**DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA - CTC – UFSC**

**INE5687 - Projeto em Ciência de Dados**

Relatório de andamento do projeto em Ciência de Dados

# Avaliação comparativa de modelos de Machine Learning na classificação de comportamento canino

Matheus Cadorin Luca

Mariana Mazzo Heitor

Yuri Rodrigues de Souza

Outubro de 2025

# Atividades Desenvolvidas

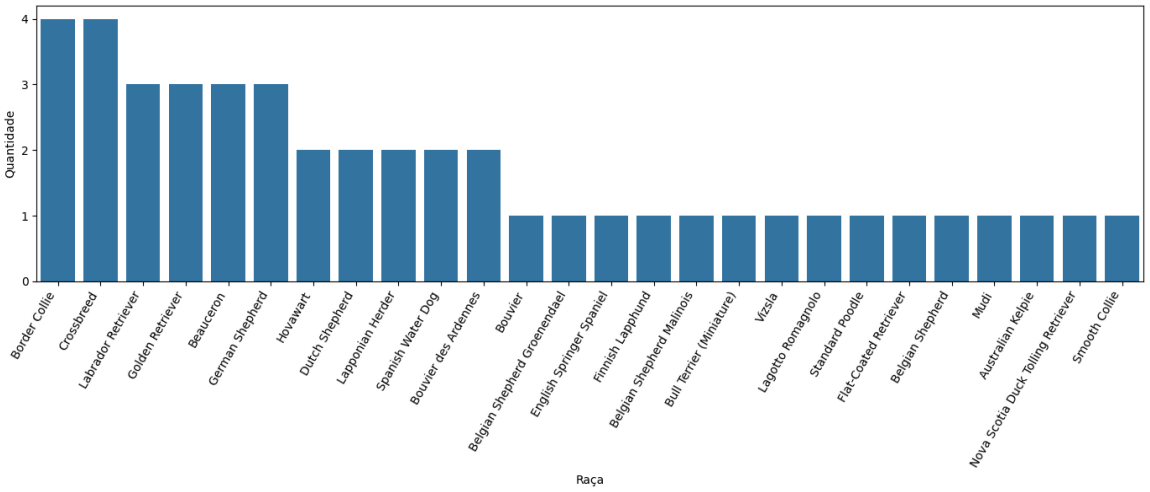
- ## Análise Exploratória dos Dados

Para a análise exploratória dos dados, utilizou-se a biblioteca *MatPlotLib* do *Python* para exibir gráficos informativos a respeito das características e dados importantes para o projeto. Para isso, foram construídos gráficos tanto em relação às informações do animal, a partir do arquivo *DogInfo.csv* quanto às informações de seus sensores e comportamentos observados, a partir do arquivo *DogMoveData.csv*.

Além disso, para possibilitar uma análise individual de cada cachorro, foram agrupados, em outra observação, os dados de *Task*, *Behavior_1*, *Behavior_2* e *Behavior_3* de acordo com o *DogID*, gerando gráficos que mostram os dados correspondentes a essas informações para cada animal.

### Análise de perfil dos cães

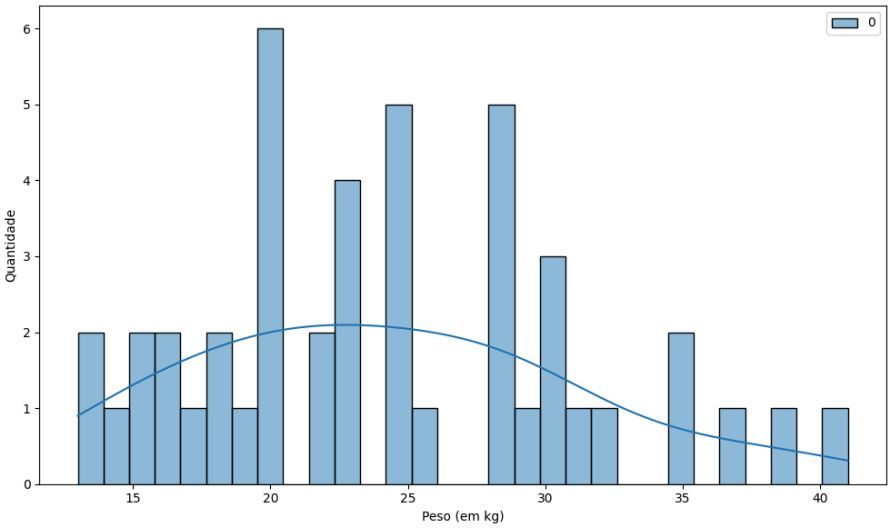
Figura 1 - Gráfico da distribuição das raças de cães




Fonte: Autores

Na Figura 1 é possível observar que existe uma variedade considerável de raças, porém possui baixa representatividade por raça. Além disso, a amplitude no número de indivíduos entre raças é de apenas 3. Logo, os modelos não devem se generalizar somente com os dados de uma raça específica.

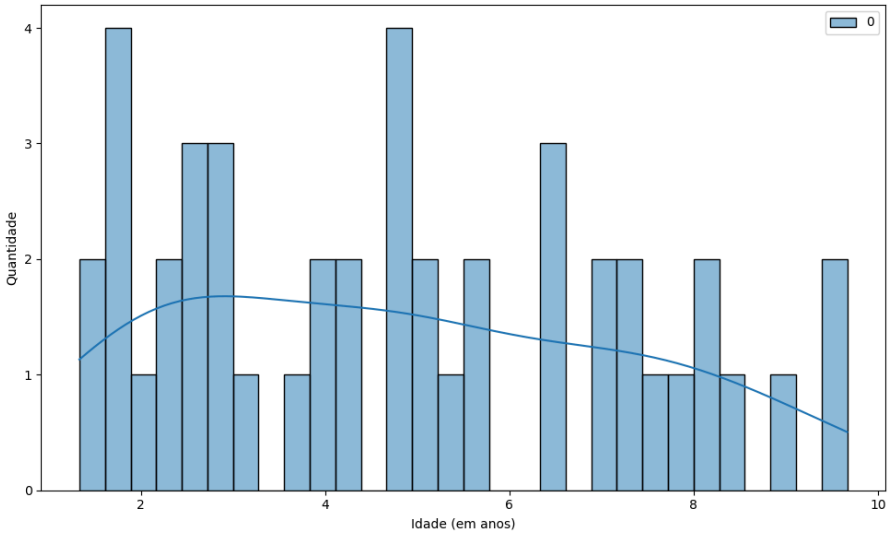
Figura 2 - Gráfico da distribuição do peso dos cães




Fonte: Autores

A Figura 2 revela uma faixa de variação de 13 kg a 41 kg. Observa-se uma concentração maior de cães em torno de 20 kg e também agrupamentos entre 23 a 25 kg e 28 a 30 kg. Isso indica que os modelos treinados tendem apresentar melhor desempenho na inferência para cães desse porte.

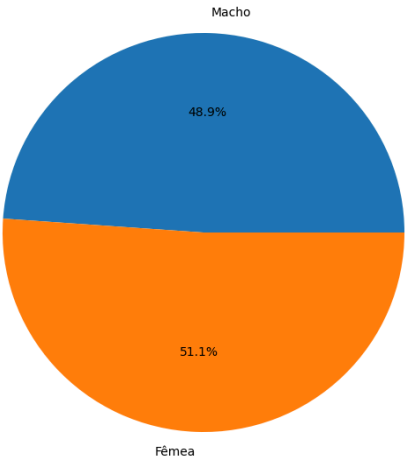
Figura 3 - Gráfico da distribuição da idade dos cães




Fonte: Autores

Já a Figura 3 demonstra uma amostra bem diversificada em termos de faixa etária. A idade se espalha de forma relativamente uniforme, desde jovens adultos (2 anos) até cães mais velhos (9 anos). Essa variedade permite que os modelos aprendam padrões de diferentes estágios da vida de um cão.

Figura 4 - Gráfico da distribuição por sexo dos cães



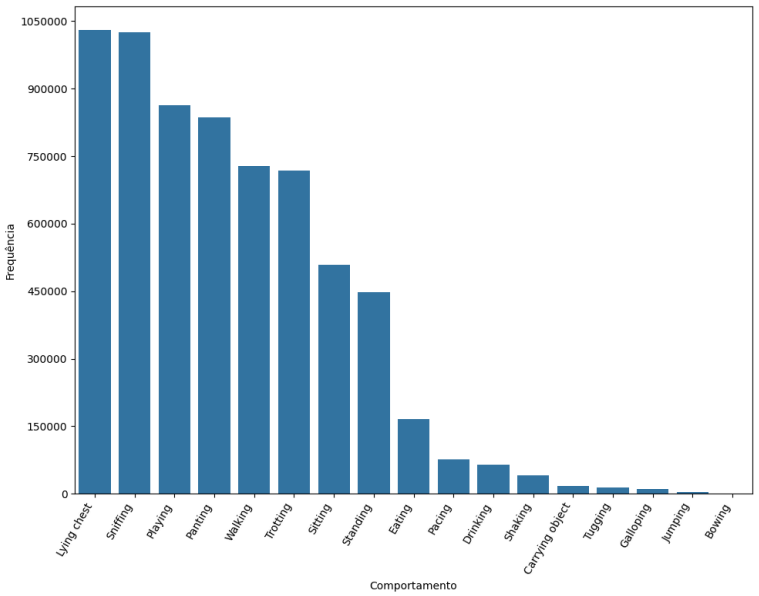

Fonte: Autores

Por fim, a Figura 4 revela um equilíbrio quase perfeito na amostra. Essa paridade é excelente, pois previne que os modelos desenvolvam vieses comportamentais baseado em dimorfismo sexual.

## Análise da variável alvo

A Figura 5 evidencia que o conjunto de dados apresenta um desbalanceamento significativo entre as classes. Sendo assim, deve se considerar o uso de técnicas de resampling, tanto undersampling das classes majoritárias, quanto oversampling das minoritárias. Além disso, pode ser necessário remover algumas classes que se mostram bastante minoritárias.

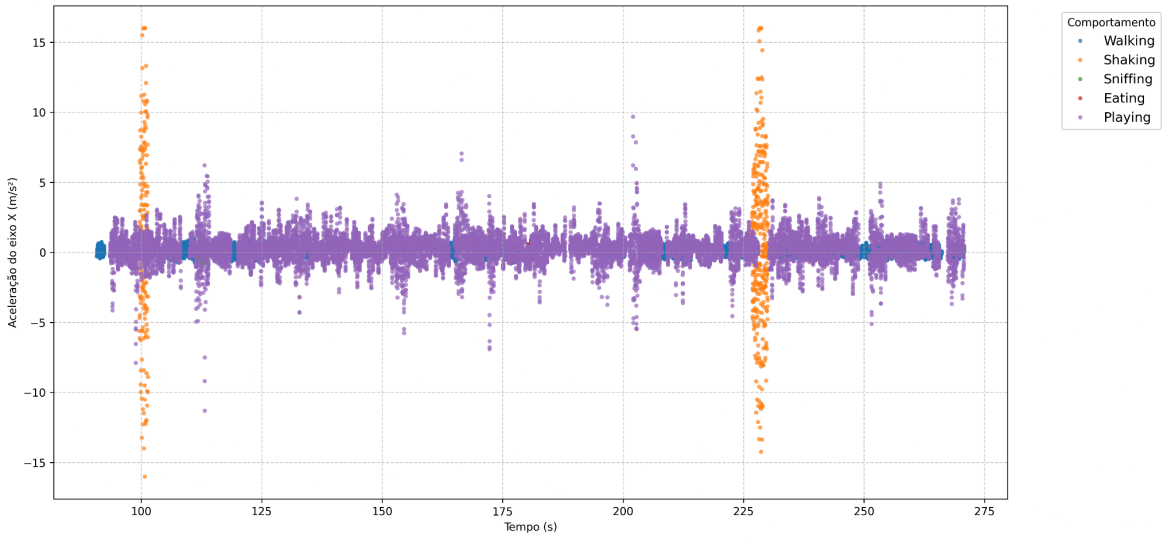
Figura 5 - Gráfico da distribuição de comportamentos - *Behavior_1*




Fonte: Autores

## Análise dos dados dos sensores

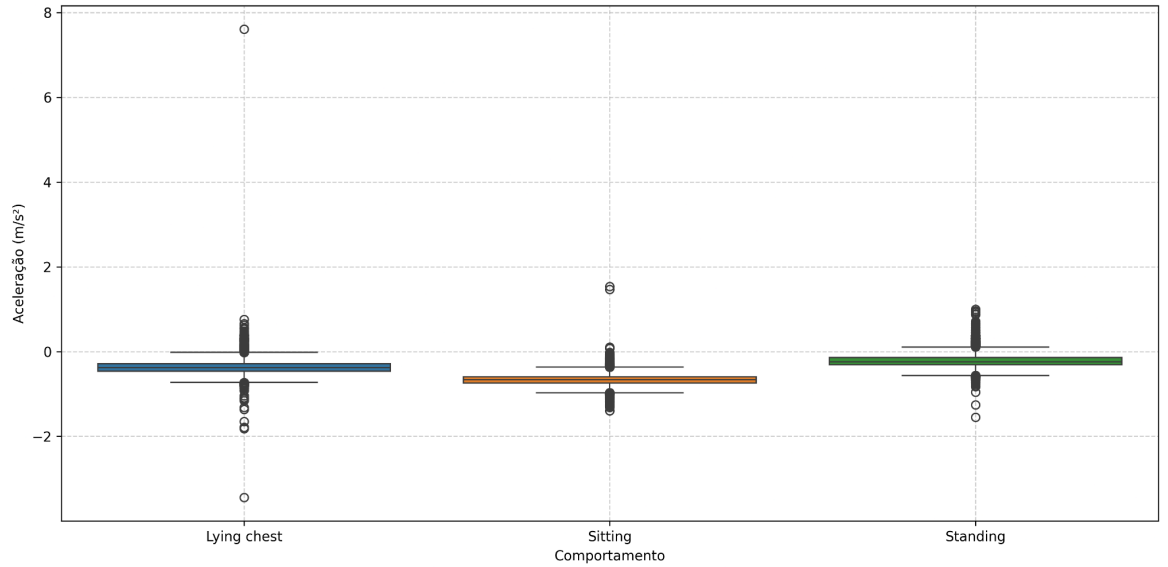
Figura 6 - Gráfico da Aceleração do eixo X das costas ao longo do tempo




Fonte: Autores

A Figura 6 trata-se de um recorte de aproximadamente 3 minutos de um dos testes de um cão e tem como objetivo mostrar os padrões de cada comportamento medido pelo eixo X do acelerômetro das costas. É possível observar visualmente que cada comportamento possui um padrão levemente distinto. Isso nos mostra que as características devem ser calculadas sobre janelas de tempo, visto que são padrões de variabilidade ao longo do tempo que permitem a distinção dos comportamentos.

Figura 7 - Distribuição da aceleração do eixo Y do sensor das costas por comportamentos estáticos




Fonte: Autores

Conforme descrito em outros trabalhos, alguns modelos têm dificuldades em distinguir padrões de comportamentos estáticos. Portanto,

buscamos encontrar variáveis que possuam esse poder discriminativo para compor as features. A figura 7 valida essa busca, sendo possível observar que as medianas estão centradas em valores nitidamente diferentes. Nos permitindo criar features para ajudar o modelo a diferenciar melhor esses comportamentos.

- **Seleção de Algoritmos de Machine Learning**

  Optou-se por utilizar sete algoritmos diferentes na análise comparativa deste projeto, sendo dividido em cinco modelos clássicos e dois modelos de Deep Learning. Para a abordagem clássica, os algoritmos selecionados foram Support Vector Machine (*SVM*), Random Forest (*RF*), Gaussian Naive Bayes (*GNB*), Gradient Boosting e K-Nearest Neighbors (*KNN*). Para a abordagem de Deep Learning, os algoritmos selecionados foram Long Short-Term Memory (*LSTM*) e Convolutional Neural Network (*CNN*). A escolha desses algoritmos foi baseada em artigos analisados que tratam de problemáticas semelhantes e buscas a respeito de algoritmos eficientes para lidar com classificação multi-classe.

- **Tratamento e Extração de Características dos Dados**

  Para o tratamento e filtragem dos dados, realizou-se a remoção, a partir dos dados originais extraídos de *DogMoveData.csv*, de valores vazios/nulos, de valores ‘<undefined>’ na coluna Task e de valores ‘<undefined>’ na coluna Behavior - removendo apenas quando o valor está presente em Behavior_1, Behavior_2 e Behavior_3. Dessa forma, a quantidade de dados a serem analisados, originalmente sendo 10.611.068, diminuiu para 6.734.323.

  Adicionalmente, verificou-se a viabilidade de remover comportamentos fora do escopo da pesquisa “*Dog behaviour classification with movement sensors placed on the harness and the collar*”, a qual classifica as ações de *galopar*, *deitar*, *sentar*, *cheirar*, *ficar em pé*, *trotar* e *caminhar*. A implementação dessa filtragem reduziu a quantidade de dados disponíveis para 5.927.923. Entretanto, optou-se por, neste momento, manter esses dados para avaliar futuramente a possibilidade de incluí-los dentro do conjunto de comportamentos possíveis de serem classificados.

## Futuras Etapas de Desenvolvimento

Para as próximas etapas de desenvolvimento do projeto, iremos revisar os dados a serem utilizados e, se necessário, realizar uma nova filtragem sobre eles. Além de dar continuidade no processo de extração de características. Dessa forma, com os dados prontos, prosseguiremos para o treinamento dos modelos selecionados, validando as predições alcançadas e avaliando o desempenho de cada um deles. Por fim, iremos realizar uma análise dos resultados obtidos, comparando a performance dos algoritmos para a classificação do comportamento canino e, assim, concluindo quais são os melhores para serem utilizados como solução da problemática.

## Referências

[1] KUMPULAINEN, P. et al. Dog behaviour classification with movement sensors placed on the harness and the collar. Applied Animal Behaviour Science, v. 241, p. 105393, ago. 2021.

[2] AZAMJON MUMINOV; MUKHRIDDIN MUKHIDDINOV; CHO, J. Enhanced Classification of Dog Activities with Quaternion-Based Fusion Approach on High-Dimensional Raw Data from Wearable Sensors. Sensors, v. 22, n. 23, p. 9471–9471, 4 dez. 2022.

[3] HUSSAIN, A. et al. Long Short-Term Memory (LSTM)-Based Dog Activity Detection Using Accelerometer and Gyroscope. Applied Sciences, v. 12, n. 19, p. 9427, 20 set. 2022.

[4] GERENCSÉR, L. et al. Identification of Behaviour in Freely Moving Dogs (Canis familiaris) Using Inertial Sensors. PLoS ONE, v. 8, n. 10, p. e77814, 18 out. 2013.

[5] PIRGE, G. Performance Comparison of Multi-Class Classification Algorithms. Disponível em: <https://gursev-pirge.medium.com/performance-comparison-of-multi-class-classification-algorithms-606e8ba4e0ee>.