Aplicação de Machine Learning no Diagnóstico da Peste des Petits Ruminants: Uma Abordagem Comparativa entre Classificação e Clusterização

Rafael L. Araújo[1,2], Vitor R. F. Da Silva[1], Francisco E. Santos[1], Anthony I. M. Luz[2], Romuere R. V. e Silva[2]

[1]Instituto Federal do Piauí (IFPI) – Picos, PI – Brasil
[2]Universidade Federal do Piauí (UFPI) – Picos, PI – Brasil

rafaluzaraujo@ifpi.edu.br, romuere@ufpi.edu.br

Abstract. *Peste des Petits Ruminants (PPR) is an infectious disease affecting goats and sheep, causing significant economic and health impacts. Traditional diagnostic methods, such as RT-qPCR, are accurate but require time and specialized infrastructure. This study evaluates the use of Machine Learning techniques to support the diagnosis of PPR based on clinical data. Classification and clustering models were applied to identify patterns associated with the presence of the disease. Gradient Boosting achieved the best predictive performance, while clustering analysis revealed relevant structures in the dataset. The findings suggest the potential of these approaches to support the detection and monitoring of PPR.*

Resumo. *A Peste dos Pequenos Ruminantes (PPR) é uma doença infecciosa que afeta caprinos e ovinos, causando impactos econômicos e sanitários significativos. Métodos tradicionais de diagnóstico, como o RT-qPCR, embora precisos, demandam tempo e infraestrutura. Este estudo avalia o uso de técnicas de Machine Learning para auxiliar no diagnóstico da PPR a partir de dados clínicos. Foram aplicados modelos de classificação e clusterização para identificar padrões associados à presença da doença. O Gradient Boosting obteve os melhores resultados preditivos, enquanto a análise por clusterização indicou estruturas relevantes nos dados. Os achados apontam para o potencial dessas abordagens no apoio à detecção e monitoramento da PPR.*

1. Introdução

A peste dos pequenos ruminantes (*Peste des Petits Ruminants* - PPR) é uma doença infecciosa altamente contagiosa que afeta cabras e ovelhas, causando significativas perdas econômicas e ameaçando a segurança alimentar em regiões vulneráveis [Banyard et al. 2010]. Identificar e controlar rapidamente os surtos de PPR é essencial para minimizar o impacto da doença. Métodos tradicionais de diagnóstico incluem a observação clínica e testes laboratoriais como o RT-qPCR, que, apesar de precisos, podem ser demorados e requerem infraestrutura laboratorial sofisticada [Banyard et al. 2010].

Nos últimos anos, técnicas de *Machine Learning* (ML) têm sido aplicadas em diagnósticos médicos para melhorar a acurácia e a eficiência. Estudos recentes demonstram o potencial dos classificadores como *Logistic Regression*, *Decision Tree* e *Random Forest*

na predição de doenças infecciosas [Nguyen et al. 2020, Kohli et al. 2017]. Além disso, algoritmos de clusterização, como o *K-Means*, são utilizados para identificar padrões em dados médicos e segmentar populações em grupos com características similares, facilitando a análise epidemiológica [Xu and Tian 2015]. No entanto, a escolha do número de clusters é crítica e métodos como *Elbow Method* e *Silhouette Score* são comumente empregados para este fim [Kodinariya and Makwana 2013].

Neste sentido, este estudo investiga a aplicação de técnicas de *Machine Learning* no diagnóstico da Peste dos Pequenos Ruminantes (PPR), por meio da comparação entre abordagens supervisionadas de classificação e métodos não supervisionados de clusterização. A proposta visa avaliar o desempenho de cada abordagem na identificação da doença com base em dados clínicos, destacando as potencialidades e limitações de cada técnica no contexto veterinário e epidemiológico.

2. Revisão da Literatura

Nos últimos anos, técnicas de *Machine Learning* (ML) têm sido exploradas para aprimorar o diagnóstico da PPR. [Myagila et al. 2023] desenvolveram modelos preditivos utilizando algoritmos como *Logistic Regression* e *Support Vector Machines*, aplicados a dados clínicos de ovinos e caprinos na Tanzânia. Os autores utilizaram a mesma base de dados deste estudo e obtiveram como melhor resultado o modelo de *Logistic Regression*, com acurácia, precisão e sensibilidade de 79%. Embora os modelos tenham mostrado desempenho promissor, os autores destacam desafios na distinção entre sintomas de PPR e outras doenças semelhantes, indicando a necessidade de abordagens mais robustas.

Apesar de uma busca sistemática na literatura, identificamos que apenas o estudo de [Myagila et al. 2023] utiliza a mesma base de dados e se propõe a resolver diretamente o problema de diagnóstico clínico da PPR por meio de aprendizado de máquina. Os demais trabalhos encontrados tratam da PPR sob diferentes perspectivas, o que impossibilita uma comparação direta com este estudo. Por exemplo, [Niu et al. 2021] integraram algoritmos de *Random Forest* com dados meteorológicos para prever surtos de PPR em escala global. Apesar de fornecerem insights valiosos sobre padrões de disseminação, o estudo não abordou diretamente a aplicação de ML no diagnóstico clínico individual da doença.

[Sobeih et al. 2020] realizaram análises de *cluster* e *hotspot* para mapear a incidência da PPR em Bangladesh, utilizando dados de registros veterinários. Embora eficazes na identificação de áreas de risco, os métodos empregados não se concentraram na predição de casos individuais com base em sintomas clínicos. Por fim, [Walsh et al. 2021] discutem o potencial da inteligência artificial na epidemiologia veterinária, destacando sua capacidade de auxiliar no monitoramento e análise de doenças animais, especialmente em situações com recursos laboratoriais limitados. No entanto, enfatizam a necessidade de validar essas abordagens em diferentes contextos epidemiológicos.

Esses estudos evidenciam o potencial das técnicas de ML no diagnóstico e monitoramento da PPR, mas também revelam lacunas, como a necessidade de modelos mais precisos para distinguir PPR de outras doenças com sintomas semelhantes e a validação dessas abordagens em diferentes contextos regionais. O presente estudo busca contribuir para essa área, explorando modelos de classificação e clusterização aplicados a dados clínicos de PPR, visando aprimorar a detecção e o monitoramento da doença.

3. Metodologia

A metodologia deste trabalho propõe uma comparação entre abordagens de classificação e clusterização aplicadas ao diagnóstico da PPR. O processo segue o fluxo ilustrado na Figura 1, iniciando pela aquisição e preparação dos dados, seguido pela aplicação e análise comparativa dos métodos de classificação e clusterização.

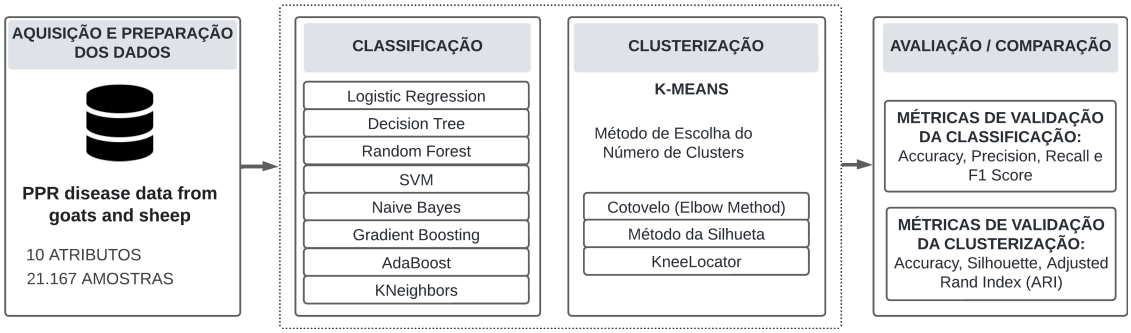


Figura 1. Metodologia proposta.

3.1. Aquisição e Preparação da Base de Dados

Para este estudo, utilizou-se a base de dados *PPR disease data from goats and sheep* [Thanyambo 2023], composta por 161 amostras clínicas coletadas na zona norte da Tanzânia. As amostras foram obtidas por especialistas em PPR, que avaliaram ovinos e caprinos quanto a sintomas clínicos e confirmaram os casos suspeitos por meio de testes rápidos RT-qPCR. As variáveis coletadas estão apresentadas na Tabela 1.

Tabela 1. Informações sobre a base de dados PPR Disease Data

Variável	Tipo	Descrição
Temperatura (°C)	Contínua	Temperatura corporal do animal
Secreção nasal	Categórica	Presença de descarga nasal (Sim/Não)
Diarreia	Categórica	Estado das fezes (Normal/Diarreia)
Respiração Difícil	Categórica	Dificuldade na respiração (Sim/Não)
Idade	Contínua	Idade do animal em anos
Descarga Ocular	Categórica	Presença de descarga ocular (Sim/Não)
Lesão Oral/Nasal	Categórica	Presença de lesões orais e nasais (Sim/Não)
Animal	Categórica	Espécie do animal (Ovelha/Cabra)
Sexo	Categórica	Sexo do animal (Macho/Fêmea)
Resultado	Categórica	Resultado do teste rápido (Positivo/Negativo)

No total, apenas 12 amostras foram positivas para PPR, evidenciando um forte desequilíbrio de classes. Para mitigar esse problema e possibilitar a construção de modelos mais robustos, [Thanyambo 2023] aplicou a técnica de geração de dados sintéticos *Conditional Tabular GAN* (CTGAN) [Xu et al. 2019], com o objetivo de aumentar a representatividade dos casos positivos. Após esse processo, o conjunto de dados foi expandido para 21.167 amostras, mantendo características estatísticas próximas ao conjunto original. Por fim, os dados foram particionados em conjuntos de treinamento e teste por meio da técnica de validação cruzada *k-fold*, garantindo uma avaliação robusta e imparcial dos modelos de classificação e clusterização.

3.2. Classificação

Nesta etapa, o objetivo foi desenvolver modelos preditivos capazes de identificar a presença da PPR com base em sintomas clínicos observados. Para isso, foram utilizados sete algoritmos de classificação, cada um com características distintas e consolidados na literatura. O modelo *Logistic Regression* foi aplicado por sua capacidade de modelar a probabilidade de ocorrência de um evento binário a partir de uma função logística, sendo especialmente eficaz em problemas lineares [Cox 1958]. O *Decision Tree* foi utilizado pela simplicidade interpretativa das regras geradas ao segmentar os dados com base em seus atributos, enquanto o *Random Forest*, composto por múltiplas árvores construídas a partir de subconjuntos aleatórios de dados e características, foi empregado por sua robustez e maior precisão [Breiman 2001].

O *Support Vector Machine (SVM)*, encontra o hiperplano ótimo de separação entre as classes e se destaca em problemas complexos [Cortes and Vapnik 1995], e o *Naive Bayes*, um modelo probabilístico baseado no Teorema de Bayes que assume independência entre os atributos [McCallum and Nigam 1998]. Além disso, foram utilizados o *Gradient Boosting* e o *AdaBoost*, técnicas baseadas no método de *boosting*, que combinam iterativamente modelos fracos com o objetivo de construir um modelo preditivo mais robusto e preciso [Friedman 2001]. Por fim, o *k-Nearest Neighbors (k-NN)*, que classifica as instâncias com base nas k amostras mais próximas no espaço de atributos, sendo eficaz em padrões não lineares [Cover and Hart 1967]. Para todos os algoritmos, foram adotadas as configurações padrão de hiperparâmetros disponíveis na biblioteca *scikit-learn*.

Utilizou-se a técnica de validação cruzada *k-fold* com k=5, que divide o conjunto de dados em cinco subconjuntos (*folds*), utilizando quatro para o treinamento e um para o teste em cada iteração. As métricas de avaliação adotadas foram a *Accuracy*, que mede a proporção de previsões corretas sobre o total de amostras; a *Precision*, que calcula a proporção de verdadeiros positivos entre todas as previsões positivas; o *Recall* (ou sensibilidade), que corresponde à proporção de verdadeiros positivos em relação ao total de casos positivos reais; e o *F1-Score*, definido como a média harmônica entre precisão e *recall*, equilibrando ambas as métricas [Kohavi et al. 1995]. Cada modelo foi avaliado múltiplas vezes, e os resultados foram expressos em termos de média e desvio padrão para garantir uma análise comparativa robusta e confiável.

3.3. Clusterização

Neste estudo, o algoritmo *K-Means* foi utilizado com o propósito de agrupar as amostras com base em similaridade clínica, possibilitando uma comparação indireta com os modelos supervisionados. Embora o *K-Means* seja tradicionalmente aplicado como uma técnica não supervisionada voltada à identificação de padrões e segmentação de dados em grupos homogêneos, aqui ele foi utilizado com diferentes valores de k, com posterior alinhamento dos rótulos gerados aos rótulos reais, permitindo assim a avaliação do agrupamento por métricas de classificação. Esse processo de mapeamento possibilita analisar, de forma adaptada, o quanto a estrutura dos dados clínicos se aproxima das classes verdadeiras da doença, mesmo sem supervisão explícita.

O algoritmo *K-Means* agrupa os dados minimizando a variância *intra-cluster*, ou seja, a distância entre os pontos e o centróide do grupo ao qual pertencem [MacQueen 1967]. A escolha do número de *clusters* k é uma etapa crítica nesse processo, sendo

realizada por meio de três métodos distintos: o *Elbow Method*, que analisa a curva da soma das distâncias quadradas *intra-cluster* (WCSS) para detectar o ponto de inflexão que indica o número ideal de agrupamentos [Thorndike 1953]; o *Silhouette Method*, que avalia a coesão e a separação entre os *clusters* com base na pontuação média da silhueta [Rousseeuw 1987]; e o algoritmo *KneeLocator*, que realiza a detecção automática do ponto de inflexão na curva WCSS [Satopaa et al. 2011].

A qualidade dos agrupamentos foi avaliada por meio de três métricas principais. A *Accuracy*, que mensura a proporção de instâncias corretamente agrupadas quando comparadas às classes verdadeiras; o *Silhouette Scor*e, que varia de -1 a 1 e expressa a definição dos *clusters* com base na distância entre elementos e centróides [Rousseeuw 1987]; e o *Adjusted Rand Index (ARI)*, que compara a clusterização gerada com uma classificação de referência, ajustando o resultado para correlação ao acaso [Hubert and Arabie 1985].

4. Resultados e Discussões

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da aplicação dos modelos de classificação e clusterização aos dados clínicos da PPR. A análise busca avaliar o desempenho preditivo de cada abordagem e compreender sua eficácia no diagnóstico da doença a partir de diferentes perspectivas de aprendizado de máquina.

4.1. Resultados da Classificação

Os resultados obtidos para os modelos de classificação são apresentados na Tabela 2. Embora os resultados gerais tenham sido próximos entre os modelos, o *Gradient Boosting* destacou-se com o melhor desempenho. Em contrapartida, os algoritmos *k-Nearest Neighbors* e *Naive Bayes* obtiveram as menores pontuações nas métricas avaliadas. O desempenho inferior do *KNeighbors* pode estar relacionado à sua sensibilidade a ruídos e ao impacto do desequilíbrio de classes na base de dados, enquanto o *Naive Bayes* possivelmente foi prejudicado pela suposição de independência entre atributos, que não se sustenta em dados clínicos com correlações relevantes entre variáveis.

Tabela 2. Resultados dos Modelos de Classificação

Modelo	Accuracy	Precision	Recall	F1 Score
Logistic Regression	80,67 ± 0,6	80,41 ± 0,6	80,67 ± 0,6	80,38 ± 0,6
Decision Tree	80,44 ± 0,7	80,20 ± 0,7	80,44 ± 0,7	80,24 ± 0,7
Random Forest	80,50 ± 0,7	80,29 ± 0,7	80,50 ± 0,7	80,34 ± 0,7
SVM	80,72 ± 0,5	80,50 ± 0,5	80,72 ± 0,5	**80,54 ± 0,5**
Naive Bayes	78,32 ± 0,5	78,05 ± 0,5	78,32 ± 0,5	78,12 ± 0,5
Gradient Boosting	**80,75 ± 0,6**	**80,52 ± 0,6**	**80,75 ± 0,6**	80,48 ± 0,6
AdaBoost	80,73 ± 0,7	80,46 ± 0,7	80,73 ± 0,7	80,41 ± 0,6
KNeighbors	77,98 ± 1,7	77,87 ± 1,5	77,98 ± 1,7	77,83 ± 1,6

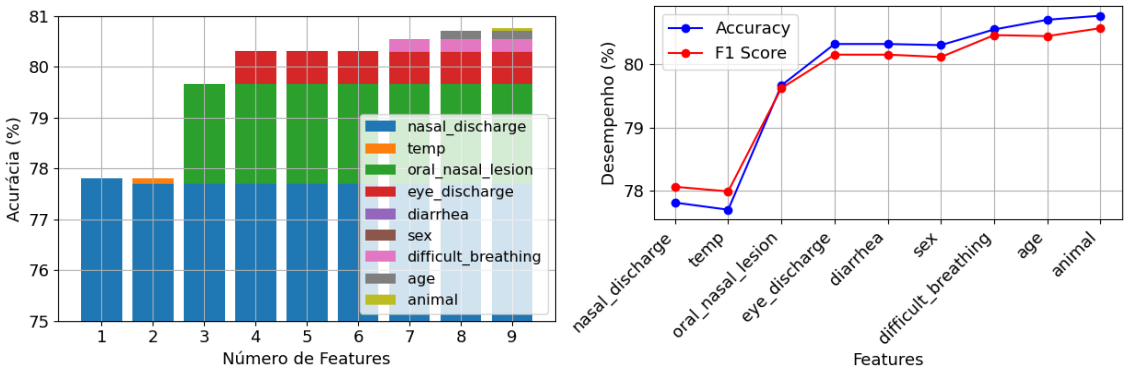
O modelo *Gradient Boosting*, que obteve o melhor desempenho entre os classificadores, identificou corretamente 2.340 casos positivos (verdadeiros positivos) e 1.083 negativos (verdadeiros negativos), com 378 falsos negativos e 438 falsos positivos. Esse resultado demonstra alta sensibilidade, reduzindo o risco de não detectar animais infectados, algo crítico no controle da PPR. A baixa taxa de falsos positivos também indica boa especificidade, evitando alarmes e intervenções desnecessárias.

Para identificar a importância de cada atributo na predição da PPR, realizou-se uma análise no ranqueamento de atributos do *Random Forest*, devido à sua ampla adoção e à facilidade de interpretação das importâncias atribuídas aos atributos. A Tabela 3 apresenta os principais atributos ordenados por sua importância segundo o *Random Forest*.

Tabela 3. Importância dos atributos segundo o Random Forest

Atributo	Importância
Secreção nasal	0,4805
Temperatura	0,1611
Lesão oral ou nasal	0,1239
Secreção ocular	0,0677
Diarréia	0,0402
Sexo	0,0391
Dificuldade respiratória	0,0382
Idade	0,0261
Animal	0,0232

A análise de importância dos atributos com *Random Forest* indicou que secreção nasal (48,05%), temperatura corporal (16,11%) e lesões orais ou nasais (12,39%) são os principais preditores da PPR, destacando-se como fortes indicadores clínicos da doença. Os demais atributos apresentaram menor relevância. A priorização dos três atributos mais relevantes pode otimizar o diagnóstico, direcionar melhor os recursos e aprimorar as intervenções veterinárias. Para validar essa hipótese, realizou-se um teste incremental com o algoritmo *Gradient Boosting*, adicionando os atributos um a um conforme sua ordem de importância, a fim de avaliar o impacto de cada variável no desempenho do modelo. Os resultados desse teste estão apresentados na Figura 2.



(a) (b)

Figura 2. Desempenho do Gradient Boosting com adição incremental de atributos: (a) accuracy (barras) e (b) accuracy e F1-score (linhas).

Os resultados do teste incremental revelaram que a secreção nasal, por si só, atingiu uma *accuracy* próxima de 78%, destacando-se como um forte indicador isolado da PPR. A lesão oral ou nasal, quando adicionada, aumentou a acurácia para próximo de 80%, reforçando sua relevância no diagnóstico da doença. Por outro lado, a temperatura, apesar de ser apontada como importante pelo modelo de *Random Forest*, mostrou-se contraproducente no modelo de *Gradient Boosting*, reduzindo a acurácia.

A discrepância observada com a temperatura pode ser explicada por alguns fatores. Primeiramente, a temperatura corporal pode ser influenciada por várias condições diferentes, não sendo específica apenas para a PPR. Isso pode introduzir ruído no modelo de *Gradient Boosting*, que é sensível a variações e *outliers*. Além disso, a combinação de temperatura com outros atributos pode não ter sinergia positiva, ao contrário do que ocorre com a secreção nasal e as lesões orais ou nasais. Esses resultados sugerem que, embora a temperatura seja um indicador relevante, sua inclusão deve ser avaliada com cautela, considerando o impacto negativo potencial na performance de alguns modelos.

Em comparação com a literatura, o estudo de [Myagila et al. 2023] obteve 79% de acurácia com o modelo de *Logistic Regression*. Os modelos desenvolvidos neste trabalho apresentaram desempenho superior, alcançando 81% de acurácia, além de resultados mais robustos devido à aplicação de validação cruzada. Esses achados reforçam a importância da seleção adequada de atributos na predição da PPR e evidenciam o *Gradient Boosting* como a abordagem de classificação mais eficaz para este conjunto de dados.

4.2. Resultados da Clusterização

Os resultados da clusterização são apresentados na Tabela 4. Ao utilizar a clusterização com 2 *clusters* para classificar os dados da PPR, a *accuracy* foi de 74,05%, consideravelmente inferior ao modelo de classificação com *Gradient Boosting*. Embora a acurácia seja relativamente alta, o coeficiente *Silhouette* de 0,1965 e o *ARI* de 0,2287 indicam que a separação dos clusters não é ideal.

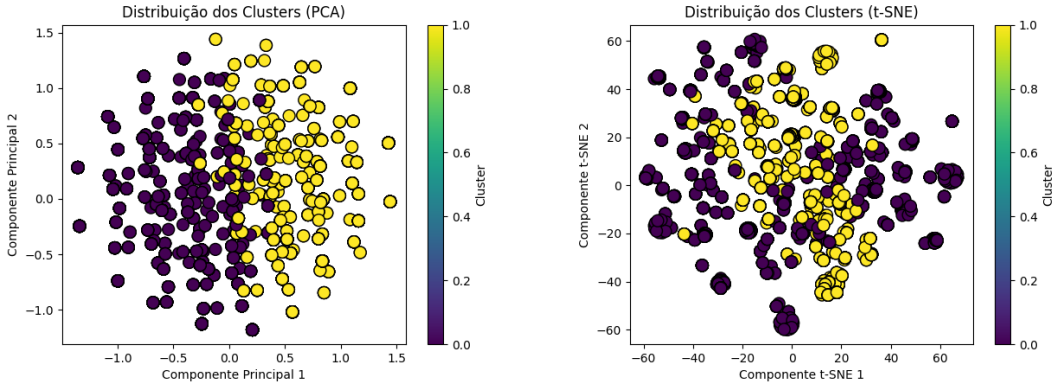
Clusters	Escolha de Clusters	Accuracy	Silhouette	ARI
2	Nº de Classes	74,05	0,1965	**0,2287**
4	Cotovelo - KneeLocator	70,14	0,2226	0,0823
10	Silhouette	**76,67**	**0,2477**	0,0569
6	Cotovelo - soma dos quadrados	75,80	0,2307	0,0716

Tabela 4. Resultados da clusterização.

Esses resultados sugerem uma sobreposição significativa entre os grupos, mostrando que a clusterização com apenas dois grupos não conseguiu capturar adequadamente as nuances e variações presentes nas classes. A clusterização, por natureza, agrupa dados baseados em similaridades e proximidade no espaço multidimensional, mas isso não necessariamente coincide com a lógica subjacente às classes de PPR, resultando em baixa acurácia. A visualização dos dados utilizando PCA e t-SNE presentes na Figura 3 revela mais detalhes sobre essa sobreposição: no PCA, as duas classes se misturam ao centro, enquanto no t-SNE, uma classe fica concentrada ao centro e a outra se divide entre a esquerda e a direita. Isso evidencia ainda mais a dificuldade de separação clara entre as classes quando se utiliza apenas dois *clusters*.

Adicionalmente, realizou-se testes para identificar o melhor número de *clusters* para a segmentação dos dados. Observou-se que o melhor resultado foi obtido com 10 *clusters*, determinado pelo método da *silhouette*. Este resultado sugere que os dados possuem subgrupos internos mais complexos que não são bem representados por apenas duas classes.

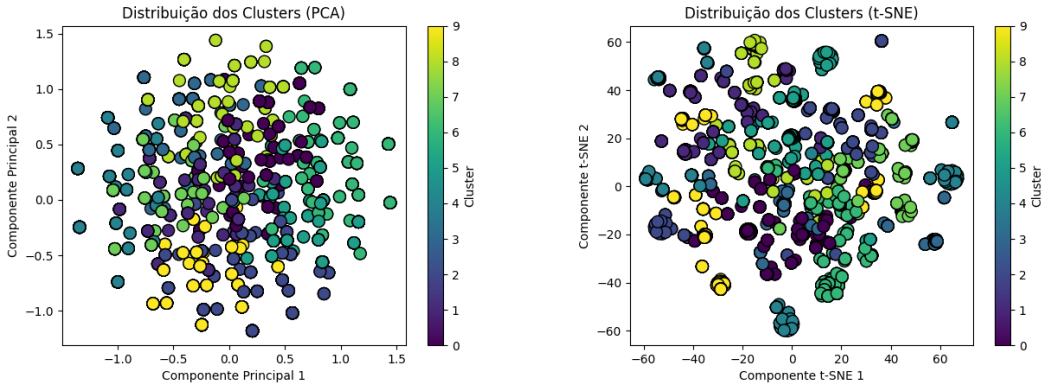
Com mais clusters, é possível identificar variações sutis e diferenças dentro das



(a) (b)

Figura 3. Visualização dos dados com dois clusters. (a) PCA e (b) t-SNE.

classes, permitindo uma segmentação mais refinada. Isso ocorre porque os dados de duas classes podem ser muito semelhantes, e a divisão em múltiplos clusters ajuda a discernir essas pequenas diferenças, resultando em uma melhor representação das estruturas internas dos dados e, consequentemente, melhor desempenho na predição. A Figura 4 mostra o PCA e o t-SNE para 10 clusters e evidencia esses subgrupos: no PCA, os subgrupos ficam mais juntos, enquanto no t-SNE, um subgrupo (como o amarelo) se divide em quatro partes. Essa divisão torna a separação entre subgrupos mais difícil, destacando a complexidade inerente dos dados.



(a) (b)

Figura 4. Visualização dos dados com dez clusters. (a) PCA e (b) t-SNE.

4.3. Limitações

Este estudo apresenta limitações relacionadas à natureza dos dados utilizados e à generalização dos resultados. A base de dados foi majoritariamente composta por amostras sintéticas, geradas por meio da técnica CTGAN a partir de um conjunto real bastante reduzido, contendo apenas 12 casos positivos de PPR. Essa escassez de dados reais compromete a representatividade e a diversidade clínica dos exemplos positivos, o que pode afetar negativamente a validade externa dos modelos desenvolvidos.

Adicionalmente, observou-se que o modelo *Gradient Boosting* demonstrou sensibilidade a *outliers* presentes em variáveis contínuas, como a temperatura corporal. Esse comportamento pode comprometer a precisão do diagnóstico em cenários clínicos com medições extremas ou ruidosas, indicando a necessidade de um tratamento mais criterioso desses valores. Por fim, destaca-se a necessidade de validação dos modelos em diferentes contextos epidemiológicos e sob variações ambientais, o que demanda investigações complementares para garantir sua robustez e aplicabilidade em situações reais.

5. Conclusão

Este estudo explorou a aplicação de técnicas de Machine Learning no contexto do diagnóstico da Peste dos Pequenos Ruminantes (PPR), utilizando algoritmos de classificação e clusterização sobre dados clínicos de animais. Entre os modelos avaliados, o Gradient Boosting apresentou desempenho promissor, com resultados consistentes nas métricas analisadas, sugerindo sua capacidade de capturar padrões relevantes a partir de sintomas como secreção nasal, temperatura e lesões orais. A análise de importância dos atributos reforçou o potencial diagnóstico dessas variáveis clínicas. Complementarmente, a aplicação do K-Means permitiu observar estruturas latentes nos dados, indicando a possível existência de subgrupos dentro das classes, o que evidencia a complexidade e a heterogeneidade do conjunto analisado. Embora os resultados de clusterização tenham sido inferiores aos dos classificadores supervisionados, a abordagem se mostrou útil como ferramenta exploratória no apoio à análise epidemiológica.

Para estudos futuros, recomenda-se a ampliação da base de dados com amostras provenientes de diferentes regiões afetadas pela PPR, o que pode contribuir para a generalização dos modelos desenvolvidos. A inclusão de variáveis geográficas e ambientais também pode fornecer uma visão mais abrangente dos fatores associados à disseminação da doença. Além disso, a adoção de técnicas de interpretação de modelos, como SHAP ou LIME, pode enriquecer a compreensão sobre a influência de cada variável na decisão dos modelos, favorecendo o desenvolvimento de estratégias mais direcionadas e fundamentadas para o controle e prevenção da PPR.

Referências

Banyard, A. C., Parida, S., Batten, C., Oura, C., Kwiatek, O., and Libeau, G. (2010). Global distribution of peste des petits ruminants virus and prospects for improved diagnosis and control. *Journal of General Virology*, 91(12):2885–2897.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Kohli, V., Arora, A., and Dagar, P. (2017). A study on disease prediction using machine learning in healthcare. *International Journal of Information Technology*, 9:119–124.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. 1(14):281–297.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48.

Myagila, K. C., Sabas, J., and Mark, P. N. (2023). Multi-model ppr disease prediction using machine learning algorithms. In *2023 First International Conference on the Advancements of Artificial Intelligence in African Context (AAIAC)*, pages 1–6. IEEE.

Nguyen, D. C., Ding, M., Pathirana, P. N., and Seneviratne, A. (2020). Artificial intelligence in the battle against coronavirus (covid-19): a survey and future research directions. *arXiv preprint arXiv:2008.07343*.

Niu, B., Liang, R., Zhou, G., Zhang, Q., Su, Q., Qu, X., and Zhang, S. (2021). Prediction for global peste des petits ruminants outbreaks based on a combination of random forest algorithms and meteorological data. *Frontiers in Veterinary Science*, 7:570829.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a "kneedle"in a haystack: Detecting knee points in system behavior. pages 166–171.

Sobeih, M. M., Rahman, A. K. M. A., Islam, S. S., Sufian, M. A., Talukder, M. H., Ward, M. P., and Martínez-López, B. (2020). Peste des petits ruminants risk factors and space-time clusters in bangladesh. *Frontiers in Veterinary Science*, 7:572432.

Thanyambo, D. (2023). Ppr disease data from goats and sheep. `https://www.kaggle.com/datasets/devothanyambo/ppr-disease-data-from-goats-and-sheep`. Acesso em: abr. 2025.

Thorndike, R. L. (1953). Who belongs to the family? *Psychometrika*, 18(4):267–276.

Walsh, M. G., Haseeb, M. A., and Mor, S. M. (2021). Prediction for global peste des petits ruminants outbreaks based on a combination of random forest algorithms and meteorological data. *Frontiers in Veterinary Science*, 7:570829.

Xu, D. and Tian, Y. (2015). A comprehensive review of clustering algorithms. *Annals of Data Science*, 2(2):165–193.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *arXiv preprint arXiv:1907.00503*.