

Universidade de Brasília – UnB  
Faculdade de Ciências e Tecnologias em Engenharia – FCTE  
Engenharia de Software

**Aplicação de Aprendizado de Máquina na  
Predição Não-Invasiva de Euploidia em  
Embriões Humanos com Base em Dados  
Morfocinéticos**

Autor: Maria Eduarda Dos Santos Abritta Ferreira e Sabrina  
Caldas Berno

Orientador: Prof Dr. George Marsicano Corrêa

Brasília, DF

2025



Maria Eduarda Dos Santos Abritta Ferreira e Sabrina Caldas Berno

**Aplicação de Aprendizado de Máquina na Predição  
Não-Invasiva de Euploidia em Embriões Humanos com  
Base em Dados Morfocinéticos**

Monografia submetida ao curso de graduação  
em Engenharia de Software da Universidade  
de Brasília, como requisito parcial para ob-  
tenção do Título de Bacharel em Engenharia  
de Software.

Universidade de Brasília – UnB

Faculdade de Ciências e Tecnologias em Engenharia – FCTE

Orientador: Prof Dr. George Marsicano Corrêa

Brasília, DF

2025

*Este trabalho é dedicado a todas as mulheres que  
sonham em poder gerar a sua família*

# Resumo

A fertilização in vitro é amplamente utilizada por casais com dificuldades reprodutivas ou que desejam postergar a gravidez, mas seu sucesso depende de vários fatores, sendo a seleção de embriões viáveis um dos mais críticos. Escolher o embrião correto aumenta as chances de implantação, reduz o risco de abortos e melhora as taxas de nascimentos saudáveis. O Teste Genético Pré-Implantacional para Aneuploidia é o método mais usado, mas, além do custo elevado, apresenta limitações de precisão diagnóstica, podendo levar à transferência de embriões inviáveis ou exclusão de viáveis. Com o avanço de tecnologias como o Time-Lapse System, que captura imagens contínuas do desenvolvimento embrionário, surge a chance de prever a euploidia usando dados morfocinéticos de forma menos invasiva e mais eficaz. Este trabalho propõe uma abordagem baseada em Machine Learning para identificar padrões nos dados morfocinéticos e prever a euploidia, oferecendo uma alternativa ao PGT-A sem intervenções invasivas. Após revisão de literatura e análise de correlação, foram identificadas as variáveis mais relevantes, sendo a idade e tb-t2b as com maior influência negativa. Os próximos passos incluem treinar os modelos de Machine Learning e validar os resultados com métricas como acurácia, sensibilidade e especificidade. A comparação com o PGT-A avaliará a eficácia da abordagem. Espera-se que a análise dos dados morfocinéticos com Machine Learning alcance alta acurácia, superior a 70%, complementando os métodos invasivos e auxiliando na escolha do melhor embrião. A aplicação dessa solução pode aumentar as taxas de sucesso da FIV, oferecendo uma alternativa mais econômica e menos prejudicial. Este estudo tem potencial para transformar a reprodução assistida, reduzindo custos e riscos, além de tornar a escolha de embriões mais acessível e eficaz.

**Palavras-chave:** Fertilização in vitro. Euploidia. Time-Lapse System. PGT-A. Inteligência Artificial. Aprendizado de máquina. Análise de dados.

# Abstract

In vitro fertilization is widely used by couples facing reproductive difficulties or those wishing to postpone pregnancy, but its success depends on several factors, with the selection of viable embryos being one of the most critical. Choosing the correct embryo significantly increases the chances of successful implantation, reduces the risk of miscarriage, and improves healthy birth rates. Preimplantation Genetic Testing for Aneuploidy is the most commonly used method, but in addition to its high cost, it has diagnostic precision limitations, which can lead to the transfer of non-viable embryos or exclusion of viable ones. With the advancement of technologies such as the Time-Lapse System, which captures continuous images of embryonic development, the possibility arises of predicting euploidy using morphokinetic data in a less invasive and more effective manner. This work proposes a Machine Learning-based approach to identify patterns in morphokinetic data and predict euploidy, offering an alternative to PGT-A without invasive interventions. After a literature review and correlation analysis, the most relevant variables were identified, with age and tb-t2b showing the highest negative influence. The next steps include training Machine Learning models and validating the results using metrics such as accuracy, sensitivity, and specificity. The comparison with PGT-A will assess the approach's effectiveness. It is expected that Machine Learning analysis of morphokinetic data will achieve high accuracy, above 70%, complementing invasive methods and assisting in the selection of the best embryo. Applying this solution may significantly increase IVF success rates, while offering a more economical and less harmful alternative. This study has the potential to transform assisted reproduction practices by reducing costs and risks, making embryo selection more accessible and effective.

**Key-words:** In vitro fertilization. Euploidy. Time-Lapse System. PGT-A. Artificial Intelligence. Machine Learning. Data analysis.

# Lista de ilustrações

- Figura 1 – Classificação dos embriões após o teste genético pré-implantacional para aneuploidias (PGT-A). O PGT-A avalia o número de cromossomos em células do embrião, permitindo classificá-los em três categorias: euploides (número normal de cromossomos), mosaicos (mistura de células com número normal e anormal de cromossomos) e aneuploides (número anormal de cromossomos em todas as células). A figura ilustra esquematicamente as diferentes classificações e as respectivas probabilidades de sucesso gestacional. . . . . 21
- Figura 2 – Representação esquemática da biópsia do trofotoderma e a influência da amostragem na determinação da ploidia embrionária. (A) Embrião euploide com células da massa celular interna (MCI) e do trofotoderma (TE) com pouca ou nenhuma aneuploidia. A biópsia resulta em uma amostra com menos de 20% de células aneuploides, classificando o embrião como euploide. (B) Embrião mosaico com proporção similar de células euploides e aneuploides. A biópsia pode resultar em uma amostra com 20-80% de células aneuploides, classificando o embrião como mosaico. (C) Embrião predominantemente aneuploide. A biópsia resulta em uma amostra com mais de 80% de células aneuploides, classificando o embrião como aneuploide. Observação: A ploidia embrionária determinada pela biópsia do trofotoderma pode variar de acordo com a região amostrada, devido à mosaicidade embrionária. A figura ilustra a incerteza associada à classificação da ploidia embrionária com base em uma pequena amostra de células. . . . . 22
- Figura 3 – Os dados já possuem seus próprios rótulos. Já é definido o que significa ser euploide e aneuplóide. Os dados rotulados (ou seja, com rótulos já atribuídos) são usados para treinar o modelo. Nesse contexto, ao se referir a "euploide" e "aneuplóide", esses rótulos podem representar classes que o modelo deve aprender a identificar com base nas características dos dados, ajudando a prever ou classificar novos casos com precisão. . . 28
- Figura 4 – Dispersão entre Idade e t4 - Coeficiente de Spearman: -0.15 . . . . . 68
- Figura 5 – Dispersão entre Idade e t5 - Coeficiente de Spearman: 0,11 . . . . . 68
- Figura 6 – Dispersão entre Idade e tSB - Coeficiente de Spearman: -0.10 . . . . . 69
- Figura 7 – Dispersão entre Idade e cc2 (t3-t2) - Coeficiente de Spearman: -0.15 . . 69
- Figura 8 – Dispersão entre Idade e s2 (t4-t3) - Coeficiente de Spearman: -0.24 . . 69
- Figura 9 – Dispersão entre Idade e s3 (t8-t5) - Coeficiente de Spearman: -0.28 . . 69
- Figura 10 – Dispersão entre Idade e tB-tSB - Coeficiente de Spearman: 0.20 . . . . 70

Figura 11 – Dispersão entre Idade e cc3 (t5-t3) - Coeficiente de Spearman: 0.20 . . .	70
Figura 12 – Dispersão entre Morfo e cc2 (t3-t2) - Coeficiente de Spearman: -0.38 . .	72
Figura 13 – Dispersão entre Morfo e cc3 (t5-t3) - Coeficiente de Spearman: -0.31 . .	72
Figura 14 – Dispersão entre t2 e t4 - Coeficiente de Spearman: 0.89 . . . . .	72
Figura 15 – Dispersão entre t3 e t2 - Coeficiente de Spearman: 0.78 . . . . .	72
Figura 16 – Dispersão entre t4 e t5 - Coeficiente de Spearman: 0.56 . . . . .	73
Figura 17 – Dispersão entre t5 e t8 - Coeficiente de Spearman: 0.52 . . . . .	73
Figura 18 – Dispersão entre tSC e t2 - Coeficiente de Spearman: 0.40 . . . . .	74
Figura 19 – Dispersão entre tSC e t3 - Coeficiente de Spearman: 0.42 . . . . .	74
Figura 20 – Dispersão entre tSC e t4 - Coeficiente de Spearman: 0.43 . . . . .	74
Figura 21 – Dispersão entre tSC e t8 - Coeficiente de Spearman: 0.35 . . . . .	74
Figura 22 – Dispersão entre tSC e tSB - Coeficiente de Spearman: 0.75 . . . . .	75
Figura 23 – Dispersão entre tSC e tB - Coeficiente de Spearman: 0.74 . . . . .	75
Figura 24 – Dispersão entre tSB e t3 - Coeficiente de Spearman: 0.56 . . . . .	76
Figura 25 – Dispersão entre tSB e t4 - Coeficiente de Spearman: 0.57 . . . . .	76
Figura 26 – Dispersão entre tSB e tB - Coeficiente de Spearman: 0.93 . . . . .	76
Figura 27 – Dispersão entre cc2 (t3-t2) e t3 - Coeficiente de Spearman: 0.80 . . . .	78
Figura 28 – Dispersão entre cc3 (t5-t3) e t5 - Coeficiente de Spearman: 0.81 . . . .	79
Figura 29 – Dispersão entre cc3 (t5-t3) e t5-t2 - Coeficiente de Spearman: 0.81 . .	79
Figura 30 – Dispersão entre s3 (t8-t5) e t8 - Coeficiente de Spearman: 0.48 . . . .	82
Figura 31 – Conjunto de Treinamento: dados_treinamento.xlsx . . . . .	88
Figura 32 – Conjunto de Validação: dados_validacao.xlsx . . . . .	88
Figura 33 – Conjunto de Teste: dados_teste.xlsx . . . . .	88

# Lista de tabelas

Tabela 1	– Fase 1: Análise e Preparação de Dados . . . . .	36
Tabela 2	– Fase 2: Desenvolvimento e Avaliação do Modelo . . . . .	40
Tabela 3	– Planilha Normalizada . . . . .	85
Tabela 4	– Normalização da Idade e Kidscore . . . . .	86
Tabela 5	– Cronograma de atividades . . . . .	95
Tabela 6	– Cronograma de atividades . . . . .	96
Tabela 7	– Interpretação do coeficiente de correlação de Spearman . . . . .	113



# Lista de abreviaturas e siglas

IA	Inteligência Artificial
ICSI	Injeção Intracitoplasmática de Espermatozoides
FIV	Fertilização In Vitro
ML	Machine Learning
PGT-A	Teste Genético Pré-Implantacional para Aneuploidia
TRA	Transferência de Embriões
TLS	Time-Lapse System

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>13</b>
<b>1.1</b>	<b>Contexto</b>	<b>13</b>
<b>1.2</b>	<b>Motivação</b>	<b>14</b>
<b>1.3</b>	<b>Problema</b>	<b>15</b>
<b>1.4</b>	<b>Objetivos</b>	<b>16</b>
1.4.1	Objetivos Gerais	16
1.4.2	Objetivos Específicos	16
<b>1.5</b>	<b>Metodologia</b>	<b>16</b>
<b>1.6</b>	<b>Composição e estrutura do trabalho</b>	<b>17</b>
<b>2</b>	<b>REFERENCIAL TEÓRICO</b>	<b>18</b>
<b>2.1</b>	<b>Fertilização In Vitro</b>	<b>18</b>
<b>2.2</b>	<b>Métodos de Avaliação Genética em Reprodução Assistida</b>	<b>19</b>
<b>2.3</b>	<b>Time-Lapse System</b>	<b>23</b>
2.3.1	Idade	23
2.3.2	t2, t3, t4, t5, t8, s2, cc2 (t3-t2), tSC, tSB, tB, cc3 (t5-t3), s3 (t8-t5), t5-t2, tSC-t8 e tB-tSB	24
2.3.3	Estágio e Morfo	25
2.3.4	KIDScore™	26
2.3.5	Ploidia	26
<b>2.4</b>	<b>Aprendizado de Máquina</b>	<b>27</b>
2.4.1	O Algoritmo K-Nearest Neighbor	29
2.4.2	Regressão Linear	30
2.4.3	Naive Bayes	30
<b>2.5</b>	<b>Identificação de Padrões Morfocinéticos e Predição de Euploidia com IA e Trabalhos Correlatos</b>	<b>31</b>
<b>3</b>	<b>METODOLOGIA</b>	<b>34</b>
<b>3.1</b>	<b>Classificação da Pesquisa</b>	<b>34</b>
3.1.1	Natureza	34
3.1.2	Método ou Abordagem Metodológica	34
3.1.3	Objetivos	34
3.1.4	Procedimentos De Pesquisa	35
<b>3.2</b>	<b>Design da Pesquisa</b>	<b>35</b>
3.2.1	Fases de Trabalho	35
3.2.1.1	<b>Objetivo Específico 1 - Identificação de Parâmetros em Embriões</b>	<b>36</b>

3.2.1.1.1	<b>Atividade 1 (A1):</b> Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia . . . . .	36
3.2.1.1.2	<b>Atividade 2 (A2):</b> Normalização dos Dados para Otimização . . . . .	37
3.2.1.1.3	<b>Atividade 3 (A3):</b> Identificação da Correlação e Atribuição de Pesos aos Parâmetros na Previsão da Ploidia do Embrião . . . . .	38
3.2.1.1.4	<b>Atividade 4 (A4):</b> Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados . . .	39
3.2.1.2	<b>Objetivo Específico 2 -</b> Treinamento e Ajuste de Modelo de Machine Learning para Predição de Euploidia . . . . .	41
3.2.1.2.1	<b>Atividade 5 (A5):</b> Desenvolvimento e Treinamento do Modelo de Machine Learning para Otimização da Predição de Euploidia, Incluindo Treinamento, Validação e Teste . .	41
3.2.1.3	<b>Objetivo Específico 3 -</b> Avaliação do modelo . . . . .	42
3.2.1.3.1	<b>Atividade 6 (A6):</b> Utilizar métricas de avaliação mais adequadas para medir o desempenho do modelo de acordo com a natureza do problema de classificação . . . . .	42
3.2.1.3.2	<b>Atividade 7 (A7):</b> Avaliar a precisão e eficácia do modelo em prever corretamente casos de euploidia e aneuploidia por meio da Matriz de Confusão e Curva ROC . . . . .	43
3.2.1.4	<b>Objetivo Específico 4 -</b> Protótipo de Interface . . . . .	45
3.2.1.4.1	<b>Atividade 8 (A8):</b> Prototipar uma interface básica para exibir as previsões de euploidia para o usuário final (médicos) . . . . .	45
3.3	<b>Passos para o Desenvolvimento de um Algoritmo de Aprendizado de Máquina . . . . .</b>	45
3.3.1	Definição do problema e análise do panorama geral . . . . .	46
3.3.2	Obtenção de Dados . . . . .	46
3.3.3	Exploração de Dados . . . . .	47
3.3.4	Preparação dos dados para os Algoritmos de Aprendizado de Máquina . . .	48
3.3.5	Seleção e treinamento do modelo . . . . .	49
3.3.6	Ajuste do modelo . . . . .	50
3.3.7	Lançamento da Solução . . . . .	51
4	<b>EXECUÇÃO DA PESQUISA . . . . .</b>	53
4.1	<b>Fase 1: Análise e Preparação de Dados . . . . .</b>	53
4.1.1	OE1 - Expansão, Processamento e Análise de Dados para Predição de Ploidia	53
4.1.1.1	Atividade 1 (A1): Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia . . . . .	53
4.1.1.2	Limpeza dos Dados . . . . .	54
4.1.1.3	Atividade 2 (A2): Identificação da Correlação entre os Parâmetros na Previsão da Ploidia do Embrião . . . . .	54
4.1.1.4	Atividade 3 (A3): Normalização dos Dados para Otimização . . . . .	57

4.1.1.5	Atividade 4 (A4): Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados . . . . .	59
4.1.1.6	Nota sobre a Apresentação dos Códigos . . . . .	64
<b>5</b>	<b>ANÁLISE DOS RESULTADOS . . . . .</b>	<b>65</b>
<b>5.1</b>	<b>Fase 1: Análise e Preparação de Dados . . . . .</b>	<b>65</b>
5.1.1	OE1 - Expansão, Processamento e Análise de Dados para Predição de Ploidia	65
5.1.1.1	Atividade 1 (A1): Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia . . . . .	65
5.1.1.2	Atividade 2 (A2): Identificação da Correlação entre os Parâmetros na Previsão da Ploidia do Embrião . . . . .	66
5.1.1.3	Atividade 3 (A3): Normalização dos Dados para Otimização . . . . .	85
5.1.1.4	Atividade 4 (A4): Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados . . . . .	87
<b>6</b>	<b>CONSIDERAÇÕES E TRABALHOS FUTUROS . . . . .</b>	<b>90</b>
<b>6.1</b>	<b>Atividade 1 (A1): Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia . . . . .</b>	<b>90</b>
<b>6.2</b>	<b>Atividade 2 (A2): Identificação da Correlação entre os Parâmetros na Previsão da Ploidia do Embrião. . . . .</b>	<b>90</b>
<b>6.3</b>	<b>Atividade 3 (A3): Normalização dos Dados para Otimização . . . . .</b>	<b>91</b>
<b>6.4</b>	<b>Atividade 4 (A4): Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados. . . . .</b>	<b>92</b>
<b>6.5</b>	<b>Conclusão da Fase 1: Análise e Preparação de Dados . . . . .</b>	<b>93</b>
<b>6.6</b>	<b>Conclusão da Fase 1: Análise e Preparação de Dados . . . . .</b>	<b>93</b>
	<b>REFERÊNCIAS . . . . .</b>	<b>97</b>
	<b>GLOSSÁRIO . . . . .</b>	<b>103</b>
	<b>APÊNDICES . . . . .</b>	<b>105</b>
	<b>APÊNDICE A – VARIÁVEIS UTILIZADAS NA ANÁLISE DA PLOIDIA EMBRIONÁRIA . . . . .</b>	<b>106</b>
	<b>APÊNDICE B – Z-SCORE NORMALIZATION (STANDARDIZATION) . . . . .</b>	<b>109</b>

APÊNDICE C – MONTE CARLO . . . . .	111
APÊNDICE D – COEFICIENTE DE SPEARMAN . . . . .	112
<b>ANEXOS</b>	<b>114</b>
ANEXO I – PARECER DO COMITÊ DE ÉTICA EM PESQUISA	115
ANEXO II – TERMO DE CONSENTIMENTO PARA UTILIZAÇÃO DE DADOS DE ENTREVISTAS, GRAVAÇÃO DE REUNIÕES E USO DE GRAVAÇÃO . . . . .	123
ANEXO III – CONTRATO DE AUTORIZAÇÃO PARA UTILIZAÇÃO DE DADOS EM PESQUISA . . . . .	128

# 1 Introdução

## 1.1 Contexto

A fertilização in vitro (FIV) é uma das técnicas mais importantes de reprodução assistida, e tem ganhado crescente relevância no campo da medicina reprodutiva, oferecendo novas possibilidades para pessoas com dificuldades para engravidar. Maria Chaves Jardim destaca que o Brasil lidera a América Latina em número de procedimentos de fertilização in vitro, evidenciando a relevância e o avanço dessa tecnologia no país.([JARDIM, 2022](#)).

O sucesso da FIV está diretamente relacionado à saúde genética dos embriões utilizados. De acordo com [Ping et al. \(2023\)](#), a qualidade genética dos embriões é um fator crucial para o êxito da fertilização, uma vez que aneuploidias estão frequentemente associadas a desfechos desfavoráveis, como abortos espontâneos ou doenças genéticas nos fetos. Cerca de 50% dos casos de abortos espontâneos no primeiro trimestre estão relacionados a alterações cromossômicas ([SILVA et al., 2023](#)). Este cenário destaca a importância de métodos de avaliação genética, que se tornam ferramentas essenciais para aumentar as taxas de sucesso da FIV e melhorar a segurança dos tratamentos.

Para avaliar a qualidade dos embriões, são comumente utilizados testes genéticos de triagem embrionária, sendo o Teste Genético Pré-Implantação (PGT) um dos mais conhecidos e amplamente aplicados, especialmente o Teste Genético Pré-Implantação de Aneuploidia (PGT-A) ([YANG et al., 2024](#)). No entanto, a realização do PGT-A apresenta alguns desafios e limitações, como o tempo necessário para obtenção dos resultados, o custo elevado e a complexidade e o risco do procedimento, que envolve a biópsia embrionária (extração de células do trofotoderma) e, por isso, pode causar dano ao embrião ([YANG et al., 2024](#)). Assim, surge a oportunidade de se desenvolver um método de avaliação embrionária mais rápido, menos custoso e menos invasivo.

Um exemplo de tecnologia que pode unir Tecnologia de Reprodução Assistida (TRA) e Inteligência Artificial (IA) é a fotografia time-lapse incorporada às incubadoras de última geração, que geram informações sobre a morfologia e a cinética do desenvolvimento embrionário, e facilita a observação de eventos dinâmicos, seus tempos e padrões, definindo-os, em conjunto, como variáveis morfocinéticas ([MESEGUER et al., 2011](#)). Postula-se que, com o maior conhecimento da dinâmica embrionária em cultivo, a identificação de marcadores do potencial de implantação poderá fornecer, em futuro próximo, informações cruciais para o processo de escolha do embrião a ser transferido para o útero materno ([LUONG; LE, 2023](#)). Entretanto, até aqui, pouco se evoluiu no que diz respeito à

aplicação dos dados ou à análise de variáveis múltiplas sobre resultados clínicos objetivos.

A proposta deste projeto é desenvolver uma abordagem de baixo custo, utilizando IA para a detecção de padrões genéticos em embriões. Mais especificamente, o objetivo é identificar a probabilidade de um embrião ser euploide, isto é, aquele que possui a quantidade correta de cromossomos (23 pares), ou seja, os cromossomos estão organizados em pares completos (ZEGERS-HOCHSCHILD et al., 2017). Ao realizar a análise dos dados gerados pelo TLS, a IA poderá oferecer uma abordagem menos invasiva, eliminando a necessidade de testes genéticos caros e invasivos, como o PGT-A. Assim, ao melhorar a seleção dos embriões com maior potencial de euploidia, a IA poderá tornar o tratamento mais acessível para um maior número de pacientes e, no futuro, aumentar as taxas de sucesso da FIV.

Para as clínicas de fertilização, a implementação dessa tecnologia pode resultar em um fluxo de trabalho mais ágil e eficiente, otimizando recursos e melhorando a satisfação dos pacientes. A integração da IA na avaliação embrionária tem o potencial de transformar os tratamentos de fertilidade, proporcionando resultados mais positivos e uma experiência aprimorada para todos os envolvidos. Assim, este estudo busca ser uma solução menos invasiva dentre os métodos de seleção embrionária em tratamentos de reprodução assistida, integrando as inovações da IA à medicina reprodutiva.

## 1.2 Motivação

A motivação para a elaboração deste Trabalho de Conclusão de Curso surge da necessidade de contruibir para a criação de métodos mais acessíveis na medicina reprodutiva, especialmente no que diz respeito à avaliação da qualidade genética dos embriões em tratamentos de fertilidade. Essa área, que lida com questões pessoais e delicadas, fez avanços notáveis na última década (PANDIT; SHARMA, 2022), mas ainda enfrenta desafios significativos que exigem soluções eficazes. O avanço das tecnologias de reprodução assistida possibilitou que muitas pessoas, antes incapazes de conceber, realizassem o sonho da gravidez. No entanto, essa evolução trouxe consigo um novo desafio: a falha na implantação e a gravidez não viável. A repetição de ciclos sem sucesso pode gerar profunda frustração e desespero, fazendo com que os casais busquem incessantemente respostas e soluções (MONTAGNINI et al., 2010).

A jornada para a maternidade através da FIV é marcada por uma intensa carga emocional, com expectativas e incertezas. A perda gestacional, além do sofrimento emocional, pode gerar um impacto significativo na saúde física e mental da mulher (MONTAGNINI et al., 2010). De acordo com o estudo de Montagnini et al. (2010), as mulheres apresentam, em comparação aos homens, níveis mais elevados de ansiedade e depressão, além de uma autoestima mais baixa, com sentimentos de culpa e vergonha relacionados à

infertilidade. A ansiedade, em particular, é um dos principais desafios enfrentados pelos casais, frequentemente ultrapassando níveis considerados normais (MONTAGNINI et al., 2010).

A escolha cuidadosa dos embriões para a transferência é um passo crucial nesse processo de tentativa de adicionar um membro à família, pois a qualidade e a saúde dos embriões influenciam diretamente a taxa de sucesso da gravidez (YANG et al., 2024). O impacto da seleção de embriões com defeitos genéticos pode ser devastador para o emocional dos casais, agravando o sofrimento causado por abortos espontâneos e falhas de implantação. Ao eliminar a necessidade de intervenções, como o exame PGT-A, que podem gerar ansiedade e desconforto, podemos proporcionar um ambiente mais acolhedor e propício à realização do sonho de ter filhos.

Portanto, a relevância deste projeto no contexto atual da medicina reprodutiva é indiscutível. Ao abordar as limitações dos métodos tradicionais e explorar as potencialidades da IA, este estudo permite que os médicos tomem decisões mais informadas, com base em análises detalhadas, resultando em melhores resultados clínicos. Além disso, este trabalho destaca a importância da Engenharia de Software como uma ferramenta essencial para impulsionar o desenvolvimento de outras áreas do conhecimento. Ao aplicar técnicas avançadas de análise de dados e algoritmos de Machine Learning, ele contribui diretamente para o crescimento da medicina reprodutiva, promovendo inovações que tornam os tratamentos mais precisos, acessíveis e eficientes. A integração de abordagens interdisciplinares e a personalização do tratamento têm o potencial de gerar avanços significativos em médio e longo prazo, além de reduzir o tempo necessário para se obter um nascimento saudável. Assim, este projeto não apenas beneficia a área médica, mas também reforça o papel da Engenharia de Software como transformadora em diferentes campos científicos.

## 1.3 Problema

A seleção de embriões euploides em procedimentos de fertilização in vitro (FIV) é determinante para o sucesso do tratamento. No entanto, esse processo ainda depende de técnicas invasivas, como o Teste Genético Pré-Implantacional de Aneuploidia, um procedimento relativamente complexo, pois requer uma biópsia do embrião, durante a qual deve ser garantido o mínimo de danos ao embrião (YANG et al., 2024).

Dessa forma, o objetivo deste trabalho é responder à seguinte pergunta: "Como utilizar a inteligência artificial para identificar padrões em dados morfocinéticos de embriões, obtidos por meio do Time-Lapse System, para prever a porcentagem de euploidia, oferecendo uma solução mais eficaz e menos invasiva do que o PGT-A?"



## 1.4 Objetivos

### 1.4.1 Objetivos Gerais

Desenvolver uma abordagem baseada em inteligência artificial para identificar padrões em dados morfocinéticos de embriões, obtidos por meio do Time-Lapse System, capaz de prever a porcentagem de euploidia, proporcionando uma solução mais eficaz e menos invasiva em comparação ao PGT-A.

### 1.4.2 Objetivos Específicos

- **OE1:** Expansão, Processamento e Análise de Dados para Predição de Ploidia.
- **OE2:** Treinamento e Ajuste de Modelo de Machine Learning para Predição de Euploidia.
- **OE3:** Avaliação do Modelo
- **OE4:** Protótipo de Interface

## 1.5 Metodologia

A metodologia do projeto está dividida em 3 fases, onde cada uma visa resolver um Objetivo Específico (OE) do projeto, que contém suas respectivas atividades para serem alcançados:

- **Fase 1: Análise e Preparação de Dados**
  - **OE1 - Expansão, Processamento e Análise de Dados para Predição de Ploidia:**
    - \* **Atividade 1 (A1):** Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia
    - \* **Atividade 2 (A2):** Normalização dos Dados para Otimização.
    - \* **Atividade 3 (A3):** Identificação da Correlação e Atribuição de Pesos aos Parâmetros na Previsão da Ploidia do Embrião
    - \* **Atividade 4 (A4):** Divisão dos Dados e aplicação de Data Augmentation
- **Fase 2: Desenvolvimento e Avaliação do Modelo**
  - **OE2 - Treinamento e Ajuste de Modelo de Machine Learning para Predição de Euploidia:**

- \* **Atividade 5 (A5):** Desenvolvimento e Treinamento do Modelo de Machine Learning para Otimização da Predição de Euploidia, Incluindo Treinamento, Validação e Teste
- **OE3 - Avaliação do Modelo:**
  - \* **Atividade 6 (A6):** Utilizar métricas adequadas para medir o desempenho do modelo
  - \* **Atividade 7 (A7):** Avaliação do Desempenho do Modelo na Predição por meio da Matriz de Confusão e Curva ROC
- **OE4 - Protótipo de Interface:**
  - \* **Atividade 8 (A8):** Prototipar uma interface

## 1.6 Composição e estrutura do trabalho

Este trabalho foi organizado da seguinte maneira:

**Capítulo 2**, intitulado "Referencial Teórico", apresenta os principais conceitos que fundamentam a contextualização deste estudo.

**Capítulo 3**, intitulado "Metodologia", descreve os procedimentos e métodos utilizados na pesquisa, incluindo o planejamento de trabalho, as atividades realizadas e os resultados esperados. A seção 1.5 será detalhada ao longo deste capítulo.

**Capítulo 4**, intitulado "Execução da Pesquisa e Análise dos Resultados", apresenta a previsão de término de cada atividade e fase propostas ao longo do trabalho.

**Capítulo 5**, intitulado "Considerações e Trabalhos Futuros", expõe os avanços e resultados obtidos durante o período de desenvolvimento deste TCC.

**Capítulo 6**, intitulado "Planejamento", apresenta todas as atividades realizadas e as atividades futuras a serem realizadas.

## 2 Referencial Teórico

Para inicializar o nosso Referencial teórico de forma concisa em que se alinhe com o que foi proposto com o capítulo 1, é imprescindível abordar o contexto e papel da Fertilização In Vitro, qualidade dos embriões e desafios associados, além de evidenciar como o Aprendizado de Máquina pode auxiliar a aprimorar as taxas de sucesso de gravidez e minimizar os fatores de risco relacionados a abortos, problemas cromossômicos, além de reduzir os impactos emocionais e físicos associados a esses desafios.

### 2.1 Fertilização In Vitro

A fertilidade tem sido um tema de grande relevância ao longo da história humana, visto como uma bênção divina em diversas culturas. Civilizações antigas, como a grega e a egípcia, realizavam rituais, usavam amuletos e talismãs, ou buscavam ajuda de divindades para garantir a continuidade de suas linhagens e prosperidade (MOURA; SOUZA; SCHEFFER, 2020). Esses métodos, embora enraizados em crenças religiosas e espirituais, refletem o desejo universal de superar desafios relacionados à reprodução.

Com os avanços científicos e médicos, a compreensão da fertilidade passou por uma profunda transformação. O primeiro marco documentado foi a inseminação artificial em animais, realizada pelos árabes em 1332 (MOURA; SOUZA; SCHEFFER, 2020). Essa técnica consiste na introdução do sêmen diretamente no sistema reprodutor da fêmea, otimizando as chances de fertilização (CORLETA, 2010). Em humanos, um dos eventos mais notáveis ocorreu em 1978, com o nascimento de Louise Brown, o primeiro "bebê de proveta" (MOURA; SOUZA; SCHEFFER, 2020). Este feito foi possível graças ao desenvolvimento da técnica de Fertilização In Vitro (FIV), criada pelo embriologista Robert Edwards e pelo ginecologista Patrick Steptoe. A técnica permitiu a fertilização de embriões fora do corpo humano e, por sua contribuição revolucionária, Edwards recebeu o Prêmio Nobel de Fisiologia ou Medicina em 2010 (CORLETA, 2010).

As Técnicas de Reprodução Assistida (TRA) compreendem um conjunto de métodos médicos especializados que buscam ajudar indivíduos com dificuldades reprodutivas a alcançarem a concepção (SOUZA, 2024). Dentre essas técnicas, a FIV se destaca como a mais avançada e amplamente utilizada (MOURA; SOUZA; SCHEFFER, 2020). O processo envolve várias etapas, como a estimulação ovariana controlada (uso de medicamentos para estimular a produção de óvulos), coleta de óvulos por punção transvaginal, fertilização em laboratório e posterior transferência dos embriões formados para o útero (MOURA; SOUZA; SCHEFFER, 2020).

Um avanço significativo no campo da TRA foi a introdução da Injeção Intracitoplasmática de Espermatozoides (ICSI), na década de 1990 (PEREIRA; ALVES, 2016). O ICSI é uma técnica complementar à FIV que consiste na injeção direta de um único espermatozoide no citoplasma do óvulo, utilizando uma micropipeta (PEREIRA; ALVES, 2016). Este método é particularmente útil em casos de infertilidade masculina severa, como baixa contagem de espermatozoides, baixa motilidade ou presença de anomalias estruturais nos espermatozoides (PEREIRA; ALVES, 2016). Embora a FIV possa ser realizada sem o uso de ICSI, este último é frequentemente empregado em casos que exigem maior precisão na fertilização (PEREIRA; ALVES, 2016).

Além de estabelecer as bases da medicina reprodutiva moderna, o nascimento de Louise Brown abriu caminho para avanços na medicina reprodutiva. Desde então, a combinação de avanços médicos e tecnológicos permitiu não apenas a realização da fertilização in vitro, mas também a análise genética detalhada dos embriões (MOURA; SOUZA; SCHEFFER, 2020). Esses exames identificam anomalias cromossômicas e genéticas, proporcionando uma maior chance de sucesso na implantação e no desenvolvimento de gestações saudáveis.

Os esforços históricos e as inovações científicas ilustram a busca contínua da humanidade por soluções eficazes contra a infertilidade. No próximo tópico, discutiremos detalhadamente a análise de embriões, um procedimento crucial para aumentar as chances de concepção saudável e seu papel na e seu papel na seleção de embriões para a FIV.

## 2.2 Métodos de Avaliação Genética em Reprodução Assistida

No contexto das Tecnologias de Reprodução Assistida, os métodos de avaliação genética desempenham um papel essencial na identificação de anomalias cromossômicas, como a aneuploidia, que é a alteração no número normal de cromossomos da espécie humana, representa a principal causa de falhas de implantação quando sua origem é embrionária. Dessa forma, a presença de aneuploidia nas células do embrião pode impactar diretamente a taxa de sucesso das técnicas de reprodução humana assistida. Além de dificultar a implantação, essa alteração cromossômica pode levar a abortos espontâneos ou até mesmo a malformações em bebês nascidos vivos (SOUZA, 2022b). Após décadas de avanços científicos, os testes genéticos tornaram-se cada vez mais precisos, com o desenvolvimento de técnicas sofisticadas e integradas à tecnologia, como a Testagem Genética Pré-implantacional.

Pelo final do século XX, o método do PGT começou a ser usado para realizar o rastreamento de doenças genéticas que possuíam uma alta taxa de incidência nas populações de amostra (YANG et al., 2024). Com ele, foi propiciada a triagem de embriões antes da implantação, permitindo a seleção de embriões que possuíam menos riscos (??). Tais

métodos incluem a Testagem Genética Pré-implantacional para Doenças Monogênicas (PGT-M), como distrofia miotônica e fibrose cística; para rearranjos estruturais cromossômicos (PGT-SR); para aneuploidias (PGT-A), como Síndrome de Down e Síndrome de Turner; e mais recentemente, o PGT-P para doenças poligênicas (YANG et al., 2024). No caso do nosso objeto de estudo, PGT-A, se destaca por sua capacidade de aumentar as chances de implantação embrionária bem-sucedida, reduzir a probabilidade de perdas gestacionais espontâneas e garantir maior probabilidade de nascimento de crianças com o número de cromossomos normais (YANG et al., 2024). Além disso, o embrião considerado “normal” (euploide) apresenta 23 pares de cromossomos (46 cromossomos no total), sendo metade proveniente do espermatozoide e metade do óvulo. Já o embrião “anormal” (aneuploide) possui uma contagem incorreta de cromossomos em uma célula, sendo a maioria dos embriões com aneuploidias não compatíveis com a vida (ZEGERS-HOCHSCHILD et al., 2017).

Os testes genéticos realizados durante a fase de blastocisto, formados cerca de 5–6 dias após a inseminação (ZEGERS-HOCHSCHILD et al., 2017), são geralmente preferidos por especialistas em reprodução assistida porque essa abordagem envolve a coleta de células do trofotoderma, a camada externa do embrião que futuramente dará origem à placenta, e não diretamente do interior do embrião (LEAVER; WELLS, 2019). Essa técnica, o PGT-A, é considerada menos invasiva em comparação com as biópsias realizadas na fase de clivagem, quando o embrião possui apenas algumas células e está em um estágio muito inicial de desenvolvimento (LEAVER; WELLS, 2019). Na fase de clivagem, a retirada de células do interior do embrião (blastômeros), pode causar danos mais significativos ao embrião, afetando sua capacidade de se desenvolver adequadamente e reduzindo as chances de implantação bem-sucedida no útero (LEAVER; WELLS, 2019).

No entanto, é fundamental mencionar que a técnica de micromanipulação necessária para realizar esses procedimentos ainda não foi completamente padronizada. Isso significa que diferentes laboratórios podem adotar práticas distintas para a biópsia embrionária, o que pode resultar em variações na segurança e eficácia dos testes. Além disso, os impactos potenciais dessa intervenção, tanto nos desfechos reprodutivos (como taxas de gravidez e nascimento) quanto na saúde a longo prazo dos bebês nascidos a partir desses embriões, ainda não estão completamente compreendidos.

O PGT-A classifica os embriões em três categorias: euploide (normal, com 46 cromossomos), mosaico (mistura de células normais e anormais) e aneuploide (todas as células com número anormal de cromossomos). A Figura ?? ilustra os resultados do PGT-A e suas respectivas probabilidades de sucesso gestacional.


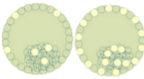

Resultados Possíveis do <b>PGT-A</b>	<b>Euplóide</b> 	<b>Mosaico</b> 	<b>Aneuplóide</b> 
Número de cromossomos	Normal	Misto: Alguns normais e outros anormais	Anormal
Probabilidade de produzir uma <b>gravidez bem-sucedida</b>	Alta	Intermediária	Baixa

Figura 1 – Classificação dos embriões após o teste genético pré-implantacional para aneuploidias (PGT-A). O PGT-A avalia o número de cromossomos em células do embrião, permitindo classificá-los em três categorias: euploides (número normal de cromossomos), mosaicos (mistura de células com número normal e anormal de cromossomos) e aneuploides (número anormal de cromossomos em todas as células). A figura ilustra esquematicamente as diferentes classificações e as respectivas probabilidades de sucesso gestacional.

Apesar de suas vantagens, o PGT-A apresenta limitações significativas, como a variabilidade nos resultados das biópsias do trofotoderma e o risco de diagnósticos falso-positivos (GLEICHER; PATRIZIO; BRIVANLOU, 2021). A The Preimplantation Genetic Diagnosis International Society (PGDIS) e a European Society of Human Reproduction and Embryology (ESHRE) apontam questões críticas, incluindo:

- As divergências no conteúdo de DNA aneuploide entre diferentes regiões da trofoectoderma e a massa celular interna demonstram que a biópsia de cinco células pode apresentar resultados variados;
- O número exato de células na biópsia nunca é conhecido, o que impossibilita determinar com precisão a porcentagem de DNA aneuploide;
- A biópsia da trofoectoderma danifica células individuais, causando vazamento de DNA e contaminação das células vizinhas, dificultando a medição precisa da aneuploidia;
- O limiar de 20% entre euploidia e mosaicismo é baseado apenas na sensibilidade atual do sequenciamento de nova geração (NGS), que não detecta níveis inferiores a 20% de DNA aneuploide. Consequentemente, qualquer mosaicismo abaixo de 20% é considerado euploide normal;
- Dentro da faixa de mosaicismo (20% a 80%), os desfechos de implantação e nascimento são semelhantes, indicando que o uso de limiares rígidos para prever resultados de FIV é incorreto.

Estudos indicam que embriões descartados como aneuploides pelo PGT-A resultaram em nascimentos normais (GLEICHER; PATRIZIO; BRIVANLOU, 2021), evidenciando a necessidade de revisar as diretrizes para evitar o desperdício de embriões viáveis. A Figura ?? ilustra a seleção de um pedaço do embrião e como isso pode influenciar a definição da ploidia do mesmo. A biópsia mencionada nas etapas do PGT-A, envolvendo a remoção de uma célula do embrião para análise genética Phillips et al. (2024), levanta a preocupação de que a remoção dessas células em crescimento possa comprometer o desenvolvimento do embrião, afetando os resultados neonatais, visto que as técnicas de micromanipulação utilizadas na biópsia não são totalmente isentas de riscos, como observado por Leaver e Wells (2019).

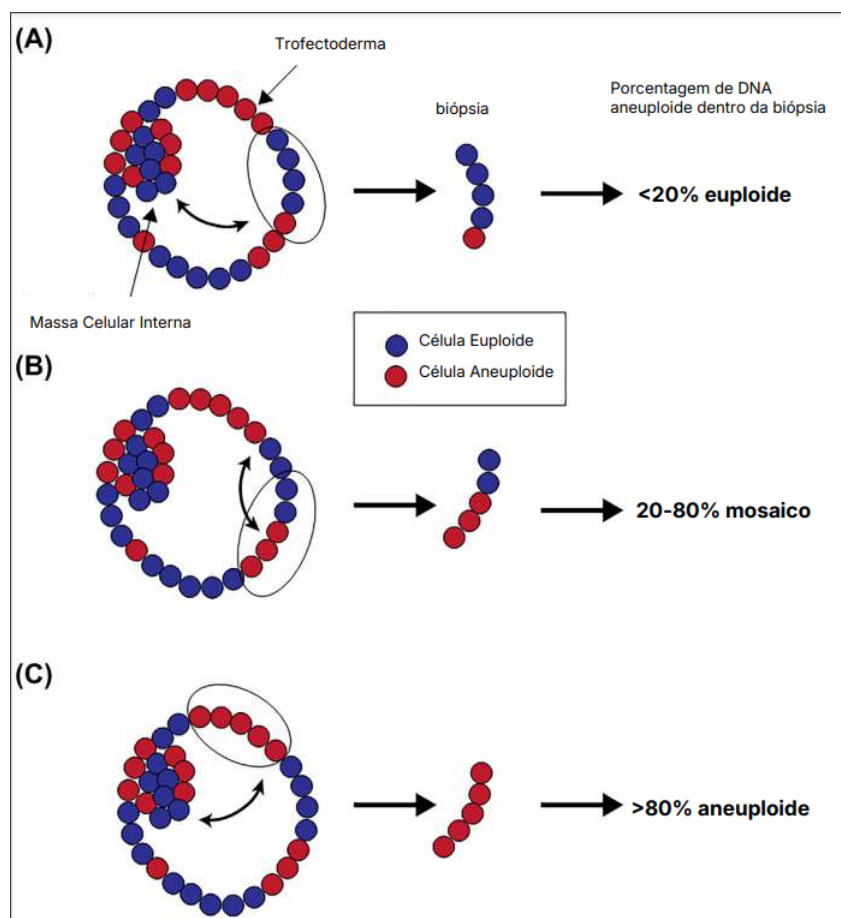


Figura 2 – Representação esquemática da biópsia do trofotoderma e a influência da amostragem na determinação da ploidia embrionária. (A) Embrião euploide com células da massa celular interna (MCI) e do trofotoderma (TE) com pouca ou nenhuma aneuploidia. A biópsia resulta em uma amostra com menos de 20% de células aneuploides, classificando o embrião como euploide. (B) Embrião mosaico com proporção similar de células euploides e aneuploides. A biópsia pode resultar em uma amostra com 20-80% de células aneuploides, classificando o embrião como mosaico. (C) Embrião predominantemente aneuploide. A biópsia resulta em uma amostra com mais de 80% de células aneuploides, classificando o embrião como aneuploide. Observação: A ploidia embrionária determinada pela biópsia do trofotoderma pode variar de acordo com a região amostrada, devido à mosaicidade embrionária. A figura ilustra a incerteza associada à classificação da ploidia embrionária com base em uma pequena amostra de células.

Embora o PGT-A permita detectar aneuploidias e aumente as chances de uma gravidez bem-sucedida, ele pode impactar negativamente o potencial de implantação do embrião (GLEICHER; PATRIZIO; BRIVANLOU, 2021). Por esses motivos, métodos não invasivos estão sendo estudados como alternativas eficientes e seguras, tornando-se cada vez mais relevantes. Um exemplo de técnica não invasiva é a análise morfocinética a partir de imagens obtidas de incubadoras de última geração equipadas com a tecnologia Time-Lapse System.

## 2.3 Time-Lapse System

O TLS é um sistema que captura imagens contínuas dos embriões em desenvolvimento, em intervalos regulares, sem alterar o ambiente de cultivo (MOUSTAKLI et al., 2024). Essa análise morfocinética, gerada pelas imagens adquiridas pelo TLS, permite o monitoramento quase contínuo do desenvolvimento do embrião, possibilitando a observação de eventos dinâmicos e frequentemente transitórios que não seriam visíveis em observações estáticas (BOUCRET et al., 2021). O uso do TLS não interrompe as condições de cultura, mantendo a viabilidade do embrião durante o processo de monitoramento (MOUSTAKLI et al., 2024).

As variáveis morfocinéticas incluem aspectos como a forma e a estrutura do embrião (morfológicas) e o movimento e o desenvolvimento do embrião ao longo do tempo (cinéticas), os quais são essenciais para uma análise detalhada de seu progresso (GLEICHER; PATRIZIO; BRIVANLOU, 2021). Com esse monitoramento contínuo, é possível observar a regularidade das divisões celulares e identificar momentos críticos do crescimento, o que pode auxiliar na diferenciação de embriões euplóides e aneuplóides com base no seu padrão de desenvolvimento (BOUCRET et al., 2021). De acordo com Moustakli et al. (2024), o TLS oferece insights valiosos sobre a saúde e o potencial de desenvolvimento dos embriões, utilizando uma abordagem não invasiva, em contraste com a biópsia de embriões. Alguns estudos indicam que, ao ser combinado com pontuações morfocinéticas, o TLS pode aumentar as taxas de implantação e gravidez clínica em comparação aos métodos tradicionais (BOUCRET et al., 2021).

As variáveis da planilha de dados, dados esses extraídos pelo TLS, que estamos utilizando para esse trabalho incluem informações sobre a qualidade morfológica e cinética dos embriões, como a taxa de divisão celular e a regularidade do desenvolvimento. Essas variáveis são essenciais para a análise morfocinética e a classificação dos embriões em diferentes estágios de desenvolvimento (BOUCRET et al., 2021). A seguir, discutiremos as variáveis cientificamente comprovadas que influenciam a ploidia.



### 2.3.1 Idade

As informações trazidas pelo Fertility and Ageing (Baird et al. (2005)) corroboram a relevância de incluir variáveis relacionadas à **idade materna**, pois estudos apontam que o aumento da aneuploidia em embriões está diretamente associado ao envelhecimento materno. O estudo de Yuan et al. (2023) aponta que a taxa de euploidia dos embriões está correlacionada com a idade feminina. À medida que a idade avança, há um declínio no número total e na qualidade dos ovócitos, um fator crítico para a fecundidade reduzida observada após os 35 anos (YUAN et al., 2023). Além disso, a resposta à estimulação ovariana e os níveis de FSH (hormônio folículo-estimulante) emergem como possíveis variáveis preditivas relevantes. Estudos indicam que a idade materna exerce maior influência sobre a qualidade embrionária do que os níveis de FSH isoladamente, reforçando que o impacto na fertilidade está mais relacionado à qualidade do oócito do que à sua quantidade (BAIRD et al., 2005).

### 2.3.2 t2, t3, t4, t5, t8, s2, cc2 (t3-t2), tSC, tSB, tB, cc3 (t5-t3), s3 (t8-t5), t5-t2, tSC-t8 e tB-tSB

Os **sistemas de Time-Lapse** ajudam a identificar **marcadores morfocinéticos**, que mostram como as células se dividem durante o desenvolvimento do embrião. Esses marcadores, junto com características físicas tradicionais, são fundamentais para selecionar o embrião mais adequado para a transferência (SOUZA, 2022a). O desenvolvimento embrionário é um processo dinâmico, com mudanças perceptíveis em um curto período (CRUZ et al., 2012). Estudos detalhados sobre o ritmo das divisões celulares, assim como características como tamanho e organização das células, demonstram que o tempo necessário para atingir certos estágios de desenvolvimento está diretamente relacionado ao potencial de implantação do embrião (SOUZA, 2022a).

Embriões que se dividem muito rapidamente apresentam menor chance de implantação quando comparados aos que seguem um ciclo celular dentro do intervalo considerado normal (CRUZ et al., 2012). Isso ocorre porque alterações no tempo de compactação inicial, na formação da blástula e na progressão até o estágio de blastocisto completo estão associadas a uma maior probabilidade de aneuploidias (CRUZ et al., 2012). O projeto *Timing of cell division in human cleavage-stage embryos is linked with blastocyst formation and quality* de Cruz et al. (2012) utilizou o sistema Time-Lapse para identificar marcadores morfocinéticos e monitorar com precisão os tempos das divisões celulares durante o desenvolvimento embrionário. Os tempos considerados ótimos para previsões de desenvolvimento embrionário foram: **t2 (24,3–27,9 horas)**, **t3 (35,4–40,3 horas)**, **t5 (48,8–56,6 horas)**, **s2 (<0,76 horas)** e **cc2 (<11,9 horas)**. A explicação detalhada dessas variáveis está disponível no **Apêndice A**.

Especificamente, no nível morfológico, **t5** destacou-se como o indicador mais relevante do potencial de implantação (CRUZ et al., 2012). Observa-se que a capacidade de diferenciar embriões viáveis daqueles não viáveis melhora significativamente quando os critérios se baseiam em eventos de divisão celular mais tardios (CRUZ et al., 2012). Embriões com **t5** entre 48,8 e 56,6 horas demonstram não apenas um maior potencial de implantação, mas também uma maior propensão a se desenvolverem em blastocistos de morfologia superior (CRUZ et al., 2012).

Ao criar um modelo de IA para a previsão de euploidia, é importante que considere as variáveis **t4**, **t8**, **tSC**, **tSB**, **tB**, **cc3** (**t5 - t3**) e **s3** (**t8 - t5**). Essas métricas são fundamentais para o desenvolvimento embrionário, conforme evidenciado no estudo que comprovou a efetividade da IA em detectar embriões viáveis com uma precisão de 70% (RIENZI et al., 2020). A incorporação dessas variáveis em um modelo de IA é crucial, pois possibilita a captura das sutilezas do desenvolvimento embrionário, que, de acordo com a pesquisa, são melhoradas através da análise automatizada fundamentada em Inteligência Artificial. Isso se torna especialmente relevante pois os sistemas de time-lapse disponibilizam informações detalhadas e contínuas, que podem ser combinadas para detectar padrões relacionados à euploidia (RIENZI et al., 2020).

A avaliação de fatores morfocinéticos, como o tempo necessário para clivagem e a extensão das fases subsequentes, tem sido alvo de pesquisa para antecipar a probabilidade de implantação embrionária. Por exemplo, um estudo publicado pelo Instituto Sapiientiae por Desai et al. (2019) destacou a importância desses parâmetros na predição do potencial de implantação embrionária. Apesar dessa pesquisa não tratar especificamente os intervalos **t5-t2**, **tSC-t8** e **tB-tSB**, ela sugere que a avaliação de intervalos de tempo entre eventos específicos no desenvolvimento embrionário pode oferecer percepções valiosas sobre a qualidade e a capacidade de desenvolvimento dos embriões (DESAI et al., 2019).

### 2.3.3 Estágio e Morfo

Modelos de IA têm sido empregados na avaliação de embriões produzidos até o quinto dia ou mais, com o objetivo de aprimorar a escolha com base em informações objetivas e de alta exatidão (LASSEN et al., 2022). Esses modelos priorizam a análise dos estágios "**Dia 5+**", ou seja, aqueles que atingem o estágio de blastocisto no quinto dia ou mais tarde, devido à maior disponibilidade de dados morfológicos e dinâmicos do desenvolvimento embrionário (LASSEN et al., 2022).

De acordo com os critérios definidos por Gardner (1999), com base na **morfologia do embrião** se tem a categorização dos blastocistos, um fator determinante para o potencial de implantação e a qualidade embrionária (CAPALBO et al., 2014). Os blastocistos são agrupados em quatro categorias principais, considerando tanto a massa celular

interna (ICM, Inner Cell Mass) quanto o TE (trophectoderma):

- **Grupo 1 (Excelente):** Blastocistos com classificação  $\geq 3AA$ . Blastocistos altamente desenvolvidos com massa celular interna densa e trophectoderma bem organizado (CAPALBO et al., 2014).
- **Grupo 2 (Bom):** Blastocistos com classificação 3, 4, 5 ou 6 e com notas AB ou BA. Apresentam características boas, mas menos consistentes em relação ao grupo excelente (CAPALBO et al., 2014).
- **Grupo 3 (Médio):** Blastocistos com classificação 3, 4, 5 ou 6 e notas BB, AC ou CA. Qualidade moderada com irregularidades tanto na ICM quanto no TE (CAPALBO et al., 2014).
- **Grupo 4 (Ruim):** Blastocistos com classificação  $\leq 3BB$ . Blastocistos de menor qualidade, com poucas células organizadas na ICM e TE menos coeso (CAPALBO et al., 2014).

O estudo de Capalbo et al. (2014) enfatizou a relação entre a morfologia padrão dos blastocistos, a euploidia e as taxas de implantação. Blastocistos de excelente morfologia, particularmente os biopsiados no dia 5, mostraram uma maior probabilidade de serem euploides e apresentaram taxas de implantação superiores.

#### 2.3.4 KIDScore<sup>TM</sup>

O KIDScore<sup>TM</sup>, um algoritmo baseado em IA aplicado à análise de imagens em sistemas Time-Lapse, tem se mostrado uma ferramenta importante na avaliação de embriões durante os tratamentos de reprodução assistida (KATO et al., 2021). O algoritmo combina variáveis morfocinéticas e parâmetros de desenvolvimento embrionário para fornecer uma pontuação que auxilia na seleção de embriões com maior potencial de implantação e viabilidade genética (GAZZO et al., 2020). A pontuação vai de 0 a 10. Pontuações baixas, entre 0 e 3, indicam embriões de qualidade inferior, com baixo potencial de implantação. Pontuações médias, de 4 a 6, correspondem a embriões de qualidade moderada, com um potencial razoável de implantação. Já pontuações altas, de 7 a 10, representam embriões de alta qualidade, com grande potencial de implantação (GAZZO et al., 2020). Kato et al. (2021) cita que o modelo apresentou uma alta precisão na previsão de resultados de gravidez, sendo especialmente útil tanto em pacientes com idade materna avançada quanto em pacientes mais jovens. Por fim, Gazzo et al. (2020) informaram que o uso do algoritmo no processo de seleção embrionária levou a um aumento expressivo nas taxas de implantação após a transferência de embriões congelados (FET). Então o modelo quando combinada com informações sobre os tempos de divisão celular, sendo **t2**, **t3**, **t5**, **s2** e **cc2** descritos por Cruz et al. (2012), possibilita um exame mais completo do crescimento

embrionário, melhorando a acurácia na seleção de embriões com maior probabilidade de êxito em tratamentos de FIV.

### 2.3.5 Ploidia

Para a elaboração do nosso modelo de previsão de euploidia, optamos por empregar a coluna de **Ploidia**, que proporciona uma categorização minuciosa dos embriões em diversas categorias de euploidia. As categorizações contidas nesta coluna são: Aneuploide complexo, Aneuploide/Triploide XXX, Caótico, Haploide, Mosaico de alto grau, Mosaico de baixo grau e Normal/Euploide. De acordo com o [Bastida et al. \(2019\)](#), considera-se os embriões com Caótico, Haplóide e Mosaico de alto grau como Aneuploides, enquanto os com mosaico de baixo grau como Euploides. Diante disso, optamos por reorganizar os valores na tabela de dados, agrupando as seguintes categorias sob o termo **Aneuploide**: Aneuploide complexo, Aneuploide/Triploide XXX, Caótico, Haploide e Mosaico de alta complexidade. Em contrapartida, os embriões categorizados como Mosaico de baixo grau e Normal/Euploide serão reunidos na categoria **Euploide**.

## 2.4 Aprendizado de Máquina

As dificuldades em definir IA não são, portanto, o resultado de alguma deficiência ou descuido, mas surgem do fato de que fomos incapazes de determinar precisamente qual inteligência desejaríamos replicar artificialmente ([SHEIKH; PRINS, 2023](#)). Dessa forma, definimos a Inteligência Artificial como sistemas que exibem comportamento inteligente ao analisar seu ambiente e tomar ações – com algum grau de autonomia – para atingir objetivos específicos ([SHEIKH; PRINS, 2023](#)).

Na medicina reprodutiva, a IA tem se mostrado promissora na melhoria de processos como a fertilização in vitro (FIV). Essa busca por imitar a inteligência humana e entender seus processos cognitivos levou ao desenvolvimento de diversas abordagens, entre elas o Aprendizado de Máquina (Machine Learning, ML), que envolve a capacidade de computadores de interpretar grandes volumes de dados, construir modelos baseados nesses dados e, assim, gerar hipóteses ou previsões sobre o mundo ao seu redor ([RUSSELL; NORVIG, 2016](#)). Na medicina, algoritmos podem ser treinados para reconhecer padrões genéticos em embriões, classificar o melhor embrião para implantação e prever características genéticas de novos embriões.

Os métodos de ML são geralmente classificados em três tipos principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Neste estudo, opta-se pelo aprendizado supervisionado como abordagem principal, dada sua eficácia na análise de dados rotulados, permitindo decisões mais precisas e embasadas para otimizar os tratamentos de FIV.

O aprendizado supervisionado consiste no treinamento de algoritmos com base em conjuntos de dados rotulados, nos quais as variáveis de entrada (inputs) e os resultados esperados (outputs) já são conhecidos. O algoritmo aprende a correlacioná-los de forma eficiente (RUSSELL; NORVIG, 2016), ajustando seus parâmetros com base nas diferenças entre previsões e resultados reais (TRASK, 2019). Uma técnica amplamente usada dentro do aprendizado supervisionado é a classificação, cujo objetivo é atribuir rótulos ou classes pré-definidas aos dados. Por exemplo, no contexto da medicina reprodutiva, um modelo pode ser treinado para diferenciar embriões euploides e aneuploides, aprendendo a reconhecer padrões associados a cada grupo (IZBICKI; SANTOS, 2020).



Figura 3 – Os dados já possuem seus próprios rótulos. Já é definido o que significa ser euploide e aneuploide. Os dados rotulados (ou seja, com rótulos já atribuídos) são usados para treinar o modelo. Nesse contexto, ao se referir a "euploide" e "aneuploide", esses rótulos podem representar classes que o modelo deve aprender a identificar com base nas características dos dados, ajudando a prever ou classificar novos casos com precisão.

De maneira geral, um algoritmo de aprendizado supervisionado separa o banco de dados em três subconjuntos: treinamento, validação e teste (IZBICKI; SANTOS, 2020). Na fase de treinamento, o algoritmo identifica padrões nos dados de entrada e os associa às classes desejadas (IZBICKI; SANTOS, 2020). Na validação, um subconjunto de dados não utilizado no treinamento avalia o desempenho do modelo (IZBICKI; SANTOS, 2020), permitindo ajustes nos hiperparâmetros. Após resultados satisfatórios, o conjunto de testes mensura métricas como acurácia, recall e precisão, garantindo o desempenho esperado (IZBICKI; SANTOS, 2020).

A seleção aleatória das amostras para treinamento, validação e teste é uma boa prática (IZBICKI; SANTOS, 2020), evitando problemas decorrentes de ordenações previamente estabelecidas nos bancos de dados. Isso assegura uma visão representativa e

imparcial dos dados, fundamental para a construção de modelos robustos e confiáveis (IZBICKI; SANTOS, 2020).

Para a exploração inicial dos dados, utilizaremos o modelo k-Nearest Neighbors (kNN) como ponto de partida, conforme apresentado no livro *Aprendizado de Máquina: Uma Abordagem Estatística* Izbicki e Santos (2020). Caso o desempenho e a confiabilidade do kNN não sejam satisfatórios, passaremos a avaliar outros modelos de classificação, como Regressão Linear e Naive Bayes, seguindo esse processo iterativo até encontrarmos um modelo que atenda aos critérios desejados. Se alguns desses modelos demonstrarem resultados satisfatórios, não será necessário explorar os demais.

### 2.4.1 O Algoritmo K-Nearest Neighbor

O algoritmo K-Nearest Neighbor (KNN) é, em essência, um dos algoritmos mais simples e populares em aprendizado supervisionado, sendo utilizado principalmente em tarefas de classificação e regressão (ZHANG, 2016). Sua principal característica é classificar uma nova observação com base nas classes de seus k vizinhos mais próximos previamente rotulados (ZHANG, 2016). O KNN pode ser considerado um modelo não paramétrico, pois não faz suposições a respeito da distribuição dos dados, o que o torna bastante versátil para uma diversidade de problemas (ZHANG, 2016). É também um algoritmo de aprendizado "preguiçoso" já que não ocorre um processo explícito de aprendizado (ZHANG, 2016). Então, o algoritmo armazena os dados de treinamento e, na hora de fazer uma predição usa a distância da nova observação para cada exemplo discretamente do conjunto de treinamento para descobrir qual classe ou valor alvo deve usar (ZHANG, 2016).

O desempenho do KNN é dependente da escolha de k, que são o número de vizinhos que são considerados na classificação (ZHANG, 2016). Valores muito pequenos de k podem levar a overfitting, já que o modelo se torna sensível aos ruídos contidos nos dados. Valores muito grandes, por sua vez, podem resultar em underfitting pela perda de padrões locais importantes (ELKAN, 2011). Nesse trabalho, utilizaremos num primeiro momento  $k = 3$  e  $k = 5$  para avaliar o desempenho do modelo. Estes valores são adequados uma vez que eles capturam padrão locais sem exagerada sensibilidade ao ruído. Caso necessário, faremos o ajuste no parâmetro k com base nos resultados do conjunto de validação, pois o conjunto de validação é empregado especificamente para otimizar hiperparâmetros.

A métrica de distância adotada foi a distância Euclidiana, uma das mais simples e mais utilizadas. Essa é uma escolha adequada para pequenos conjuntos de dados e de baixa dimensionalidade, como o utilizado neste estudo. Para tais condições a distância euclidiana capta similaridades com eficiência, além de facilitar a interpretação dos resultados (ELKAN, 2011).

O funcionamento do KNN pode ser resumido em três etapas principais de acordo com o [Elkan \(2011\)](#):

- Armazenar os exemplos de treinamento com suas respectivas etiquetas.
- Calcular a distância entre uma nova observação e todos os exemplos armazenados, utilizando uma métrica de similaridade, como a distância Euclidiana.
- Selecionar os  $k$  vizinhos mais próximos e determinar a classe mais frequente entre eles, atribuindo-a à nova observação.

Essa abordagem de classificação por maioria contribui para mitigar o impacto de ruídos ou valores extremos ([ELKAN, 2011](#)). Uma prática comum para definir o valor de  $k$  é utilizar a raiz quadrada do número total de observações no conjunto de treinamento, embora ajustes específicos sejam feitos dependendo do problema e do conjunto de dados ([ELKAN, 2011](#)).

### 2.4.2 Regressão Linear

A regressão linear é uma técnica estatística utilizada em aprendizado de máquina para expressar relações entre variáveis e para realizar previsões baseadas em dados passados ([RODRIGUES, 2012](#)). É uma abordagem essencial em aprendizado supervisionado, o que visa detectar padrões e criar modelos generalizáveis para dados até então inexistentes ([SOTO, 2021](#)). Segundo [Santos \(2007\)](#), é usada para quantificar a associação entre as variáveis, usando frequentemente o coeficiente de correlação de Pearson para avaliar a força e a direção dessa relação. A regressão linear pode ser dividida em dois tipos principais: regressão linear simples e regressão linear múltipla ([SOTO, 2021](#)).

Em um contexto de IA, a regressão linear simples estabelece uma relação matemática entre uma variável dependente  $Y$  e uma única variável independente  $X$ , descrita pela equação:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

onde  $\beta_0$  representa o intercepto,  $\beta_1$  é o coeficiente angular que expressa a taxa de variação de  $Y$  em relação a  $X$ , e  $\epsilon$  é o termo de erro aleatório ([RODRIGUES, 2012](#)). Modelos mais complexos, como a regressão linear múltipla, permitem a inclusão de mais de uma variável independente, aumentando assim a capacidade preditiva e explicativa do modelo ([RODRIGUES, 2012](#)).

A regressão linear é um método paramétrico que requer a suposição de uma relação linear entre as variáveis, o que pode torná-la restritiva quando os padrões não são lineares. Entretanto, técnicas de pré-processamento, como transformações de variáveis e inclusão de termos polinomiais, podem contornar essa limitação e aumentar a capacidade de generalizar do modelo ([MONTGOMERY; RUNGER, 2009](#)).



### 2.4.3 Naive Bayes

O Naive Bayes é um algoritmo de classificação que se baseia no Teorema de Bayes para estimar a classe mais provável de uma instância, utilizando a probabilidade condicional dos atributos observados (RISH et al., 2001). Ele é chamado de "naive" (ou ingênuo) porque parte da suposição de que os atributos são condicionalmente independentes dado a classe. Em outras palavras, presume que a presença ou ausência de um atributo não afeta os outros atributos. Embora essa suposição não seja totalmente realista em muitas situações do mundo real, o algoritmo ainda apresenta resultados impressionantes em diversas aplicações práticas (RISH et al., 2001).

De acordo com Zhang (2004), o Naive Bayes determina a probabilidade de uma instância pertencer a uma classe específica analisando a distribuição condicional dos atributos. Mesmo com a simplificação imposta pela suposição de independência, o algoritmo muitas vezes alcança resultados comparáveis a métodos mais avançados. O Naive Bayes é particularmente eficaz em problemas de classificação de texto, como análise de sentimentos e classificação de documentos, e em tarefas de diagnóstico médico, como detecção de doenças com base em sintomas ou resultados de exames (RISH et al., 2001).

Segundo Zhang (2004), A ideia central do Naive Bayes é calcular a probabilidade de uma instância  $X = \{x_1, x_2, \dots, x_n\}$  pertencer a uma classe  $C_k$  utilizando a fórmula:

$$P(C_k|X) = \frac{P(X|C_k) \cdot P(C_k)}{P(X)}$$

Como o denominador  $P(X)$  é constante para todas as classes, o foco do cálculo está em maximizar o numerador, que é proporcional a  $P(C_k) \cdot \prod_{i=1}^n P(x_i|C_k)$ , onde  $P(x_i|C_k)$  representa a probabilidade condicional de cada atributo  $x_i$ , dado a classe  $C_k$ . O algoritmo calcula essas probabilidades para cada classe e atribui a instância à classe com a maior probabilidade (ZHANG, 2004).

Em resumo, o Naive Bayes é uma solução simples, rápida e incrivelmente eficaz para muitos problemas de classificação, mesmo quando a suposição de independência não é completamente verdadeira. Essa combinação de praticidade e desempenho o torna uma escolha popular em áreas como processamento de linguagem natural e diagnóstico médico.

## 2.5 Identificação de Padrões Morfocinéticos e Predição de Euploidia com IA e Trabalhos Correlatos

Os dados que são coletados pela tecnologia do TLS, são chamados de “dados morfocinéticos”, que são definidos como dados do desenvolvimento dos embriões (MOUSTAKLI et al., 2024). Essa informação reunida proporciona noções detalhadas sobre o



padrão do desenvolvimento e divisão celular embrionário. Atualmente, após recorrentes estudos sobre esses dados, sabe-se que as características morfocinéticas dos embriões têm sido associadas à avaliação de sua potência de desenvolvimento, ou seja, se um embrião analisado pelo TLS tenha um melhor desenvolvimento, ele terá mais probabilidade de ser euploide, pois um bom desempenho de um embrião é capaz de prever a implantação (YUAN et al., 2023).

Os modelos de TLS, de acordo com Yuan et al. (2023), tem uma avaliação contínua na etapa do desenvolvimento embrionário por meio de suas imagens e, por observações estáticas, monitora as características do embrião, como tempo e padrões de divisão celular, fornecendo uma base para prever a euploidia. O TLS por si só, não opera com a IA, mas é frequentemente mesclado com essa tecnologia para maiores análises. Um exemplo é o estudo do Yuan et al. (2023), o artigo “Development of an artificial intelligence based model for predicting the euploidy of blastocysts in PGT-A treatments” o qual teve como objetivo utilizar o TLS e desenvolver um modelo de IA usando uma técnica de regressão logística, para predizer a euploidia de blastocistos—fase do desenvolvimento embrionário que ocorre após a clivagem do óvulo fertilizado—em tratamentos de PGT-A, ajudando a identificar embriões com maiores possibilidades de serem geneticamente normais antes da etapa de transferência. O modelo foi avaliado com uma boa precisão, indicando que ele consegue distinguir entre embriões euploides e aneuploides.

Outro estudo é o de Souza (2022a), “Análise da ploidia de embriões humanos por meio da inteligência artificial com o uso de variáveis de morfologia, morfocinética e variáveis relacionadas com a paciente”, que também utiliza IA, mas combinando dados morfológicos, morfocinéticos e clínicos para prever a ploidia dos embriões, utilizando uma rede neural artificial para classificá-los como euploides ou aneuploides. Em contraste, o modelo proposto neste projeto busca prever a porcentagem de aneuploidia, oferecendo uma análise quantitativa detalhada, em vez de uma classificação binária. Essa abordagem visa proporcionar uma compreensão mais profunda da saúde genética dos embriões, permitindo decisões mais precisas durante a seleção.

Divergente do trabalho de Yuan et al. (2023) e de Souza (2022a), que focam na predição binária de euploidia (ou seja, identificar se um embrião é euploide ou aneuploide), o modelo proposto neste projeto visa prever a porcentagem de aneuploidia dos embriões. Essa abordagem oferece um indicador quantitativo em vez de uma simples classificação binária, permitindo uma análise mais detalhada e informativa sobre a saúde genética dos embriões. Com isso, embriologistas poderão avaliar não apenas se um embrião é geneticamente normal, mas também entender o grau de aneuploidia presente, possibilitando decisões mais precisas durante o processo de seleção.

Além disso, nosso modelo busca ser uma alternativa menos invasiva e mais acessível. Embora técnicas atuais, como o PGT-A (Testagem Genética Pré-implantacional),

forneçam informações precisas sobre a euploidia, elas dependem de métodos invasivos, como a biópsia embrionária, e de infraestrutura avançada, o que pode limitar o acesso a essas tecnologias em algumas clínicas. Com o uso de IA e dados de morfocinética obtidos via TLS (Time-Lapse System), o objetivo é desenvolver uma solução que permita uma avaliação robusta sem a necessidade de procedimentos invasivos, ampliando o alcance e a aplicação da tecnologia.

Um exemplo de modelo amplamente utilizado na prática clínica é o *KIDScore<sup>TM</sup> Day 5*, que classifica embriões com base no seu potencial de implantação, sendo frequentemente usado em dispositivos como o *TM* (REIGNIER et al., 2019). Esse modelo utiliza grandes bancos de dados multicêntricos para atribuir uma pontuação ao embrião, ajudando a ranquear os embriões do mesmo ciclo, facilitando a escolha do embrião com maior potencial para transferência (REIGNIER et al., 2019).

Apesar dos avanços e do sucesso clínico do *KIDScore<sup>TM</sup>*, algumas limitações importantes foram destacadas na literatura. Segundo Reignier et al. (2019), o desempenho desses modelos pode ser influenciado por variáveis como idade materna, número de oócitos, origem dos oócitos e características específicas de cada centro de FIV, como as condições de embrionária e uso de oxigênio reduzido. Isso destaca a necessidade de criar modelos robustos fundamentados em grandes conjuntos de dados multicêntricos e amplamente representativos, a fim de assegurar maior generalização e confiabilidade nos resultados.

Dentro desse cenário, o sistema *CHLOE<sup>TM</sup>* (Cultivating Human Life through Optimal Embryos), desenvolvido pela *Fairtility<sup>TM</sup>*, se destaca como uma ferramenta inovadora no campo da FIV por usar IA e visão computacional para prever com precisão a implantação e o desenvolvimento dos blastocistos, oferecendo uma análise transparente das decisões da IA (FAIRILITY<sup>TM</sup>, 2020). Enquanto nosso modelo é mais quantitativo e menos invasivo, o CHLO oferece uma solução mais integrada e transparente, visando otimizar a seleção de embriões de formas complementares, com foco na acurácia e confiança no processo de FIV (FAIRILITY<sup>TM</sup>, 2020).

Com essas diferenças em mente, nosso objetivo é criar uma alternativa menos invasiva e mais acessível para a seleção de embriões na FIV, utilizando um modelo preditivo capaz de estimar a porcentagem de euploidia. Essa abordagem, ao contrário dos modelos binários, oferece uma análise quantitativa detalhada da qualidade genética dos embriões, o que pode contribuir significativamente para a melhoria das taxas de sucesso nos tratamentos de fertilização in vitro.

## 3 Metodologia

### 3.1 Classificação da Pesquisa

#### 3.1.1 Natureza

Em relação à natureza desta pesquisa, trata-se de uma pesquisa aplicada. Temos como objetivo principal gerar um conhecimento que possa ter um impacto direto e uma utilidade prática, ambos em contextos reais ([NASCIMENTO, 2016](#)).

Com esta pesquisa, a aplicação do modelo de IA tem a potencialidade de melhorar os sucessos dos tratamentos de FIV, fazendo o processo de seleção embrionária mais eficaz, menos invasivo e mais acessível a um maior número de pessoas.

#### 3.1.2 Método ou Abordagem Metodológica

A metodologia ou abordagem metodológica dessa pesquisa é quantitativa ([NASCIMENTO, 2016](#)). Nosso foco é a análise dos dados numéricos referentes aos padrões morfocinéticos de embriões. Esses dados serão utilizados para o desenvolvimento da IA, que será capaz de prever a porcentagem de euploidia, auxiliando na seleção de embriões com maior possibilidade de saúde genética.

Escolher a abordagem quantitativa nos ajudará a atingir os objetivos desta pesquisa, permitindo explorar e validar os dados com precisão, oferecendo resultados objetivos.

#### 3.1.3 Objetivos

Quanto aos objetivos, o objetivo desta pesquisa é exploratório ([NASCIMENTO, 2016](#)). Este trabalho procura identificar e investigar padrões em dados morfocinéticos de embriões, usando o TLS, explorando a possibilidade de realizar essa predição juntamente com as tecnologias de IA.

Ao focar na concepção de um modelo que terá a capacidade de identificar padrões nos dados, exploraremos a relação entre esses dados e a importância de cada padrão para o resultado desejado, compreendendo os fatores que influenciam a saúde genética dos embriões, mas sem um conhecimento prévio estabelecido que explique completamente essas relações.

### 3.1.4 Procedimentos De Pesquisa

O procedimento adotado neste trabalho é experimental. Definimos este procedimento por causa do objetivo de investigar as relações entre as variáveis, o que é uma característica da pesquisa experimental (NASCIMENTO, 2016). Buscamos estabelecer uma relação de causa e efeito entre as características morfocinéticas dos embriões, mais especificamente a porcentagem de euploidia.

## 3.2 Design da Pesquisa

Esse estudo adotará o uso de IA para realizar a análise dos dados morfogenéticos dos embriões, desenvolvendo um modelo de predição baseado em machine learning. O modelo será treinado para identificar os padrões nos dados coletados pelo TLS, com foco em prever a porcentagem de euploidia, o que indica a saúde genética dos embriões.

Para desenvolver e testar o modelo, utilizaremos a linguagem de programação Python, aproveitando as bibliotecas disponíveis para a construção da IA. Quanto à validação do modelo, será elaborada uma fase experimental, na qual o modelo será testado com dados reais de embriões já classificados, a fim de compararmos e testarmos seu desempenho, refinando-o quando necessário.

Nesta pesquisa, buscaremos identificar e mapear os padrões em um campo que ainda está em desenvolvimento. A prática será testada em um ambiente controlado com dados obtidos pelo TLS, avaliando a efetividade do modelo com base na sua capacidade de prever, em porcentagem, a ploidia do embrião.

### 3.2.1 Fases de Trabalho

As fases do nosso trabalho se dividem em duas etapas: **Fase 1: Análise e Preparação de Dados** (Tabela 1), com o objetivo de compreender e organizar os dados para realizar a análise da influência dos parâmetros na porcentagem de euploidia, e a **Fase 2: Desenvolvimento e Avaliação do Modelo** detalhada na Tabela (Tabela 2), que foca no desenvolvimento, ajuste e avaliação de um modelo de ML para efetuar a predição de euploidia, finalizando com a entrega de um protótipo de uma interface a ser evoluída em trabalhos futuros, realizando a criação e junção dos dois.

Nas seções a seguir, demonstraremos os objetivos de cada fase, mostrando suas atividades, nas quais estão descritos resumidamente o que será feito, qual método será utilizado e o resultado esperado. Em seguida, detalharemos cada parte.

Tabela 1. Fase 1: Análise e Preparação de Dados

Fase 1: Análise e Preparação de Dados			
Objetivos Específicos	Atividades	Método de Pesquisa	Resultados Esperados
<b>OE1</b> Expansão, Processamento e Análise de	<b>Atividade 1 (A1)</b> Análise, Revisão e Seleção de Variáveis para Predição de Euploidia	- Pesquisa bibliográfica  - Python - Biblioteca Pandas	- Analisar, Revisar e Selecionar as variáveis - Limpeza dos dados
	<b>Atividade 2 (A2)</b> Normalização dos Dados para Otimização	- Z-Score	- Normalização das variáveis numéricas
	<b>Atividade 3 (A3)</b> Identificação da Correlação e Atribuição de Pesos aos Parâmetros na Previsão da Ploidia do Embrião	- Coeficiente de correlação de Spearman	- Análise das correlações entre variáveis pelo Gráfico de dispersão
	<b>Atividade 4 (A4)</b> Divisão de Dados e Aplicação de Data Augmentation	- Divisão do conjunto de dados - Data augmentation com o Algoritmo de Monte Carlo	- Dados para treinamento, validação e teste - Aumento do conjunto de dados para treinamento

### 3.2.1.1 Objetivo Específico 1 - Identificação de Parâmetros em Embriões

#### 3.2.1.1.1 Atividade 1 (A1): Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia

Começaremos com a verificação da pertinência das variáveis já existentes na planilha de dados dos embriões e com a avaliação da introdução de outras variáveis que possam aprimorar a exatidão da análise, ou até mesmo com a eliminação de variáveis, se necessário. Em seguida, faremos a limpeza dos dados, substituindo valores nulos por valores mais apropriados em alguns casos, como nas colunas que realizam cálculos entre outras colunas, onde será possível identificar valores e substituí-los.

Para realizar a verificação da pertinência das variáveis já existentes e avaliar a viabilidade de introduzir novas variáveis, utilizaremos a pesquisa bibliográfica, nos norteando pelos estudos apresentados no capítulo 2: o de [Yuan et al. \(2023\)](#), o artigo 'Development of an artificial intelligence-based model for predicting the euploidy of blastocysts in PGT-A treatments' e o de [Souza \(2022a\)](#), 'Análise da ploidia de embriões humanos por meio da inteligência artificial com o uso de variáveis de morfologia, morfocinética e variáveis rela-

cionadas com a paciente’. Ambos os artigos descrevem o uso de IA para fazer a predição da ploidia de embriões, o que se assemelha com o que queremos propor, com a diferença de que temos o objetivo final de prever a porcentagem de aneuploidia.

Dessa forma, analisaremos os estudos feitos por ambos pesquisadores e utilizaremos para entender o poder que cada variável tem para o objetivo final e se é necessário adicionar outras variáveis que não existem na nossa planilha, ou até mesmo excluí-las. Além disso, conduziremos entrevistas com o especialista encarregado de fornecer os dados, para uma análise prática da pertinência das variáveis disponíveis e para debater possíveis variáveis extras que possam ser relevantes para a análise. As entrevistas possibilitarão alinhar a seleção das variáveis ao conhecimento clínico e experiência prática do profissional, garantindo que as variáveis selecionadas sejam aplicáveis no cenário real de previsão de euploidia. Em relação a limpeza dos dados, utilizaremos a linguagem *Python* que possui a biblioteca *Pandas*, que permite o carregamento de planilhas do Excel, do formato .xlsx, como a que possuímos. De acordo com [Chen \(2018\)](#): “O Pandas é uma biblioteca Python de código aberto para análise de dados. Ele dá a Python a capacidade de trabalhar com dados do tipo planilha, permitindo carregar, manipular, alinhar e combinar dados rapidamente, entre outras funções.” Usaremos as funções *isnull()* e *info()*, da biblioteca *Pandas*.

Inicialmente, analisar as variáveis possibilitando a detecção de variáveis irrelevantes ou redundantes, aprimorando a correlação entre os parâmetros estudados e a porcentagem de euploidia. Também a detecção e correção de inconsistências, como valores em branco ou discrepâncias, garantindo que o conjunto de dados esteja organizado e pronto para futuras análises, prevenindo distorções nos resultados.

#### 3.2.1.1.2 **Atividade 2 (A2):** Normalização dos Dados para Otimização

As variáveis numéricas são normalizadas, por meio de uma técnica de normalização, o Z-Score, assegurando os intervalos de valores de cada coluna. A normalização é de grande importância que façamos a normalização dos dados, visto que, de acordo com [Milewski et al. \(2016\)](#) estudos anteriores demonstraram que a incorporação de dados morfológicos normalizados para avaliação da qualidade do embrião aumenta consideravelmente o poder preditivo dos modelos criados. A normalização é uma forma de dimensionar recursos, transformando o intervalo deles em uma escala padrão ([JAISWAL, 2024](#)). Também é importante citar que dados normalizados também são fáceis de interpretar e, portanto, mais fáceis de entender. Quando todos os recursos de um conjunto de dados estão na mesma escala, também se torna mais fácil identificar e visualizar as relações entre diferentes recursos e fazer comparações significativas. ([JAISWAL, 2024](#)). Utilizaremos o método Z-Score, detalhado no APÊNDICE B.

Por fim, a normalização das variáveis assegurará que todas estejam dentro do mesmo intervalo, eliminando vieses numéricos e aprimorando a exatidão dos algoritmos de aprendizado de máquina que serão implementados posteriormente.

#### 3.2.1.1.3 **Atividade 3 (A3):** Identificação da Correlação e Atribuição de Pesos aos Parâmetros na Previsão da Ploidia do Embrião

O objetivo desta atividade é identificar as relações entre os diferentes parâmetros presentes na planilha de dados dos embriões, avaliando a intensidade e o sentido dessas relações, com foco em sua influência na porcentagem de euploidia. Após a pesquisa bibliográfica, utilizaremos o coeficiente de correlação de Spearman, que mede a relação monótona entre duas variáveis, considerando as ordens atribuídas às observações em vez dos valores originais (SOUSA, 2019). A correlação será calculada para todas as combinações possíveis de variáveis, possibilitando uma análise mais detalhada de suas interações (SOUSA, 2019). Um gráfico de dispersão será gerado para complementar a análise visual, facilitando a identificação de padrões. Também com o conhecimento adquirido pela A1, determinaremos a relevância relativa de cada parâmetro na previsão da ploidia do embrião, atribuindo pesos que reflitam sua influência, identificando a relevância de cada parâmetro.

Para realizar a análise de correlação entre os parâmetros será utilizado o coeficiente de Spearman, um método estatístico amplamente empregado para avaliar a intensidade e o sentido da relação monótona entre duas variáveis. Inicialmente, as variáveis do conjunto de dados serão classificadas em ordem crescente, atribuindo-lhes ranks que serão usados para o cálculo do coeficiente. Essa abordagem permite capturar relações tanto lineares quanto não lineares entre as variáveis (SOUSA, 2019), explicado com mais detalhes no APÊNDICE D.

A fórmula do coeficiente de correlação de Spearman será aplicada, utilizando as ordens atribuídas, assegurando que o método se adapte a diferentes formatos de relação, como curvas monótonas crescentes ou decrescentes. Além disso, será realizada uma análise complementar com gráficos de dispersão, que ajudarão a identificar a inclinação dos dados e o sentido da correlação, sendo positiva (próximas ao valor 1) quando as variáveis variam no mesmo sentido e negativa (próximas ao valor -1) quando variam em sentidos opostos (SOUSA, 2019). O resultado numérico do coeficiente será avaliado em relação à sua magnitude, indicando se a correlação é forte, moderada ou fraca, e seu sinal indicará o tipo de relação (positiva ou negativa).

Usaremos a biblioteca Pandas para manipulação dos dados e a SciPy para calcular o coeficiente de Spearman. A metodologia para a definição e atribuição de pesos específicos aos parâmetros relevantes para a ploidia do embrião combina análise teórica e



prática. Inicialmente, será realizada uma pesquisa bibliográfica em publicações científicas e revisões sistemáticas que explorem a influência dos parâmetros na ploidia embrionária.

Os cálculos realizados com o coeficiente de Spearman permitirão identificar quais parâmetros possuem uma relação mais forte (positiva ou negativa) com a porcentagem de euploidia, além de explorar como esses parâmetros se relacionam entre si. Os gráficos gerados deverão destacar as variáveis mais influentes, os padrões visuais que reforcem as relações monótonas, áreas onde a ausência de correlação linear não exclua outras formas de relação. Também resultará em uma lista detalhada dos parâmetros identificados, acompanhada das respectivas justificativas para os pesos atribuídos a cada um, com base em sua influência na ploidia do embrião. Em vez de valores numéricos, os resultados trarão explicações fundamentadas cientificamente, apresentando as razões pelas quais cada parâmetro exerce determinada influência. O resultado final fornecerá uma visão qualitativa e consistente dos fatores que impactam a ploidia, sendo uma base essencial para análises e decisões futuras no estudo.

#### 3.2.1.1.4 **Atividade 4 (A4):** Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados

A separação do conjunto de dados será feita para garantir que as etapas de treinamento, validação e testes sejam feitas de forma organizada. Iremos dividir em 3 conjuntos: Treinamento, para ensinar o modelo; Validação, para ajustar os parâmetros de forma adequada e evitar o overfitting, ocorre quando um algoritmo reduz o erro por meio da memorização de exemplos de treinamento em vez de aprender a verdadeira relação geral entre os dados ([BASHIR et al., 2020](#)); e Teste, para avaliar o desempenho final. A divisão será produzida com uma distribuição de 70% dos dados para treinamento, 15% para validação e 15% para teste. O nosso conjunto de dados original é relativamente pequeno, tendo 84 linhas, assim, aplicaremos a técnica de aumento de dados (data augmentation) usando o algoritmo Monte Carlo (APÊNDICE C), que gera novas amostras a partir de dados existentes. Utilizaremos a técnica exclusivamente nos dados de treinamento para evitar interferências ruins na criação e desempenho do modelo de aprendizado de máquina, preservando as características e padrões já existentes na tabela, contribuindo para a capacidade do modelo ([KIAR et al., 2021](#)).

Para a realização, usaremos uma abordagem padrão em aprendizado de máquina, dividindo o conjunto de dados em três, como citado. A divisão é feita para garantir que cada conjunto seja representativo e que o modelo seja avaliado imparcialmente ([BASHIR et al., 2020](#)). Para aumentar o conjunto de dados de treinamento, será aplicada uma técnica de aumento de dados (data augmentation) com Monte Carlo, que é a geração



de dados artificiais de alta qualidade por meio da manipulação de amostras de dados existentes (WANG et al., 2024).

O resultado esperado é a divisão do conjunto de dados em três subconjuntos: treinamento, validação e teste e a ampliação do conjunto de treinamento com o uso de data augmentation, aplicando o algoritmo de Monte Carlo.

Tabela 2. Fase 2: Desenvolvimento e Avaliação do Modelo

Fase 2: Desenvolvimento e Avaliação do Modelo			
Objetivos Es-pecíficos	Atividades	Método de Pesquisa	Resultados Esperados
OE2 Treinamento e Ajuste de Modelo de Machine Learning para Predição de Euploidia	<b>Atividade 5 (A5)</b> Desenvolvimento e Treinamento do Modelo de Machine Learning para Otimização da Predição de Euploidia, Incluindo Treinamento, Validação e Teste	Python (Bibliotecas scikit-learn)	- Treinamento do modelo bem-sucedido com no mínimo 70% de precisão, usando KNN, regressão linear e naive bayes
	<b>Atividade 6 (A6)</b> Utilizar métricas adequadas para medir o desempenho do modelo	Python (Biblioteca scikit-learn e pandas)	- Acurácia - Precisão - Recall - F1-Score
OE3 Avaliação do Modelo	<b>Atividade 7 (A7)</b> Avaliação do Desempenho do Modelo na Predição por Meio da Matriz de Confusão e Curva ROC	Matriz de confusão (Random Forest - scikit-learn)	- Verdadeiros positivos (TP) - Verdadeiros negativos (TN) - Falsos positivos (FP) - Falsos negativos (FN) - Gráfico exibindo a relação entre a sensibilidade e a especificidade para diferentes valores de limiar.
	<b>Atividade 8 (A8)</b> Prototipar uma interface	Interface básica desenvolvida no FIGMA	- Interface básica - Coleta de opiniões

### 3.2.1.2 **Objetivo Específico 2** - Treinamento e Ajuste de Modelo de Machine Learning para Predição de Euploidia

#### 3.2.1.2.1 **Atividade 5 (A5):** Desenvolvimento e Treinamento do Modelo de Machine Learning para Otimização da Predição de Euploidia, Incluindo Treinamento, Validação e Teste

O desenvolvimento do modelo de Machine Learning para a predição de euploidia começará com a implementação do algoritmo k-Nearest Neighbors (KNN), com o objetivo de classificar os embriões como euploides ou aneuploides. O KNN foi escolhido inicialmente devido à sua simplicidade e eficácia em problemas de classificação, especialmente quando se tem dados com distribuições bem definidas. Caso o modelo de KNN não forneça os resultados esperados, a regressão linear e o Naive Bayes serão testados como alternativas. O objetivo principal é alcançar uma precisão de 70%, e se esse objetivo não for atingido com o KNN, esses modelos alternativos serão ajustados até que o objetivo seja atingido.

O processo de desenvolvimento será realizado em várias etapas. Na Atividade 3, o pré-processamento dos dados já terá sido realizado, incluindo a limpeza e normalização dos dados, sendo a normalização especialmente importante para o KNN, pois este modelo é sensível à escala dos dados (ZHANG, 2016). Com os dados prontos, passaremos para a fase de treinamento do modelo.

Inicialmente, utilizaremos o *KNeighborsClassifier* do *Scikit-learn*. O primeiro passo será definir o número de vizinhos (`n_neighbors`), um dos parâmetros mais importantes do KNN (ZHANG, 2016). Para isso, utilizaremos um valor de `k` igual a 3 para o número de vizinhos, devido à sua simplicidade e aplicabilidade comprovada em problemas semelhantes. Se o modelo não alcançar a precisão de 70% no conjunto de teste, ajustaremos o valor de `k` para 5, a fim de observar se uma maior quantidade de vizinhos pode melhorar o desempenho do modelo.

A regressão linear será uma das alternativas a ser explorada. Embora tradicionalmente voltada para problemas de regressão, a regressão linear pode ser adaptada para tarefas de classificação binária (RODRIGUES, 2012). Usaremos o *LogisticRegression* do *Scikit-learn* para esse fim. Assim como o KNN, o modelo de regressão linear será treinado com os dados de treinamento, e seus parâmetros serão ajustados, se necessário, para maximizar a precisão. A avaliação será feita com base nos resultados obtidos no conjunto de teste.

Outra alternativa será o Naive Bayes, que é particularmente eficiente para problemas de classificação, especialmente quando as características são independentes entre si (RISH et al., 2001). Usaremos a implementação do *NaiveBayes* disponível no *Scikit-learn*, treinando o modelo da mesma forma que os anteriores, e ajustando seus parâmetros

conforme necessário.

Os dados utilizados para treinar e testar os modelos serão divididos em três conjuntos: treinamento, validação e teste. O conjunto de treinamento será utilizado para ajustar os parâmetros do modelo, o de validação ajudará a ajustar os hiperparâmetros como o valor de  $k$  no KNN, e o conjunto de teste será usado para avaliar a precisão final do modelo.

A avaliação final do modelo será feita na Atividade 6, onde a precisão dos diferentes modelos será comparada. Nosso objetivo é alcançar uma precisão de 70% na classificação dos embriões como euploides ou aneuploides. Caso isso não seja alcançado com o KNN, a regressão linear e o Naive Bayes serão testados até que o modelo que melhor se ajuste aos dados seja encontrado, com a precisão desejada.

Ao final, espera-se que o modelo final seja capaz de classificar corretamente os embriões com pelo menos 70% de precisão, permitindo uma análise confiável da euploidia, que pode ser utilizada como apoio em estudos científicos e aplicações clínicas.

### 3.2.1.3 Objetivo Específico 3 - Avaliação do modelo

#### 3.2.1.3.1 Atividade 6 (A6): Utilizar métricas de avaliação mais adequadas para medir o desempenho do modelo de acordo com a natureza do problema de classificação

O objetivo dessa atividade é aplicar e avaliar métricas adequadas para medir a confiança e o desempenho do modelo de IA, considerando as especificidades e objetivos do problema abordado. As métricas escolhidas incluem Acurácia, Precisão, Recall (Sensibilidade) e F1-Score. Cada métrica será implementada e analisada com base em um conjunto de dados previamente definido, utilizando ferramentas de análise estatística e frameworks de aprendizado de máquina. A análise será realizada para garantir que as métricas reflitam de forma eficaz o desempenho do modelo em termos de sua capacidade de generalização e identificação de padrões no conjunto de teste.

A definição e cálculo das métricas de avaliação serão conduzidos utilizando *Python* e bibliotecas amplamente empregadas em aprendizado de máquina, como *scikit-learn*, *pandas*, para análise, visualização e manipulação dos resultados. O processo será realizado para fornecer uma análise detalhada e equilibrada do classificador. As métricas escolhidas incluem acurácia, precisão, *recall* e *F1-Score*. A acurácia mede a proporção de previsões corretas em relação ao total de previsões realizadas (JUNIOR et al., 2022). Essa métrica é útil para problemas onde as classes estão balanceadas e não há uma preocupação maior com erros específicos, como falsos positivos ou falsos negativos (JUNIOR et al., 2022). Ela será calculada usando a função *accuracy\_score* do *scikit-learn*. A precisão avalia a proporção de exemplos classificados como positivos que realmente pertencem à classe positiva (JUNIOR et al., 2022). É crucial em problemas onde falsos positivos têm

consequências severas, como em diagnósticos médicos (JUNIOR et al., 2022). Será calculada com a função `precision_score` do *scikit-learn*. O *recall* mede a capacidade do modelo de identificar corretamente os exemplos pertencentes à classe positiva (JUNIOR et al., 2022). É especialmente importante em situações onde falsos negativos têm maior impacto (JUNIOR et al., 2022), como na detecção de aneuploidia de embriões. Será calculado com a função `recall_score` do *scikit-learn*. O *F1-Score* combina precisão e *recall*, fornecendo uma visão equilibrada entre ambos (JUNIOR et al., 2022). É particularmente relevante em casos onde as classes estão desbalanceadas e há necessidade de avaliar o desempenho geral do modelo (JUNIOR et al., 2022). O cálculo será realizado utilizando a função `f1_score` do *scikit-learn*.

Espera-se que a análise detalhada das métricas forneça uma visão abrangente do desempenho do modelo, destacando seus pontos fortes e fracos em diferentes cenários. A Acurácia deverá apresentar um panorama geral da desempenho, enquanto Precisão, Recall e F1-Score deverão evidenciar aspectos específicos de classificação positiva e negativa (JUNIOR et al., 2022). A métrica *ROC-AUC* permitirá avaliar a capacidade geral do modelo de distinguir entre classes (JUNIOR et al., 2022). Os resultados deverão ser utilizados para refinar e ajustar o modelo, garantindo maior confiabilidade e aderência aos objetivos propostos no problema de classificação. Além disso, a escolha criteriosa das métricas será essencial para orientar decisões estratégicas relacionadas ao modelo, especialmente em aplicações sensíveis a erros de classificação.

#### 3.2.1.3.2 **Atividade 7 (A7):** Avaliar a precisão e eficácia do modelo em prever corretamente casos de euploidia e aneuploidia por meio da Matriz de Confusão e Curva ROC

O objetivo desta atividade é avaliar o desempenho do modelo de classificação desenvolvido para prever corretamente os casos de euploidia e aneuploidia, utilizando a Matriz de Confusão e a Curva ROC (Receiver Operating Characteristic). A Matriz de Confusão será construída após o treinamento e teste do modelo, permitindo identificar as taxas de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. Essa análise ajudará a compreender o desempenho geral do modelo e a identificar possíveis áreas de melhoria. Adicionalmente, será gerada a Curva *ROC* para avaliar o desempenho do modelo na separação das duas classes: euploide (classe positiva) e aneuploide (classe negativa) (JUNIOR et al., 2022). A curva será analisada com base em diferentes valores de limiar (threshold), fornecendo uma visão detalhada sobre a sensibilidade e a especificidade do modelo em cada ponto. A métrica AUC (Área sob a Curva) será utilizada como indicador global da capacidade do modelo de distinguir entre as classes, sendo especialmente útil para avaliar problemas de classificação desbalanceada (JUNIOR et al., 2022).

A avaliação do desempenho do modelo será realizada em duas etapas principais: a

construção da Matriz de Confusão e a geração da Curva ROC. A Matriz de Confusão é uma ferramenta essencial para entender o desempenho do modelo de classificação. De acordo com [Sathyanarayanan e Tantri \(2024\)](#), essa matriz é uma tabela de dimensão  $N \times N$ , onde  $N$  representa o número de classes. Cada linha da matriz indica a quantidade de instâncias previstas em uma classe, enquanto cada coluna representa a quantidade de instâncias reais da classe. Para este estudo, a matriz permitirá a análise de predições corretas e incorretas, classificando-as como verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN).

A partir dessas classificações, podemos calcular diversas métricas importantes para medir a precisão do modelo, como a acurácia, precisão, *recall* e *F1-score*, que nos ajudam a identificar as áreas de melhoria no modelo de predição. Além disso, será utilizada uma Curva ROC para avaliar o desempenho do modelo de forma mais detalhada. O modelo de classificação Random Forest será implementado utilizando a biblioteca *scikit-learn*, que permite gerar probabilidades de pertença à classe positiva (euploide). A Curva ROC, que é fundamental para problemas de classificação binária, é baseada nessas probabilidades, em vez de apenas classificações binárias ([JUNIOR et al., 2022](#)).

A curva será gerada variando o limiar de decisão, (*threshold*), do modelo. O limiar define a probabilidade a partir da qual uma instância será classificada como pertencente à classe positiva (euploide) ([JUNIOR et al., 2022](#)). O *threshold* será ajustado para diferentes valores, e para cada um, será calculada a sensibilidade (taxa de verdadeiros positivos) e a especificidade (1 - taxa de falsos positivos) ([JUNIOR et al., 2022](#)). Para isso, utilizaremos o método `predict_proba()` do *scikit-learn* para obter as probabilidades previstas pelo modelo. A função `roc_curve()` da biblioteca também será utilizada para calcular os valores de falso positivo e verdadeiro positivo para os diferentes limiares, gerando o gráfico da Curva ROC, com a especificidade no eixo x e a sensibilidade no eixo y.

Finalmente, a métrica *AUC* (Área sob a Curva) será calculada utilizando a função `roc_auc_score()` do *scikit-learn*. A *AUC*, que varia de 0 a 1, fornecerá uma avaliação quantitativa do modelo, sendo que valores mais próximos de 1 indicam um melhor desempenho na classificação. A Curva ROC, junto com a *AUC*, nos ajudará a entender o comportamento do modelo para diferentes limiares e a selecionar o melhor ponto de corte, equilibrando os *trade-offs* entre a taxa de verdadeiros positivos e falsos positivos.

Ao construir a Matriz de Confusão, espera-se obter uma análise detalhada do desempenho do modelo, identificando os verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Essa análise permitirá avaliar não apenas a precisão global do modelo, mas também as taxas de erro em diferentes categorias, como os casos de euploidia erroneamente classificados como aneuploidia (FP) e os de aneuploidia erroneamente classificados como euploidia (FN). Essa avaliação fornecerá subsídios para aprimoramentos no modelo de classificação. Para a Curva ROC, o gráfico

gerado mostrará a relação entre a sensibilidade e a especificidade em diferentes valores de limiar. Espera-se que o modelo apresente uma curva ascendente, indicando sua capacidade de identificar corretamente os positivos (euploide) sem gerar muitos falsos positivos. O valor da AUC (Área sob a Curva) será calculado para quantificar a habilidade do modelo em distinguir entre as classes (JUNIOR et al., 2022). Valores de AUC próximos de 1 indicam um excelente desempenho do modelo, enquanto valores próximos de 0,5 sugerem que o modelo apresenta desempenho similar a uma escolha aleatória (JUNIOR et al., 2022). A partir da Curva ROC, será possível selecionar o limiar mais adequado para balancear os erros de falso positivo e falso negativo, permitindo uma análise criteriosa dos trade-offs envolvidos (JUNIOR et al., 2022). Esses resultados fornecerão uma base sólida para avaliar a eficácia do modelo e sua aplicabilidade no contexto do estudo.

#### 3.2.1.4 Objetivo Específico 4 - Protótipo de Interface

##### 3.2.1.4.1 Atividade 8 (A8): Prototipar uma interface básica para exibir as predições de euploidia para o usuário final (médicos)

A finalidade é desenvolver o protótipo de uma interface básica que possibilite aos médicos visualizar as previsões de euploidia produzidas pelo modelo. O protótipo incluirá componentes cruciais como campos de preenchimento para os dados necessários à previsão, botões de interação para envio de informações, além de uma área de apresentação dos resultados.

Para criar uma interface fácil de usar que permita aos médicos visualizar as previsões de euploidia, empregaremos o Figma, um software colaborativo baseado na web para a criação de interfaces. O Figma disponibiliza funcionalidades que simplificam a criação de interfaces de usuário e experiências do usuário, priorizando a colaboração em tempo real, por meio de uma gama de ferramentas de edição vetorial e prototipagem (Figma, 2024).

O resultado é que o protótipo da interface básica possibilite aos médicos uma visualização clara e intuitiva das previsões de euploidia. A interface precisa ser funcional e esteticamente atraente, assegurando a compreensão simples dos resultados exibidos. Serão coletadas opiniões valiosas sobre o design, a navegação e a clareza das informações apresentadas, possibilitando as alterações necessárias.

### 3.3 Passos para o Desenvolvimento de um Algoritmo de Aprendizado de Máquina

O desenvolvimento de sistemas baseados em aprendizado de máquina é um processo que envolve uma série de etapas bem definidas, cada uma desempenhando um papel

crucial para o sucesso do modelo final. Essas etapas vão desde a definição inicial do problema até a implementação e manutenção do sistema em um ambiente de produção. Segundo [Géron \(2017\)](#), em *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, seguir uma abordagem estruturada permite que os desenvolvedores abordem os desafios de forma sistemática, garantindo não apenas a eficácia técnica do modelo, mas também sua aplicabilidade prática. O processo começa com a definição do problema e o entendimento do contexto geral. Depois, os dados são coletados, explorados para descobrir padrões e preparados para melhorar os resultados dos algoritmos. Em seguida, diferentes modelos são testados, e os melhores são ajustados para formar uma solução final. Essa solução é apresentada, implementada em produção e monitorada continuamente para garantir que funcione bem ao longo do tempo ([GÉRON, 2017](#)).

### 3.3.1 Definição do problema e análise do panorama geral

A definição clara do problema e a análise do panorama geral são etapas essenciais para o sucesso de projetos de aprendizado de máquina, pois orienta todas as decisões subsequentes, desde a coleta de dados até a implementação final. Conforme abordado por [Géron \(2017\)](#) e [Müller \(2017\)](#), essas etapas fornecem a base para decisões estratégicas ao longo do desenvolvimento do modelo. Isso envolve determinar como a solução será utilizada, identificar o tipo de problema (como classificação ou regressão) e escolher métricas de desempenho que estejam alinhadas aos objetivos esperados ([GÉRON, 2017](#)). O primeiro passo é estabelecer claramente o objetivo em termos de negócio e pesquisa.

Além disso, é importante avaliar as soluções existentes ou alternativas em uso, que podem servir como ponto de comparação para medir o impacto do modelo. Também é necessário validar hipóteses iniciais sobre os dados e a abordagem, garantindo que o problema esteja bem estruturado e que as limitações sejam compreendidas desde o início. Segundo [Müller \(2017\)](#), uma compreensão profunda dos dados e suas características é essencial para a escolha dos algoritmos e para o sucesso do projeto. Perguntas como “Quantos dados possuo?”, “Há dados faltantes?” e “Esses dados são suficientes para responder às perguntas do projeto?” guiam essa análise ([MÜLLER, 2017](#)).

Por fim, essa etapa também exige atenção ao alinhamento entre o problema técnico e os resultados esperados em termos de negócio ou impacto social. Isso garante que o desenvolvimento não seja apenas tecnicamente sólido, mas também relevante e eficaz em seu contexto de aplicação.

### 3.3.2 Obtenção de Dados

A etapa de obtenção de dados é um dos pilares fundamentais para o sucesso de um projeto de aprendizado de máquina, pois a qualidade e a relevância das informações co-



letadas impactam diretamente a eficácia do modelo (GÉRON, 2017). O processo começa com a identificação e a listagem dos dados necessários, levando em conta sua quantidade e suas características, como formato, tipo e origem (GÉRON, 2017). É igualmente importante garantir que as fontes de dados sejam documentadas, que haja espaço suficiente para armazenamento e que os dados estejam acessíveis de maneira eficiente.

Nessa etapa, é imprescindível observar as obrigações legais e éticas, especialmente as previstas na Lei Geral de Proteção de Dados (LGPD) no Brasil. Isso abrange a obtenção de consentimentos adequados para o uso das informações e a aplicação de medidas de segurança, como a anonimização, para proteger dados sensíveis (MÜLLER, 2017). Conforme destacado por Müller (2017), além de cumprir os requisitos legais, é fundamental assegurar a integridade e a privacidade dos dados, principalmente em situações que envolvam informações pessoais ou confidenciais.

O livro *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* do Géron (2017) aborda que a qualidade dos dados é fundamental para o sucesso de qualquer projeto de Machine Learning. Modelos treinados com dados insuficientes, não representativos ou de baixa qualidade tendem a apresentar problemas como overfitting, underfitting ou generalizações imprecisas. Dados não representativos podem introduzir vieses que prejudicam a aplicação do modelo em cenários reais, enquanto dados de baixa qualidade, como informações inconsistentes, incompletas ou ruidosas, comprometem diretamente a capacidade do modelo de identificar padrões úteis.

Géron (2017) também destaca que a coleta e o preparo dos dados vão além do aspecto técnico. Essas atividades envolvem práticas organizacionais, como a colaboração entre equipes de negócios e engenharia para identificar as fontes de dados mais relevantes, e uma análise cuidadosa para garantir que os dados refletem a realidade do problema a ser resolvido. Além disso, questões legais e éticas desempenham um papel central, principalmente em situações que envolvam informações sensíveis ou pessoais. A conformidade com legislações como a GDPR na Europa e a LGPD no Brasil não apenas protege os dados, mas também promove a confiança dos usuários e das partes interessadas (GÉRON, 2017).

Para mitigar os desafios relacionados à qualidade dos dados, Géron (2017) sugere o uso de técnicas como a limpeza, normalização e engenharia de atributos. Além disso, ele ressalta a importância de práticas contínuas de validação, como a separação adequada entre conjuntos de treinamento, validação e teste, garantindo que o modelo seja testado em dados que nunca encontrou antes. Dessa forma, a abordagem holística para o gerenciamento de dados reforça o desenvolvimento de sistemas robustos, confiáveis e éticos em projetos de Aprendizado de Máquina (GÉRON, 2017).



### 3.3.3 Exploração de Dados

A obtenção e a exploração de dados representam etapas essenciais para o sucesso de projetos de aprendizado de máquina, pois determinam a base sobre a qual os modelos serão desenvolvidos (GÉRON, 2017). Antes de preparar os dados para os algoritmos, é essencial entender as características e as relações existentes no conjunto de dados, iniciando a exploração com uma cópia dos dados originais, reduzindo sua escala, se necessário, para facilitar a análise inicial (GÉRON, 2017).

A exploração dos dados deve seguir uma abordagem sistemática que envolve:

1. **Estudo das características dos atributos:** Identificar o nome, o tipo (categórico, numérico, texto, etc.), o percentual de valores ausentes, o nível de ruído (outliers, erros de arredondamento), a utilidade potencial para a tarefa e o tipo de distribuição (gaussiana, uniforme, logarítmica) dos dados disponíveis.
2. **Identificação de atributos-alvo:** No caso de aprendizado supervisionado, determinar qual atributo será o alvo da predição.
3. **Visualização de dados:** Criar gráficos de dispersão, histogramas ou outros métodos visuais para identificar padrões, correlações e tendências. Géron (2017) sugere, por exemplo, experimentar combinações de atributos, como comparar o número de quartos por domicílio, em vez de analisar apenas o número total de quartos.
4. **Correlação entre atributos:** Analisar as relações entre as variáveis para identificar combinações promissoras que possam melhorar a precisão do modelo.

Müller (2017) destaca que a inspeção visual dos dados é essencial para compreender sua estrutura e identificar inconsistências, como unidades de medida divergentes ou valores inesperados, comuns em cenários reais. Ele recomenda o uso de gráficos de dispersão para analisar relações entre dois atributos ou gráficos de pares para explorar múltiplas combinações quando o número de variáveis é pequeno. Essa etapa também permite verificar se o problema pode ser resolvido manualmente, validando se as informações necessárias estão presentes no conjunto de dados.

Além disso, Géron (2017) ressalta a importância de aplicar transformações aos atributos, criando variáveis derivadas mais relevantes, como "população por domicílio", em vez de usar dados brutos. Ele também enfatiza a necessidade de documentar aprendizados e, se necessário, ajustar o escopo do projeto para incluir dados adicionais que possam melhorar os resultados. Essas práticas tornam a análise de dados uma etapa crucial para preparar modelos robustos e aumentar as chances de sucesso no projeto.

### 3.3.4 Preparação dos dados para os Algoritmos de Aprendizado de Máquina

A preparação dos dados para algoritmos de aprendizado de máquina é uma etapa essencial que envolve diversas transformações. Para tornar o processo mais eficiente, é recomendável criar funções específicas para realizar essas transformações, o que permite aplicá-las facilmente em novos conjuntos de dados e reutilizá-las em projetos futuros (GÉRON, 2017). Entre as principais transformações, destaca-se a limpeza dos dados, que envolve lidar com valores ausentes. Para atributos com valores ausentes existem três opções: eliminar os registros correspondentes, excluir o atributo inteiro ou substituir os valores faltantes por uma constante, como a média ou a mediana (GÉRON, 2017).

Outro aspecto crucial na preparação dos dados é a escalabilidade dos atributos. Muitos algoritmos de aprendizado de máquina não funcionam bem quando as variáveis numéricas têm escalas muito diferentes (GÉRON, 2017). Recomenda-se o uso de técnicas de escalonamento de atributos, como a normalização (min-max scaling), que ajusta os valores para um intervalo de 0 a 1, ou a padronização (standardization), que ajusta os dados para ter média zero e variância unitária (GÉRON, 2017). A escolha entre essas duas técnicas depende das características do algoritmo utilizado, já que a padronização é menos afetada por outliers e pode ser mais indicada em certos casos, como em redes neurais. A partir desse processo de preparação, o próximo passo é a seleção de modelos promissores, onde o uso de validação cruzada e a análise dos erros dos modelos ajudam a refinar a escolha do modelo mais adequado para o problema em questão (GÉRON, 2017).

### 3.3.5 Seleção e treinamento do modelo

A seleção e o treinamento de modelos em aprendizado de máquina são etapas essenciais para desenvolver soluções eficazes (GÉRON, 2017). Após a preparação dos dados, que inclui a exploração e a transformação, o próximo passo é escolher e treinar um modelo adequado. Para isso, um modelo simples como a *Regressão Linear* pode ser treinado no conjunto de dados, permitindo observar sua performance inicial (GÉRON, 2017). No entanto, para uma avaliação mais precisa, a validação cruzada é uma abordagem melhor. Essa técnica divide o conjunto de treinamento em K subconjuntos e, em seguida, treina e avalia o modelo várias vezes, usando um subconjunto diferente para validação a cada vez, o que gera uma estimativa mais confiável da performance do modelo (GÉRON, 2017).

Uma alternativa eficaz para melhorar o desempenho do modelo é utilizar *Random Forests*, que combinam múltiplas árvores de decisão para melhorar a acurácia e reduzir o risco de overfitting (GÉRON, 2017). O aprendizado de conjunto (ensemble learning) é uma técnica poderosa que utiliza a combinação de diversos modelos para aumentar a robustez e a precisão do modelo final (GÉRON, 2017). A abordagem de treinar múltiplos modelos com parâmetros padrão e avaliá-los usando validação cruzada permite selecionar

as melhores opções para o problema em questão (GÉRON, 2017).

Além disso, para aprimorar a performance do modelo, é fundamental realizar uma análise das variáveis mais relevantes e ajustar as características dos dados (GÉRON, 2017). A engenharia de atributos e a seleção de features permitem que os modelos se ajustem para cometer diferentes tipos de erros, o que ajuda a melhorar a precisão geral. Essas etapas devem ser repetidas de forma iterativa, ajustando o modelo com base nas análises de erros e nas mudanças nos dados, o que possibilita a evolução do desempenho do modelo ao longo do processo (GÉRON, 2017).

No caso de problemas de classificação, o *algoritmo k-vizinhos mais próximos (k-NN)* é uma das opções mais simples e eficazes (MÜLLER, 2017). Esse modelo faz previsões baseadas na proximidade dos dados de treinamento, atribuindo o rótulo da classe mais comum entre os vizinhos mais próximos de um ponto desconhecido (MÜLLER, 2017). A definição do parâmetro 'k', que determina quantos vizinhos são considerados, pode ser ajustada para otimizar os resultados. A implementação do *KNeighborsClassifier* no *Scikit-learn* facilita a criação e a avaliação do modelo, tornando-o uma excelente escolha para tarefas de classificação simples (MÜLLER, 2017).

### 3.3.6 Ajuste do modelo

O processo de ajuste fino de modelos (fine-tuning) é uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina, especialmente após a seleção de modelos promissores (GÉRON, 2017). Uma das abordagens mais comuns para realizar esse ajuste é o uso de *GridSearchCV* do *Scikit-Learn*, que automatiza a busca pelos melhores hiperparâmetros do modelo. Em vez de testar manualmente combinações de valores para os hiperparâmetros, o *GridSearchCV* avalia todas as possibilidades de uma lista de valores, utilizando validação cruzada para escolher a melhor configuração (GÉRON, 2017). Isso facilita o processo, economizando tempo e aumentando a precisão na escolha dos parâmetros ideais para o modelo.

No entanto, para espaços de busca de hiperparâmetros maiores, a *RandomizedSearchCV* pode ser uma alternativa mais eficiente. Em vez de testar todas as combinações possíveis, essa abordagem realiza uma busca aleatória, selecionando valores aleatórios para cada hiperparâmetro em cada iteração (GÉRON, 2017). Essa técnica oferece duas grandes vantagens: a possibilidade de explorar um maior número de combinações de hiperparâmetros dentro de um orçamento computacional limitado, além de permitir um controle mais flexível sobre o número de iterações realizadas (Géron, 2017). Dessa forma, a *RandomizedSearchCV* é especialmente útil quando o espaço de busca é grande e as combinações possíveis são muitas (GÉRON, 2017).

Além do ajuste de hiperparâmetros, outra técnica importante para aprimorar o

modelo é o uso de *Métodos de Ensemble*. Esses métodos combinam os melhores modelos individuais, muitas vezes resultando em um desempenho superior ao de qualquer modelo isolado (GÉRON, 2017). A combinação de modelos com erros diferentes pode reduzir a variabilidade e melhorar a precisão geral. Um exemplo clássico é o *Random Forest*, que utiliza múltiplas árvores de decisão para obter melhores resultados do que uma única árvore. A estratégia de ensemble pode ser fundamental para melhorar o desempenho do modelo, especialmente em tarefas complexas (GÉRON, 2017).

Após o ajuste fino, é importante analisar os melhores modelos e seus erros para entender melhor o desempenho do sistema. Inspeccionar os atributos mais importantes para a previsão pode revelar insights valiosos sobre o problema (MÜLLER, 2017). Além disso, entender os tipos de erros cometidos pelo modelo e as razões por trás deles pode ajudar a ajustar o modelo, adicionando ou removendo features, tratando outliers ou refinando a transformação dos dados (MÜLLER, 2017). Finalmente, após realizar todas essas melhorias, o modelo deve ser avaliado no conjunto de teste para estimar sua capacidade de generalização. A avaliação no conjunto de teste fornece uma medida objetiva da performance do modelo em dados não vistos, sendo fundamental para garantir que o modelo não esteja superajustado aos dados de treinamento (MÜLLER, 2017).

### 3.3.7 Lançamento da Solução

Após a aprovação do lançamento de um projeto, é crucial preparar a solução para produção, conectando as fontes de dados de entrada e escrevendo os testes necessários para garantir que o sistema funcione conforme esperado. A integridade e a qualidade do sistema, bem como sua adaptação ao ambiente de produção, são essenciais para que ele se mantenha eficiente. Além disso, é necessário implementar códigos de monitoramento que acompanhem a performance do sistema em tempo real, verificando se ele continua operando de maneira eficiente e acionando alertas sempre que houver uma queda na performance (GÉRON, 2017). Isso é especialmente importante, pois os modelos de aprendizado de máquina podem sofrer uma degradação gradual ao longo do tempo devido à mudança nos dados, o que é conhecido como "data drift" (GÉRON, 2017).

Para garantir a qualidade das previsões, o monitoramento também deve incluir uma avaliação humana periódica, muitas vezes realizada por analistas ou por meio de plataformas de *crowdsourcing*, como o *Amazon Mechanical Turk* ou o *CrowdFlower*. Essa análise ajuda a validar se o modelo continua atendendo às expectativas e fornece informações cruciais sobre possíveis melhorias (MÜLLER, 2017). A qualidade dos dados de entrada também deve ser constantemente monitorada, já que problemas como sensores defeituosos ou dados desatualizados podem impactar significativamente a acuracidade do modelo, prejudicando sua performance e a confiança nos resultados gerados (MÜLLER, 2017).

A manutenção do sistema é outro aspecto crítico após o lançamento, exigindo que o modelo seja re-treinado regularmente com dados frescos (GÉRON, 2017). Isso é importante para evitar que a performance do sistema flutue de forma inesperada, o que pode acontecer se o modelo for atualizado de maneira esporádica (GÉRON, 2017). A automação desse processo de re-treinamento é essencial para garantir que o modelo seja atualizado sempre que necessário, sem depender de intervenções manuais.

Portanto, o lançamento de um sistema de aprendizado de máquina não se limita à integração inicial dos dados e à validação do modelo. O sucesso contínuo depende de um monitoramento eficiente e de uma manutenção constante, garantindo que o modelo se adapte às mudanças nos dados ao longo do tempo. Implementando uma infraestrutura de monitoramento e re-treinamento robusta, as equipes podem assegurar que o sistema continue atendendo aos objetivos de negócios de forma eficaz e precisa, minimizando riscos e maximizando a performance ao longo de sua vida útil.

## 4 Execução da Pesquisa

Antes de começar a "Execução da Pesquisa", é crucial enfatizar que todas as fases deste estudo foram realizadas de acordo com as normas éticas e legais pertinentes à investigação científica. A documentação que comprova isso está devidamente incluída nos anexos deste estudo, incluindo os seguintes documentos essenciais:

- **Parecer da Plataforma Brasil:** Inclui a permissão ética que valida a utilização dos dados dos embriões para a execução deste estudo, garantindo a conformidade com os padrões éticos definidos para pesquisas que envolvem seres humanos (disponível no Anexo I)
- **Contrato de Autorização para Utilização de Dados em Pesquisa:** Assinado pelo Dr. Bruno Ramalho, que autoriza o uso dos dados clínicos de suas pacientes, essenciais para a modelagem e análise realizadas durante esta pesquisa (disponível no Anexo III)
- **Termo de Consentimento para Utilização de Dados de Entrevistas, Gravação de Reuniões e Uso de Gravação:** Contrato que permite o uso de discursos, dados e gravações provenientes das reuniões e entrevistas conduzidas com o Dr. Bruno Ramalho, garantindo que as informações e contribuições fornecidas por ele fossem utilizadas com total autorização (disponível no Anexo II).

Estes documentos evidenciam o comprometimento ético e a transparência desta pesquisa, enfatizando que todas as informações empregadas foram obtidas de maneira responsável e autorizada.

### 4.1 Fase 1: Análise e Preparação de Dados

#### 4.1.1 OE1 - Expansão, Processamento e Análise de Dados para Predição de Ploidia

##### 4.1.1.1 Atividade 1 (A1): Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia

O desenvolvimento dessa atividade teve início com a elaboração do referencial teórico, abordando os conceitos fundamentais relacionados ao aprendizado supervisionado e às tecnologias aplicadas no contexto da fertilização in vitro, além de nos aprofundarmos em pesquisas relacionadas a área da medicina reprodutiva. A pesquisa teórica foi direcionada ao entendimento dos padrões morfocinéticos de embriões capturados por sistemas de

Time-Lapse e de como essas informações podem ser utilizadas para prever a porcentagem de euploidia, um indicador crítico da saúde genética do embrião.

Os dados utilizados neste estudo foram obtidos por meio do especialista, [Ramalho \(2024\)](#), e de suas pacientes, que autorizaram o uso dessas informações para fins de pesquisa, contratos esses que estão na sessão de Anexo. As informações foram coletadas na clínica Bruno Ramalho Reprodução Humana, bem como na clínica Genesis, onde o [Ramalho \(2024\)](#) também atua.

A primeira etapa dessa atividade consiste em verificar a pertinência das variáveis já existentes na “**Planilha Original de Dados dos Embriões**”, que se encontra no Anexo ??, analisando seu impacto no desempenho do modelo preditivo. Essa análise também busca identificar se há necessidade de introduzir novas variáveis que possam melhorar a precisão do modelo ou excluir aquelas que se mostram irrelevantes.

#### 4.1.1.2 Limpeza dos Dados

Ao lidar com os dados ausentes, utilizamos o método de Análise de Casos Completos (ACC), que envolve a remoção de observações que possuem pelo menos um valor ausente ([CAMARGOS et al., 2011](#)). Este procedimento é frequentemente empregado quando o número de dados ausentes é reduzido, assegurando que a remoção de algumas observações não interfira de forma significativa na análise estatística e preserva a consistência do modelo ([CAMARGOS et al., 2011](#)). Em nosso cenário, das 84 linhas de dados disponíveis, apenas 2 apresentavam valores ausentes. As células sem dados estão em preto na “**Planilha com Indicação de Dados Ausentes**” no Anexo ??.

#### 4.1.1.3 Atividade 2 (A2): Identificação da Correlação entre os Parâmetros na Previsão da Ploidia do Embrião

O coeficiente de **correlação de Spearman**, explicado no APÊNDICE D, foi escolhido para esta atividade devido à necessidade de reconhecer relações entre os parâmetros documentados nos dados dos embriões e a porcentagem de euploidia. A metodologia leva em conta a classificação ordinal das observações, minimizando os efeitos de valores atípicos e permitindo a análise de variáveis com distribuições não normais ([SOUSA, 2019](#)). Além disso, a sua facilidade de uso em pesquisas científicas e a sua vasta aplicação em estudos destacam sua utilidade na análise de ploidia dos embriões.

O programa em Python criado para essa análise foi desenvolvido para ser modular, eficaz e gerar resultados claros tanto em formato visual quanto textual. Ele emprega as bibliotecas *Pandas*, *SciPy*, *Matplotlib* e *python-docx* para manipular dados, determinar correlações, produzir gráficos e elaborar relatórios em Word. A seguir, detalha-se a lógica do código:

- **Carregamento e Preparação dos Dados:** As informações dos embriões foram processadas a partir da Planilha de Dados Refinada, encontrada no Anexo ??, utilizando todas as colunas disponíveis. Essa etapa garante que todas as variáveis relevantes sejam consideradas.
- **Cálculo da Correlação de Spearman:** O coeficiente de Spearman foi calculado para todas as combinações possíveis de variáveis, utilizando a função `spearmanr` da biblioteca SciPy. O método "omit" foi empregado para lidar com valores ausentes, garantindo maior robustez mesmo que não existam dados faltantes nesta planilha.
- **Criação de Gráficos de Dispersão:** Para cada par de variáveis, foram criados gráficos de dispersão que auxiliam na análise numérica e permitem a identificação visual de padrões. As cores foram padronizadas, com a variável 1 (`var1`) em verde claro e marcador "o", e a variável 2 (`var2`) em azul escuro e marcador "x".
- **Elaboração de Relatório Automatizado:** O relatório gerado contém descrições textuais dos coeficientes e gráficos correspondentes. Este documento automatizado facilita a comunicação visual e escrita dos resultados.

O código em Python utilizado para realizar a análise está apresentado a seguir:

```

1  import pandas as pd
2  from scipy.stats import spearmanr
3  import matplotlib.pyplot as plt
4  from docx import Document
5  from docx.shared import Inches
6  import os
7
8  file_path = "PlanilhaDadosRefinados.xlsx"
9  data = pd.read_excel(file_path)
10
11 numerical_columns = [
12     "Idade", "Estágio", "Morfo", "KIDScore", "t2", "t3", "t4",
13     "t5", "t8", "tSC", "tSB", "tB", "cc2 (t3-t2)",
14     "cc3 (t5-t3)", "t5-t2", "s2 (t4-t3)", "s3 (t8-t5)", "tSC-t8",
15     "tB-tSB", "Ploidia"
16 ]
17
18 numerical_columns = [col for col in numerical_columns if col in
    data.columns]
19
20 results = []

```



```
19     for col1 in numerical_columns:
20         for col2 in numerical_columns:
21             if col1 != col2:
22                 coef, _ = spearmanr(data[col1], data[col2],
23                                     nan_policy="omit")
24                 results.append({"Variable 1": col1, "Variable 2":
25                                 col2, "Spearman Coefficient": coef})
26
27 correlation_df = pd.DataFrame(results)
28
29 output_file = "correlation_results.xlsx"
30 correlation_df.to_excel(output_file, index=False)
31 print(f"\nResultados salvos no arquivo: {output_file}")
32
33 doc = Document()
34 doc.add_heading('Correlação de Spearman - Embriões', 0)
35
36 output_dir = "gráficos"
37 os.makedirs(output_dir, exist_ok=True)
38
39 color_var1 = "lightgreen"
40 color_var2 = "darkblue"
41
42 for index, row in correlation_df.iterrows():
43     var1, var2 = row['Variable 1'], row['Variable 2']
44     coef = row['Spearman Coefficient']
45
46     plt.figure(figsize=(6, 4))
47     plt.scatter(data[var1], data[var2], alpha=0.6, c=color_var1
48                 , label=f'{var1}', marker='o')
49     plt.scatter(data[var2], data[var1], alpha=0.6, c=color_var2
50                 , label=f'{var2}', marker='x')
51
52     plt.title(f"Dispersão entre {var1} e {var2} (Coeficiente de
53               Spearman: {coef:.2f})")
54     plt.xlabel(var1)
55     plt.ylabel(var2)
56
57     img_path = os.path.join(output_dir, f"grafico_{var1}_{var2}
58                               }.png")
59     plt.savefig(img_path)
60     plt.close()
```

```
55
56     doc.add_heading(f'Correlação entre {var1} e {var2}', level
57                     =1)
58     doc.add_paragraph(f'Coeficiente de Spearman: {coef:.2f}')
59     doc.add_picture(img_path, width=Inches(5.0))
60
61     output_word = "relatorio_correlacoes.docx"
62     doc.save(output_word)
63
64     print(f"\nRelatório gerado e salvo em: {output_word}")
```

O código analisa a correlação entre as variáveis numéricas em uma sequência de dados, empregando o coeficiente de Spearman. Ele inicia carregando a planilha Excel com as informações dos embriões e seleciona todas as colunas relevantes para a análise. O coeficiente de Spearman é calculado para cada par de variáveis, avaliando a intensidade e a direção da relação monótona entre elas. Os resultados são armazenados em um `DataFrame`, que é uma estrutura de dados bidimensional do Pandas, semelhante a uma tabela, permitindo fácil manipulação e análise. Esta tabela é então exportada para um arquivo Excel chamado `correlation_results.xlsx`.

Após isso, o programa gera gráficos de dispersão para observar as correlações, ressaltando os padrões das variáveis associadas. Esses gráficos, juntamente com os coeficientes de correlação, são automaticamente incorporados a um documento Word. No final, temos um relatório completo com as análises e visualizações, salvo como um arquivo chamado `relatorio_correlacoes.docx`.

#### 4.1.1.4 Atividade 3 (A3): Normalização dos Dados para Otimização

A etapa de normalização dos dados é um passo fundamental na criação de modelos de aprendizado de máquina, particularmente em situações onde as variáveis têm diferentes escalas e distribuições. Esta tarefa foi executada com o uso do método Z-Score. Esta metodologia modifica os dados de forma que cada variável possua uma média de 0 e um desvio padrão de 1, explicado na APÊNDICE B.

A normalização foi conduzida com Python e a biblioteca Pandas. O procedimento foi automatizado para simplificar a análise em grande escala dos dados morfocinéticos dos embriões. A normalização foi aplicada especificamente às variáveis numéricas do conjunto de dados, previamente selecionadas na fase de escolha das variáveis. O código utilizado para realizar essa atividade foi:

```
1     import pandas as pd
2     import numpy as np
```

```
3
4
5 file_path = "PlanilhaDeDadosDosEmbriões4.xlsx"
6 data = pd.read_excel(file_path)
7
8
9 numeric_columns = data.select_dtypes(include=[np.number]).
    columns
10
11
12 normalized_data = data.copy()
13
14
15 for col in numeric_columns:
16     mean = data[col].mean()
17     std = data[col].std()
18     normalized_data[col] = (data[col] - mean) / std
19
20
21 output_file = "Planilha_normalizada.xlsx"
22 normalized_data.to_excel(output_file, index=False)
23
24
25 print(f"Normalização concluída. Arquivo salvo como: {
    output_file}")
```

A sequência de passos executados pelo código para a normalização dos dados foi:

1. **Análise dos Dados:** Os dados foram inicialmente importados do arquivo Excel que possui os dados dos embriões, usando a função `pd.read_excel()`.
2. **Identificação das Colunas Numéricas:** Em seguida, foram identificadas as colunas numéricas do *DataFrame* através do método `select_dtypes()`. Essa etapa é muito importante, já que a normalização com Z-Score deve ser realizada apenas em variáveis contínuas. As variáveis normalizadas foram: Idade, Kidscore, t2, t3, t4, t5, t8, tSC, tSB, tB, cc2 (t3-t2), cc3 (t5-t3), t5-t2, s2 (t4-t3), s3 (t8-t5), tSC-t8 e tB-tSB.
3. **Cálculo da Média e Desvio Padrão:** Para cada variável numérica, a média e o desvio padrão foram calculados. Esses valores são fundamentais para a utilização da fórmula do Z-Score, apresentada e explicada no **APÊNDICE B**.

4. **Utilização do Z-Score:** Para cada valor de uma variável, a média é subtraída e o desvio padrão é dividido, de acordo com a fórmula. Isso modifica os dados para que todas as variáveis apresentem média zero e desvio padrão igual a 1.
5. **Armazenamento dos Dados Normalizados:** Depois de aplicar o Z-Score, os dados normalizados foram guardados em uma nova planilha Excel, denominada “Planilha\_normalizada.xlsx”. A planilha está em “Atividade 3” no GitHub do nosso projeto: [GitHub-TCC](#).

A opção pelo Z-Score se deve à sua habilidade de converter os dados para uma escala padrão, preservando as informações pertinentes. Pesquisas apontam que a normalização melhora a precisão dos algoritmos de aprendizado de máquina ao remover os efeitos de escalas distintas entre as variáveis (JAISWAL, 2024). Além disso, essa técnica simplifica a comparação entre variáveis, auxiliando na interpretação e validação dos modelos criados.

#### 4.1.1.5 Atividade 4 (A4): Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados

Nessa atividade, apresentamos a implementação do procedimento de segmentação do conjunto de dados em três subconjuntos diferentes: treinamento, validação e teste. Esta distribuição obedece às proporções estabelecidas na metodologia (70%, 15% e 15%, respectivamente) e foi feita através de um *script* em Python. Utilizamos a planilha criada pela Atividade 3, a “Planilha\_normalizada.xlsx”, pois utilizaremos os dados já normalizados e tratados a partir de agora.

A segmentação do conjunto de dados foi planejada para assegurar que cada subconjunto tenha uma representação adequada para seus respectivos propósitos, com já citado na Capítulo 3 em Metodologia. A distribuição foi feita de maneira aleatória para prevenir viés e assegurar a aplicabilidade geral do modelo. O *script* calculou o tamanho de cada subconjunto com base em um total 82 linhas de amostra, seguindo as porcentagens estabelecidas.

Abaixo, descrevemos o código implementado, seguido de sua explicação detalhada:

```
1 import pandas as pd
2 import numpy as np
3
4
5 df = pd.read_excel("Planilha_normalizada.xlsx")
6
```

```
7
8 df = df.sample(frac=1, random_state=42).reset_index(drop=True)
9
10
11 n_total = len(df)
12 n_train = int(n_total * 0.7)
13 n_val = int(n_total * 0.15)
14 n_test = n_total - n_train - n_val
15
16
17 train_data = df.iloc[:n_train]
18 val_data = df.iloc[n_train:n_train + n_val]
19 test_data = df.iloc[n_train + n_val:]
20
21
22 train_data.to_excel("dados_treinamento.xlsx", index=False)
23 val_data.to_excel("dados_validacao.xlsx", index=False)
24 test_data.to_excel("dados_teste.xlsx", index=False)
25
26
27 print(f"Conjunto de Treinamento: {len(train_data)} linhas")
28 print(f"Conjunto de Validação: {len(val_data)} linhas")
29 print(f"Conjunto de Teste: {len(test_data)} linhas")
```

- **Importação das Bibliotecas:** A biblioteca `pandas` foi empregada para a manipulação de dados e `numpy` para realizar operações matemáticas.
- **Carregamento dos Dados:** O documento Excel que continha a base de dados dos embriões foi carregado em um *DataFrame* do `pandas` para simplificar as operações subsequentes.
- **Embaralhamento dos Dados:** Utilizamos a função `sample` com os parâmetros (`frac=1` e `random_state=42`) para embaralhar os dados de maneira aleatória, assegurando a remoção de qualquer padrão sequencial existente no conjunto original. O argumento `random_state` garante a reprodutibilidade, possibilitando que o mesmo resultado seja alcançado em futuras execuções.
- **Cálculo das Proporções:** As proporções foram estabelecidas para os conjuntos de treinamento, validação e teste, assegurando que a soma das dimensões corresponda ao número total de amostras. Isso é crucial para prevenir perdas de informações ou desbalanceamento nos conjuntos.

- **Divisão dos Dados:** A escolha dos subconjuntos foi feita com base nos índices do *DataFrame*, que foram divididos em intervalos definidos com base nos tamanhos previamente calculados.
- **Armazenamento dos Dados:** Os subconjuntos resultantes foram salvos em arquivos Excel separados utilizando a função `to_excel`, com o parâmetro `index=False` para evitar a inclusão de índices desnecessários nos arquivos finais. Os arquivos foram nomeados para refletir suas respectivas funções: `dados_treinamento.xlsx`, `dados_validacao.xlsx` e `dados_teste.xlsx`.

O script priorizou a simplicidade, a replicabilidade e a eficácia. Através deste método, garantimos que os conjuntos produzidos sejam apropriados para as fases seguintes de treinamento, validação e teste do modelo de IA, mantendo a integridade dos dados originais.

### Aumento de Dados (*Data Augmentation*) utilizando o Método de Monte Carlo:

A técnica de aumento de dados (*data augmentation*) foi utilizada para melhorar o desempenho do modelo ao gerar mais exemplos de treinamento a partir dos dados existentes. Este processo visa aumentar a capacidade de generalização do modelo e evitar o *overfitting*, tornando-o mais robusto. Abaixo, descrevemos o código implementado, seguido de sua explicação detalhada:

```
1  import pandas as pd
2  import numpy as np
3
4
5  train_data_path = 'dados_treinamento.xlsx'
6  df_train = pd.read_excel(train_data_path)
7
8
9  numerical_columns = df_train.select_dtypes(include=['float64',
10     'int64']).columns
11
12  categorical_columns = df_train.select_dtypes(include=['object',
13     ]).columns
14
15
16  augmentation_factor = 3
17  num_new_samples = len(df_train) * augmentation_factor
```

```
17 new_samples = {}
18 for col in numerical_columns:
19     mean = df_train[col].mean()
20     std = df_train[col].std()
21     new_samples[col] = np.random.normal(loc=mean, scale=std,
22                                         size=num_new_samples)
23
24 df_new_samples = pd.DataFrame(new_samples)
25
26
27 for col in categorical_columns:
28     replicated_values = np.random.choice(df_train[col], size=
29                                         num_new_samples, replace=True)
30     df_new_samples[col] = replicated_values
31
32 df_augmented = pd.concat([df_train, df_new_samples],
33                           ignore_index=True)
34
35 augmented_data_path = 'dados_treinamento_aumentado.xlsx'
36 df_augmented.to_excel(augmented_data_path, index=False)
37
38
39 print(f"Conjunto de dados aumentado salvo em: {
    augmented_data_path}")
```

O programa emprega distribuições estatísticas autênticas para produzir novas amostras que seguem os padrões dos dados originais. A distinção entre dados numéricos e categóricos possibilita o tratamento adequado de cada variável, garantindo a integridade do conjunto de dados ampliado.

- **Importação das Bibliotecas:** As bibliotecas `pandas` e `numpy` foram utilizadas para manipular dados e realizar operações matemáticas: o `pandas` foi empregado para carregar, manipular e armazenar os conjuntos de dados, enquanto o `numpy` foi responsável pela criação de valores aleatórios para as colunas numéricas e pela replicação de valores categóricos.
- **Carregamento do Conjunto de Dados de Treinamento:** O documento com os dados de treinamento foi inserido num *DataFrame*, que atua como uma estrutura para a organização e manipulação eficaz dos dados. A variável `train_data_path`

definiu o caminho do arquivo, enquanto a função `pd.read_excel` foi empregada para acessar os dados.

- **Identificação das Colunas Numéricas e Categóricas:** Foram escolhidas colunas numéricas (como `float64` ou `int64`) para produzir novos valores com base em distribuições estatísticas, enquanto as colunas categóricas (do tipo `object`) para duplicar valores já existentes, preservando as proporções originais. Esta divisão foi feita por meio dos métodos `select_dtypes`, que possibilitaram a identificação e categorização das colunas com base no seu tipo de informação.
- **Definição do Fator de Aumento:** O fator de ampliação foi estipulado em 3, o que sugere que o volume do conjunto de treinamento será triplicado. A quantidade de amostras adicionais foi determinada ao multiplicar o tamanho original do conjunto (`len(df_train)`) pelo fator de ampliação. Isso estabeleceu o número de novos exemplos a serem produzidos para ampliar a base de dados de treinamento.
- **Geração de Novas Amostras para Colunas Numéricas:** Para cada coluna numérica, foram criados novos valores por meio da distribuição normal. Os parâmetros utilizados foram o valor médio (`mean`) e o desvio padrão (`std`) da coluna original, assegurando que os novos valores mantenham as propriedades estatísticas da distribuição original. Este procedimento garante que os dados expandidos para as colunas numéricas mantenham a mesma distribuição da amostra inicial.
- **Replicação de Valores para Colunas Categóricas:** Os valores já existentes nas colunas categóricas foram replicados de maneira aleatória por meio da função `np.random.choice`, com reposição. Esta estratégia mantém a proporção original dos valores categóricos, assegurando a uniformidade nos padrões e possibilitando a manutenção da distribuição das categorias nos novos exemplos criados.
- **Combinação dos Dados Originais com os Novos:** A meta era unir os dados originais aos novos dados produzidos, resultando em um conjunto de treinamento expandido. Com o parâmetro `ignore_index=True`, a função `pd.concat` foi empregada para recomençar os índices no novo conjunto de dados, assegurando que o *DataFrame* gerado apresentasse uma sequência ininterrupta de índices, sem duplicações ou sobreposições.
- **Armazenamento do Conjunto de Dados Ampliado:** O arquivo Excel com os dados ampliados foi nomeado como `dados_treinamento_aumentado.xlsx`. O parâmetro `index=False` foi empregado para prevenir a criação de índices nos arquivos finais, facilitando sua utilização em fases subsequentes do trabalho e assegurando que o arquivo gerado contenha apenas os dados, sem colunas de índice adicionais.



A aplicação da média e do desvio padrão assegura que os novos dados numéricos correspondam às características do conjunto original, ao mesmo tempo que as categorias mantêm suas proporções, mantendo os padrões originais. A produção de dados por meio de distribuições estatísticas amplia a diversidade do conjunto sem adicionar ruído excessivo ou padrões inverossímeis, assegurando uma diversificação controlada. Este procedimento garante que o conjunto de treinamento ampliado seja robusto o suficiente para treinar o modelo, prevenindo sobrecargas e aprimorando a generalização, levando a um modelo mais eficiente e seguro.

#### 4.1.1.6 Nota sobre a Apresentação dos Códigos

Os códigos expostos no Capítulo 4 foram esclarecidos de maneira clara e direta, visando tornar eles compreensíveis para todos os leitores, independentemente do grau de habilidade técnica. A estratégia selecionada tem como objetivo assegurar que até mesmo aqueles sem formação em Tecnologia da Informação (TI) possam entender os princípios das implementações feitas.

Os leitores que desejam uma explicação mais minuciosa ou um nível técnico mais avançado podem encontrar os códigos com explicação completa no repositório do projeto no GitHub. O repositório pode ser acessado através do link a seguir: [GitHub do TCC](#).

## 5 Análise dos Resultados

### 5.1 Fase 1: Análise e Preparação de Dados

#### 5.1.1 OE1 - Expansão, Processamento e Análise de Dados para Predição de Ploidia

##### 5.1.1.1 Atividade 1 (A1): Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia

Na fase inicial da tarefa (A1), realizamos uma análise, revisão e escolha das variáveis para o nosso modelo de IA. Através do estudo bibliográfico, conseguimos verificar um grupo de variáveis da planilha de dados que possui fundamentação científica que evidenciam sua importância para a evolução do modelo. As variáveis selecionadas para o nosso estudo serão: **Idade, t2, t3, t4, t5, t8, s2, cc2, tSC, tSB, tB, cc3 (t5-t3), s3 (t8-t5), t5-t2, tSC-t8, tB-tSB, Estágio, Morfo e KIDScore.**

A coluna de **Plodia** também foi selecionada, pois nos possibilita agrupar os embriões em duas categorias claras, distinguindo entre aqueles com euploidia normal e aqueles com alterações cromossômicas, o que é crucial para a elaboração de um modelo sólido.

Não identificamos estudos que estabelecem uma ligação direta entre o parâmetro **st2** e a previsão de euploidia. Apesar do movimento citoplasmático antes da citocinese ser um marco significativo no desenvolvimento embrionário, a ausência de provas científicas que liguem esse movimento à qualidade do embrião e à euploidia nos levou a remover o **st2** da lista de variáveis para o modelo. Igualmente, não foram identificados estudos que analisassem especificamente o intervalo entre **t2** (o instante em que o embrião alcança a fase de duas células) e **st2** (movimento citoplasmático pré-citocinese) para prever a euploidia. Como **st2** foi eliminada, também removemos o parâmetro **t2-st2** do nosso grupo de variáveis

A partir do estudo bibliográfico minucioso realizado na seção de **Atividade 1 do Capítulo 4**, conseguimos determinar quais variáveis são fundamentais para a elaboração do nosso modelo de previsão de euploidia. Depois de examinar e revisar as variáveis, modificamos a planilha para espelhar os dados mais significativos para o modelo, que se encontra no Anexo ?? como "**Planilha Revisada - Colunas Excluídas**". Portanto, as colunas **Id, Data da biópsia e Embrião n.** foram eliminadas, uma vez que não contribuem para o valor do modelo. Adicionalmente, as variáveis **st2** e **t2-st2** foram eliminadas, conforme mencionado anteriormente. Portanto, a planilha revisada agora inclui somente as variáveis que possuem uma ligação comprovada com a previsão de

euploidia, fundamentada nas evidências científicas revisadas.

Ao tratar de **dados ausentes** em conjuntos de dados, utilizando o método Análise de Casos Completos (ACC), possuíamos 84 linhas de dados, das quais apenas 2 têm dados ausentes. Na "**Planilha com Indicação de Dados Ausentes**" (Anexo ??), os campos que não possuem dados estão em preto. Por esse motivo, excluimos essas duas linhas de modo manual, já que é um número muito pequeno para fazer um código de limpeza de dados, assim, a planilha se modifica se tornando a "**Planilha de Dados Refinados**" presente no Anexo ??, resultando em 82 linhas, que está nos anexos.

#### 5.1.1.2 Atividade 2 (A2): Identificação da Correlação entre os Parâmetros na Previsão da Ploidia do Embrião

A análise de correlação foi realizada com o objetivo de identificar as interações mais significativas entre os parâmetros relacionados ao desenvolvimento embrionário e a ploidia. Para isso, foi utilizado o coeficiente de Spearman, explicado no APÊNDICE D. A análise das relações entre as variáveis possibilitou a identificação das variáveis que se afetam e afetam a porcentagem de euploidia, fornecendo um alicerce para a melhoria de processos e tomada de decisões da IA. A consolidação dos resultados resultou em dois documentos: um arquivo Excel (*correlation\_results.xlsx*) que apresenta uma visão aprofundada dos coeficientes calculados, transformado em um PDF mais compacto para facilitar a visualização, disponível nos anexos, e um documento Word (*relatorio\_correlacoes.docx*) que contém todos os gráficos de dispersão para as combinações examinadas e valores de correlação relacionados, totalizando 172 páginas. Caso haja interesse em consultar o documento completo, siga os passos descritos na "Atividade 2" no GitHub do projeto: [GitHub-TCC](#). Executando o código, é possível gerar o documento completo.

#### correlation\_results

A função `spearman` da biblioteca `SciPy` foi empregada para gerar o coeficiente de Spearman no código, calculando tanto o coeficiente de correlação quanto o valor-p para cada par de variáveis. Inicialmente, o código importou as informações do arquivo "*Planilha de Dados Refinados*" presente no Anexo ??, reconheceu as colunas numéricas pertinentes e iterou sobre todas as combinações possíveis de variáveis, assegurando que uma variável não fosse comparada a si mesma.

Os resultados gerados na planilha *correlation\_results* incluem três colunas:

- **Variable 1:** A primeira variável da análise.
- **Variable 2:** A segunda variável correlacionada.
- **Spearman Coefficient:** O coeficiente calculado para o par de variáveis.

Os coeficientes de correlação de Spearman, foram calculados a partir de todas as combinações possíveis de variáveis escolhidas da base de dados, que está presente de forma mais enxuta nos anexos. Cada linha identifica um par de variáveis examinadas, com os seus resultados apontando a intensidade e a direção (positiva ou negativa) da correlação entre elas. Ressaltando que, uma correlação positiva perfeita é representada por um valor de +1 que indica que o crescimento de uma variável está sempre ligado ao crescimento da outra. Por outro lado, um valor de -1 indica uma correlação negativa perfeita, em que o aumento de uma variável corresponde consistentemente à diminuição da outra. Quando o coeficiente é zero, indica que as variáveis não possuem uma relação monotônica evidente, isto é, não existe um padrão consistente de variação conjunta.

## relatorio\_correlacoes

O documento *relatorio\_correlacoes.docx* foi gerado automaticamente usando a biblioteca `python-docx`, que permite a criação e manipulação de documentos Word. Depois de obter os coeficientes de Spearman para cada par de variáveis como descrito e apresentado no documento “*correlation\_results*” explicado acima, o código itera sobre essas combinações.

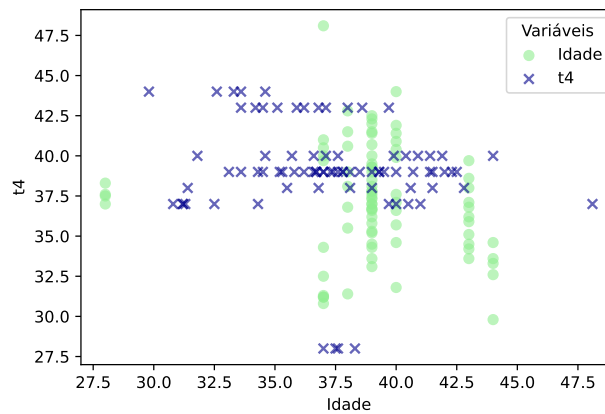
Este documento fornece uma avaliação minuciosa das correlações, com os gráficos de dispersão para cada par de variáveis e os valores dos coeficientes de Spearman, em que esses valores estão escritos com seu sinal de positivo ou negativo, seguido de duas casas decimais. O documento é necessário para entender padrões particulares e enfatizar relações relevantes de forma visual.

Em seguida, apresentaremos a avaliação dos resultados mais relevantes obtidos a partir dos dois documentos produzidos pelo código, ressaltando as variáveis que possuem ligações mais sólidas entre si, assim como as que possuem pouca ou nenhuma interação recíproca. Esses resultados serão comparados com as referências bibliográficas anteriormente discutidas.

## Idade

A correlação com **t4 (-0,15)** na **Figura 4** sugere que a **idade exerce uma influência leve sobre eventos específicos do desenvolvimento embrionário**. Mulheres mais velhas podem apresentar embriões com ligeiro atraso no tempo necessário para atingir o estágio *t4*. Embora a influência da idade no ritmo inicial de desenvolvimento seja limitada, há uma leve tendência de atraso.

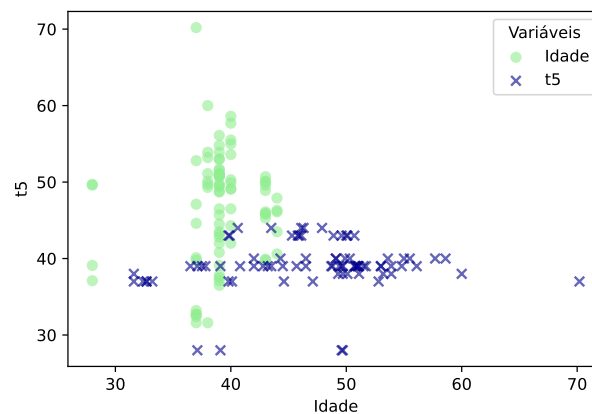
Figura 4. Dispersão entre Idade e t4 - Coeficiente de Spearman: -0.15



Fonte: Autoras (2025)

Em relação ao tempo para 5 células,  $t5$  (0.11), obteve uma correlação positiva. Isso sugere uma tendência muito sutil de que embriões em fases mais avançadas estejam ligados a mães de mais idade, como observado na Figura 5.

Figura 5. Dispersão entre Idade e t5 - Coeficiente de Spearman: 0,11

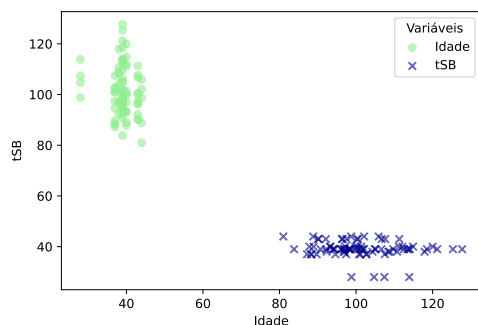


Fonte: Autoras (2025)

Ao reparar em alguns índices negativos, observamos os índices **tSB** (-0.10), **cc2** (**t3-t2**) (-0.17), **s2** (**t4-t3**) (-0.24), **s3** (**t8-t5**) (-0.28) e a **Ploidia** (-0.50). Nota-se que os coeficientes mais próximos de 0 (como -0,10 a -0,28) indicam uma correlação negativa fraca. Isso traduz que há uma tendência muito sutil de que, quando uma variável aumenta, a idade, a outra diminui. Em **tSB** Figura 6, o coeficiente negativo insinua que o tempo de formação inicial da blastocisto tende a ser menor em embriões provenientes de mães mais velhas. Na variável **cc2**, Figura 7, a correlação desfavorável sugere uma maior irregularidade no intervalo entre a segunda e a terceira divisão celular ( $t2$  para  $t3$ ) em embriões de mães mais velhas. Em **s2** (Figura 8) e **s3** (Figura 9), mostra que a idade materna também está ligada a uma diminuição na eficácia do intervalo entre

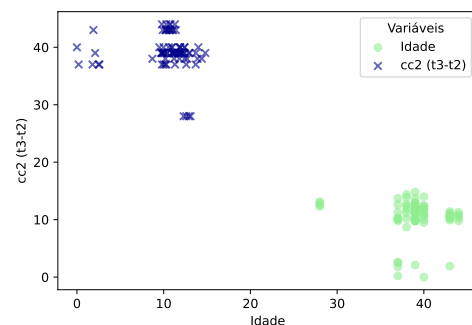
as divisões celulares de t3 para t4. Isso indica um efeito na fase intermediária do ciclo celular. Nos embriões de mães mais velhas, o período entre a fase de 8 células e a formação do blastocisto final é estendido, sinalizando obstáculos no progresso dessas fases. Todos esses atrasos podem ser cruciais, já que fases iniciais bem sincronizadas são fundamentais para um desenvolvimento embrionário adequado, mostrando como uma idade avançada pode afetar o desenvolvimento embrionário, afirmando a bibliografia estudada, fato já citado por ??), que reitera que **a idade materna exerce maior influência sobre a qualidade embrionária.**

Figura 6. Dispersão entre Idade e tSB -  
Coeficiente de Spearman: -0.10



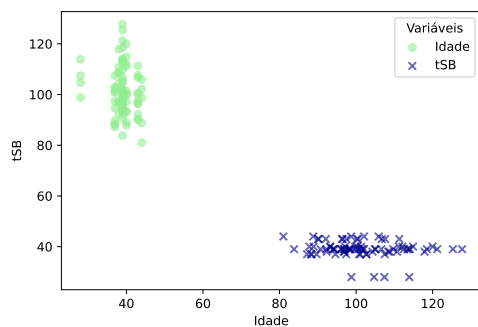
Fonte: Autoras (2025)

Figura 7. Dispersão entre Idade e cc2 (t3-t2) -  
Coeficiente de Spearman: -0.15



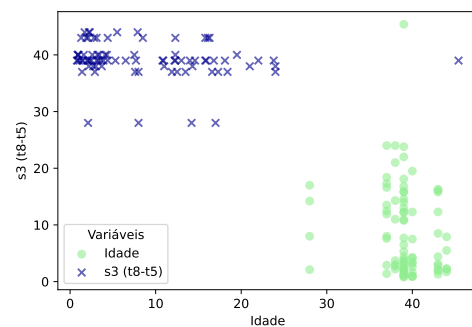
Fonte: Autoras (2025)

Figura 8. Dispersão entre Idade e s2 (t4-t3) -  
Coeficiente de Spearman: -0.24



Fonte: Autoras (2025)

Figura 9. Dispersão entre Idade e s3 (t8-t5) -  
Coeficiente de Spearman: -0.28

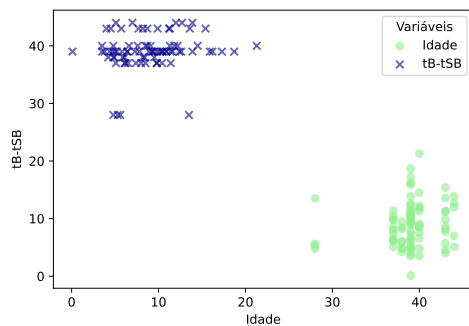


Fonte: Autoras (2025)

Analisando os índices positivos, temos **tB-tSB (0.20)** e **cc3 (t5-t3) (0.20)** com valores que sugerem que a elevação de uma variável está de forma sutil ligada ao crescimento da outra. Em relação ao intervalo entre o estágio de **cc3** (Figura 11) e **tB-tSB** (Figura 10) nos embriões de mulheres mais velhas aumenta levemente. Este crescimento pode sugerir que, mesmo com atrasos em etapas posteriores, o embrião busca se ajustar para compensar o desenvolvimento inicial mais lento.

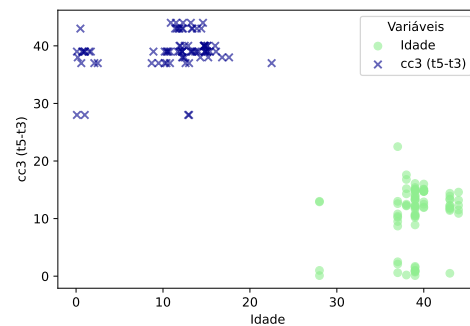
No gráfico de dispersão entre as variáveis Idade e tSB, se nota que o eixo Y, que simboliza a variável tSB, alcança valores próximos a 120. Esta característica está ligada à origem dos dados, onde o tSB varia de 81 a 127,7, enquanto a Idade se situa em uma escala mais limitada (28 a 44 anos). A diferença visual no gráfico é previsível e espelha fielmente os valores reais contidos na planilha, sem sinalizar qualquer erro.

Figura 10. Dispersão entre Idade e tB-tSB -  
Coeficiente de Spearman: 0.20



Fonte: Autoras (2025)

Figura 11. Dispersão entre Idade e cc3 (t5-t3)  
- Coeficiente de Spearman: 0.20



Fonte: Autoras (2025)

E por fim temos a *Ploidia*  $(-0,50)$ , a correlação negativa mais relevante entre todas as outras variáveis, que sugere uma relação negativa, demonstrando que **a alta idade materna está ligada a uma diminuição na proporção de embriões euploides. Esta informação indica que embriões de mães mais velhas contêm uma proporção reduzida de euploidia, o que pode estar ligado a uma queda na qualidade genética dos embriões.** Portanto, a idade materna é um dos elementos chave na diminuição da euploidia embrionária, e conseguimos comprovar esse dado juntamente com as informações que possui na literatura, como citado por *Fertility and Ageing* (BAIRD et al., 2005), que também reitera que o aumento da aneuploidia em embriões está diretamente associado ao envelhecimento materno.

## Estágio

Os coeficientes de correlação de Spearman entre os *estágios* e os tempos de transição celular ( $t2$ ,  $t3$ ,  $t4$ ,  $t5$ ,  $t8$ ) revelam relações positivas moderadas. Esses dados indicam que, **conforme o embrião progride para estágios mais avançados, os tempos associados às divisões celulares iniciais tendem a aumentar de forma sutil.** Por exemplo,  $t2$   $(0,30)$ ,  $t3$   $(0,25)$  e  $t4$   $(0,32)$  apresentam as correlações mais significativas, evidenciando que o avanço no estágio está vinculado a uma maior duração das transições celulares iniciais. Por outro lado,  $t5$   $(0,15)$  e  $t8$   $(0,17)$ , apesar de apresentarem correlações mais baixas, também corroboram essa tendência de relação positiva. Esses resultados sugerem que o progresso do estágio embrionário está associado a um padrão

de desenvolvimento, possivelmente, mais metódico nos estágios iniciais e intermediários do ciclo celular.

As correlações positivas observadas entre o *estágio* embrionário e os tempos  $tSC$  (0,50),  $tSB$  (0,53) e  $tB$  (0,59) indicam uma conexão significativa entre o progresso do estágio embrionário e o desenvolvimento contínuo das estruturas celulares. O incremento nos índices de correlação sugere que, conforme **o embrião avança para fases mais desenvolvidas, há uma maior regularidade no cumprimento dos marcos temporais dessas transições**. Isso implica que o estágio não apenas representa uma condição de desenvolvimento estrutural, mas também abriga informações significativas sobre a dinâmica temporal do ciclo celular. Assim, essas correlações ressaltam que o estágio embrionário atua como um indicador abrangente da qualidade e da evolução do desenvolvimento embrionário.

A relação inversa entre o *estágio* de desenvolvimento e a *ploidia* (-0,24) indica que, **embora essa correlação seja considerada tênue, sugere que o estágio de desenvolvimento pode ter um impacto negativo na ploidia**, enfatizando a necessidade de avaliar não apenas a morfologia do embrião, mas também sua genética ao longo do ciclo celular.

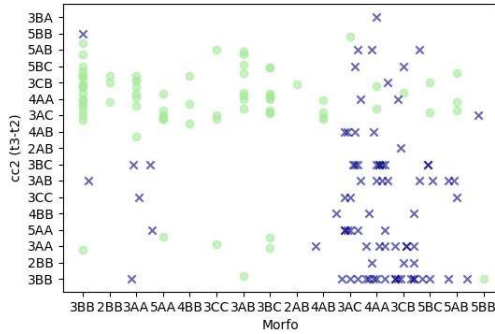
## Morfo

A avaliação da variável *Morfo*, que categoriza a expansão morfológica dos embriões em fases iniciais, revelou correlações moderadamente negativas com variáveis temporais como **t2 (-0.38)**, **t3 (-0.43)** e **t5 (-0.45)**. Esses achados sugerem que **embriões com características morfológicas menos expansivas costumam apresentar atrasos nos primeiros tempos de divisão celular**. Assim, mudanças na morfologia podem ser influenciadas por variações no ritmo de desenvolvimento celular. Dessa forma, conseguimos evidenciar a fala de [Capalbo et al. \(2014\)](#), que cita que a morfologia do embrião é um fator determinante para o potencial de implantação e a qualidade embrionária.

Ao examinarmos os intervalos  $cc2$  ( $t3-t2$ ) (-0,38) e  $cc3$  ( $t5-t3$ ) (-0,31), os gráficos 12 e 13, respectivamente, corroboram essa tendência, destacando que embriões que apresentam alterações morfológicas menos favoráveis quando enfrentam flutuações no ritmo de desenvolvimento celular, a medida que uma variável aumenta, a outra vem a diminuir. Esses resultados enfatizam a relevância de uma divisão celular sincronizada para preservar características morfológicas ideais, evidenciando como o ritmo do ciclo celular influencia diretamente a qualidade estrutural do embrião. A variável “Morfo” está representada pelo marcador “o” em azul escuro e as outras variáveis “cc2” e “cc3”, estão identificadas com o marcador “x” em verde claro.

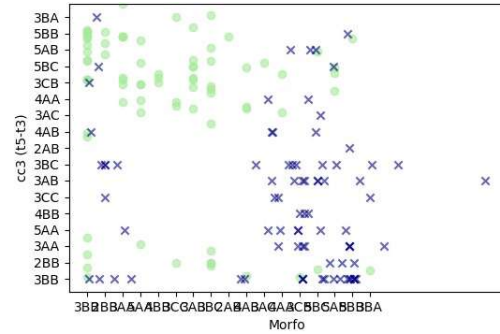


Figura 12. Dispersão entre Morfo e cc2 (t3-t2)  
- Coeficiente de Spearman: -0.38



Fonte: Autoras (2025)

Figura 13. Dispersão entre Morfo e cc3 (t5-t3)  
- Coeficiente de Spearman: -0.31



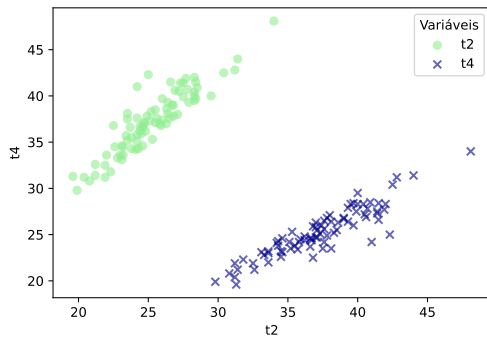
Fonte: Autoras (2025)

A correlação entre *Morfo* e *Ploidia* exibe um coeficiente de correlação de  $0.05$ , sugerindo uma ligação positiva, porém extremamente fraca. Isso sugere que, neste conjunto de dados, as características morfológicas dos embriões não parecem estar diretamente ligadas à euploidia.

## t2, t3, t4, t5 e t8

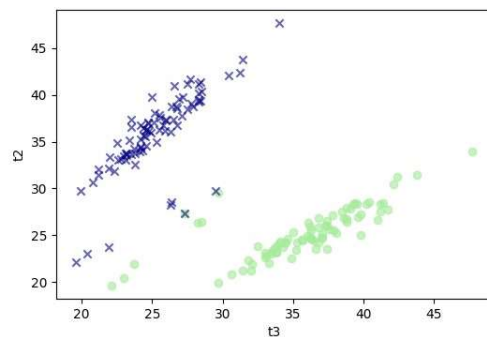
Existe uma forte interdependência entre os tempos de transição celular, como evidenciado pela correlação de  $0.89$  entre  $t2$  e  $t4$ , indicando que esses eventos de desenvolvimento estão fortemente associados, à medida que uma variável aumenta, a outra também aumenta de forma consistente. A correlação entre  $t3$  e  $t2$  ( $0.78$ ),  $t4$  e  $t5$  ( $0.56$ ) e entre  $t5$  e  $t8$  ( $0.52$ ) também demonstra alinhamento nas fases iniciais e intermediárias do desenvolvimento embrionário.

Figura 14. Dispersão entre t2 e t4 -  
Coeficiente de Spearman: 0.89



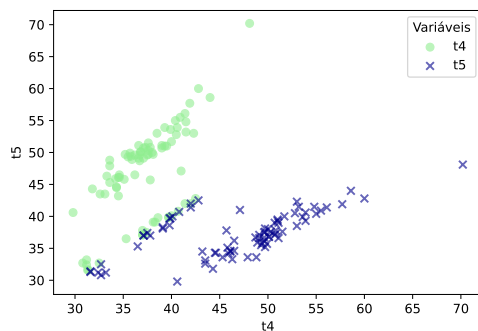
Fonte: Autoras (2025)

Figura 15. Dispersão entre t3 e t2 -  
Coeficiente de Spearman: 0.78



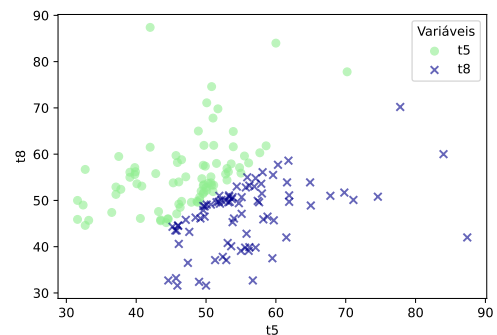
Fonte: Autoras (2025)

Figura 16. Dispersão entre  $t_4$  e  $t_5$  -  
Coeficiente de Spearman: 0.56



Fonte: Autoras (2025)

Figura 17. Dispersão entre  $t_5$  e  $t_8$  -  
Coeficiente de Spearman: 0.52



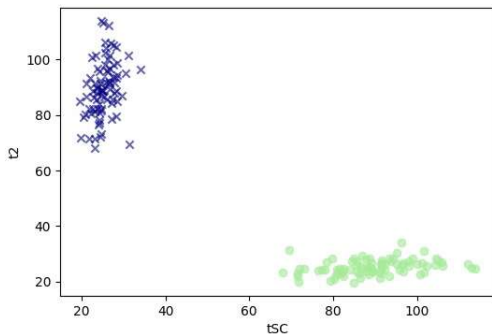
Fonte: Autoras (2025)

A correlação quase nula entre  $t_2$  ( $-0.08$ ),  $t_3$  ( $-0.06$ ),  $t_4$  ( $-0.02$ ),  $t_8$  ( $-0.07$ ) e a *ploidia*, demonstram uma falta de correlação entre essas variáveis para a determinação de um embrião euploidia ou aneuplóide. Já  **$t_5$  e *ploidia* ( $-0.24$ )** indica que atrasos nesse estágio de desenvolvimento, estão ligados a uma qualidade genética inferior (menor *ploidia*). Algo que já foi evidenciado por Cruz et al. (2012), pois destacou  $t_5$  como o indicador mais relevante do potencial de implantação.

## tSC

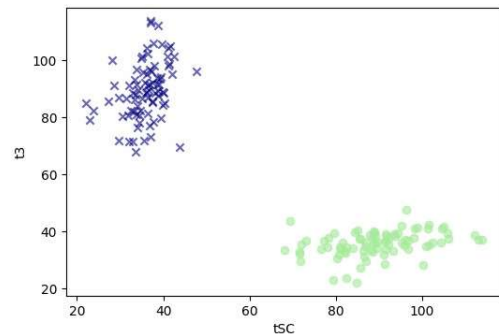
Ao relacionar a variável de Tempo de formação do estágio de clivagem sincronizada (Time to Synchronized Compaction) com as variáveis de tempos de transição celular,  $t_2$  ( $0.40$ ),  $t_3$  ( $0.42$ ),  $t_4$  ( $0.43$ ),  $t_5$  ( $0.35$ ) e  $t_8$  ( $0.35$ ), indicam uma relação de intensidade moderada e positiva durante o desenvolvimento embrionário, sendo mais acentuada nos estágios iniciais e intermediários. A conexão com  $t_2$ , Figura 18, sugere uma interdependência inicial que persiste em  $t_3$ , Figura 19, e se intensifica um pouco em  $t_4$ , Figura 20, indicando um alinhamento mais intenso nas etapas intermediárias do ciclo celular. As correlações com  $t_5$  e  $t_8$ , Figura 21, apresentam uma ligeira redução, sugerindo que o impacto do *tSC* nos eventos embrionários começa a se desvanecer em fases mais avançadas. Esses padrões indicam que o **tSC tem um papel mais significativo nas etapas iniciais e intermediárias do desenvolvimento, diminuindo seu impacto progressivamente conforme o tempo passa**. Os gráficos abaixo mostram o retrato da explicação, em que *tSC* está identificado pelo marcador “o” em verde claro. Nota-se que, à medida que uma variável cresce, a outra cresce de forma conjunta.

Figura 18. Dispersão entre tSC e t2 -  
Coeficiente de Spearman: 0.40



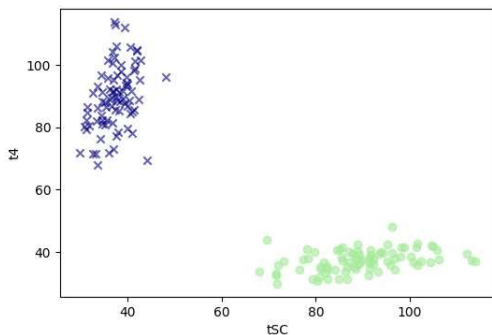
Fonte: Autoras (2025)

Figura 19. Dispersão entre tSC e t3 -  
Coeficiente de Spearman: 0.42



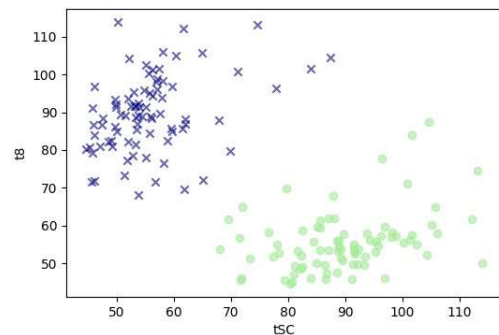
Fonte: Autoras (2025)

Figura 20. Dispersão entre tSC e t4 -  
Coeficiente de Spearman: 0.43



Fonte: Autoras (2025)

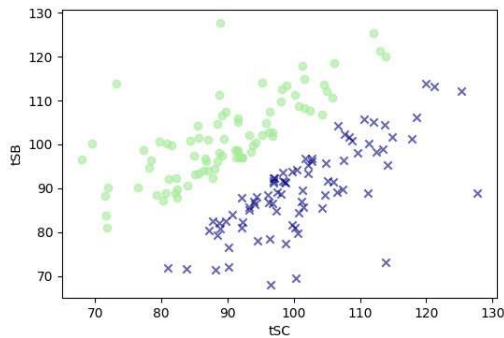
Figura 21. Dispersão entre tSC e t8 -  
Coeficiente de Spearman: 0.35



Fonte: Autoras (2025)

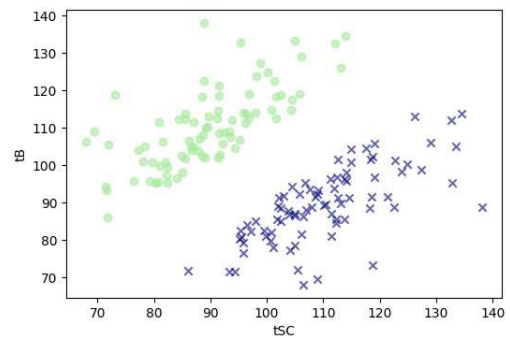
As correlações de  $tSC$  com  $tSB$  ( $0.75$ ) e  $tB$  ( $0.74$ ), Figura 22 e Figura 23 respectivamente, demonstram uma forte ligação entre essas variáveis, indicando que estão fortemente conectadas, em que a medida que uma variável aumenta, a outra também aumenta de forma consistente. Isso indica uma relação quase linear, em que mudanças no  $tSB$  impactam diretamente no  $tSC$ , evidenciando uma ligação funcional direta entre os dois acontecimentos. A variável  $tSC$  permanece sendo identificada pelo marcador “o” nos gráficos a seguir.

Figura 22. Dispersão entre tSC e tSB -  
Coeficiente de Spearman: 0.75



Fonte: Autoras (2025)

Figura 23. Dispersão entre tSC e tB -  
Coeficiente de Spearman: 0.74



Fonte: Autoras (2025)

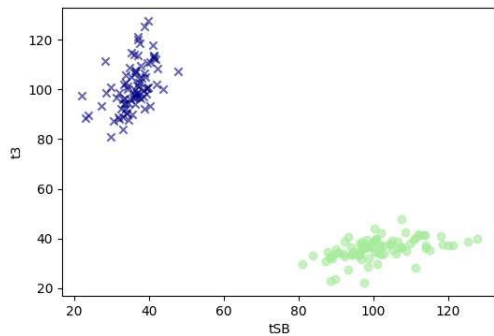
Os coeficientes de correlação de tSC com os intervalos de tempo *cc2* ( $t3-t2$ ) (0.41), *cc3* ( $t5-t3$ ) (0.24) e *t5-t2* (0.34) apresentam relações positivas que variam de moderadas a fracas, sinalizando variados níveis de concordância com esses tempos acumulados. A correlação moderada com *cc2* ( $t3-t2$ ) indica que o intervalo entre *t3* e *t2* tem uma influência moderada no tSC, evidenciando uma ligação mais clara neste estágio. Em contrapartida, a correlação fraca com *cc3* ( $t5-t3$ ) sugere que as alterações no intervalo entre **t5 e t3** exercem uma influência restrita sobre o tSC. Por outro lado, a relação moderada com *t5-t2* indica que o tempo acumulado entre esses eventos afeta o tSC, embora em menor grau do que com *cc2*. Estes achados indicam que a influência do tSC se **reduz** conforme os períodos de tempo se estendem.

Por fim, a correlação com a *Ploidia* (-0.04), é fraca e negativa, sugerindo que não existe uma conexão relevante entre essas variáveis. Este resultado indica que *Ploidia* pode estar ligada a processos diferentes que não afetam diretamente *tSC*.

## tSB

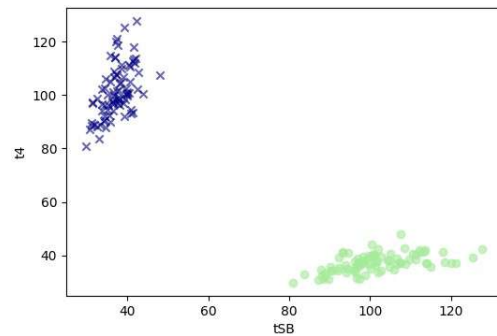
Ao analisar sua realção com *t2* (0.43), *t3* (0.56), *t4* (0.57), *t5* (0.39) e *t8* (0.40) mostram relações positivas de intensidade moderada, conforme uma variável aumenta, a outra segue um aumento consistente, aumentando nos estágios iniciais e intermediários e apresentando um ligeiro declínio nos estágios posteriores. A relação com *t2* **indica que os eventos iniciais exercem um efeito significativo no comportamento de tSB**. A forte ligação com *t3* e *t4*, Figura 24 e Figura 25 respectivamente, indica que **esses estágios intermediários têm uma influência mais significativa sobre tSB**. Em contrapartida, as correlações moderadas com **t5 e t8** sugerem uma redução na conexão nos estágios subsequentes, evidenciando um efeito menos significativo de *tSB* conforme os processos progridem no desenvolvimento. Nos gráficos a seguir, tSB está definido em cor verde claro.

Figura 24. Dispersão entre  $tSB$  e  $t3$  -  
Coeficiente de Spearman: 0.56



Fonte: Autoras (2025)

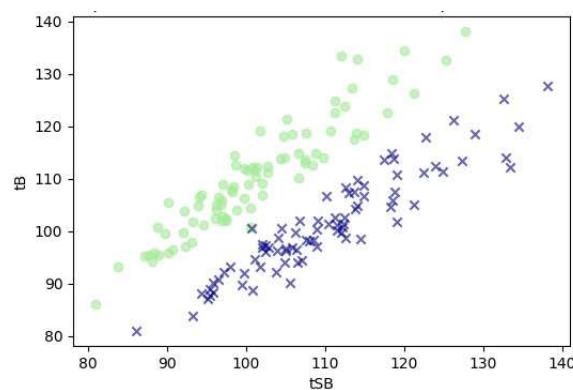
Figura 25. Dispersão entre  $tSB$  e  $t4$  -  
Coeficiente de Spearman: 0.57



Fonte: Autoras (2025)

As correlações entre  $tSB$  e  $tSC$  (0.75),  $tB$  (0.93) e  $cc2$  ( $t3-t2$ ) (0.63) evidenciam ligações fortes e relevantes, indicando uma interdependência entre essas variáveis. A forte ligação com  $tSC$  sugere que alterações em uma variável têm um **impacto direto** na outra. **A correlação com  $tB$  indica uma dependência quase total entre as duas variáveis**, indicando que são praticamente comparáveis em termos de sua interação. Ademais, a correlação robusta com  $cc2$  ( $t3-t2$ ) indica que **o período entre  $t3$  e  $t2$  tem um impacto considerável sobre  $tSB$ , possivelmente por causa do efeito direto dos eventos temporais iniciais no comportamento dessa variável**. A Figura 26 ilustra a forma que  $tSB$  e  $tB$  possuem uma relação de crescimento mútuo entre si, quase se espelhando. Dessa forma, alterações em  $tB$  podem ser significativas para mudanças em  $tSB$ .

Figura 26. Dispersão entre  $tSB$  e  $tB$  - Coeficiente de Spearman: 0.93



Fonte: Autoras (2025)

O coeficiente de correlação de  $tSB$  com  $Ploidia$  (-0.11) é fraco e negativo, sugerindo que não existe uma conexão evidente ou relevante entre essas variáveis. Este resultado indica que  $Ploidia$  é afetada por elementos diferentes, que não estão diretamente ligados aos processos ligados à  $tSB$ .

## tB

As correlações entre *tB* e os tempos de desenvolvimento *t2* (0.40), *t3* (0.48), *t4* (0.50), *t5* (0.35) e *t8* (0.37) exibem relações moderadamente positivas, sugerindo que *tB* está ligado a eventos temporais durante o ciclo embrionário, sendo mais intensa nos estágios intermediários. A forte ligação com *t3* e *t4* sugere que *tB* está mais em sintonia com eventos desses estágios, o que pode indicar uma interação mais intensa com processos de divisão celular intermediários. O crescimento da correlação em *t4* pode indicar uma etapa crucial para o desenvolvimento do embrião. Por outro lado, as correlações mais baixas com *t5* e *t8* indicam uma redução da influência do *tB* em fases mais avançadas, ressaltando que sua importância é maior nos estágios iniciais e intermediários do ciclo celular.

As correlações de *tB* com os intervalos de tempo *cc2* (*t3-t2*) (0.54) e *tSC-t8* (0.47) ressaltam a influência de processos particulares no comportamento dessa variável. A correlação moderada-alta com *cc2* (*t3-t2*) indica que **os períodos de transição entre t3 e t2 têm uma influência relevante no comportamento de tB**, sugerindo que as dinâmicas nos estágios iniciais influenciam diretamente o desenvolvimento subsequente. Em contrapartida, a correlação moderada entre *tSC* e *t8* indica que a diferença temporal entre **tSC e t8** está positivamente ligada a *tB*.

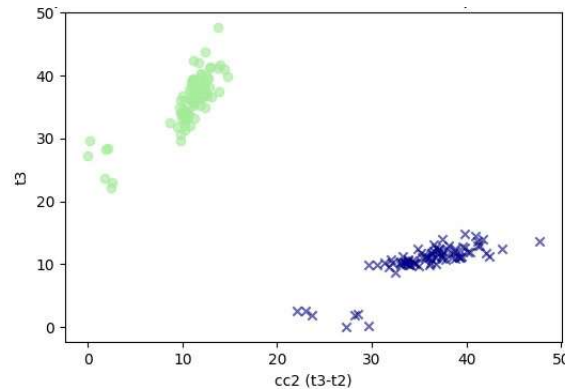
A fraca e negativa correlação entre *tB* e *Ploidia* (-0.17) indica uma conexão fraca entre essas variáveis. A direção negativa sugere que aumentos em *tB* podem estar ligados a pequenas diminuições em *Ploidia*. Esta associação tênue enfatiza que *tB* não desempenha um papel crucial no comportamento da *Ploidia*, porém o afeta fracamente de forma negativa.

## cc2 (t3-t2)

A relação entre o intervalo *cc2* e as diversas fases de crescimento do embrião, simbolizadas pelos tempos *t2*, *t3*, *t4*, *t5* e *t8*, mostra variações de intensidade. A correlação com *t2* (0.39) é moderada e positiva, indicando que acontecimentos iniciais, como o surgimento das primeiras células embrionárias, possuem uma ligação direta, porém não tão relevante com *cc2*. Com *t3* (0.80), se nota, na figura 27, uma correlação positiva significativa, demonstrando um **alinhamento quase imediato entre cc2** e o momento de desenvolvimento correspondente a *t3*, sugerindo uma **conexão robusta entre ambos**. Por outro lado, a correlação com **t4 (0.59)** é moderadamente alta, indicando que **cc2 é também impactado por eventos subsequentes representados por t4**. A conexão com *t5* (0.64), também moderada-alta, **evidencia a persistência da influência de eventos temporais**, evidenciando que *t5* desempenha um papel significativo na progressão de *cc2*. Por fim, com *t8* (0.46), a correlação é moderada, porém o efeito de *t8* sobre *cc2* é menos significativo do que nos estágios anteriores, **sugerindo uma dependência**

que diminui conforme o embrião progride para fases mais avançadas de desenvolvimento. O gráfico abaixo mostra a maior relação que cc2 possui durante as fases de crescimento, que é na fase **t3 (0.80)**, em que a variável cc2, em verde, cresce de forma consistente à medida que a outra também cresce.

Figura 27. Dispersão entre cc2 (t3-t2) e t3 - Coeficiente de Spearman: 0.80



Fonte: Autoras (2025)

A correlação com *cc3* (0.34) é moderada, sugerindo uma relação restrita entre cc2 e cc3. Por outro lado, a correlação com *t5-t2* (0.65) é forte e positiva, indicando que o intervalo *cc2* se alinha fortemente ao intervalo completo entre *t5* e *t2*. Isso indica que o comportamento de cc2 pode ser significativamente elucidado pela combinação das variáveis que compõem esses intervalos de tempo, evidenciando uma relação mais sólida entre eles.

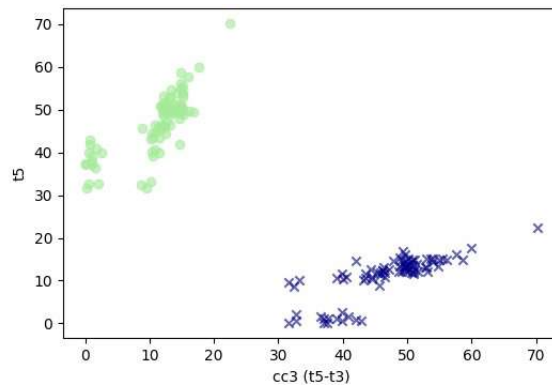
A correlação entre *cc2* e *Ploidia* (-0.03) é muito fraca e negativa, sugerindo que não existe uma conexão relevante entre esses dois componentes. Isso indica que **o período temporal entre t3 e t2 não tem um impacto significativo na Ploidia**, o que sugere que o desenvolvimento temporal do embrião, simbolizado por cc2, não está diretamente ligado à ploidia do embrião.

### cc3 (t5-t3)

A correlação com *t2* (0.15) é fraca e positiva, sugerindo que pequenas alterações em t2 não provocam alterações significativas em cc3. Com *t3* (0.25), a correlação é moderada e positiva, mostrando um ligeiro crescimento em cc3 à medida que t3 se eleva. A correlação com *t4* (0.26) é igualmente moderada e positiva. A conexão entre t4 e cc3 é parecida com a de t3, porém um pouco mais intensa, indicando que t4 desempenha um papel moderado nas alterações percebidas em cc3. A correlação com *t5* (0.81), presente na figura 28, é extremamente forte e positiva, sugerindo que **conforme t5 cresce, cc3 também cresce quase de maneira linear**, como se consegue observar no gráfico abaixo, em que cc3 está definido pelo marcador “o”. Isso é coerente, já que cc3 é determinado como a diferença entre t5 e t3, e qualquer alteração em t5 afeta diretamente cc3. Por fim, a correlação com

o **t8 (0.40)** é moderada e positiva, indicando que, conforme o t8 se eleva, o cc3 tende a crescer. O t8 tem um impacto maior sobre o cc3 do que o t2, t3, e t4, mas ainda distante do impacto de t5.

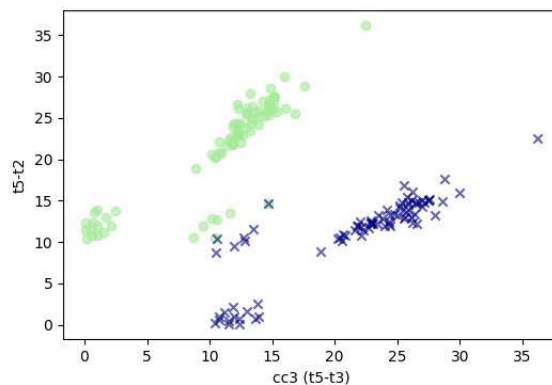
Figura 28. Dispersão entre cc3 (t5-t3) e t5 - Coeficiente de Spearman: 0.81



Fonte: Autoras (2025)

A correlação entre o intervalo *t5-t3* (*representado por cc3*) e o intervalo *t5-t2* (*0.90*) é extremamente forte e positiva como pode ser visto na figura 29, **sugerindo que o comportamento de cc3 está fortemente ligado ao intervalo t5-t2**. Isso corrobora a noção de que as alterações de t5 em relação a outras variáveis temporais, como t2, têm um impacto significativo em cc3. Portanto, o intervalo t5-t2 pode ser um indicador relevante para prever o comportamento de cc3, indicando que **a evolução do embrião entre os estágios t5 e t2 desempenha um papel relevante no comportamento de cc3 (t5-t3)**. Isso indica que alterações nesse intervalo temporal afetam significativamente a diferença de tempo entre os estágios t5 e t3.

Figura 29. Dispersão entre cc3 (t5-t3) e t5-t2 - Coeficiente de Spearman: 0.81



Fonte: Autoras (2025)

A correlação entre cc3 e o intervalo *s3* (*-0.36*) é moderada e negativa, sugerindo que conforme s3 se eleva, a tendência é que cc3 diminua. Esta correlação indica que um **aumento no intervalo entre t8 e t5 pode estar inversamente ligado a alterações em cc3**, o que poderia indicar um comportamento de compensação entre as



variáveis. Ao estender o tempo entre os estágios t8 e t5, pode haver uma diminuição no comportamento observado em cc3 (t5-t3), indicando uma interação entre esses momentos de desenvolvimento do embrião.

A correlação de cc3 com a *Ploidia* (-0.281) é bastante elevada quando comparada a outras variáveis, como a correlação de Ploidia com t2 (-0.075) ou t5 (-0.237). Isso sugere que **cc3 exerce uma influência moderada e negativa sobre a Ploidia, indicando que alterações na cc3 podem ter um impacto mais relevante sobre a ploidia do que em outras variáveis temporais**. Este efeito adverso indica que alterações em cc3 estão ligadas a uma diminuição no valor da ploidia. No entanto, **a correlação entre cc3 e Ploidia é evidente**, indicando uma influência mais significativa em relação a outras variáveis temporais.

## t5-t2

A correlação com *Morfo* (-0.38) é moderadamente negativa, sugerindo que, conforme o valor de Morfo cresce, o intervalo t5-t2 tende a se estreitar. Isso indica que **as características morfológicas do embrião podem influenciar o comportamento temporal entre os estágios t5 e t2**, possivelmente indicando um desenvolvimento mais devagar ou desordenado conforme as características morfológicas se modificam.

Com *t2* (0.21), a correlação é fraca e positiva, sugerindo uma pequena ligação entre o crescimento de t2 e o intervalo t5-t2, porém essa relação é pouco relevante. Com *t3* (0.51), a correlação é moderada e positiva, sugerindo uma correlação mais intensa entre os intervalos temporais dos estágios t5 e t3. A correlação com *t4* (0.39) também é moderada e positiva, embora um pouco menos intensa que com t3, sugerindo que t4 também afeta o intervalo entre t5 e t2, porém em uma escala menor. Com *t5* (0.92), a correlação é extremamente intensa e positiva, sinalizando que **o período entre t5 e t2 está fortemente ligado ao tempo de t5. Isso é esperado, pois t5 representa o término desse intervalo**. A correlação robusta indica que **alterações em t5 afetam consideravelmente o intervalo entre t5 e t2**. A correlação com *t8* (0.45) é positiva e moderada, indicando uma conexão entre t8 e o intervalo t5-t2, embora não seja tão intensa quanto com t5.

Com *tSB* (0.40) e *tB* (0.37), as correlações são moderadas e positivas, indicando que **a mudança nos tempos tSB e tB também afeta o intervalo entre t5 e t2**. No caso de *s3* (indicado por *t8-t5*, -0.41), a correlação é moderada e negativa, indicando que, **conforme o intervalo t8-t5 cresce, o intervalo t5-t2 tende a se reduzir, indicando um possível comportamento compensatório entre os dois períodos de tempo**.

O coeficiente de correlação de **Ploidia** (-0.23) com t5-t2 é moderado e negativo,

sugerindo **uma relação inversa entre as alterações na ploidia e o intervalo entre t5 e t2**. Isso indica que, **conforme t5-t2 evolui, a ploidia tende a se reduzir, indicando uma possível restrição da ploidia no comportamento temporal entre esses estágios**.

### s2 (t4-t3)

A correlação com  $t3$  ( $-0.22$ ) é moderada e negativa, sugerindo que conforme o valor de  $t3$  cresce, a distância s2 (t4-t3) tende a se afilar. Isso indica que uma alteração no tempo  $t3$  pode diminuir a distância entre  $t4$  e  $t3$ , sugerindo que as alterações no tempo  $t3$  estão um pouco inversamente ligadas a esse intervalo. A correlação de  $t4$  ( $0.16$ ) é fraca e positiva, indicando uma pequena ligação entre o aumento de  $t4$  e o intervalo s2 (t4-t3).

A correlação com  $cc2$  ( $t3-t2$ ) ( $-0.21$ ) é moderada e negativa. Isso pode sugerir uma relação inversa entre o comportamento temporal dos intervalos t4-t3 e t3-t2, com alterações em  $cc2$  impactando negativamente o intervalo s2, possivelmente por causa de um comportamento mais rápido no começo do desenvolvimento do embrião.

Com uma correlação moderada e positiva de  $s3$  ( $t8-t5$ ) ( $0.25$ ), indica que um crescimento no intervalo s3 está ligado a um crescimento no intervalo s2. Isso pode sugerir que **o aumento do intervalo entre t8 e t5 está ligado ao aumento do intervalo entre t4 e t3**, o que pode representar uma compensação nos tempos entre as diversas fases do embrião.

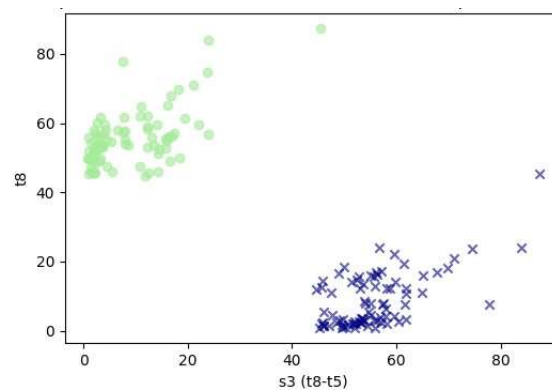
A correlação com *Ploidia* ( $0.11$ ) é leve e positiva, sugerindo que alterações em s2 exercem um efeito levemente positivo na ploidia. Apesar do efeito ser mínimo, ele propõe que **uma alteração em s2 pode levar a uma pequena alteração na ploidia**.

### s3 (t8-t5)

A correlação com  $t5$  ( $-0.39$ ) é moderada e negativa. Isso indica que **o período entre t8 e t5 diminui à medida que o estágio t5 progride, sugerindo que uma elevação em t5 pode levar a uma aceleração do progresso**, reduzindo o intervalo subsequente até t8.

Com  $t8$  ( $0.48$ ) na Figura 30, a correlação é positiva e moderada mais alta que  $t5$ , indicando que, conforme  $t8$  se eleva, o intervalo s3 também tende a se expandir. Isso sugere que um aumento no intervalo de tempo entre  $t8$  e  $t5$  está ligado a um incremento no valor de  $t8$ , sinalizando um processo de evolução mais extenso ou mais lento em comparação com outros intervalos. **Este resultado indica que, conforme o embrião se aproxima do estágio t8, o período de tempo entre os estágios t8 e t5 aumenta**, assim demonstrado no gráfico abaixo, em que s3 está em ver claro.

Figura 30. Dispersão entre s3 (t8-t5) e t8 - Coeficiente de Spearman: 0.48



Fonte: Autoras (2025)

A correlação com *cc3* ( $t5-t3$ ) ( $-0.36$ ) é moderada e negativa. Este efeito indica que uma extensão prolongada entre t5 e t3 pode estar ligada a uma diminuição no tempo entre t8 e t5, sugerindo que **o progresso no período intermediário (t5-t3) pode impactar diretamente os estágios finais**, acelerando o período entre t5 e t8.

A correlação com *tSC-t8* ( $-0.35$ ) é moderada e negativa. Isso indica uma **relação de compensação entre as fases intermediárias e finais do desenvolvimento embrionário**, onde o tempo entre tSC e t8 diminui, enquanto o tempo entre t8 e t5 aumenta.

Finalmente, a correlação com a *Ploidia* ( $0.25$ ) é moderada e positiva, indicando que, conforme a ploidia cresce, o intervalo s3 (t8-t5) também tende a se expandir. Este efeito pode sugerir que a **aneuploidia está ligado a uma diminuição no tempo de desenvolvimento, estendendo o intervalo entre t8 e t5**. Este comportamento pode indicar variações no crescimento celular devido à composição genética do embrião, onde **aneuploidias podem resultar em um desenvolvimento mais lento**.

### tSC-t8

A relação com *t8* ( $-0.29$ ) é moderadamente negativa sugere que, conforme o intervalo de tempo entre tSC e t8 cresce, o tempo de t8 também tende a crescer. Este resultado indica que, **ao aumentar o intervalo entre tSC e t8, pode ocorrer uma diminuição na velocidade do desenvolvimento embrionário entre tSC e t8, estendendo o período até t8**. Isso pode indicar uma etapa de crescimento mais lenta ou prolongada, influenciando o ritmo geral de avanço do embrião.

A correlação com *tSC* ( $0.70$ ) é forte e positiva, que indica que o intervalo tSC-t8 tem uma ligação direta com o intervalo tSC. **Conforme o intervalo de tempo entre tSC e t8 se amplia, o intervalo tSC também tende a se expandir**. Isso mostra uma relação em que o começo do estágio tSC afeta diretamente o tempo entre tSC e t8,

indicando que um intervalo mais extenso entre tSC e t8 pode ser parcialmente determinado pela duração do estágio tSC em si. **Este efeito indica uma relação temporal na qual o estágio tSC desempenha um papel significativo no progresso do embrião no período subsequente até t8.**

Em *tSB* (0.44), a correlação é moderadamente positiva e sugere que, conforme o intervalo tSB se amplia, o intervalo tSC-t8 também tende a crescer. Isso pode sugerir que o período entre tSB e tSC pode estar ligado a um prolongamento até t8, o que poderia sinalizar a continuidade do desenvolvimento embrionário em um estágio mais prolongado. Este resultado indica que **um período mais extenso entre tSB e tSC pode se estender até o intervalo tSC-t8, estendendo o processo.**

A correlação de *tB-tSB* (0.20) é positiva e indica que um crescimento no intervalo entre tB e tSB pode estar estreitamente ligado a um crescimento no intervalo tSC-t8. Isso indica que o avanço do desenvolvimento embrionário de tB para tSB pode afetar o intervalo de tempo entre tSC e t8, embora não de maneira tão marcante.

A correlação muito tênue (0.04) e positiva sugere que Ploidia exerce uma influência mínima no intervalo tSC-t8.

## tB-tSB

A correlação com a *Idade* (0.20) sendo positiva e fraca indica que, conforme a idade avança, o intervalo tB-tSB também tende a crescer, mesmo que de maneira sutil. Isso pode sugerir que a idade da paciente pode ter um pequeno impacto no desenvolvimento embrionário, possivelmente indicando uma variação no ritmo de desenvolvimento.

A correlação moderadamente positiva do *Estágio* (0.38) indica que, conforme o estágio de desenvolvimento progride, o intervalo tB-tSB tende a se estender. Este efeito sugere que **fases mais avançadas do desenvolvimento embrionário podem estar ligadas a um prolongamento no intervalo entre tB e tSB**, o que pode indicar um avanço mais lento ou um intervalo mais extenso entre os eventos celulares que caracterizam tais fases.

A correlação moderadamente positiva de *tB* (0.38) indica que, conforme o intervalo tB se amplia, o intervalo tB-tSB também tende a se expandir. Isso é coerente, já que **o período tB é um estágio inicial do desenvolvimento embrionário, e sua duração pode impactar diretamente o tempo até o estágio seguinte**, o tSB, impactando a evolução do desenvolvimento embrionário. Uma maior extensão do tB geralmente indica um período mais extenso até o tSB.

A correlação de *tSC-t8* (0.20) é fraca e positiva correlação sugere uma pequena ligação entre o intervalo tB-tSB e o intervalo tSC-t8. Isso indica que alterações no intervalo tB-tSB podem estar ligadas a alterações no tempo entre tSC e t8, mesmo que essa conexão

seja restrita.

A correlação com *Ploidia* ( $-0.284$ ) exerce um efeito significativo, visto que **é a segunda correlação de maior magnitude entre as variáveis examinadas**. Desempenha um papel significativo, uma vez que indica que a qualidade genética do embrião, avaliada por tB e tSB, pode impactar diretamente a ploidia. Isso pode ser interpretado como um sinal de que **embriões com predisposições genéticas mais favoráveis, tais como ploidia euploide ou mosaico de baixo grau, podem exibir um ritmo acelerado de desenvolvimento, evidenciado pelo intervalo mais breve entre essas fases cruciais do seu desenvolvimento**. Essa correlação negativa sugere que a qualidade genética não só afeta a saúde do embrião, mas também pode influenciar a temporização dos eventos celulares.

## Ploidia

A avaliação das variáveis relacionadas à ploidia indica que **a idade materna tem o maior efeito na qualidade genética do embrião**. A idade materna, **com uma correlação de -0,50**, apresenta uma correlação inversa significativa com a ploidia. Isso significa que, **conforme a mulher envelhece, a quantidade de embriões euploides diminui, o que indica um crescimento na taxa de aneuploidia**. Isso está alinhado com o que a literatura já aponta sobre o impacto do envelhecimento materno na qualidade genética dos embriões, onde embriões de mulheres mais idosas têm maior probabilidade de conter erros cromossômicos. Assim, a idade materna é um elemento crucial na avaliação da ploidia embrionária e deve ser levada em conta em procedimentos de fertilização in vitro.

Adicionalmente, as variáveis **tB-tSB e cc3 (t5-t3)** possuem uma correlação moderada de  $-0,28$  com a ploidia, sugerindo que o período entre as fases de desenvolvimento **t5, t3, tB e tSB podem afetar diretamente a qualidade genética do embrião**. Este intervalo de tempo está ligado ao ritmo de divisão celular e ao comportamento dos cromossomos. Mudanças nesse processo podem comprometer a criação de embriões euploides, levando a um aumento na probabilidade de aneuploidia. Essa correlação sugere que a sincronização adequada do desenvolvimento embrionário entre esses estágios é fundamental para a formação de embriões geneticamente saudáveis.

Outros fatores, como o **estágio de desenvolvimento (-0,24)** e o **t5 (-0,24)**, também mostraram uma correlação negativa com a ploidia. Isso indica que **o avanço para fases mais avançadas de desenvolvimento e o atraso no t5 podem estar ligados a uma queda na qualidade genética dos embriões**. Essas variáveis enfatizam a necessidade de um acompanhamento exato do desenvolvimento embrionário, pois pequenas mudanças no intervalo entre as fases podem influenciar o comportamento genético do embrião, levando a embriões com menor probabilidade de implantação.

5.1.1.3 Atividade 3 (A3): Normalização dos Dados para Otimização

O método Z-Score, explicado no Apêndice B, foi aplicado com êxito para normalizar os dados morfocinéticos dos embriões, conforme explicado no Capítulo 4. Esta fase foi importante para garantir que todas as variáveis numéricas, sem importar suas unidades ou escalas originais, estivessem na mesma base de comparação. Este procedimento, crucial para a implementação de algoritmos de aprendizado de máquina, eliminou vieses de amplitude e permitiu uma avaliação mais consistente e comparável entre os dados.

Análise da Planilha Normalizada

Depois de realizar a normalização, a Tabela 3 gerada apresenta os dados convertidos para que cada variável possua uma média de 0 e um desvio padrão de 1.

Tabela 3. Planilha Normalizada

Kidscore	t2	t3	t4
-0,1651468441	-1,04331772	-0,2061446738	-0,1365241993
-0,3621219891	0,0828611265	0,1102377935	-0,04951654895
1,509141888	-0,7805426557	-0,4547308981	-0,8035828516
-0,1651468441	0,233018306	0,08769304581	-0,1365241993
1,75536082	-0,4426890018	-0,4095334028	-0,8909505013
1,410654316	-0,9306998352	-0,6129221317	-1,151613453
0,8689726671	-1,231014194	-0,5677246364	-1,064605802
-0,3128782029	-0,8180819506	-0,4547308981	-0,7745803015
0,9182164534	-1,11839631	-0,9067058513	-1,586651704
-0,06665927163	0,833647024	0,607410242	0,7625458359
-0,5095833479	0,833647024	1,217576429	1,197593105
-1,888679363	0,3831754855	0,0650429815	-0,07851909905

Fonte: Autoras (2024)

Isso reduz o efeito das variações de escala entre as variáveis e simplifica a compreensão dos dados. A tabela normalizada foi verificada e todos os valores foram ajustados de acordo com a fórmula do Z-Score, com a média e o desvio padrão calculados para cada linha. Por exemplo, a variável "t2" tinha valores que oscilavam entre 19 e 38 horas antes da normalização, indicando uma ampla dispersão na faixa de dados. Depois de normalizados, os dados da variável foram convertidos em uma escala com média zero e desvio padrão de 1. Isso permitiu que todos os valores da variável estivessem comparáveis, mesmo com a grande variação original.

### Efeito da Normalização na Comparabilidade das Variáveis

O principal benefício de normalizar as variáveis é que agora é possível compará-las de forma mais justa. Antes da normalização, variáveis como o "t8"(que apresentava uma amplitude maior) poderiam impactar de forma desproporcional o modelo de aprendizado de máquina, enquanto variáveis como o "t2", com uma amplitude menor, poderiam ser desconsideradas. Com a normalização, todas as variáveis estão na mesma escala, possibilitando uma avaliação justa de cada uma durante o treinamento do modelo.

Este efeito é especialmente relevante em algoritmos que lidam com escalas de variáveis, tais como redes neurais. Com os dados normalizados, as conexões entre as variáveis podem ser examinadas de forma mais nítida, simplificando a elaboração de modelos preditivos mais confiáveis.

### Impacto na Qualidade dos Modelos de Previsão

Modelos elaborados a partir de dados normalizados costumam ter um desempenho superior, já que são menos afetados por variáveis de escalas mais elevadas (Jaiswal, 2024). Ademais, a normalização tem o potencial de intensificar a convergência de algoritmos de otimização, como o gradiente descendente, que pode ser mais lento ou até mesmo não funcionar se os dados não forem padronizados (Milewski et al., 2016).

Agora que as variáveis estão em uma escala comparável, os modelos de previsão da porcentagem de euploidia dos embriões têm maior chance de gerar resultados mais exatos e seguros.

### Análise dos Resultados Normalizados

**Idade e Kidscore:** A normalização destas variáveis (Tabela 4) mostrou que ambas possuem distribuições semelhantes em termos de amplitude, sugerindo uma maior homogeneidade nos dados demográficos e nos escores atribuídos aos embriões.

Tabela 4. Normalização da Idade e Kidscore

Idade	Kidscore
-0,3353649796	-0,1651468441
1,209574814	-0,3621219891
1,209574814	1,509141888
1,209574814	-0,1651468441
1,209574814	1,75536082

Fonte: Autoras (2024)

**Parâmetros Temporais:** Após a normalização, ficou evidente que os tempos relacionados à divisão celular possuem uma distribuição com pouca variabilidade, o que pode refletir padrões biológicos comuns nos embriões analisados.

**Durações e Intervalos:** Os intervalos normalizados mostraram diferenças mais acentuadas nessas variáveis em relação aos tempos absolutos. Isso pode sugerir que os intervalos entre eventos celulares são mais propensos a alterações individuais.

#### 5.1.1.4 Atividade 4 (A4): Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados

Nessa atividade, compartilhamos os resultados alcançados ao realizar o processo de subdivisão do conjunto de dados em três subconjuntos: treinamento, validação e teste. Esta fase é crucial para assegurar que o modelo de aprendizado de máquina seja formado, ajustado e avaliado de forma justa e representativa.

A distribuição dos dados foi feita por meio de um programa Python, apresentado e explicado no capítulo 4, que respeita as proporções estabelecidas na metodologia: 70% para treinamento, 15% para validação e 15% para teste. A tabela de dados normalizada gerada pela Atividade 3 contém 82 linhas de dados, as quais foram distribuídas de maneira aleatória para assegurar a representatividade e prevenir viés de classificação. Os tamanhos dos subconjuntos formados foram os seguintes:

- **Treinamento:** 57 linhas
- **Validação:** 12 linhas
- **Teste:** 13 linhas

Foi utilizado um gerador pseudoaleatório controlado pelo parâmetro `random_state` para embaralhar os dados, assegurando a reprodutibilidade dos resultados.

## Arquivos Gerados

Os conjuntos de dados foram convertidos em arquivos Excel, cada um denominado conforme sua função. Cada subconjunto reflete a diversidade presente no conjunto original e o embaralhamento garante que os subconjuntos não compartilhem padrões específicos devido à ordenação prévia dos dados. Caso haja a vontade de ver esses documentos, siga os passos descritos da “Atividade 4” no GitHub do nosso projeto: [GitHub-TCC](#), executando o código, se obtém todos os arquivos. Os arquivos produzidos incluem as figuras 31, 32 e 33.



Figura 31. Conjunto de Treinamento: dados\_treinamento.xlsx

Idade	Estágio	Morfo	Kidscore	t2	t3	t4	t5	t8	tSC	tSB	tB	cc2 (t3-t2)	cc3 (t5-t3)	t5-t2	s2 (t4-t3)	s3 (t8-t5)	tSC-t8	tB-tSB	Ploidia
-0.03	D5	3BB	0.52	0.98	0.68	0.59	0.58	1.49	-0.18	-0.96	-0.61	0.12	0.23	0.25	-0.34	1.01	-1.33	0.69	Aneuploide
-0.34	D5	3BB	-0.17	-1.04	-0.21	-0.14	0.46	1.90	1.11	0.77	0.43	0.62	0.83	0.99	0.16	1.54	-0.41	-0.73	Aneuploide
1.21	D5	3BB	0.43	0.68	0.45	0.21	0.44	-0.35	-1.10	-0.53	-0.50	0.06	0.23	0.22	-0.45	-0.76	-0.80	-0.06	Aneuploide
1.21	D5	3BC	-0.56	-0.41	-0.32	-0.63	-0.15	-1.05	-0.86	-0.12	0.11	-0.11	0.07	0.00	-0.27	-0.95	-0.01	0.61	Aneuploide
0.28	D6	3CC	-1.79	0.91	1.33	1.34	1.52	0.57	1.52	1.12	2.16	1.15	0.99	1.39	-0.45	-0.78	1.03	3.19	Aneuploide
-0.03	D6	3BB	-0.51	1.17	1.26	1.23	1.11	0.18	0.92	1.25	1.59	0.82	0.47	0.79	-0.48	-0.82	0.75	1.28	Aneuploide
-0.03	D6	3BB	-0.51	0.83	1.22	1.20	1.30	0.30	0.85	1.16	1.27	1.05	0.77	1.16	-0.45	-0.86	0.59	0.64	Aneuploide
-0.64	D5	5AA	1.16	-1.83	-2.90	-1.76	-1.92	-1.22	-1.02	-1.35	-1.35	-2.63	-0.17	-1.44	2.40	0.47	-0.03	-0.39	Euploide
1.21	D5	5AA	1.76	-0.44	-0.41	-0.89	-0.13	0.31	-1.29	-1.17	-1.35	-0.21	0.17	0.04	-0.45	0.44	-1.50	-0.83	Aneuploide
-0.03	D5	3AA	0.92	-0.14	0.04	0.15	0.55	-0.29	-0.09	-0.34	-0.22	0.19	0.75	0.71	0.12	-0.79	0.14	0.25	Aneuploide
-0.03	D5	3AA	1.26	-0.33	-0.03	-0.31	-0.05	-0.72	0.26	-0.47	-0.42	0.25	-0.05	0.09	-0.34	-0.69	0.81	0.00	Euploide
-0.03	D6	3AC	-1.49	-0.10	0.90	1.46	0.86	-0.15	-0.07	2.74	2.60	1.41	0.43	1.06	0.37	-0.93	0.05	0.38	Aneuploide
1.52	D5	5AB	0.03	-1.53	-0.86	-1.35	-0.47	-1.21	0.16	-0.28	0.21	0.09	0.09	0.12	-0.31	-0.82	1.10	1.28	Euploide
0.28	D6	4BB	-2.78	-0.22	0.29	-0.05	0.31	-0.68	2.40	1.94	2.26	0.62	0.19	0.46	-0.52	-0.98	2.86	1.44	Aneuploide
0.28	D5	5BB	-0.36	0.76	-1.92	1.20	-0.68	0.72	-0.39	-0.84	-0.79	-3.49	0.73	-1.13	4.51	1.35	-0.94	-0.11	Euploide
-3.43	D5	3BB	-0.61	-0.22	0.27	-0.08	-1.37	-0.53	-1.62	1.30	0.78	0.59	-2.18	-1.51	-0.52	0.68	-1.15	-1.06	Euploide
-0.03	D5	3BB	0.03	1.92	1.42	1.52	-0.57	0.02	0.55	0.07	-0.33	0.39	-2.06	-1.51	-0.38	0.53	0.52	-1.09	Euploide
-0.34	D5	3BC	-1.00	2.22	1.49	1.60	1.84	3.48	1.19	0.73	0.21	0.22	1.30	1.19	-0.38	1.92	-1.56	-1.22	Euploide
-0.64	D5	3BC	-0.66	-2.13	-3.10	-1.73	-2.14	-0.69	-0.46	-0.42	-0.73	-2.66	-0.31	-1.58	2.76	1.21	0.09	-0.98	Euploide
-0.03	D6	3CC	-0.66	0.42	-1.65	0.24	-1.09	-0.08	0.18	-0.29	0.39	-2.79	-0.09	-1.46	2.90	0.90	0.24	1.80	Aneuploide
0.28	D5	3BB	-0.07	0.83	0.61	0.76	0.95	0.28	0.42	-0.17	0.16	0.15	0.81	0.74	-0.02	-0.57	0.19	0.85	Aneuploide
-0.03	D5	3AA	1.46	-0.33	0.16	-0.17	0.26	-0.72	0.17	-0.47	-0.77	0.52	0.23	0.45	-0.45	-0.97	0.72	-0.96	Euploide
1.52	D5	5AA	1.41	-0.93	-0.61	-1.15	-0.08	-0.88	-0.74	-1.31	-0.89	-0.08	0.43	0.32	-0.45	-0.83	-0.03	0.79	Aneuploide
-3.43	D5	5AB	0.52	-0.41	0.22	0.10	0.40	-0.47	-1.21	-0.27	-0.58	0.68	0.37	0.65	-0.23	-0.85	-0.81	-0.93	Euploide

Fonte: Autoras (2025)

Figura 32. Conjunto de Validação: dados\_validacao.xlsx

Idade	Estágio	Morfo	Kidscore	t2	t3	t4	t5	t8	tSC	tSB	tB	cc2 (t3-t2)	cc3 (t5-t3)	t5-t2	cc2 (t4-t3)	cc3 (t8-t5)	tSC-t8	tB-tSB	Ploidia
-3.43	D5	3BC	-1.05	-0.03	0.52	0.30	-1.09	0.06	-0.11	0.35	0.74	0.78	-2.00	-1.26	-0.45	1.04	-0.15	1.18	Euploide
-0.03	D5	3BC	-0.51	1.13	1.22	1.37	-0.68	3.90	1.28	0.67	0.78	-2.04	-1.30	-0.23	4.63	-1.61	-1.32	-1.32	Aneuploide
-0.03	D5	3BC	-0.51	-0.67	-0.50	-0.60	0.40	0.77	-0.15	-0.73	-0.31	-0.14	1.01	0.76	0.05	0.43	-0.74	0.95	Euploide
-0.64	D5	3BB	-0.12	0.27	0.34	0.70	-0.99	0.18	0.65	0.14	0.08	0.25	-1.70	-1.28	0.34	1.08	0.49	-0.11	Euploide
-0.64	D5	5AA	1.56	-1.68	-1.18	-1.88	-1.99	-1.36	-0.92	-1.48	-1.41	-0.24	-1.78	-1.59	-0.45	0.39	0.17	-0.24	Euploide
-0.03	D6	3CC	-1.89	0.38	0.07	-0.08	-1.37	-0.34	0.56	1.33	2.10	-0.24	-2.00	-1.77	-0.20	0.89	0.81	2.52	Aneuploide
1.21	D6	3BC	-1.49	0.38	-1.72	0.39	-0.99	-0.01	1.05	1.03	1.36	-2.86	0.11	-1.33	3.19	0.89	1.02	1.21	Aneuploide
-0.03	D5	3AB	0.33	1.13	0.79	0.65	0.58	-0.46	0.39	-0.32	-0.25	0.15	0.13	0.18	-0.45	-1.00	0.73	0.10	Euploide
0.28	D5	5AA	1.76	-0.25	-0.27	-0.77	-0.05	0.39	-0.71	-1.42	-1.39	-0.18	0.17	0.05	-0.52	0.44	-0.99	-0.34	Euploide
-0.34	D5	3AB	-0.02	-0.67	0.36	0.24	0.40	-0.28	-0.40	0.30	0.33	1.12	0.25	0.76	-0.27	-0.64	-0.18	0.15	Euploide
-0.64	D5	3AB	-0.81	1.58	-1.38	0.79	-0.95	-0.25	-0.26	-0.03	0.11	-3.42	-0.13	-1.81	3.15	0.60	-0.05	0.38	Euploide
-0.03	D5	4AA	1.46	-1.01	-0.64	-0.80	-0.51	-0.99	-0.11	-0.55	-0.73	-0.04	-0.17	-0.16	0.01	-0.55	0.67	-0.65	Aneuploide

Fonte: Autoras (2025)

Figura 33. Conjunto de Teste: dados\_teste.xlsx

Idade	Estágio	Morfo	Kidscore	t2	t3	t4	t5	t8	tSC	tSB	tB	cc2 (t3-t2)	cc3 (t5-t3)	t5-t2	s2 (t4-t3)	s3 (t8-t5)	tSC-t8	tB-tSB	Ploidia
-0.03	D5	4AB	1.16	-0.82	-0.64	-1.21	-0.47	-1.23	-1.78	-1.83	-1.59	-0.21	-0.11	-0.19	-0.48	-0.85	-0.76	0.15	Aneuploide
1.21	D5	2BB	-0.36	0.08	0.11	-0.05	0.54	-0.07	1.26	0.65	0.79	0.09	0.67	0.60	-0.23	-0.55	1.28	0.59	Aneuploide
-0.03	D6	3BC	-1.15	0.95	0.74	0.99	-0.85	-0.31	0.20	0.39	1.03	0.25	-1.86	-1.41	0.05	0.44	0.44	1.88	Euploide
1.21	D5	3AB	0.57	-0.18	0.09	-0.31	0.38	-0.51	-0.05	-0.41	-0.78	0.29	0.47	0.53	-0.52	-0.87	0.35	-1.14	Aneuploide
1.21	D5	3AA	1.51	-0.78	-0.45	-0.80	-0.11	-0.79	-0.84	-0.97	-0.98	0.02	0.25	0.22	-0.27	-0.72	-0.19	-0.31	Aneuploide
0.28	D5	3BB	-2.78	1.21	1.04	1.05	1.21	0.49	-0.45	-0.85	-1.15	0.45	0.81	0.89	-0.34	-0.59	-0.82	-1.06	Aneuploide
-0.03	D5	3AA	0.97	0.53	0.68	0.50	0.54	-0.23	-0.04	0.54	-0.02	0.52	0.17	0.40	-0.45	-0.72	0.14	-1.40	Euploide
-0.03	D6	2BB	-0.95	0.42	0.68	0.59	0.99	0.73	2.22	2.49	2.08	0.62	0.81	0.97	-0.34	-0.14	1.58	-0.42	Euploide
1.21	D5	3BC	-0.31	-0.29	-0.05	-0.40	0.29	1.15	-1.74	-1.18	-0.45	0.19	0.45	0.46	-0.41	0.92	-2.58	1.67	Aneuploide
-0.34	D5	4AA	1.36	0.57	0.63	0.50	0.60	-0.29	0.21	-0.46	-0.69	0.42	0.29	0.45	-0.38	-0.83	0.43	-0.78	Euploide
-0.64	D5	3AB	0.08	0.72	0.90	0.94	0.83	0.01	-0.51	-0.06	0.18	0.68	0.39	0.66	-0.27	-0.74	-0.50	0.64	Euploide
-0.03	D5	3AB	0.33	1.13	0.83	0.79	0.68	1.74	-0.98	-0.08	-0.90	0.22	0.23	0.30	-0.34	1.18	-2.30	-2.27	Aneuploide
-0.03	D6	3BC	-0.71	0.20	0.22	-0.08	-1.27	-0.40	0.20	-0.31	-0.16	0.15	-2.00	-1.58	-0.45	0.73	0.50	0.33	Euploide

Fonte: Autoras (2025)

O documento "dados\_treinamento.xlsx" será o único empregado nas próxima fase da atividade, de aumento de dados, para evitar interferências ruins na criação e desempenho do modelo de aprendizado de máquina (KIAR et al., 2021)

## Aumento de Dados (data augmentation)

A utilização da técnica de ampliação de dados (data augmentation) através do método de Monte Carlo levou a uma ampliação considerável do conjunto de dados de treinamento, fortalecendo sua solidez e habilidade de generalização. O procedimento de ampliação de dados foi posto em prática para criar novos exemplos a partir dos dados originais, visando aprimorar a performance do modelo e prevenir overfitting.

Depois de executar o código apresentado e explicado na Atividade 4, a quantidade de dados de treinamento foi triplicada, de acordo com o fator de ampliação estabelecido em 3, utilizado em situações onde o conjunto de dados inicial é extremamente pequeno. O programa produziu novos valores numéricos através da distribuição normal, usando a média e o desvio padrão das colunas originais, assegurando que os novos valores preservassem as propriedades estatísticas do conjunto original. Por outro lado, as colunas categóricas tiveram seus valores replicados de forma aleatória, mantendo as proporções originais e a disposição das categorias.

Esta operação resultou na geração de 228 linhas, o que equivale a três vezes o total de linhas existentes no conjunto de dados. Se desejar acessar esse documento, basta seguir as instruções da "Atividade 4" no GitHub do nosso projeto: [GitHub-TCC](#). O conjunto ampliado de dados foi guardado em um novo documento Excel, denominado `dados_treinamento_aumentado.xlsx`, pronto para ser aplicado em fases subsequentes do trabalho.

Assim, a implementação do procedimento de ampliação de dados levou à formação de um conjunto de dados de treinamento ampliado, preservando as características estatísticas e a consistência das distribuições originais, enquanto ampliava a variedade e a quantidade de exemplos disponíveis para o treinamento do modelo. Isso garantiu um fundamento de treinamento mais robusto, auxiliando na conquista de melhores resultados no modelo preditivo.

## 6 Considerações e Trabalhos Futuros

### 6.1 Atividade 1 (A1): Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia

A Atividade 1 (A1) foi crucial para a realização deste estudo, ao estabelecer os fundamentos para o modelo de Inteligência Artificial (IA) sugerido. Ao analisar, revisar, escolher e limpar as variáveis, conseguimos identificar os atributos mais significativos para a previsão de euploidia. A seleção cuidadosa das variáveis foi respaldada por uma pesquisa bibliográfica detalhada, assegurando que cada componente escolhido possuísse uma base científica robusta.

As variáveis finais abrangem parâmetros de tempo como t2, t3, t4, t5, t6 e t8, indicadores de desenvolvimento embrionário (s2, cc2, s3, tSC, tSB, tB), características qualitativas (Estágio, Morfo, KIDScore), além da coluna Plodia, crucial para classificar embriões com euploidia normal ou com alterações cromossômicas. Variáveis sem comprovação científica adequada, como st2 e t2-st2, foram excluídas do conjunto.

A partir das fases de limpeza e organização de dados, a planilha original passou por uma revisão, resultando na "Planilha de Dados Refinados", que está nos anexos, com 83 linhas, assegurando a consistência e a qualidade do conjunto de dados para as próximas etapas. Este estudo também ressalta a relevância da precisão ética e técnica ao manusear dados delicados, aumentando a confiabilidade das análises conduzidas.

Os resultados alcançados até agora indicam que o conjunto de dados aprimorado e a escolha meticulosa de variáveis estabeleceram uma fundação sólida para a criação de um modelo de IA eficiente e seguro.

### 6.2 Atividade 2 (A2): Identificação da Correlação entre os Parâmetros na Previsão da Ploidia do Embrião.

Com base nos resultados da Atividade 2, conseguimos identificar padrões significativos que oferecem perspectivas sobre os elementos que afetam a criação de embriões geneticamente saudáveis. Além disso, conseguimos avaliar a capacidade preditiva das variáveis analisadas para a elaboração de um modelo fundamentado em inteligência artificial.

O fator idade se mostrou como a variável mais importante na previsão da ploidia, com uma correlação negativa significativa de -0,50. Esta conclusão está alinhada com

a literatura científica, que indica o envelhecimento como um elemento fundamental no crescimento de erros cromossômicos. Portanto, os embriões provenientes de mulheres com mais idade têm maior probabilidade de sofrer aneuploidia.

A diferença entre os tempos tB e tSB (-0,28) e o ciclo celular CC3 (-0,28) também se sobressaíram como variáveis significativas na previsão da ploidia. Essas correlações indicam que o ritmo e a sincronização do crescimento celular desempenham um papel crucial na criação de embriões euploides. Mudanças nesses momentos podem indicar falhas cromossômicas que levam a aneuploidias.

Variáveis como o estágio de desenvolvimento e o tempo t5 apresentam correlações moderadas e negativas (-0,24), sugerindo que atrasos em etapas particulares do ciclo embrionário podem afetar adversamente a qualidade genética. Possivelmente, esses atrasos indicam falhas no comportamento cromossômico durante as divisões celulares.

Os achados indicam que a qualidade genética do embrião não se limita ao comportamento individual de cada variável, mas também à interação dinâmica entre elas. Por exemplo, a correlação significativa entre cc2 e t3 (0,80) destaca a relevância do sincronismo nos estágios iniciais da divisão celular. Ademais, a correlação negativa entre tB (-0,17) e ploidia sublinha que as variáveis temporais podem exercer impactos discretos, porém significativos, no comportamento genético embrionário.

### 6.3 Atividade 3 (A3): Normalização dos Dados para Otimização

A normalização dos dados, feita através do método Z-Score, foi um passo crucial na realização deste estudo. Esta metodologia ofereceu uma base comparativa sólida entre as variáveis, removendo efeitos de escalas desiguais e assegurando maior solidez aos modelos de aprendizagem de máquina. A avaliação minuciosa das variáveis normalizadas forneceu percepções valiosas sobre as informações morfocinéticas dos embriões, preparando o conjunto de dados para a próxima fase da modelagem.

Antes do processo de normalização, variáveis como "t8" exibiam amplitudes consideravelmente maiores, enquanto variáveis como "t2" apresentavam amplitudes mais baixas. Esta diferença complicava a avaliação equitativa das variáveis e poderia afetar de forma negativa o rendimento dos algoritmos de aprendizado de máquina. Depois da normalização, todas as variáveis foram ajustadas para uma escala homogênea, com média zero e desvio padrão igual a 1, possibilitando uma análise mais equilibrada e comparativa.

A normalização evidenciou a uniformidade de certas variáveis, tais como "Idade" e "KIDScore", que apresentaram distribuições parecidas após o ajuste. Isso indica que a variabilidade relativa dos dados e dos escores atribuídos aos embriões é reduzida, o que pode simplificar a interpretação e aplicação dessas variáveis nos modelos preditivos.

## 6.4 Atividade 4 (A4): Separar o conjunto de dados em conjuntos de treinamento, validação e teste, fazendo uma distribuição dos dados e aplicar técnica de aumento de dados.

A divisão e o aumento dos dados foram passos cruciais para assegurar a solidez e a fiabilidade do modelo preditivo criado neste estudo. A realização dessas tarefas ajudou a estabelecer uma base robusta, variada e estatisticamente consistente para o treinamento, validação e teste do modelo, ao mesmo tempo que abordou os desafios associados ao reduzido volume de dados inicial.

A segmentação das informações em subconjuntos de treinamento, validação e teste, de acordo com a proporção de 70%, 15% e 15%, garantiu a neutralidade e representatividade dos dados. A utilização de um gerador pseudoaleatório, regulado pelo parâmetro `random_state`, assegurou a replicabilidade dos resultados, possibilitando a replicação do procedimento em análises futuras.

A aleatoriedade na distribuição dos dados impediu vieses e garantiu que os subconjuntos espelhassem a variedade existente no conjunto original. A elaboração de arquivos Excel estruturados e de fácil acesso forneceu um alicerce claro e reutilizável para as fases seguintes.

A decisão de concentrar o aumento de dados apenas no conjunto de treinamento impediu qualquer impacto negativo nos subconjuntos de validação e teste, garantindo a integridade da análise do modelo.

Utilizar a técnica de Monte Carlo para expandir o conjunto de treinamento foi uma abordagem eficiente para superar as restrições de amostragem decorrentes do tamanho limitado do dataset inicial. A metodologia produziu exemplos sintéticos a partir de características estatísticas dos dados originais, preservando sua consistência e assegurando que os dados expandidos fossem representativos.

As variáveis numéricas foram expandidas de acordo com distribuições normais, ao passo que as categóricas mantiveram seus valores originais. Isso assegurou que o conjunto expandido não só mantivesse as propriedades originais, como também ampliasse a gama de exemplos à disposição do modelo.

O acréscimo de 228 linhas ao conjunto de treinamento aumentou consideravelmente a variedade e a quantidade de dados disponíveis, reforçando a habilidade do modelo de se adaptar a novos cenários e diminuir a probabilidade de superfaturamento.

A fusão da segmentação estratégica dos dados com a expansão do conjunto de treinamento gerou uma base de dados mais sólida para o treinamento do modelo. Este procedimento melhorou a preparação dos dados e estabeleceu condições para que o modelo

atinja maior exatidão e confiabilidade.

## 6.5 Conclusão da Fase 1: Análise e Preparação de Dados

Com o cumprimento das quatro tarefas sugeridas na Fase 1: Análise e Preparação de Dados, concluímos com sucesso a primeira fase do projeto, voltada para a expansão, processamento e análise dos dados para a previsão de ploidia embrionária.

A Primeira Fase nos possibilitou criar um conjunto de dados sólido, estruturado e aprimorado para uso em algoritmos de aprendizado de máquina. Este marco simboliza um progresso importante no progresso do projeto, confirmando as suposições iniciais e estabelecendo uma fundação sólida para a próxima fase.

## 6.6 Conclusão da Fase 1: Análise e Preparação de Dados

Após a finalização da Fase 1 do projeto, começamos a Fase 2, que se concentra no desenvolvimento, análise e criação da solução preditiva. Nesta fase, os esforços serão direcionados para três metas específicas: a capacitação e ajuste do modelo de aprendizado de máquina, a análise minuciosa de sua performance e o desenvolvimento de um protótipo inicial de interface direcionada ao usuário final, neste caso, os médicos. Cada tarefa será organizada e realizada para assegurar resultados consistentes, fiáveis e relevantes para a prática clínica.

O objetivo inicial, OE2, diz respeito ao treinamento e ao ajuste do modelo de aprendizado de máquina. Na Quinta Atividade, será feita a escolha e ajuste dos algoritmos que serão desenvolvidos para o desafio de classificação de euploidia. Esta fase envolverá a verificação do modelo com base nos dados coletados na Fase 1. A meta principal é determinar a estratégia que proporcione os resultados mais precisos e robustos, registrando as configurações perfeitas para aplicações futuras.

Depois da elaboração inicial do modelo, avançaremos para uma análise minuciosa de sua efetividade, conforme previsto no OE3. A Atividade 6 se concentrará na aplicação de métricas como acurácia, precisão, recall e F1-score para avaliar a performance do modelo em situações de classificação binária. Na Atividade 7, será conduzida uma análise minuciosa com o uso da matriz de confusão e da curva ROC, possibilitando a avaliação da habilidade do modelo em diferenciar corretamente as categorias de euploidia e aneuploidia. Essas avaliações são fundamentais para assegurar a fiabilidade do modelo em cenários reais de uso.

Finalmente, na Atividade 8 do OE4, começaremos a desenvolver uma interface básica que será implementada para simplificar o acesso e a compreensão dos resultados preditivos. Este protótipo será desenvolvido para mostrar as previsões de euploidia de

forma compreensível e compreensível, possibilitando que os médicos usem o instrumento de maneira eficaz no processo decisório clínico.

Com as bases estabelecidas na Fase 1, a Fase 2 é um momento crucial para a realização das metas do projeto. O objetivo será desenvolver um modelo preditivo eficiente e uma interface útil, sempre procurando harmonizar a inovação tecnológica com as demandas da medicina reprodutiva. Este progresso será crucial para obter resultados clínicos mais precisos e acessíveis, proporcionando vantagens diretas para pacientes e profissionais do setor.

Tabela 5 – Cronograma de atividades

<b>Atividades</b>	<i>Capítulo 6.</i>
A1 - Análise, Revisão, Seleção e Limpeza de Variáveis para Predição de Euploidia.	
A2 - Normalização dos Dados para Otimização.	<i>Considerações</i>
A3 - Identificação da Correlação e Atribuição de Pesos aos Parâmetros na Previsão da Ploidia do Embrião.	
A4 - Divisão dos Dados e aplicação de Data Augmentation.	
Apresentação do TCC1.	
A5 - Desenvolvimento e Treinamento do Modelo de Machine Learning para Otimização da Predição de Euploidia, Incluindo Treinamento, Validação e Teste.	<i>Trabalhos Futuros</i>
A6 - Utilizar métricas adequadas para medir o desempenho do modelo.	
A7 - Avaliação do Desempenho do Modelo na Predição por meio da Matriz de Confusão e Curva ROC.	
A8 - Prototipar uma interface.	



Atividades	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
A1			X								
Apresentação do TCC1.				X							
A2							X				
A3										X	
A4								X			
A5									X		
A6										X	
Apresentação do TCC2.											X

Tabela 6 – Cronograma de atividades

## Referências

- BAIRD et al. Fertility and ageing. *Human Reproduction Update*, v. 11, n. 3, p. 261–276, 2005. Disponível em: <<https://academic.oup.com/humupd/article-abstract/11/3/261/759255>>. Citado 3 vezes nas páginas 23, 24 e 70.
- BASHIR, D. et al. An information-theoretic perspective on overfitting and underfitting. In: *The 33rd Australasian Joint Conference on Artificial Intelligence (AJCAI 2020)*. [s.n.], 2020. Disponível em: <<https://doi.org/10.48550/arXiv.2010.06076>>. Citado na página 39.
- BASTIDA, A. M. et al. Impact of different degrees of genetic mosaicism in the kinetic profile of the human embryo. *Fertility and Sterility*, v. 112, n. 6, p. 1069–1078, 2019. Disponível em: <[https://www.fertstert.org/article/S0015-0282\(19\)31316-0/fulltext](https://www.fertstert.org/article/S0015-0282(19)31316-0/fulltext)>. Citado na página 26.
- BOUCRET, L. et al. Change in the strategy of embryo selection with time-lapse system implementation—impact on clinical pregnancy rates. *Journal of Clinical Medicine*, v. 10, n. 18, p. 4111, 2021. Disponível em: <<https://www.mdpi.com/2077-0383/10/18/4111>>. Citado na página 23.
- CAMARGOS, V. P. et al. Imputação múltipla e análise de casos completos em modelos de regressão logística: uma avaliação prática do impacto das perdas em covariáveis. *Cadernos de Saúde Pública*, v. 27, n. 12, p. 2299–2313, 2011. Disponível em: <<https://www.scielo.br/j/csp/a/P4qJKtyRhKFfd3grkbfsL6y>>. Citado na página 54.
- CAPALBO, A. et al. Correlation between standard blastocyst morphology, euploidy and implantation: An observational study in two centers involving 956 screened blastocysts. *Human Reproduction*, v. 29, n. 6, 2014. Disponível em: <<https://academic.oup.com/humrep/article-abstract/29/6/1173/624854?login=false>>. Citado 3 vezes nas páginas 25, 26 e 71.
- CHEN, D. Y. *Análise de dados com Python e Pandas*. Novatec Editora, 2018. Disponível em: <<https://books.google.com.br/books?hl=pt-BR&lr=&id=ILFwDwAAQBAJ&oi=fnd&pg=PA31&dq=biblioteca+Pandas&ots=sQ0dSaigr&sig=lcdNNbH67MKJuXccwqIuisdtua4#v=onepage&q=biblioteca%20Pandas&f=false>>. Citado 2 vezes nas páginas 37 e 109.
- CORLETA, H. von E. Fertilização in vitro: mais de 4 milhões de crianças nascidas e um prêmio nobel. *Clinical and Biomedical Research*, v. 30, n. 4, p. 451–455, 2010. Disponível em: <<https://seer.ufrgs.br/index.php/hcpa/article/view/17351>>. Citado na página 18.
- CRUZ, M. et al. Timing of cell division in human cleavage-stage embryos is linked with blastocyst formation and quality. *Reproductive Biomedicine Online*, v. 25, n. 4, p. 371–381, 2012. Disponível em: <[https://www.sciencedirect.com/science/article/pii/S1472648312004099?casa\\_token=Q7hp0XipKegAAAAA:TQWmUPtUYztYC0Vp\\_cgchEh27w3nJpHB49gyD9b06zb4af2hO0jzrcCBWRbpPCLGghHSHhATOQ](https://www.sciencedirect.com/science/article/pii/S1472648312004099?casa_token=Q7hp0XipKegAAAAA:TQWmUPtUYztYC0Vp_cgchEh27w3nJpHB49gyD9b06zb4af2hO0jzrcCBWRbpPCLGghHSHhATOQ)>. Citado 5 vezes nas páginas 24, 26, 73, 106 e 107.

- DESAI, N. et al. Odds of euploidy are significantly associated with not only age but blastocyst morphokinetic parameters and icm/trophectoderm characteristics. *Fertility and Sterility*, v. 112, n. 6, p. 1016–1025, 2019. Disponível em: <[https://www.fertstert.org/article/S0015-0282\(19\)31117-3/fulltext](https://www.fertstert.org/article/S0015-0282(19)31117-3/fulltext)>. Citado na página 25.
- ELKAN, C. Nearest neighbor classification. v. 11, p. 3, 2011. Disponível em: <<https://cseweb.ucsd.edu/~elkan/250Bwinter2010/nearestn.pdf>>. Citado 2 vezes nas páginas 29 e 30.
- FAIRILITY™. Fairtility's ai-powered embryo quality assessment assistant. 2020. Disponível em: <<https://fairtility.com/chloe/>>. Citado na página 33.
- Figma. Design colaborativo online. 2024. Disponível em: <<https://www.figma.com/pt-br/design/>>. Citado na página 45.
- GARDNER, D. In vitro culture of human blastocysts. In: JANSEN, R.; MORTIMER, D. (Ed.). *Toward Reproductive Certainty: Fertility and Genetics Beyond*. Carnforth, UK: Parthenon Publishing, 1999. p. 378–388. Disponível em: <[https://journals.lww.com/co-obgyn/fulltext/1999/06000/culture\\_and\\_transfer\\_of\\_human\\_blastocysts.13.aspx](https://journals.lww.com/co-obgyn/fulltext/1999/06000/culture_and_transfer_of_human_blastocysts.13.aspx)>. Citado na página 25.
- GAZZO, E. et al. The kidscore d5 algorithm as an additional tool to morphological assessment and pgt-a in embryo selection: A time-lapse study. *JBRA Assisted Reproduction*, v. 24, p. 55–60, 2020. Disponível em: <<https://pmc.ncbi.nlm.nih.gov/articles/PMC6993168/>>. Citado 2 vezes nas páginas 26 e 106.
- GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: "O'Reilly Media, Inc.", 2017. Citado 7 vezes nas páginas 45, 46, 47, 48, 49, 50 e 51.
- GLEICHER, N.; PATRIZIO, P.; BRIVANLOU, A. Preimplantation genetic testing for aneuploidy—a castle built on sand. *Trends in molecular medicine*, Elsevier, v. 27, n. 8, p. 731–742, 2021. Disponível em: <[https://www.cell.com/trends/molecular-medicine/fulltext/S1471-4914\(20\)30313-0](https://www.cell.com/trends/molecular-medicine/fulltext/S1471-4914(20)30313-0)>. Citado 3 vezes nas páginas 21, 22 e 23.
- IZBICKI, R.; SANTOS, T. M. D. *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki, 2020. Disponível em: <[https://books.google.com.br/books?hl=pt-BR&lr=&id=6O8OEAAAQBAJ&oi=fnd&pg=PR13&dq=related:sXAZNf4-S38J:scholar.google.com/&ots=V9g1n31Nqk&sig=fjPChwYESoj3P8Jr6dbKM2\\_HAbA#v=onepage&q&f=false](https://books.google.com.br/books?hl=pt-BR&lr=&id=6O8OEAAAQBAJ&oi=fnd&pg=PR13&dq=related:sXAZNf4-S38J:scholar.google.com/&ots=V9g1n31Nqk&sig=fjPChwYESoj3P8Jr6dbKM2_HAbA#v=onepage&q&f=false)>. Citado 2 vezes nas páginas 27 e 28.
- JAISWAL, S. What is normalization in machine learning? a comprehensive guide to data rescaling. 2024. Disponível em: <<https://www.datacamp.com/tutorial/normalization-in-machine-learning>>. Citado 3 vezes nas páginas 37, 59 e 109.
- JARDIM, M. C. Política e emoções. *Revista de Ciências Sociais*, UNESP, v. 21, n. 51, 2022. Publicado em 2023-06-21. Disponível em: <<https://doi.org/10.5007/2175-7984.2022.e91402>>. Citado na página 13.
- JUNIOR, G. B. V. et al. Métricas utilizadas para avaliar a eficiência de classificadores em algoritmos inteligentes. *Centro de Pesquisas Avançadas em Qualidade de Vida*, v. 14, n. 2, 2022. ISSN 2178-7514. Disponível em: <[https://www.researchgate.net/publication/359541310\\_METRICAS\\_UTILIZADAS\\_PARA\\_AVALIAR\\_A\\_EFICIENCIA\\_DE\\_](https://www.researchgate.net/publication/359541310_METRICAS_UTILIZADAS_PARA_AVALIAR_A_EFICIENCIA_DE_)

CLASSIFICADORES\_EM\_ALGORITMOS\_INTELIGENTES>. Citado 4 vezes nas páginas 42, 43, 44 e 45.

KALOS, M. H.; WHITLOCK, P. A. *Monte Carlo Methods*. John Wiley & Sons, 2009. Disponível em: <<https://books.google.com.br/books?hl=pt-BR&lr=&id=5z-AI0pbNsYC&oi=fnd&pg=PR5&dq=Monte+Carlo+Methods+Malvin&ots=7rAR9jBwDl&sig=41VtfDAcq7Wg>>. Citado na página 111.

KATO, K. et al. Comparing prediction of ongoing pregnancy and live birth outcomes in patients with advanced and younger maternal age patients using kidscore<sup>TM</sup> day 5: A large-cohort retrospective study with single vitrified-warmed blastocyst transfer. *Reproductive Biology and Endocrinology*, v. 19, p. 98, 2021. Disponível em: <<https://pmc.ncbi.nlm.nih.gov/articles/PMC8252298/>>. Citado 2 vezes nas páginas 26 e 107.

KIAR, G. et al. Data augmentation through monte carlo arithmetic leads to more generalizable classification in connectomics. Montreal Neurological Institute, McGill University; Department of Computer Science and Computer Engineering, Concordia University, 2021. Disponível em: <<https://arxiv.org/pdf/2109.09649>>. Citado 3 vezes nas páginas 39, 88 e 111.

LASSEN, J. et al. Development and validation of deep learning-based embryo selection across multiple days of transfer. *arXiv preprint arXiv:2210.02120*, 2022. Disponível em: <<https://arxiv.org/abs/2210.02120>>. Citado na página 25.

LEAVER, M.; WELLS, D. Non-invasive preimplantation genetic testing (npgt): the next revolution in reproductive genetics? *Human Reproduction Update*, Oxford University Press on behalf of the European Society of Human Reproduction and Embryology, v. 25, n. 2, p. 241–255, 2019. Disponível em: <<https://academic.oup.com/humupd/article/26/1/16/5643748?login=false>>. Citado 2 vezes nas páginas 20 e 22.

LUONG, T.-M.-T.; LE, N. Q. K. Artificial intelligence in time-lapse system: advances, applications, and future perspectives in reproductive medicine. *Journal of Assisted Reproduction and Genetics*, v. 40, n. 12, p. 1205–1221, 2023. Disponível em: <<https://link.springer.com/article/10.1007/s10815-023-02973-y>>. Citado na página 13.

MESEGUER, M. et al. The use of morphokinetics as a predictor of embryo implantation. *Human reproduction*, Oxford University Press, v. 26, n. 10, p. 2658–2671, 2011. Disponível em: <<https://academic.oup.com/humrep/article/26/10/2658/611030?login=true>>. Citado na página 13.

MILEWSKI, R. et al. Morphokinetic parameters as a source of information concerning embryo developmental and implantation potential. *Ginekologia Polska*, v. 87, n. 10, p. 677–684, 2016. Disponível em: <<https://pubmed.ncbi.nlm.nih.gov/27958618/>>. Citado na página 37.

MONTAGNINI, H. M. L. et al. Estados emocionais de casais submetidos à fertilização in vitro. *Revista Brasileira de Ginecologia e Obstetrícia*, 2010. Disponível em: <<https://www.scielo.br/j/estpsi/a/hHphXxpTdNzZHt3PGqL3c7j/>>. Citado na página 14.

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística aplicada e probabilidade para engenheiros*. 4. ed. Rio de Janeiro: LTC, 2009. Citado na página 30.

- MOURA, M. D. de; SOUZA, M. do Carmo Borges de; SCHEFFER, B. B. Reprodução assistida. um pouco de história. *Revista da Sociedade Brasileira de Psicologia Hospitalar*, v. 12, n. 2, p. 1–7, 2020. Disponível em: <<https://pepsic.bvsalud.org/pdf/rsbph/v12n2/v12n2a04.pdf>>. Citado 2 vezes nas páginas 18 e 19.
- MOUSTAKLI, E. et al. Evolution of minimally invasive and non-invasive preimplantation genetic testing: An overview. *Journal of Clinical Medicine*, v. 13, n. 8, p. 2160, 2024. Disponível em: <<https://www.mdpi.com/2077-0383/13/8/2160>>. Citado 2 vezes nas páginas 23 e 31.
- MÜLLER, A. *Introduction to machine learning with python*. 2017. Citado 5 vezes nas páginas 46, 47, 48, 50 e 51.
- NASCIMENTO, F. P. do. Classificação da pesquisa: Natureza, método ou abordagem metodológica, objetivos e procedimentos. In: \_\_\_\_\_. *Metodologia da Pesquisa Científica: teoria e prática – como elaborar TCC*. Brasília: Thesaurus, 2016. cap. 6. Citado 2 vezes nas páginas 34 e 35.
- PANDIT, S.; SHARMA, R. Non invasive assessment of human oocytes and embryos in assisted reproduction: Review on present practices and future trends. *Medical Journal Armed Forces India*, v. 78, n. 1, p. 7–16, January 2022. Disponível em: <<https://pubmed.ncbi.nlm.nih.gov/35035038/>>. Citado na página 14.
- PEREIRA, K. K. P. C. S.; ALVES, O. de F. As principais técnicas de reprodução humana assistida. *Saúde & Ciência em ação*, v. 2, n. 1, p. 26–37, 2016. Disponível em: <<https://www.revistas.unifan.edu.br/index.php/RevistaICS/article/view/182>>. Citado 2 vezes nas páginas 18 e 19.
- PHILLIPS, K. R. B. et al. Temporal evaluation of a minimally invasive method of preimplantation genetic testing for aneuploidy (mi-pgt-a) in human embryos. *Reproductive Medicine*, v. 5, n. 3, p. 97–112, 2024. Disponível em: <<https://doi.org/10.3390/reprodmed5030011>>. Citado na página 22.
- PING, P. et al. Association of embryo aneuploidy and sperm dna damage in unexplained recurrent implantation failure patients under ngs-based pgt-a cycles. *Archives of Gynecology and Obstetrics*, v. 308, p. 997–1005, 2023. Disponível em: <<https://link.springer.com/article/10.1007/s00404-023-07098-2>>. Citado na página 13.
- RAMALHO, D. B. *Entrevistas concedidas a Maria Abritta e Sabrina Berno*. 2024. Citado 4 vezes nas páginas 54, 106, 107 e 108.
- REIGNIER, A. et al. Performance of day 5 kidscore<sup>TM</sup> morphokinetic prediction models of implantation and live birth after single blastocyst transfer. *Journal of assisted reproduction and genetics*, Springer, v. 36, p. 2279–2285, 2019. Disponível em: <<https://pubmed.ncbi.nlm.nih.gov/31444634/>>. Citado 2 vezes nas páginas 32 e 33.
- RESTREPO, L. F.; GONZÁLEZ, J. De pearson a spearman. *Revista Colombiana de Ciencias Pecuarias*, v. 20, n. 2, p. 183–192, 2007. Disponível em: <[http://www.scielo.org.co/scielo.php?pid=S0120-06902007000200010&script=sci\\_arttext](http://www.scielo.org.co/scielo.php?pid=S0120-06902007000200010&script=sci_arttext)>. Citado na página 112.

- RIENZI, L. et al. Time of morulation and trophectoderm quality are predictors of a live birth after euploid blastocyst transfer: a multicenter study. *Fertility and Sterility*, v. 113, n. 5, p. 991–998, 2020. Disponível em: <[https://www.fertstert.org/article/S0015-0282\(19\)31930-2/fulltext](https://www.fertstert.org/article/S0015-0282(19)31930-2/fulltext)>. Citado na página 25.
- RISH, I. et al. An empirical study of the naive bayes classifier. In: SEATTLE, WA, USA;. *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001. v. 3, n. 22, p. 41–46. Disponível em: <<https://www.dors.it/documentazione/testo/201911/10.1.1.330.2788.pdf>>. Citado 3 vezes nas páginas 30, 31 e 41.
- RODRIGUES, S. C. A. Modelo de regressão linear e suas aplicações. 2012. Disponível em: <<https://ubibliorum.ubi.pt/bitstream/10400.6/1869/1/Tese%20Sandra%20Rodrigues.pdf>>. Citado 2 vezes nas páginas 30 e 41.
- RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3rd. ed. Upper Saddle River, NJ: Pearson, 2016. Citado na página 27.
- SANTOS, C. Estatística descritiva—manual de auto-aprendizagem. *Edições Silabo: Lisboa*, 2007. Disponível em: <[https://www.researchgate.net/publication/311103840\\_Estatistica\\_Descritiva\\_Manual\\_de\\_auto-aprendizagem](https://www.researchgate.net/publication/311103840_Estatistica_Descritiva_Manual_de_auto-aprendizagem)>. Citado 2 vezes nas páginas 30 e 113.
- SATHYANARAYANAN; TANTRI, R. Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, November 30 2024. Disponível em: <<https://www.africanjournalofbiomedicalresearch.com/index.php/AJBR/article/view/4345>>. Citado na página 43.
- SHEIKH, H.; PRINS, C. Artificial intelligence: Definition and background. In: \_\_\_\_\_. *Mission AI*. Springer, 2023. Disponível em: <[https://link.springer.com/chapter/10.1007/978-3-031-21448-6\\_2](https://link.springer.com/chapter/10.1007/978-3-031-21448-6_2)>. Citado na página 27.
- SILVA, A. E. et al. Casais com abortamento espontâneo recorrente: participação das translocações cromossômicas / couples with recurrent miscarriages: contributions of chromosome translocations. *Revista de Ciências Biológicas e da Saúde*, UNESP – Universidade Estadual Paulista, Campus de São José do Rio Preto-SP, v. 14, n. 4, p. e249, 2023. Disponível em: <[https://repositorio-racs.famerp.br/racs\\_ol/vol-14-4/ID249.pdf](https://repositorio-racs.famerp.br/racs_ol/vol-14-4/ID249.pdf)>. Citado na página 13.
- SOTO, T. Regression analysis. In: *Encyclopedia of Autism Spectrum Disorders*. Cham: Springer International Publishing, 2021. p. 3906–3906. Disponível em: <[https://link.springer.com/referenceworkentry/10.1007/978-3-319-91280-6\\_251](https://link.springer.com/referenceworkentry/10.1007/978-3-319-91280-6_251)>. Citado na página 30.
- SOUSA Áurea. Coeficiente de correlação de pearson e coeficiente de correlação de spearman. o que medem e em que situações devem ser utilizados? *Correio dos Açores: Matemática*, p. 19, mar 2019. Disponível em: <[https://repositorio.uac.pt/bitstream/10400.3/5365/1/Sousa\\_CA\\_21%20Mar%20c3%a7o%202019.pdf](https://repositorio.uac.pt/bitstream/10400.3/5365/1/Sousa_CA_21%20Mar%20c3%a7o%202019.pdf)>. Citado 4 vezes nas páginas 38, 54, 109 e 112.
- SOUZA, M. C. de. As técnicas de reprodução assistida. a barriga de aluguel. a definição da maternidade e da paternidade. bioética. *Revista EMERJ*, v. 50, p. 348–366, 2024. Disponível em: <[https://www.emerj.tjrj.jus.br/revistaemerj\\_online/edicoes/revista50/Revista50\\_348.pdf](https://www.emerj.tjrj.jus.br/revistaemerj_online/edicoes/revista50/Revista50_348.pdf)>. Citado na página 18.



SOUZA, R. C. M. de. *Análise da ploidia de embriões humanos por meio da Inteligência Artificial com o uso de variáveis de morfologia, morfocinética e variáveis relacionadas com a paciente*. Tese (Doutorado) — UNESP, 2022. Disponível em: [https://bdtd.ibict.br/vufind/Record/UNSP\\_2209d12b9be14edf04d513c039a770d8](https://bdtd.ibict.br/vufind/Record/UNSP_2209d12b9be14edf04d513c039a770d8). Citado 3 vezes nas páginas 24, 32 e 36.

SOUZA, R. C. M. de. *Dissertação de mestrado: Farmacologia e Biotecnologia*. Dissertação (Mestrado) — Universidade Estadual Paulista (Unesp), June 2022. Disponível em: <https://repositorio.unesp.br/items/02fb8a4e-c577-4aa1-aa7e-f259f23e8353>. Citado na página 19.

TRASK, A. W. *Grokking: Deep Learning*. New York: Manning Publications Co., 2019. Disponível em: <https://edu.anarcho-copy.org/Algorithm/grokking-deep-learning.pdf>. Citado na página 27.

WANG, Z. et al. A comprehensive survey on data augmentation. *arXiv preprint arXiv:2405.09591*, 2024. Disponível em: <https://arxiv.org/abs/2405.09591>. Citado na página 39.

YANG, H. et al. Preimplantation genetic testing for aneuploidy: challenges in clinical practice. *Nature Reviews Genetics*, v. 25, n. 1, p. 30–45, 2024. Disponível em: <https://link.springer.com/article/10.1186/s40246-022-00442-8>. Citado 4 vezes nas páginas 13, 15, 19 e 20.

YUAN, Z. et al. Development of an artificial intelligence based model for predicting the euploidy of blastocysts in pgt-a treatments. *Scientific Reports*, v. 13, 2023. Disponível em: <https://www.nature.com/articles/s41598-023-29319-z>. Citado 6 vezes nas páginas 23, 24, 31, 32, 36 e 107.

Z-SCORE: saiba o que é e como funciona. jul 2022. Disponível em: <https://maisretorno.com/portal/termos/z/z-score>. Citado na página 109.

ZEGERS-HOCHSCHILD, F. et al. The international glossary on infertility and fertility care, 2017. *Human reproduction*, Oxford University Press, v. 32, n. 9, p. 1786–1801, 2017. Disponível em: <https://academic.oup.com/humrep/article/32/9/1786/4049537?login=false>. Citado 2 vezes nas páginas 14 e 20.

ZHANG, H. The optimality of naive bayes. *Aa*, v. 1, n. 2, p. 3, 2004. Disponível em: <https://cdn.aaai.org/FLAIRS/2004/Flairs04-097.pdf>. Citado 2 vezes nas páginas 30 e 31.

ZHANG, Z. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, v. 4, n. 11, 2016. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4916348/>. Citado 2 vezes nas páginas 29 e 41.

## GLOSSÁRIO

Aneuploidia	Condição em que o número de cromossomos não é múltiplo exato de 23.
Blastocisto	Estágio do desenvolvimento embrionário, que ocorre cerca de 5 a 6 dias após a fecundação, apresentando uma cavidade interna e as primeiras divisões celulares mais complexas.
Blastômeros	Células resultantes das divisões iniciais do embrião, multiplicando-se durante as primeiras fases de desenvolvimento.
Biópsia	Procedimento médico para retirar uma pequena amostra de tecido ou células para análise. Utilizado em reprodução assistida para avaliar a saúde do embrião.
Citoplasma	Parte do conteúdo celular que envolve o núcleo e onde ocorrem várias funções vitais da célula, como metabolismo e síntese de proteínas.
Clivagem	Processo de divisão celular do embrião, onde uma célula inicial se divide sucessivamente em células menores chamadas blastômeros.
Cromossomos	Estruturas presentes no núcleo celular que carregam a informação genética. Os humanos têm 23 pares de cromossomos, totalizando 46.
Dados Morfocinéticos	Análise dos parâmetros morfológicos e cinéticos do embrião, como sua estrutura e o movimento/desenvolvimento ao longo do tempo.
Euploidia	Condição em que o número de cromossomos é múltiplo exato de 23.
Gravidez Clínica	Definição de uma gestação confirmada por ultrassonografia, com presença de embrião e batimento cardíaco fetal no útero.
Implantação de Embrião	Processo no qual o embrião se fixa e se insere na parede do útero, iniciando o desenvolvimento da gravidez.
Massa Celular Interna do Embrião	Conjunto de células do blastocisto que dará origem ao feto durante o desenvolvimento da gestação.



---

Mosaico	Tipo de aneuploidia em que algumas células têm o número correto de cromossomos, enquanto outras têm um número alterado.
Ploidias	Refere-se ao número de conjuntos de cromossomos em uma célula ou organismo.
Trofectoderma	Camada externa do blastocisto que dará origem à placenta, sendo essencial para a implantação do embrião no útero.

## Apêndices

## APÊNDICE A – Variáveis utilizadas na análise da ploidia embrionária

Este anexo descreve as variáveis utilizadas na planilha de dados referente ao desenvolvimento embrionário e sua relação com a ploidia. As variáveis incluem informações demográficas, temporais e morfológicas do embrião, bem como métricas de qualidade baseadas no sistema Gardner. Todos os tempos descritos abaixo são expressos em horas.

### • Variáveis Gerais

- **Id:** Identificador numérico de cada paciente.
- **Idade:** Idade da paciente no momento do procedimento.
- **Data da biópsia:** Data em que foi realizada a biópsia embrionária.
- **Embrião n.:** Identificação numérica do embrião dentro do ciclo de fertilização.
- **Estágio:** Dia de evolução no cultivo (5º dia ou 6º dia) ([RAMALHO, 2024](#)).
- **Morfo:** Classificação morfológica dos embriões baseada no estágio de expansão da blástula, estágio inicial do desenvolvimento embrionário que ocorre após a segmentação (divisões celulares iniciais) do zigoto. As notas de 1 a 5 são a expansão do embrião, com 1 sendo o menos expansivo e 5 o mais expansivo ([RAMALHO, 2024](#)).
- **KIDScore™:** Algoritmo combina variáveis morfocinéticas e parâmetros de desenvolvimento embrionário. A pontuação vai de 0 a 10 por conta de estarmos utilizando embriões de estado de blastocisto ([GAZZO et al., 2020](#)).

### • Variáveis Temporais

As variáveis temporais são baseadas nos intervalos de tempo entre eventos específicos durante o desenvolvimento do embrião:

- **st2:** Primeiro indício de movimentos citoplasmáticos antes da primeira citocinese. É a fase final da divisão celular, onde o citoplasma é dividido entre as duas células filhas ([RAMALHO, 2024](#)).
- **t2:** Tempo para 2 células. Tempo necessário para o embrião completar a primeira clivagem, ou seja, a divisão da célula-ovo (zigoto) em duas células, entre 24,3 – 27,9 horas após a fertilização ([CRUZ et al., 2012](#)).
- **t3:** Tempo para 3 células. Marca o momento em que o embrião se divide de duas para três células, entre 35,4 – 40,3 horas após a fertilização ([CRUZ et al., 2012](#)).

- **t4**: Tempo para 4 células. Transição do embrião para o estágio de quatro células.
- **t5**: Tempo para 5 células. Momento em que o embrião alcança o estágio de cinco células, entre 48,8 – 56,6 horas após a fertilização (CRUZ et al., 2012).
- **t8**: Tempo para 8 células.
- **tSC**: Tempo de formação do estágio de clivagem sincronizada (Time to Synchronized Compaction). Representa o tempo necessário para que o embrião, após atingir o estágio de 8 células, comece a apresentar compactação sincronizada. A compactação consiste na união mais forte das células embrionárias (KATO et al., 2021).
- **tSB**: Tempo para o início da blastulação (Time to Start Blastulation). Tempo necessário para o embrião iniciar a formação do blastocisto (KATO et al., 2021).
- **tB**: Tempo para o Blastocisto (Time to Blastocyst). Tempo que leva para o embrião alcançar o estágio de blastocisto completo, que é o último estágio de desenvolvimento embrionário antes da implantação no útero (YUAN et al., 2023).
- **t2-st2**: Intervalo de tempo entre o T2 e o ST2. Esse intervalo é usado para avaliar a regularidade e a qualidade das divisões celulares no estágio inicial do desenvolvimento embrionário (RAMALHO, 2024).
- **cc2 (t3-t2)**: Tempo necessário para que o embrião passe da divisão de 2 células (T2) para a divisão de 3 células (T3). É utilizado para avaliar a regularidade e a dinâmica do ciclo celular inicial do embrião (RAMALHO, 2024).
- **cc3 (t5-t3)**: Tempo necessário para que o embrião passe da divisão de 3 células (T3) para a divisão de 5 células (T5). É utilizado como um indicador da dinâmica do ciclo celular (RAMALHO, 2024).
- **t5-t2**: Intervalo de tempo entre o estágio de 2 células (T2) e o estágio de 5 células (T5). É uma métrica importante para avaliar a eficiência das divisões celulares iniciais (RAMALHO, 2024).
- **s2 (t4-t3)**: Intervalo de tempo necessário para que o embrião passe do estágio de 3 células (T3) para o estágio de 4 células (T4). Este parâmetro é usado para avaliar a regularidade da divisão celular (RAMALHO, 2024).
- **s3 (t8-t5)**: Intervalo de tempo necessário para que o embrião passe do estágio de 5 células (T5) para o estágio de 8 células (T8). É usado para avaliar a sincronização e a regularidade do ciclo celular (RAMALHO, 2024).
- **tSC-t8**: Intervalo de tempo entre o estágio em que o embrião atinge a compactação inicial (tSC) e o estágio de 8 células (T8). É usado para avaliar a

transição do embrião das divisões celulares iniciais para o início da compactação (RAMALHO, 2024).

- **tB-tSB:** Intervalo de tempo entre o estágio em que o embrião atinge o blastocisto inicial (tSB) e o estágio de blastocisto expandido (tB). Avalia o tempo necessário para que o embrião progrida do início da formação do blastocisto até a sua expansão completa (RAMALHO, 2024).

- **Variável de Resultado**

- **Ploidia:** Estado de ploidia do embrião, resultado final da análise.

## APÊNDICE B – Z-score normalization (standardization)

O Z-Score é um indicador numérico que ilustra a conexão entre uma determinada quantia e a média de um conjunto de valores, expressa em termos de desvios padrão. Quando o Z-Score de um dado é zero, isso sugere que ela coincide com a média do conjunto. Um Z-Score de 1 indica que o valor está um desvio padrão acima da média, ao passo que valores negativos sugerem que estão abaixo da média. A normalização pelo Z-Score, também conhecida como padronização, pressupõe que os dados sigam uma distribuição gaussiana (em forma de sino) e transforma os valores para que tenham uma média ( $\mu$ ) de 0 e um desvio padrão ( $\sigma$ ) de 1, facilitando a análise e a comparação entre variáveis com diferentes escalas (JAISWAL, 2024; Z-SCORE... , 2022). A fórmula para padronização é:

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

Em que:

- $x$ : Valor original do dado;
- $\mu$ : Média da variável;
- $\sigma$ : Desvio padrão da variável.

O Z-Score é um instrumento valioso para a identificação se uma pontuação é típica ou atípica em comparação a um conjunto de dados previamente definido. Esta métrica permite a comparação entre pontuações de variados conjuntos de dados, o que torna as análises mais exatas e uniformes. O Z-Score, apesar de ser sensível a outliers (valores atípicos que estão muito distantes da maioria dos outros dados) e depender do intervalo de dados, é benéfico em circunstâncias onde a preservação do intervalo original é crucial (SOUSA, 2019; Z-SCORE... , 2022).

Também é crucial reconhecer as variáveis numéricas que serão convertidas, já que o Z-Score só se aplica a variáveis contínuas. Assim, cada valor na variável será transformado em um Z-Score, representando quantos desvios padrão ele está acima ou abaixo da média da variável.

Para normalizar a tabela de dados usando o método Z-Score, utilizaremos o Pandas, que é uma biblioteca Python de código aberto para análise de dados (CHEN, 2018). Inicialmente, importaremos a planilha Excel com os dados originais. Depois, são identificadas e ordenadas manualmente as colunas numéricas a serem normalizadas. Uma réplica do original DataFrame é gerada para manter os dados inalterados, enquanto as variáveis selecionadas são modificadas utilizando a fórmula do Z-Score: cada valor é ajustado pela subtração da média da variável e divisão pelo seu desvio padrão. Este procedimento assegura que as variáveis sejam ajustadas para uma média de 0 e um desvio padrão de 1, removendo variações de escala entre elas. Assim, a tabela normalizada é armazenada em um novo arquivo Excel. Após a aplicação do Z-score, se faz a verificação da média e do desvio padrão das variáveis transformadas, em que a média ( $\mu$ ) deve ser próxima de 0 e o desvio padrão ( $\sigma$ ) deve ser próximo de 1.

## APÊNDICE C – Monte Carlo

O Método de Aritmética de Monte Carlo (MCA) é uma técnica matemática e computacional amplamente utilizada para resolver problemas que envolvem incerteza e variabilidade (KALOS; WHITLOCK, 2009). Baseia-se no uso de números aleatórios para simular processos estocásticos, ou seja, processos que evoluem de maneira dependente de eventos aleatórios (KALOS; WHITLOCK, 2009). Esse método é especialmente eficaz para lidar com problemas de alta complexidade matemática, onde métodos analíticos tradicionais podem ser inviáveis. Além de que “o aumento do conjunto de dados foi amplamente demonstrado como uma técnica eficaz para melhorar a generalização de modelos de aprendizado” (KIAR et al., 2021).

A Simulação de Monte Carlo trabalha com o princípio de geração de valores aleatórios dentro de um intervalo previamente definido para variáveis que apresentam incerteza (KALOS; WHITLOCK, 2009). Esses valores aleatórios são extraídos de uma distribuição de probabilidade específica, como a distribuição uniforme ou normal, dependendo do problema. O método consiste em repetir o cálculo de um modelo várias vezes, cada vez utilizando um conjunto diferente de valores aleatórios como entrada (KALOS; WHITLOCK, 2009).

Diferentemente de modelos de previsão tradicionais, que trabalham com valores fixos, o Monte Carlo oferece uma gama de resultados possíveis e a probabilidade associada a cada um. Isso permite uma análise mais detalhada e uma maior flexibilidade para lidar com incertezas.

Embora os números usados no método sejam chamados de aleatórios, em implementações computacionais eles são, na verdade, gerados por algoritmos que criam números pseudoaleatórios (KALOS; WHITLOCK, 2009). Esses números imitam propriedades de números verdadeiramente aleatórios, mas são derivados de um processo determinístico (KALOS; WHITLOCK, 2009). É “importante ressaltar que essa técnica produz uma gama de resultados igualmente plausíveis, onde nenhuma observação é mais ou menos válida do que as outras — incluindo aquelas que não foram perturbadas” (KIAR et al., 2021).

Para aplicar a técnica de Monte Carlo na base de dados, iniciaremos analisando as variáveis numéricas do conjunto de treinamento, determinando suas distribuições. Após isso, usaremos essas distribuições para gerar valores aleatórios utilizando funções como *numpy.random.normal()* para distribuições normais. Esses valores serão usados para criar pequenas variações nas variáveis originais, aumentando a diversidade do conjunto de dados. Para finalizar, combinaremos os dados originais com os novos dados gerados, ga-



---

rantindo que os padrões estatísticos sejam mantidos e validaremos o conjunto expandido para assegurar que as novas amostras respeitam as características do conjunto original.

## APÊNDICE D – Coeficiente de Spearman

O coeficiente de correlação de Spearman é uma ferramenta estatística bastante útil quando trabalhamos com dados que não seguem uma distribuição normal ou apresentam outliers (valores atípicos que estão muito distantes da maioria dos outros dados). Isso ocorre porque o coeficiente de Spearman não usa os valores originais dos dados, mas sim as ordens ou posições em que as observações são classificadas (SOUSA, 2019). Essa abordagem torna o coeficiente mais robusto (menos sensível) a distorções causadas por assimetria nos dados (quando os dados se distribuem de forma desigual) e por outliers (SOUSA, 2019).

O coeficiente de Spearman é uma medida não paramétrica, o que significa que ele não depende de pressupostos sobre a distribuição dos dados, como a necessidade de os dados serem normalmente distribuídos (um tipo de distribuição comum em estatísticas) (RESTREPO; GONZÁLEZ, 2007). Ele é usado para medir o grau de associação monotônica entre duas variáveis (RESTREPO; GONZÁLEZ, 2007).

Uma relação monotônica entre duas variáveis é uma relação em que uma variável tende a aumentar ou diminuir à medida que a outra também aumenta ou diminui, mas essa mudança não precisa ser em uma linha reta (RESTREPO; GONZÁLEZ, 2007). Em outras palavras, a direção da mudança nas variáveis é constante, mas não necessariamente linear.

O coeficiente de Spearman atribui um posto (ou ranking) a cada valor das variáveis (SOUSA, 2019). Isso significa que, ao invés de olhar diretamente para os valores das variáveis, ele compara a posição relativa dos dados em cada variável. Por exemplo, se temos uma variável que mede a idade e outra que mede a altura, em vez de comparar diretamente a idade e a altura, o coeficiente compara as posições relativas (ranks) de cada dado nas duas variáveis. Depois de classificar os dados dessa forma, o coeficiente de Spearman avalia a correlação entre essas classificações, ou seja, verifica o quanto as posições (ranks) das variáveis estão relacionadas entre si (SOUSA, 2019).

A fórmula do coeficiente de Spearman é dada por:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Em que:

- **di**: Diferença entre os postos de cada par de observações,
- **n**: Número de observações.

Os valores do coeficiente de Spearman variam entre -1 e +1. Utilizaremos esses valores para interpretar a força e a direção da relação entre as variáveis segundo Santos (2007) na tabela 7:

Tabela 7. Interpretação do coeficiente de correlação de Spearman

Coeficiente de Correlação	Correlação
$R_{xy} = 1$	Perfeita positiva
$0,8 \leq R_{xy} < 1$	Forte positiva
$0,5 \leq R_{xy} < 0,8$	Moderada positiva
$0,1 \leq R_{xy} < 0,5$	Fraca positiva
$0 \leq R_{xy} < 0,1$	Infima positiva
0	Nula
$-0,1 \leq R_{xy} < 0$	Infima negativa
$-0,5 \leq R_{xy} < -0,1$	Fraca negativa
$-0,8 \leq R_{xy} < -0,5$	Moderada negativa
$-1 \leq R_{xy} < -0,8$	Forte negativa
$R_{xy} = -1$	Perfeita negativa

Fonte: (SANTOS, 2007)

Na base de dados, o coeficiente de Spearman será aplicado ao transformar os valores originais das variáveis em ranks e calcular a correlação entre eles usando a função `spearmanr()` da biblioteca *SciPy*. Isso permitirá identificar relações monotônicas entre as variáveis de forma robusta contra outliers. Com os resultados obtidos, determinaremos quais variáveis têm maior influência na previsão de euploidia e justificaremos com base nos coeficientes e nas visualizações geradas.

# Anexos

# ANEXO I – Parecer do Comitê de Ética em Pesquisa



## PARECER CONSUBSTANCIADO DO CEP

### DADOS DO PROJETO DE PESQUISA

**Título da Pesquisa:** Parâmetros do time-lapse relacionados à ploidia de embriões humanos: investigando o uso da tecnologia para a seleção não invasiva de embriões euploides

**Pesquisador:** BRUNO RAMALHO DE CARVALHO

**Área Temática:**

**Versão:** 2

**CAAE:** 71313923.1.0000.0023

**Instituição Proponente:** Centro Universitário de Brasília - UNICEUB

**Patrocinador Principal:** Financiamento Próprio

### DADOS DO PARECER

**Número do Parecer:** 6.313.392

#### **Apresentação do Projeto:**

As informações elencadas nos campos "Apresentação do Projeto", "Objetivo da Pesquisa" e "Avaliação dos Riscos e Benefícios" foram retiradas do documento de Informações Básicas da Pesquisa postado na Plataforma Brasil.

#### **- TIPO DO ESTUDO:**

Estudo observacional. Pretende-se analisar retrospectivamente os dados da fotografia time-lapse de blastocistos humanos biopsiados e, portanto, destinados ao diagnóstico genético pré-implantação (para pesquisa de aneuploidias), em ciclos de tratamento ocorridos entre 1º de janeiro de 2022 e 31 de dezembro de 2023.

#### **- NÚMERO PARTICIPANTE DAS PESQUISA: 50**

**- DESCRIÇÃO DOS PARTICIPANTES:** Os participantes da pesquisa serão pacientes da clínica Bruno Ramalho Reprodução Humana, que realiza tratamentos de reprodução assistida em parceria com a GENESIS Centro de Assistência em Reprodução Humana, em Brasília, Distrito Federal.

**- FORMA DE RECRUTAMENTO DOS PARTICIPANTES:** Os participantes das pesquisa serão abordados

**Endereço:** SEPN 707/907 - Bloco 6, sala 6.205, 2º andar

**Bairro:** Setor Universitário

**CEP:** 70.790-075

**UF:** DF

**Município:** BRASÍLIA

**Telefone:** (61)3966-1511

**E-mail:** cep.uniceub@uniceub.br

Continuação do Parecer: 6.313.392

pessoalmente pelo pesquisador responsável, Bruno Ramalho de Carvalho, por contato telefônico (retrospectivamente) e durante a consulta médica (prospectivamente), antes de iniciarem tratamento de reprodução assistida destinado a teste pré-implantação para pesquisa de aneuploidias (PGT-A). O termo de consentimento livre e esclarecido (TCLE) será enviado por meio digital, via e-mail, pelo sistema de prontuário da clínica Bruno Ramalho Reprodução Humana, e será, da mesma forma, assinado digitalmente pelos participantes. O sistema não permite adição do timbre ao documento. O armazenamento do TCLE poderá ser feito por ambas as partes em formato Portable Document Format (PDF).

- CRITÉRIOS DE INCLUSÃO: Serão incluídos na pesquisa todos os embriões cultivados em incubadora com tecnologia time-lapse e biopsiados para PGT-A, entre 01 de janeiro de 2023 e 31 de dezembro de 2023, sem restrições quanto a idade, classe social ou outras quaisquer.

- CRITÉRIOS DE EXCLUSÃO: Não informado.

- LOCAL ONDE SERÁ REALIZADO O ESTUDO: clínica Bruno Ramalho Reprodução Humana e GENESIS Centro de Assistência em Reprodução Humana.

- PROCEDIMENTOS QUE SERÃO REALIZADOS COM OS PARTICIPANTES: Coleta de dados de prontuários.

- MÉTODO DE COLETA DE DADOS/INFORMAÇÕES:

Pretende-se analisar retrospectivamente os dados da fotografia time-lapse de blastocistos humanos biopsiados e, portanto, destinados ao diagnóstico genético pré-implantação (para pesquisa de aneuploidias), em ciclos de tratamento ocorridos entre 1º de janeiro de 2022 e 31 de dezembro de 2023.

Variáveis a serem analisadas: início de t2 (st2), correspondendo aos primeiros movimentos citoplasmáticos anteriores à primeira citocinese (o primeiro frame detectável dos seguintes movimentos: desaparecimento do halo, ondas citoplasmáticas ou movimentos do citoplasma circulante); t3; t5; tSB; tb; cc3; e t5–t2, e outros parâmetros eventualmente identificados como relevantes; e o resultado da inteligência artificial KIDSCORE D5.

- METODOLOGIA DE ANÁLISE DE DADOS: Os dados serão tabulados e analisados estatisticamente,

**Endereço:** SEPN 707/907 - Bloco 6, sala 6.205, 2º andar

**Bairro:** Setor Universitário

**CEP:** 70.790-075

**UF:** DF

**Município:** BRASILIA

**Telefone:** (61)3966-1511

**E-mail:** cep.uniceub@uniceub.br

Continuação do Parecer: 6.313.392

com auxílio de software especializado.

**Objetivo da Pesquisa:**

Objetivo Primário: Identificar se existe correlação positiva entre o parâmetro de clivagem precoce t2e a ploidia de embriões humanos.

Objetivo Secundário: Existe correlação positiva entre os parâmetros de clivagem precoce t3, t5, tSB, tB, cc3 e t5-t2, e outros a serem identificados, e a ploidia de embriões humanos.

**Avaliação dos Riscos e Benefícios:**

RISCOS: Não há riscos às pacientes, uma vez que a análise será retrospectiva, ou seja, não haverá interferência da pesquisa sobre o desenvolvimento dos tratamentos. Os riscos potenciais relacionados à pesquisa envolvem a quebra da confidencialidade e, por esse motivo, os pesquisadores envolvidos, sob coordenação do pesquisador responsável, comprometem-se a manter em sigilo quanto à identidade dos genitores dos embriões incluídos, bem como a quaisquer dados que possibilitem a quebra do anonimato.

BENEFÍCIOS: Desenvolver ferramentas não invasivas de avaliação embrionária (quanto à ploidia), que possam substituir de forma satisfatória a biopsia embrionária para pesquisa de aneuploidias.

**Comentários e Considerações sobre a Pesquisa:**

- Devido à natureza do estudo, considera-se a pesquisa com risco mínimo.
- Houve indicação correta das medidas protetivas para o risco apresentado.
- Orçamento: os gastos serão custeados pelo pesquisador.
- Cronograma: A coleta de dados está prevista para iniciar-se em agosto de 2023.

**Considerações sobre os Termos de apresentação obrigatória:**

- Apresentou a Folha de Rosto devidamente preenchida e assinada.
- Apresentou o Termo de Aceite Institucional devidamente preenchido e assinado.
- O Termo de Consentimento Livre e Esclarecido (TCLE) foi apresentado.

**Recomendações:**

Ao final do estudo os pesquisadores devem enviar o Relatório de Finalização da Pesquisa ao CEP. O envio de relatórios deverá ocorrer pela Plataforma Brasil, por meio de notificação.

**Endereço:** SEPN 707/907 - Bloco 6, sala 6.205, 2º andar

**Bairro:** Setor Universitário

**CEP:** 70.790-075

**UF:** DF

**Município:** BRASILIA

**Telefone:** (61)3966-1511

**E-mail:** cep.uniceub@uniceub.br



Continuação do Parecer: 6.313.392

**Conclusões ou Pendências e Lista de Inadequações:**

O pesquisador atendeu às solicitações indicadas pelo CEP:

- O pesquisador deverá inserir no TCLE o esclarecimento sobre os riscos de quebra de confidencialidade e as medidas protetivas, conforme informado nas informações básicas do projeto que constam na Plataforma Brasil: "Os riscos potenciais relacionados à pesquisa envolvem a quebra da confidencialidade e, por esse motivo, os pesquisadores envolvidos, sob coordenação do pesquisador responsável, comprometem-se a manter em sigilo quanto à identidade dos genitores dos embriões incluídos, bem como a quaisquer dados que possibilitem a quebra do anonimato."

O pesquisador realizou as seguintes adequações ao projeto:

1. Informar o local onde será realizado o estudo (nome da clínica) e inserir o TERMO DE ACEITE INSTITUCIONAL DEVIDAMENTE PREENCHIDO E ASSINADO. Conforme estabelecido na resolução 466/2012 do Conselho Nacional de Saúde, a instituição coparticipante de pesquisa é a organização, pública ou privada, legitimamente constituída e habilitada, na qual alguma das fases ou etapas da pesquisa se desenvolve. PENDÊNCIA ATENDIDA

2. Descrever a forma de abordagem dos participante das pesquisa. O pesquisador deverá informar como e em que momento serão abordados os convidados a participantes de pesquisa, com a descrição do processo e do registro do consentimento. A Norma Operacional Nº 001/2013 do Conselho Nacional de Saúde / Ministério da Saúde que dispõe sobre a organização e funcionamento do Sistema CEP/CONEP e sobre os procedimentos para submissão, avaliação e acompanhamento da pesquisa e de desenvolvimento envolvendo seres humanos no Brasil define, no item 3.4.1, que todos os protocolos de pesquisa devem conter, obrigatoriamente, A DESCRIÇÃO DA FORMA DE ABORDAGEM OU PLANO DE RECRUTAMENTO DOS POSSÍVEIS INDIVÍDUOS PARTICIPANTES. PENDÊNCIA ATENDIDA

3. Inserir a descrição dos participantes da pesquisa. Para analisar a eticidade do projeto de pesquisa, a pesquisa deverá conter as características da população a ser estudada, como faixa etária, sexo, grupos sociais e outras que sejam pertinentes à descrição da população e que

**Endereço:** SEPN 707/907 - Bloco 6, sala 6.205, 2º andar

**Bairro:** Setor Universitário

**CEP:** 70.790-075

**UF:** DF

**Município:** BRASILIA

**Telefone:** (61)3966-1511

**E-mail:** cep.uniceub@uniceub.br

Continuação do Parecer: 6.313.392

possam ser significativas para a caracterização da amostra, população e análise ética da pesquisa.  
PENDÊNCIA ATENDIDA

4. Informar os CRITÉRIOS DE INCLUSÃO e de EXCLUSÃO dos participantes da pesquisa. PENDÊNCIA ATENDIDA. O pesquisador informou o critério de inclusão. Entende-se que os critérios de exclusão poderão ser estabelecidos após a consulta aos dados dos prontuários.

5. O Termo de Consentimento Livre e Esclarecido (TCLE) não foi apresentado de forma adequada. O TCLE necessita de adequações.

5.1. Colocar o nome e logotipo da instituição proponente (CEUB); JUSTIFICATIVA APRESENTADA

5.2. As Resoluções em Pesquisa com seres humanos do Conselho Nacional de Saúde asseguram aos participantes da pesquisa a possibilidade de contato com o CEP que aprovou o projeto de pesquisa. O TCLE deve conter o contato do CEP UniCEUB para que o participante da pesquisa possa fazer considerações, tirar dúvidas ou informar ocorrências relacionadas ao estudo. Para isso o pesquisador deve inserir os dados do CEP do UniCEUB no TCLE: Comitê de Ética em Pesquisa do Centro Universitário de Brasília – CEP/UniCEUB. Telefone 3966.1511 E-mail cep.uniceub@uniceub.br. PENDÊNCIA ATENDIDA

5.3. Inserir o contato dos pesquisadores no TCLE. Conforme especificado na resolução 466/2012 do CNS, no TCLE deve constar o endereço e contato telefônico dos responsáveis pela pesquisa. Considerando o disposto na Resolução nº 466/2012, do Conselho Nacional de Saúde, do Ministério da Saúde, no que se refere ao PROCESSO DE CONSENTIMENTO LIVRE E ESCLARECIDO, consta em seu item IV.5 d) que o TCLE deve “ser elaborado em duas vias, rubricadas em todas as suas páginas e assinadas, ao seu término, pelo convidado a participar da pesquisa, ou por seu representante legal, assim como pelo pesquisador responsável, ou pela(s) pessoa(s) por ele delegada(s), devendo as páginas de assinaturas estar na mesma folha. EM AMBAS AS VIAS DEVERÃO CONSTAR O ENDEREÇO E CONTATO TELEFÔNICO DOS RESPONSÁVEIS PELA PESQUISA E DO CEP LOCAL E DA CONEP, QUANDO PERTINENTE.” PENDÊNCIA ATENDIDA.

----

O CEP-UniCEUB ressalta a necessidade de desenvolvimento da pesquisa, de acordo com o

**Endereço:** SEPN 707/907 - Bloco 6, sala 6.205, 2º andar

**Bairro:** Setor Universitário

**CEP:** 70.790-075

**UF:** DF

**Município:** BRASÍLIA

**Telefone:** (61)3966-1511

**E-mail:** cep.uniceub@uniceub.br

Continuação do Parecer: 6.313.392

protocolo avaliado e aprovado, bem como, atenção às diretrizes éticas nacionais quanto aos incisos XI.1 e XI.2 da Resolução nº 466/12 CNS/MS concernentes às responsabilidades do pesquisador no desenvolvimento do projeto:

XI.1 - A responsabilidade do pesquisador é indelegável e indeclinável e compreende os aspectos éticos e legais.

XI.2 - Cabe ao pesquisador:

- c) desenvolver o projeto conforme delineado;
- d) elaborar e apresentar os relatórios parciais e final;
- e) apresentar dados solicitados pelo CEP ou pela CONEP a qualquer momento;
- f) manter os dados da pesquisa em arquivo, físico ou digital, sob sua guarda e responsabilidade, por um período de 5 anos após o término da pesquisa;
- g) encaminhar os resultados da pesquisa para publicação, com os devidos créditos aos pesquisadores associados e ao pessoal técnico integrante do projeto; e
- h) justificar fundamentadamente, perante o CEP ou a CONEP, interrupção do projeto ou a não publicação dos resultados.

#### Considerações Finais a critério do CEP:

Protocolo previamente avaliado, com parecer homologado na 16ª Reunião Ordinária do CEP-UniCEUB do ano em 15 de setembro de 2023.

#### Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_2138177.pdf	21/08/2023 12:51:16		Aceito
Recurso Anexado pelo Pesquisador	Carta_de_Envio_de_Pendencias_assinado.pdf	21/08/2023 12:51:06	BRUNO RAMALHO DE CARVALHO	Aceito
Projeto Detalhado / Brochura Investigador	Projeto_integra.docx	21/08/2023 12:34:41	BRUNO RAMALHO DE CARVALHO	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLETimeLapse.docx	21/08/2023 12:32:05	BRUNO RAMALHO DE CARVALHO	Aceito
TCLE / Termos de	Anuencia_BRUNO.pdf	21/08/2023	BRUNO RAMALHO	Aceito

**Endereço:** SEPN 707/907 - Bloco 6, sala 6.205, 2º andar

**Bairro:** Setor Universitário

**CEP:** 70.790-075

**UF:** DF

**Município:** BRASILIA

**Telefone:** (61)3966-1511

**E-mail:** cep.uniceub@uniceub.br



Continuação do Parecer: 6.313.392

Assentimento / Justificativa de Ausência	Anuencia_BRUNO.pdf	11:57:20	DE CARVALHO	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	Anuencia_GENESIS.pdf	21/08/2023 11:57:08	BRUNO RAMALHO DE CARVALHO	Aceito
Declaração de Pesquisadores	DECLARACAO_DE_CONFIDENCIALID ADE_PARA_PESQUISA_CIENTIFICA_ modeloassinado.pdf	05/07/2023 08:50:07	BRUNO RAMALHO DE CARVALHO	Aceito
Folha de Rosto	folhaDeRostoAssinada.pdf	22/06/2023 17:01:17	BRUNO RAMALHO DE CARVALHO	Aceito

**Situação do Parecer:**

Aprovado

**Necessita Apreciação da CONEP:**

Não

BRASILIA, 20 de Setembro de 2023

---

**Assinado por:**  
**Marilia de Queiroz Dias Jacome**  
**(Coordenador(a))**

**Endereço:** SEPN 707/907 - Bloco 6, sala 6.205, 2º andar

**Bairro:** Setor Universitário

**CEP:** 70.790-075

**UF:** DF

**Município:** BRASILIA

**Telefone:** (61)3966-1511

**E-mail:** cep.uniceub@uniceub.br

## ANEXO II – Termo de Consentimento para Utilização de Dados de Entrevistas, Gravação de Reuniões e Uso de Gravação

## **Termo de Consentimento para Utilização de Dados de Entrevistas, Gravação de Reuniões e Uso de Gravação**

Pelo presente instrumento, de um lado, o Dr. **Bruno Ramalho de Carvalho**, inscrito no CPF sob o número **539.662.101-00**, inscrito no CRM DF sob o número **16335** – Ginecologia e Obstetrícia (RQE 13573) e Reprodução Assistida (RQE 15521), representante da **Clínica Bruno Ramalho Reprodução Humana**, em parceria com a **GENESIS Centro de Assistência em Reprodução Humana**, situada em Brasília, Distrito Federal, doravante denominado **AUTORIZANTE**; e, de outro lado, as alunas **Maria Eduarda Dos Santos Abritta Ferreira**, inscrita no CPF sob o número **081.776.091-14**, e **Sabrina Caldas Berno**, inscrita no CPF sob o número **033.782.041-41**, ambas estudantes da Universidade de Brasília (UnB), doravante denominadas **PESQUISADORAS**, resolvem firmar o presente contrato mediante as cláusulas e condições abaixo estabelecidas:

### **1. Objeto do Contrato**

O presente contrato tem como objeto a autorização para a utilização das falas do **AUTORIZANTE**, extraídas das entrevistas realizadas, como referencial teórico para a pesquisa das **PESQUISADORAS**, no contexto do Trabalho de Conclusão de Curso (TCC) intitulado "Desenvolver uma abordagem baseada em inteligência artificial para identificar padrões em dados morfocinéticos de embriões obtidos por meio do Time-Lapse System, capaz de prever a porcentagem de euploidia, proporcionando uma solução mais eficaz e menos invasiva em comparação ao PGT-A", elaborado pelas **PESQUISADORAS**.

Além disso, o **AUTORIZANTE** autoriza a gravação de reuniões, seja em áudio, vídeo ou qualquer outra forma de registro, para fins de estudo, análise e documentação, conforme os objetivos da pesquisa.

### **2. Autorização para Gravação de Entrevistas e Reuniões**

O **AUTORIZANTE** autoriza a gravação completa da entrevista, bem como de reuniões subsequentes, seja em áudio, vídeo ou qualquer outra forma de registro, para fins de estudo, análise e documentação. As gravações poderão ser utilizadas integralmente, incluindo trechos ou citações, nas etapas da pesquisa conduzida pelas **PESQUISADORAS**.

### **3. Uso das Respostas e Trechos das Gravações**

O **AUTORIZANTE** concede permissão para que suas respostas fornecidas durante a entrevista, bem como as falas extraídas das reuniões, sejam utilizadas como referencial teórico do Trabalho de Conclusão de Curso (TCC). Tais falas poderão ser incorporadas, integralmente ou em trechos relacionados ao estudo realizado pelas **PESQUISADORAS**.

### **4. Confidencialidade e Identificação**

O **AUTORIZANTE** reconhece que, na utilização das gravações, sua identidade e informações pessoais poderão ser reveladas, caso seja necessário para a pesquisa. O **AUTORIZANTE** autoriza a associação de sua identidade às suas falas ou respostas, conforme o objetivo acadêmico ou científico da pesquisa. Caso o **AUTORIZANTE** deseje que sua identidade não seja revelada, poderá solicitar a anonimização dos dados.

## 5. Revogação do Consentimento

O **AUTORIZANTE** tem pleno conhecimento de que pode revogar este consentimento a qualquer momento, mediante solicitação por escrito. A revogação não afetará o uso das falas ou dados já coletados até o momento da solicitação.

## 6. Direitos sobre o Material

O **AUTORIZANTE** reconhece que o material resultante das gravações das entrevistas e reuniões, incluindo suas respostas e falas, poderá ser armazenado, analisado e divulgado exclusivamente para fins acadêmicos ou científicos. Concorde que não receberá compensação financeira ou qualquer outra forma de benefício pelo uso do material gerado.

## 7. ASSINATURAS

Por estarem de pleno acordo com as condições deste contrato, as partes o assinam.  
**24 de dezembro de 2024**  
**Brasília, \_\_\_\_\_**

**AUTORIZANTE:** Dr. Bruno Ramalho de Carvalho

CRM DF 16335

Assinatura: \_\_\_\_\_



**PESQUISADORAS:** Maria Eduarda Dos Santos Abritta Ferreira

CPF: 081.776.091-74

Assinatura: \_\_\_\_\_



Sabrina Caldas Berno

CPF: 033.782.041-41

Assinatura: \_\_\_\_\_



### TESTEMUNHA:

1. Nome: **George Marsicano Correa**

2. CPF: **818.841.921-49**

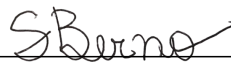
3. Assinatura: \_\_\_\_\_



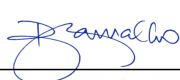
## Página de assinaturas



**Maria Ferreira**  
081.776.091-14  
Signatário



**Sabrina Berno**  
033.782.041-41  
Signatário











**Bruno Carvalho**  
539.662.101-00  
Signatário



**George Correa**  
818.841.921-49  
Signatário

## HISTÓRICO

24 dez 2024 09:07:42		<b>Maria Eduarda Dos Santos Abritta Ferreira</b> criou este documento. ( Email: eduardaabritta@gmail.com, CPF: 081.776.091-14 )
24 dez 2024 09:07:42		<b>Maria Eduarda Dos Santos Abritta Ferreira</b> (Email: eduardaabritta@gmail.com, CPF: 081.776.091-14) visualizou este documento por meio do IP 177.158.135.161 localizado em Brasília - Federal District - Brazil
24 dez 2024 09:08:23		<b>Maria Eduarda Dos Santos Abritta Ferreira</b> (Email: eduardaabritta@gmail.com, CPF: 081.776.091-14) assinou este documento por meio do IP 177.158.135.161 localizado em Brasília - Federal District - Brazil
24 dez 2024 11:10:10		<b>Sabrina Caldas Berno</b> (Email: sabrinacberno@gmail.com, CPF: 033.782.041-41) visualizou este documento por meio do IP 177.174.208.14 localizado em Brasília - Federal District - Brazil
24 dez 2024 11:15:29		<b>Sabrina Caldas Berno</b> (Email: sabrinacberno@gmail.com, CPF: 033.782.041-41) assinou este documento por meio do IP 189.61.13.144 localizado em Brasília - Federal District - Brazil
24 dez 2024 11:21:30		<b>Bruno Ramalho de Carvalho</b> (Email: bruno.ramalho@ceub.edu.br, CPF: 539.662.101-00) visualizou este documento por meio do IP 189.6.30.39 localizado em Brasília - Federal District - Brazil
24 dez 2024 11:22:24		<b>Bruno Ramalho de Carvalho</b> (Email: bruno.ramalho@ceub.edu.br, CPF: 539.662.101-00) assinou este documento por meio do IP 189.6.30.39 localizado em Brasília - Federal District - Brazil
06 jan 2025 14:17:53		<b>George Marsicano Correa</b> (Email: georgemarsicano@unb.br, CPF: 818.841.921-49) visualizou este documento por meio do IP 179.214.113.5 localizado em Brasília - Federal District - Brazil





06 jan 2025

14:18:46



**George Marsicano Correa** (Email: [georgemarsicano@unb.br](mailto:georgemarsicano@unb.br), CPF: 818.841.921-49) assinou este documento por meio do IP 179.214.113.5 localizado em Brasília - Federal District - Brazil



## ANEXO III – Contrato de Autorização para Utilização de Dados em Pesquisa

# CONTRATO DE AUTORIZAÇÃO PARA UTILIZAÇÃO DE DADOS EM PESQUISA

## IDENTIFICAÇÃO DAS PARTES

Pelo presente instrumento, de um lado, o Dr. **Bruno Ramalho de Carvalho**, inscrita no CPF sob o número **539.662.101-00** inscrito no CRM DF sob o número **16335** – Ginecologia e Obstetrícia (RQE 13573) e Reprodução Assistida (RQE 15521), representante da **Clínica Bruno Ramalho Reprodução Humana**, em parceria com a **GENESIS Centro de Assistência em Reprodução Humana**, situada em Brasília, Distrito Federal, doravante denominado **AUTORIZANTE**; e, de outro lado, as alunas **Maria Eduarda Dos Santos Abritta Ferreira**, inscrita no CPF sob o número **081.776.091-14**, e **Sabrina Caldas Berno**, inscrita no CPF sob o número **033.782.041-41**, ambas estudantes da Universidade de Brasília (UnB), doravante denominadas **PESQUISADORAS**, resolvem firmar o presente contrato mediante as cláusulas e condições abaixo estabelecidas.

## 1. CLÁUSULA PRIMEIRA – OBJETO DO CONTRATO

O presente contrato tem como objeto a autorização para a utilização de dados anônimos de pacientes da Clínica Bruno Ramalho Reprodução Humana, com o objetivo exclusivo de realizar análise estatística e desenvolver modelos no contexto do Trabalho de Conclusão de Curso (TCC) intitulado “**Desenvolver uma abordagem baseada em inteligência artificial para identificar padrões em dados morfocinéticos de embriões obtidos por meio do Time-Lapse System, capaz de prever a porcentagem de euploidia, proporcionando uma solução mais eficaz e menos invasiva em comparação ao PGT-A**”, elaborado pelas PESQUISADORAS.

Os dados que serão utilizados incluem: **Idade, Data da biópsia, Embrião n., Estágio, Morfo, Kidscore, st2, t2, t3, t4, t5, t8, tSC, tSB, tB, t2-st2, cc2 (t3-t2), cc3 (t5-t3), t5-t2, s2 (t4-t3), s3 (t8-t5), tSC-t8, tB-tSB, Ploidia.**

## 2. CLÁUSULA SEGUNDA – TERMO DE CONSENTIMENTO

O AUTORIZANTE declara que possui autorização vigente junto à Plataforma Brasil para a utilização dos dados das pacientes incluídos no presente contrato, garantindo que tais dados foram obtidos com o devido consentimento informado e em conformidade com as normas éticas aplicáveis.

## 3. CLÁUSULA TERCEIRA – RESPONSABILIDADE E CONFIDENCIALIDADE

1. Os dados serão fornecidos em formato anonimizado, identificados apenas por códigos (ID), impossibilitando a associação com a identidade das pacientes.
2. As PESQUISADORAS comprometem-se a manter o sigilo e a confidencialidade dos dados fornecidos, utilizando-os exclusivamente para os fins especificados neste contrato.
3. Fica vedada qualquer tentativa de reidentificação dos dados ou de uso para outros fins alheios ao projeto descrito na Cláusula Primeira.

#### 4. CLÁUSULA QUARTA – PRAZO DE UTILIZAÇÃO

A utilização dos dados estará autorizada até a finalização do TCC 2 das PESQUISADORAS, prevista para o final do semestre 2025.1 da Universidade de Brasília.

#### 5. CLÁUSULA QUINTA – DIREITOS DE AUTORIA E PUBLICAÇÃO

O AUTORIZANTE concorda que os dados utilizados pelas PESQUISADORAS poderão ser mencionados em publicações acadêmicas ou apresentações públicas relacionadas ao TCC, desde que sejam preservados o anonimato das pacientes e os créditos à Clínica Bruno Ramalho Reprodução Humana e à GENESIS Centro de Assistência em Reprodução Humana.

#### 6. CLÁUSULA SEXTA – REVOGAÇÃO DA AUTORIZAÇÃO

O AUTORIZANTE poderá revogar a autorização a qualquer momento, desde que notifique as PESQUISADORAS por escrito. Contudo, a revogação não obrigará as PESQUISADORAS a excluir os estudos ou análises já realizados, mas impedirá o uso dos dados para qualquer outro estudo futuro.

#### 7. CLÁUSULA SÉTIMA – DISPOSIÇÕES GERAIS

1. Este contrato não prevê cláusulas de indenização.
2. Qualquer situação omissa será resolvida em comum acordo entre as partes, respeitando-se as legislações vigentes.

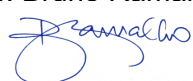
#### 8. ASSINATURAS

Por estarem de pleno acordo com as condições deste contrato, as partes o assinam.  
Brasília, 24 de dezembro de 2024

**AUTORIZANTE:** Dr. Bruno Ramalho de Carvalho

CRM DF 16335

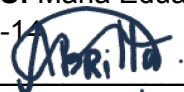
Assinatura: \_\_\_\_\_



**PESQUISADORAS:** Maria Eduarda Dos Santos Abritta Ferreira

CPF: 081.776.091-14

Assinatura: \_\_\_\_\_



Sabrina Caldas Berno

CPF: 033.782.041-41

Assinatura: \_\_\_\_\_



#### TESTEMUNHA:

1. Nome: George Marsicano Correa

2. CPF: 818.841.921-49

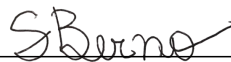
3. Assinatura: \_\_\_\_\_



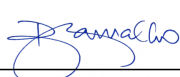
## Página de assinaturas



**Maria Ferreira**  
081.776.091-14  
Signatário



**Sabrina Berno**  
033.782.041-41  
Signatário











**Bruno Carvalho**  
539.662.101-00  
Signatário



**George Correa**  
818.841.921-49  
Signatário

## HISTÓRICO

24 dez 2024 09:06:20		<b>Maria Eduarda Dos Santos Abritta Ferreira</b> criou este documento. ( Email: eduardaabritta@gmail.com, CPF: 081.776.091-14 )
24 dez 2024 09:06:20		<b>Maria Eduarda Dos Santos Abritta Ferreira</b> (Email: eduardaabritta@gmail.com, CPF: 081.776.091-14) visualizou este documento por meio do IP 177.158.135.161 localizado em Brasília - Federal District - Brazil
24 dez 2024 09:06:25		<b>Maria Eduarda Dos Santos Abritta Ferreira</b> (Email: eduardaabritta@gmail.com, CPF: 081.776.091-14) assinou este documento por meio do IP 177.158.135.161 localizado em Brasília - Federal District - Brazil
24 dez 2024 10:22:27		<b>Sabrina Caldas Berno</b> (Email: sabrinacberno@gmail.com, CPF: 033.782.041-41) visualizou este documento por meio do IP 189.61.13.144 localizado em Brasília - Federal District - Brazil
24 dez 2024 10:22:43		<b>Sabrina Caldas Berno</b> (Email: sabrinacberno@gmail.com, CPF: 033.782.041-41) assinou este documento por meio do IP 189.61.13.144 localizado em Brasília - Federal District - Brazil
24 dez 2024 10:37:29		<b>Bruno Ramalho de Carvalho</b> (Email: bruno.ramalho@ceub.edu.br, CPF: 539.662.101-00) visualizou este documento por meio do IP 189.6.30.39 localizado em Brasília - Federal District - Brazil
24 dez 2024 10:43:03		<b>Bruno Ramalho de Carvalho</b> (Email: bruno.ramalho@ceub.edu.br, CPF: 539.662.101-00) assinou este documento por meio do IP 189.6.30.39 localizado em Brasília - Federal District - Brazil
06 jan 2025 14:13:30		<b>George Marsicano Correa</b> (Email: georgemarsicano@unb.br, CPF: 818.841.921-49) visualizou este documento por meio do IP 179.214.113.5 localizado em Brasília - Federal District - Brazil



06 jan 2025

14:17:18



**George Marsicano Correa** (Email: [georgemarsicano@unb.br](mailto:georgemarsicano@unb.br), CPF: 818.841.921-49) assinou este documento por meio do IP 179.214.113.5 localizado em Brasília - Federal District - Brazil

