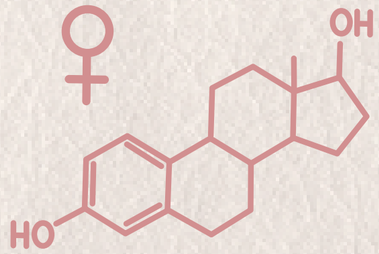
Cracking the PCOS Code

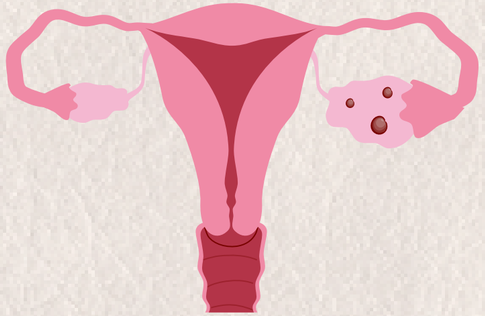
A Semester Project Proposal

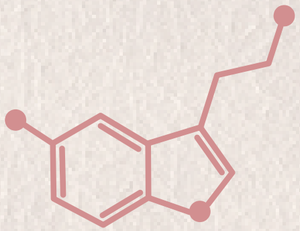
INST 414 (0102) Group 1: Adrien Rozario & Lakshya Sajal Kumar

What is PCOS?

- PCOS (Polycystic Ovary Syndrome) is a common endocrine disorder in women
- It is linked to hormonal imbalance, irregular periods, infertility, and metabolic issues
- Yet for such a widespread condition, PCOS is often underdiagnosed
- Symptoms and diagnostic criteria vary, causing many women spend years seeking answers



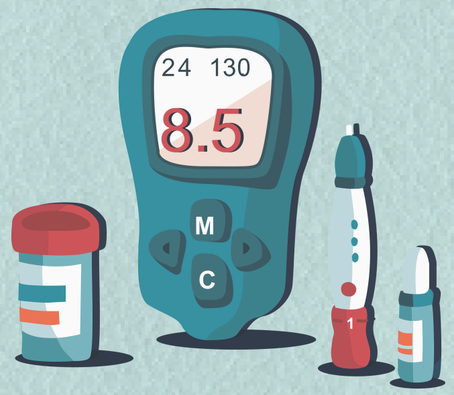




The Big Question

“What Health Indicators Best Predict PCOS Risk?”

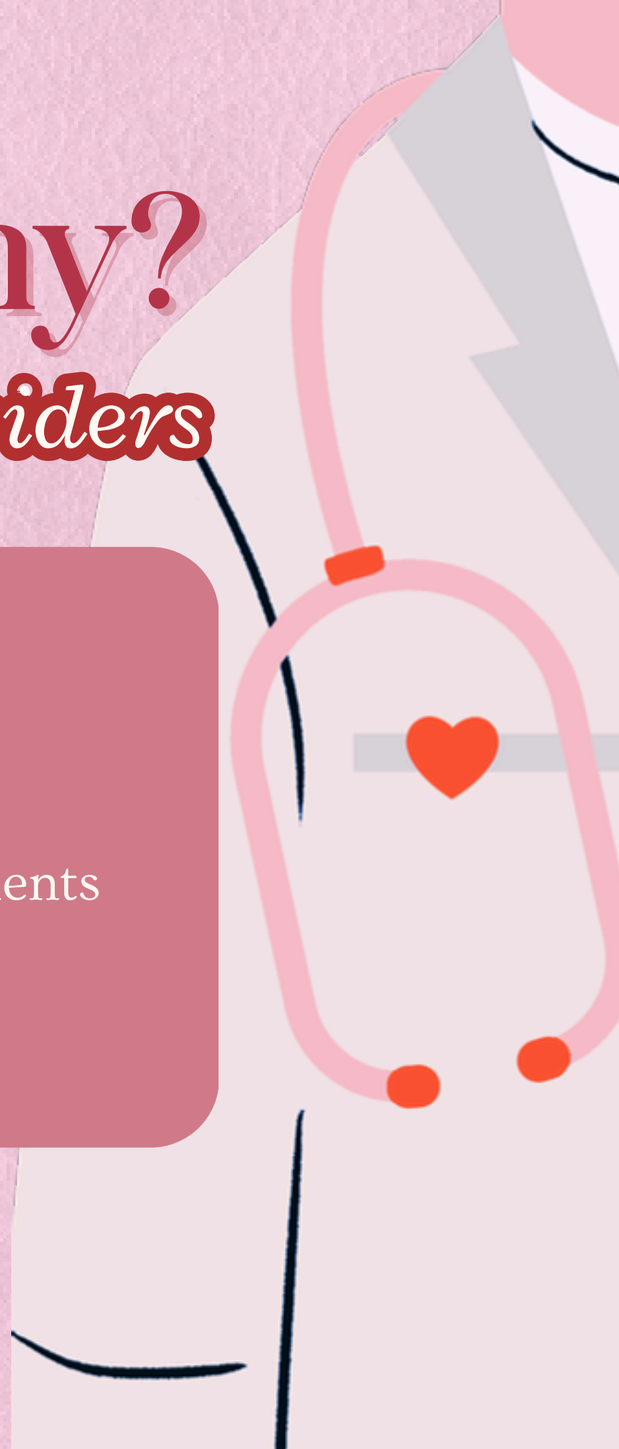
Using statistical and machine learning techniques, we aim to identify the the strongest symptoms and biomarkers of PCOS to support earlier detection and personalized care



Who Needs This & Why?

Primary Stakeholder: *Healthcare Providers*

- Clinicians (OB/GYNs, endocrinologists, and primary care providers) who sees patients with PCOS symptoms
- The insights from this analysis will help them prioritize patients for further testing, intervene earlier, and provide more personalized medical and lifestyle recommendations



Subject Matter Expertise

In addition to data science techniques, this analysis relies on clinical knowledge to ensure the results are meaningful

This includes the following fields of expertise:

Reproductive Health

Understanding hormonal biomarkers, menstrual/reproductive patterns, and infertility indicators relevant to PCOS

Biostatistics

Interpreting biomarker ranges, population variability, and clinical significance

Clinical Communication

Translating predictive features into insights for OB/GYNs and primary care providers

Original Project

Data

PCOS_data_without_infertility.csv

- 541 patients from 10 hospitals in Kerala, India
- Clinician-verified PCOS diagnoses
- 13 clinically relevant features:
  - Symptoms: age, weight gain, hair growth, skin darkening, hair loss, pimples, fast food, regular exercise
  - Hormonal indicators: AMH, LH, FSH, PRL, TSH, PRG
  - Target: PCOS (0 = no, 1 = yes)

Original Project

Analysis

Medium

Supervised Learning Model

- Built random forest classification model to predict PCOS diagnosis

Evaluation

- Evaluated accuracy, precision, recall, and F1-score
- Generated:
  - Classification report
  - Confusion matrix
  - Feature importance + distributions
  - Error analysis

Results: 82% accurate model

- The top predictors of PCOS are skin darkening, weight gain, and hair growth, followed by AMH levels
- Physical symptoms are often stronger predictors than lifestyle or hormone levels
- Error analysis revealed critical faults in the sample

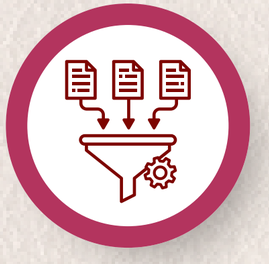
Project Extension Plan

1

Extend Feature Columns

2

Data Cleaning

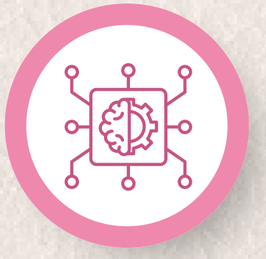


3

Cluster Analysis

4

Supervised Learning Model



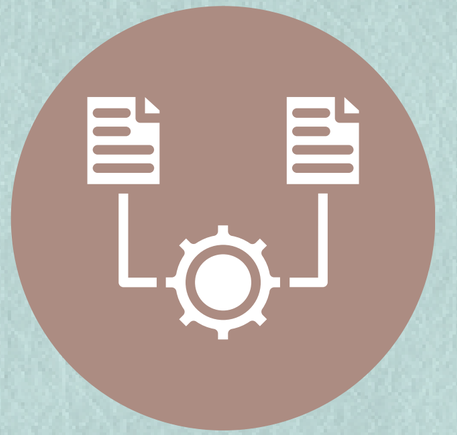
5

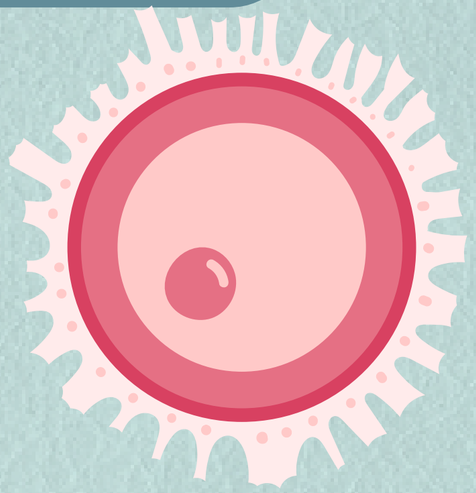
Model Performance Evaluation

Extend Feature Columns

Add 3 fertility-related biomarkers to feature columns:

- I beta-HCG, II beta-HCG, AMH
- Patient IDs match the main dataset -> can be merged to assess hormonal indicators related to infertility

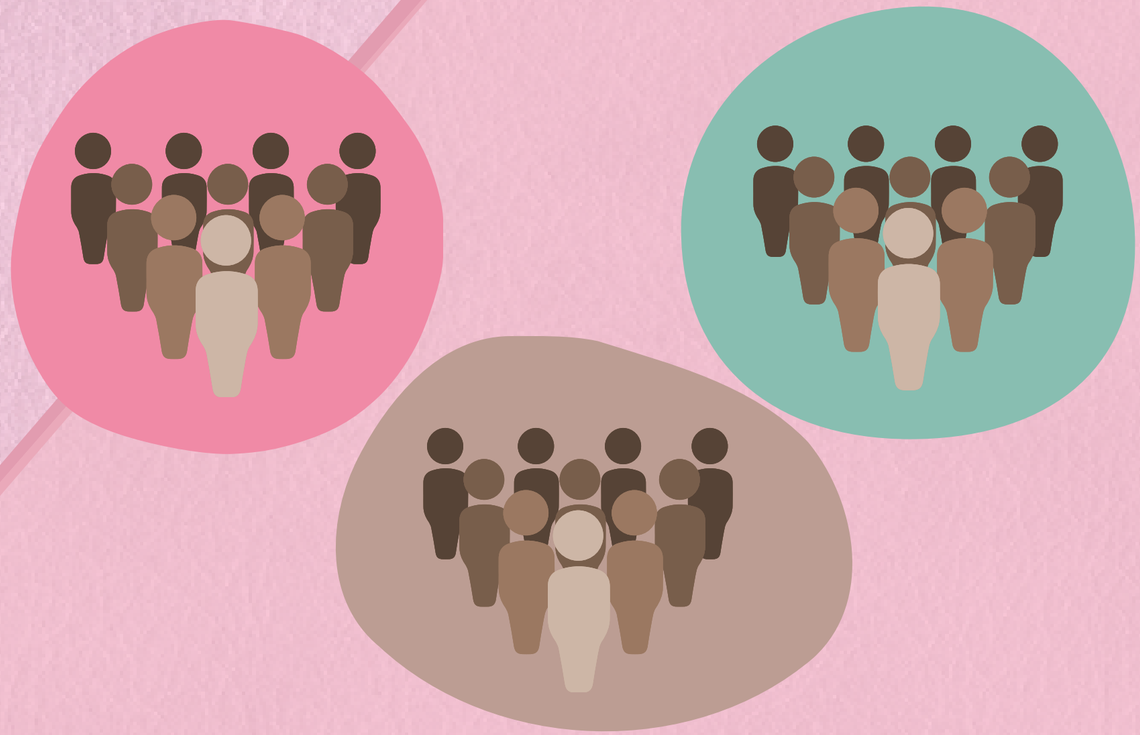




Cluster Analysis

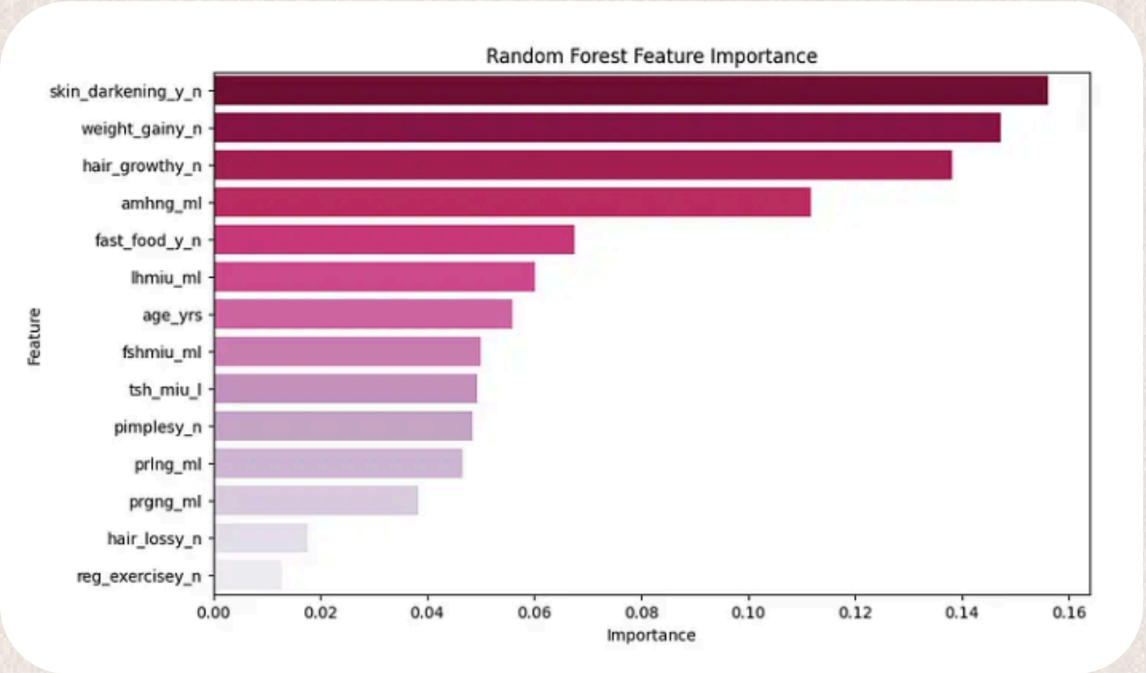
“What health patterns naturally group PCOS and non-PCOS patients?”

- Use unsupervised clustering (k-means) to identify patient groups
- Explore patterns of PCOS risk across symptoms and lifestyle patterns
- Examine which features define clusters
- Compare those features with predictive model results



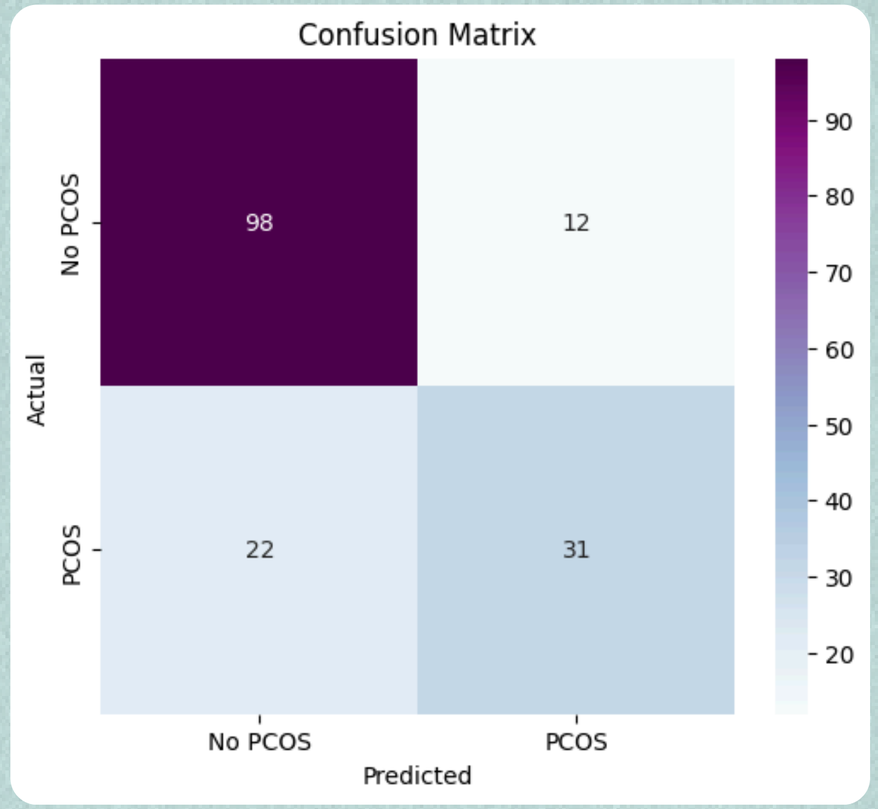
Supervised Learning Model

- Apply random forest classification model on merged dataset
- Predict PCOS risk and identify key health indicators
- Incorporate insights from clustering to inform feature selection
- Focus on interpretable and clinically relevant predictors



Model Performance Evaluation

- Accuracy, sensitivity (recall), specificity, precision
- Confusion matrix to visualize correct vs incorrect predictions
- Examine feature importance and top feature distributions
- Highlight consistent indicators of PCOS risk for clinicians




Thank you!