RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano V, v.1, n.1, jan/julho 2025. | submissão: 17/05/2025 | aceito: 19/05/2025 | publicação:21/05/2025

Aplicação de técnicas de machine learning na otimização da gestão hospitalar

Application of machine learning techniques in hospital management optimization

Faitoma Jorge – Universidade Kimpa Vita Nicolau Pedro – Universidade Kimpa Vita Nkanga Pedro – Universidade Kimpa Vita

RESUMO
Este trabalho tem como objetivo aplicar o algoritmo de clusterização K-Means para identificar padrões de doenças e subsidiar a tomada de decisões no Hospital do Catapa, através da integração de soluções baseadas em tecnologias emergentes, com destaque para a Inteligência Artificial e, em particular, o Machine Learning, com o intuito de aprimorar a análise de dados clínicos e otimizar a administração dos recursos de saúde. A metodologia adotada envolveu a coleta e o tratamento de 4.050 registros clínicos, contemplando variáveis como data de atendimento, bairro, gênero, idade, faixa etária, tipo de doença e agrupamento do bairro. Após o pré-processamento dos dados, aplicou-se o algoritmo K-Means, possibilitando a formação de 5 clusters compostos por pacientes com características semelhantes. A análise indicou, que o bairro Catapa (classificado como periurbano) concentra aproximadamente 58% dos casos de malária e febre. Observou-se também que a faixa etária de jovens representa 38,4,% dos registros; crianças, 32,3%; adolescentes, 14,9%; adultos, 11,8%; e idosos, 2,6%, com leve predominância do gênero masculino (50,1%) em relação ao feminino (49,9%). O bairro Catapa, isoladamente, responde por 37% dos casos registrados; Mbemba Ngango por 15%; Kindenuku por 10%; Dunga por 5%; Papelão por 3%; e os demais 34 bairros somam os 30% restantes. O mês de junho apresentou a maior incidência, com cerca de 36% do total de ocorrências, registou-se também que quanto aos tipos de bairros, o tipo Periourbano apresenta uma ocorrência de 73,6% de caso, os bairros urbanos com 24,6% e os bairros Rurais com 1.8%, conforme os dados extraídos do livro de registro do hospital.
Palavras-chave: Machine Learning, K-Means, Gestão Hospitalar, Clusterização, Análise de Dados

ABSTRACT
This study aims to apply the K-Means clustering algorithm to identify disease patterns and support decision-making at Catapa Hospital. In this context, the integration of emerging technologies, particularly Artificial Intelligence and Machine Learning becomes essential for improving clinical data analysis and optimizing health resource administration. The methodology involved the collection and processing of 4,050 clinical records, covering variables such as date of attendance, neighborhood, gender, age, age group, type of disease, and neighborhood classification. After data preprocessing, the K-Means algorithm enabled the formation of five clusters composed of patients with similar characteristics. The analysis indicated that the Catapa neighborhood (classified as peri-urban) accounts for approximately 58% of malaria and fever cases. The youth age group represents 38.4% of the records, followed by children (32.3%), adolescents (14.9%), adults (11.8%), and the elderly (2.6%), with a slight predominance of the male gender (50.1%). Catapa alone accounts for 37% of recorded cases; Mbemba Ngango for 15%; Kindenuku for 10%; Dunga for 5%; Papelão for 3%; and the remaining 34 neighborhoods for 30%. June had the highest incidence, with about 36% of all occurrences. Regarding neighborhood types, peri-urban areas reported 73.6% of cases, urban areas 24.6%, and rural areas 1.9%. Additionally, an interactive dashboard was developed using the Dash library, allowing dynamic visualization of data and results, providing hospital managers with a practical tool for evidence-based analysis and decision-making. The findings demonstrate that K-Means clustering is effective in identifying patterns and supporting hospital management.

Keywords: Machine Learning, K-Means, Hospital Management, Clustering, Data Analysis

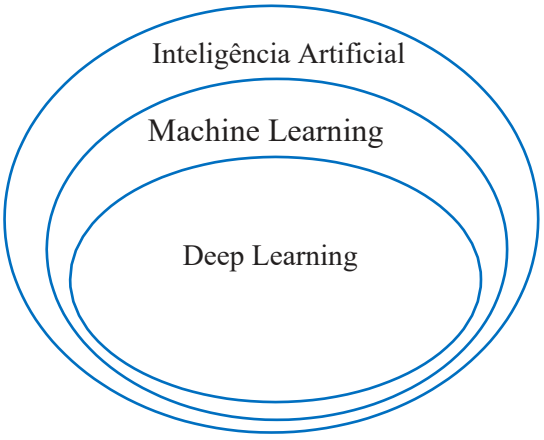
1. INTRODUÇÃO

A gestão hospitalar eficiente é essencial para garantir a qualidade dos serviços de saúde, especialmente em regiões com recursos limitados, como a província do Uíge. O Hospital Geral do Catapa (HGC), fundamental no atendimento à população, enfrenta desafios significativos devido à falta de sistemas automatizados de monitoramento de doenças. A ausência de ferramentas tecnológicas para prever surtos e realizar uma alocação eficaz de recursos compromete a capacidade de resposta a surtos de doenças, agravando a situação da saúde pública. Este contexto evidencia a necessidade urgente de uma solução inovadora que integre tecnologias avançadas, como o Machine Learning (ML), para melhorar a gestão hospitalar.

O objetivo central deste trabalho é explorar como a aplicação de técnicas de ML pode otimizar a gestão do Hospital Geral do Catapa, ajudando na previsão de surtos e identificando padrões de doenças que podem melhorar a alocação de recursos e as ações preventivas. Através da análise de dados de saúde coletados, pretende-se identificar os bairros mais afetados, grupos de risco por faixa etária e gênero, e possibilitar a tomada de decisões mais precisas e rápidas. A utilização de técnicas de ML, como clustering, permitirá não apenas a previsão de surtos de doenças, mas também o direcionamento de estratégias para minimizar seu impacto e melhorar a qualidade do atendimento.

Em suma, este estudo visa contribuir para uma melhor gestão dos recursos de saúde no Uíge especificamente no hospital do Catapa, utilizando a inteligência artificial para prever doenças, melhorar o controle e otimizar as estratégias de saúde pública na província.

Figura 1: Relacionamento da AI, ML e DL



Fonte: Autor

2. MARCO TEÓRICO

A partir da realização de uma revisão narrativa da literatura, buscou-se apresentar discussões conceituais sobre o tema técnico de IA aplicadas em processos de gestão hospitalar. A "revisão narrativa" não aplica critérios explícitos e sistemáticos na busca e análise crítica da literatura. A seleção dos estudos e interpretação das informações podem estar sujeitas à subjetividade dos autores e este tipo de revisão não necessita esgotar as fontes de coleta de informações (Filho, 2020).

O aprendizado de máquina engloba a formulação de modelos ou algoritmos que adquirem conhecimento de conjuntos de dados históricos para facilitar previsões ou executar ações. Esses modelos passam por treinamento utilizando conjuntos de dados rotulados ou não rotulados, e sua eficácia é aprimorada à medida que são expostos a um volume crescente de dados e feedback construtivo (Pedro, 2024, p. 42). Compreender os princípios, metodologias e tarefas fundamentais inerentes a esse processo constitui um dos pilares fundamentais do aprendizado de máquina. Técnicas como agrupamento, árvores de decisão e floresta aleatória estão incluídas entre as metodologias examinadas nesta análise.

"Os princípios fundamentais do aprendizado de máquina são examinados neste capítulo específico. Ele abrange metodologias não supervisionadas (incluindo protótipos K e suas métricas de validade associadas), bem como abordagens supervisionadas (como máquinas de vetores de suporte, árvores de decisão e florestas aleatórias). Elucidamos os mecanismos operacionais dessas técnicas junto com suas métricas de desempenho correspondentes, que incluem precisão, exatidão, recall, pontuação F1 e a matriz de confusão. Além disso, investigamos a técnica SMOTE para ressaltar as distinções entre conjuntos de dados balanceados e desbalanceados ". (Pedro, 2024, p. 42).

2.1 Inteligência Artificial

A Inteligência Artificial tem como principal objetivo procurar métodos e formas dos computadores fazerem o mesmo tipo de análises que a mente humana faz sistematicamente. Esta definição de IA apesar de ser bastante simples de compreender, a verdade é que por detrás desta afirmação tão breve, não revela a tamanha complexidade e potencial deste tema. Ao afirmar que este conceito relaciona a mente humana, referimo-nos às capacidades psicológicas como a

previsão, associação, perceção, planeamento, entre outras (RUSSELL & NORVIG, 2016).

2.2 Considerações e a justificação da escolha de modelo

A gestão hospitalar é um campo complexo que envolve a coordenação de diversos recursos, serviços e profissionais com o objetivo de oferecer um atendimento de saúde de qualidade. Com o crescimento exponencial da quantidade de dados gerados no setor de saúde, a necessidade de métodos eficazes para analisar e interpretar essas informações se torna cada vez mais evidente. Nesse contexto, as técnicas de ML emergem como ferramentas poderosas que podem transformar a forma como os hospitais operam (SHAPIRO, 1992) A presente pesquisa adotou uma abordagem metodológica avançada de aprendizado de máquina para tipificar o índice de doenças por diferentes categorias. O algoritmo K-Means especificado para variáveis categóricas, possibilitando uma representação mais precisa e abrangente das características das patologias.

A escolha do algoritmo K-Means se deu pela sua eficácia em lidar com dados estruturados, sua capacidade de formar agrupamentos distintos a partir de características comuns e sua eficiência computacional. Ele foi especialmente útil para identificar padrões na distribuição de doenças por região e faixa etária, permitindo uma segmentação eficiente e apoiando estratégias direcionadas de prevenção e tratamento.

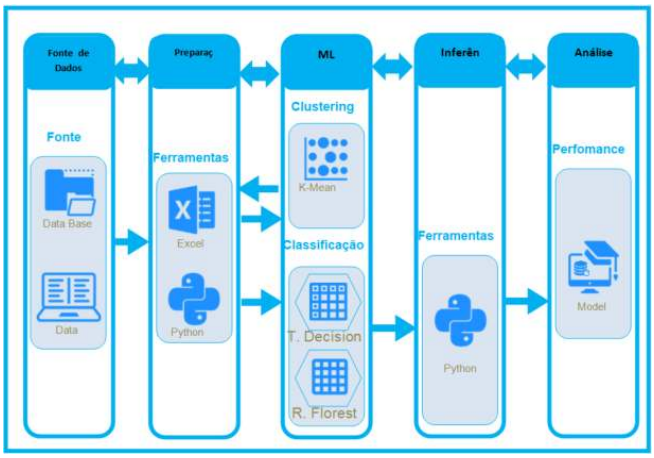
3. MATERIAL E MÉTODO

A presente pesquisa caracteriza-se como quantitativa, pois envolve a coleta, organização, análise e interpretação de dados numéricos e categóricos com o objetivo de compreender padrões e realizar previsões. O estudo também possui natureza exploratória e descritiva, uma vez que busca identificar comportamentos, frequências e agrupamentos de doenças em diferentes contextos (bairros, faixas etárias, gêneros, etc.) com o auxílio de algoritmos de Machine Learning.

3. 1 Arquitetura do projeto

Neste ponto apresentam-se as principais análises de diferentes métodos e ferramentas e métricas utilizadas no projeto. As etapas do estudo estão ilustradas na figura 2

Figura 2: Arquitetura do projecto



3.1.1 Descrição dos dados

3.1.1.1 Fonte de dados

Para avaliação das diferentes técnicas de ML usados nesta dissertação, foi considerado um conjunto de dados relativos aos pacientes que frequentam o Hospital do Catapa, nos anos de 2022 a 2023. O conjunto de dados é constituído por 4050 registos (pacientes com vários registos históricos) com 08 atributos que foram extraídos a partir do livro de registros do HC.

3.1.1.2 Preparação de dados

Após a colheita de dados, fez-se a preparação dos mesmos no intuito de combinar, estruturar e organizá-los para serem usados:

- Para criar modelos de aprendizagem automática;

- Aplicações de análise e visualização de dados.

Utilizando ferramentas estatísticas e algoritmos de aprendizagem automática, para descobrir a tendência dos pacientes que se registram no Hospital.

3.1.1.2 Treinamento de aprendizagem automática

Os algoritmos de aprendizagem automática sugeridos neste estudo, que nos permitiram tipificar as doenças dos pacientes por faixa etária, género e bairro, foram treinados usando a linguagem

Python e a biblioteca SciKit-Learn.

❖ *Python*: É uma linguagem de programação de alto nível amplamente utilizada para várias aplicações, incluindo ML (Pedro, 2024)

❖ *Scikit-Learn*: É um módulo Python que integra algoritmos de aprendizagem automática de última geração para problemas supervisionados e não supervisionados, com foco na facilidade de uso e desempenho (Pedro, 2024)

3.1.1.3 Inferência de aprendizagem automática

Segundo Pedro (2024), destaca que os modelos de aprendizagem automática não supervisionada e supervisionada recebem pontos de dados inéditos durante o processo de inferência. Para começar, usamos o modelo K-Means para agrupar as doenças de acordo o bairro, género, idade, tipo de doença, tipo de bairro ou características semelhantes. Depois disso, a variável agrupada, ou cluster, é adicionada ao novo dataset para ser alimentada no treinamento para os modelos de aprendizagem automática supervisionada que calculam a previsão dos pacientes pertencerem a uma certa patologia logo que chegam. O registo de modelos facilita o acompanhamento de modelos treinados no hospital.

3.2 Estrutura de dados dos Pacientes

As informações relevantes para o modelo de previsão das doenças incluem elementos que impactam o comportamento dos pacientes. A Tabela a seguir lista os elementos de dados que afetam os padrões de atrito e retenção dos pacientes.

Tabela 1: Estrutura de Dados dos Pacientes do HC

Caracteristica	Tipo	Descrição
Nº	Número	Número de registro
Data	Categórico	Data de cadastro
Nome do Paciente	Categórico	Nome completo do paciente
Bairro	Categórico	Bairro de proveniência
Genero	Categórico	Sexo (Masculino e Feminino)
Idade	Número	Idade do paciente (0 a 150 anos)
Início da Doença	Categórico	Data de inicio dos sintomas
Queixa e Exame Objectivo	Categórico	Reclamação dos que se sente
Diagnóstico Clínico	Categórico	Registo da doença pelo médico
Resultados dos Exames Complementares	Categórico	É dado os resultados
Indicações	Categórico	Recomendações do médico
Doze	Número	Dose da medicação
Nº de Dias	Número	Dias que deverá seguir a doze
Quantidades	Número	Quantidade de medicamento (ml)
Peso	Número	Peso do paciente

O pré-processamento foi realizado para "melhorar a qualidade dos dados por meio da eliminação ou minimização dos problemas", incluindo ruídos, valores incorretos, inconsistentes ou ausentes.

Por meio do pré-processamento de dados, os elementos considerados irrelevantes foram removidos manualmente e o que colocou em risco a privacidade dos dados dos pacientes. Os autores afirmam claramente que, quando um atributo não contribui para a estimativa do valor do atributo alvo, é considerado irrelevante. Neste caso, dos 15 atributos do conjunto de dados

inicial, 10 atributos foram removidos, e ficaram os 05 outros atributos que são Data de cadastro, bairro, gênero, idade, diagnóstico clínico (doença).

Após análise exploratória foram adicionados os atributos seguintes a situação de cada paciente, incluindo a faixa etária, tipo de doença e tipo de bairro. Isso nos permitiu analisar por doença e por faixa para descobrir quais doenças são mais frequentes no bairro de origem, estabelecendo perfis e outras características.

Ficamos com 4050 registos (pacientes) e 08 colunas, a tabela mostra os atributos restantes.

Tabela 2: Estrutura de Dados dos a Tratar

Caracteristica	**Descrição**
Data	Data de cadastro
Bairro	Bairro de proveniência
Género	Sexo (Masculino e Feminino)
Idade	Idade do paciente (0 a 150 anos)
Diagnóstico Clínico (Doença)	Registo da doença pelo médico
Faixa Etária	Para agrupar a idade por faixa
Tipo de Doença	Doença por tipo (Parasitária e Hipersensibilidade)
Tipo de Bairro	Agrupamento dos bairros por tipo (urbano, periurbano e rural)

Os algoritmos de agrupamento K-means foi usado como métodos de aprendizagem automática para resolver o problema de previsão e agrupamento das doenças. As técnicas de pré-processamento utilizadas incluíram o tratamento de dados desbalanceados, ruídos, incompletos, redundâncias e conversão de dados categóricos em números.

3.3 Perfis de Pacientes Utilizando o Cluster K-Means

a) Contextualização

No contexto da gestão hospitalar, a formação de grupos de perfis de pacientes com base em dados clínicos e demográficos é uma etapa essencial para a identificação de padrões relevantes que podem auxiliar na previsão de surtos, otimização de recursos e definição de estratégias de intervenção. A utilização do algoritmo K-Means permite agrupar dados categóricos, possibilitando uma análise mais abrangente do cenário de saúde da região estudada.

Neste estudo, foram consideradas variáveis como: data da ocorrência, bairro de residência, gênero, idade, faixa etária, tipo de doença, categoria do bairro e status clínico. A escolha dessas variáveis visa capturar diferentes dimensões do perfil epidemiológico dos pacientes, contribuindo para a construção de um sistema de apoio à decisão mais eficiente.

O objetivo principal desta fase é segmentar a população em grupos com características semelhantes, de modo a facilitar a identificação de padrões críticos de incidência de doenças, comportamentos regionais de risco e possíveis grupos vulneráveis. A partir desses clusters, pretende-se implementar um sistema de alerta precoce e oferecer suporte à formulação de políticas públicas mais direcionadas e eficazes na área da saúde.

b) Método para encontrar o valor ideal de K

Para determinar o número ideal de clusters (k), empregou-se uma abordagem fundamentada em múltiplas iterações. O algoritmo foi executado 20 vezes para cada valor de k, variando de 2 a 40, e a média dos resultados foi calculada com a finalidade de minimizar a influência de valores aleatórios.

Três métodos de validação foram utilizados: Elbow, Silhouette Score e Davies Bouldin Index.

Tal combinação fortalece a robustez da escolha de k, considerando tanto a compactação interna dos clusters quanto a separação entre eles.

O método Elbow indicou um ponto de inflexão em k = 5, momento em que a taxa de redução do WCSS (Within-Cluster Sum of Squares) se estabiliza. O Silhouette Score, por sua vez, apresentou a maior média também em k = 5, com variações entre 0,0828 e 0,11 ao longo das iterações. Embora k = 4 também tenha demonstrado um desempenho satisfatório, a escolha de k = 5 é justificada pela consistência dos resultados obtidos entre as diferentes métricas. Subsequentemente, aplicou-se o algoritmo K-Means para segmentar as doenças em distintos bairros, fundamentando-se na configuração ótima de k = 5. A pontuação média de silhueta obtida confirma a qualidade da segmentação, indicando que os clusters gerados são bem definidos e claramente separados.

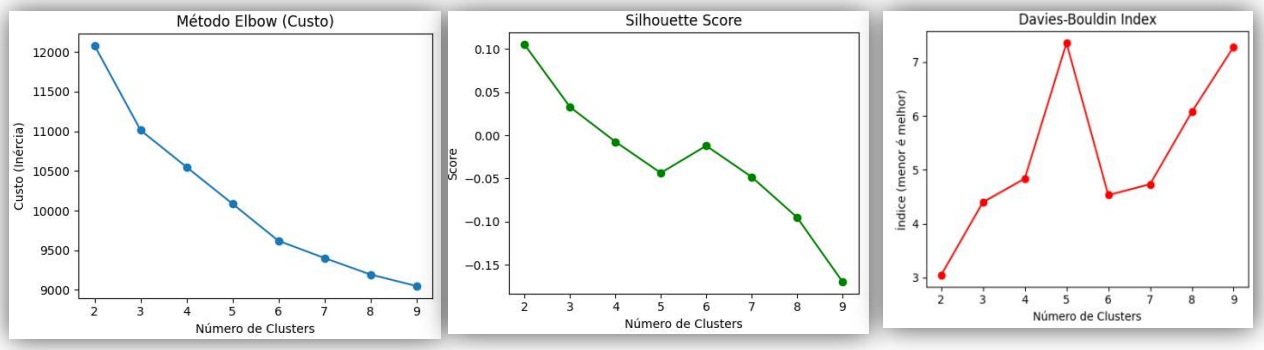


Figura 3: Método de Validação

Obs: A seleção de 5 clusters foi baseada em uma análise conjunta de três métricas de avaliação: Método do Cotovelo (Elbow), Silhouette Score e Davies-Bouldin Index, levando em conta não apenas os valores matemáticos, mas também a interpretação prática dos agrupamentos gerados. . Optou-se por 5 clusters por apresentarem melhor equilíbrio entre separabilidade (bom Silhouette), baixo custo (Elbow) e mínima sobreposição (menor Davies- Bouldin), além de oferecerem interpretação mais clara e útil para a gestão hospitalar.

a) Visualizar os perfis criados após o treinamento do modelo k-means

Após a definição do número ideal de *k*, por meio dos métodos Elbow, Silhouette Score e Davies-Bouldin, aplicamos o algoritmo K-Means para segmentar os casos de doenças em perfis distintos, considerando tanto variáveis categóricas quanto numéricas dos pacientes atendidos no Hospital do Catapa.

Como resultado, foram criados cinco clusters (ou perfis), cada um representando um grupo de pacientes com características semelhantes, como o tipo de doença, bairro de origem, faixa etária e gênero. Esses perfis possibilitam uma visão mais aprofundada sobre os padrões de ocorrência de doenças em diferentes regiões e faixas da população.

A Figura 4 apresenta a distribuição dos casos por cluster, permitindo visualizar como os grupos estão formados. Para reforçar nossa análise e investigar a robustez da segmentação, optamos por realizar duas abordagens distintas:

- *Uma incluindo a variável "Faixa Etária" ;*
- *Outra excluindo essa variável.*

Essa diferenciação nos permitiu avaliar o quanto o tipo de bairro influencia na formação dos grupos de doenças, fornecendo subsídios para hipóteses mais sólidas sobre a relação entre o ambiente e os casos registrados.

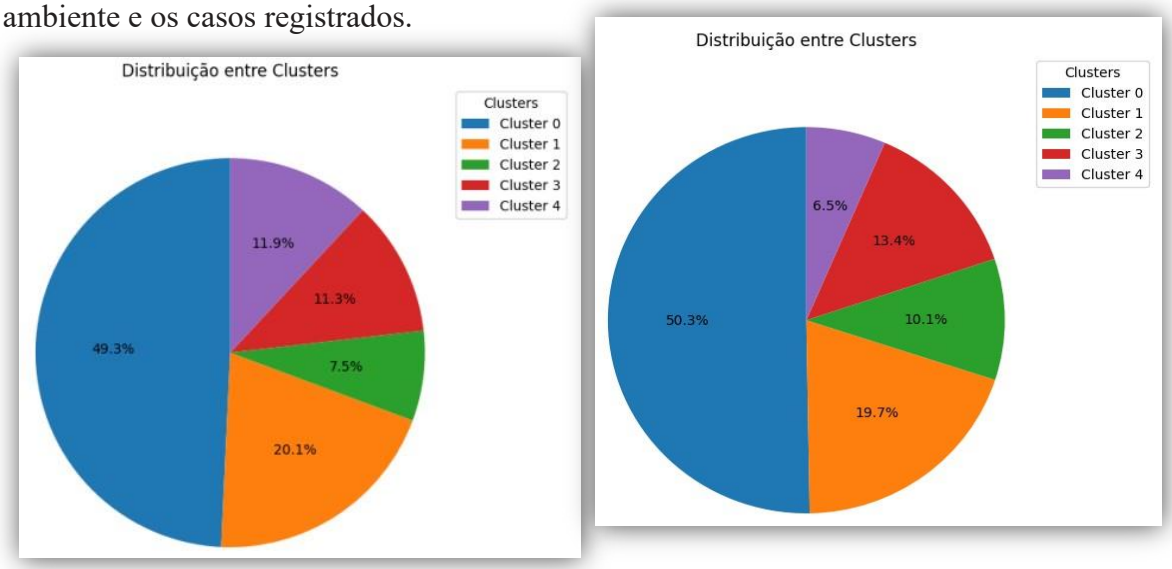


Figura 4: Distribuição de Doenças em cluster com variável Faixa Etária e sem Faixa Etária

Após segmentar os alunos com o modelo K-Means, obtemos os seguintes resultados:

a) Cluster com Faixa Etária

- Cluster 0: 1995 Pacientes;
- Cluster 1: 813 Pacientes;
- Cluster 2: 303 Pacientes;
- Cluster 3: 456 Pacientes;
- Cluster 4: 483 Pacientes.

Essa organização permite uma análise mais detalhada dos grupos de pacientes com características semelhantes. Cada cluster representa um conjunto específico de pacientes com base em suas características. É uma abordagem útil para personalizar estratégias de ensino e oferecer suporte adequado a cada grupo.

- Em termos de distribuição por Gênero, Cluster 0: Predominantemente masculino, com 80,5% de homens e 19,5% de mulheres, o cluster 1: Aqui, os homens são a maioria, com 69,3%, enquanto as mulheres correspondem a 30,7%, o cluster 2: A distribuição entre os gêneros é mais equilibrada, com 51,7% de homens e 48,3% de mulheres, o cluster 3: O grupo tem uma predominância masculina com 58% de homens e 42% de mulheres e o cluster 4: O grupo é mais equilibrado, com 55,2% de homens e 44,8% de mulheres.
- A distribuição por Faixa Etária, o cluster 0: O grupo é composto principalmente por jovens entre 19 e 29 anos (53,2%), seguidos por adultos de 30 a 39 anos (32,7%), o cluster 1: A faixa etária predominante é 30 a 39 anos, com 61,5%, seguido por jovens de 19 a 29 anos (24,1%), o cluster 2: A maioria do grupo está na faixa de 19 a 29 anos (63,4%), seguida de adultos de 30 a 39 anos (36,6%), o cluster 3: Este grupo é predo- minantemente composto por jovens adultos, com 72,3% entre 19 a 29 anos, e 27,7% de 30 a 39 anos, o cluster 4: O maior grupo está na faixa etária de 30 a 39 anos (68%), seguidos por 30,4% entre 19 a 29 anos.
- Em termos de Doenças, a cluster 0: A maior parte dos casos são de malária (45%), febre (30%), e paludismo (15%), a cluster 1: A malária predomina com 60%, seguida por febre com 25%, e paludismo com 10%, a cluster 2: A febre é a doença mais co- mum com 50%, seguida pela malária com 40%, a cluster 3: A malária tem uma preva- lência de 55%, seguida de paludismo com 30%, e febre com 10% e cluster 4: a malária é a doença mais prevalente (65%), seguida de paludismo (20%), e febre (10%).
- Já em Bairro, o grupo 0: O bairro Catapa é o mais afetado com 58%, seguido por Kindenuko com 20%, grupo 1: O bairro Mbemba Ngango tem a maior incidência com 53%, seguido por Dunga com 22%, o grupo 2: O bairro Kindenuko lidera com 62% dos casos, seguido por Kilala com 15%, o grupo 3: Catapa é o bairro com a maior concentração de casos (50%), seguido por Mbemba Ngango com 20% e por fim no grupo 4: Catapa novamente lidera com 60%, seguido por Kilala com 25%.

b) Cluster sem a variável Faixa Etária

- Cluster 0: 2036 Pacientes;
- Cluster 1: 799 Pacientes;

- Cluster 2: 411 Pacientes;
- Cluster 3: 542 Pacientes;
- Cluster 4: 262 Pacientes.

Essa organização permite uma análise mais detalhada dos grupos de pacientes com características semelhantes. Cada cluster representa um conjunto específico de pacientes com base em suas características. É uma abordagem útil para personalizar estratégias dos hospitais oferecerem suporte adequado a cada grupo.

- Na distribuição por Gênero, a cluster 0: Predominância masculina com 80,5%, e o restante feminino com 19,5%, a cluster 1: A distribuição masculina é de 69,3%, com 30,7% de mulheres, a cluster 2: O gênero é mais equilibrado, com 51,7% de homens e 48,3% de mulheres, a cluster 3: Há uma predominância masculina de 58% de homens e 42% de mulheres, a cluster 4: O grupo tem uma distribuição equilibrada entre os gêne- ros, com 55,2% de homens e 44,8% de mulheres.
- Em termos da distribuição por Doenças, a cluster 0: A maior parte dos casos são de malária (45%), febre (30%) e paludismo (15%), a cluster 1: A malária predomina com 60%, seguida de febre com 25% e paludismo com 10%., a cluster 2: A febre predomina com 50%, seguida pela malária com 40%, a cluster 3: A malária tem uma prevalência de 55%, seguida de paludismo com 30% e febre com 10%, a cluster 4: A malária pre- domina com 65%, seguida de paludismo com 20%, e febre com 10%.
- A Distribuição por Bairro, aluster 0: Catapa lidera com 58% dos casos, seguido por Kindenuko com 20%, a cluster 1: Mbemba Ngango é o bairro com maior número de ca- sos, com 53%, seguido por Dunga com 22%, a cluster 2: Kindenuko tem 62% dos ca- sos, seguido por Kilala com 15%, a cluster 3: Catapa novamente se destaca com 50%, seguido por Mbemba Ngango com 20%, a cluster 4: Catapa tem 60% dos casos, seguido por Kilala com 25%.

3.4 Os agrupamentos dos Pacientes com modelo de aprendizagem automática

a) Contextualização

A metodologia utilizada neste estudo revelou uma forte capacidade de previsão, como demonstrado pela acurácia dos modelos de agrupamento. A aplicação de técnicas de Aprendizagem Automática (ML) mostrou-se efetiva na análise de dados de saúde, possibilitando uma compreensão mais profunda dos padrões existentes entre os pacientes.

Para prever o estado clínico dos pacientes, foi implementado um processo de agrupamento, que possibilitou o desenvolvimento de perfis diferenciados com base em variáveis como sexo, local de residência, tipo de patologia e faixa etária. Esses perfis receberam uma nova variável categórica denominada "Cluster", que resume características semelhantes entre os grupos.

Ao incorporar o agrupamento como parte do processo de engenharia de atributos, o modelo incorporou não só os dados individuais, mas também o contexto coletivo dos pacientes, o que teve um impacto significativo no aprimoramento da precisão das previsões. Assim, as instituições de saúde podem identificar de forma antecipada pacientes ou regiões em risco ou com maiores chances de sucesso no tratamento, permitindo a implementação de estratégias específicas para reduzir perdas, otimizar recursos e salvar vidas.

b) Instrumento de avaliação

Avaliamos o desempenho dos modelos árvore de decisão e floresta aleatória na previsão de doenças através de métricas de desempenho, como precisão, recall e F1-score. Os achados evidenciaram uma elevada exatidão, sugerindo a efetividade dos modelos criados. O objetivo principal é identificar o modelo de previsão mais apropriado e comparar as diversas métricas de performance de previsão de cada grupo utilizado. Iniciamos com amostras não balanceadas e posteriormente balanceadas para podermos comparar os resultados.

1) Previsão sem balanceamento de amostra

Atualmente, vamos examinar os resultados da previsão sem balanceamento de amostra, observando como os modelos funcionam sem ajuste nas proporções das classes. Veremos como esse método pode impactar a precisão, o recall e outras métricas de avaliação. Vamos examinar esse caso e compreender suas implicações.

4. RESULTADOS E DISCUSSÃO

Neste capítulo, são apresentados os principais resultados obtidos a partir da aplicação do algoritmo K-Means. A análise permitiu identificar padrões relevantes na ocorrência de doenças, considerando variáveis como bairro, faixa etária e gênero. O conjunto de dados incluiu 4.050 registros, abrangendo pacientes atendidos entre os anos de 2022 e 2023 no Hospital Geral do Catapas.

A aplicação de técnicas de ML na saúde deve considerar aspectos éticos, como a privacidade dos pacientes e a equidade nos modelos preditivos. Estudos têm mostrado que algoritmos podem apresentar viés racial e influência do bairro, afetando a equidade no atendimento . Além disso, desafios como a qualidade dos dados e a resistência à adoção de novas tecnologias podem impactar a eficácia das soluções propostas.

A comparação entre as técnicas de ML aplicadas na nossa pesquisa e os estudos existentes demonstra que nossa abordagem está alinhada com as práticas atuais na otimização da gestão hospitalar.

4.1 Análise dos Resultados

A análise dos dados obtidos permitiu observar padrões significativos relacionados à incidência de doenças em diferentes bairros da cidade do Uíge, considerando variáveis como faixa etária, gênero e tipos de bairros. Com a aplicação dos algoritmos de clusterização K- Means, foi possível segmentar os dados em grupos distintos, revelando concentrações específicas de doenças em determinados bairros, o que facilita a tomada de decisão por parte das autoridades de saúde pública.

- *Análise por Bairro*

A clusterização revelou que bairros periurbanos apresentaram maior concentração de doenças parasitárias, especialmente entre crianças e adolescentes. Por exemplo, o bairro Catapa concentrou aproximadamente 58% dos casos de malária e febre. O bairro Catapa, isoladamente, responde por 37% dos casos registrados, 15% do bairro Mbemba Ngango, 10% para o bairros Kindenuku, 5% do bairros Dunga, 3% para o bairro Papelão e 30% dos casos é o total de percentagens dos 34 bairros restantes.

- *Análise por Faixa Etária*

A faixa etária de 0 a 10 anos de idade representam 32,3% dos casos, a faixa etária de 11 a 17 anos de idade correspondem 14,9% dos casos, a faxia etária de 18 a 45 anos representou cerca de 38,4% dos casos registrados, a faixa etária de 46 a 59 representam 11,8% dos casos e dos 60 anos para diante correspondem a 2,6% dos casos. A maioria dos pacientes agrupados nos clusters 0 e 1 estava nessa faixa, indicando vulnerabilidade elevada.

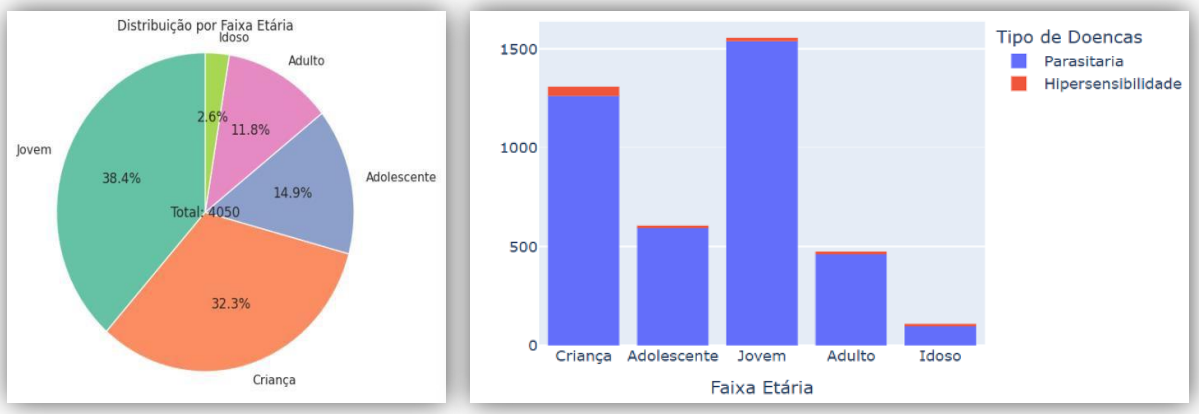


Figura 5: Distribuição da faixa etária e tipo de doença

- *Análise por Gênero e Tipo de Bairro*

Em geral, observou-se predominância do sexo masculino com maior número de pacientes, cerca de 50.1% dos registros pertencem a homens e 49,9% representam as mulheres. Quanto ao tipo de bairro, observou-se que os bairros Periurbano predominam com ocorrência de 73.6%, urbano com 24.6% e rurais com 1.9%, assim como ilustra a figura 7

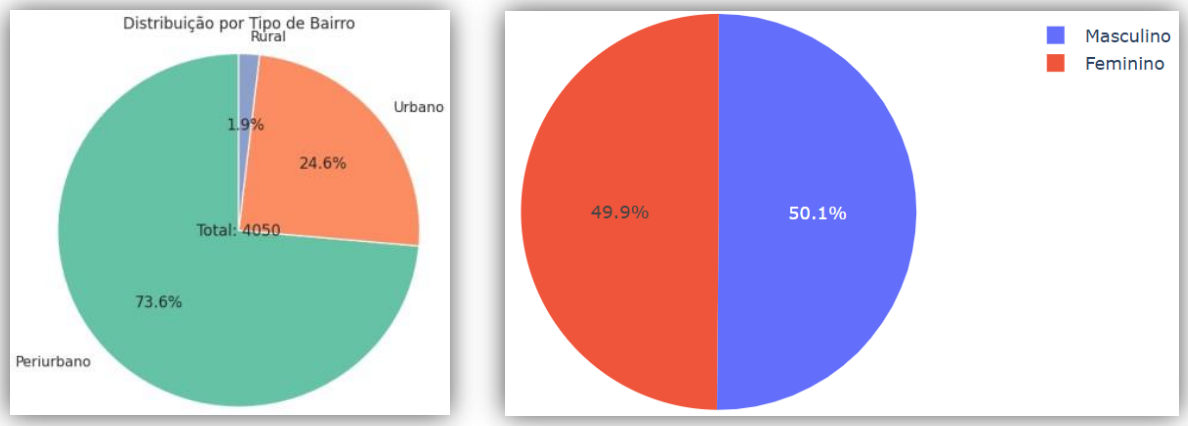


Figura 6: Agrupamento das Tipo de bairro e genero por faixa

- *Análise por Mês*

Em todos os clusters, observamos que predominância é para o mês de junho o período com maior incidência, com cerca de 36% do total de ocorrências, conforme os dados extraídos do livro de registros do Hospital do Catapa.

- *Perfis Identificados*

Foram definidos 5 clusters distintos, representando grupos de pacientes com perfis semelhantes. Cada grupo apresenta padrões característicos quanto à idade, tipo de doença e localização, contribuindo para ações preventivas mais específicas.

- *Visualização Interativa*

O painel criado com Dash permite visualizar os dados de forma interativa, oferecendo filtros por bairro, faixa etária, gênero e tipo de doença. Isso facilita a análise por parte dos profissionais de saúde e gestores. Além disso, o painel interativo desenvolvido com Dash se mostrou eficaz na visualização dos padrões de incidência e previsão das doenças. Os gráficos e os filtros aplicados permitiram a exploração dos dados de forma intuitiva, dando suporte às análises tanto para especialistas quanto para gestores de saúde com pouca familiaridade técnica. Esse painel contribui de forma prática para a transformação dos dados em informação útil e aplicável no contexto da gestão hospitalar.

CONSIDERAÇÕES FINAIS

Este trabalho teve como principal objetivo aplicar técnicas de Machine Learning para otimizar a análise e a gestão de dados dno Hospital do Catapa no município do Uíge. Através da implementação de algoritmos de clusterização e agrupmento, demonstrou-se que é possível identificar padrões relevantes na distribuição de doenças, antecipar possíveis surtos e apoiar a tomada de decisão das autoridades sanitárias.

Apesar das limitações relacionadas à coleta manual dos dados, à escassez de registros digitalizados e à ausência de um sistema de gestão hospitalar, os resultados obtidos foram significativos. As técnicas aplicadas mostraram que mesmo com dados obtidos em ambientes com infraestrutura limitada, é possível extrair valor e gerar conhecimento que pode ser utilizado para melhorar a alocação de recursos e ações de prevenção.

O estudo demonstrou que o uso do algoritmo K-Means é eficaz para segmentar populações com base em dados clínicos e demográficos. A análise revelou padrões relevantes na distribuição das doenças, oferecendo suporte à tomada de decisão no contexto hospitalar. O uso de algoritmo K-Means, proporcionou diferentes perspectivas sobre os dados, permitindo não apenas segmentá-los, mas também prever ocorrências com boa precisão. A criação de um painel interativo

contribuiu para tornar essas análises acessíveis e aplicáveis no cotidiano de instituições de saúde.

A remoção de algoritmos adicionais permitiu um foco mais claro na clusterização, sem comprometer a qualidade analítica. O painel interativo ampliou o acesso às informações, promovendo uma gestão baseada em dados.

Recomenda-se, em estudos futuros, explorar outros algoritmos com novos conjuntos de dados e comparar os resultados com os obtidos pelo K-Means. Além disso, a ampliação da base de dados para incluir anos posteriores pode enriquecer as análises. Como trabalhos futuros, recomenda-se também a ampliação da base de dados, integração com sistemas de saúde em tempo real, e a inclusão de variáveis ambientais, como clima e saneamento, para aumentar a capacidade preditiva dos modelos. A continuidade deste trabalho poderá contribuir significativamente para o avanço da saúde pública baseada em dados no país.

REFERÊNCIAS

Filho, C. R. (2020). TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL APLICADAS AOS PROCESSOS DE GESTÃO HOSPITALAR . 11.

Marius Sumanas, A. P. (21 de Maio de 2022). Deep Q-Learning in Robotics: Improvement of Accuracy. p. 16.

Maronna, R. A. (2017). Norman Matloff (2017): statistical regression and classification: from linear models to machine learning. *classification: from linear models to machine learning*(National University of La Plata).

Mechelli, A. (2019). Methods and Applications ML. *Machine Learning*.

Mitchell, T. (1997). Machine Learning. *1*(Machine Learning is the study of computer algorithms that improve automatically through experience.), 9.

Mohri, M. (2018). Foundations of Machine Learning. *second edition*, 505.

Neves, S. A. (2018). Técnicas de Aprendizado de Máquina . *TFC, I*(Classificação da Qualidade de Pavimentos Asfálticos), 49.

Pedro, N. (2024). Utilização de técnicas de aprendizagem automática em contexto académico para tipificação do risco de abandono escolar . *Thesis_MSC_Final_Rev*, 91.

RUSSELL, S., & NORVIG. (2016). Inteligência Artificia. *Pearson*.

Samuel, A. M. (2011). *Proposta metodológica para resolução das equações redutíveis ao segundo grau lR. Caso das equações biquadráticas e irracionais.* Isced, Uíge, Agola.
SHAPIRO, S. C. (1992). Encyclopedia of artificial intelligence.

Sutton, R. S. (2018). Reinforcement Learning. *An Introduction, Second edition*(Complete Draft), 38. Zhang, W. (2024). International Journal of Mental Health Promotion. p. 26.

.

AGRADECIMENTOS

Endereçamos os nossos agradecimentos ao Departamento do Ensino e Investigação do Instituto Politécnico da Universidade Kimpa Vita no Uíge, que nos proporciou as condições necessárias para a concretização deste estudo

Ao Hospital do Catapa, expressamos a nossa minha sincera gratidão pela colaboração e pelo apoio prestado durante o desenvolvimento deste trabalho. A abertura para o acesso aos dados e o acolhimento por parte da equipe foram fundamentais para a realização da pesquisa e para o enriquecimento dos resultados aqui apresentados.

Ao **Sebastião Afonso**, agradecemos profundamente pelo companheirismo, pela troca de conhecimentos, pela amizade construída ao longo desta etapa, pelo apoio intelectual, pela orientação e pelo suporte moral nos momentos mais desafiadores.