ENHANCING EARLY PCOS DETECTION THROUGH FEATURE-DRIVEN PREDICTION MODELS

Jayshree Majumdar[1,2], Lakshmikant Dhruv[1], Debarati Biswas[3], Nanda Kishor Jana[4]

1: Department of Agricultural Food Engineering, Indian Institute of Technology Kharagpur
2: Department of Food technology, Guru Nanak Institute of Technology, Kolkata
3: School of Medical Science and Technology, Indian Institute of Technology Kharagpur
4: Analyst, Tata Consultancy Services, Kolkata

Polycystic Ovarian Syndrome (PCOS) is one of the most prevalent endocrine ailments that affect women in childbearing years, traditionally defined by hormonal abnormality, irregular menstruation, infertility, and metabolic disorder. Proper and timely diagnosis is indispensable for impending treatment. These addresses exploring the feasibility of applying a range of machine learning (ML) approaches towards PCOS diagnosis, ranging from clinical, physiological, and lifestyle factors.

A publicly accessible dataset is pre-processed first by dealing with missing values and normalisation, and then later feature selection using Recursive Feature Elimination (RFE) and mutual information. Multiple ML algorithms such as Logistic Regression, Support Vector Machines, Random Forest, Decision Trees, and Gradient Boosting are compared on the performance metrics such as precision, accuracy, recall, and F1-score. The ensemble models are illustrated to outperform due to their energy benefits and potential for identifying complex patterns. The guidelines include utilizing ML in the initial diagnosis of PCOS and confidentiality for sensitive information, with a progress towards the ethical application of forecasting technology in the health sector.

Decision Tree classifier was the top-classifier in diagnosing PCOS with near-perfect accuracy rate of 99.68% and 100% precision and F1-score. The highest-ranked was also K-Nearest Neighbors with 97.89% accuracy and high sensitivity. SVM also performed well with 93.67% accuracy rate, which was excellent in identifying negative cases, then Logistic Regression and Naive Bayes with medium-rankings. The K-Nearest Neighbors (KNN) model also performed well, at 97.89% accuracy and 98% F1-score, and as such is a strong contender for a genuine alternative since it has high specificity and sensitivity. SVM followed closely behind at 93.67% accuracy and 94% F1-score, having good generalization across various PCOS profiles. Logistic Regression was acceptable with 91.38% accuracy and a 91% F1-score, best trading performance for interpretability if features are linearly correlated with each other.

ISBN Number : 978-81-995582-9-8 Crossref