UNIVERSIDADE FEDERAL DE SANTA CATARINA

CAMPUS TRINDADE

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E ENGENHARIA DE MATERIAIS

Henrique de Medeiros Back

Aprendizado de Máquina Aplicado à Identificação de Microplásticos no Oceano Utilizando Espectroscopia de Infravermelho - FTIR

Florianópolis

2020

Henrique de Medeiros Back

Aprendizado de Máquina Aplicado à Identificação de Microplásticos no Oceano Utilizando Espectroscopia de Infravermelho - FTIR

Dissertação de Mestrado submetida ao Programa de Pós-graduação em Ciência e Engenharia de Materiais da Universidade Federal de Santa Catarina para obtenção do título de mestre em Engenharia de Materiais
Orientador: Prof. Orestes Estevam Alarcon, Dr.

Florianópolis
2020

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Back, Henrique de Medeiros
Aprendizado de Máquina Aplicado à Identificação de Microplásticos no Oceano Utilizando Espectroscopia de Infravermelho – FTIR / Henrique de Medeiros Back ; orientador, Orestes Estevam Alarcon, 2020.
68 p.

Dissertação (mestrado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência e Engenharia de Materiais, Florianópolis, 2020.

Inclui referências.

1. Ciência e Engenharia de Materiais. 2. Microplásticos. 3. Poluição Marinha. 4. Aprendizado de máquina. 5. Espectroscopia. I. Alarcon, Orestes Estevam. II. Universidade Federal de Santa Catarina. Programa de Pós Graduação em Ciência e Engenharia de Materiais. III. Título.

Henrique de Medeiros Back

Aprendizado de Máquina Aplicado à Identificação de Microplásticos no Oceano Utilizando Espectroscopia de Infravermelho - FTIR

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Orestes Estevam Alarcon, Dr.
Universidade Federal de Santa Catarina

Prof. Edson Cilos Vargas Júnior, Dr
Universidade Federal de Santa Catarina

Prof. Guilherme Mariz de Oliveira Barra, Dr.
Universidade Federal de Santa Catarina

Prof.(a) Marilena Valadares Folgueiras, Dr.(a)
Universidade do Estado de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Engenharia de Materiais

Documento assinado digitalmente
Joao Batista Rodrigues Neto
Data: 25/02/2022 14:01:08-0300
CPF: 645.810.989-04
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Orestes Estevan Alarcon
Data: 25/02/2022 09:04:39-0300
CPF: 295.644.417-49
Verifique as assinaturas em https://v.ufsc.br

Prof. Orestes Estevam Alarcon, Dr.
Orientador

Florianópolis, 2020.

Agradecimentos

Agradeço, em primeiro lugar, à minha família, em especial minha companheira, pelo apoio e incentivo, sem os quais a conclusão desta tarefa seria muito mais difícil.

Agradeço ao Professor Orestes Estevam Alarcon pela orientação e pela confiança. Sem isso, aquilo que era potencial não se tornaria concreto neste trabalho.

Agradeço aos colegas e amigos na UFSC, com os quais compartilhei meus dias de trabalho e cuja companhia tornou mais leve a tarefa posta à frente.

Agradeço e reconheço o privilégio de continuar meus estudos na UFSC e de poder buscar o título de mestre nesta grande instituição.

Agradeço, de forma geral, a todos aqueles que dedicam e dedicaram suas vidas a busca de conhecimento e que reconhecem seu papel fundamental na construção de uma sociedade mais consciente.

Agradeço à vida.

A grande Via não tem porta. Milhares de estradas desembocam nela.

(Provérbio Zen)

Resumo

Polímeros são materiais cada vez mais amplamente utilizados em uma enorme diversidade de aplicações, desde componentes industriais a objetos de consumo diário, especialmente embalagens. Sua produção, proveniente majoritariamente de fontes fósseis atingiu 350 milhões de toneladas em 2017, representando um crescimento de 40% em relação à 2009. Além de ser um produto gerador de gases de efeito estufa, contribuindo para o aquecimento global do planeta, seu descarte tem produzido grandes quantidades de resíduos, os quais não são adequadamente geridos: acumulam na natureza, degradam-se em microplásticos e afetam negativamente os ecossistemas. O estudo e caracterização de microplásticos têm importância fundamental para melhor compreender a procedência, o destino e os impactos desse poluente na natureza. Entretanto, o processo de análise convencional por busca por similaridade em bibliotecas espectrais e análise visual de um *expert* é laborioso e pouco preciso, reduzindo a quantidade e qualidade de informações disponíveis. Este trabalho visa aplicar ferramentas de inteligência artificial, mais notadamente de aprendizado de máquina, para análise dos dados de Espectroscopia no Infravermelhos com Transformada de Fourier (FTIR), com o intuito de classificar polímeros de maneira mais rápida e com maior confiabilidade. Os dados utilizados para desenvolvimento dos modelos são resultados de coletas e análises de FTIR feitas pelo veleiro de pesquisa Francês Tara em diversas posições geográficas no Mar Mediterrâneo, durante os meses de maio e novembro de 2014. A partir desses dados, disponibilizados na nuvem de computadores, foram gerados 24 modelos de classificação e os mesmos foram comparados por uma medida de acurácia da sua capacidade em prever 13 classes de polímeros. O classificador Random Forest aplicado aos dados pré-processados por t-SNE foi o que resultou numa melhor capacidade preditiva. Entretanto, a vantagem foi pouco significativa em relação a outras metodologias avaliadas.

Palavras-chave: Microplásticos. Poluição marinha. Aprendizado de máquina. Espectroscopia

Abstract

Polymers are materials increasingly used in a wide variety of applications, from industrial components to objects of daily consumption, especially packaging. Its production, mostly from fossil sources, reached 350 million tons in 2017, representing a 40% growth compared to 2009. In addition to being a product that generates greenhouse gases, contributing to the global warming of the planet, its disposal has produced large quantities of waste, which are not properly managed: they accumulate in nature, degrade in microplastics and negatively affect ecosystems. The study and characterization of microplastics is of fundamental importance to better understand the source, destination and impacts of this pollutant on nature. However, the conventional analysis process by looking for similarity in spectral libraries and visual analysis by an expert is laborious and imprecise, reducing the quantity and quality of information available. This work aims to apply artificial intelligence tools, most notably machine learning, in the analysis of infrared spectroscopy (FTIR) data from marine microplastic samples in order to classify polymers more quickly and with greater reliability. The data used to develop the models are the result of FTIR collections and analysis carried out by the French research sailboat Tara in different geographical positions in the Mediterranean Sea during the months of May and November 2014. 24 classification models were generated and their ability to predict 13 classes of polymers were assessed and compared using a measure of accuracy. The Random Forest classifier applied to the data pre-processed by t-SNE was the one that resulted in a better predictive capacity. However, the advantage was not significant in relation to other evaluated methodologies.

Keywords: Microplastics. Marine pollution. Machine Learning. Spectroscopy.

Lista de figuras

Lista de tabelas

Lista de abreviaturas

ABRELPE – Associação Brasileira de Empresas de Limpeza Pública e Resíduos Especiais

AM – Aprendizado de Máquina

ATR – *Attenuated Total Reflectance* (Refletância Total Atenuada)

DNA – Á**cido Desoxirribonucleico**

EPS – Poli(estireno) Expandido

FTIR – Espectroscopia no Infravermelho por Transformada de Fourier

IA – Inteligência Artificial

KNN – *K-nearest Neighbours*

LLDPE – Poli(etileno) linear de baixa densidade

PC – Componente Principal

PCA – Análise de Componentes Principais

PE - Poli(etileno)

PEAD – Poli(etileno) de Alta Densidade

PEBD – Poli(etileno) de Baixa Densidade

PET – Poli(tereftalato de etileno)

PEVA – Poli(acetato-vinilo de etileno)

POP – Poluente Orgânico Persistente

PP – Poli(propileno)

PS – Poli(estireno)

PUR – Poli(uretano)

PVC – Poli(cloreto de vinila)

RD – Redução de Dimensionalidade

SVM – *Support Vector Machine*

t-SNE – *t-distributed Stochastic Neighbor Embedding*

Sumário

1. INTRODUÇÃO

Os polímeros compõe uma classe de materiais com enorme diversidade de composições e propriedades correspondentes, sendo utilizados amplamente em inúmeros setores e aplicações devido à sua versatilidade e relativa facilidade de produção. Não há dúvida que seu desenvolvimento trouxe uma série de benefícios à sociedade (ANDRADY; NEAL, 2009).

Desde meados do século XX vêm sendo produzidos em larga escala e os volumes de produção anuais só crescem, chegando próximo aos 350 milhões de toneladas em 2017 (PLASTICS EUROPE, 2018). Entretanto, cerca de 40% da produção é destinada ao uso em embalagens, produtos de vida média muito curta e que contribuem enormemente para a geração de resíduos (GEYER; JAMBECK; LAW, 2017). Devido às baixas taxas de reciclagem, à má gestão de resíduos e à durabilidade desses materiais, grandes quantidades acabam por acumular na natureza, mais notadamente nos oceanos (THOMPSON et al., 2004b).

Produtos poliméricos constituem a maior parte do macrolixo marinho (ERIKSEN et al., 2014). O contato destes com a vida selvagem gera uma série de danos à saúde dos organismos vivos. Já foram relatados emaranhamentos, sufocamentos, interrupção do trato digestivo, etc (COLE et al., 2011a). Porém, essa talvez não constitua a maior ameaça. O macroplástico degrada quando exposto às intempéries e se fragmenta em microplásticos que causam danos químicos além de danos físicos (BRANDON; GOLDSTEIN; OHMAN, 2016). O seu potencial em causar impactos negativos nos rios e oceano, associado a incrível ubiquidade desse poluente, o fez ser considerado um dos mais importantes poluentes ambientais.

Muito esforço vem sendo empregado pela comunidade científica a fim de entender os impactos no meio ambiente que os materiais plásticos já devem estar causando, mas também para melhor compreender como eles chegam ao oceano, qual a sua distribuição, quanto tempo permanecem lá e como poder-se-ia evitar que ele escapasse ou até mesmo em como retirá-los da natureza (SHARMA; CHATTERJEE, 2017). Manter saudáveis os oceanos é um compromisso firmado entre 193 países como parte da Agenda 2030 e um dos 17 Objetivos do Desenvolvimento Sustentável.

Muitos desafios se apresentam ao lidar com a questão do lixo marinho, em especial devido à inerente interdisciplinaridade do tema, sendo a colaboração entre as diversas áreas de extrema importância. Dentre as competências necessárias ao bom entendimento dos riscos

que emergem da poluição por plásticos e da busca por potenciais soluções, está a caracterização dos polímeros encontrados na natureza. A partir dela, espera-se ser possível traçar um perfil da origem do lixo marinho, o seu destino uma vez que escapa para o meio ambiente e quais os impactos decorrentes da exposição às substâncias associadas a ele (HIDALGO-RUZ et al., 2012b).

As técnicas de caracterização por espectroscopia (FTIR e Raman) têm sido utilizadas para este fim como complemento à análise visual, pois permitem identificar a composição química e confirmar a origem sintética de microplásticos (que podem ser confundidos com partículas naturais), classificar amostras por tipo de polímero e ainda indicar algum grau de degradação (GESAMP, 2019). Entretanto, a atual capacidade em analisar os dados adquiridos é limitada. A identificação dos espectros é parcialmente automatizada por um processo de buscas por similaridade em bases de dados que são computacionalmente intensivas, produzem alto índice de erros e requerem a confirmação visual de um *expert*, resultando em demora e baixa confiabilidade (HUFNAGL et al., 2019; KEDZIERSKI et al., 2019).

A proposta deste trabalho é de incorporar ferramentas de aprendizado de máquina na análise de dados espectroscópicos com o objetivo de automatizar a caracterização dos espectros coletados de amostras de microplástico marinho, melhorar a qualidade dos resultados das análises e acelerar o processo. O aprendizado de máquina é a principal área a fazer avançar o desenvolvimento da inteligência artificial e é caracterizada pelo aprendizado de algoritmos baseados em dados (experiência prévia) e sem a utilização de instruções explícitas (como é o caso da programação convencional), dependendo apenas da identificação de padrões na estrutura dos dados e fazendo inferências a partir deles (BEN-DAVID, 2014). O crescente desenvolvimento e incorporação dessas novas tecnologias nos mais diversos campos de aplicação traz luz ao que pode ser alcançado também no estudo de microplásticos ambientais (CHARRINGTON, 2019; BERLINSKI, 2001; CLARK, 2015).

O aumento da quantidade e disponibilidade de dados de poluição por polímeros e a necessidade de melhor compreender o problema, fez surgir o interesse da incorporação de ferramentas de aprendizado de máquina (AM) na análise de dados espectrais de microplásticos. Melhorar a capacidade de extrair informações dos dados disponíveis, em última análise, trará benefícios à compreensão do amplo problema do lixo marinho. Com o uso de métodos analíticos têm-se a vantagem de não haver a necessidade de grandes investimentos para aquisição de equipamentos, sendo possível fazer mais com a infraestrutura já disponível, o que é particularmente interessante no contexto da ciência brasileira.

Este trabalho buscou comparar diferentes ferramentas de redução de dimensionalidade, que reduzem o número de variáveis necessárias para descrever uma amostra, com pouca ou nenhuma perda de informação, e classificadores (modelos estatísticos de previsão para classificação) a fim de identificar qual metodologia resultaria em melhor capacidade preditiva, observando duas métricas: acurácia e velocidade de implementação dos modelos. Esperava-se que a aplicação das técnicas de redução pudesse influenciar positivamente ambas as métricas. Para tal, foram utilizados dados de Espectroscopia no Infravermelho por Transformada de Fourier (FTIR) de um conjunto de cerca de 1000 amostras de microplástico previamente identificadas. Todos os dados foram coletados em expedições do veleiro TARA no Mar Mediterrâneo e disponibilizados online. Além disso, a aplicação das funções estatísticas foi baseada em bibliotecas de código aberto como *scikit-learn* e *pandas*.

2. OBJETIVOS

Objetivo Geral:

- Identificar metodologia com melhor performance na classificação dos dados espectroscópicos de amostras de microplásticos encontrados no oceano, comparando 24 combinações de técnicas de redução de dimensionalidade e modelos de classificação.

Objetivos específicos:

- Encontrar antecedentes na bibliografia que embasem a seleção e aplicação de metodologias de aprendizado de máquina ao contexto da caracterização de microplásticos.
- Obter dados espectroscópicos de amostras de microplástico marinho suficientes e coerentes para a validação estatística dos modelos.
- Avaliar a razão de redução no número de variáveis descritivas obtida por técnica de RD.
- Gerar mapeamentos utilizando técnicas de RD, a fim de buscar padrões nos dados das amostras, com foco em identificar semelhanças e diferenças entre amostras e classes de polímeros para melhor compreender as limitações da metodologia proposta, bem como das análises convencionais.
- Avaliar a melhora da eficiência dos classificadores após a implementação de diferentes técnicas de redução de dimensionalidade (RD), comparando a velocidade de implementação e a acurácia atingida por metodologia.
- Comparar resultados encontrados com a metodologia de análise convencional

3. ESTADO DA ARTE

Foi feita de forma exploratória para a revisão da maioria dos temas relacionados ao tema central, que é Aprendizado *de máquina para caracterização de microplásticos* (descrito na seção 43), enquanto neste caso, foi feita uma revisão bibliográfica sistemática (KITCHENHAM; CHARTERS, 2007), descrita com detalhes no item 47.

3.1. MATERIAIS POLIMÉRICOS

Polímeros naturais vêm sendo usados há milênios por seres humanos para diversos fins. Couro, fibras vegetais, e borracha natural, são alguns exemplos. Entretanto, muitos desses materiais têm limitações quanto às suas propriedades, durabilidade, disponibilidade e processos de fabricação. Foi buscando alternativas que em meados do século XIX os primeiros materiais poliméricos sintéticos foram criados. Em 1839 Goodyear inventou a borracha vulcanizada e o poliestireno (PS) foi descoberto. Durante o mesmo século alguns outros polímeros foram sintetizados, como o celuloide, o policloreto de vinila (PVC) e a viscose. Entretanto, seu uso ainda era limitado e havia pouca produção comercial (ANDRADY; NEAL, 2009).

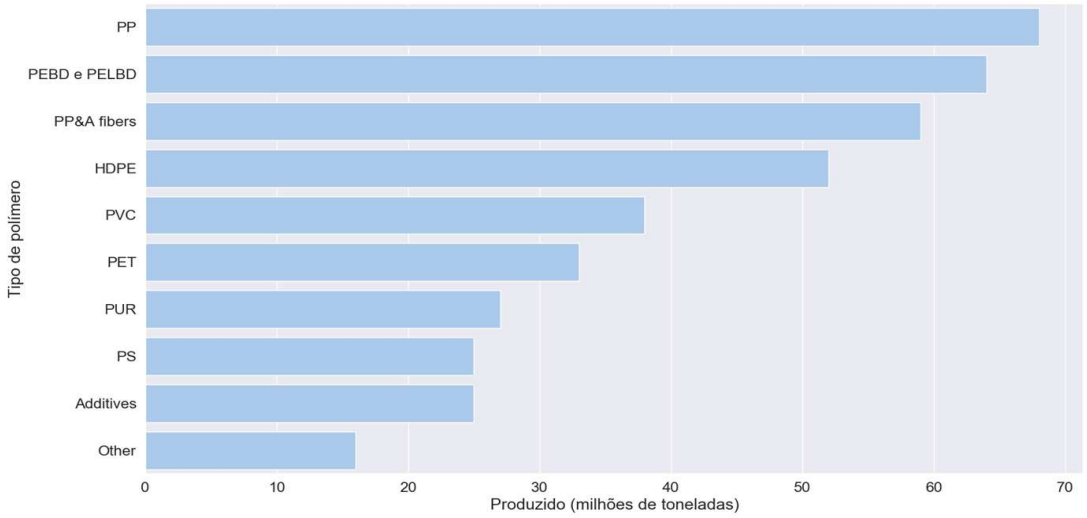
Durante as primeiras décadas do século XX muitos polímeros novos surgiram, a exemplo do polietileno (PE) (1933), do politereftalato de etileno (PET) (1941) e do polipropileno (PP) (1954) (ANDRADY; NEAL, 2009). Isso se deu com o grande desenvolvimento da indústria química, principalmente em função das duas grandes guerras. Porém, foi só por volta do fim da Segunda Guerra Mundial (1945) que a produção em massa desses novos materiais começou (COLE et al., 2011b).

Hoje, existem centenas de polímeros disponíveis no mercado, especializados para diversas utilizações. A variedade de composições e a facilidade com que podem ser alteradas suas propriedades faz com que os polímeros sintéticos sejam muito interessantes para o uso em inumeráveis produtos. De fato, os polímeros são muito versáteis. São leves, resistentes mecanicamente, duráveis e resistentes à corrosão. Ainda, o baixo custo desses materiais tornou acessíveis muitos bens de consumo que antes eram restritos apenas a um público mais rico e assim contribuiu significativamente para o crescimento do seu uso e com isso, também, o surgimento de um grande mercado consumidor.

Em 2017 a produção mundial de polímeros chegou próximo de 350 milhões de toneladas, um aumento de 40% em relação à 2009 (PLASTICS EUROPE, 2018). Com a

presente taxa de crescimento, a produção pode duplicar nos próximos 20 anos. Mas apesar da grande diversidade de resinas poliméricas existentes, alguns poucos polímeros têm um uso amplo em diferentes setores e aplicações e, portanto, têm um volume de produção muito mais elevado. São os polímeros *commodities*. Entre eles estão: PEBD (polietileno de baixa densidade), LLDPE (polietileno linear de baixa densidade), PEAD (polietileno de alta densidade), PP, PVC, PS, EPS (poliestireno expandido), PET, PUR (poliuretano) (COLE et al., 2011b).

Figura 1 – Volume de produção de polímeros commodities em 2015



Fonte: Adaptado de GEYER; JAMBECK, 2017

As poliolefinas, nomeadamente o polietileno e o polipropileno, dominam o cenário da produção. Podem ser produzidas a partir da fração líquida do gás natural ou da fração gasosa de baixo valor do refino do petróleo, ambas matérias-primas muito baratas. São muito leves, quimicamente resistentes, hidrofóbicos, bioinertes e são facilmente moldados em produtos.

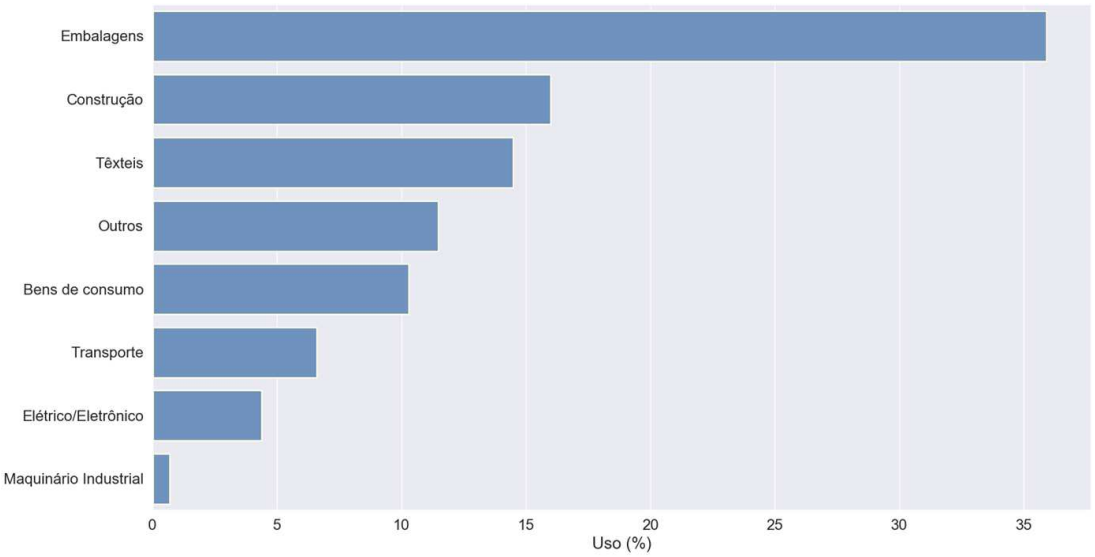
3.2. POLÍMEROS COMO POLUENTES AMBIENTAIS

3.2.1. Fontes

Em certa medida, as características das poliolefinas são compartilhadas por outros polímeros *commodities* e permitem a produção de bons produtos com pouco valor comercial e que muitas vezes são usados apenas uma vez. Isso pode ser observado na Figura 2, que

evidencia o fato de que mais de um terço da produção mundial de polímeros é utilizada pelo setor de embalagens.

Figura 2 – Percentual de uso de polímeros por setor de aplicação



Fonte: Adaptado de GEYER; JAMBECK 2017

Isso levou a uma imensa transformação em direção ao uso de produtos descartáveis, inclusive, mas não somente no setor de embalagens. Cada vez mais, produtos de consumo como roupas e eletrônicos têm vida útil mais curta. Entretanto, ela não foi acompanhada de melhores práticas e alternativas ao descarte. De fato, até recentemente o destino no fim de vida desses produtos não causava grande preocupação e não era um problema para designers de produtos ou consumidores. Isso tornou difícil o reaproveitamento de materiais descartados que junto a outros empecilhos de cunho tecnológico e econômico, impediram o crescimento da reciclagem como alternativa ao descarte de materiais.

No Brasil, bem como em grande parte dos países não desenvolvidos ou em desenvolvimento, o acesso da população a serviços de saneamento básico ainda é precário. Nesses casos, grandes parcelas das populações vivem em áreas sem coleta de lixo e mesmo quando há coleta, o lixo não têm o destino adequado. No Brasil, por exemplo, mesmo com a proibição dos chamados *lixões*, em vigor desde 2014, em 2018 havia quase 3 mil ainda funcionando e o percentual do resíduo que têm esse destino é crescente, segundo relatório da Associação Brasileira de Empresas de Limpeza Pública e Resíduos Especiais (ABRELPE) (ABRELPE, 2017). O mesmo relatório, apresentou dados de reciclagem no país, mas se limitou ao percentual reciclado de resíduos recicláveis de papel/papelão, alumínio e plástico sendo respectivamente 52,3%, 87,2%, e 8,2%.

Devido em grande parte ao mal gerenciamento de resíduos sólidos, mas também a outras formas de perda para o ambiente, como o desgaste de materiais em uso e emissões diretas, por exemplo, por produtos de higiene e beleza, grandes quantidades de polímeros escapam para o meio ambiente todos os anos. Estimativas apontam para *inputs* globais de cerca de 100 milhões de toneladas anuais até 2020 (JAMBECK et al., 2015).

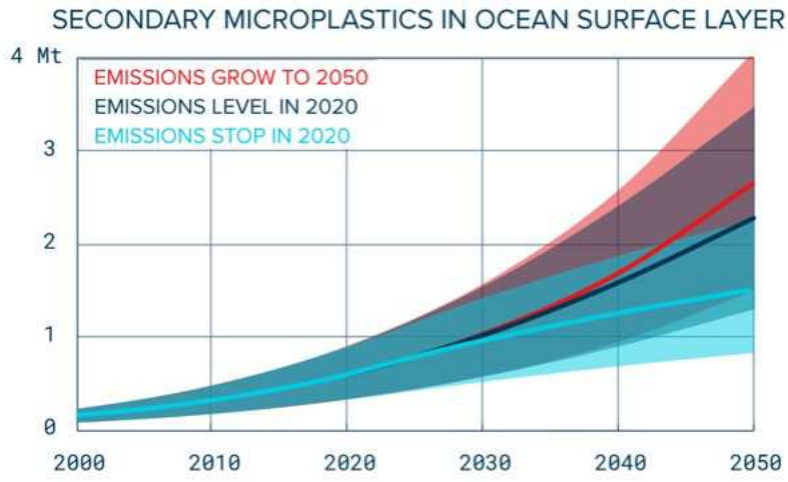
Os polímeros são materiais leves e facilmente transportados por água e vento. Assim, bacias de drenagem próximas a grandes centros urbanos, em especial em países não desenvolvidos, captam grandes quantidades de lixo provenientes do uso e mal gerenciamento de produtos descartados. Já foi mostrado que os rios são os maiores responsáveis por transportar resíduos plásticos para os mares. Com isso, estimou-se que 20 rios, localizados principalmente na Ásia, são responsáveis por 67% dos plásticos despejados no mar via fluvial (LEBRETON et al., 2017). Tanto o tamanho da bacia de captação, como a densidade populacional e os percentuais de resíduos mal geridos são de extrema relevância para tais resultados.

3.2.2. Destinos

Duas das propriedades mais importantes dos materiais poliméricos durante seu uso, leveza e durabilidade, são também propriedades muito problemáticas quando resíduos desses materiais são mal geridos (RYAN et al., 2012). A leveza tida nos polímeros como uma vantagem sobre outras classes de materiais, faz com que o plástico mal gerido possa ser transportado longas distâncias pelo vento e pela água. Já a durabilidade faz com que esses poluentes sejam reconhecidos, hoje, por serem um dos mais persistentes e abundantes detritos marinhos (BRANDON; GOLDSTEIN; OHMAN, 2016b)

Em 2004, um artigo publicado na revista *Science* com o título *Lost at sea: Where is all the plastic?* (“Perdido no mar: Onde está todo o plástico?”, em tradução livre), sugeriu que o lixo plástico parecia não acumular no oceano apenas por não estarmos detectando-o. De fato, nesse artigo, foram reportadas coletas em sedimentos e águas superficiais do oceano que continham micropartículas de material plástico e que tais partículas pareciam ter-se originado da fragmentação de resíduos plásticos maiores. Apesar de já haverem sido relatados pequenos resíduos poliméricos em habitats naturais anteriormente, o artigo mencionado é um marco na pesquisa desses poluentes (THOMPSON et al., 2004a).

Figura 3 - Projeção do acúmulo de microplásticos oriundos da fragmentação de macroplásticos nos oceanos até 2050 em três casos: crescimento das emissões, emissões continuam ao nível das atuais e emissões cessam em 2020.



Fonte: (LEBRETON; EGGER; SLAT, 2019).

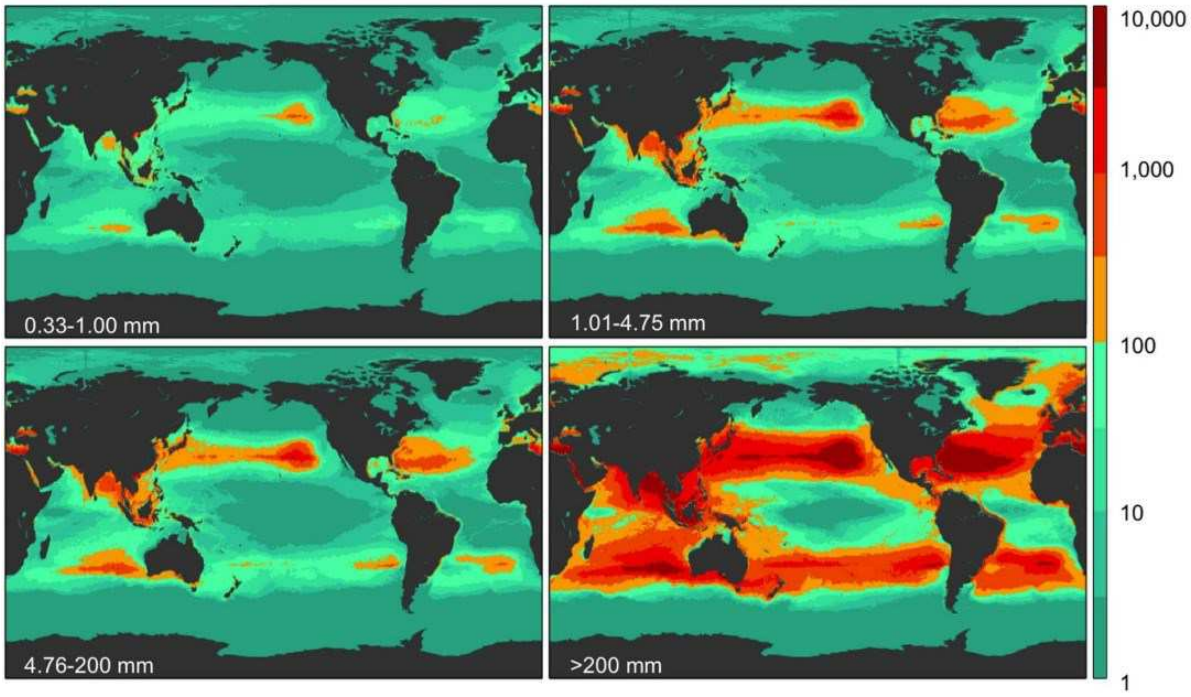
Não se sabe ao certo qual o tempo de degradação completa dos polímeros na natureza, mas as taxas de degradação parcial por processos físicos, principalmente por ação de raios UV, é rápida o suficiente para permitir a perda de propriedades mecânicas. Assim, por ação mecânica dos ventos, das ondas ou mesmo da vida animal, o plástico se fragmenta em partículas cada vez menores. Não havendo um mecanismo de biodegradação completa, que resulte apenas em moléculas de substâncias simples, como água e dióxido de carbono, tais micropartículas não se degradam por completo e acabam por acumular nos ecossistemas.

O fato é que os microplásticos já são encontrados em grande quantidade na natureza e considerados hoje como um importante poluente ambiental. Sua presença já foi confirmada em lagos remotos de altitude (FREE et al., 2014), nas mais profundas fossas oceânicas (SAITO et al., 2018), nos solos (WANG et al., 2019), na água que bebemos e no ar que respiramos (BARROWS; CATHEY; PETERSEN, 2018). Onde quer que se busque por plástico, pode-se encontrá-lo.

Entretanto sua distribuição não é aleatória ou homogênea. O lixo plástico tende a ser carregado por ação dos ventos, chuvas e correntes para locais de menor energia potencial gravitacional. Em última instância, isso significa que acumulam principalmente nos oceanos. De fato, o mar é o local mais estudado para presença de lixo plástico, principalmente sua superfície. Já foi visto que ele tende a acumular nos giros oceânicos, zonas de convergência de correntes. Na Figura 4, pode ser observada a distribuição desses resíduos nos oceanos, com as zonas de maior concentração coincidindo com os giros (ERIKSEN et al., 2014;

SHERMAN et al., 2015). Corpos de água fechados, como lagos, também tendem a concentrar esses poluentes, visto que os *inputs* são maiores que os *outputs*.

Figura 4: Densidade de acumulação por peso estimada de plástico nos oceanos do mundo, divididas por tamanho em 4 classes. Foram utilizados dados coletados in situ e extrapolações utilizando um modelo de circulação de correntes.



Fonte: (ERIKSEN et al., 2014)

3.2.3. Impactos

A ubiquidade dos plásticos no meio ambiente os torna acessíveis a seres vivos de todos os níveis tróficos, desde produtores primários até predadores de topo de cadeia, como é o caso dos humanos. Já foram reportados mais de 44.000 casos de contato entre polímeros e organismos vivos de mais de 1.400 espécies (AGAMUTHU et al., 2019). Surge então, a necessidade de compreender quais os efeitos causados pela exposição a estes poluentes.

Os impactos podem ser diversos. Danos físicos podem ocorrer por emaranhamento de animais em macroplásticos, como redes de pesca, canudos ou garrafas. Além disso, é comum a ingestão de micro ou macroplásticos, por animais que os confundem com alimento, o que pode causar feridas internas ou mesmo bloquear o trato digestivo e levar à inanição (GREGORY, 2009).

Ademais, efeitos fisiológicos da exposição ao plástico podem ser importantes. Apesar de os polímeros em geral serem atóxicos, eles podem conter substâncias

potencialmente nocivas. Isso porque podem haver em sua composição, monômeros residuais (resultado da polimerização incompleta), aditivos (misturados à matriz a fim de variar suas propriedades), ou ainda, outros poluentes orgânicos persistentes (POPs) presentes em ambientes aquáticos, sejam eles fármacos, agrotóxicos ou resíduos industriais, que são absorvidos e adsorvidos por esses materiais, devido sua característica altamente hidrofóbica e especial na sua forma micro ou nano (GOUIN et al., 2011). Muitas dessas substâncias são conhecidas por causar disrupção do sistema endócrino e/ou câncer. Assim, os organismos marinhos têm uma tendência maior a morrerem após ingerirem plásticos (AGAMUTHU et al., 2019).

Ao nível ecossistêmico, os plásticos podem servir como vetores para a introdução de espécies invasoras (GREGORY, 2009). Microorganismos crescendo na *plastisfera* podem ser transmissores de doenças ou diretamente prejudiciais a outros organismos. Apesar do desenvolvimento de microorganismos na superfície dos plásticos poder facilitar a degradação desses detritos (SHAH et al., 2008), isso pode significar que as substâncias tóxicas ali presentes estarão mais acessíveis aos outros organismos que futuramente entrem em contato direto com eles. Ademais, sua presença os torna mais semelhantes a alimentos naturais e pode favorecer a ingestão (MOORE, 2008).

A exposição humana ao microplástico pode se dar via aérea, principalmente microfibras de tecidos sintéticos em ambientes fechados, e por ingestão. Microplásticos já foram reportados em fezes humanas, e já foram encontrados em amostras de água engarrafada, de sal e em diversos peixes, crustáceos e moluscos de consumo humano, sendo que por esses últimos a ingestão é maior, pois são organismos filtradores e são consumidos inteiros, sem remoção do trato digestivo (AGAMUTHU et al., 2019; PEIXOTO et al., 2019).

Porém, apesar das evidências quanto à exposição e à toxicidade dos microplásticos, poucos estudos diretos dos efeitos destes na saúde humana já foram realizados. Portanto, a corrente compreensão científica do problema ainda é insuficiente (AGAMUTHU et al., 2019).

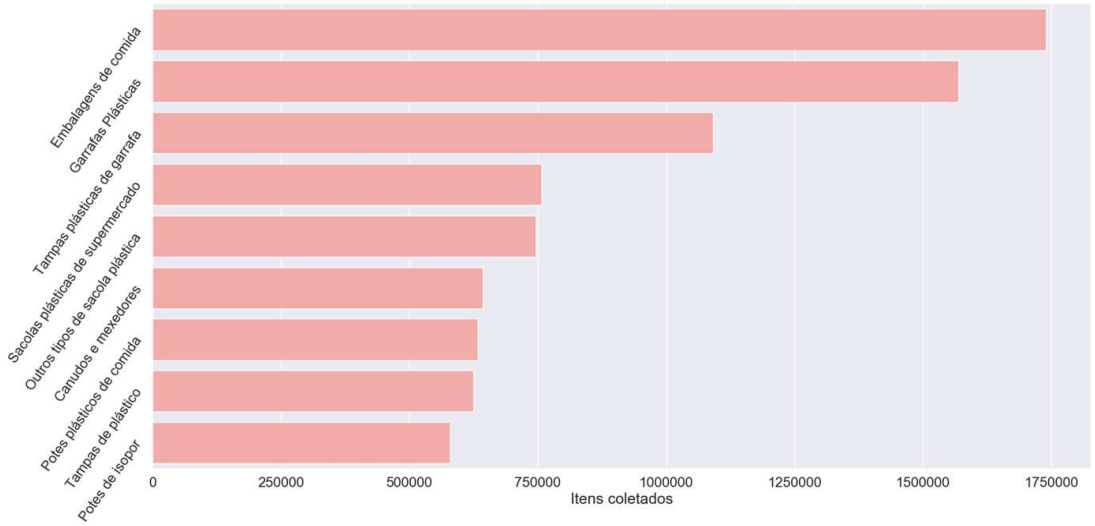
3.2.4. Caracterização

Como foi descrito anteriormente, os polímeros formam uma classe de materiais bastante diferentes quimicamente. Assim, cada um deve apresentar comportamentos ambientais e interações fisiológicas diversas. Portanto, uma etapa fundamental na compreensão das fontes, do destino e dos impactos dos polímeros no meio ambiente é a caracterização do resíduo.

Muitas formas de caracterização foram descritas. Macroplásticos encontrados em limpezas de praia, por exemplo, costumam ser classificados por tipo de produto. A partir daí pode-se inferir as condições que levaram tal resíduo a acumular no ambiente, os hábitos de consumo associados a ele, e talvez até o tipo de material, o tempo de degradação e os impactos ecológicos relacionados (OCEAN CONSERVANCY; INTERNATIONAL COASTAL CLEANUP, 2018).

No relatório anual publicado pela ONG *Ocean Conservancy* sobre as limpezas de praia realizadas globalmente, é possível identificar os 10 itens mais encontrados (ver Figura 5)(OCEAN CONSERVANCY; INTERNATIONAL COASTAL CLEANUP, 2018). Embalagens de comidas, embalagens e sacolas plásticas estão entre os mais coletados.

Figura 5 - Top 10 itens mais coletados em limpezas de praia pelo mundo.

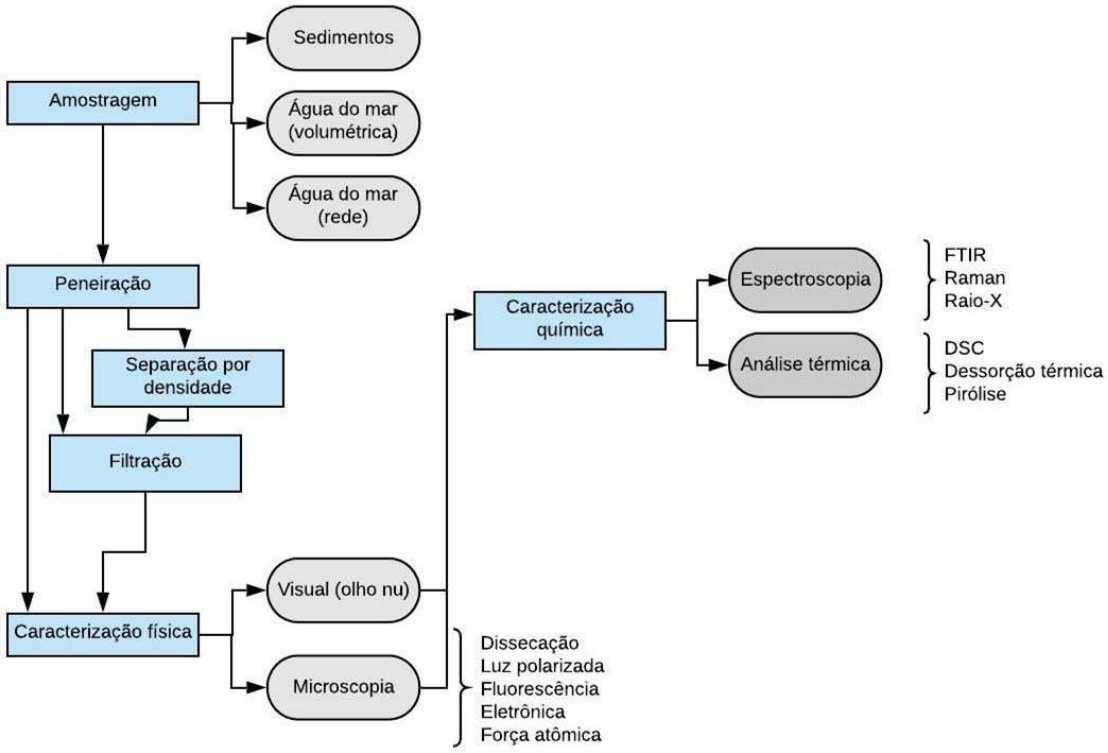


Fonte: Adaptado de OCEAN CONSERVANCY; INTERNATIONAL COASTAL CLEANUP, 2018)

A caracterização do macrolixo é comumente feita também incluindo outros aspectos do resíduo encontrado, como: forma, tamanho, cor, material e grau de degradação (GESAMP, 2019). Entretanto, essa análise é limitada no que concerne ao ambiente de coleta, ao tamanho dos resíduos que é capaz de identificar e coletar e também à própria caracterização da composição química do material encontrado, tendo em vista que um mesmo produto pode ser fabricado com diversos materiais, inclusive distintos polímeros. Assim, para uma compreensão mais aprofundada do problema do lixo plástico marinho, é necessário preencher tais lacunas. Um foco especial vem sendo dado aos microplásticos por apresentarem grande abundância e distribuição, um potencial de risco maior e razoável possibilidade de detecção (AUTA; EMENIKE; FAUZIAH, 2017).

Dada a diversidade de ambientes de coleta, de materiais encontrados e de rigor na identificação de microplásticos, muitas metodologias já foram utilizadas para amostragem e caracterização. Entretanto, procedimentos padronizados são importantes para comparação de resultados e identificação de padrões de distribuição geográficos e temporais. Na Figura 6 vê-se um fluxograma geral das principais etapas de caracterização de microplásticos (adaptado de (HIDALGO-RUZ et al., 2012a) e (GESAMP, 2019)).

Figura 6: Fluxograma dos procedimentos de amostragem, preparação e identificação de microplásticos marinhos.



Fonte: Elaborado pelo autor

A etapa de caracterização física é um importante passo para a separação de outros tipos de detritos, inclusive os de origem natural. Mas tendo em vista que no caso dos microplásticos a degradação tende a ser predominante, informações anteriores sobre o produto e material de origem podem ser inacessíveis somente por uma análise visual. Assim, há uma tendência em superestimar ou subestimar a quantidade de microplásticos (KYOUNG et al., 2015; LENZ et al., 2015). Dificuldade que, em certa medida, pode ser contornada pela caracterização química diminuindo a frequência de erros (BERGMANN; GUTOW; KLAGES, 2015).

A amostragem de microplásticos descrita na metodologia deste trabalho apresenta uma configuração que vai ao encontro das melhores práticas e está descrita com detalhes na seção 47.

É interessante também ressaltar a aplicação de técnicas de caracterização convencionais em engenharia, tanto em pesquisa e desenvolvimento de materiais como no controle de qualidade de produção. Muito do conhecimento necessário ao estudo de polímeros como poluentes vêm da base de conhecimento em ciência e engenharia de materiais.

3.2.4.1. *Espectroscopia*

Nesta seção serão discutidas algumas técnicas de análise espectroscópica utilizadas na caracterização de microplásticos. Será dada ênfase na técnica de Espectroscopia no Infravermelho por Transformada de Fourier com Refletância Total Atenuada (FTIR-ATR), utilizada na identificação das amostras deste trabalho.

Como já foi visto, a análise química tem um papel importante na caracterização dos microplásticos. Primeiro, permite distinguir polímeros sintéticos entre os vários tipos de detritos, inclusive biológicos. Segundo, permite identificar o tipo de polímero. Entretanto, tal análise apresenta desafios, uma vez que há uma gama bastante variada de polímeros com estruturas químicas, composições e estágios de degradação diferentes (KYOUNG et al., 2015).

As técnicas mais comuns para identificação de microplástico são FTIR e a espectroscopia Raman (KÄPPLER et al., 2016). Ambas são técnicas de espectroscopia vibracional, ou seja, excita-se as ligações moleculares de uma amostra para que possam ser detectadas.

Em FTIR, a amostra é irradiada com um feixe de luz na faixa do infravermelho. Parte do feixe é absorvido, e a parte que não é pode ser detectada por transmissão ou reflexão. O modo ATR (*Attenuated Total Reflectance*), utilizado neste trabalho, opera por reflexão e permite análise de amostras com morfologia mais variada.

A espectroscopia Raman, por sua vez, é um método de espalhamento onde um feixe de laser monocromático incide na amostra e interage com as ligações moleculares. Isso faz uma parte dos fótons variarem sua energia e fornece informações sobre os tipos de ligação molecular ali presentes.

Há diferenças fundamentais no mecanismo de cada uma das técnicas, porém ambas são úteis para o fim de caracterizar microplásticos, apenas podendo apresentar resultados um pouco diferentes. Dessa maneira, recomenda-se sua utilização complementar (KÄPPLER et

al., 2016). Entretanto, o processo de aquisição e análise dos espectros é laborioso e os resultados obtidos por apenas uma delas costumam ser satisfatórios (BRANDON; GOLDSTEIN; OHMAN, 2016; HALSTEAD et al., 2018; JUNG et al., 2018; KEDZIERSKI et al., 2019; VEERASINGAM et al., 2016).

Após a aquisição dos dados no equipamento, eles devem ser tratados para gerar um gráfico, ou espectro, mais limpo e com picos mais bem definidos. Assim, o especialista poderá mais facilmente classificar as amostras por tipo de polímero. Comumente, utiliza-se algum *software* para identificação mais rápida. Com eles, pode-se gerar os gráficos e é possível realizar uma busca em uma biblioteca espectral (previamente organizada e disponível ao programa) para identificar à quais polímeros um espectro determinado é mais similar. As bibliotecas costumam conter uma gama de espectros de substâncias químicas sintéticas ou não, inclusive dezenas, ou até centenas de polímeros (PRIMPKE et al., 2018). Entretanto, bibliotecas tão completas costumam custar caro para serem montadas e o acesso a elas acaba por ser restrito.

Outra deficiência deste método de identificação espectral reside no fato de os percentuais de similaridade espectral serem geralmente baixos para o caso dos microplásticos ambientais, além de necessitar alta capacidade computacional para serem produzidos (HUFNAGL et al., 2019). O estágio de degradação dos polímeros, bem como contaminações e dificuldades na limpeza e manipulação das amostras, alteram os espectros de modo que sua caracterização fica complicada. Em geral, os resultados fornecidos pelo *software* precisam ser confirmados pela análise visual do especialista (JUNG et al., 2018). Essa etapa de revisão demanda tempo de um especialista bem treinado e retarda o processo de identificação. Quando a quantidade de amostras é grande, tais dificuldades se tornam grandes empecilhos à aquisição de dados da poluição marinha, o que mais globalmente afeta nossa capacidade de compreensão do problema.

3.3. INTELIGÊNCIA ARTIFICIAL

Inteligência Artificial é a área de estudo e desenvolvimento de *agentes inteligentes*. Em seu sentido mais literal, um *agente* é algo que *age* em um ambiente, seja ele uma bactéria, um ser humano, um Estado, um avião ou mesmo o vento. Já um agente inteligente age de maneira inteligente. Isso quer dizer que age de acordo com as suas circunstâncias e seus objetivos, sendo seu comportamento mutável com a variação do seu ambiente e com a aprendizado sobre ações passadas. Ainda, tal agente toma decisões apropriadas dada sua

percepção e computação limitados (POOLE; MACKWORTH; GOEBEL, 1998). O objetivo primeiro desse campo de estudos é compreender quais as condições necessárias para o surgimento de comportamento inteligente, seja em sistemas naturais ou artificiais.

Tendo como central a hipótese de que raciocínio é computação, pretende-se desenvolver métodos de replicar o comportamento inteligente em artefatos artificiais, ou sintéticos (POOLE; MACKWORTH; GOEBEL, 1998). O termo Inteligência Computacional é preferido por certos autores visto que a palavra *artificial* pode causar confusão e tende a limitar ao estudo de sistemas não naturais.

O conceito de agentes (e não seres vivos) inteligentes é muito antigo e pode ser encontrado em registros egípcios, gregos e chineses muitos anos antes de Cristo. Entretanto, o evento mais recente que levou, inclusive ao desenvolvimento do computador como *máquina inteligente*, é a tese de Church-Turing (1938), que enuncia a possibilidade de qualquer raciocínio formal poder ser reproduzido por uma máquina computacional.

Já o campo de estudos moderno em inteligência artificial foi fundado em 1956 durante um *workshop* em Darthmouth College (Estados Unidos da América), por professores do *Massachussets Institute of Technology (John McCarthy e Marvin Minsky)*, da *Carnegie Mellon University* (Allen Newell e Herbert Simon) e um pesquisador da IBM (Arthur Samuel) (CREVIER, 1993).

Durante os anos que se seguiram, muitos resultados foram obtidos e aplicações surgiram. Tanto o público como os pesquisadores estavam animados com incríveis avanços que a nova área logo traria. Entretanto, no fim da década de 60, o progresso começava a demorar e os financiamentos começaram a ser cortados.

Foi só no fim dos anos 90, quando a capacidade computacional das máquinas já estava mais bem desenvolvida e o famoso robô *DeepBlue* venceu Garry Kasparov, o então campeão mundial de xadrez em uma partida, que a inteligência artificial voltou a ganhar notoriedade (BERLINSKI, 2001). Desde então, o interesse pela área vem crescendo, especialmente por sua aplicabilidade aos mais diferentes nichos.

Segundo um artigo publicado na revista Bloomberg em 2016, o ano de 2015 foi um marco no desenvolvimento da inteligência artificial. O autor Jack Clark atribui isso ao surgimento de redes neurais, da computação na nuvem e do crescimento de ferramentas de pesquisa e disponibilidade de dados (CLARK, 2015)

Algoritmos de inteligência artificial têm sido usados para detecção prévia de falhas em sistemas mecânicos, na detecção de câncer em análises clínicas, em algoritmos de recomendação de conteúdo virtual e no desenvolvimento de tradutores simultâneos

automáticos e carros autônomos, apenas para citar alguns exemplos. Essas “inteligências” já realizam tarefas com mais precisão do que humanos e até tarefas nas quais humanos seriam incapazes de realizar.

3.3.1. Aprendizado de Máquina

Em um mundo onde os dispositivos eletrônicos têm se tornado cada vez menores e mais presentes no cotidiano das pessoas, sensores de todo o tipo vêm coletando enormes quantidades de dados sobre diversos parâmetros físicos e virtuais. A enorme disponibilidade de dados, fez crescer o interesse por ferramentas que possam ajudar a gerir e extrair informações relevantes de maneira eficiente.

O aprendizado de máquina é a principal área a fazer avançar o desenvolvimento da inteligência artificial, mais recentemente impulsionada pela aprendizagem profunda. É uma área dentro de IA (ver Figura 7) caracterizada pelo aprendizado de algoritmos baseados em dados (experiência prévia) e sem a utilização de instruções explícitas (como é o caso da programação convencional), dependendo apenas da identificação de padrões e fazendo inferências a partir deles. Um modelo matemático baseado em conceitos e operações estatísticas é construído baseado em dados de treino e é utilizado para fazer predições ou informar decisões (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Figura 7: Relação entre aprendizado de máquina e inteligência artificial

Fonte: Elaborado pelo autor

Por exemplo, o programa de tradução do Google, o Google Tradutor, continha cerca de 500 mil linhas de código. Hoje, ele depende de apenas 500 linhas e com grande melhora em seu desempenho (CHARRINGTON, 2019). Isso tudo devido à grande disponibilidade de dados e ao avanço nas ferramentas de aprendizado de máquina.

Os modelos de AM podem ser classificados em três grandes classes:

- Aprendizado supervisionado: quando os dados de treino contém *inputs* e *outputs*, que podem ser considerados como as variáveis a partir das quais o modelo é criado e o resultado esperado para cada amostra de treino, respectivamente. Por exemplo: deve-se classificar amostras de DNA (ácido desoxirribonucleico) de espécimens pertencentes a espécies conhecidas, baseado no DNA de outros espécimens conhecidos dessas espécies
- Aprendizado não-supervisionado: quando há apenas *inputs* e o algoritmo deve tentar encontrar alguma estrutura intrínseca aos dados, como aglomerados de amostras. Por exemplo: deve-se identificar espécies e não se sabe quais e/ou quantas espécies existem num conjunto de dados de DNA.
- Aprendizado por reforço: quando um agente é modelado de tal maneira a maximizar uma certa noção de "recompensa" ao agir em um ambiente. Por exemplo: um agente em um jogo de computador aprende a vencer uma partida baseado em seu histórico buscando fazer mais pontos que seu oponente.

Um conceito muito recorrente e que deve ser esclarecido é o de Aprendizado Profundo (do inglês, *deep learning*), que é uma área dentro de AM que utiliza redes neurais, baseadas num entendimento simplificado do funcionamento do cérebro, para gerar aprendizado de maneira mais eficiente em modelos de AM. Entretanto, como não serão usadas técnicas desse tipo neste trabalho, o conceito não será discutido com maior profundidade.

3.3.2. Linguagens de Programação

Linguagens de programação são linguagens formais usada para escrever programas que por sua vez são especificações de computações e algoritmos. Existem inúmeras linguagens, algumas das mais utilizadas são Java, C, Python e R. Cada uma tem suas próprias particularidades, sendo melhores ou piores para certas tarefas. Neste trabalho, foi usada a linguagem Python, descrita a seguir com mais detalhes.

Python foi criada em 1991 por Guido van Rossum que buscou desenvolver uma linguagem simples de fácil leitura de código e com poucas funções embutidas em seu núcleo. Ao contrário, optou por permitir a extensão de suas bibliotecas para fácil adaptação às emergentes necessidades de seus usuários.

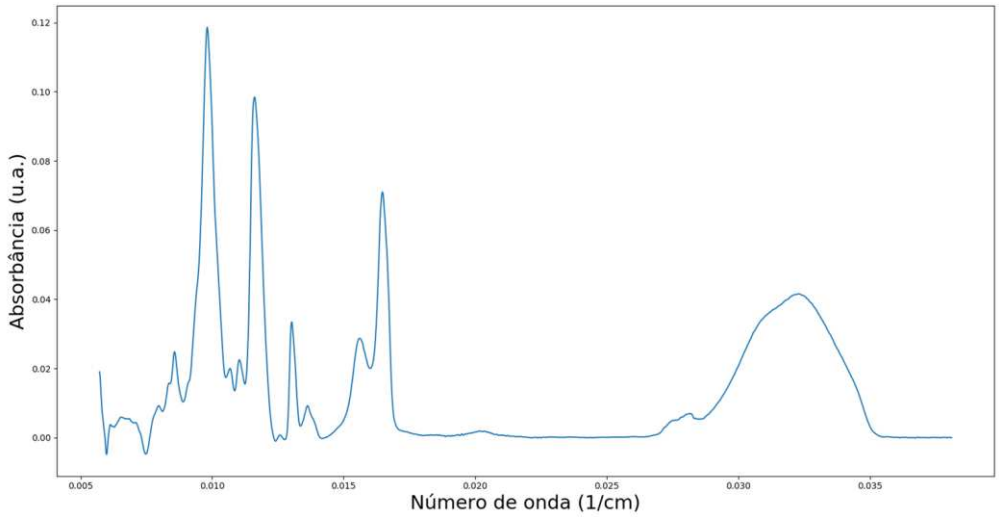
Justamente por isso foi adotada para a realização deste trabalho. Bibliotecas muito completas de funções dedicadas à exploração de dados e aprendizado de máquina foram

desenvolvidas e já embutidas na linguagem. Dessa forma, sempre que necessário operar uma função a uma matriz de dados, é possível aplicá-la em poucas linhas de código, ao invés das muitas linhas necessárias para descrever a função em si. Alguns exemplos das bibliotecas utilizadas são: *pandas*, *scikit-learn* e *matplotlib*.

3.3.3. Redução de Dimensionalidade

Em análise de dados, é interessante possuir mais dados quanto possível sobre uma amostra ou classe, o que naturalmente leva a melhor aprendizado, e portanto, melhor aplicação dos modelos. Entretanto, já foi observado na literatura que, após um certo número de variáveis adicionais, as generalizações oferecidas pelos algoritmos tendem a ser menos significativas (TRUNK, 1979). Esse fenômeno é muitas vezes referido como *curse of dimensionality* ("maldição da dimensionalidade", em tradução livre). Isso acontece, por exemplo, em dados espectroscópicos, onde há longas faixas do espectro que contém apenas ruído, ou seja, dados que não agregam informação (ver Figura 8). Manter frequências desse tipo pode fazer classes de polímeros diferentes terem espectros muito parecidos, visto que para a maioria das variáveis (ou frequências) o sinal é o mesmo. Além disso, num espaço alto-dimensional as amostras ficam muito esparsas e é mais difícil estabelecer limites de classificação (BEN-DAVID, 2014).

Figura 8: Espectro de microplástico identificado como acetato de celulose



Fonte: Elaborado pelo autor

Dessa forma, os objetivos da redução de dimensionalidade são: melhorar a capacidade preditiva dos modelos de aprendizado de máquina, gerar modelos mais rapidamente e com melhor custo-benefício e, ainda, permitir um melhor entendimento dos dados utilizados. Portanto, a RD foi incorporada no método de classificação de microplásticos desenvolvido neste trabalho.

3.3.3.1. Seleção de atributos

Uma forma de realizar a redução da dimensionalidade é por seleção de atributos*, (*do inglês *feature selection)*. Essa técnica consiste em identificar as variáveis importantes para identificação das amostras, nesse caso, as frequências características dos polímeros, e manter apenas essas, descartando os dados das frequências que contém apenas ruído.

Há diversas formas de realizar esse tipo de redução, como *backward elimination, forward selection e* filtros de alta correlação. Entretanto, a técnica adotada nesse trabalho foi o filtro de baixa variância. Este filtro consiste em identificar variáveis com pouca variância e descartá-las, mantendo apenas as que variam significativamente (contém picos), e que portanto devem carregar informação.

3.3.3.2. Extração de atributos

Ao contrário da seleção de atributos, a extração de atributos consiste em apreender a informação contida nos dados das variáveis originais e armazená-las em novas variáveis. Essas técnicas são particularmente interessantes, pois permitem a visualização do conjunto das amostras em gráficos bi ou tridimensionais, com cada uma das dimensões contendo um percentual significativo da informação.

Novamente, há uma série de técnicas dedicas à extração de atributos de um conjunto de dados. Entretanto, este trabalho limitou-se a avaliar apenas dois: Análise de Componentes Principais (PCA) (mais utilizada) e *t-stochastic neighbour embedding* (t-SNE) (mais moderna).

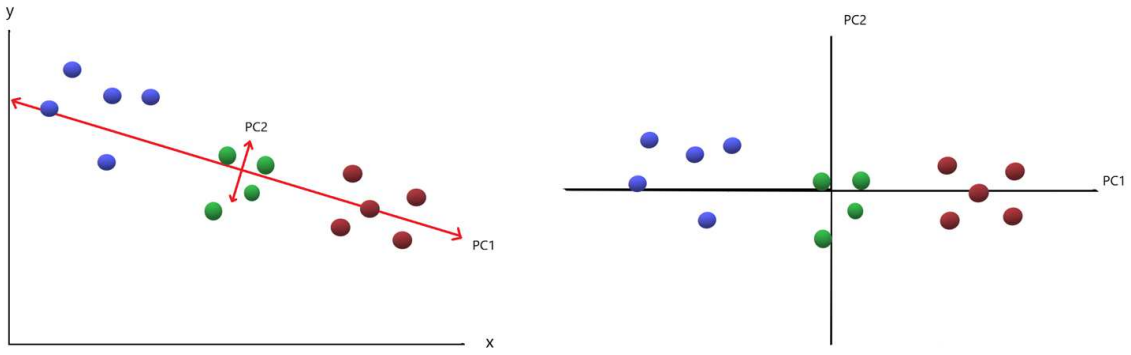
3.3.3.2.1. Análise de Componentes Principais

A análise de componentes principais é uma das técnicas mais bem-sucedidas e provavelmente a mais comum para redução de dimensionalidade. Desenvolvida por Harold Hotteling em 1933 (VAN DER MAATEN; HINTON, 2008), ganhou evidência com o

desenvolvimento e popularização dos computadores. Atualmente é utilizada nos mais diversos campos de aplicação, dentre eles: ciências sociais, finanças, *marketing,* ciências da terra, astronomia, biologia, bioquímica e medicina. Inclusive no estudo de microplásticos marinhos, como forma de análise exploratória de dados (JUNG et al., 2018)

O PCA consiste em uma transformação ortogonal da matriz de dados original. É utilizado para decompor uma matriz de dados de múltiplas variáveis em sucessivas componentes ortogonais, chamadas de componentes principais (PC). Tal decomposição é feita de forma a ordenar as componentes por variância explicada, ou seja, a primeira componente principal é a que contém a maior parte da informação, em seguida a segunda e assim por diante. Geometricamente, tem-se que a soma dos quadrados das distâncias entre os pontos e uma linha é mínima quando a linha é igual à PC1. Igualmente, pode-se dizer que a primeira componente principal determina a direção de maior variância dos dados.

Figura 9: PCA - Transformação ortogonal de variáveis.



Fonte: Elaborado pelo autor

É verdade que uma PC pode ser computada para cada dimensão original do conjunto de dados, entretanto, como as primeiras componentes principais contém a maior parte da informação, é conveniente manter somente algumas delas. Assim, o PCA é considerado uma ferramenta de redução de dimensionalidade, pois permite reduzir a quantidade de dimensões com pouca perda de informação.

3.3.3.2.2. *t- distributed Stochastic Neighbor Embedding* (*t-SNE*)

Assim como PCA, *t-SNE* é usado para representar um conjunto de dados multidimensional em 2 ou 3 dimensões, portanto pode ser usado como uma ferramenta de redução de dimensionalidade. Entretanto, ela é tida como uma técnica mais avançada, visto

que foi desenvolvida mais recentemente, apenas em 2008, pouco mais de 70 anos depois do PCA, por Laurens van der Maaten e Geoffrey Hinton (VAN DER MAATEN; HINTON, 2008).

A mais evidente diferença entre PCA e t-SNE se encontra no fato de que a primeira é uma combinação linear de variáveis, já a segunda é uma combinação não linear. De certa forma, o PCA tende a manter pontos dissimilares distantes entre si, enquanto t-SNE tende a aproximar pontos similares, o que em última análise facilita a identificação de aglomerados de classes.

Matematicamente, pode ser dito que t-SNE converte a distância euclidiana entre pontos no espaço alto-dimensional em probabilidades condicionais que representam similaridades. A similaridade entre um ponto x_j e outro x_i é a probabilidade condicional p_{ji}, que x_i escolheria x_j como vizinho se vizinhos fossem escolhidos em relação a sua densidade de probabilidade em uma curva Gaussiana centrada em x_i. Assim, para pontos próximos, p_{ji} é grande, enquanto para pontos distantes é pequena (dessa forma, p_{ii} seria máxima e igual a 1, entretanto para evitar distorções atribui-se o valor 0). A probabilidade condicional é então descrita pela equação (1) (VAN DER MAATEN; HINTON, 2008):

$$pji = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2) / 2\sigma_i^2} \tag{1}$$

Onde σ é a variância da Gaussiana centrada no ponto x_i.

O mesmo pode ser feito para o mapeamento em baixa dimensionalidade. Neste caso, tem-se a probabilidade condicional q_{ji}:

$$qji = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} \tag{2}$$

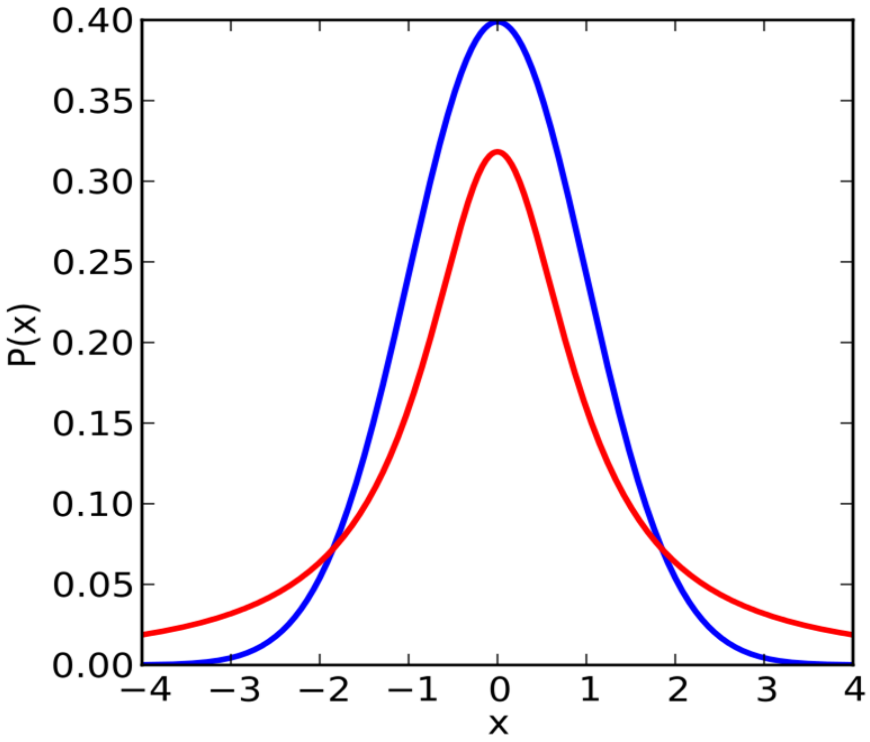
Onde y_i e y_j são os correspondentes em baixa dimensionalidade de x_i e x_j.

Se y_i e y_j mapeiam corretamente x_i e x_j em baixa dimensionalidade, então p_{ij} e q_{ij} serão iguais.

Um mapeamento tão preciso é dificilmente encontrado, visto que poucas dimensões costumam não abarcar toda a complexidade de dados de muitas dimensões, entretanto, pode-se buscar aproximar tal mapeamento da igualdade. E é exatamente o que faz o algoritmo de *t-SNE*. Realiza-se um processo iterativo até que p e q sejam o mais similares quanto possível, otimizando, assim, o resultado do algoritmo.

O método *t-SNE* apresentado em 2008, entretanto, consiste em uma variação do método *SNE* já utilizado previamente. A mudança consistiu na substituição da curva de Gauss por uma curva *t*, que deu o nome à nova técnica. Elas diferem um pouco na sua forma, tendo, a curva *t* um pico mais baixo e extremidades mais altas (ver Figura 10). Essa mudança visa reduzir a aglomeração de pontos no centro do mapa e melhorar a função de otimização do mapeamento responsável por aproximar as duas funções de probabilidade (VAN DER MAATEN; HINTON, 2008).

Figura 10: Exemplo da diferença entre uma curva t (vermelho) e uma curva Normal (azul) para um caso contendo apenas um grau de liberdade.



Fonte: Elaborado pelo autor

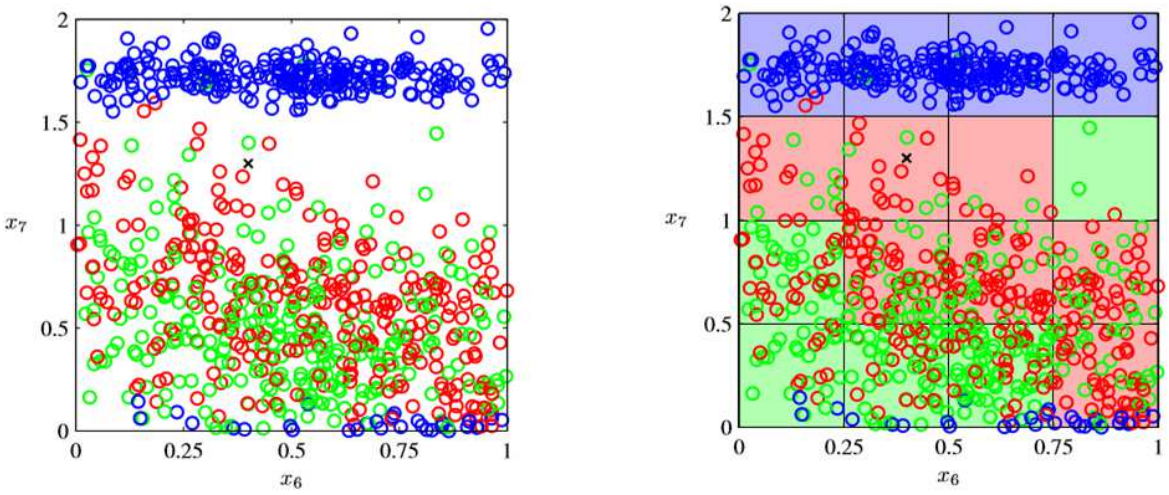
3.3.4. Modelos de Classificação

Modelos de classificação são uma classe importante de algoritmos de aprendizado de máquina. Se distinguem por conter uma variável alvo que é categórica. Modelos de regressão têm alvos contínuos, podendo ser qualquer número real entre 0 e 100, por exemplo. Já modelos de aprendizado não supervisionado não têm uma classe alvo definida (não explicitada pelo conjunto de dados original). Dessa forma, utiliza-se algoritmos de

classificação quando pretende-se atribuir uma classe conhecida a cada amostra desconhecida de um conjunto novo de dados, baseando-se em dados de treino.

Por exemplo, pode se fornecer informações sobre o estilo de vida e histórico da família de um paciente e atribuir a ele uma classe entre “alto”, “médio” ou “baixo” risco de hipertensão. Imagine que cada uma das características seja uma dimensão no espaço e cada classe tem uma cor associada a ela. Nos gráficos da Figura 11, têm-se que cada círculo representa um paciente com um risco conhecido de ter a doença no espaço bidimensional das variáveis conhecidas. No gráfico da direita, o espaço foi dividido em regiões e cada uma delas foi atribuída a uma classe de acordo com a classe das amostras contidas em seu interior. Uma região com amostras verdes e vermelhas é atribuída à classe vermelha se essa estiver em maior número dentro da região. Uma nova amostra desconhecida pode ser inserida, por exemplo, o “x” no gráfico da direita, e o algoritmo atribuirá a ela então a classe vermelha.

Figura 11: Exemplo que ilustra a ideia por trás dos algoritmos de classificação



Fonte: (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Apesar de ilustrar a ideia por trás dos algoritmos de classificação, esse modelo é limitado em sua capacidade preditiva e modelos baseados em operações estatísticas mais complexas têm melhor aplicação em conjuntos de dados reais. As subseções a seguir esperam elucidar os mecanismos de algumas delas.

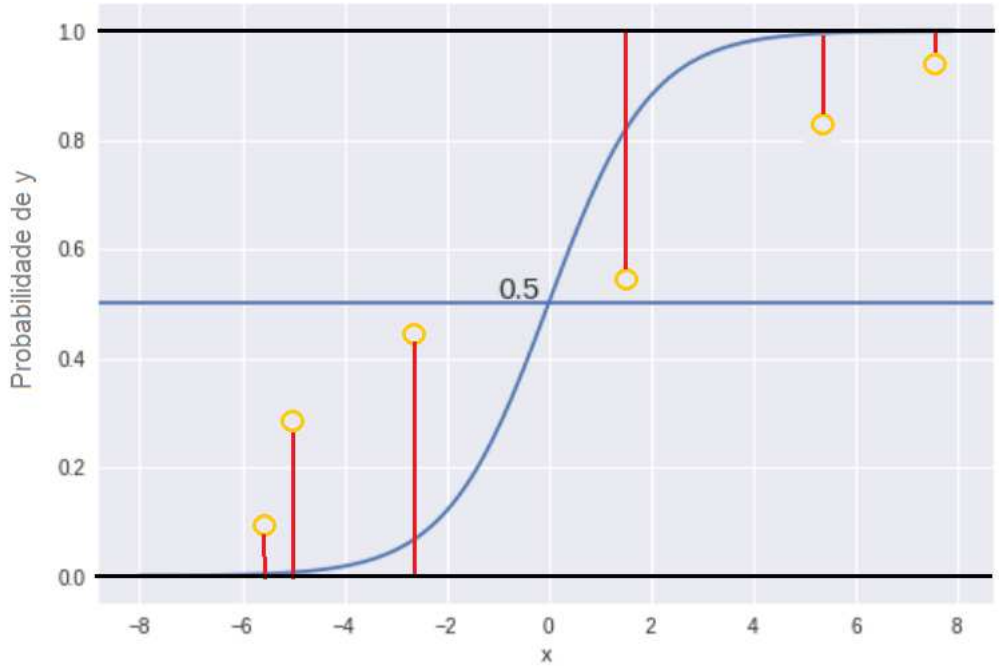
3.3.4.1. Regressão Logística

Regressão Logística é um modelo linear generalizado de aprendizado de máquina que busca minimizar uma função custo definida como função logística, que tem a forma de

uma curva sigmoide. Ela mapeia valores de probabilidade (que variam de 0 a 1) em classes. Dado um *input* X qual a probabilidade de pertencer à classe Y. Porém a resposta deve ser binária. Portanto, toda vez que o valor é superior ou inferior a um limite definido lhe é atribuído um valor de 0 ou 1 (BEN-DAVID, 2014).

Por exemplo, pode-se buscar classificar pessoas em 2 classes, “homem” ou “mulher” baseando-se no seu peso e altura. A função logística atribui um valor de probabilidade que cada pessoa têm de pertencer a uma determinada classe. Pode-se, então definir um limite em 0,5, por exemplo, em que valores abaixo dele pertencem à classe “mulher” e a cima pertencem à classe “homem” (Conforme Figura 12).

Figura 12: Regressão Logística: mapeamento de duas classes baseado na probabilidade em função de uma variável x.



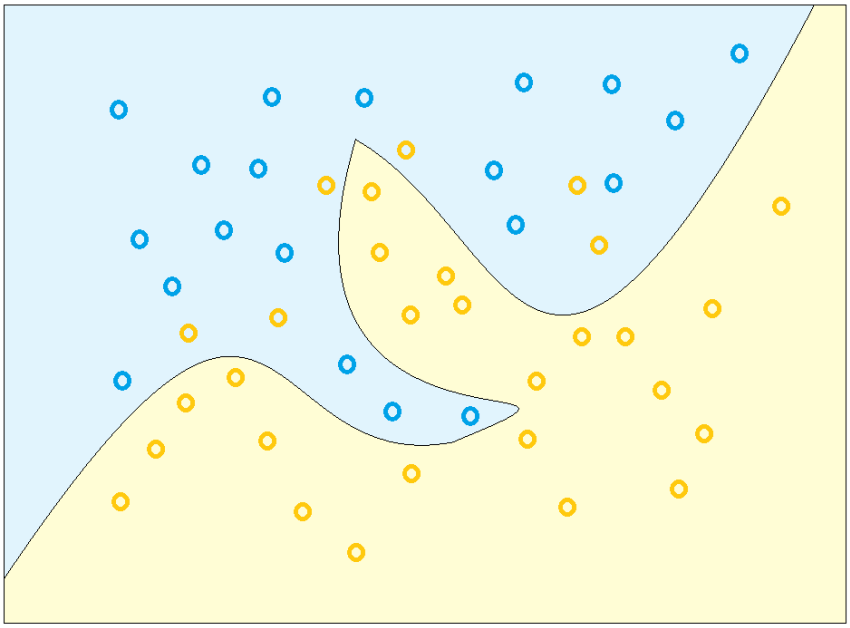
Fonte: Elaborado pelo autor

Entretanto, esse modelo não é limitado à classificação binária, podendo ser utilizado em problemas multi-classes. Neste caso, há duas abordagens: one vs. All (um contra todos), em que havendo k classes são criados k modelos para comparar as diversas classes com uma classe de referência escolhida aleatoriamente; e one vs one (um contra um) que cria k*(k-1)/2 modelos. A segunda abordagem costuma resultar em melhores modelos para casos com muitas classes, porém é mais cara computacionalmente.

3.3.4.2. K-nearest Neighbours

Traduzido como “vizinhos mais próximos”, essa técnica consiste na regra básica de que amostras que se parecem devem pertencer a uma mesma classe. Dessa forma, o algoritmo calcula a distância entre um ponto e um número definido *k* de vizinhos. Um ponto no espaço tem, por exemplo, 15 vizinhos próximos pertencentes a duas classes. Àquela que tiver mais representantes será atribuída ao ponto. Isso pode ser feito para todos os pontos num espaço multidimensional e a partir daí é possível traçar uma linha limite de decisão, que define regiões de classificação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Um parâmetro importante é a quantidade de vizinhos próximos levada em conta durante o aprendizado. Um número muito grande pode gerar *underfitting*, ou seja, ser muito genérico e não representar fielmente os dados de treino. Já um número muito pequeno (1 vizinho, no limite) pode gerar *overfitting*, ou seja, ele representa muito fielmente o conjunto de dados de treino e é pouco generalizável a novos conjuntos (ver Figura 13).

Figura 13: Ilustração da definição de um contorno de decisão em 2 dimensões para 2 classes utilizando o método K-nearest neighbours (KNN)



Fonte: Elaborado pelo autor

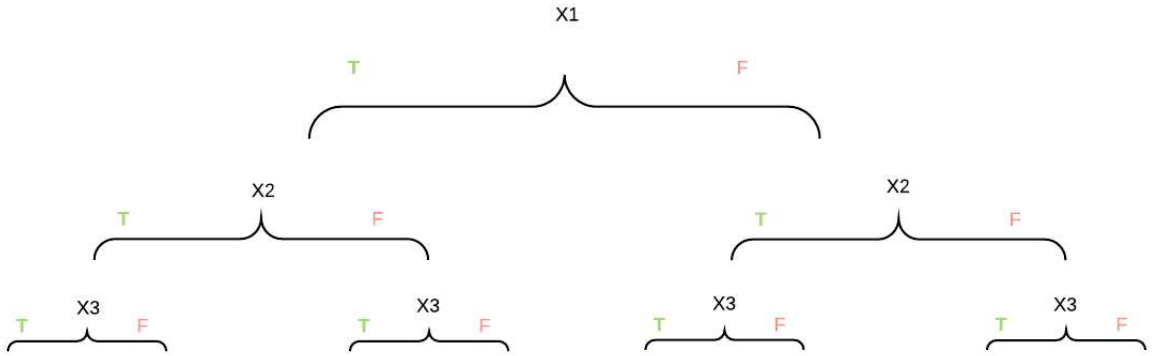
Voltando a Figura 11, se o espaço fosse dividido em 2 quadrados, teríamos um caso claro de *underfitting*, já se fosse dividido em um número de regiões que fosse igual ao número de amostras de treino, teríamos *overfitting*. Apesar da sua simplicidade, este tipo de

classificador é bem-sucedido em uma série de problemas de classificação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.3.4.3. *Decision Trees*

Decision Trees (árvores de decisão) é um método de aprendizado de máquina não-linear que utiliza condicionais ("se isso então aquilo") para definir regras de decisão. Tem este nome pois sua estrutura é semelhante a galhos que se ramificam em uma árvore. Para ramificações que tenham como resposta valores contínuos e não binários, pode-se definir um valor limite a partir do qual o nodo se ramifica entre "acima" e "abaixo" do limite (similarmente ao processo de classificação por regressão logística), dessa forma binarizando a solução.

Figura 14: Gráfico de uma árvore de decisão simples. T e F significam verdadeiro e falso em cada um dos nós de ramificação.



Fonte: Elaborado pelo autor

No nível 1 (x1) o algoritmo divide o conjunto de treino em dois subconjuntos baseado em uma única característica *k* e um limiar *t_k*. A escolha de *k* e *t_k* é feita de forma a produzir subconjuntos mais puros, ou seja, que produzam maior separação entre as classes. A medida utilizada nessa ordenação é o ganho de informação. Se após um nodo há pouca separação entre classes, tal ramificação contribui pouco para uma resposta adequada do modelo, portanto o ganho de informação é pequeno (BEN-DAVID, 2014). Após essa etapa, o processo é repetido nos níveis subsequentes (x2, x3...). Entretanto, o algoritmo de *Decision Tree* é conhecido por ser ganancioso. Isso quer dizer que a otimização da separação em cada nível não leva em conta a pureza dos subconjuntos posteriores nos níveis mais abaixo na árvore. A solução ideal configura um problema NP-completo, cujo tempo de processamento

aumenta exponencialmente em relação à quantidade de dados. Por isso, costuma-se aceitar a primeira solução com resultados razoavelmente bons.

Na Figura 14, cada ramificação leva a caminhos diferentes e num espaço de atributos define uma superfície de separação de regiões ou limite de decisão (similar aos métodos anteriores). No caso das árvores de decisão, estas superfícies tendem a ser ortogonais umas às outras, o que dificulta a classificação se o conjunto de dados estiver rotacionado. A utilização do PCA pode contribuir para resolver essa dificuldade. Outras dificuldades são à sua propensão ao *overfitting* e a instabilidade frente a variações pequenas no conjunto de treino.

3.3.4.4. *Random Forests*

Random Forests ou *Random Decision Forests* (Florestas de decisão aleatória, em tradução livre) é um outro modelo de aprendizado de máquina baseado em árvores de decisão. De forma análoga a uma floresta, este modelo consiste de muitas árvores desse tipo. Ele computa os resultados de classificação para diversas árvores e oferece como resposta o resultado que for mais recorrente entre as árvores individuais.

Entretanto, é preciso que as árvores sejam diferentes umas das outras. Isso é possível fazendo-se duas alterações na computação de cada árvore de decisão: primeiro, a escolha da ordem dos atributos é aleatória, mas o limiar do ganho de informação não; segundo, o conjunto de amostras de treino de tamanho N é reordenado contendo possíveis repetições de amostras mas mantendo o tamanho N original (conhecido na literatura por *bagging*) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

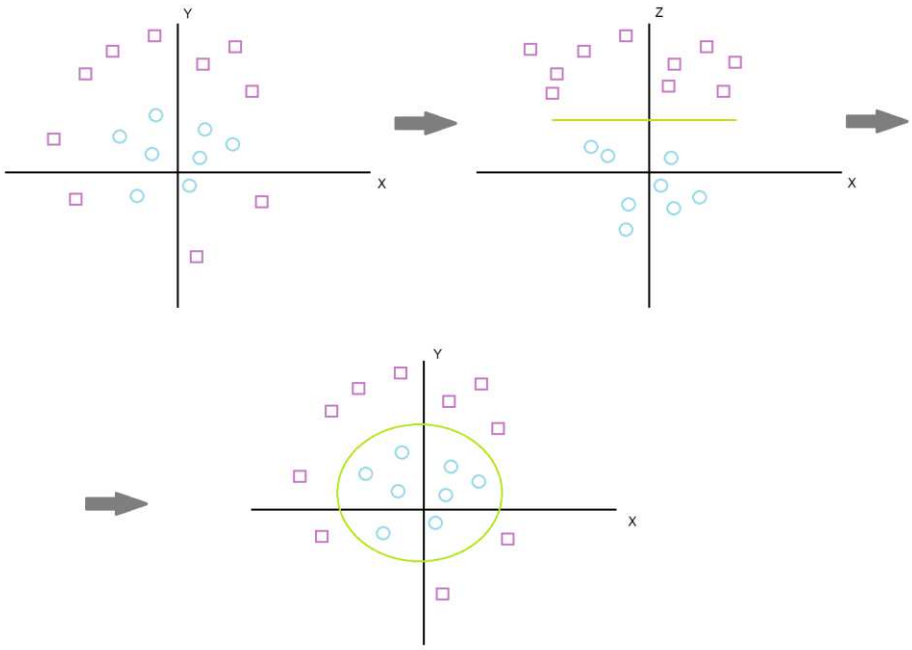
Como o algoritmo de *decision trees* é bastante sensível aos dados de treino, pequenas variações podem resultar em árvores completamente diferentes. Dessa forma, uma floresta aleatória de árvores pode levar em consideração dados de treino diferentes e melhorar a capacidade de generalização do modelo.

3.3.4.5. *Support Vector Machine*

Support Vector Machine (SVM) é um classificador estatístico formalmente descrito por um hiperplano em um espaço n-dimensional. Em duas dimensões isso equivale a traçar uma linha, definindo duas regiões distintas no gráfico, cada uma pertencente a uma classe. Foi originalmente proposto em 1963, mas foi aperfeiçoado em 1992 com a inclusão dos *kernels (CORTES; VAPNIK, 1995)*.

Pode ser entendido como uma generalização de classificadores lineares para casos onde as classes se sobrepõe e não são linearmente separáveis. O método define um contorno de decisão linear em um espaço dimensional maior que o original e ao retornar a ele descreve um contorno não linear (Figura 15). O hiperparâmetro *kernel* é usado nessa transformação e pode ser de diversos tipos, como por exemplo, linear, polinomial ou gaussiano. É ele quem define a distribuição dos dados no espaço dimensional maior que o original (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Figura 15: SVM - Aplicação de um classificador linear em maior dimensão a classes não linearmente separáveis.



Fonte: Elaborado pelo autor

3.3.4.6. *Naive Bayes*

Naive Bayes Classifier, ou classificador ingênuo de Bayes é uma ferramenta estatística muito utilizada em inteligência artificial, especialmente em categorização de textos e que apesar do nome, manteve alta sua reputação desde seu surgimento nos anos 1960. Este método utiliza o Teorema de Bayes (a) para calcular a probabilidade de um evento baseado nas variáveis de entrada. O método é dito “ingênuo” pois assume duas hipóteses importantes como verdadeiras: a primeira de que não há correlação entre as variáveis de entrada e a

segunda de que todas elas têm a mesma importância em definir uma classe (HAND et al., 2019). Aplicada à equação (3), tal hipótese resulta em (4).

$$P(y \mid X_n) = \frac{P(y)P(X_n|y)}{P(XN)} \tag{3}$$

$$P(y \mid X_n) = \frac{P(y)\prod_{i=1}^{n} P(X|y)}{P(X_n)} \tag{4}$$

Para melhor compreensão, segue um exemplo: um jogador de golfe define um bom dia para praticar o esporte de acordo com algumas variáveis climáticas, digamos temperatura, umidade, velocidade do vento e volume de chuva. A primeira hipótese implica em considerar que um dia quente não significa um dia com pouco vento (as variáveis são independentes). A segunda hipótese implica dizer, por exemplo, que o volume de chuva é tão relevante quanto a umidade do ar para classificar o dia como bom para a prática.

No exemplo, a aplicação da equação (a) equivale a dizer que dadas as condições do dia a probabilidade de este ser um dia bom ou ruim para a prática do esporte é P(y|X), ou seja, y (classe do dia) está condicionado à X (condição climática). O cálculo é feito para todas as possíveis classes (binárias ou não) e a que obtiver uma probabilidade maior é obtida como resposta.

3.3.5. *Avaliação e seleção de modelos*

Existem diversas maneiras de avaliar modelos de classificação. A acurácia é uma medida simples e objetiva de apresentar os resultados, mas que ainda assim contém a informação necessária para avaliação e escolha de um modelo. Mesmo assim, há algumas questões que devem ser consideradas.

Durante a etapa de pré-processamento dos dados, antes da aplicação de um modelo, é possível que haja algum viés ao dividir um *dataset* em duas partes. Como as divisões são aleatórias, cada partição é diferente e produz modelos com resultados específicos àquela divisão. Por exemplo, uma classe pode ser bem populosa em uma e pouco em outra, o que pode gerar resultados diferentes cada vez que o modelo é aplicado. Entretanto, uma análise de

resultados apropriada deve permitir verificar a capacidade de generalização de um modelo e evitar tais variações (BEN-DAVID, 2014).

O método *k-fold* de validação cruzada é utilizado para melhor avaliar a performance de classificadores de aprendizado de máquina. Ele consiste em subdividir o *training set* em *k* partes, gerar um modelo baseado em *k-1* e avaliar em 1 (uma espécie de subdivisão treino/teste). Entretanto, o processo é iterativo e a cada iteração uma parte é escolhida como *test set*. Assim, obtém-se *k* valores de acurácia. A partir deles se obtém uma média e um desvio padrão. Dessa forma, pode-se dizer que o modelo possui uma acurácia que fica dentro de uma faixa de valores e que é mais generalizável a outros *datasets.* O que é especialmente importante em modelos mais sensíveis a variações nos dados, como é caso de *Decision Trees (BEN-DAVID, 2014).*

Além disso, como já foi visto, cada modelo de classificação possui parâmetros que precisam ser definidos pelo usuário (o número *k* de vizinhos em *KNN*, por exemplo) são os chamados hiperparâmetros (diferentes dos parâmetros que são aprendidos pelo modelo). Entretanto, testar todas as possíveis combinações de parâmetros seria dispendioso demais. A ferramenta *Grid Search,* permite avaliar essas combinações de maneira automatizada, servindo como uma forma de otimização dos modelos de classificação. Assim pode-se fazer uma comparação entre modelos em suas melhores performances.

3.4. *Aprendizado de máquina para caracterização de microplásticos*

Devido à ubiquidade dos resíduos plásticos no oceano, em particular os microplásticos e aos seus efeitos e potenciais impactos na vida, um número crescente de publicações têm focado no estudo desses poluentes e buscado preencher lacunas de pesquisa.

As técnicas de espectroscopia são frequentemente utilizadas para a identificação de partículas poliméricas e para sua caracterização por tipo de resina. Entretanto, devido à grande quantidade de microplásticos presentes no ambiente, amostragens podem conter centenas ou até milhares de partículas desse tipo.

A análise de dados espectroscópicos é feita visualmente por um *expert* ou então por busca em bibliotecas, que frequentemente precisa ser confirmada visualmente. Mesmo a última é demorada, pois os espectros de polímeros ambientais costumam ter baixa correlação com os da base de dados, devido principalmente à degradação, e os resultados ainda precisam ser confirmados por análise visual. Portanto, esse método é trabalhoso e muito limitado: é pouco automatizável, requer a análise de um *expert* treinado em espectroscopia, é demorado,

tem baixo índice de acerto, assim frequentemente apenas uma parte do conjunto de amostras coletadas é de fato identificada.

Surge então o interesse em aplicar ferramentas analíticas mais modernas à análise de microplásticos ambientais. Como já foi visto, o aprendizado de máquinas vem sendo utilizado em diversos nichos de aplicação nas mais diversas áreas, para tratar eficientemente problemas que tenham grande volume de dados disponíveis.

Para entender o estado da arte da automatização da caracterização de microplásticos foi feita uma revisão bibliográfica sistemática conforme descrito por (KITCHENHAM; CHARTERS, 2007) utilizando a *query* "Microplastic AND (Characterization OR Identification) AND (FTIR OR Raman) AND Automat*" para busca na base de dados *Scopus.* Foram encontradas 13 publicações. Os resultados da busca por espectroscopia Raman foram incluidos devido à similaridade desta técnica com o FTIR e a potencial aplicação de ferramentas de automatização de identificação à ambas.

Uma síntese da revisão pode ser vista na Tabela 1:

Tabela 1: Revisão da literatura e identificação dos métodos de automação

Publicação	Ano	Jornal/Revista	Espectroscopia	Tamanho mínimo MP	Método de Automação
Automated identification and quantification of microfibres and microplastics	2019	Analythical Methods	μ-FTIR Imaging	11μm	PCA
Multi-temporal surveys for microplastic particles enabled by a novel and fast application of SWIR imaging spectroscopy e Study of an urban watercourse traversing the city of Berlin, Germany	2018	Environmental Pollution	SWIR Imaging	450μm	Busca em biblioteca
A Methodology for the Fast Identification and Monitor- ing of Microplastics in Environmental Samples using Random Decision Forest Classifier	2019	Analytical Methods	μ-FTIR Imaging	10μm	*Spectral Descriptor + Random Decision Forests Classifier*
A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the Mediterranean Sea	2019	Chemosphere	FTIR	315μm	*KNN Classifier*
Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy	2018	Analytical and Bioanalythical Chemistry	μ-FTIR Imaging	11μm	Busca em biblioteca
Robust Automatic Identification of Microplastics in Environmental Samples Using FTIR Microscopy	2019	Analytical Chemistry	ATR-FTIR + μ-FTIR Imaging	75μm	Isolar bandas do IV + Busca em biblioteca
FTIR and Raman imaging for microplastics analysis: State of the art, challenges and prospects	2019	Trends in Analytical Chemistry	FTIR + Raman	indefinido	Revisão - sem menção a aprendizado de máquina
Implementation of an open source algorithm for particle recognition and morphological characterisation for microplastic analysis by means of Raman microspectroscopy	2019	Analytical Methods	μRaman	indefinido	Método Otsu
Development of an optimal filter substrate for the identification of small microplastic particles in food by micro-Raman spectroscopy	2017	Analytical and Bioanalytical Chemistry	μRaman	1μm	não
A semi-automated Raman micro-spectroscopy method for morphological and chemical characterizations of microplastic litter	2016	Marine Pollution Bulletin	μRaman	335μm	Seleção automática de partículas
Molecular identification of polymers and anthropogenic particles extracted from oceanic water and fish stomach e A Raman micro- spectroscopy study	2018	Environmental Pollution	μRaman	100μm	Busca em biblioteca
Identification of microplastics using Raman spectroscopy: Latest developments and future prospects	2018	Water Research	Raman + μRaman	indefinido	Seleção automática de partículas
An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis	2017	Analytical Methods	μ-FTIR Imaging	11μm	Busca em biblioteca

Fonte: Elaborado pelo autor

A primeira informação que fica clara é que apesar da pouca extensão da literatura dedicada ao assunto, as publicações são bastante recentes, muitas datando inclusive do ano

2019. É interessante observar que com exceção de um artigo (KEDZIERSKI et al., 2019), todos os outros abordaram a automatização em técnicas de espectroscopia por imagem (μFTIR ou μRaman), capaz de identificar amostras de tamanho bastante reduzido. Isso pode ser devido ao fato de a espectroscopia por imagem gerar um número muito maior de espectros, sendo que um espectro é gerado para cada *pixel* em uma imagem e cada uma pode conter de centenas a milhões de *pixels*, a depender do equipamento utilizado. Assim, fica claro que a automatização se faz necessária e é de fato útil quando cresce a quantidade de dados.

Entretanto, é possível perceber que grande parte dos estudos se dedicam ao aprimoramento das técnicas de busca em bibliotecas espectrais. Um artigo buscou a inclusão de espectros degradados no ambiente para melhorar as correspondências das amostras com a referências durante as buscas em bibliotecas (PRIMPKE et al., 2018). Apenas 2 artigos apresentaram classificadores como alternativa, o que mostra que já existe um interesse em aplicar ferramentas de aprendizado de máquina para a automação do processo de caracterização de microplásticos, mas deixa lacunas na forma como o tema pode ser abordado (HUFNAGL et al., 2019; KEDZIERSKI et al., 2019). Um terceiro artigo apresentou um método que inclui a redução de dimensionalidade por PCA, mostrando que essa etapa pode ser crucial para reduzir o tempo de análise e melhorar os resultados (DIAS, 2019).

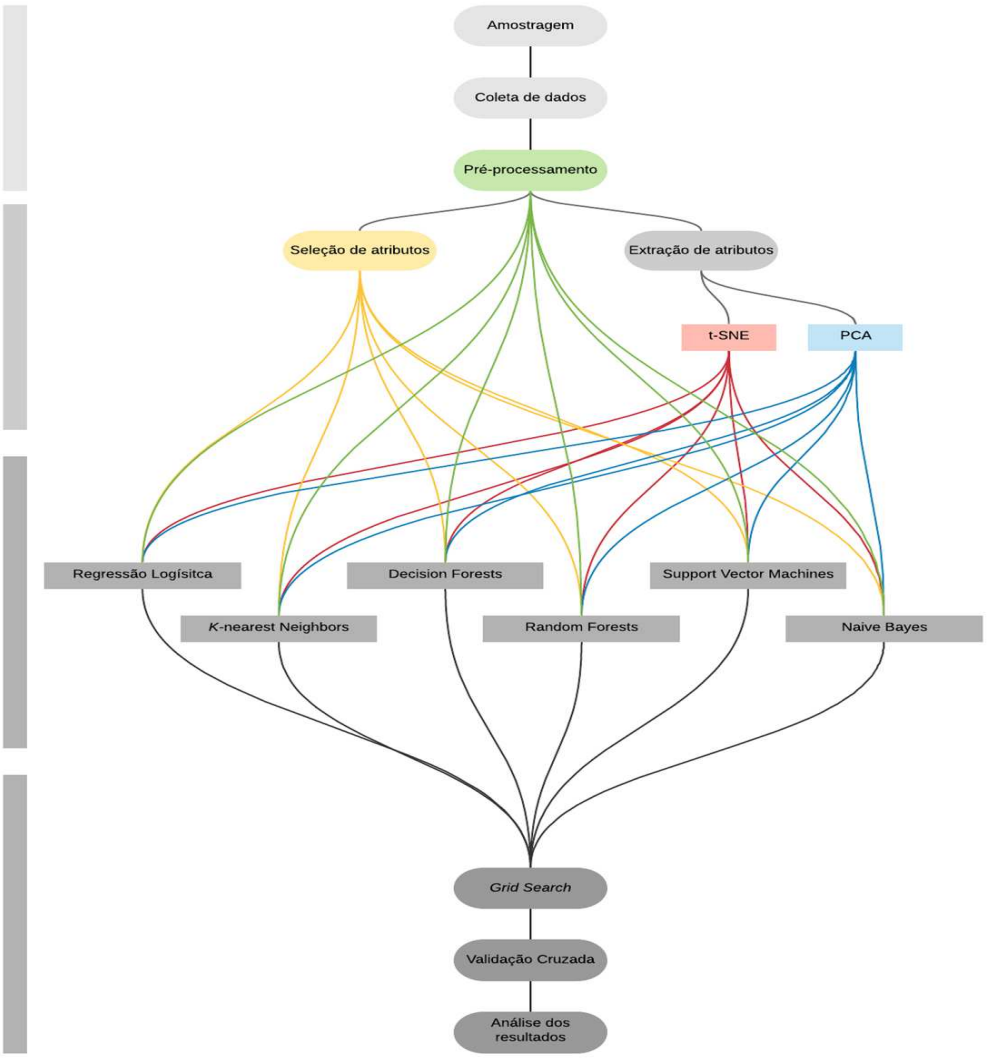
Atenção especial deve ser dada ao artigo *A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the Mediterranean Sea*, pois os dados espectroscópicos utilizados no presente trabalho foram obtidos e disponibilizados nesta publicação (mais detalhes na seção Materiais e Métodos). Com a aplicação de um classificador *KNN*, os autores foram capazes de automatizar a identificação de mais de 4000 espectros com acurácia de 90%, percentual significativamente maior que os cerca de 75% comuns às buscas em bibliotecas.

Apesar do sucesso da metodologia descrita no artigo mencionado, não foi avaliada a acurácia de classificação de outros modelos de AM e nem de técnicas de RD. Portanto, segue relevante o objetivo deste trabalho, visando desenvolver uma metodologia mais eficiente em relação às já descritas na literatura.

4. MATERIAIS E MÉTODOS

Essa seção discorrerá sobre os materiais utilizados e métodos aplicados no desenvolvimento do trabalho. Um fluxograma da metodologia proposta para classificação de espectros de infravermelho de amostras de microplástico marinho está apresentado na Figura 16.

Figura 16: Fluxograma da metodologia de caracterização de microplásticos desenvolvida neste trabalho



Fonte: Elaborado pelo autor

Tanto o pré-processamento dos dados como a aplicação das ferramentas de aprendizado de máquina em si foram implementadas utilizando a linguagem de programação Python, devido à grande disponibilidade de bibliotecas dedicadas ao processamento de dados

e ao aprendizado de máquina. Foi utilizada a versão 3.7.4 do Python, a versão 0.21.3 do scikit learn e a versão 0.25.1 do pandas.

O código utilizado para o tratamento dos dados, modelagem e análise dos resultados será anexado ao trabalho final. Entretanto, estará apresentada somente uma versão do código, restrita apenas ao modelo que obtiver melhor performance.

4.1. REVISÃO BIBLIOGRÁFICA SISTEMÁTICA

Com o intuito de reunir evidências científicas que concernem ao uso de ferramentas de aprendizado de máquina para identificação de poluentes poliméricos e situar apropriadamente este trabalho e seus objetivos no contexto da área, foi utilizada a base de dados bibliográfica Scopus. Foram feitas buscas utilizando combinações de termos relevantes ao tópico. Rigor maior foi dado à seção que trata de inteligência artificial para caracterização de microplásticos, tema específico deste trabalho, onde utilizou-se a *query*: “Microplastic AND (Characterization OR Identification) AND FTIR AND Automat*”. Então, buscou-se identificar os artigos relevantes para a compreensão do tópico em questão por meio da leitura dos títulos dos trabalhos, com atenção também ao ano de publicação e número de citações. Em seguida, foram lidos os resumos dos trabalhos escolhidos na etapa anterior e mantidos só os que provaram ser importantes para a revisão. Por fim, o corpo dos trabalhos filtrados foi lido com maior atenção. Os resultados desta revisão, porém, já foram apresentados na seção específica no capítulo de revisão bibliográfica.

4.2. AMOSTRAGEM

As amostras foram coletadas a bordo do veleiro de pesquisa científica francês TARA em uma expedição pelo Mar Mediterrâneo entre maio e novembro de 2014, por pesquisadores não relacionados ao autor do presente trabalho. As datas e posições geográficas das coletas estão disponíveis em *Pangaea Data Publisher* (https://www.pangaea.de), mas não foram incorporadas na análise deste trabalho.

Foi usada uma rede manta longa de 440 cm de comprimento, abertura de 16 x 60 cm e mesh de 333 μm (Figura 17). Tal equipamento é tradicionalmente utilizado para coleta de plâncton na superfície da água e vêm sendo útil para amostragem de microplásticos. Um total de 120 locais foram selecionados baseados em análises de cor em imagens de satélite. Em

cada um, a rede foi arrastada por cerca de 60 minutos a uma velocidade de 2,5 nós, o que permitiu a filtragem de em torno de 507 m3 de água. Um total de 13.374 amostras foram coletadas e separadas (utilizando um microscópio de dissecação) nessas expedições.

Figura 17: Rede manta durante coleta

Fonte: (BERGMANN, GUTOW; KLAGES, 2015)

4.3. COLETA DE DADOS ESPECTROSCÓPICOS

Os dados utilizados neste trabalho, são fruto da análise espectroscópica por FTIR-ATR de 4.064 amostras selecionadas aleatoriamente do universo amostral inicial. Dessas, em torno de 1000 haviam sido previamente identificadas por tipo de polímero, num total de 17 classes. Devido à necessidade de amostras classificadas para o aprendizado supervisionado somente estas foram utilizadas. Os espectros foram coletados em 1762 frequências de onda na faixa do espectro eletromagnético entre 4000 e 600 cm^{-1}.

Estes dados foram publicados junto ao artigo *A machine learning algorithm for high throughput identification of FTIR spectra: Application on microplastics collected in the Mediterranean Sea,* publicado pela revista Chemosphere no ano de 2019 (KEDZIERSKI et al., 2019) e utilizados neste trabalho com o consentimento dos autores. É preciso ressaltar que os dados podem ser enviesados para a condição da poluição por microplásticos da região onde foram coletados e não refletem, necessariamente, as condições de regiões diversas, portanto sua aplicação em outros contextos pode ser limitada.

4.4. PRÉ-PROCESSAMENTO

Os dados foram pré-processados utilizando operações comuns em análises espectrais. Começou-se pela remoção dos picos associados à água e ao dióxido de carbono, ruído que pode gerar distorções. Em seguida, foram feitos: correção da linha base, *smoothing* da curva (arredondamento dos picos) utilizando o método Savitszky-Golay, e então a normalização.

Posteriormente verificou-se que 4 classes possuíam poucas amostras em seu conjunto, o que dificultaria o aprendizado e criação dos modelos, portanto elas foram descartadas. Assim, a base de dados de referência para o aprendizado de máquina consistiu em 959 espectros de amostras de 13 tipos de polímeros.

4.5. REDUÇÃO DE DIMENSIONALIDADE

A redução de dimensionalidade foi utilizada antes da aplicação dos modelos de classificação com o intuito de reduzir a capacidade e tempo de processamento necessários ao aprendizado dos modelos de classificação e, também, a fim de avaliar a possível melhora dos resultados de acurácia dos classificadores. Além disso, o mapeamento das amostras apresentados em gráficos bidimensionais pretendia a visualização dos dados, identificação de padrões entre as amostras e, portanto, serviriam também como análise exploratória.

Foram utilizados 3 métodos diferentes de redução de dimensionalidade descritos na seção 31 e a aplicação foi feita utilizando a linguagem Python. A eficiência de redução foi calculada para cada técnica pela equação (5) e a influência nos resultados de acurácia dos classificadores foi feita comparativamente após a aplicação dos modelos.

$$\frac{\Sigma Variáveis iniciais - \Sigma Variáveis finais}{\Sigma Variáveis iniciais} \tag{5}$$

4.6. CLASSIFICAÇÃO

Foram implementados 6 modelos de classificação, ou classificadores. Como cada modelo realiza um conjunto específico de operações matemáticas, esperava-se verificar qual deles é mais adequado ao problema da identificação de espectros de FTIR de microplásticos ambientais.

Os modelos foram aplicados nos dados completos após o pré-processamento e em seguida nos dados após cada método de redução de dimensionalidade. Assim, foram obtidos 4

conjuntos de resultados, cada um contendo os 6 modelos, totalizando 24 resultados (ver Figura 16).

4.7. AVALIAÇÃO E SELEÇÃO DE MODELOS

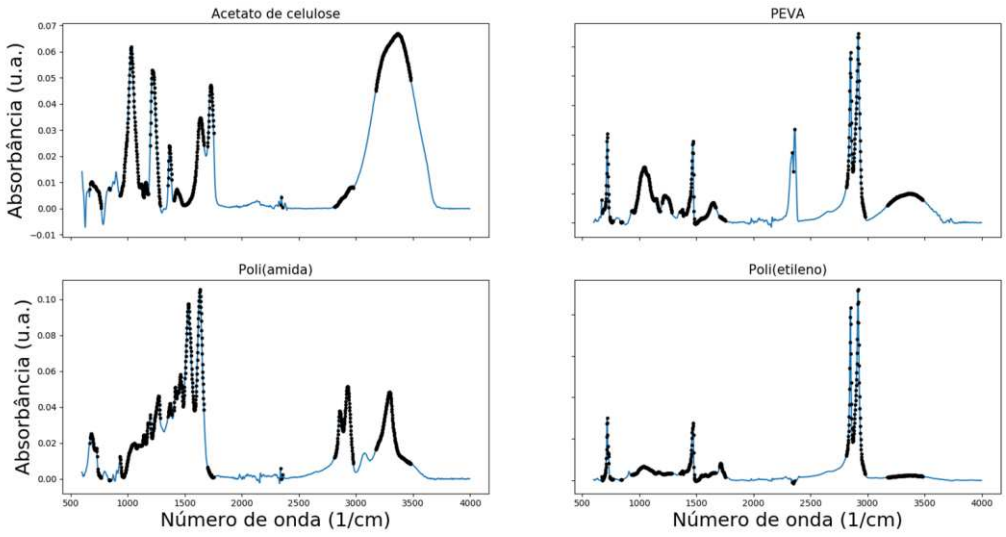
Cada modelo possui hiper-parâmetros passíveis de alteração, o que pode fazer variar o resultado de acurácia. Entretanto, a alteração manual dos parâmetros produziria um número incrivelmente grande de combinações e resultados. Para superar essa dificuldade foi utilizada a ferramenta *Grid Search*, que permite identificar a melhor combinação dos hiper-parâmetros testados para cada modelo de forma iterativa e automática. Por fim, o resultado de cada modelo aplicado a cada conjunto de dados sem RD e com RD otimizados por *Grid Search* foi verificado por validação cruzada *k-fold* (descrita no item 42) *com k = 10* e avaliado pelo seu percentual de acurácia.

5. RESULTADOS E DISCUSSÃO

Nesta seção será feita a apresentação dos resultados obtidos ao final da aplicação das metodologias, mas primeiro serão apresentados os resultados parciais e de análise exploratória obtidos pela aplicação das técnicas de redução de dimensionalidade.

Como já foi dito, a etapa de RD foi realizada com a finalidade de reduzir o tempo de computação da aplicação dos modelos de classificação, bem como de avaliar a melhora da acurácia nos resultados finais e realizar análise exploratória. Para a técnica de redução por filtro de baixa variância, a redução foi feita de forma a excluir as frequências que variavam pouco. Um limite de variância foi definido em 0.0001, observando os valores dessa grandeza para as frequências que continham alguma informação. Quando uma frequência não variava acima desse limite, ela deveria conter apenas ruído e era eliminada. A Figura 18 ilustra o resultado do filtro superpondo os pontos remanescentes após o filtro ao espectro original da mesma amostra para quatro exemplos de polímeros diferentes.

Figura 18: Espectros originais (azul) e filtrados (preto) para quatro polímeros de classes diferentes selecionados aleatoriamente do conjunto de dados



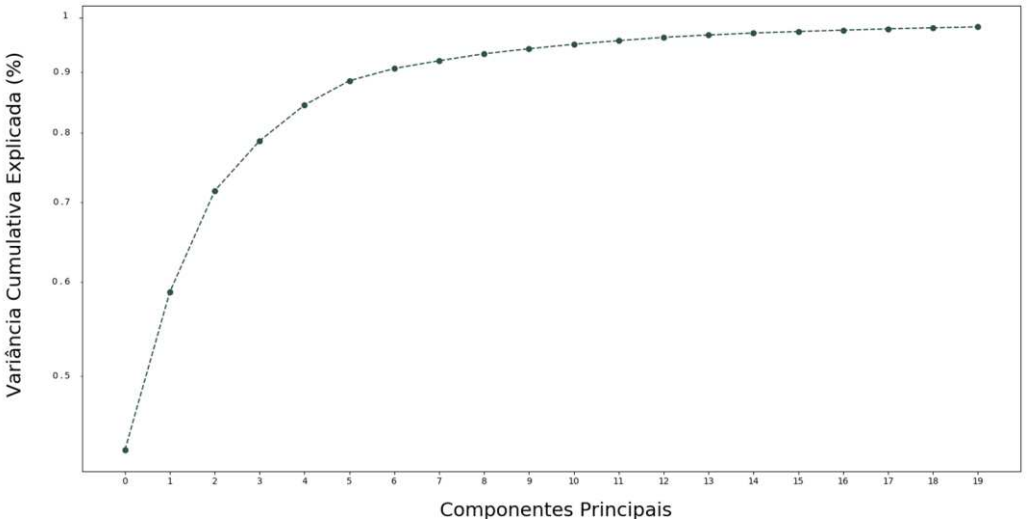
Fonte: Elaborado pelo autor

Pode-se observar que as frequências mantidas correspondem em geral aos picos característicos dos polímeros em questão. Para alguns polímeros, frequências do espectro que foram mantidas contém apenas ruído, porém as mesmas frequências são características de outros polímeros. Essa diferença permite que os algoritmos de classificação façam distinções entre as classes, mas sem a necessidade de incluir frequências que não são relevantes para a

caracterização de nenhuma delas, dessa forma, reduzindo o tempo necessário para a computação dos modelos e possivelmente melhorando sua acurácia.

Para o PCA foram escolhidas as 20 primeiras componentes principais que representaram 98% da informação contida no conjunto de dados (conforme Figura 19). Dessa forma, foram mantidas apenas 20 PCs (1,13% das variáveis) que contém quase a totalidade da informação. Novamente, esta decisão é arbitrária, sendo que o limite de variância cumulativa explicada adotado poderia ser outro. Por exemplo, por volta da oitava PC, o gráfico aproxima-se de um patamar onde o incremento em y é pequeno para cada avanço em x.

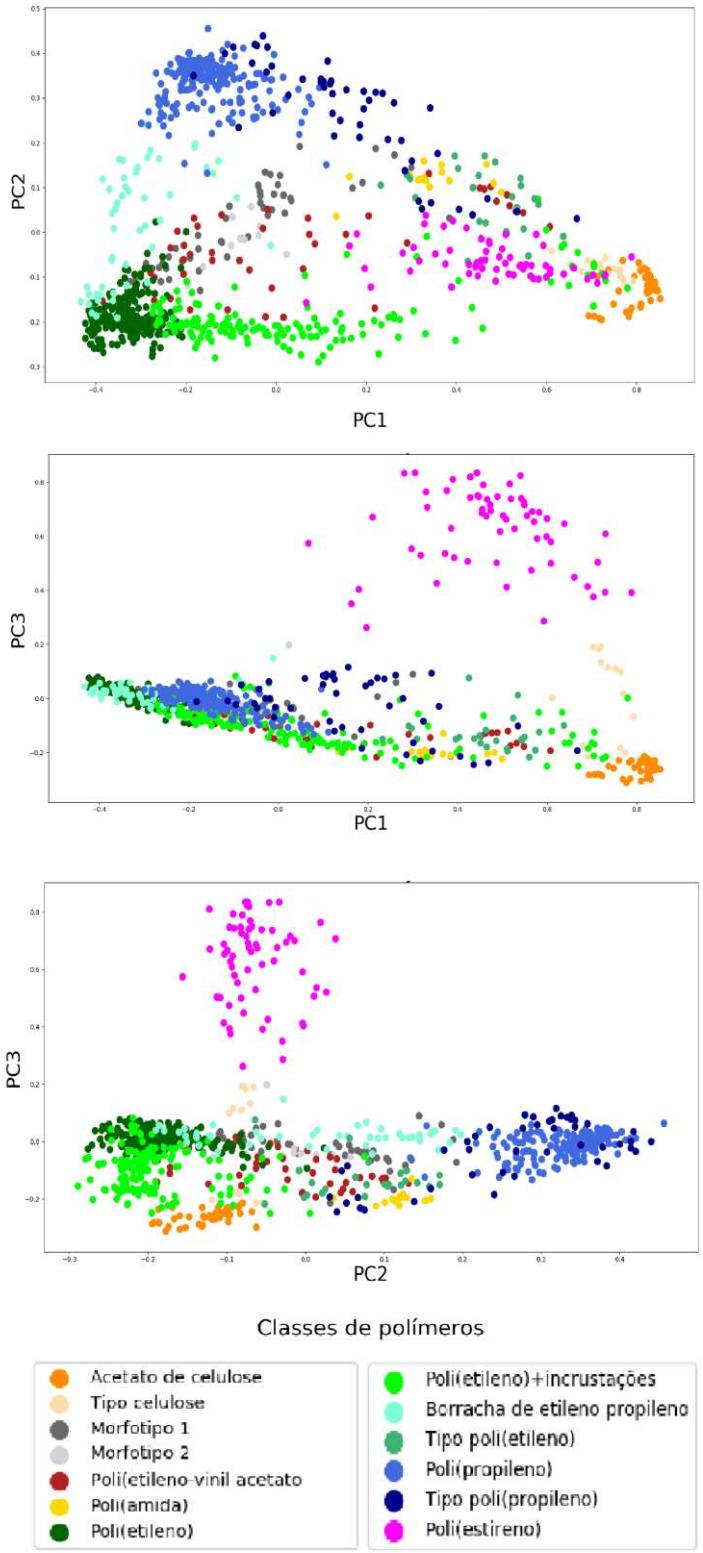
Figura 19: Variância cumulativa explicada pelas primeiras 20 componentes principais (Nota: PC1 está representada por 0 no eixo das abscissas)



Fonte: Elaborado pelo autor

Utilizando-se as novas variáveis, é possível visualizar as amostras em gráficos de dimensões bastante representativas. A Figura 20 mostra os mapeamentos de todas as amostras no conjunto de dados para as três primeiras componentes principais, contendo 71,5% da informação no total. Pode-se perceber que amostras (representadas por pontos) de um mesmo material tendem a formar aglomerados. Isso porque seus espectros são bastante semelhantes (porém não iguais) e tendem a ter um mapeamento (de frequências para PCs) também parecido. Pode-se dizer que existe uma estrutura nos dados, referente à similaridade dos espectros de uma mesma classe e que esta pode ser visualizada nestes mapeamentos.

Figura 20: Mapeamentos das amostras do conjunto de dados para PC1, PC2 e PC2

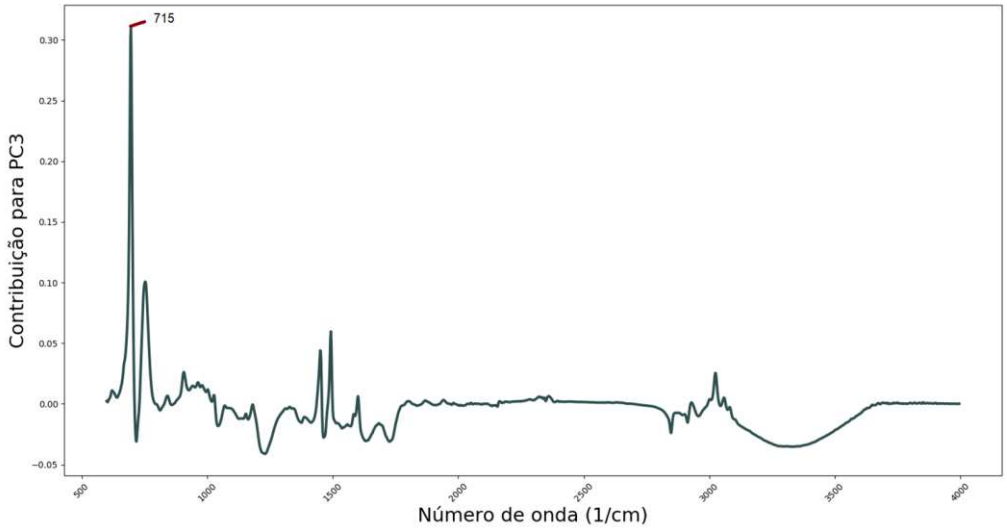


Fonte: Elaborado pelo autor

É possível notar, que algumas classes de polímeros são mais distintas das outras em determinadas PCs. Dito de outra forma, algumas classes se diferenciam mais das outras em

certos eixos. O poli(estireno), por exemplo, diferencia-se bastante das outras classes pela PC3. Neste sentido, é interessante observar o gráfico da Figura 21 que mostra a contribuição de cada frequência original para essa PC.

Figura 21: Contribuição das frequências dos espectros de infravermelho para a terceira componente principal. O pico em 715 está associado ao anel aromático do poli(estireno)



Fonte: Elaborado pelo autor

Percebe-se claramente um grande pico por volta de 715cm-1. Esta banda passou pelo filtro de baixa variância (Figura 18) e está associada a presença do anel aromático na composição química de um material, sendo este o caso, justamente, do poli(estireno). De forma análoga, pode-se identificar as frequências que mais contribuíram para PC1, a componente contendo a maior parte da informação. Na Figura 22, foram destacadas algumas frequências importantes para essa componente, mas igualmente importantes para a caracterização dos polímeros presentes no conjunto de dados. De fato, quando um *expert* faz a caracterização visual de um espectro, é justamente nestas frequências que ele está interessado. Os picos em 2910 cm^{-1}, 1725 cm^{-1}, 1475 cm^{-1} e 1025 cm^{-1}, por exemplo, são geralmente associados às ligações C-H, C=O, C=C do anel aromático e C-O respectivamente (JUNG et al., 2018). Os gráficos das Figura 21 e Figura 22 podem ser entendido como sendo a representação de todos os espectros do conjunto de dados condensados em apenas um e podem servir como um indicativo dos materiais presentes em um conjunto de dados ainda a ser analisado.

Figura 22: Contribuição das frequências dos espectros de infravermelho para a primeira componente principal. (Algumas frequências de interesse foram destacadas)

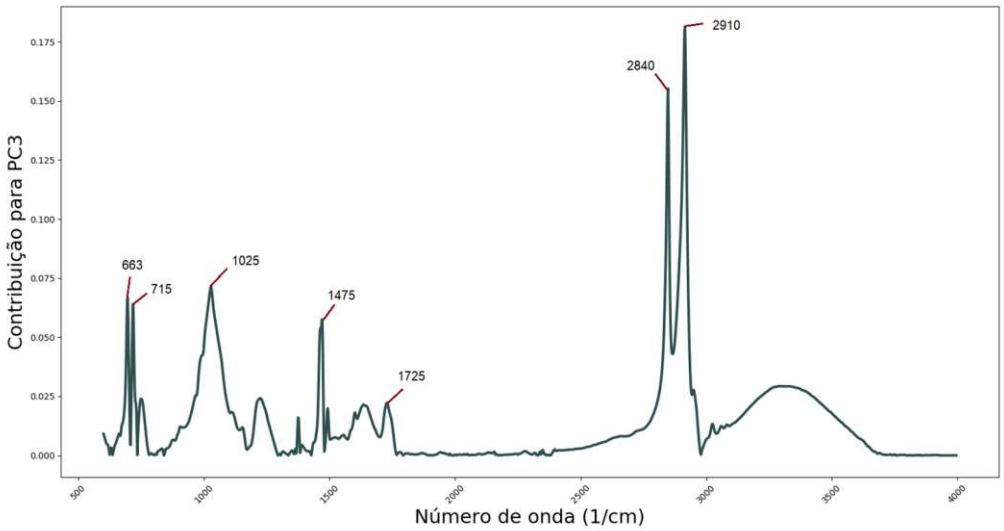
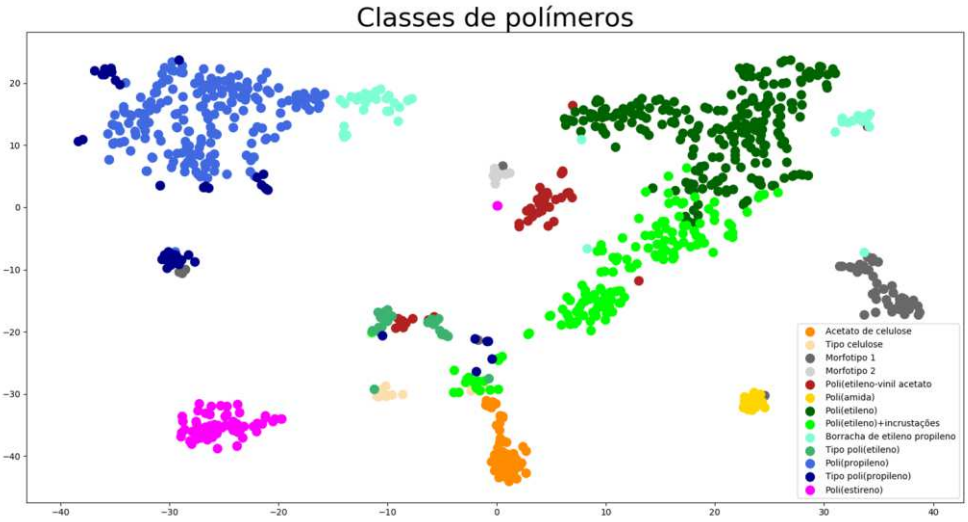


Figura 22: Fonte: Elaborado pelo autor

A extração de variáveis pela técnica t-SNE resultou em apenas 2 dimensões responsáveis pelo mapeamento e visualização das amostras na Figura 23. Tal mapeamento pode ser entendido de forma análoga aos mapeamentos do PCA, porém no caso desta técnica a otimização da função de custo busca reter a estrutura local dos dados, o que aproxima amostras semelhantes em vez de afastar amostras diferentes, como é o caso da anterior. Os eixos do gráfico t-SNE não têm significado físico, como é o caso das componentes principais.

Como era esperado, as amostras de mesma classe ficaram mais próximas no mapeamento t-SNE, formando aglomerados mais densos se comparados aos gráficos do PCA. Isso se dá possivelmente em razão da função de otimização usada nesta técnica (descrita na seção 33). Entretanto, pode-se perceber alguns *outliers*, amostras de uma classe próximas a aglomerados de classes diferentes. Por exemplo, as amostras de borracha de etileno propileno não formaram um aglomerado coerente, ao contrário ficaram dispersas próximo aos aglomerados do poli(etileno) e do poli(propileno), possivelmente devido às diferentes composições dessas borrachas. Além do mais, alguns espectros de amostras das classes “tipo” poli(etileno) e “tipo” poli(propileno) se misturaram às amostras de PEVA e poli(etileno) + incrustações, respectivamente. A evidência da similaridade entre estes espectros pode indicar possíveis fontes de erro no processo convencional de identificação dos espectros. Por exemplo, incrustações em microplásticos de poli(etileno) podem modificar o espectro a ponto de este ser confundido com o de poli(propileno) tanto por buscas automatizadas em bases de dados como por um *expert*.

Figura 23: Mapeamento das amostras pela técnica t-SNE



Fonte: Elaborado pelo autor

Pode-se concluir que o PCA e t-SNE permitem visualizações das amostras, identificar uma estrutura nos dados relacionada ao espectro de FTIR e melhor identificar espectros ambíguos. Além disso, o filtro de baixa variância é eficiente em excluir as frequências de ruído que não agregam informação e não contribuem para a classificação, mesmo sem conhecimento prévio das classes de polímeros e das frequências características dos espectros. Com isso, esperava-se observar uma redução no tempo de implementação dos classificadores, especialmente os mais computacionalmente intensivos. Para avaliar a eficiência de redução de cada técnica, foi utilizada a equação (I) apresentada na seção 49.

Os resultados estão apresentados na Tabela 2 e como esperado, as duas técnicas de extração de atributos foram muito mais eficientes na redução do número de variáveis. O PCA, por exemplo, permitiu concentrar 98% da informação em apenas 1,13% das variáveis. Já o Filtro de Baixa Variância não teve um resultado tão impressionante, mas que não deixa de ser expressivo, visto que não houve perda de informação relevante. Entretanto, vale ressaltar que há uma limitação metodológica quanto à definição do limite de variância que é arbitrária, assim diferentes autores podem adotar limites distintos e chegarem a resultados diferentes. A aproximação de um limite aceitável para uma amostra futura de composição desconhecida é um desafio a ser encarado.

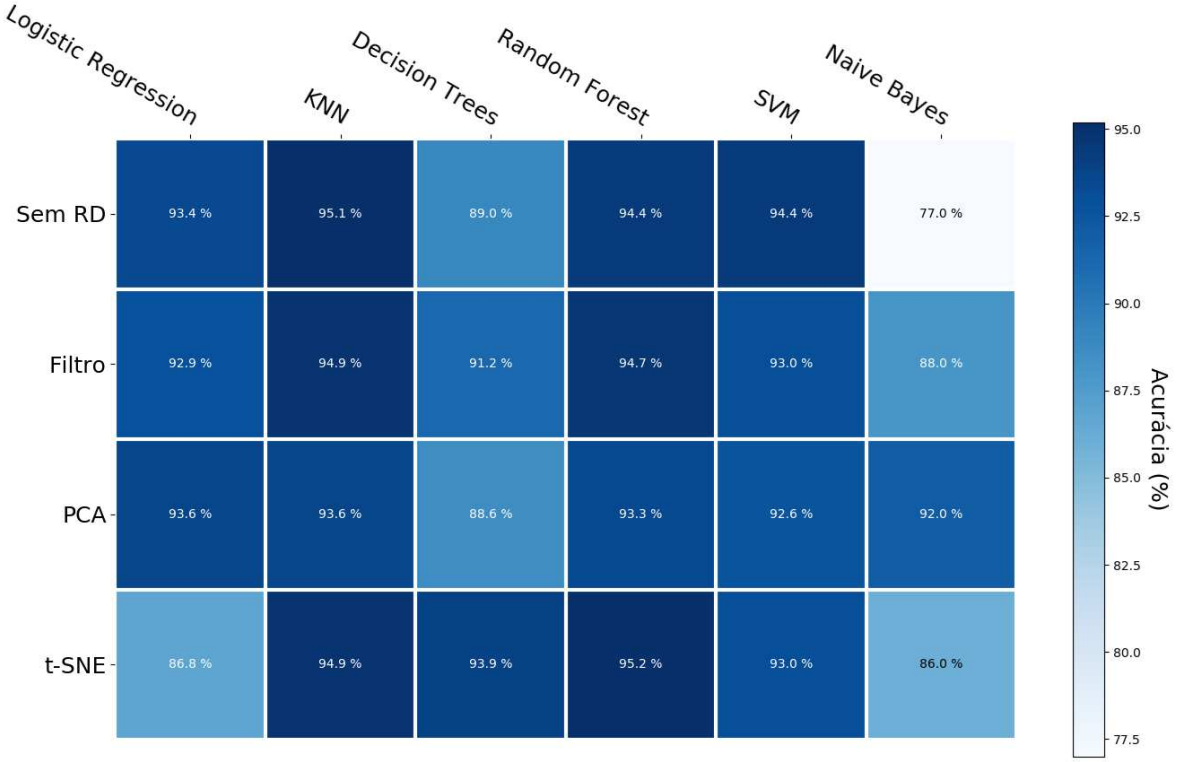
Tabela 2: Razão de redução para cada técnica de redução de dimensionalidade

Técnica de redução	Razão de redução (%)
Sem RD	0
Filtro	62,38
PCA	98,87
t-SNE	99,89

Fonte: Elaborado pelo autor

Ao fim da etapa de RD, foram obtidos 4 novos conjuntos de dados, cada um com variáveis descritivas diferentes para as mesmas amostras. Os 6 modelos de classificação apresentados na seção 35 foram então implementados para cada um desses conjuntos. Os resultados podem ser observados na Figura 24, em que cada célula do mapa de calor contém um valor de acurácia resultante da aplicação de um modelo (colunas) a um conjunto de dados após o pré-processamento (linhas). A intensidade da cor das células aumenta com a melhora do resultado segundo um gradiente explicitado pela barra à direita do gráfico.

Figura 24: Mapa de calor contendo os resultados da aplicação dos modelos de classificação



Fonte: Elaborado pelo autor

Em primeiro lugar, pode ser dito que a maioria das combinações retornou bons resultados, sendo capazes de classificar as amostras com cerca de 90% de acurácia. Resultado tal qual já havia sido apresentado por outros autores (KEDZIERSKI et al., 2019) (HUFNAGL et al., 2019) e significativamente mais acurado do que o método convencional de caracterização por busca em bibliotecas espectrais (JUNG et al., 2018). Entretanto, alguns classificadores se destacam: os modelos K-Nearest Neighbours, Random Forests e Support Vector Machine resultaram em melhores percentuais de acurácia, em geral. A combinação Random Forest/t-SNE foi a que obteve o melhor resultado. Entretanto, a vantagem sobre outras combinações é muito pequena e a decisão pelo melhor modelo não pode ser suportada estatisticamente, visto que pequenas alterações tanto no pré-processamento como na base de dados podem variar sutilmente tais resultados. Tendo em vista a "navalha de Occam", que afirma que a explicação para qualquer fenômeno deve pressupor o menor número de premissas, deve-se optar pelo modelo mais simples, neste caso o KNN, sabendo-se que este algoritmo é significativamente mais simples, inclusive em sua compreensão intuitiva.

É interessante comparar os resultados obtidos aqui com os apresentados por (KEDZIERSKI et al., 2019) que utilizou o classificador KNN neste mesmo conjunto de dados e obteve 90,5% de acurácia. Este resultado difere um pouco dos 95,1% encontrados pela mesma metodologia neste trabalho. Tal diferença pode ser devida à exclusão das 4 classes pouco representadas do conjunto de dados original. Além do mais, supõe-se que manter somente os espectros mais representativos de cada polímero com a exclusão dos espectros ambíguos poderia melhorar ainda mais a qualidade dos classificadores, definindo contornos mais precisos de regiões pertencentes a cada classe.

No gráfico da Figura 24 percebe-se que há uma tendência de variação maior entre classificadores do que entre as técnicas de RD. Na verdade, elas tiveram pouca influência no resultado dos modelos. Pode-se dizer que isso é, na verdade, esperado, visto que as informações contidas em todos os 4 conjuntos é aproximadamente a mesma. Entretanto, esperava-se perceber algum efeito da chamada "maldição da dimensionalidade" para o conjunto de dados original sem RD, o que resultaria em valores de acurácia menores para todos os modelos aplicados sobre esses dados, e especialmente ao KNN que utiliza uma medida de similaridade baseada em cálculos de distância entre as amostras (HUFNAGL et al., 2019). Como foi dito na seção 35, num espaço de muitas dimensões, amostras diferentes de uma mesma classe podem ficar esparsas e não formarem aglomerados coerentes. Entretanto, para este conjunto de dados espectroscópicos de microplásticos ambientais, é possível dizer

que a quantidade de variáveis originais não tenha sido um empecilho para a classificação das amostras em relação a quantidade de exemplos para cada classe.

Os melhores classificadores sem RD resultaram em valores de acurácia próximos de 95%, o que deixou pouca margem para melhora com a aplicação das técnicas de RD. De fato, para os classificadores *Naive Bayes* e *Decision Trees* que tiveram resultados piores com o conjunto original de dados, houve uma melhora significativa com a aplicação dessas técnicas. Da mesma forma, cada classificador pareceu responder melhor a uma técnica de redução específica. A aplicação do PCA resultou na melhora da acurácia do classificador *Naive Bayes*, mas piora de *Decision Trees*, enquanto o contrário é verdadeiro para o filtro de baixa variância e o t-SNE.

Os hiper-parâmetros para o treino dos modelos KNN/Sem RD e Random Forest/t-SNE foram os seguinte:

- KNN - n_neighbors = 3, p = 4, weights = 'distance'
- Random Forest - n_estimators = 100, criterion = 'entropy'

Vale avaliar qual foi de fato a influência das técnicas de RD no tempo de treino dos melhores modelos (Tabela 3). Deve-se salientar que o tempo de computação foi calculado para o treino do modelo final, sem considerar o tempo de importação e pré-processamento dos dados ou *GridSearch* e levando em conta a configuração do computador e as versões do software e bibliotecas (descritos na seção MATERIAIS E MÉTODOS).

Tabela 3: Tempo de implementação dos 2 melhores classificadores para cada técnica de redução de dimensionalidade

Técnica de Redução	Random Forests	KNN
Sem RD	1,39 s	0,04 s
Filtro	0,90 s	0,02 s
PCA	0,19 s	0,008 s
T-SNE	0,14 s	0,001 s

Fonte: Elaborado pelo autor

Verifica-se que utilizando os dados reduzidos por t-SNE, há uma redução de uma ordem de grandeza no tempo de computação se comparado aos dados sem redução. Portanto, apesar de pouco significativa, houve melhora de acurácia (0,8%) com redução do tempo de computação. Tal resultado está de acordo com o esperado e proposto nos objetivos deste trabalho. No entanto, esperava-se um tempo maior para implementação do modelo KNN, pois

é tido na bibliografia como computacionalmente intensivo. Apesar disso, o modelo Random Forests se mostrou mais lento em todos os casos. Apesar da melhora em termos percentuais para ambos os modelos, o tempo absoluto de implementação ainda é muito pequeno para este universo amostral.

É importante lembrar que a ferramenta *GridSearch* foi utilizada para avaliar diversas combinações de hiper-parâmetros e o resultado mostrado na figura é apenas um deles, sendo na verdade o melhor. Qualquer escolha arbitrária de hiper-parâmetros provavelmente retornaria um valor de acurácia menor do que o apresentado. Assim, destaca-se a importância da utilização desta ferramenta para uma comparação mais fidedigna da capacidade de cada modelo em classificar os dados espectroscópicos deste trabalho.

A validação cruzada, ou seja, a verificação da acurácia de um modelo em múltiplos conjuntos de teste, permitiu a generalização mais adequada dos classificadores, sendo o resultado final de acurácia uma média de múltiplos resultados para o mesmo modelo. Apesar dos cuidados para permitir uma melhor generalização a dados desconhecidos, deve-se ter cautela ao aplicar esses resultados a dados novos. É preciso lembrar que o valor de acurácia encontrado para todos os modelos só é válido para o conjunto de dados utilizado e pode ser diferente em conjuntos de dados novos.

6. Conclusão

Foi possível desenvolver uma metodologia para caracterização de microplásticos encontrados no oceano por FTIR utilizando ferramentas de aprendizado de máquina atingindo-se assim o objetivo geral deste trabalho. Ainda, os resultados foram superiores aos métodos previamente descritos na literatura. A combinação da técnica t-SNE de redução de dimensionalidade com o classificador Random Forest resultou na maior acurácia, seguida pelo classificador K-Nearest Neighbors aplicado aos dados sem RD. O tempo de implementação destes classificadores foi menor que 1 s e foi, portanto, muito mais rápido (e acurado) do que as metodologias convencionais de busca em bibliotecas e confirmação visual por um *expert*.

As técnicas KNN e Random Forests já haviam sido propostas em outros trabalhos para classificar espectros de amostras de microplástico (HUFNAGL et al., 2019; KEDZIERSKI et al., 2019). Os resultados apresentados aqui vêm corroborar tais metodologias e trazer novos *insights* sobre a estrutura dos dados espectroscópicos de amostras de microplástico. O que foi possível especialmente devido às três técnicas de RD implementadas, que permitiram identificar frequências importantes na classificação das amostras e mapear todo o universo amostral em gráficos bidimensionais. Por meio destes, foi possível identificar semelhanças entre classes e entre amostras de uma mesma classe e localizar *outliers.* Além disso, os mapeamentos podem ser interessantes para a apresentação dos resultados de classificação. Com eles, fica fácil observar as classes mais abundantes.

As técnicas de RD não melhoraram significativamente a acurácia dos classificadores, como era esperado. No entanto, cada classificador respondeu melhor a uma técnica de redução específica, o que pode ser investigado futuramente visando a melhoria do processo. Apesar disso, a maioria dos classificadores apresentaram bons resultados mesmo sem esta etapa. Além disso, mesmo sendo possível reduzir o tempo de implementação dos modelos com a aplicação da etapa de RD, o tempo já era muito pequeno e essa vantagem não é tão importante. Resultados mais significativos podem ser esperados no futuro com a aplicação desta metodologia a conjuntos de dados mais extensos, por exemplo com a adição de amostras na base de espectros de referência ou possivelmente até na aplicação em dados de técnicas como μFTIR ou μRaman que produzem conjuntos de dados muito maiores.

Já foi mencionado na literatura que uma base de dados de referência contendo amostras de microplásticos ambientais aumenta significativamente a capacidade de caracterização devido à maior semelhança entre ambos (PRIMPKE et al., 2018), uma vantagem também

deste método. Entretanto, deve-se mencionar que o resultado automático do classificador é dado dentro de um conjunto de apenas 13 classes, ou tipos, de polímeros. Sabe-se que existe uma variedade muito maior de materiais poliméricos, sendo esses apenas os encontrados nos dados disponíveis para criação da base de dados de referência. Esse fator é limitante, pois o algoritmo não consegue prever classes que ele não conhece. Apesar disso, os mapeamentos podem ser utilizados para identificar amostras de polímeros desconhecidos para o algoritmo que apareceriam como *outliers*. A colaboração entre pesquisadores da área poderá melhorar a base de dados no sentido de incluir novas classes e mais amostras das classes, o que terá efeitos na capacidade preditiva real do classificador.

A metodologia desenvolvida neste trabalho está apresentada no Apêndice A na forma de códigos Python e poderá ser aplicada por qualquer pesquisador para a identificação de microplásticos coletados no ambiente como uma maneira rápida, acurada e gratuita de caracterizar tais materiais. Mas apesar de promissora, não espera-se que essa metodologia seja definitiva para a caracterização de microplásticos. O interesse em estudos deste tipo é recente, portanto, ainda são esperados grandes avanços na área.

REFERÊNCIAS

ABRELPE. Panorama dos resíduos Sólidos no Brasil - 2017. 2017.

AGAMUTHU, P. et al. Marine debris : A review of impacts and global initiatives. Waste Management, v. 37, 2019.

ANDRADY, A. L.; NEAL, M. A. Applications and societal benefits of plastics. Philosophical Transactions of the Royal Society B: Biological Sciences, v. 364, n. 1526, p. 1977–1984, 2009.

AUTA, H. S.; EMENIKE, C. .; FAUZIAH, S. . Distribution and importance of microplastics in the marine environment: A review of the sources, fate, effects, and potential solutions. Environment International, v. 102, p. 165–176, 1 maio 2017.

BARROWS, A. P. W.; CATHEY, S. E.; PETERSEN, C. W. Marine environment microfiber contamination: Global patterns and the diversity of microparticle origins. Environmental Pollution, v. 237, p. 275–284, 2018.

BEN-DAVID, S. Understanding Machine Learning : From Theory to Algorithms. [s.l: s.n.].

BERGMANN, M.; GUTOW, L.; KLAGES, M. Marine anthropogenic litter. [s.l: s.n.].

BRANDON, J.; GOLDSTEIN, M.; OHMAN, M. D. Long-term aging and degradation of microplastic particles : Comparing in situ oceanic and experimental weathering patterns. Marine Pollution Bulletin, v. 110, n. 1, p. 299–308, 2016.

COLE, M. et al. Microplastics as contaminants in the marine environment: A review. Marine Pollution Bulletin, v. 62, n. 12, p. 2588–2597, 2011a.

COLE, M. et al. Microplastics as contaminants in the marine environment: A review. Marine Pollution Bulletin, v. 62, n. 12, p. 2588–2597, 1 dez. 2011b.

CORTES, C.; VAPNIK, V. Support-Vector Networks. v. 297, p. 273–297, 1995.

DIAS, P. A. Automated identification and quantification of microfibres and microplastics. Analytical Methods, p. 2138–2147, 2019.

ERIKSEN, M. et al. Plastic Pollution in the World's Oceans: More than 5 Trillion Plastic Pieces Weighing over 250,000 Tons Afloat at Sea. PLoS ONE, v. 9, n. 12, p. 1–15, 2014.

FREE, C. M. et al. High-levels of microplastic pollution in a large , remote , mountain lake. Marine Pollution Bulletin, v. 85, n. 1, p. 156–163, 2014.

GESAMP. Guidelines for the monitoring and assessment of plastic litter in the ocean. 2019.

GEYER, R.; JAMBECK, J. R.; LAW, K. L. Production, use, and fate of all plastics ever made. Science Advances, v. 3, n. 7, 2017.

GOUIN, T. et al. A Thermodynamic Approach for Assessing the Environmental Exposure of Chemicals Absorbed to Microplastic. Environmental Science & Technology, p. 1466–1472, 2011.

GREGORY, M. R. Environmental implications of plastic debris in marine settings-entanglement, ingestion, smothering, hangers-on, hitch-hiking and alien invasions. Philosophical Transactions of the Royal Society B: Biological Sciences, v. 364, n. 1526, p. 2013–2025, 2009.

HALSTEAD, J. E. et al. Assessment tools for microplastics and natural fibres ingested by fish in an urbanised estuary. Environmental Pollution, v. 234, p. 552–561, 2018.

HAND, D. J. et al. Idiot ' s Bayes-Not So Stupid After All ? v. 69, n. 3, p. 385–398, 2019.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. [s.l: s.n.].

HIDALGO-RUZ, V. et al. Microplastics in the Marine Environment: A Review of the Methods Used for Identification and Quantification. Environmental Science Technology, v. 46, p. 3060–3075, 2012a.

HIDALGO-RUZ, V. et al. Microplastics in the marine environment: A review of the methods used for identification and quantification. Environmental Science and Technology, v. 46, n. 6, p. 3060–3075, 20 mar. 2012b.

HUFNAGL, B. et al. A methodology for the fast identification and monitoring of microplastics in environmental samples using random decision forest classifiers. Analytical Methods, v. 11, n. 17, p. 2277–2285, 2019.

JAMBECK, J. R. et al. Plastic waste inputs from land into the ocean. Science, v. 347, n. 6223, p. 768–771, 13 fev. 2015.

JUNG, M. R. et al. Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms. Marine Pollution Bulletin, v. 127, n. December 2017, p. 704–716, 2018.

KÄPPLER, A. et al. Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both? Analytical and Bioanalytical Chemistry, v. 408, n. 29, p. 8377–8391, 8 nov. 2016.

KEDZIERSKI, M. et al. A machine learning algorithm for high throughput identification of FTIR spectra. Chemosphere, 2019.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. 2007.

KYOUNG, Y. et al. A comparison of microscopic and spectroscopic identification methods for analysis of microplastics in environmental samples. Marine Pollution Bulletin, v. 93, n. 1–2, p. 202–209, 2015.

LEBRETON, L. C. M. et al. River plastic emissions to the world ' s oceans. Nature Communications, v. 8, p. 1–10, 2017.

LENZ, R. et al. A critical assessment of visual identi fi cation of marine microplastic using Raman spectroscopy for analysis improvement. Marine Pollution Bulletin, v. 100, n. 1, p. 82–91, 2015.

MOORE, C. J. Synthetic polymers in the marine environment: A rapidly increasing, long-term threat. Environmental Research, v. 108, n. 2, p. 131–139, 2008.

OCEAN CONSERVANCY; INTERNATIONAL COASTAL CLEANUP. Building a clean swell - 2018 Report. [s.l: s.n.].

PEIXOTO, D. et al. Microplastic pollution in commercial salt for human consumption : A review. Estuarine, Coastal and Shelf Science, v. 219, n. January 2018, p. 161–168, 2019.

PLASTICS EUROPE. Plastics – the Facts. [s.l: s.n.].

POOLE, D.; MACKWORTH, A.; GOEBEL, R. Computational Intelligence and Knowledge - Chapter 1. In: Computational Intelligence: A Logical Approach. [s.l: s.n.]. p. 1–22.

PRIMPKE, S. et al. Reference database design for the automated analysis of microplastic samples based on Fourier transform infrared (FTIR) spectroscopy. Analytical and Bioanalytical Chemistry, p. 5131–5141, 2018.

SAITO, H. et al. Human footprint in the abyss : 30 year records of deep-sea plastic debris Sanae Chiba. Marine Policy, v. 96, n. April, p. 204–212, 2018.

SHAH, A. A. et al. Biological degradation of plastics : A comprehensive review. Biotechnology Advances, v. 26, p. 246–265, 2008.

SHARMA, S.; CHATTERJEE, S. Microplastic pollution, a threat to marine ecosystem and human health: a short review. Environmental Science and Pollution Research, v. 24, n. 27, p. 21530–21547, 2017.

SHERMAN, P. et al. A global inventory of small floating plastic debris. Environmental Research, 2015.

THOMPSON, R. C. et al. Lost at Sea: Where Is All the Plastic? Science, v. 304, n. 5672, p. 838, 2004a.

THOMPSON, R. C. et al. Lost at Sea: Where Is All the Plastic? Science, v. 304, n. 5672, p. 838, 7 maio 2004b.

TRUNK, G. V. A Problem of Dimensionality: A Simple Example given. IEEE Transactions on Pattern Analysis and Machine Intelligence, n. 3, p. 306–307, 1979.

VAN DER MAATEN, L.; HINTON, G. Visualizing Data using t-SNE. New Scientist, v. 164, n. 2210, p. 10, 2008.

VEERASINGAM, S. et al. Characteristics, seasonal distribution and surface degradation features of microplastic pellets along the Goa coast, India. Chemosphere, v. 159, p. 496–505, 2016.

WANG, J. et al. Microplastics as contaminants in the soil environment : A mini-review. Science of the Total Environment, v. 691, p. 848–857, 2019.

Apêndice A

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Importing the dataset
dataset = pd.read_csv('D4_4_publication.csv', header = None)
dataset = dataset.drop(dataset.index[[181, 182, 183]])
X = dataset.iloc[2:961, 2:1764].values #Excluded least represented classes
y = dataset.iloc[2:961, 1].values

# Encoding the Dependent Variable
from sklearn.preprocessing import LabelEncoder
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)

#Smoothing
from scipy.signal import savgol_filter
X = savgol_filter(X, 11, 10)

# Normalization
from sklearn.preprocessing import Normalizer
n = Normalizer()
X = n.fit_transform(X)

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)

# Fitting K-NN to the Training set
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 3, p = 4, weights = 'distance')
classifier.fit(X_train, y_train)

# Time it
import timeit

code_to_test = """
# Fitting K-NN to the Training set
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 3, p = 4, weights = 'distance')
classifier.fit(X_train, y_train)
"""
setup_code = """
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```
# Importing the dataset
dataset = pd.read_csv('D4_4_publication.csv', header = None)
dataset = dataset.drop(dataset.index[[181, 182, 183]])
X = dataset.iloc[2:961, 2:1764].values #Excluded least represented classes
y = dataset.iloc[2:961, 1].values

# Encoding the Dependent Variable
from sklearn.preprocessing import LabelEncoder
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)

#Smoothing
from scipy.signal import savgol_filter
X = savgol_filter(X, 11, 10)

# Normalization
from sklearn.preprocessing import Normalizer
n = Normalizer()
X = n.fit_transform(X)

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
"""
elapsed_time = timeit.timeit(stmt=code_to_test, setup=setup_code, number=10)/10
print(elapsed_time)

# Predicting the Test set results
y_pred = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(classification_report(y_test, y_pred))

# Applying Grid Search to find the best model and the best parameters
from sklearn.model_selection import GridSearchCV
parameters = [{'n_neighbors': [1, 3, 5, 7, 9, 11, 13], 'weights': ['uniform', 'distance'], 'p': [1, 2,
3, 4]}]
grid_search = GridSearchCV(estimator = classifier,
                           param_grid = parameters,
                           scoring = 'accuracy',
                           cv = 10,
                           n_jobs = -1)
grid_search = grid_search.fit(X_train, y_train)
best_accuracy = grid_search.best_score_
best_parameters = grid_search.best_params_
```