Predição de Ciclos Menstruais com o Apoio de Técnicas de *Machine Learning*: Uma aplicação com dados sintéticos

Paulina Julia Costa de Oliveira [1], Rosana C. B. Rego[2]

[1] Graduanda em Tecnologia da Informação, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
(paulina.oliveira@alunos.ufersa.edu.br)
[2] Departamento de Engenharia e Tecnologia, Universidade Federal Rural do Semi-Árido, Pau dos Ferros, Brasil
(rosana.rego@ufersa.edu.br)

Resumo: Neste trabalho, exploramos o uso de técnicas de aprendizado de máquina para prever ciclos menstruais. A previsão precisa do ciclo menstrual é importante para a saúde feminina, pois permite aos indivíduos tomar medidas preventivas para minimizar os desconfortos associados ao ciclo. Além disso, a previsão precisa pode ser útil para planejar eventos importantes na vida da mulher. Para atingir nossos objetivos, dados sintéticos foram gerados. Os resultados mostraram que é possível prever com precisão o início e a duração dos ciclos reprodutivos com o uso de técnicas de aprendizado de máquina.

Palavras-chave: Machine learning; Ciência de dados; Predição; Saúde

INTRODUÇÃO

O ciclo menstrual é um processo natural que ocorre no corpo feminino, com duração média de 28 dias, embora possa variar entre 21 e 35 dias em mulheres adultas. Ele é controlado por hormônios produzidos pelos ovários, hipotálamo e hipófise e consiste em uma série de eventos que preparam o corpo para a gravidez (Trickey, 2004).

O ciclo começa no primeiro dia da menstruação, ou seja, primeiro dia de sangramento, e dura até o início da próxima menstruação. Durante esse período, o corpo feminino passa por diversas fases, tais como a fase folicular, a ovulação e a fase lútea. A fase folicular é caracterizada pelo desenvolvimento dos folículos ovarianos, que contêm os óvulos imaturos. A ovulação ocorre quando um desses óvulos maduros é liberado do ovário e segue em direção às trompas de falópio. A fase lútea é o período que se segue à ovulação, em que o corpo se prepara para a possível gravidez (Owen, 1975; Mihm, 2011). Se a fecundação não ocorrer, o corpo se prepara para a menstruação, eliminando o revestimento do útero. Esse processo pode causar sintomas como cólicas, inchaço e alterações de humor em algumas mulheres. Já em caso de fecundação, o óvulo fertilizado se implanta no revestimento do útero, dando início à gestação (Trickey, 2004; Barbieri, 2014).

Saber a data exata do ciclo menstrual é importante por diversas razões. Em primeiro lugar, permite que a mulher possa prever quando ocorrerá a próxima menstruação, o que é útil tanto para se preparar para o período quanto para identificar possíveis irregularidades no ciclo. Além disso, conhecer a data exata do ciclo é fundamental para o planejamento familiar, seja para evitar ou para buscar uma gravidez (Carmichael, 2021).

Por outro lado, em termos de saúde, conhecer a data exata do ciclo menstrual pode auxiliar na identificação de possíveis desequilíbrios hormonais e outros problemas ginecológicos, como a síndrome dos ovários policísticos. Também pode ser útil para monitorar a eficácia de tratamentos para problemas como a endometriose e a dismenorreia (Edelman, 2022; Findlay, 2020).

Por fim, em alguns casos, a data exata do ciclo menstrual pode ser importante para a prevenção de doenças, como o câncer de colo de útero, já que o período menstrual pode afetar o resultado dos exames de rastreamento. Entretanto, quando o ciclo é irregular, pode ser mais difícil determinar a data exata do ciclo menstrual. Isso pode ocorrer por diversos motivos, como variações hormonais, estresse, doenças ou outros fatores (Trickey, 2004).

Os algoritmos de machine learning podem ajudar a prever ciclos irregulares através da análise de dados de saúde e de sintomas da paciente. Para isso, é preciso construir um modelo de machine learning que leve em consideração as variáveis relevantes para a predição do ciclo menstrual, como histórico de menstruação, duração do ciclo, data de início e fim da menstruação e sintomas relacionados ao ciclo, entre outros (Arunkumar, 2023; Ogidan, 2023).

Com base nos dados coletados, o modelo pode então ser treinado para identificar padrões e correlações entre as variáveis, e, assim, fazer previsões mais precisas sobre o ciclo menstrual, mesmo em casos de irregularidade. O modelo pode ser atualizado e melhorado continuamente com novos dados, tornando as previsões cada vez mais precisas. Essas previsões podem ser valiosas para pacientes que têm dificuldade em prever a data exata do ciclo menstrual devido à irregularidade, permitindo um planejamento mais efetivo de atividades e auxiliando no monitoramento da saúde feminina (Thakur, 2023; Arunkumar, 2023).

Neste trabalho buscamos explorar a aplicação de dois modelos de aprendizado de máquina, sendo estes o modelo Random Forest, que é um algoritmo baseado em árvores de decisão (Fawa Greh, 2014) e o modelo de Regressão Linear que é baseado na relação linear entre as variáveis independentes e a variável dependente (Aiken et al., 2003). Exploramos o uso destes com o objetivo de analisar o seu desempenho e a forma como estes se aplicam na predição de ciclos regulares. Desta forma buscamos identificar o melhor modelo que se adapte ao que buscamos na previsão de ciclos menstruais.

Dessa forma, o trabalho está dividido como segue: na seção conjunto de dados, características dos dados gerados são descritas. Na seção modelos de machine learning, os algoritmos utilizados são apresentados, bem como os hiperparâmetros selecionados. Na seção resultados e discussão, as previsões obtidas com as aplicações dos modelos são apresentadas. Por fim, na seção conclusões, as considerações finais do trabalho são apresentadas.

CONJUNTO DE DADOS

Dados sintéticos foram gerados para permitir a realização de testes e comparações entre os modelos de predição implementados. Os dados sintéticos foram gerados com base na equação:

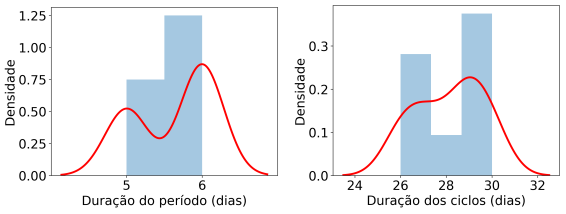
$$ciclo = ciclo_{média} + ciclo_{std} * \epsilon_{ciclo} \quad (1)$$

em que, $ciclo_{média}$ é a média do ciclo, $ciclo_{std}$ é o desvio padrão do ciclo e ϵ_{ciclo} é um parâmetro de incerteza, isto é, um número aleatório. O mesmo procedimento foi utilizado para gerar os valores dos períodos. Portanto,

$$p = p_{média} + p_{std} * \epsilon_{p} \quad (2)$$

em que, $p_{média}$ é a média em dias do período, p_{std} é o desvio padrão do período e ϵ_{p} é a incerteza em dias.

As Figura 1 (a) e (b), mostram a densidade da distribuição dos dados gerados, tanto para o ciclo total como para os períodos em dias. Foram gerados 24 ciclos, isto é, dados referente a dois anos, com duração de 26 a 30 dias. Os períodos desses ciclos possuem uma duração de 5 a 6 dias, como mostrado nas Figuras 1 (a) e (b).



(a) (b)

Figura 1. Dados gerados: (a) duração dos períodos e (b) duração dos ciclos.

As Figuras 2 (a) e (b), mostram os *box plots* dos dados gerados. Pode ser observado, que o box plot para a duração dos ciclos, possui um comportamento assimétrico, ou seja, os ciclos não estão distribuídos de maneira equilibrada ao redor da mediana. Dessa forma, a distribuição dos dados é assimétrica. Já para o período, pode ser observado que a média está em torno de 6 dias.

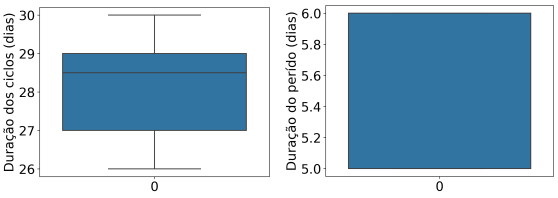


Figura 2. Dados gerados: (a) *box plot* da duração dos ciclos e (b) *box plot* da duração dos períodos.

Com os dados gerados e analisados, o conjunto foi dividido em 80% para treino e 20% para teste dos modelos.

MODELOS DE *MACHINE LEARNING*

Os modelos de machine learning são ferramentas poderosas para a análise de dados e previsão de resultados. Nesta seção, os dois modelos utilizados, Random Forest e Regressão Linear, serão apresentados brevemente.

A. RANDOM FOREST

O Random Forest é um algoritmo de aprendizado de máquina baseado em árvores de decisão (Fawagreh, 2014). Ele funciona construindo várias árvores de decisão a partir de amostras aleatórias dos dados de treinamento e combinando suas previsões para melhorar a precisão (Figura 3). Este modelo é conhecido por ser robusto e eficaz em lidar com problemas complexos, tais como a classificação e a regressão em dados com muitas variáveis (Cutler, 2012).

Esse algoritmo pode ser usado para fazer previsões sobre dados de séries temporais, como é o caso dos ciclos menstruais. Uma série temporal é uma série de pontos de dados coletados em intervalos regulares ao longo do tempo (Masini, 2021). O objetivo de um algoritmo de previsão de séries temporais é usar dados passados para fazer previsões sobre pontos de dados futuros.

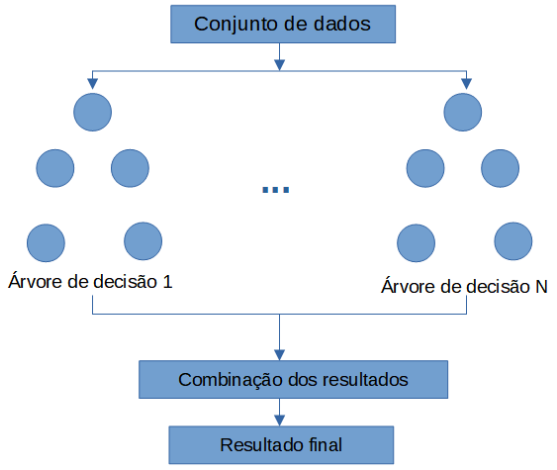


Figura 3. Representação do modelo *Random Forest*.

Para implementação do modelo, o algoritmo regressor Random Forest foi implementado na biblioteca *scikit learn*. Para medir a qualidade da divisão do conjunto de dados, a função do erro médio quadrático foi utilizada. O número de árvores foi de 50 árvores de decisão para geração da floresta.

B. REGRESSÃO LINEAR

A Regressão Linear é um modelo de previsão baseado na relação linear entre as variáveis independentes e a variável dependente (Aiken et al., 2003). A ideia principal da Regressão Linear é encontrar a melhor linha de tendência que explique a relação entre as variáveis e permita prever valores futuros (James et al., 2021). Este modelo é mais simples e fácil de interpretar do que o Random Forest, mas pode ser menos preciso em problemas mais complexos.

O algoritmo de regressão pode ser aplicado a dados de séries temporais para fazer previsões sobre valores futuros (Hope, 2020). Nesse caso, a variável dependente é a série temporal do ciclo e as variáveis independentes são os valores passados do ciclo e período em dias. O objetivo do algoritmo é encontrar os coeficientes da equação linear que melhor se ajustam aos dados, de modo que a equação possa ser usada para fazer previsões sobre os valores futuros do ciclo.
De modo geral, a equação da regressão linear é dada por:

$$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + \ldots + \beta n * xn + \varepsilon \quad (3)$$

em que, y é a variável dependente, x1,...,xn são as variáveis independentes, $\beta 0, \beta 1, \beta 2, \ldots, \beta n$ são os coeficientes associados a cada variável independente e ε é termo erro que captura a variabilidade não explicada do modelo. Dessa forma, o objetivo da regressão linear é ajustar os coeficientes $\beta 0, \beta 1, \beta 2, \ldots, \beta n$ de maneira que a equação ajustada seja a melhor representação possível da relação entre y e x. Isso é feito minimizando a soma dos quadrados dos erros entre os valores previstos e os valores observados (Weisberg, 2005).

RESULTADOS E DISCUSSÃO

Nesta seção, os resultados obtidos a partir da aplicação dos algoritmos de machine learning, tais como o Random Forest e a Regressão Linear, para previsão da série temporal do ciclo são apresentados. Sabendo que a acurácia não é uma métrica apropriada para avaliar o desempenho de um modelo de regressão, a performance dos modelos foi comparada com base nas métricas de desempenho RMSE (*Root Mean Squared Error*) e o MAE (*Mean Absolute Error*). O RMSE é uma medida de erro que representa a média dos erros ao quadrado entre os valores previstos e os valores observados, enquanto o MAE é uma medida de erro que representa a média dos erros absolutos entre os valores previstos e os valores observados.

As Figuras 4 e 6 mostram os resultados de previsão com os modelos implementados. Analisando em particular o modelo de regressão linear, na Figura 4, observa-se que os valores preditos pelo o modelo são bem próximos dos valores dos dados reais. Além disso, na Figura 4, é possível observar a predição para o tempo futuro (*t+1*), isto é, o ciclo 5, em que o modelo previu que no ciclo 5 terá a duração de 30 dias e o período será de 6 dias.

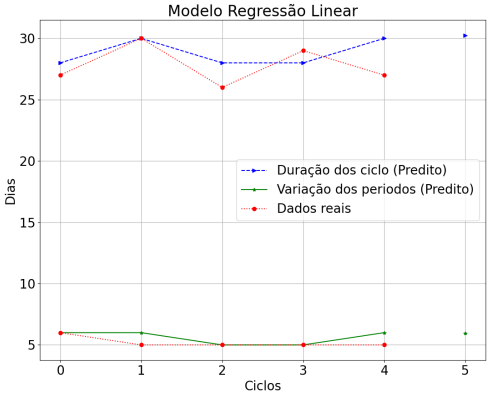


Figura 4. Predição realizada com o modelo de regressão linear.

Na Figura 5, tem-se o erro de predição para o modelo de regressão linear. O erro de predição é a diferença entre o valor predito pelo modelo e o valor desejado ou observado na prática.

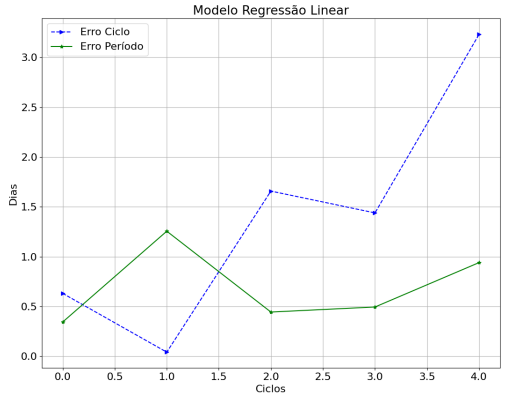


Figura 5. Erro de predição realizado com o modelo de regressão linear.

De acordo com a Figura 5, observa-se que no ciclo zero o erro foi baixo, podendo ser arredondado para zero, se considerarmos o dia como um valor inteiro. Já nos demais ciclos, o erro foi de aproximadamente 1 dia, exceto no ciclo 4, em que o erro foi de 3 dias. Já quando o período é analisado, o erro flutua próximo de 1 dia.

Na Figura 6, tem-se os resultados obtidos com a aplicação do modelo do *Random Forest*. Observar-se que para alguns ciclos, o modelo acerta a quantidade de dias. Analisando a previsão do modelo realizado no tempo futuro (*t+1*), ou seja, no ciclo 5, o modelo fez a previsão de 29 dias para o ciclo 5 e 6 dias para a duração do período. Dessa forma, é possível observar que a variação em dias de um modelo para o outro foi de apenas mais/menos 1 dia.

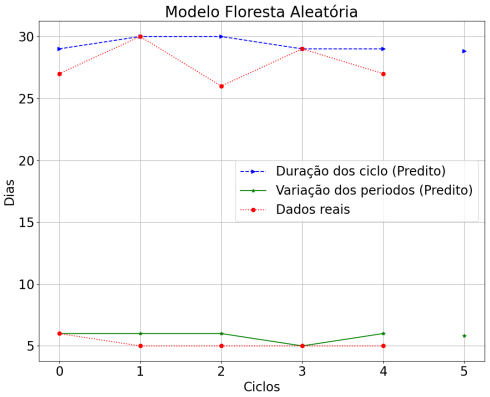


Figura 6. Predição realizada com o modelo Random Forest.

Na Figura 7, tem-se o erro de predição para o modelo *Random Forest*. É possível observar que o erro do ciclo com o modelo *Random* flutua mais entre 2 e 3 dias de diferença, diferentemente do erro apresentado com o modelo de regressão linear. No entanto, este modelo em questão, apresenta uma menor flutuação na predição do dia exato do período.

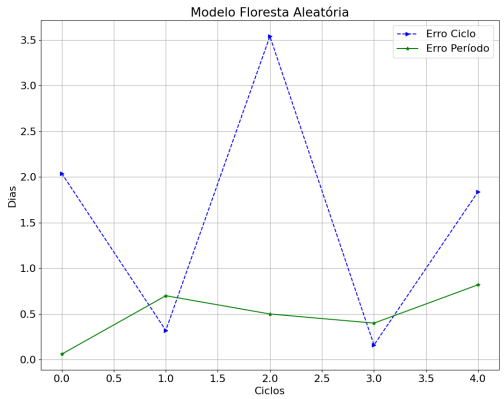


Figura 7. Erro de predição realizada com o modelo *Random Forest*.

Conforme a Tabela 1, o modelo de Regressão Linear apresentou um RMSE de 1,3667 e um MAE de 1,0476, enquanto o modelo de Random Forest apresentou um RMSE de 1,4759 e um MAE de 1,0379. Isso sugere que o modelo de Regressão

Linear se saiu melhor que o modelo de Random Forest na previsão dos dados quanto ao RMSE e pior quanto a MAE. No entanto, é importante ressaltar que o RMSE por si só não garante que o modelo de Regressão Linear seja o melhor para todos os casos, uma vez que a escolha do modelo depende de muitos fatores, como a complexidade do problema, a quantidade de dados disponíveis, e a necessidade de interpretabilidade. Em geral, o MAE é uma métrica mais robusta a outliers (valores extremos). Já o RMSE é mais sensível a outliers.

Tabela 1. Métricas de avaliação dos modelos.

Modelo	RMSE	MAE
Regressão Linear	1,3667	1,0476
Random Forest	1,4759	1,0379

O Random Forest é conhecido por ser robusto e eficaz em lidar com séries temporais complexas, mas pode ser menos fácil de interpretar devido à sua natureza baseada em árvores de decisão. A Regressão Linear, por outro lado, é uma abordagem mais simples e fácil de interpretar, mas pode ser menos precisa em séries temporais com muitas variáveis ou sazonalidades. No caso da predição de ciclos, o modelo de Regressão Linear foi capaz de capturar a relação linear entre as variáveis independentes e a variável dependente. Dessa forma, o modelo pode ser usado para prever ciclos regulares com boa precisão.

CONCLUSÃO

A regressão linear é um algoritmo simples e rápido que pode ser aplicado a dados de séries temporais para fazer previsões. No contexto apresentado no trabalho, o resultado obtido com o modelo foi satisfatório para previsão no tempo $t+1$. No entanto, nem sempre o método pode fornecer os melhores resultados, especialmente se a relação entre as variáveis dependentes e independentes for mais complexa do que uma linha reta. Nesses casos, algoritmos mais sofisticados, como ARIMA ou SARIMA, podem ser mais apropriados.

A partir dos resultados apresentados, pode-se concluir que o Random Forest e a Regressão Linear são modelos diferentes, cada um com suas próprias vantagens e desvantagens, e a escolha do melhor modelo para um determinado problema depende das características dos dados e do objetivo da análise. Para trabalhos futuros, pretende-se aplicar modelos mais sofisticados, como o ARIMA em conjunto de dados reais.

AGRADECIMENTOS

Ao grupo de pesquisa CILab (*Computational Intelligence Laboratory*) da Universidade Federal Rural do Semi-Árido (UFERSA).

REFERÊNCIAS

AIKEN, Leona S.; WEST, Stephen G.; PITTS, Steven C. Multiple linear regression. Handbook of psychology, p. 481-507, 2003.

ARUNKUMAR, P. et al. Application using Machine Learning to Promote Women's Personal Health. In: 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2023. p. 908-914.

BARBIERI, Robert L. The endocrinology of the menstrual cycle. Human fertility: methods and protocols, p. 145-169, 2014.

CARMICHAEL, Mikaeli Anne et al. The impact of menstrual cycle phase on athletes' performance: a narrative review. International journal of environmental research and public health, v. 18, n. 4, p. 1667, 2021.

CUTLER, Adele; CUTLER, D. Richard; STEVENS, John R. Random forests. Ensemble machine learning: Methods and applications, p. 157-175, 2012.

EDELMAN, Alison et al. Association between menstrual cycle length and coronavirus disease 2019 (COVID-19) vaccination: a US cohort. Obstetrics and gynecology, v. 139, n. 4, p. 481, 2022.

FAWAGREH, Khaled; GABER, Mohamed Medhat; ELYAN, Eyad. Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal, v. 2, n. 1, p. 602-609, 2014.

FINDLAY, Rebekka J. et al. How the menstrual cycle and menstruation affect sporting performance: experiences and perceptions of elite female rugby players. British journal of sports medicine, v. 54, n. 18, p. 1108-1113, 2020.

JAMES, Gareth et al. Linear regression. An introduction to statistical learning: with applications in R, p. 59-128, 2021.

HOPE, Thomas MH. Linear regression. In: Machine Learning. Academic Press, 2020. p. 67-81.

MASINI, Ricardo P.; MEDEIROS, Marcelo C.; MENDES, Eduardo F. Machine learning advances for time series forecasting. Journal of economic surveys, 2021.

MIHM, M.; GANGOOLY, S.; MUTTUKRISHNA, S. The normal menstrual cycle in women. Animal reproduction science, v. 124, n. 3-4, p. 229-236, 2011.

OGIDAN, Olugbenga Kayode; OGUNNIYI, Julius Olasunmibo; TEDIMOLA, Abisola. Development of an Automated Temperature Measuring Device: A Potential Tool for Ovulation Detection. ABUAD Journal of Engineering Research and Development, v. 6, n. 1, p. 13-21, 2023.

OWEN, John A. Physiology of the menstrual cycle. The American journal of clinical nutrition, v. 28, n. 4, p. 333-338, 1975.

THAKUR, Tanmay et al. Machine Learning in Period, Fertility and Ovulation Tracking Application. 2023.

TRICKEY, Ruth. Women, Hormones and the Menstrual Cycle: herbal and medical solutions from adolescence to menopause. Allen & Unwin, 2004.

WEISBERG, Sanford. Applied linear regression. John Wiley & Sons, 2005.