MEDI504B_ML_Project

Jacob and Hanwei

2023-02-15

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). PCOS is an endocrine (hormonal) disorder that affects females of a reproductive age. Given the widespread nature of this condition and the troubling symptoms which accompany it, including infertility, it would be helpful for physicians to be able to predict individuals more likely to experience PCOS thereby enabling them to therapeutically intervene and provide care and support in a timely manner.

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 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 (Teede et al., 2018).

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, given considerations surrounding false positive and false negatives.

Ethics

** include something about the target population and who is at risk ** what to do if someone is pregnant –> follow up with study doctor* **what to do if you have a suspicion that the patient is at risk for a life threatening medical disorder?** people taking hormone therapy ** identification of patients given the data

Labels and predictors

1.3 Data

The dataset consists of physical and clinical parameters collected from 10 hospitals across Kerala, India, 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.

The rationale for objective one was based on the motivation to provide a data-based method of diagnosis that is cheaper and less invasive than current methods. 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 advantaegous for reasons related to clinical care and resource utilization. This influenced our variable selection, as we did not consider the inclusion of predictors that can not be measured in the above stated context in model development.

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

Following our EDA (please refer to EDA section), we start by splitting our dataset into the training set and the validation set, 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.

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 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.

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.

Results

4. Results

4a. Exploratory Data Analysis

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 refer to bullet points, figure titles and figure captions 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. We find that our dataset has the following variables: Sl. No, Patient File No., PCOS (Y/N), Age (yrs), Weight (Kg), Height(Cm), BMI, Blood Group, Pulse rate(bpm), RR(breaths/min), Hb(g/dl), Cycle(R/I), Cycle length(days), Marriage Status (Yrs), Pregnant(Y/N), No. of aborptions, I beta-HCG(mIU/mL), II beta-HCG(mIU/mL), FSH(mIU/mL), LH(mIU/mL), FSH/LH, Hip(inch), Waist(inch), Waist:Hip Ratio, TSH (mIU/L), AMH(ng/mL), PRL(ng/mL), Vit D3 (ng/mL), PRG(ng/mL), RBS(mg/dl), Weight gain(Y/N), hair growth(Y/N), Skin darkening (Y/N), Hair loss(Y/N), Pimples(Y/N), Fast food (Y/N), Reg.Exercise(Y/N), BP _Systolic (mmHg), BP _Diastolic (mmHg), Follicle No. (L), Follicle No. (R), Avg. F size (L) (mm), Avg. F size (R) (mm), Endometrium (mm).

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. Here, we see a plot showing our variables and the percentage of missing rows per variable. The plot 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.

4a.2 Bivariate associations

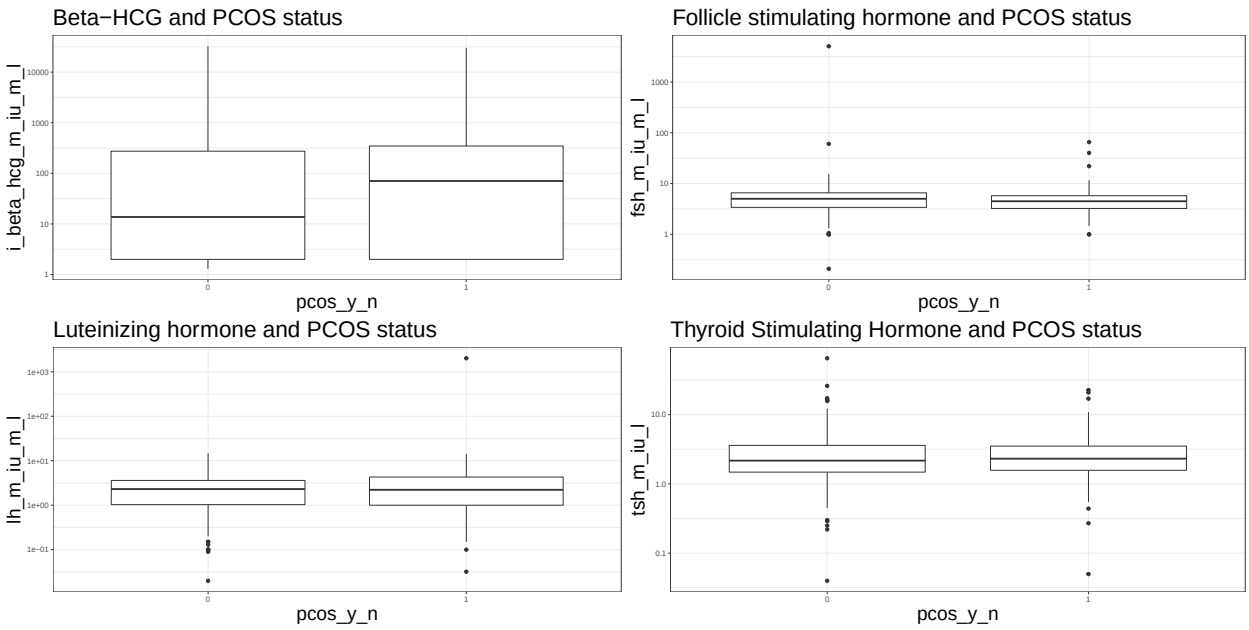
We examined associations for categorical predictors and the outcome (PCOS) by constructing bivariate plots for the predictors stratified by PCOS status[4].

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 hormones and PCOS status



Modelling

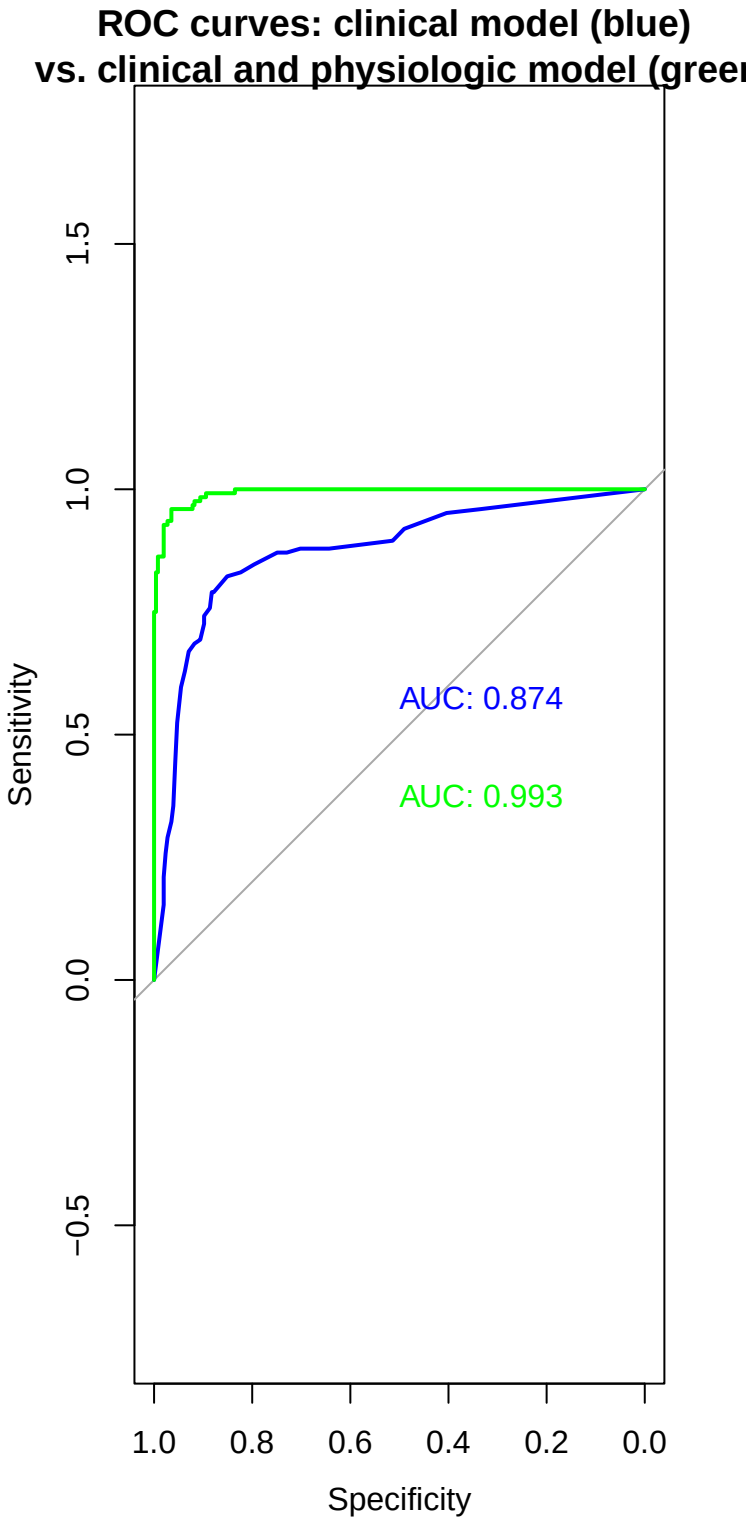
Based on these results, we will compare the clinical model to the full clinical and physiological model. While the model including clinical predictors only is more simple both technically (as evidenced by the AIC) and practically (less inputs and no blood testing required), there is an approximately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

3.3 Model cross-validation (without RIDGE or LASSO) (using the caret package)

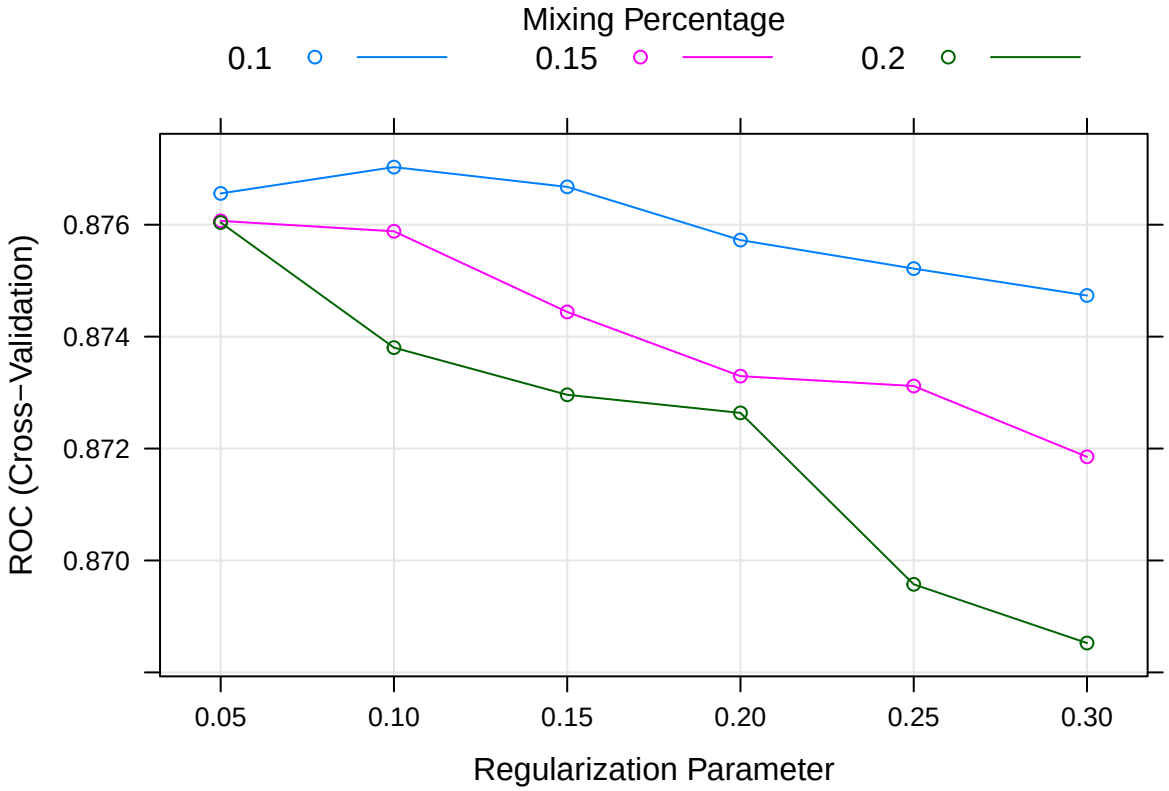
- We performed model cross-validation on our three logistic regression models and generated a summary of the cross-validation results.

3.4 Assessing model classification performance

- Assessing model performance using ROC curves


ROC curves: clinical model (blue)
vs. clinical and physiologic model (greer
Sensitivity
1.5
1.0
0.5
0.0
−0.5
AUC: 0.874
AUC: 0.993
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

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




```
# library(bestglm)

#train.data.covariates.1 <- train.data %>%
#  select(
#     pcos_y_n,
#     i_beta_hcg_m_iu_m_l,
#     fsh_m_iu_m_l,
#     lh_m_iu_m_l,
#     tsh_m_iu_l,
#     amh_ng_m_l ,
#     weight_gain_y_n ,
#     hair_growth_y_n,
#     skin_darkening_y_n,
#     hair_loss_y_n ,
#     pimples_y_n
#   ) %>%
#   mutate_at(c("pcos_y_n",
#               "weight_gain_y_n",
#               "hair_growth_y_n",
#               "skin_darkening_y_n",
#               "hair_loss_y_n",
#               "pimples_y_n"),as.factor) %>%
#   mutate_at(c(
#     "i_beta_hcg_m_iu_m_l",
```

```
#     "fsh_m_iu_m_l" ,
#     "lh_m_iu_m_l" ,
#     "tsh_m_iu_l" ,
#     "amh_ng_m_l"), as.numeric)
#
# x_tr <- train.data.covariates.1
# # x_tr$pcos_labelled <- ifelse(x_tr$pcos_labelled=="yes",1,0)
# colnames(x_tr)[1] <- "y"
# res.bestglm <-bestglm::bestglm(Xy = x_tr,
#             family = binomial,
#             IC = "AIC",                 # Information criteria for
#             method = "exhaustive")
#
# ## Show top 5 models
# res.bestglm$BestModels
#
# summary(res.bestglm$BestModel)
```

1.4 Modelling our data using Trees and Forests

Let's select only the predictor columns we decided we wanted to include earlier. From logistic model 3, here is the list of covariates: * pcos_y_n * i_beta_hcg_m_iu_m_l * fsh_m_iu_m_l * lh_m_iu_m_l * tsh_m_iu_l * amh_ng_m_l * weight_gain_y_n * hair_growth_y_n * skin_darkening_y_n * hair_loss_y_n * pimples_y_n

```
train.data.covariates.model2 <- train.data %>%
  select(
    pcos_y_n,
    weight_gain_y_n ,
    hair_growth_y_n,
    skin_darkening_y_n,
    hair_loss_y_n ,
    pimples_y_n
  ) %>% mutate(
  pcos_labelled = as.factor(ifelse(pcos_y_n==1, "yes", "no")))

train.data.covariates.model2
```

```
## # A tibble: 379 x 7
##    pcos_y_n weight_gain_y_n hair_growth_y_n skin_darkening_y_n hair_loss_y_n
##    <fct>    <fct>           <fct>           <fct>              <fct>
##  1 0        0               0               0                  0
##  2 0        0               0               0                  0
##  3 0        0               0               0                  0
##  4 0        1               0               0                  0
##  5 0        0               0               0                  0
##  6 0        0               0               0                  0
##  7 0        0               0               0                  0
##  8 1        1               1               1                  1
##  9 0        0               0               0                  1
## 10 0        1               0               0                  0
## # ... with 369 more rows, and 2 more variables: pimples_y_n <fct>,
## #   pcos_labelled <fct>
```

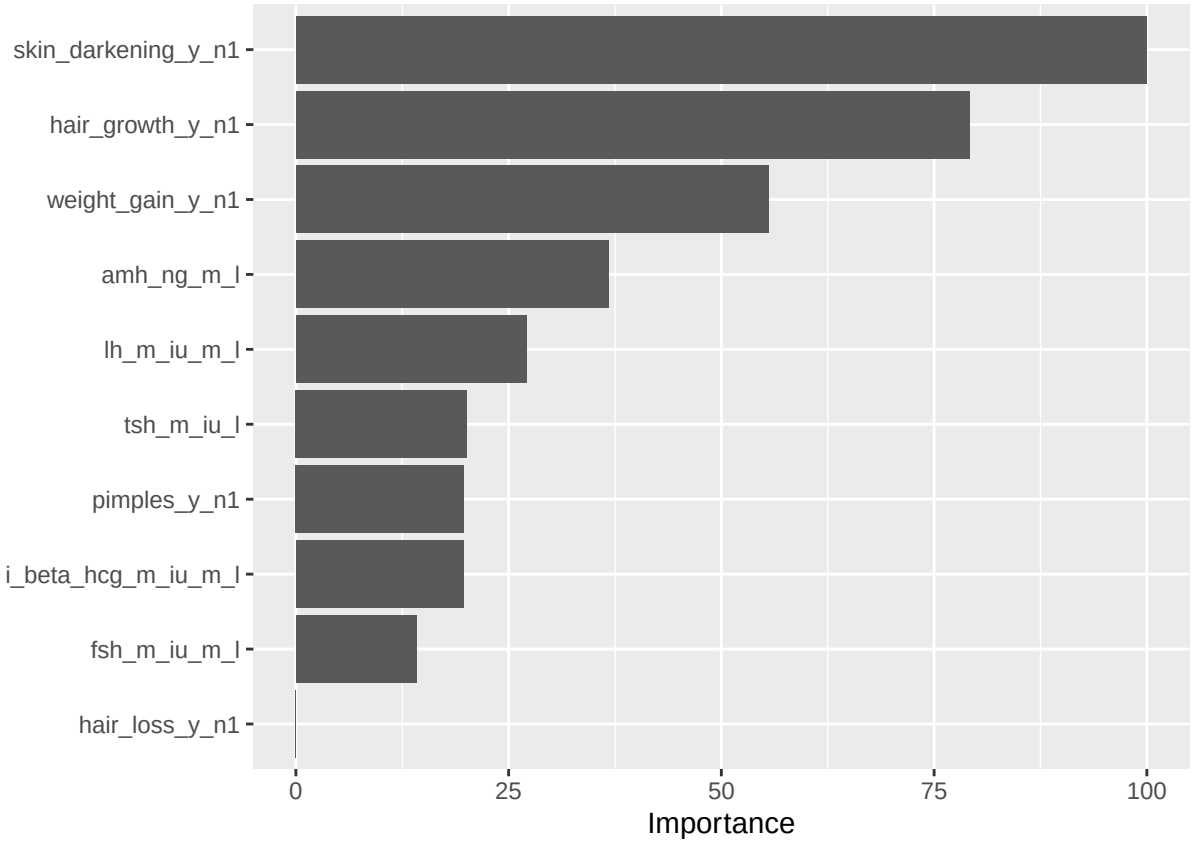
```
train.data.covariates.model3 <- train.data %>%
  select(
    pcos_y_n,
    i_beta_hcg_m_iu_m_l,
    fsh_m_iu_m_l ,
    lh_m_iu_m_l ,
    tsh_m_iu_l ,
    amh_ng_m_l ,
    weight_gain_y_n ,
    hair_growth_y_n,
    skin_darkening_y_n,
    hair_loss_y_n ,
    pimples_y_n
  ) %>% mutate(
  amh_ng_m_l = as.numeric(amh_ng_m_l),
  pcos_labelled = as.factor(ifelse(pcos_y_n==1, "yes", "no")))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

Random Forest – note that I've had to remove one missing ob from amh_ng_m_l for now.

RF Variable Importance

```
vip::vip(caret_rf_model3)
```



XGboost

Comparison code for later

5. Discussion and Conclusions

5.1 Discussion

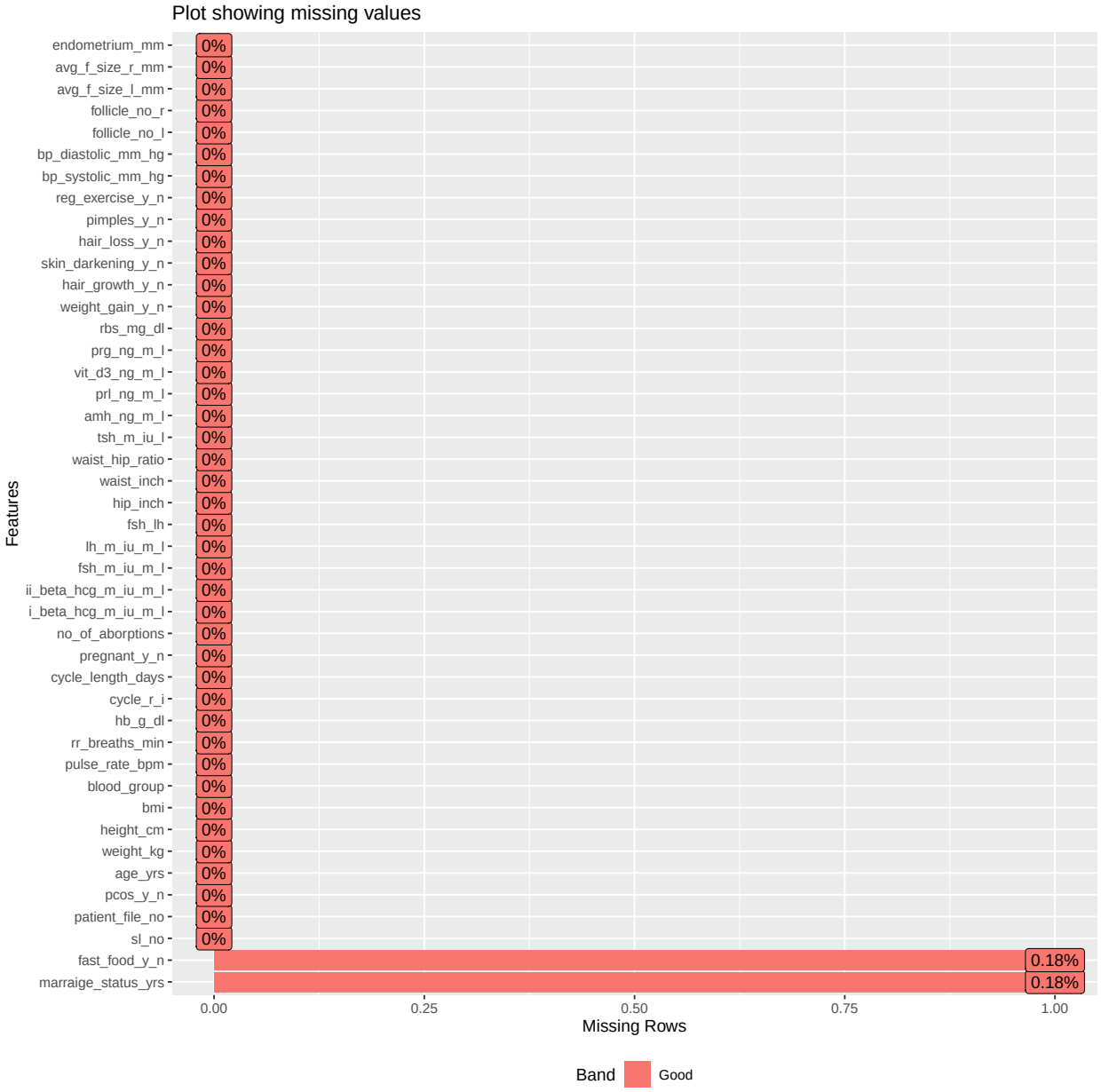
- Ethics statement

References

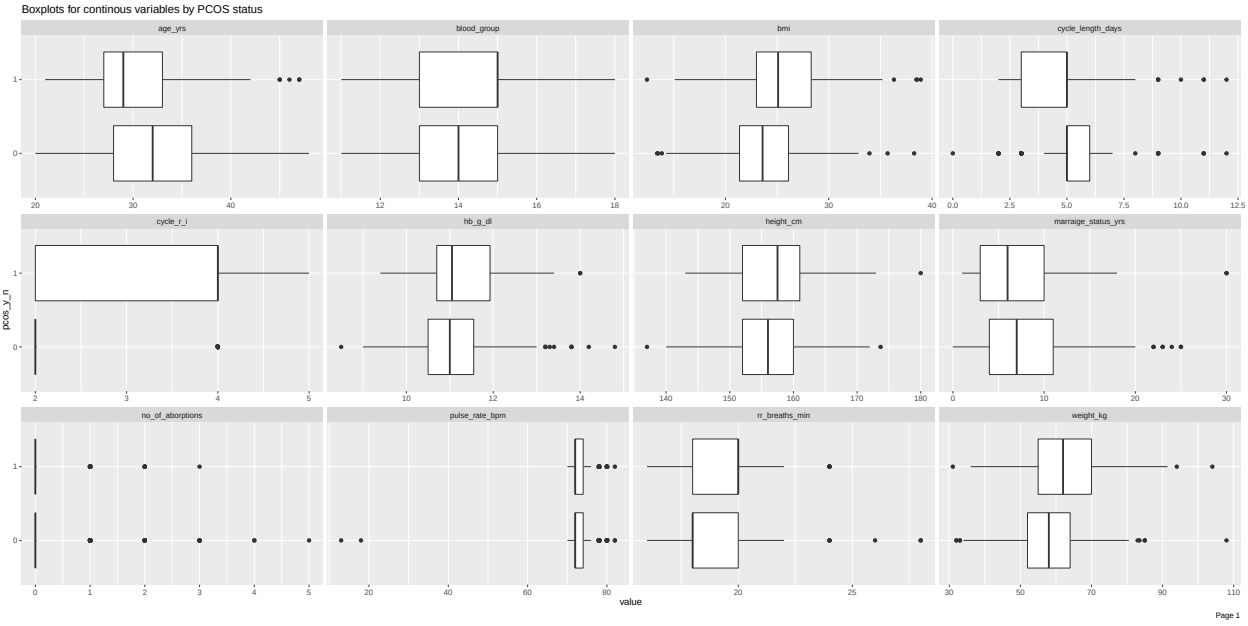
1. Teede, H., Misso, M., Tassone, E. C., Dewailly, D., Ng, E. H., Azziz, R., ... & Norman, R. J. (2018). Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Human Reproduction, 33(9), 1602-1618. https://doi.org/10.1093/humrep/dey256
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.
6.
7.
8.

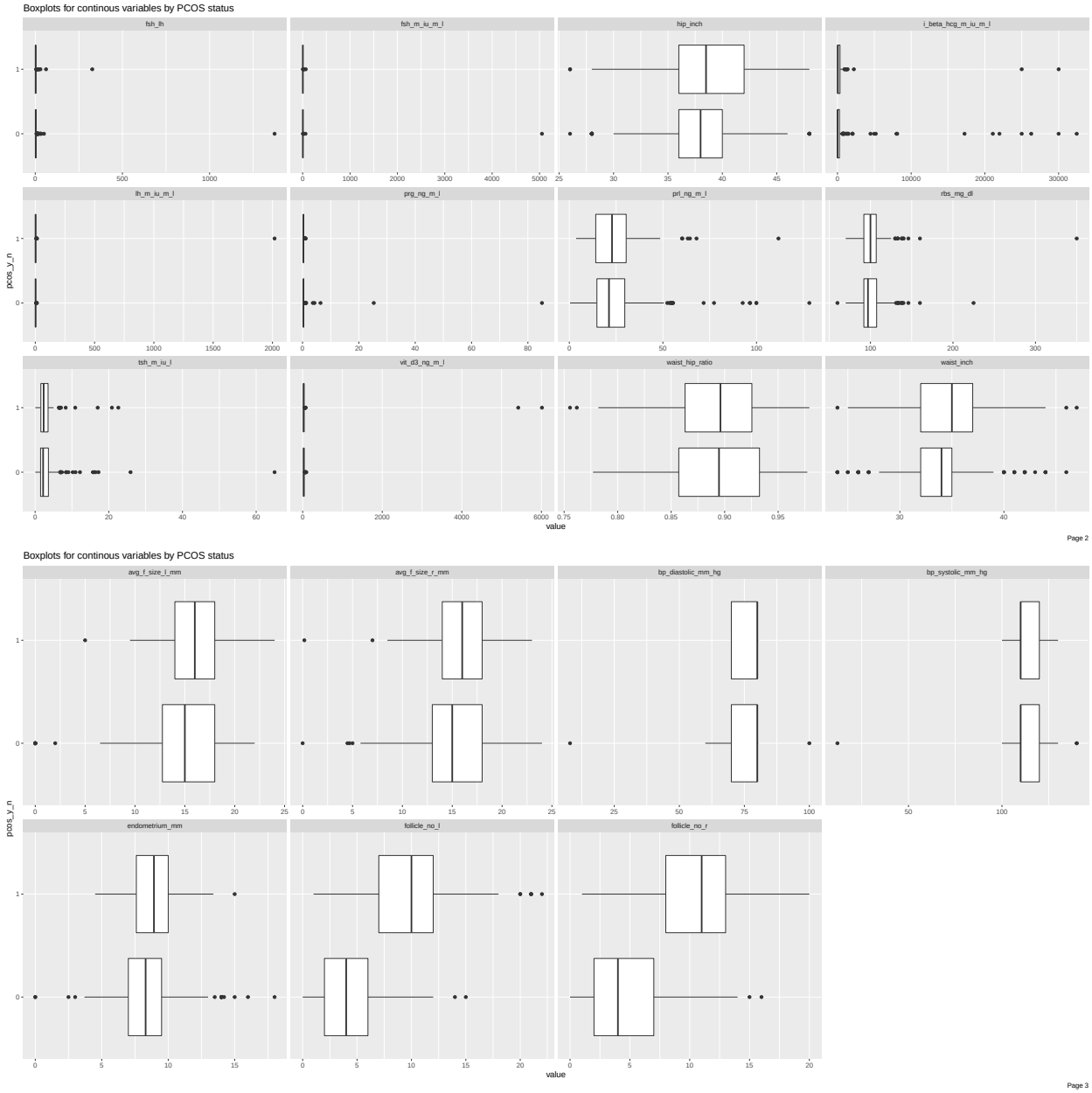
Appendix

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


Plot showing missing values
endometrium_mm
avg_f_size_r_mm
avg_f_size_l_mm
follicle_no_r
follicle_no_l
bp_diastolic_mm_hg
bp_systolic_mm_hg
reg_exercise_y_n
pimples_y_n
hair_loss_y_n
skin_darkening_y_n
hair_growth_y_n
weight_gain_y_n
rbs_mg_dl
prg_ng_m_l
vit_d3_ng_m_l
prl_ng_m_l
amh_ng_m_l
tsh_m_iu_l
waist_hip_ratio
waist_inch
hip_inch
fsh_lh
lh_m_iu_m_l
fsh_m_iu_m_l
ii_beta_hcg_m_iu_m_l
i_beta_hcg_m_iu_m_l
no_of_aborptions
pregnant_y_n
cycle_length_days
cycle_r_i
hb_g_dl
rr_breaths_min
pulse_rate_bpm
blood_group
bmi
height_cm
weight_kg
age_yrs
pcos_y_n
patient_file_no
sl_no
fast_food_y_n
marraige_status_yrs
0%
0.18%
Features
0.00
0.25
0.50
0.75
1.00
Missing Rows
Band
Good

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





4a.2 Univariable distributions for continuous variables

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).