Advanced Mining Techniques Project
PCOS Prediction

Introduction

In recent times, there has been a lot of conversations around various diseases that affect a person's health physically and emotionally at a very slow pace. Some of these are clinical depression, anxiety issues, PCOS, etc. PCOS is a disorder that occurs in females. There is still a lot of research happening on this particular topic. Research about the causes, symptoms, effects, health issues, etc. There are many variables to determine if a person is suffering from PCOS or not. The most difficult part is to predict if the person is suffering from PCOS because initially there may be only minor effects and it may gradually increase resulting in a serious disorder called the PCOD(Polycystic Ovary Disorder).

Project Abstract

This project has been created to help solve the above problem. We have tried to create a model that will be able to predict if the person is suffering from PCOS or not. The data that has been used and the variables that we have taken for prediction are more towards the daily lifestyle of a person. This project helps a person to predict if she is suffering from PCOS by just answering some questions that are related to the symptoms and causes of this. We have created a form, where the user will fill in their respective details and in the backend, the prediction model will take those details as the inputs and run through the model and help us to predict if the person is suffering or not. The accuracies which we have obtained are fairly good, but still, the model cannot detect successfully each time.

Polycystic ovary syndrome (PCOS):

Polycystic ovary syndrome (PCOS) is a hormonal disorder common among women of reproductive age. Women with PCOS may have infrequent or prolonged menstrual periods or excess male hormone (androgen) levels. The ovaries may develop numerous small collections of fluid (follicles) and fail to regularly release eggs.

The exact cause of PCOS is unknown. Early diagnosis and treatment along with weight loss may reduce the risk of long-term complications such as type 2 diabetes and heart disease.

Symptoms:

Signs and symptoms of PCOS often develop around the time of the first menstrual period during puberty. Sometimes PCOS develops later, for example, in response to substantial weight gain.

Signs and symptoms of PCOS vary. A diagnosis of PCOS is made when you experience at least two of these signs:

1. **Irregular periods**: Infrequent, irregular or prolonged menstrual cycles are the most common sign of PCOS. For example, you might have fewer than nine periods a year, more than 35 days between periods, and abnormally heavy periods.
2. **Excess androgen**: Elevated levels of male hormones may result in physical signs, such as excess facial and body hair (hirsutism), and occasionally severe acne and male-pattern baldness.
3. **Polycystic ovaries:** Your ovaries might be enlarged and contain follicles that surround the eggs. As a result, the ovaries might fail to function regularly.

PCOS signs and symptoms are typically more severe if you're obese.

Dataset explanation:

The Dataset collected from the google forms had following feature columns:

1.**age:**Age of the Candidate.

2.**residence_area:**Candidate's Residence Area from the following choices:

a.Rural area

b.Urban area

C.semi-urban area

3.**job_type:**Candidate's Job type from the following choices:

a.Working women

b.Student

c.Housewife

4.**job_physical_activity:**Does the Candidate's Job involve Physical Activity?

5.**Stress Levels**:What is the level of stress a candidate experiences on an everyday basis on a rating from 0-5?

6.**Sleep Durations :**How many hours Does the Candidate sleep, ranging from 4-10 hours/day?

7.**sleep ratings:** How would the candidate rate her quality of sleep on a rating from 0-5 ?

8.**Sleep Timings :**At what time does the Candidate sleep from the given time ranges?

9.**Exercise Hours:** How many hours does the candidate exercise in a week?

10.**Junk Food :**How many times does the candidate eat junk food in a Week?

11.**smoking:**Does the Candidate Smoke?

12.**alcohol:**Does the Candidate Consume Alcohol?

13. **sedentary age:**At what age did the Candidate's physical activity(playing, dancing, swimming , etc) significantly reduced? {click on continued if you are still doing any physical activity}

14.**Period_startage :** The age at which the Candidate's period started?

15.**Period Status :** Is the Menstrual Cycle Regular of the Candidate?

16.**periodDuration:**What is the average duration of your menstrual periods of the Candidate?

17.**pms:**Does the Candidate suffer from premenstrual syndrome?

18.**MalePattern_you:**Is the Candidate suffering from Male-pattern baldness or thinning hair?

19.**delayInPeriods**:What is the period of delay in the Candidate's menstrual cycle?

20.**MalePattern_heredity**:Does the Candidate have a family history of Male Pattern Balding in Women?

21.**periodstatus_heredity:**Does the Candidate have a family history of Irregular Periods in Women?

22.**currentMedications**:Is the Candidate currently on any Medications?

23.**milkType**:What is the Milk Type the Candidate consumes?

24.**milkQuantity**:What is the Quantity of Milk the Candidate consumes in a day (approx. in ml)?

25.**exerciseAfterDiagnosis**:Has the Candidate started Exercising After being Diagnosed with PCOS?

26.**consumptionOfOrganicFood**:Does the Candidate Consume Organic food?

27.**height:**Candidate's height {in cm's}?

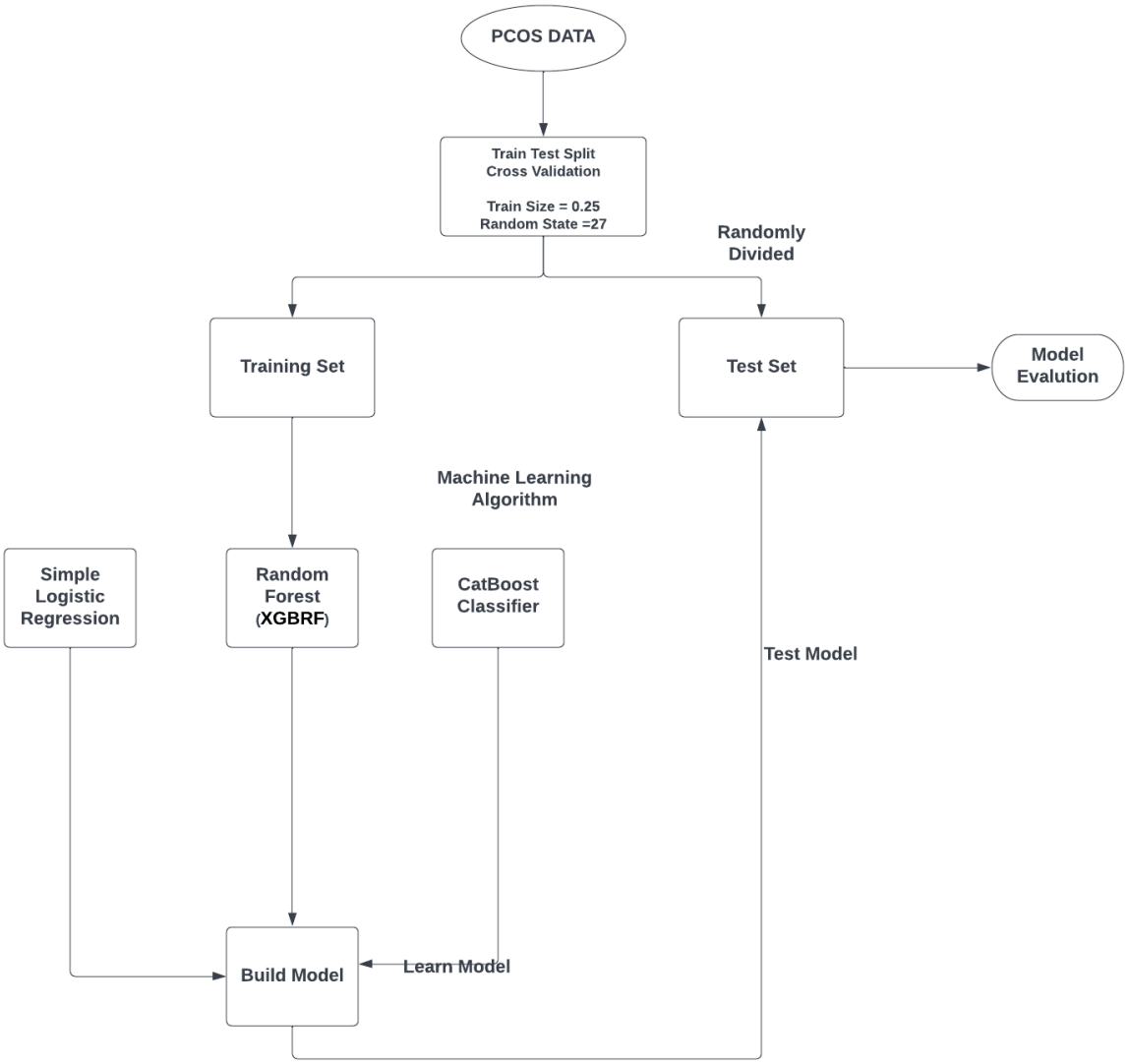
28.**weight**:Candidate's weight {in kg's}?

29:**Bmi:**Bmi has been calculated from the height and weight of the candidate.

Target Column:**sufferingFrompcos:**Is the Candidate Suffering From PCOS?

Modelling

In this step, nine different models are used on the pre-processed data. We implemented machine algorithms such as Simple Logistic Regression, Random Forest, CatBoost Classifier as baseline approaches on the pre-processed PCOS dataset. Random Forest(XGBRF) and CatBoost is the novelty of this paper for detecting PCOS.



- **Simple Logistic Regression:**
  - ➢ This model regulate the relationship among independent variables and binary outcomes based on probability as forecast value of dependent variable.
  - ➢ In this paper, every feature is tested and allocated a probability which is used to classify the PCOS as Normal women or PCOS Women.
  - ➢ If the probability is higher than threshold it is PCOS women else Normal women.
  - ➢ The equation of Logistic Regression is as follows:

$$\Pi(x) = 1/1 + e^{-y}$$

➢ Here y represents coefficients of variable and e is Euler's number. If Π(x) is higher than 0.5 then it is considered as Home win else as Away win.

- **Random Forest (XGBRF):-**
  - ➢ This model was developed by Breiman in 2001. It initiates both the procedure of random feature selection and bagging idea.
  - ➢ The construction of Bagging method is done to calculate the distribution of estimator based on sampling and with replacing from real dataset.
  - ➢ In bagging model, n sample size is taken from training data, bagging model produce new data using the sampling and replacing the actual dataset with n sample size.
  - ➢ On the other hand, procedure of random feature selection authorise random feature subsets in every node during splitting in the trees in such a way that diversity of base method may be observed.
  - ➢ Both, Bagging and Random feature selection improve accuracy during prediction. The variance of Random Forest is calculated as follows:

$$\rho\sigma^2 + \frac{1-\rho}{K}\sigma^2$$

  - ➢ Here σ 2 denotes tree variance, ρ denotes the correlation between trees, K represents total trees.

- **CatBoost Model:**
  - ➢ CatBoost is a Machine learning model which uses gradient boosting on decision trees. It uses a schema of estimating leaf values when choosing a tree structure, which helps to overcome the over-fitting problem.

- ➢ It has four principal merits, first one is creative model for computing the categorical features which means there is no need for processing features on your own - it is constructed out of the box.
- ➢ For a dataset having categorical features results like accuracy if greater than other algorithms [Li et al., 2020]. Implementation of direct boosting, a permutation - driven different to other classic boosting models.
- ➢ On small datasets gradient boosting causes over-fitting while there is special modification based on CatBoost for such cases.
- ➢ CatBoost makes it fast and easy use of GPU implementation training and at last it produces missing value great support visualisation.

Technologies used

- **VSCode:**
  - ➢ Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, macOS and Linux.
  - ➢ It comes with built-in support for JavaScript, TypeScript and Node.js and has a rich ecosystem of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity).
- **Python**
  - ➢ Python is an interpreted high level programming language used for a wide host of functionalities.
- **Streamlit**
  - ➢ Streamlit is an open-source python library for creating and sharing web apps for data science and machine learning projects. The library can help you create and deploy your data science solution in a few minutes with a few lines of code.

➢ Streamlit can seamlessly integrate with other popular python libraries used in Data science such as NumPy, Pandas, Matplotlib, Scikit-learn and many more.

➢ Note: Streamlit uses React as a frontend framework to render the data on the screen.

Libraries used:

Numpy, Pandas: Data preparation, data cleaning and data manipulation.
Matplotlib, Plotly, Seaborn: For plotting Graphs and data visualization
Sklearn: used for splitting data and creating machine learning models
xgboost and catboost: for using respective models.

Code Snippet:

```
[ ]  1 # data preparation and cleaning
     2
     3 df['sufferingFrompcos']
     4 df.sufferingFrompcos.replace(('Yes', 'No'), (1, 0), inplace=True)
     5 df.smoking.replace(('Yes', 'No'), (1, 0), inplace=True)
     6 df.MalePattern_you.replace(('Yes', 'No'), (1, 0), inplace=True)
     7 df.job_physical_activity.replace(('A little', 'No'), (1, 0), inplace=True)
     8 df.exercisehours.replace(('regular', 'No'), (1, 0), inplace=True)
     9 df.smoking.replace(('Yes', 'No'), (1, 0), inplace=True)
    10 df.alcohol.replace(('Yes', 'No', 'Maybe'), (1, 0, 1), inplace=True)
    11 df.pms.replace(('Yes', 'No', 'Sometimes'), (1, 0, 1), inplace=True)
    12 df.MalePattern_you.replace(('Yes', 'No', 'A little'), (1, 0, 1), inplace=True)
    13 df.periodstatus.replace(('Yes', 'No', 'Sometimes', 'No (Irregular and Absent Periods)'), (1, 0, 1, 0), inplace=True)
    14 df['periodstatus'] = pd.to_numeric(df['periodstatus'])
    15
    16
```

```
[ ]  1 x = df[['age', 'stresslevels', 'sleepduration', 'sleeprating', 'smoking','alcohol', 'delayInPeriods', 'periodstatus', 'MalePattern_you', 'pms']]
     2 y = df.sufferingFrompcos
```

```
[ ]  1 x_train, x_test, y_train, y_test = train_test_split(x.values, y.values, test_size = 0.3)
     2
     3
```

```
[ ]
# Simple Logistic Regression

result = []
lr = LogisticRegression(random_state = 42)
lr.fit(x_train, y_train)
logreg = LogisticRegression()
logreg.fit(x_train, y_train)
acc_log_train = round(logreg.score(x_train, y_train) * 100, 2)
acc_log_test = round(logreg.score(x_test, y_test) * 100, 2)

result.append(acc_log_train)

print("Training Accuracy :  % {}".format(acc_log_train))
print("Testing Accuracy :  % {}".format(acc_log_test))
```

```
Training Accuracy :  % 78.18
Testing Accuracy :  % 79.17
```

```

from pandas.core.common import random_state
#XGBRF (Random Forest)
random_state = 42
xgb_clf = xgboost.XGBRFClassifier(max_depth=4, random_state=random_state)
xgb_clf.fit(x_train, y_train)

acc_xgb_clf_train = round(xgb_clf.score(x_train, y_train) *100, 2)
acc_xgb_clf_test = round(xgb_clf.score(x_test, y_test) *100, 2)

result.append(acc_xgb_clf_train)

print("Training Accuracy :  % {}".format(acc_xgb_clf_train))
print("Testing Accuracy :  % {}".format(acc_xgb_clf_test))
```

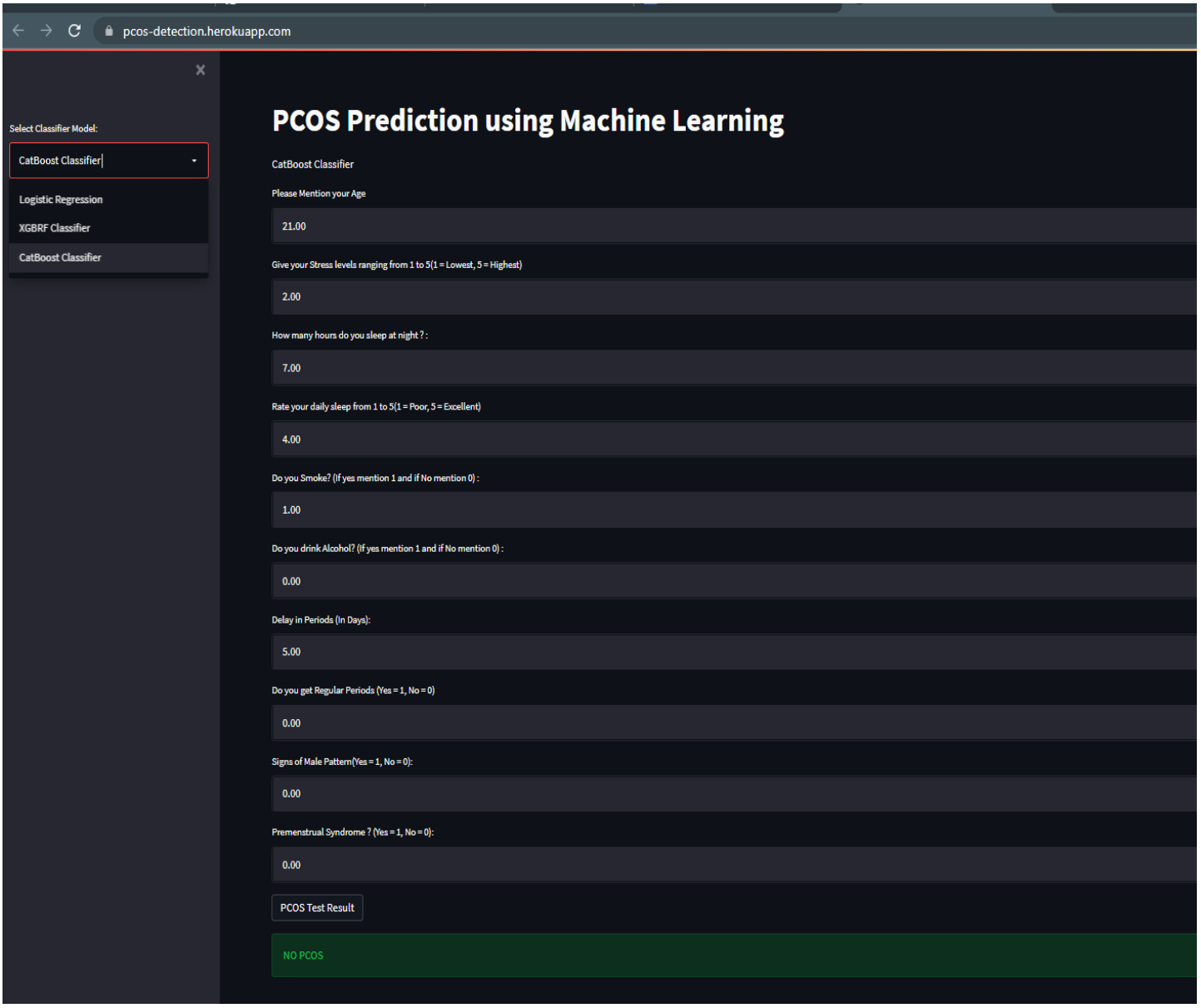
```
Training Accuracy :  % 85.45
Testing Accuracy :  % 79.17
```

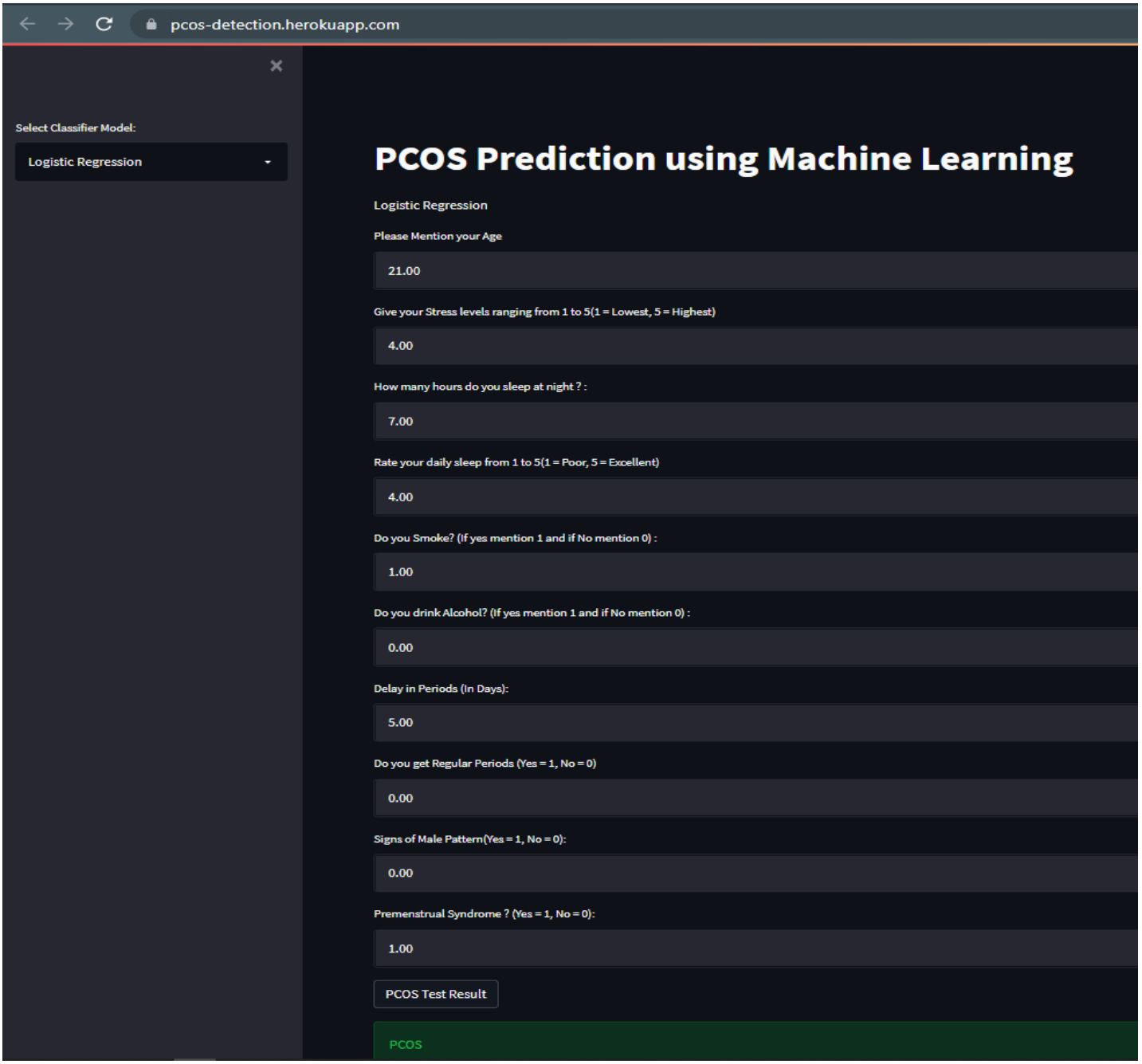
```
[ ]  1 # CatBoost Classifier
     2 cat_clf = CatBoostClassifier()
     3 cat_clf.fit(x_train, y_train)
     4 acc_cat_clf_train = round(cat_clf.score(x_train, y_train) * 100, 2)
     5 acc_cat_clf_test = round(cat_clf.score(x_test, y_test) * 100, 2)
     6
     7 result.append(acc_cat_clf_train)
     8
     9 print("Training Accuracy :  % {}".format(acc_cat_clf_train))
    10 print("Testing Accuracy :  % {}".format(acc_cat_clf_test))
```

```
473:    learn: 0.3034456        total: 324ms    remaining: 359ms
474:    learn: 0.3030460        total: 324ms    remaining: 358ms
475:    learn: 0.3026354        total: 325ms    remaining: 358ms
476:    learn: 0.3024393        total: 325ms    remaining: 357ms
477:    learn: 0.3019201        total: 326ms    remaining: 356ms
478:    learn: 0.3013980        total: 326ms    remaining: 355ms
479:    learn: 0.3013147        total: 327ms    remaining: 354ms
480:    learn: 0.3009064        total: 327ms    remaining: 353ms
481:    learn: 0.3004723        total: 328ms    remaining: 352ms
482:    learn: 0.3001097        total: 328ms    remaining: 351ms
483:    learn: 0.2997281        total: 329ms    remaining: 351ms
484:    learn: 0.2995136        total: 329ms    remaining: 350ms
485:    learn: 0.2990664        total: 330ms    remaining: 349ms
486:    learn: 0.2989050        total: 330ms    remaining: 348ms
487:    learn: 0.2985932        total: 331ms    remaining: 347ms
488:    learn: 0.2983936        total: 331ms    remaining: 346ms
489:    learn: 0.2980933        total: 332ms    remaining: 345ms
490:    learn: 0.2978373        total: 332ms    remaining: 344ms
491:    learn: 0.2976385        total: 333ms    remaining: 343ms
492:    learn: 0.2971027        total: 337ms    remaining: 347ms
493:    learn: 0.2967461        total: 337ms    remaining: 346ms
494:    learn: 0.2964621        total: 338ms    remaining: 345ms
495:    learn: 0.2960128        total: 340ms    remaining: 345ms
496:    learn: 0.2956725        total: 340ms    remaining: 344ms
497:    learn: 0.2954152        total: 341ms    remaining: 343ms
498:    learn: 0.2949699        total: 341ms    remaining: 343ms
499:    learn: 0.2946099        total: 342ms    remaining: 342ms
500:    learn: 0.2943096        total: 342ms    remaining: 341ms
501:    learn: 0.2941594        total: 343ms    remaining: 340ms
502:    learn: 0.2937991        total: 343ms    remaining: 339ms
```

```
987:    learn: 0.2037410        total: 644ms    remaining: 7.83ms
988:    learn: 0.2036147        total: 645ms    remaining: 7.17ms
989:    learn: 0.2035746        total: 645ms    remaining: 6.51ms
990:    learn: 0.2033681        total: 645ms    remaining: 5.86ms
991:    learn: 0.2033097        total: 650ms    remaining: 5.24ms
992:    learn: 0.2031958        total: 650ms    remaining: 4.58ms
993:    learn: 0.2030852        total: 651ms    remaining: 3.93ms
994:    learn: 0.2029671        total: 651ms    remaining: 3.27ms
995:    learn: 0.2028031        total: 651ms    remaining: 2.62ms
996:    learn: 0.2027479        total: 652ms    remaining: 1.96ms
997:    learn: 0.2027186        total: 656ms    remaining: 1.31ms
998:    learn: 0.2025912        total: 657ms    remaining: 657us
999:    learn: 0.2023589        total: 658ms    remaining: 0us
Training Accuracy :  % 94.55
Testing Accuracy :  % 87.5
```

Output:


pcos-detection.herokuapp.com
Select Classifier Model:
CatBoost Classifier
Logistic Regression
XGBRF Classifier
CatBoost Classifier
PCOS Prediction using Machine Learning
CatBoost Classifier
Please Mention your Age
21.00
Give your Stress levels ranging from 1 to 5(1 = Lowest, 5 = Highest)
2.00
How many hours do you sleep at night ? :
7.00
Rate your daily sleep from 1 to 5(1 = Poor, 5 = Excellent)
4.00
Do you Smoke? (If yes mention 1 and if No mention 0) :
1.00
Do you drink Alcohol? (If yes mention 1 and if No mention 0) :
0.00
Delay in Periods (In Days):
5.00
Do you get Regular Periods (Yes = 1, No = 0)
0.00
Signs of Male Pattern(Yes = 1, No = 0):
0.00
Premenstrual Syndrome ? (Yes = 1, No = 0):
0.00
PCOS Test Result
NO PCOS



Conclusion

PCOS is a serious syndrome that should be handled with utmost care. A proper discussion should be held regarding this and if detected consultation with a professional and appropriate medication is advised. This project was built to help common people to know if they are suffering from this syndrome. The models that have been used in this project have yielded decent to good accuracies for prediction.

Google Colab Link:

https://colab.research.google.com/drive/1lhLjhSHGI_WJpsytTYSPlNuO17FYwfvQ?usp=sharing

Working Model Live Link:

http://pcos-detection.herokuapp.com/