ABSTRACT

Polycystic Ovarian Syndrome (PCOS) Detection using Gradient Boosted Decision Trees

by

Caroline Chan
19/442467/PA/19216

Polycystic Ovarian Syndrome (PCOS) is one of the most common diseases in women of reproductive age, affecting 5 to 10 percent. It is a common condition affecting how a woman's ovary work, characterized by irregular periods, excess androgen (male hormone), and the absence of polycystic ovaries. Women with PCOS tend to be insulin resistant, which can lead to diabetes and overweight or even obesity. According to one study, women with PCOS also have twice as likely a risk of a future cardiovascular event. It also affects mental because women that suffer from PCOS tend to have high depression scores and great body dissatisfaction.

Unfortunately, 50 to 70 percent struggle with undiagnosed PCOS because PCOS diagnosis is still tricky. Doctor finds it difficult to diagnose PCOS because there is no universal definition, meaning there is no universal diagnostic test to assess the patients. It became even more complicated because symptoms vary between women and might not necessarily point to PCOS.

Recently, machine learning has been a trending topic and has proven efficient in diagnosing diseases. Machine learning can help diagnose by analyzing collected clinical data and might become a powerful tool to help women struggle with undiagnosed PCOS. The output of this research will be a classification model to detect PCOS by implementing Gradient Boosted Decision Tree and using data collected from ten different hospitals in Kerala, India. There are 45 clinical and physical features from the collected data. For model construction, this study will implement feature selection algorithm to rank the existing features, hyperparameter optimization and data resampling. The models will be built with different number of features from the maximum number to only ten features. Ultimately, these metrics will evaluate the final classification models: accuracy, precision, recall, and F1-score. In the end, the proposed model shows a great performance with 98,57% for all evaluation metrics while implementing ANOVA F-test to rank the dataset's features.

Keywords: Polycystic Ovarian Syndrome (PCOS), Machine Learning, Gradient Boosted Decision Tree

INTISARI

Polycystic Ovarian Syndrome (PCOS) Detection using Gradient Boosted Decision Trees

by

Caroline Chan
19/442467/PA/19216

Polycystic Ovarian Syndrome (PCOS) adalah salah satu penyakit yang paling umum pada wanita usia reproduksi, mempengaruhi 5 sampai 10 persen. Ini adalah kondisi umum yang mempengaruhi cara kerja ovarium wanita, ditandai dengan menstruasi yang tidak teratur, kelebihan androgen (hormon pria), dan tidak adanya ovarium polikistik. Wanita dengan PCOS cenderung resisten terhadap insulin, yang dapat menyebabkan diabetes dan kelebihan berat badan atau bahkan obesitas. Menurut sebuah studi, wanita dengan PCOS juga memiliki risiko dua kali lebih besar untuk kejadian kardiovaskular di masa depan. Ini juga mempengaruhi mental karena wanita yang menderita PCOS cenderung memiliki skor depresi yang tinggi dan ketidakpuasan tubuh yang besar.

Sayangnya, 50 hingga 70 persen berjuang dengan PCOS yang tidak terdiagnosis karena diagnosis PCOS masih rumit. Dokter sulit mendiagnosis PCOS karena tidak ada definisi universal, artinya tidak ada tes diagnostik universal untuk menilai pasien. Ini menjadi lebih rumit karena gejala bervariasi antara wanita dan belum tentu mengarah ke PCOS.

Baru-baru ini, *machine learning* telah menjadi trending topik dan terbukti efisien dalam mendiagnosis penyakit. *Machine learning* dapat membantu diagnosis dengan menganalisis data klinis yang dikumpulkan dan mungkin menjadi alat yang ampuh untuk membantu wanita yang berjuang dengan PCOS yang tidak terdiagnosis. Hasil akhir dari penelitian ini adalah model klasifikasi untuk mendeteksi PCOS dengan mengimplementasikan Gradient Boosted Decision Tree dan menggunakan data yang dikumpulkan dari sepuluh rumah sakit berbeda di Kerala, India. Ada 45 gambaran klinis dan fisik dari data yang terkumpul. Untuk konstruksi model, penelitian ini akan mengimplementasikan algoritma pemilihan fitur untuk menentukan peringkat fitur yang ada, optimasi *hyperparameter* dan *resampling* data. Model-model tersebut akan dibangun dengan jumlah fitur yang berbeda dari jumlah maksimum hingga hanya sepuluh fitur. Pada akhirnya, metrik ini akan mengevaluasi model klasifikasi akhir: akurasi, presisi, daya ingat, dan skor F1. Model yang diajukan menunjukkan performa yang baik dengan 98,57% untuk semua metrik evaluasi dengan mengimplementasikan ANOVA F-test untuk mengurutkan fitur pada dataset.

Kata kunci: *Polycystic Ovarian Syndrome (PCOS), Machine Learning, Gradient Boosted Decision Tree*