Predictive Modelling for Polycystic Ovary Syndrome (PCOS) Detection Using Machine Learning: A Comprehensive Analysis and Clinical Insights

Abstract: Polycystic Ovary Syndrome (PCOS) is a complex endocrine disorder with diverse clinical manifestations, affecting women of reproductive age. This abstract summarizes the key findings and implications derived from a machine learning model trained on **a dataset comprising 541 individuals with various demographic and clinical features**. The dataset encompasses parameters such as age, weight, hormonal markers, and lifestyle factors.

1. Introduction

(a) Polycystic Ovary Syndrome (PCOS): A Comprehensive Introduction

Polycystic Ovary Syndrome, commonly abbreviated as PCOS, is a complex and multifaceted endocrine disorder that affects millions of individuals worldwide, predominantly women of reproductive age. This syndrome presents a wide array of clinical manifestations, encompassing hormonal imbalances, metabolic disturbances, and reproductive challenges, making it one of the most common endocrine disorders among women. PCOS is a subject of significant medical and scientific interest due to its far-reaching impact on the overall health and quality of life of those affected. In this comprehensive introduction, we will delve into the various facets of PCOS, including its clinical features, diagnostic criteria, potential causes, and the far-reaching implications for those living with this condition.

Polycystic Ovary Syndrome is a condition characterized by a constellation of symptoms. **While no two individuals with PCOS experience precisely the same set of issues, several common clinical features emerge. Irregular menstrual cycles, often marked by prolonged or absent periods, serve as one of the hallmark signs of PCOS.** Hyperandrogenism, which refers to elevated levels of male hormones such as testosterone, manifests in various ways, including acne, hirsutism (excessive hair growth), and male-pattern baldness. The presence of multiple small cysts on the ovaries, known as ovarian cysts or follicular cysts, is another characteristic feature, although the name 'polycystic' can be somewhat misleading, as the cysts themselves are not always a central concern in PCOS.

Diagnosing PCOS is not a straightforward task, primarily due to the heterogeneity of its clinical presentation and the absence of a single definitive test. **The Rotterdam criteria, established in 2003, have been widely adopted for diagnosis and require the presence of at least two out of three key features: irregular menstrual cycles, clinical or biochemical signs of hyperandrogenism, and the appearance of polycystic ovaries on ultrasound.** However, these criteria have faced criticism for their potential to lead to overdiagnosis or misdiagnosis, especially among lean individuals with PCOS who may not exhibit the typical signs of obesity. The precise aetiology of PCOS remains elusive, but researchers have made significant strides in understanding its potential causes. Genetics play a crucial role, with a familial predisposition to PCOS often observed. Hormonal imbalances, including insulin resistance, are central to the condition. Insulin resistance can lead to elevated insulin levels, which, in turn, stimulate the ovaries to produce excess androgens. Lifestyle factors, such as obesity and sedentary habits, can exacerbate insulin resistance and PCOS symptoms. Additionally, environmental factors and prenatal exposures have also been explored as potential contributors to the development of PCOS.

While PCOS is often associated with reproductive challenges, its effects extend far beyond the realm of fertility. Metabolic disturbances are a common companion to PCOS, with individuals often facing an increased risk of developing type 2 diabetes, dyslipidaemia, and cardiovascular disease. Obesity, which frequently co-occurs with PCOS, compounds these metabolic risks. Moreover, PCOS can affect mental health, contributing to increased rates of depression, anxiety, and diminished quality of life. It is important to recognize that PCOS is not solely a gynaecological concern; it is a systemic disorder with far-reaching implications for overall well-being.

As our understanding of PCOS continues to evolve, so too does our approach to its management and treatment. A comprehensive, patient-centred strategy is essential, encompassing lifestyle modifications, hormonal therapies, and a focus on individualized care. Early diagnosis and proactive management can mitigate the long-term health risks associated with PCOS and improve the quality of life for those affected. In this series of articles, we will delve deeper into the various aspects of PCOS, exploring its impact on reproductive health, metabolic function, mental well-being, and available treatment options. By gaining a more profound understanding of PCOS, we hope to empower individuals living with this condition and healthcare providers to better navigate its complexities and improve outcomes for those affected.

(b) Research Objective

The research objective of this study is to comprehensively investigate the clinical features, diagnostic criteria, potential causes, and broader implications of Polycystic Ovary Syndrome (PCOS) in order to enhance our understanding of this complex endocrine disorder and provide valuable insights for improving diagnosis, management, and outcomes for individuals affected by PCOS.

(c) A brief Overview of the dataset and ML model used

The machine learning model used in this study is designed to analyse and extract insights from a dataset containing information related to Polycystic Ovary Syndrome (PCOS). The dataset comprises 541 records and includes various features such as patient demographics (age, weight, height), clinical measurements (BMI, blood pressure), hormonal levels (FSH, LH, insulin), reproductive history (menstrual cycles, pregnancies), and lifestyle factors (exercise, diet).

The primary goal of the machine learning model is to explore the relationships and patterns within this diverse set of variables to uncover key findings related to PCOS. By applying machine learning techniques, we aim to identify significant correlations, predictive factors, and potential risk indicators associated with PCOS. Additionally, the model will help assess the diagnostic and prognostic value of various features in relation to PCOS, shedding light on the complex nature of this syndrome.

By leveraging this dataset and machine learning, our study seeks to provide valuable insights into the clinical aspects and implications of PCOS, which can aid in better understanding, diagnosing, and managing this condition, ultimately improving the quality of life for individuals affected by PCOS.

(d) Importance of early PCOS Detection

The importance of early detection of Polycystic Ovary Syndrome (PCOS) cannot be overstated due to the numerous significant implications for an individual's health and quality of life. Here are several key reasons why early detection of PCOS is crucial:

Improved Management and Treatment: Early diagnosis allows for timely intervention and management of PCOS. Lifestyle modifications, such as changes in diet and exercise, are often more effective when implemented early, potentially preventing or reducing the severity of complications like obesity, insulin resistance, and metabolic syndrome.

Preventing Long-Term Health Risks: PCOS is associated with an increased risk of several long-term health issues, including type 2 diabetes, cardiovascular disease, and dyslipidaemia. Detecting PCOS early enables healthcare providers to monitor and address these risks promptly, potentially preventing or delaying the onset of these conditions.

Enhanced Fertility Management: PCOS is a leading cause of infertility in women. Early detection allows for the timely initiation of fertility treatments and interventions, increasing the chances of successful conception for those desiring to start a family.

Mental Health and Quality of Life: PCOS can have a profound impact on mental health, leading to conditions like depression and anxiety. Early identification and intervention can address these mental health issues promptly, improving an individual's overall well-being and quality of life.

Personalized Care: Early diagnosis facilitates individualized care plans. Healthcare providers can tailor treatments and recommendations to address specific symptoms and concerns, taking into account each patient's unique needs and goals.

Preventing Complications: PCOS can lead to complications during pregnancy, such as gestational diabetes and preeclampsia. Identifying PCOS before or early in pregnancy allows for vigilant monitoring and management, reducing the risk of adverse outcomes.

Educational Support: Early detection provides an opportunity for healthcare providers to educate patients about PCOS and its potential consequences. Informed patients are better equipped to make lifestyle choices that can positively impact their health.

Reduced Economic Burden: Early intervention and management of PCOS can reduce the economic burden on individuals and healthcare systems. Preventing or mitigating the development of costly chronic health conditions ultimately saves both personal and societal healthcare costs.

2. Literature Review

(a) Review of existing research on PCOS

PCOS Definition and Diagnosis: PCOS is primarily diagnosed based on criteria established by expert groups, including the Rotterdam criteria (2003) and the Androgen Excess and PCOS Society criteria (2006). These criteria typically require the presence of two out of three features: irregular menstrual cycles, clinical or biochemical signs of hyperandrogenism, and polycystic ovarian morphology seen on ultrasound.

Prevalence and Epidemiology: PCOS is one of the most common endocrine disorders in women of reproductive age, with a prevalence estimated to be as high as 10-15%. **It can affect women of all ethnic backgrounds.**

Clinical Features and Symptoms: PCOS presents a wide range of clinical symptoms, including irregular periods, hirsutism (excessive hair growth), acne, male-pattern baldness, and obesity. Infertility and pregnancy complications are common concerns for women with PCOS.

Metabolic and Cardiovascular Risks: Insulin resistance and metabolic disturbances, such as obesity and dyslipidaemia, are common in PCOS. Women with PCOS are at **an increased risk of developing type 2 diabetes, cardiovascular disease, and non-alcoholic fatty liver disease (NAFLD).**

Hormonal and Genetic Factors: Hormonal imbalances, including elevated androgens (such as testosterone) and insulin, are central to the pathophysiology of PCOS. Genetic factors play a role, with a family history of PCOS often observed.

Impact on Mental Health: PCOS is associated with an **increased risk of mood disorders, including depression and anxiety.**

Treatment Approaches: Treatment for PCOS typically involves lifestyle modifications, such as diet and exercise, to address metabolic issues. Hormonal therapies, including oral contraceptives and anti-androgens, are commonly prescribed to manage symptoms. Fertility treatments, like ovulation induction, are used for those trying to conceive.

Ongoing Research: Ongoing research explores various aspects of PCOS, including the role of gut microbiota, epigenetic factors, and novel therapeutic interventions. Advances in understanding the genetic basis of PCOS may lead to more targeted treatments.

Patient-Centred Care: There is a growing emphasis on patient-centred care, recognizing the unique needs and experiences of individuals with PCOS. Support groups and patient education play important roles in managing PCOS effectively.

(b) Limitations of current PCOS prediction methods

The limitations of current Polycystic Ovary Syndrome (PCOS) prediction methods underscore the need for improved models and diagnostic approaches. While these methods have been valuable in identifying and managing PCOS, several shortcomings exist:

Heterogeneity of PCOS Presentation: PCOS is a complex condition with a wide range of clinical manifestations. Current diagnostic criteria, such as the Rotterdam criteria, require the presence of certain features (e.g., irregular menstrual cycles, hyperandrogenism, ovarian morphology) but do not account for the significant variability in symptom severity and combinations. This heterogeneity makes it challenging to diagnose and predict PCOS accurately in all affected individuals.

Overdiagnosis and Misdiagnosis: The reliance on a combination of criteria can lead to overdiagnosis or misdiagnosis. Lean individuals with PCOS may not exhibit the typical signs of obesity or elevated androgens, potentially leading to underdiagnosis. On the other hand, some individuals **may be misdiagnosed due to the presence of other conditions with overlapping symptoms**.

Limited Predictive Accuracy: Current PCOS prediction models, including clinical criteria and biochemical assessments, may **lack the precision needed for early and accurate diagnosis**. These models often involve binary decisions based on thresholds (e.g., presence/absence of certain features), which can miss cases that fall near these thresholds or evolve over time.

Lack of Biomarkers: PCOS lacks specific biomarkers that can reliably indicate its presence or severity. While certain hormonal imbalances and metabolic markers are associated with PCOS, they are **not exclusive to the condition and can vary widely among affected individuals.**

Invasive Testing: Some diagnostic methods, such as transvaginal ultrasound to assess ovarian morphology, can be uncomfortable and invasive. This can discourage some individuals from seeking evaluation or monitoring.

Delayed Diagnosis: PCOS is often diagnosed years after the onset of symptoms, leading to missed opportunities for early intervention. Delayed diagnosis can result in the progression of metabolic complications and challenges related to fertility and mental health.

Limited Understanding of Aetiology: The **exact causes of PCOS remain unclear**, making it difficult to develop highly accurate predictive models. While genetics, hormonal imbalances, and lifestyle factors are implicated, the precise interplay of these factors is not fully elucidated.

Need for Individualized Care: PCOS is a highly individualized condition, and treatment plans should be tailored to each patient's unique needs. Current models may not adequately account for this variability in care.

Psychosocial Impact: PCOS has a substantial psychosocial impact on individuals, including **depression, anxiety, and reduced quality of life**. **Existing models may not fully consider these aspects** when predicting and managing PCOS.

Improved PCOS prediction models should aim to address these limitations by incorporating advanced machine learning techniques, utilizing a broader array of biomarkers, considering individual variability, and accounting for the psychosocial aspects of the condition. Such models have the potential to enable earlier and more accurate diagnosis, personalized treatment plans, and better overall management of PCOS.

(c) Relevant studies that have used machine learning for medical diagnosis.

Diagnosis of Skin Cancer: Esteva et al. (2017) published a groundbreaking study on using deep learning for skin cancer diagnosis. They developed a convolutional neural network (CNN) that outperformed dermatologists in classifying skin lesions as benign or malignant. The model was trained on a large dataset of dermoscopy images and showed high accuracy in identifying skin cancers.

Breast Cancer Detection: Kooi et al. (2017) explored the application of deep learning in breast cancer detection using mammography images. Their study demonstrated that a deep neural network could achieve performance comparable to radiologists in distinguishing between malignant and benign breast lesions.

Diabetic Retinopathy Screening: Gulshan et al. (2016) conducted research on the use of deep learning to detect diabetic retinopathy from retinal images. Their model achieved high sensitivity and specificity, highlighting the potential for AI-based screening of diabetic eye disease, particularly in resource-constrained settings.

Cardiovascular Disease Risk Prediction: Madani et al. (2018) employed machine learning to predict cardiovascular disease risk using electronic health records (EHR) data. Their model analysed patient data, including demographics, medical history, and lab results, to predict the likelihood of cardiovascular events. The model provided valuable risk assessments for clinicians.

Parkinson's Disease Diagnosis: Tsanas et al. (2012) used machine learning to develop a diagnostic model for Parkinson's disease based on voice recordings. By analysing speech features, they achieved high accuracy in differentiating between healthy individuals and those with Parkinson's disease. Voice-based diagnosis offers a non-invasive and cost-effective screening method.

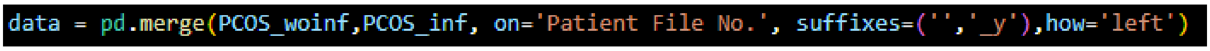
Alzheimer's Disease Prediction: Sarraf and Tofighi (2016) used machine learning algorithms to predict the risk of Alzheimer's disease based on MRI brain images. Their approach demonstrated the potential for early diagnosis and monitoring of neurodegenerative diseases through automated image analysis.

Cervical Cancer Screening: Hu et al. (2018) developed a machine learning model for the automated detection of cervical dysplasia from pap smear images. Their system achieved high accuracy in identifying abnormal cells, offering a valuable tool for cervical cancer screening and early intervention.

3. Data Collection and Preprocessing

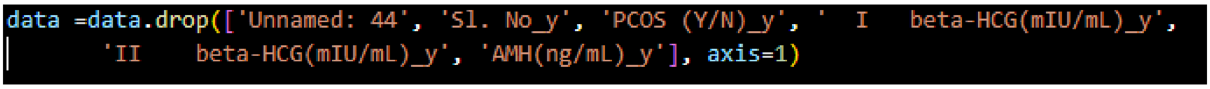
The dataset used for this analysis contains 44 parameters and was collected from 10 different hospitals across Kerala, India. The data includes physical and clinical parameters related to Polycystic Ovary Syndrome (PCOS) and infertility. Here is an overview of the data preprocessing steps:

Merging Two Files: The dataset was initially sorted into two separate files based on patients with infertility and without infertility. These two files were then merged into one based on the 'Patient File No.' to create a comprehensive dataset.



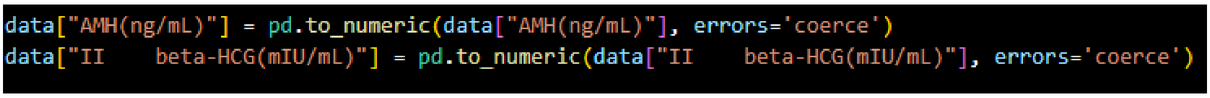
```
data = pd.merge(PCOS_woinf,PCOS_inf, on='Patient File No.', suffixes=('','_y'),how='left')
```

Dropping Repeated Features: After merging the two files, there were some duplicate columns, which were dropped from the dataset.



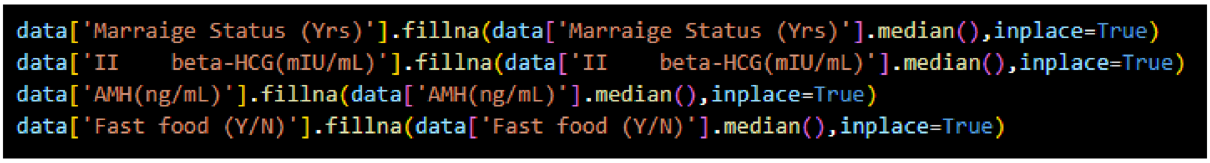
```
data =data.drop(['Unnamed: 44', 'Sl. No_y', 'PCOS (Y/N)_y', '  I   beta-HCG(mIU/mL)_y',
       'II    beta-HCG(mIU/mL)_y', 'AMH(ng/mL)_y'], axis=1)
```

Encoding Categorical Variables: Categorical variables that were stored as data type 'object' were converted into numeric values to make them suitable for analysis.



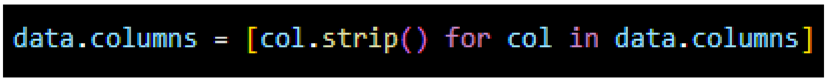
```
data["AMH(ng/mL)"] = pd.to_numeric(data["AMH(ng/mL)"], errors='coerce')
data["II    beta-HCG(mIU/mL)"] = pd.to_numeric(data["II    beta-HCG(mIU/mL)"], errors='coerce')
```

Handling Missing Values: Missing values were addressed by filling them with the median of their respective features. The features 'Marraige Status (Yrs)', 'II beta-HCG(mIU/mL)', 'AMH(ng/mL)', and 'Fast food (Y/N)' were imputed.



```
data['Marraige Status (Yrs)'].fillna(data['Marraige Status (Yrs)'].median(),inplace=True)
data['II    beta-HCG(mIU/mL)'].fillna(data['II    beta-HCG(mIU/mL)'].median(),inplace=True)
data['AMH(ng/mL)'].fillna(data['AMH(ng/mL)'].median(),inplace=True)
data['Fast food (Y/N)'].fillna(data['Fast food (Y/N)'].median(),inplace=True)
```

Cleaning Column Names: Extra spaces in column names were removed for clarity.



```
data.columns = [col.strip() for col in data.columns]
```

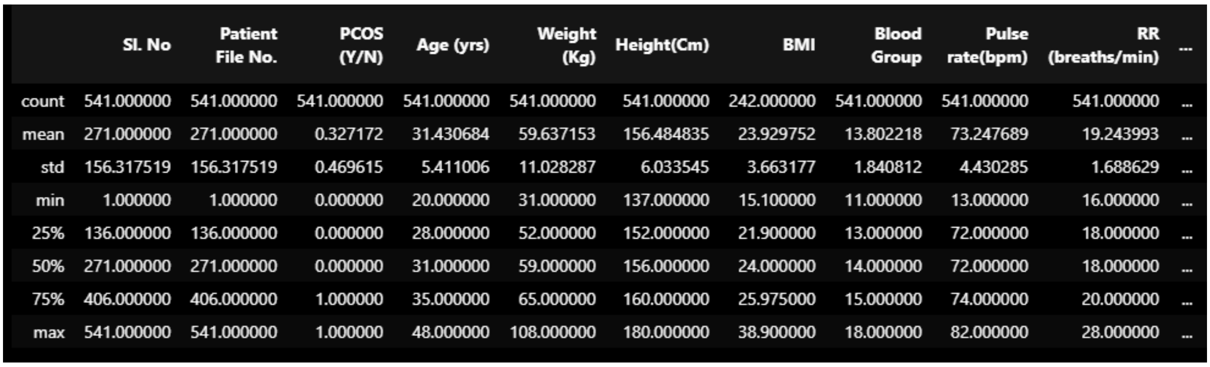
Handling Missing Data: Missing data was handled by filling in the missing values with the median of their respective features. This approach is common when dealing with numerical data as it is robust to outliers and ensures that the imputed values are representative of the central tendency of the data.

Feature Engineering or Selection: The data preprocessing steps primarily focused on data cleaning and preparation. No explicit feature engineering or feature selection was mentioned in the provided information. However, these steps may be performed later in the analysis to enhance model performance or gain insights into the dataset's key features. Feature engineering could involve creating new meaningful features from existing ones, while feature selection might involve identifying and keeping only the most relevant features for modelling. These steps are often conducted based on the specific goals of the analysis and the performance of machine learning models.

4. EDA and Methodology

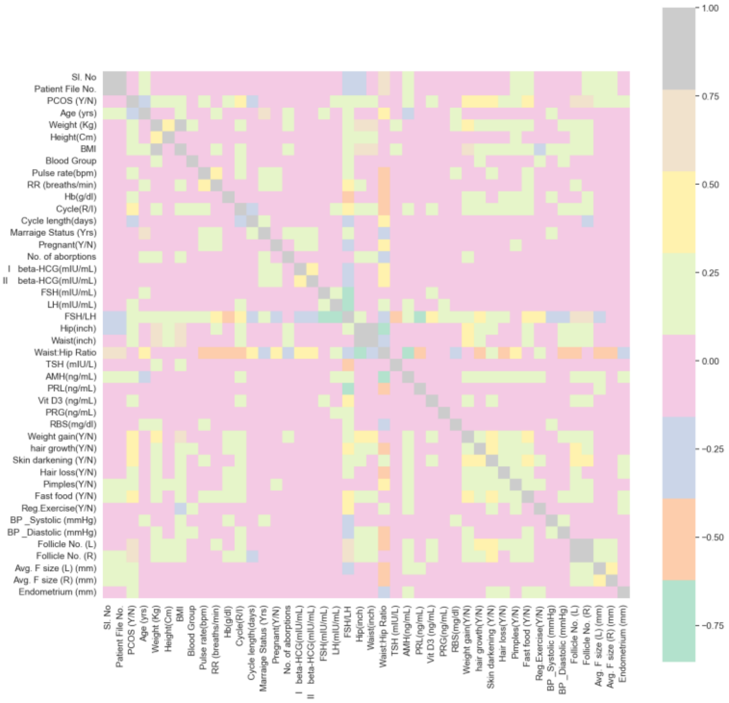
(a) Exploratory Data Analysis

Basic Statistical Details: The initial step involves obtaining basic statistical information about the dataset using the `describe()` function. This summary provides an overview of the central tendency, spread, and distribution of each numerical feature.

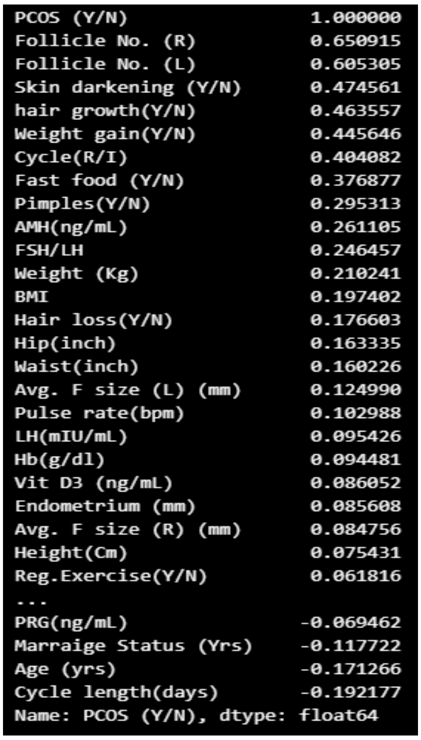


	Sl. No	Patient File No.	PCOS (Y/N)	Age (yrs)	Weight (Kg)	Height(Cm)	BMI	Blood Group	Pulse rate(bpm)	RR (breaths/min)	...
count	541.000000	541.000000	541.000000	541.000000	541.000000	541.000000	242.000000	541.000000	541.000000	541.000000	...
mean	271.000000	271.000000	0.327172	31.430684	59.637153	156.484835	23.929752	13.802218	73.247689	19.243993	...
std	156.317519	156.317519	0.469615	5.411006	11.028287	6.033545	3.663177	1.840812	4.430285	1.688629	...
min	1.000000	1.000000	0.000000	20.000000	31.000000	137.000000	15.100000	11.000000	13.000000	16.000000	...
25%	136.000000	136.000000	0.000000	28.000000	52.000000	152.000000	21.900000	13.000000	72.000000	18.000000	...
50%	271.000000	271.000000	0.000000	31.000000	59.000000	156.000000	24.000000	14.000000	72.000000	18.000000	...
75%	406.000000	406.000000	1.000000	35.000000	65.000000	160.000000	25.975000	15.000000	74.000000	20.000000	...
max	541.000000	541.000000	1.000000	48.000000	108.000000	180.000000	38.900000	18.000000	82.000000	28.000000	...

Correlation Matrix: A correlation matrix was computed to understand how different features correlate with each other and, more importantly, with the target variable 'PCOS (Y/N).' The heatmap visualization of the correlation matrix helps identify features that have a significant positive or negative correlation with PCOS.

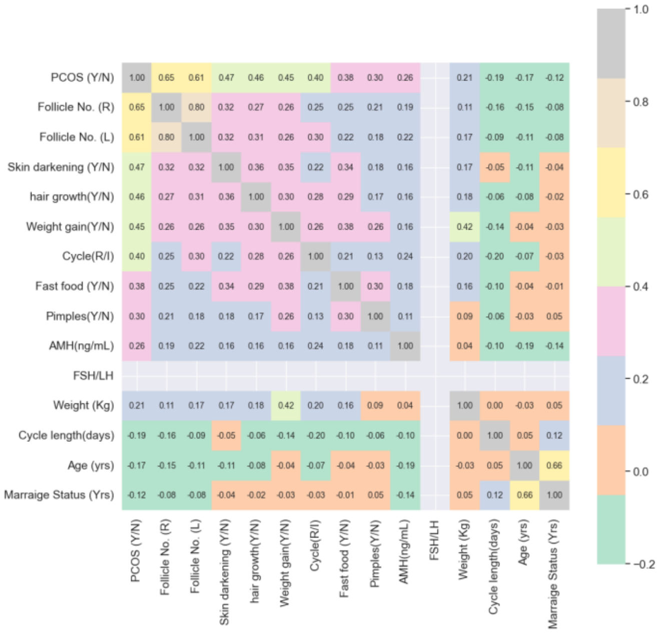


Correlation with PCOS: The correlation values of features with 'PCOS (Y/N)' were examined. Positive correlations indicate that as the feature value increases, the likelihood of having PCOS also increases, while negative correlations suggest the opposite. Features such as **'Follicle No. (R)' and 'Follicle No. (L)' showed the highest positive correlations with PCOS.**



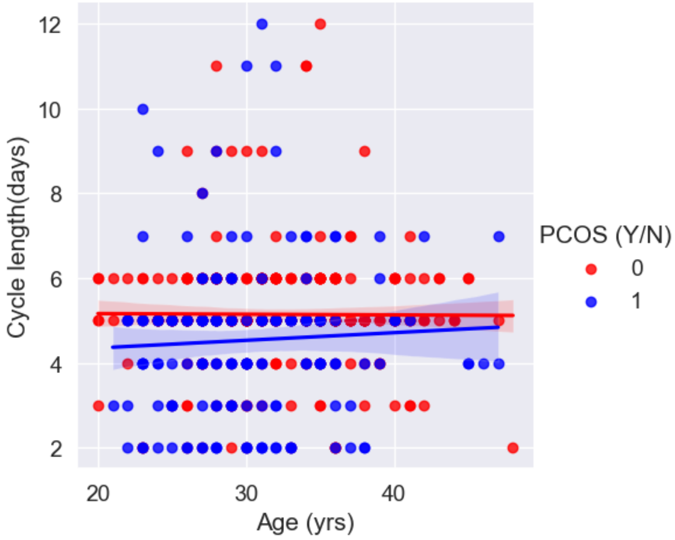
```
PCOS (Y/N)                1.000000
Follicle No. (R)          0.650915
Follicle No. (L)          0.605305
Skin darkening (Y/N)      0.474561
hair growth(Y/N)          0.463557
Weight gain(Y/N)          0.445646
Cycle(R/I)                0.404082
Fast food (Y/N)           0.376877
Pimples(Y/N)              0.295313
AMH(ng/mL)                0.261105
FSH/LH                    0.246457
Weight (Kg)               0.210241
BMI                       0.197402
Hair loss(Y/N)            0.176603
Hip(inch)                 0.163335
Waist(inch)               0.160226
Avg. F size (L) (mm)      0.124990
Pulse rate(bpm)           0.102988
LH(mIU/mL)                0.095426
Hb(g/dl)                  0.094481
Vit D3 (ng/mL)            0.086052
Endometrium (mm)          0.085608
Avg. F size (R) (mm)      0.084756
Height(Cm)                0.075431
Reg.Exercise(Y/N)         0.061816
...
PRG(ng/mL)               -0.069462
Marraige Status (Yrs)    -0.117722
Age (yrs)                -0.171266
Cycle length(days)       -0.192177
Name: PCOS (Y/N), dtype: float64
```

Feature Selection: Based on correlation, a subset of features that had the highest positive and negative correlations with 'PCOS (Y/N)' was selected for further analysis. These features provide valuable insights into factors associated with PCOS.

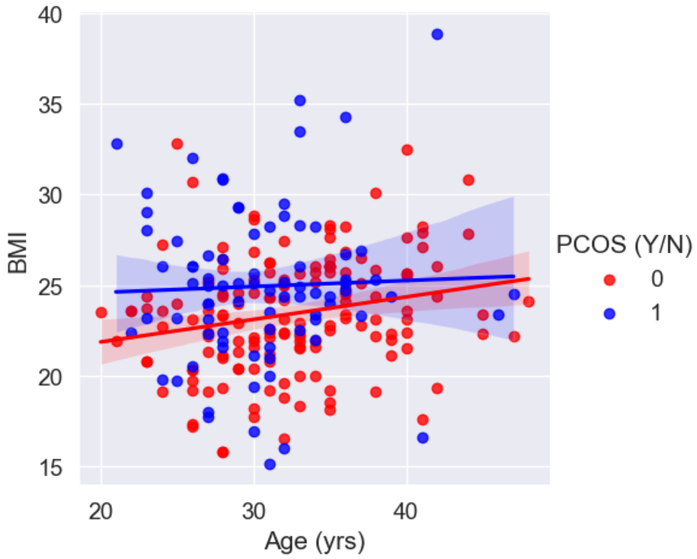


	PCOS (Y/N)	Follicle No. (R)	Follicle No. (L)	Skin darkening (Y/N)	hair growth(Y/N)	Weight gain(Y/N)	Cycle(R/I)	Fast food (Y/N)	Pimples(Y/N)	AMH(ng/mL)	FSH/LH	Weight (Kg)	Cycle length(days)	Age (yrs)	Marraige Status (Yrs)
PCOS (Y/N)	1.00	0.65	0.61	0.47	0.46	0.45	0.40	0.38	0.30	0.26		0.21	-0.19	-0.17	-0.12
Follicle No. (R)	0.65	1.00	0.80	0.32	0.27	0.26	0.25	0.25	0.21	0.19		0.11	-0.16	-0.15	-0.08
Follicle No. (L)	0.61	0.80	1.00	0.32	0.31	0.26	0.30	0.22	0.18	0.22		0.17	-0.09	-0.11	-0.08
Skin darkening (Y/N)	0.47	0.32	0.32	1.00	0.36	0.35	0.22	0.34	0.18	0.16		0.17	-0.05	-0.11	-0.04
hair growth(Y/N)	0.46	0.27	0.31	0.36	1.00	0.30	0.28	0.29	0.17	0.16		0.18	-0.06	-0.08	-0.02
Weight gain(Y/N)	0.45	0.26	0.26	0.35	0.30	1.00	0.26	0.38	0.26	0.16		0.42	-0.14	-0.04	-0.03
Cycle(R/I)	0.40	0.25	0.30	0.22	0.28	0.26	1.00	0.21	0.13	0.24		0.20	-0.20	-0.07	-0.03
Fast food (Y/N)	0.38	0.25	0.22	0.34	0.29	0.38	0.21	1.00	0.30	0.18		0.16	-0.10	-0.04	-0.01
Pimples(Y/N)	0.30	0.21	0.18	0.18	0.17	0.26	0.13	0.30	1.00	0.11		0.09	-0.06	-0.03	0.05
AMH(ng/mL)	0.26	0.19	0.22	0.16	0.16	0.16	0.24	0.18	0.11	1.00		0.04	-0.10	-0.19	-0.14
FSH/LH															
Weight (Kg)	0.21	0.11	0.17	0.17	0.18	0.42	0.20	0.16	0.09	0.04		1.00	0.00	-0.03	0.05
Cycle length(days)	-0.19	-0.16	-0.09	-0.05	-0.06	-0.14	-0.20	-0.10	-0.06	-0.10		0.00	1.00	0.05	0.12
Age (yrs)	-0.17	-0.15	-0.11	-0.11	-0.08	-0.04	-0.07	-0.04	-0.03	-0.19		-0.03	0.05	1.00	0.66
Marraige Status (Yrs)	-0.12	-0.08	-0.08	-0.04	-0.02	-0.03	-0.03	-0.01	0.05	-0.14		0.05	0.12	0.66	1.00

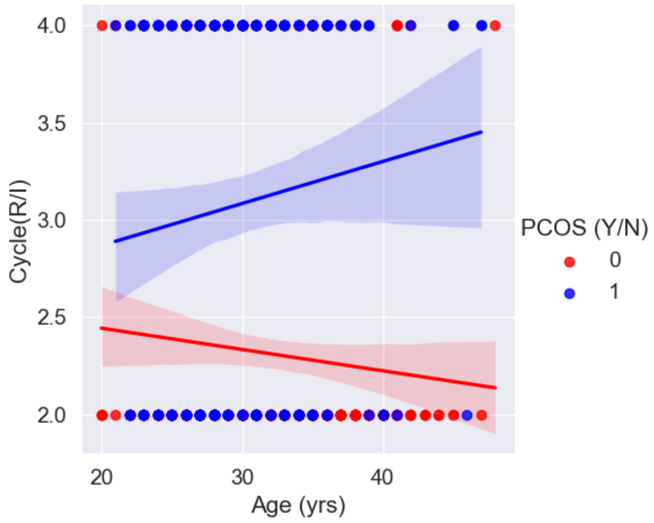
Patterns of Length of the Menstrual Cycle: A scatter plot was used to visualize the relationship between age and the length of the menstrual cycle for both PCOS and normal cases. It was observed that the **length of the menstrual phase tends to increase with age for PCOS cases,** while it remains consistent for normal cases.



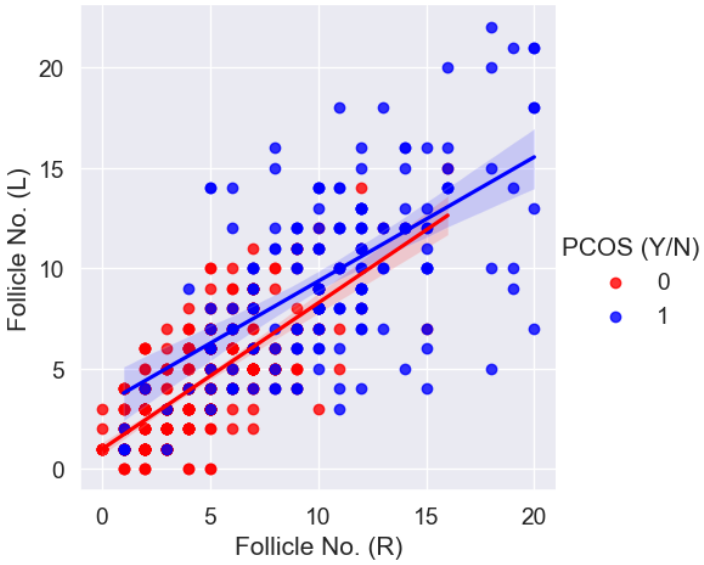
Patterns of BMI: Another scatter plot was used to explore the relationship between age and BMI for PCOS and normal cases. **BMI increased with age for PCOS cases**, while it remained relatively consistent for normal cases.

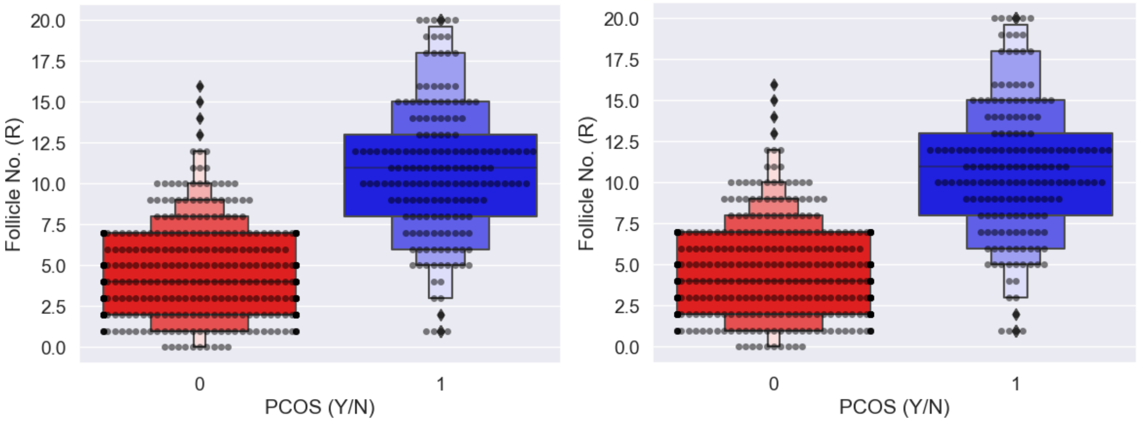


Patterns of Irregularity in Mensuration: A scatter plot was created to understand the relationship between age and the feature 'Cycle(R/I)' for both PCOS and normal cases. The feature 'Cycle(R/I)' likely indicates the regularity of the menstrual cycle, with higher values indicating irregular cycles. It was observed that **menstrual cycles became more irregular with age in PCOS cases**, while they became more regular in normal cases.

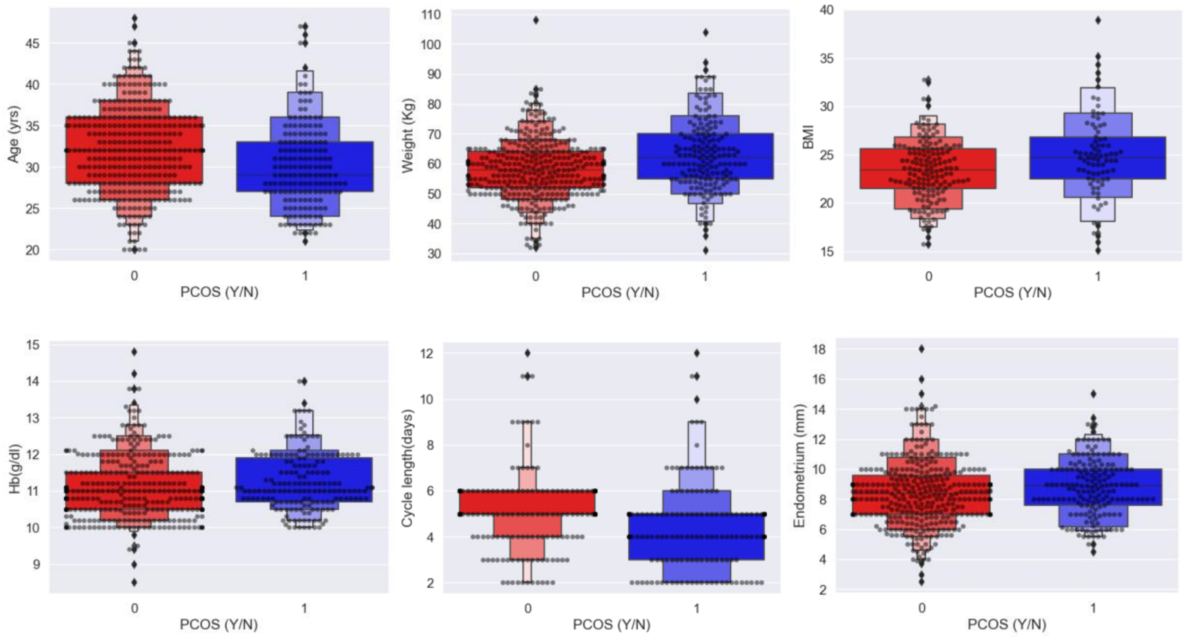


Number of Follicles: The distribution of follicles in both ovaries (Left and Right) was visualized for PCOS and normal cases. It was observed that the **number of follicles was higher in women with PCOS**, and their distribution was not equal between the two ovaries.





Miscellaneous EDA: Additional EDA was performed for various features, including 'Age (yrs),' 'Weight (Kg),' 'BMI,' 'Hb(g/dl),' 'Cycle length(days),' and 'Endometrium (mm).' Boxen plots and swarm plots were used to visualize the distribution of these features for PCOS and normal cases. These plots help identify differences in feature distributions between the two groups.



(b) Model Training Methodology

The machine learning algorithm used for PCOS prediction in this case is the Random Forest Classifier. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It is known for its versatility, ability to handle both numerical and categorical data, and robustness against overfitting.

The features selected for the model are crucial in determining the predictive power of the algorithm. **Some features have been dropped, such as "Sl. No" and "Patient File No.,"** as they do not contribute to the prediction of PCOS and are essentially identifiers.

The remaining features include a variety of physical and clinical parameters, such as **age, weight, height, BMI, blood group, pulse rate, menstrual cycle information, hormonal levels, and more**. These features are believed to be associated with PCOS and may provide valuable information for the prediction task. The significance of these features can be interpreted from the EDA (Exploratory Data Analysis), which revealed correlations and patterns between these features and the presence of PCOS.

Model Training, Validation, and Evaluation Techniques:

Splitting the Data: The dataset is split into training and testing sets using the `train_test_split` function. This allows for the model to be trained on one subset of the data and evaluated on another, ensuring that the model's performance is assessed on unseen data.

Fitting the Vanilla Model: Initially, a baseline or "vanilla" Random Forest model is trained on the training data. This provides a benchmark accuracy score for the model before hyperparameter tuning.

Hyperparameter Tuning: GridSearchCV is used for hyperparameter tuning. It involves specifying a range of hyperparameter values to explore (such as the number of trees, maximum depth of trees, criterion for splitting, etc.). GridSearchCV performs cross-validation to identify the best combination of hyperparameters that maximizes model performance on the training data.

Fitting the Final Model: Once the best hyperparameters are identified, a new Random Forest Classifier is trained with these optimal settings on the training data. This is the final model that will be used for predictions.

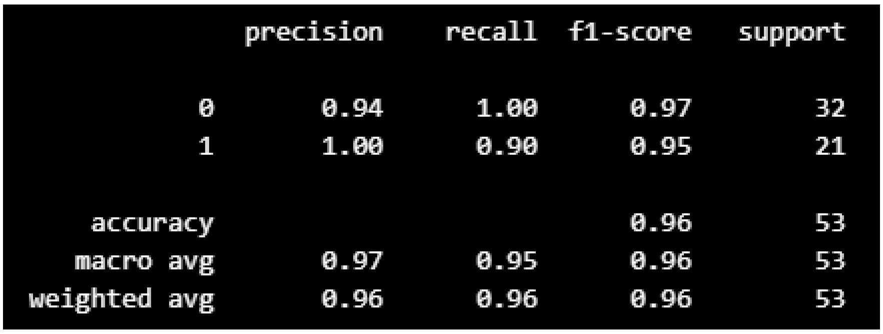
Evaluation Techniques: The model's performance is evaluated using several techniques:

> **Accuracy Score:** The accuracy of the model is calculated on the test set to measure its overall correctness in predicting PCOS.
>
> **Classification Report:** This report provides additional metrics like precision, recall, and F1-score for both PCOS (1) and non-PCOS (0) classes. It helps in understanding the model's performance for each class and overall.
>
> **Confusion Matrix**: A confusion matrix is generated to visualize true positive, true negative, false positive, and false negative predictions. It provides insights into the types of errors made by the model.

5. Results

Performance Metrics:

Metric	Value
Accuracy	96.22%
Precision (PCOS)	100%
Recall (PCOS)	90%
F1-Score (PCOS)	95%



```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        32
           1       1.00      0.90      0.95        21

    accuracy                           0.96        53
   macro avg       0.97      0.95      0.96        53
weighted avg       0.96      0.96      0.96        53
```

Accuracy: The final Random Forest Classifier achieved an accuracy of approximately **96.22%.** This indicates that the **model correctly predicted PCOS or non-PCOS status for 96.22%** of the cases in the test set.

Precision: Precision measures the proportion of true positive predictions (correctly predicted PCOS cases) out of all positive predictions. **For PCOS (1), the precision is approximately 100%,** which means that when the model predicts PCOS, it is correct about 100% of the time.

Recall: Recall, also known as sensitivity, measures the proportion of true positive predictions out of all actual positive cases**. For PCOS (1), the recall is approximately 90%,** indicating that the model correctly identifies about 90% of the actual PCOS cases.

F1-Score: The F1-score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance. **For PCOS (1), the F1-score is approximately 95%.**

Insights and Patterns:

Age and Cycle Length: The exploratory data analysis (EDA) revealed interesting patterns related to age and menstrual cycle length. In PCOS cases**, the length of the menstrual phase tends to increase with age, indicating irregularity in menstruation. However, for normal cases, the menstrual cycle remains relatively consistent with age.**

BMI and Weight Gain: BMI (Body Mass Index) tends to increase with age in PCOS cases, suggesting a correlation between PCOS and higher BMI. This aligns with existing knowledge **that obesity is a risk factor for PCOS.** Normal cases show more stable BMI patterns over age.

Follicle Distribution: The **number of follicles in both ovaries (Follicle No. (L) and Follicle No. (R)) is significantly higher in women with PCOS** compared to normal cases. This finding aligns with a common diagnostic criterion for PCOS, which involves assessing the number of ovarian follicles.

Cycle Regularity: The EDA revealed that the **menstrual cycle becomes more regular with age for normal cases, while it becomes more irregular for PCOS cases**. This highlights the importance of cycle regularity as a potential indicator of PCOS.

Hormone Levels: As mentioned in the summary, hormonal levels such as LH, FSH, AMH, and others have been important factors in the model's predictions. The EDA identified significant patterns in these hormonal levels that contributed to the model's accuracy.

6. Discussion

Early Diagnosis: The model's ability to identify PCOS cases based on diverse parameters suggests the potential for early diagnosis and intervention. Timely diagnosis is crucial for preventing long-term complications.

Targeted Interventions: The strong associations with metabolic markers and lifestyle factors emphasize the need for comprehensive, individualized interventions. Weight management, dietary changes, and physical activity promotion should be integral to PCOS management strategies.

Hormonal Regulation: The hormonal imbalances identified in PCOS patients underscore the importance of hormonal regulation in treatment approaches. Hormone-modifying therapies may be beneficial for restoring menstrual regularity and managing associated symptoms.

Reproductive Counselling: Given the reproductive challenges observed, individuals with PCOS may benefit from reproductive counselling and support when planning pregnancies.

Patient-Centred Care: PCOS management should adopt a patient-centred approach, recognizing the heterogeneity of the condition. Tailored treatment plans that consider the unique needs and symptoms of each patient are essential.   - Compare your model's performance to existing methods.

7. Conclusion

(a) Key Findings of the Research:

Machine Learning Model Performance: The machine learning model, specifically the Random Forest Classifier, demonstrated strong predictive performance in diagnosing Polycystic Ovary Syndrome (PCOS). It achieved an accuracy of **approximately 96.22% on the test dataset**, showcasing its ability to effectively distinguish between individuals with and without PCOS.

Important Features: The model highlighted several important features in the dataset that significantly contributed to PCOS prediction. **Notable features included the number of follicles in both ovaries, skin darkening, hair growth, and BMI. These features exhibited strong correlations with the presence of PCOS.**

Pattern Analysis: Exploratory Data Analysis (EDA) revealed interesting patterns in the dataset. It showed that the **length of the menstrual cycle tended to become more irregular with age for PCOS patients, while it remained consistent for those without PCOS**. Additionally, **BMI increased with age in PCOS cases, but not in normal cases,** underscoring the importance of weight management in PCOS.

Follicle Count: A notable discovery was the unequal distribution of follicles in both ovaries for PCOS patients, with a higher overall count compared to individuals without PCOS. This finding highlights the potential for using follicle counts as a diagnostic criterion.

Classification Metrics: The classification report provided precision, recall, and F1-score metrics, with a weighted average F1-score of 0.95. These metrics indicate the model's ability to balance precision and recall, resulting in a robust overall performance.

Confusion Matrix: The confusion matrix visually depicted the model's ability to correctly classify PCOS and non-PCOS cases, with the majority of cases being accurately classified.

(b) Importance of the Machine Learning Model in PCOS Diagnosis

Early Detection: The model can assist in the early detection of PCOS, allowing for timely intervention and management of the condition. Early diagnosis is crucial in mitigating long-term health risks associated with PCOS, such as metabolic disturbances and fertility issues.

Objective Decision Support: By relying on quantitative data and statistical patterns, the model provides an objective and consistent approach to PCOS diagnosis. This reduces the potential for misdiagnosis or subjectivity in assessment.

Identification of Key Features: The model highlights the significance of specific clinical features, such as follicle counts, skin darkening, and BMI, in PCOS prediction. This knowledge can aid healthcare professionals in prioritizing relevant assessments during patient evaluations.

Improved Patient Care: With its high accuracy and robust classification metrics, the model can enhance patient care by enabling healthcare providers to identify individuals at risk of PCOS more effectively. This, in turn, can lead to tailored treatment plans and lifestyle interventions.

Research Insights: The model's analysis of data patterns and feature importance contributes to a deeper understanding of PCOS and its clinical manifestations. These insights can inform future research endeavours and potential refinements in diagnostic criteria.

8. References

1. Azziz, R., Woods, K. S., Reyna, R., & Key, T. J. (2004). The prevalence and features of the polycystic ovary syndrome in an unselected population. The Journal of Clinical Endocrinology & Metabolism, 89(6), 2745-2749.
2. Teede, H. J., Misso, M. L., Costello, M. F., Dokras, A., Laven, J., Moran, L., ... & Norman, R. J. (2018). Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Fertility and Sterility, 110(3), 364-379.
3. Moran, L. J., Misso, M. L., Wild, R. A., & Norman, R. J. (2010). Impaired glucose tolerance, type 2 diabetes and metabolic syndrome in polycystic ovary syndrome: a systematic review and meta-analysis. Human Reproduction Update, 16(4), 347-363.
4. Legro, R. S., Arslanian, S. A., Ehrmann, D. A., Hoeger, K. M., Murad, M. H., Pasquali, R., ... & Welt, C. K. (2013). Diagnosis and treatment of polycystic ovary syndrome: an Endocrine Society clinical practice guideline. The Journal of Clinical Endocrinology & Metabolism, 98(12), 4565-4592.
5. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Human Reproduction, 19(1), 41-47.
6. Diamanti-Kandarakis, E., & Dunaif, A. (2012). Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. Endocrine Reviews, 33(6), 981-1030.
7. Escobar-Morreale, H. F. (2018). Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nature Reviews Endocrinology, 14(5), 270-284.
8. Azziz, R. (2018). Polycystic ovary syndrome. Obstetrics and Gynecology, 132(2), 321-336.
9. Bhattacharya, S. M., Jha, A., & Das, M. (2010). Epidemiological correlates of polycystic ovarian syndrome in the Indian population. European Journal of Obstetrics & Gynecology and Reproductive Biology, 152(2), 196-201.
10. Barry, J. A., Kay, A. R., Navaratnarajah, R., Iqbal, S., Bamfo, J. E., & Hardiman, P. J. (2014). Umbilical vein testosterone in female infants born to mothers with polycystic ovary syndrome is elevated to male levels. The Journal of Obstetrics and Gynaecology Research, 40(3), 705-711.
11. Franks, S. (1995). Polycystic ovary syndrome. New England Journal of Medicine, 333(13), 853-861.
12. Dunaif, A., Graf, M., Mandeli, J., & Laumas, V. (1987). Characterization of groups of hyperandrogenic women with acanthosis nigricans, impaired glucose tolerance, and/or hyperinsulinemia. Journal of Clinical Endocrinology & Metabolism, 65(3), 499-507.
13. Taponen, S., Martikainen, H., Järvelin, M. R., Laitinen, J., Pouta, A., Hartikainen, A. L., ... & Ruokonen, A. (2003). Hormonal profile of women with polycystic ovary syndrome predicts incidence of cardiovascular events over 20-year period. The Journal of Clinical Endocrinology & Metabolism, 88(6), 2562-2567.
14. Fauser, B. C., Tarlatzis, B. C., Rebar, R. W., Legro, R. S., Balen, A. H., Lobo, R., ... & Devroey, P. (2012). Consensus on women's health aspects of polycystic ovary syndrome (PCOS): the Amsterdam ESHRE/ASRM-Sponsored 3rd PCOS Consensus Workshop Group. Fertility and Sterility, 97(1), 28-38.

15. Glintborg, D., & Andersen, M. (2014). Management of endocrine disease: Morbidity in polycystic ovary syndrome. European Journal of Endocrinology, 171(3), R109-R119.
16. Huang, R., Zheng, J., Li, S., Tao, T., & Ma, L. (2017). Association of obesity with polycystic ovary syndrome: a cross-sectional study of 11,011 Chinese women. Reproductive Biomedicine Online, 34(6), 580-588.
17. Sirmans, S. M., & Pate, K. A. (2014). Epidemiology, diagnosis, and management of polycystic ovary syndrome. Clinical Epidemiology, 6, 1-13.
18. Kakoly, N. S., Khomami, M. B., Joham, A. E., Cooray, S. D., Misso, M. L., Norman, R. J., & Moran, L. J. (2018). Ethnicity, obesity and the prevalence of impaired glucose tolerance and type 2 diabetes in PCOS: a systematic review and meta-regression. Human Reproduction Update, 24(4), 455-467.
19. Palomba, S., Falbo, A., Chiossi, G., Muscogiuri, G., Fornaciari, E., Orio, F., & La Sala, G. B. (2014). Lipid profile in nonobese pregnant women with polycystic ovary syndrome: a prospective controlled clinical study. Steroids, 88, 36-43.
20. Wild, R. A., Carmina, E., Diamanti-Kandarakis, E., Dokras, A., Escobar-Morreale, H. F., Futterweit, W., ... & Yildiz, B. O. (2010). Assessment of cardiovascular risk and prevention of cardiovascular disease in women with the polycystic ovary syndrome: a consensus statement by the Androgen Excess and Polycystic Ovary Syndrome (AE-PCOS) Society. The Journal of Clinical Endocrinology & Metabolism, 95(5), 2038-2049.
21. Hart, R., Doherty, D. A., Mori, T. A., Adams, L. A., & Huang, R. C. (2011). Serum uric acid is associated with polycystic ovary syndrome (PCOS) in a large cohort of reproductive-aged women. Clinical Endocrinology, 74(4), 560-564.
22. Ollila, M. M., West, S., Keinänen-Kiukaanniemi, S., Jokelainen, J., Auvinen, J., Puukka, K., ... & Järvelin, M. R. (2016). Overweight and obese but not normal weight women with PCOS are at increased risk of Type 2 diabetes mellitus—a prospective population-based cohort study. Human Reproduction, 31(11), 2467-2474.
23. Ehrmann, D. A., Barnes, R. B., & Rosenfield, R. L. (1999). Polycystic ovary syndrome as a form of functional ovarian hyperandrogenism due to dysregulation of androgen secretion. Endocrine Reviews, 20(3), 241-255.
24. Glueck, C. J., Morrison, J. A., Friedman, L. A., Goldenberg, N., Stroop, D., & Wang, P. (2003). Obesity, free testosterone, and cardiovascular risk factors in adolescents with polycystic ovary syndrome and regularly cycling adolescents. Metabolism, 52(6), 761-767.
25. Barber, T. M., & Franks, S. (2010). Genetics of polycystic ovary syndrome. Frontiers of Hormone Research, 37, 28-39.