Polycystic Ovary Syndrome Analysis Using Machine Learning Algorithms

Namitha T S
Department of Computer Applications
Amal Jyothi College of Engineering, Koovapally
Kottayam, Kerala.
namithats2022@mca.ajce.in

Meera Rose Mathew
Asst. Professor in Computer Applications
Amal Jyothi College of Engineering, Koovapally
Kottayam, Kerala
meerarosemathew@amaljyothi.ac.in

Abstract **- Polycystic ovary syndrome or PCOS is a hormone A common disease among women of reproductive age. This once diagnosed, it cannot be cured. Help to avoid its effects. The exact cause of PCOS is still unknown. But there are some factors that illustrate the possibility of PCOS. The factors that cause this syndrome are : obesity and insulin, immunity , blood pressure , depression , inflammation. Symptoms include : hirsutism, oligo-ovulation, acne, heavy bleeding, darkening of the skin. Causes and uses a model is developed to accept the symptoms and their characteristics. Machine Learning models used for supervised classification are K-Nearest Neighbor and logistic regression. The reason multiple models were built behind the scenes to identify the best one for a given dataset , the known extent of knowledge.**

Keywords **- Machine Learning , KNN Algorithm , Logistic Regression, SVM, Decision tree , Random forest, CatBoost Classifier.**

I INTRODUCTION

Polycystic ovary syndrome Women during childbirth .
The female reproductive organs, known as the ovaries, generate progesterone and estrogen-hormones, which regulate and are affected by the menstrual cycle. Androgen, sometimes known as male hormones, is produced in minute amounts in the ovaries.
The basic features of PCOS are,
Cysts in the ovaries.
High levels of the hormone: androgen.
Irregular periods
Excessive body hair growth
Hair loss
Darkening of skin
Headaches
Acne or Oily skin
Since this condition is a syndrome, it has a collection symptoms indicating its presence. These symptoms play an important role in diagnosing this condition. With these symptoms, the cause may also increase the risk of acquiring the illness in question. Early detection of PCOS is essential due to the risk of infertility and diabetes, this is a possibility.
Endometrial cancer and heart disease followed
Stage of the condition .Here, two machine learning models are built to determine the presence of PCOS. Supervised machine learning techniques are utilised because the dataset classifies whether or not the condition occurs calling:K-Nearest Neighbor (K-NN), Logistics Regression. The first is distance-based technology the second is based on probability, which is why both uses polarization techniques, their compares accuracy . Logistic regression is high The accuracy is 92% and the KNN is 90.74%. This paper is as follows: Describes methodology, Results obtained by method and conclusions made with references respectively. We check the best algorithm and accurate value. This paper we use ML models used for Supervised classification are K-NN algorithm, Logistic regression ,SVM , Desion tree, Random forest and CatBoost Classifier.

II MACHINE LERANING ALGORITHMS

Machine Knowledge is a sort of Artificial Intelligence that enables computers to learn and develop without being explicitly programmed.Machine learning is concerned with the creation of computer programmes that can pierce data and analyse it.The knowledge process begins with obediences or data, analogous as direct experience or suggestion, to look for patterns in the data and make better opinions in the unborn predicated on the samples we give. The primary thing is to allow computers to learn on their own without mortal intervention or backing and to adjust their exertion accordingly. Logistic Retrogression, It's used to calculate discriminational values, double values similar as0/1, yes/ no, true/ false grounded on a set of free variable.Simply put, By linking data to a logit function, it estimates the liability of an occurrence. As a result, logit retrogression is another name for it. Because it predicts probability, with values ranging from 0 to 1, it is referred regarded as a probability forecaster.
The decision tree approach is a type problem-solving supervised knowledge tool. It can now be used to calculate continuous and racial dependent variables.The population in this method is divided into two or more clothing sets.
To create as many different groups as possible, this is done based on the most essential attributes/independent variables.
This is a type system, and it's called SVM for short. Each data item is compassed as a point in n-dimensional space in this procedure, and the value of each point represents the value of each data item.Still, it's further extensively used in

DOI: 10.5281/zenodo.6362178
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

bracket problems in assiduity. K Neighbors is a simple algorithm that uses the maturity votes of its k neighbours to store all available and new examples.
A set of decision trees known as Random Forest is a trademarked name.In the Random Forest, we've a collection of evanescent trees (known as" timbers"). Each tree is given a bracket to classify a new item based on attributes, and the tree" votes"for that order. The timber selects the bracket that receives the most votes.
Gradient Boosting Algorithms are,XGBoost
The decisive choice between palm and master in some Kaggle matches. XGBoost has a veritably high prophetic power, which makes it an excellent choice for delicacy in events due to its direct model and tree literacy algorithm, making the algorithm 10 times faster than being grade supporter ways.
Yandex introduced CatBoost, an open-source machine literacy algorithm.It can be fluently integrated with in-depth literacy fabrics similar as Google's TenserFlow and Apple's Core ML.
The stylish part about CatBoost is that it, like other ML models, doesn't bear expansive data training and can operate on a variety of data formats; It doesn't weaken how strong it's before pacing with the perpetration, make sure that the lost data is handled well .

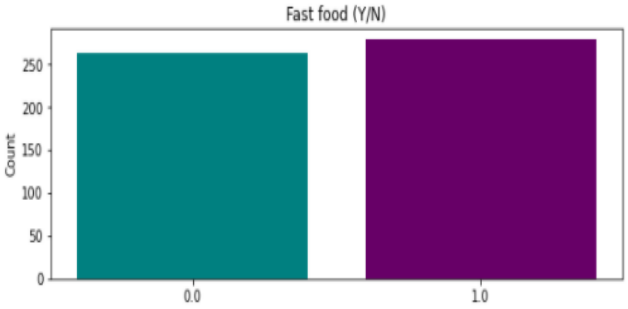
III DATASET DESCRIPTION

To build an appropriate machine learning model, the data of a dataset must pass through a sequence of steps. It takes time to become one filtered and silent input to the algorithm
Data Collection

1. Data Analysing
2. Feature Selection
3. Fitting in to model
4. Making Predictions
5. Evaluvation of models
6. Comparison
7. Selection of best model

a)Data Collection
This is a necessary step before beginning the data collection.Various Platforms are available for this. It contains Samples from different hospitals in Kerala, India (Downloaded from Kaggle)Two datasets are here, PCOS with fertility and PCOS without fertility.

b)Data analysing
Data analysing based on the target value. Analyzing based on the Age ,Weight ,Height, BMI Blood group ,pulse rate(bpm) , RR(breaths/min) ,Hb(g/dl) cycle(R/I) ,pimples (Y/N) ,fast food (Y/N) , Reg. Exercise(Y/N) ,BP-Systolic(mmHg), BP-Diastolic, follicle no.(L) ,follicle no.(R) ,Avg F Size(R) , Avg F Size (L) ,Endometrium (mm).
Here is a sample,

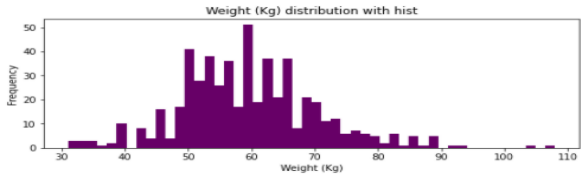


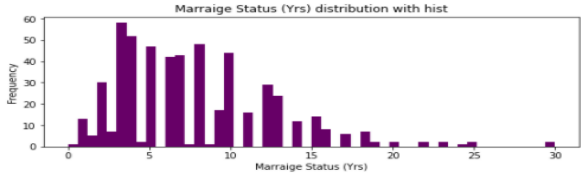
Fast Food (Y/N)
1.0 278
1.0 262
Name: Fast Food (Y/N) , dtype :int64

c)Features of selection
And to improve the performance of the model Reduce computational costs, with only selected attributes.Acts as a feature of samples. The filter method is used.Find the weight of the features to determine.Which of them has the highest relationship with the goal.

Features	weights
fastfood	0.3802
Weight gain	0.4417
Follicle (L)	0.6010
Follicle (R)	0.6506
Skin darkening	0.4796
Cycle(R/I)	0.3997
Pimples	0.2869
Hair growth	0.4646
AMH(ng/ml)	0.2602
Weight(kg)	0.2060
BMI	0.1955
Hair loss	0.1750
Avg. F Size (L)	0.1265

Some features and analysis are :



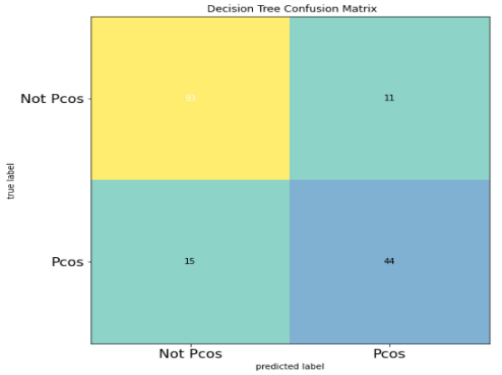


From the list above, we can Parameter follicle number (R) and Follicle number (L) is the highest weight,
i.e. Determine the number of follicles on the right and left Ovaries independently.Features such as darkening

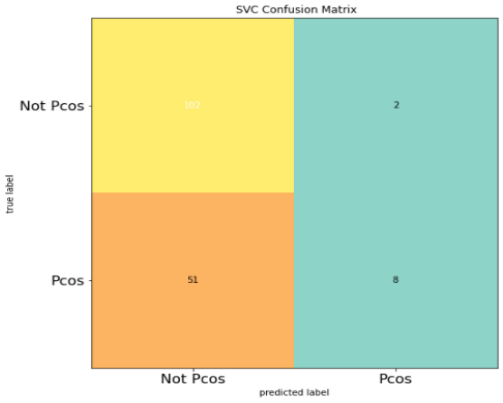
DOI: 10.5281/zenodo.6362178
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

of the skin , Hair Growth , Weight Loss, Fast Food , Acne,
Hair loss contains values of 0 or 1, 0
1 indicates the absence of that particular feature Presence. Uses AMH or anti-mullerian horm one Indicator of egg number. Its unit is the nanogram per Ml. Below, BMI indicates body mass index respectively. The CF patients' average weight and height were.Inches, along with waist and waist sizes the size of the follicle on the left ovary
and measured in millimeters.

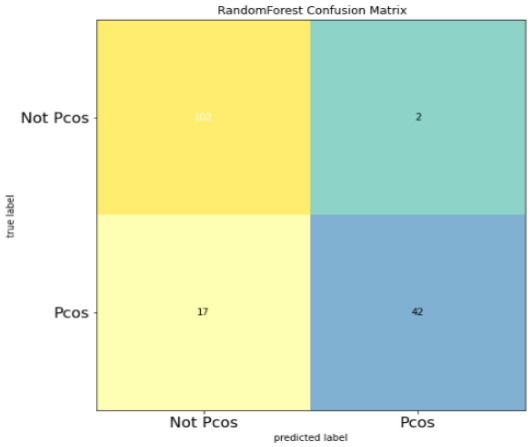
d) Fitting in to models predictions
Decision tree Analysis
It's a type of supervised knowledge algorithm. Suddenly, it also works for racial and non stop dependent variables. In this algorithm, we divide the population into two or farther outfit sets. This is done predicated on the most important attributes/ independent variables to make as multitudinous different groups as possible.



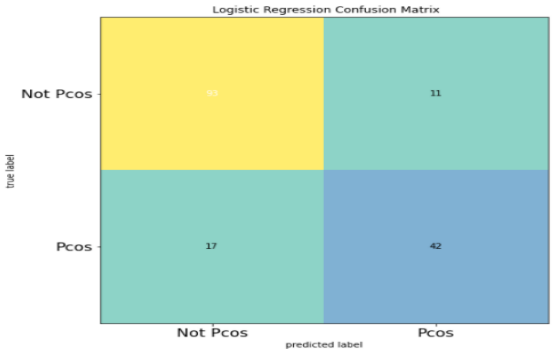
SVC Analysis
A collection of n points in m-dimensional space, where x-a denotes the value of point I on axis a. This formula is straightforward to apply and can be used to calculate distances over a wide variety of lengths.



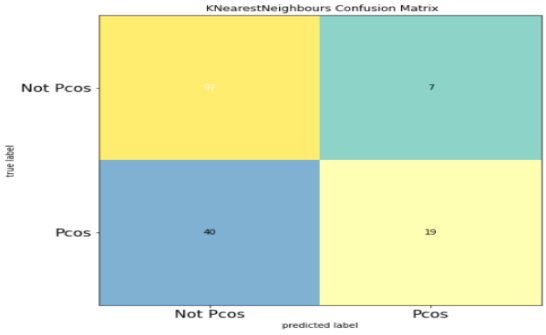
Random Forest Analysis
A cluster of decision trees known as a RF is a trade marked name. A group of deciduous trees can be found in the Random Forest (known as" timbers"). Each tree is given a type in order to classify a new object based on attributes, and we say the tree" votes"for that order. The timber chooses the type that has received the most votes.



Analyze the data using logistic regression
It's used to calculate discriminational values, which are double values such as 0/1, yes/no, true/false, and are based on a set of free variables (s). Simply described, it attaches data to a logit function in order to estimate the liability of an occurrence. As a result, logistic retrogression is a term used to describe it. Its affair values range from 0 to 1 because it forecasts probability. P =1/1+e-(a+bx) Sigmoid function. P – probability and a,b are model parameters.



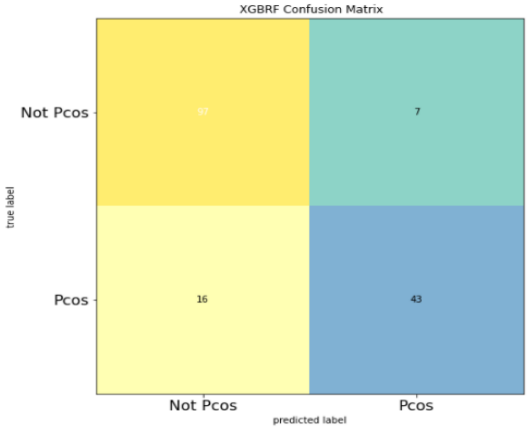
KNN Analysis
It can help with type and retrogression issues. Despite this, it's commonly used in assiduity-related disorders. KNN is an algorithm that maintains track of all open cases and categorises new ones based on their maturity votes.The distance function near the K neighbour is most commonly used to measure the case allocated to the class.



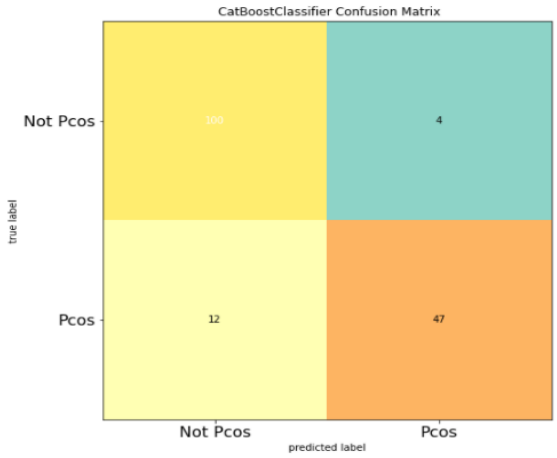
XGBRF analysis,
Grade Boosting Algorithms are XGBoost
In various Kagglematches, the final decision between Random palm and master. Because of its direct model and tree literacy algorithm, XGBoost has a very high

DOI: 10.5281/zenodo.6362178
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

predictive power, making it an ideal choice for events with a lot of nuance.

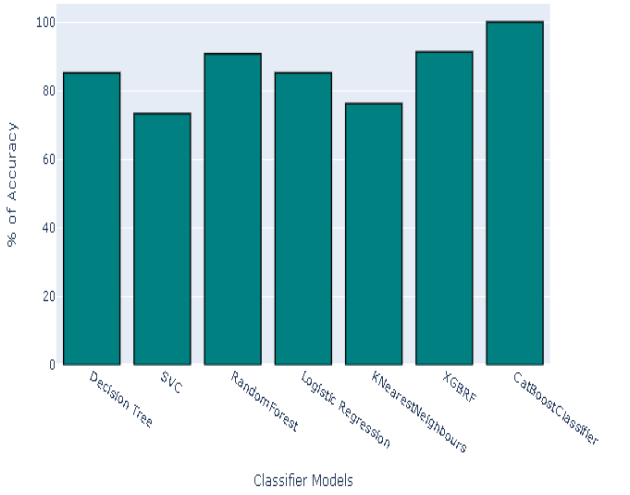


CatBoost an open source ml algorithm introduced by Yandex.It can be fluently integrated with in- depth literacy fabrics similar as Google's TenserFlow and Apple's Core ML. The stylish part about CatBoost is that it, like other ML models, doesn't bear expansive data training and can operate on a variety of data formats; It doesn't weaken how strong it is .Before proceeding with the perpetration, make sure that the lost data is handled well Catboost can handle sort variables automatically without showing any conversion error, which allows you to concentrate on tuning your model more rather of fixing minor crimes.



IV ACCURACY OF DIFFERENT MODELS

Result of the analysis



V CONCLUSION

In this paper we have applied different Machine Learning algorithms in order to analyze PCOS from the data set.We were able to create different classifier models using different machine learning algorithms.With this analysis we were able to analyse the model with the highest accuracy.Of all the models , the CatBoost had the highest accuracy.Therefore, with methodology and techniquesMachine learning, we determined successfully Best supervised classification model for detection Polycystic ovary syndrome.

VI REFERENCES

[1] DETECTION OF POLYCYSTIC OVARIAN SYNDROME USING FOLLICLE RECOGNITION TECHNIQUE IN SCIENCE DIRECTHTTPS://WWW.SCIENCEDIRECT.COM/SCIENCE/JOURNAL/2666285X

[2] Int. J. Trend Sci. Res. Dev. (IJTSRD) (2019) Dipamoni Morang, Pankaj Chasta, K. Kaushal Chandrul A Review on “Polycystic Overy Syndrome (PCOS)

[3] Supervised Learning-Classification Using K-Nearest Neighbors (KNN). (2019). Python® Machine Learning, 205–220. doi:10.1002/9781119557500.ch9

[4] Bichler, Martin and Kiss, Christine, "A Comparison of Logistic Regression, k-Nearest Neighbor, and Decision Tree Induction for Campaign Management" (2004). AMCIS 2004 Proceedings. 230

[5] Research Gate https//www.researchgate.net/publication/354245156 An Analysis on the Implementation of PCOS

DOI: 10.5281/zenodo.6362178
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam