**Documentation: Unveiling PCOS: Data-Driven Insights**

**1. Introduction**

Polycystic Ovary Syndrome (**PCOS**) is a **common hormonal disorder** that affects individuals with ovaries, primarily during their reproductive years. It is one of the leading causes of **menstrual irregularities, infertility, and metabolic complications.** Despite being widespread, PCOS remains **underdiagnosed** due to its diverse symptoms and varying manifestations.

This project explores **lifestyle, health, and medical factors associated with PCOS** using data analysis. The dataset primarily consists of **categorical variables (Yes/No responses)** related to **diet, sleep, exercise, symptoms, and medical history.**

**1.1. About PCOS (Polycystic Ovary Syndrome)**

PCOS is characterized by hormonal imbalances that can lead to:

* **Irregular Menstrual Cycles** (skipped or prolonged periods)
* **Excess Androgen Levels** (causing acne, hair growth, or hair thinning)
* **Insulin Resistance & Weight Gain** (increasing the risk of diabetes)
* **Hirsutism** (excessive hair growth on the face and body
* **Mental Health Challenges** (such as anxiety, stress, and depression)

The exact cause of PCOS is unknown, but factors like **genetics, insulin resistance, and lifestyle choices** play a significant role in its development and severity. Early diagnosis and lifestyle interventions can help manage symptoms effectively.

**1.2. Project Objectives**

Through **Exploratory Data Analysis (EDA),** this project aims to:

* Identify **key trends and prevalence** of PCOS in the dataset
* Analyse **associations between lifestyle factors and PCOS**
* Derive **meaningful insights** that can contribute to awareness and further research

The findings from this analysis can help in understanding the impact of **diet, exercise, sleep, and medical history** on PCOS.

**2. Project Overview**

**Objective:** Analyse the dataset to uncover key insights related to **PCOS prevalence, symptoms, lifestyle factors, and correlations**.  
**Scope:** Focused on **Exploratory Data Analysis (EDA)**; predictive modeling was not pursued due to the categorical nature of the dataset.

**3. Dataset Information**

* **Total Records:** *173*
* **Features:** *36* (Mostly Yes/No categorical data)
* **Key Columns:**
  + **Target Variable:** PCOS (Yes/No)
  + **Major Factors:** Menstrual irregularity, Hormonal Imbalance, Lifestyle factors (Diet, Sleep, Exercise), BMI

**4. Data Processing Steps**

✔ **Phase 1: Data Preparation & Feature Engineering**

* **Encoding:** Converted categorical values (Yes/No) into numerical form.
* **Feature Engineering:** Created **Diet Score, Sleep Score, Exercise Score, Healthy Lifestyle Score**.
* **Normalization:** Applied to numerical columns for uniformity.

✔ **Phase 2: Exploratory Data Analysis (EDA)**

* Identified **PCOS prevalence** and its relationship with **age, BMI, symptoms, and lifestyle factors**.
* **Correlation analysis** using Cramér's V to find strong and moderate associations.
* Derived **meaningful insights** from trends in symptoms, lifestyle, and medical history.

**5. Key Findings (EDA Summary)**

* **PCOS Prevalence:** 22% of the dataset, most common in **20-25 age group** and **unmarried women**.
* **Top Symptoms:** Menstrual irregularity, hormonal imbalance, hirsutism.
* **Lifestyle Impact:** PCOS patients show **less stable diet & exercise habits**.
* **BMI & PCOS:** No strong correlation but **higher BMI is commonly observed.**
* **Childhood Trauma:** **Slight increase in PCOS prevalence** among individuals reporting past trauma.
* **Cramér's V Analysis:** Strong associations with **hormonal imbalance, menstrual issues, and hyperandrogenism**.

**6. Conclusion & Future Scope**

✅ **Why No Predictive Modeling?**

* Dataset is **highly categorical (Yes/No format)**, making predictive modeling complex and less meaningful.
* Converting all columns into numerical format would make the model **messy and difficult to interpret**.
* After encoding, the dataset becomes very wide (many columns but fewer numerical features). This leads to the curse of dimensionality, making model training difficult.

✅ **Potential Future Work:**

* **Collect more continuous numerical data** (e.g., hormone levels, glucose levels).
* Consider a **survey-based study** to analyse patient history and risk factors in-depth.

**7. References & Acknowledgments**

* **Dataset Source:** *Kaggle (*[*Dataset Link*](file:///C:\Users\dell\AppData\Roaming\Microsoft\Word\(https:\www.kaggle.com\datasets\hasaanrana\diet-exercise-and-pcos-insights)) *)*
* **Tools Used:** Python, Pandas, Matplotlib, Seaborn, SciPy
* **Author:** *Gurpreet Kaur*