Classification for Predicting PCOS Using Red Deer Algorithm with XGBoost Classifier

J Fathima Kaleema[1], D Usha Rani[2]

[1,2]Department of Computer Science, Thassim Beevi Abdul Kader College for Women, Tamil Nadu, India

Abstract:

Polycystic Ovary Syndrome (PCOS) is a hormonal imbalance disorder that has common among women reproductive age. It affects 20% of women of bearing age. PCOS affected the age between 15-44 .PCOS enlarge ovaries with small cysts in the ovaries. It can lead to unregulated hormonal cycle and also trigger periods , high blood pressure, diabetes, acne, Infertility and growth of hair on face , PCOS cause type 2 diabetes. In this paper propose combination of red deer algorithm with XGBoost algorithm to classify and predict the earlier stage of ovaries to diagnosis and treatment can be used to prevent the long-termproblems.

Keyword: PCOS, Type 2 Diabetes ,Red Deer Algorithm, XGBoost.

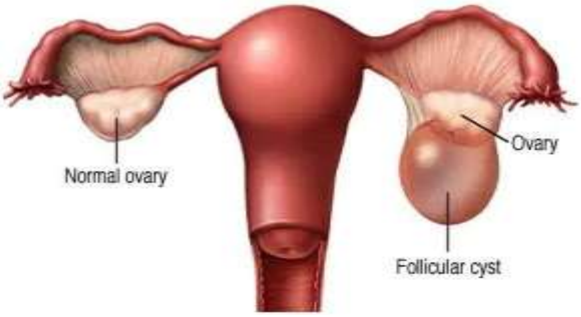
Introduction:

Polycystic Ovary Disease (PCOD) or Polycystic Ovary Syndrome (PCOS) is a hormonal imbalance issue found in women reproductive age. In today’s generation, there has been a constant rise in the occurrence of this disease due to the modern lifestyle and fast food habits. There are three types of ovaries and is classified as normal ovary, cystic ovary and polycystic ovary.(1] Poly Cystic Ovarian Syndrome or PCOS is a complex hormonal plights distressing up to around 1 in every 5 women at their conceptive age. Peripherally inside the ovary, fluid-filled sacs are present which are called follicles or cysts. A polycystic ovary (PCO) can be characterized by twelve or more follicles with a diameter of 2-9 mm(Jarrett etal., 2020). PCOS affects the health of women's life. The symptoms include cardiovascular diseases, infertility, late menopause, type 2 diabetes, acne, baldness, hair loss, obesity, anxiety, depression, and stress(2).

The lack of ovulation changes the level of estrogen, progesterone, follicle stimulating hormone (FSH) and Leutinizing hormone(LH). Small amount of androgen, the male hormone present in women’s body. PCOS can both cause too much androgen production. It causes to derange menstrual cycle it leads to getting fewer menstruation and ovulation of women with normal PCOS. Cyst of ovaries severity of the disease regarding the size and measure of follicles in ovaries and to verify the thickness of uterus lining patients should definitely undergo ultrasound scanning or further diagnostic steps. The ovaries are 1½ to 3 times larger than normal ovaries.

Fig 1 Before Cyst form in Ovary

Fig 2: Follicular Cyst in Ovary



Proposed work:

The proposed work for classification comprises a pre-processing subsystem .The preprocessing subsystem uses the zscore for reducing the noise of that dataset. The wrapper approach comprising Red Deer Algorithm and XGBoosting classifier is used for performing feature selection. After selecting optimal feature ,XGBoost classifies is used to classified the dataset into training and testing dataset. The PCOS dataset collected from kaggle repository is used in this work.The dataset containing the details of 541 patients and 38 features .From the instance of data divided into testing and training data set. 70% of PCOS data instance in the data set are assigned for the training set and the remaining 30% are selected for testing data set its related to the model which include different kind of patterns for evaluating that dataset accuracy using highly similar but different instance from the test data.

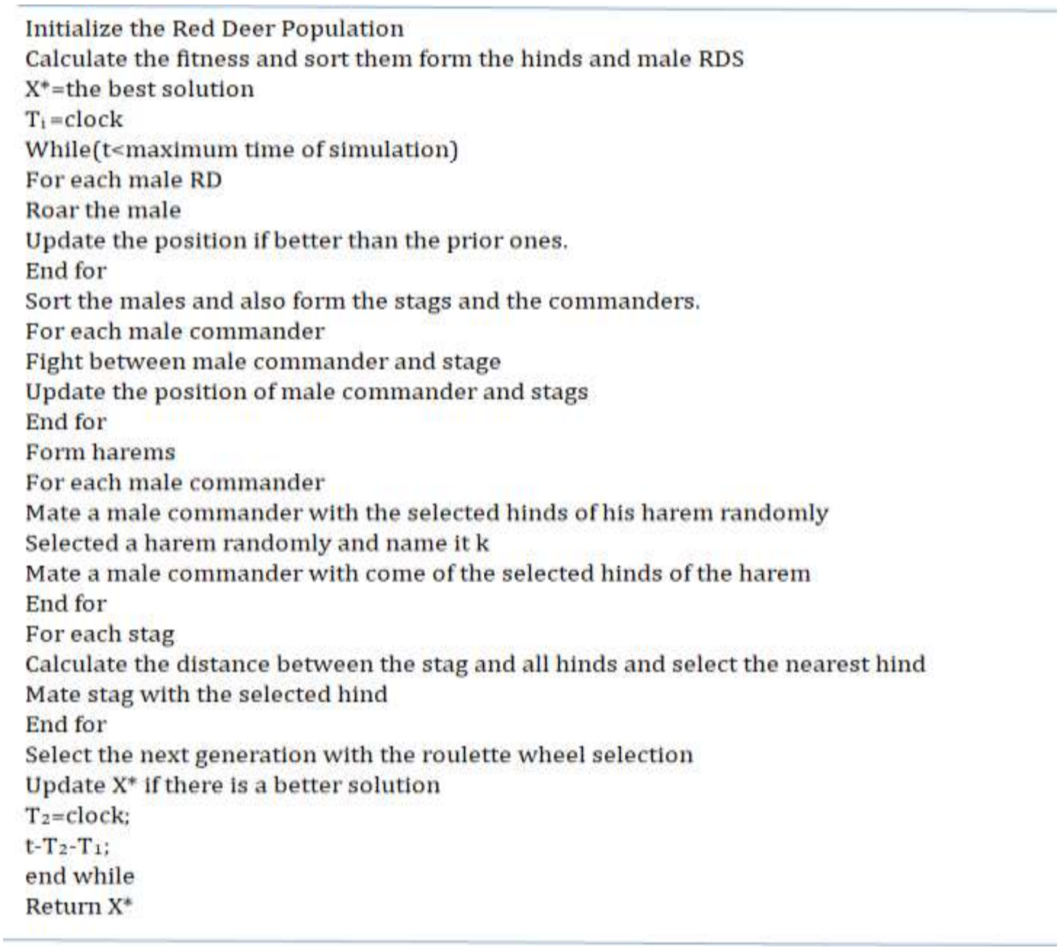
Red Deer Algorithm

It Similar to other meta-heuristics, the RDA starts with an initial random population that is the counterpart to RD's. A number of the best RDs among the population is selected and named the ''male RD'' and the rest of them are called ''hinds.'[7]' First of all, the male RD should roar. Based on the power of a roaring phase, they are divided into two groups(i.e., commanders and stags). After that commanders and stags of each harem fight together in order to own their harem. Besides, harems are formed by commanders. The number of hinds in harems is directly related to the commanders' abilities in roaring and fighting process. Consequently, commanders mate with a number of hinds in harems. Note that the other males (i.e., stags) mate with the nearest hind without considering the limitation of the harem[8]

Generally, the mentioned steps of the RDA are designed in a way to consider the exploitation and exploration phases satisfactorily. The user can tune the phases regarding the used parameters and mathematical formulation. Accordingly, the roaring of male RD is the counter-part of local search in solution space to improve the exploitation properties. Similarly, the fighting between commanders and stags is also considered as local search however, in this process, we only accept the better observed solutions. This step mainly considered the exploitation characteristics as well. After that harems are formed and allocated to the commanders according to their power. This step helps the algorithm to do the exploration phase. Accordingly, the commander of a harem mates with a percentage of hinds in his harem and also with a percentage of hinds in another harem.

These stages have also been improved the exploration properties. Note that regarding the breading season, all stags should mate with the nearest hind, that is, a stag mate with the hind with the minimum distance without considering of the limitation of the harem. This step also focuses on both exploration and exploitation phases, simultaneously. Another main phase of the RDA is the mating process, which leads to generating offspring of RD's. This phase is the counterpart of making new solutions in solution space. Finally, the next generation of the algorithm is done by giving a chance for weak solutions regarding the classification of the algorithm as an evolutionary one.

Red Deer Algorithm



```
Initialize the Red Deer Population
Calculate the fitness and sort them form the hinds and male RDS
X*=the best solution
T1 =clock
While(t<maximum time of simulation)
For each male RD
Roar the male
Update the position if better than the prior ones.
End for
Sort the males and also form the stags and the commanders.
For each male commander
Fight between male commander and stage
Update the position of male commander and stags
End for
Form harems
For each male commander
Mate a male commander with the selected hinds of his harem randomly
Selected a harem randomly and name it k
Mate a male commander with come of the selected hinds of the harem
End for
For each stag
Calculate the distance between the stag and all hinds and select the nearest hind
Mate stag with the selected hind
End for
Select the next generation with the roulette wheel selection
Update X* if there is a better solution
T2=clock;
t-T2-T1;
end while
Return X*
```

XGBoosting

Extreme Gradient Boosting or XGBoost is a decision tree based ensemble ML Algorithm that is used in the library of gradient boosting Some of the major benefits of XGBoost are that its highly scalable/parallelizable, easy to visualize, quick to execute, and typically outperforms other algorithms are used to more regularized model formalization, to control over-fitting, which gives it better performance and the variety of hyper parameters that can fine tune. XGBoost used for regression, classification (binary and multiclass), and ranking problems.

Performance metrics:

In this section the most common evaluation metrics to investigate performance of classifies the accuracy ,TP,FP,FN,TN

Accuracy: It indicate the total proportion of correctly predicted

Accuracy:

$$\frac{TP+TN}{TP+FP+FN+TN}$$

Sensitivity : Indicate positive labels

$$\frac{TP}{TP+FN}$$

Specificity: Indicate negative labels

$$\frac{TN}{TN+FP}$$

Precision: Predict all positive classified labels

$$\frac{TP}{TP+FP}$$

$$F\text{-measure}=2 \text{ x } \frac{\text{Sensitivity x precision}}{\text{Sensitivity x precision}}$$

ROC is one of the evaluation tools to performance analysis of the classified in which X-axis represent the FP-rate and Y-axis represent TP rate ROC curve between 0 and 1

Classification

The dataset with the optimal feature subset is then trained and tested using a XGBoost classifier. XGBoost is a collective classifier which assigns class label to a test data. The process begin with the root node. It uses sequentially-built shallow decision trees to provide accurate results and a highly-scalable training method and avoiding over fitting .

Results

The google colab was used to implement this algorithms. The PCOS dataset do not have any issing values and so no imputation technique is in this work. However it can be observed that dataset contains noise in the form of outliers. The outlier analysis done using z-score has revealed the number of outliers in the dataset[6].

The dataset is then divided into labelled and unlabelled datasets sets using 10 fold cross validation. RDA is then applied to the labelled datasets for selecting the feature subsets and a XGBoost classifier is used to evaluate the subsets[4]. The fitness value started to converge after 87 iterations, the best training accuracy achieved is 93.2% and the number of attributes selected is 30.

The performance of the proposed work is evaluated using the performance measures viz. Accuracy, Sensitivity and Specificity. Table 1 shows the results of the proposed classification framework on the PCOS dataset.The Accuracy distribution and MCC distribution are shown in Fig.1 and Fig. 2 respectively

No of Runs	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	MCC
1	63	30	2	7	91.17	90	93.75	0.82
2	73	25	1	3	96.07	96.05	96.15	0.84
3	66	24	1	10	89.10	86.84	96	0.69
4	70	25	0	6	94.05	92.10	100	0.82
5	64	28	2	11	87.61	85.33	93.33	0.76
6	64	26	2	13	85.71	83.11	92.85	0.72
7	66	30	3	5	92.30	92.95	90.90	0.72
8	68	27	5	7	88.78	90.66	84.37	0.52
9	64	26	5	6	89.10	91.42	83.87	0.52
10	75	25	3	4	93.45	94.93	89.28	0.73
Average SD					90.734	90.339	92.05	0.714

Fig 1 Accuracy Distribution

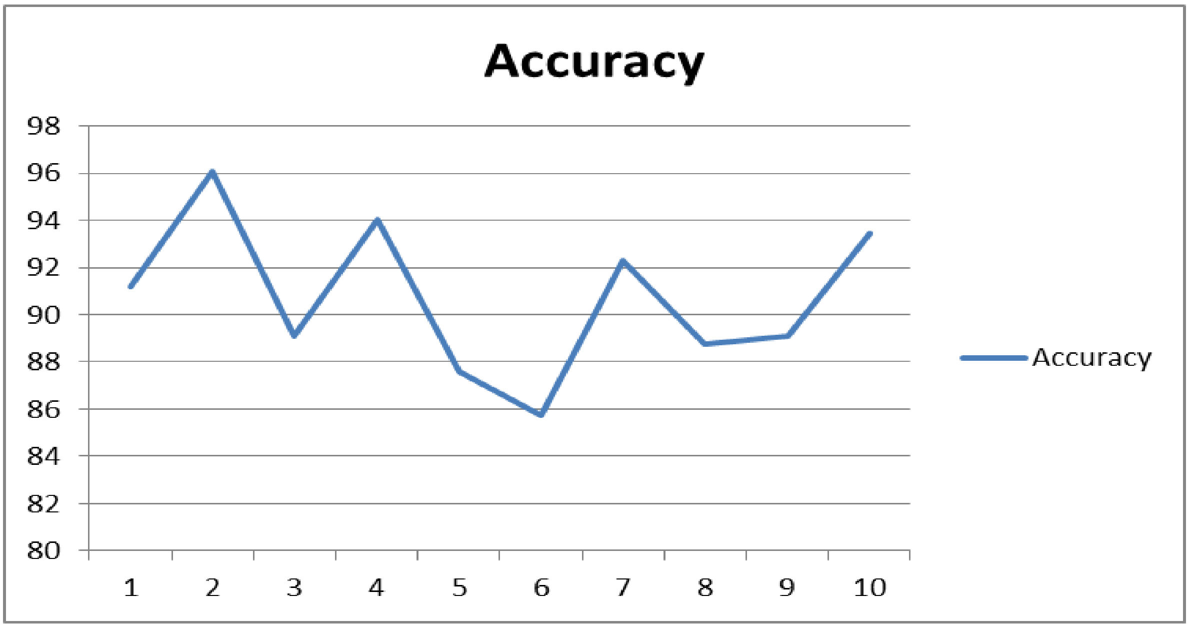
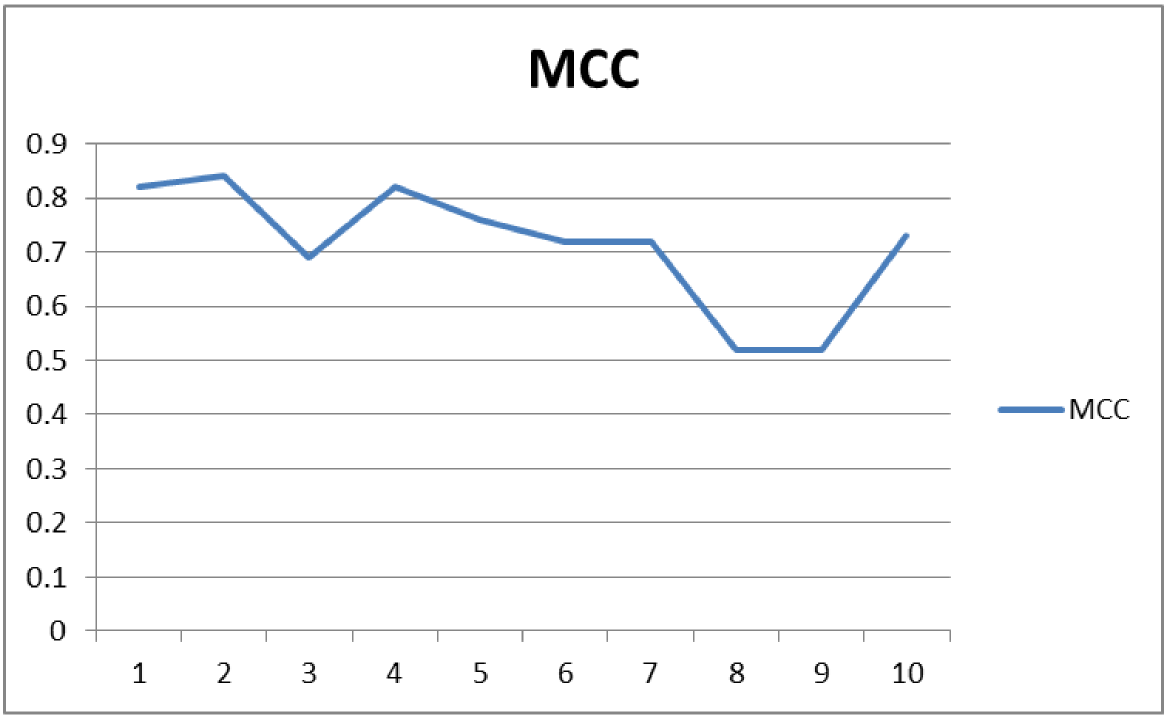


Fig 2 MCC distribution



Conclusion:

The PCOS dataset with 37 features obtained from kaggle repository wrapped feature selection methods.Red Deer Algorithm is used for selecting the feature of subsets and XGBoosting algorithm is evaluating the dataset.The proposed approach attained an accuracy 90.73% ,specificity of 90.33, sensitivity of .The proposed method is perform better in PCOS XGBoosting with Red Algorithm algorithm.

Reference:

1. Glueck C.J., *et al.* Incidence and treatment of metabolic syndrome in newly referred women with confirmed polycystic ovarian syndrome
2. Marika Mikola, *et al.*Obstetric outcome in women with polycystic ovarian syndrome Hum. Reproduct., 16 (2) (2001), pp. 226-229
3. Cindy L. Weiner, Primeau Margaret, Ehrmann David A. Androgens and mood dysfunction in women: comparison of women with polycystic ovarian syndrome to healthy controls
4. Saman Sarraf, Tofighi Ghassem Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks
5. Behrouz Samieiyan, *et al.* Novel optimized crow search algorithm for feature selection Expert Syst. Appl. (2022), Article 117486
6. Darling Frank Fraser A Herd of Red Deer: A Study in Animal Behaviour Luath Press Ltd (2008) Fathollahi-Fard Amir Mohammad, Hajiaghaei-Keshteli Mostafa, Tavakkoli- Moghaddam Reza Red deer algorithm (RDA): a new nature-inspired meta-heuristicSoft Comput. (2020), pp. 1-29
7. Kottarathil Prasoon Polycystic ovraian syndrome (PCOS) dataset retrieved 4 2020 (2019) from https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos