Machine learning to screen polycystic ovary syndrome in South Asian women in primary care: Model development and validation

MEDI 504B Project Report - Erick Navarro & Timo Tolppa

Introduction

Polycystic ovary syndrome (PCOS) is the most common endocrine disorder of menstruating people of reproductive age [1,2]. This complex multisystem disease is characterized by irregular menstrual periods, androgen excess and enlarged polycystic ovaries, and is associated with infertility, mental health disorders, obesity, diabetes, low vitamin D3 and a higher risk of cardiovascular disease [2-4]. The disease has a substantial negative effect on quality of life and health, and thus early diagnosis and treatment is crucial [1].

Diagnosis of PCOS is confirmed by the presence of two out of three criteria: irregular cycles, polycystic ovary morphology on transvaginal ultrasound, and hyperandrogenism based on either blood tests or clinical history suggestive of acne, hair loss and excess hair growth [2]. These diagnostic criteria were established largely based on studies on White women and limited existing evidence suggests that clinical features vary between different ethnic groups [1,3]. For example, South Asian women with PCOS have been shown to have lower body mass index, higher central obesity and more hair growth than White women [3]. Therefore, a greater understanding of clinical features of PCOS from a more diverse ethnic population is required.

Access to specialist care in many areas of the world, including the state of Kerala in India, is poor [5]. Thus, menstruating people in primary care are assessed based solely on their clinical history and ones suspected of PCOS are referred to specialist testing, including hormone tests and transvaginal ultrasound. Travel to specialist centers can incur substantial costs for patients, which may deter them from attending appointments [5]. To avoid unnecessary testing, referrals and cost, a diagnostic tool to help predict the need for further confirmatory testing for patients highly likely to have PCOS would be beneficial [1]. Furthermore, a model using solely information obtained through a telephone consultation (i.e. patient history) would mean the prediction could be performed by non-physicians remotely, reducing the burden on overworked physicians and the need for menstruating people to unnecessarily attend consultations in-person.

Machine learning models have been used to create diagnostic prediction tools in various disciplines [6,7]. These models are particularly helpful in diseases where uncertainty exists about the importance of symptoms and tests in the diagnosis, as is the case with PCOS. Prior PCOS machine learning studies have identified tests and gene biomarkers relevant to PCOS, however, this information is often not available to healthcare workers in primary care deciding whether to refer menstruating people for further testing [8,9].

Aim and objectives

This project aims to develop and validate a predictive diagnostic model using machine learning methods to help non-physicians predict the need for South Asian patients to undergo further testing based on their clinical history to diagnose PCOS. Prediction models developed using logistic regression, elastic net and random forest were compared based on their ability to identify all PCOS cases (i.e. sensitivity) that need to be referred for confirmatory testing. Prediction models based on patient history were compared to models developed using information obtained from clinical examination, blood tests and transvaginal ultrasound for overall accuracy (i.e. F_1 score and area under the receiver operating characteristic curve) to understand the benefit of additional information for diagnostic prediction. Finally, we explored variable importance to gain insights into the most relevant discriminating features of PCOS in a South Asian population.

Methods

Data source and participants

The data used to develop and validate the predictive diagnostic model was obtained from a publicly available dataset collected across 10 hospitals in the state of Kerala in India [10], with information regarding 44 parameters related to PCOS and infertility (supplementary table 1.1).

Outcome and predictors

The outcome of interest was a diagnosis of PCOS. We conducted initial variable selection based on existing literature and pathophysiology of PCOS. In summary, features in the original dataset that were related to infertility (pregnancy, human chorionic gonadotropin levels) and those without clear evidence of association with PCOS (marital status, blood group, thyroid stimulating hormone levels, respiratory and heart rate, hemoglobin) were not used in the predictive model. Variables kept after this feature selection were grouped into variables related to PCOS that can be collected through clinical history, clinical examination, blood tests and transvaginal ultrasound (**Table 1**). These include the PCOS diagnostic criteria; cycle irregularity, symptoms of hair loss, excess hair growth and pimples, and ovarian follicle number and size [1,2]. The predictors also include other presenting features of PCOS (multiple miscarriages, skin darkening), laboratory variables related to the pathophysiology of PCOS (luteinizing hormone, follicle stimulating hormone, prolactin, progesterone, anti-müllerian hormone), metabolic features (obesity, blood sugar), hypothesized lifestyle related factors (reduced physical activity, diet quality) and potential consequences of PCOS (low vitamin D3 levels, elevated blood pressure, thickened endometrium) [1-4, 11].

Clinical history	**Clinical examination**	**Blood tests**	**Ultrasound**
Age	Weight	Follicle stimulating hormone (FSH)	Number of follicles (Left)
Cycle regularity	Height	Luteinizing hormone (LH)	Average follicle size (Left)
Cycle length	Body mass index	FSH/LH ratio	Number of follicles (Right)
Miscarriage	Hip width	Anti-müllerian hormone	Average follicle size (Right)
Weight gain	Waist width	Prolactin	Endometrial thickness
Hair growth	Waist-hip ratio	Vitamin D3	*+ Clinical history, examination and blood tests*
Skin darkening	Systolic blood pressure	Progesterone	
Hair loss	Diastolic blood pressure	Random blood sugar	
Pimples	*+ Clinical history*	*+ Clinical history and examination*	
Regular exercise			
Fast food consumption			

Table 1. Variables used in the predictive model grouped based on the collection method

Data exploration and pre-processing

Prior to model development, an exploratory data analysis was conducted to understand variation, explore unusual and missing values, determine correlation and covariation between variables as well as illustrate relationships between predictors and outcome. Univariate statistical tests between women with and without PCOS were conducted for specific variables of interest with a significance level set at 0.05, continuous variables were tested using the independent t-test and categorical variables using the chi-square test, with Fisher's exact test as appropriate. Continuous variables with a skewed distribution were log-10 transformed for model development and imputation, and biologically implausible values were removed.

Data was subsequently divided into training and validation datasets using a 70:30 split at random that ensured balance in the number of cases with a positive outcome (i.e. a diagnosis of PCOS). Following the split, missing data in the training dataset was replaced through conditional multiple imputation using the *mice* (Multivariate Imputation by Chained Equations) package [12]. We used the default imputation

arguments, which are predictive mean matching for numeric data, logistic regression imputation for binary data or factors with 2 levels, polytomous regression imputation for un-ordered categorical data and proportional odds model for ordered categorical data with more than 2 levels. Missing values in the validation dataset were not imputed, but rather, listwise deletion was performed.

Model development

Models were built using four different sets of variables outlined in table 1. The main models of interest were the ones built using variables collected through clinical history, as per the aim of this project, which is to develop a diagnostic prediction tool to determine the need for patients to undergo further testing based on their clinical history. Models were built using three supervised machine learning algorithms using the *caret* package in R: logistic regression, elastic net regression and random forest.

Logistic regression is a commonly used method in classification that produces a probability model of the outcome. The strengths of logistic regression are simplicity, ease of implementation, and interpretability [13].

Elastic net is a regularization technique, which simultaneously performs variable selection and continuous shrinkage [14]. Thus, elastic net was used to conduct further feature selection with an embedded method. A grid search was used to find the best tuning parameters (i.e. alpha and lambda).

Random forest was selected for its efficiency in the classification of tabular data, ability to avoid overfitting and provide insights into the importance of individual predictors [15]. Random forest is an ensemble method where a large collection of de-correlated individual tree predictions are built with each producing a classification. The final prediction is the classification that is selected by most individual trees. Detailed hyperparameter tuning was not performed for random forest, as the classifier performs well out-of-the-box [16]. However, hyperparameters m_{try}, minimum node size and split rule were optimized to sensitivity using *caret* package's inbuilt default tuning grid.

Model evaluation

All models were trained using 5-fold cross-validation to avoid overfitting. Models were compared using three metrics: sensitivity, F_1-score, and area under the receiver operating characteristic curve (AUC-ROC). Sensitivity is the number of true positives expressed as a percentage of all positive cases. Sensitivity was chosen as the main metric for comparing different classifiers (i.e. logistic regression, elastic net and random forest), as we aimed our tool to effectively identify all likely PCOS cases that would benefit from confirmatory testing. The classifier with the best performance on sensitivity was chosen as the final model. The F_1-score is the weighted average of precision and sensitivity, and is more useful than accuracy in the case of unbalanced class distribution, as is the case with our data. The AUC-ROC represents the ability of a model to distinguish between classes. The F_1-score and AUC-ROC were used to compare models built using the four different sets of variables in their overall performance. This information was used to understand the benefit of gathering increasingly more invasive information for diagnostic prediction.

Model interpretation and additional analyses

After identifying the method and model with the best performance, variables in the final model were ranked by their coefficients to understand their relative importance. This was compared between the chosen model and rest of the models trained with the same method but different groups of variables. The effect of class imbalance in the chosen model was explored using down-sampling. The effect of multiple imputation was explored by comparing the performance of the model with its counterpart generated using list-wise deletion.

Promoting interoperability and replicability

Models were trained using RStudio (version 2022.12.0+353) and the caret package, and this report has been reported according to the TRIPOD statement [17]. All code used to develop and validate the predictive

models have been released in a public GitHub repository alongside our comprehensive exploratory data analysis, data preprocessing and model development reports [18].

Results

After data pre-processing, the study cohort consisted of 541 women aged between 20 and 48 years, of which 177 (32.7%) had a diagnosis of PCOS and 364 (67.3%) did not have a diagnosis of PCOS.

Exploratory data analysis

The cleaned dataset presented 14 missing values, which belonged to different individuals. Additionally, no more than 2 data points were missing for any parameter (supplementary table 2.1), supporting the conclusion that the data was missing at random and that multiple imputation could be conducted.

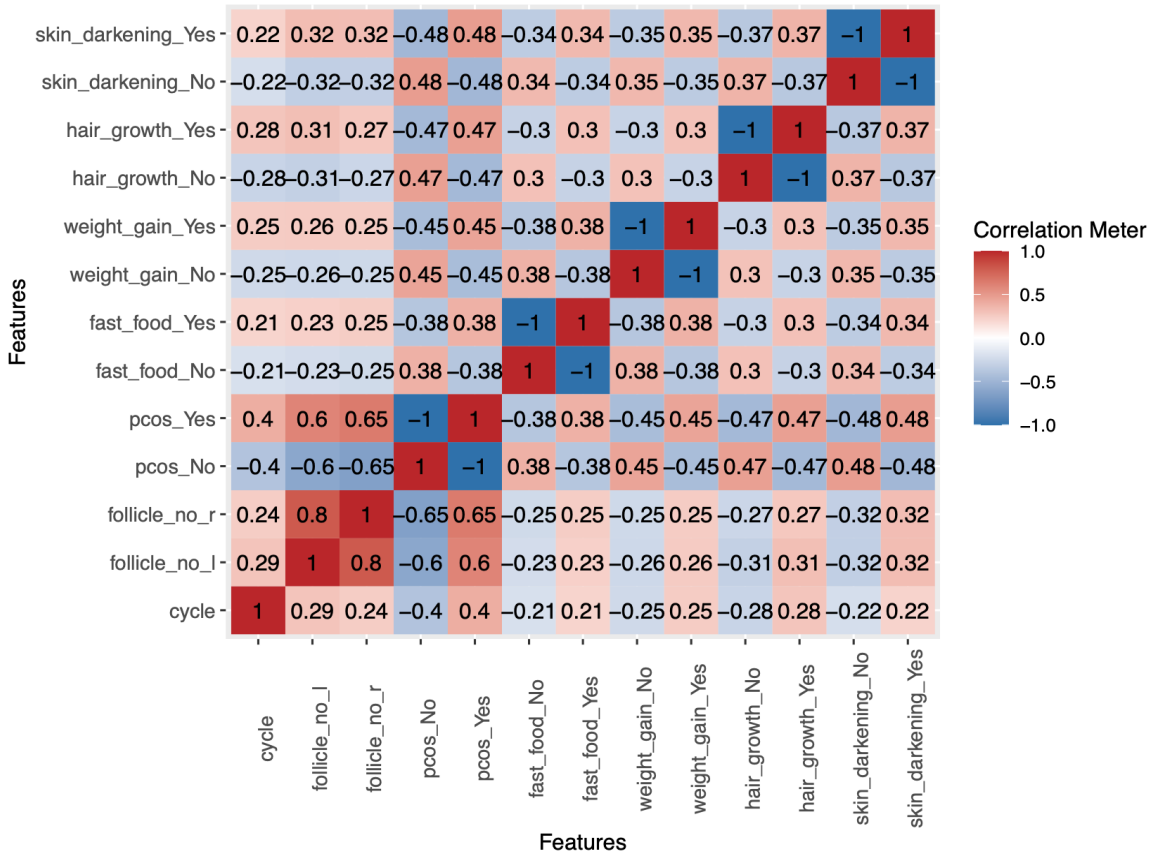


Figure 1. Correlation matrix for variables with a correlation coefficient > 0.35 with PCOS. Features with 'Yes' and 'No' denote presence or absence, respectively, of symptoms or a PCOS diagnosis. Full details in supplementary table 1.1.
Abbreviations: cycle = cycle length, follicle_no_l = number of follicles (left), follicle_no_r = number of follicles (right)

Correlation analysis revealed several variables that were correlated with PCOS (**Figure 1**), namely symptoms of skin darkening, hair growth and weight gain, length of the menstruation cycle, consumption of fast food and number of follicles in each ovary. All these variables, with the exception of the number of follicles, can be obtained through clinical history, which was encouraging for our aim of building a predictive model using just clinical history variables. The strongest correlation was between the number of follicles and PCOS. The mean number of follicles in the left and right ovary for women with PCOS was 9.8 (±4.3) and 10.8 (±4.2), respectively, whereas the number for women without PCOS was 4.3 (±2.8) and 4.6 (±2.9). This difference was statistically significant ($p<0.001$) and is illustrated in supplemental figure 2.1.

Variables not used in the predictive models were analyzed for differences between women with and without PCOS (supplementary table 2.2). Most variables were not statistically significantly different ($p>0.05$), which further supports their removal from predictive modeling. A statistical difference between these groups existed for years of marriage, hemoglobin and pulse rate. However, these were not added to the models due to lack of scientific rationale for their role in PCOS. Additionally, both hemoglobin and pulse rate would not contribute to our primary model that uses variables obtained through patient history.

Comparing model performance

For each model family (logistic regression, elastic net and random forest), we trained four models corresponding to the groups detailed in **Table 1**. Briefly, model 1 included variables from clinical history, model 2 incorporated clinical examination, model 3 added blood tests and model 4 incorporated ultrasound results. Following training, we predicted the presence of PCOS in the validation data set, and computed sensitivity, specificity, F1 and AUC metrics to test their performance. We found that in each model family, model 4 generally outperformed models 1, 2 and 3. Also, we observed that model 1 outperformed or performed similarly to models 2 and 3 across all metrics. The main classification model of interest (model 1 using only clinical history variables) most effectively balanced out performance and data collection ease. Finally, by comparing model 1 across different model families, the final chosen model was the one trained with elastic net because overall it had the highest sensitivity, as well as the highest F1 score (**Figure 2**).

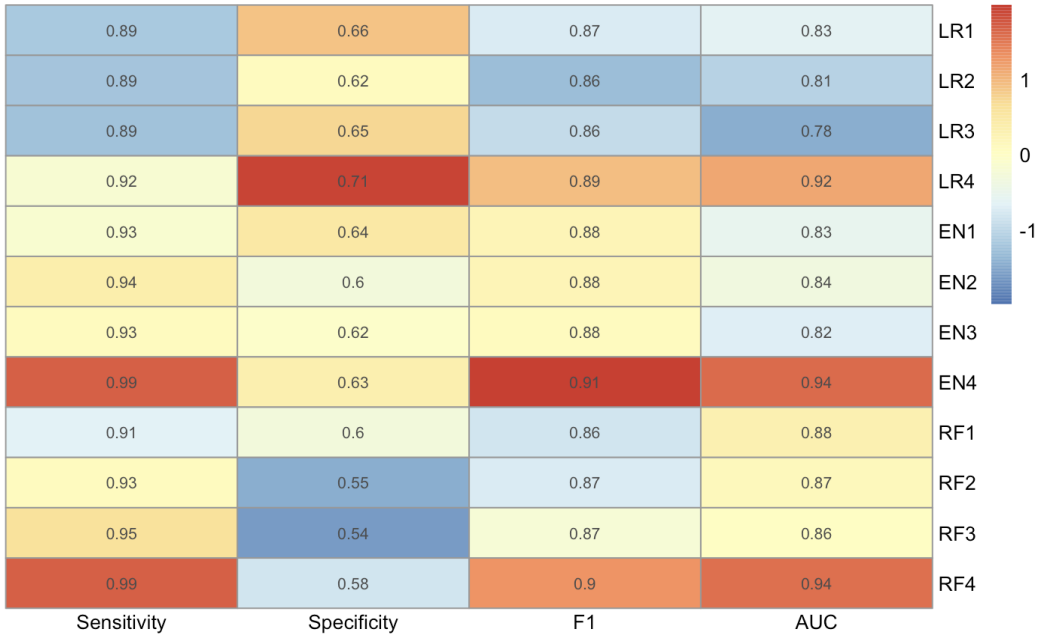


Figure 2. Performance metrics of 4 different models trained with 3 classifier types. The heatmap is scaled by column, with highest Z-scores in a column in red and lowest in blue. Numbers in the cells correspond to raw metric values
Abbreviations: LR = Logistic Regression, EN = Elastic Net, RF = Random Forest.

Exploration of class imbalance and imputation effect

After selecting the elastic net model 1 (EN1), we aimed to explore if using a different method to deal with missing values and correcting class imbalance had any significant impact on our model. We trained 2 additional models with elastic net using clinical history variables on the training set: one model using complete cases (EN1 complete) and another using imputed values for missing data with downsampling to correct for class imbalance (EN1 down). After computing their performance metrics in the validation set (**Figure 3A**), we observed that the downsampled model 'EN1 down' had a lower sensitivity and F1 score than the final chosen model 'EN1'. 'EN1 complete' performed similarly to 'EN1' in specificity, sensitivity and F1, but had a higher AUC by 0.004. Since this difference was very small, we kept the original EN1 model. In

summary, addressing class imbalance with down sampling led to a model with lower performance, and using a different method to deal with missing values didn't have any significant effect, which is what we would expect due to the low missingness present in our dataset.

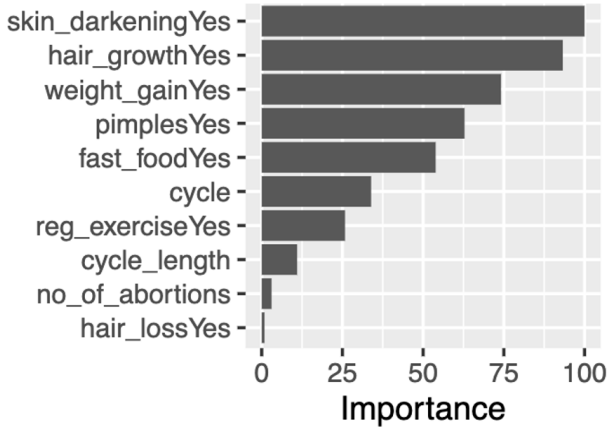
Variable importance and coefficients of the final model

Variable importance of our final model was explored to get an insight into the features that were most relevant in the diagnosis of PCOS. We found skin darkening, hair growth and weight gain to be the 3 variables with the highest weight in the model (**Figure 3B**). We report the coefficients for each variable make the model more usable by others (**Figure 3C**).

A

Name	Specificity	Sensitivity	F1	AUC
EN1	0.642	0.927	0.882	0.832
EN1 complete	0.642	0.927	0.882	0.836
EN1 down	0.679	0.872	0.860	0.842

B



C

Variable	Coeficient
(Intercept)	-1.128
age	-0.060
cycle	0.425
cycle_length	-0.178
no_of_abortions	-0.093
weight_gainYes	0.858
hair_growthYes	1.064
skin_darkeningYes	1.136
hair_lossYes	-0.069
pimplesYes	0.736
fast_foodYes	0.640
reg_exerciseYes	0.338

Figure 3. Final PCOS prediction model. A) Exploration of Elastic Net model 1 sensitivity to using a method to correct for class imbalance and a different way to deal with missingness. B) Variable importance. C) Variable coefficients. *Abbreviations: reg_exercise = regular exercise*

Discussion

In this study, we trained models with logistic regression, elastic net and random forest using variables that can be obtained through clinical history. After evaluating model performance based on sensitivity, our final model was selected to be the one trained with elastic net (sensitivity = 0.93, F_1 = 0.88, AUC = 0.83). We prioritized the model with the highest sensitivity since our aim was to create a model primarily used for screening. Our model could be used for identifying women with risk of having PCOS through online or telephone consultation in South Asia. As only people at high risk of PCOS are directed for further clinical testing using this model, the rate of unnecessary invasive procedures in people with low probability of having this syndrome can be decreased.

Our study also compared models built using information obtained through increasingly more invasive methods of data collection: clinical history, clinical examination, blood tests and transvaginal ultrasound. We found that adding information to clinical history from clinical examination and blood tests did not improve predictive ability of the models, whereas data from transvaginal ultrasound did. This seems to suggest that people at risk of a diagnosis of PCOS based on our final model could proceed directly to confirmatory transvaginal ultrasound without having unnecessary clinical visits and expensive hormonal blood tests.

This work confirms previous study results that suggest skin darkening and hair growth are important diagnostic features of PCOS on South Asian women [3]. The importance of consumption of fast food and regular exercise in our final predictive model for PCOS was somewhat surprising (**Figure 3B**) considering that these factors are often not taken into account in clinical diagnosis and assessment [1]. However, some authors have suggested that increasing consumption of processed foods and ever more sedentary lifestyles are contributing to the rising prevalence of PCOS [11]. Furthermore, dyslipidemia has been linked to PCOS pathophysiology, particularly in South Asian women, and both exercise and diet have effects on the lipid profile of individuals [1-3]. Therefore, our results suggest that these factors warrant further investigation in the pathophysiology of PCOS, particularly in South Asian women. Variable importance in our final model is consistent with previously published papers on the same dataset, which supports the training of our classifier [19].

Interestingly, our exploratory data analysis demonstrated a statistically significant difference between women with and without PCOS in terms of years of marriage, hemoglobin and pulse rate. Whilst years of marriage is unlikely to play any role in the pathophysiology of PCOS, both hemoglobin and pulse rate could be explored further. They were not included in our model due to lack of evidence for their role in PCOS and the fact that they would not have contributed to our primary model using clinical history variables.

Strengths, limitations and ethical considerations

The knowledge-based feature selection and grouping is one of the strengths of our work, since previous models using the same data [19-21] have trained the model using all available variables, even ones highly unlikely to be connected to PCOS pathophysiology (i.e. length of marriage). and others need a clinical test, thus their applicability in a screening context is highly decreased. Though previously published models have a higher performance compared to the one we propose, they would not be suitable for our aim as they utilized variables not available in a screening context (e.g. blood tests). Finally, exploratory data analyses have not been as comprehensive as the one shown in this report, and has not removed biologically non-plausible values have not been removed, which could be a limitation in alternative models.

The major limitations of this project are the lack of detail regarding the publicly available data and lack of external validation. The data source does not specify key study dates, elements of the study setting, eligibility criteria of participants, blinding of outcome assessment, ethical approval for data collection/use, or variables and their codes. Without this information, the model cannot be accurately reproduced and the interpretability of the model is in question. This is compounded by the lack of external validation of the model, which needs to be conducted to ensure adequate model performance. Subsequently, the model should be evaluated concurrent to routine care to ensure real-world applicability and safety monitoring should be implemented. Until such time that the model is externally validated and tested in clinical practice, the model should not be used to guide clinical care. The model is also targeted at women in a South Asian population, and may not be generalizable to other populations or groups of menstruating people. Finally, all effort has been made to ensure our model development is transparent and reproducible through sharing our code and model in a github repository, and the predictive tool uses only variables supported by scientific literature to ensure sensibility and interpretability.

Conclusion

We propose a predictive model for PCOS trained using elastic net with easily collected variables from clinical history. This classifier can be used as a virtual screening tool to help non-physicians predict the need for South Asian patients to undergo further testing based on their probability of having PCOS. This could help reduce in-person burden of physicians and the need for menstruating people to unnecessarily attend consultations in South Asia. However, the model requires further validation prior to clinical use.

References

[1] Joham AE, Norman RJ, Stener-Victorin E, et al. Polycystic ovary syndrome. Lancet Diabetes Endocrinol. 2022;10(9):668-680.

[2] Teede HJ, Misso ML, Costello MF, et al. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Fertil Steril. 2018;110(3):364-379.

[3] Sendur SN, Yildiz BO. Influence of ethnicity on different aspects of polycystic ovary syndrome: a systematic review. Reprod Biomed Online. 2021;42(4):799-818.

[4] Morgante G, Darino I, Spanò A, et al. PCOS Physiopathology and Vitamin D Deficiency: Biological Insights and Perspectives for Treatment. J Clin Med. 2022;11(15):4509.

[5] Muraleedharan M, Chandak AO. Emerging challenges in the health systems of Kerala, India: qualitative analysis of literature reviews. J Health Res. 2022;36(2):242-54.

[6] Álvarez JD, Matias-Guiu JA, Cabrera-Martín MN, et al. An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders. BMC Bioinformatics. 2019;20(1):491.

[7] Vaid A, Somani S, Russak AJ, et al. Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients With COVID-19 in New York City: Model Development and Validation. J Med Internet Res. 2020;22(11):e24018

[8] Xie NN, Wang FF, Zhou J, Liu C, Qu F. Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network. Biomed Res Int. 2020;2020:2613091.

[9] Silva IS, Ferreira CN, Costa LBX, et al. Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models. J Endocrinol Invest. 2022;45(3):497-505.

[10] Kottarathil P. Polycystic ovary syndrome (PCOS). Kaggle; 2020 [cited 2023 February 18]. Available from: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[11] Parker J, O'Brien C, Hawrelak J, Gersh FL. Polycystic Ovary Syndrome: An Evolutionary Adaptation to Lifestyle and the Environment. Int J Environ Res Public Health. 2022;19(3):1336.

[12] van Buuren S, Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45(3):1–67

[13] James G, Witten D, Hastie T, et al. An Introduction to Statistical Learning: with Applications in R. New York; Springer: 2013.

[14] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. J R Stat Soc Series B Stat Methodol. 2005;67:301-20

[15] Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.

[16] Probst P, Boulesteix AL, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. J Mach Learn Res. 2019;20:1-32.

[17] Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-W73.

[18] Navarro E, Tolppa T. PCOS-ML-Prediction. GitHub; 2023 [cited 2023 february 18]. Available from: https://github.com/ErickNavarroD/MachineLearningPCOS

[19] Zigarelli A, Jia Z, Lee H. Machine-Aided Self-diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study. JMIR Form Res. 2022 Mar 15;6(3):e29967.

[20] Khanna VV, Chadaga K, Sampathila N, et al. A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. Appl Syst Innov. 2023;6(2):32.

[21] Nasim S, Almutairi MS, Munir K, Raza A, Younas F. A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics. IEEE Access. 2022;10:97610-97624.

Supplementary Appendix 1: Data Dictionary

Variable name	Units	Data codes	Description
Sl. No	N/A	N/A	Patient identification number
Patient file no.	N/A	N/A	Patient file number
PCOS	N/A	0 = No 1 = Yes	Diagnosis of polycystic ovary syndrome
Age	yrs	N/A	Patient age in years
Weight	kg	N/A	Patient weight in kilograms
Height	cm	N/A	Patient height in centimeters
BMI	kg/m2	N/A	Patient body mass index
Blood group	N/A	11 = A+ 12 = A- 13 = B+ 14 = B- 15 = O+ 16 = O- 17 = AB+ 18 = AB-	Patient blood type
Pulse rate	bpm	N/A	Pulse rate in beats per minute
RR	breaths/min	N/A	Respiration rate in breaths per minute
Hb	g/dl	N/A	Hemoglobin level grams per deciliter
Cycle	N/A	Unclear	Cycle regularity
Cycle length	days	N/A	Cycle length
Marriage status	yrs	N/A	Number of years of marriage
Pregnant	N/A	0 = No 1 = Yes	Pregnancy
No of abortions	N/A	N/A	Number of abortions or miscarriages
I beta-HCG	mIU/mL	N/A	Test 1 for human chorionic gonadotropin (hCG)
II beta-HCG	mIU/mL	N/A	Test 2 for human chorionic gonadotropin (hCG)
FSH	mIU/mL	N/A	Follicle stimulating hormone (FSH)
LH	mIU/mL	N/A	Luteinizing hormone (LH)
FSH/LH	N/A	N/A	Ratio between FSH and LH

Variable name	Units	Data codes	Description
Hip	inch	N/A	Hip width in inches
Waist	inch	N/A	Waist width in inches
Waist/Hip Ratio	N/A	N/A	Waist to hip ratio
TSH	mIU/L	N/A	Thyroid stimulating hormone
AMH	ng/mL	N/A	Anti-Müllerian hormone
PRL	ng/mL	N/A	Serum prolactin
Vit D3	ng/mL	N/A	Vitamin D3
PRG	ng/mL	N/A	Progesterone
RBS	mg/dl	N/A	Random blood sugar (glucose) test
Weight gain	N/A	0 = No 1 = Yes	Weight gain
Hair growth	N/A	0 = No 1 = Yes	Hair growth
Skin darkening	N/A	0 = No 1 = Yes	Skin darkening
Hair loss(Y/N)	N/A	0 = No 1 = Yes	Hair loss
Pimples(Y/N)	N/A	0 = No 1 = Yes	Pimples
Fast food (Y/N)	N/A	0 = No 1 = Yes	Fast food
Reg.Exercise(Y/N)	N/A	0 = No 1 = Yes	Regular exercise
BP Systolic	mmHg	N/A	Systolic blood pressure
BP Diastolic	mmHg	N/A	Diastolic blood pressure
Follicle No. (L)	N/A	N/A	Number of follicle (left ovary)
Follicle No. (R)	N/A	N/A	Number of follicle (right ovary)
Avg. F size (L)	mm	N/A	Average follicle size in left ovary
Avg. F size (R)	mm	N/A	Average follicle size in rightt ovary
Endometrium	mm	N/A	Thickness of the endometrium

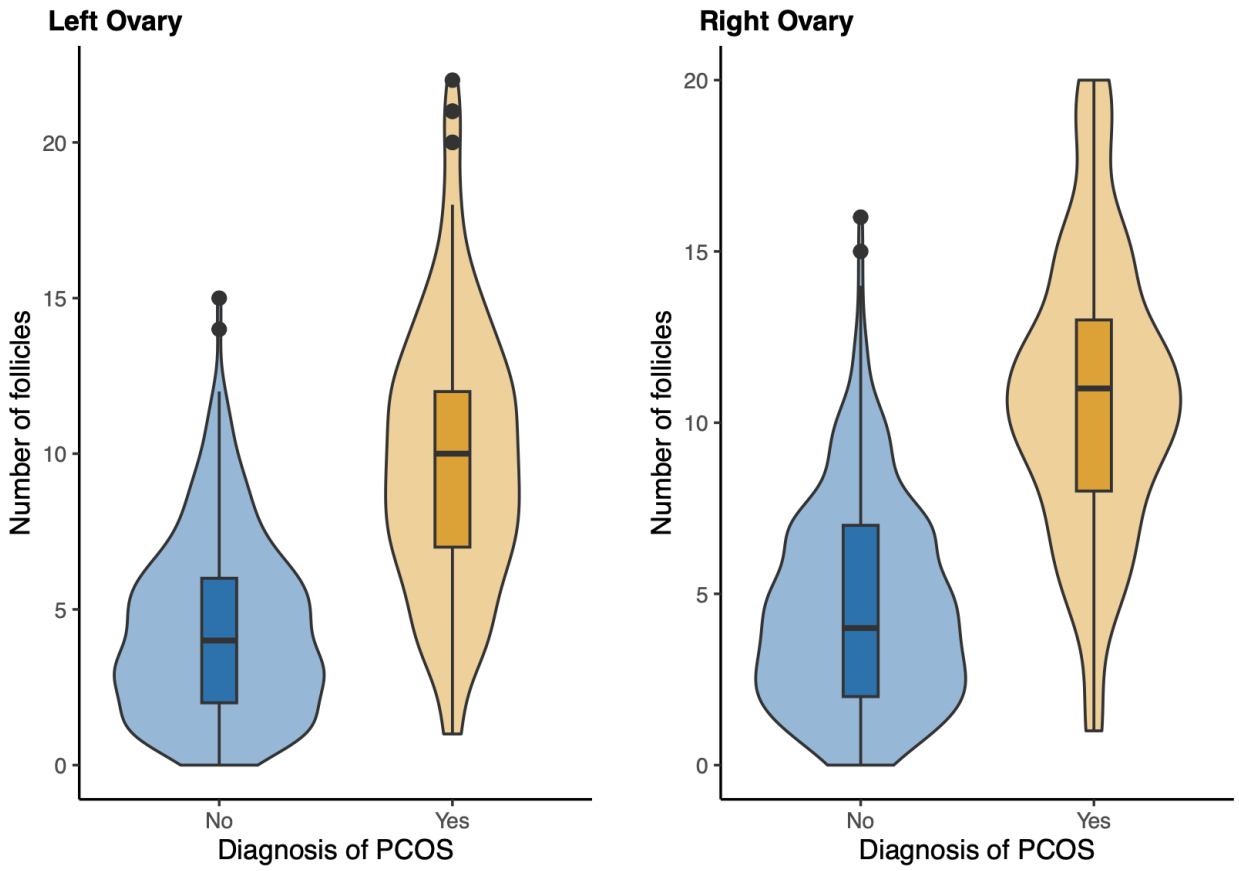
Supplementary table 1.1. Data dictionary for the dataset, adapted from [10]

Supplementary Appendix 2: Exploratory Data Analysis

The full exploratory analysis is provided in a publicly available repository as a PDF and with the original code [17]. Below are tables and figures most relevant to this project report.

Variable	Number of missing values, n (%)	Source
Pulse rate	2 (0.4%)	Outlier analysis
Marriage status	1 (0.2%)	Original data
FSH	1 (0.2%)	Outlier analysis
LH	1 (0.2%)	Outlier analysis
FSH/LH ratio	2 (0.4%)	Outlier analysis
Vit D3	2 (0.4%)	Outlier analysis
AMH	2 (0.4%)	Original data and outlier analysis
Fast food	1 (0.2%)	Original data
Systolic blood pressure	1 (0.2%)	Outlier analysis
Diastolic blood pressure	1 (0.2%)	Outlier analysis

Supplementary table 2.1. Number and source of missing variables (N=541)



Supplementary figure 2.1. Comparison of the number of follicles in each ovary in women with and without PCOS

Variable	Total (N = 541)	PCOS Positive (N = 177)	PCOS Negative (N = 364)	p-value for difference
Pregnant, n (%)				0.521
Yes	206 (38.1)	64 (36.2)	142 (39.0)	
No	335 (61.9)	113 (63.8)	222 (61.0)	
i-beta-HCG (mIU/mL), mean (SD)	664.6 (3348.9)	532.0 (2922.9)	729.9 (3539.6)	0.494
ii-beta-HCG (mIU/mL), mean (SD)	238.2 (1603.83)	267.6 (1905.7)	224 (1437.0)	0.788
Length of marriage (years), mean (SD)	7.7 (4.8)	6.9 (4.7)	8.0 (4.8)	0.008*
Blood group, n (%)				0.932
A+	108 (20.0)	34 (19.2)	74 (20.3)	
A-	13 (2.4)	4 (2.3)	9 (2.5)	
B+	135 (24.9)	42 (23.7)	93 (25.5)	
B-	16 (2.9)	6 (3.4)	10 (2.7)	
O+	206 (38.1)	66 (37.3)	140 (38.5)	
O-	19 (3.5)	8 (4.5)	11 (3.0)	
AB+	42 (7.8)	16 (9.0)	26 (7.1)	
AB-	2 (0.4)	1 (0.6)	1 (0.3)	
TSH (mIU/L), mean (SD)	3.0 (3.7)	2.9 (2.8)	3.0 (4.1)	0.789
Respiratory rate, mean (SD)	19.2 (1.7)	19.3 (1.6)	19.2 (1.7)	0.385
Pulse rate, mean (SD)	73.5 (2.7)	73.8 (2.7)	73.3 (2.7)	0.028*
Hemoglobin (g/dl), mean (SD)	11.2 (0.9)	11.3 (0.8)	11.1 (0.9)	0.039*

Supplementary table 2.2. Comparison of variables not included in predictive models between women with and without PCOS. * denotes statistical significance for a difference