Using Machine Learning Algorithms to Predict the likelihood of PCOS based on Demographic, Clinical and Lifestyle Factors

Leticia Salazar

May 16, 2023

Contents

Abstract:

Polycystic Ovary Syndrome (PCOS) stands as a prevalent endocrine disorder affecting women of reproductive age worldwide, presenting a confluence of hormonal imbalances, reproductive irregularities, and potential metabolic complications. Characterized by irregular menstrual cycles, hyperandrogenism, and polycystic ovaries, PCOS poses multifaceted challenges that extend beyond reproductive health, encompassing metabolic disturbances and psychological implications. This complex syndrome, rooted in intricate interplays of genetic predispositions, hormonal dysregulation, and environmental factors, manifests variably among affected individuals. The clinical landscape of PCOS often requires a comprehensive, multidisciplinary approach, incorporating lifestyle modifications, pharmacological interventions, and personalized treatments to address symptoms and reduce associated health risks. Despite ongoing research efforts, elucidating the precise etiology and optimal management strategies for PCOS remains a dynamic area of exploration in contemporary medicine. This research project investigates the current studies on PCOS diagnosis, assessing the effectiveness of various machine learning algorithms including Linear Regression (LM), Logistic Regression, Decision Tree (rpart), Random Forest (randomForest), Gradient Boosting Machines (xgboost or gbm), Support Vector Machines (e1071), Neural Networks (neuralnet), and K-Nearest Neighbors (kknn or class). From my research

Key words:

Polysyctic Ovary Syndrome, Polycystic Ovarian Syndrome, PCOS, Women's Health, Machine Learning, Linear Regression, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Super Vector Machine, Neural Networks, K-Nearest Neighbors

The Problem:

Polycystic Ovary Syndrome (PCOS) is a hormonal imbalance disorder affecting women of reproductive age. Determining the precise global count of women affected by PCOS poses challenges due to many cases remaining undiagnosed. However, the World Health Organization (WHO) estimates that approximately 3.4% of women are affected [1]. While this percentage might seem relatively small, considering that women constitute 49.7% of today's population [2], nearly 13% of them fall within the reproductive age bracket [3]. This data suggests that approximately 17.5 million women report suffering from PCOS. Beyond the physical and emotional toll PCOS exacts, it also disrupts ovarian function, leading to challenges in maintaining a healthy menstrual cycle and can result in the formation of cysts, ultimately impacting fertility. It's important to note that the prevalence of PCOS varies by region and ethnic groups, with some studies suggesting higher rates of PCOS in certain populations [4]. Early identification of risk factors associated with PCOS can assist in timely interventions and lifestyle adjustments. Tailoring suggestions or treatments based on individualized risk profiles has the potential to enhance patient outcomes. Employing predictive models can play a pivotal role in increasing awareness regarding PCOS risk factors and preventive measures. However, limitations may exist in accessing comprehensive and varied datasets containing accurate demographic, clinical, and lifestyle data. It is crucial to ensure the model's transparency and interpretability to facilitate well-informed decisions and recommendations. Additionally, it's imperative that the model demonstrates proficiency across diverse demographic groups and populations. Addressing these challenges involves the development of a robust predictive model using machine learning techniques. Such an endeavor holds the promise of significantly contributing to the identification of individuals at risk of PCOS, thereby enabling early interventions and guiding personalized healthcare strategies for improved management of the condition. This project will investigate publicly available datasets that will be used to develop machine learning models that predicts the probability or risk of an individual having or developing PCOS based on demographic information (age, ethnicity, geographical location), clinical data (hormonal levels, BMI, menstrual irregularities), and lifestyle factors (dietary habits, exercise routine, stress levels).

Objectives:

Construct predictive models using machine learning techniques that utilize a dataset comprising demographic, clinical, and lifestyle variables as features and PCOS diagnosis as the target variable. Machine learning algorithms have shown promise in advancing our understanding of the disease and improving its diagnosis and treatment.

I anticipate answering the following questions with my data:

1. Are there commonalities women with and without PCOS have that can be easily dismissed as normal?
2. Are there differences for women of different race/ethnic background when it comes to having PCOS? What about women without PCOS?
3. What is the likelihood of a woman developing PCOS based on her age, ethnicity, and BMI history?
4. Can we predict the risk of insulin resistance, diabetes, and cardiovascular disease in women with PCOS based on their medical history, hormone levels, and lifestyle factors?

5. Can we predict the likelihood of successful pregnancy outcomes in women with PCOS based on their age, weight, hormone levels, and treatment history?

6. Can we predict the long-term health outcomes and quality of life of women with PCOS based on their age, lifestyle factors, hormone levels, and treatment history?

Literature Review:

For this project, my emphasis was on discovering literature reviews that validate the relevance of the dataset utilized alongside the chosen machine learning algorithms. In pursuit of this goal, I identified several articles outlined below, including studies related to PCOS and its findings in order to better understand the data I am working with.

The literature reviews collectively delve into various aspects of Polycystic Ovarian Syndrome (PCOS), employing diverse methodologies and approaches for diagnosis, classification, understanding clinical manifestations, and proposing potential treatments. Researchers across these studies have primarily utilized data mining, machine learning, and clinical investigations to address the complexity of PCOS. Below are the key characteristics, achievements, advantages, and drawbacks across these reviews:

Common Focus Areas: Multiple studies emphasize the use of machine learning algorithms, such as Naïve Bayes, Decision Trees, Artificial Neural Networks, Support Vector Machines, and ensemble methods, for PCOS diagnosis. They explore the accuracy and predictive power of these models using diverse datasets, including clinical, lifestyle, and physiological parameters. Investigations into clinical parameters and anthropometric measures aim to identify potential predictors or indicators of PCOS, such as hormonal imbalances, insulin resistance, metabolic traits, obesity, and associated risks like infertility and cardiovascular issues. Studies examine the diverse clinical presentations and phenotypic variations within PCOS, shedding light on how different subgroups may manifest the syndrome and respond to various treatments.

Achievements: Machine learning models, particularly those utilizing Convolutional Neural Networks (CNNs) and ensemble methods, have shown high accuracy in diagnosing PCOS. These models effectively utilize diverse datasets, ranging from ultrasound images to clinical parameters. Research has identified several potential predictive factors for PCOS, including hormonal markers, lifestyle attributes, and metabolic indicators, offering insights into early detection and tailored treatment strategies. Studies evaluating PCOS awareness among women have highlighted the importance of education and awareness campaigns in enhancing understanding and facilitating early diagnosis.

Advantages: Machine learning algorithms offer promising accuracy rates, particularly CNNs and ensemble models, in diagnosing PCOS using various non-invasive parameters. Understanding phenotypic variations aids in tailoring treatments based on specific subgroups, potentially improving patient outcomes and management. Efforts toward identifying early markers or predictive factors can facilitate early intervention and lifestyle modifications, mitigating long-term health risks associated with PCOS.

Drawbacks and Recommendations: Some studies may suffer from limited sample sizes or datasets, impacting the generalizability of findings. Larger and more diverse datasets are recommended for robust model development and validation. While certain machine learning models showcase high accuracy, the variability in dataset characteristics and preprocessing methods could influence their effectiveness across different populations or settings. The multifaceted nature of PCOS, influenced by both genetic and environmental factors, presents challenges in pinpointing a singular cause or standard diagnostic criteria.

The collective body of literature reviews signifies advancements in PCOS diagnosis, understanding clinical manifestations, and potential avenues for tailored treatments. The use of machine learning models, especially those employing CNNs and ensemble methods, showcases significant promise in accurate PCOS diagnosis based on non-invasive parameters. However, further research is necessary, emphasizing larger and more diverse datasets, refined models, and a multidisciplinary approach to fully comprehend the complexity of PCOS and enhance diagnostic and treatment strategies.

Datasets:

The following datasets will be used in this project:

1. DataSet for PCOS: These datasets provide the results of an untargeted metabolomic survey was conducted on the metabolites in the FF of 35 patients with PCOS and 37 age-matched individuals as control
2. Polycystic Ovary Syndrome (PCOS): PCOS dataset contains all physical and clinical parameters of patients from 10 different hospitals across Kerala, India.
3. Menstrual Cycle Data: Randomized Comparison of Two Internet –Supported Natural Family Planning Methods.

Methodology:

Data Collection and Preparation: Gather and preprocess a comprehensive dataset containing demographic details, clinical measurements (hormone levels, BMI, menstrual history), and lifestyle information (diet, exercise, stress levels) from a diverse population.

Feature Selection and Engineering: Identify the most relevant features through exploratory analysis and feature engineering techniques, ensuring that the model focuses on key predictors for PCOS.

Model Training and Evaluation: Train the machine learning model using appropriate algorithms (such as Logistic Regression, Random Forest, Support Vector Machines) on a subset of the dataset, validate its performance using cross-validation techniques, and evaluate its accuracy, precision, recall, and F1-score.

Prediction and Risk Assessment: Deploy the trained model to predict the likelihood or risk of PCOS in new, unseen data, providing valuable insights into individuals who might be predisposed to or already have the condition.

Interpretation and Recommendations: Interpret the model's findings, analyze the significance of various factors contributing to the prediction, and offer recommendations or interventions based on identified risk factors to potentially prevent or manage PCOS.

This project will investigate mainly publicly available datasets that will be used to create predictive models on markers in routine test results to make a diagnosis. Some variables that are included in these datasets are:

- Age
- Weight
- BMI
- Race/ethnicity
- Family history of PCOS
- Menstrual cycle irregularity
- Hormone levels (e.g., testosterone, LH, FSH)

- Insulin resistance
- Physical activity level
- Diet

I will be using R (statistical performing language) to perform exploratory data analysis to process and analyze the data to check for structural errors and be able to create graphs and perform tests with minimal errors. Once the data is ready to use, I will be splitting the data into a training and test set to be able to use a machine learning algorithm such as logistic regression to create a predictive model. The predictive model will be based on markers (variables mentioned above) used to identify individuals who are at high risk for PCOS and target interventions to manage the condition.

To answer my research question: * The datasets publicly available and the NICHD Dash looking to obtain all have PCOS and non-PCOS patients including demographic information, medical history, and laboratory tests. Preprocess the data by removing missing values, outliers, and redundant variables. Perform feature selection to identify the most informative variables for prediction. * Split the dataset into training, validation, and testing sets. The training set is used to train the machine learning algorithm, the validation set is used to tune hyperparameters and prevent overfitting, and the testing set is used to evaluate the performance of the final model. * Select an appropriate machine learning algorithm for the task at hand, such as logistic regression, decision trees, random forests, support vector machines, or neural networks. Train the algorithm on the training set using various techniques, such as cross-validation and regularization, to optimize its performance. * Evaluate the performance of the trained model on the validation set using various metrics, such as accuracy, precision, recall, F1 score, and area under the curve. Use feature importance analysis to identify the most influential variables for prediction. * Tune the hyperparameters of the machine learning algorithm using grid search, random search, or Bayesian optimization to improve its performance on the validation set. * Select the final model based on its performance on the validation set. Evaluate its performance on the testing set to assess its generalization ability. * Interpret the results of the machine learning algorithm using various techniques, such as decision trees, feature importance analysis, and partial dependence plots. Visualize the results using graphs, charts, and heatmaps to facilitate understanding and communication. * Deploy the trained model on new data and disseminate the findings through scientific publications, presentations, and online platforms.

Note: this methodology plan is not exhaustive and may vary depending on the specific research question, dataset, and machine learning algorithm used.

Note 2: data has already been collected, there is no need for me to gather participants, perform exams (such as bloodwork), use medical equipment to collect the data, perform surveys, have a location to perform a study, etc. I will be the sole person studying the data set and conducting the analysis.

Assumptions:

While there's limited information available in the medical field and even less data sets available to analyze, I have some concerns on being successful in predicting a PCOS diagnosis. Yet there are justifications exploring PCOS in depth:

- Identify diagnostic biomarkers that can distinguish PCOS patients from healthy individuals or those with other disorders. These biomarkers can aid in earlier diagnosis and better management of the disease.
- Predict the likelihood of disease progression and the risk of developing complications, such as diabetes and cardiovascular disease, in PCOS patients. This information can guide treatment decisions and improve patient outcomes.
- Develop personalized treatment plans for PCOS patients based on their individual characteristics and medical history. This approach can lead to more effective and targeted interventions.

- Integrate data from various sources, such as electronic health records, imaging studies, and genetic analyses, to provide a more comprehensive understanding of PCOS. This can help identify new pathways involved in the disease and potential targets for therapy.
- Aid in the design and analysis of clinical trials, leading to more efficient and informative studies. This can accelerate the development of new treatments for PCOS.

Early diagnosis and management of PCOS can lead to better health outcomes, improved quality of life, and reduced long-term health risks. Therefore, predicting PCOS diagnosis can have several societal benefits, including:

- Predicting PCOS diagnosis can help healthcare providers identify women at risk of developing PCOS and intervene early with appropriate treatment, such as lifestyle modifications and medication, to prevent or minimize the long-term health consequences of the disorder.
- Early diagnosis and treatment of PCOS can help manage symptoms such as irregular periods, infertility, acne, and excess hair growth, leading to improved physical and mental health outcomes for affected women.
- By predicting PCOS diagnosis and intervening early, healthcare providers can prevent or reduce the need for more expensive treatments or surgeries later in life, resulting in cost savings for individuals, healthcare systems, and society.
- Predicting PCOS diagnosis can increase awareness of the disorder among healthcare providers, patients, and the public, leading to more education, research, and advocacy efforts aimed at improving PCOS diagnosis, treatment, and management.
- Early intervention and management of PCOS can improve the quality of life for affected women, leading to increased productivity, better mental health, and greater overall well-being.

Overall, I'll be able to explore the insights into PCOS pathophysiology, diagnosis, and treatment. Their use in PCOS research can lead to more personalized and effective care for patients with this complex disorder.

Experimentation and Results:

Conclusion:

My initial assumption was that while there's limited information available in the medical field and even less data sets available to analyze, I have some concerns on being successful in predicting a PCOS diagnosis.

References:

1. Bulsara, J., Patel, P., Soni, A., & Acharya, S. (2021, February 10). A review: Brief insight into polycystic ovarian syndrome. Endocrine and Metabolic Science. Retrieved February 23, 2023, from https://www.sciencedirect.com
2. World female population, 1960-2022. Knoema. (2022). Retrieved February 24, 2023, from https://knoema.com.

3. MarchofDimes. (2022). Population of women 15-44 years by age: United States, 2020. March of Dimes | PeriStats. Retrieved February 24, 2023, from https://www.marchofdimes.org

4. Engmann, L., Jin, S., Sun, F., Legro, R. S., Polotsky, A. J., Hansen, K. R., Coutifaris, C., Diamond, M. P., Eisenberg, E., Zhang, H., Santoro, N., & Reproductive Medicine Network (2017). Racial and ethnic differences in the polycystic ovary syndrome metabolic phenotype. American journal of obstetrics and gynecology, 216(5), 493.e1–493.e13. https://doi.org

5. Fehring, Richard J., "Menstrual Cycle Data" (2012). Randomized Comparison of Two Internet-Supported Methods of Natural Family Planning. 7. https://epublications.marquette.edu

6. Khan, M. J., Ullah, A., & Basit, S. (2019). Genetic Basis of Polycystic Ovary Syndrome (PCOS): Current Perspectives. The application of clinical genetics, 12, 249–260. https://doi.org/

Appendices:

Appendix A - Figures:

Appendix B - Tables:

Appendix C - R Code: