International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

A Prediction Model for Evaluating the Risk of Developing PCOS

Pushkarini H[1], M A Anusuya[2]

[1]*Student, Department of Computer Science and Engineering, JSS S&TU, Mysuru, India*
[2]*Associate professor, Department of Computer Science and Engineering, JSS S&TU, Mysuru, India*

Abstract - *Polycystic ovary syndrome (PCOS) is very common in women these days mainly due to poor lifestyle choices. It is discovered that PCOS, an endocrine disorder found among the females of childbearing age has become a critical reason for infertility. PCOS can induce abnormalities in the ovaries, with high danger of abortion, infertility, heart problems, diabetes, uterus cancer etc. The signs of PCOS include cysts in ovaries, overweight, menstrual disorder, high levels of male hormones, pimples, hair fall and hirsutism. It is not easy to determine PCOS because of its different combinations of symptoms in different women and various criteria involved in the diagnosis. The time needed for various biochemical tests and ovary scanning; also the financial expenses have become a hardship to the patients. To solve this issue, the paper proposes a symptom evaluating and monitoring system using which the risk of developing PCOS can be predicted and the patients can be advised to consult a doctor and undergo tests and scanning only when the risk is high. Parameters selected to evaluate the risk of developing PCOS are optimal and low cost. Training and testing of models are performed on the PCOS dataset using the following features i.e., Testosterone, Hirsutism, Family history, BMI, Fast food, Menstrual disorder, Risk (target). Linear regression, KNN and Random forest models are implemented and evaluated on various performance metrics like R^2, MAE, and RMSE using python 3. Random forest algorithm outperforms the other two algorithms by achieving less error values and highest R^2 value.*

Key Words: PCOS, Machine learning, Risk prediction, Linear regression, KNN, Random forest, Performance metrics.

1. INTRODUCTION

Ovaries are an important part of the female reproductive system; they're located in the lower belly on either side of the uterus. Women have 2 ovaries that grow eggs and secrete the hormones oestrogen and progesterone. During a woman's menstrual cycle, an egg grows in a sac called a follicle. This sac is located inside the ovaries. In normal cases, this follicle or sac breaks open and releases an egg. But if the follicle doesn't tear open, the fluid inside the follicle can form a cyst on the ovary (known as polycystic ovaries i.e., PCO/PCOD). This is also associated with hormonal imbalance and many women may develop polycystic ovary syndrome (PCOS) [1].

Polycystic ovary syndrome (PCOS) is a common endocrine disorder with a world-wide prevalence of 5-10% and is a major cause for chronic menstrual disorder and infertility in women. Women with PCOS have high levels of male hormones and insufficient female hormones, leading to changes in their menstrual cycle. With PCOS, the ovaries are bulged and sometimes have multiple small cyst formations (immature follicles). PCOS is identified by menstrual irregularity, signs of hyperandrogenism such as pimples, hirsutism, hair fall and infertility. In addition, PCOS is linked to various chronic health issues like heart diseases, obesity, infertility, uterus cancer and diabetes. Many women will develop minimum one cyst in their lifetime. In many cases, cysts are silent i.e., show no signs which makes it difficult to diagnose [2].

Though the exact reason for PCOS is not fully understood, it is considered to be dependent on various factors. Mostly due to hormonal imbalances like high levels of androgens, luteinizing hormone (LH) and normal or suppression of follicle stimulating hormone (FSH) resulting in imbalanced LH/FSH ratio. Also the clinical features of hyperandrogenism are related to hyperinsulinemia and insulin resistance. It is not clear that which factors can put woman in risk of developing PCOS, however it was observed in some cases that PCOS is genetic in nature, also various life style factors and obesity was found to contribute for PCOD and hyperinsulinemia there by risking the individuals for developing PCOS [3].

There are many published researches that have investigated the prevalence and common clinical parameters of PCOS in different regions but there is no research that explains the link between some common parameters with PCOS. Through this project which uses machine learning algorithms, I have made an attempt to understand the parameters that contribute in the development of PCOS and to assess the risk associated with it, so that symptoms of PCOS can be identified and tracked at earlier stages and also to avoid further complications.

1.1 Motivation

Identifying PCOS is tricky due to several manifestations, gynecological, clinical and metabolic parameters involved in diagnosing it. The time needed for various clinical tests and ovary scanning, also the financial expenses has become a hardship to the patients with PCOS. This is one of the main reasons for women to neglect the symptoms in initial stages and then later suffer from complications created by PCOS.
It is not possible for everyone to afford these tests and scanning. To address this problem this paper proposes a

symptom evaluating and monitoring system using which the risk of developing PCOS can be predicted and the patients can be advised to consult a doctor and undergo tests and scanning only when the risk is high. This is achieved by applying various machine learning algorithms. Clinical parameters selected to train the model for evaluating the risk of developing PCOS are optimal and low cost.

2. LITERATURE SURVEY

This chapter presents a thorough survey of studies on PCOS, various image processing and machine learning methods to support its diagnosis and also to automate it. Literature says that about 5-10% of Indian women in child-bearing age are impacted by this endocrine disorder called Polycystic Ovary Syndrome (PCOS). It is a significant cause of anovulatory infertility, and heightened risk for diabetes, obesity, heart disease and psychosocial disorders [2,3]. Two way relationships exist between obesity and PCOS. Few recent researches are exploring PCOS associated factors such as obesity [4] and heredity [5].

Table 1: Criteria given by various medical associations to diagnose PCOS

Medical associations	1.Hyperandrogenism	2. Menstrual disorder	3. PCO
NIH criteria,1990 (Should have both of the criteria)	Yes	Yes	-
Rotterdam Criteria, 2003(Should have any two of the marked criteria)	Yes	Yes	Yes
AES criteria,2009 (Should have 1 along with 2 or 3)	Yes	Yes	Yes

Gulam Saidunnisa Begum et al. have investigated the general factors of PCOS such as heredity, fast food diet habits, involvement in physical exercise, BMI and waist size of study participants as possible risk factors for development of PCOS. They found that women with heredity of PCOS, fast food diet habits, and obesity are at higher risk of PCOS compared to participants without these factors.[6]

Amsy Denny et al. This paper proposes system for detecting and predicting PCOS, using various clinical and metabolic features. Out of the 23 features from clinical and metabolic features, 8 potential features were identified by them using SPSS V 22.0 tool based on their significance. The dimensionality of the feature set was reduced using Principal Component Analysis (PCA) and various classification machine learning techniques such as Naïve Bayes classifier method, logistic regression, K-Nearest neighbor (KNN), Classification and Regression Trees (CART), Random Forest Classifier, Support Vector Machine (SVM) are applied. PCOS prediction using RFC obtained accuracy of 89.02%.[7]

Sharvari S. Deshpande et al. In this paper, automated detection of pcos is done by calculating number of follicles in ovarian ultrasound image and also considering clinical, biochemical and imaging criteria to classify patients in two groups i.e. normal and pcos affected. Number of follicle are detected by ovarian ultrasound image processing using image preprocessing like contrast enhancement and filtering, feature extraction using Multiscale morphological approach and segmentation. Support Vector Machine algorithm is applied for classification which inputs all the parameters such as body mass index (BMI), hormonal levels, menstrual cycle length and number of follicles detected in ultrasound image. The accuracy obtained for this method is 95%.[8]

Palak Mehrotra et al. proposed an algorithm that involves selection of feature vector based on the clinical and metabolic features. And also statistically significant features for discriminating between normal and PCOS groups are selected based on two sample t-test. To classify the selected features Bayesian and Logistic Regression (LR) classifiers are used. Performance of Bayesian classifier is better than the logistic regression. The overall accuracy of Bayesian classifier is 93.93% as compared with logistic regression i.e. 91.04%.[9]

Shruthi Mahalingaiah et al. have proposed and evaluated the performance of a rules-based classifier and a gradient boosted tree model for automatic feature extraction and classification of polycystic ovary morphology (PCOM) in pelvic ultrasounds. The accuracy of rules-based classifier (RBC) was 97.6% and 96.1% for the gradient boosted tree model (GBT). [10]

2.1 Drawbacks and observations

Many authors have worked on the detecting the ovarian follicles from the ultra-sound images of the ovaries, by applying various image processing algorithms like noise removal and segmentation. They have considered morphological and stereological features for this purpose. Many have used various classification algorithms like SVM, logistic regression, naïve bayes, decision trees etc to classify the images into PCO or non-PCO classes based on the features extracted from image pre-processing. Some authors have used morphological and clinical parameters for classification. But the parameters selected by them require the patient/women to undergo various clinical tests and ultrasound scanning, which can be hectic for some patients in terms of time and money.

To address the above mentioned drawback, we have tried to understand various clinical and morphological parameters required for diagnosing PCOS and then selected only those parameters which are low cost and also optimal for early detection of PCOS or can act as a caution for the women to

not ignore the symptoms and consult a doctor or make the required diet or lifestyle changes to prevent further complications. These features are then used to train and test the prediction model.

3. METHODOLOGY

For building an efficient machine learning model for PCOS risk prediction, comparison of performance of various existing algorithms on the dataset must be carried out. Steps that are involved in building an efficient model and tuning it for improving the model's performance, is given below.

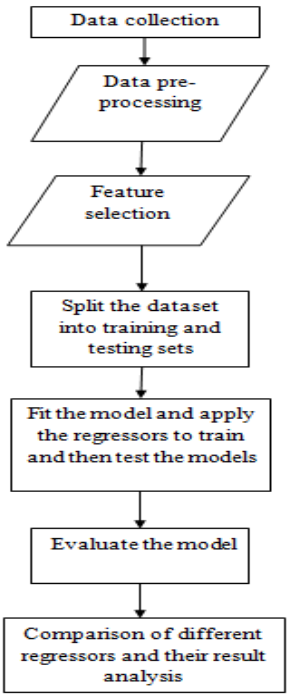

Fig -1: Workflow of the proposed prediction model

Step 1: Dataset collection
Related data was collected from an infertility clinic and research centre.

Step 2: Data pre-processing and feature selection
The data could have some null values and some values which may not be reliable to train the model in a correct way. So to handle such values data cleaning must be done.

Step 3: Feature selection
For the detection of PCOS various clinical and morphological criteria are involved, each person with PCOS can have different combination of symptoms and some symptoms may be acute which makes the diagnosis of PCOS difficult. There is no one particular test to confirm whether the person has PCOS or not, it requires various tests and scans to diagnose and confirm. Studying all such criteria we have selected some important features that are also low cost and optimal to detect PCOS in its early stages. It may not be sufficient to perfectly diagnose PCOS but is enough to warn the person on whether to be concerned about their symptoms and whether they require medical intervention. It can act as a guide to track their symptoms and in detecting PCOS in an early stage.

Step 4: Split the dataset into training and testing sets
Split the pre-processed dataset into train and test sets, for example 80% of the dataset as training set and 20% as test. Training sets are used to train and tune the models. Training set will be used for training the model for predicting the risk of PCOS. Cross-validation is a method for checking the model performance using only the training data. Test sets will be set aside as "new" data to evaluate the models. And after the training phase is complete, the testing set will be used for testing the model on its prediction ability when it is introduced with fresh data.

Step 5: Apply the regressors to train and fit the model and then test the models
Apply various regressors (linear regression, KNN, random forest) on the previously split train set to train and fit the model. Model fitting is a gauge of how well a machine learning model relates to similar data to that on which it was trained. A well fitted model will have high accuracy. And then test the model to check its prediction capability. Verify the predicted output value by comparing it with the actual output value.

Step 6: Evaluate the model
Model evaluation aims to estimate the generalization accuracy of a model on future (out-of-sample) data. Model evaluation metrics are necessary to check how well the model performs. Some of the metrics used are R^2, RMSE, MAE.

Step 7: Result analysis - comparison of the models and selecting the best model
Select the best model after analyzing the results of performance metrics applied on each model.

4. IMPLEMENTATION

Data collected from infertility clinic and research centre contains physiological and metabolic features that are the signs and risk factors of PCOS and are given as input into the regression model. The proposed risk prediction model is demonstrated in Figure 2.

After performing the literature survey, the below mentioned features were selected as they were found to be optimal and low cost for evaluating the risk of developing PCOS.

Independent features:
CASE 1**:** Hirsutism, Family history of PCOS, BMI, Fast food (per week), Menstrual disorder.
CASE 2: Testosterone, Hirsutism, Family history of PCOS, BMI, Fast food (per week), Menstrual disorder

Dependent feature: Predicted risk (in terms of %)
These features/symptoms do not need any kind of clinical or ultrasound tests to be identified, whereas Poly-Cystic Ovary morphology (PCOM - Ultrasound scan findings) criteria was dropped as it would need the person to go under ultrasound scanning and would add more to the cost and time.

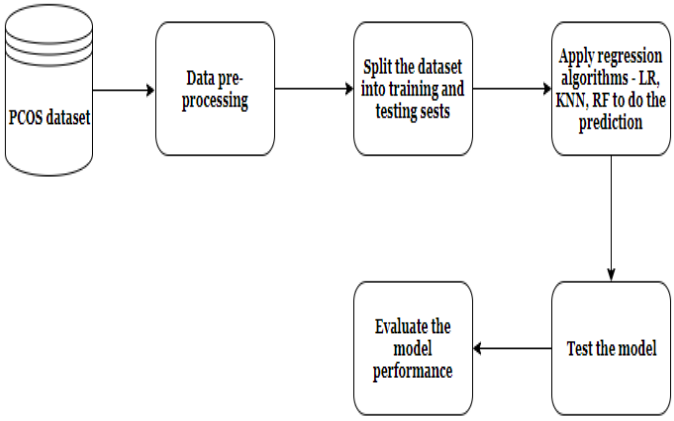

Fig -2: Proposed system model

If it is convenient for the individual, she can get a clinical test done to measure the testosterone level in her body; this can add more accuracy in identifying hyperandrogenism. This can also be added as one of the independent features while training the model. The developed and trained model is capable of predicting (i.e. "risk" in terms of %) that how likely a woman is to develop PCOS in the future, based on the collected information on her symptoms, lifestyle and genetic factors.

Testoster	Hirsutism	Family his	BMI	Fast food	Menstrual	Predicted risk
72.43	yes	yes	30.3	5	yes	100
62.3	no	no	21.6	0	no	0
75.97	yes	no	35.2	4	yes	87
82.62	yes	yes	38	4	yes	100
37.35	no	no	18.1	2	no	0
46.38	no	no	20.5	1	no	0
69.71	no	yes	22.7	2	no	13
67.83	no	no	22.3	1	no	0
74.33	yes	yes	29.2	6	yes	100
61.26	no	no	23.9	2	no	0

Fig -3: PCOS dataset – selected features

Various regression models vary based on the kind of relationship between output and input variables that are considered and the number of input variables given. We have used the following regression algorithms – linear regression, KNN and random forest for performing the prediction task.

4.1 Linear regression

Linear regression is commonly used algorithm for finding linear relationship between output variable and one/more input variables. In this PCOS dataset, the independent variables have a linear relation with the dependent variable so multiple-linear regression can be applied on this dataset for predicting the risk. This technique uses statistical calculations to plot a regression line that fits best in a set of data points. It may seem as there are several other popular algorithms available, but linear regression is more useful and commonly used statistical learning method. There are 2 types of linear regression models: simple & multiple. Simple linear regression describes the relationship between single output and single input variable. Multiple-Linear Regression describes the relationship between one continuous output variable and 2 or more input variables [11]. Since the dependent variable "risk" is dependent on 6 independent variables i.e, Testosterone, Hirsutism, Family history, BMI, Fast food, Menstrual disorder. So here multiple-linear regression algorithm has to be used for training the model. Mathematical representation of linear regression is as follows:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + \varepsilon$$

Where, $\hat{Y}$ = dependent variable, ε is the error
X_1 to X_p are p distinct independent variables
b_0 is the value of Y when all independent variables equal to zero and b_1 to b_p are the estimated regression coefficients.
While training the model we are given dependent variable and independent variables used for training. Model fits the best line to forecast the value of target variable for the given value of input variable. The model produces the best regression fit line by finding the best intercept and coefficient values.

Algorithm steps:

Step 1: The model assumes that output is a linear function of the input variable.
Step 2: For every data point calculate the respective intercept (b_0) value and co-efficient (b_p) for each independent variable which determines the slope of the regression line.
Step 3: Repeat step 2 until cost function is minimized and best values for b_0 and b_p are found.
Step 4: Finally the best fit regression line for all the data points is obtained.
Step 5: Predict the output value for the new input data using the obtained values.

4.2 K-nearest neighbors

The k-nearest neighbors (KNN) algorithm is a non-parametric, easy-to-implement supervised machine learning algorithm that can be used for both classification and regression problems. It works on the idea that similar things exists in close proximity, this helps in predicting the output value based on the similarity in the inputs [12]. There are various measures like Euclidean, Manhattan, Hamming

distance available to calculate the distance between two data points.

Algorithm steps:

Step 1: The gap between the new data point and every training data point is found using Euclidean distance formula.

$$Euclidean = \sum_{i=1}^{k}(u_i - v_i)^2$$

Where, u and v are data points

Step 2: The distance values obtained are sorted in ascending order.
Step 3: The nearest "k" data points are selected (top k values from the sorted list).
Step 4: The mean of these data points is the ultimate output value for the new data point.

4.3 Random forest

Random Forest is one of the most mainstream models, which can be used for regression and classification problems. The random forest model is a type of additive model that calculate predictions by combining and merging different decisions from a set of base models (an approach known as ensemble model). Base models are nothing but set of decision trees generated for the input data. In the Random Forest, all base trees are constructed independently using different features and data points. Decision tree is a graphical representation of all possible solutions to a decision based on certain condition. It is a tree where each node represents a feature, each branch represents a decision and each leaf represents an output [13, 14].

The Iterative Dichotomizer (ID3) algorithm is used in constructing the decision tree. The algorithm iteratively separates the attributes into 2 groups i.e., the most dominant attributes and the others to construct a tree. The entropy and information gain of feature is calculated and the feature with greater information gain is chosen as the most dominant feature and is placed as the root node in the tree. Then, entropy and information gain is calculated again among other attributes so the next most dominant attribute is found. This process continues till a decision for that branch is reached [15, 16]. Entropy is the variance present in the feature values.
Entropy can be calculated as:

$$Entropy = -\sum p(x) log_2 p(x)$$

where p(x) is the fraction of examples of a given class.

Information gain is a factual strategy that estimates how well a given feature isolates the training examples according to their target value. A feature with greater information gain is considered to be the best because it can divide the training examples into their respective classes very well, so it can be chosen as the attribute we can base our decision upon. The algorithm performs splitting on attributes till entropy becomes zero; that is when it has reached the final outcome. Information gain at any node is defined as:

*Entropy (parent) – [(weighted average) * entropy (children)]*

Algorithm steps:

Step1: Pick N arbitrary instances from the dataset.
Step 2: Without pruning construct a decision tree based on these N instances
Step 3: Determine the number of trees you desire in your forest and repeat steps 1 and 2.
Step 4: In regression algorithms, when a new instance is given, every tree produces an output value.
Step 5: The ultimate output for target variable is obtained by taking the average of all the values produced by every trees in the forest.

4.4 Model evaluation using performance metrics

After training the model, it is important to test and check how well the model is performing. Various metrics are available for this purpose. Regression algorithms has evaluation measures like, Mean Absolute Error (MAE), R-squared and Root Mean Squared Error (RMSE), Cross-validation score (CV score) etc to check the performance of the model.

Co-efficient of determination known as R-squared (R^2) is a measure used in statistics to see how good a fit of a set of predictions to the actual values is. R^2 value ranges from 0 to 1. Higher the R-squared value, the better.

$$R^2 = 1 - \frac{SSerror}{SStotal}$$

Where, SSerror is the sum of squares of prediction errors. SStotal is the overall sum of squares of errors from the average model.

RMSE is the square root of the mean of squared errors (MSE). Here the error implies the distinction between the real and anticipated values. Root Mean Square Error (RMSE) is the standard deviation of the prediction errors.
In linear models, prediction errors are a proportion of how far the data points lie from the regression line. Lesser the RMSE value, prediction will be more precise. The lower the RMSE better is the performance of the model. [17]

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Where, N is the number of observations

MAE known as Mean Absolute Error is utilized for summing up and evaluating the nature of machine learning model. It alludes to the mean of the absolute values of each prediction error on all instances of the test data-set. Lesser the MAE value, prediction will be more precise.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | x_i - x |$$

Where,
n = number of errors,
$\Sigma |x_i - x|$ = summation of all the absolute errors.

5. RESULTS AND ANALYSIS

The three trained models are evaluated using performance measures and the results are compared to find the best model for PCOS risk prediction. In both cases, it is observed from the below table that Random forest regression has the lowest MAE, RMSE values and higher R^2 which implies that it is the best among all the three algorithms. But this result was obtained after scaling the data and tuning certain parameters. Therefore, if used wisely, Random forest regression can outperform other two models used in this experiment.

Table 2: Results comparison for case 1

CASE 1	Algorithm	R^2	MAE	RMSE
i.	Linear regression	0.978	3.282	3.930
ii.	KNN	0.970	1.81	4.54
iii.	Random forest	0.985	1.556	3.079

Table 3: Results comparison for case 2

CASE 1	Algorithm	R^2	MAE	RMSE
i.	Linear regression	0.980	3.079	3.909
ii.	KNN	0.985	1.789	3.463
iii.	Random forest	0.986	2.436	3.132

6. CONCLUSION AND FUTURE WORK

Diagnosing PCOS (especially in its initial stages) is a crucial real-time medical problem. Methodical steps are followed in planning this prediction model, which evaluates the risk of an individual in developing PCOS in the future. It also helps in detecting PCOS in its early stages. Training and testing of models are performed on PCOS dataset with the following features i.e., Testosterone, Hirsutism, Family history, BMI, Fast food, Menstrual disorder, Risk. Three machine learning regression algorithms - linear regression, KNN and Random forest, are studied, applied and evaluated on various metrics like R^2, MAE, and RMSE. Random forest algorithm outperforms the other two algorithms by achieving less error values i.e. average of 1.99 (MAE) and 3.10 (RMSE), and highest R^2 value i.e. average of 0.985. Therefore it can be concluded that Random forest algorithm can be a powerful algorithm if used wisely by appropriately tuning its parameters.

This developed prediction model can be improved by applying deep learning algorithms and can also be applied on live data.

ACKNOWLEDGEMENT

I would like to thank my college JSS S&TU and my guide Dr.M.A.Anusuya for her valuable words of advice and support. I am also thankful to my friends without whom this would not have been possible.

REFERENCES

1. https://www.healthline.com/health/ovarian-cysts
2. https://www.thehindu.com/sci-tech/health/one-in-five-indian-women-suffers-from-pcos/article29513588.ece
3. https://www.huffpost.com/entry/frustrating-facts-about-pcos_b_7686030
4. Sachdeva, G., Gainder, S., Suri, V., Sachdeva, N. and Chopra, S.,2019. "Obese and non-obese polycystic ovarian syndrome: Comparison of clinical, metabolic, hormonal parameters, and their differential response to clomiphene". Indian journal of endocrinology and metabolism, 23(2), p.257.
5. Zhang, X.Z., Pang, Y.L., Wang, X. and Li, Y.H., 2018. "Computational characterization and identification of human polycystic ovary syndrome genes". Scientific reports, 8(1), p.12949.
6. Gulam Saidunnisa Begum, Atiqulla Shariff, Ghufran Ayman, Bana Mohammad, Raghad Housam, Noura Khaled, "Assessment of risk factors for development of polycystic ovarian syndrome" International Journal of Contemporary Medical Research 2017;4(1):164-167.
7. Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram C, Remya George. "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques" 2019 IEEE Region 10 Conference (TENCON 2019)
8. Sharvari S. Deshpande ,Asmita Wakankar, "Automated Detection of Polycystic Ovarian Syndrome Using Follicle Recognition", 2014 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

9. Palak Mehrotra, Jyotirmoy Chatterjee,Chandan Chakraborty, "Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques" in 2011 Annual IEEE India Conference.
10. J. Jojo Cheng and Shruthi Mahalingaiah, "Data mining polycystic ovary morphology in electronic medical record ultrasound reports" in Fertility Research and Practice (2019)
11. https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86.
12. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
13. https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/
14. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
15. https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567
16. https://www.geeksforgeeks.org/decision-tree-introduction-example/
17. https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/