

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
CENTRO TECNOLÓGICO  
DEPARTAMENTO DE ENGENHARIA ELÉTRICA E ELETRÔNICA  
CURSO DE GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Lucas Fiamoncini de Souza

**Predição da Fertilidade de Reprodutores Suínos Através de Parâmetros de  
Cinética Espermática e Aprendizado de Máquina**

Florianópolis

2022



Lucas Fiamoncini de Souza

**Predição da Fertilidade de Reprodutores Suínos Através de Parâmetros de Cinética Espermática e Aprendizado de Máquina**

Trabalho de Conclusão de Curso de Graduação em Engenharia Elétrica do Centro Tecnológico da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia Elétrica.

Orientador: Prof. Márcio Holsbach Costa, Dr.

Coorientador: Prof. Rafael da Rosa Ulguim, Dr.

Florianópolis

2022

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

Souza, Lucas Fiamoncini de

Predição da fertilidade de reprodutores suínos através de  
parâmetros de cinética espermática e aprendizado de máquina  
/ Lucas Fiamoncini de Souza ; orientador, Márcio Holsbach  
Costa, coorientador, Rafael da Rosa Ulguim, 2022.  
53 p.

Trabalho de Conclusão de Curso (graduação) -  
Universidade Federal de Santa Catarina, Centro Tecnológico,  
Graduação em Engenharia Elétrica, Florianópolis, 2022.

Inclui referências.

1. Engenharia Elétrica. 2. aprendizado de máquina. 3.  
classificação. 4. fertilidade de suínos. 5. inseminação  
artificial. I. Holsbach Costa, Márcio. II. da Rosa Ulguim,  
Rafael. III. Universidade Federal de Santa Catarina.  
Graduação em Engenharia Elétrica. IV. Título.

Lucas Fiamoncini de Souza

## Predição da Fertilidade de Reprodutores Suínos Através de Parâmetros de Cinética Espermática e Aprendizado de Máquina

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Engenharia Elétrica e aprovado em sua forma final pelo Curso de Engenharia Elétrica.

Florianópolis, 02 de agosto de 2022.



Documento assinado digitalmente  
**Miguel Moreto**  
Data: 02/08/2022 16:33:55-0300  
CPF: 948.850.100-63  
Verifique as assinaturas em <https://v.ufsc.br>

---

Prof. Miguel Moreto, Dr.  
Coordenador do Curso

**Banca Examinadora:**



Documento assinado digitalmente  
**Marcio Holsbach Costa**  
Data: 02/08/2022 15:25:14-0300  
CPF: 572.170.680-53  
Verifique as assinaturas em <https://v.ufsc.br>

---

Prof. Márcio Holsbach Costa, Dr.  
Orientador  
Universidade Federal de Santa Catarina



Documento assinado digitalmente  
**Cesar Ramos Rodrigues**  
Data: 02/08/2022 18:25:11-0300  
CPF: 562.557.990-34  
Verifique as assinaturas em <https://v.ufsc.br>

---

Prof. César Ramos Rodrigues, Dr.  
Avaliador  
Universidade Federal de Santa Catarina



Documento assinado digitalmente  
**Eduardo Luiz Ortiz Batista**  
Data: 02/08/2022 16:37:13-0300  
CPF: 036.521.889-85  
Verifique as assinaturas em <https://v.ufsc.br>

---

Prof. Eduardo Luiz Ortiz Batista, Dr.  
Avaliador  
Universidade Federal de Santa Catarina



## **AGRADECIMENTOS**

Agradeço à minha família, pais e irmãos, por me apoiarem e darem suporte durante toda a minha vida. Aos meus amigos e colegas de curso, pelos anos de estudo e diversão que acompanham a graduação. Ao professor Rafael Ulguim, pelo fornecimento do banco de dados e viabilização da interdisciplinaridade com a medicina veterinária. Ao meu orientador, professor Márcio Costa, pelas aulas, pela oportunidade de iniciação científica e estágio, e por toda as orientações, correções e ajuda ao longo dos últimos anos de trabalho. À UFSC, que possibilitou um ambiente de ensino e vivência que contribuiu muito para meu desenvolvimento pessoal. Por fim, agradeço a todos que fizeram parte da minha trajetória ao longo da graduação e que contribuíram para minha formação pessoal, acadêmica e profissional.



## RESUMO

O Brasil é o quarto maior produtor mundial de carne suína, tendo elevada participação no mercado de exportações. De forma a se elevar os níveis de produtividade, a biotécnica de inseminação artificial é amplamente utilizada, visto que possibilita um melhoramento genético da prole de reprodutores de alto potencial genético. Nesse viés, é de grande importância assegurar a qualidade dos sêmens utilizados para o programa de inseminação artificial. No mesmo sentido, é de grande interesse econômico a predição da fertilidade dos reprodutores, de forma a se garantir maiores índices de fertilidade e avanço genético. Este trabalho apresenta um estudo de técnicas de aprendizado de máquina e o desenvolvimento de um classificador da fertilidade de reprodutores suínos através de parâmetros de cinética espermática. O presente trabalho apresenta os desafios de lidar com um banco de dados reduzido e desbalanceado, obtendo resultados moderados ( $F1 = 0,66$ ) de classificação de 12 machos em férteis ou subférteis. Como resultado, o trabalho se apresenta como um ponto de partida para exploração de técnicas de aprendizado de máquina para classificação de reprodutores suínos através de parâmetros de cinética espermática.

**Palavras-chave:** Fertilidade de suínos, aprendizado de máquina, classificação, inseminação artificial.



## ABSTRACT

Brazil is the fourth largest producer of pork in the world, with a high share of the export market. In order to raise productivity levels, the biotechnique of artificial insemination is widely used, since it allows a genetic improvement of the offspring of high genetic potential boars. Therefore, it is of great importance to ensure the quality of the semen used for the artificial insemination program. Furthermore, it is of great economic interest to predict the fertility of boars, in order to guarantee higher rates of fertility and genetic improvement. This work presents a study of machine learning techniques and the development of fertility classifier of boars through sperm kinetics parameters. This work presents the challenges of dealing with a reduced and imbalanced dataset, obtaining moderate results ( $F1 = 0.66$ ) of classifying 12 males as fertile or subfertile. As a result, the work presents itself as a starting point for exploring machine learning techniques for classifying boars through sperm kinetics parameters.

**Keywords:** Boar fertility, machine learning, classification, artificial insemination.



## LISTA DE FIGURAS

Figura 1 – Terminologia do sistema CASA. . . . .	25
Figura 2 – Exemplo de classificação binária. . . . .	27
Figura 3 – Exemplo de classificação binária com dados desbalanceados. . . . .	28
Figura 4 – Exemplo de sobreamostragem. . . . .	28
Figura 5 – Exemplo de geração de amostras sintéticas SMOTE. . . . .	29
Figura 6 – SVM de margens suaves. . . . .	31
Figura 7 – Matriz de confusão para um problema binário. . . . .	33
Figura 8 – Exemplo de curva ROC. . . . .	35
Figura 9 – Ilustração das etapas de pré-processamento, treinamento e validação. . . . .	38
Figura 10 – Matriz de confusão da classificação das amostras para o melhor classificador obtido . . . . .	44
Figura 11 – Matriz de confusão da classificação dos machos para o melhor classificador obtido . . . . .	44
Figura 12 – Curva ROC da classificação dos machos para o melhor classificador obtido . . . . .	45



## LISTA DE TABELAS

Tabela 1 – Descrição dos parâmetros de cinética espermática do sistema CASA. . . . .	25
Tabela 2 – Valores de hiperparâmetros definidos para o algoritmo de otimização . . . . .	40
Tabela 3 – Vetores de atributos para cálculo de média, variância, valor máximo e mínimo . . . . .	41
Tabela 4 – Vetores de atributos para treinamento (após normalização) . . . . .	41
Tabela 5 – Classificadores das amostras, com vetor de atributos utilizado, valores de hiperparâmetros selecionados por otimização e o valor da medida F1 de validação. . . . .	43
Tabela 6 – Medidas de desempenho da classificação dos machos para o melhor classificador obtido . . . . .	45



## LISTA DE ABREVIATURAS E SIGLAS

AUC	<i>Area Under Curve</i>
CASA	<i>Computer Assisted Sperm Analysis</i>
LOO	<i>Leave-one-out</i>
ROC	<i>Receiver Operating Characteristic</i>
SMOTE	<i>Synthetic Minority Over-sampling Technique</i>
SVM	<i>Support Vector Machine</i>



## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>19</b>
<b>2</b>	<b>OBJETIVOS</b>	<b>21</b>
2.1	OBJETIVOS GERAIS	21
2.2	OBJETIVOS ESPECÍFICOS	21
<b>3</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>23</b>
3.1	SUINOCULTURA E INSEMINAÇÃO ARTIFICIAL	23
3.2	ANÁLISE ESPERMÁTICA AUXILIADA POR COMPUTADOR (CASA)	24
3.3	APRENDIZADO DE MÁQUINA	24
<b>3.3.1</b>	<b>Aprendizado supervisionado</b>	<b>25</b>
<b>3.3.2</b>	<b>Aprendizado não supervisionado</b>	<b>26</b>
3.4	CLASSIFICAÇÃO BINÁRIA	26
3.5	DADOS DESBALANCEADOS	27
<b>3.5.1</b>	<b>Sobreamostragem</b>	<b>27</b>
3.5.1.1	SMOTE	28
3.6	MODELOS DE APRENDIZADO DE MÁQUINA	29
<b>3.6.1</b>	<b>Regressão logística</b>	<b>29</b>
<b>3.6.2</b>	<b>Máquina de vetores de suporte (SVM)</b>	<b>30</b>
3.6.2.1	Caso de classes linearmente separáveis	30
3.6.2.2	Caso de classes não linearmente separáveis	30
3.6.2.3	SVM não linear	31
<b>3.6.3</b>	<b>K-means clustering</b>	<b>31</b>
3.7	VALIDAÇÃO CRUZADA	32
<b>3.7.1</b>	<b>K-Fold</b>	<b>32</b>
<b>3.7.2</b>	<b>Leave-one-out</b>	<b>32</b>
3.8	MÉTRICAS DE AVALIAÇÃO	32
<b>3.8.1</b>	<b>Matriz de Confusão</b>	<b>33</b>
<b>3.8.2</b>	<b>Curva ROC</b>	<b>34</b>
3.9	TRABALHOS RELACIONADOS	35
<b>4</b>	<b>MÉTODOS</b>	<b>37</b>
4.1	BANCO DE DADOS	37
4.2	ROTULAÇÃO DAS AMOSTRAS	37
4.3	ANÁLISE PRELIMINAR	38
4.4	PRÉ-PROCESSAMENTO, TREINAMENTO E VALIDAÇÃO	38
4.5	AVALIAÇÃO DE DESEMPENHO	42
<b>5</b>	<b>RESULTADOS</b>	<b>43</b>
<b>6</b>	<b>DISCUSSÃO E CONCLUSÃO</b>	<b>47</b>
	<b>REFERÊNCIAS</b>	<b>49</b>



## 1 INTRODUÇÃO

A produção e comercialização de carne suína no Brasil tem apresentado crescimento nos últimos anos, tendo lugar de destaque entre os produtores mundiais. De forma a acompanhar esse crescimento, a utilização da biotécnica de inseminação artificial se mostra como uma ferramenta útil para atualização do potencial genético do plantel. Isso proporciona uma melhoria dos índices de eficiência e produtividade e, por conseguinte, maior retorno econômico.

Atualmente, a inseminação artificial é um procedimento de grande importância econômica na suinocultura (KHALIFA *et al.*, 2014), sendo responsável por cerca de 95% das fecundações em granjas comerciais (LUCCA *et al.*, 2019). Devido ao fato do macho ter um impacto maior do que a fêmea na produtividade (RUIZ-SÁNCHEZ *et al.*, 2006), é necessário garantir a qualidade do sêmen utilizado para inseminação artificial, de forma a assegurar elevados índices de produção. Nesse quesito, a avaliação da capacidade de fertilização de ejaculados — amostras de sêmen coletadas — de machos suínos é essencial para a otimização da inseminação artificial no que diz respeito à produção de leitões (SCHULZE *et al.*, 2013). A avaliação do sêmen é realizada a partir de alguns parâmetros, tais como concentração, cinética e morfologia espermática, os quais possuem ligação com a fertilidade do macho suíno. Isso permite selecionar ejaculados adequados para o processamento de doses inseminantes (LUCCA *et al.*, 2021). Com o avanço do desenvolvimento tecnológico e de métodos computadorizados, a análise do sêmen passou a ser realizada a partir de sistemas computadorizados de auxílio à análise espermática conhecidos como CASA (do inglês - *Computer Assisted Sperm Analysis*), de forma a remover o caráter subjetivo de uma análise visual ao microscópio (WABERSKI; SUAREZ; HENNING, 2021).

Para o processo de inseminação artificial em suínos, é importante assegurar a adequada qualidade das doses de sêmen utilizadas a fim de maximizar a taxa de fertilidade e o número total de leitões nascidos (KUMMER *et al.*, 2013). Nesse processo, laboratórios de inseminação artificial utilizam o sistema CASA para obter informações a respeito das amostras e, a partir da sua análise, decidir quais doses são adequadas para serem processadas no programa de inseminação artificial e quais delas devem ser descartadas (KAMPHUIS *et al.*, 2020). O principal parâmetro utilizado para selecionar uma dose é a comparação do índice de motilidade com um valor de referência. Pesquisas estimam que cerca de 5% a 7% dos machos são subfêrteis e que 10% a 17% dos reprodutores no plantel que satisfazem os padrões normais para a qualidade do ejaculado estariam classificados como subfêrteis (LUCCA, 2020). A partir da análise das doses *in natura* e após armazenamento de cada macho suíno, pode ser possível inferir sobre quais machos seriam subfêrteis e que necessitariam ser substituídos, de forma a se manter o nível elevado de fertilidade da granja.

Seguindo essa vertente de classificação para tomada de decisão, o aprendizado de

máquina, que é uma área da inteligência artificial que tem por objetivo o estudo de técnicas e métodos capazes de extrair conhecimento a partir de um conjunto de dados, encontra um importante ramo de aplicação. Nos últimos anos, com o crescente desenvolvimento de tecnologias e massiva geração de dados, inúmeros setores — tais como financeiro, tecnológico e médico — passaram a adotar medidas computacionais para auxiliar no tratamento de diversos problemas. Nesse quesito, o aprendizado de máquina propicia uma ferramenta automatizada capaz de reconhecer padrões em dados através de algoritmos computacionais e, inclusive, generalizar esse conhecimento para tomar ações como a classificação de novos dados.

Nesse contexto, os setores de agropecuária e medicina veterinária também têm buscado uso de ferramentas computacionais para resolução de problemas antes totalmente dependentes da subjetividade da análise humana (PILES *et al.*, 2021; SMITH *et al.*, 2020; GARCÍA *et al.*, 2020; NEETHIRAJAN, 2020; KAMPHUIS *et al.*, 2020).

Visando a contribuir com tecnologias para o ramo de suínos, o presente trabalho utiliza o banco de dados e os resultados obtidos no trabalho de Lucca *et al.* (2021) de forma a auxiliar na predição da fertilidade de suínos. O trabalho de Lucca *et al.* (2021) apresenta uma análise do efeito de diferentes tempos de armazenamento na motilidade espermática de ejaculados suínos, baseado no conhecimento prévio de que a preservação da motilidade das células espermáticas após seu armazenamento apresenta informação relevante para predição da fertilidade individual de machos suínos. Os resultados obtidos indicam que machos suínos com baixa preservação da motilidade após 120 horas da realização da coleta resultam em um número menor de leitões nascidos quando comparados à sua contrapartida.

## 2 OBJETIVOS

O presente trabalho, motivado pelos resultados de Lucca *et al.* (2021), visa a utilizar técnicas de aprendizado de máquina para auxiliar a predição da fertilidade de suínos. As principais informações utilizadas neste trabalho são as medidas de avaliação do sêmen fornecidas pelo sistema CASA e a utilização dessas mesmas medidas 120 horas após a coleta.

### 2.1 OBJETIVOS GERAIS

Este trabalho tem por objetivo estudar e explorar técnicas de aprendizado de máquina para a classificação da fertilidade de reprodutores suínos (em férteis ou subférteis) através de parâmetros de cinética espermática. A principal contribuição desse trabalho é avaliar de forma preliminar a possibilidade de identificação de reprodutores de baixa fertilidade.

### 2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste trabalho são:

- a) Realizar um estudo sobre métodos de aprendizado de máquina;
- b) Desenvolver um classificador da fertilidade de reprodutores suínos utilizando dados referentes à cinética espermática;
- c) Avaliar o desempenho do método desenvolvido utilizando um banco de dados com informações de análise espermática auxiliada por computador;
- d) Analisar a viabilidade da utilização dos resultados deste trabalho como um sistema de recomendação para a substituição de machos suínos subférteis em um plantel reprodutivo.



### 3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentada a fundamentação teórica necessária para a compreensão e desenvolvimento deste trabalho.

#### 3.1 SUINOCULTURA E INSEMINAÇÃO ARTIFICIAL

O Brasil é o quarto maior produtor mundial de carne suína, responsável por cerca de 3,2% da produção mundial até o ano de 2014 (ABCS, 2014); sendo que de 2017 a 2021 houve um crescimento médio de 7% na produção brasileira de carne suína (ABCS, 2021). No mercado de exportações, o setor de suínos registrou um aumento médio de 6,6% entre os anos de 2014 e 2019 (MARTINS; TALAMINI; SANTOS FILHO, 2019). Dessa forma, é notório o desejo por maiores níveis de produtividade.

De acordo com Moreira *et al.* (2013), a inseminação artificial é uma biotécnica que pode contribuir significativamente para o aumento da produtividade, possibilitando melhoria dos índices reprodutivos. Isso se dá pelo fato de haver um melhoramento genético da prole de reprodutores de alto potencial genético. Na suinocultura, um macho possui maior impacto na eficiência e produtividade do que uma fêmea, sendo esse impacto ainda maior com a utilização de inseminação artificial (RUIZ-SÁNCHEZ *et al.*, 2006). Nas centrais de inseminação artificial, é comum a produção de doses inseminantes formadas pela mistura de sêmen proveniente de mais de um macho. Dessa forma, a participação de reprodutores subfêrteis diminui a eficiência produtiva e, por conseguinte, a margem de lucro do produtor (MOREIRA *et al.*, 2013).

Portanto, a relação entre a qualidade de sêmen e os resultados de fertilidade é um indicador importante para as centrais de processamento de sêmen (LUCCA, 2020). Para um melhor aproveitamento da inseminação artificial, é muito comum a análise da qualidade seminal dos ejaculados, de forma a auxiliar no processo de seleção das amostras que integrarão o programa de inseminação artificial. De acordo com Lucca (2020), uma das maneiras de verificar a fertilidade do reprodutor é através da avaliação rotineira dos ejaculados *in vitro*.

Devido à influência do macho no desempenho reprodutivo do plantel, é de grande interesse econômico a predição da fertilidade do reprodutor. Nesse sentido, Lucca (2020) afirma que a utilização dos recursos de um sistema computadorizado de análise de sêmen para obtenção de dados mais objetivos de cinética espermática, associado com a análise dos mesmos dados em diferentes tempos de armazenamento e utilização de inseminações artificiais únicas (uma dose contendo conteúdo seminal de apenas um macho) pode ser uma ferramenta adequada para identificar machos subfêrteis.

### 3.2 ANÁLISE ESPERMÁTICA AUXILIADA POR COMPUTADOR (CASA)

A análise espermática por microscopia data dos anos 1600, tendo como principal parâmetro a motilidade, que é utilizada até hoje nos centros de inseminação artificial para a seleção de amostras (VALVERDE; BARQUERO; SOLER, 2020).

A partir do início dos anos 80, motivados pela subjetividade envolvida na determinação visual da motilidade, foram iniciados estudos buscando métodos que diminuíssem o viés de uma avaliação visual humana (VALVERDE; BARQUERO; SOLER, 2020). Os estudiosos reconheceram que a utilização de imagens, gravações e tecnologia computadorizada são importantes no desenvolvimento de um sistema automatizado para medição de cinética espermática (AMANN; WABERSKI, 2014).

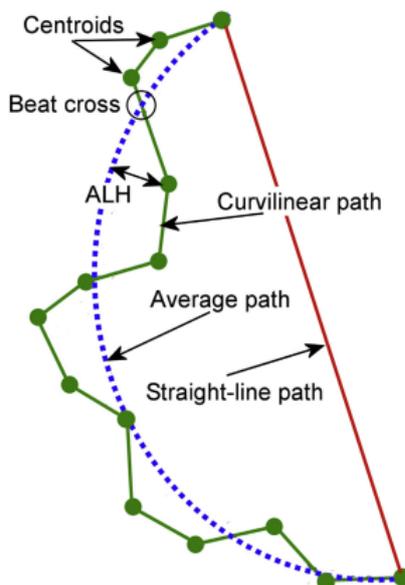
Dessa forma, nasceu o sistema conhecido como *Computer Assisted Sperm Analysis* (CASA), traduzido por análise espermática auxiliada por computador, que utiliza ferramentas computacionais de processamento de imagens e vídeo para uma avaliação objetiva de parâmetros de cinética espermática. Nesse quesito, laboratórios de inseminação artificial utilizam o sistema CASA para obter informações a respeito da qualidade das amostras de sêmen de suínos, como também de outros animais, e, a partir da análise dessas informações, decidir quais doses são adequadas para serem utilizadas no programa de inseminação artificial e quais delas devem ser descartadas por não atenderem aos requisitos mínimos de valores de cinética espermática (AMANN; WABERSKI, 2014), (KAMPHUIS *et al.*, 2020). Além disso, os parâmetros de cinética espermática fornecidos pelo CASA também podem ser utilizados como ferramenta para predição da fertilidade de machos suínos, sendo que alguns autores (KUMMER *et al.*, 2013), (LUCCA *et al.*, 2021) sugerem também levar em consideração a análise do sêmen após armazenamento.

A terminologia do sistema CASA é apresentada na Figura 1. Inicialmente, algoritmos de processamento de imagem são aplicados para cada quadro de uma gravação de vídeo de forma a fornecer um centroide associado à localização da cabeça de um espermatozoide. Posteriormente, é feito o traçado ligando os centroides obtidos para todos os quadros da gravação de vídeo, que informa a trajetória real do espermatozoide, denominada de trajetória curvilínea. A partir dessa trajetória, são calculadas algumas medidas, apresentadas na Tabela 1.

### 3.3 APRENDIZADO DE MÁQUINA

O aprendizado de máquina é uma área da inteligência artificial que tem por objetivo o estudo de técnicas e métodos capazes de extrair conhecimento a partir de um conjunto de dados. Nesta seção, serão abordados alguns tipos de aprendizado de máquina e as tarefas envolvidas em uma tarefa de classificação binária.

Figura 1 – Terminologia do sistema CASA.



Fonte: Retirado de Amann e Waberski (2014).

Tabela 1 – Descrição dos parâmetros de cinética espermática do sistema CASA.

Parâmetros	Unidade	Descrição
DAP	$\mu\text{m}$	Distância percorrida do trajeto médio
DCL	$\mu\text{m}$	Distância percorrida do trajeto real
DSL	$\mu\text{m}$	Distância percorrida do trajeto em linha reta
VAP	$\mu\text{m/s}$	Velocidade percorrida do trajeto médio
VCL	$\mu\text{m/s}$	Velocidade percorrida do trajeto real
VSL	$\mu\text{m/s}$	Velocidade percorrida do trajeto em linha reta
ALH	$\mu\text{m}$	Amplitude de deslocamento lateral da cabeça
BCF	Hz	Frequência com que o trajeto real cruza o trajeto médio
STR	%	Retilinearidade (VSL/VAP)
LIN	%	Linearidade (VSL/VCL)
WOB	%	Coefficiente de oscilação (VAP/VCL)

Fonte: Retirado de Lucca (2020).

### 3.3.1 Aprendizado supervisionado

O aprendizado supervisionado é um tipo de aprendizado de máquina caracterizado pelo uso de dados rotulados. Nesse tipo de aprendizado, os rótulos são utilizados para treinar algoritmos para classificar ou prever o valor de saída de um vetor de atributos de entrada cujo rótulo é conhecido.

Tomando-se como exemplo os dados de cinética espermática fornecidos pelo sistema CASA, temos um conjunto de dados de amostras de sêmen com diferentes atributos cujo rótulo seria “próprio” (caracterizando um sêmen apropriado para o programa de inseminação artificial) e “impróprio” (caracterizando sêmen que deve ser descartado).

No aprendizado supervisionado, é comum subdividir o conjunto de dados em três: treinamento, validação e teste. A etapa de treinamento consiste em treinar o algoritmo com dados de entrada que possuem a informação do rótulo. A validação tem o papel de fornecer ao algoritmo uma estimativa de sua capacidade de aprendizado durante o treinamento. Já a etapa de teste, é responsável por fornecer métricas de desempenho do algoritmo em um conjunto de dados até então desconhecido, ou seja, validar a capacidade de generalização do aprendizado.

### 3.3.2 Aprendizado não supervisionado

No aprendizado não supervisionado, não há conhecimento do rótulo dos dados do conjunto. Nesse caso, os algoritmos de aprendizado de máquina agrupam os dados de acordo com uma métrica ou relação especificada.

Em um problema de classificação associado ao sistema CASA, os dados de entrada poderiam ser agrupados de acordo com níveis de motilidade.

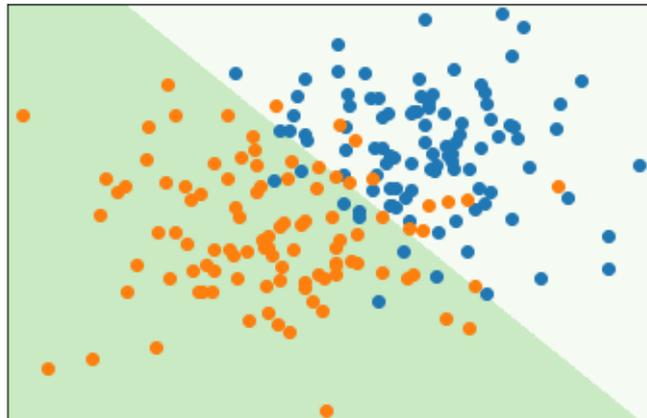
## 3.4 CLASSIFICAÇÃO BINÁRIA

A classificação é um caso específico de aprendizado supervisionado em que, dado um conjunto de dados, deseja-se treinar um classificador que possa determinar o rótulo (classe) de um vetor de atributos de entrada. A classificação binária consiste na situação em que só há duas classes possíveis para o vetor de entrada. Por exemplo, dado um vetor de atributos de cinética espermática de determinado macho suíno, deseja-se saber se o macho é fértil ou subfértil. No caso de um problema de classificação, o objetivo é treinar um algoritmo de aprendizado de máquina de forma que implemente uma função que obtenha a melhor classificação a partir de uma métrica escolhida associada ao problema.

A Figura 2 apresenta um exemplo de classificação binária, em que os pontos laranjas e azuis representam dados de classes distintas e a fronteira entre diferentes tons de verde representa o limiar de decisão de um classificador genérico. Nota-se a existência de erros de classificação representados pela existência de pontos laranjas na região verde acinzentada e pontos azuis na região verde.

Em uma tarefa de classificação, é comum que o banco de dados utilizado seja dividido em treinamento e teste. Como já mencionado, os dados de treinamento serão utilizados para que o algoritmo encontre uma função que possibilite classificar os dados nas classes do problema. Durante o processo de treinamento, este conjunto é subdividido, originando o subconjunto de validação, utilizado para validar o aprendizado do algoritmo antes da etapa de teste. O conjunto de teste é utilizado como elemento de desfecho da avaliação de desempenho do classificador.

Figura 2 – Exemplo de classificação binária.



Fonte: do autor.

### 3.5 DADOS DESBALANCEADOS

A questão de dados desbalanceados é uma problemática associada a uma diferença considerável entre as ocorrências das classes de um determinado problema (SANTOS *et al.*, 2018). Supondo um cenário de classificação binária com dados desbalanceados, uma classe possui uma representação maior sobre a outra; denominadas, respectivamente, de classe majoritária e minoritária (HE; GARCIA, 2009).

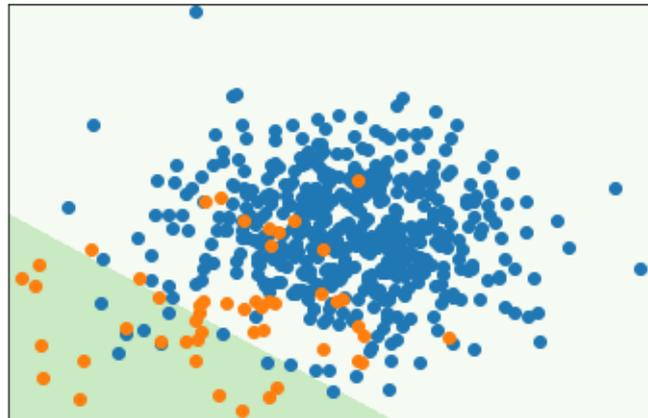
Tarefas de classificação que envolvem dados desbalanceados (Figura 3) são comumente enviesadas em relação à classe majoritária, devido a uma maior quantidade de informação presente nessa classe. Isso influencia o desempenho do classificador em reconhecer corretamente novos dados, especialmente da classe minoritária. A Figura 3 ilustra essa situação, em que há um maior erro de classificação na classe minoritária (pontos laranjas) do que na classe majoritária (pontos azuis).

Algumas alternativas podem ser tomadas na etapa de treinamento e validação de um modelo de classificação para lidar com os dados desbalanceados. Uma das estratégias mais conhecidas é a sobreamostragem (SANTOS *et al.*, 2018).

#### 3.5.1 Sobreamostragem

A estratégia de sobreamostragem (Figura 4) consiste em elevar a quantidade de ocorrências na classe minoritária produzindo novos dados ou repetindo alguns deles (MOHAMMED; RAWASHDEH; ABDULLAH, 2020). Uma das técnicas de sobreamostragem mais utilizadas é a técnica de sobreamostragem minoritária sintética, do inglês *Synthetic Minority Over-sampling Technique* (SMOTE) (CHAWLA *et al.*, 2002).

Figura 3 – Exemplo de classificação binária com dados desbalanceados.



Fonte: do autor.

Figura 4 – Exemplo de sobreamostragem.

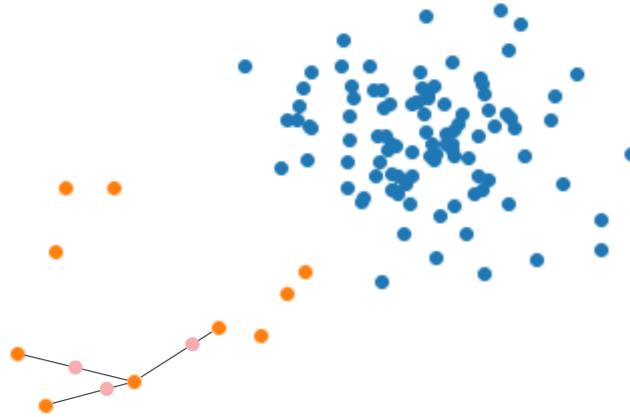


Fonte: do autor.

### 3.5.1.1 SMOTE

O algoritmo SMOTE, proposto por Chawla *et al.* (2002), é de grande popularidade nos problemas de classificação binária com dados desbalanceados. O algoritmo atua no espaço de atributos; inicialmente, é calculada a distância de  $k$  vizinhos mais próximos do vetor de atributos analisado em questão. Em seguida, multiplica-se essa distância por um valor aleatório entre zero e um e soma-se o resultado obtido ao vetor de atributos de referência para gerar uma nova amostra sintética (CHAWLA *et al.*, 2002). Dessa forma, é possível aumentar a representatividade da classe minoritária. A Figura 5 apresenta um exemplo de geração de amostras sintéticas pelo algoritmo SMOTE na classe minoritária (pontos laranjas), em que os novos pontos gerados sinteticamente para essa classe são representados em rosa claro.

Figura 5 – Exemplo de geração de amostras sintéticas SMOTE.



Fonte: do autor.

### 3.6 MODELOS DE APRENDIZADO DE MÁQUINA

Existem diversos modelos de aprendizado de máquina. A seguir, serão apresentados os modelos utilizados na realização deste trabalho.

#### 3.6.1 Regressão logística

A regressão logística é um modelo de classificação associado à probabilidade *a posteriori* de um vetor de atributos pertencer a uma determinada classe. A função que modela este classificador é dada por:

$$\tilde{y} = \varphi(\mathbf{w}^T \mathbf{x}) \quad (1)$$

Em que  $\tilde{y}$  representa a classe predita do problema,  $\mathbf{x}$  indica o vetor de atributos,  $\mathbf{w}$  o vetor de pesos que representa o hiperplano separador das classes, e  $\varphi(\cdot)$  a função sigmoide logística, dada por:

$$\varphi(a) = \frac{1}{1 + e^{-a}} \quad (2)$$

O treinamento do modelo se dá a partir da otimização de uma função perda, responsável por medir o desempenho do classificador durante a etapa de treino. Uma função perda comum é a entropia cruzada com regularização  $l_2$  (SILVA, 2021), dada por:

$$\min_{\mathbf{w}, C} \left( \frac{1}{2} \mathbf{w}^T \mathbf{w} - C \sum_{i=1}^N [y_i \log \tilde{y}_i + (1 - y_i) \log(1 - \tilde{y}_i)] \right) \quad (3)$$

Em que  $y$  é a classe verdadeira do problema,  $N$  é o número de amostras e  $C$  é um termo que controla o nível de regularização, responsável por garantir uma melhor generalização

do modelo; isto é, melhorar o desempenho na classificação de novas amostras fora do conjunto de treinamento.

### 3.6.2 Máquina de vetores de suporte (SVM)

Assim como a regressão logística, a máquina de vetores de suporte, do inglês *Support Vector Machine* (SVM), é um classificador amplamente utilizado em problemas de classificação binária.

#### 3.6.2.1 Caso de classes linearmente separáveis

Considere  $\mathbf{x}$  o vetor de atributos de determinado conjunto de treinamento, em que tais atributos pertencem a uma de duas classes (linearmente separáveis) de um problema de classificação binária. O objetivo é projetar um hiperplano que classifique corretamente todos os vetores de treinamento. Todavia, mais de uma solução é possível para o hiperplano; o que gera diferentes linhas de decisão (caso linear) (THEODORIDIS; KOUTROUMBAS, 2009). A equação do hiperplano é dada por:

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (4)$$

em que  $b$  é o termo de *bias*.

Dessa forma, diferentes hiperplanos podem realizar a tarefa de classificação. Entretanto, nosso interesse se dá pelo hiperplano que possui a maior margem de segurança no processo de separação de classes, possibilitando uma maior variação nos dados sem acarretar erros de classificação e, conseqüentemente, uma maior generalização do classificador.

Para encontrar o hiperplano ótimo, o SVM tem como objetivo:

$$\text{minimizar } J(\mathbf{w}, b) \equiv \frac{1}{2} \|\mathbf{w}\|^2 \quad (5)$$

$$\text{sujeito a } y_i(\mathbf{w}^T \mathbf{x} + b) \geq 1, \quad i = 1, 2, \dots, N \quad (6)$$

Em problemas de classificação binária, é comum denominarmos uma classe como sendo a positiva ( $y = 1$ ) e a outra classe como sendo a negativa ( $y = -1$ ).

#### 3.6.2.2 Caso de classes não linearmente separáveis

Em situações reais, é muito difícil encontrar problemas em que as classes são linearmente separáveis. Como solução, são introduzidas as chamadas variáveis de folga  $\xi_i$ , responsáveis por relaxar as imposições impostas ao problema de otimização. Dessa forma, dá-se origem ao SVM de margens suaves (Figura 6) (FACELI *et al.*, 2011). Agora, o objetivo de otimização é:

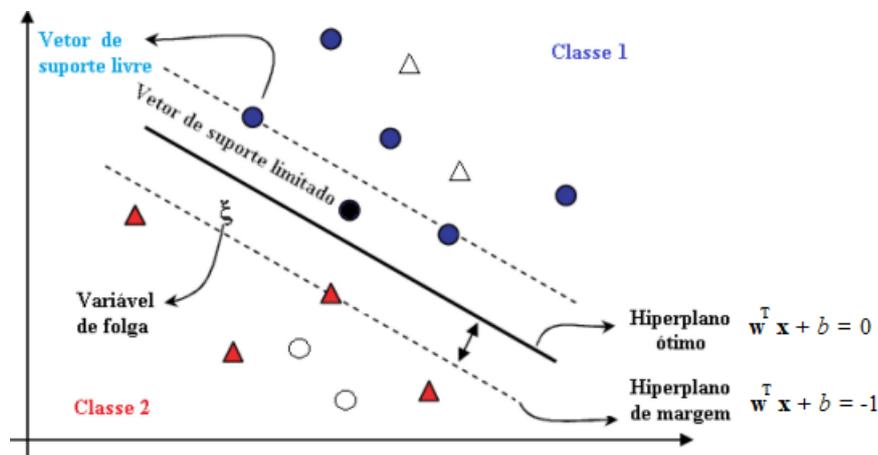
$$\text{minimizar } J(\mathbf{w}, b, \boldsymbol{\xi}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \quad (7)$$

$$\text{sujeito a } y_i(\mathbf{w}^T \mathbf{x} + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N \quad (8)$$

Em que  $C$  é um parâmetro que controla a flexibilização das margens.

Na Figura 6 os pontos azuis representam uma classe do problema, enquanto que os triângulos vermelhos a outra classe. Os pontos e triângulos sem cores representam dados classificados erroneamente.

Figura 6 – SVM de margens suaves.



Fonte: adaptado de (GUBERT, 2019)

### 3.6.2.3 SVM não linear

Os SVMs não lineares fazem um mapeamento do conjunto de dados de treinamento original para um novo espaço. A escolha apropriada da função de mapeamento possibilita que o problema seja resolvido por um classificador SVM linear no espaço transformado (FACELI *et al.*, 2011).

De acordo com Faceli *et al.* (2011), não é necessário que saibamos explicitamente a função de mapeamento utilizada, que pode ser muito complexa. Nosso interesse se dá pelo cálculo do produto escalar entre os objetos a serem classificados no espaço transformado, o que possibilita sua separação. Isso se dá através das funções de Kernel, sendo a polinomial, a RBF (do inglês - *Radial Basis Function*) e a sigmoide as mais utilizadas.

### 3.6.3 K-means clustering

O algoritmo de agrupamento *K-means* é um dos mais famosos de aprendizado não supervisionado. Inicialmente, especifica-se o número  $k$  de *clusters* em que se deseja agrupar

os dados do problema. A partir disso, o algoritmo inicializa com  $k$  centroides no espaço dos dados. O objetivo é minimizar a soma da distância quadrática de cada ponto do espaço ao centroide mais próximo (ARTHUR; VASSILVITSKII, 2007). Tem-se, então, que:

$$\sum_{i=1}^N \min(\|x_i - \mu_k\|^2) \quad (9)$$

Em que  $N$  é o número de dados  $x$  do problema, e  $\mu_k$  é o centroide mais próximo do dado  $x_i$ .

O algoritmo *K-means* se apresenta como uma técnica para classificação de dados não rotulados em um conjunto de  $k$  classes especificadas.

### 3.7 VALIDAÇÃO CRUZADA

A validação cruzada é uma técnica de avaliação de desempenho muito utilizada durante o processo de treinamento de modelos de aprendizado de máquina.

#### 3.7.1 K-Fold

A validação cruzada *K-fold* consiste em dividir o conjunto total de dados em  $K$  subconjuntos, em que  $K-1$  conjuntos serão utilizados para treinamento e o conjunto restante será utilizado para testar o modelo de classificação treinado. Em seguida, há uma rotação dos *folds* e o procedimento é realizado novamente. Isso é feito  $K$  vezes, até que todos os *folds* tenham sido utilizados para treinar e testar o modelo. Por fim, o desempenho do classificador é dado pela média de seu desempenho nos  $K$  *folds* de teste.

#### 3.7.2 Leave-one-out

A validação *Leave-one-out* (LOO) pode ser entendida como um caso especial da validação *K-fold*, em que o número de *folds* corresponde ao número total de dados. Sendo assim, a lógica do procedimento é a mesma, em que todos os dados serão utilizados para treinar e testar o modelo e o desempenho do modelo será a média de todos os dados utilizados para teste.

### 3.8 MÉTRICAS DE AVALIAÇÃO

A avaliação de desempenho de um classificador é fundamental em um projeto de aprendizado de máquina. Esta avaliação consiste em analisar a rotulação dada pelo classificador a dados até então desconhecidos por ele.

Dado um conjunto de  $N$  dados, é comum dividirmos esse conjunto em subconjuntos de treinamento e teste. Os dados de treinamento, como o próprio nome insinua, serão utilizados na etapa de treinamento (aprendizado) do classificador. Já os dados de teste,

são utilizados para verificar o desempenho do classificador treinado. Algumas formas de análise são a matriz de confusão e a curva ROC.

### 3.8.1 Matriz de Confusão

A matriz de confusão é uma tabela que representa a quantidade de dados classificados em cada uma das classes do problema em questão. Por exemplo, em um problema binário (duas classes, uma positiva e outra negativa), temos uma matriz de dimensão  $2 \times 2$ , em que as linhas representam as duas classes reais do problema; enquanto que as duas colunas representam as predições realizadas pelo classificador. A Figura 7 ilustra a matriz de confusão para um problema binário.

Figura 7 – Matriz de confusão para um problema binário.

Classe verdadeira	1	VP	FN
	-1	FP	VN
		1	-1
		Classe predita	

Fonte: do autor.

Na Figura 7, temos as seguintes abreviações:

- VP: número de verdadeiros positivos (dados da classe positiva classificados corretamente);
- VN: número de verdadeiros negativos (dados da classe negativa classificados corretamente);
- FP: número de falsos positivos (dados da classe negativa classificados como sendo da classe positiva);
- FN: número de falsos negativos (dados da classe positiva classificados como sendo da classe negativa);

A partir da matriz de confusão, uma série de medidas de desempenho podem ser obtidas, tais como:

- Acurácia: avalia os dados corretamente classificados.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (10)$$

- Precisão: proporção de dados positivos classificados corretamente entre todos os preditos como positivos.

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (11)$$

- Sensibilidade: taxa de acerto na classe positiva.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \quad (12)$$

- Medida F: a medida F é uma combinação das medidas de precisão e sensibilidade. Dessa forma, é possível analisar as duas métricas em conjunto, evitando uma análise equivocada das duas métricas isoladamente.

$$F_l = \frac{(l + 1) \cdot \text{Sensibilidade} \cdot \text{Precisão}}{\text{Sensibilidade} + (l \cdot \text{Precisão})} \quad (13)$$

A constante  $l$  permite alterar o grau de importância de uma das medidas. Escolhendo-se  $l = 1$  definimos ambas com o mesmo grau de importância e temos a chamada medida  $F_1$ .

$$F_1 = \frac{2 \cdot \text{Sensibilidade} \cdot \text{Precisão}}{\text{Sensibilidade} + \text{Precisão}} \quad (14)$$

- Coeficiente de correlação de Matthews (MCC): medida para avaliação de classificador binário. É um valor pertencente ao intervalo  $[-1,1]$ , sendo 1 indicativo de um classificador perfeito, 0 um classificador aleatório e -1 um classificador totalmente desconexo (GUBERT, 2019).

$$\text{MCC} = \frac{VP \cdot VN - FP \cdot FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \quad (15)$$

### 3.8.2 Curva ROC

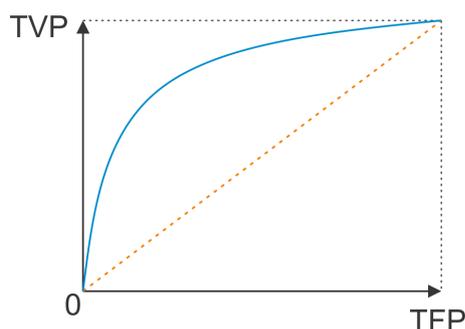
A curva de característica de operação do receptor, do inglês *Receiver Operating Characteristic* (ROC), é uma outra forma de avaliar o desempenho de classificadores binários (FACELI *et al.*, 2011). A curva ROC (Figura 8) é obtida a partir de um gráfico bidimensional em que a abscissa é representada pela taxa de falsos positivos (TFP) e a ordenada pela taxa de verdadeiros positivos (TVP), também chamada de sensibilidade, tal que:

$$TFP = 1 - \frac{VN}{VN + FP} \quad (16)$$

$$TVP = \text{Sensibilidade} = \frac{VP}{VP + FN} \quad (17)$$

Na Figura 8, a curva em azul representa um classificador hipotético para ilustração do exemplo, enquanto que a linha pontilhada em laranja representa a curva de um classificador aleatório.

Figura 8 – Exemplo de curva ROC.



Fonte: do autor.

Uma medida obtida a partir da curva ROC é a *Area Under Curve* (AUC), traduzida por área sob a curva. O valor da AUC pode variar entre 0 e 1, sendo que o valor 0,5 indica que o classificador tem o mesmo desempenho de um classificador aleatório, enquanto que valores de AUC próximos de 1 indicam que o classificador tem desempenho próximo ao de um classificador ideal. Valores abaixo de 0,5 indicam que o classificador possui um desempenho inadequado e seu projeto deve ser refeito.

### 3.9 TRABALHOS RELACIONADOS

A utilização de algoritmos de regressão linear e logística é abordada nos trabalhos de Quintero-Moreno, Rigau e Rodríguez-Gil (2004) e Gadea, Sellés e Marco (2004) como ferramenta de análise de correlação dos parâmetros de sêmen de machos suínos com resultados *in vivo* como, por exemplo, taxa de parto e tamanho da leitegada. Na mesma vertente, o trabalho desenvolvido por Tsakmakidis, Lymberopoulos e Khalifa (2010) também estuda a relação de parâmetros de qualidade espermática com fertilidade, porém carece de informações de análise estatística.

Em Kummer *et al.* (2013), é utilizado um processo de *cluster* para rotulação das amostras, além da utilização da análise de discriminante múltipla como técnica de seleção de atributos e classificação. Uma acurácia de aproximadamente 80% é apresentada na correta classificação dos ejaculados em suas respectivas classes. Entretanto, a medida de acurácia apresentada é obtida nos mesmos dados utilizados para o desenvolvimento do

modelo e não em um conjunto separado. Além disso, apenas a acurácia é apresentada, medida que tende a ser enviesada em se tratando de dados desbalanceados. O trabalho de Kummer *et al.* (2013) utiliza dados pós armazenamento.

O estudo realizado por Kuhlitz *et al.* (2019) apresenta um modelo de regressão utilizado para estimar características de qualidade espermática de suínos. Nesse trabalho, é utilizado o método *K-fold* como medida de validação do modelo. Esse método de validação é recomendado para banco de dados pequenos, que é o caso deste trabalho.

Em Kamphuis *et al.* (2020), um classificador GBM (*Gradient Boosting Machine*) é utilizado para decidir quais ejaculados utilizar em doses de inseminação artificial. Apesar de possuir um grande banco de dados de resultados *in vivo*, os classificadores projetados apresentam *overfitting*, sendo que o melhor dos classificadores apresenta uma AUC=0,624 para predição de tempo de gestação a partir de ejaculados.

O trabalho realizado em Lucca *et al.* (2021), apesar de não fazer uso de técnicas de aprendizado de máquina, proporciona resultados e discussões interessantes que viabilizam o uso de algoritmos de classificação.

## 4 MÉTODOS

Nesta seção, são descritos o banco de dados, a metodologia de pré-processamento dos dados, treinamento, validação e avaliação de desempenho do classificador proposto.

A partir dos resultados *in vivo* apresentados no trabalho de Lucca *et al.* (2021), é desenvolvido um modelo de aprendizado de máquina que utiliza parâmetros de cinética espermática para auxílio na identificação da fertilidade de machos suínos sem a necessidade da obtenção dos dados *in vivo* (taxa de parto e número total de leitões nascidos).

### 4.1 BANCO DE DADOS

Nesta seção é apresentado o banco de dados desenvolvido por Lucca *et al.* (2021). O objetivo do estudo de Lucca *et al.* (2021) foi avaliar a fertilidade de machos suínos através de informações relacionadas à resistência do sêmen ao armazenamento. Esse trabalho foi aprovado pelo Comitê de Ética no Uso de Animais da Universidade Federal do Rio Grande do Sul, sob número 34329.

Esse banco de dados é constituído por informações referentes a 12 machos e 519 fêmeas. A coleta de sêmen foi realizada pelo mesmo técnico uma vez por semana, durante três semanas, para cada macho. As amostras de sêmen foram avaliadas quanto à motilidade e concentração utilizando um sistema CASA com 24h e 120h de armazenamento, gerando 24 medidas de cinética espermática no total. As fêmeas foram inseminadas com doses armazenadas por 24h pertencentes a um único macho, com uma média de 43 fêmeas para cada macho.

Após o período de gestação, foram documentados o número total de leitões nascidos (NT) e a taxa de parto (TP), indicadores *in vivo* mais recomendados para análise da fertilidade de machos suínos (LUCCA *et al.*, 2021).

### 4.2 ROTULAÇÃO DAS AMOSTRAS

A rotulação de cada uma das amostras coletadas foi feita a partir da rotulação dos 12 machos utilizados no experimento. Utilizando os dados de NT e TP, os machos foram rotulados em “fértil” e “subfértil” através do algoritmo *K-means* com  $k = 2$ . A partir desse resultado, todas as amostras de um determinado macho foram rotuladas conforme a classificação do mesmo macho. Por exemplo, um macho considerado fértil teve todas as suas amostras rotuladas como férteis; enquanto que um macho subfértil teve todas as suas amostras rotuladas como subférteis. Para facilitar o uso de algoritmos de aprendizado de máquina, as classes “fértil” e “subfértil” foram substituídas por 0 e 1, respectivamente.

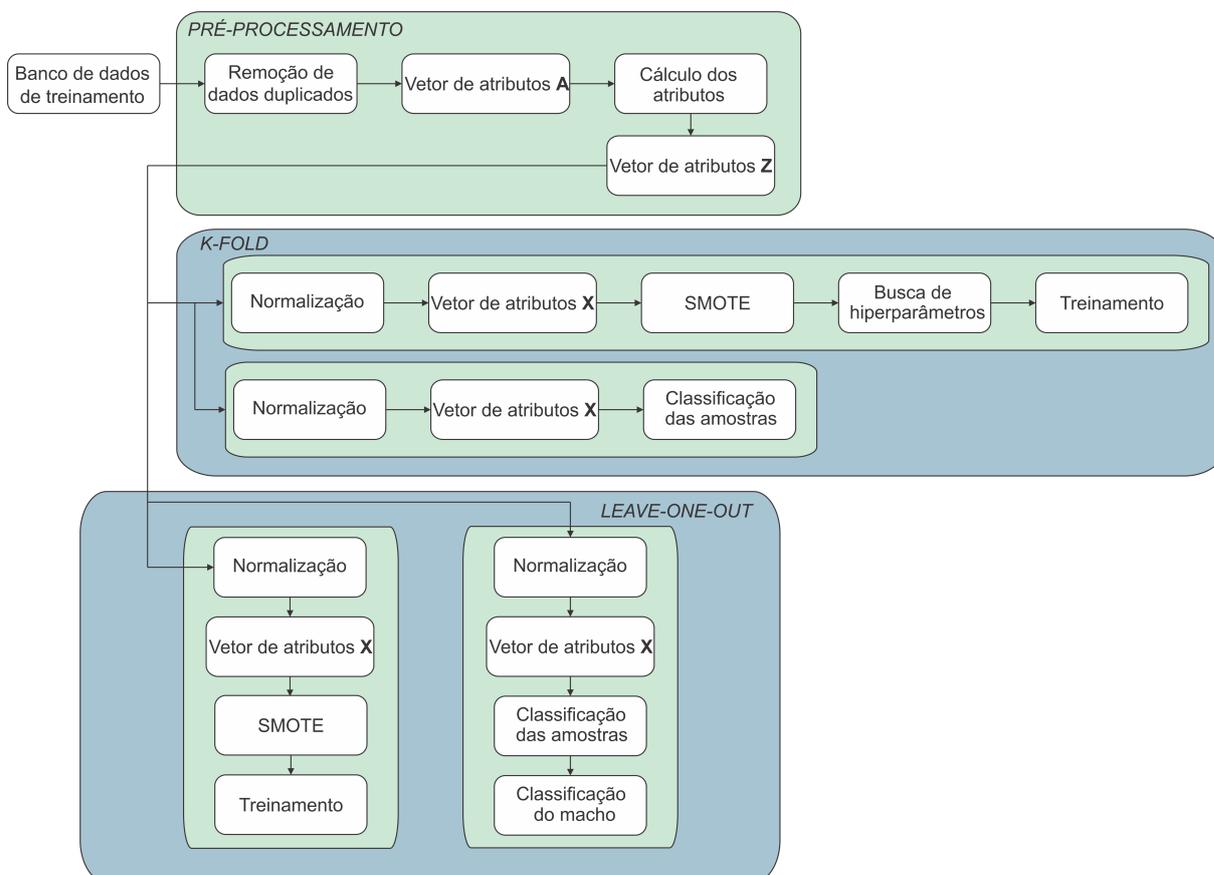
### 4.3 ANÁLISE PRELIMINAR

A primeira abordagem foi separar o conjunto de dados em treinamento e teste. A ideia seria utilizar os dados de teste para avaliar o desempenho do classificador projetado pelo conjunto de treinamento e, assim, obtermos medidas de desempenho a respeito da capacidade de generalização do classificador. Essa metodologia é muito empregada em aplicações de aprendizado de máquina, entretanto, durante o desenvolvimento do trabalho, percebeu-se que o banco de dados é escasso demais para que se utilize essa abordagem. O conjunto de teste possuiria poucos dados e, portanto, uma avaliação de desempenho nesses dados não teria significância estatística. Por conseguinte, optou-se por avaliar o desempenho do classificador apenas durante a etapa de validação.

### 4.4 PRÉ-PROCESSAMENTO, TREINAMENTO E VALIDAÇÃO

A Figura 9 apresenta um diagrama de blocos da metodologia empregada para o pré-processamento, treinamento e validação dos dados.

Figura 9 – Ilustração das etapas de pré-processamento, treinamento e validação.



Fonte: do autor.

A etapa de pré-processamento é realizada antes do treinamento do classificador, de forma a tratar os dados originais, deixando-os prontos para o aprendizado do algoritmo de classificação.

Inicialmente, são descartados os vetores de atributos duplicados. Isso acontece pelo fato de uma mesma amostra de sêmen de determinado macho ser utilizada para inseminar mais de uma fêmea, de forma a minimizar o efeito da fêmea nos dados *in vivo*. Pelo fato da amostra ser a mesma, os parâmetros de cinética espermática são idênticos. Dessa forma, no banco de dados, há linhas em que os valores de parâmetros de cinética espermática são idênticos entre si, em que a diferença se dá nos dados da fêmea e da prole gerada. Como estamos interessados apenas nos dados de cinética espermática, as linhas idênticas entre si são descartadas, deixando apenas uma delas.

Feita a remoção dos valores duplicados, é calculada a média, a variância, o valor máximo e o valor mínimo dos atributos que serão utilizados para o treinamento do classificador em conjuntos de três amostras (linhas da tabela do banco de dados). Por exemplo, as linhas de número 1 a 3 são utilizadas para se calcular a média e variâncias das medidas de cinética espermática. Em seguida, utilizam-se as linhas de número 4 a 6 e assim por diante, até que todas as linhas sejam utilizadas. Esse procedimento tem o objetivo de acrescentar informações de mais de uma amostra de um mesmo macho para a classificação deste. Portanto, o novo conjunto de dados passa a ser composto das médias e variâncias dos atributos selecionados para a classificação, originando o vetor de atributos  $\mathbf{Z}$ . A escolha do número de amostras foi determinada a partir de um conjunto de testes com duas, três e quatro amostras, de forma a ser possível combinar informações de diferentes amostras e ainda possuir amostras suficientes para serem classificadas. Nesse sentido, o número de três amostras obteve melhor desempenho e, portanto, foi escolhido.

Em seguida, os dados são normalizados pela pontuação padrão, dada por:

$$\mathbf{X} = \frac{\mathbf{Z} - \mu}{\sigma} \quad (18)$$

em que  $\mathbf{X}$  representa a pontuação padrão (novo valor do dado),  $\mathbf{Z}$  é o dado cujo valor desejamos normalizar,  $\mu$  é o valor médio dos dados do conjunto e  $\sigma$  o desvio padrão.

Essa transformação é realizada de forma a diminuir a influência de *outliers* na etapa de treinamento do classificador, além de alguns algoritmos de aprendizado de máquina requererem que os dados de treinamento sejam normalizados.

Após a normalização, os dados são submetidos ao algoritmo SMOTE para geração de amostras sintéticas, de forma a se igualar a quantidade de amostras da classe minoritária à da majoritária. Para a geração de amostras sintéticas pelo algoritmo SMOTE, foram utilizados  $k = 3$  vizinhos.

O procedimento de validação *K-fold* consiste em dividir o conjunto de dados em  $K$  subconjuntos, denominados de *folds*, em que  $K-1$  *folds* são utilizados para treinamento e o *fold* restante para validação. Isso é feito  $K$  vezes, até que todos os *folds* tenham

sido utilizados como validação, sendo o desempenho de validação a média da métrica de avaliação utilizada em todos os *folds*. Foram utilizados 3 *folds* estratificados para validação, isto é, mantendo-se a mesma proporção dos dados minoritários em cada um dos *folds*. A escolha do número de *folds* se deu por um conjunto de testes com diferentes valores de  $K$ , em que  $K = 3$  obteve o melhor desempenho. Juntamente com a validação  $K$ -*fold*, foi utilizado um algoritmo de otimização de hiperparâmetros (*grid search*) para a busca dos hiperparâmetros dos modelos, de forma a selecionar os hiperparâmetros que obtiveram melhor desempenho na média dos  $K$  *folds* de validação. A métrica de validação utilizada foi a medida  $F_1$ , métrica recomendada para problemas de classificação desbalanceados (ALI; SHAMSUDDIN; RALESCU, 2015). A Tabela 2 apresenta os valores de hiperparâmetros definidos para a otimização. Os valores foram definidos inicialmente a partir da heurística e, conforme o desempenho que apresentavam, os intervalos eram restringidos ou aumentados. Após vários experimentos, foram definidos os valores que se apresentam na Tabela 2.

Para a etapa de treinamento, foram utilizados dois classificadores; um classificador de regressão logística e outro SVM.

Tabela 2 – Valores de hiperparâmetros definidos para o algoritmo de otimização

Regressão Logística	
Parâmetros	Valores
C	[0.01, 0.1, 1, 10, 100]
penalty	'L2'
class_weight	'balanced'
max_iter	[300, 400, 500]
solver	'liblinear'
SVM	
Parâmetros	Valores
C	[0.01, 0.1, 1, 10, 100]
kernel	['rbf', 'poly', 'sigmoid']
gamma	[0.08, 0.12, 0.16, 0.20, 0.24]
class_weight	'balanced'

Fonte: do autor.

Os vetores de atributos selecionados para os cálculos de média, variância, valor máximo e mínimo são apresentados na Tabela 3. A seleção desses atributos se deu a partir das sugestões de Kummer *et al.* (2013) e Lucca *et al.* (2021), que apontam motilidade total, motilidade parcial, amplitude de deslocamento lateral da cabeça e preservação dos níveis de motilidade após armazenamento como possíveis indicadores da fertilidade dos reprodutores. Pode-se perceber que metade dos vetores de atributos utiliza dados de armazenamento de 120 horas, enquanto que a outra não. Nesse sentido, também podemos comparar os resultados e inferir sobre a influência da utilização de dados de armazenamento para a classificação da fertilidade dos machos. O treinamento dos classificadores foi realizado para

quatro vetores de atributos (Tabela 4) compostos por associações dos valores da média, variância, valor máximo e mínimo de cada um dos atributos presentes na Tabela 3.

Tabela 3 – Vetores de atributos para cálculo de média, variância, valor máximo e mínimo

Vetor de atributos	Atributos
A0	MT, MP, ALH, MT120, MP120
A1	MT, MP, ALH

Fonte: do autor.

Tabela 4 – Vetores de atributos para treinamento (após normalização)

Vetor de atributos	Atributos
X0	$\overline{MT}$ , $\overline{MP}$ , $\overline{ALH}$ , $\overline{MT120}$ , $\overline{MP120}$ , $\sigma^2(MT)$ , $\sigma^2(MP)$ , $\sigma^2(ALH)$ , $\sigma^2(MT120)$ , $\sigma^2(MP120)$
X1	$\overline{MT}$ , $\overline{MP}$ , $\overline{ALH}$ , $\sigma^2(MT)$ , $\sigma^2(MP)$ , $\sigma^2(ALH)$
X2	$\overline{MT}$ , $\overline{MP}$ , $\overline{ALH}$ , $\overline{MT120}$ , $\overline{MP120}$ , $\max(MT)$ , $\max(MP)$ , $\max(ALH)$ , $\max(MT120)$ , $\max(MP120)$ , $\min(MT)$ , $\min(MP)$ , $\min(ALH)$ , $\min(MT120)$ , $\min(MP120)$
X3	$\overline{MT}$ , $\overline{MP}$ , $\overline{ALH}$ , $\max(MT)$ , $\max(MP)$ , $\max(ALH)$ , $\min(MT)$ , $\min(MP)$ , $\min(ALH)$

Fonte: do autor.

O treinamento do modelo foi realizado com o objetivo de classificar o vetor de atributos das amostras dos machos de acordo com a classificação do macho. Por exemplo, uma amostra que foi rotulada como “fértil” é proveniente de um macho classificado como “fértil”, portanto, o algoritmo tentará classificá-la nessa classe.

Após a validação *K-fold* para busca dos hiperparâmetros do modelo de classificação, é feita uma segunda etapa de validação para decidir a regra de classificação dos machos do conjunto de treinamento. Nessa etapa, é utilizado o classificador com melhor desempenho na primeira etapa de validação (já definidos o vetor de atributos e os parâmetros). A segunda etapa de validação consiste em encontrar uma regra de classificação dos machos a partir da classificação dada pelo modelo de suas amostras. A técnica de validação empregada nessa etapa é a LOO (do inglês - *Leave one out*), que consiste em separar um macho para testar e utilizar o restante para treinar o modelo. Esse procedimento é repetido até que todos os machos tenham sido utilizados para teste e, dessa forma, obtém-se a classificação de cada um dos machos do conjunto feita pelo classificador treinado a partir da classificação das amostras. A regra de decisão empregada nessa etapa consiste em classificar como fértil um macho que tenha a maior parte de suas amostras classificadas como férteis. Caso as amostras sejam igualmente classificadas em férteis e subférteis ou majoritariamente em subférteis, o macho é classificado como subfértil. Essa regra de decisão foi obtida a partir da análise da classificação das amostras dos machos, em que observou-se que as amostras de machos férteis são majoritariamente classificadas como férteis, enquanto que as amostras de machos subférteis são igualmente classificadas nas duas classes.

#### 4.5 AVALIAÇÃO DE DESEMPENHO

Pelo fato de não termos dados suficientes para serem divididos em treinamento e teste, a análise de desempenho do classificador se deu pela validação LOO. Dessa forma, a classificação dos machos ocorre a partir da classificação de suas amostras como pertencentes a férteis ou subférteis seguida pela regra de decisão descrita anteriormente.

## 5 RESULTADOS

Neste capítulo, serão apresentados os resultados obtidos pelo classificador das amostras e dos machos para os vetores de atributos com e sem dados de armazenamento de 120 horas.

A Tabela 5 apresenta os classificadores das amostras, com seus respectivos valores de hiperparâmetros, e a medida F1 obtida na validação *K-fold*.

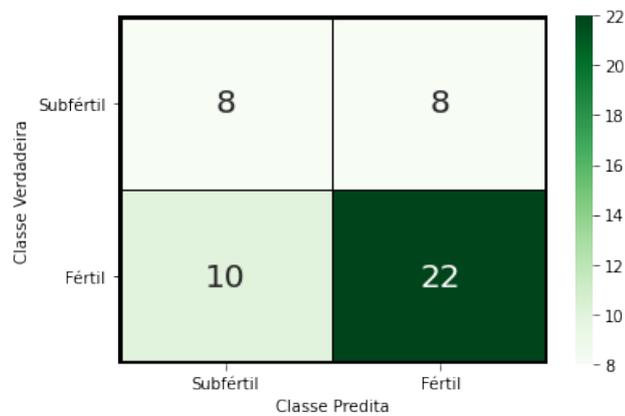
Tabela 5 – Classificadores das amostras, com vetor de atributos utilizado, valores de hiperparâmetros selecionados por otimização e o valor da medida F1 de validação.

Classificador	Vetor de Atributos	Valores de hiperparâmetros	F1 validação
Regressão Logística	X0	C = 10 penalty = 'L2' class_weight = 'balanced' max_iter = 300 solver = 'liblinear'	0,57
Regressão Logística	X1	C = 100 penalty = 'L2' class_weight = 'balanced' max_iter = 300 solver = 'liblinear'	0,60
Regressão Logística	X2	C = 100 penalty = 'L2' class_weight = 'balanced' max_iter = 300 solver = 'liblinear'	0,65
Regressão Logística	X3	C = 10 penalty = 'L2' class_weight = 'balanced' max_iter = 300 solver = 'liblinear'	0,64
SVM	X0	C = 10 kernel = 'rbf' gamma = 0,08 class_weight = 'balanced'	0,70
SVM	X1	C = 10 kernel = 'rbf' gamma = 0,08 class_weight = 'balanced'	0,73
SVM	X2	C = 10 kernel = 'rbf' gamma = 0,08 class_weight = 'balanced'	0,59
SVM	X3	C = 10 kernel = 'sigmoid' gamma = 0,08 class_weight = 'balanced'	0,59

Fonte: do autor.

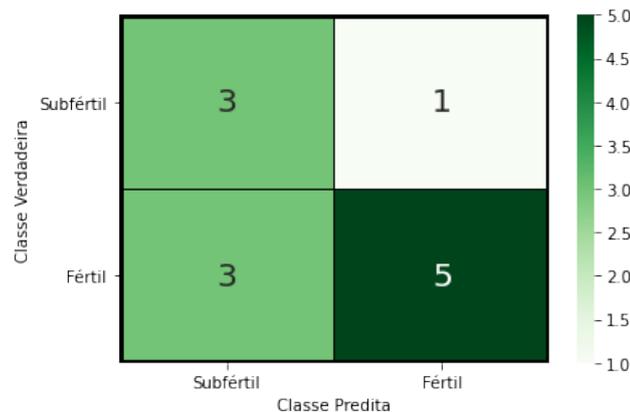
O classificador com maior medida F1 para a classificação das amostras foi também o que apresentou melhor desempenho na classificação dos machos na etapa de validação LOO. A Figura 10 apresenta a matriz de confusão da classificação das amostras de todos os machos na validação LOO, de onde foi extraída a regra de decisão para a classificação dos machos. A Figura 11 ilustra a matriz de confusão da classificação dos machos, enquanto que as medidas de desempenho da classificação são apresentadas na Tabela 6 e a curva ROC (juntamente com AUC) na Figura 12.

Figura 10 – Matriz de confusão da classificação das amostras para o melhor classificador obtido



Fonte: do autor.

Figura 11 – Matriz de confusão da classificação dos machos para o melhor classificador obtido



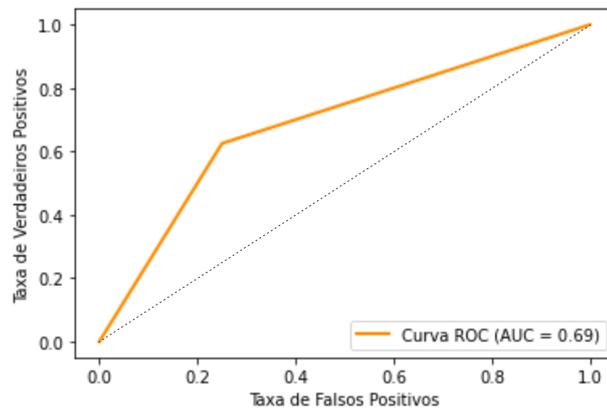
Fonte: do autor.

Tabela 6 – Medidas de desempenho da classificação dos machos para o melhor classificador obtido

Classificador	Vetor de Atributos	Acurácia	Precisão	Sensibilidade	F1	MCC
SVM	X1	0,67	0,67	0,69	0,66	0,35

Fonte: do autor.

Figura 12 – Curva ROC da classificação dos machos para o melhor classificador obtido



Fonte: do autor.



## 6 DISCUSSÃO E CONCLUSÃO

A seleção de reprodutores favorece o desempenho produtivo da granja, sendo a qualidade de sêmen, idade do reprodutor e condições de saúde fatores intimamente relacionados a essa questão (KNECHT; JANKOWSKA-MAKOSA; DUZIŃSKI, 2017), (BENNEMANN *et al.*, 2020).

No sentido de auxiliar na identificação da fertilidade de suínos, este trabalho teve como propósito utilizar técnicas de aprendizado de máquina e parâmetros de cinética espermática para a classificação de machos suínos em férteis e subférteis, o que se mostra de grande interesse para a suinocultura. Motivado pelos resultados obtidos por Lucca *et al.* (2021) e visando a contribuir com uma inferência sobre a classificação de fertilidade suína a partir da qualidade seminal, o presente estudo fez uso do banco de dados gerado em Lucca *et al.* (2021) para avaliar de forma preliminar a identificação da fertilidade de reprodutores suínos.

Há poucas abordagens parecidas com essa na literatura, sendo o trabalho de Kummer *et al.* (2013) o que mais se assemelha com a proposta aqui apresentada. A técnica de *cluster* para rotulação dos ejaculados utilizando dados de taxa de parto e número total de nascidos foi inspirada no trabalho de Kummer *et al.* (2013). Nesse trabalho, os autores tentam classificar cada um dos ejaculados em sua respectiva classe (de acordo com a classe dos próprios reprodutores). Essa foi a primeira abordagem testada no presente trabalho, porém os resultados apresentaram uma medida  $F1 = 0,45$  para a regressão logística e  $F1 = 0,41$  para o SVM.

No presente trabalho, foi proposta uma técnica de cálculo de média, variância, valor máximo e mínimo dos atributos em um conjunto de três amostras, de forma a ter um vetor de atributos que melhor representasse determinado reprodutor. Para sua realização, foram encontradas duas grandes limitações, um banco de dados pequeno e com desbalanceamento amostral. Essas duas problemáticas, em conjunto, dificultam o treinamento e análise de desempenho dos classificadores e regra de decisão propostos. A abordagem clássica de separação do conjunto de dados em treinamento e teste, de forma a analisar a capacidade de generalização do classificador, não pôde ser implementada devido a essas limitações. Entretanto, este trabalho traz uma inovação no que diz respeito a técnicas de aprendizado de máquina para classificação de machos suínos através de parâmetros de cinética espermática, o que não foi encontrado na literatura. Os resultados obtidos são modestos em termos percentuais quando comparados a outras aplicações de aprendizado de máquina, mas tornam-se significativos em virtude da incapacidade humana de classificação somente a partir dos parâmetros de cinética espermática disponíveis.

Um ponto interessante a se destacar é a utilização de dados de 120h de armazenamento para a classificação dos machos. A literatura aponta que podem ser bons indicadores; entretanto, este trabalho concluiu que os melhores resultados obtidos não levam em con-

sideração esses dados. Uma possibilidade para essa divergência é a necessidade de uma maior quantidade de dados para que seja possível o reconhecimento da influência dos dados de 120h de armazenamento por parte dos classificadores.

Dessa forma, o presente estudo se apresenta como um ponto de partida para a exploração de técnicas de aprendizado de máquina para classificação de reprodutores suínos através de parâmetros de cinética espermática. Além disso, também contribui com informações para auxílio na identificação da fertilidade dos reprodutores a partir somente da avaliação da qualidade seminal, apresentando desempenho elevado frente à incapacidade humana de identificação apenas com dados de parâmetros de cinética espermática.

## REFERÊNCIAS

ABCS. **Dados de Mercado de Suínos 2021**. 2021. Disponível em: <https://abcs.org.br/dados-do-setor/>. Acesso em: 1 jun. 2022.

ABCS. **Produção de suínos: teoria e prática**. Brasília: ABCS, 2014.

ALI, Aida; SHAMSUDDIN, Siti Mariyam; RALESCU, Anca. Classification with class imbalance problem: A review. v. 7, p. 176–204, jan. 2015.

AMANN, Rupert P.; WABERSKI, Dagmar. Computer-assisted sperm analysis (CASA): Capabilities and potential developments. **Theriogenology**, v. 81, n. 1, 5–17.e3, 2014. ISSN 0093-691X. DOI: <https://doi.org/10.1016/j.theriogenology.2013.09.004>.

ARTHUR, David; VASSILVITSKII, Sergei. K-means++: the advantages of careful seeding. *In: IN Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*. [S.l.: s.n.], 2007.

BENNEMANN, Paulo Eduardo; BRAGANÇA, José Francisco Manta; WALTER, Marina Patricia; BOTTAN, Juliano; MACHADO, Sérgio Abreu. Characterization of Boar Studs in Brazil. **Ciência Rural**, FapUNIFESP (SciELO), v. 50, n. 11, 2020. DOI: [10.1590/0103-8478cr20190998](https://doi.org/10.1590/0103-8478cr20190998).

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, AI Access Foundation, v. 16, p. 321–357, jun. 2002. Disponível em: <https://doi.org/10.1613/jair.953>.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, A. C. P. L. F. de; ALMEIDA, T. A. de. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: LTC, 2011.

GADEA, J; SELLÉS, E; MARCO, MA. The Predictive Value of Porcine Seminal Parameters on Fertility Outcome under Commercial Conditions. **Reproduction in Domestic Animals**, Wiley, v. 39, n. 5, p. 303–308, set. 2004. DOI: [10.1111/j.1439-0531.2004.00513.x](https://doi.org/10.1111/j.1439-0531.2004.00513.x).

GARCÍA, Rodrigo; AGUILAR, Jose; TORO, Mauricio; PINTO, Angel; RODRÍGUEZ, Paul. A systematic literature review on the use of machine learning in

precision livestock farming. **Computers and Electronics in Agriculture**, v. 179, p. 105826, 2020. DOI: <https://doi.org/10.1016/j.compag.2020.105826>.

GUBERT, Paulo Henrique. **Contribuição à Matriz de Informações em Sistemas de Imagética Motora Baseadas no Método de CSP**. 2019. Mestrado em Engenharia Elétrica – Universidade Federal de Santa Catarina.

HE, Haibo; GARCIA, Edwardo A. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009. DOI: 10.1109/TKDE.2008.239.

KAMPHUIS, Claudia; DUENK, Pascal; VEERKAMP, Roel Franciscus; VISSER, Bram; SINGH, Gurnoor; NIGSCH, Annette; DE MOL, Rudi Maria; BROEKHUIJSE, Marleen Leonarda Wilhelmina Johanna. Machine learning to further improve the decision which boar ejaculates to process into artificial insemination doses. **Theriogenology**, v. 144, p. 112–121, 2020. DOI: <https://doi.org/10.1016/j.theriogenology.2019.12.017>.

KHALIFA, T.; REKKAS, C.; SAMARTZI, F.; LYMPEROPOULOS, A.; KOUSENIDIS, K.; DOVENSKI, T. Highlights on Artificial Insemination (AI) Technology in the Pigs. **Macedonian Veterinary Review**, v. 37, mar. 2014. DOI: 10.14432/j.macvetrev.2013.09.001.

KNECHT, Damian; JANKOWSKA-MAKOSA, Anna; DUZIŃSKI, Kamil. Analysis of the lifetime and culling reasons for AI boars. **Journal of Animal Science and Biotechnology**, Springer Science e Business Media LLC, v. 8, n. 1, jun. 2017. DOI: 10.1186/s40104-017-0179-z.

KUHLGATZ, D.A.; KUHLGATZ, C.; AEPLI, M.; SCHUMANN, B.; GROSSFELD, R.; BORTFELDT, R.; JAKOP, U.; JUNG, M.; SCHULZE, M. Development of predictive models for boar semen quality. **Theriogenology**, Elsevier BV, v. 134, p. 129–140, ago. 2019. DOI: 10.1016/j.theriogenology.2019.05.024.

KUMMER, ABHP; GAGGINI, TS; BERNARDI, ML; MCMANUS, C; GONÇALES, EM; WENTZ, I; BORTOLOZZO, FP. Multivariate Analyses for Determining the Association of Field Porcine Fertility With Sperm Motion Traits Analysed by Computer-Assisted Semen Analysis and With Sperm Morphology. **Reproduction in Domestic Animals**, Wiley, v. 48, n. 5, p. 747–754, mar. 2013. DOI: 10.1111/rda.12155.

LUCCA, Matheus. **Indicadores in vitro e in vivo da fertilidade de machos que diferem em relação a resistência ao resfriamento das doses de sêmen**. 2020. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul.

LUCCA, Matheus Schardong; GIANLUPPI, Rafael; PEREIRA, Victória; MELLAGI, Ana Paula; ULGUIM, Rafael; BORTOLOZZO, Fernando. Avaliação in vitro e in vivo de reprodutores suínos com diferentes sensibilidades ao armazenamento a 17 °C. *In: SIMPÓSIO Internacional de Suinocultura XII: Avanços em sanidade, produção e reprodução de suínos*. [S.l.: s.n.], mai. 2019.

LUCCA, Matheus Schardong; GIANLUPPI, Rafael Dal Forno; MELLAGI, Ana Paula Gonçalves; BORTOLOZZO, Fernando Pandolfo; WENTZ, Ivo; ROSA ULGUIM, Rafael da. Effects of the classification of boars according to progressive sperm motility and the extender type on the reproductive performance of a single fixed-time insemination. **Theriogenology**, v. 161, p. 120–125, 2021. DOI: <https://doi.org/10.1016/j.theriogenology.2020.11.018>.

MARTINS, F. M.; TALAMINI, D. J. D.; SANTOS FILHO, J. I. **Conjuntura econômica da suinocultura brasileira**. 2019. Disponível em: <https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1117392>. Acesso em: 1 jun. 2022.

MOHAMMED, Roweida; RAWASHDEH, Jumanah; ABDULLAH, Malak. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *In: 2020 11th International Conference on Information and Communication Systems (ICICS)*. [S.l.: s.n.], 2020. P. 243–248. DOI: [10.1109/ICICS49469.2020.239556](https://doi.org/10.1109/ICICS49469.2020.239556).

MOREIRA, Fabiana; FERREIRA, Carlos Eduardo Ranquetat; PANZARDI, Andrea; CORCINI, Carine Dahl. TÉCNICAS DE INSEMINAÇÃO ARTIFICIAL E UTILIZAÇÃO DE DIFERENTES DOSES INSEMINANTES EM SUÍNOS. **Science And Animal Health**, Universidade Federal de Pelotas, v. 1, n. 1, p. 50, mai. 2013. DOI: [10.15210/sah.v1i1.485](https://doi.org/10.15210/sah.v1i1.485). Disponível em: <https://doi.org/10.15210/sah.v1i1.485>.

NEETHIRAJAN, Suresh. The role of sensors, big data and machine learning in modern animal farming. **Sensing and Bio-Sensing Research**, v. 29, p. 100367, 2020. DOI: <https://doi.org/10.1016/j.sbsr.2020.100367>.

PILES, Miriam; BERGSMA, Rob; GIANOLA, Daniel; GILBERT, H el ene; TUSELL, Llibertat. Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. **Frontiers in Genetics**, v. 12, p. 137, 2021. DOI: 10.3389/fgene.2021.611506.

QUINTERO-MORENO, Armando; RIGAU, Teresa; RODR IGUEZ-GIL, Joan E. Regression analyses and motile sperm subpopulation structure study as improving tools in boar semen quality analysis. **Theriogenology**, Elsevier BV, v. 61, n. 4, p. 673–690, fev. 2004. DOI: 10.1016/s0093-691x(03)00248-6.

RUIZ-S ANCHEZ, A.L.; O'DONOGHUE, R.; NOVAK, S.; DYCK, M.K.; COSGROVE, J.R.; DIXON, W.T.; FOXCROFT, G.R. The predictive value of routine semen evaluation and IVF technology for determining relative boar fertility. **Theriogenology**, Elsevier BV, v. 66, n. 4, p. 736–748, set. 2006. DOI: 10.1016/j.theriogenology.2005.12.012. Dispon ivel em: <https://doi.org/10.1016/j.theriogenology.2005.12.012>.

SANTOS, Miriam Seoane; SOARES, Jastin Pompeu; ABREU, Pedro Henriques; ARAUJO, Helder; SANTOS, Joao. Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. **IEEE Computational Intelligence Magazine**, v. 13, n. 4, p. 59–76, 2018. DOI: 10.1109/MCI.2018.2866730.

SCHULZE, M.; RUEDIGER, K.; MUELLER, K.; JUNG, M.; WELL, C.; REISSMANN, M. Development of an in vitro index to characterize fertilizing capacity of boar ejaculates. **Animal Reproduction Science**, Elsevier BV, v. 140, n. 1-2, p. 70–76, jul. 2013. DOI: 10.1016/j.anireprosci.2013.05.012.

SILVA, Henrique Pickler da. **Predic o de Incid ncia de Les o por Press o em UTI atrav s de Registros Eletr nicos e Aprendizado de M quina**. 2021. TCC (Gradua o) - Curso de Engenharia El trica – Universidade Federal de Santa Catarina, Florian polis.

SMITH, R.P.; GAVIN, C.; GILSON, D.; SIMONS, R.R.L.; WILLIAMSON, S. Determining pig holding type from British movement data using analytical and machine learning approaches. **Preventive Veterinary Medicine**, v. 178, p. 104984, 2020. DOI: <https://doi.org/10.1016/j.prevetmed.2020.104984>.

THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. **Pattern Recognition**. 4<sup>a</sup> ed. [S.l.: s.n.], 2009.

TSAKMAKIDIS, I. A.; LYMBEROPOULOS, A. G.; KHALIFA. Relationship between sperm quality traits and field-fertility of porcine semen. **Journal of Veterinary Science**, The Korean Society of Veterinary Science, v. 11, n. 2, p. 151, 2010. DOI: 10.4142/jvs.2010.11.2.151.

VALVERDE, A.; BARQUERO, V.; SOLER, C. The application of computer-assisted semen analysis (CASA) technology to optimise semen evaluation. A review. **Journal of Animal and Feed Sciences**, v. 29, n. 3, p. 189–198, 2020. ISSN 1230-1388. DOI: 10.22358/jafs/127691/2020.

WABERSKI, Dagmar; SUAREZ, Susan S.; HENNING, Heiko. Assessment of sperm motility in livestock: Perspectives based on sperm swimming conditions in vivo. **Animal Reproduction Science**, p. 106849, 2021. DOI: <https://doi.org/10.1016/j.anireprosci.2021.106849>.