

Universidade Federal de Mato Grosso do Sul
Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia
Dissertação de Mestrado do Curso de Engenharia Elétrica

João Marcos Soares Anjos

Sistema de Apoio à Decisão Baseado em Inteligência Artificial para Gerenciamento de Resíduos Sólidos

Campo Grande - MS

14 de Janeiro de 2023

Universidade Federal de Mato Grosso do Sul
Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia
Dissertação de Mestrado do Curso de Engenharia Elétrica

João Marcos Soares Anjos

Sistema de Apoio à Decisão Baseado em Inteligência Artificial para Gerenciamento de Resíduos Sólidos

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFMS para obtenção do Grau de Mestre em Engenharia Elétrica na área de Sistemas de Decisão Baseados em Computação Flexível.

Orientador: Professor Doutor Marcio Luiz Magri Kimpara

Coorientador: Professor Doutor João Onofre Pereira Pinto

Campo Grande - MS

14 de Janeiro de 2023

Dedico este trabalho aos meus pais, amigos e professores, por acreditarem em mim e me inspirarem a buscar sempre o melhor, e acima de tudo agradeço à Deus, por me conceder as bênçãos necessárias para chegar até aqui.

Agradecimentos

O presente trabalho foi realizado com apoio financeiro do Projeto Resíduos Sólidos: Disposição Legal como parte do Convênio de Cooperação Técnica e Científica celebrado entre o Ministério Público de Mato Grosso do Sul (MPMS) e a Universidade Estadual do Mato Grosso do Sul (UEMS) especificado pelo Termo de Convênio n. 1076/2020-UEMS/MPMS.

Agradeço especialmente aos alunos que participaram ativamente deste estudo, dedicando tempo e esforço para coletar dados e contribuir para as discussões. A colaboração de vocês foi essencial para o sucesso desta pesquisa. Agradeço pela dedicação e empenho em transformar a teoria em prática e fornecer informações valiosas para a comunidade acadêmica e para o projeto.

Também quero agradecer aos meus familiares e amigos que me apoiaram durante todo o decorrer da pesquisa. Seus encorajamentos, motivações e compreensão foram de grande ajuda para que eu pudesse superar os momentos mais desafiadores e manter minha energia e inspiração. Agradeço profundamente pelo apoio constante que me foi dado. Não teria sido possível chegar até aqui sem o amor e o suporte que me proporcionaram durante esta jornada de aprendizado.

Por fim, expresso minha gratidão a todos os profissionais que contribuíram para esta pesquisa. Sua participação foi fundamental para o sucesso deste estudo e para o avanço do conhecimento.

Resumo

Este trabalho descreve o desenvolvimento de um sistema de apoio à decisão para auxiliar na gestão de resíduos sólidos urbanos. O sistema foi criado com base em algoritmos de inteligência artificial/aprendizado de máquina, o que o torna capaz de realizar previsões precisas sobre a quantidade e a caracterização dos resíduos gerados, de acordo com a Política Nacional de Resíduos Sólidos (PNRS).

Para a criação deste sistema, foi realizada uma pesquisa de campo na cidade de Campo Grande, Mato Grosso do Sul, onde foram coletados dados referentes à geração de resíduos sólidos em 158 domicílios. Os dados coletados foram usados para construir modelos de predição por meio de técnicas de regressão e classificação.

Os modelos desenvolvidos foram utilizados para estimar a quantidade de resíduos sólidos gerados por domicílio, e sua caracterização de acordo com as diretrizes da PNRS. O uso desses modelos permitiu que o sistema de apoio à decisão pudesse fornecer informações importantes e precisas sobre a gestão de resíduos sólidos.

O sistema de apoio à decisão também foi projetado para ser escalável e adaptável, permitindo sua implementação em outras cidades e regiões. A implementação bem-sucedida deste sistema de apoio à decisão pode ajudar a melhorar a gestão de resíduos sólidos urbanos em todo o país, contribuindo para a preservação do meio ambiente e a promoção de uma vida mais sustentável.

Palavras-chaves: inteligência artificial, aprendizado de máquina, sistema de apoio à decisão, gerenciamento de resíduos sólidos, caracterização de resíduos.

Abstract

This work describes the development of a decision support system to assist in the management of urban solid waste. The system was created based on artificial intelligence/machine learning algorithms, which makes it capable of making accurate predictions about the quantity and characterization of waste generated, according to the National Solid Waste Policy (PNRS).

To create this system, a field survey was conducted in the city of Campo Grande, Mato Grosso do Sul, where data on solid waste generation in 158 households were collected. The collected data was used to build prediction models through regression and classification techniques.

The developed models were used to estimate the quantity of solid waste generated per household and their characterization according to the PNRS guidelines. The use of these models allowed the decision support system to provide important and accurate information on solid waste management.

The decision support system was also designed to be scalable and adaptable, allowing its implementation in other cities and regions. The successful implementation of this decision support system can help improve the management of urban solid waste throughout the country, contributing to environmental preservation and promoting a more sustainable life.

Keywords: artificial intelligence, machine learning, decision support system, solid waste management, waste characterization.

Lista de ilustrações

Figura 1	–	Estrutura de uma rede neural.	22
Figura 2	–	Representação de uma árvore de decisão	24
Figura 3	–	Hiperplanos dividindo um conjunto de dados	27
Figura 4	–	Suavização da margem de um SVM	28
Figura 5	–	Panfleto de divulgação e orientação	37
Figura 6	–	Cronograma para o primeiro levantamento de dados.	39
Figura 7	–	Cronograma para o segundo levantamento de dados.	39
Figura 8	–	Processo de coleta dos resíduos.	40
Figura 9	–	Mapa com rotas entre domicílios	41
Figura 10	–	Disposição das amostras de resíduos sólidos	41
Figura 11	–	Fluxograma da pesagem e classificação dos resíduos	42
Figura 12	–	Planilha de registro da classificação	43
Figura 13	–	Equipe realizando a classificação dos resíduos	43
Figura 14	–	Correlação entre os dados	44
Figura 15	–	Diagrama de caixa para os plásticos	46
Figura 16	–	Divisão do diagrama de caixa	46
Figura 17	–	Matriz de confusão	48
Figura 18	–	Arquitetura de um sistema de apoio à decisão.	50
Figura 19	–	Rede neural densa.	52
Figura 20	–	Resultado do treinamento da rede	52
Figura 21	–	Saída da rede para os dados de teste	53
Figura 22	–	Saída da rede para uma quantidade menor de épocas	53
Figura 23	–	Histograma com intervalos iguais	55
Figura 24	–	Histograma combinado com a FDP	56
Figura 25	–	Árvore de decisão modelada	58
Figura 26	–	Matriz de confusão para modelo baseado em árvore de decisão . . .	58
Figura 27	–	Matriz de confusão para modelo baseado em naive bayes	59
Figura 28	–	Matriz de confusão para modelo baseado no knn	59
Figura 29	–	Matriz de confusão para modelo baseado no svm	60
Figura 30	–	Tela inicial do sistema de apoio à decisão	63
Figura 31	–	Página de predição do sistema	64
Figura 32	–	Página de visualização do sistema	65
Figura 33	–	Página de visualização de projeções	65
Figura 34	–	Proporção de resíduos sólidos domiciliares	66
Figura 35	–	Composição dos resíduos sólidos urbanos	67

Figura 36 –	Proporção estimada de embalagens	68
-------------	--	----

Lista de tabelas

Tabela 1	–	Quantidade de amostras por classe	57
Tabela 2	–	Margem de erro para os tipos de resíduos	60
Tabela 3	–	Pecisão dos modelos de classificação	62
Tabela 4	–	Pecisão dos modelos nos piores casos	62
Tabela 5	–	Valores de F1 Score	62
Tabela 6	–	Estimação da geração de resíduos	66
Tabela 7	–	Média da geração de resíduos por classe	69

Abreviações

ABRELPE	Associação Brasileira das Empresas de Limpeza Pública
AG	algoritmo genético
ANFIS	sistemas de inferência neurofuzzy adaptativo
DWT	transformada wavelet discreta
FDP	função densidade de probabilidade
IA	inteligência artificial
MAE	erro médio absoluto
MARE	erro relativo médio absoluto
ML	aprendizagem de máquina
MLP	perceptron multicamadas
PERS	Plano Estadual de Resíduos Sólidos
PNRS	Política Nacional de Resíduos Sólidos
RMSE	raiz quadrada média do erro
RNA	rede neural artificial
RSD	resíduos sólidos domiciliares
RSU	resíduos sólidos urbanos
SAD	sistema de apoio à decisão
SGBD	sistema de gerenciamento de banco de dados
SSE	<i>sum of square errors</i>
SVM	máquina de vetores de suporte

Sumário

Abreviações	9
1 Introdução	12
1.1 Objetivos	13
1.1.1 Objetivo geral	13
1.1.2 Objetivos Específicos	13
1.2 Organização do trabalho	13
2 Revisão Bibliográfica	15
2.1 Modelos de estimativa	15
2.2 Sistemas de apoio à decisão	17
3 Fundamentação Teórica	20
3.1 Inteligência artificial	20
3.1.1 Redes Neurais Artificiais	21
3.1.2 Árvores de decisão	24
3.1.3 Naive bayes	26
3.1.4 Support Vector Machine (SVM)	27
3.1.5 K-Nearest Neighbors (KNN)	30
3.2 Sistemas de apoio à decisão	31
4 Metodologia	34
4.1 Pesquisa de campo	34
4.1.1 Definição do escopo	35
4.1.2 Seleção dos domicílios	37
4.1.3 Treinamento técnico	38
4.1.4 Coleta dos dados	39
4.1.5 Classificação e pesagem	41
4.2 Análise dos dados	43
4.2.1 Preparação dos dados	45
4.3 Modelagem	47
4.4 Sistema de apoio à decisão	49
5 Resultados	51
5.1 Modelos de regressão	51
5.2 Modelos de classificação	54
5.3 Escolha do modelo	61

5.4 Sistema de apoio à decisão	63
6 Conclusão	70
Referências	72
 Apêndices	 76
APÊNDICE A Códigos	77
A.1 Preparação dos dados	77
A.2 Modelo de regressão	81
A.3 Modelos de classificação	84
A.4 Sistema de apoio à decisão	85

1 Introdução

A geração de resíduos sólidos urbanos têm sido um dos principais problemas ambientais que atinge toda a população mundial. Na Rio + 20, a gestão de resíduos sólidos foi um dos temas incluídos entre as nove dimensões em que houve foco. Para contornar esse problema, diversas soluções vêm sendo discutidas como a logística reversa, e outras metodologias para gestão e operação na área de resíduos sólidos.

Segundo o Panorama dos Resíduos Sólidos ([ABRELPE, 2019](#)), produzido pela [Associação Brasileira das Empresas de Limpeza Pública \(ABRELPE\)](#), em 2018 foram gerados no Brasil 79 milhões de toneladas de resíduos sólidos. Esses resíduos podem ser classificados como: resíduos secos e resíduos úmidos.

Em Mato Grosso do Sul o [Plano Estadual de Resíduos Sólidos \(PERS\)](#) foi construído para estabelecer políticas de gestão e gerenciamento dos resíduos, bem como para promover ações de inclusão social e emancipação econômica dos catadores de materiais recicláveis. Além disso a logística reversa está sendo implantada de forma a compartilhar a responsabilidade com o setor empresarial de reintroduzir os resíduos e embalagens na cadeia produtiva, ou garantir a destinação final ambientalmente adequada.

Conforme [SINIR \(2020\)](#), em relação à responsabilidade compartilhada, compete ao cidadão descartar os resíduos nos locais estabelecidos e ao setor privado o gerenciamento adequado dos resíduos sólidos e a adoção de inovações que tragam benefícios socioambientais. Por outro lado, cabe ao Poder Público a fiscalização do processo, e a conscientização do cidadão.

Segundo [Khan, Kumar e Samadder \(2016\)](#), a quantificação da taxa de geração de resíduos e a caracterização de sua composição são essenciais para planejar e projetar sistemas eficazes de gestão de resíduos sólidos de qualquer região, assim como para estimar seus impactos ambientais, selecionar tecnologias de tratamento, planejar infraestruturas e formular políticas. Outra finalidade desta quantificação é dimensionar o volume de aterro necessário para realizar a disposição adequada dos resíduos.

De acordo com [Melaré et al. \(2017\)](#) a gestão de resíduos é complexa e engloba um conjunto de ações realizadas direta ou indiretamente nas etapas de planejamento, coleta, transporte, transbordo e processamento dos mesmos. Sendo assim, é de fundamental importância que as pessoas responsáveis pelas tomadas de decisões sobre os aspectos gerenciais em resíduos sólidos estejam bem informadas com dados representativos do cenário em foco.

Os parâmetros mais importantes que afetam a geração de resíduos sólidos, de

acordo com [Ali \(2018\)](#) são a densidade populacional, consumo médio de materiais manufaturados, produção industrial e fatores socioeconômicos (renda per capita, entre outros). Aliado a esses parâmetros, o crescimento populacional associado à migração da população para áreas urbanas e ao desenvolvimento industrial tem levado a uma relação de consumo que resulta em problemas ambientais, sociais e econômicos.

Atualmente, ainda faltam dados nacionais confiáveis sobre a geração e composição de resíduos sólidos. Assim, o propósito desse trabalho é realizar um levantamento de dados sobre a caracterização dos resíduos sólidos domiciliares em uma amostra de cidades do Estado de Mato Grosso do Sul e desenvolver um sistema computacional de apoio à tomada de decisão, baseado em inteligência artificial, que estime a quantidade de resíduos sólidos por tipo de resíduo, conforme previstos na [Política Nacional de Resíduos Sólidos \(PNRS\)](#), para qualquer município.

1.1 Objetivos

1.1.1 Objetivo geral

Desenvolver um sistema de apoio à decisão utilizando técnicas de inteligência artificial/aprendizado de máquina para estimação da quantidade de resíduos sólidos, total e por tipo de resíduos, por município, para o apoio ao gerenciamento de resíduos sólidos urbanos.

1.1.2 Objetivos Específicos

- Realizar uma pesquisa de campo com a finalidade de levantar dados sobre a geração e caracterização de resíduos sólidos, bem como dados socioeconômicos de uma amostra de municípios do Estado de Mato Grosso do Sul.
- Avaliar diferentes algoritmos de aprendizado de máquina para identificar modelos capazes de realizar predição e generalizar o conhecimento adquirido para diferentes cenários.
- Desenvolver um sistema de apoio à decisão integrado aos modelos de predição para auxiliar tomadores de decisão da área de resíduos sólidos e stakeholders a se posicionarem frente a determinadas perspectivas.

1.2 Organização do trabalho

Os capítulos seguintes estão organizados da seguinte forma:

- **Capítulo 2:** é apresentada uma revisão de literatura relacionada a área de resíduos sólidos e técnicas de inteligência artificial dirigidas ao desenvolvimento de modelos de predição.
- **Capítulo 3:** refere-se à fundamentação teórica utilizada na construção desse estudo e é dividida em: Inteligência Artificial, onde são apresentados os algoritmos principais utilizados, e Sistemas de Apoio à Decisão, detalhando seu funcionamento e arquitetura.
- **Capítulo 4:** são apresentadas as etapas elaboradas para o desenvolvimento da metodologia para a solução dos problemas de pesquisa abordados.
- **Capítulo 5:** são apresentados os resultados dos modelos e do sistema de apoio à decisão construídos.
- **Capítulo 6:** relata as contribuições para a área e as propostas de trabalhos futuros.

2 Revisão Bibliográfica

2.1 Modelos de estimativa

Em [Melaré et al. \(2017\)](#) é apresentado uma revisão dos trabalhos na área de gerenciamento de resíduos sólidos, avaliando as soluções apresentadas e realizando uma discussão sobre as tecnologias e métodos envolvidos. Dentre os algoritmos mais utilizados estão: regressão linear, [rede neural artificial \(RNA\)](#), [máquina de vetores de suporte \(SVM\)](#), [algoritmo genético \(AG\)](#), lógica difusa, análise estatística, mineração de dados e outros algoritmos de [aprendizagem de máquina \(ML\)](#).

Em geral, as técnicas de [inteligência artificial \(IA\)](#) dependem de dados representativos do problema para que o modelo desenvolvido seja confiável, por esta razão em muitas áreas a IA ainda não tem sido amplamente utilizada.

Segundo [Abdallah et al. \(2020\)](#) os modelos de IA são orientados principalmente por extensos conjuntos de dados para fins de treinamento e calibração. As pesquisas atuais são frequentemente prejudicadas pela falta, incompletude e/ou imprecisão dos dados sobre resíduos. Isso se deve parcialmente ao fato de as indústrias de gerenciamento de resíduos sólidos estarem desatualizadas, com registros confiáveis limitados e dados sensoriais escassos, especialmente em países em desenvolvimento.

Uma rede neural foi utilizada em [Ma et al. \(2020\)](#) para estabelecer modelos quantitativos relacionados a fatores socioeconômicos e estimar a composição física dos resíduos sólidos municipais na China. A análise da correlação entre os fatores socioeconômicos e a composição dos resíduos sólidos indicam que a localização geográfica e a classe social são fatores determinantes. A regressão desenvolvida utilizou dados provenientes da literatura científica e obteve resultados com alta fidelidade, avaliados por indicadores estatísticos como o coeficiente de determinação R^2 e o erro quadrático médio.

A caracterização de resíduos sólidos foi estudada em [Bernache-Pérez et al. \(2001\)](#). Os objetivos do trabalho eram estimar a taxa de geração diária de [resíduos sólidos urbanos \(RSU\)](#) e de [resíduos sólidos domiciliares \(RSD\)](#), caracterizar e comparar sua composição por tipo de material e determinar a proporção que o RSD contribui para o RSU. Para isso, foi realizada a amostragem direta de resíduos sólidos por um processo de triagem, classificação e pesagem. Uma amostra de 300 domicílios de Guadalajara foi obtida por meio de uma amostra estratificada, em dois estágios. Cada domicílio selecionado foi visitado cinco vezes e um questionário foi utilizado para obter informações socioeconômicas básicas. O estudo constatou que os RSD representaram 55,9% dos RSU, e a principal diferença com relação à composição dos resíduos foi a proporção de materiais orgânicos,

53% à nível domiciliar e apenas 16,5% à nível municipal.

Em [Dangi et al. \(2008\)](#) foi demonstrada uma abordagem estatística e eficiente para caracterizar RSU em nível domiciliar. O estudo utilizou amostragem estratificada em seis setores da cidade de Kathmandu para medir RSU por 2 semanas. A caracterização de resíduos constatou que cerca de dois terços dos resíduos são dominados por resíduos de cozinha e vegetais, seguidos por entulhos de construção.

Os estudos desenvolvidos em [Bernache-Pérez et al. \(2001\)](#) e [Dangi et al. \(2008\)](#) se mostraram eficazes no processo de amostragem, coleta e caracterização dos resíduos sólidos urbanos. De posse dos dados coletados, e com o propósito de enriquecer os estudos, poderiam ser utilizadas outras técnicas, como redes neurais ou regressão linear múltipla, para realizar a estimação e caracterização dos resíduos sólidos.

O estudo desenvolvido em [Ali \(2018\)](#) realiza uma estimativa dos resíduos sólidos urbanos a serem produzidos durante os próximos 20 anos para a cidade de Bagdá e utiliza redes neurais artificiais para prever o volume do aterro com base na população e na quantidade de resíduos sólidos gerados. O modelo de RNA desenvolvido para a predição do volume de aterro obteve um coeficiente de determinação R^2 igual a 0,996 com erro quadrático médio de $3,6 \times 10^{-5}$. Os resultados mostraram que as redes neurais artificiais podem realizar a predição de forma adequada, enquanto alguns modelos de regressão linear não são suficientemente precisos.

Em [Araiza-Aguilar, Rojas-Valencia e Aguilar-Vera \(2020\)](#) foi desenvolvido um modelo de previsão para determinar a taxa de geração de resíduos sólidos urbanos no México. Foi utilizada regressão linear múltipla com variáveis explicativas sociais e demográficas correspondentes aos anos 2010 a 2015. As variáveis mais importantes para predizer a taxa de geração de RSU na área de estudo foram a população de cada município, a população proveniente de outros municípios e a densidade populacional. Apesar da regressão ter gerado um modelo de previsão com variáveis explicativas sociais e demográficas, algumas das variáveis coletadas foram eliminadas através do teste VIF, que mede a multicolinearidade entre variáveis preditoras. Isso implica em perda de informações coletadas no estudo, que poderiam influenciar de alguma forma o resultado se fosse utilizada alguma técnica ou transformação nos dados.

O trabalho desenvolvido em [Khan e Burney \(1989\)](#) utilizou dados socioeconômicos e a composição de resíduos sólidos de 28 cidades para desenvolver modelos de regressão usando regressão linear múltipla. Os modelos desenvolvidos foram aplicados à previsão de quatro tipos de resíduos: papel, metal, matéria orgânica e vidro. Um dos resultados encontrados diz respeito ao clima e a renda afetarem os hábitos de vida e alimentação das pessoas, o que consequentemente afeta a geração dos resíduos. Outro resultado revela que o percentual de papel nos resíduos está diretamente relacionado e é afetado pelas faixas de renda. Além disso, a geração de vidro e matéria orgânica é mais afetada pela taxa de

ocupação domiciliar.

Uma pesquisa bibliográfica foi realizada em Yetilmezsoy, Ozkaya e Cakmakci (2011) para avaliar as aplicações recentes de estudos de modelagem baseados em inteligência artificial no campo da engenharia ambiental. A revisão indicou que, entre os vários tipos de modelos avaliados, as redes neurais de três camadas foram consideradas como os tipos de rede mais simples e amplamente utilizados.

No estudo realizado em Jalali e Nouri (2008), foi utilizado uma rede neural feed-forward, *perceptron multicamadas* (MLP), para prever a quantidade, em toneladas, de resíduos gerados em Mashhad, no Irã. Usando a RNA com uma camada oculta e alterando o número de neurônios da camada, diferentes modelos foram criados e testados. Então, de acordo com os índices de *erro médio absoluto* (MAE), *erro relativo médio absoluto* (MARE), *raiz quadrada média do erro* (RMSE), e R^2 , estruturas com 10 e 16 neurônios na camada oculta foram selecionadas como os modelos adequados e com base no índice TS (que representa a distribuição de erros de predição), foi escolhida a estrutura com 16 neurônios na camada oculta para a previsão da geração de resíduos.

Em Soni et al. (2019) foram comparados diferentes modelos de inteligência artificial, como a RNA, *sistemas de inferência neurofuzzy adaptativo* (ANFIS), *transformada wavelet discreta* (DWT), algoritmos genéticos, entre outros, para examinar e avaliar sua capacidade de prever a quantidade de geração de resíduos sólidos urbano. O modelo híbrido de algoritmo genético e rede neural artificial apresentou menores erros e foi considerado o mais preciso dos modelos investigados.

A caracterização e quantificação de resíduos sólidos municipais foi abordada em Miezah et al. (2015). Foi realizado uma pesquisa domiciliar em Gana para obter dados sobre a taxa de geração e composição física dos resíduos, a eficiência da triagem e separação, bem como a geração per capita. Foi utilizada a regressão linear para verificar a relação entre a taxa de geração de resíduos e a renda domiciliar, e também a relação entre a geração de resíduos e o número de moradores. De acordo com o estudo, áreas com classes socioeconômicas mais altas geraram a maior quantidade de resíduos, seguidas pelas de classe média e baixa, respectivamente. Assim como na grande maioria de outros estudos a fração orgânica nos resíduos foi a mais alta e variou de 48% a 69%.

2.2 Sistemas de apoio à decisão

Sistemas de apoio à decisão são sistemas baseados em conhecimento amplamente utilizados em organizações no processamento de informações, controle de processos e sistemas colaborativos. Devido a isso, existem muitos trabalhos e estudos sobre seu desenvolvimento e aplicação nas mais diversas áreas de pesquisa.

Em [Zeng et al. \(2012\)](#) um sistema de apoio à decisão (SAD) baseado na web foi desenvolvido para apoiar a gestão dos problemas de recursos hídricos. Foram utilizadas aplicações de redes neurais para a previsão da demanda de água e algoritmos genéticos para a alocação de recursos hídricos, e esses modelos foram incorporados ao SAD. O SAD desenvolvido apresentava interface simples e acessível, facilitando a visualização dos dados para os usuários.

No trabalho desenvolvido em [Perraju \(2013\)](#) é apresentado a organização de um sistema de apoio a decisão clássico, suas características e aplicações e é definido um sistema de apoio à decisão baseado em conhecimento, e suas diferenças em relação ao SAD clássico - sendo elas principalmente a utilização de técnicas de inteligência artificial.

A gestão de resíduos sólidos é um desafio complexo que envolve muitos aspectos técnicos, ambientais, sociais e econômicos. Para lidar com essas complexidades, os sistemas de apoio à decisão são cada vez mais utilizados como ferramentas para ajudar os gestores a tomar decisões eficazes.

O estudo desenvolvido em [Jiao et al. \(2021\)](#) aplicou a técnica KNN para prever a quantidade de resíduos sólidos gerados em uma determinada região. No estudo, a técnica foi aplicada para prever a quantidade de resíduos gerados em uma região com base em dados históricos e dados demográficos. O modelo de previsão desenvolvido pode ajudar os gestores a tomar decisões informadas sobre o gerenciamento de resíduos, como determinar a capacidade necessária para aterros sanitários e a frequência de coleta de lixo em determinadas regiões.

Uma estrutura de apoio à decisão foi projetada em [Sharma et al. \(2019\)](#) com o objetivo de ajudar a aumentar a eficiência do processo de tomada de decisão para a gestão sustentável de resíduos sólidos. A estrutura desenvolvida é modular, com um módulo para a coleta de resíduos, outro para transporte e outro para classificação e destinação de resíduos por meio de uma abordagem integrada.

Em um outro estudo, conduzido por [Lee, Lee e Park \(2020\)](#), a técnica SVM foi empregada para estimar a quantidade de resíduos de construção e demolição gerados em uma cidade coreana. Nessa pesquisa, a técnica foi utilizada para realizar previsões sobre a produção de resíduos, permitindo aos gestores desenvolver ações de reciclagem e reutilização. O modelo se provou um valioso instrumento para aprimorar o planejamento e o controle da gestão de resíduos sólidos.

No estudo realizado por [Amoako, Yeo e Zhang \(2020\)](#), foi demonstrado que a técnica de Árvore de Decisão pode ser aplicada com sucesso para prever a quantidade de resíduos sólidos urbanos gerados em Gana. Com base em variáveis como densidade populacional, renda per capita e atividades econômicas, o modelo desenvolvido alcançou uma precisão de 78,5%. Esses resultados podem auxiliar estrategicamente a tomada de

decisão da gestão de resíduos de forma mais eficiente e sustentável.

3 Fundamentação Teórica

3.1 Inteligência artificial

A inteligência artificial é definida em McCarthy (2007) como a ciência e a engenharia de construir máquinas inteligentes, especialmente programas de computador inteligentes. Ela está relacionada a utilização de computadores para entender a inteligência humana, mas não precisa se limitar a métodos biologicamente observáveis.

Segundo Nilsson e Nilsson (1998) a inteligência artificial preocupa-se com o comportamento inteligente, que por sua vez compreende a percepção, raciocínio, aprendizagem, comunicação e atuação em ambientes complexos. Além disso, ela tem como um de seus objetivos desenvolver máquinas capazes de desempenhar esses comportamentos tão bem quanto humanos.

O trabalho desenvolvido em Russell e Norvig (2002) investigou os objetivos ou definições potenciais de IA, que diferencia os sistemas com base na racionalidade e pensamento em contrapartida à ação. Essas definições variam em duas dimensões principais: abordagem humana e abordagem ideal.

A abordagem humana é definida como sistemas que pensam, ou agem, como humanos em certos aspectos. Enquanto que a abordagem ideal faz referência a sistemas que pensam, ou agem, de forma racional. Em Bellman (1978) a abordagem humana da IA é definida como a automação de atividades que associamos ao pensamento humano, atividades como tomada de decisão, resolução de problemas e aprendizagem, por exemplo. Por outro lado, Winston (1992) define a abordagem idealizada da IA como o estudo das faculdades mentais através do uso de modelos computacionais.

De forma simples, pode-se assumir de acordo com Haenlein e Kaplan (2019) que a inteligência artificial combina a capacidade de um sistema de interpretar dados externos corretamente, de aprender com esses dados e de usar os aprendizados para atingir metas e tarefas específicas para permitir a resolução de problemas.

Muitos sistemas usuais do dia a dia vem se tornando inteligentes e adaptáveis no sentido de utilizaram técnicas de inteligência artificial para auxiliar no desenvolvimento de suas tarefas. Essa forma de inteligência artificial mais conhecida é baseada na otimização, e considerada como uma proposta de inteligência artificial do tipo "fraca" entre os teóricos da área. Por outro lado, a inteligência artificial "forte" apresenta uma ideia mais radical onde considera aspectos como consciência e ética no desenvolvimento efetivo de uma máquina inteligente. Esse trabalho examina a aplicação de técnicas de otimização e resolução de problemas centrado a sua investigação na elaboração de algoritmos capazes de aprender

um comportamento em certos conjuntos de dados e generalizá-lo para outros cenários.

As seções seguintes apresentam algumas das técnicas, algoritmos e métodos de inteligência artificial que integram os atributos específicos de várias disciplinas, como matemática, estatística, física, ciência da computação e, recentemente, aplicações em gerenciamento de resíduos sólidos.

3.1.1 Redes Neurais Artificiais

Uma rede neural artificial, segundo [Fausett \(2006\)](#) é um sistema de processamento de informações que possui certas características de desempenho em comum com as redes neurais biológicas. Redes neurais artificiais foram desenvolvidas como generalizações de modelos matemáticos de cognição humana ou biologia neural, com base nas premissas de que:

- O processamento da informação ocorre em muitos elementos simples chamados neurônios.
- Os sinais são passados entre os neurônios por meio de links de conexão.
- Cada link de conexão tem um peso associado, que, em uma rede neural típica, multiplica o sinal transmitido.

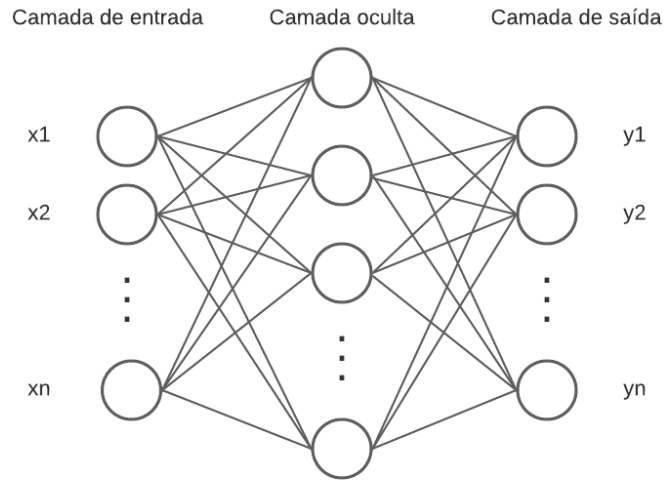
As RNAs têm sido amplamente utilizadas para modelar vários processos de gerenciamento de resíduos sólidos devido à sua robustez, tolerância a falhas e adequação para representar as relações complexas entre variáveis em sistemas multivariados.

De acordo com [Yegnanarayana \(2009\)](#), não se espera que esses modelos cheguem perto do desempenho da rede biológica por vários motivos. Em primeiro lugar, não entendemos totalmente o funcionamento de um neurônio biológico e as interconexões neurais. Além disso, é quase impossível simular o número de neurônios e suas interconexões como existem em uma rede biológica, e suas operações no modo naturalmente assíncrono.

Um modelo básico de RNA consiste em três camadas, e é representado de forma simplificada como um grafo na Figura 1. A camada de entrada é constituída por um conjunto de dados de entrada do ambiente externo. A camada oculta recebe e executa as entradas ponderadas transferidas da camada de entrada, em seguida a saída é transferida para a camada de saída. A camada de saída fornece o resultado do processamento realizado pelas camadas anteriores.

Percebe-se que cada neurônio está conectado a vários de seus vizinhos, com coeficientes ou pesos variáveis (simbolizado pelas arestas do grafo) que representam a influência relativa das diferentes entradas de neurônios para outros neurônios. A soma ponderada das entradas é transferida para os neurônios ocultos, onde é transformada usando uma

Figura 1 – Estrutura de uma rede neural.



Fonte: Elaborado pelo autor, 2022

função de ativação, como por exemplo a função tangente sigmóide. Por sua vez, as saídas dos neurônios ocultos atuam como entradas para o neurônio de saída, onde passam por outra transformação.

Considerando a camada oculta na Figura 1, as informações de todas as entradas (ou seja, x_k) são passadas para os primeiros nós através dos links de conexão. A força desta transferência de informação é medida por pesos de conexão (ou seja, w_{ki}), e o sinal de saída dos nós da camada oculta, y_i é obtido avaliando o valor de uma função de ativação, f , como na Equação 3.1.

$$y_i = f \left(\sum_{k=1}^m (x_k w_{ki}) - \Theta_i \right) \quad (3.1)$$

Onde θ_i é o valor limite, um limiar que indica se o neurônio deve ou não ser ativado.

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3.2)$$

Onde x representa o resultado da soma da Equação 3.1.

Da mesma forma, o vetor do sinal de saída da camada de saída é calculado de acordo com a Equação 3.3.

$$O_t = f \left(\sum_i T_{li} y_i - \Theta_l \right) \quad (3.3)$$

Onde O_t é o vetor do sinal de saída da camada de saída, T_{li} são os pesos de conexão entre a camada oculta e de saída e Θ_l é o valor limite.

A adequação dos parâmetros é avaliada usando uma função de erro da rede, como a soma dos erros quadrados (do inglês, *sum of square errors*) entre o vetor de saída previsto e desejado:

$$E = \frac{1}{n \times m} \sum_{j=1}^n \sum_{l=1}^m (t_l - O_t)^2 \quad (3.4)$$

Onde E é a função de erro da rede e t_l é o vetor de saída desejado.

O vetor de saída é obtido com base em pesos da rede inicialmente aleatórios e, depois de comparado com a saída desejada, o erro computado é propagado de volta pela rede e os pesos são atualizados através do processo de aprendizagem da rede, que visa minimizar a função de erro entre o vetor de saída previsto e o vetor de saída desejado, alterando o peso da camada de entrada e o valor limite ao longo da direção do gradiente. Essa etapa será repetida até que o erro de treinamento satisfaça a precisão exigida, especificada como critério de parada. Todo esse processo é conhecido como treinamento do modelo da rede neural.

Em geral, as RNAs são sensíveis ao número de neurônios em suas camadas ocultas. Poucos neurônios podem levar a um ajuste insuficiente da rede para gerar uma solução adequada. Por outro lado, muitos neurônios podem contribuir para o sobreajuste (do inglês, *overfitting*). Neste caso, o erro no conjunto de dados de treinamento é muito pequeno, porém, quando novos dados são apresentados à rede, obtêm um erro grande. Em resumo, embora a rede tenha memorizado os exemplos de treinamento, ela não aprendeu a generalizar o aprendizado para novos cenários.

Após o treinamento, uma RNA pode gerar as respostas apropriadas para dados diferentes daqueles utilizados na aprendizagem, o que é conhecido como generalização. Assim, quando o aprendizado é concluído a rede neural pode ser usada para realizar inferências sobre um novo conjunto de dados.

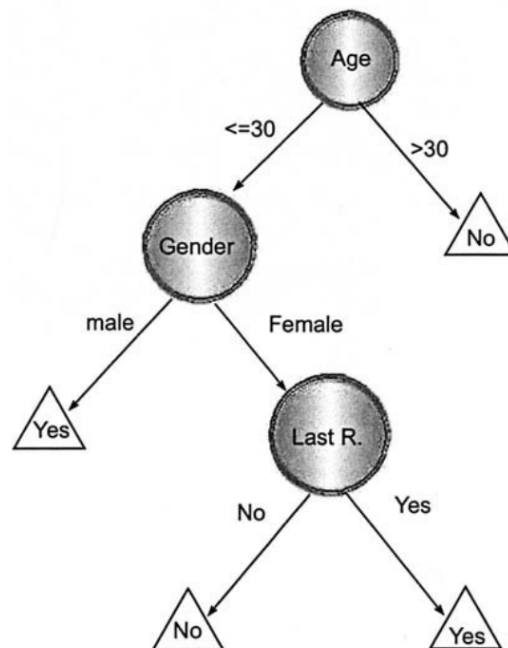
Segundo Yetilmezsoy, Ozkaya e Cakmakci (2011) acredita-se que as RNAs podem fornecer uma boa alternativa à utilização de técnicas estatísticas clássicas quando aplicadas a problemas ambientais devido à sua velocidade e capacidade de aprendizagem. Além disso, as RNAs também podem ser utilizadas em problemas iterativos, que geram uma sequência de soluções que melhoram a cada repetição.

3.1.2 Árvores de decisão

Árvores de decisão são modelos de aprendizado supervisionado que podem ser utilizados para realizar a classificação de dados, e que utilizam regras de decisão baseadas em atributos, sob a forma de árvores (estrutura de dados formada por elementos que armazenam informações representados de forma hierárquica).

Na Figura 2 observa-se a representação de uma estrutura que, de acordo com [Rokach e Maimon \(2005\)](#), descreve uma árvore de decisão para determinar se um cliente em potencial responderá ou não a um e-mail de publicidade.

Figura 2 – Representação de uma árvore de decisão



Fonte: Rokach e Maimon, 2005

Na representação sugerida, os elementos que armazenam as informações são conhecidos como nós, e cada um deles estão conectados à outros. A árvore também é constituída por arestas, ou ramos, que referem-se à atributos categóricos, ou numéricos, utilizados nas regras de decisão. Os nós que possuem ramos de saída são conhecidos como internos e os que não possuem são os nós folha (ou simplesmente folha).

No exemplo da Figura 2, os nós são rotulados com os atributos os quais serão testados e os ramos recebem o valor resultante, que no caso da classificação refere-se à classe. Tendo como base as definições e estruturas apresentadas, verifica-se que se um cliente tem idade menor ou igual à 30 anos, e é do gênero masculino, então ele possivelmente irá responder a um e-mail de publicidade.

Logo, a tarefa principal desse algoritmo é encontrar quais os atributos que serão inseridos em cada nó da árvore, e qual sua posição na hierarquia. Para isso, é utilizado

o cálculo do ganho de informação e a entropia (não uniformidade), como medidas de impureza de um conjunto de dados.

Seja $p(x_i)$ a proporção de uma variável que pode assumir i valores no conjunto, então a entropia pode ser calculada como um somatório das probabilidades, de acordo com a Equação 3.5.

$$Entropia(X) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (3.5)$$

Consequentemente, o ganho de informação mede o quanto a entropia foi reduzida ao particionar o conjunto de dados de acordo com o atributo que está sendo analisado. Assim sendo, dado um atributo A do conjunto de dados, o ganho de informação pode ser calculado conforme Equação y:

$$Ganho(X, A) = Entropia(X) - \sum_{i=1}^n p(A_i) \cdot Entropia(A_i) \quad (3.6)$$

Onde, X é um conjunto de dados e A_i é a proporção da frequência de i em relação ao conjunto.

Então, após a realização dos cálculos é verificado qual o atributo com maior ganho de informação, e este é utilizado como o nó raiz da árvore de decisão. Depois disso, para selecionar os atributos para os nós da esquerda e direita da raiz da árvore, são realizados novos calculos de entropia e ganho de informação. Portanto, o algoritmo é definido como recursivo, isto é, ele é executado de forma repetida para cada nível de profundidade da árvore.

Para avaliar a qualidade da divisão em uma árvore de decisão, é utilizado o índice Gini, que mede a probabilidade de classificação incorreta de uma amostra aleatória em uma população. Esse valor é calculado para cada divisão da árvore e a divisão com o menor índice Gini é escolhida como a melhor divisão. Uma divisão com um índice Gini mais baixo significa que as amostras resultantes são mais homogêneas em relação à classe. A fórmula para o cálculo do índice Gini é dada conforme 3.7.

$$Gini(p) = 1 - \sum_{i=1}^J p_i^2 \quad (3.7)$$

Onde p_i é a proporção de amostras que pertencem à classe i e J é o número total de classes.

O uso do índice Gini na construção de árvores de decisão tem várias vantagens. Ele é fácil de calcular e interpretar, adequado para dados categóricos e numéricos, e a árvore

de decisão construída é fácil de entender e visualizar. No entanto, é importante lembrar que a árvore de decisão pode sofrer de sobreajuste se o conjunto de dados de treinamento for muito pequeno ou desequilibrado. É fundamental avaliar cuidadosamente a qualidade da árvore de decisão construída antes de usá-la para tomada de decisões em um problema real.

3.1.3 Naive bayes

De acordo com [Webb, Keogh e Mäkeläinen \(2010\)](#) Naive Bayes é um algoritmo de aprendizado simples que utiliza a regra de Bayes junto com uma forte suposição de que os atributos de uma classe são condicionalmente independentes. Essa suposição geralmente é contrariada na prática, visto que os atributos geralmente costumam ter um determinado grau de correlação entre eles. Por isso ele recebe o nome "*naive*" (do inglês, ingênuo).

O algoritmo Naive Bayes é treinado usando um conjunto de dados de treinamento que contém as características de entrada e as classes correspondentes. Ele usa essas informações para calcular a probabilidade de cada classe dado um conjunto de características de entrada. Isso é feito usando a regra de Bayes, que leva em consideração a probabilidade a priori de cada classe e a probabilidade condicional de cada característica dada a classe.

Por ser um classificador probabilístico, ele utiliza as informações no conjunto de dados para retornar a probabilidade de uma classe y , dado um conjunto de dados x . Portanto, utilizando o teorema de Bayes, calcula-se $P(y|x)$:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (3.8)$$

Para realizar o cálculo da Equação 3.8, primeiramente é elaborado uma tabela de frequências para todos os preditores do conjunto de dados. Depois deve-se determinar a densidade condicional de classe, para cada valor de y , da seguinte forma:

$$P(x|y) = \prod_{i=1}^n P(x_i|y) \quad (3.9)$$

Onde, x_i é o i -ésimo atributo em x , e n é o número de atributos.

Então, utiliza-se a Equação 3.10 para calcular $P(x)$.

$$P(x) = \prod_{i=1}^k P(c_i) P(x|c_i) \quad (3.10)$$

Onde, c_i é a i -ésima classe e k é o número de classes.

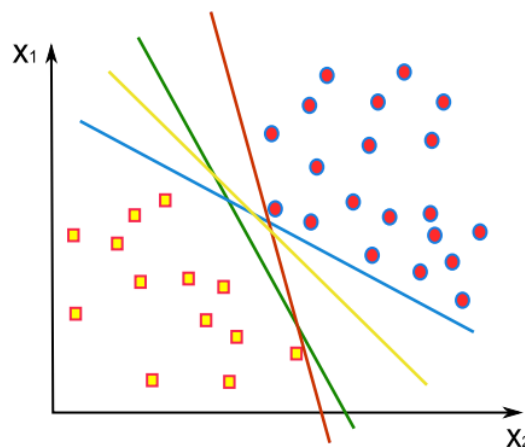
Após o cálculo das probabilidades, o algoritmo classifica cada conjunto de características de entrada na classe com a maior probabilidade condicional. Isso é feito comparando as probabilidades calculadas para cada classe e selecionando a classe com a maior probabilidade. Se houver empate nas probabilidades, o algoritmo pode usar uma estratégia de desempate, como selecionar a classe com a menor variância.

Naive Bayes tem várias vantagens em relação a outros algoritmos de classificação. Ele é rápido, eficiente em termos de memória e funciona bem com conjuntos de dados grandes. Além disso, segundo [Murphy et al. \(2006\)](#) o Naive Bayes pode verificar a incerteza sobre a previsão, isto é, o resultado sobre a classificação para um determinado valor é representado por uma probabilidade de o valor ser rotulado com a classe correta. No entanto, a suposição de independência das características de entrada pode não ser válida em todos os casos, o que pode levar a uma performance inferior em alguns conjuntos de dados.

3.1.4 Support Vector Machine (SVM)

De acordo com [Cervantes et al. \(2020\)](#) o SVM foi introduzido por Vapnik como um modelo de aprendizado de máquina baseado em kernel para a realização de tarefas de classificação e regressão. Seu objetivo é maximizar a habilidade de generalização de um modelo utilizando uma superfície N-dimensional para dividir um conjunto de dados em classes distintas. A Figura 3 ilustra a divisão de um conjunto de exemplo por quatro hiperplanos de separação.

Figura 3 – Hiperplanos dividindo um conjunto de dados



Fonte: Cervantes et al., 2020

Segundo [Patle e Chouhan \(2013\)](#) o SVM tem quatro conceitos básicos:

1. O hiperplano de separação;
2. O hiperplano de maximização da margem;

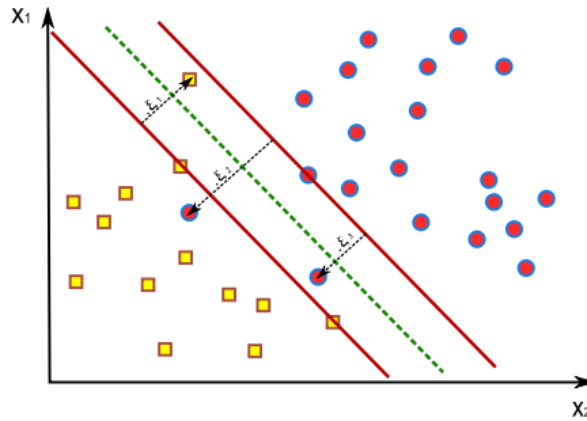
3. A suavização da margem;
4. A função kernel.

O hiperplano de maximização (H) da margem é aquele que possui a maior distância para o exemplo mais próximo de cada uma das classes (margem), enquanto que a suavização da margem consiste em encontrar dois hiperplanos (H_1 e H_2) paralelos à H que permitam que alguns pontos de dados atravessem a margem do hiperplano de separação sem afetar o resultado final da classificação. Pontos que estejam sobre os planos H_1 e H_2 são considerados vetores de suporte que auxiliam na definição de H .

A suavização da margem é necessária para lidar com problemas em que as classes estão sobrepostas, ou com ruídos existentes no conjunto de dados, permitindo que alguns pontos pertencentes à uma dada classe esteja no lado errado do hiperplano de separação sem que o resultado da classificação seja afetado.

Na Figura 4 verifica-se um exemplo da suavização da margem, onde existem dois pontos que estão entre H_1 e H_2 .

Figura 4 – Suavização da margem de um SVM



Fonte: Cervantes et al., 2020

Portanto, para cada ponto (x_i, y_i) , no caso de uma separação linear, constituído pelo valor de um atributo x_i e pela sua classe $y_i \in (+1, -1)$, o hiperplano de separação H é definido pela Equação 3.11.

$$\mathbf{w}^T \mathbf{x}_i + b = 0 \quad (3.11)$$

Onde, w é normal ao hiperplano.

Então, adotando d^+ como a menor distância entre H e os exemplos mais próximos da classe positiva, e d^- a menor distância entre H e os exemplos mais próximos da classe negativa, assume-se as seguintes restrições:

$$w \cdot x^+ + b = 1 \quad (3.12)$$

$$w \cdot x^- + b = -1 \quad (3.13)$$

Combinando a Equação 3.12 e a Equação 3.13, obtém-se:

$$y_i(w \cdot x_i + b) - 1 \geq 0 \quad \forall_i \quad (3.14)$$

Então o hiperplano de separação linear pode ser encontrado solucionando-se o seguinte problema de otimização:

$$\begin{aligned} \min_{w,b} (w \cdot w) &= \|w\|^2 \\ \text{sujeito a : } y_i(w \cdot x_i + b) - 1 &\geq 0 \quad \forall_i \end{aligned} \quad (3.15)$$

Onde, $\|w\|$ é a norma Euclidiana de w .

Em casos em que existe a intersecção entre as classes, como na Figura 4 a restrição dada pela Equação 3.14 não pode ser satisfeita, logo é necessário a introdução de variáveis de folga, não negativas $\xi_i (\geq 0)$. Dessa forma, obtém-se a seguinte restrição:

$$y_i(w^T \cdot x_i + b) \geq 1 - \xi_i \quad \forall_i \quad (3.16)$$

Consequentemente, a largura da margem suavizada pode ser controlada através de um parâmetro de penalização C que determina a relação entre o erro no treinamento e um erro empírico. Portanto, $\sum_{i=1}^N \xi_i$ é o limite para a quantidade de erros no treinamento do modelo. De outro modo, o hiperplano de separação agora é encontrado de acordo com o problema de otimização da Equação 3.17:

$$\begin{aligned} \min_{w,b} (w \cdot w) &= C \sum_{i=1}^l \xi_i^2 \\ \text{sujeito a : } y_i(w \cdot x_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0 \end{aligned} \quad (3.17)$$

Finalmente, para implementar o SVM em um espaço multidimensional é necessário a utilização de uma função kernel que, de acordo com Patle e Chouhan (2013) realiza a projeção de um espaço de baixa dimensão para um com dimensão superior. Isso seria semelhante ao mapeamento de um espaço de entrada X para um espaço equivalente, como na Equação 3.18:

$$X = (x_1, x_2) \rightarrow K(x_1, x_2) = (\Phi(x_1) \cdot \Phi(x_2)) \quad (3.18)$$

Ainda de acordo com [Patle e Chouhan \(2013\)](#), o SVM é preferível por lidar com mais de duas variáveis preditoras, lidar com classificações com mais de duas categorias e suportar uma grande quantidade de atributos em um conjunto de dados.

3.1.5 K-Nearest Neighbors (KNN)

O KNN é um método de classificação não-paramétrico que, de acordo com [Guo et al. \(2003\)](#) classifica um registro t em um conjunto de dados recuperando seus k vizinhos mais próximos, formando uma vizinhança de t . A classificação desse registro é decidida de acordo com a maioria dos registros vizinhos de t de forma à considerar, ou não, a ponderação da distância.

Existem dois conceitos importantes ao analisar o método KNN. São eles:

1. O método utilizado para calcular a distância entre as classes;
2. A escolha apropriada de k .

Com relação ao primeiro conceito, por padrão é utilizado a distância Euclidiana, apresentada na Equação 3.19, por ser intuitiva, fácil de calcular e funcionar bem em muitos problemas de classificação. No entanto, em algumas situações, outras medidas de distância, como a distância de Manhattan ou a distância de Chebyshev, podem ser mais adequadas. Por exemplo, em problemas onde as características dos dados são categóricas em vez de numéricas, outras medidas de distância podem ser mais apropriadas.

$$D(p, q) = \sqrt{\sum_{i=1}^N (p_i - q_i)^2} \quad (3.19)$$

Onde, $P = (p_1, \dots, p_n)$ e $Q = (q_1, \dots, q_n)$.

Quanto ao segundo conceito, para [Zhang \(2016\)](#) a escolha apropriada de k tem um impacto significativo no desempenho de diagnóstico do algoritmo KNN. Um k grande reduz o impacto da variância causada pelo erro aleatório, mas corre o risco de ignorar um padrão pequeno, mas importante.

Após encontrar as k amostras mais próximas do novo ponto de dados no espaço de características, o algoritmo KNN determina a classe do novo ponto de dados pela classe mais frequente entre essas k amostras. Isso é feito contando o número de amostras em cada classe e selecionando a classe com o maior número de amostras. Por exemplo, se dos k vizinhos mais próximos, 3 pertencem à classe A, 5 pertencem à classe B e 2 pertencem à classe C, o novo ponto de dados será atribuído à classe B.

No caso de uma situação de igualdade nas probabilidades condicionais para duas ou mais classes, o algoritmo pode usar uma estratégia para resolver a concorrência entre elas, como selecionar a classe com a menor variância. Outra opção é determinar a classe do novo ponto de dados atribuindo pesos às amostras com base em uma medida de distância entre o ponto de dados e seus k vizinhos mais próximos. Em seguida, a classe com o maior peso é selecionada como a classe do novo ponto de dados.

Sendo assim, para classificar um exemplo utilizando o KNN realiza-se as seguintes etapas:

1. Selecione o número K de elementos vizinhos;
2. Para cada exemplo nos dados, calcule a distância entre um exemplo de consulta e os K vizinhos mais próximos;
3. Adicione as distâncias calculadas em uma coleção ordenada;
4. Classifique a coleção em ordem crescente de acordo com as distâncias;
5. Escolha as k primeiras entradas da coleção e retorne a moda das classes.

Pode-se verificar a precisão do modelo desenvolvido utilizando a Equação 3.20:

$$Precisao\ media = \sum_{i=1}^l \frac{VP_i + VN_i}{VP_i + FN_i + FP_i + VN_i} / l \quad (3.20)$$

Onde, VP são os verdadeiros positivos, VN são os verdadeiros negativos, FP são os falsos positivos, FN os falsos negativos e l o total de classes.

O algoritmo KNN é um algoritmo simples e fácil de implementar, mas pode ser computacionalmente caro para grandes conjuntos de dados e não lida bem com características irrelevantes ou correlacionadas. Além disso, a escolha do valor de k pode ter um grande impacto na precisão do modelo.

3.2 Sistemas de apoio à decisão

Segundo [Espinasse e Pascot \(1987\)](#) sistemas de apoio à decisão referem-se à sistemas de computador construídos para auxiliar os gerentes na tomada de decisões semiestruturadas ou mal estruturadas. Eles devem auxiliar os responsáveis pela tomada de decisões diante de diferentes problemas combinando o uso de modelos ou técnicas analíticas com acesso à um bancos de dados.

Entende-se que a utilização de sistemas de apoio à decisão origina-se da necessidade de um sistema de gerenciamento de informações que apresente possibilidades dinâmicas e integração com processos interligados a uma demanda específica.

De acordo com [Melaré et al. \(2017\)](#) os sistemas de apoio à decisão são ferramentas valiosas para auxiliar os gestores a garantir a conformidade com os regulamentos de gestão de resíduos sólidos propostos pelos governos. Um SAD pode ser desenvolvido usando tecnologias de informação e algoritmos de otimização.

Segundo [Perraju \(2013\)](#) são características de um SAD:

- Eles tendem a visar os problemas menos bem estruturados e subespecificados que os gerentes de nível superior normalmente enfrentam;
- Eles tentam combinar o uso de modelos ou técnicas analíticas com funções tradicionais de acesso e recuperação de dados;
- Eles se concentram especificamente em recursos que os tornam fáceis de serem utilizados por seus usuários;
- Eles enfatizam a flexibilidade e adaptabilidade para acomodar mudanças no ambiente e na abordagem de tomada de decisão do usuário.

De acordo com [Sprague e Carlson \(1982\)](#) os SADs são compostos por três componentes principais: dados, modelos e interface com o usuário. Cada um desses componentes desempenha um papel crucial no processo de tomada de decisões.

Os dados são a base de informações para a tomada de decisões. Eles podem incluir dados históricos, dados em tempo real, informações sobre o mercado ou qualquer outra informação relevante para a decisão em questão. A qualidade dos dados é fundamental para a precisão das decisões tomadas com base nesses dados. Os dados são normalmente armazenados em um banco de dados que pode ser acessado pelo sistema de apoio à decisão.

Os modelos fornecem a estrutura e os métodos para analisar e processar os dados. Eles podem ser modelos estatísticos, modelos de simulação ou modelos de otimização, entre outros. Os modelos são projetados para ajudar a extrair informações dos dados e fornecer uma base para a tomada de decisões. Eles podem ser ajustados e refinados à medida que novos dados são incorporados ao sistema.

A interface do usuário permite que os usuários interajam com o sistema, inserindo informações e visualizando os resultados. É através da interface que os usuários podem acessar o sistema, fornecer informações e instruções, e receber respostas sobre as decisões tomadas. A interface pode ser um *software* de *desktop* ou baseado na web, dependendo da implementação específica do SAD.

Em conjunto, esses três componentes trabalham juntos para fornecer um sistema de apoio à decisão eficaz. Os dados fornecem a base de informações, os modelos fornecem as ferramentas para analisar e processar essas informações e a interface permite que os usuários interajam com o sistema e recebam retorno sobre as decisões tomadas.

A utilização de técnicas de mineração de dados em sistemas de apoio à decisão tem sido cada vez mais comum nos últimos anos. De acordo com [Sharma et al. \(2019\)](#), as ferramentas de mineração de dados são utilizadas em diversas áreas de negócios e indústrias, incluindo finanças, marketing, medicina e transporte. Essas ferramentas permitem extrair informações valiosas de grandes conjuntos de dados, identificar padrões e tendências ocultas, e fornecer conhecimento para suportar a tomada de decisões.

Existem várias técnicas e algoritmos de mineração de dados que podem ser utilizados para análise de dados e previsão de resultados em um sistema de apoio à decisão. Cada técnica é adequada para diferentes tipos de dados e problemas específicos, e é importante selecionar a mais adequada para cada caso. As técnicas de mineração de dados mais comuns utilizadas em SADs incluem:

- Árvore de decisão;
- Clusterização;
- Redes neurais
- Regras de associação.

As ferramentas de mineração de dados também são capazes de lidar com diferentes tipos de dados, incluindo dados estruturados e não estruturados. Além disso, de acordo com [Alsabti, Ranka e Singh \(1998\)](#), muitas ferramentas de mineração de dados incluem recursos de visualização de dados para facilitar a interpretação dos resultados.

4 Metodologia

A metodologia para o desenvolvimento desse trabalho leva em consideração quatro etapas: a primeira delas é a pesquisa de campo ([seção 4.1](#)) a qual trata do processo de realização dos levantamentos de dados a partir do questionário que foi utilizado para a obtenção de informações qualitativas da população amostrada. Em seguida, a segunda etapa ([seção 4.2](#)) contempla a análise dos dados. A terceira etapa compreende o desenvolvimento do modelo baseado em inteligência artificial ([seção 4.3](#)) utilizado para a predição do volume e da caracterização dos resíduos sólidos urbanos. Por fim o desenvolvimento do sistema de apoio à decisão ([seção 4.4](#)) fornece, de forma organizada, todos os dados necessários, apresentando informações para auxiliar no processo de tomada de decisão no gerenciamento de resíduos sólidos.

4.1 Pesquisa de campo

Para o desenvolvimento de um modelo baseado em inteligência artificial que seja capaz de realizar a predição da geração de resíduos sólidos domiciliares de um município de maneira razoável, é necessária uma grande quantidade de dados que expliquem da melhor forma possível o cenário em questão. De acordo com [Miezahl et al. \(2015\)](#), faltam dados nacionais confiáveis sobre a geração e composição de resíduos sólidos para o planejamento e gestão eficaz. Tais dados fornecem um recurso abrangente para uma avaliação extensiva, crítica e informativa das opções na gestão de resíduos sólidos, como o planejamento da coleta seletiva e a diminuição da quantidade de resíduos enviados para os aterros sanitários.

Infelizmente, esses dados fundamentais estão em falta em muitos países em desenvolvimento segundo [Buenrostro, Bocco e Bernache \(2001\)](#), e quando estão disponíveis geralmente são inconsistentes por serem retirados de muitas fontes que não podem ser validadas e muitas vezes não são plausíveis por serem baseados em suposições não científicas.

No caso do estado de Mato Grosso do Sul, têm-se pouca informação sobre pesquisas voltadas a avaliar a geração de resíduos sólidos domiciliares diretamente da fonte geradora, ou seja, realizando a coleta nos domicílios. A maior parte das informações disponíveis é referente à quantidade de resíduos recolhida pela empresa responsável pela coleta de lixo.

Existem algumas dificuldades na apuração da geração de resíduos sólidos domiciliares, como por exemplo a participação voluntária dos moradores, disponibilidade de pessoal treinado, disposição de veículo para realização das coletas, local para segregação

dos materiais, entre outros.

Levando em conta os obstáculos relacionados à aquisição dessas informações fez-se necessário a realização de uma pesquisa de campo para o levantamento dos dados que foram utilizados para o treinamento de um algoritmo de inteligência artificial capaz de realizar a predição de aspectos importantes relacionados ao gerenciamento de resíduos sólidos, como a quantificação da geração e caracterização dos materiais.

4.1.1 Definição do escopo

A pesquisa de campo foi realizada com base em alguns procedimentos metodológicos apresentados em [Berríos \(1997\)](#), que trata das técnicas de amostragem de resíduos sólidos urbanos. Foram realizados alguns ajustes para que as técnicas pudessem ser aplicadas no âmbito domiciliar.

No que se refere aos procedimentos, inicialmente foi realizado um estudo para levantar os dados socioeconômicos do município de Campo Grande, MS. Esse município foi utilizado como modelo para a realização da pesquisa de campo, que futuramente poderá ser estendida para demais municípios do estado, enriquecendo a base de dados. Tal levantamento apurou dados do IBGE, que incluem o número de domicílio, população urbana e quantidade de moradores por domicílio, e da Prefeitura Municipal, como lista de bairros por região e valor do IPTU por domicílio.

Em seguida, foi calculado o tamanho da amostra necessária para um erro de estimação de 8% sobre o número total de domicílios da cidade (considerando uma margem de 95% de confiabilidade), que de acordo com os dados fornecidos pela Prefeitura são de 303.433 domicílios. O cálculo do tamanho da amostra para uma população infinita (quando o desvio padrão é desconhecido) pode ser realizado com base na Equação 4.1.

$$n_0 = \left(\frac{z_{\frac{\alpha}{2}}}{2 \cdot e} \right)^2 \quad (4.1)$$

Onde n_0 corresponde ao tamanho da amostra para uma população infinita, $z_{\frac{\alpha}{2}}$ é referente ao Z crítico e, e é o erro de estimação.

Em seguida, foi necessário realizar a correção desse valor, visto que a população, isto é, o número de domicílios, é finita. Para tal, utiliza-se a Equação 4.2, onde:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \quad (4.2)$$

Onde n é o tamanho da amostra corrigido, e N é o tamanho da população.

Dessa forma, a partir das equações anteriores, encontrou-se o tamanho da amostra como sendo igual a 150 domicílios. De acordo com o tamanho da amostra, foi estipulada a lista de materiais que seriam utilizados no decorrer da pesquisa para a coleta dos dados e classificação dos resíduos (segregação e pesagem), o número de pessoas necessárias para o andamento do estudo e a quantidade de veículos para realizar a coleta dos resíduos no tempo planejado. Em geral, foram utilizados:

1. Balanças
2. Pares de luvas de proteção
3. Recipientes para separação
4. Lonas
5. Sacolas plásticas
6. Etiquetas de identificação
7. Panfletos para divulgação da pesquisa 5

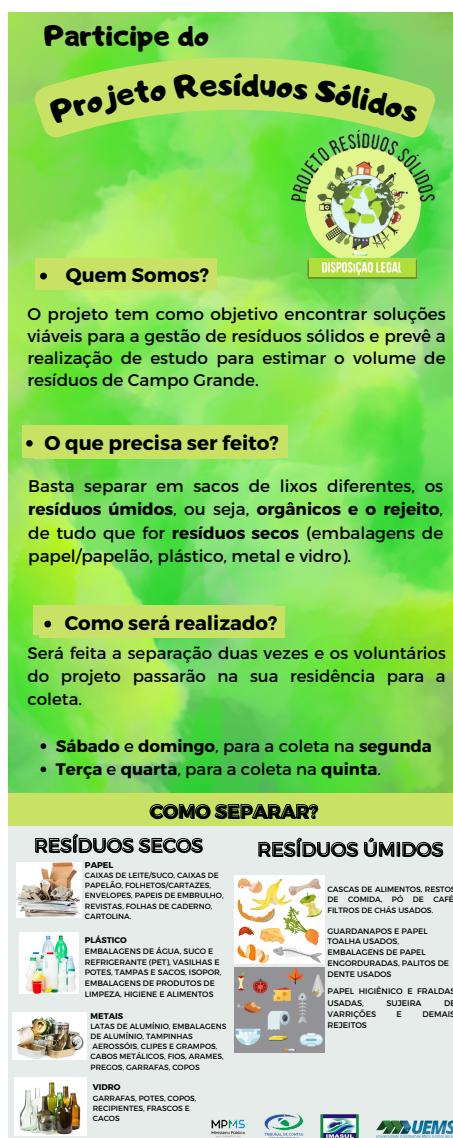
Em relação as pessoas que participaram da pesquisa, foram selecionados alunos da UFMS e bolsistas do convênio entre UEMS e MPMS, como voluntários, para visitar os domicílios e auxiliar nas coletas, bem como realizar a gravimetria dos resíduos sólidos (segregação e pesagem). Para o transporte dos alunos para os locais de coleta, foram utilizados dois veículos (Lifan x60 da UFMS e um veículo pessoal) de acordo com a quantidade de domicílios que seriam visitados no dia.

Para esta etapa de coleta, foi produzido um material de apoio para ajudar os moradores a identificar os dois diferentes grupos de resíduos que seriam coletados pela equipe de pesquisa, diferenciados pela sua natureza física (resíduos secos e resíduos úmidos), os quais deveriam ser separados pelo morador em sacolas distintas.

Os resíduos secos são representados por plásticos, papeis e papelão, metais, entre outros tipos de resíduos que podem ser reciclados, enquanto que os resíduos úmidos são constituídos principalmente por restos de comida, resíduos oriundos da varrição (matéria orgânica), fraldas, papel higiênico (rejeitos).

A Figura 5 apresenta o panfleto utilizado para expor os objetivos da pesquisa, e instruir os moradores a realizar a separação de todo o lixo domiciliar entre os dois grupos de resíduos, assim como lembrá-los dos dias em que a equipe realizaria a coleta.

Figura 5 – Panfleto de divulgação e orientação



Fonte: Elaborado pelo autor, 2022

4.1.2 Seleção dos domicílios

Foram consideradas duas condições no momento da seleção dos domicílios participantes da pesquisa. A primeira delas é de que a amostragem deveria levar em consideração os diversos estratos da população municipal, agrupados de acordo com a renda domiciliar (em salários mínimos). Essa divisão com base na renda reflete a variação nos padrões de consumo advinda do poder de compra da população.

A amostra por estratos é um método de amostragem em que uma população é dividida em subgrupos diferentes com o objetivo de melhorar a precisão da amostra, reduzindo o erro amostral, visto que a variação da amostra como um todo seria reduzida para pequenas variações referentes a cada estrato. Para isso considerou-se a seguinte divisão de estratos:

- Estrato 1: até 1 salário mínimo;
- Estrato 2: mais de 1 à 5 salários mínimos;
- Estrato 3: mais de 5 à 10 salários mínimos;
- Estrato 4: mais de 10 à 20 salários mínimos.

Também foram estipuladas as respectivas porcentagens de participação sobre os domicílios totais: 40% deveriam abranger o primeiro estrato, 30% o segundo, 20% o terceiro e 10% o quarto. Dessa forma o estrato mais popular receberia uma maior significância sobre os demais, transparecendo o cenário social.

A segunda condição se refere à aleatoriedade na seleção dos domicílios dentro de cada estrato. Com esse propósito, foi estudada a possibilidade de divulgar a pesquisa gravimétrica para a população, e também realizar visitas em alguns bairros de estratos em que houvesse menor participação de voluntários. Nessas visitas (de porta em porta), a equipe apresentou os objetivos da pesquisa e os moradores foram convidados a se voluntariar.

Também foi elaborado um formulário onde cada voluntário pudesse se inscrever e responder algumas questões relacionadas ao perfil domiciliar, que incluem, endereço, renda, quantidade de moradores, valor do IPTU e informações para contato).

De acordo com essas condições, e com a finalidade de diminuir o erro amostral derivado do período do ano em que as amostras seriam recolhidas, definiu-se um cronograma para a realização de dois levantamentos de dados. O primeiro teve início na última semana de maio de 2022 e teve duração de duas semanas, e o segundo, iniciando na última semana de agosto desse mesmo ano, e com duração de cinco semanas.

Durante o primeiro levantamento a equipe conseguiu recolher dados alusivos à 46 domicílios, enquanto que no segundo foram coletados dados de 112 domicílios, somando 158 domicílios para integrar a amostra.

4.1.3 Treinamento técnico

Foi realizado um treinamento técnico coletivo com os membros da equipe para orientar sobre como iria se desenvolver as etapas da pesquisa, bem como instruir sobre como deveriam ser passadas as informações para os moradores sobre a separação, e também prepará-los para os procedimentos da segregação.

No que se refere a abordagem aos moradores, no caso das visitas domiciliares, a equipe foi orientada quanto à abordagem na explicação dos objetivos e importância da pesquisa e indagar sobre a disposição dos mesmos para participar como voluntários.

Com relação à segregação, os integrantes da equipe receberam instruções sobre os tipos de resíduos, e foram acompanhados por profissionais da área ambiental enquanto

Diante disso e com base no dia planejado para realizar a coleta, a equipe responsável pela pesquisa se dirigiu até a residência do morador e realizou a coleta das sacolas, as quais foram identificadas com um identificador único, para posteriormente serem associadas aos dados socioeconômicos do respectivo domicílio.

A Figura 8 apresenta o processo de coleta, onde um membro da equipe chega no domicílio, conversa com o morador, identifica as sacolas plásticas com os resíduos e conduz os mesmos para a carroceria do veículo.

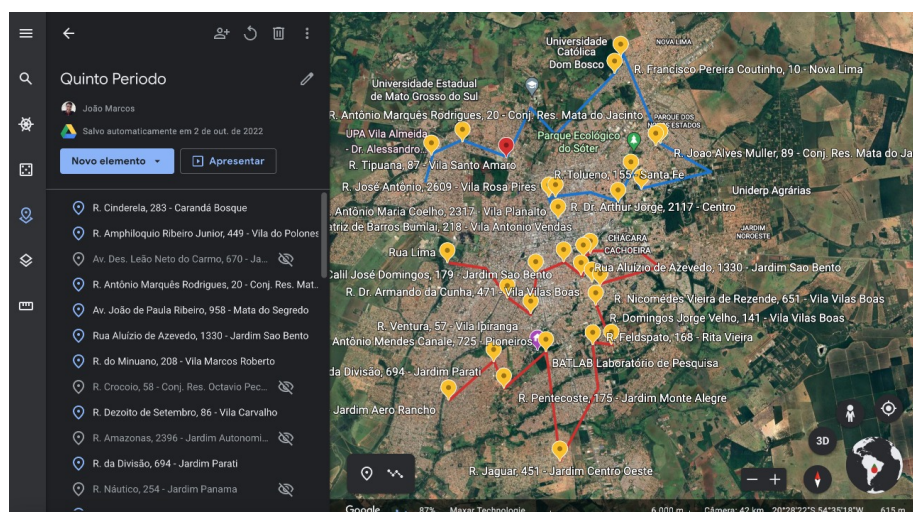
Figura 8 – Processo de coleta dos resíduos.



Fonte: Elaborado pelo autor, 2022

Foi desenvolvido um mapa para auxiliar nas rotas entre os domicílios selecionados para serem visitados no seu devido dia. Para tal foi utilizada a ferramenta Google Earth. A Figura 9 apresenta a ferramenta descrita, com a rota utilizada pela equipe na quinta semana de coleta do levantamento realizado em agosto.

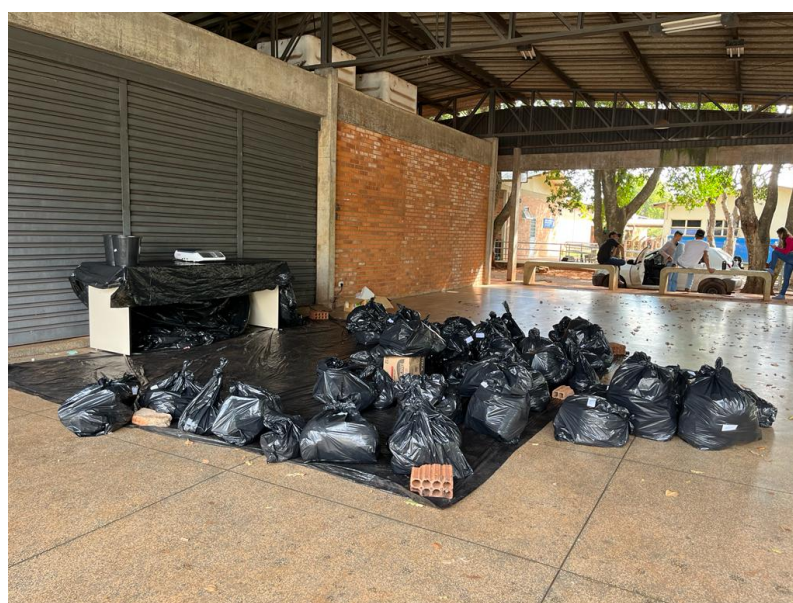
Figura 9 – Mapa com rotas entre domicílios



Fonte: Elaborado pelo autor, 2022

Após a coleta e identificação das sacolas com resíduos dos domicílios, as amostras foram levadas para um espaço na universidade para que posteriormente fosse realizada a segregação e pesagem dos resíduos, e também o registro dos dados (Figura 10).

Figura 10 – Disposição das amostras de resíduos sólidos



Fonte: Elaborado pelo autor, 2022

4.1.5 Classificação e pesagem

Com as sacolas plásticas no local definido para realizar a segregação, a equipe de pesquisa classificou os resíduos conforme a PNRS indica: plástico, papel e papelão, vidro, metais, embalagens multicamadas, têxteis couros e borrachas, matéria orgânica e rejeitos. Também foi considerada a distinção entre embalagens e não embalagens durante essa etapa.

Com o auxílio dos equipamentos já citados anteriormente, foi pesado todo o resíduo de um determinado domicílio para que se pudesse ter o valor total da geração. Depois disso, foi aberto cada uma das sacolas (resíduos secos e úmidos) e os resíduos eram separados manualmente. Após a separação dos resíduos provenientes de uma mesma residência, foi utilizado um recipiente para armazenar cada um dos tipos de resíduos individualmente e então a pesagem foi realizada. Os valores foram registrados em uma planilha, conforme Figura 12, e o fluxograma do procedimento para a pesagem e classificação é apresentado na Figura 11.

Figura 11 – Fluxograma da pesagem e classificação dos resíduos

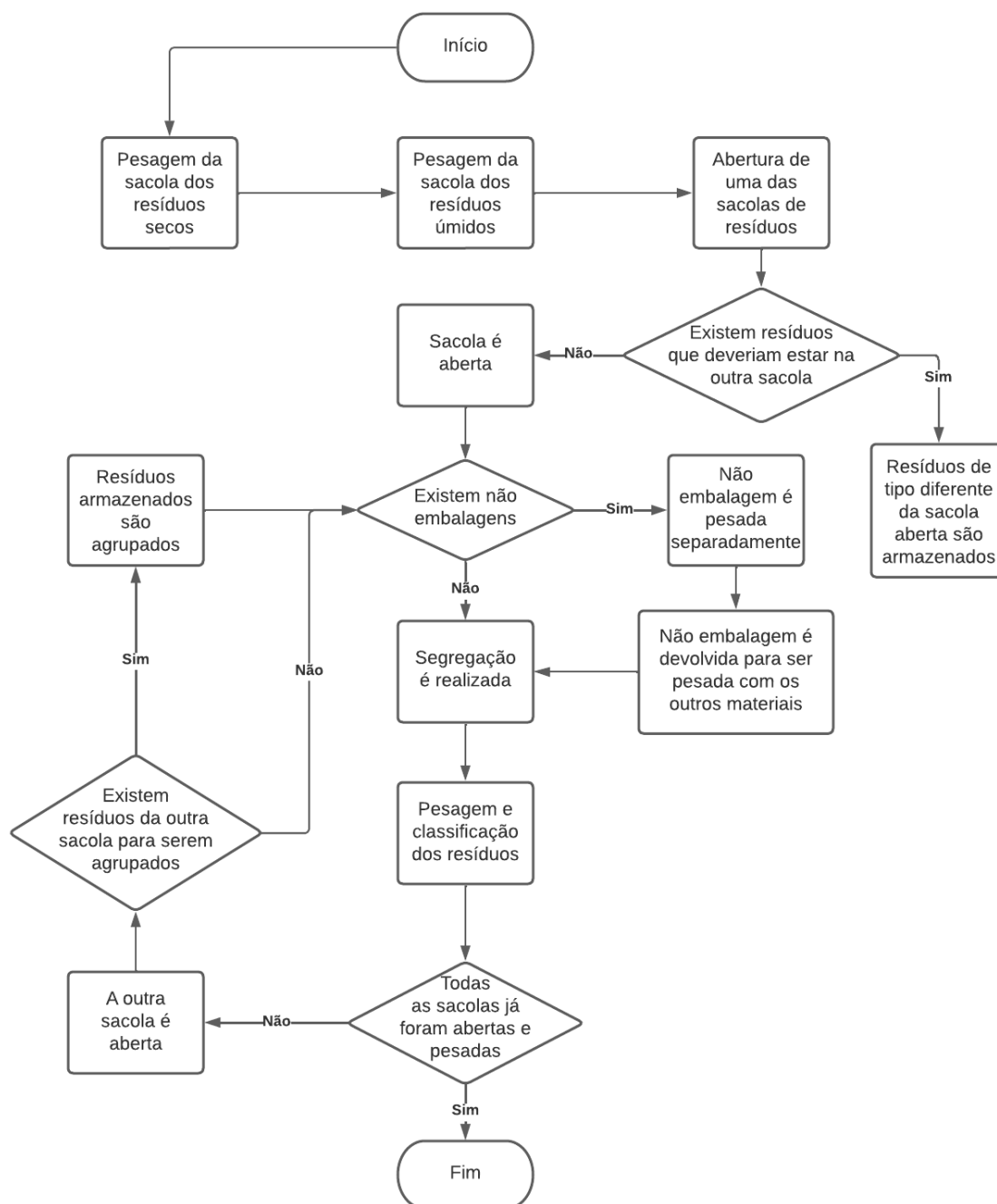


Figura 12 – Planilha de registro da classificação

Amostra	Plástico	Papel e Papelão	Vidro	Metais	Emb. Mult.	Tex. Cour. Bor.	Isopor	Outros	Mat. Org.	Rejeitos	Não Emb.	Moradores	Renda	IPU
158	150	240	0	0	34	0	0	308	1094	0	0	2	Mais de 10 à 20	1994.35
160	170	60	0	0	0	4	13	0	1130	186	16	2	Mais de 5 à 10	1,248.82
161	382	972	2152	16	64	130	68	0	90	0	10	2	Mais de 10 à 20	3,000.00
162	306	492	902	426	0	0	34	850	1442	88	68	2	Mais de 10 à 20	800.00
163	418	530	868	0	2	48	156	1014	2162	286	0	3	Mais de 2 à 5	
164	362	174	804	148	26	36	18	1314	3328	142	10	3	Mais de 10 à 20	
165	524	214	0	0	364	0	0	0	1302	1338	8	2	Mais de 10 à 20	1,318.23
167	178	104	0	40	76	0	0	2934	664	62	0	4	Mais de 10 à 20	
169	836	184	0	444	0	30	2	0	4710	0	0	2	Mais de 1/2 à 1	650.00
171	1040	1848	918	44	180	0	0	0	2776	0	10	4	Mais de 20	4,000.00
174	472	430	730	172	0	0	162	4040	670	3212	0	3	Mais de 5 à 10	1,200.00
175	458	373	636	104	10	4	6	0	16	144	68	3	Mais de 5 à 10	2,000.00
176	480	214	432	320	30	0	18	6134	4502	0	74	3	Mais de 20	4,000.00
178	428	1550	1068	226	0	0	86	1226	1787	4596	8	3	Mais de 5 à 10	6,000.00
179	1202	944	0	174	122	110	20	2618	3558	342	24	4	Mais de 5 à 10	
180	76	244	0	0	32	0	2	834	880	0	40	3	Mais de 1 à 2	100.00
181	100	18	0	38	74	0	0	0	1026	54	0	1	Mais de 1 à 2	450.00
182	682	424	0	48	30	32	22	1190	6778	658	44	4	Mais de 10 à 20	3,864.44
183	1050	1230	4608	506	278	292	0	1870	612	0	60	5	Mais de 1 à 2	1,200.00
185	108	158	0	10	0	0	0	0	550	154	58	3	Mais de 1 à 2	1,019.50
186	2938	458	3634	536	114	0	28	0	4652	42	0	3	Mais de 2 à 5	300.00
187	674	764	250	142	0	0	6	2956	400	2024	28	4	Mais de 5 à 10	1,280.00
191	124	346	0	0	74	46	2	1020	1362	862	4	4	Mais de 5 à 10	3,500.00
192	86	122	0	0	88	0	6	1054	138	676	22	2	Mais de 1 à 2	1,700.00
193	610	124	0	22	278	0	0	2318	8526	1832	0	3	Mais de 10 à 20	
194	256	70	0	2	0	0	0	1152	1424	268	4	1	Mais de 5 à 10	1,032.00
195	278	282	374	32	14	0	14	3790	130	2644	26	3	Mais de 5 à 10	5,400.00
196	572	896	0	4	0	0	104	0	300	868	102	2	Mais de 2 à 5	1481.14
197	72	88	0	0	0	0	0	0	2434	0	0	2	Mais de 2 à 5	1,644.00
198	156	422	0	0	0	0	48	1304	4158	350	12	3	Mais de 10 à 20	5467.44
199	172	122	0	0	0	0	0	1340	1632	94	0	3	Mais de 1 à 2	2,400.00

Fonte: Elaborado pelo autor, 2022

A Figura 13 mostra os processos de classificação e pesagem sendo realizados pela equipe de pesquisa.

Figura 13 – Equipe realizando a classificação dos resíduos



Fonte: Elaborado pelo autor, 2022

4.2 Análise dos dados

A análise dos dados é um processo fundamental para descobrir informações sobre os dados provenientes da pesquisa de campo. Com esse propósito, foi utilizado o Excel e o Python (uma linguagem de programação licenciada pela Python Software Foundation) para realizar os processos de limpeza, análise exploratória e visualização de dados.

O processo de limpeza de dados (do inglês, data cleaning) consiste na prevenção ou correção de erros resultantes da forma como os dados são inseridos em um sistema, ou armazenados. De forma geral, os dados brutos podem ser imprecisos, duplicados ou segmentados.

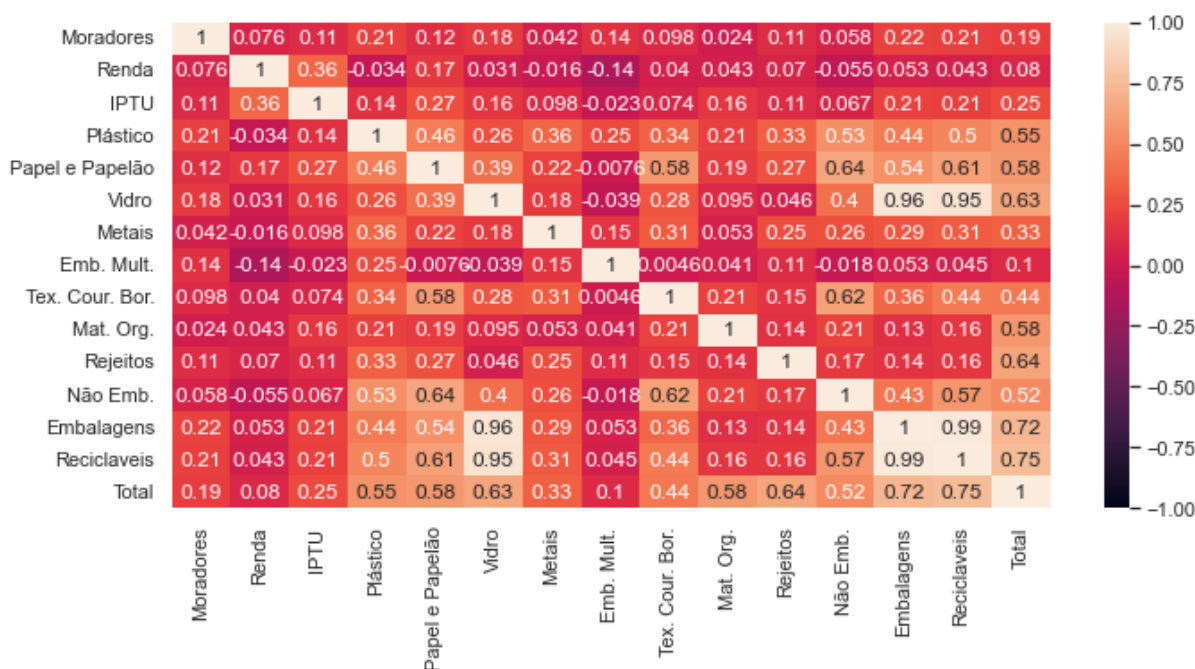
Uma vez que o banco de dados esteja ajustado, existem diferentes técnicas que podem ser utilizadas para analisar os dados. Um exemplo de técnica utilizada na análise exploratória são as estatísticas descritivas, como a análise univariada, que descreve, por exemplo a medida central, ou a dispersão de uma única variável.

Outra técnica muito utilizada quando se realiza a análise de um conjunto de dados é a visualização de dados. Esta, por sua vez, permite examinar diversas características dos dados de uma forma gráfica, o que facilita a identificação de informações sobre o conjunto e também a validação de certas propriedades.

Sendo assim, foi realizada a análise da correlação entre todos os dados do conjunto. Essa análise possibilita que seja verificado o grau de dependência entre as variáveis. A Figura 14 apresenta um gráfico com os coeficientes de correlação de Pearson, que podem ser interpretados da seguinte forma:

- Correlação maior que zero: as variáveis estão diretamente relacionadas
- Correlação menor que zero: as variáveis estão inversamente relacionadas
- Correlação igual a zero: não existe dependência linear entre elas

Figura 14 – Correlação entre os dados



Fonte: Elaborado pelo autor, 2022

É importante destacar que, embora a dependência linear não exista quando a correlação é igual a zero, não significa que as variáveis não tenham algum tipo de dependência não-linear entre elas. Outra característica importante para se analisar quanto ao valor dos coeficientes é a magnitude da correlação, sendo que quanto maior o valor (positivamente ou negativamente) maior a magnitude da relação entre as variáveis.

De acordo com a Figura 14 percebe-se que a magnitude da correlação é baixa entre os diferentes atributos do conjunto de dados, com exceção da correlação entre os materiais, ou os tipos de resíduos, e a sua produção total.

4.2.1 Preparação dos dados

Os dados coletados foram registrados em diferentes planilhas, e correspondem à períodos diferentes da pesquisa de campo. Em um primeiro momento as planilhas foram agrupadas de acordo com os levantamentos realizados, e os valores da pesagem foram divididos pela quantidade de dias correspondentes aos períodos em que os moradores armazenaram os resíduos. A combinação das planilhas resultou em apenas um arquivo, contendo a geração por dia de resíduos de 158 domicílios.

A partir disso, a planilha foi carregada em um *script* (conjunto de instruções para que uma função seja executada) em Python desenvolvido para realizar a preparação dos dados. Esse *script* juntamente com demais arquivos utilizados nessa etapa podem ser encontrados no Apêndice A.1.

Logo, foi utilizado o Excel para visualizar os dados adquiridos na pesquisa de campo. Percebeu-se que alguns moradores não preencheram o campo destinado ao valor do IPTU. Alguns dos motivos foram a isenção do imposto, ou pelo fato de o morador não lembrar do valor cobrado no momento do preenchimento dessa informação.

Para solucionar esse problema, foi utilizado uma função da biblioteca Pandas para o Python, que realiza a interpolação linear de dados faltantes. A interpolação linear é um método para deduzir um valor entre um conjunto de valores de acordo com o seu contexto. A Equação 4.3 abaixo, apresenta o cálculo da interpolação linear entre dois pontos $A(x_0, y_0)$ e $B(x_1, y_1)$.

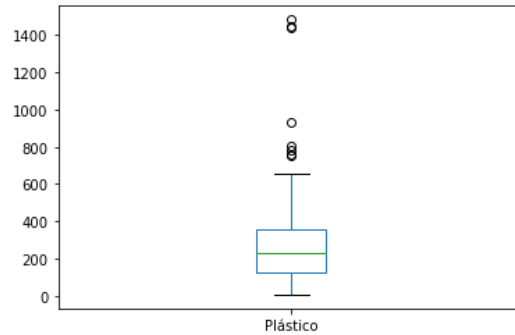
$$y = y_0 + (y_1 - y_0) \cdot \frac{x - x_0}{x_1 - x_0} \quad (4.3)$$

Onde, x e y são os valores resultantes da interpolação.

Após o preenchimento dos valores faltantes do IPTU, foi desenvolvido o diagrama de caixa dos dados (apresentado por John W. Tukey em 1969). Para exemplificar, a Figura 15 apresenta o diagrama referente a apenas um material, no caso o plástico. Esse

diagrama foi utilizado para realizar a remoção de valores discrepantes (do inglês, *outliers*) dos dados.

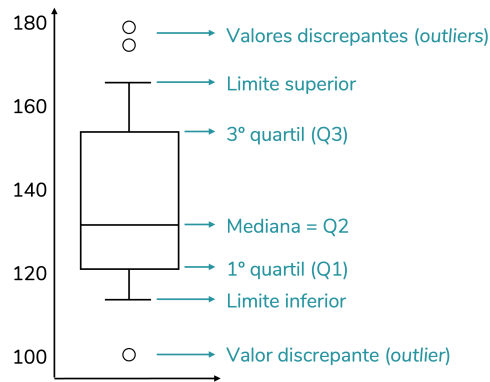
Figura 15 – Diagrama de caixa para os plásticos



Fonte: Elaborado pelo autor, 2022

Primeiramente identificou-se os quartis do diagrama de caixa, onde o primeiro (Q1) compreende 25% dos dados, o segundo (Q2) é equivalente à mediana do conjunto, e o terceiro (Q3) corresponde à 25% dos dados à partir da mediana. Além disso, a diferença entre o Q3 e o Q1 é identificada como amplitude interquartil. A Figura 16 ilustra os componentes do diagrama de caixa.

Figura 16 – Divisão do diagrama de caixa



Fonte: Fernanda Peres, 2022

Então, foi calculado o limite inferior e superior do diagrama (correspondentes ao menor e maior valor do conjunto de dados que não sejam outliers), utilizando a Equação 4.4 e Equação 4.5, respectivamente:

$$L_i = Q_1 - 1,5 \cdot (Q_3 - Q_1) \quad (4.4)$$

$$L_s = Q_3 + 1,5 \cdot (Q_3 - Q_1) \quad (4.5)$$

Portanto, os domicílios que tinham o valor dos materiais abaixo ou acima dos limites inferior e superior do diagrama de caixa foram considerados outliers, pois seus

valores eram discrepantes do restante do conjunto. Dessa forma foram identificados 63 outliers analisando todos os respectivos tipos de resíduos sólidos nos dados coletados.

Devido ao grande número de exemplos que seriam removidos do conjunto de dados, foi necessário encontrar um outro modo de remover os outliers de forma mais flexível. Logo, foi utilizado o kNN para realizar a detecção de outliers por conta dos seguintes pontos:

- Simplicidade: é um algoritmo simples de entender e implementar.
- Não paramétrico: é um algoritmo não paramétrico, o que significa que ele não faz nenhuma suposição sobre a distribuição dos dados.
- Robustez: é um algoritmo robusto que pode lidar com ruído e dados ausentes sem afetar significativamente sua capacidade de detectar outliers.

O funcionamento do kNN para detecção de outliers é relativamente simples. Para cada valor de um determinado material no conjunto de dados, o algoritmo encontra os K valores mais próximos e calcula a distância média desses valores ao que está sendo analisado. Os valores com as maiores distâncias médias são considerados outliers.

Logo, o conjunto de dados resultante contemplou 142 domicílios, 16 dos 158 foram classificados como outliers, divididos entre dados de treinamento e dados de teste, para a realização da validação cruzada, técnica utilizada durante o treinamento dos modelos de aprendizagem de máquina para melhorar a representatividade dos conjuntos e a capacidade de generalização.

Os dados de teste geralmente costumam representar de 20 à 30% do conjunto de dados. Como a quantidade de domicílios não era tão grande, foi selecionado a proporção de 20% dos dados para realizar os testes, possibilitando uma quantidade maior de domicílios para treinamento, e estes foram utilizados no desenvolvimento dos modelos apresentados no capítulo seguinte.

4.3 Modelagem

Existem diferentes maneiras de estimar as taxas de geração de resíduos sólidos. Segundo [Jalali e Nouri \(2008\)](#), com base em alguns elementos como população e fatores socioeconômicos, é possível calcular o coeficiente de geração de resíduos por pessoa.

Modelos baseados em inteligência artificial têm melhores habilidades de previsão do que outros modelos que utilizam regressão linear, de acordo com [Soni et al. \(2019\)](#). Alguns dos modelos desenvolvidos neste trabalho, utilizam redes neurais artificiais com o objetivo de realizar a predição da caracterização dos resíduos sólidos com base nos dados

socioeconômicos, como a renda domiciliar, quantidade de moradores e valor do imposto predial e territorial urbano (IPTU).

Outros modelos, que alcançaram melhores resultados, buscam realizar a classificação dos domicílios entre diferentes classes, e a partir do resultado da classificação utilizam a caracterização dos resíduos de cada classe para inferir a geração dos mesmos.

Durante a realização desse trabalho, os modelos foram desenvolvidos através do Spyder (um ambiente de desenvolvimento integrado, do inglês *Integrated Development Environment* - IDE, multiplataforma para programação na linguagem Python), e seus resultados foram analisados quanto à precisão e taxa de acerto para futuramente ser inserido no sistema de apoio à decisão.

A matriz de confusão foi utilizada para comparar os modelos desenvolvidos. Ela é uma ferramenta comumente usada para avaliar a precisão de um modelo de classificação através de uma tabela que mostra a frequência com que as amostras foram classificadas em cada classe do modelo em relação a classe real a que pertencem.

Supondo um modelo de classificação com 3 classes possíveis (A, B e C), a matriz de confusão para esse modelo teria a forma apresentada na Figura 17.

Figura 17 – Matriz de confusão

Matriz de Confusão			
A B C	A	B	C
	TP	FN	FN
	FP	TP	FN
A B C	FP	FP	TP
	A	B	C

Fonte: Fernanda Peres, 2022

Onde:

- TP (*True Positive*) representa o número de amostras que foram classificadas corretamente como pertencentes à classe A, B e C.
- FP (*False Positive*) representa o número de amostras que foram classificadas incorretamente como pertencentes a classe A, B e C, respectivamente.

- FN (*False Negative*) representa o número de amostras que foram erroneamente classificadas como não pertencentes a uma classe.

A diagonal principal representa as classificações corretas, enquanto as células fora da diagonal principal indicam classificações incorretas. A soma das células na linha de uma classe representa quantas vezes essa classe apareceu no conjunto de dados e a soma das células na coluna de uma classe representa quantas vezes o modelo a classificou nessa classe.

A partir da matriz de confusão, é possível calcular várias métricas de avaliação do modelo, como a acurácia, a precisão e o recall. A acurácia é a proporção de amostras classificadas corretamente pelo modelo, enquanto a precisão mede a proporção de amostras classificadas como positivas que realmente pertencem à classe positiva e o recall é uma medida que indica a proporção de exemplos positivos que foram corretamente identificados pelo modelo em relação ao número total de exemplos positivos na base de dados.

4.4 Sistema de apoio à decisão

Um sistema de apoio à decisão bem projetado, segundo [Rybnytska et al. \(2018\)](#) é baseado em aprendizado de máquina avançado e em sistemas baseados em conhecimento, para usar, classificar os dados e resolver os problemas, bem como aumentar a precisão do processo de decisão.

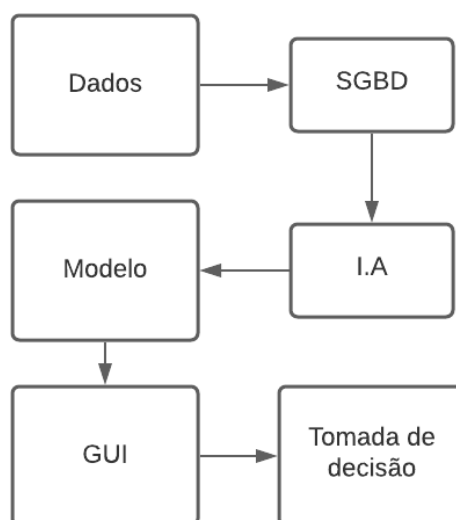
Algumas características importantes de um SAD bem desenvolvido, conforme [Rybnytska et al. \(2018\)](#):

- O SAD deve ser amigável e fácil para os usuários realizarem tarefas básicas.
- O design do SAD deve consistir em várias ferramentas e técnicas avançadas para acessar facilmente as informações.
- A função do SAD deve ser simples para os usuários finais.
- O SAD deve ser destacado pela adotabilidade e flexibilidade frente à um ambiente de incertezas.

Percebe-se que a usabilidade é o atributo principal a ser trabalhado em um sistema de apoio à decisão para garantir que o usuário tenha sucesso na execução de suas tarefas. Além disso, a confiabilidade e o desempenho devem ser características que refletem as exigências do processo, nesse caso relacionado ao gerenciamento de resíduos sólidos.

A arquitetura proposta para o sistema de apoio à decisão é apresentada na Figura 18 e consiste em três funcionalidades principais, sendo elas o armazenamento dos dados, a análise/modelagem dos dados e a apresentação das informações.

Figura 18 – Arquitetura de um sistema de apoio à decisão.



Fonte: Elaborado pelo autor, 2022

O armazenamento das informações geralmente é realizado através de um [sistema de gerenciamento de banco de dados \(SGBD\)](#) que possui a responsabilidade de gerenciar o acesso, a persistência, a manipulação e a organização dos dados.

A análise de dados consiste na utilização de várias técnicas baseadas em estatística e inteligência artificial aplicadas a diversos cenários abordados pelo SAD para projetar modelos que auxiliarão no processo de tomada de decisão.

O SGBD armazenará os dados enviados pelo usuário. Esses dados serão submetidos ao algoritmo de inteligência artificial, desenvolvido previamente. Em seguida, o modelo de estimativa desenvolvido utiliza os dados do usuário para realizar a predição. O resultado é então organizado e apresentado ao usuário pela interface gráfica.

Por último, a apresentação das informações por meio de uma interface gráfica é a funcionalidade responsável por tornar o sistema adotável. Ela deve ser capaz de permitir a interação do usuário com os outros componentes do sistema.

Dentre as várias tecnologias disponíveis para o desenvolvimento de um sistema de apoio à decisão a escolhida para esse trabalho é a aplicação web, por ser a mais acessível e escalável em relação à expansão estratégica do sistema. Um outro ponto importante ao se utilizar uma plataforma web é a facilidade de desenvolvimento e manutenção, sendo que existem várias linguagens de programação suportadas pelos navegadores atuais.

5 Resultados

Após o desenvolvimento da pesquisa de campo, e com os dados referentes a geração e caracterização de resíduos sólidos domiciliares, bem como os dados socioeconômicos levantados já preparados, foi possível o desenvolvimento de modelos baseados em inteligência artificial para realizar a predição e generalizar a caracterização dos resíduos.

5.1 Modelos de regressão

Foram criados diferentes modelos de regressão utilizando redes neurais, variando entre eles os parâmetros da própria rede, assim como utilizando combinações diferentes de preditores para buscar um resultado razoável para a saída.

De acordo com a própria natureza do problema, que consiste na predição da geração de resíduos sólidos domiciliares de um município com base em dados socioeconômicos, a modelagem através de uma rede neural considera que as variáveis preditoras e a saída da rede sejam contínuas, para que possa se estabelecer uma concordância na estimação da saída.

Dentre os modelos de regressão desenvolvidos, o que apresentou melhor resultado para a generalização do total de resíduos foi construído sob uma estrutura de rede neural artificial, conforme explicado na seção 3.1.1, com duas camadas ocultas, ambas com dez neurônios artificiais, uma camada de entrada com três neurônios, de acordo com o número de variáveis de entrada da rede, e uma camada de saída, com um neurônio.

Nas camadas ocultas da rede neural foi utilizada a função de ativação sigmoide (não-linear, e com saída entre 0 e 1), enquanto que na camada de saída foi utilizada a função linear, devido ao fato desta não alterar a saída do neurônio. A Figura 19 apresenta a rede neural desenvolvida:

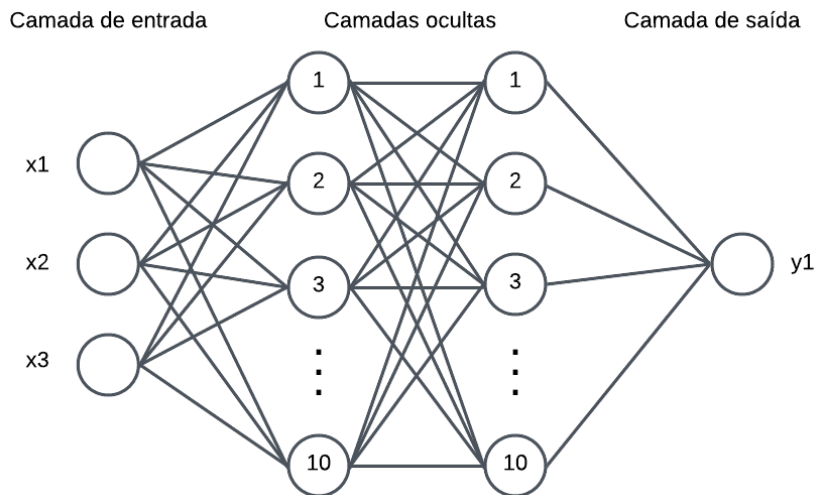
Além da remoção de *outliers*, apresentada na seção 4.2.1, com a intenção de que todas as variáveis preditores recebessem a mesma importância durante o processo de aprendizagem do modelo, também foi realizado a normalização dos dados entre 0,1 e 0,9.

Os seguintes parâmetros foram utilizados no desenvolvimento da rede:

- Otimizador: Adam;
- Função de perda: MSE;
- Número de épocas: 50.000;

- Número de exemplos por iteração (batch size): 100;
- Atributos previsores: Moradores, Renda e IPTU.

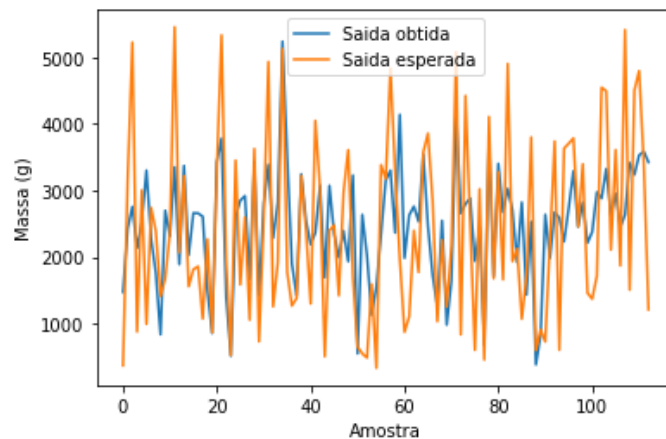
Figura 19 – Rede neural densa.



Fonte: Elaborado pelo autor, 2022

A Figura 20 apresenta o resultado do treinamento da rede, e a Figura 21 demonstra a saída da rede quando os dados de teste são utilizados para realizar a predição da geração total de resíduos.

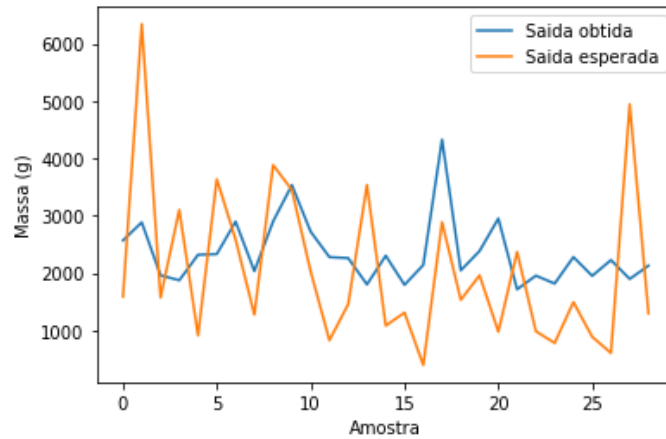
Figura 20 – Resultado do treinamento da rede



Fonte: Elaborado pelo autor, 2022

Observa-se que o modelo desenvolvido sofre com o sobreajuste, ou seja, o modelo se adapta muito bem ao conjunto de treinamento, porém se mostra ineficiente quanto à capacidade de realizar generalizações. Mesmo realizando mudanças em cada um dos parâmetros essa característica se mantém (até mesmo com a diminuição da quantidade de épocas no treinamento).

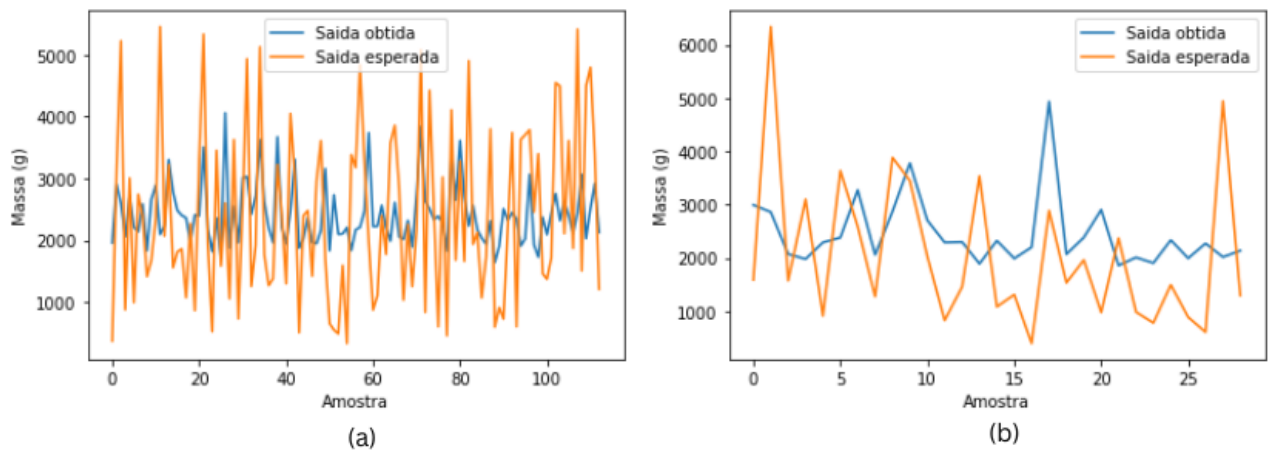
Figura 21 – Saída da rede para os dados de teste



Fonte: Elaborado pelo autor, 2022

A Figura 22 mostra o resultado do modelo utilizando 10.000 épocas no processo de treinamento (a) e teste (b), desenvolvido com a intenção de evitar o sobreajuste. Percebe-se que mesmo com a diminuição na quantidade de épocas na fase de treinamento, não houve melhora no desempenho do modelo.

Figura 22 – Saída da rede para uma quantidade menor de épocas



Fonte: Elaborado pelo autor, 2022

Dessa maneira, a utilização de um modelo baseado em regressão não foi adequado ao problema, possivelmente devido à fatores aleatórios (não identificados) presentes nos dados. Logo, seria necessário um número maior de amostras para que se pudesse analisar melhor esse comportamento.

Além disso, depois de testado vários modelos de regressão, percebeu-se que para esse problema a granularidade nos dados de entrada não representava uma granularidade na saída. Logo, fez-se necessário uma outra abordagem para o problema.

5.2 Modelos de classificação

A partir dos resultados alcançados pelos modelos de regressão, tornou-se necessário encontrar outra solução para estimar a geração dos resíduos sólidos domiciliares utilizando os dados obtidos através da pesquisa de campo. Foram desenvolvidos então, quatro modelos de classificação através das seguintes abordagens: Árvore de decisão, SVM, KNN e Naive Bayes.

Uma tarefa importante no momento de dimensionar um problema para o desenvolvimento de um modelo de classificação competente é determinar o número de classes. Para isso pode ser utilizado um histograma, uma representação gráfica de um conjunto de dados dividida em classes ou frequências.

Contudo, selecionar aleatoriamente uma quantidade de classes e desenhar o gráfico do histograma até encontrar uma parcela que se ajuste ao conjunto pode ser trabalhoso, além de ineficiente. Para contornar isso, existem algumas fórmulas para o cálculo da quantidade de classes, dentre elas a mais efetiva é a Regra de Sturges, definida conforme a Equação 5.2:

$$k = 1 + 3,322 \cdot \log N \quad (5.1)$$

Onde N é a quantidade de amostras e k é o número de classes

A partir dessa regra, calcula-se o intervalo de cada classe através da Equação 5.2:

$$h = \frac{H}{k} \quad (5.2)$$

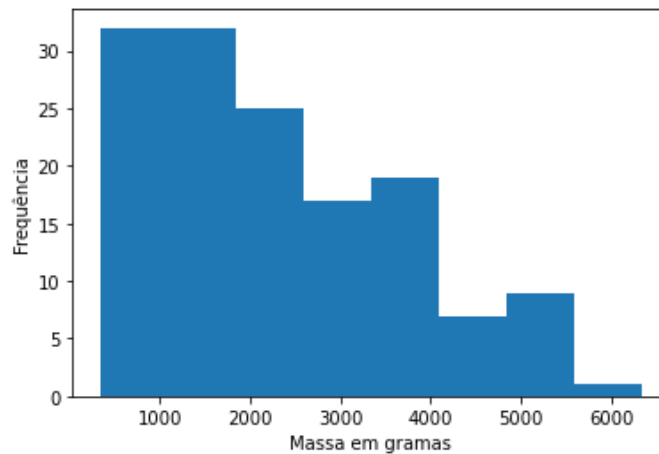
Onde h é a amplitude do intervalo e H é a amplitude total (subtração entre o maior e o menor valor do conjunto de dados).

Aplicando essa regra, foi encontrado os seguintes valores de classe para o conjunto de dados contendo o total de resíduos gerados:

- $k = 8,15$
- $H = 6.007$
- $h = 737,06$

A Figura 23 apresenta o histograma gerado a partir do número de classes e da amplitude do intervalo calculados acima:

Figura 23 – Histograma com intervalos iguais



Fonte: Elaborado pelo autor, 2022

Percebe-se que essa representação das classes não possibilita a identificação do real comportamento das médias da geração de resíduos, bem como a dispersão dos valores, variância e outras propriedades. Sendo assim, não é possível utilizar intervalos igualmente distribuídos, pois os mesmos podem levar à diferentes interpretações dos dados.

Uma outra forma, mais adequada para abordar o problema em questão, é a utilização da [função densidade de probabilidade \(FDP\)](#). Essa função relaciona uma variável aleatória com a sua probabilidade de modo que seja possível identificar a probabilidade de qualquer valor dentro do conjunto de dados e, portanto, estimar a quantidade de classes baseada nessa densidade.

Para isso, primeiro é necessário definir se a distribuição do conjunto de dados é conhecida. Com esse propósito pode ser realizado diferentes plotes de histogramas variando o número de classes. Além do que, é possível que os dados correspondam à uma distribuição já conhecida, sendo necessário apenas a realização de alguma transformação nos dados.

No conjunto utilizado, foi possível identificar que os dados variam muito em relação à uma mesma variável, sendo mais difícil encontrar uma transformação que o aproxime de uma distribuição conhecida. Logo, foi utilizado uma abordagem não paramétrica para estimar a função densidade de probabilidade.

A abordagem utilizada foi a estimação da densidade de kernel, que retorna uma probabilidade para um determinado valor do conjunto de dados realizando a interpolação das probabilidades de modo que a soma das probabilidades seja igual a 1.

Um parâmetro, conhecido como largura de banda, é utilizado para limitar a janela de observações para estimar a probabilidade de uma determinada amostra do conjunto de dados. Ele deve ser escolhido de maneira a representar da melhor forma a granularidade nos dados. Além desse parâmetro, também é necessário definir qual será a função

base do kernel, como a função gaussiana apresentada na Equação 5.3 (que descreve uma distribuição normal quando a média é 0 e o desvio padrão é unitário).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (5.3)$$

Onde μ é a média da distribuição, σ é o desvio padrão e e corresponde ao número de Euler.

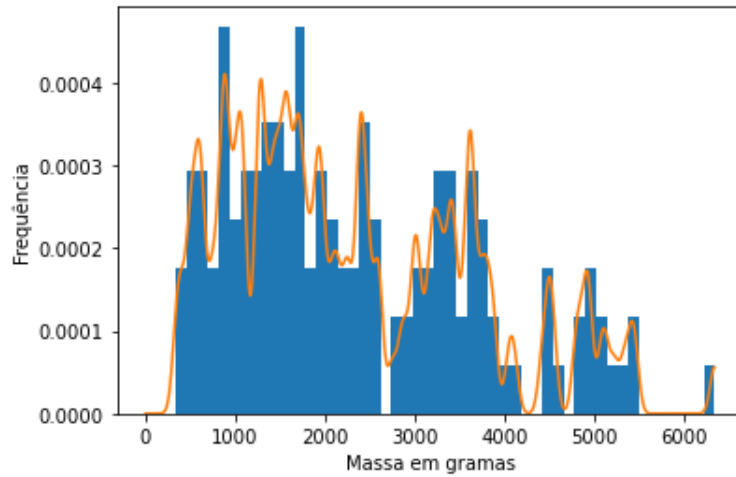
Desta maneira, sendo (x_1, x_2, x_n) os dados do conjunto, estimou-se a densidade de probabilidade não paramétrica através da Equação 5.5:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) \quad (5.4)$$

Onde K é a função de gauss e h a largura de banda (igual a 50, escolhida de acordo com a variância dos dados através de experimentação).

Logo, foi utilizado a função densidade de probabilidade em conjunto com o histograma para gerar o gráfico da Figura 24:

Figura 24 – Histograma combinado com a FDP



Fonte: Elaborado pelo autor, 2022

Verifica-se que existem três aglomerados (do inglês, *clusters*) onde a função densidade de probabilidade se mantém maior que 0. Sendo assim, cada um desses *clusters* correspondem a uma classe, e o intervalo de cada uma delas é definido pelos valores das amostras onde a função densidade de probabilidade é equivalente aos menores mínimos locais, em conformidade com os intervalos do histograma.

Então, as classes utilizadas nos modelos de classificação desenvolvidos e seus limites foram estabelecidos como:

- Classe 1 = $[0, 2.700[$
- Classe 2 = $[2.700, 4.265[$
- Classe 3 = $[4.265, m]$

Onde m é o maior valor do conjunto de dados, 6.342.

A partir do conjunto de dados e da determinação das classes, verificou-se as proporções, conforme a Tabela 1.

Tabela 1 – Quantidade de amostras por classe

Classe	Amostras
1	91
2	35
3	16

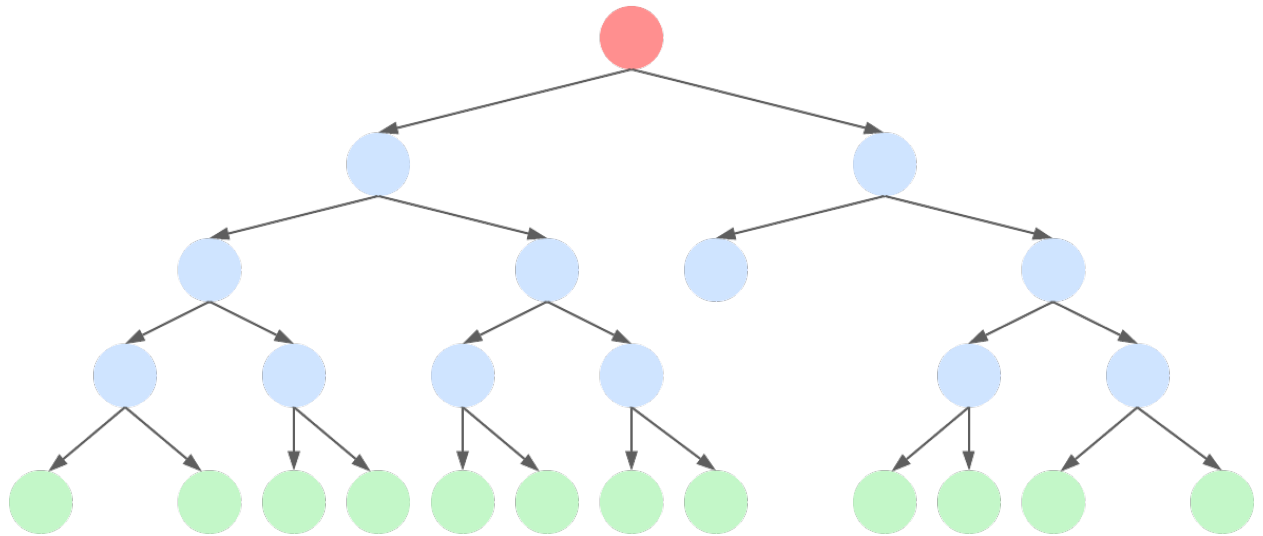
Fonte: Elaborado pelo autor, 2022

É válido observar que como os conjuntos de dados selecionados para a validação cruzada são escolhidos de maneira aleatória, dependendo da proporção entre dados de treinamento e de teste, o resultado do modelo pode ser heterogêneo, ou seja, em um dado momento o modelo vai apresentar resultados melhores que em outro, para conjuntos de dados de teste diferentes. Nessa seção serão apresentados apenas os resultados dos melhores casos, em que os modelos tiveram melhor desempenho.

Utilizando o algoritmo da árvore de decisão, explicado anteriormente no Capítulo 3, foi definido como parâmetros a profundidade máxima da árvore em quatro nós, e então o modelo foi treinado utilizando o mesmo conjunto de dados de treinamento dos modelos anteriormente desenvolvidos.

A estrutura modelada pode ser verificada na Figura 25. Para cada nó interno da árvore, representado pela cor azul, existe uma condição a ser testada, no caso, se o valor do IPTU, atributo previsor, é menor ou igual a um determinado valor. Os nós que não possuem essa condição são considerados nós folha, representados pela cor verde.

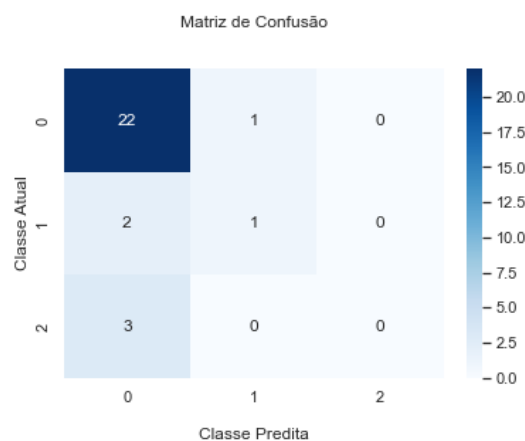
Figura 25 – Árvore de decisão modelada



Fonte: Elaborado pelo autor, 2022

Após o treinamento, o modelo foi utilizado para realizar a predição do conjunto de testes, onde obteve-se os resultados ilustrados pela matriz de confusão da Figura 26. Note que o modelo atingiu uma precisão de 79,31%, acertando 23 das 29 amostras totais. Isso pode ser verificado analisando os valores da diagonal principal da tabela, que correspondem às previsões corretas do classificador.

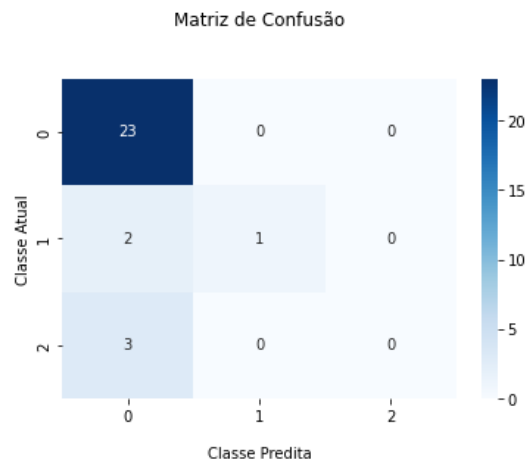
Figura 26 – Matriz de confusão para modelo baseado em árvore de decisão



Fonte: Elaborado pelo autor, 2022

Outro modelo de classificação desenvolvido, utilizou o algoritmo Naive Bayes e atingiu uma precisão superior à da árvore de decisão (82,75%). A matriz de confusão resultante pode ser verificada através da Figura 27. Percebe-se que no melhor caso apurado a matriz é muito parecida com a do modelo da árvore de decisão, se diferenciando apenas no caso em que houve um erro de classificação para a primeira classe, identificada na matriz como a de número zero.

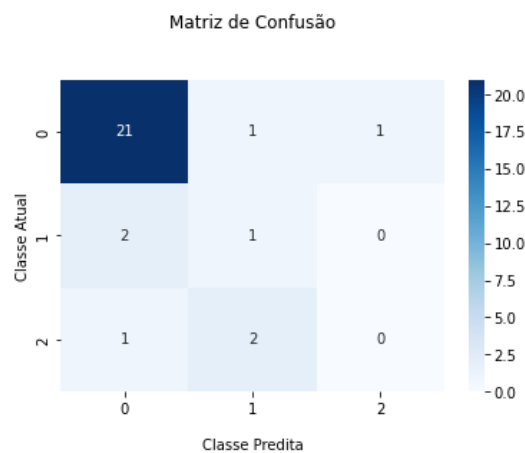
Figura 27 – Matriz de confusão para modelo baseado em naive bayes



Fonte: Elaborado pelo autor, 2022

Também foi construído um modelo baseado no classificador kNN, com o parâmetro $k = 5$ vizinhos. A precisão apresentada por esse modelo foi de 75,86% acertando 22 amostras, conforme Figura 28.

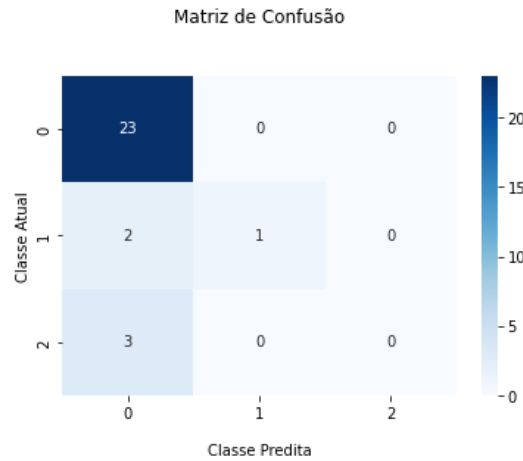
Figura 28 – Matriz de confusão para modelo baseado no knn



Fonte: Elaborado pelo autor, 2022

Por fim, foi desenvolvido um modelo baseado no SVM adotando a função de base radial como kernel, e o valor 0,5 para o parâmetro de otimização 'C'. Conforme a Figura 29, a precisão também foi de 82,75%, acertando 24 amostras, assim como o modelo que utiliza o algoritmo Naive Bayes.

Figura 29 – Matriz de confusão para modelo baseado no svm



Fonte: Elaborado pelo autor, 2022

Com base nos modelos desenvolvidos, nota-se que eles são capazes de realizar uma inferência sobre a classificação de um dado domicílio de maneira satisfatória, tendo em vista a baixa correlação entre as variáveis preditoras e a saída do modelo, sendo que, de acordo com os resultados, a melhor precisão atingida ultrapassou 82%.

Para finalizar, após a classificação do domicílio, foi calculado a média de todos os tipos de materiais por classe, e para garantir que a média fosse realmente representativa da população, foram calculados os intervalos para uma probabilidade de 95% de confiança.

Logo, a Tabela 2 apresenta o percentual de erro, com relação à média de resíduos produzidos por domicílio, calculado à partir da margem de erro para cada tipo de resíduo das diferentes classes de domicílios. Com base nelas, pode-se calcular os intervalos de confiança conforme a Equação 5.5 e o percentual de erro, segundo a Equação 5.6:

$$I_c = \mu \pm e \quad (5.5)$$

$$p_e = \frac{e}{\mu} \quad (5.6)$$

Onde μ é a média e e é a margem de erro.

Tabela 2 – Margem de erro para os tipos de resíduos

Classe	Plástico	Papel	Vidro	Metais	Emb.	Tex.	Org.	Rejeitos	Total
1	19,5%	20,1%	44,1%	34,6%	28,6%	76,2%	16,8%	24,0%	9,0%
2	18,2%	34,9%	43,2%	40,2%	56,4%	100%	15,7%	34,7%	3,4%
3	20,8%	39,8%	62,9%	72,9%	33,3%	88,9%	19,8%	46,2%	5,1%

Fonte: Elaborado pelo autor, 2022

Percebe-se que o percentual de erro é muito maior na classe 3, a qual possui apenas 16 amostras. Além disso, com exclusão da classe 3, a maior margem de erro fica por conta dos ‘têxteis, couros e borracha’, que chega em 100% na classe 2. Isso ocorre devido ao valor referente à esse material ser extremamente baixo, sendo apenas 11,05g a sua média. Além disso, a maioria dos domicílios não produzem esse tipo de resíduos.

Em geral, verifica-se que quanto menor o valor de cada material, maior é a porcentagem da margem de erro. Sendo assim, materiais como ‘Matéria orgânica’ e ‘Rejeitos’ possuem uma porcentagem menor referente à margem de erro, pois são os tipos de resíduos sólidos domiciliares mais produzidos.

De acordo com a complexidade do problema, as margens de erro encontradas foram consideradas aceitáveis, e a precisão pode ser melhorada conforme um número maior de domicílios forem adicionados à amostra, visto que o sistema desenvolvido comportará a adição posterior de mais domicílios, caso seja necessário aumentar a exatidão das inferências.

5.3 Escolha do modelo

O modelo utilizado para ser integrado ao sistema de apoio à decisão desenvolvido foi escolhido a partir da comparação direta entre os melhores e piores resultados dos modelos apresentados na subseção anterior, além da análise do valor de *F1Score*, uma métrica para avaliar o desempenho de modelos de classificação, especialmente quando as classes não têm o mesmo número de exemplos. O *F1Score* é calculado de acordo com a Equação 5.8, e utiliza para tal outra métrica conhecida como *Recall*, Equação 5.7, que mede a proporção de exemplos positivos corretamente identificados pelo modelo, em relação ao número total de exemplos positivos presentes nos dados.

$$Recall = \frac{Verdadeiros\ positivos(VP)}{Verdadeiros\ positivos(VP) + Falsos\ negativos(FN)} \quad (5.7)$$

$$F1 = \frac{2 * precisao * recall}{precisao + recall} \quad (5.8)$$

Sendo assim, os modelos foram executados 5 vezes consecutivas, e os resultados foram avaliados com base na precisão alcançada, e também na quantidade de vezes em que o modelo errou na classificação. Então, a Tabela 3 apresenta a comparação entre a precisão dos modelos de classificação desenvolvidos.

Tabela 3 – Precisão dos modelos de classificação

Algoritmo	1ª Exec.	2ª Exec.	3ª Exec.	4ª Exec.	5ª Exec.	Média
Árvore de Decisão	75,86%	58,62%	68,96%	75,51%	79,31%	71,65%
Naive Bayes	68,96%	72,41%	86,20%	68,96%	82,75%	75,85%
KNN	65,51%	62,06%	58,62%	62,06%	75,86%	64,82%
SVM	68,96%	75,86%	72,41%	58,62%	58,27%	66,82%

Fonte: Elaborado pelo autor, 2022

Percebe-se que os modelos baseados na árvore de decisão e no naive bayes foram o que atingiram a maior precisão na média. Porém como ambos os algoritmos utilizados foram muito próximos (variando em apenas 4,2% entre a maior e a menor precisão alcançada) deve-se analisar também os piores casos para escolher entre um deles.

Logo, a Tabela 4 apresenta o resultado da precisão nos piores casos (casos em que a divisão entre dados de treinamento e teste apresentou valores para os quais a etapa de treinamento não obteve resultados satisfatório) para os mesmos parâmetros utilizados anteriormente, variando apenas a aleatoriedade na seleção dos conjuntos de treinamento e teste dos modelos.

Tabela 4 – Precisão dos modelos nos piores casos

Algoritmo	Precisão	Acertos	Erros
Árvore de Decisão	55,17%	16	13
Naive Bayes	62,07%	18	11
KNN	48,27%	14	15
SVM	51,72%	15	14

Fonte: Elaborado pelo autor, 2022

É possível verificar que os modelos que utilizam o naive bayes e a árvore de decisão atingiram uma maior precisão nos piores casos, de 62,07% e 55,17% respectivamente. Portanto, comparando o valor médio da precisão na execução dos modelos juntamente com a melhor precisão nos piores casos para analisar os modelos, verifica-se que os dois alcançaram um bom resultado, quando comparados com os demais modelos desenvolvidos.

Diante disso, a Tabela 5 apresenta os valores de *F1 Score* para todos os modelos de classificação desenvolvidos, proporcionando mais um parâmetro para a definição do melhor modelo.

Tabela 5 – Valores de F1 Score

Algoritmo	F1 Score
Árvore de Decisão	0,7393
Naive Bayes	0,7670
KNN	0,7382
SVM	0,7670

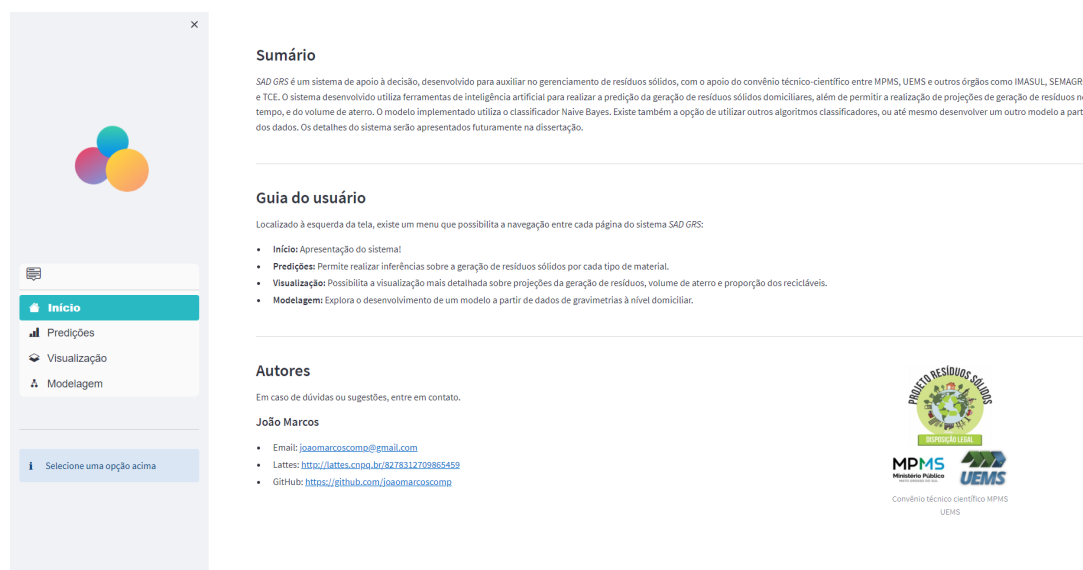
Fonte: Elaborado pelo autor, 2022

Analisando os resultados alcançados pelos modelos, os baseados no naive bayes e na árvore de decisão poderiam ser implementados, com resultados muito semelhantes em relação à precisão. Porém, foi decidido incorporar ao sistema de apoio à decisão o modelo de classificação naive bayes, por possuir o maior valor de $F1\ Score$ e pela sua vantagem em poder escalar linearmente de acordo com a dimensão do conjunto de dados, isto é, como a complexidade da sua computação é resumida apenas na multiplicação das probabilidades, o algoritmo é capaz de manipular um conjunto maior de atributos de maneira mais eficiente que a árvore de decisão.

5.4 Sistema de apoio à decisão

Após o modelo de classificação ser selecionado, foi desenvolvido um sistema de apoio à decisão, como descrito na Seção 4.4, com o objetivo de proporcionar uma interface de usuário para a visualização das estimativas da geração e classificação dos resíduos sólidos, bem como possibilitar a realização de projeções sobre as informações fornecidas pelo modelo. A Figura 30 mostra a tela inicial do SAD. Ela apresenta um sumário explicando o objetivo do sistema e um guia rápido para auxiliar os usuários na usabilidade.

Figura 30 – Tela inicial do sistema de apoio à decisão



Fonte: Elaborado pelo autor, 2022

A interface foi desenvolvida utilizando um framework (estrutura de software que possibilita o desenvolvimento de forma simplificada, agrupando pedaços de código escritos para realizar uma funcionalidade específica) muito popular que utiliza bibliotecas de código aberto em Python para proporcionar um ambiente em que cientistas de dados possam colocar seus sistemas em produção.

A escolha do framework se deu devido à sua popularidade, que simplifica a manutenção do sistema e na facilidade ao implementar novos componentes. Além disso, por ser

desenvolvido em Python, tal como os modelos apresentados anteriormente, a integração entre eles se resume na adaptação do código fonte utilizado na construção dos modelos.

Com base nisso, foram construídos três componentes principais para o SAD:

1. Predição
2. Visualização
3. Modelagem

Cada um dos componentes foi dividido em páginas diferentes dentro do sistema. A página destinada à predição (Figura 31) é a principal. Nela o usuário pode selecionar o modelo desejado para realizar as estimativas, assim como enviar o arquivo contendo a relação dos IPTU's para o cálculo das inferências. Esse arquivo é necessário pois os modelos de classificação desenvolvidos consideram o valor do IPTU como atributo previsor.

Após o cálculo das estimativas, o resultado fica armazenado em disco para futuramente ser utilizado para realizar as diferentes projeções que o usuário possa solicitar através do componente de visualização.

Figura 31 – Página de predição do sistema

Estimativa da geração de resíduos

Informe os respectivos dados para realizar a predição:

Selecione o município
Campo Grande

Selecione o modelo
Classificacao.sav

Enviar relação dos domicílios

Selecione o arquivo contendo a relação com os valores de IPTU para cada domicílio do município:

Enviar arquivo .XLSX

Drag and drop file here
Limit: 200MB per file - XLSX

Browse files

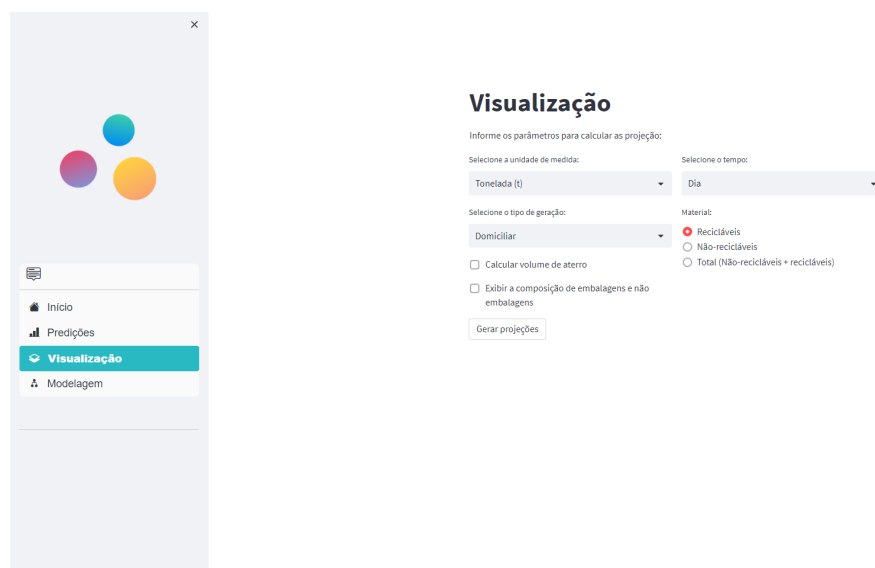
☐ Usar relação padrão (Campo Grande)

Fonte: Elaborado pelo autor, 2022

Logo, a página exibida na Figura 32, permite que o usuário selecione o tipo de geração (domiciliar ou por pessoa), a unidade de medida, a quantidade de tempo que deseja visualizar as projeções, os tipos de materiais (reciclável, não reciclável ou ambos) e também possibilita o cálculo do volume de aterro.

É possível verificar a tela de projeção dos dados na Figura 33. Nela o utilizador é apresentado ao resultado da estimativa e aos gráficos de proporção dos resíduos e de

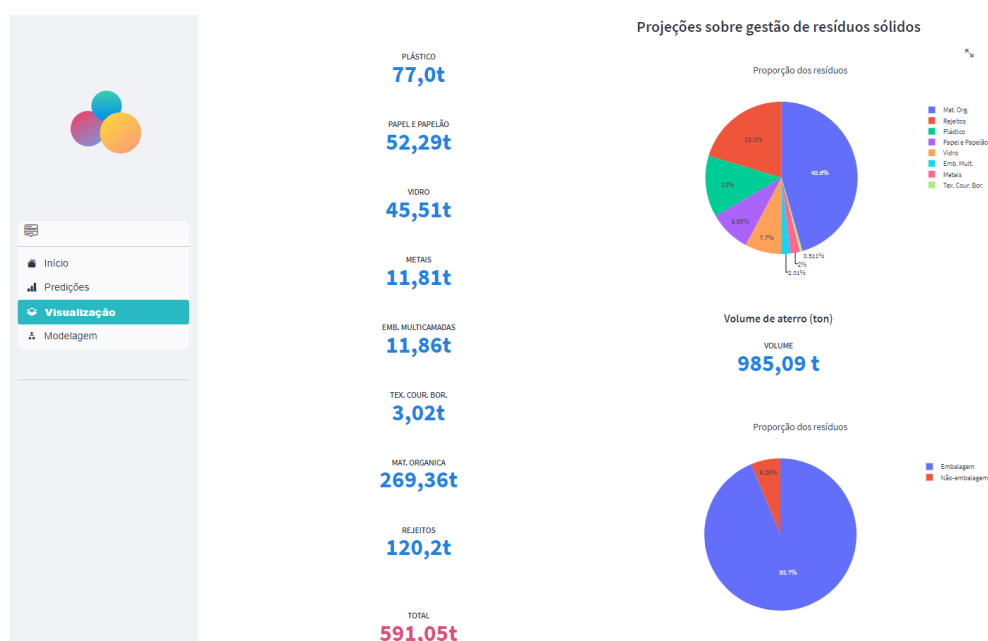
Figura 32 – Página de visualização do sistema



Fonte: Elaborado pelo autor, 2022

proporção de embalagens. Também está representado o volume de aterro estimado, para comportar a geração projetada.

Figura 33 – Página de visualização de projeções



Fonte: Elaborado pelo autor, 2022

Por fim, o componente de modelagem (que ainda está em desenvolvimento) permitirá ao utilizador enviar novos dados de geração de resíduos e configuração de parâmetros para que o sistema atualize o modelo de classificação, tornando-o ainda mais preciso.

Utilizando a relação padrão de Campo Grande (dados referentes ao valor do IPTU por domicílio para cada domicílio do município) e o modelo de classificação selecionado, o

sistema apresentou os valores estimados para a geração domiciliar de materiais (por dia) e seus intervalos de confiança conforme a Tabela 6.

Tabela 6 – Estimação da geração de resíduos

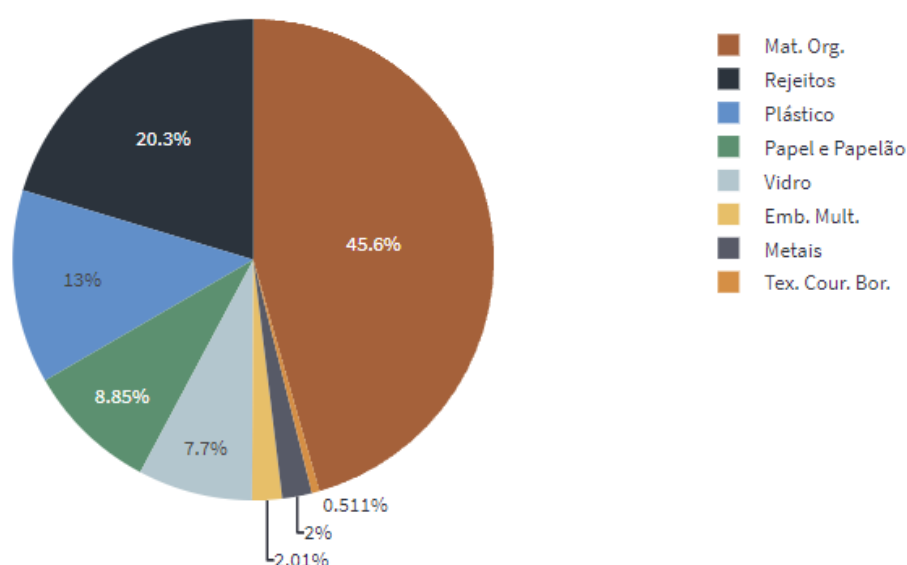
Tipo	Valor estimado (t)	Intervalos de confiança (t)
Plástico	64,45	51,89 — 77,01
Papel e papelão	43,28	34,27 — 52,29
Vidro	31,58	17,65 — 45,51
Metal	8,75	5,69 — 11,81
Emb. Mult.	9,13	6,41 — 11,85
Tex. Cour. Bor.	1,71	0,39 — 3,03
Mat. Orgânica	230,53	191,71 — 269,35
Rejeitos	96,52	72,84 — 120,2
Total	485,95	380,85 — 591,05

Fonte: Elaborado pelo autor, 2022

Os cálculos dos intervalos de confiança para a estimativa das médias de geração de resíduos foram realizados pelo sistema de acordo com a média ponderada das margens de erro. Sendo assim, os intervalos de confiança foram calculados considerando a quantidade de cada classe de domicílio que o modelo de classificação inferiu.

Ao acessar a página de visualização e selecionar ambos os materiais (reciclável e não reciclável) para gerar as projeções, pode-se verificar por meio do gráfico de setores, a proporção dos resíduos (conhecida também como composição gravimétrica) ilustrada na Figura 34.

Figura 34 – Proporção de resíduos sólidos domiciliares



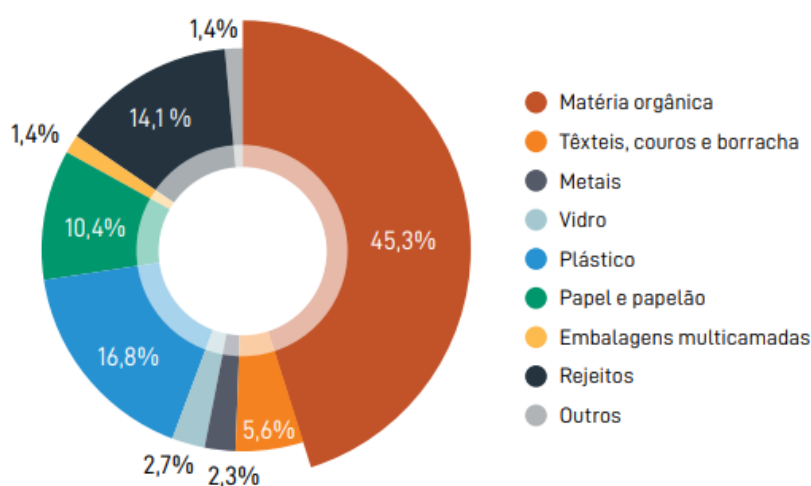
Fonte: Elaborado pelo autor, 2022

Percebe-se que a matéria orgânica constitui aproximadamente 45% do total de resíduos sólidos domiciliares gerados, seguida pelos rejeitos com proporção próxima à

20%. Plásticos, com 13% e, papel e papelão, com 8,85% ocupam o terceiro e quarto setor do gráfico, respectivamente.

Ao realizar a comparação da estimativa da composição dos resíduos de Campo Grande (MS) gerada pelo sistema com a composição derivada da ABRELPE (Figura 35), divulgada em 2020, verifica-se a correspondência da maioria dos tipos de materiais, onde a maior diferença é advinda dos rejeitos, que correspondem à apenas 14,1%. Essa diferença se deve principalmente ao fato de que na gravimetria realizada para a obtenção dos dados utilizados nesse trabalho, o tipo de material "Outros" e os "Rejeitos" foram ambos categorizados em um único tipo, "Rejeitos", enquanto que na composição apresentada pela ABRELPE existe a distinção entre esses tipos de materiais.

Figura 35 – Composição dos resíduos sólidos urbanos



Fonte: Panorama dos resíduos sólidos no Brasil - ABRELPE, 2020

É importante lembrar que a composição gravimétrica divulgada pela ABRELPE refere-se a uma estimativa baseada na média ponderada da geração de resíduos sólidos urbanos considerando a população e a renda per capita dos municípios brasileiros.

Além disso, deve-se levar em consideração as diferenças entre o cenário estadual e nacional, bem como a época em que os estudos foram realizados, e também os efeitos da pandemia de COVID-19 na geração de resíduos sólidos. Portanto, as divergências entre as composições são esperadas, e possibilitam estudos futuros sobre a influência dos resíduos sólidos domiciliares na geração de resíduos sólidos urbanos.

O sistema de apoio à decisão desenvolvido possibilita a realização de várias análises sobre a geração e gestão dos resíduos sólidos. Um outro exemplo é a comparação entre a composição de Campo Grande com um estudo de gravimetria realizado em Dourados (MS) onde, segundo [Souza et al. \(2022\)](#), a matéria orgânica corresponde à 53,1% dos resíduos sólidos urbanos.

Outras análises possíveis são referentes ao cálculo do volume de aterro necessário

para comportar os resíduos gerados pelos municípios. Esse cálculo é feito de acordo com a Equação 5.9:

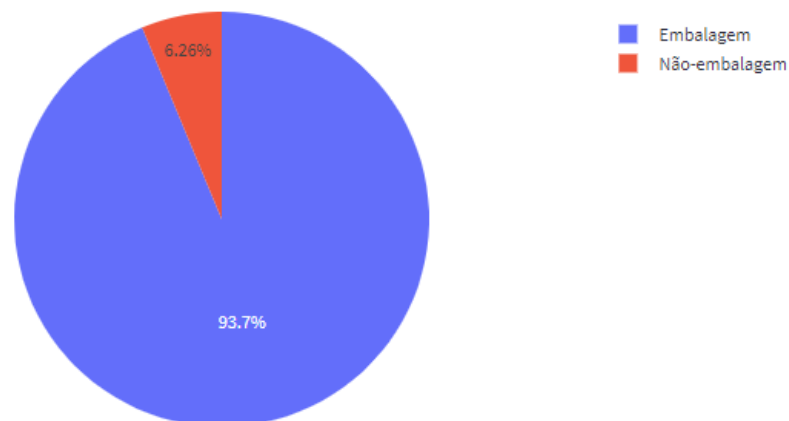
$$volume = \frac{P}{D} \quad (5.9)$$

Onde P é o peso dos resíduos (ton/ano) e D é a densidade compactada (ton/ m^3).

A densidade compactada de resíduos sólidos varia de 600 - 800 kg/ m^3 de acordo com os procedimentos de resíduos sólidos urbanos em IBRAOP (2020). Adotando a densidade mínima de 0,6 ton/ m^3 , o SAD estimou o volume de aterro em 359.557,5 toneladas.

Para terminar, o sistema também fornece a proporção estimada de embalagens e não embalagens para os resíduos sólidos domiciliares. A Figura 36 apresenta essa proporção para a cidade de Campo Grande.

Figura 36 – Proporção estimada de embalagens



Fonte: Elaborado pelo autor, 2022

Percebe-se que mais de 90% dos resíduos domiciliares secos produzidos são constituídos de embalagens. De acordo com a ABRELPE (2019), o índice de reciclagem do Brasil é de apenas 4%. Isso significa que muitas das embalagens que vão para o lixo não são reaproveitadas. Sendo assim, a logística reversa das embalagens é de fundamental importância para diminuir os impactos causados ao meio ambiente.

Para finalizar, pode-se comparar a geração de resíduos sólidos domiciliares inferida pelo sistema de apoio à decisão com um relatório de pesagem dos resíduos sólidos domiciliares da SOLURB (concessionária responsável pela gestão da limpeza urbana e o manejo de resíduos sólidos do município de Campo Grande). Segundo ela, em 2022 foram coletadas em média 760 toneladas de resíduos. Comparando com o valor total de geração estimado pelo SAD, considerando o limite superior do intervalo de confiança, 591 toneladas, verifica-se que a precisão atingida foi de aproximadamente 77,8%.

Tabela 7 – Média da geração de resíduos por classe

Classe	Plástico	Papel	Vidro	Metais	Emb.	Tex.	Org.	Rejeitos	Total
1	198,9	134,2	87,8	26,6	28,2	4,7	682,5	290,0	1.453,2
2	314,4	197,8	368,9	50,7	42,9	20,2	1.772,3	646,7	3.414,3
3	452,3	421,0	674,3	83,7	54,9	26,8	2.442,1	865,0	5.020,5
Total	255,9	182,2	223,1	39,0	34,8	11,0	1.149,4	442,7	2.338,5

Fonte: Elaborado pelo autor, 2022

Essa diferença no resultado da geração de resíduos sólidos domiciliares pode estar relacionada, ou não, com a influência dos resíduos comerciais e de prestadores de serviço, que estão inseridos na coleta de resíduos sólidos domiciliares pela SOLURB, enquanto que a gravimetria realizada para o desenvolvimento desse trabalho não considerou essas parcelas.

6 Conclusão

A aplicação de técnicas de inteligência artificial têm se provado eficiente na estimação da geração de resíduos sólidos. O trabalho desenvolvido levantou dados da geração de resíduos em 159 domicílios no município de Campo Grande, Mato Grosso do Sul, e apresentou resultados estatisticamente satisfatórios no que diz respeito à estimação da geração e caracterização dos resíduos sólidos através do uso de modelos de classificação.

A utilização de modelos baseados em redes neurais para realizar a regressão da geração de resíduos não se mostrou adequado, possivelmente devido ao tamanho da amostra, que não foi o suficiente para que o modelo capturasse as variações nos padrões de geração. Sendo assim, trabalhos futuros podem ser desenvolvidos buscando melhorar a amostragem com uma quantidade maior de domicílios de forma a diminuir a variabilidade nos dados. Além disso, os modelos de classificação poderiam ser desenvolvidos utilizando uma ferramenta, como algoritmos genéticos, para a seleção dos melhores parâmetros.

Os modelos que apresentaram melhores resultados realizam a classificação dos domicílios com base no valor do IPTU e, a partir do resultado da classificação, utilizam a caracterização dos resíduos de cada classe para inferir a geração dos mesmos. Outras variáveis preditoras além do valor do IPTU podem ser analisadas para incorporar um modelo de aprendizado, e suas influências na estimação da geração e caracterização dos resíduos podem ser comparadas.

Outros trabalhos podem ser desenvolvidos com a finalidade de analisar a influência dos resíduos sólidos domiciliares sobre a geração de resíduos sólidos urbanos e suas componentes, comercial e de serviços, bem como realizar novas gravimetrias para o enriquecimento da base de dados. Além disso outros modelos podem ser desenvolvidos, explorando novas características do problema como a seleção de diferentes atributos preditores.

É importante ressaltar que prever o comportamento de geração de resíduos é uma tarefa complexa, que demanda esforços significativos. Desenvolver um modelo preciso e eficiente é um desafio, especialmente em um contexto em que os padrões de geração podem variar amplamente. Nesse sentido, o sistema desenvolvido apresenta importantes contribuições, permitindo a melhoria contínua do modelo à medida que mais dados forem adicionados.

Além disso, o valor social e ambiental do sistema é inegável, uma vez que a gestão adequada de resíduos sólidos é essencial para a preservação do meio ambiente e da saúde pública. A aplicação de técnicas de inteligência artificial na gestão de resíduos é uma tendência em expansão, e o sistema desenvolvido oferece uma alternativa promissora em relação a métodos tradicionais.

O sistema de apoio à decisão permite projeções precisas e visualização simplificada de resultados para auxiliar na tomada de decisão em diferentes contextos da gestão de resíduos sólidos, contribuindo para o desenvolvimento de novas estratégias e soluções. Sendo baseado em componentes, o sistema é de fácil manutenção e possibilita a implementação de novas funcionalidades.

Referências

- ABDALLAH, M.; TALIB, M. A.; FERROZ, S.; NASIR, Q.; ABDALLA, H.; MAHFOOD, B. Artificial intelligence applications in solid waste management: A systematic research review. **Waste Management**, Elsevier, v. 109, p. 231–246, 2020.
- ABRELPE. **Panorama dos resíduos sólidos no Brasil 2018/2019**. [S.l.]: Ass. Bras. Empr. Limp. Públ. Resíd. Esp. S. Paulo, 2019.
- ALI, S. Estimation of municipal solid waste generation and landfill volume using artificial neural networks. **International Journal of Civil Engineering and Technology**, v. 9, p. 2123–2130, 10 2018.
- ALSABTI, K.; RANKA, S.; SINGH, V. An efficient k-means clustering algorithm. ipps. In: **SPDP Workshop on High Performance Data Mining**. [S.l.: s.n.], 1998.
- AMOAKO, J. K.; YEO, S. Y.; ZHANG, X. Decision tree-based modelling for predicting municipal solid waste generation in developing countries. **Journal of Cleaner Production**, Elsevier, v. 276, p. 123267, 2020.
- ARAIZA-AGUILAR, J.; ROJAS-VALENCIA, M.; AGUILAR-VERA, R. Forecast generation model of municipal solid waste using multiple linear regression. **Global Journal of Environmental Science and Management**, GJESM Publisher, v. 6, n. 1, p. 1–14, 2020.
- BELLMAN, R. **An Introduction to Artificial Intelligence: Can Computers Think?** Boyd & Fraser Publishing Company, 1978. ISBN 9780878350667. Disponível em: <<https://books.google.com.br/books?id=84xQAAAAMAAJ>>.
- BERNACHE-PÉREZ, G.; SÁNCHEZ-COLÓN, S.; GARMENDIA, A. M.; DÁVILA-VILLARREAL, A.; SÁNCHEZ-SALAZAR, M. E. Solid waste characterisation study in the guadalajara metropolitan zone, mexico. **Waste Management & Research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 19, n. 5, p. 413–424, 2001.
- BERRÍOS, M. R. Técnicas de amostragem de resíduos sólidos. In: **Indicadores ambientais**. [S.l.: s.n.], 1997. p. 233–43.
- BUENROSTRO, O.; BOCCO, G.; BERNACHE, G. Urban solid waste generation and disposal in mexico: a case study. **Waste management & research**, Sage Publications, v. 19, n. 2, p. 169–176, 2001.
- CERVANTES, J.; GARCIA-LAMONT, F.; RODRÍGUEZ-MAZAHUA, L.; LOPEZ, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, Elsevier, v. 408, p. 189–215, 2020.
- DANGI, M. B.; URYNOWICZ, M. A.; GEROW, K. G.; THAPA, R. B. Use of stratified cluster sampling for efficient estimation of solid waste generation at household level. **Waste Management & Research**, Sage Publications Sage UK: London, England, v. 26, n. 6, p. 493–499, 2008.

- ESPINASSE, B.; PASCOT, D. Decision support systems (dss): A knowledge oriented approach. In: **Economics and Artificial Intelligence**. [S.l.]: Elsevier, 1987. p. 105–108.
- FAUSETT, L. V. **Fundamentals of neural networks: architectures, algorithms and applications**. [S.l.]: Pearson Education India, 2006.
- GUO, G.; WANG, H.; BELL, D.; BI, Y.; GREER, K. Knn model-based approach in classification. In: SPRINGER. **OTM Confederated International Conferences"On the Move to Meaningful Internet Systems"**. [S.l.], 2003. p. 986–996.
- HAENLEIN, M.; KAPLAN, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. **California management review**, SAGE Publications Sage CA: Los Angeles, CA, v. 61, n. 4, p. 5–14, 2019.
- IBRAOP. **IBRAOP. Instituto Brasileiro de auditoria de obras públicas**. 2020. <<http://www.ibraop.org.br/procedimientos-para-residuos-solidos-urbanos/>>. Acessado em: 12 ago. 2020.
- JALALI, G. Z. M.; NOURI, R. E. Prediction of municipal solid waste generation by use of artificial neural network: A case study of mashhad. **INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH (IJER)**, 2008.
- JIAO, X.; WANG, Q.; LIN, L.; WEI, S.; ZHANG, H. Predicting municipal solid waste generation in china using a k-nearest neighbors model. **Journal of Cleaner Production**, Elsevier, v. 279, p. 123628, 2021.
- KHAN, D.; KUMAR, A.; SAMADDER, S. Impact of socioeconomic status on municipal solid waste generation rate. **Waste Management**, Elsevier, v. 49, p. 15–25, 2016.
- KHAN, M. A.; BURNEY, F. Forecasting solid waste composition—an important consideration in resource recovery and recycling. **Resources, Conservation and Recycling**, Elsevier, v. 3, n. 1, p. 1–17, 1989.
- LEE, K.-M.; LEE, S.-R.; PARK, S.-H. Prediction model for construction and demolition waste generation in korea using support vector machine with consideration of national policies. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 12, n. 12, p. 5043, 2020.
- MA, S.; ZHOU, C.; CHI, C.; LIU, Y.; YANG, G. Estimating physical composition of municipal solid waste in china by applying artificial neural network method. **Environmental Science & Technology**, ACS Publications, v. 54, n. 15, p. 9609–9617, 2020.
- MCCARTHY, J. What is artificial intelligence. 2007.
- MELARÉ, A. V. de S.; GONZÁLEZ, S. M.; FACELI, K.; CASADEI, V. Technologies and decision support systems to aid solid-waste management: a systematic review. **Waste management**, Elsevier, v. 59, p. 567–584, 2017.
- MIEZAH, K.; OBIRI-DANSO, K.; KÁDÁR, Z.; FEI-BAFFOE, B.; MENSAH, M. Y. Municipal solid waste characterization and quantification as a measure towards effective waste management in ghana. **Waste Management**, Elsevier, v. 46, p. 15–27, 2015.

- MURPHY, K. P. et al. Naive bayes classifiers. **University of British Columbia**, v. 18, n. 60, p. 1–8, 2006.
- NILSSON, N. J.; NILSSON, N. J. **Artificial intelligence: a new synthesis**. [S.l.]: Morgan Kaufmann, 1998.
- PATLE, A.; CHOUHAN, D. S. Svm kernel functions for classification. In: IEEE. **2013 International Conference on Advances in Technology and Engineering (ICATE)**. [S.l.], 2013. p. 1–9.
- PERRAJU, T. Artificial intelligence and decision support systems. **International Journal of Advanced Research in IT and Engineering**, v. 2, n. 4, p. 17–26, 2013.
- ROKACH, L.; MAIMON, O. Decision trees. In: **Data mining and knowledge discovery handbook**. [S.l.]: Springer, 2005. p. 165–192.
- RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 2002.
- RYBNEYTSKA, O.; BURSTEIN, F.; RYBIN, A. V.; ZASLAVSKY, A. Decision support for optimizing waste management. **Journal of Decision Systems**, Taylor & Francis, v. 27, n. sup1, p. 68–78, 2018.
- SHARMA, N.; LITORIYA, R.; SHARMA, D.; SINGH, H. P. Designing a decision support framework for municipal solid waste management. **Int J Emerg Technol**, v. 10, n. 4, p. 374–379, 2019.
- SINIR. **SINIR. Sistema nacional de informações sobre a gestão dos resíduos sólidos**. 2020. <<https://sinir.gov.br/perfis/logistica-reversa/logistica-reversa/>>. Acessado em: 12 jan. 2023.
- SONI, U.; ROY, A.; VERMA, A.; JAIN, V. Forecasting municipal solid waste generation using artificial intelligence models—a case study in india. **SN Applied Sciences**, Springer, v. 1, n. 2, p. 162, 2019.
- SOUZA, R. H. M. de; RIBEIRO, V. de O.; DIODATO, J. O.; SANTOS, A. S. dos. Análise gravimétrica dos resíduos sólidos urbanos do município de dourados, ms. **Multitemas**, p. 215–237, 2022.
- SPRAGUE, R. H.; CARLSON, E. D. **Building effective decision support systems**. [S.l.]: Prentice-Hall, 1982.
- WEBB, G. I.; KEOGH, E.; MIKKULAINEN, R. Naïve bayes. **Encyclopedia of machine learning**, v. 15, p. 713–714, 2010.
- WINSTON, P. H. **Artificial Intelligence**. 3. ed. [S.l.]: Addison-Wesley, 1992.
- YEGNANARAYANA, B. **Artificial neural networks**. [S.l.]: PHI Learning Pvt. Ltd., 2009.
- YETILMEZSOY, K.; OZKAYA, B.; CAKMAKCI, M. Artificial intelligence-based prediction models for environmental engineering. **Neural Network World**, Institute of Computer Science, v. 21, n. 3, p. 193, 2011.

ZENG, Y.; CAI, Y.; JIA, P.; JEE, H. Development of a web-based decision support system for supporting integrated water resources management in daegu city, south korea. **Expert Systems with Applications**, Elsevier, v. 39, n. 11, p. 10091–10102, 2012.

ZHANG, Z. Introduction to machine learning: k-nearest neighbors. **Annals of translational medicine**, AME Publications, v. 4, n. 11, 2016.

Apêndices

APÊNDICE A – Códigos

A.1 Preparação dos dados

Para análise exploratória dos dados foi utilizado o seguinte código:

```

1 # -*- coding: utf-8 -*-
2 """
3 Created on Wed Nov 16 13:35:23 2022
4
5 @author: joaom
6 """
7 import seaborn as sns
8 import sweetviz as sv
9 import pandas as pd
10 import numpy as np
11 from matplotlib import pyplot as plt
12 from sklearn.preprocessing import LabelEncoder
13 from mpl_toolkits import mplot3d
14 from pyod.models.knn import KNN
15
16 # Carregar dados
17 file_loc = 'all_dados.xlsx'
18 base = pd.read_excel(file_loc)
19
20 # Preencher dados faltantes
21 base.loc[0, 'IPTU'] = 500
22 base['IPTU'].fillna(base['IPTU'].interpolate(), inplace=True)
23
24 previsores = pd.concat([base['Moradores'], base['Renda'], base['
    IPTU']], axis=1)
25
26 # Codificar a renda
27 labelencoder_renda = LabelEncoder()
28 previsores['Renda'] = labelencoder_renda.fit_transform(previsores
    ['Renda'])

```

```
29 materiais = base.iloc[:,6:14]
30
31 plastico = base['Plastico']
32 papel = base['Papel e Papel o']
33 vidro = base['Vidro']
34 metais = base['Metais']
35 embal = base['Mat. Org.']
36 texteis = base['Tex. Cour. Bor.']
37 materia = base['Emb. Mult.']
38 rejei = base['Rejeitos']
39
40 saida = pd.concat([plastico, papel, vidro, metais, embal, texteis,
                    materia, rejei], axis=1)
41 classe = base['Total']
42 classe_2 = base['Total Reciclveis']
43
44 dados = pd.concat([previsores, saida, classe], axis=1)
45 dados_values = dados.copy()
46
47 # Boxplot
48 def plot_boxplot(df, ft):
49     df.boxplot(column=[ft])
50     plt.grid(False)
51     plt.show()
52
53 plot_boxplot(dados, 'Plastico')
54
55 # Quartis
56 def outliers(df, ft):
57     Q1 = df[ft].quantile(0.25)
58     Q3 = df[ft].quantile(0.75)
59     IQR = Q3 - Q1
60
61     lower_bound = Q1 - 1.5 * IQR
62     upper_bound = Q3 + 1.5 * IQR
63
64     ls = df.index[(df[ft] < lower_bound) | (df[ft] >
        upper_bound)]
65     return ls
```

```
66
67 # Lista para armazenar os índices
68 index_list = []
69 for feature in ['Plástico', 'Papel e Papelão', 'Vidro', 'Metais',
70                'Emb. Mult.', 'Tex. Cour. Bor.', 'Mat. Org.', 'Rejeitos']:
71     #for feature in ['Total']:
72     index_list.extend(outliers(dados, feature))
73
74 # Remove outliers using boxplot
75 def remove(df, ls):
76     ls = sorted(set(ls))
77     df = df.drop(ls)
78     return df
79
80 df_cleaned = remove(dados, index_list)
81 df_cleaned = df_cleaned.reset_index(drop=True)
82
83 # Remove outliers using pyod
84 detector = KNN()
85 detector.fit(materiais)
86 previsoes_detec = detector.labels_
87 np.unique(previsoes_detec, return_counts=True)
88 confianca_previsoes = detector.decision_scores_
89 out_detec = []
90 for i in range(0, len(previsoes_detec)):
91     if previsoes_detec[i] == 1:
92         out_detec.append(i)
93
94 lista_outliers = dados.iloc[out_detec, :]
95 dados_result = dados.copy().drop(out_detec).reset_index(drop=
96     True)
97
98 previsores = pd.concat([dados_result['IPTU']], axis=1)
99 nao_emb = dados_result['Não Emb.']
100 emb = dados_result['Embalagens']
101 recic = dados_result['Recicláveis']
102 classe = dados_result['Total']
103 classe_orig = classe.copy()
104 materiais = dados_result.iloc[:, 3:11]
```

```

103
104 # Relatório analise dos dados
105 report = sv.analyze(dados_values)
106 report.show_html()
107
108 # Gráfico de correlação
109 plt.figure(figsize=(10, 6))
110 sns.set(font_scale=1)
111 heatmap = sns.heatmap(df_cleaned.corr(), vmin=-1, vmax=1, annot=
    True)
112 plt.savefig('corr_boxplot_3classes.png', format='png')
113
114 %matplotlib notebook
115 fig = plt.figure()
116 plt.plot(1,1)
117 ax = plt.axes(projection='3d')
118 ax.set_xlabel("Plástico")
119 ax.set_ylabel("Moradores")
120 ax.set_zlabel("IPTU")
121
122 ax.scatter(plastico, previsores['Moradores'], previsores['IPTU'
    ])
123 plt.show();

```

O código seguinte prepara os dados para serem utilizados nos modelos:

```

1 # fit density
2 sample = classe.values
3 model = KernelDensity(bandwidth=30, kernel='gaussian')
4 sample = sample.reshape((len(sample), 1))
5 model.fit(sample)
6
7 # sample probabilities for a range of outcomes
8 min_val = int(min(classe))
9 max_val = int(max(classe))
10 values = np.asarray([value for value in range(0, max_val)])
11 values = values.reshape((len(values), 1))
12 probabilities = model.score_samples(values)
13 probabilities = np.exp(probabilities)
14

```

```
15 # plot the histogram and pdf
16 fig, ax = plt.subplots(1, 1)
17 ax.hist(classe, bins=20, density=True)
18 ax.plot(probabilities)
19 ax.set_xlabel('Massa em gramas')
20 ax.set_ylabel('Frequência')
21 plt.show()
22
23 # for local minima
24 ind_x_min = argrextrema(probabilities, np.less)
25 lim = [min_val, ind_x_min[0][1], ind_x_min[0][2], ind_x_min[0][3],
        max_val]
26
27 # Definição das classes
28 for i in range(len(classe)):
29     if (classe[i] < 1389) and (classe[i] >= 0):
30         classe[i] = 1
31     elif (classe[i] < 2237) and (classe[i] >= 1389):
32         classe[i] = 2
33     elif (classe[i] < max_val) and (classe[i] >= 2237):
34         classe[i] = 3
35     else:
36         classe[i] = 3
37
38 qnts = classe.value_counts()
39
40 # Melhor caso
41 X_treino, X_teste, y_treino, y_teste = train_test_split(
    previsores, classe, test_size=0.2, random_state=64)
42 # Pior caso
43 #X_treino, X_teste, y_treino, y_teste = train_test_split(
    previsores, classe, test_size=0.2, random_state=10)
44 # Aleatorio
45 #X_treino, X_teste, y_treino, y_teste = train_test_split(
    previsores, classe, test_size=0.2)
```

A.2 Modelo de regressão

Código para Rede Neural Artificial:

```
1 # Normalizacao entre 0.1 e 0.9
2 def calculate_coef(z,w,df,ft):
3     min = np.min(df[ft])
4     max = np.max(df[ft])
5     a = (z - w) / (max - min)
6     b = w - (a * min)
7
8     return a, b
9
10 def map_values(df,ft,a,b):
11     values = pd.DataFrame(columns=[ft])
12
13     for i in range(len(df)):
14         y = (a * df[ft].loc[i]) + b
15         values.loc[i] = y
16
17     return values
18
19 def normalize(df):
20     z = 0.9
21     w = 0.1
22     col_ls = []
23     ab_ls = []
24
25     for ft in ['Moradores', 'Renda', 'IPTU', 'Total']:
26         a, b = calculate_coef(z,w,df,ft)
27         ab_ls.append([a,b])
28         col_ls.append(map_values(df,ft,a,b))
29
30     return col_ls,ab_ls
31
32 dados = pd.concat([df_cleaned['Moradores'], df_cleaned['Renda'],
33                    df_cleaned['IPTU'], df_cleaned['Total']], axis=1)
34
35 dados_ls, ab = normalize(dados)
36
37 previsores = pd.concat([pd.DataFrame(dados_ls[0]),pd.DataFrame(
38     dados_ls[1]),
39                          pd.DataFrame(dados_ls[2])], axis=1)
40
41 saida = pd.DataFrame(dados_ls[3])
```

```
38
39 # Divisao treino teste
40 X_treino, X_teste, y_treino, y_teste = train_test_split(
    previsoires, saida, test_size=0.2, random_state=1)
41
42 # Rede neural
43 camada_entrada = Input(shape=(3,))
44 camada_oculta1 = Dense(units = 10, activation='sigmoid')(
    camada_entrada)
45 camada_oculta2 = Dense(units = 10, activation='sigmoid')(
    camada_oculta1)
46 camada_saida1 = Dense(units = 1, activation='linear')(
    camada_oculta2)
47
48 regressor = Model(inputs = camada_entrada, outputs = [
    camada_saida1])
49
50 regressor.compile(optimizer = 'adam', loss='mse')
51 regressor.fit(X_treino, [y_treino],
52               epochs=5000, batch_size=100)
53
54 pv_saida = regressor.predict(X_treino)
55
56 a = ab[3][0]
57 b = ab[3][1]
58 new_saida = []
59 new_teste = []
60 new_treino = []
61
62 numpy_teste = y_teste.to_numpy()
63 numpy_treino = y_treino.to_numpy()
64
65 for i in range(len(pv_saida)):
66     new_saida.append((pv_saida[i][0] - b) / a)
67
68 for i in range(len(numpy_teste)):
69     new_teste.append((numpy_teste[i][0] - b) / a)
70
71 for i in range(len(numpy_treino)):
```

```
72     new_treino.append((numpy_treino[i][0] - b) / a)
73
74 # Comparacao teste
75 plt.plot(new_saida)
76 plt.plot(new_teste)
77 plt.legend(["Saida obtida", "Saida esperada"])
78
79 # Comparacao treino
80 plt.plot(new_saida)
81 plt.plot(new_treino)
82 plt.legend(["Saida obtida", "Saida esperada"])
```

A.3 Modelos de classificação

Código para Árvore de Decisão:

```
1 # Arvore de Decisao
2 dtree_model = DecisionTreeClassifier(max_depth = 3).fit(X_treino
    , y_treino)
3 #dtree_model = DecisionTreeClassifier().fit(X_treino, y_treino)
4 dtree_predictions = dtree_model.predict(X_teste)
5
6 # Precisao
7 print("Accuracy:", metrics.accuracy_score(y_teste,
    dtree_predictions))
8
9 # Matriz de confusao
10 cm = confusion_matrix(y_teste, dtree_predictions)
```

Código para SVM:

```
1 # SVM
2 svm_model_linear = SVC(kernel = 'rbf', C = 0.5).fit(X_treino,
    y_treino)
3 svm_predictions = svm_model_linear.predict(X_teste)
4
5 # Precisao
6 accuracy = svm_model_linear.score(X_teste, y_teste)
7 print(accuracy)
8
9 # Matriz de confusao
```

```
10 cm = confusion_matrix(y_teste, svm_predictions)
```

Código para KNN:

```
1 # KNN
2 knn = KNeighborsClassifier(n_neighbors = 5).fit(X_treino,
        y_treino)
3
4 # Precisao
5 accuracy = knn.score(X_teste, y_teste)
6 print (accuracy)
7
8 # Matriz de confusao
9 knn_predictions = knn.predict(X_teste)
10 cm = confusion_matrix(y_teste, knn_predictions)
```

Código para Naive Bayes:

```
1 # Naive Bayes
2 gnb = GaussianNB().fit(X_treino, y_treino)
3 gnb_predictions = gnb.predict(X_teste)
4
5 # Precisao
6 accuracy = gnb.score(X_teste, y_teste)
7 print (accuracy)
8
9 # Matriz de confusao
10 cm = confusion_matrix(y_true=y_teste, y_pred=gnb_predictions)
```

A.4 Sistema de apoio à decisão

O sistema desenvolvido encontra-se disponível no GitHub do autor:

github.com/joaomarcoscomp