DATA-DRIVEN SOLUTIONS FOR PCOS: EXPLORING MACHINE LEARNING IN DIAGNOSTIC PROCESSES

Raaga Likhitha Musunuri

AGENDA OVERVIEW

01

BACKGROUND OF THE STUDY

02

PROBLEM STATEMENT- MOTIVATION AND RESEARCH QUESTIONS

03

ANALYSIS

04

STATISTICAL TESTS

05

ML MODELS

06

ML MODELS

07

DISCUSSION

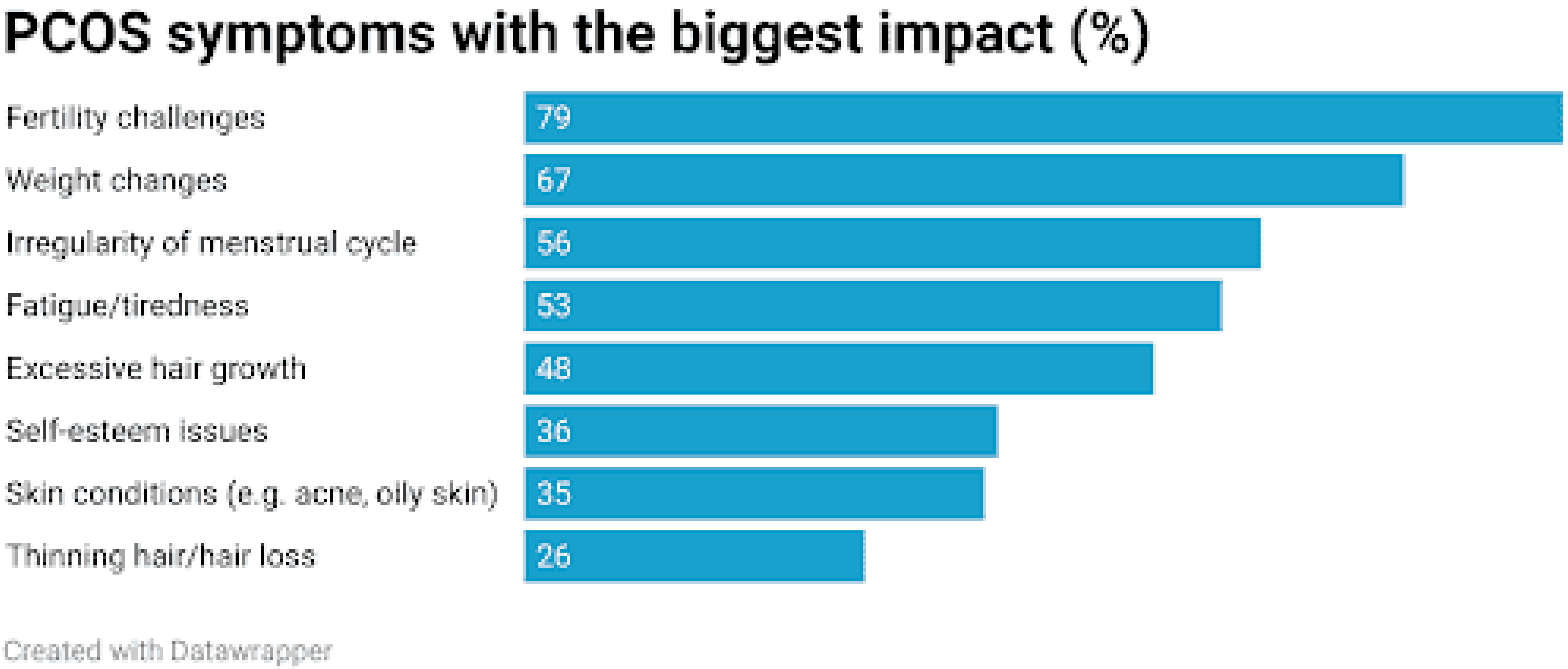
08

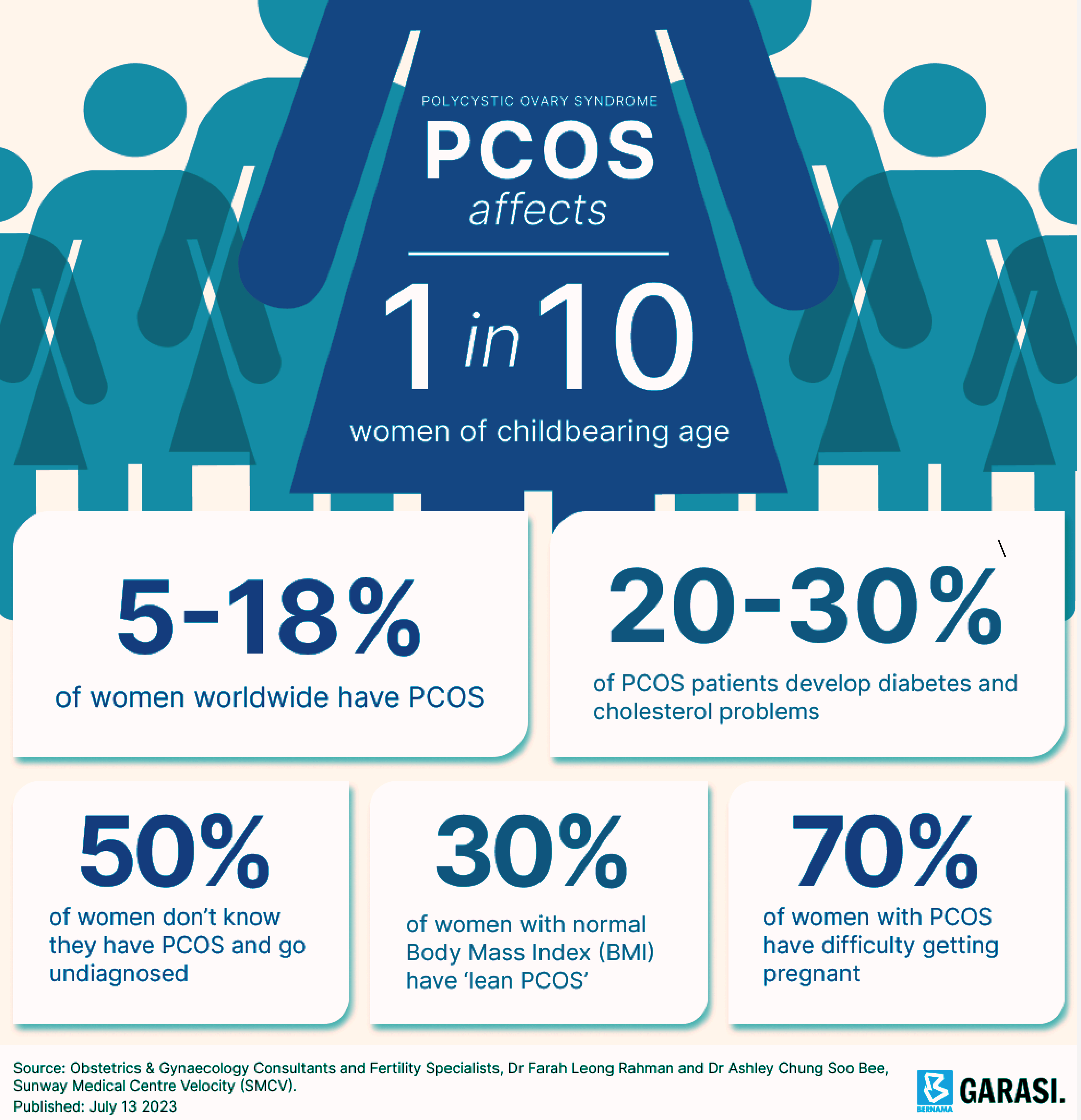
FUTURE PLAN OF ACTION

But first, what is PCOS?

Polycystic ovary syndrome, commonly known as PCOS, often comes with a high number of small follicles, resembling cysts, situated just beneath the ovary’s surface. These underdeveloped follicles contain eggs.

The syndrome is associated with various health risks including infertility, metabolic syndrome, type 2 diabetes, cardiovascular disease, and mental health issues(Singh S et al., 2023, Rasquin L.I et. al., 2022).





BACKGROUND

- Despite its prevalence, PCOS remains underdiagnosed, with up to 70% of affected women undiagnosed globally(WHO, 2023).
- 2 in 5 people (39%) received no treatment after their initial consultation
- Over 2 in 3 people waited over a year for their PCOS diagnosis
- 1 in 10 people waited over 5 years for treatment, with 7% waiting over 10 years
- Whilst PCOS impacts up to 10% of women of childbearing age, it's a condition that often goes undiagnosed. This is usually because the symptoms can be indicators of a variety of different conditions and can vary dramatically from person to person.

Motivation

This study focuses on the early diagnostics of PCOS by understanding different symptoms, and their interactions with one another. Through this research, we aim to develop a robust, accurate, and interpretable tool for early PCOS detection and management, improving women's health outcomes worldwide.

Enhanced Pattern Recognition

Potential for earlier and more accurate diagnoses, enabling timely interventions

Personalized Medicine

Accounts for individual variations in symptoms and risk factors

Addressing Healthcare Disparities

Automated screening tools provide consistent, objective assessments

Integration with EHR Systems

Provides clinicians with valuable decision support tools

Cost-Effective Screening

Offers potential for large-scale, cost-effective PCOS screening

Global Health implications

Advancements in Diagnostic Methods

RESEARCH QUESTIONS

This project aims to develop accurate machine-learning models for PCOS prediction using health data, potentially surpassing traditional diagnostics **(ROC AUC 73-100% in previous studies)**.

Operating efficiency

By identifying key predictive features and comparing algorithms, it seeks to enhance PCOS understanding, inform interventions, and integrate AI into EHR systems, leading to earlier diagnoses and improved patient care.

Productivity

How accurately can machine learning models predict PCOS based on commonly available health data?

Optimization

Which features (symptoms, health parameters) are most predictive of PCOS?

Streamlining

How do different machine learning algorithms (Naive Bayes, Logistic Regression, etc.) compare in their predictive performance for PCOS?

Performance Metrics

Can machine learning models identify subtypes or clusters within PCOS patients that might inform personalized treatment approaches?

Automation

How does the performance of machine learning models compare to traditional diagnostic criteria for PCOS?

ANALYSIS

This framework provides a structured approach to analyzing PCOS prediction using machine learning, from data preparation to model selection and interpretation that has been used in the project

01

DATA PREPARATION AND EXPLORATION

Data set Overview
Data Preprocessing
Exploratory Data Analysis

02

STATISTICAL ANALYSIS

Chi-Square Tests
Lifestyle Factor Analysis

03

MACHINE LEARNING MODELS

Logistic Regression
SVM
Naives Bayes
Decision Trees

04

MODEL REFINEMENT AND INTERPRETATION

Feature selection
Model Retraining
Evaluate performance improvements
Final Model Selection

05

CONCLUSIONS AND IMPLICATIONS

Summarize key findings
Discuss potential applications in clinical settings
Address limitations and areas for future research

DATASET

The dataset contains information from 267 individuals. It was collected through a survey conducted primarily in Mumbai and surrounding areas in India. The dataset includes both PCOS-positive and PCOS-negative cases, providing a comprehensive view of factors potentially associated with PCOS.

Key Variables

Target Variable: PCOS (Yes/No); Age: Categorized into groups (e.g., 18-25, 26-30, 31-35); Symptoms: Various binary indicators for PCOS-related symptoms; Lifestyle factors

Geographical Focus

Primarily centered on Mumbai, with some participants from other Indian cities

Age Distribution

Includes a wide range of age groups, from below 18 to above 45

Unique Aspects

- Includes both physical symptoms and lifestyle factors
- Captures information on the age of PCOS onset for diagnosed cases

Sample Size

267 entries

Features

Menstrual cycle characteristics; Demographic information; PCOS diagnosis and onset age; Physical symptoms; Lifestyle factors - 25 columns

Data Types

- 22 columns are categorical
- 2 columns are numerical- exercise per week, eat outside per week
- 1 column is float - PCOS_from (age of PCOS onset)

Missing Data

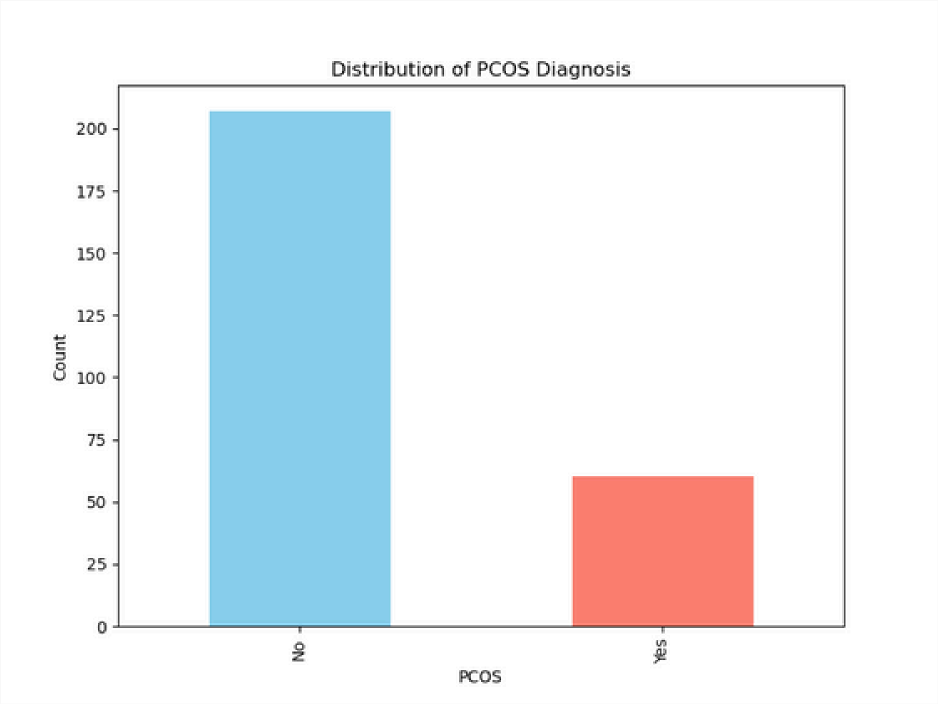
- PCOS_from: Only 55 non-null values (likely because it's only applicable to PCOS-positive cases)
- Overweight: 153 non-null values

Data Preprocessing:

- Removed unnecessary columns (Timestamp, PCOS tested, mood swing information)
- Standardized city names to lowercase
- Renamed 'PCOS from age of' to 'PCOS_from'
- Extracted numeric values from the 'PCOS_from' column
- Handled missing values in the 'PCOS_from' column

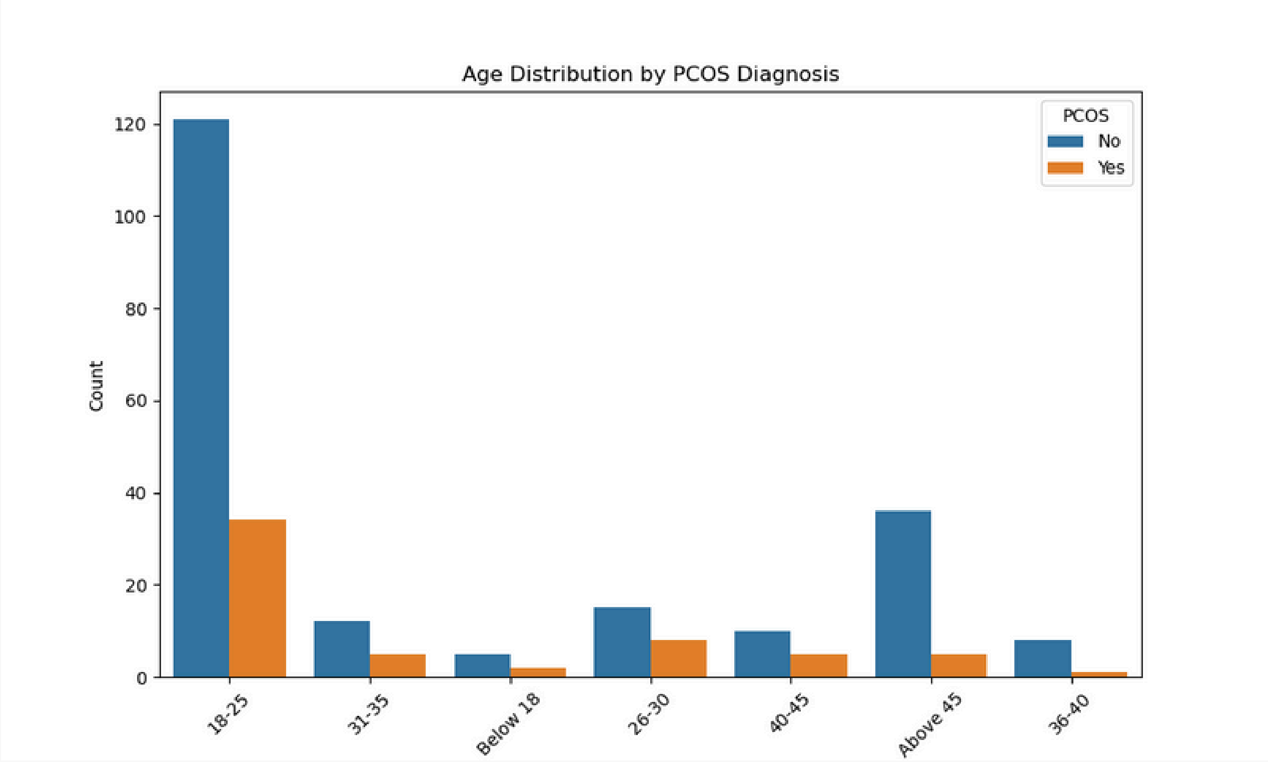
Exploratory data analysis

PCOS Prevalence



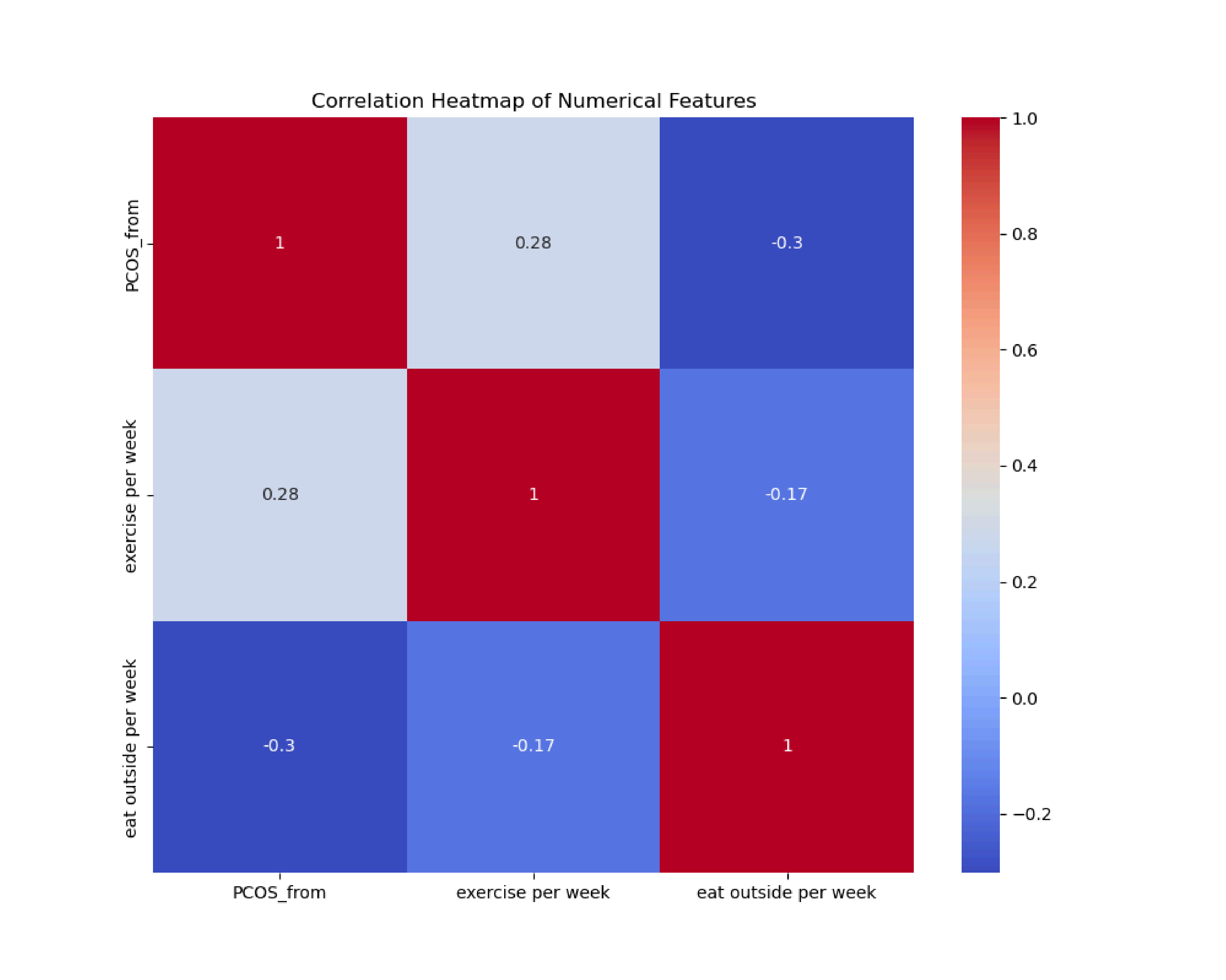
- This chart suggests that within the given dataset, there is a significantly higher number of individuals without PCOS compared to those diagnosed with it.
- The disparity in the counts may reflect the prevalence of PCOS within the studied population or sample, with fewer individuals diagnosed compared to those without the condition.

Age Distribution



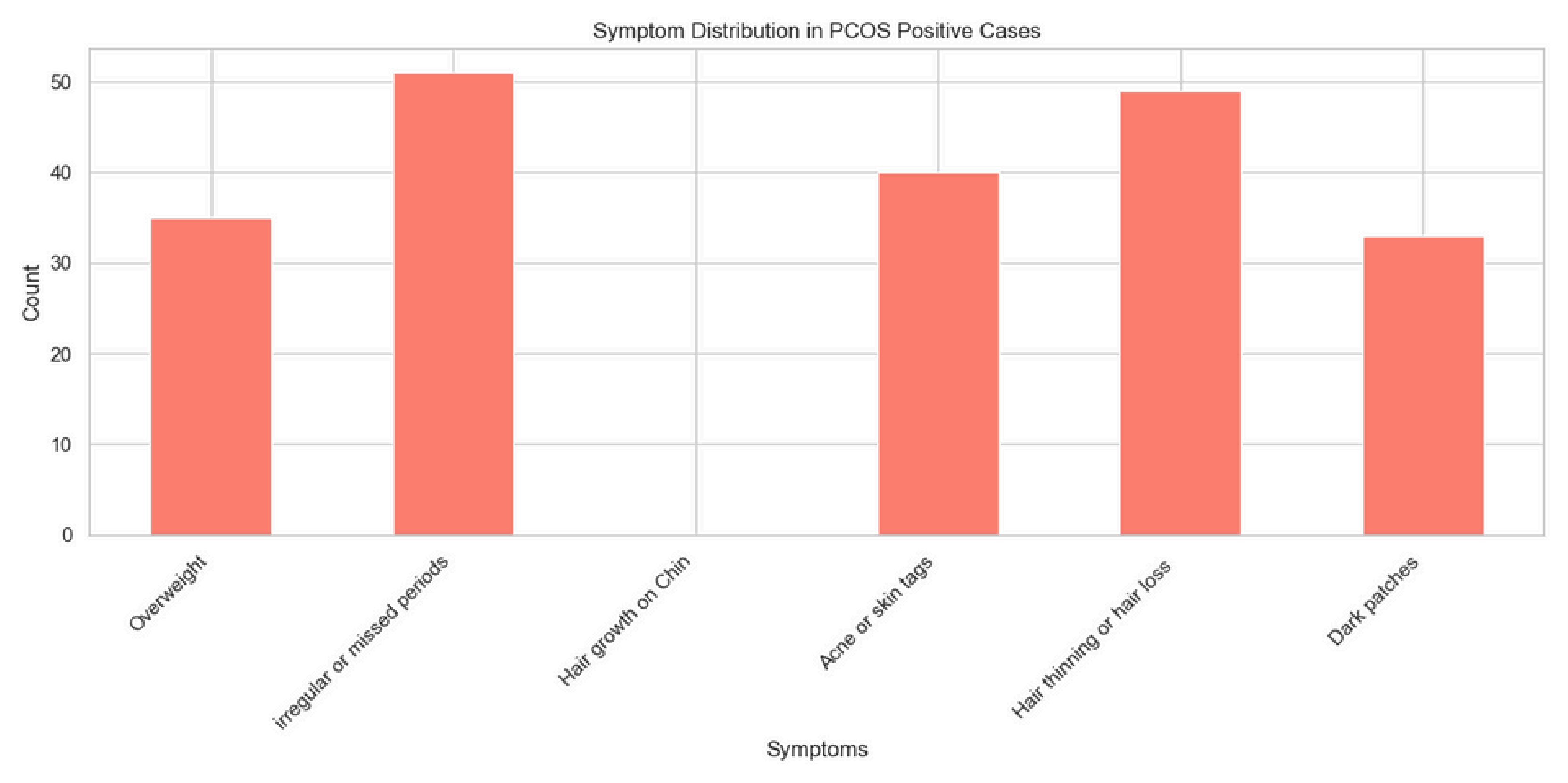
- The 18-25 age group has the highest count of individuals, with a significant difference between those without PCOS (over 120) and those with PCOS (around 40).
- The population above 45 has a notable count of individuals without PCOS but a very low count with PCOS.
- Age groups "Below 18" and "36-40" show a minimal number of individuals with and without PCOS.

Symptom Analysis



Between three variables: "PCOS_from" (likely indicating the diagnosis or duration of PCOS), "exercise per week," and "eat outside per week." Each cell in the heatmap shows the correlation coefficient between two variables, ranging from -1 to 1. A value close to 1 indicates a strong positive correlation, close to -1 indicates a strong negative correlation, and values around 0 indicate no correlation.

- "PCOS_from" and "exercise per week" have a mild positive correlation (0.28).
- "PCOS_from" and "eat outside per week" have a moderate negative correlation (-0.3), suggesting that eating out may be inversely related to PCOS diagnosis or severity.
- "exercise per week" and "eat outside per week" have a weak negative correlation (-0.17).

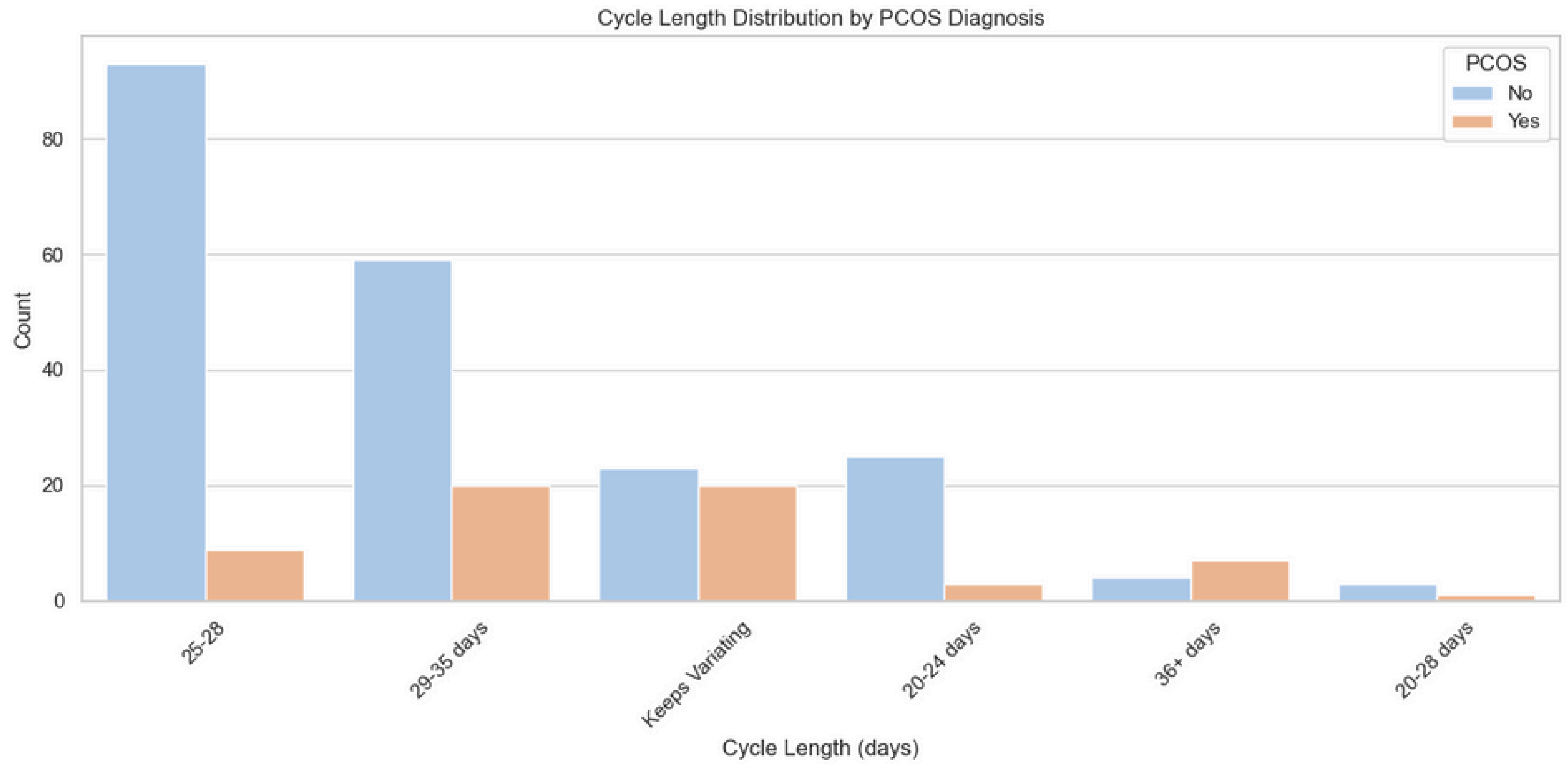


Symptom Distribution in PCOS Positive Cases

"Irregular or missed periods" and "Hair thinning or hair loss" are the most common symptoms among individuals with PCOS, with counts above 50. "Overweight" and "Dark patches" are less common but still significant symptoms.

Cycle Length Analysis

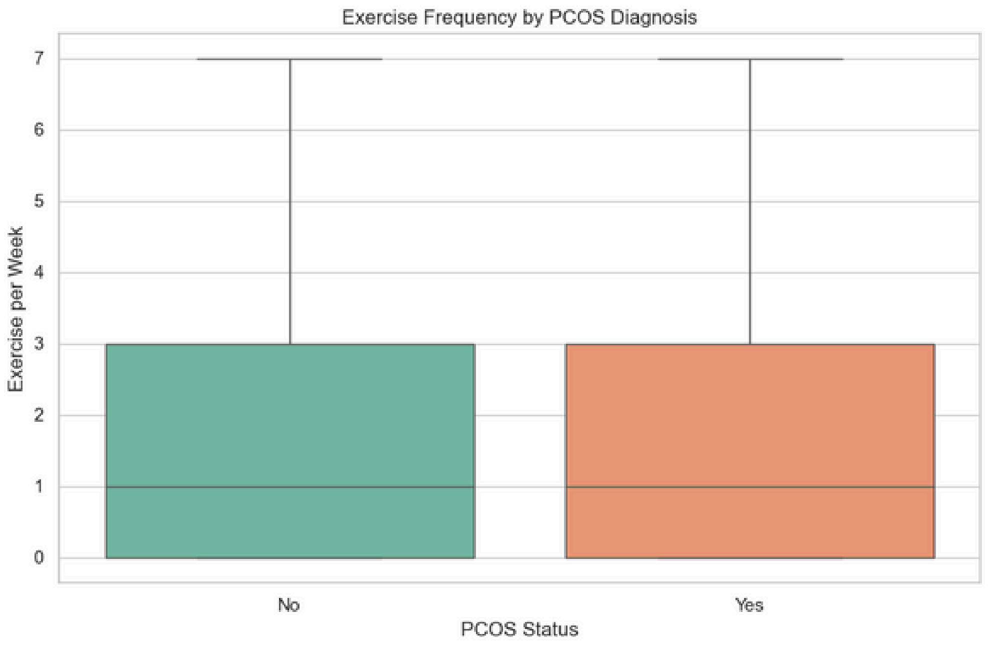
Most individuals without PCOS fall within the regular 25-28 day range, while those with PCOS are more likely to experience irregular or extended cycles, as shown by the higher counts in categories like "Keeps Varying" and "36+ days." This aligns with the known association between PCOS and irregular menstrual cycles.



STATISTICAL ANALYSIS

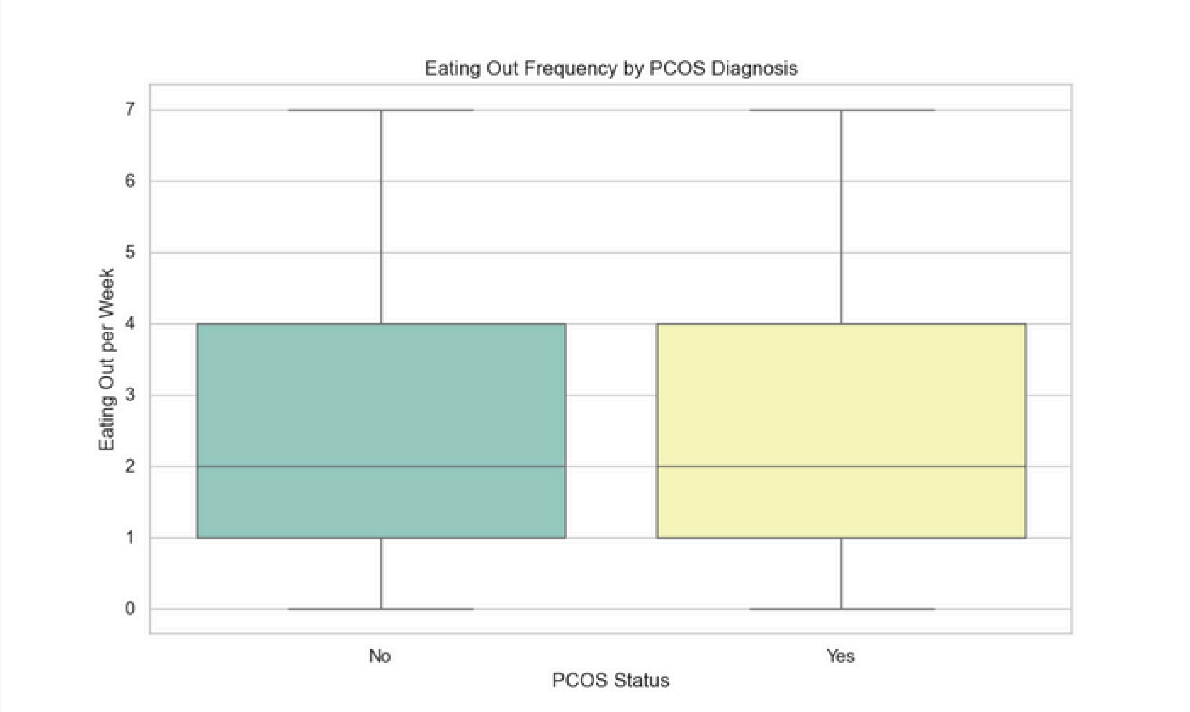
T-Test for Lifestyle Factors

Exercise frequency



H0: There is no significant difference in exercise frequency between PCOS-positive and PCOS-negative groups.
HA: There is a significant difference in exercise frequency between PCOS-positive and PCOS-negative groups.
p-value: 0.625
Fail to reject the null hypothesis. There is no statistically significant difference in exercise frequency between PCOS-positive and PCOS-negative groups ($p > 0.05$).

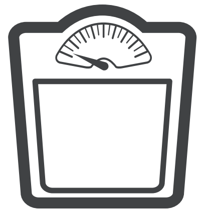
Eating Out frequency



H0: There is no significant difference in eating-out habits between PCOS-positive and PCOS-negative groups.
HA: There is a significant difference in eating-out habits between PCOS-positive and PCOS-negative groups.R
p-value: 0.669
Fail to reject the null hypothesis. There is no statistically significant difference in eating-out habits between PCOS-positive and PCOS-negative groups ($p > 0.05$).

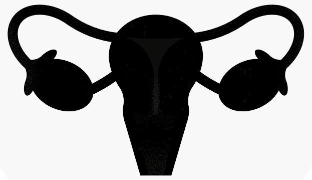
Chi-Square Tests for Symptoms

Overweight



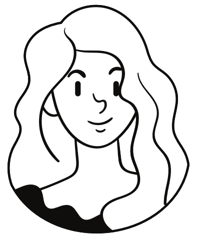
H0: There is no association between being overweight and PCOS.
HA: There is an association between being overweight and PCOS.Results:
Chi-square statistic: 35.21
p-value: 2.26e-08
Reject the null hypothesis. There is a statistically significant association between being overweight and PCOS ($p < 0.05$).

Irregular or Missed Periods



H0: There is no association between irregular or missed periods and PCOS.
HA: There is an association between irregular or missed periods and PCOS.Results:
Chi-square statistic: 42.08
p-value: 7.28e-10
Reject the null hypothesis. There is a statistically significant association between irregular or missed periods and PCOS ($p < 0.05$).

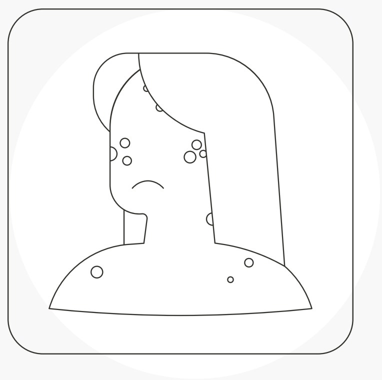
Hair Growth on Chin



H0: There is no association between hair growth on chin and PCOS.
HA: There is an association between hair growth on chin and PCOS.Results:
Chi-square statistic: 22.83
p-value: 1.10e-05
Reject the null hypothesis. There is a statistically significant association between hair growth on chin and PCOS ($p < 0.05$).

Chi-Square Tests for Symptoms

Acne or Skin Tags



H0: There is no association between acne or skin tags and PCOS.
HA: There is an association between acne or skin tags and PCOS.
Chi-square statistic: 9.51
p-value: 0.002
Reject the null hypothesis. There is a statistically significant association between acne or skin tags and PCOS ($p < 0.05$).

Hair Thinning or Hair Loss

H0: There is no association between hair thinning or hair loss and PCOS.
HA: There is an association between hair thinning or hair loss and PCOS.
Chi-square statistic: 2.60
p-value: 0.107
Fail to reject the null hypothesis. There is no statistically significant association between hair thinning or hair loss and PCOS ($p > 0.05$).

Dark Patches

H0: There is no association between dark patches and PCOS.
HA: There is an association between dark patches and PCOS.
Chi-square statistic: 7.53
p-value: 0.006
Reject the null hypothesis. There is a statistically significant association between dark patches and PCOS ($p < 0.05$).

MACHINE LEARNING APPROACH

SCOPE

Prepare the data for machine learning models: Encode categorical variables

IMPLEMENT MODELS

Logistic regression, SVM, Naives bayes, Decision trees

EVALUATE

Evaluate model performance using metrics like accuracy, precision, recall, and F1-score

FEATURE IMPORTANCE

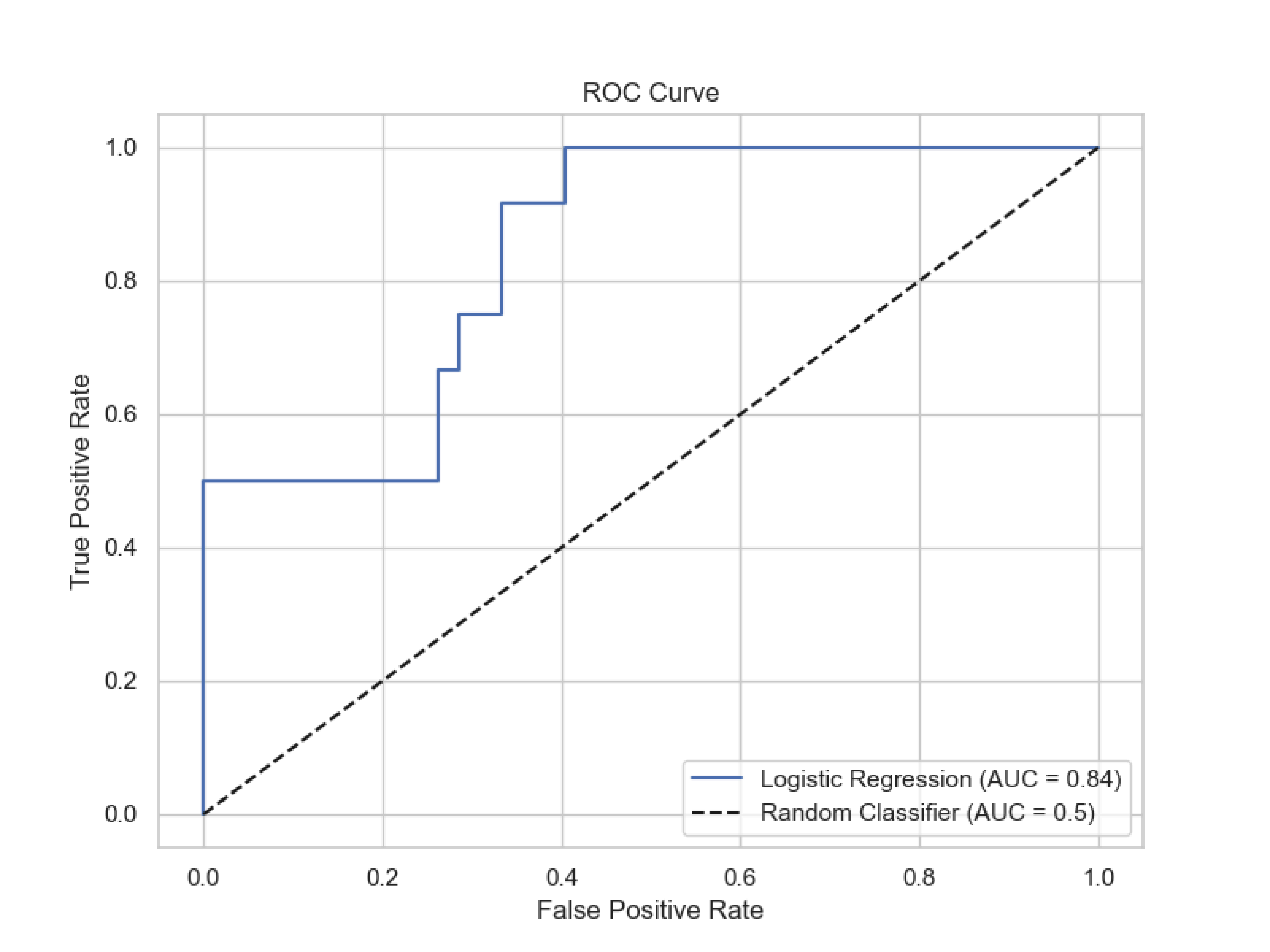
Analyze which features are most predictive of PCOS

Logistic regression

```python
# Handle missing values by filling or dropping (depending on your approach)
# Fill numeric columns with the median and categorical with the mode
data['PCOS_from'] = data['PCOS_from'].fillna(data['PCOS_from'].median())
data['Overweight'] = data['Overweight'].fillna(data['Overweight'].mode()[0])


# Convert categorical variables to numerical format
label_encoders = {}
for column in data.select_dtypes(include='object').columns:
    le = LabelEncoder()
    data[column] = le.fit_transform(data[column])
    label_encoders[column] = le  # store the encoder for each column

# Define features (X) and target (y)
X = data.drop(columns=['PCOS'])  # Drop the target column from features
y = data['PCOS']  # Target column

# Standardize the features
scaler = StandardScaler()
X = scaler.fit_transform(X)

# Add constant for statsmodels logistic regression
X_sm = sm.add_constant(X)

# Logistic regression using statsmodels for feature significance
logit_model = sm.Logit(y, X_sm)
result = logit_model.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.313682
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                   PCOS   No. Observations:                  267
Model:                          Logit   Df Residuals:                      242
Method:                           MLE   Df Model:                           24
Date:                Sun, 01 Dec 2024   Pseudo R-squ.:                  0.4113
Time:                        15:03:28   Log-Likelihood:                -83.753
converged:                       True   LL-Null:                       -142.26
Covariance Type:            nonrobust   LLR p-value:                 3.286e-14
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.1384      0.286     -7.482      0.000      -2.699      -1.578
x1             0.2665      0.204      1.305      0.192      -0.134       0.667
x2             0.2825      0.211      1.336      0.181      -0.132       0.697
x3            -0.0885      0.256     -0.346      0.729      -0.589       0.412
x4             0.0016      0.228      0.007      0.995      -0.446       0.449
x5            -0.0736      0.159     -0.464      0.643      -0.384       0.237
x6             0.8168      0.215      3.791      0.000       0.395       1.239
x7             0.5573      0.249      2.234      0.025       0.068       1.046
x8             0.7646      0.257      2.979      0.003       0.262       1.268
x9             0.2122      0.240      0.884      0.377      -0.258       0.683
x10           -0.5395      0.282     -1.911      0.056      -1.093       0.014
x11            0.4899      0.286      1.711      0.087      -0.071       1.051
x12           -0.1942      0.227     -0.856      0.392      -0.639       0.251
x13           -0.0744      0.273     -0.273      0.785      -0.609       0.460
x14            0.4561      0.314      1.451      0.147      -0.160       1.072
x15           -0.2942      0.296     -0.994      0.320      -0.874       0.286
x16            0.3838      0.224      1.711      0.087      -0.056       0.823
x17            0.0233      0.234      0.100      0.921      -0.435       0.481
x18           -0.0110      0.221     -0.050      0.960      -0.443       0.421
x19            0.4577      0.238      1.919      0.055      -0.010       0.925
x20           -0.0197      0.242     -0.081      0.935      -0.495       0.455
x21            0.0804      0.237      0.338      0.735      -0.385       0.546
x22           -0.1832      0.240     -0.764      0.445      -0.653       0.287
x23           -0.1425      0.245     -0.581      0.561      -0.623       0.338
x24           -0.2017      0.206     -0.981      0.327      -0.605       0.201
==============================================================================
```

the Significant Predictors were The coefficients with p-values less than 0.05.x6 (coef = 0.8168, p < 0.001), x7 (coef = 0.5573, p = 0.025) and x8 (coef = 0.7646, p = 0.003) are the features have a statistically significant association with the likelihood of PCOS,
suggesting they contribute meaningfully to the model.



Model Setup:
Features: All columns except PCOS
Target: PCOS (Yes/No)
Data split: 80% training, 20% testing
Feature standardization applied
Model Performance:
AUC (Area Under ROC Curve): 0.83
Indicates strong ability to distinguish between PCOS and non-PCOS cases
ROC Curve Analysis:
Clear separation between true positive and false positive rates
Model curve well above random classifier line (0.5 AUC)

Most impactful features: Overweight, Always tired, Irregular/missed periods

Other significant features: Cycle length, Hair growth in arms

Positive SHAP values indicate higher likelihood of PCOS diagnosis

Model Effectiveness:

Good classification power with low false positive/negative rates

Solid foundation for PCOS prediction, with potential for further refinement

SMOTE (Synthetic Minority Over-sampling Technique)

```
PCOS
0    165
1     48
Name: count, dtype: int64
PCOS
0    165
1    165
Name: count, dtype: int64
```

```
## based on the SHAP insights,  I want to retain the top impactful features such as Overweight,
##always tired,irregular or missed periods, Cycle Length, Hair growth on Chin
## and Remove less impactful features like more Mood Swings, canned food often, relocated city
```

I retrained the model, which gave me performance metrics of Performance Metrics such as

accuracy: 85.2% - The refined model predicts PCOS cases correctly for 85% of the test set.

Precision: 83.3% - When the model predicts PCOS, it is correct 83.3% of the time.

Recall: 41.7% - The model captures 41.7% of actual PCOS cases in the test set.

F1-Score: 55.6% - Balances precision and recall, but shows moderate performance due to lower recall.

AUC (ROC): 81.2%

	Metric	Logistic Regression	SVM	Naive Bayes	Decision Tree
0	Accuracy	0.851852	0.814815	0.870370	0.833333
1	Precision	0.833333	0.625000	0.727273	0.636364
2	Recall	0.416667	0.416667	0.666667	0.583333
3	F1-Score	0.555556	0.500000	0.695652	0.608696
4	ROC-AUC	0.811508	0.575397	0.847222	0.716270

```
##Decision Tree offers a good balance between precision(63.6%) and recall(58.3%) but doesn't outperform Naive Bayes
```

Model Performance Overview

Naive Bayes: Best overall performer

Highest accuracy (87.0%)

Highest ROC-AUC (84.7%)

Strong recall (66.7%)

Logistic Regression: Strong all-around performance

High precision (83.3%)

Robust accuracy (85.2%)

Good ROC-AUC (81.2%)

Decision Tree: Balanced performance

Good precision (63.6%)

Decent recall (58.3%)

SVM: Underperformed compared to other models

Lowest precision (62.5%)

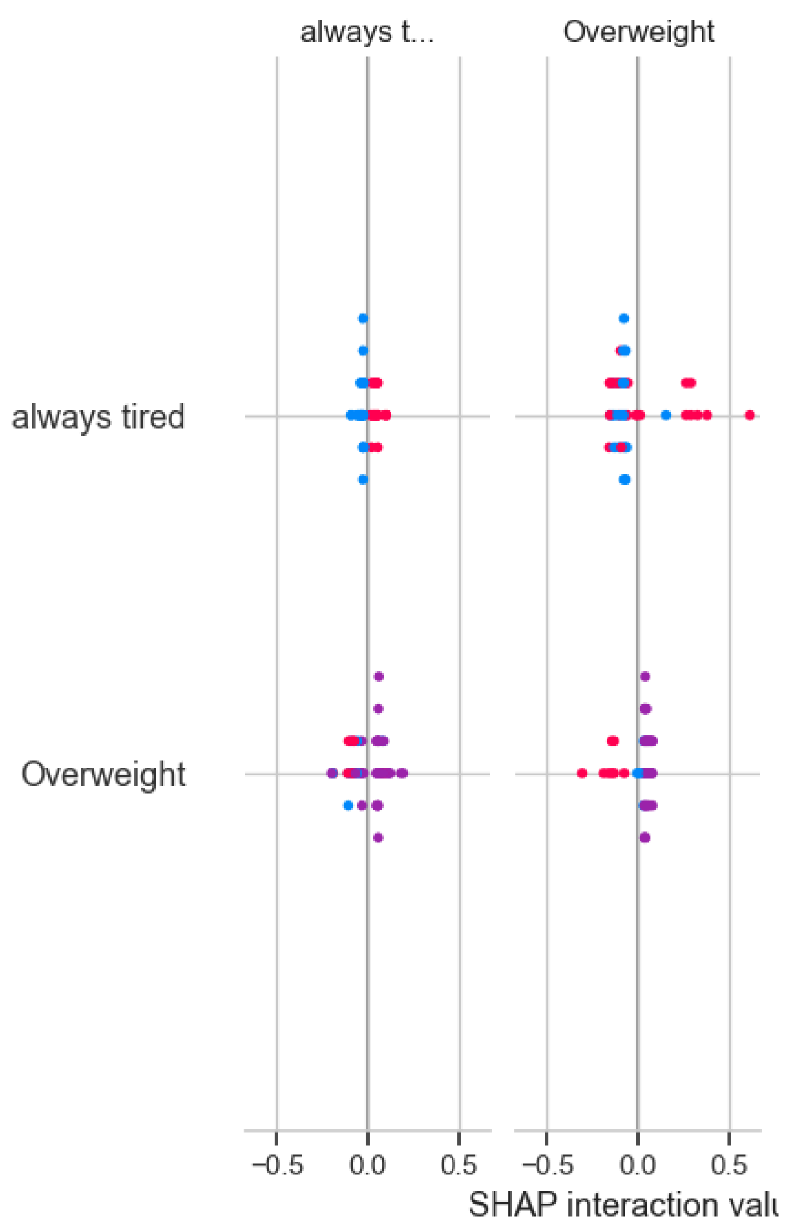
Lowest ROC-AUC (57.5%)

```python
## feature engineering
## to do this I am beginning with creation of composite features and exploring potential interaction terms
# Create a composite feature: Symptom Severity Score
##Symptom_Severity_Score represents the cumulative severity of symptoms like being overweight, irregular periods, hair growth, acne, and dark patches.
##Higher scores indicate greater symptom severity and potentially stronger associations with PCOS.
##Overweight_Cycle_Interaction Captures the interaction between being overweight and cycle length,
##which could provide insights into combined effects on PCOS diagnosis.
data['Symptom_Severity_Score'] = (
    data['Overweight'] +
    data['irregular or missed periods'] +
    data['Hair growth on Chin'] +
    data['Acne or skin tags'] +
    data['Dark patches']
)

# Add an interaction feature: Overweight × Cycle Length
data['Overweight_Cycle_Interaction'] = data['Overweight'] * data['Cycle Length']

# Display the first few rows to check the new features
data[['Symptom_Severity_Score', 'Overweight_Cycle_Interaction']].head()
```

	Symptom_Severity_Score	Overweight_Cycle_Interaction
0	5	5
1	2	1
2	7	6
3	3	3
4	6	1

```
{'Accuracy': 0.8518518518518519,
 'Precision': 0.7,
 'Recall': 0.5833333333333334,
 'F1-Score': 0.6363636363636364,
 'ROC-AUC': 0.8670634920634921}
```

Naive Bayes Model Performance with New Features:
Accuracy: 85.2% (Unchanged)
Model maintains strong overall predictive power
Precision: 70.0% (Slight decrease)
Minor increase in false positives
Recall: 58.3% (Improved)
Model now captures more PCOS cases
F1-Score: 63.6% (Increased)
Better balance between precision and recall
ROC-AUC: 86.7% (Improved)
Enhanced class separation and prediction confidence
Key Insights:
Symptom_Severity_Score and Overweight_Cycle_Interaction features improved model performance
Trade-off: Slight precision decrease for improved recall
Prioritizing recall is beneficial for identifying more PCOS cases



<function summary_legacy at 0x0000022149FAB7E0>

Most Influential Features:
Overweight
Always tired
Irregular or missed periods
Hair growth on chin
Least Influential Features:
Mood swings
Canned food consumption
Relocation
Feature Impact:
Top features show wide spread of SHAP values, indicating significant impact on predictions
Less important features have minimal effect on model output
Implications:
Focus on key symptoms for PCOS diagnosis
Consider reducing emphasis on less impactful factors in future iterations

Insights

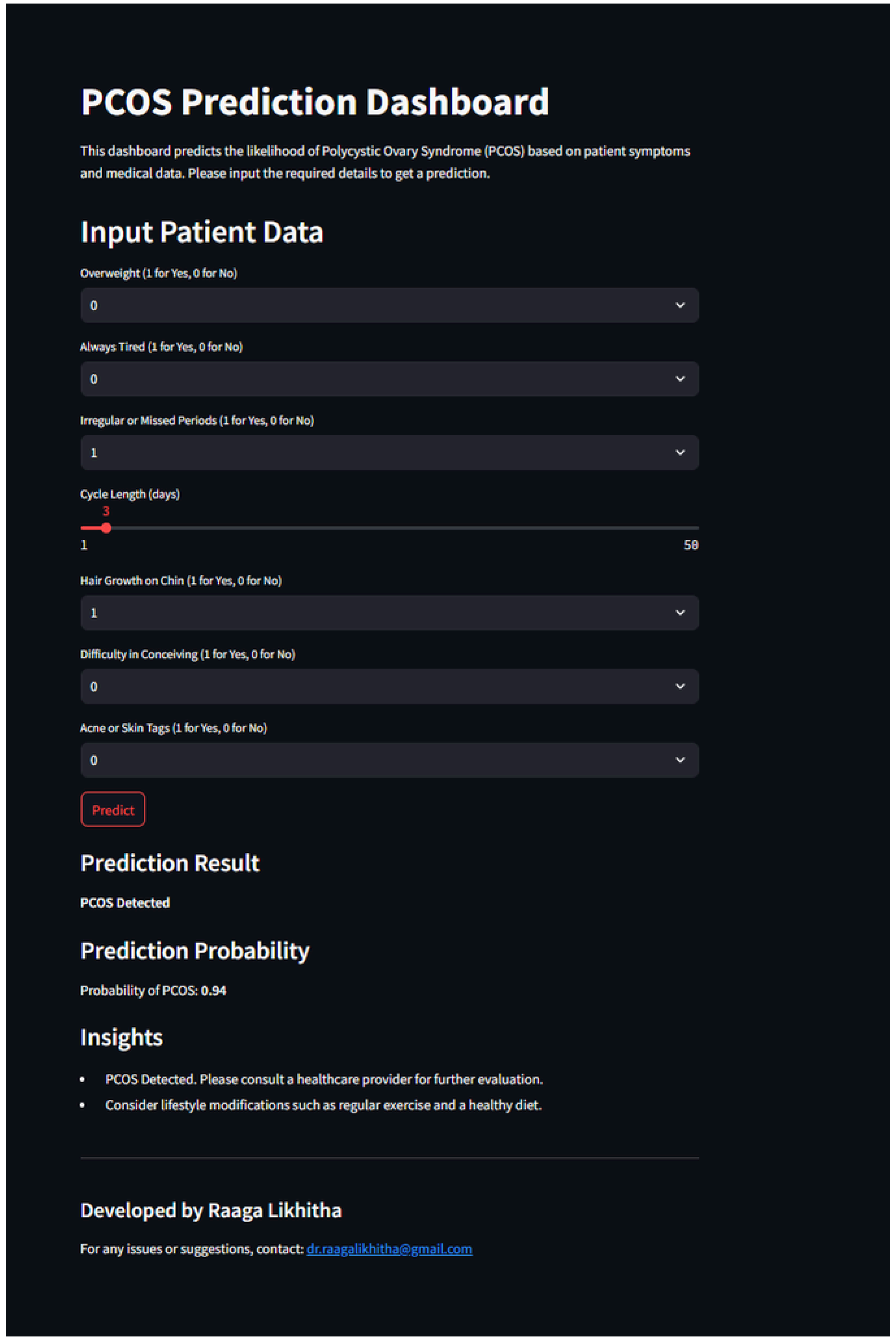
- Critical predictors: Irregular periods, overweight, hair growth on chin
- Composite features (e.g., Symptom Severity Score) enhanced predictive power
- Lifestyle factors showed nuanced relationships with PCOS prevalence
- Higher PCOS prevalence among younger women and urban populations
- Data gaps highlight need for robust preprocessing techniques

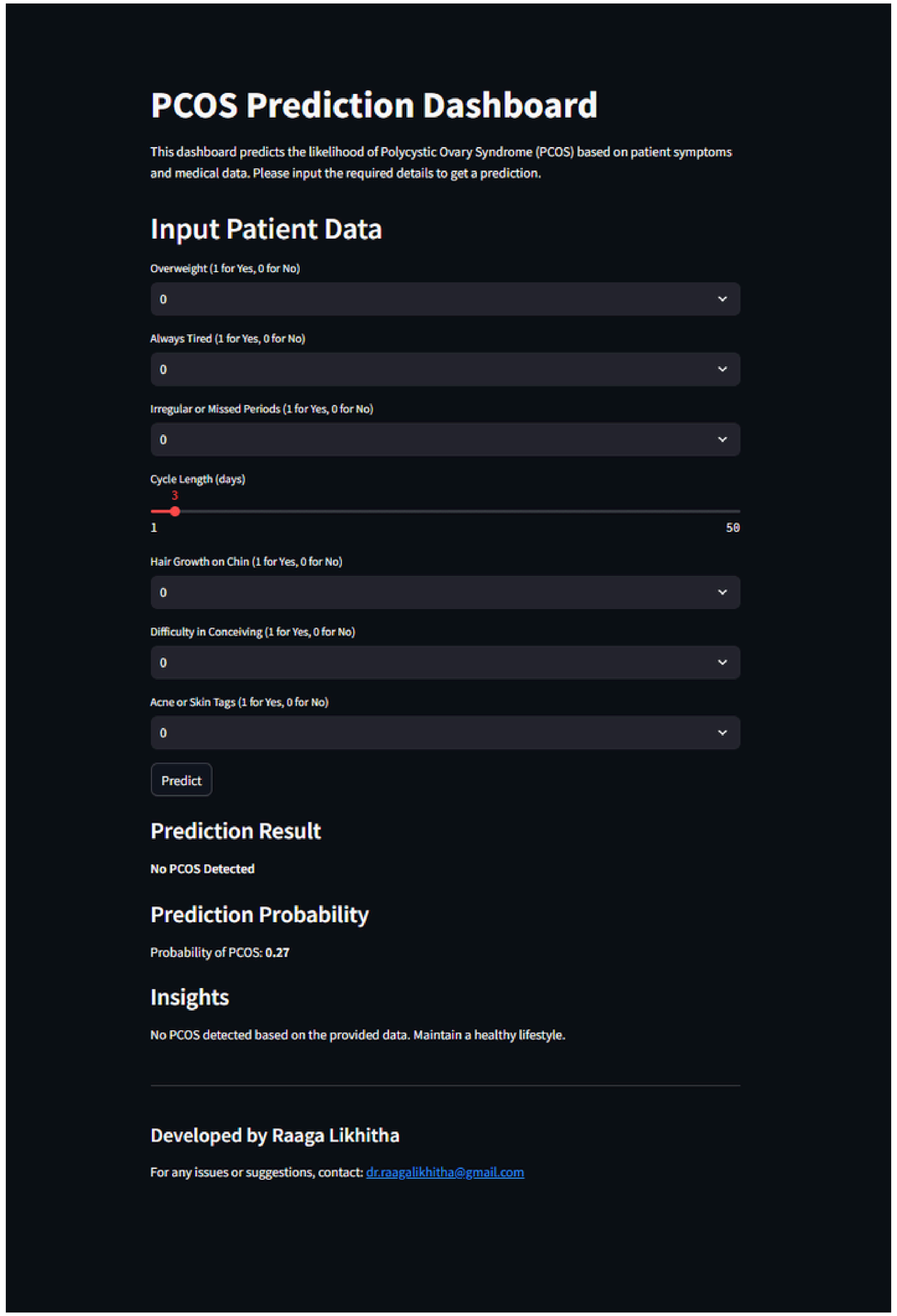
Potential Applications

- AI-enabled diagnostic tools integrated into EHR systems
- Risk scoring and case prioritization for clinicians
- Targeted public health awareness campaigns
- Educational tools for women's self-assessment

Expected Impact

- Reduced diagnostic delays and earlier interventions
- Improved quality of life for women with PCOS
- Enhanced healthcare efficiency and cost reduction
- Addressing healthcare inequities in underserved regions
- Support for personalized treatment approaches





Future Directions

- Validation across larger, diverse datasets
- Development of user-friendly interfaces (dashboards, mobile apps)
- Integration into public health strategies and clinical workflows
- Creation of self-training models on extensive external datasets
- Development of live, accurate dashboards for real-time insights

References

- Barrera, F. J., Brown, E. D. L., Rojo, A., Obeso, J., Plata, H., Lincango, E. P., Terry, N., Rodríguez-Gutiérrez, R., Hall, J. E., & Shekhar, S. (2023). Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. Frontiers in endocrinology, 14, 1106625. https://doi.org/10.3389/fendo.2023.1106625
- Elmannai, H., El-Rashidy, N., Mashal, I., Alohali, M. A., Farag, S., El-Sappagh, S., & Saleh, H. (2023). Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence. Diagnostics (Basel, Switzerland), 13(8), 1506. https://doi.org/10.3390/diagnostics13081506
- Rasquin, L. I., Anastasopoulou, C., & Mayrin, J. V. (2022, November 15). Polycystic ovarian disease. In StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing. Available from https://www.ncbi.nlm.nih.gov/books/NBK459251/
- Ganie MA, Chowdhury S, Malhotra N, et al. Prevalence, Phenotypes, and Comorbidities of Polycystic Ovary Syndrome Among Indian Women. JAMA Netw Open. 2024;7(10):e2440583. doi:10.1001/jamanetworkopen.2024.40583
- India unprepared for PCOS crisis. (n.d.). Think Global Health. Retrieved November 17, 2024, from https://www.thinkglobalhealth.org/article/india-unprepared-pcos-crisis
- Pruthi, B. (2019, September 26). One in five Indian women suffers from PCOS: Peak in hormonal issues draws attention to poor lifestyle habits; doctors say 80% of patients are obese. The Hindu. Retrieved November 17, 2024, from https://www.thehindu.com/sci-tech/health/one-in-five-indian-women-suffers-from-pcos/article29513588.ece
- Singh, S., Pal, N., Shubham, S., Sarma, D. K., Verma, V., Marotta, F., & Kumar, M. (2023). Polycystic Ovary Syndrome: Etiology, Current Management, and Future Therapeutics. Journal of clinical medicine, 12(4), 1454. https://doi.org/10.3390/jcm12041454
- Suha, S. A., & Islam, M. N. (2022). An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Scientific Reports, 12(1), 17123. https://doi.org/10.1038/s41598-022-21724-0
- World Health Organization. (2023, June 28). Polycystic ovary syndrome. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome
- Zad, Z., Jiang, V. S., Wolf, A. T., Wang, T., Cheng, J. J., Paschalidis, I. C., & Mahalingaiah, S. (2024). Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records. Frontiers in Endocrinology. https://doi.org/10.3389/fendo.2024.1298628

THANK YOU