Building a Predictive Machine Learning Model to Identify Polycystic Ovary Syndrome Using Easily Measured Clinical or Physiological Parameters

Jacob and Hanwei

2023-02-27

1. Introduction and Background

1.1 Introduction

Our overarching goal for this project is to produce a machine learning (ML) model which can accurately predict polycystic ovary syndrome (PCOS) status (presence or absence of disease) using predictors which may be easily acquired from common clinical settings, for instance, information acquired from a standard blood test and routine clinical examination. PCOS is an endocrine (hormonal) disorder that affects females of a reproductive age[1]. Given the widespread nature of this condition among women of reproductive age and the troubling symptoms which accompany it, including infertility, it would be helpful for physicians to be able to predict, using inexpensive, minimally-invasive and readily available methods, individuals more likely to experience PCOS thereby enabling them to therapeutically intervene, support, advise or provide care in a timely manner, especially considering PCOS has been associated with other conditions such as endometriosis and endometrial cancer[2,5].

1.2 Background

Polycystic ovarian syndrome (PCOS) is a common endocrine disorder affecting approximately 10-15% of reproductive-age women worldwide[1]. The condition is characterized by a complex set of symptoms, including hyperandrogenism, menstrual irregularities, and polycystic ovaries[2]. The diagnosis of PCOS is typically based on clinical and biochemical assessments, as well as ultrasound imaging of the ovaries[1]. However, the diagnosis of PCOS can be challenging due to the heterogeneous presentation of symptoms and the lack of a single diagnostic criterion.

Machine learning models have shown promise as a potential tool for the accurate prediction of PCOS. Compared to traditional diagnostic methods, machine learning models can utilize large amounts of data from various sources and provide more accurate predictions. This is particularly beneficial in the case of PCOS, as traditional diagnostic methods such as tissue biopsy or ultrasound imaging can be expensive and invasive. Furthermore, machine learning models can assist clinicians in identifying patients who may benefit from early intervention, which can improve long-term health outcomes[1].

Our aim for the PCOS predictive model we are building would be to assist clinicians with identifying women between the ages of 21-47 at highest risk or likely to develop PCOS. Since current methods of PCOS diagnosis often require specialized equipment, we sought to develop a model that used a subset of easily measurable clinical and physiological parameters to predict PCOS, thus enabling early diagnoses to be made and enabling specialized resources to be used in fewer patients as a confirmatory test.

The appropriate ethical approval from Ethics Research Boards was obtained for this ML project. To help ensure privacy fo this ML project, we worked with de-identified anonymized data. In the event an individual

was identified as being pregnant or may be at risk of a life-threatening condition, we followed-up immediately with the individual's physician.

The rationale for developing this predictive ML model is based on the motivation to (1) provide a data-driven method of diagnosis that is cheaper and less invasive than current methods, (2) develop a method which may be applied without expensive diagnostic equipment (such as an ultrasound imaging device), (3) develop a method which may be used with just a blood test and the review of clinical symptoms and (4) construct a method which relies minimally on self-reported data (as this data can be highly variable and sometimes unreliable). As discussed in the background section of this report, current diagnosis of PCOS can be time-consuming and invasive, and replacing these methods with a model is advantageous for reasons related to clinical care, resource utilization and accessibility. This influenced our variable selection, as we did not consider the inclusion of predictors that cannot be measured in the above stated context in model development. We therefore, for example, exclude self-reported variables or variables which required ultrasound imaging (for example, endometrium thickness).

While accuracy was a key consideration for evaluating model performance, we also focused on optimizing the specificity of the model predictions. The rationale for optimizing specificity was based on the relatively high prevalence of PCOS[7] as well as the relatively low mortality associated with the disease[1,2,5]. Since patients flagged for high risk of PCOS would receive a confirmatory ultrasound, we wanted to select a model that would be biased towards reducing false positives, as one of our goals was reducing the cost of care.

Overall, the use of machine learning models to predict PCOS has the potential to improve the accuracy of diagnosis and reduce the cost and invasiveness of traditional diagnostic methods. We sought to evaluate the comparative accuracy of multiple machine learning classifiers to select the optimal model for predicting PCOS based on variables obtained via a blood test and routine clinical examination, given considerations surrounding false positive and false negatives.

1.3 Data

The dataset consists of physical and clinical parameters collected from 10 hospitals across Kerala, India, and from 541 women, to determine PCOS and infertility-related issues. The dataset contains information that can be used to analyze and understand the diagnosis and treatment of PCOS and infertility.

2. Objectives

Our objectives were twofold:

1: Develop a simple model that can predict PCOS status using clinical and physiologic data that can be acquired using a routine blood test and assessment by a general practitioner clinician.

2: Optimize the model for specificity, thus minimizing the risk of a false positive in model predictions.

Optimizing model specificity for the prediction of PCOS is important because PCOS is associated with several negative health outcomes, including infertility, insulin resistance, and metabolic disorders. Early diagnosis and treatment of PCOS can help prevent or manage these conditions, which can ultimately improve the overall health and quality of life of those affected. However, given that PCOS is not a life-threatening condition in and of itself, it is important to balance the trade-off between maximizing sensitivity and specificity in order to minimize the number of false positives and prevent unnecessary and potentially invasive follow-up testing. By optimizing model specificity, we can ensure that those who are diagnosed with PCOS are more likely to truly have the condition, while also reducing the risk of unnecessary medical interventions for those who do not have PCOS.

3. Methods

3.1 Splitting our dataset into training and testing datasets

To ensure data privacy, we worked with de-identified anonymized data. Following our Exploratory Data Analysis (EDA, please refer to Results section for EDA), we started by splitting our dataset into the training set and the testing dataset, followed by building a simple model aimed at using our data to find factors that predict PCOS status (our outcome variable). We chose to partition our data into only two sets based on the relatively small overall sample size of our data and small effective sample size of our data. Further, to help determine generalizability, we believe it would be ideal for our predictive model to be validated on an external dataset derived from a different population than our training and test data. For example, an external validation could be performed on PCOS hospital data from other southern Indian states such as Tamil Nadu or Karnataka. For an even better gauge of model generalizability, the external validation could be performed on PCOS hospital data from other countries.

3.2 Variable selection and building logistic regression models

We first attempted to model our data with PCOS status as the outcome variable using three different logistic regression models. At this stage of variable selection, we undertook a combined approach to informing variable selection. The Akaike Information Criteria (AIC) was used to assess model fit, the Area Under the Receiving Operator Characteristic (AUC) was used to assess model accuracy, and the model with the best balance between both of these parameters, as well as clinical utility was used. The three tested models utilized the following construction:

- (1) Model 1 is the physiological model which includes five hormones of interest, namely hormone measurements of n=5 different types of hormones commonly included or easily measured/acquired as a part of a standard routine blood test.
- (2) Model 2 is the clinical model, including as variables clinical symptoms associated with PCOS, and specifically clinical symptoms which may be easily determined as a part of a routine examination by a physician. These are the variables included in the clinical model: weight gain, hair growth + skin darkening, hair loss and pimples.
- (3) Model 3 includes both the physiological and clinical variables, that is the n=5 the hormone measurement levels we as well as the clinical symptoms that have been known to be associated with PCOS: weight gain, hair growth + skin darkening, hair loss and pimples (clinical symptoms which may be easily identified or tracked in a basic clinical setting).

Following this initial variable selection, we selected the two best performing models and applied various machine learning training methods to optimize model performance. We then selected the best performing model as the final model.

Of our three logistic regression models, Model 2 had the lowest AIC score of 317, followed by Model 3 with an AIC score of 566. However, the predictive performance as assesed by AUC was higher for model 3 (0.99) than model 2 (0.88). While model 3 was better performing, it also required more parameters for prediction that require blood testing. The rationale for testing both models with machine learning optimization was to assess if a simpler model with clinical parameters only could approximate the classification performance of a more complex model.

4. Results

4a. Exploratory Data Analysis (EDA)

This section contains the steps and output for the EDA performed on the PCOS dataset. The section proceeds sequentially with each step of EDA (please see Appendix for more details). The first steps in our

EDA involved getting a high-level overview of our dataset and determining the dimension of our data: n=541 rows and n=45 columns. Further checking of the dataset reveals an additional column (column 45). This column does not contain any useful information and is not one of our 44 features, therefore the column was removed.

4a.1 Data Wrangling

The data was formatted to ensure the variables are the appropriate class type, this will enable us to perform our EDA. Specifically we converted binary variables (1 or 0) into the character class type.Next in our EDA we sought to identify missing values in our dataset. Figure A1 in Appendix A shows the missing values in our dataset. Our EDA identified some missing values for two variables: fast food and marriage status. The analysis indicates that only 0.18% of the rows for these variables are missing. Therefore, as this is below the generally used threshold of 5 %, missing values were simply removed for our subsequent analyses.

Table 1: Clinical and sociodemographic characteristics of the study cohort (n=541, women from Kerala, India)

Now that the data is cleaned, we can produce the standard table 1, common in biomedical research for describing the study cohort.

Pruned table1

For this table, we have included only the parameters we deemed eligible for the model. Thus, we have excluded self-report variables, expensive and/or invasive tests, and clinical variables not easily measured by a routine clinical examination or a blood test.

	Negative	Positive	Overall
	(N=364)	(N=177)	(N=541)
age_yrs			
Mean (SD)	32.1 (5.36)	30.1 (5.29)	31.4 (5.41)
Median [Min, Max]	32.0 [20.0, 48.0]	29.0 [21.0, 47.0]	31.0 [20.0, 48.0]
bmi			
Mean (SD)	23.7 (3.76)	25.5 (4.40)	24.3 (4.06)
Median [Min, Max]	23.6 [13.4, 38.3]	25.1 [12.4, 38.9]	24.2 [12.4, 38.9]
pulse_rate_bpm			
Mean (SD)	73.0 (5.03)	73.8 (2.73)	73.2 (4.43)
Median [Min, Max]	72.0 [13.0, 82.0]	72.0 [70.0, 82.0]	72.0 [13.0, 82.0]
rr_breaths_min			
Mean (SD)	19.2 (1.71)	19.3 (1.65)	19.2 (1.69)
Median [Min, Max]	18.0 [16.0, 28.0]	20.0 [16.0, 24.0]	18.0 [16.0, 28.0]
hb_g_dl			
Mean (SD)	11.1 (0.880)	11.3 (0.831)	11.2 (0.867)
Median [Min, Max]	11.0 [8.50, 14.8]	11.0 [9.40, 14.0]	11.0 [8.50, 14.8]
pregnant_y_n			
Mean (SD)	0.390 (0.488)	0.362 (0.482)	0.381 (0.486)
Median [Min, Max]	0 [0, 1.00]	0 [0, 1.00]	0 [0, 1.00]
as.numeric(i_beta_hcg_m_iu_m_l)			
Mean (SD)	729 (3540)	532 (2920)	665 (3350)
Median [Min, Max]	13.7 [1.30, 32500]	70.5 [1.92, 30000]	20.0 [1.30, 32500]
as.numeric(fsh_m_iu_m_l)			
Mean (SD)	19.2 (265)	5.17 (5.74)	14.6 (217)

	Negative	Positive	Overall
Median [Min, Max]	5.01 [0.210, 5050]	4.48 [1.00, 65.4]	4.85 [0.210, 5050]
as.numeric(lh_m_iu_m_l)			
Mean (SD)	2.61 (2.10)	14.4 (151)	6.47 (86.7)
Median [Min, Max]	2.31 [0.0200, 14.7]	2.22 [0.0320, 2020]	2.30 [0.0200, 2020]
as.numeric(tsh_m_iu_l)			
Mean (SD)	3.01 (4.14)	2.93 (2.82)	2.98 (3.76)
Median [Min, Max]	2.17 [0.0400, 65.0]	2.31 [0.0500, 22.6]	2.26 [0.0400, 65.0]
as.numeric(amh_ng_m_l)			
Mean (SD)	4.54 (4.29)	7.84 (7.79)	5.62 (5.88)
Median [Min, Max]	3.20 [0.160, 26.8]	5.90 [0.100, 66.0]	3.70 [0.100, 66.0]
Missing	1 (0.3%)	0 (0%)	1 (0.2%)
as.numeric(prl_ng_m_l)			
Mean (SD)	24.3 (15.5)	24.4 (13.9)	24.3 (15.0)
Median [Min, Max]	21.2 [0.400, 128]	22.9 [3.64, 112]	21.9 [0.400, 128]
as.numeric(vit_d3_ng_m_l)			
Mean (SD)	29.3 (12.4)	92.3 (604)	49.9 (346)
Median [Min, Max]	26.3 [9.01, 90.0]	25.5 [0, 6010]	25.9 [0, 6010]
as.numeric(prg_ng_m_l)			
Mean (SD)	0.727 (4.64)	0.372 (0.174)	0.611 (3.81)
Median [Min, Max]	0.310 [0.110, 85.0]	0.320 [0.0470, 1.10]	0.320 [0.0470, 85.0]
as.numeric(rbs_mg_dl)			
Mean (SD)	99.2 (15.5)	101 (23.6)	99.8 (18.6)
Median [Min, Max]	96.0 [60.0, 225]	100 [70.0, 350]	100 [60.0, 350]
weight_gain_y_n			
0	281 (77.2%)	56 (31.6%)	337 (62.3%)
1	83 (22.8%)	121 (68.4%)	204 (37.7%)
hair_growth_y_n			
0	317 (87.1%)	76 (42.9%)	393 (72.6%)
1	47 (12.9%)	101 (57.1%)	148 (27.4%)
skin_darkening_y_n			
0	308 (84.6%)	67 (37.9%)	375 (69.3%)
1	56 (15.4%)	110 (62.1%)	166 (30.7%)
hair_loss_y_n			
0	221 (60.7%)	75 (42.4%)	296 (54.7%)
1	143 (39.3%)	102 (57.6%)	245 (45.3%)
pimples_y_n			
0	222 (61.0%)	54 (30.5%)	276 (51.0%)
1	142 (39.0%)	123 (69.5%)	265 (49.0%)

From table 1, we can observe some differences among women with and without PCOS in the following parameters: Clinical: cycle_r_i, cycle_length_days, weight_gain_y_n, hair_growth_y_n, skin_darkening_y_n, hair_loss_y_n, pimples_y_n, reg_exercise_y_n

Physiologic: i_beta_hcg_m_iu_m_l (Human chorionic gonadotropin), fsh_m_iu_m_l (Follicle Stimulating hormone), lh_m_iu_m_l (Lutenizing hormone), amh_ng_m_l (Anti-mullerian hormone), vit_d3_ng_m_l (Vitamin D), prg_ng_m_l (Progesterone).

We can visualize these distributions in the following figures.

4a.2 Bivariate associations

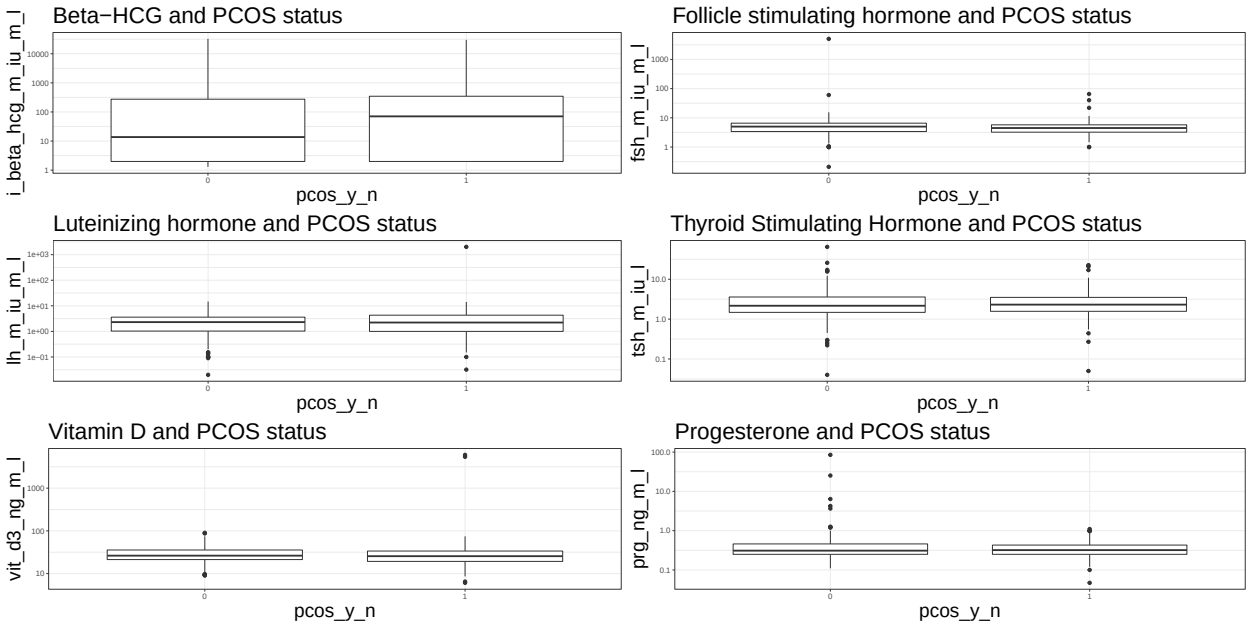
We visualized associations for selected categorical predictors and the outcome (PCOS) by constructing bivariate plots for the predictors stratified by PCOS status[4].

From the figure, we can see that the proportion of individuals with weight gain and a positive PCOS status is greater than the proportion of individuals with no weight gain and a positive PCOS status. We observe similar associations for hair growth and skin darkening. We also observe a similar but much weaker association for the categorical variables of hair loss and pimples. Interestingly, we observe the inverse trend for the categorical variable for fast food consumption. As these predictors appear to be visually associated with the outcome, as well as physiologically feasible to be associated with PCOS, they should be considered for inclusion in the predictive model.

The bar plots for regular exercise and pregnancy do not appear to show a significant difference in proportion of positive PCOS cases, but we would need to perform a statistical analyses (such as an unpaired t-test) to know for sure. To ascertain the relationship between PCOS status and continuous predictors, we also plotted boxplots to visualize any associations between PCOS status and a continuous variable. Please refer to the appendix for these boxplots.

For this section of the EDA, we looked more closely at the relationships between selected variables and our outcome variable of interest (for this project: yes or no for PCOS), based on how our boxplots looked initially (please refer to Appendix for all boxplots) and on some knowledge gained via our literature survey, we will highlight the relatiionship of select variables with our outcome variable (PCOS status). While the etiology for PCOS is not known, our literature survey suggests PCOS is associated with abnormal hormone levels. Thus, to look into this further as part of the EDA, we compared measurements of hormone levels and PCOS status to get a sense of the relationships between these variables.

Boxplots looking at relationship of selected physiologic parameters and PCOS status (note log-scale)



To get a sense of the variation in our dataset, we plotted histograms for each continuous variable using the plot_histogram() from the DataExplorer Package. We see that the age distribution in our dataset reveals most individuals are in the range of 20 to 40 years. No individuals in the dataset are younger than 20 or older than 50. As we would expect, the BMI values follow an approximately normal distribution. The most common blood type we observe is O+, which is consistent with the fact that O+ is the most frequent bloodtype globally. The most common cycle length is 5 days. The endometrium thickness data suggests there are two most most common thickness values (two clear peaks in the distribution).

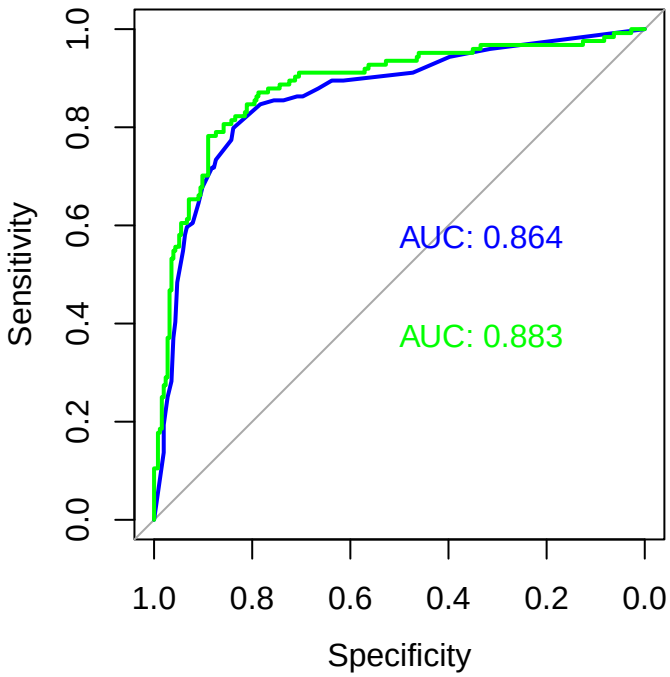
Modelling

Based on these results, we will compare the clinical model to the full clinical and physiological model. While the model including clinical predictors only is more simple both technically (as evidenced by the AIC) and practically (less inputs and no blood testing required), there is an approximately 10% drop in accuracy when excluding the physiological data. We will select these two models and compare performance in applying various machine learning methods below.

4.1 Assessing model classification performance

- Assessing model performance using ROC curves

Figure 1. ROC curves: clinical model (blue) vs. clinical and physiologic model (green)



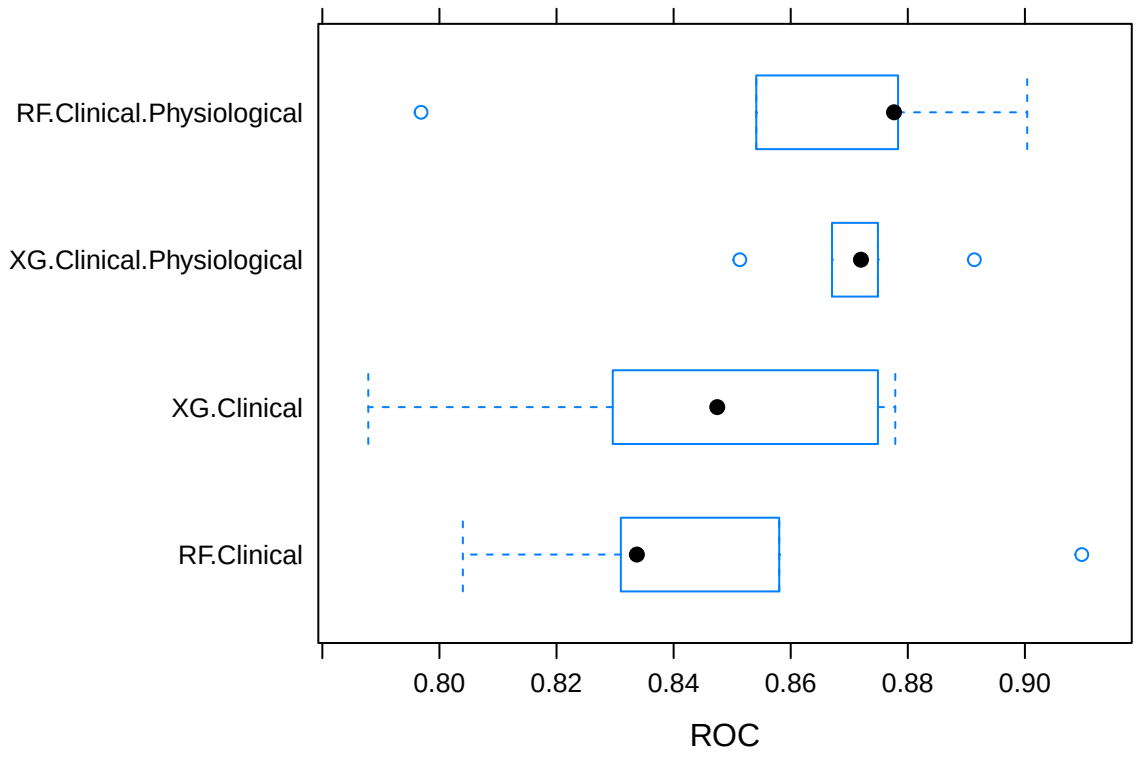
4.2 Modelling our data using Trees and Forests

We next pursued a Random Forest approach for our selection of variables, using both the Clinical and Clinical + Physiological Parameters for our Random Forest Modelling. Random Forest modelling was also used to examine variable importance.

Random Forest

Model	ROC	Sensitivity	Specificity
Clinical and Physiologic Model	0.8614706	0.8938039	0.6616667
Clinical Model	0.8472935	0.8973333	0.6466667

Figure X. Comparative performance on the training data



Summarised performance

Table 2. Comparative performance of two models across three different modelling methodologies

Model	Accuracy	Sensitivity	Specificity
Clinical Parameters: Random Forest	0.88125	0.7115385	0.9629630
Clinical and Physiologic Parameters: Random Forest	0.87500	0.6923077	0.9629630
Clinical Parameters: XGBoost	0.83125	0.5961538	0.9444444
Clinical and Physiologic Parameters: XGBoost	0.81875	0.4807692	0.9814815
LASSO: Clinical Parameters	0.69375	0.5882353	0.7062937
LASSO: Clinical and Physiologic Parameters	0.85625	0.8918919	0.8455285

5. Discussion and Conclusion

5.1 Discussion

We found the best ML model which balances accuracy, specificity and clinical utility to be the Random Forest model with clinical parameters (n=5 clinical symptoms: weight gain, hair growth + skin darkening, hair loss and pimples). This model performed with the following accuracy = 88.1%, sensitivity = 71.2% and specificity = 96.3% when applied to our test dataset. Not only did this model have the highest accuracy of

the six models we compared, the model also performed well on sensitivity and specificity. We opted to use a Random Forest approach as part of trying different ML models as a recent study in the literature indicating Random Forest models perform the most accurately on a wide variety of datasets[6]. One of the drawbacks of the Random Forest is its black box nature, making it harder to determine and/or interpret how the model is making its predictions. This could pose a challenge to clinicians trying to troubleshoot our improve upon the model. Given our use of physiological parameters such as the hormones human chorionic gonadotropin (HCG) and anti-mullerian hormone (AMH), which may be used as proxies to determine pregnancy status and fertility respectively, it is important that privacy is maintained and also physicians have access to the model's de-anonymized input data.

5.2 Conclusions

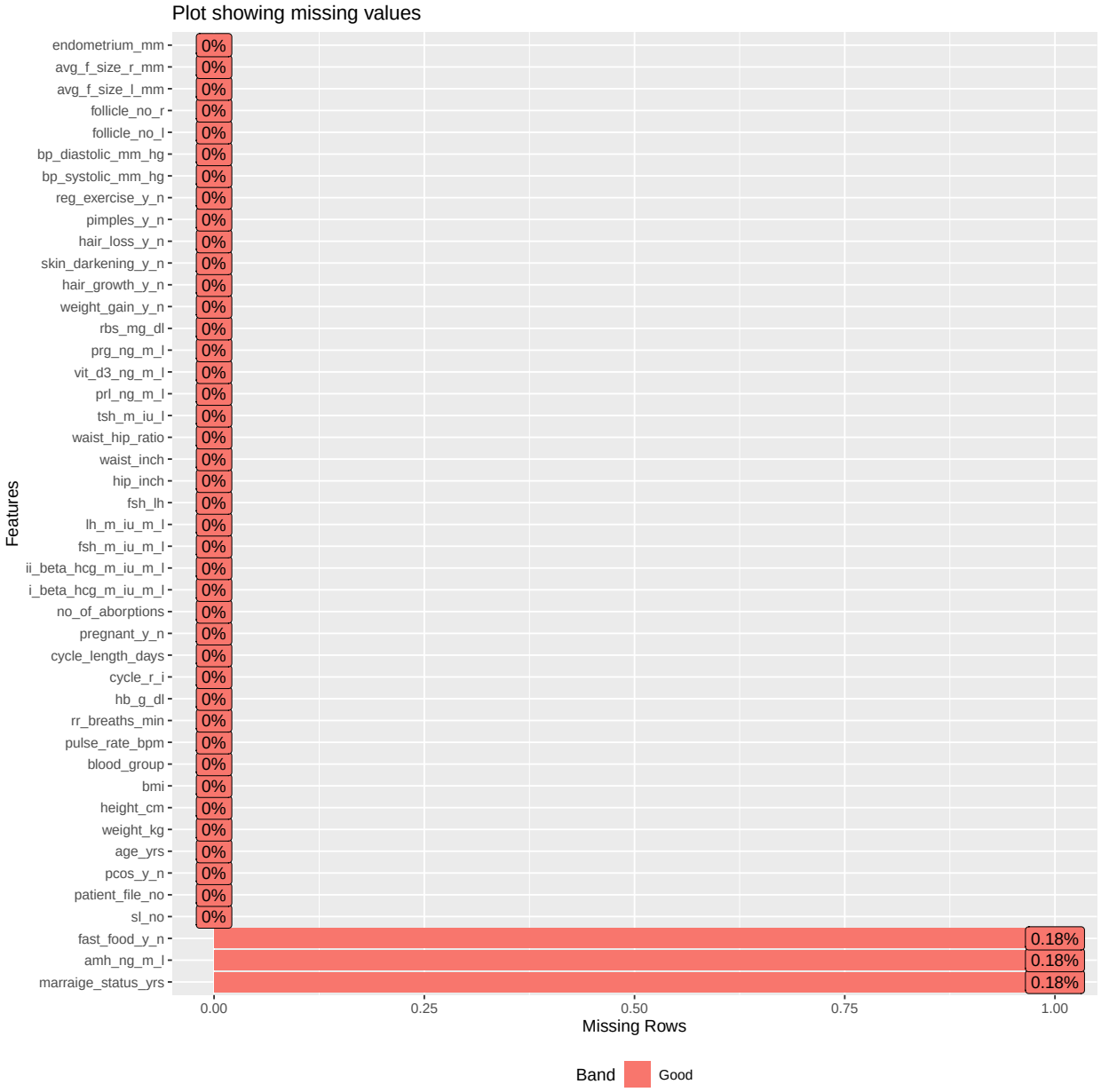
PCOS status can be predicted with a reasonably high level of accuracy (88.1% with our model) using a combination of clinical and physiological parameters that are easily measurable in a general clinical practice setting. By optimizing for specificity, patients likely to be PCOS positive can be selected for more expensive and invasive testing, thus reducing the cost of care and improving patient experience for the population of at-risk women. Implementation of model-based screening should be considered particularly in resource-limited settings, where access to advanced diagnostics is challenging due to cost or geography.

References

1. Teede, H., Misso, M., Tassone, E. C., Dewailly, D., Ng, E. H., Azziz, R., ... & Norman, R. J. (2018). Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Human Reproduction, 33(9), 1602-1618. https://doi.org/10.1093/humrep/dey256
2. Bozdag, G., Mumusoglu, S., Zengin, D., & Karabulut, E. (2016). The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis. Human Reproduction, 31(12), 2841–2855. https://doi.org/10.1093/humrep/dew218
3. R for Data Science by Hadley Wickham (https://r4ds.had.co.nz)
4. Ajmal N, Khan SZ, Shaikh R. Polycystic ovary syndrome (PCOS) and genetic predisposition: A review article. Eur J Obstet Gynecol Reprod Biol X. 2019 Jun 8;3:100060. doi: 10.1016/j.eurox.2019.100060. PMID: 31403134; PMCID: PMC6687436.
5. Hart R, Doherty DA. The potential implications of a PCOS diagnosis on a woman's long-term health using data linkage. J Clin Endocrinol Metab. 2015 Mar;100(3):911-9. doi: 10.1210/jc.2014-3886. Epub 2014 Dec 22. Erratum in: J Clin Endocrinol Metab. 2015 Jun;100(6):2502. PMID: 25532045.
6. Fernández-Delgado, M., Cernadas E., Barro S., Amorim D. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research. Volume 15. Issue 1, pp 3133–3181.
7. Deswal R, Narwal V, Dang A, Pundir CS. The Prevalence of Polycystic Ovary Syndrome: A Brief Systematic Review. J Hum Reprod Sci. 2020 Oct-Dec;13(4):261-271. doi: 10.4103/jhrs.JHRS_95_18. Epub 2020 Dec 28. PMID: 33627974; PMCID: PMC7879843.

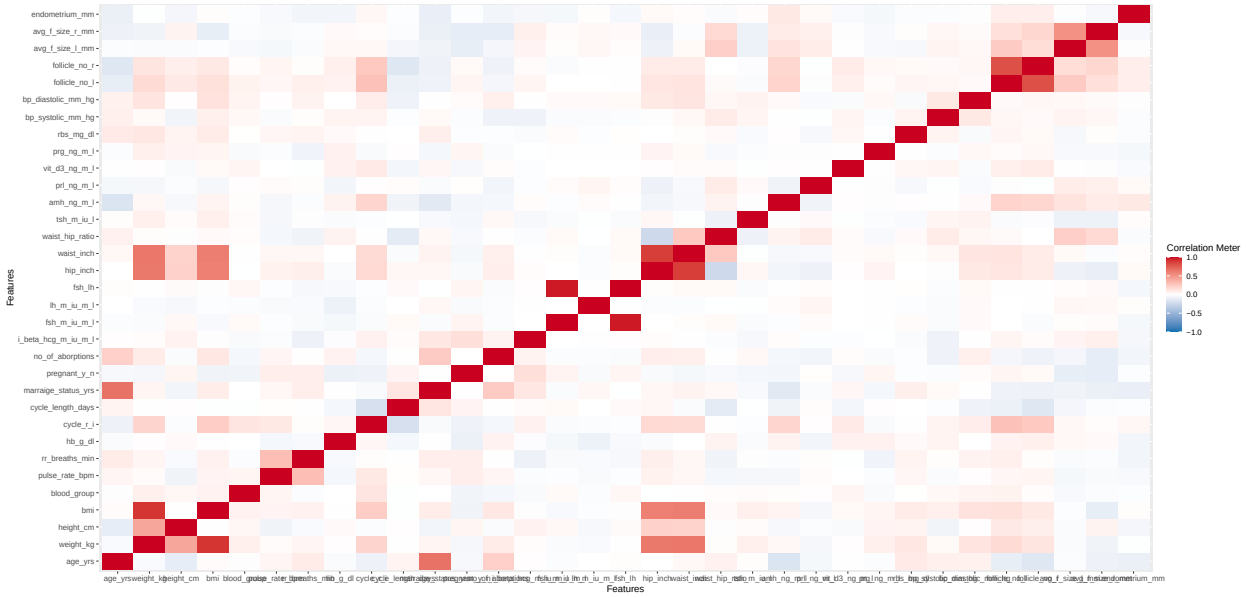
Appendix

- Appendix Plot 1, as part of EDA, showing missing values in our dataset



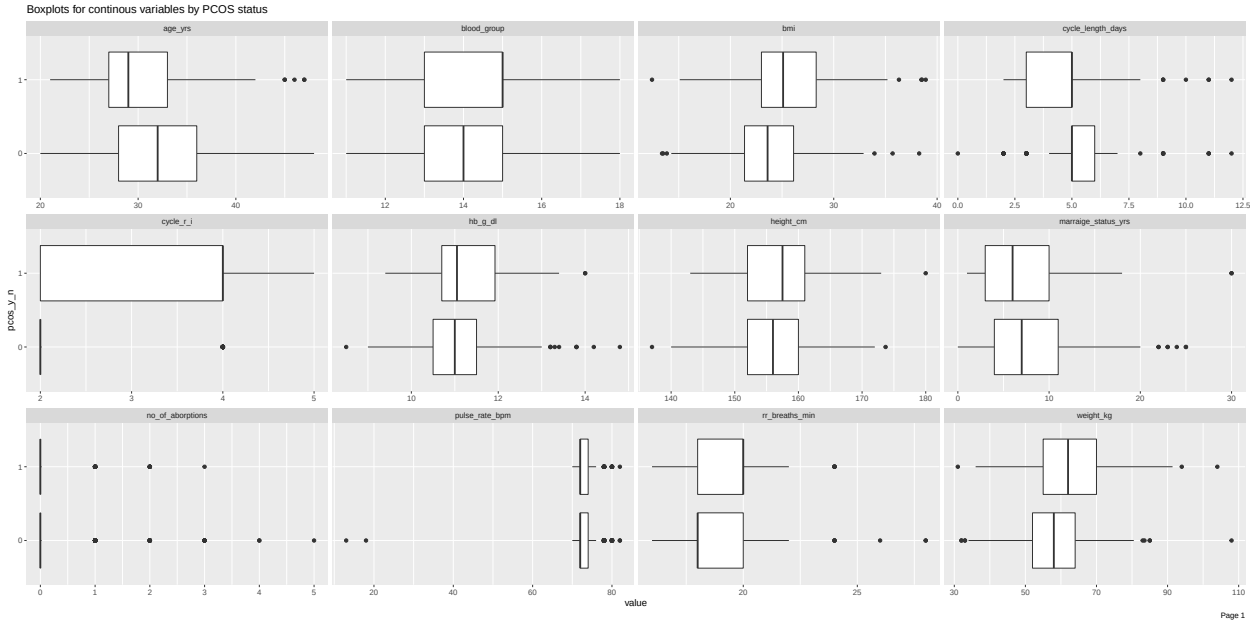
- Appendix plot 2, as part of EDA, examining correlations between continuous features

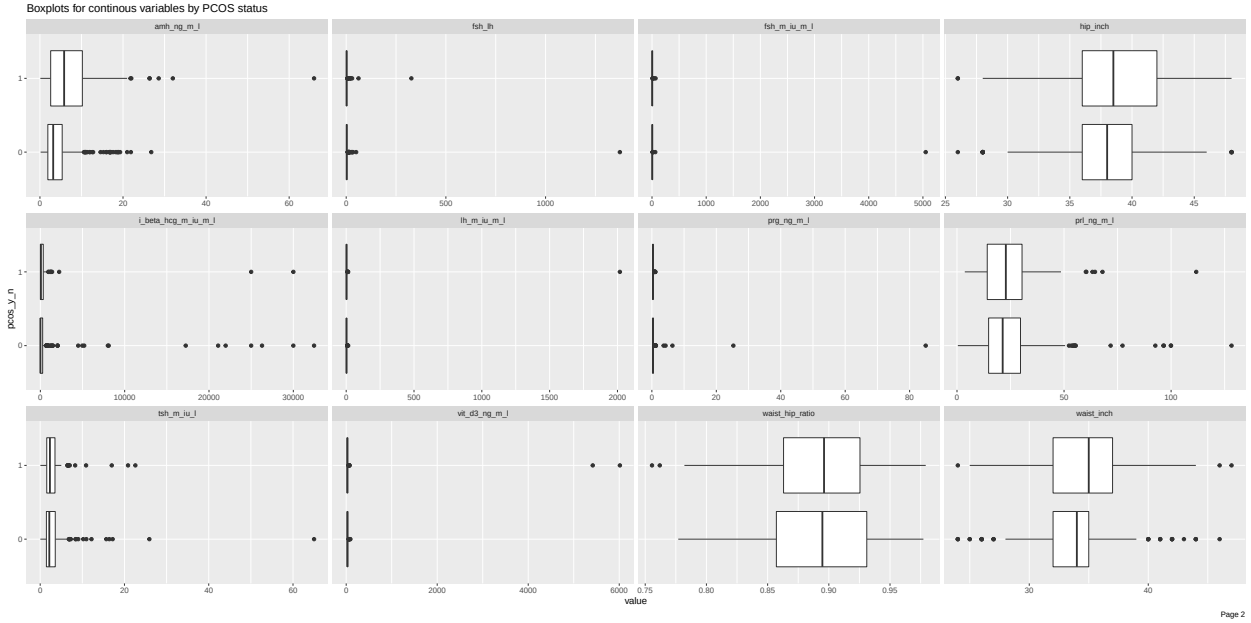
Bivariate analysis to visualize our covariation between our continuous variables

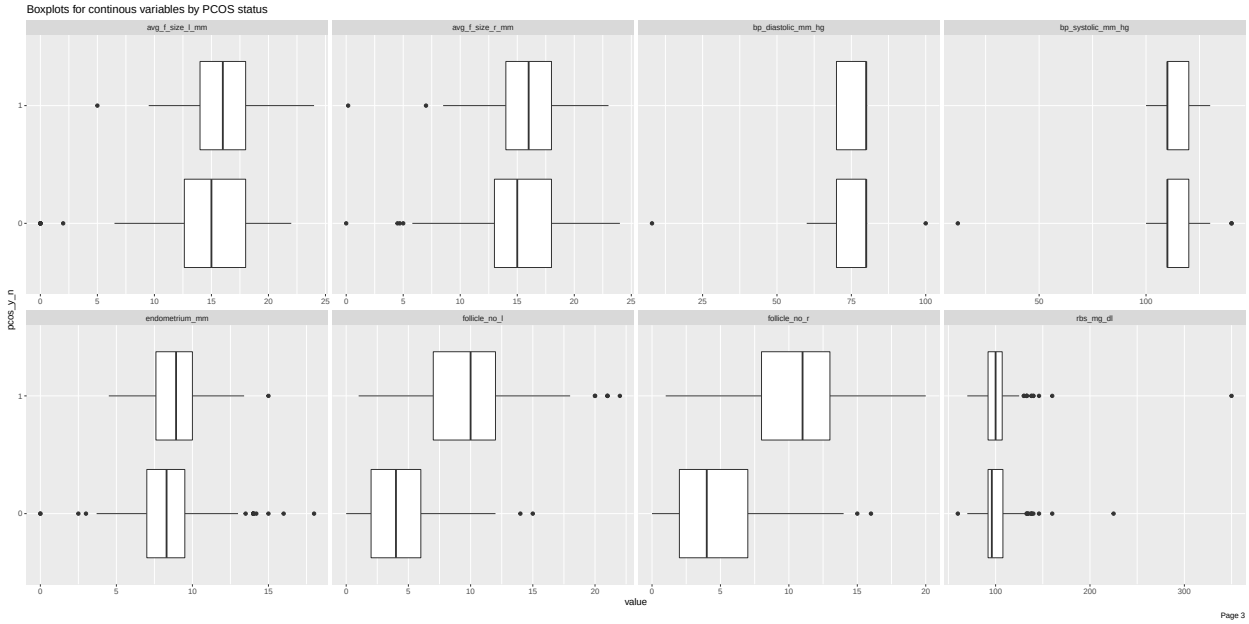


- From this correlation plot, we find that several continuous variables do co-vary with one another.
- Specifically, as we would expect, we find a positive correlation between the variables waist and hip (in inches) with weight. We find the same positive correlation for BMI.
- Another obvious correlation we observe is that between age (in years) and marriage (in years)
- Appendix Plot 3 showing boxplots generated to investigate potential associations between continuous variables and PCOS status.

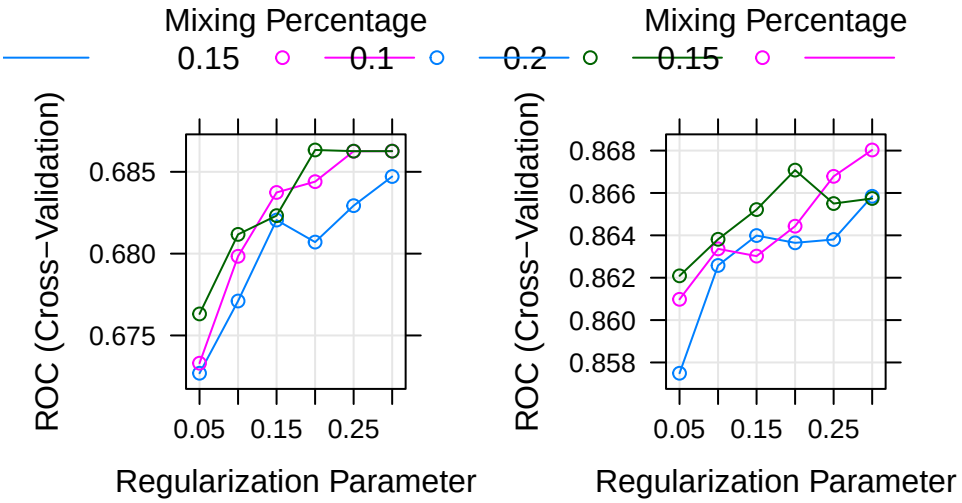
Appendix Plot 3 - Univariable distributions for continuous variables







Appendix Plot 4 - Penalized logistic regression with cross-validation (RIDGE and LASSO)



Appendix Section 5 - Model cross-validation (without RIDGE or LASSO) (using the caret package)

- We performed model cross-validation on our three logistic regression models and generated a summary of the cross-validation results.

Appendix Plot 6 - Variable importance (based on Random Forest model)