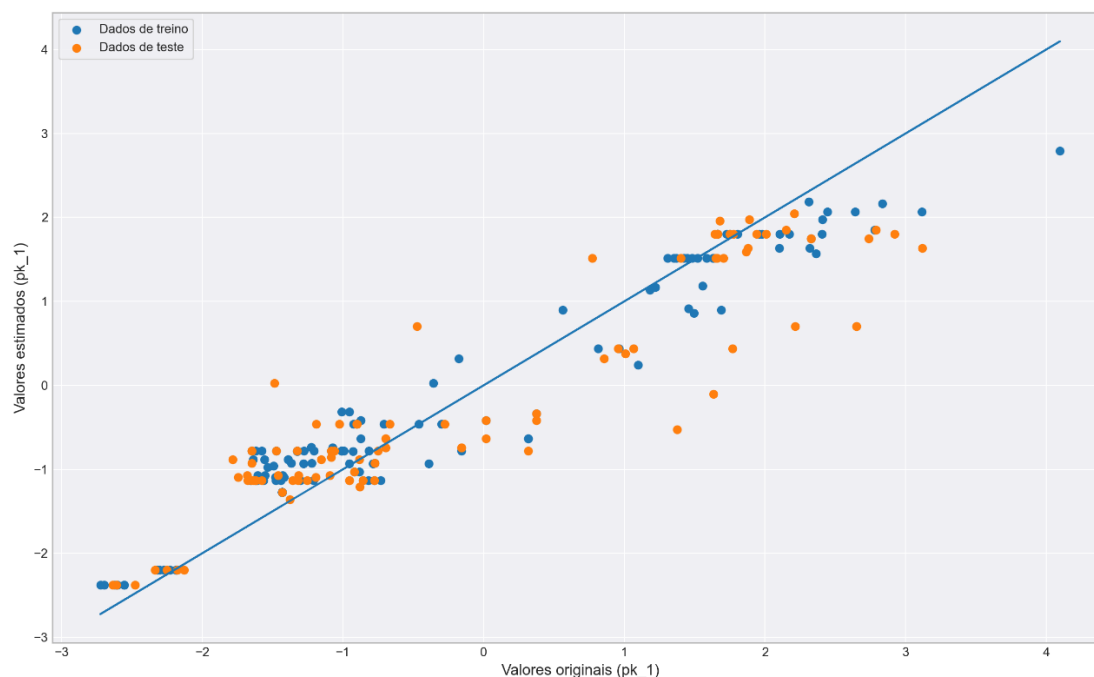


Figura 10 - Dados originais de pk_1 e valores previstos pelo modelo.

De seguida, foi realizado o levantamento das variáveis mais importantes no valor obtido, recorrendo-se à técnica de permutação.

Tabela 3 - Importância por permutação das 10 variáveis mais relevantes do modelo de Random Forest.

Variável	Média IP	Desvio-padrão IP	Importância de permuta (%IP)
% _o	0.09903	0.01213	17.1
m_{cat}	0.05616	0.01322	9.7
XRD030	0.0468	0.0124	8.1
D_{part}	0.0356	0.0063	6.2
m_{sub}	0.03316	0.0031	5.7
t	0.03222	0.00279	5.6
V	0.02974	0.00750	5.1
XRD031	0.02388	0.00438	4.1
% _N	0.01917	0.00101	3.3
% _C	0.01829	0.00147	3.2

A Tabela 3 representa os valores médios, e respetivos desvios-padrão do valor de importância de permuta (*Permutation Importance*; IP) para as 10 variáveis mais importantes do modelo de RF fornecido. Apresenta também o valor médio da IP em percentagem da soma das IPs de todas as variáveis, que de acordo com o método, representam 68.1% da importância de permuta total. A informação acerca das variáveis com mais impacto na resposta do modelo é relevante para que depois seja feita a apreciação da dependência parcial de cada uma e de alguns dos pares, de modo que não seja necessário fazê-lo para todas as variáveis. Pela Tabela 3 observa-se que em termos de condições, a quantidade de catalisador e substrato é importante na velocidade da reação, bem como o tempo de reação e o volume de solução. Já as características do catalisador mais relevantes são as percentagens de oxigénio, azoto e carbono na composição do material catalítico, o diâmetro médio das suas partículas e dois picos de XRD, localizados a $2\theta = 30^\circ$ e $2\theta = 31^\circ$.

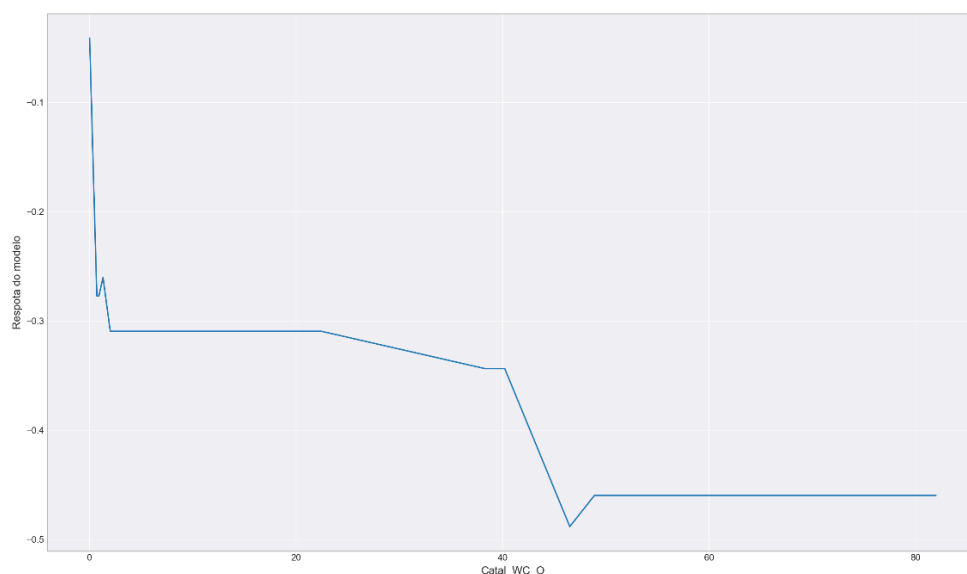


Figura 11 - Resposta parcial do modelo em função da percentagem de oxigénio no catalisador (%).

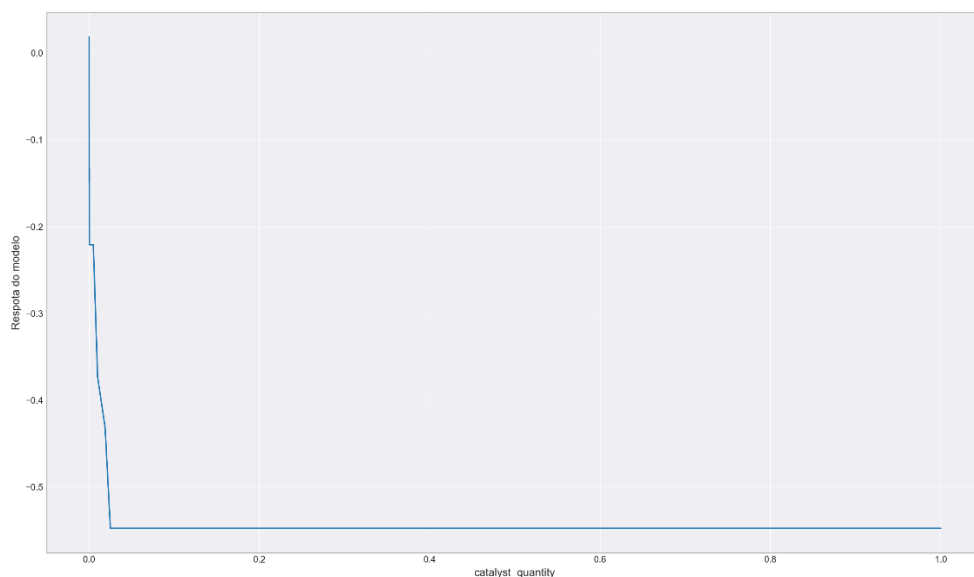


Figura 12 – Resposta parcial do modelo em função da massa de catalisador (g).

A Figura 11 representa a resposta parcial do modelo (pk_1) em função dos valores existentes nos dados, da variável $\%O_2$. Pelo gráfico vemos que cerca de 45% de composição de oxigénio poderá ser a composição ideal para um destes catalisadores, assumindo a não interferência das restantes variáveis. Já a Figura 12 demonstra que a quantidade de catalisador só pode diminuir o k_1 se não for utilizado catalisador suficiente. A partir de cerca de 0.02 g, adicionar mais catalisador não parece influenciar o k_1 .

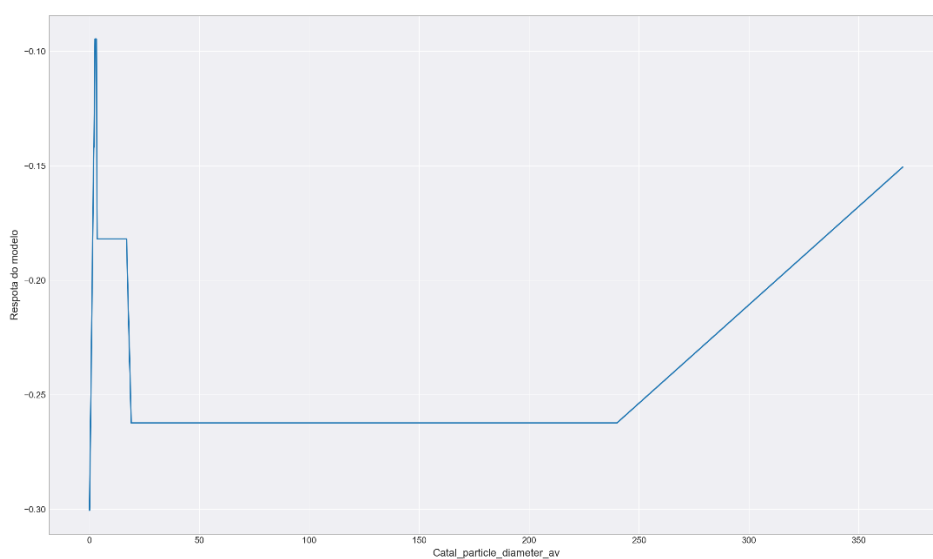


Figura 13 - Resposta parcial do modelo em função da média do diâmetro de partícula do catalisador (nm).

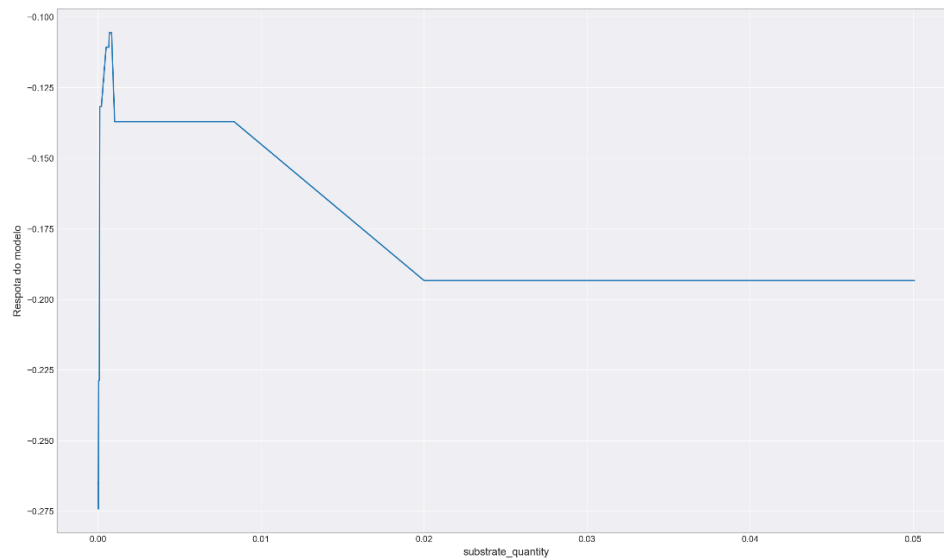


Figura 14 - Resposta parcial do modelo em função da quantidade de substrato (g).

Pela Figura 13 deduz-se que para valores muito pequenos de diâmetro de partícula obtêm-se os melhores valores de k_1 , contudo, pioram drasticamente com um pequeno aumento do tamanho, o que pode significar que isto se deve a uma incidência de melhores resultados para catalisadores nos quais o valor do diâmetro médio das partículas não estava definido. Da mesma forma a Figura 14, apresenta os melhores valores para quantidades muito baixas de substrato, diminuindo de forma abrupta o k_1 , mas retomando valores melhores para quantidades superiores de substrato.

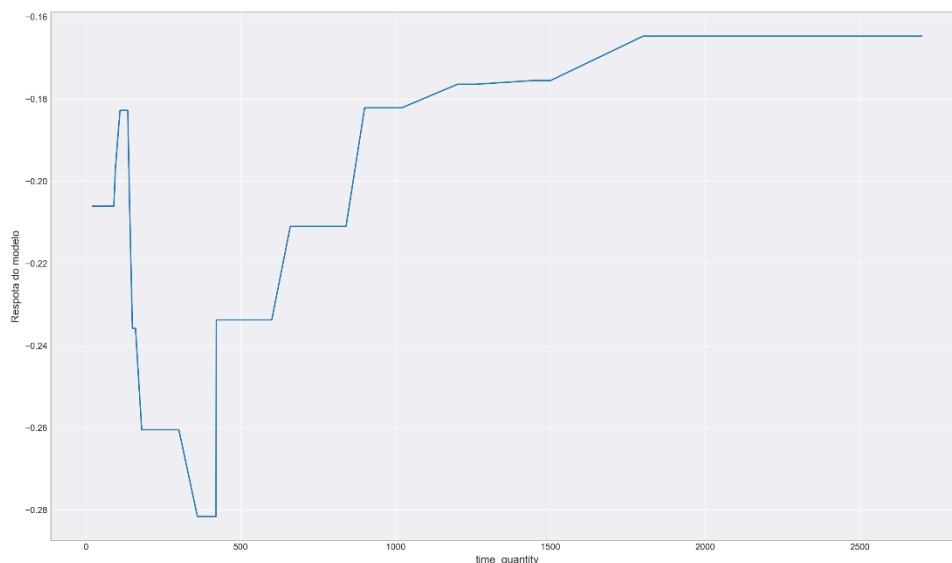


Figura 15 - Resposta parcial do modelo em função do tempo de reação (s).

O gráfico do tempo de reação é particularmente interessante, uma vez, que dá uma ideia quanto ao tempo em que a reação ocorre de forma praticamente completa. Mesmo sendo o k_1 calculado pelo declive inicial da curva de conversão de aminoarenos em nitroarenos, este aparenta ter um início de reação lento, o que faz com que tempos ligeiramente superiores permitam um melhor cálculo do k_1 . Pela análise da Figura 15 parece que o tempo ideal destas reações de redução se encontra entre os 400 e 450 s.

Na Figura 16, vê-se relatado o impacto do volume de solução no k_1 . O modelo parece indicar que a partir dos 44 mL se observa volume de solução suficiente para que a reação ocorra sob condições ideais.

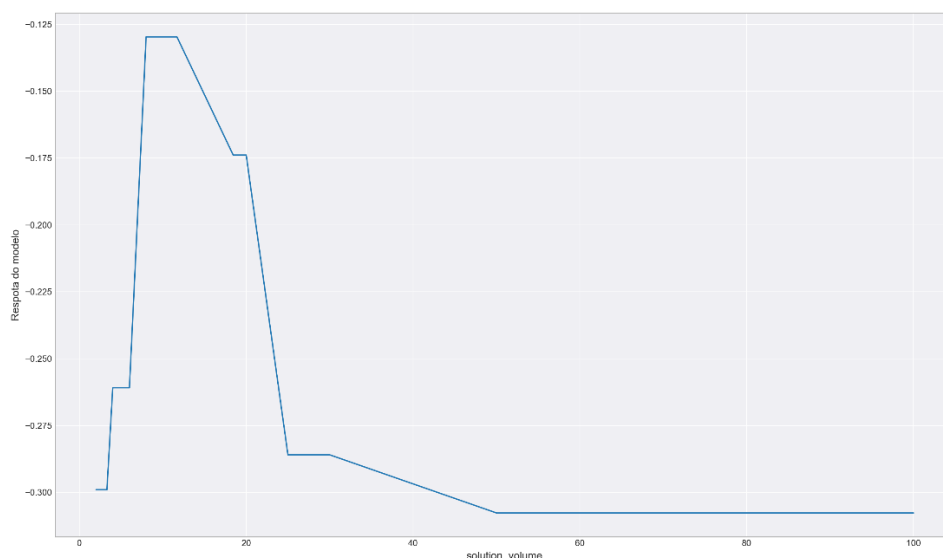


Figura 16 - Resposta parcial do modelo em função do volume de solução (mL).

A percentagem de azoto na composição do catalisador aparenta ter uma região ótima entre os 0,5% e 2%. Após a análise da Figura 17, indica que quaisquer outras percentagens de azoto presentes na composição conduzem a uma menor eficácia do catalisador. Exibindo um comportamento semelhante, a percentagem de carbono presente na composição do catalisador, origina uma zona ótima a cerca de 96 a 98%. Pela literatura é conhecido que, geralmente, catalisadores metálicos e, especialmente, compostos por metais nobres, são particularmente mais eficientes nesta reação de redução.⁷⁵ Neste aspeto, o presente trabalho introduz a possibilidade de que a utilização de catalisadores, cujo constituinte principal é carbono, na presente reação pode efetivamente, demonstrar-se mais eficaz do que o que a literatura relata.

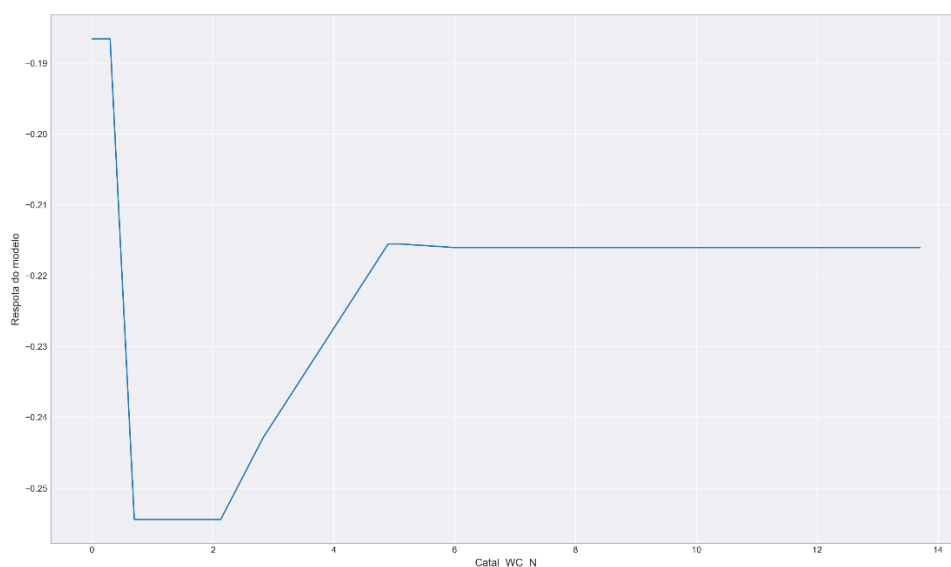


Figura 17 - Resposta parcial do modelo em função da percentagem de azoto na composição do catalisador (%).

Além das características já listadas, é também relevante analisar algumas características do catalisador que estavam nas 20 variáveis mais importantes, nomeadamente, S_{BET} e D_{pore} .

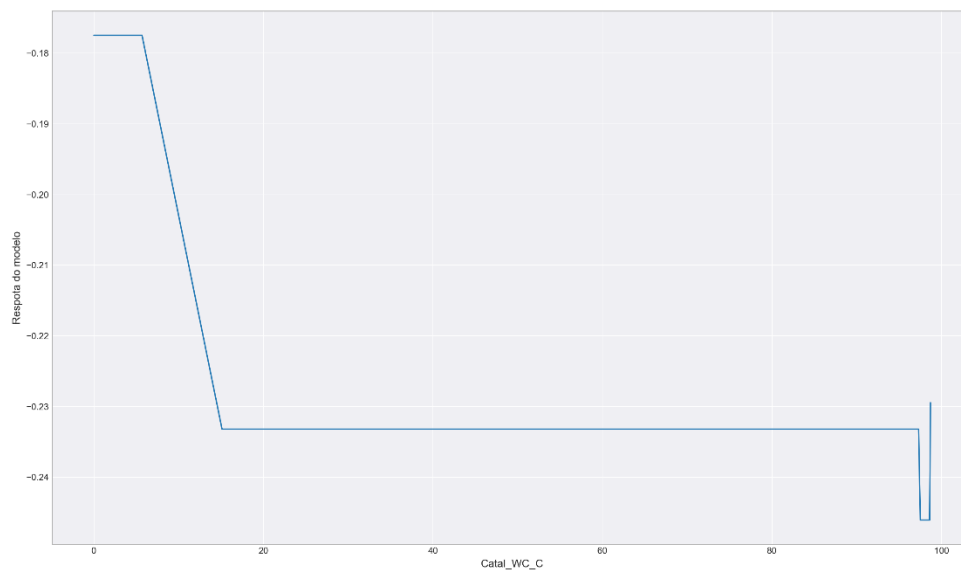


Figura 18 - Resposta parcial do modelo em função da percentagem de carbono na composição do catalisador (%).

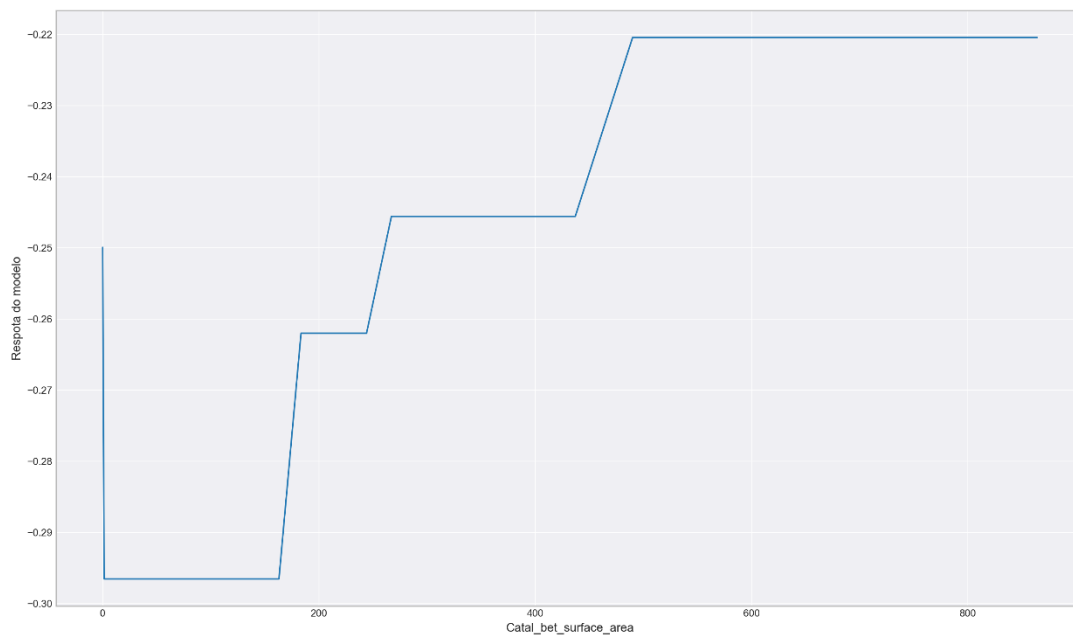


Figura 19 - Resposta parcial do modelo em função da área de superfície do catalisador (m^2/g).

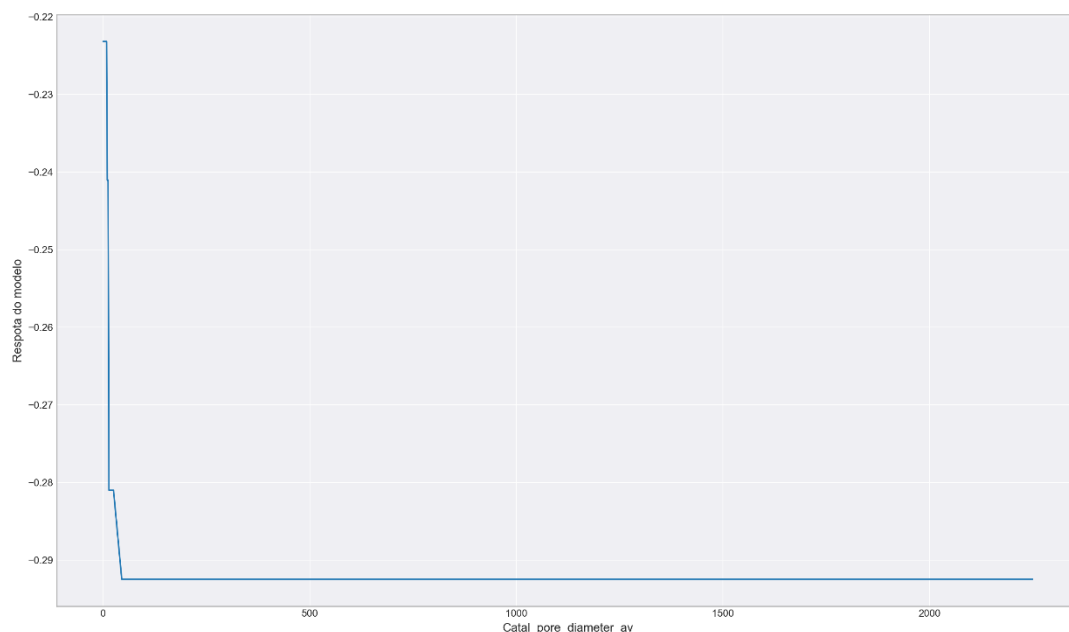


Figura 20 - Resposta parcial do modelo em função do diâmetro médio dos poros do catalisador (nm).

A Figura 19 indica que a área de superfície ideal de um catalisador para esta redução, será de, no máximo, 170 m²/g. Já o diâmetro dos poros do catalisador deve apenas ser superior a cerca de 50 nm, segundo a Figura 20.

É também relevante mencionar que ao longo da caracterização das dependências parciais do modelo, nota-se um decréscimo da amplitude do k_1 verificado nos gráficos. Isto reflete o menor impacto que cada variável vai tendo na resposta do modelo.

Uma vez terminada a caracterização dos dados e do respetivo modelo, resta agora discutir o comportamento dos otimizadores usados para alcançar o objetivo a que o presente trabalho se propõe.

3.3. Otimização de Monte Carlo

O trabalho da presente dissertação iniciou-se pela implementação do algoritmo de otimização de Monte Carlo (MC). Uma vez que é possível que o algoritmo tendo por base uma otimização de Monte Carlo tenha algumas variantes, estas variantes foram testadas e aferida qual a que melhor se adequava à função.

3.3.1. Otimização de parâmetros

Relativamente ao passo realizado em cada iteração, foram testadas três abordagens diferentes:

1. Por cada iteração era alterado o valor de cada dimensão iterativamente, de modo que no final fossem alterados todos os valores de todas as dimensões o mesmo número de vezes.
2. A segunda abordagem consistiu em alterar um valor de uma dimensão aleatória, fazendo com que por cada iteração fosse escolhida uma dimensão aleatoriamente, e cujo valor era alterado e avaliado o novo resultado.
3. Por fim, desenhou-se também um algoritmo que tinha em vista alterar os valores de todas as dimensões em simultâneo. Isto resultava na atribuição de um vetor aleatório que era somado ao ponto em que o otimizador se encontrava.

Cada uma destas abordagens do otimizador foi testada na função com um número reduzido de iterações para avaliar a forma como o algoritmo iria percorrer o domínio da função. Idealmente, o domínio da função seria percorrido com maior intensidade nos pontos que apresentavam valores de k_1 mais elevados.

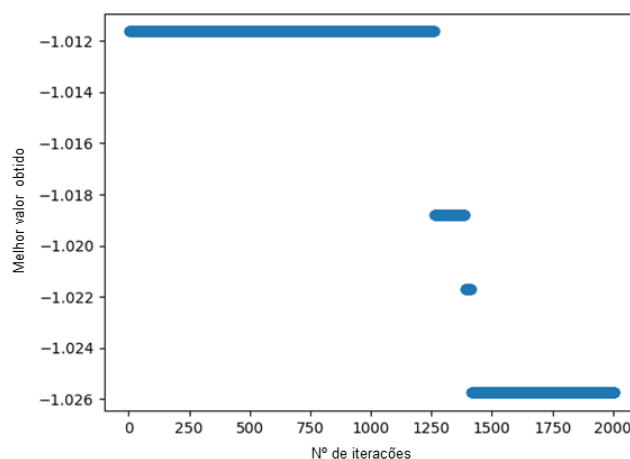


Figura 21 - 1ª abordagem de otimização, $P_{acpt} = 0.45$.

A Figura 21, representa o melhor valor de resposta da função encontrado pelo otimizador em função do número de iterações. Este gráfico pertence apenas à primeira

abordagem tendo sido testados vários valores de P_{acpt} para avaliar a performance do algoritmo em função do mesmo valor. Os gráficos para os restantes valores de P_{acpt} serão remetidos para anexo.

Ainda assim, é relevante referir que o algoritmo foi incapaz de melhorar o valor inicial nas primeiras 1000 iterações. Uma vez que a função sobre a qual é aplicado o otimizador tem nove dimensões, este pode ter alguma dificuldade em encontrar valores que reduzam a função de resposta. Isto pode levar a concluir que o algoritmo possa ser pouco eficiente na presente função.

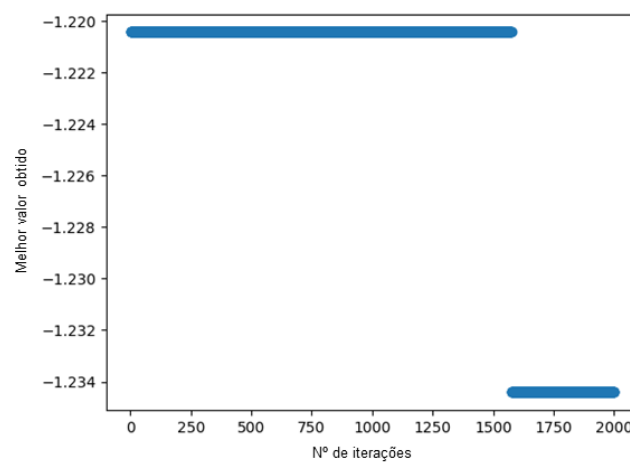


Figura 22 - 2ª abordagem de otimização, $P_{acpt} = 0.45$.

Na Figura 22 representa-se o gráfico das primeiras iterações da variante 2 mencionada anteriormente e os melhores valores de resposta encontrados para valores de P_{acpt} de 45 %. Ainda que a Figura 22 apresente um aumento do k_1 com menor número de iterações, tal como na abordagem anterior, nenhuma alteração no parâmetro apresentado parece causar um impacto muito relevante nesta variante do otimizador.

As restantes figuras até à Figura 24, representam os resultados dos ensaios exploratórios do otimizador de MC implementado de acordo com a variante 3. Na maior parte dos casos, esta parece oferecer uma solução que explora consideravelmente melhor o domínio da função uma vez que tanto a Figura 23 como a Figura 24, apresentam vários valores da função de resposta. De todas as formas, para um P_{acpt} de 45 %, o valor de resposta da função parece não só diminuir mais no total, mas também de forma mais gradual, o que indica que pode existir uma melhor exploração do domínio da função. Por este motivo, foi utilizado no algoritmo final um valor de P_{acpt} de 45%.

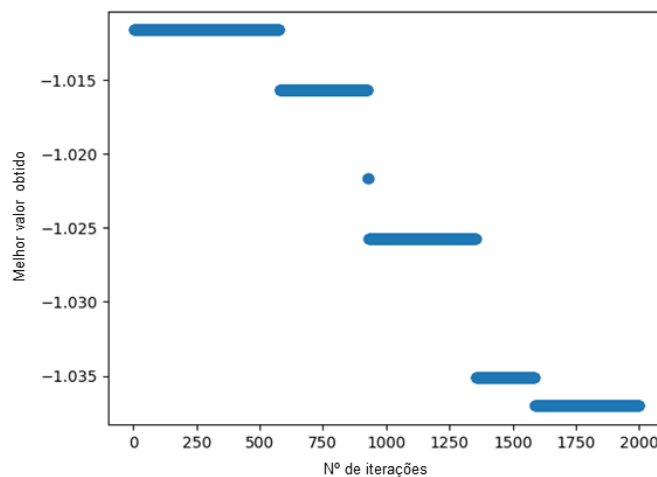


Figura 23 - 3ª abordagem do otimizador, $P_{acpt} = 0.45$.

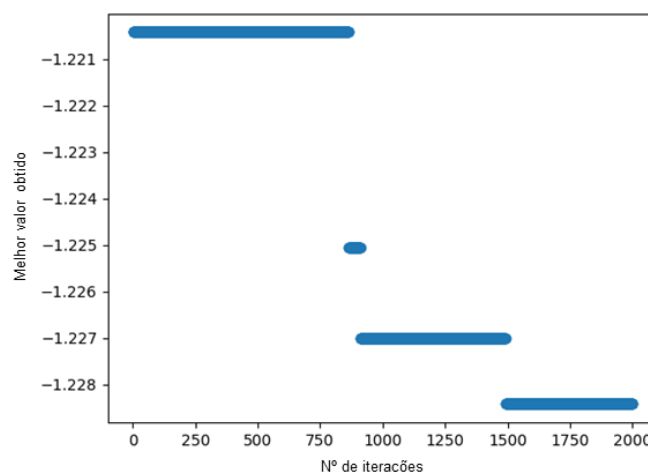


Figura 24 - 3ª abordagem do otimizador, $P_{acpt} = 0.80$.

3.3.2. Otimização de condições de reação

Depois de encontrados os melhores parâmetros para o otimizador, deu-se início às corridas de otimização das condições de reação para cada um dos 97 catalisadores catalogados na base de dados. Quando o otimizador não apresentou melhorias do k_1 em três corridas consecutivas, assumiu-se que as condições estavam otimizadas para esse catalisador. A melhoria do k_1 foi calculada de acordo com a Equação 7:

$$Melhoria = [10^{-pk_1}]_{ot} - [10^{-pk_1}]_{exp} = (k_1)_{ot} - (k_1)_{exp}$$

Equação 7 - Fórmula do cálculo da melhoria do k_1 após a otimização.

Pela Figura 25, o otimizador de Monte Carlo parece exibir um melhor desempenho para catalisadores metálicos do que para os de carbono equivalentes.

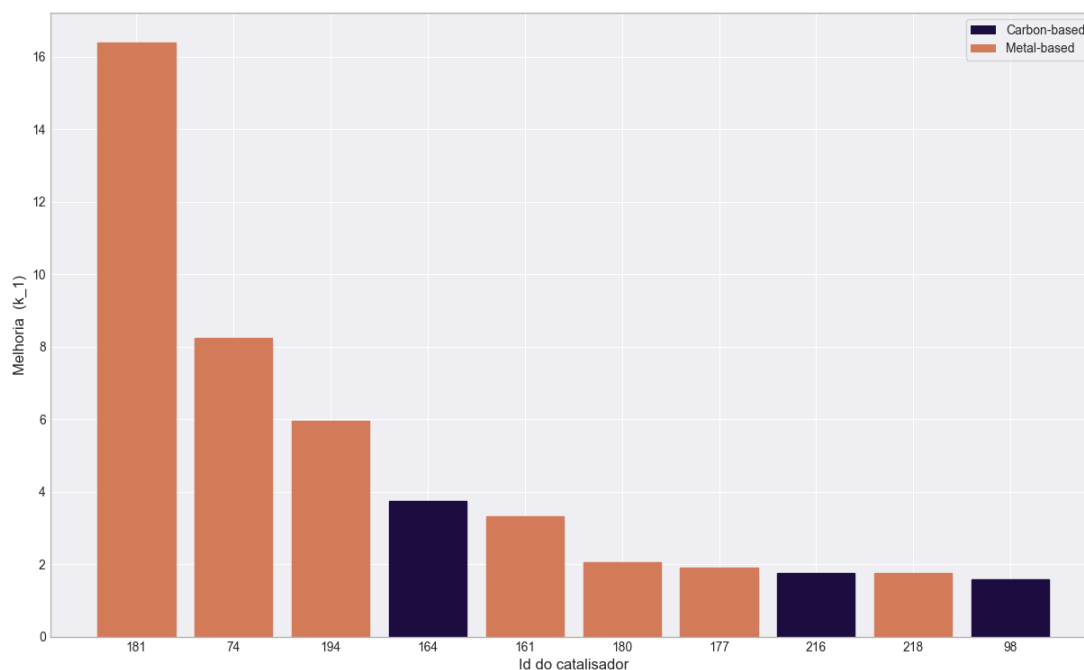


Figura 25 - Diferença entre o valor inicial e final de k_1 para os 10 catalisadores com maiores melhorias registadas após as corridas do otimizador de Monte Carlo.

Iniciou-se a caracterização dos resultados da otimização das condições experimentais usando o método de MC, pela matriz de correlação das condições de reação otimizadas, tal como demonstrado na Figura 26.

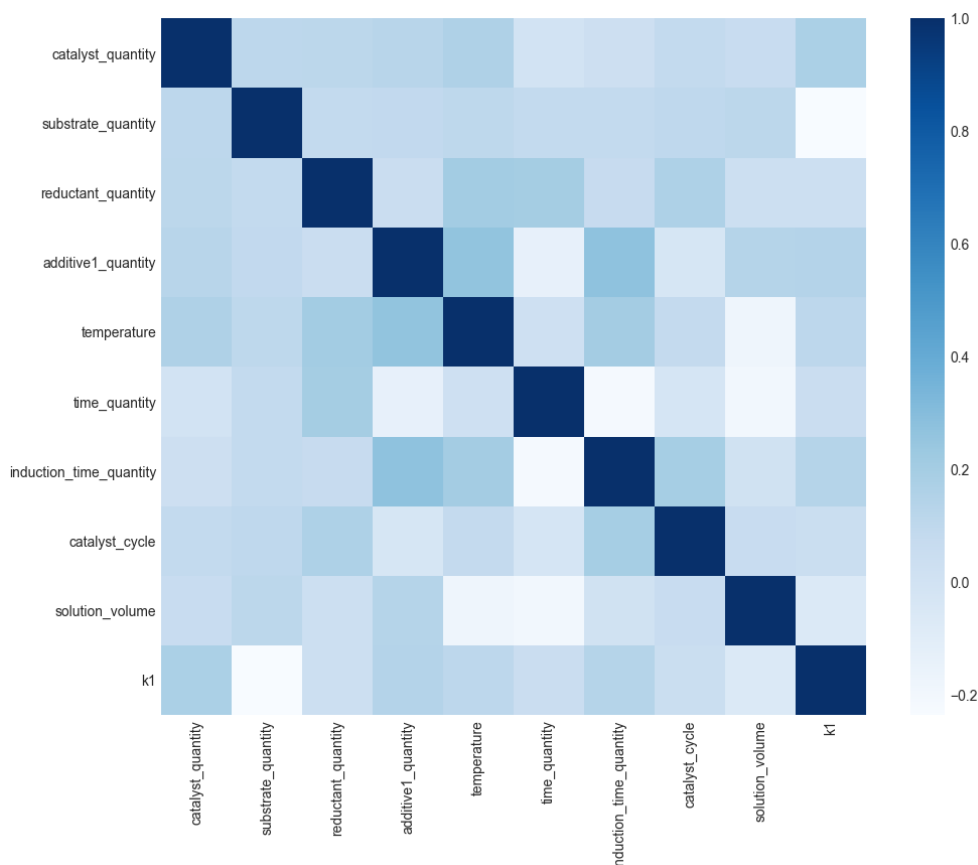


Figura 26 - Matriz de correlação das condições otimizadas e do k_1 .

A matriz de correlação apresentada na Figura 26, não apresenta grandes correlações entre as variáveis. Este resultado suporta a construção do modelo e do otimizador, uma vez que a eficiência de um catalisador não depende só das condições de reação, e não era espectável que as condições estivessem correlacionadas entre si.

3.3.2.1. Correlação das caraterísticas de catalisadores e condições ótimas

O objetivo do presente trabalho passa, principalmente, por otimizar as variáveis de maneira que estas se pudessem traduzir em condições e caraterísticas físicas passíveis de ser implementadas experimentalmente. Ainda assim, através de um processamento e tratamento de dados mais intensivo, foi desenhado, como objetivo secundário, a identificação de correlação entre as caraterísticas otimizadas, e as caraterísticas dos catalisadores. Para tal efeito, aglomeraram-se os dados das condições otimizadas com os dados das caraterísticas do catalisador, e a estes dados foram feitas novas análises de componentes principais.

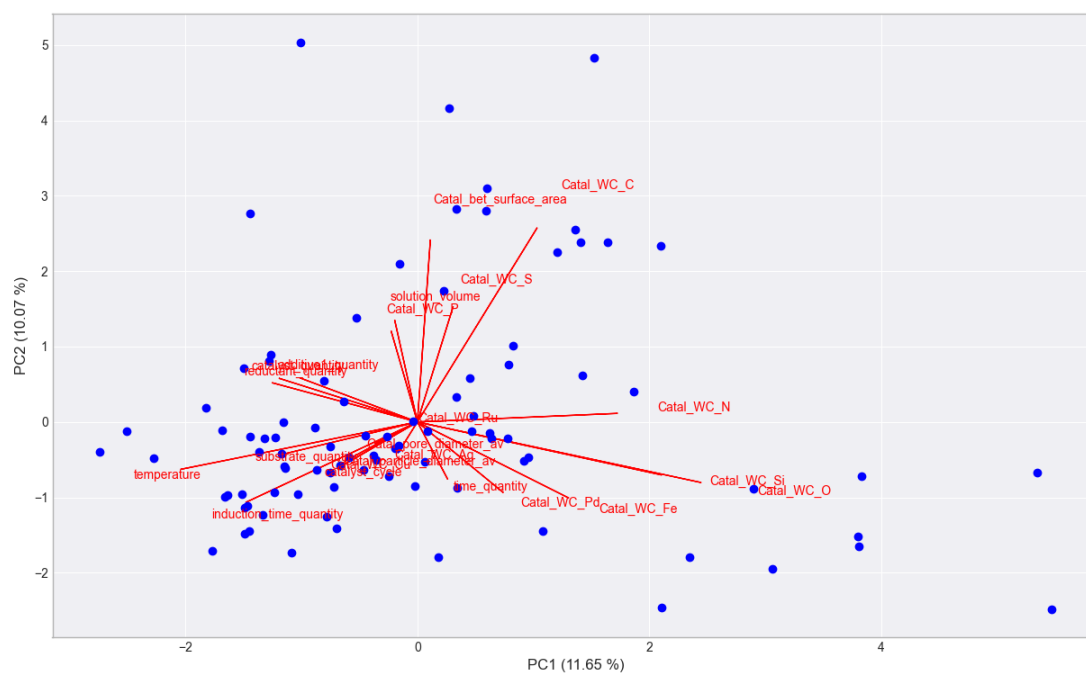


Figura 27 - Projeção dos catalisadores e condições otimizadas nas primeiras duas componentes principais.

A Figura 27 não demonstra uma grande aglomeração de dados, contudo, é relevante olhar para a matriz cumulativa de explicação de variância das componentes principais.

Tabela 4 - Variância explicada pelas componentes principais.

Componente principal	Variância cumulativa explicada	Componente principal	Variância cumulativa explicada
1	0.11655	12	0.76542
2	0.21722	13	0.80197
3	0.29889	14	0.83350
4	0.36687	15	0.86442
5	0.42964	16	0.89251
6	0.48609	17	0.91759
7	0.53884	18	0.94015
8	0.58940	19	0.95894
9	0.63703	20	0.97614
10	0.68252	21	0.98846
11	0.72681	22	1.00000

Dos valores apresentados na Tabela 4, verifica-se que duas componentes principais explicam pouco menos que 22% da variância total dos dados. Posto isto, e de forma a identificar correlações mais facilmente, começaram a ser utilizadas mais gráficos visando representar graficamente maior variância.

Para identificar as correlações inerentes aos dados seria necessário empregar um novo método que conseguisse fazer esse tipo de inferência. Uma vez que se pretendiam criar aglomerações de pontos de condições/caraterísticas foi empregue o método de *k*-médias aos dados apresentados.

Tratando-se este de um método não-supervisionado, para identificar o número de *clusters* recorreu-se ao “método do cotovelo”, cujo gráfico é apresentado na Figura 28.

Pelo gráfico da Figura 28 optou-se por escolher 16 *clusters* para o método de *k*-médias. Para evidenciar correlações foram desenhadas as projeções nas componentes principais e atribuídas cores diferentes a cada *cluster*.

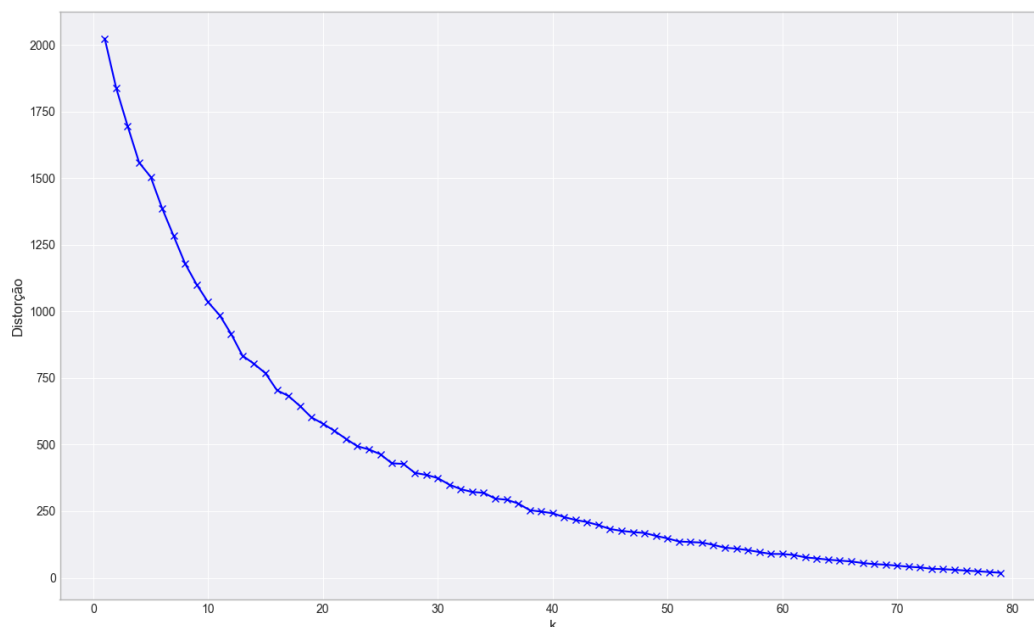


Figura 28 - Método do cotovelo, distorção do espaço pelo número de clusters de *k*-médias.

De forma a tentar explicar maior percentagem de variância, nas figuras seguintes até à Figura 30, encontram-se representadas as projeções em 8 componentes principais.

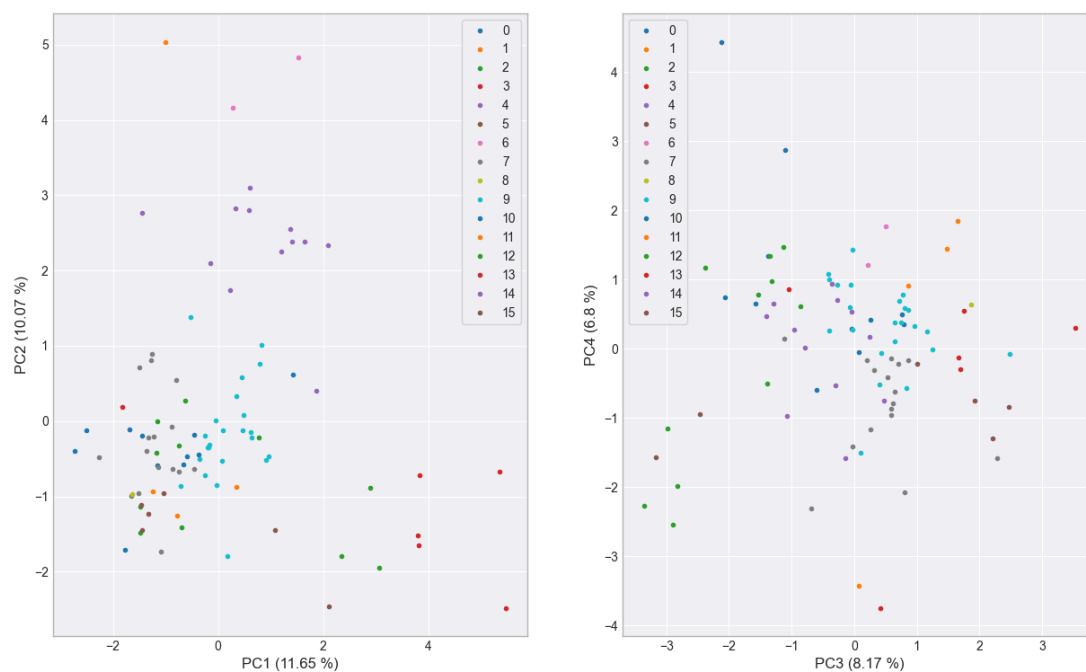


Figura 29 - PC 1 a 4 e representação dos clusters obtidos.

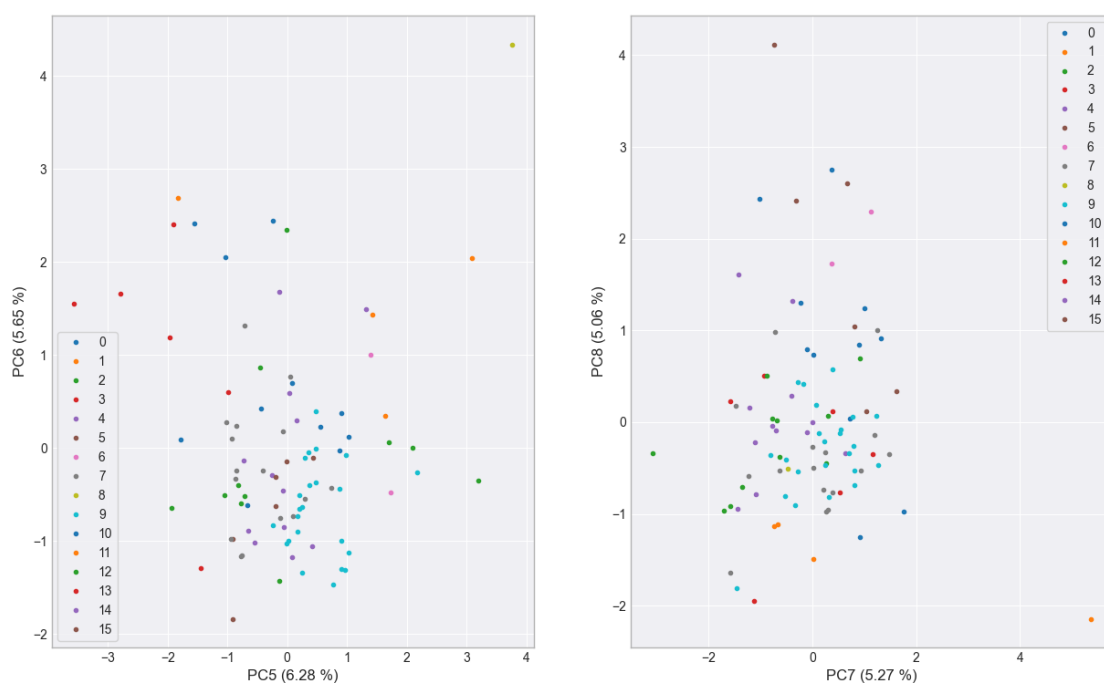


Figura 30 - PC 5 a 8 e representação dos clusters obtidos.

Pelos gráficos apresentados, observa-se que ainda que exista alguma separação dos *clusters*, alguns destes aparentam estar muito dissimulados entre si. Por

este motivo, e apesar de os métodos de aglomeração sugerirem o contrário, tentou reproduzir-se gráficos com menos *clusters*, para que todos fossem bem evidenciados. Como tal, a Figura 31 representa a mesma projeção com 5 clusters.

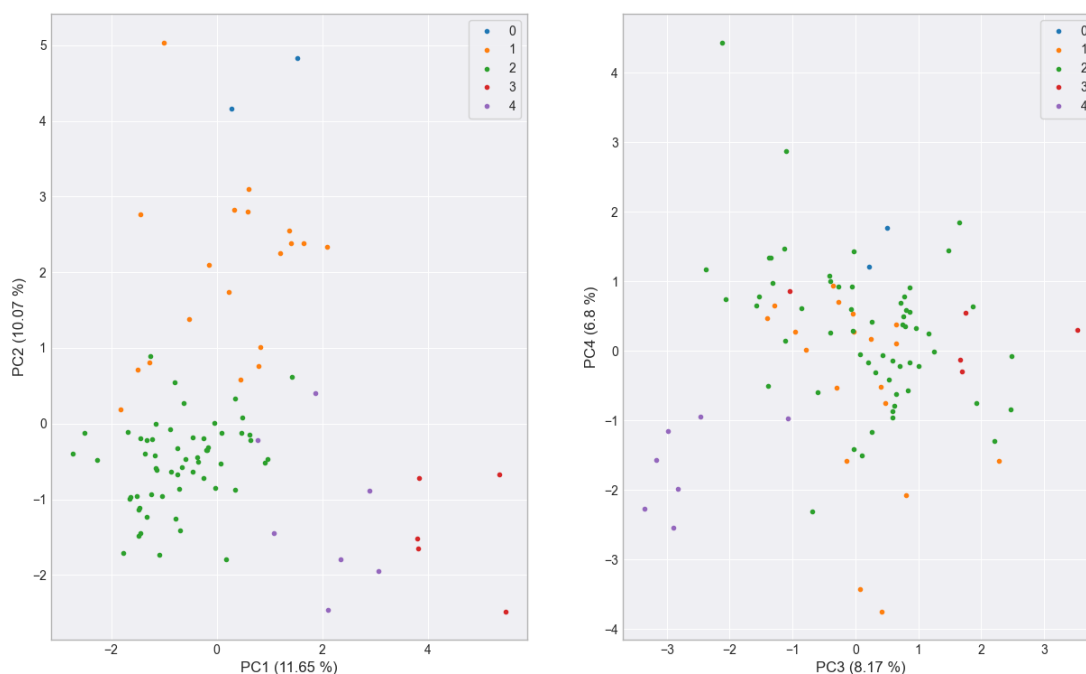


Figura 31 - Projeção dos dados nas primeiras 4 componentes principais, subdivididos em 5 clusters.

Apesar de ser muito provável que esteja a ser perdida informação em termos de aglomeração, agrupamento e generalização dos dados, é notória a melhoria em termos de separação gráfica mesma através das componentes principais. O motivo para não ser possível inferir mais acerca de um maior número de *clusters* pode passar pelo facto de existirem dados insuficientes acerca de outros *clusters*. Também é possível que outros métodos de *clustering* fornecessem mais informação.

Através da análise de *k*-médias, e da subsequente definição de *clusters* anteriormente apresentada, conclui-se que efetivamente os catalisadores foram efetivamente otimizados consoante a sua composição. Na Tabela 5 apresentam-se os centroides encontrados pelo método, e os respetivos pk_1 e k_1 .

Tabela 5 - Centroides dos clusters obtidos em k-médias.

Nº do <i>cluster</i>	0	1	2	3	4
m_{cat}	0.185	0.208	0.164	0.099	0.072
m_{sub}	0.013	0.008	0.006	0.006	0.007
m_{red}	0.202	0.054	0.181	0.054	0.090
m_{adt}	0.013	0.017	0.013	0.009	0.016
T	309.6	306.6	298.0	290.7	304.3
t	398	284	224	625	177
T_{ind}	147.6	51.8	57.8	73.6	69.9
cdr	2.6	1.2	2.0	1.9	3.2
V	27.67	35.67	44.28	21.54	34.77
D_{part}	21.39	4.75	4.12	7.40	5.53
D_{pore}	10.05	0.00	4.35	0.00	19.21
S_{BET}	41.87	325.50	373.75	12.77	34.01
% _{Ag}	0.11	0.00	0.01	0.00	0.00
% _C	0.00	97.85	55.02	16.22	0.00
% _{Cu}	4.36	0.00	0.39	0.23	0.00
% _{Fe}	0.03	0.00	0.00	4.91	30.14
% _N	0.00	0.00	0.96	2.26	4.76
% _O	0.07	1.35	2.77	50.14	10.80
% _p	0.00	0.00	0.01	0.00	0.00
% _{Pd}	0.29	0.00	0.00	0.00	3.39
% _{Ru}	0.00	0.00	0.00	0.00	0.00
% _S	0.00	0.80	0.00	0.00	0.00
% _{Si}	0.00	0.00	0.00	13.10	0.00
pk_1	-0.5180	-1.1636	-0.9180	-1.0914	-1.0738
k_1	3.296	14.576	8.280	12.342	11.852

Uma análise profunda à Tabela 5, permite aferir algumas propriedades relativamente às características dos catalisadores e às respetivas condições otimizadas. Em primeiro lugar, é relevante referir que os *clusters* parecem estar divididos também por composição de catalisador, isto é, por percentagens de elementos na sua constituição. Destes, separam-se os 5 grupos da seguinte forma:

- 0) *Cluster* nº 0 – Catalisadores não-orgânicos, cuja composição é desconhecida.
- 1) *Cluster* nº 1 – Catalisadores orgânicos cuja percentagem de composição é quase exclusivamente carbono.
- 2) *Cluster* nº 2 – Catalisadores orgânicos ou organometálicos cuja percentagem maioritária é carbono, mas alguma da restante composição contém elementos não listados.
- 3) *Cluster* nº 3 – Catalisadores orgânicos ou organometálicos cuja percentagem maioritária são sílicas, mas alguma da restante composição contém elementos não listados.
- 4) *Cluster* nº 4 – Catalisadores metálicos cujo elemento predominante conhecido é ferro.

Na mesma tabela, apresentam-se os resultados de k_1 estimados para cada um dos centroides. Naturalmente, estas estimativas são apenas indicativas, uma vez que não dizem respeito a catalisadores concretos. Além do valor de k_1 da reação existem outros fatores que devem ser considerados para a preferência de um catalisador em relação a outro, como por exemplo a possibilidade de reutilização de um tal catalisador ou a temperatura ótima a que estes funcionam.

Catalisadores compostos essencialmente por carbono (*cluster* nº 1) aparentam requerer as quantidades mais elevadas de catalisador, mas, por outro lado, funcionam bem com poucas quantidades de redutor. Também parecem requerer temperaturas ligeiramente superiores à temperatura ambiente, e em simultâneo, ser os menos recicláveis. Este grupo de catalisadores apresenta também uma das maiores áreas superficiais. O k_1 , estimado pelo modelo do centroide dos catalisadores suprarreferidos, parece ser o mais elevado dos 5 grupos.

Catalisadores pertencentes ao grupo 2 precisam de uma quantidade elevada de redutor, quando comparados aos restantes grupos e também não admitem quantidades de substratos particularmente elevadas. É relevante referir que este grupo é o que

apresenta o volume de solução mais elevado, e mais próximo do valor considerado ótimo nas dependências parciais (Figura 16).

Os catalisadores do *cluster 3*, maioritariamente sílicas, apresentam as menores quantidades de todos os intervenientes da reação, bem como, o menor volume de solução. Estes catalisadores parecem ser os que necessitam de maior tempo de reação. São também os que possuem a maior superfície de contacto. Tal pode dever-se ao facto de a maior parte das sílicas catalogadas na base de dados serem nanopartículas.

Catalisadores metálicos, pertencentes ao grupo 4, para serem utilizados em condições ótimas, podem requerer temperaturas superiores à temperatura ambiente. Este tipo de catalisadores requer menor quantidade de catalisador, possuem as reações mais rápidas, assumindo os valores mais baixos de tempos de reação de todos os grupos, e parecem ser também os catalisadores mais passíveis de serem reciclados.

Finalmente, os catalisadores de outras composições, demonstram o menor valor indicativo de k_1 . Estes requerem as quantidades mais elevadas de redutor, mas também aceitam as quantidades mais elevadas de substrato. Parecem também recicláveis, exibindo o segundo maior valor de ciclos de catalisador.

3.3.2.2. Seleção de catalisadores e otimização das condições

De entre todos os catalisadores, foram selecionados seis que tivessem as características mais distintas entre si, e as otimizações de k_1 mais variadas. Foi tido em consideração nesta seleção, que os catalisadores fossem, contudo, suficientemente representados nos dados. De entre todos os catalisadores representados na base de dados foram selecionados os seguintes:

- Dos catalisadores sintetizados e caracterizados por Bruno Jarrais, Alexandra Guedes e Cristina Freire, foram selecionados os catalisadores denominados “GF_S6” e “MWCNT_S6” sendo estes compostos quase exclusivamente por carbono, e não se tendo observado uma melhoria muito acentuada pelo otimizador de Monte Carlo.²³ Estes catalisadores são doravante referidos como 0 e 18 respetivamente.
- Dos catalisadores preparados e caracterizados por Mariana Rocha *et al.* foi selecionado o catalisador ao qual é atribuído o nome “K10_A@Au” pelos autores. Este catalisador é um silicato e foi selecionado por possuir um k_1 consideravelmente inferior aos restantes, contendo nanopartículas de ouro na composição.¹⁸ Este catalisador é aqui descrito como 65.

- As características do catalisador referido aqui como 73, foram retiradas do trabalho de *Sravanthi et al.* do catalisador “B-ZVIN”. Este catalisador foi selecionado por ser um silicato e conter também metais na sua composição.
- Do trabalho de *Shouchun Ma et al.* retira-se a informação do catalisador “Pd-Fe₃O₄”, doravante reconhecido como 120. Este foi selecionado por se tratar de um catalisador metálico que viu uma otimização média.
- Dos catalisadores descritos por *Fang Lin et al.* foi selecionado e otimizado o catalisador “Au-Fe₃O₄”, identificado aqui como 181, também este um catalisador metálico, que viu a maior melhoria na corrida inicial do algoritmo de Monte Carlo.

A Tabela 6 representa o índice dos catalisadores selecionados e as respetivas características. Para estes catalisadores, o otimizador de Monte Carlo foi corrido novamente, e desta vez, foram registados os tempos de corrida do otimizador. A segunda característica a ser desconsiderada foi o ciclo do catalisador, pois em ambiente experimental, não fará sentido “gastar” catalisadores para atingirem apenas o ciclo ótimo e aplicá-los a partir desse ponto à reação pretendida, mas sim utilizá-los todos desde o momento que são sintetizados.

Tabela 6 - Características dos catalisadores selecionados.

Nº Catalisador	0²³	18²³	65¹⁸	73¹⁹	120²¹	181²⁰
<i>D_{part}</i>	0	9.5	0	30	4.8	4.5
<i>D_{pore}</i>	0	0	0	0	20.9	14
<i>S_{BET}</i>	341	310	0	62.47	56	0
% _{Ag}	0	0	0	0	0	0
% _C	97.1	98.6	38.3	5.68	0	0
% _{Cu}	0	0	0	0	0	0
% _{Fe}	0	0	0.3	20.13	28.5	59
% _N	0	0	5	0	4.9	0
% _O	2	0.7	38.6	38.28	10.9	22
% _P	0	0	0	0	0	0
% _{Pd}	0	0	0	0	10.6	0
% _{Ru}	0	0	0	0	0	0
% _S	0.9	0.7	0	0	0	0
% _{Si}	0	0	14.2	14.77	0	0

Esta seleção de seis catalisadores tinha por objetivo comparar os otimizadores entre si, sem recorrer ao dispêndio de tanto poder computacional como o que é necessário na otimização dos 97 catalisadores. Mais ainda, com as diferentes composições, pretende-se também realizar algumas comparações em termos de propriedades de catalisador e as suas condições ótimas.

A Tabela 7 e Tabela 8 apresentam as condições antes e após a otimização das condições dos catalisadores, respetivamente. Nos resultados desta segunda otimização, pode observar-se que os catalisadores tendem, para valores próximos dos que foram verificados anteriormente pela análise de k -médias. Isto é particularmente interessante, uma vez que foram retiradas algumas das variáveis, mas o modelo continua a atribuir melhores valores de k_1 para as condições estimadas. Isto parece indicar que as variáveis retiradas não têm, por este motivo, grande impacto na eficiência da reação.

Tabela 7 - Valores de condições de partida do otimizador.

Nº de ID do catalisador	0	18	65	73	120	181
m_{cat}	0.0100	0.0100	0.0030	0.0006	1×10^{-9}	0.0020
m_{sub}	0.0002	0.0002	0.0000	0.0008	0.0001	1×10^{-9}
m_{red}	0.0567	0.0567	0.0057	0.2270	0.0015	0.0006
m_{adt}	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
T	298.0	298.0	273.2	298.0	298.0	300.0
t	420	420	40	1200	360	55
V	50.00	50.00	50.00	60.00	3.00	2.16
k_1	9.9815	4.3822	0.1932	1.5830	12.0642	11.3527

Tabela 8 - Resultados da última corrida do otimizador de Monte Carlo.

Nº de ID do catalisador	0	18	65	73	120	181
m_{cat}	0.0100	0.5733	0.2757	1×10^{-9}	1×10^{-9}	0.0014
m_{sub}	0.0002	0.0199	0.0156	0.0005	0.0100	0.0022
m_{red}	0.0567	0.0602	0.0982	0.0592	0.3818	0.0160
m_{adt}	0.0000	0.0279	0.0202	0.0096	0.0046	0.0033
T	298.0	308.8	317.0	309.6	330.5	330.1
t	420	302	2329	167	153	2605
V	50.00	32.92	30.53	19.28	31.49	95.99
k_1	9.9815	12.8086	0.4349	3.1189	14.6454	15.6243

O catalisador número 0 não verificou qualquer melhoria, de facto, as condições permaneceram inalteradas. Isto pode dever-se a qualquer tipo de erro que tenha ocorrido durante a programação da otimização das condições, ou porque o otimizador não foi, efetivamente, capaz de encontrar um valor melhor para as condições deste catalisador. Já o catalisador número 18, cuja composição é maioritariamente carbono, viu enaltecida a quantidade de catalisador, bem como a quantidade de substrato. A quantidade de redutor aumentou apenas ligeiramente, tal como foi previsto pelo modelo anterior.

Para o catalisador nº 65 o comportamento parece assemelhar-se mais ao dos catalisadores sem composição conhecida. Pouco acima de 91% da composição deste catalisador é conhecida, logo não seria de esperar este tipo de comportamento pela análise de *clustering*. Contudo, nota-se que existe pouca informação adicional acerca deste catalisador, sendo que os valores das restantes variáveis apresentadas na Tabela 6 são iguais a zero, indicando que não existe essa informação acerca do catalisador. Isto pode ajudar a explicar um pouco o *cluster* 0 criado anteriormente, estando definido não só pela composição química do catalisador, mas também pelas suas propriedades

físicas. Não obstante, observa-se um comportamento que se assemelha ao do centroide descrito.

À exceção da temperatura, o catalisador 73 apresenta resultados que se aproximam muito do *cluster* 3, o grupo das sílicas, algo que é coerente com a composição do catalisador. Os restantes catalisadores (120 e 181) exibem um comportamento pouco semelhante ao *cluster* em que se deveriam inserir, o grupo de catalisadores metálicos.

É relevante referir que para todas as zonas ótimas verificadas na seção “Caraterização do modelo de *Random Forest*” pelo método de permutação, foram encontradas outras consideradas melhores pelo otimizador.

A Figura 32 apresenta as melhorias de k_1 em termos absolutos, dos seis catalisadores selecionados. É notável a diferença entre esse gráfico e o da Figura 25, que representa os resultados da aplicação do mesmo otimizador, mas a todos os catalisadores e com a otimização de mais variáveis. A remoção de duas variáveis, juntamente com a natureza aleatória do otimizador, pode causar estas diferenças significativas no comportamento do otimizador.

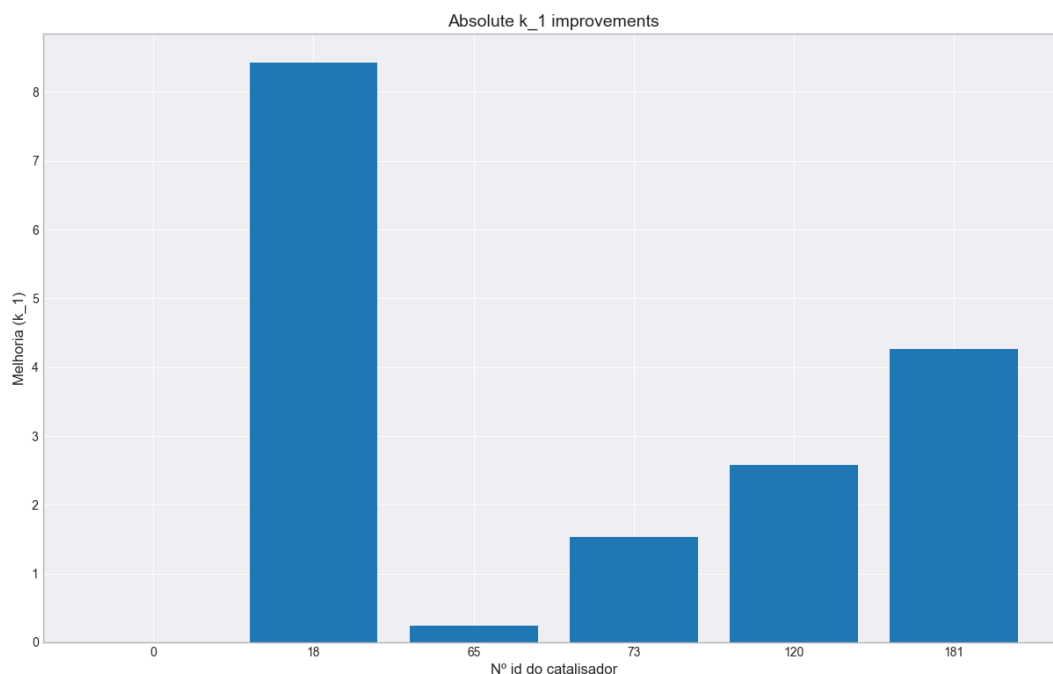


Figura 32 - Melhorias de k_1 para os catalisadores selecionados.

O primeiro ponto a referir acerca do gráfico da Figura 32, será a considerável melhoria do catalisador 18, que se trata de um catalisador de carbono, especialmente considerando que este catalisador não viu grandes melhorias na primeira otimização. Aparentemente, a eliminação de algumas variáveis permitiu, uma melhoria mais acentuada de catalisadores não-metálicos. Esta premissa torna mais plausível a hipótese de que nem o otimizador, nem o modelo, possuem uma tendência para aumentar artificialmente o k_1 de catalisadores metálicos. Não obstante, é de referir que todas as outras melhorias observadas, com a exceção do catalisador 18, foram maiores com o aumento da percentagem de ferro na sua composição.

Tabela 9 - Valores de k_1 estimados para as condições iniciais e otimizadas.

Nº de ID do catalisador	k_1 estimado nas condições utilizadas	k_1 estimado das condições otimizadas
0	9.9821	9.9821
18	4.3822	12.808
65	0.19330	0.43489
73	1.5831	3.1192
120	12.066	14.647
181	11.354	15.623

3.4. Otimização de enxame

Uma vez concluída a otimização de Monte Carlo, procede-se à análise da otimização através de algoritmos de enxame, de forma que se possam comparar os resultados, e verificar que as condições de utilização foram efetivamente otimizadas.

3.4.1. Otimização de parâmetros

A principal diferença entre estes otimizadores e o de Monte Carlo, reside no número de agentes e na sua correlação nos algoritmos de enxame. Devido à implementação do algoritmo de abelhas artificiais, 20 foi o menor número de agentes possível a utilizar, sendo corroborado na Figura 33.

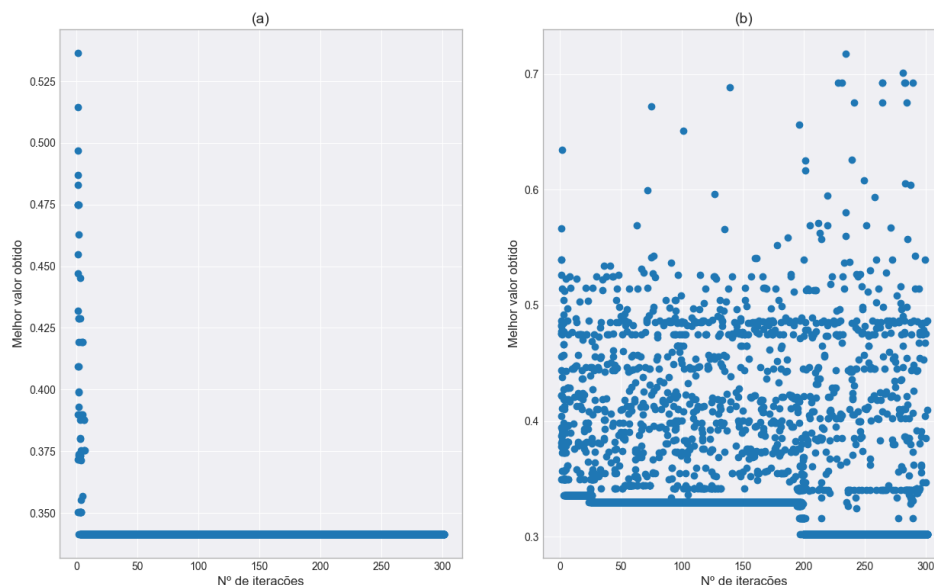


Figura 33 - Teste de otimização para: (a) 15 agentes; (b) 20 agentes.

Como se observa na Figura 33, o k_1 para 20 agentes é otimizado continuamente, ou seja, mesmo depois de ser encontrado um mínimo local, o otimizador continua à procura de valores melhores, uma vez que nem todos os agentes se encontram sempre no mesmo valor. Já para 15 agentes, nota-se que depois de uma melhoria inicial de k_1 , o otimizador estagna nesse ponto, não procurando mais nenhum em torno do melhor ponto encontrado. Isto deve-se à construção do otimizador que apenas aceita atribuir a função de abelhas batedoras a alguns agentes, a partir de um total de 20 agentes. Por este motivo utilizam-se 20 agentes nas otimizações das condições da reação para ambos os algoritmos de enxame.

3.4.2. Otimização de condições de reação – Abelhas artificiais

Os catalisadores a otimizar são os mesmos que foram utilizados com o otimizador de Monte Carlo, cujas características se encontram apresentadas na Tabela 6. Foram utilizadas três versões do algoritmo de abelhas artificiais: início aleatório; colmeia e colmeia dispersa.

3.4.2.1. Abelhas artificiais – Início aleatório

Começando pelo algoritmo de início aleatório, é importante mencionar que como o início não parte sempre do melhor ponto encontrado anteriormente, a última corrida

não apresentará necessariamente o melhor valor encontrado. Para tal, apresentam-se os gráficos na Figura 34 que representam as diferenças entre os resultados do melhor ponto encontrado pelas 10 corridas do otimizador e os pontos iniciais.

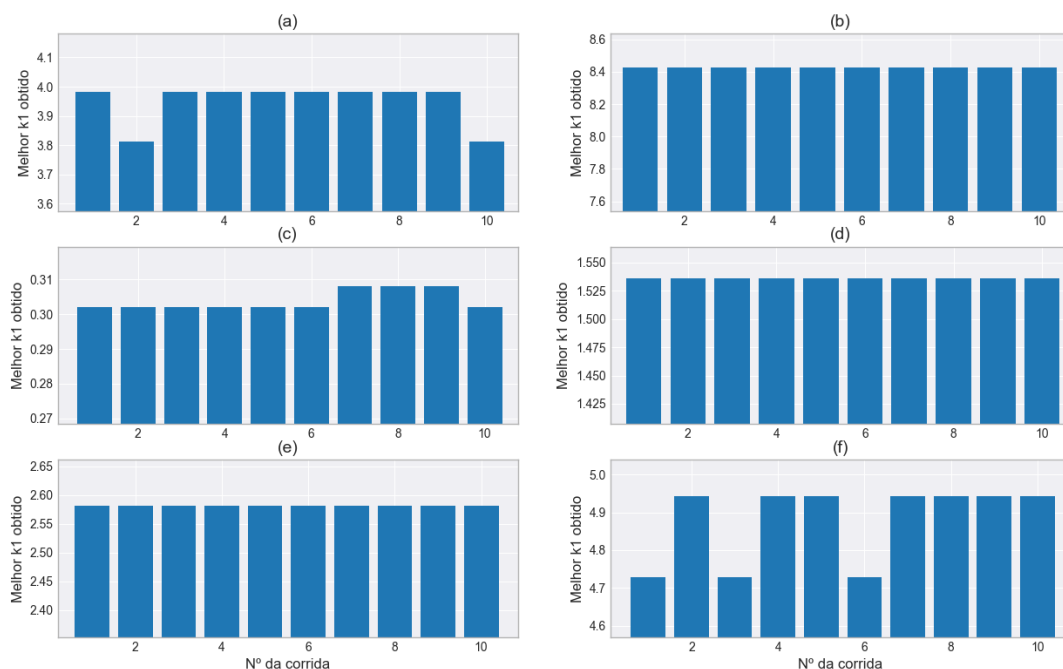


Figura 34 - Melhoria do k_1 obtido por corrida do algoritmo das abelhas artificiais para os catalisadores: (a) 0; (b) 18; (c) 65; (d) 73; (e) 120; (f) 181.

Pela Figura 34, observa-se que todos os catalisadores viram as melhores condições nas corridas 7, 8 e 9. Como tal, selecionar qualquer uma destas resultará no mesmo valor de k_1 . Contudo, é relevante referir que, tratando-se este de um modelo de *Random Forest*, os mesmos valores da resposta da função, podem ser obtidos para diferentes pontos de domínio da função.

Da Tabela 10 à Tabela 12 apresentam-se as diferentes condições em que foram obtidos os melhores valores de k_1 para todos os catalisadores. Numa primeira análise a todas as condições obtidas, observa-se que muitas podem diferir drasticamente de entre corridas. Contudo, numa análise mais profunda, regista-se que outras condições, como a quantidade de substrato, não diferem muito (exceto no caso do catalisador 65). Na maioria destas corridas notam-se desvios de algumas das variáveis, muito significativos em relação aos ideais assumidos pelos centroides de k -médias. É exemplo disto, a temperatura dos catalisadores metálicos (120 e 181) que apresenta, nestas condições otimizadas, valores inferiores à temperatura ambiente. Para estes catalisadores, nota-

se também que o tempo tende a corresponder àquele dos centroides, apresentando-se como o tempo de reação mais baixo.

Tabela 10 - Condições experimentais obtidas após a 7ª corrida do algoritmo de abelhas artificiais, para os catalisadores selecionados.

Nº de ID do catalisador	0	18	65	73	120	181
m_{cat}	0.848	0.294	0.664	1×10^{-9}	1×10^{-9}	1.27×10^{-3}
m_{sub}	4.32×10^{-4}	0.0441	4.17×10^{-7}	0.0501	3.39×10^{-3}	0.0369
m_{red}	0.0596	0.0652	1.14×10^{-6}	0.0940	0.169	0.0105
m_{adt}	0.060	0.047	0.043	0.060	0.002	0.042
T	283.5	299.3	313	287.1	273.2	293.6
t	167	302	595	218	101	20
V	30.52	34.90	2.45	19.54	98.35	29.13
k_1	13.9637	12.8089	0.5013	3.1189	14.6454	16.2967

Tabela 11 - Condições experimentais obtidas após a 8ª corrida do algoritmo de abelhas artificiais, para os catalisadores selecionados.

Nº de ID do catalisador	0	18	65	73	120	181
m_{cat}	0.404	0.857	1	1×10^{-9}	1×10^{-9}	1.05×10^{-3}
m_{sub}	3.90×10^{-4}	0.0375	4.17×10^{-7}	0.015	1.50×10^{-3}	0.0496
m_{red}	0.0452	0.0824	1.14×10^{-6}	0.0772	0.611	0.0104
m_{adt}	0.056	0.017	0.060	0.035	0.015	0.059
T	323.2	300.8	311.3	333	282.2	274.3
t	173	308	577	300	203	32
V	34.00	34.75	2.35	20.56	68.44	93.73
k_1	13.9637	12.8089	0.5013	3.1189	14.6454	16.2967

Tabela 12 - Condições experimentais obtidas após a 9ª corrida do algoritmo de abelhas artificiais, para os catalisadores selecionados.

Nº de ID do catalisador	0	18	65	73	120	181
m_{cat}	0.983	0.311	0.1016	1×10^{-9}	1×10^{-9}	1.11×10^{-3}
m_{sub}	3.69×10^{-4}	0.0493	1.70×10^{-6}	0.0501	5.01×10^{-2}	0.0218
m_{red}	0.0689	0.0837	2.21×10^{-5}	0.0643	0.329	0.0102
m_{adt}	0.019	0.009	0.023	0.025	0.023	0.000
T	328.7	300.9	314.7	305.5	320.1	273.3
t	165	311	537	358	20	21
V	32.97	30.53	2.33	25.07	28.77	33.70
k_1	13.9637	12.8089	0.5013	3.1189	14.6454	16.2967

A Figura 35 apresenta os valores de k_1 obtidos para a sétima corrida do algoritmo. Foram utilizados os valores da sétima corrida uma vez que estes demonstraram ter os maiores incrementos de k_1 , juntamente com os da 8ª e 9ª corrida. Os valores de incremento aqui apresentados são comparados às condições de reação utilizadas experimentalmente representados na Tabela 7.

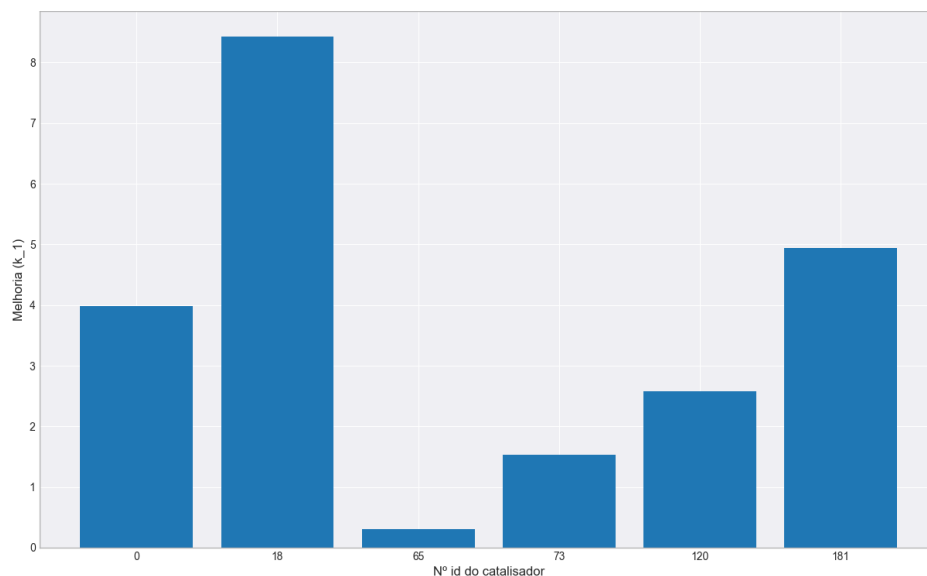


Figura 35 - Melhorias absolutas por catalisador do algoritmo de abelhas artificiais de início aleatório na corrida 7.

3.4.2.2. Abelhas artificiais – Colmeia

A versão da colmeia do algoritmo de abelhas artificiais, introduz um ponto inicial, que se traduz, habitualmente, pelo melhor ponto obtido em corridas anteriores. Na primeira corrida realizada, como este ponto não existe, o otimizador parte das condições utilizadas experimentalmente. Ao contrário da versão do algoritmo que tem inícios aleatórios, este vai apenas obter um ponto ótimo, uma vez que quando atingir o melhor valor de k_1 , não vai registar novos valores de condições que correspondam a um valor de k_1 idêntico.

Tabela 13 - Melhores condições obtidas pelo algoritmo de abelhas artificiais de início predefinido.

Nº de ID do catalisador	0	18	65	73	120	181
m_{cat}	0.060	0.604	1	1.0009	2.32×10^{-5}	1.40×10^{-3}
m_{sub}	4.31×10^{-4}	0.041	4.17×10^{-7}	0.0053	2.56×10^{-3}	0.0497
m_{red}	0.053	0.0734	1.14×10^{-6}	0.0757	0.321	0.01
m_{adt}	0.002	0.034	0.009	0.034	0.003	0.060
T	318.6	274.3	333	333	273.4	273.6
t	162	313	489	187	161	67
V	33.17	30.63	2.60	25.41	96.77	98.02
k_1	13.9637	12.8086	0.5013	3.1189	14.6454	16.2967

As melhorias do k_1 aqui observadas são exatamente iguais às da versão com o início aleatório e como tal, não será apresentado o gráfico de melhorias de k_1 . Já nas condições da Tabela 13, são observadas algumas diferenças para as condições anteriormente apresentadas, como os valores de temperatura. Por outro lado, existem também consistências às condições apresentadas até então, como o tempo de reação e o volume de solução.

3.4.2.3. Abelhas artificiais – Colmeia dispersa

O princípio deste algoritmo é em tudo semelhante ao seu par da versão de colmeia, contudo, acrescenta incrementos aleatórios ao inicial, ou seja, um dos pontos assume o melhor valor obtido, enquanto os restantes dispersam-se, de forma aleatória, em torno deste ponto inicial.

O conjunto de condições obtido por esta versão do algoritmo é apresentado na Tabela 14, é semelhante àquele da versão apresentada anteriormente. Isto pode sustentar a hipótese de que algumas condições devem ser mantidas para obter resultados ótimos, mas outras podem ser alteradas para determinados intervalos. Uma vez que também aqui, as melhorias são idênticas às apresentadas na Figura 35, o gráfico será omitido.

Tabela 14 - Condições ótimas obtidas pelo algoritmo de abelhas artificiais - colmeia dispersa.

Nº de ID do catalisador	0	18	65	73	120	181
m_{cat}	0.247	0.896	1	1×10^{-9}	1×10^{-9}	1.05×10^{-3}
m_{sub}	3.92×10^{-4}	0.038	4.17×10^{-7}	0.0302	4.23×10^{-2}	3.91×10^{-3}
m_{red}	0.0635	0.0952	1.14×10^{-6}	0.0778	0.677	9.64×10^{-3}
m_{adt}	0.007	0.041	0.006	0.054	0.007	0.059
T	305.5	332.6	323.8	273.2	328.5	273.8
t	155	306	599	243	152	71
V	29.22	34.60	2.09	7.65	85.75	86.34
k_1	13.9637	12.8086	0.5013	3.1189	14.6454	16.2967

3.4.3. Otimização de condições de reação – Lobos cinzentos

Aqui discutem-se as otimizações realizadas pelas duas versões do algoritmo de lobos cinzentos utilizadas na otimização de condições de utilização dos seis catalisadores previamente selecionados. Importa salientar, que neste otimizador não foi implementada uma versão em que os agentes partiam todos do mesmo ponto. Isto deve-se ao facto de que o papel desempenhado pelos agentes é apenas escolhido em função da sua posição relativa aos outros agentes. Assim listam-se as duas diferentes implementações: Covil que consiste em atribuir um agente igual ao melhor anteriormente obtido, ou às condições experimentais; Início aleatório (correspondente ao algoritmo original).

3.4.3.1. Lobos cinzentos – Covil

Esta versão do algoritmo difere da original apenas na inicialização do algoritmo em que este aceita um ponto ótimo do qual partir. Fornecido este ponto, o algoritmo força um dos agentes a tomar esta posição e os demais a situarem-se em torno do mesmo. O restante funcionamento do algoritmo é em tudo igual ao original.

A Tabela 15 apresenta as condições ótimas encontradas pelo otimizador.

Tabela 15 - Condições ótimas obtidas pelo algoritmo de lobos cinzentos – covil.

Nº ID do catalisador	0	18	65	73	120	181
m_{cat}	0.541	0.234	0.6347	1×10^{-9}	1×10^{-9}	1.02×10^{-3}
m_{sub}	4.34×10^{-4}	0.0267	2.09×10^{-5}	2.92×10^{-4}	1.10×10^{-3}	4.53×10^{-2}
m_{red}	0.0644	0.0724	1.14×10^{-6}	0.0458	0.015	1.39×10^{-2}
m_{adt}	0.032	0.006	0.024	0.015	0.003	0.003
T	273.2	278.2	317.9	274.9	333	275.8
t	276	315	490	250	259	54
V	29.03	29.70	2.61	10.80	30.21	46.95
k_1	13.9637	12.8086	0.5013	3.1189	14.6454	16.2967

Verificando uma melhoria, para todos os otimizadores, idêntica às verificadas pelo algoritmo de abelhas artificiais, dispensar-se-á o gráfico da melhoria do k_1 uma vez que será igual ao já representado na Figura 35.

3.4.3.2. Lobos cinzentos – Início aleatório

Este algoritmo corresponde à versão original do algoritmo *Grey Wolf Optimizer* tal como é fornecido no repositório de SwarmPackagePy. Denomina-se de início aleatório, tal como o original das abelhas, por não aceitar pontos iniciais e atribuir valores aleatórios iniciais a cada agente aquando da inicialização da função.

Tal como para o algoritmo das abelhas aleatórias é importante verificar qual (ou quais) das corridas tem os melhores resultados e podem, portanto, ser utilizadas.

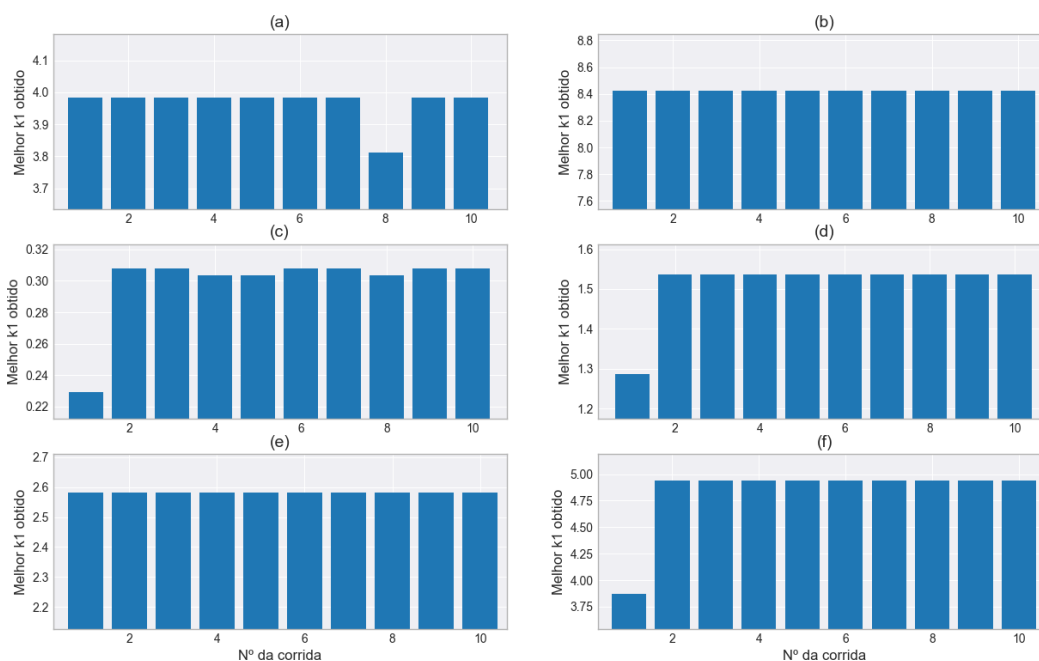


Figura 36 - Melhor k_1 obtido por corrida do algoritmo de lobos cinzentos para os catalisadores (a) 0; (b) 18; (c) 65; (d) 73; (e) 120; (f) 181.

Como se observa pela Figura 36, a maior parte das corridas apresenta os melhores resultados obtidos para todos os catalisadores em simultâneo, neste otimizador. Verifica-se então, que o melhor valor de k_1 será idêntico para as corridas 2, 3, 6, 7, 9 e 10.

Uma vez que existem diversas corridas a apresentar os melhores valores de k_1 obtidos, e que por sua vez podem apresentar diferentes condições, em vez de se apresentar todas as diferentes condições aqui obtidas opta-se por apresentar a média e o desvio-padrão de cada condição entre as corridas em que foram atingidas as condições ótimas. Tal como anteriormente, não é apresentado o gráfico de melhorias de k_1 , pois será idêntico ao da Figura 35.

Tabela 16 - Médias e desvios-padrões de todos os catalisadores para as corridas 2, 3, 6, 7, 9 e 10 da versão de início aleatório de lobos cinzentos.

Nº ID do catalisador		0	18	65	73	120	181
m_{cat}	μ	0.49	0.67	0.58	1×10^{-9}	1×10^{-9}	0.0012
	σ	0.31	0.32	0.33	0	0	0.0001
m_{sub}	μ	0.00041	0.035	4.17×10^{-7}	0.043	0.0266	0.030
	σ	2×10^{-5}	0.010	0	0.010	0.0063	0.014
m_{red}	μ	0.076	0.079	1.14×10^{-6}	0.068	0.31	0.0129
	σ	0.018	0.014	0	0.013	0.26	0.0015
m_{adt}	μ	0.017	0.017	0.017	0.046	0.029	0.012
	σ	0.017	0.012	0.005	0.014	0.013	0.016
T	μ	273.2	279.2	329.9	291.5	294.0	275.9
	σ	0.0	8.8	5.0	19.7	15.8	2.9
t	μ	186	305	538	323	69	64
	σ	52	3	25	59	71	38
v	μ	31.98	31.01	2.54	20.23	55.75	78.19
	σ	1.85	1.42	0.18	4.75	17.10	24.21

3.5. Comparação das condições otimizadas

Nesta secção não serão apresentadas novamente tabelas ou gráficos que tenham sido apresentados anteriormente. Sempre que possível é apresentado um novo item que permita a fácil verificação da análise a ser executada. Nos restantes casos, remeter-se-á às tabelas apresentadas em capítulos anteriores.

Pela Tabela 16, observa-se que a massa de aditivo, como era esperado, não parece ter um valor ideal. Uma vez que para que o aditivo surta efeito é apenas necessário que seja adicionado certa massa do mesmo, não é expectável ter um impacto muito significativo na resposta do modelo, o que ajuda a explicar a grande variância desta variável nos resultados de todos os otimizadores.

No caso da temperatura não se observam grandes desvios para os catalisadores **0**, **65** e **181**, enquanto os restantes parecem apresentar variações maiores, para os resultados deste otimizador. Comparando também com os resultados anteriormente obtidos, nota-se uma clara tendência ao longo de todos os otimizadores de enxame para cumprir estes intervalos de temperatura, no caso dos catalisadores **65** e **181**, havendo alguns desvios para o catalisador **65** no otimizador de início aleatório de abelhas artificiais. Já o catalisador **0** apresenta valores de temperatura totalmente díspares em todos os outros otimizadores, apresentando intervalos de 273 K a 323 K. isto pode indicar que a temperatura pode ser de maior importância para certos catalisadores e de importância muito baixa para outros. Ainda assim, é peculiar este otimizador em particular exibir evidente preferência deste catalisador para valores claramente inferiores de temperatura. Isto pode dever-se à topografia da função ser consideravelmente mais favorável a que os valores ótimos deste catalisador sejam encontrados em zonas de temperatura inferiores.

Para o tempo de reação nota-se que todos os catalisadores têm um intervalo ótimo de funcionamento. Para alguns o intervalo relativo pode ser maior, como o catalisador **120** que apresenta valores mínimos de 70 s e máximos de 277 s. Por outro lado, o catalisador **65**, apesar de apresentar um intervalo de valores mínimos e máximos ligeiramente menor (489 s a 620 s), em todas as corridas de todos os otimizadores, apresenta sempre os valores de tempo de reação mais elevados.

Na análise à Tabela 16, o volume de solução apresenta um desvio-padrão superior para catalisadores metálicos. Isto verifica-se analisando os resultados das condições obtidas pelos restantes otimizadores, em que os volumes de solução dos catalisadores **0**, **18** e **65** alteram-se muito pouco, enquanto os catalisadores **120** e **181** sofrem as alterações maiores. Esta premissa leva à conclusão de que o volume de solução pode ser um fator mais importante em catalisadores de carbono e sílicas, do que em catalisadores metálicos.

Concluída a análise a estas quatro variáveis resta executar a mesma análise às três restantes variáveis, referentes à massa dos constituintes da reação. Contudo, nota-se que estas variáveis assumem valores ótimos consideravelmente distintos. Por outro lado, também é relevante referir que o catalisador **73** apresenta variância 0 em todos os ensaios de todos os otimizadores para a massa de catalisador. Por outras palavras, sempre que foram atingidas as condições ótimas, a massa de catalisador era 1.0×10^{-9} g. Ainda assim, para ajudar a entender possíveis correlações entre as variáveis que pudessem ajudar a explicar estas variações, recorreu-se a gráficos de correlação de variáveis para as condições otimizadas no otimizador de início aleatório de lobos cinzentos.

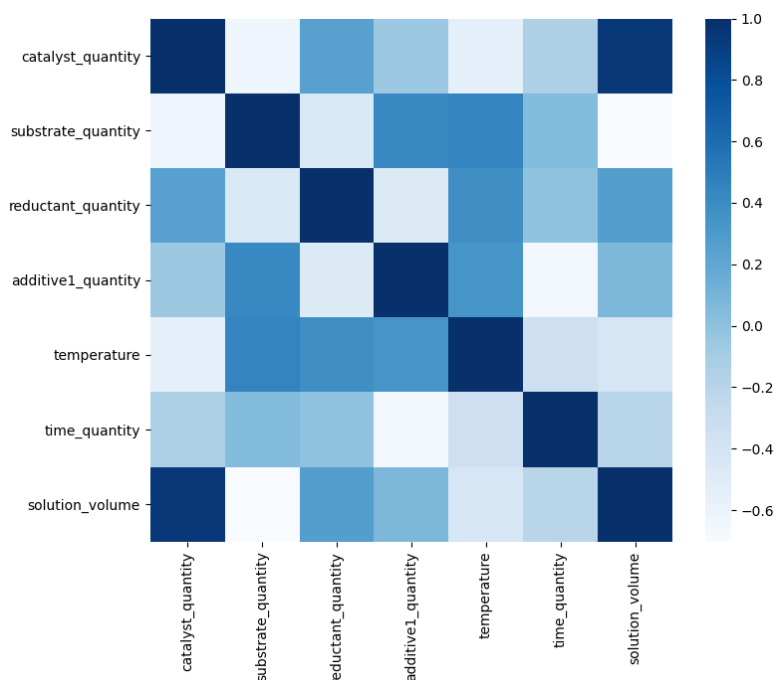


Figura 37 - Gráfico de matriz de correlação das variáveis obtidas pelo otimizador de lobos cinzentos de início aleatório para o catalisador **0**.

No gráfico da Figura 37 observam-se algumas correlações que podem explicar a variância da quantidade mássica de catalisador para o catalisador **0**. A correlação mais elevada será a da quantidade com o volume de solução. Esta correlação assume um valor de aproximadamente 0.94, um valor que não pode ser desconsiderado. Isto aponta para a hipótese de o otimizador não estar a tentar encontrar uma gama de quantidade mássica ótima, mas sim um valor ou intervalo de concentração de catalisador. Este facto indica que parece existir uma zona de concentração de catalisador ótima para a execução da reação. Pela análise dos resultados obtidos pelo algoritmo de abelhas artificiais, realça-se que com o aumento da temperatura observa-se uma tendência para a alteração da massa de catalisador ótima a adicionar, sem se verificarem grandes alterações no volume de solução. Isto pode alterar a hipótese previamente apresentada, e adicionar que a gama de concentrações ótimas pode variar largamente com a temperatura.

No caso deste catalisador, observa-se que tanto a massa de substrato como a massa de redutor não variam muito, nem dentro deste algoritmo de otimização nem nos demais. Isto indica que estas duas variáveis têm intervalos ótimos estreitos. Para a massa de substrato, o mínimo observado foi de 3.7×10^{-4} g e o máximo 4.34×10^{-4} g. Já a massa de redutor encontrou-se sempre entre 0.0530 g e 0.0689 g para o catalisador **0**.

O gráfico da Figura 38 representa a matriz de correlação das condições otimizadas pelo algoritmo de lobos cinzentos para o catalisador **18**. As correlações mais significativas aqui serão aquelas entre as quantidades de catalisador e de redutor; e a correlação inversa do tempo com o volume de solução. Ambas se verificam nos resultados de todos os outros otimizadores. A correlação entre as massas de catalisador de redutor pode ser explicada pela necessidade de uma quantidade de catalisador mínima mediante a presença de uma certa quantidade de redutor, de modo que todos os reagentes sejam catalisados.

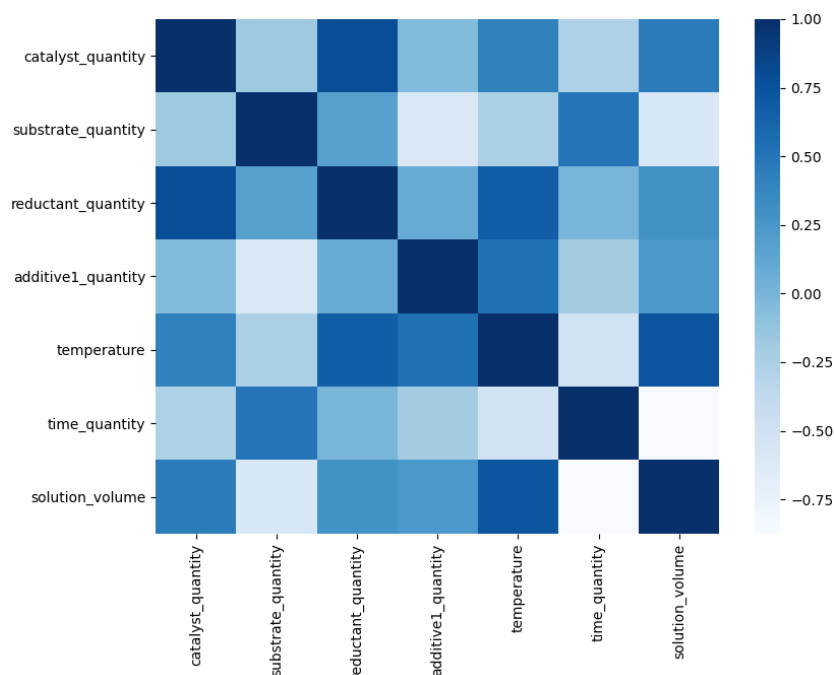


Figura 38 - Gráfico de matriz de correlação das variáveis obtidas pelo otimizador de lobos cinzentos de início aleatório para o catalisador **18**.

Para o catalisador **65**, a massa de catalisador não parece ter grande influência nos valores de k_1 encontrados. Essa tem na verdade uma grande variância e pequena correlação com as demais. Isto verifica-se em alguns dos restantes catalisadores, em que não parece existir correlação entre nenhuma outra variável como a quantidade mássica de substrato para o catalisador **73**, que possui elevada variância entre os diferentes ensaios de otimização, mas não apresenta correlação com outras variáveis. Alternativamente, a variável pode apresentar valores muito semelhantes ao longo de todas as otimizações, como a massa de catalisador para o **120**.

Tendo sido extraídos os resultados mais importantes destas otimizações, a nível de condições reacionais para cada catalisador, procede-se à comparação dos optimizadores para que se pudesse optar por qual, ou quais, utilizar na otimização das características do catalisador.

3.6. Comparação do processo de otimização

Nesta secção comparam-se os melhores valores de k_1 obtidos por cada optimizador, bem como os tempos de corrida de cada um.

Tabela 17 - Valores de k_1 estimados para as condições utilizadas experimentalmente e para as melhores condições encontradas pelos algoritmos de enxame.

Nº de ID do catalisador	k_1 estimado nas condições utilizadas	k_1 estimado das condições otimizadas
0	9.9821	13.965
18	4.3822	12.808
65	0.19330	0.50127
73	1.5831	3.1192
120	12.066	14.647
181	11.354	16.297

Pela análise da Tabela 9 e da Tabela 17, verifica-se que todos os otimizadores de enxame obtiveram os mesmos valores de k_1 para os seis catalisadores, indicando assim, que estes serão provavelmente os melhores valores de k_1 possíveis para estes catalisadores. Além de não se ter observado melhorias no catalisador **0** para o otimizador de Monte Carlo, as melhorias observadas para o **65** e para o **181**, são menores que aquelas realizadas pelos algoritmos de enxame. Tendo sido realizadas as mesmas avaliações de função por cada algoritmo, conclui-se que, para esta função, os algoritmos de enxame parecem funcionar melhor. Ainda assim, será importante aferir cada algoritmo quanto ao tempo de cada corrida.

Tabela 18 - Tempo médio de corrida de cada tipo de algoritmo.

	Tempo médio por corrida (horas:minutos)
Monte Carlo	11:15
Abelhas	31:09
Lobos	11:58

Pela Tabela 18, denota-se que o otimizador de Monte Carlo demonstrou ser o mais rápido. Contudo, uma vez que os resultados ficam aquém dos restantes otimizadores, opta-se por excluí-lo de análises subsequentes. Verifica-se também que os algoritmos de abelhas artificiais, têm tempos de execução consideravelmente mais longos. Por estes motivos, o único algoritmo utilizado para a otimização de características do catalisador é o de lobos cinzentos.

3.7. Otimização de características de catalisador

Para otimizar as características de catalisador, foram utilizadas as condições descritas na Tabela 1. Além das características, como já foi referido, o tempo de reação foi também otimizado, uma vez que não havia um valor específico que fosse mais conveniente. Realiza-se, então, a discussão para as variáveis referentes a características de catalisador.

Como os desempenhos previamente observados entre as versões do algoritmo de covil e início aleatório foram equiparáveis em termos de resultados obtidos, prosseguiu-se assim, com a utilização de ambas as versões.

3.7.1. Lobos cinzentos – Início aleatório

Tal como foi feito previamente, é necessário também aqui descrever qual foi o k_1 obtido por corrida, uma vez que todas partem de pontos aleatórios.

A Figura 39 mostra os valores de k_1 obtidos para cada substrato nas 10 corridas do otimizador de início aleatório. Ainda que seja notada alguma concordância entre os valores de k_1 obtidos, nota-se que o máximo foi atingido apenas numa corrida, indicando que existem melhorias que podem ainda ser efetuadas. Não obstante, deste máximo obtido apresentam-se as características na Tabela 19.

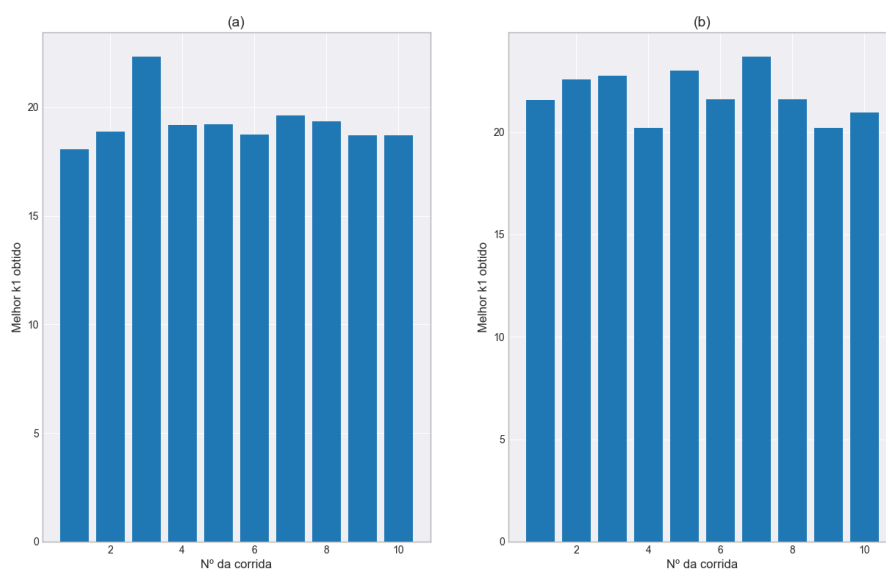


Figura 39 - Melhor k_1 obtido por corrida, para ambos os substratos (a) 2-Nitrofenol; (b) 4-Nitrofenol.

Tabela 19 - Características ótimas, para ambos os substratos, obtidas pelo otimizador de início aleatório.

Substrato	2-Nitrofenol	4-Nitrofenol
t	200	219
D_{part}	133	160
D_{pore}	419	0.00
S_{BET}	343	344
$\%_{Ag}$	0.00	0.00
$\%_C$	90.36	82.91
$\%_{Cu}$	0.00	0.00
$\%_{Fe}$	6.96	15.31
$\%_N$	0.76	0.68
$\%_O$	0.58	0.59
$\%_P$	0.00	0.00
$\%_{Pd}$	0.00	0.00
$\%_{Ru}$	1.34	0.51
$\%_S$	0.00	0.00
$\%_{Si}$	0.00	0.00
k_1	22.3310	23.6771

Pela análise à Tabela 19, os catalisadores ótimos obtidos parecem ser catalisadores organometálicos, compostos maioritariamente por carbono. Isto pode dever-se às condições pré-estabelecidas, uma vez que já tinha sido observado antes (Cf. Tabela 5) que este tipo de catalisadores parecia ser mais eficaz em condições de reação semelhantes.

Tabela 20 - Média e desvio-padrão de cada característica de catalisador para o 2-Nitrofenol, para todas as corridas

Variável	Média	Desvio padrão
t	249	98
D_{part}	119	29
D_{pore}	277	250
S_{BET}	245	137
$\%_{Ag}$	0.21	0.60
$\%_C$	9.46	26.99
$\%_{Cu}$	0.77	0.40
$\%_{Fe}$	14.87	18.30
$\%_N$	0.41	0.49
$\%_O$	20.32	25.34
$\%_P$	0.00	0.00
$\%_{Pd}$	0.00	0.00
$\%_{Ru}$	12.95	17.70
$\%_S$	2.83	5.35
$\%_{Si}$	29.89	21.49
k_1	19.2535	0.9475

Tabela 21 - Média e desvio-padrão de cada característica de catalisador para o 4-Nitrofenol, para todas as corridas.

Variável	Média	Desvio padrão
t	273	62
D_{part}	168	50
D_{pore}	271	319
S_{BET}	334	7
%Ag	0.02	0.06
%C	14.66	29.59
%Cu	0.82	0.71
%Fe	24.72	18.15
%N	0.81	0.47
%O	12.03	23.05
%P	0.00	0.00
%Pd	0.00	0.00
%Ru	10.21	13.66
%S	0.05	0.15
%Si	18.46	19.71
k_1	21.78	0.95

A esta informação é apenas relevante acrescentar que o tempo médio de corrida do otimizador foi de 3 horas e 35 minutos.

Este otimizador atinge, nas diferentes corridas, vários catalisadores diferentes com valores de k_1 semelhantes, como se depreende pela Tabela 20 e pela Tabela 21. De uma forma geral, todas as variáveis apresentam um desvio-padrão igual ou superior à média o que implica uma grande variação. A sua verificação experimental está, à data da escrita, pendente. Realça-se apenas que é provável, que mesmo que algumas destas características falhem ou não sejam possíveis de sintetizar, basta que um destes conjuntos funcione, para que o trabalho desenvolvido tenha também, o sucesso experimental comprovado.

3.7.2. Lobos cinzentos – Covil

Aqui apresentam-se apenas os resultados obtidos na última corrida, uma vez que serão os que apresentarão melhor valor de k_1 .

Tabela 22 - Características ótimas, para ambos os substratos, obtidas pelo otimizador de covil.

Substrato	2-Nitrofenol	4-Nitrofenol
t	181	350
D_{part}	292	150
D_{pore}	161	627
S_{BET}	344	344
% _{Ag}	0.00	0.01
% _C	0.00	0.64
% _{Cu}	0.00	0.93
% _{Fe}	5.62	4.27
% _N	1.17	0.02
% _O	0.51	41.79
% _P	0.00	0.00
% _{Pd}	0.00	0.00
% _{Ru}	4.52	1.96
% _S	0.00	0.39
% _{Si}	32.06	50.00
k_1	19.7775	22.9045

Pela Tabela 22 e também pela Tabela 19, observa-se que as melhorias registadas pelo otimizador de início aleatório são no total mais elevadas que aquelas vistas no otimizador de covil. Ainda assim, as melhorias médias do otimizador início aleatório são ligeiramente inferiores, apresentando um k_1 médio esperado de 19.2535 s⁻¹ para o 2-Nitrofenol e 21.7756 s⁻¹ para o 4-Nitrofenol. Mais ainda, o tempo médio das corridas deste otimizador foi de 3 horas e 42 minutos. Uma vez que demoram, sensivelmente o mesmo tempo, e para 10 corridas as melhorias mais significativas verificadas foram no algoritmo de início aleatório, conclui-se que este deverá ser o melhor algoritmo para esta função.

4. Conclusões

O trabalho propunha-se a realizar a otimização de condições de utilização de catalisadores, em modelos de *Random Forest*. Para este efeito, foram utilizados vários tipos de optimizadores e avaliados os diferentes resultados. Numa segunda fase, foi tido

por objetivo a otimização das características de catalisadores para um conjunto de variáveis.

Tendo-se dado início ao trabalho pela caracterização do modelo de *Random Forest* fornecido, verificam-se algumas das maiores importâncias parciais pelo método de permutação. Após as otimizações, confirma-se que a interação de outras variáveis afeta consideravelmente os intervalos considerados como ótimos pelo método de permutação.

Num plano secundário, foi aplicado o método de *clustering* de *k*-médias às condições otimizadas pelo método de Monte Carlo, juntamente com as características do catalisador. Isto resultou numa definição de 5 *clusters* visivelmente distintos quando projetados nas componentes principais. Estes *clusters* podem ajudar a definir conjuntos de condições ótimas mediante o tipo de catalisadores a utilizar.

Pela otimização aos seis catalisadores selecionados, observa-se que o otimizador de Monte Carlo, exibiu um desempenho ligeiramente pior que os otimizadores de enxame. Não obstante, foi suficientemente demonstrado que as condições de utilização dos seis catalisadores foram otimizadas, sendo improvável que exista algum ponto que apresente um k_1 mais elevado para qualquer um dos seis catalisadores considerados. Verifica-se aqui também, que otimizadores diferentes, podem tender para zonas diferentes do domínio da função, obtendo os mesmos valores da função de resposta. Inversamente, registam-se algumas características nas quais todos os otimizadores concordaram num determinado intervalo. Isto pode ajudar a adicionar informação às condições a utilizar para certos tipos de catalisador. Pelos resultados obtidos e pelos tempos de corrida médios dos otimizadores, conclui-se que o mais adequado, para este caso, é o de lobos cinzentos. A versão de início aleatório parece ser a que melhor funciona nesta função em particular.

Na otimização das características do catalisador, confirma-se que vários catalisadores diferentes podem originar resultados semelhantes em condições idênticas. Ao contrário do que foi feito para as condições de utilização dos catalisadores, não pode ser inferido que as características ótimas dos catalisadores foram encontradas, por haver poucas corridas a confirmá-lo. Para que isto seja possível deverão ser realizadas mais iterações pelos algoritmos, de modo a confirmar que o melhor valor de k_1 foi encontrado.

De todos os trabalhos que podem ainda vir a ser desenvolvidos no mesmo âmbito, tendo em consideração o descrito, o de maior relevância será a confirmação

experimental das conclusões retiradas, e subsequente verificação dos valores de k_1 previstos pelo modelo para essas condições/características.

Poderão eventualmente ser testados otimizadores que não foram testados no presente trabalho, e variantes dos mesmos. Utilizando os otimizadores descritos ou outros, será relevante produzir corridas suficientes, que atinjam os resultados ótimos para que se possa depois fazer uma análise estatística (ANOVA, por exemplo) para verificar se os resultados obtidos entre otimizadores são significativamente diferentes, ou se podem ser considerados parte do mesmo intervalo de zonas de variável ótimas.

Adicionalmente, uma análise semelhante àquela realizada para as condições otimizadas pelo método de Monte Carlo, poderá ser realizada para condições descritas por outro método. Isto permitirá aferir mais informação não só acerca dos intervalos das condições e das características, mas também do número possível de *clusters* existente e da razão de aglomeração desses mesmo *clusters*.

Bibliografia

- (1) Dickey, P. A. The First Oil Well. *J. Pet. Technol.* **1959**, 11 (1), 14–26. <https://doi.org/10.2118/1195-g>.
- (2) Clark, J.; Macquarrie, D. *Handbook of Green Chemistry and Technology*; Wiley Blackwell, 2007. <https://doi.org/10.1002/9780470988305>.
- (3) Frost, A.; Pearson, R. Kinetics and Mechanism, Second Edition. *J. Phys. Chem.* **1961**, 65 (2), 2–384. <https://doi.org/10.1021/j100820a601>.
- (4) Mourato, C.; Martins, M.; da Silva, S. M.; Pereira, I. A. C. A Continuous System for Biocatalytic Hydrogenation of CO₂ to Formate. *Bioresour. Technol.* **2017**, 235, 149–156. <https://doi.org/10.1016/j.biortech.2017.03.091>.
- (5) Begum, R.; Farooqi, Z. H.; Aboo, A. H.; Ahmed, E.; Sharif, A.; Xiao, J. Reduction of Nitroarenes Catalyzed by Microgel-Stabilized Silver Nanoparticles. *J. Hazard. Mater.* **2019**, 377, 399–408. <https://doi.org/10.1016/j.jhazmat.2019.05.080>.
- (6) Wong, Y. C.; Szeto, Y. S.; Cheung, W. H.; McKay, G. Pseudo-First-Order Kinetic Studies of the Sorption of Acid Dyes onto Chitosan. *J. Appl. Polym. Sci.* **2004**, 92 (3), 1633–1645. <https://doi.org/10.1002/app.13714>.
- (7) Lewtas, J.; Nishioka, M. G. Nitroarenes: Their Detection, Mutagenicity and Occurrence in the Environment. In *Nitroarenes*; Springer US, 1990; pp 61–72. https://doi.org/10.1007/978-1-4615-3800-4_5.
- (8) Xue, Y.; Wu, T.; Cui, Y.; Gong, B.; Li, X.; Qin, X.; Cao, X.; Liu, X.; Ai, Y.; Han, J.; Jin, T. Energy Consumption and Pollutant Emission of Diesel-Fired Combustion from 2009 to 2018 in Beijing, China. *J. Environ. Manage.* **2021**, 285, 112137. <https://doi.org/10.1016/j.jenvman.2021.112137>.
- (9) Laghrib, F.; Ajermoun, N.; Hrioua, A.; Lahrich, S.; Farahi, A.; El Haimouti, A.; Bakasse, M.; El Mhammedi, M. A. Investigation of Voltammetric Behavior of 4-Nitroaniline Based on Electrodeposition of Silver Particles onto Graphite Electrode. *Ionics (Kiel)*. **2019**, 25 (6), 2813–2821. <https://doi.org/10.1007/s11581-018-2735-8>.
- (10) Song, J.; Huang, Z. F.; Pan, L.; Li, K.; Zhang, X.; Wang, L.; Zou, J. J. Review on Selective Hydrogenation of Nitroarene by Catalytic, Photocatalytic and

- Electrocatalytic Reactions. *Applied Catalysis B*. Elsevier B.V. July 5, 2018, pp 386–408.
- (11) Wang, X.; Li, Y. Chemoselective Hydrogenation of Functionalized Nitroarenes Using MOF-Derived Co-Based Catalysts. *J. Mol. Catal. A Chem.* **2016**, *420*, 56–65. <https://doi.org/10.1016/j.molcata.2016.04.008>.
- (12) Noschese, A.; Buonerba, A.; Canton, P.; Milione, S.; Capacchione, C.; Grassi, A. Highly Efficient and Selective Reduction of Nitroarenes into Anilines Catalyzed by Gold Nanoparticles Incarcerated in a Nanoporous Polymer Matrix: Role of the Polymeric Support and Insight into the Reaction Mechanism. *J. Catal.* **2016**, *340*, 30–40. <https://doi.org/10.1016/j.jcat.2016.05.005>.
- (13) Nie, R.; Wang, J.; Wang, L.; Qin, Y.; Chen, P.; Hou, Z. Platinum Supported on Reduced Graphene Oxide as a Catalyst for Hydrogenation of Nitroarenes. *Carbon* **2012**, *50* (2), 586–596. <https://doi.org/10.1016/j.carbon.2011.09.017>.
- (14) Begum, R.; Rehan, R.; Farooqi, Z. H.; Butt, Z.; Ashraf, S. Physical Chemistry of Catalytic Reduction of Nitroarenes Using Various Nanocatalytic Systems: Past, Present, and Future. *J. Nanopart. Res.* **2016**, *18* (8), 1–24. <https://doi.org/10.1007/s11051-016-3536-5>.
- (15) Ibrahim, I.; Ali, I. O.; Salama, T. M.; Bahgat, A. A.; Mohamed, M. M. Synthesis of Magnetically Recyclable Spinel Ferrite (MFe₂O₄, M=Zn, Co, Mn) Nanocrystals Engineered by Sol Gel-Hydrothermal Technology: High Catalytic Performances for Nitroarenes Reduction. *Appl. Catal. B Environ.* **2016**, *181*, 389–402. <https://doi.org/10.1016/j.apcatb.2015.08.005>.
- (16) Junge, K.; Wendt, B.; Shaikh, N.; Beller, M. Iron-Catalyzed Selective Reduction of Nitroarenes to Anilines Using Organosilanes. *Chem. Commun.* **2010**, *46* (10), 1769–1771. <https://doi.org/10.1039/b924228g>.
- (17) Schneider, M.; Demoulin, P.; Sussmann, R.; Notholt, J. *Fourier Transform Infrared Spectrometry*; Wiley-Interscience, 2013; Vol. 10. https://doi.org/10.1007/978-1-4614-3909-7_6.
- (18) Rocha, M.; Costa, P.; Sousa, C. A. D.; Pereira, C.; Rodríguez-Borges, J. E.; Freire, C. L-Serine-Functionalized Montmorillonite Decorated with Au Nanoparticles: A New Highly Efficient Catalyst for the Reduction of 4-Nitrophenol. *J. Catal.* **2018**, *361*, 143–155. <https://doi.org/10.1016/j.jcat.2018.02.027>.
- (19) Sravanthi, K.; Ayodhya, D.; Swamy, P. Y. Green Synthesis, Characterization and

- Catalytic Activity of 4-Nitrophenol Reduction and Formation of Benzimidazoles Using Bentonite Supported Zero Valent Iron Nanoparticles. *Mater. Sci. Energy Technol.* **2019**, 2 (2), 298–307. <https://doi.org/10.1016/j.mset.2019.02.003>.
- (20) Lin, F. H.; Doong, R. A. Bifunctional Au-Fe₃O₄ Heterostructures for Magnetically Recyclable Catalysis of Nitrophenol Reduction. *J. Phys. Chem. C* **2011**, 115 (14), 6591–6598. <https://doi.org/10.1021/jp110956k>.
- (21) Ma, S.; Jia, W.; Tong, X.; Yang, Y.; Zhang, X.; Yao, T.; Wu, J. A Simple Approach for Synthesis of Hollow Mesoporous Nanotubes Loaded with Metallic and Magnetic Nanoparticles: Only One Step Is Required. *Appl. Organomet. Chem.* **2019**, 33 (5), e4849. <https://doi.org/10.1002/aoc.4849>.
- (22) Reizman, B. J.; Jensen, K. F. Simultaneous Solvent Screening and Reaction Optimization in Microliter Slugs. *Chem. Commun.* **2015**, 51 (68), 13290–13293. <https://doi.org/10.1039/c5cc03651h>.
- (23) Jarrais, B.; Guedes, A.; Freire, C. Heteroatom-Doped Carbon Nanomaterials as Metal-Free Catalysts for the Reduction of 4-Nitrophenol. *ChemistrySelect* **2018**, 3 (6), 1737–1748. <https://doi.org/10.1002/slct.201702706>.
- (24) Xia, Y.; Liu, Y.; Shi, N.; Zhang, X. Highly Efficient Reduction of 4-Nitrophenolate to 4-Aminophenolate by Au/γ-Fe₂O₃@HAP Magnetic Composites. *RSC Adv.* **2019**, 9 (18), 10272–10281. <https://doi.org/10.1039/c9ra00345b>.
- (25) Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **1996**, 17 (3), 37–53.
- (26) Azevedo, A.; Santos, M. F. KDD, Semma and CRISP-DM: A Parallel Overview. In *MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008*; 2008; pp 182–185.
- (27) Cano, J. R.; Herrera, F.; Lozano, M. Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study. *IEEE Trans. Evol. Comput.* **2003**, 7 (6), 561–575. <https://doi.org/10.1109/TEVC.2003.819265>.
- (28) Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques: Concepts and Techniques (3rd Edition)*; Elsevier, 2012.
- (29) Min, H. Developing the Profiles of Supermarket Customers through Data Mining. *Serv. Ind. J.* **2006**, 26 (7), 747–763. <https://doi.org/10.1080/02642060600898252>.

- (30) Chandrasekar, P.; Qian, K.; Shahriar, H.; Bhattacharya, P. Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing. *Proc. - Int. Comput. Softw. Appl. Conf.* **2017**, 2, 481–484. <https://doi.org/10.1109/COMPSAC.2017.146>.
- (31) Rousseeuw, P. J.; Van Driessen, K. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* **1999**, 41 (3), 212–223. <https://doi.org/10.1080/00401706.1999.10485670>.
- (32) Karczmarek, P.; Kiersztyn, A.; Pedrycz, W.; Al, E. K-Means-Based Isolation Forest. *Knowledge-Based Syst.* **2020**, 195, 15. <https://doi.org/10.1016/j.knosys.2020.105659>.
- (33) Carbonell, J. G.; Michalski, R. S.; Mitchell, T. M. AN OVERVIEW OF MACHINE LEARNING. In *Machine Learning*; Elsevier, 1983; pp 3–23. <https://doi.org/10.1016/b978-0-08-051054-5.50005-4>.
- (34) Poole, D. L.; Mackworth, A.; Goebel, R. G. Computational Intelligence and Knowledge. *Comput. Intell. A Log. Approach* **1998**, No. Ci, 1–22.
- (35) Jain, A. K.; Murty, M. N.; Flynn, P. J. *Data Clustering: A Review*; ACM, 2000.
- (36) Likas, A.; Vlassis, N.; J. Verbeek, J. The Global K-Means Clustering Algorithm. *Pattern Recognit.* **2003**, 36 (2), 451–461. [https://doi.org/10.1016/S0031-3203\(02\)00060-2](https://doi.org/10.1016/S0031-3203(02)00060-2).
- (37) Langley, P. The Changing Science of Machine Learning. *Mach. Learn.* **2011**, 82, 275–279. <https://doi.org/10.1007/s10994-011-5242-y>.
- (38) Bonaccorso, G. *Machine Learning Algorithms*, 2nd ed.; Dhandre, P., Dsouza, E., Eds.; Packt: Birmingham, 2018.
- (39) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, 374 (2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>.
- (40) Naik, G. R. *Advances in Principal Component Analysis: Research and Development*; Springer, Singapore, 2017. <https://doi.org/10.1007/978-981-10-6704-4>.
- (41) Hastie, T.; Tibshirani, R.; Friedman, J. Linear Methods for Regression. In *The Elements of Statistical Learning*; Springer, New York, NY, 2009; pp 43–99. https://doi.org/10.1007/978-0-387-84858-7_3.

- (42) Walker, S. H.; Duncan, D. B. Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika* **1967**, 54 (1/2), 167. <https://doi.org/10.2307/2333860>.
- (43) Hastie, T.; Tibshirani, R.; Friedman, J. Support Vector Machines and Flexible Discriminants. In *The Elements of Statistical Learning*; Springer, New York, NY, 2009; pp 417–458. https://doi.org/10.1007/978-0-387-84858-7_12.
- (44) Lorena, A. C.; De Carvalho, A. C. P. L. F. Uma Introdução Às Support Vector Machines. *Revista de Informática Teórica e Aplicada* **2007**, 14 (2), 43–67. <https://doi.org/10.22456/2175-2745.5690>
- (45) Kapil, S.; Chawla, M.; Ansari, M. D. On K-Means Data Clustering Algorithm with Genetic Algorithm. In *2016 4th International Conference on Parallel, Distributed and Grid Computing, PDGC 2016*; Institute of Electrical and Electronics Engineers Inc., 2016; pp 202–206. <https://doi.org/10.1109/PDGC.2016.7913145>.
- (46) Hamerly, G.; Elkan, C. Alternatives to the K-Means Algorithm That Find Better Clusterings; In *CIKM '02: Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 2002; pp 600–607. <https://doi.org/10.1145/584792.584890> 2002.
- (47) Bair, E. Semi-Supervised Clustering Methods. *Wiley Interdiscip. Rev. Comput. Stat.* **2013**, 5 (5), 349–361. <https://doi.org/10.1002/wics.1270>.
- (48) Gu, Y.; Li, K.; Guo, Z.; Wang, Y. Semi-Supervised K-Means DDoS Detection Method Using Hybrid Feature Selection Algorithm. *IEEE Access* **2019**, 7, 64351–64365. <https://doi.org/10.1109/ACCESS.2019.2917532>.
- (49) Kamiński, B.; Jakubczyk, M.; Szufel, P. A Framework for Sensitivity Analysis of Decision Trees. *Cent. Eur. J. Oper. Res.* **2018**, 26 (1), 135–159. <https://doi.org/10.1007/s10100-017-0479-6>.
- (50) Quinlan, J. R. Simplifying Decision Trees. *Int. J. Hum. Comput. Stud.* **1999**, 51 (2), 497–510. <https://doi.org/10.1006/ijhc.1987.0321>.
- (51) Rokach, L.; Maimon, O. *Data Mining with Decision Trees: Theory and Applications*, 2nd ed.; Yun, A., Ed.; World Scientific: Singapore, 2015; Vol. 81.
- (52) Loh, W. Y. Classification and Regression Trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2011**, 1 (1), 14–23. <https://doi.org/10.1002/widm.8>.
- (53) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature

- Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* **2017**, *121* (46), 8939–8954. <https://doi.org/10.1021/acs.jpca.7b08750>.
- (54) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2014**, *89* (9). <https://doi.org/10.1103/PhysRevB.89.094104>.
- (55) Schlexer Lamoureux, P.; Winther, K. T.; Garrido-Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine Learning for Computational Heterogeneous Catalysis. *ChemCatChem* **2019**, *11* (16), 3581–3601. <https://doi.org/10.1002/cctc.201900595>.
- (56) Odabaşı, Ç.; Günay, M. E.; Yildirim, R. Knowledge Extraction for Water Gas Shift Reaction over Noble Metal Catalysts from Publications in the Literature between 2002 and 2012. In *Int. J. Hydrog. Energy* **2014**, *39* (11), 5733–5746. <https://doi.org/10.1016/j.ijhydene.2014.01.160>.
- (57) Hastie, T.; Tibshirani, R.; Friedman, J. Ensemble Learning. In *The Elements of Statistical Learning*; Springer, New York, NY, 2009; pp 605–624. https://doi.org/10.1007/978-0-387-84858-7_16.
- (58) Han, J.; Kamber, M.; Pei, J. Classification. In *Data Mining*; Elsevier, 2012; pp 327–391. <https://doi.org/10.1016/B978-0-12-381479-1.00008-3>.
- (59) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. <https://doi.org/10.1023/A:1010933404324>.
- (60) Efron, B. Second Thoughts on the Bootstrap. *Statist. Sci.* **2003**, *18* (2), 135–140. <https://doi.org/10.1214/ss/1063994968>.
- (61) Ho, T. K. Random Decision Forests. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*; IEEE Computer Society, 1995; Vol. 1, pp 278–282. <https://doi.org/10.1109/ICDAR.1995.598994>.
- (62) Hastie, T.; Tibshirani, R.; Friedman, J. Random Forests. In *The Elements of Statistical Learning*; Springer, New York, NY, 2009; pp 587–604. https://doi.org/10.1007/978-0-387-84858-7_15.
- (63) Zhang, C.; Cao, L.; Romagnoli, A. On the Feature Engineering of Building Energy Data Mining. *Sustain. Cities Soc.* **2018**, *39*, 508–518.

<https://doi.org/10.1016/j.scs.2018.02.016>.

- (64) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *Resampling Methods*; Springer, New York, NY, 2013; pp 175–201. https://doi.org/10.1007/978-1-4614-7138-7_5.
- (65) Hastie, T.; Tibshirani, R.; Friedman, J. Model Inference and Averaging. In *The Elements of Statistical Learning*; Springer, New York, NY, 2009; pp 261–294. https://doi.org/10.1007/978-0-387-84858-7_8.
- (66) Athey, S.; Tibshirani, J.; Wager, S. Generalized Random Forests. *Ann. Stat.* **2019**, 47 (2), 1179–1203. <https://doi.org/10.1214/18-AOS1709>.
- (67) Ming, D.; Zhou, T.; Wang, M.; Tan, T. Land Cover Classification Using Random Forest with Genetic Algorithm-Based Parameter Optimization. *J. Appl. Remote Sens.* **2016**, 10 (3), 035021. <https://doi.org/10.1117/1.JRS.10.035021>.
- (68) Ibrahim, I. A.; Khatib, T. A Novel Hybrid Model for Hourly Global Solar Radiation Prediction Using Random Forests Technique and Firefly Algorithm. *Energy Convers. Manag.* **2017**, 138, 413–425. <https://doi.org/10.1016/j.enconman.2017.02.006>.
- (69) Bonnans, J. F.; Gilbert, J. C.; Lemaréchal, C.; Sagastizábal, C. A. Newtonian Methods. In *Numerical Optimization*; Springer, Berlin, Heidelberg, 2006; pp 51–66. https://doi.org/10.1007/978-3-540-35447-5_4.
- (70) Kroese, D. P.; Brereton, T.; Taimre, T.; Botev, Z. I. Why the Monte Carlo Method Is so Important Today. *Wiley Interdiscip. Rev. Comput. Stat.* **2014**, 6 (6), 386–392. <https://doi.org/10.1002/WICS.1314>.
- (71) Bonabeau, E.; Dorigo, M.; Theraulaz, G. *Swarm Intelligence: From Natural to Artificial Systems*; Oxford Scholarship Online, November 2020. <https://doi.org/10.1093/oso/9780195131581.001.0001>
- (72) Hu, J.; Turgut, A. E.; Lennox, B.; Arvin, F. Robust Formation Coordination of Robot Swarms with Nonlinear Dynamics and Unknown Disturbances: Design and Experiments. *IEEE Trans. Circuits Syst. II Express Briefs* **2021**. <https://doi.org/10.1109/TCSII.2021.3074705>.
- (73) Wu, H. S.; Zhang, F. M. Wolf Pack Algorithm for Unconstrained Global Optimization. *Math. Probl. Eng.* **2014**, 2014. <https://doi.org/10.1155/2014/465082>.
- (74) GitHub - SISDevelop/SwarmPackagePy: Library of swarm optimization algorithms. <https://github.com/SISDevelop/SwarmPackagePy#readme> (accessed

Aug 17, 2021).

- (75) Martín-Jimeno, F. J.; Suárez-García, F.; Paredes, J. I.; Martínez-Alonso, A.; Tascón, J. M. D. Nickel Nanoparticle/Carbon Catalysts Derived from a Novel Aqueous-Synthesized Metal-Organic Framework for Nitroarene Reduction. *J. Alloys Compd.* **2021**, 853, 157348. <https://doi.org/10.1016/j.jallcom.2020.157348>.

Anexos

A. Caraterização de dados e modelo de *Random Forest*

A 1 - Loadings da análise de componentes principais às características de catalisador, com XRD.

Variável	PC1	PC2
D_{part}	-0.01707	0.014342
D_{pore}	-0.03822	-0.06091
S_{BET}	0.004439	-0.01546
%Ag	0.036757	-0.03564
%C	0.002883	-0.01135
%Cu	0.039318	0.015168
%Fe	0.132187	-0.06319
%N	0.191595	0.028177
%O	0.140144	0.252175
%P	-0.02048	-0.04002
%Pd	0.14206	0.25465
%Ru	0.035725	-0.00455
%S	0.180077	-0.03756
XRD_000	-0.03056	-0.03021
XRD_001	-0.03255	-0.0322
XRD_002	-0.03657	-0.03625
XRD_003	-0.03941	-0.03916
XRD_004	-0.03486	-0.03475
XRD_005	-0.02993	-0.02989
XRD_006	0.263454	-0.13153
XRD_007	0.263477	-0.1315
XRD_008	0.263477	-0.1315
XRD_009	0.263479	-0.1315
XRD_010	0.023002	0.027696
XRD_011	0.045623	0.02172
XRD_012	0.074001	-0.015
XRD_013	0.037435	-0.01769
XRD_014	0.071991	-0.01678
XRD_015	0.071899	-0.01674
XRD_016	0.073274	-0.0174
XRD_017	0.170382	-0.06853
XRD_018	0.280209	0.015917
XRD_019	0.0963	-0.03259
XRD_020	0.19693	-0.04768
XRD_021	0.100578	-0.05182
XRD_022	0.036619	-0.01982
XRD_023	0.046683	-0.02308
XRD_024	0.055849	0.003804
XRD_025	0.243283	-0.12773

XRD_026	0.105109	-0.08933
XRD_027	0.10271	-0.0581
XRD_028	-0.05107	0.111212
XRD_029	0.097421	0.161899
XRD_030	0.005241	-0.01689
XRD_031	0.03692	0.007226
XRD_032	-0.00229	0.143182
XRD_033	-0.0523	0.106867
XRD_034	0.1907	0.035941
XRD_035	0.084652	0.062359
XRD_036	-0.0157	0.057218
XRD_037	0.023073	-0.0357
XRD_038	0.13137	-0.05281
XRD_039	-0.0069	-0.01786
XRD_040	-0.03002	-0.04765
XRD_041	-0.04012	0.018295
XRD_042	0.098387	0.163583
XRD_043	-0.02923	-0.04571
XRD_044	-0.04273	-0.01929
XRD_045	0.237759	-0.12843
XRD_046	-0.01566	-0.01831
XRD_047	-0.0759	0.075098
XRD_048	-0.00333	0.134557
XRD_049	-0.02777	0.058934
XRD_050	-0.04192	0.002571
XRD_051	-0.03611	-0.02087
XRD_052	0.144938	0.244039
XRD_053	0.002536	-0.02427
XRD_054	0.058604	-0.03022
XRD_055	-0.01718	0.158647
XRD_056	0.048401	0.165299
XRD_057	-0.00331	0.119616
XRD_058	-0.0664	0.14359
XRD_059	-0.07365	0.105174
XRD_060	0.013301	0.161768
XRD_061	0.070197	0.205298
XRD_062	-0.02585	-0.00533
XRD_063	0.050669	-0.02504
XRD_064	0.00271	0.017073
XRD_065	-0.00959	0.134575
XRD_066	-0.00271	0.060276
XRD_067	-0.02298	0.012054
XRD_068	-0.06908	0.106692
XRD_069	0.130198	0.273184
XRD_070	0.141831	0.259405
XRD_071	0.017837	0.109569
XRD_072	0.144384	0.257405
XRD_073	-0.0266	0.045199
XRD_074	-0.00966	0.142212

XRD_075	-0.09533	0.059346
XRD_076	-0.01436	-0.03119
XRD_077	-0.01161	-0.02978
XRD_078	-0.07035	0.093914
XRD_079	-0.01263	0.048793
XRD_080	-0.01301	-0.01056
XRD_081	-0.03457	-0.03543
XRD_082	-0.02528	-0.03185
XRD_083	-0.02487	-0.03144
XRD_084	-0.02596	-0.02044
XRD_085	-0.02596	-0.02044
XRD_086	-0.02417	-0.01723
XRD_087	0.123515	0.224842
XRD_088	-0.02731	0.043988
XRD_089	-0.00678	0.038344

A 2 - Loadings da análise de componentes principais às características de catalisador, sem XRD.

Variável	PC1	PC2
D_{part}	-0.06172	0.13808
D_{pore}	-0.02036	0.01126
S_{BET}	-0.04439	-0.16193
%Ag	0.18859	-0.15275
%C	0.10056	-0.13147
%Cu	-0.10140	0.06298
%Fe	0.27505	0.05776
%N	0.19065	-0.07041
%O	0.54554	-0.22451
%P	0.38584	0.53366
%Pd	0.05604	0.04595
%Ru	0.38757	0.53603
%S	0.22430	-0.32703
%Si	0.41916	-0.41247

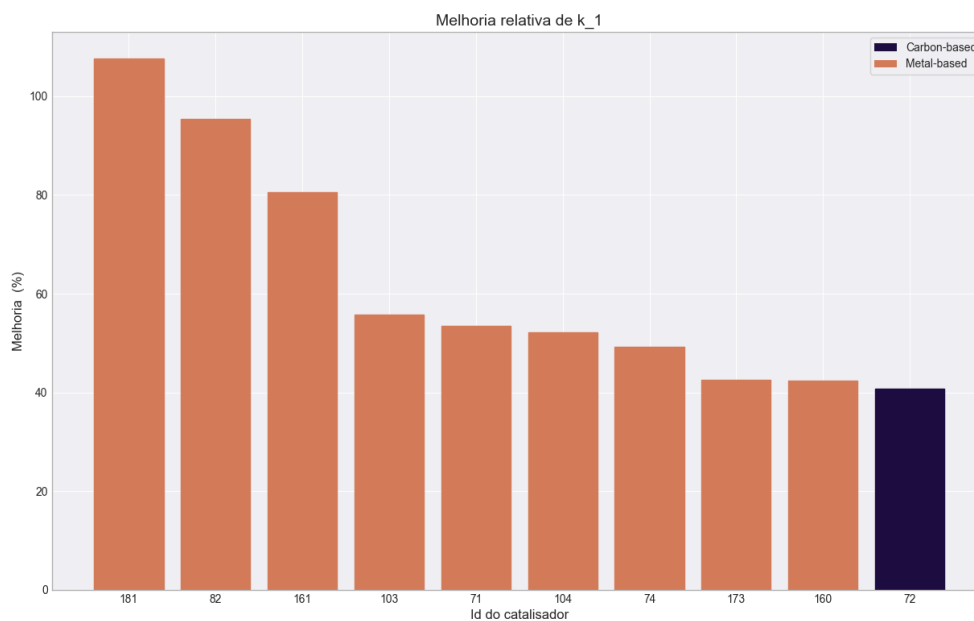
A 3 - Loadings da análise de componentes principais às características de catalisador, sem XRD após exclusão de outliers.

Variável	PC1	PC2
D_{part}	-0.05076	-0.19150
D_{pore}	-0.02301	-0.11636
S_{BET}	-0.28490	0.42781
%Ag	-0.04038	-0.13674
%C	-0.15671	0.60329
%Cu	-0.05910	-0.17201
%Fe	0.41636	0.01109
%N	0.35481	0.22663
%O	0.51073	0.23174
%P	-0.11946	0.26535
%Pd	0.30004	-0.03776
%Ru	0.00000	0.00000
%S	-0.17324	0.36597
%Si	0.43777	0.20875

B. Otimização de Monte Carlo

A 4 - Loadings da análise de componentes principais às características de catalisador e condições otimizadas, sem XRD após exclusão de outliers.

Variável	PC1	PC2	PC3	PC4
m_{cat}	-0.21775	0.11479	0.15803	-0.10293
m_{sub}	-0.21317	-0.08527	-0.01748	0.04073
m_{red}	-0.22844	0.10301	0.14367	-0.44561
m_{adt}	-0.18434	0.11551	-0.18432	-0.09626
T	-0.37273	-0.12451	0.03292	-0.16365
t	0.04755	-0.14990	0.50994	-0.16634
T_{ind}	-0.27003	-0.21124	-0.25694	0.13655
cdr	-0.12263	-0.11852	-0.20707	-0.35929
V	-0.03566	0.26708	-0.33966	0.10619
D_{part}	-0.09771	-0.09521	-0.18925	0.43528
D_{pore}	-0.06667	-0.05662	0.11197	0.05293
S_{BET}	0.02043	0.47881	0.01496	-0.17023
%Ag	-0.03024	-0.08069	0.13752	0.17659
%C	0.18855	0.51041	0.00542	0.02759
%Cu	-0.11441	-0.10330	0.23158	-0.12328
%Fe	0.23832	-0.19984	-0.40876	-0.26314
%N	0.31499	0.02256	-0.18506	-0.19847
%O	0.44694	-0.15928	0.09453	-0.03115
%P	-0.04123	0.23858	0.00456	-0.24025
%Pd	0.13510	-0.18535	-0.25330	-0.32735
%Ru	0.00000	0.00000	0.00000	0.00000
%S	0.05626	0.30317	0.03106	0.15094
%Si	0.38433	-0.13925	0.19536	0.02807

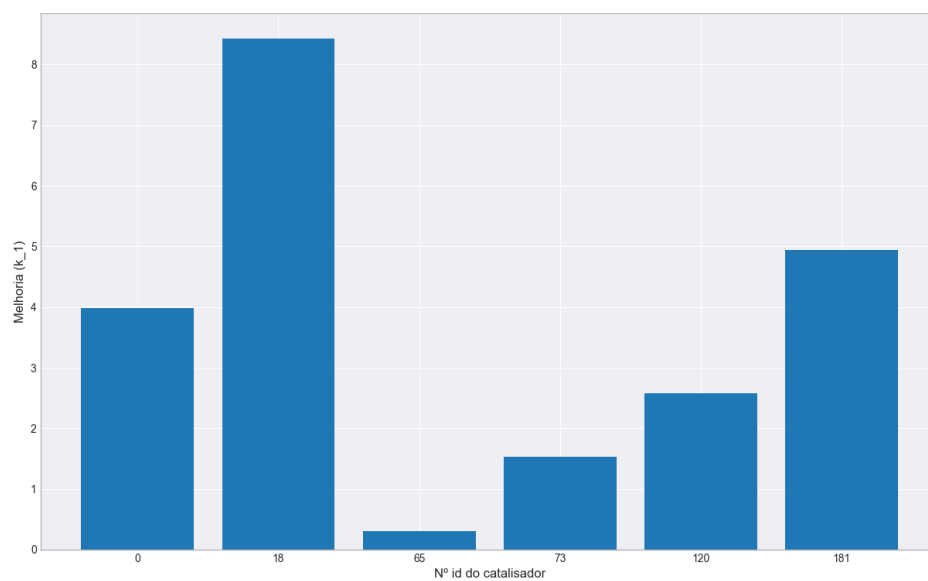


A 5 - Melhorias relativas do otimizador de Monte Carlo.

C. Otimização de enxame

A 6 - Melhores resultados obtidos a partir da otimização das condições usando os ambos os algoritmos de enxame.

Nº de ID do catalisador	k_1 estimado das condições otimizadas
0	13.965
18	12.808
65	0.50127
73	3.1192
120	14.647
181	16.297



A 7 - Representação gráfica das melhorias mais elevadas obtidas pelos dois algoritmos de enxame.