PCOS Data Mining Project

ALY 6015

Janica St Ville and Jeff Hackmeister

2025-05-05

# Introduction

For this project, we will be working with a dataset from Kaggle, the “**Polycystic Ovary Syndrome PCOS”.** This dataset contains health data for 3,000 women and is generated according to the Rotterdam criteria. It sparked our interest because it is a prevalent hormonal disorder that can significantly impact women’s health and quality of life. According to the World Health Organization PCOS represents the leading cause of infertility worldwide and nearly 70% of women go undiagnosed. By better understanding the common characteristics of PCOS patients, medical professionals will be better able to target testing efforts and provide treatment. There is no current cure for PCOS, but there are a variety of treatments available to improve symptoms and increase quality of life.

The goal is to explore how various physiological indicators — such as BMI, hormone levels, and menstrual regularity — relate to the presence of PCOS. We hope to identify patterns that may help in early detection or risk assessment. We will be implementing a Support Vector Machine to predict a PCOS diagnosis from the available variables within the dataset.

## Data Preparation and Exploration

To begin, we will load the data from Kaggle and our needed packages. Once, the data is read in, we will briefly explore the dataset

library(e1071)  
library(arules)  
library(caret)  
library(tidyverse)  
library(patchwork)  
library(corrplot)  
library(GGally)  
data <- read.csv("pcos\_rotterdam\_balanceado.csv")  
  
str(data)

'data.frame': 3000 obs. of 6 variables:  
 $ Age : int 29 20 23 19 19 23 21 26 25 31 ...  
 $ BMI : num 21.2 20.5 23.1 32.7 25.9 20.6 25.3 25.2 25.7 29 ...  
 $ Menstrual\_Irregularity : int 0 0 0 1 0 0 0 0 1 0 ...  
 $ Testosterone\_Level.ng.dL.: num 46.1 59.4 69.3 77.7 49.4 36.7 44.2 32.6 96.8 58.3 ...  
 $ Antral\_Follicle\_Count : int 9 6 10 37 5 5 4 4 37 6 ...  
 $ PCOS\_Diagnosis : int 0 0 0 1 0 0 0 0 1 0 ...

The data set contains 3,000 records, 6 variables which are variables of interest including:

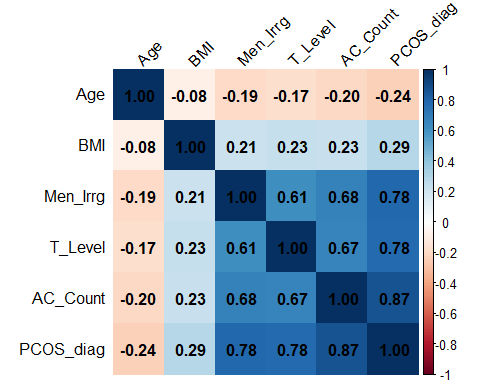
* Age – Age of the individual
* BMI – Body Mass Index
* Menstrual\_Irregularity – Indicator of menstrual Irregularities
* Testosterone\_Level – measured Testosterone level
* Antral\_Follicle\_count – Count of antral follicles
* PCOS\_Diagnosis- PCOS diagnosis (likely 0- NO. 1= yes)

This dataset does not contain missing variables, so our data cleaning is minimal. We will shorten the variable names for easy of use in the analysis.

data <- data %>%  
 rename(T\_Level = Testosterone\_Level.ng.dL.,  
 Men\_Irrg = Menstrual\_Irregularity,  
 AC\_Count = Antral\_Follicle\_Count,  
 PCOS\_diag = PCOS\_Diagnosis)

To understand the interaction of the variables, we will construct a correlation matrix

cormat <- cor(data)  
corrplot(cormat,   
 method = "color",  
 addCoef.col = "black",  
 tl.col = "black",  
 tl.srt = 45,  
 diag = TRUE)



From this, we can see a very strong correlation between AC count and PCOS, menstrual irregularities and testosterone levels are also strong correlations, while age and BMI show far weaker relationships.

To make futher analysis easier, we will convert PCOS diagnosis and menstrual irregularities to factor variables with yes/no labels.

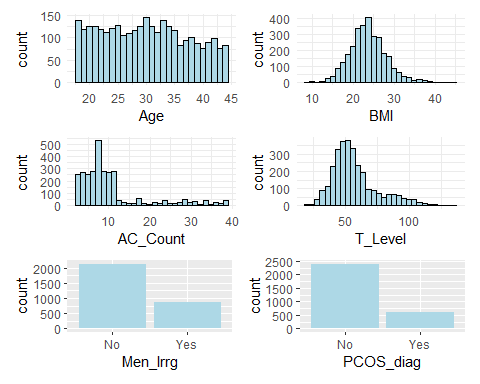
data$PCOS\_diag <- factor(data$PCOS\_diag, levels = c(0, 1), labels = c("No", "Yes"))  
data$Men\_Irrg <- factor(data$Men\_Irrg, levels = c(0, 1), labels = c("No", "Yes"))  
head(data)

Age BMI Men\_Irrg T\_Level AC\_Count PCOS\_diag  
1 29 21.2 No 46.1 9 No  
2 20 20.5 No 59.4 6 No  
3 23 23.1 No 69.3 10 No  
4 19 32.7 Yes 77.7 37 Yes  
5 19 25.9 No 49.4 5 No  
6 23 20.6 No 36.7 5 No

## Variable Distribution

Next, we take a look at the distribution of all the variables in the dataset.

unique\_ages <- seq(min(data$Age) - 0.5, max(data$Age) + 0.5, by = 1)  
age\_plot <- ggplot(data, aes(x=Age)) +   
 geom\_histogram(breaks = unique\_ages,  
 fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
bmi\_plot <- ggplot(data, aes(x = BMI)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
mi\_plot <- ggplot(data, aes(x=Men\_Irrg)) +  
 geom\_bar(fill = "lightblue")  
  
t\_plot <- ggplot(data, aes(x = T\_Level)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
ac\_plot <- ggplot(data, aes(x = AC\_Count)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
pcos\_plot <- ggplot(data, aes(x=PCOS\_diag)) +  
 geom\_bar(fill = "lightblue")  
  
combined\_plot <- (age\_plot + bmi\_plot) / (ac\_plot + t\_plot) / (mi\_plot + pcos\_plot) +   
 plot\_layout(guides = "collect")  
  
print(combined\_plot)

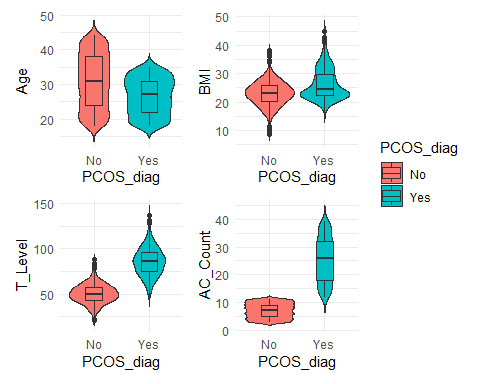


To better understand the relationship between the variables and a PCOS diagnosis, we split the dataset into yes and no groups.

y\_diag <- data %>%  
 filter(data$PCOS\_diag == "Yes")  
n\_diag <- data %>%  
 filter(data$PCOS\_diag == "No")

Then construct a series of plots to see how they interact with the detection of PCOS.

bmi\_pcos <- ggplot(data, aes(x = PCOS\_diag, y = BMI, fill= PCOS\_diag)) +  
 geom\_violin(trim=FALSE) +  
 geom\_boxplot(width = 0.3) +  
 theme\_minimal()  
  
ac\_pcos <- ggplot(data, aes(x = PCOS\_diag, y = AC\_Count, fill= PCOS\_diag)) +  
 geom\_violin(trim=FALSE) +  
 geom\_boxplot(width = 0.3) +  
 theme\_minimal()  
  
age\_pcos <- ggplot(data, aes(x = PCOS\_diag, y = Age, fill= PCOS\_diag)) +  
 geom\_violin(trim=FALSE) +  
 geom\_boxplot(width = 0.3) +  
 theme\_minimal()  
  
t\_pcos <- ggplot(data, aes(x = PCOS\_diag, y = T\_Level, fill= PCOS\_diag)) +  
 geom\_violin(trim=FALSE) +  
 geom\_boxplot(width = 0.3) +  
 theme\_minimal()  
  
(age\_pcos + bmi\_pcos) / (t\_pcos + ac\_pcos) +   
 plot\_layout(guides = "collect")



This view further illustrates the importance of AC count and testosterone levels in PCOS detection and diagnosis.

# Support Vector Machine

There are several viable classification methodologies for this type of study and many were considered for the project. Given the current dataset, a decision tree model could be appropriate. However, with the hopes of expanding the study to additional variables and more observations, a support vector machine approach gives more flexibility as complexity increases.

First, we fit the model using the e1701 library and review our results.

model <- svm(PCOS\_diag ~ ., data = data)  
summary(model)

Call:  
svm(formula = PCOS\_diag ~ ., data = data)  
  
  
Parameters:  
 SVM-Type: C-classification   
 SVM-Kernel: radial   
 cost: 1   
  
Number of Support Vectors: 64  
  
 ( 32 32 )  
  
  
Number of Classes: 2   
  
Levels:   
 No Yes

Now we use the model to make predictions on our dataset.

predictions <- predict(model, data)  
head(predictions)

1 2 3 4 5 6   
 No No No Yes No No   
Levels: No Yes

Now that we have predictions, we can take a look at the confusion matrix, which compares the actual PCOS diagnoses to the predictions from our SVM.

table(Actual = data$PCOS\_diag, Predicted = predictions)

Predicted  
Actual No Yes  
 No 2399 1  
 Yes 2 598

This shows very strong predictive strength for the model with only 1 false positive and 2 false negatives. To further examine the strength of the model, we partition the dataset into training and test sets.

set.seed(123)   
train\_index <- createDataPartition(data$PCOS\_diag, p = 0.7, list = FALSE)  
train\_data <- data[train\_index, ]  
test\_data <- data[-train\_index, ]

With these new datasets, we refit the model and test our results.

model\_train <- svm(PCOS\_diag ~ ., data = train\_data)  
test\_pred <- predict(model\_train, test\_data)  
confusionMatrix(test\_pred, test\_data$PCOS\_diag)

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 719 2  
 Yes 1 178  
   
 Accuracy : 0.9967   
 95% CI : (0.9903, 0.9993)  
 No Information Rate : 0.8   
 P-Value [Acc > NIR] : <2e-16   
   
 Kappa : 0.9896   
   
 Mcnemar's Test P-Value : 1   
   
 Sensitivity : 0.9986   
 Specificity : 0.9889   
 Pos Pred Value : 0.9972   
 Neg Pred Value : 0.9944   
 Prevalence : 0.8000   
 Detection Rate : 0.7989   
 Detection Prevalence : 0.8011   
 Balanced Accuracy : 0.9938   
   
 'Positive' Class : No

Once again, we see very strong results for the model, with 99.67% accuracy and a 95% confidence interval of 99.03 to 99.93%. There were also only 2 false negatives in this test - which is very important for this particular subject matter.

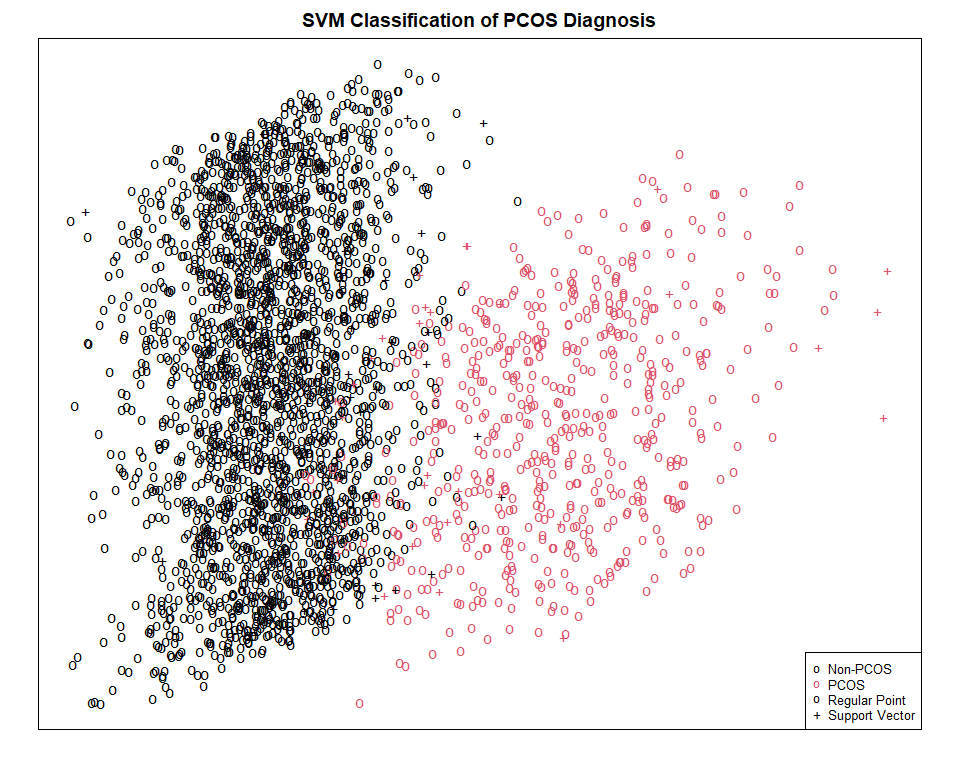
## Visualizations

Next, we will use the dist function from arules to create a distance matrix. This measures the distance between the data points in a multidimensional space. To do so, we must first limit the data to only numerical values. Once we have those distances calculated, we use the cmdscale function to convert those values into 2D coordinates that we can plot.

# Extract only numeric columns for distance calculation  
numeric\_features <- data[, sapply(data, is.numeric)]  
# Create distance matrix  
dist\_matrix <- dist(numeric\_features)  
# Apply MDS to get 2D coordinates  
mds\_coords <- cmdscale(dist\_matrix)

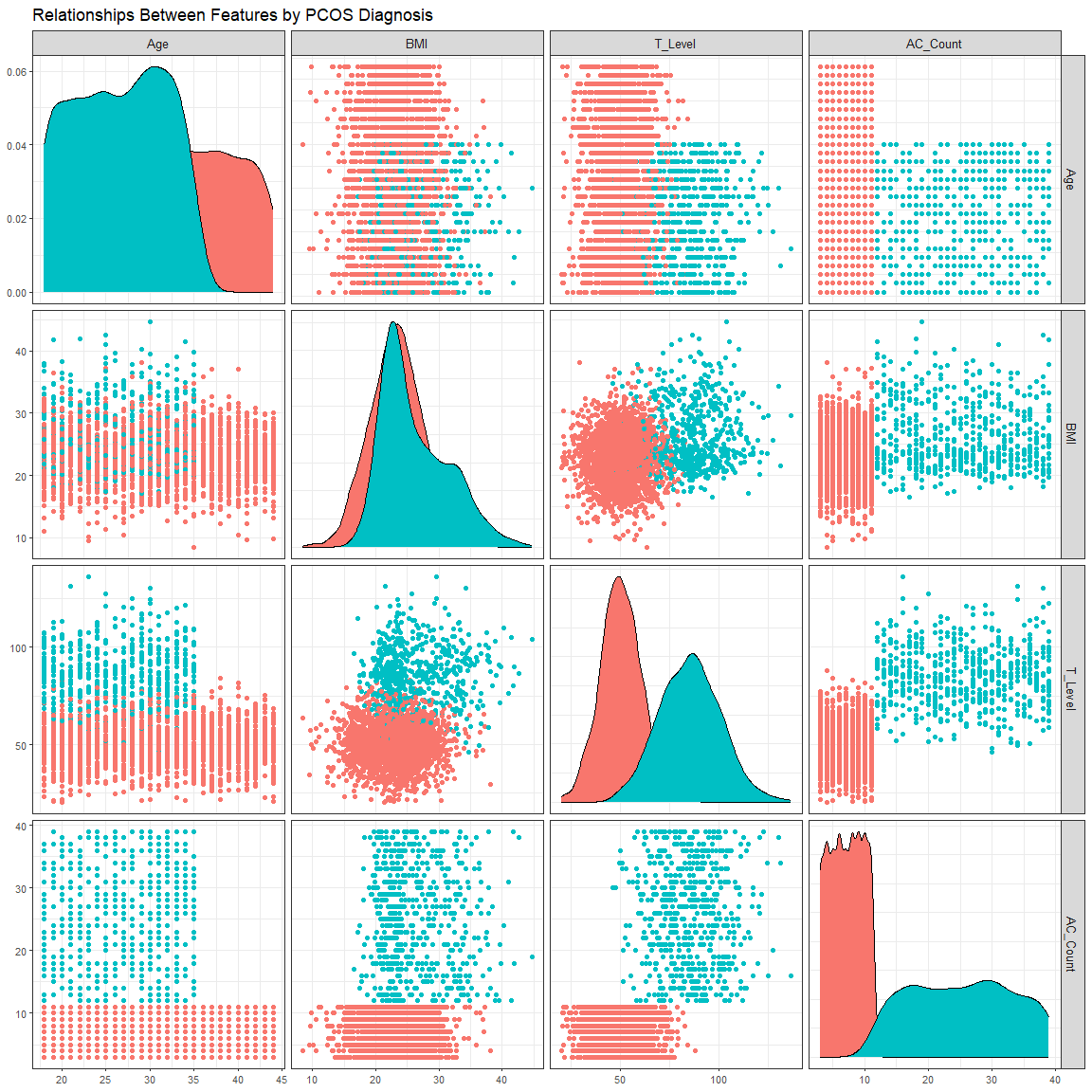
And then we will visualize the results

par(mar = c(2, 2, 2, 2), xaxt = 'n', yaxt = 'n')  
plot(mds\_coords,  
 col = as.integer(data$PCOS\_diag),  
 pch = c("o", "+")[1:nrow(data) %in% model$index + 1],  
 main = "SVM Classification of PCOS Diagnosis",  
 xlab = "", ylab = "") # This removes the axis labels  
legend("bottomright",  
 legend = c("Non-PCOS", "PCOS", "Regular Point", "Support Vector"),  
 col = c(1, 2, 1, 1),  
 pch = c("o", "o", "o", "+"),  
 cex = 0.8)



The resulting plot shows the distribution of the observations from the dataset. Positive diagnoses are in red, while negatives are in black. The 64 support vectors identified in the model are indicated by a “+” while the other observations are shown with a “o”. Despite the strong results we saw in the model above, there is not as clean of a delineation in the data from this plot. To look for a cleaner separation, we will look at each of the variable pairs independently.

plot\_data <- cbind(numeric\_features, PCOS\_diag = data$PCOS\_diag)  
ggpairs(plot\_data,   
 columns = 1:(ncol(plot\_data)-1),  
 mapping = ggplot2::aes(color = PCOS\_diag),  
 upper = list(continuous = "points"),  
 lower = list(continuous = "points"),  
 diag = list(continuous = "densityDiag"),  
 progress = FALSE) +  
 theme\_bw() +  
 theme(  
 axis.text = element\_text(size = 8),  
 strip.text = element\_text(size = 10),  
 legend.position = "bottom"  
 ) +  
 labs(title = "Relationships Between Features by PCOS Diagnosis")



This gives us a much better view into which variable combination provides the best distinction in diagnosis outcomes. Age has the weakest predictive influence, while BMI, testosterone and AC count produce much cleaner results.

# Concussion

The results of our SVM model were very strong, indicating that these variables are highly useful in predicting a PCOS diagnosis. Age appears to be of limited use in the model, which makes some logical sense given the relatively narrow age band included in a study of reproductive adult women, 18 to 44 in this dataset. Ideally, we could use this model on a new dataset collected from a different geographical area to test the effectiveness on different demographic samples. It would also be quite helpful to add additional demographic variables such as race, socioeconomic status, and access to regular healthcare as they tend to correlate strongly with many other health outcomes.

The strong results seen from the testing does lead to a concern around over fitting for the model. Adding additional observations and variables to the model would allow for better refinement of the model and the ability to hone in on the most consequential variables at play in the diagnosis and eventual treatment of PCOS in women around the world.

# References

[1] *arules Package | R Documentation*. (2012). Rdocumentation.org.

https://www.rdocumentation.org/packages/arules/versions/1.6-4

[2] Christ, J. P., & Cedars, M. I. (2023). Current Guidelines for Diagnosing PCOS. *Diagnostics*, *13*(6), 1113. https://doi.org/10.3390/diagnostics13061113

[3] *Classifying Data Using Support Vector Machines(SVMs) in R*. (2018, August 28). GeeksforGeeks.

https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/

[4] *cmdscale Function - RDocumentation*. (n.d.). Www.rdocumentation.org.

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cmdscale

[5] *e1071 Package - RDocumentation*. (2024). Rdocumentation.org.

https://www.rdocumentation.org/packages/e1071/versions/1.7-16

[6] Soares, L. S. (2025). *Polycystic Ovary Syndrome PCOS*. Kaggle.com. https://www.kaggle.com/datasets/lucass0s0/polycystic-ovary-syndrome-pcos/data

[7] Tierney, N. (2025, April 9). *Quarto for Scientists*. Njtierney.com. https://qmd4sci.njtierney.com/

[8] World Health Organization. (2025, February 7). *Polycystic ovary syndrome*. World Health Organization; World Health Organization.

https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome