Polycystic Ovary Syndrome (PCOS) Diagnosis Prediction

CSC 532 Machine Learning: Final Project

Rachel Ratajczak

April 21, 2025

1 Abstract

Polycystic ovary syndrome (PCOS) is a chronic hormonal disorder affecting many women of reproductive age, often remaining undiagnosed and leading to further complications. This project aims to diagnose PCOS using machine learning models trained on common risk factors. Several classification models were evaluated, including Lasso, Ridge, and Elastic Net logistic regression, random forest, gradient-boosted trees, support vector machines (SVM), and a neural network. The dataset, obtained from Kaggle, was structured for binary classification. Among the models tested, the radial basis function SVM achieved the best performance, with 97% accuracy and a Kappa value of 0.94, indicating strong predictive capability.

2 Problem Definition and Project Goals

Polycystic ovarian syndrome is a significant health issue that disrupts hormones in women of reproductive age. Women with PCOS can experience irregular periods, excess androgen levels, and cysts on their ovaries. Irregular periods are associated with a lack of ovulation, which makes it difficult for women struggling with PCOS to become pregnant. The exact cause of PCOS is unknown and cannot be cured. Approximately 6-13% of women are affected by this condition, but unfortunately, about 70% of women have undiagnosed cases of PCOS.[2] The goal of this project is to bring awareness to this chronic condition and understand what features can contribute to a positive diagnosis of PCOS.

The original PCOS diagnosis dataset was obtained from the Kaggle public database and contains 1000 entries representing patient data. The dataset includes five key features typically associated with the diagnosis of PCOS.

The key attributes are:

1. Age (years): This refers to the patient's age. The reproductive age range is 18 to 45.
2. BMI (kg/m2): This is the Body Mass Index, a measure of body fat based on the patient's height and weight. The BMI range is between 18 and 35.
3. Menstrual Irregularity (binary): This is a binary indicator of whether the patient has irregular menstrual cycles (0 = No, 1 = Yes).
4. Testosterone Level (ng/dL): This is the level of testosterone in the patient's blood. It is used as a hormonal indicator for PCOS diagnosis. The range is 20 – 100 ng/dL.
5. Antral Follicle Count: This is the number of antral follicles detected from an ultrasound. The antral follicle count is used to observe PCOS diagnosis. The range is from 5 – 30 antral follicles.

The target variable is:

1. PCOS Diagnosis (binary): This is a binary indicator of whether the patient has been diagnosed with Polycystic Ovary Syndrome (0 = No, 1 = Yes).

This project aims to use the target variables commonly associated with PCOS and train various machine learning models to predict the target variable (PCOS diagnosis). It also aims to find which model performs best at predicting a PCOS diagnosis.

3 Related Work

Since this dataset is publicly available on Kaggle, numerous projects have addressed the problem of predicting PCOS diagnosis. These projects vary in terms of the number and type of risk factors used, as well as the machine learning methods applied. Most commonly, logistic regression models are employed, and several projects also explore neural networks. While Python is the dominant programming language in these implementations, the variety in model selection remains somewhat limited. In contrast, this project incorporated a broader range of models, including support vector machines (SVM), random forest, and gradient boosted trees. For example, in one Kaggle notebook by Kim Kijun, the random forest model achieved the highest accuracy for PCOS prediction.[3] However, in this project, the SVM with a radial basis function kernel outperformed all other models, achieving 97% accuracy and a Kappa value of 0.94.

4 Data Exploration and Preprocessing

The first step of this project was to explore the feature attributes of the PCOS dataset and see if there are any relationships among the data we present. All of the feature attributes are numeric, two of which are continuous, two are discrete, and one is binary. Below is a summary of each of the feature attributes and the statistical tests performed on them to visualize their distribution and find relationships.

4.1 Age Attribute

Age is a continuous numeric variable since it increases continuously over time. The age attribute is measured in years and ranges from 18 to 45, representing this patient sample's reproductive ages. A histogram was used to analyze the distribution of PCOS data over the age intervals. Figure 1. shows the distribution of positive PCOS counts over the age intervals. The distribution appears to be multimodal, with some ages having distinct peaks of positive PCOS counts. These distinct peaks are around 20 years and younger, 30-35 years, and 43 years and older. Figure 2. shows the boxplot of age and PCOS diagnosis. The mean ages for positive and negative PCOS diagnoses are relatively close. The mean age of a positive PCOS diagnosis is around 30 years.

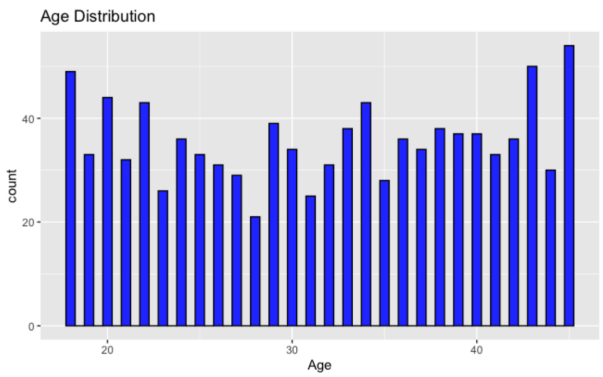

Figure 1. Histogram of Age Distribution

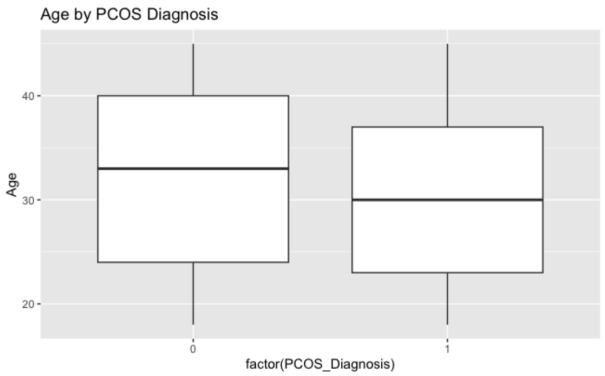

Figure 2. Boxplot of Age and PCOS Diagnosis

The histogram and boxplot suggest no significance between the attribute age and the target variable PCOS diagnosis. The multi-modal histogram suggests a range of age values are associated with a positive diagnosis. The boxplot means are very close; both are around 30 years of age for both positive and negative diagnoses. A correlation coefficient analysis was performed on age vs. PCOS diagnosis to confirm no significance between these two variables. The coefficient was -0.0646, which is a weak negative correlation. Since this coefficient is near zero, the tests suggest no significance between age and PCOS diagnosis.

4.2 BMI Attribute

BMI is a continuous variable since it can fluctuate continuously over time. It is the body mass index of the patient, which is measured based on the patient's height and weight. BMI is broken into categories to determine whether or not the patient is at a healthy weight.

BMI categories for adults

- Underweight. Less than 18.5.
- Healthy Weight. 18.5 to less than 25.
- Overweight. 25 to less than 30.
- Obesity. 30 or greater.

A histogram was used to analyze the distribution of PCOS data over the BMI intervals. Figure 3 shows the distribution of PCOS counts over the BMI intervals. The distribution appears to be multimodal, with underweight and overweight BMI values having higher PCOS counts. Figure 4 shows the boxplot for BMI and PCOS diagnosis. The mean BMI for a positive PCOS diagnosis is about 31, which is considered obese. The mean BMI for a negative PCOS diagnosis is about 24, which is considered a healthy weight.

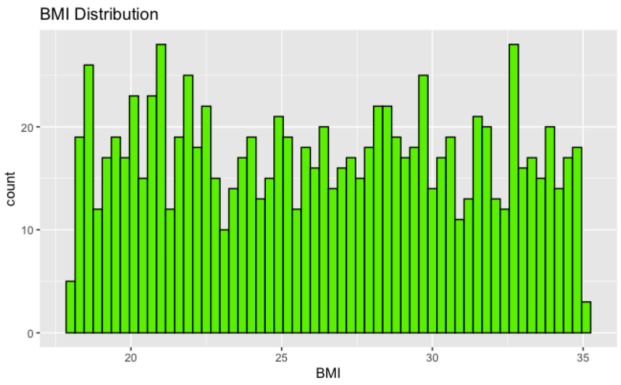

Figure 3. Histogram of BMI Distribution

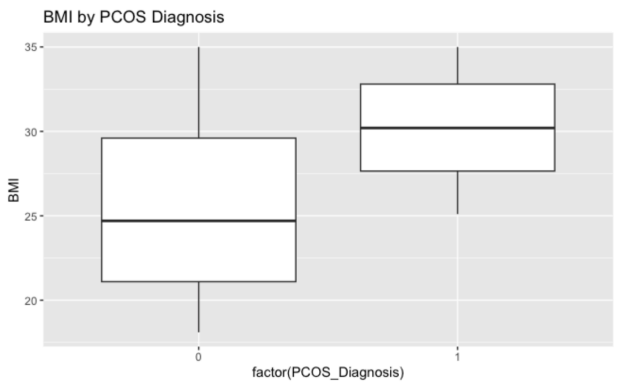

Figure 4. Boxplot of BMI and PCOS Diagnosis

The histogram does not show a major significance between BMI and PCOS diagnosis. The peaks of the histogram are on either the low end or high end of the graph, which suggests that an unhealthy BMI has higher PCOS counts. The boxplot shows the significance between a higher BMI and a positive PCOS diagnosis. However, there is much variability of BMI in the negative PCOS diagnosis, as indicated in the larger box. A correlation coefficient was performed on the BMI and PCOS diagnosis data to check further

for significance. The correlation coefficient was 0.3778, a moderate positive correlation between BMI and PCOS diagnosis. The box plot and the correlation coefficient suggest moderate significance between BMI and PCOS diagnosis.

4.3 Testosterone Level

Testosterone level is a discrete numeric variable representing the amount of testosterone in the patient's blood. This level is used as a hormonal indicator for PCOS diagnosis. The level range is from 20 – 100 ng/dL. Figure 5. shows the histogram for testosterone levels. The distribution appears to be multimodal. The highest levels and lowest levels show more PCOS diagnosis counts. Figure 6. shows the boxplot for testosterone level and PCOS diagnosis. The mean values are not significantly different between positive and negative PCOS diagnosis. A positive PCOS diagnosis has a higher mean, and a negative PCOS diagnosis has more variability.

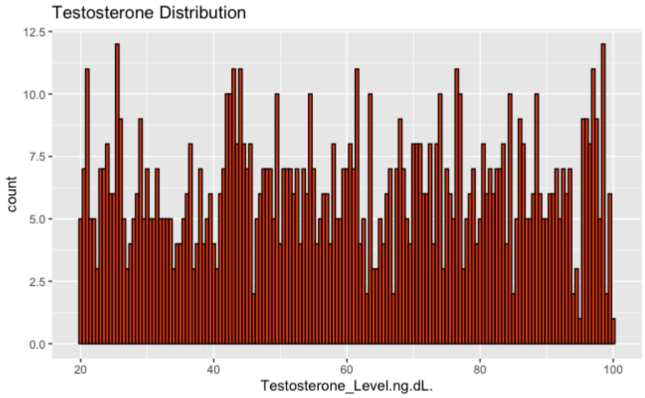

Figure 5. Histogram of Testosterone Distribution

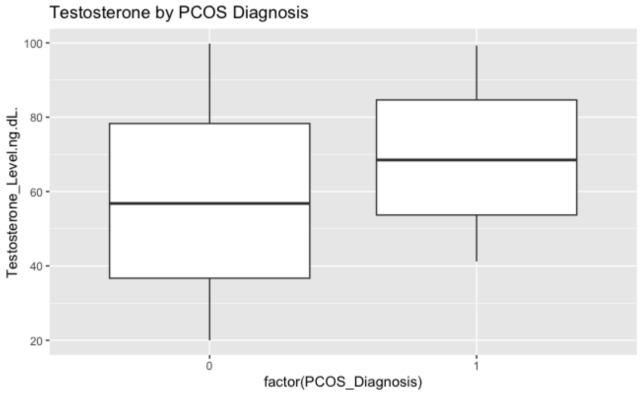

Figure 6. Boxplot of Testosterone and PCOS Diagnosis

The histogram does not suggest any significance between testosterone and PCOS diagnosis. The peak testosterone levels are at the low end and the high end of the scale. The box plot shows the slight significance that an increase in testosterone level is associated with a positive PCOS diagnosis. However, both of these tests do not clearly visualize significance. A Welch two-sample t-test was performed to further analyze the significance. The p-value was 1.693e-13, which indicates that there is not a significant correlation between testosterone and PCOS diagnosis.

4.4 Antral Follicle Count

Antral follicles are found in the ovaries, and their count is an indicator of a woman's ovary reserve. They are critical for ovulation as they have the potential to mature into an egg. It is important to note that antral follicle counts decrease with age.

General guidelines for normal AFC by age:

- 20-24 years: 15-30 follicles
- 25-34 years: 13-25 follicles
- 35-40 years: 10-15 follicles
- 41-45 years: 3-10 follicles
- 46+ years: 0-3 follicles

Figure 7. shows the distribution of antral follicle count distribution. Antral follicle counts of 12 and 21 have the highest PCOS diagnosis counts. The histogram is multimodal, with no apparent skewness in the distribution. Figure 8. shows the box plot between antral follicle count and PCOS diagnosis. The mean antral follicle count is slightly higher in the positive PCOS diagnosis. The negative PCOS diagnosis has more variability since the box is larger.

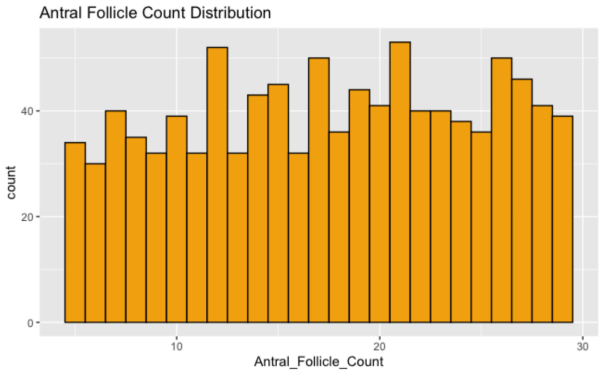

Figure 7. Histogram of Antral Follicle Count

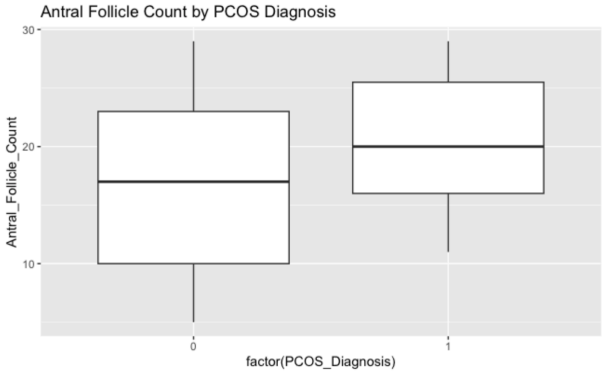

Figure 8. Boxplot of Antral Follicle Count and PCOS Diagnosis

These two tests do not show any significance between antral follicle count and PCOS diagnosis. A Welch two-sample t-test was used to confirm this further. The p-value is 1.447e-12, which shows no significance between antral follicle count and PCOS diagnosis.

4.5 Menstrual Irregularity

Menstrual irregularity is a binary indicator of whether the patient has irregular menstrual cycles (0 = No, 1 = Yes). This variable is an integer that represents a categorical value. Figure 9. shows a bar plot of menstrual irregularity for positive and negative PCOS diagnosis. Positive menstrual irregularity is associated with higher levels of PCOS diagnosis.

To confirm this observation, a Pearson's Chi-squared test was performed. The p-value is < 2.2 e-16, which is extremely small. This confirms that there is a significance between menstrual irregularity and PCOS diagnosis. Patients with menstrual irregularity are more likely to have a positive PCOS diagnosis.

The mean value of menstrual irregularity is 0.6638, which suggests that approximately 66.38% of the individuals in the dataset have irregular menstrual cycles.

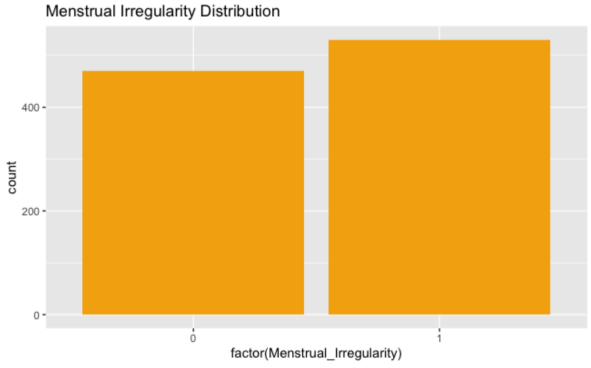

Figure 9. Bar plot of Menstrual Irregularity and PCOS Diagnosis

4.6 Further Exploring the Relationships among the Variables

A scatter plot was performed on testosterone levels and antral follicle counts to analyze their association with PCOS diagnosis. Figure 10. shows that when antral follicle count and testosterone levels are both

high, there are more positive PCOS diagnosis. This joint distribution shows a possible relationship with a positive PCOS diagnosis when the combination of these attributes is high.

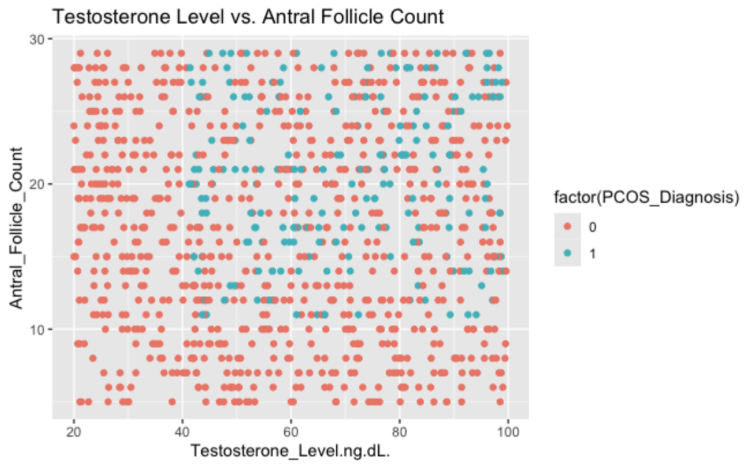

Figure 10. Scatterplot of Testosterone vs. Antral Follicle Count and PCOS Diagnosis

4.7 PCOS Diagnosis - The Target Variable

PCOS Diagnosis is a binary variable (0 = No, 1 = Yes) representing a categorical value of either a negative or positive PCOS Diagnosis.

Figure 11. shows the distribution of PCOS diagnosis in the dataset. Majority of the data set does not have a PCOS diagnosis. Further processing will need to be performed to ensure the data is balanced. To understand the attributes and how they relate to the target variable PCOS diagnosis, a summary statistic was performed on the numeric attributes. Table 1. Shows the mean values for Age, BMI, Testosterone, and Antral Follicle Count. Positive PCOS diagnosis is associated with higher mean values in BMI, Testosterone, and Antral follicle count.

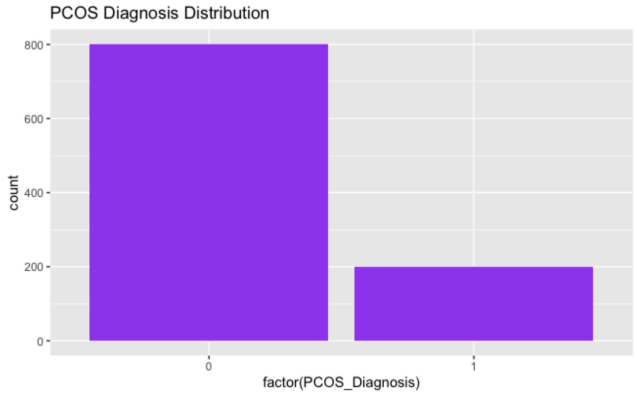

Figure 11. Bar plot of PCOS Diagnosis

PCOS Diagnosis	Mean Age	Mean BMI	Mean Testosterone	Mean Antral Follicle Count
0	32.04	25.45	57.84	16.79
1	30.67	30.12	69.48	20.19

Table 1. Summary Statistic of Numeric Variables by PCOS Diagnosis

4.7 Does the PCOS Diagnosis dataset have missing variables?

Missing data can create bias in machine learning models, affecting their ability to make accurate predictions. If missing values are present, it is crucial to check and fix them. The PCOS diagnosis dataset does not contain missing data, so no imputation methods will be needed.

4.8 Is PCOS Diagnosis Dataset Balanced?

Figure 11 shows that the majority of the patients in the PCOS data set have a negative diagnosis. An imbalance ratio was calculated on the PCOS dataset to confirm this further. Out of the 1000 observations, 801 obs. have a negative PCOS diagnosis, and 199 obs. have a positive PCOS diagnosis. The ratio calculation confirms that the dataset is imbalanced. The imbalance ratio is about 80:20, which is significant enough for an imbalance technique to be used.

Since the minority class has to do with a positive PCOS Diagnosis, it is important to have enough positive samples for the different machine learning models to make accurate predictions. To solve this problem, the Synthetic Minority Over-sampling Technique (SMOTE) technique was used to double the amount of PCOS-positive samples. After the SMOTE calculation, the balanced dataset is now 801 obs. for negative PCOS diagnosis and 597 obs. for positive PCOS diagnosis. Now, the PCOS data is ready for further processing.

4.9 Scaling the Numeric Data

Scaling the data is important so that all variables are treated the same. This way, machine learning models can compare results from an equal scale and make accurate predictions. The key attributes that should be scaled are Age, BMI, Testosterone Level, and Antral Follicle Count. Menstrual irregularity is already in binary form, as well as the target variable, PCOS Diagnosis. A min-max normalization was used on these variables to preserve the original distribution. This is done by subtracting the minimum value of each column from the corresponding values and then dividing by the column's range (maximum value - minimum value). This scales the values in each column to the range of [0, 1]. After normalizing the numeric columns, the code converts the normalized columns back to factors. Now, the PCOS dataset contains six normalized factor variables.

5 Data Analysis and Experimental Results

Now that the PCOS dataset is cleaned, scaled, and balanced, it can be split into train and test subsets, and the training of different models to predict a PCOS diagnosis can begin. The data was split into an 80:20 ratio, where 80% of the observations will be used for the training of the models.

5.1 Simple Heuristic Benchmark

A simple benchmark to predict the majority class was created for this classification problem. This will be compared against the more complex models trained in the succeeding sections. The benchmark's accuracy in predicting a diagnosis was 57.35% correct. The precision score was 0, meaning the benchmark correctly identified none of the positive cases. The recall score was 1, indicating that the benchmark correctly identified all the positive cases. This suggests that the benchmark is highly sensitive to the minority class but lacks precision in identifying the positive cases. The AUC is 0.5, which indicates that

the benchmark does not distinguish between positive and negative PCOS diagnosis. The RMSE is 0.6530875, which suggests the benchmark is not ideal for making accurate predictions.

5.2 Lasso Logistic Regression Model

The first model trained was a Lasso logistic regression model. This model aims to shrink the coefficients of less important predictors towards zero. This model performs feature selection to improve prediction accuracy and prevent overfitting. The Lasso model was tuned using the train() function from the caret package. The tuning process involved setting the "glmnet" method, the "alpha" parameter to 1 for Lasso, and a grid of 100 "lambda" values from 10^-3 to 10^3 to find the optimal regularization. 5-fold cross-validation was used to determine the best model configuration with the Kappa statistic as the performance metric.

The Lasso model demonstrated excellent performance on the binary classification task, with an accuracy of 0.9749 and a 95% confidence interval of (0.949 and 0.9899). The Kappa statistic 0.949 indicated excellent agreement between the predicted and actual classes. The model exhibited high sensitivity (0.9625) and specificity (0.9916), correctly identifying 96.25% of positive and 99.16% of negative instances. The positive predictive value (precision) was 0.9935, meaning that 99.35% of the instances predicted as positive were positive, while the negative predictive value was 0.9516. The F1-score of 0.977 and the AUC-ROC of 0.9770483 confirmed the model's strong performance, suggesting excellent discriminative power in distinguishing between the two classes. Overall, the Lasso model demonstrated robust and reliable classification capabilities for the given problem.

The coefficients of the lasso model were calculated to look for key factors. The coefficient results suggest that menstrual irregularity, high BMI, elevated testosterone levels, and increased antral follicle counts are the most important factors identified by the LASSO model for predicting PCOS diagnosis.

5.3 Ridge Logistic Regression Model

The next model trained was a Ridge logistic regression model. The purpose of this model is to prevent overfitting and address multicollinearity. This model adds a penalty term to the loss function, which shrinks the coefficients of predictions to zero without eliminating them. The Ridge model was tuned using the train() function from the caret package. The tuning process involved setting the "glmnet" method, the "alpha" parameter to 0 for Ridge, and a grid of 100 "lambda" values from 10^-3 to 10^3 to find the optimal regularization. 5-fold cross-validation was used to determine the best model configuration with the Kappa statistic as the performance metric.

The Ridge regression model achieved strong performance on the binary classification task, with an accuracy of 0.9283 and a 95% confidence interval of (0.8915, 0.9557). The Kappa statistic of 0.8566 suggested excellent agreement between the predicted and actual classes, and the model exhibited high sensitivity (0.8750), specificity (1.0000), and positive predictive value (1.0000), correctly identifying the majority of positive and negative instances. The F1-score of 0.933 and the AUC-ROC of 0.9375 further confirmed the model's robust and reliable classification capabilities, demonstrating its excellent discriminative power in distinguishing between the two classes. Overall, the Ridge regression model proved effective for predicting PCOS diagnosis.

5.4 Elastic Net Logistic Regression Model

The next model trained was an Elastic net logistic regression model. The purpose of this model is to perform regularization and variable selection. The Enet model combines the strengths of the Lasso and

Ridge models by shrinking coefficients to zero and grouping correlated variables. The Elastic Net model was tuned using the train() function from the caret package, using the "glmnet" method. The tuning process involved exploring a range of values for the alpha parameter, which controls the balance between Lasso and Ridge regularization, and the lambda parameter, which controls the overall strength of the regularization. The alpha values were set to a sequence from 0 to 1 in 10 steps, while the lambda values were set to a grid of 100 values ranging from 10^-3 to 10^3 on a logarithmic scale. The model was evaluated using 5-fold cross-validation to ensure the tuning process did not overfit the training data, allowing the model to find the optimal configuration for the PCOS diagnosis task.

The Elastic Net model demonstrated excellent performance on the binary classification task, achieving an accuracy of 0.9785 with a 95% confidence interval of (0.9538, 0.9921). The Kappa statistic of 0.9563 indicated that the model had excellent agreement between the predicted and actual classes beyond what would be expected by chance. The model exhibited high sensitivity (0.9625), specificity (1.0000), and positive predictive value (1.0000), correctly identifying the majority of positive and negative instances. The F1-score of 0.9808917 and the AUC-ROC of 0.98125 further confirmed the model's robust and reliable classification capabilities, with excellent discriminative power in distinguishing between the two classes. Overall, the Elastic Net model proved to be a highly effective tool for predicting PCOS diagnosis.

5.5 Random Forest Tree-Ensemble Model

The next model trained was a Random forest tree ensemble model. This model improves accuracy by combining predictions from multiple decision trees. The 'caret' package was used to auto-tune the random forest model.

The random forest model for predicting PCOS diagnosis performed well, with an accuracy of 93.55% and a kappa statistic of 0.8698, indicating excellent agreement between the model's predictions and the actual class labels. The model had high sensitivity (90.62%) and specificity (97.48%), with a positive predictive value of 97.97% and a negative predictive value of 88.55%. The F1-score was 0.9415584, and the AUC-ROC was 0.94052, demonstrating the model's strong predictive performance. Overall, the results suggest that this random forest model is reliable for diagnosing PCOS based on the given set of predictors.

The varImp function was used to view the variable importance and highlights. The key factors that the random forest model used to predict PCOS diagnosis are menstrual irregularity and various BMI-related measures.

5.6 Gradient Boosted Tree-Ensemble Model

The next model trained was a Gradient-boosted tree ensemble model. The purpose of this model is to combine discretion trees in a stepwise manner sequentially. This iterative process gradually improves the accuracy. The gradient-boosted model is auto-tuned with the 'caret' package.

The gradient boosting model for predicting PCOS diagnosis performed well, with an accuracy of 94.62% and a kappa statistic of 0.8918, indicating excellent agreement between the model's predictions and the actual class labels. The model had high sensitivity (90.62%) and perfect specificity (100%), with a positive predictive value of 100% and a negative predictive value of 88.81%. The F1-score was 0.9508197, and the AUC-ROC was 0.953125, demonstrating the model's strong predictive performance. The optimal model was achieved with 150 trees, an interaction depth of 3, a shrinkage value of 0.1, and a minimum of 10 observations in each node. These tuning parameters allowed the gradient-boosting model

to capture the complex relationships between the predictor variables and the PCOS diagnosis, resulting in a highly accurate and reliable classification model.
Overall, the results suggest that this gradient-boosting model is an effective tool for diagnosing PCOS. Menstrual irregularity, testosterone levels, and various BMI-related measures were the most influential factors in the model's predictions.

5.7 Support Vector Machines

The following two models are trained as support vector machines using the linear and radial methods. The purpose of this model is to separate data points into different classes on a hyperplane. Both the linear method and the radial method use the 'caret' package to auto-tune the models. These models are trained using 5-fold cross-validation with centering and scaling to standardize features, enhancing consistency and performance.
The linear SVM model, trained without pre-processing, used 5-fold cross-validation with sample sizes of around 895 per fold. It achieved 97.49% accuracy and a Kappa of 0.9491, indicating strong agreement. Sensitivity was 95.63%, specificity 100%, and the positive predictive value 100%, highlighting effective classification. The tuning parameter C was fixed at 1.
The radial SVM model, trained without pre-processing, used 5-fold cross-validation across 1119 samples and five predictors. The best model had sigma = 0.0147 and C = 4, achieving 97.49% accuracy and Kappa of 0.9491, indicating strong agreement. Sensitivity was 95.63%, specificity 100%, and the positive predictive value 100%, demonstrating high classification performance.

5.8 Neural Network

The next mode trained was a neural network. This method aims to improve decision-making by learning from data to make predictions. This method involves more preprocessing in order for the neural network to work. First, a new data frame was created with a copy of the balanced Pcos data frame to protect the integrity of the original data. The data was then split into test and train sets, and the training data was further split into train and validation sets. Neural networks rely on numerical data to process correctly, so the factor variables were one-hot encoded so the model could process them.
First, the hyperparameters were defined for the neural network (including the number of nodes, batch size, activation function, and learning rate). The number of nodes (128) determines the model's capacity to learn complex patterns, with more nodes allowing for greater complexity and increasing the risk of overfitting. The batch size (32) affects the stability and speed of training, with smaller batches leading to more frequent updates but potentially slower convergence. The activation function ("rel") introduces non-linearity, which is important for the model's learning ability. The learning rate (0.01) controls the step size during optimization, with higher rates enabling faster convergence but potentially overshooting the optimal solution.

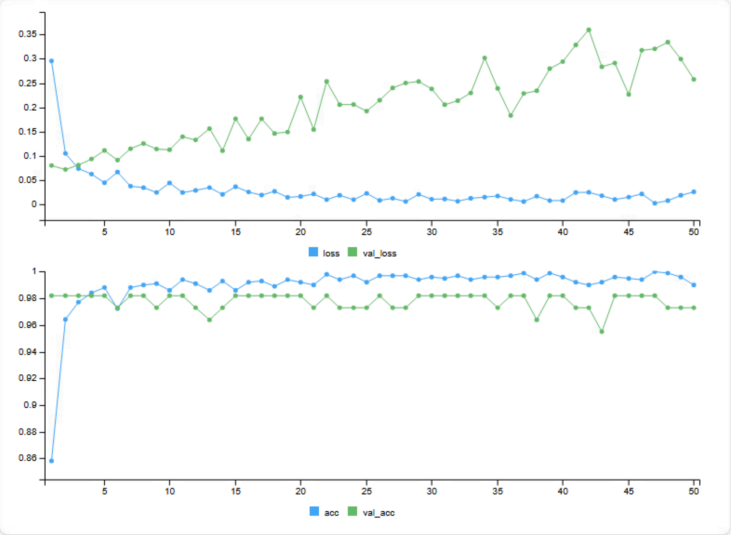

Figure 12. Neural Network Results

Then, a sequential model was created with two dense layers, each followed by a dropout layer to prevent overfitting. The final layer has a single unit with a sigmoid activation function for binary classification. The model is compiled with the binary cross-entropy loss function, the Adam optimizer with the specified learning rate, and the accuracy metric. The model is then trained for 50 epochs using the training and validation data, and the model's predictions on the validation data are obtained and converted to binary values. This ensures that the neural network is trained and evaluated using the appropriate data splits and hyperparameters, with measures in place to prevent overfitting and ensure the model's generalization performance.

5.9 Which is the Best Model?

To find the best model, the performance of all the models except the neural network was compared using the resamples() function. The results show that the Support Vector Machine with a radial kernel (svmR) is the best-performing model. Then, the performance metrics of the same model and the neural network were compared to find the best model overall. The results are presented in a data frame where the same model outperforms the Neural Network model in terms of Accuracy (0.9749 vs. 0.9729730), AUC (0.9781 vs. 0.9736066), and Precision (1.0000 vs. 0.9607843). The only exception is Recall, where the Neural Network model performs slightly better (0.98 vs. 0.9563). In conclusion, the support vector machine with the radial kernel model was the best-performing model out of all the models trained.

5.10 Compare the Best Model to the Benchmark

The results show that the SVM-Radial model performs well, with a recall of 1 (indicating it correctly identifies all positive cases), a precision of 0.9444, and an F1-score of 0.971428571428571. This suggests that the model can balance precision and recall effectively. In comparison, the benchmark model has a recall of 1 but a much lower precision of 0, resulting in an F1 score of 0. This indicates that the benchmark model correctly identifies positive cases but makes many false optimistic predictions. Overall, the SVM-Radial model demonstrates superior performance to the benchmark, with high precision and an excellent F1 score, making it a promising choice for PCOS diagnosis.

5.11 Conclusion

This project provided valuable insights into the predictive modeling of PCOS using various machine learning algorithms. The support vector machine with the radial basis method (SVM-Radial) was the best performing model, outperforming the other models in accuracy, AUC, precision, and F1 score. Notably, while the neural network model showed slightly higher recall, the SVM-Radial model demonstrated a better overall balance between sensitivity and specificity.

One of the most interesting findings was how well the SVM-Radial model handled the classification task compared to more commonly used models such as logistic regression and random forest, which dominated other Kaggle projects. Additionally, comparing the best model to a benchmark showed the importance of looking at multiple metrics like precision and F1 score, especially with a task like diagnosing a medical prediction.

The primary goal of this project was to bring awareness to polycystic ovarian syndrome (PCOS) and to better understand the features that contribute to a positive diagnosis. By applying different machine learning models to the dataset, this project identifies the key risk factors that are the

most predictive for PCOS. This project highlighted that certain features—such as irregular menstruation patterns, elevated androgen levels, and body mass index—consistently contributed to accurate diagnoses.

For future work, expanding the dataset to include more diverse populations or a larger dataset can improve model generalization. Overall, this study highlights the potential of machine learning, particularly SVMs, in aiding early detection and diagnosis of PCOS.

By finding the best-performing model (SVM-Radial), this project demonstrated the potential of machine learning as a supportive tool in early PCOS detection. This is particularly important given that an estimated 70% of women with PCOS remain undiagnosed. The findings can support clinicians and researchers in identifying high-risk individuals more effectively, contributing both to awareness and to improved diagnostic processes.

References:

1. Dalvi, S. (2025, February 19). PCOS diagnosis dataset. Kaggle. https://www.kaggle.com/datasets/samikshadalvi/pcos-diagnosis-dataset

2. World Health Organization. “Polycystic Ovary Syndrome.” World Health Organization, World Health Organization, 7 Feb. 2025, www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome.

3. kimkijun7. “PCOS Diagnosis ML Prediction with R.” Kaggle.com, Kaggle, 3 Mar. 2025, www.kaggle.com/code/kimkijun7/pcos-diagnosis-ml-prediction-with-r/notebook. Accessed 4 May 2025.