"Dissemination of Education for Knowledge, Science and Culture"
-Shikshanmaharshi Dr. Bapuji Salunkhe

(स्वायत्त) कोल्हापूर

VIVEKANAND COLLEGE KOLHAPUR (Empowered Autonomous)

DEPARTMENT OF STATISTICS

A PROJECT REPORT
on

"Analysis of the Prevalence and Health Impact of PCOS or PCOD Among Women"

Submitted by

Miss. Redekar Divyarani Pandurang

Miss. Kumbhar Shamal Sambhaji

Miss. Patil Priyanka Sayaji

Mrs. Chavan Shubham Suresh

in partial fulfillment for the award ofthe degree of

MASTER OF SCIENCE

in

STATISTICS

2023-24

CERTIFICATE

This is to Certify that,

Sr. No.	Name	Roll No.
1	Ms. Redekar Divyarani Pandurang	1421
2	Ms. Kumbhar Shamal Sambhaji	1411
3	Ms. Patil Priyanka Sayaji	1419
4	Mr. Chavan Shubham Suresh	1402

Have satisfactorily completed the project work on **"Analysis of the Prevalence and Health Impact on PCOS or PCOD Among Women"** as a part of practical evaluation course for **M.Sc. II**, prescribed by the Department of Statistics, ***Vivekanand College, Kolhapur (Empowered Autonomous)*** in the academic year **2023-24.**

This project has been completed under our guidance and supervision. To the best of our knowledge and belief, the matter presented in this project report is original and has not been submitted elsewhere for any other purpose.

Date:

Place: Kolhapur

27/05/24

Project Guide

(Mr. Bhosale. A. B.)

30/05/24

Examiner

Head

(Mrs. Shinde.V.C.)

HEAD
DEPARTMENT OF STATISTICS
VIVEKANAND COLLEGE, KOLHAPUR
(EMPOWERED AUTONOMOUS)

ACKNOWLEDGEMENT

As we approach the culmination of this enormous endeavor, it brings immense satisfication to reflect upon and express gratitude for the collective efforts and support from all those who enthusiastically contributed. Together, we have transformed what once seemed like a far-off aspiration for industrial training into a reality.

We are thankful to Principal Dr. R. R. Kumbhar for his kind cooperation and valuable support. We are also thankful to Mrs.V.C.Shinde, Head ,Department of Statistics. Vivekanand College, Kolhapur (Empowered Autonomous). We are indeed grateful to OJT Coordinator Mr.A.Pawar, Assistant Professor, Departrnent of Statistics, Vivekanand College, Kolhapur (EmpoweredAutonomous) for his kind and valuable support.

We are extremely thankful to project guide Mr. A.B.Bhosale sir. Assistant Professor, Department of Statistics, Vivekanand Colleges Kolhapur (Empowered Autonomus) for his valuable guidance and mentorship throughout this project work given to us during the study.We are also thankful toall the teaching and non-teaching staff of our department for their direct and indirect support. We also thanks to all the members of Next Level Design Academy.

Additionally, we sincerely thanks to our parents, friends and colleagues for their direct and indirect contributions.

Your's Sincerely
Msc.II
Department of Statistics

INDEX

INTRODUCTION

- One of the common diseases in today's world is polycystic ovarian disease (PCOD), Polycystic ovary syndrome (PCOS) which particularly affects the women ofage 12–45 years. In this disease, hormones are imbalanced.

- This disease affects both health and the quality of women's life. The symptoms include cardiovascular diseases, failure to ovulate and infertility, late menopause, type2 diabetes, acne, darkness, hair loss, hirsutism, obesity, anxiety, depression, and stress.

- The early diagnosis and treatment can be used to control based on the symptomsand by the prevention of long-term problems.

- PCOD/PCOS can be detected through ultrasonography by a doctor by counting thenumber and size of follicles in the ovaries. However, this process takes a long time, need good image quality and high accuracy to detect the presence of PCOD/PCOS.

- Another approach for PCOD/PCOS detection is through biochemical parameters such as hormone levels examination. Since hormone examination is very expensive, other clinical parameters such as body mass index (BMI), menstrual cycle length, etc.are taken into consideration for the detection of PCOD/PCOS.

- In recent years, machine learning (ML) classification and feature selection algorithms have been used by researchers and clinicians for the prediction of diseasesas a non-invasive method.

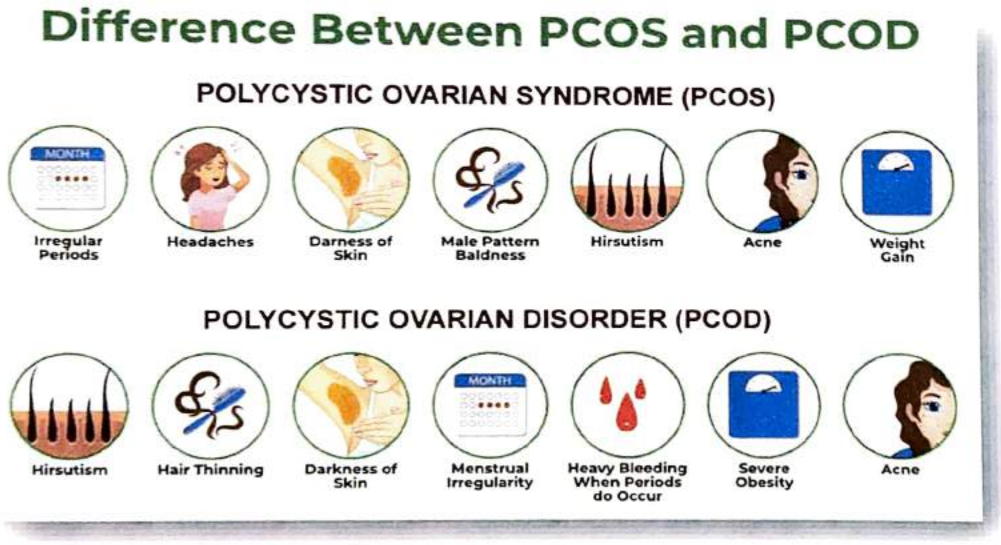
- The prevalence, diagnosis, etiology, management, clinical practices, psychologicalissues, and prevention are some of the most confusing aspects associated with PCOS.Statistical analysis played important role in such kind of study. So, we are conductinga statistical study on prevalence of obesity and depression in subjects with PCOD.

SIGNIFICANCE

- The present world women population is widely affected by preterm abortions, infertility, anovulation etc. It is observed that PCOD/PCOS, a condition seen among the women of reproductive age is having a major influence in the cause of infertility. Over five million women world wide in their reproductive age PCOD/PCOS.

- To address this problem, this study proposes model for the early detection and classificationof PCOD/PCOS from an optimal and minimal but promising clinical and metabolic parameter,which act as an early marker for this disease.

- Remarkably, our investigation helps in female reproductive health and may benefit in timely diagnosis of PCOD/PCOS which may further improve the management of reproductive health and fertility.



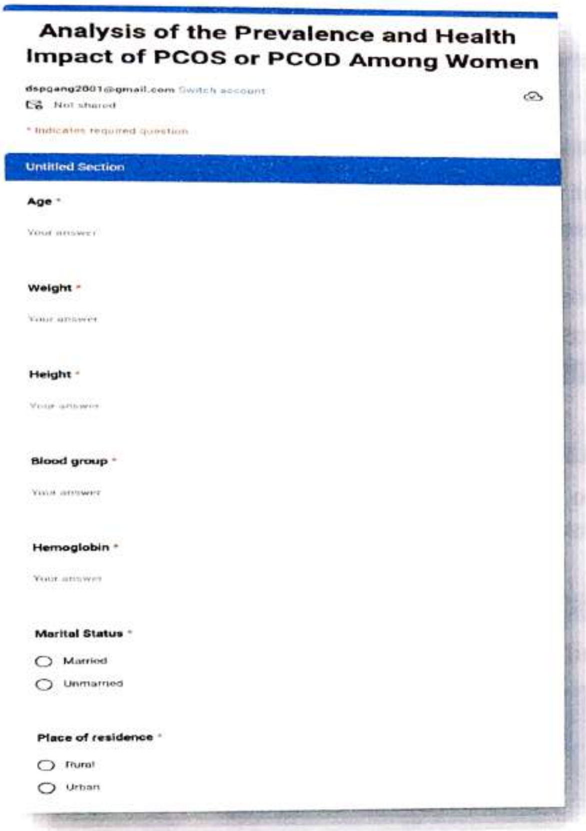
OBJECTIVES

- ✓ To study the prevalence of PCOD/PCOS in women on irregular period.
- ✓ To predict the women's with PCOD/PCOS based risk factor.
- ✓ To determine which reasons are most responsible for PCOD/PCOS.
- ✓ To check the classification of PCOD/PCOS in women's using various machine learning algorithms.

DATA COLLECTION

For this project, we have collected primary data.

- Target Population: Women in age group 12-45.
- Dataset consists total 412 number of sample observations.
- Among 91 women diagnosed with PCOD/PCOS.
- For conducting dataset, we prepared well-structured questionnaire which consists 30 numberof questions related to PCOD/PCOS in women.



Analysis of the Prevalence and Health Impact of PCOS or PCOD Among Women

dspgang2001@gmail.com Switch account

Not shared

* Indicates required question

Untitled Section

Age *

Your answer

Weight *

Your answer

Height *

Your answer

Blood group *

Your answer

Hemoglobin *

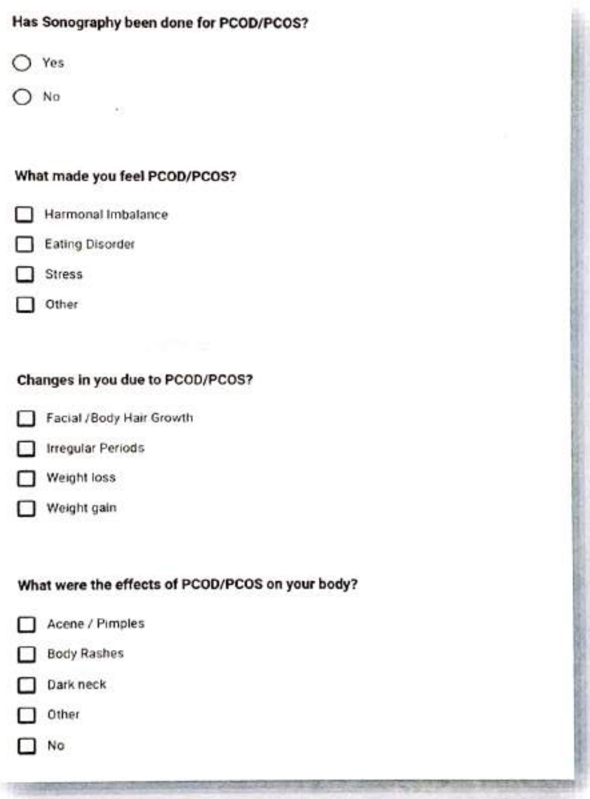
Your answer

Marital Status *

○ Married
○ Unmarried

Place of residence *

○ Rural
○ Urban



Has Sonography been done for PCOD/PCOS?

○ Yes
○ No

What made you feel PCOD/PCOS?

□ Harmonal Imbalance
□ Eating Disorder
□ Stress
□ Other

Changes in you due to PCOD/PCOS?

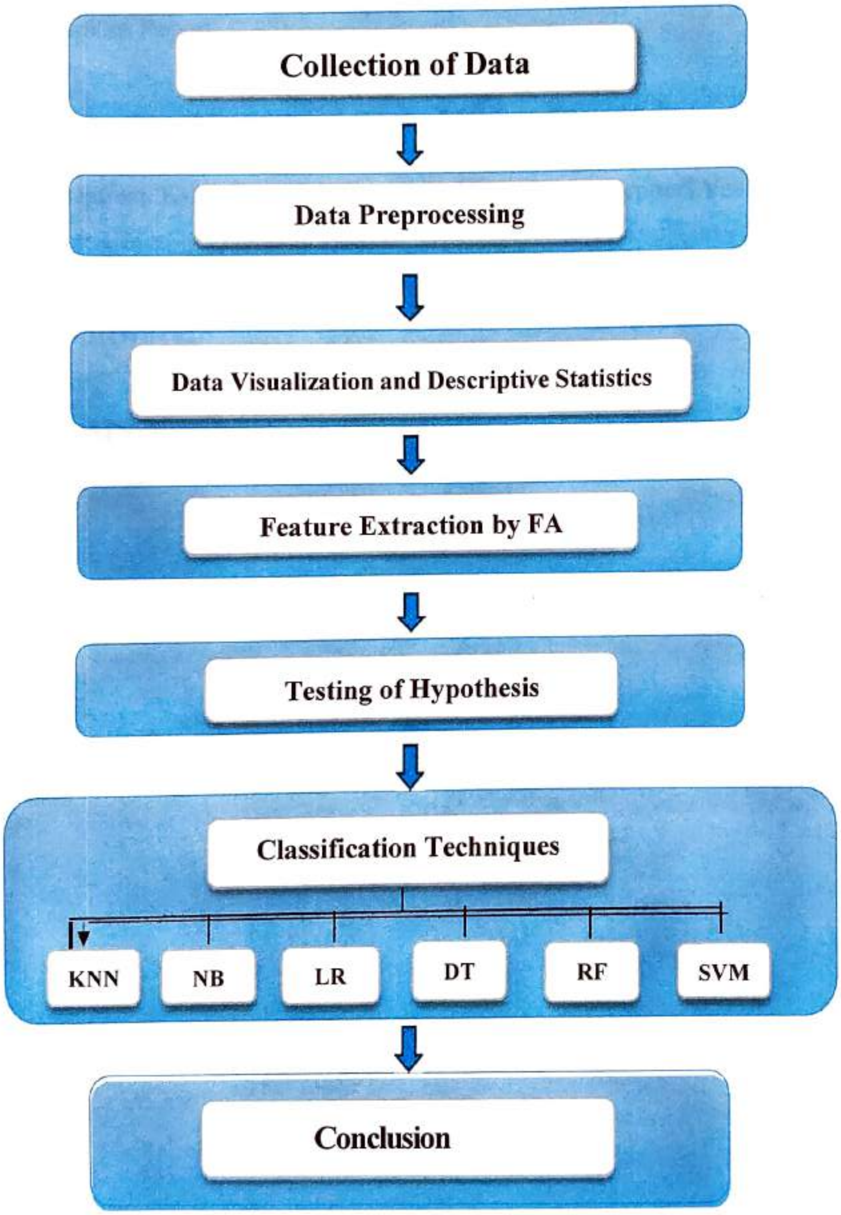
□ Facial /Body Hair Growth
□ Irregular Periods
□ Weight loss
□ Weight gain

What were the effects of PCOD/PCOS on your body?

□ Acene / Pimples
□ Body Rashes
□ Dark neck
□ Other
□ No

METHODOLOGY

In this study, machine learning classification technique has been proposed, trained and tested aimsto differentiate between PCOS and non-PCOS ovaries. The ovary disease variables have been usedas the input of the study from which the suggested method would determine whether or not the women having PCOD/PCOS. Along with this we performed comparative analysis to identify whichclassification technique is well suited for our data. The framework of the methodology usedin thisresearch is illustrated below:



STATISTICAL TOOLS

Exploratory Data Analysis:

- Bar charts, Pie Charts, Correlation Heatmap
- Chi-square test
- Feature Extraction by FA
- Classifications Report

Machine Learning Algorithms (Data Mining Classifiers) :

- Random Forest, K-Nearest Neighborhood, Naïve Bayes, Support Vector Machine, Logistic Regression,Decision Tree.

Statistical Software:

MS-Excel **Python**

GRAPHICAL REPRESENTATION

I. Reasons behind PCOD/PCOS

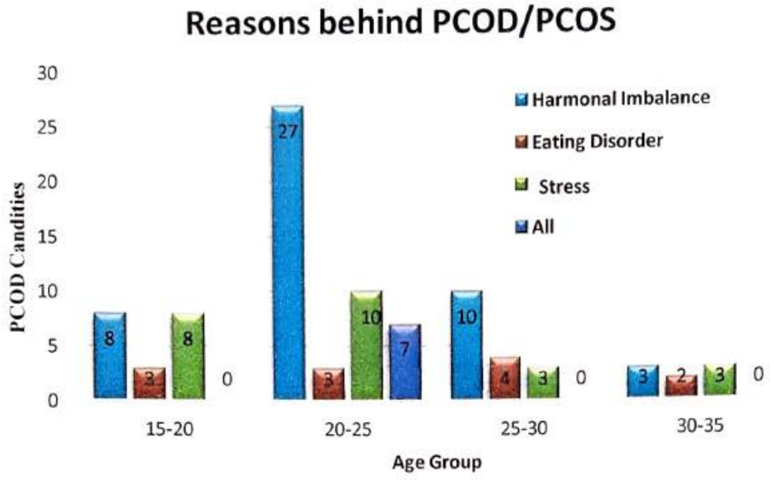


Fig.1.1

Conclusion: The Prevalence of Hormonal Imbalance and Stress is highest in the group 20-25,while eating disorders shows a more consistent distribution across different group.

II. Days of Flow in PCOD/PCOS

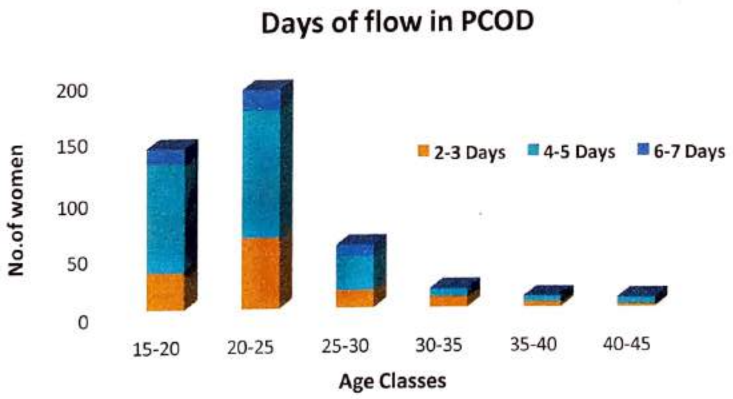


Fig 1.2

Conclusion: The younger age groups(15-25) tend to take more sick leave as compared to older age groups. The majority of sick leaves across all age groups are taken for 4-5days, followed by 2-3 days, and then 6-7 days.

III. Age wise representation of regular& irregular days of cycle length

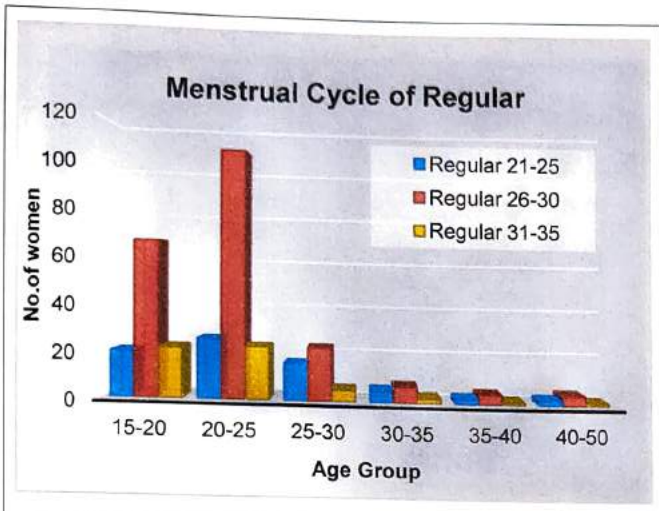


Fig 1.3

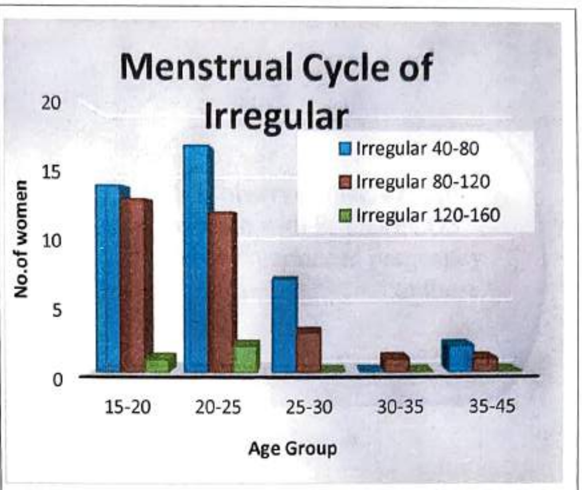


Fig 1.4

Conclusion: From Fig 1.3 and Fig 1.4, the age group 20-25 shows the highest number of individuals engaging in both regular and irregular periods.

IV. Impact of BMI on PCOD/PCOS Women

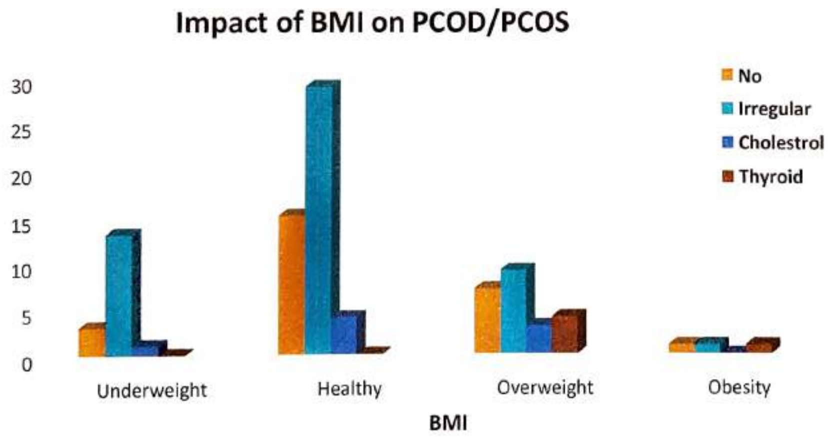
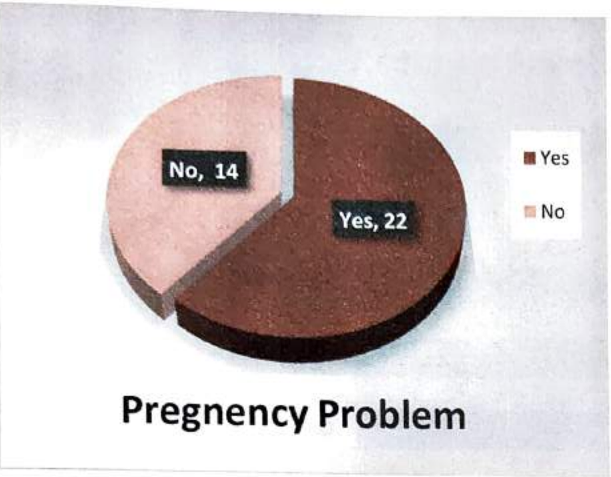


Fig 1.5

Conclusion: The majority of individual across all BMI categories do not have any of the specified health Conditions (No).

V. Pie chart of pregnancy problem in PCOS: -



It is observed that, 61% of women with PCOD/PCOS have experienced pregnancy problems compared to those who.

Fig 1.6 Pregnancy problem in PCOS

VI. Pie Chart of eating habit in PCOD/PCOS women: -

Women those who have PCOD/PCOS, only 10% of consumedbalanced diet.

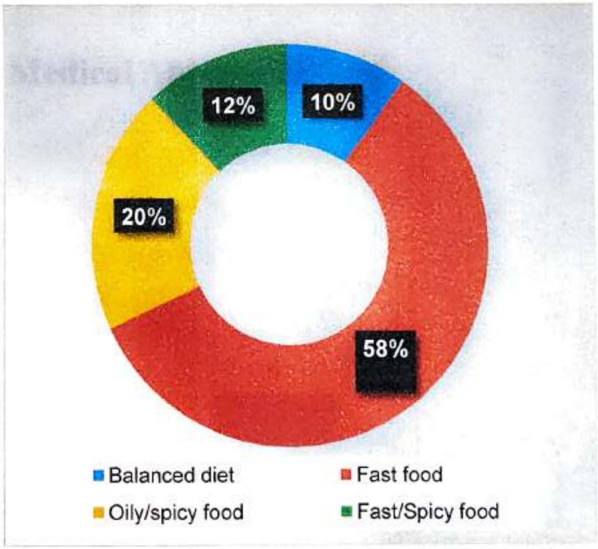


Fig 1.7 Eating habit in PCOD/PCOS women

VII. Impact of PCOD/PCOS during period.

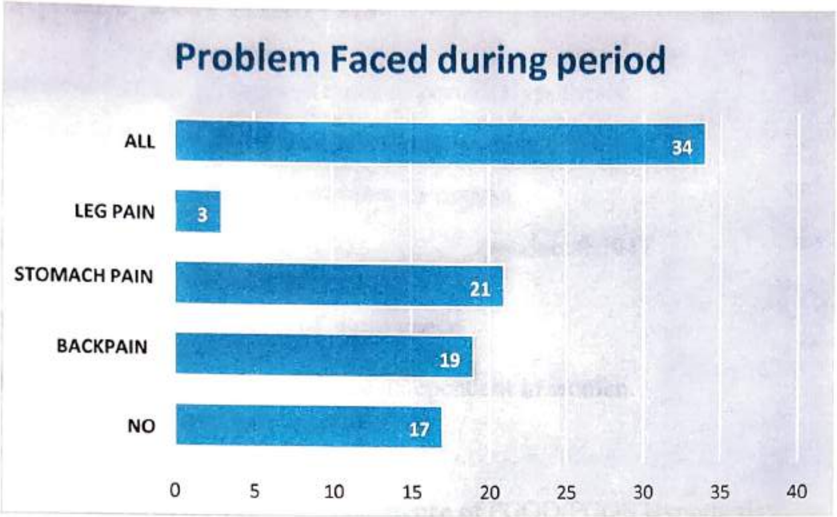


Fig 1.8

Conclusion: The most of the women faces all listed problems during periods. Leg Pain and No Pain are the least reported.

VIII. Comparison of Medical Approaches.

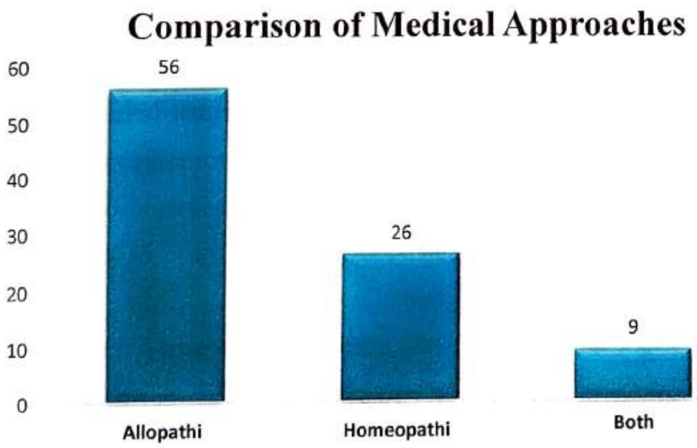


Fig 1.9 Healthcare Choices

Conclusion: The majority of women with PCOD/PCOS prefer allopathic over homepathic treatment. Additionally, 9 out of 91 Women with PCOD/PCOS prefers both types of treatment.

STATISTICAL ANALYSIS

Chi- square Test Analysis

- Testing independence of age groups and types of periodsHypothesis:

  H_0: Age and type of periods are independent in women.

  H_1: Age and type of periods are dependent in women.

  Chi-square statistic: 4.6104 α=0.05 p-value: 0.5947
  $p\text{-value} > 0.05$

  Therefore, Accept H0 at 5% level of significance

  Conclusion: Age and type of periods are independent in women.

- Testing independence of age groups and occurrence of PCOD/PCOS Hypothesis:

  H_0: Age is independent on occurrence of PCOD/PCOS in women.

  H_1: Age is dependent on occurrence of PCOD/PCOS in women.

  Chi-square statistic: 13.2777 α=0.05 p-value: 0.0388
  $p\text{-value} < 0.05$

  Therefore, Reject H0 at 5% level of significance

  Conclusion: Age is dependent on occurrence of PCOD/PCOS women.

- Testing independence of marital status and occurrence of PCOD/PCOSHypothesis:

  H_0: Marital status and occurrence of PCOD/PCOS are independent.

  H_1: Marital status and occurrence of PCOD/PCOS are dependent.

  Chi-square statistic: 9.0735 α=0.05 p-value: 0.0026

  $p\text{-value} < 0.05$

  Therefore, Reject H0 at 5% level of significance.

  Conclusion: Marital status and occurrence of PCOD/PCOS are dependent.

Testing proportion on occurrence of PCOD/PCOS

❖ Rate of PCOD/PCOS per hundred women: -

Total _observations = 412
Positive_ observations = 91

The proportion of positive observations relative to the total proportion = (positive _observations / total _observations) * 100

```
("Proportion relative to 100:", proportion)Proportion relative to 100: 22.0873
```

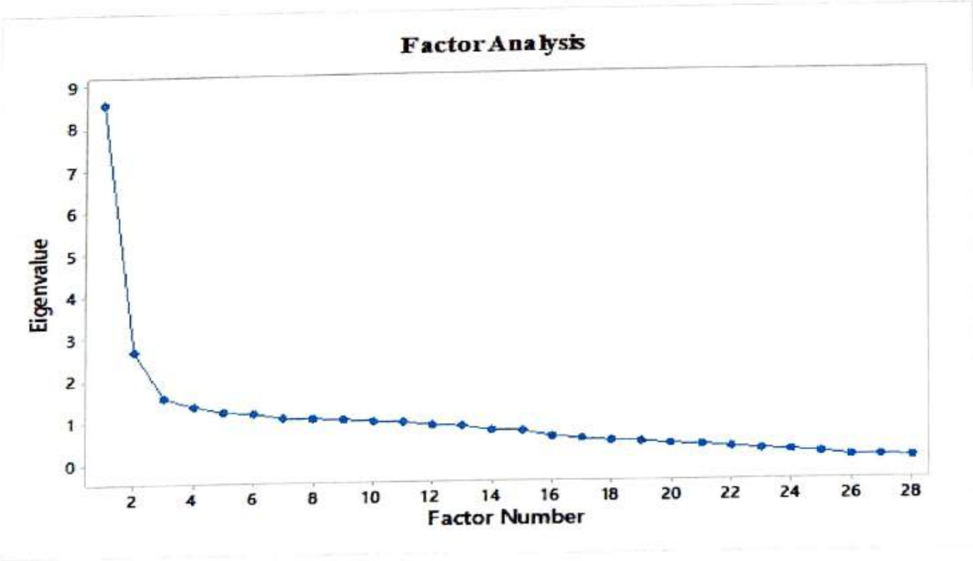
Conclusion: - Every 22 of 100 women face problem of PCOD/PCOS.

Feature Extraction by Factor Analysis

It is a feature extraction technique. The percentage indicates that how much each variable contribute to variation in the dataset.

Features	Variation in %
Age	30.5
BMI	9.5
Blood Group	5.6
Marital Status	4.8
Places of residence	4.3
Occupation	4.1
Vegetarian or non-vegetarian	3.8
Food type	3.7
Exercise Type	3.5
Sleep hours	3.4
Addiction Status	3.3
Year of starting periods	3
Nature of Period	2.9

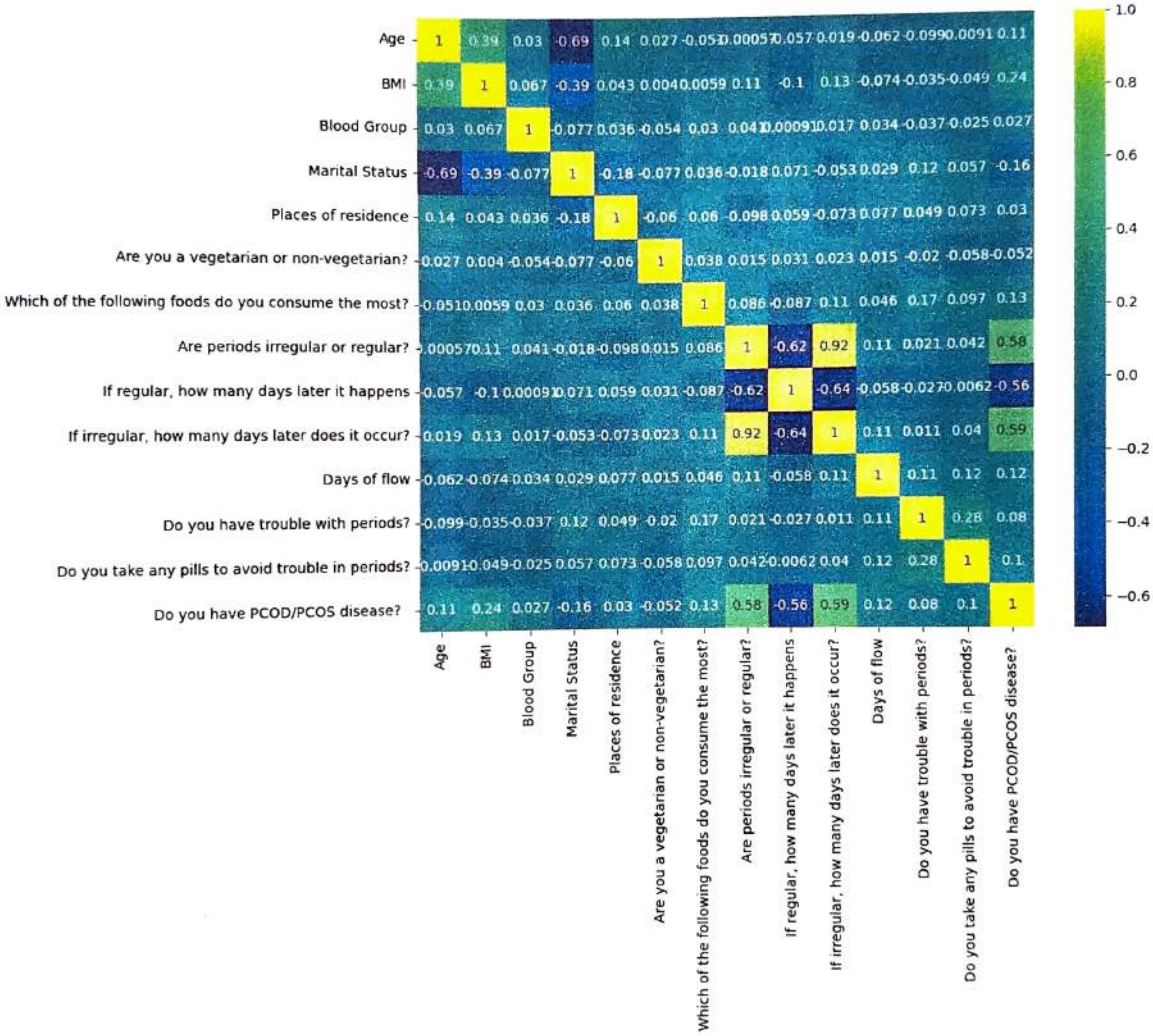
Table: Variation covered by features



Conclusion: We extracted 13 components from 28 component which covers maximumvariation (82%) within dataset

Correlation Heatmap

Correlation heatmap are a type of plot that visualize the strength of relationship between numerical variables. Correlation plots are used to understand which variable are related to each other and the strength of this relationship.



Conclusion: The above map shows the correlation between different study variable

K-Nearest Neighbours (KNN) Classifier

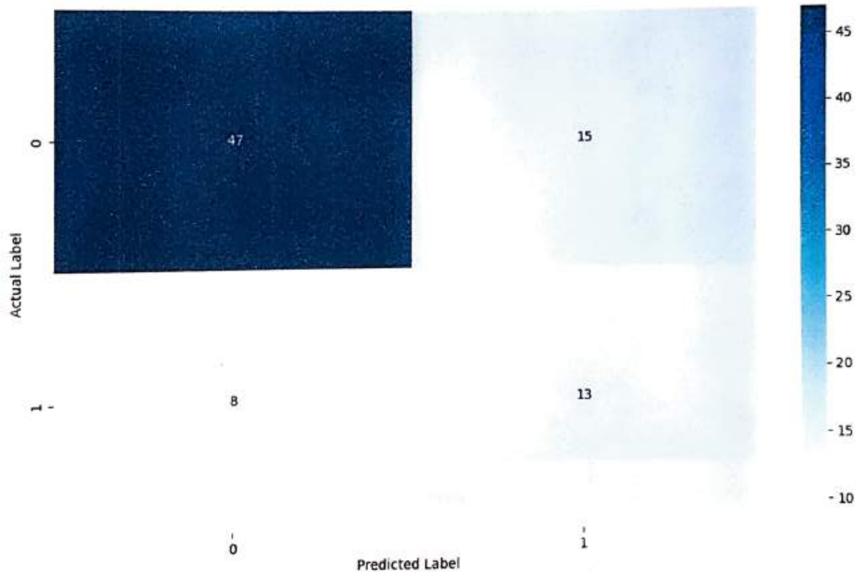
The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

Classification Report:

Precision	Recall	F1-Score	Accuracy	Roc Curve
0.76	0.72	0.73	0.72	0.84

Confusion Matrix: -



Conclusion: 72% predicted values are correctly classified with 28% misclassification rate by the KNN classifier.

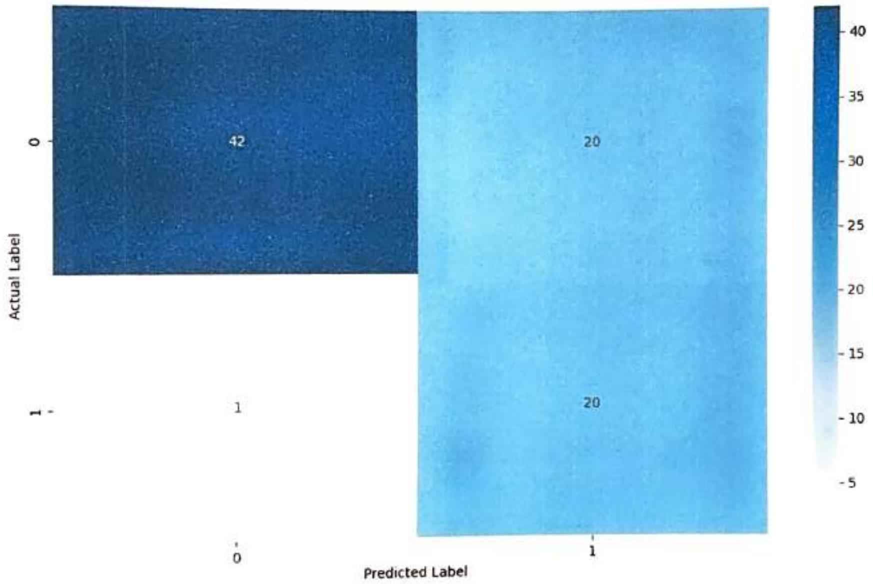
Naive Bayes Classifier

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification problem. There are three types of naive bayes classifier, viz. Multinomial naïve bayes, binomial naïve bayes and Gaussian naive bayes classifiers. In this project, we use Multinomial Naïve Bayes.

Classification Report:

Precision	Recall	F1-Score	Accuracy	Roc Curve
0.86	0.75	0.76	0.75	0.89

Confusion Matrix: -



Conclussion: 75% predicted values are correctly classified with 25% misclassification rate by the Naive Bayes classifier.

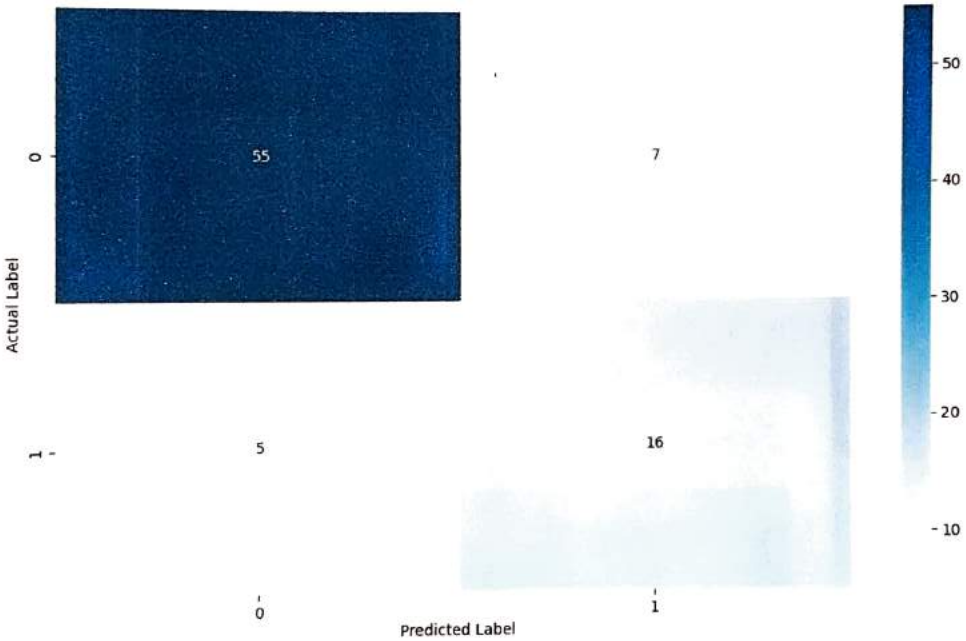
Logistic Regression Classifier

Logistic regression is the generalization of linear regression. It is used primarily for predicting binary or multi class dependent variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. While logistic regression is a powerful modeling tool,it assumes that the response variable is linear in the coefficients of the predictor variable.

Classification Report:

Precision	Recall	F1-Score	Accuracy	Roc Curve
0.86	0.86	0.86	0.86	0.91

Confusion Matrix:-



Conclusion: 86% predicted values are correctly classified with 14% misclassification rate by the Logistic regression classifier.

Decision Tree Classifier

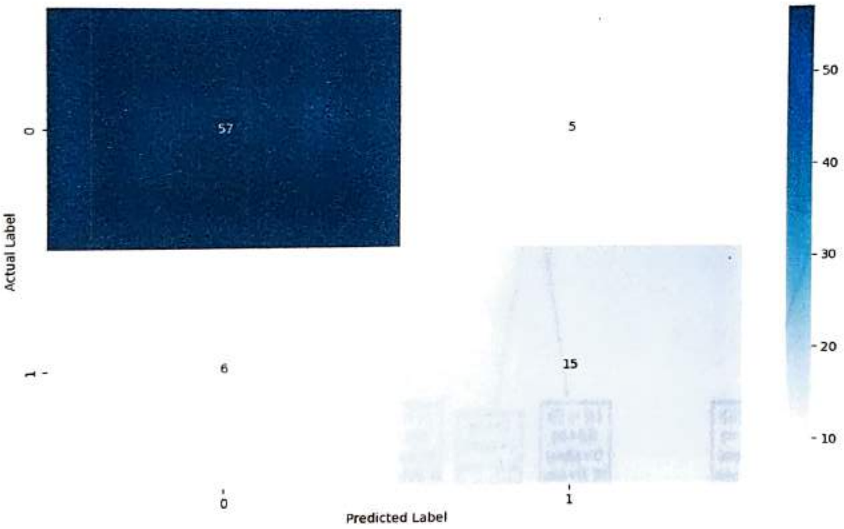
Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation tosolve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. In Decision Tree the major challenge is to identification of the attribute further root node in each level. This process is known as attributeselection. We have two popular attribute selection measures:

1. Information Gain
2. Gini Index

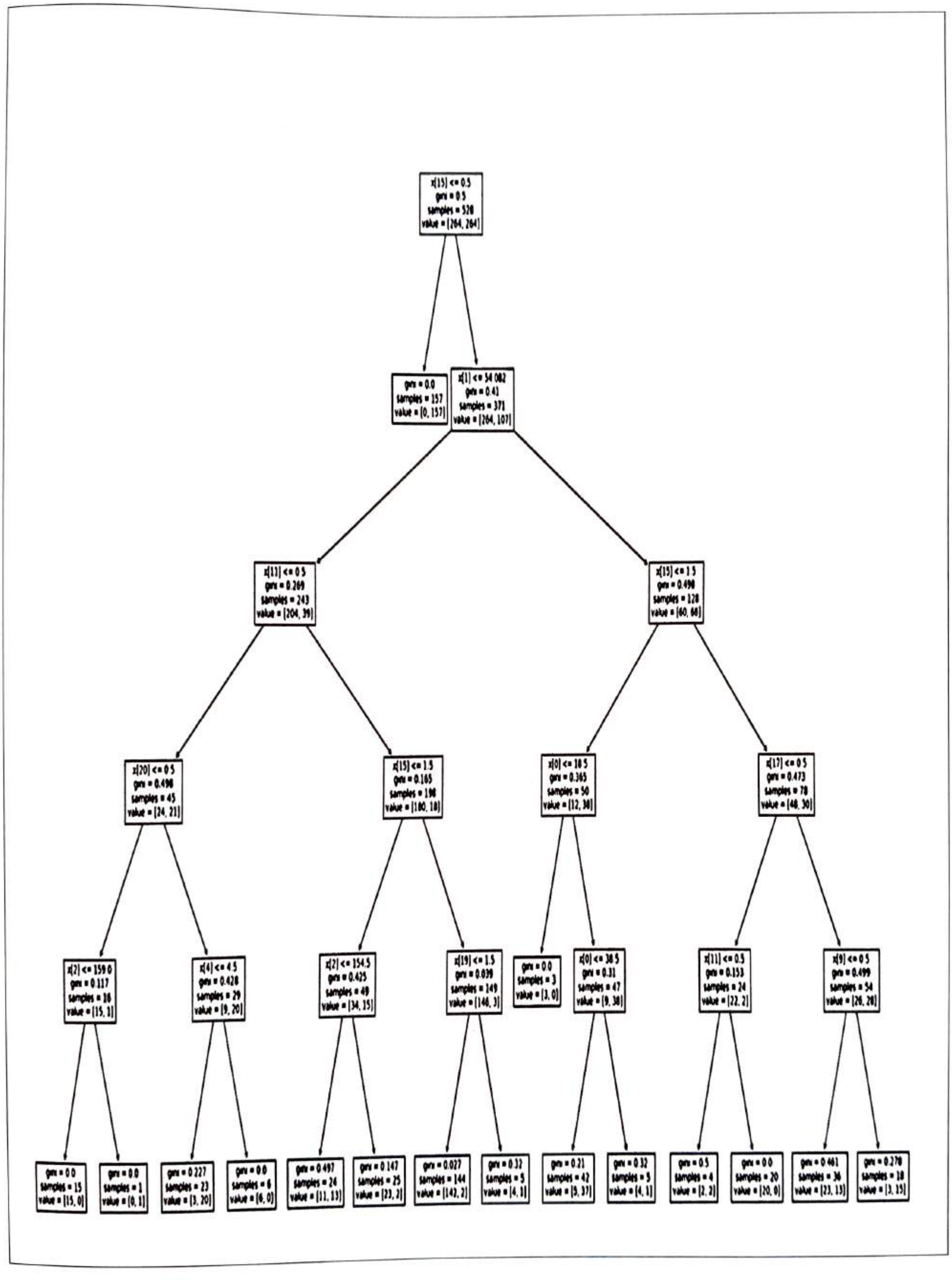
Classification Report

Precision	Recall	F1-Score	Accuracy	Roc curve
0.87	0.87	0.87	0.72	0.84

Confusion Matrix: -



Conclusion: 72% predicted values are correctly classified with 28% misclassification rate by the Logistic regression classifier.


x[15] <= 0.5
gini = 0.5
samples = 528
value = [264, 264]
gini = 0.0
samples = 157
value = [0, 157]
x[1] <= 54.082
gini = 0.41
samples = 371
value = [264, 107]
x[11] <= 0.5
gini = 0.269
samples = 243
value = [204, 39]
x[15] <= 1.5
gini = 0.498
samples = 128
value = [60, 68]
x[20] <= 0.5
gini = 0.498
samples = 45
value = [24, 21]
x[15] <= 1.5
gini = 0.165
samples = 198
value = [180, 18]
x[0] <= 18.5
gini = 0.365
samples = 50
value = [12, 38]
x[17] <= 0.5
gini = 0.473
samples = 78
value = [48, 30]
x[2] <= 159.0
gini = 0.117
samples = 16
value = [15, 1]
x[4] <= 4.5
gini = 0.428
samples = 29
value = [9, 20]
x[2] <= 154.5
gini = 0.425
samples = 49
value = [34, 15]
x[19] <= 1.5
gini = 0.039
samples = 149
value = [146, 3]
gini = 0.0
samples = 3
value = [3, 0]
x[0] <= 38.5
gini = 0.31
samples = 47
value = [9, 38]
x[11] <= 0.5
gini = 0.153
samples = 24
value = [22, 2]
x[9] <= 0.5
gini = 0.499
samples = 54
value = [26, 28]
gini = 0.0
samples = 15
value = [15, 0]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.227
samples = 23
value = [3, 20]
gini = 0.0
samples = 6
value = [6, 0]
gini = 0.497
samples = 24
value = [11, 13]
gini = 0.147
samples = 25
value = [23, 2]
gini = 0.027
samples = 144
value = [142, 2]
gini = 0.32
samples = 5
value = [4, 1]
gini = 0.21
samples = 42
value = [5, 37]
gini = 0.32
samples = 5
value = [4, 1]
gini = 0.5
samples = 4
value = [2, 2]
gini = 0.0
samples = 20
value = [20, 0]
gini = 0.461
samples = 36
value = [23, 13]
gini = 0.278
samples = 18
value = [3, 15]

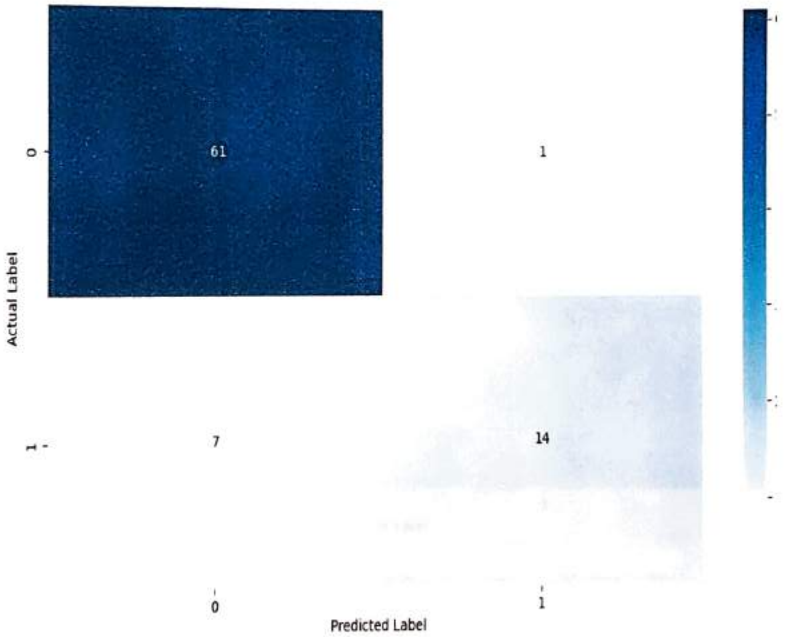
Random Forest Classifier

Random Forest is ensemble learning method where it combine multiple decision tree during training and predict model. Random forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictiveaccuracy of that dataset.

Classification Report

Precision	Recall	F1-Score	Accuracy	Roc curve
0.91	0.90	0.90	0.90	0.92

Confusion Matrix: -



Conclusion: 90% predicted values are correctly classified with 10% misclassification Rate by the Random Forest classifier.

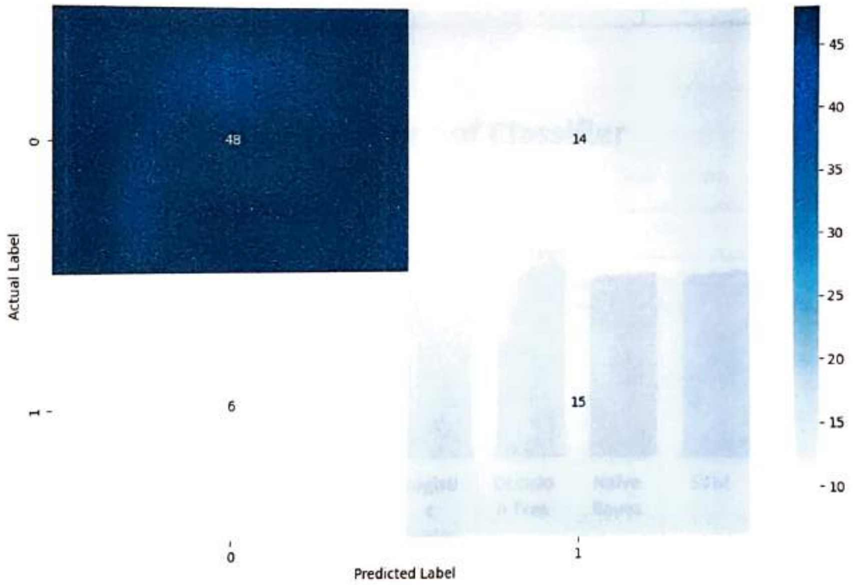
Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In two- dimensional space this hyperplane is a line dividing a plane intwo parts where in each class lay in either side.

Classification Report

Precision	Recall	F1-Score	Accuracy	Roc curve
0.79	0.76	0.77	0.75	0.91

Confusion Matrix: -



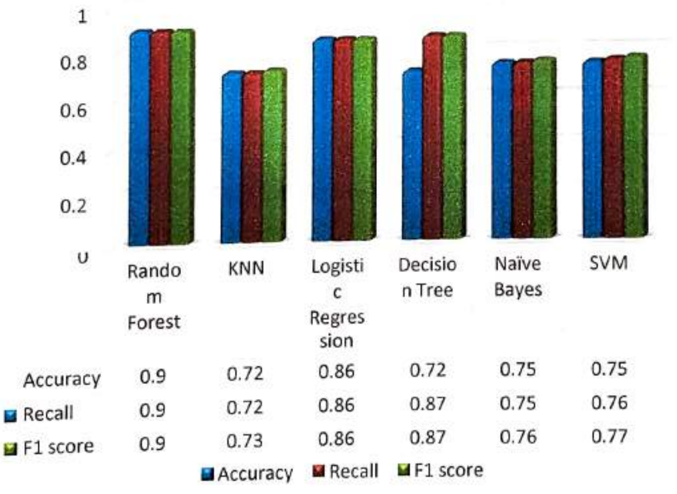
Conclusion: 75% predicted values are correctly classified with 25% misclassification rateby the Support Vector Machine classifier.

Classification results

To classify women's with PCOD/PCOS from given variables, we applied following techniques.

Performance			
Classifiers	Accuracy	Recall	F1 score
Random Forest	0.90	0.90	0.90
KNN	0.72	0.72	0.73
Logistic Regression	0.86	0.86	0.86
Decision Tree	0.72	0.87	0.87
Naïve Bayes	0.75	0.75	0.76
SVM	0.75	0.76	0.77

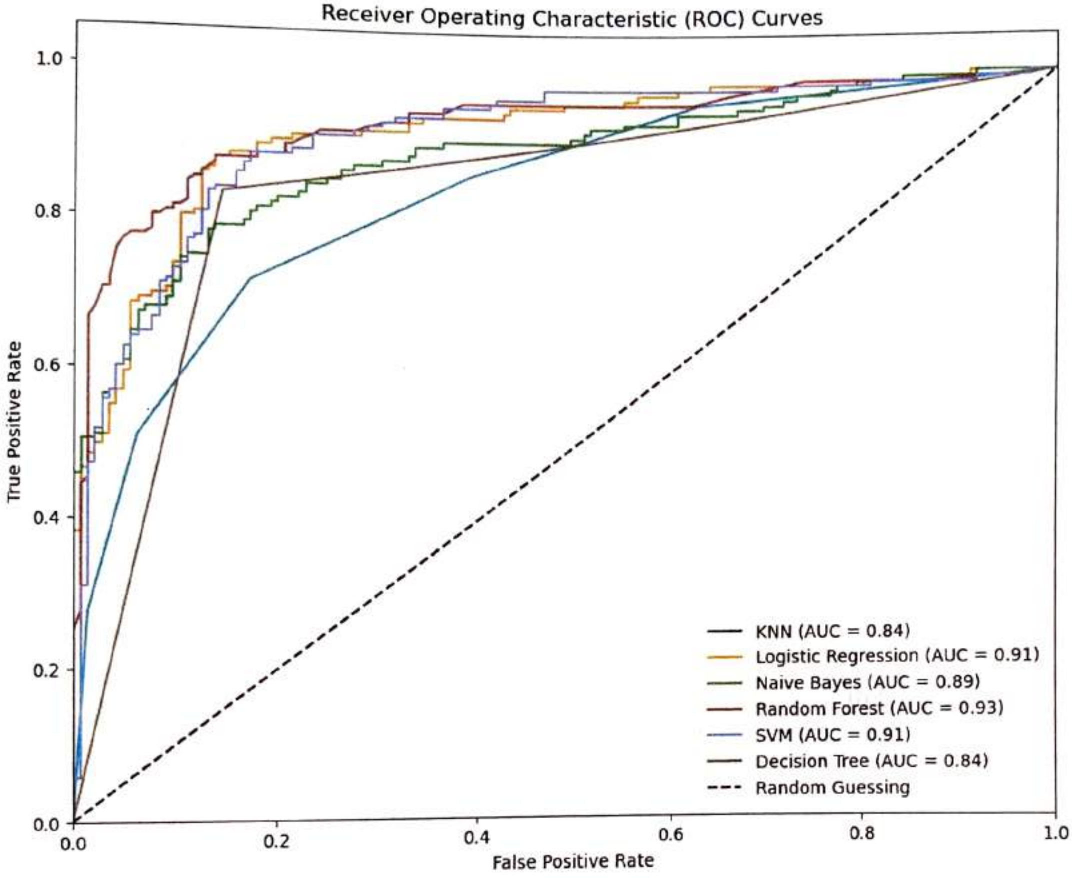
Performence of Classifier



Conclusion : From above, we can conclude that random forest model gives highest (86%) accuracy of classification, were as logistic regression gives 81% accuracy.

Receiver Operating Characteristic (ROC)

ROC curve is commonly used to visualize the performance of the classifier.



Conclusion: From above ROC curve, the AUC is 0.93 which indicates that random forest is best model which correctly classified observations into categories.

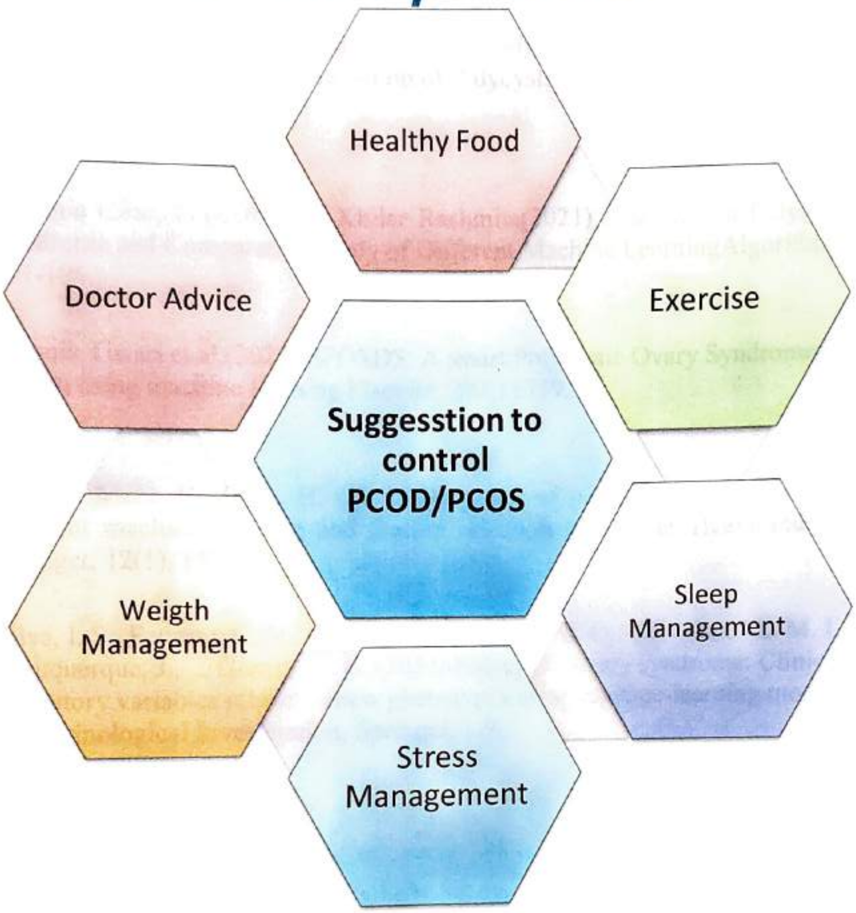
MAJOR FINDING

- Maximum number of women experienced problems in pregnancy due to PCOD/PCOS.
- PCOD appears to be more prevalent among younger age groups, with the highest numberof cases observed in the age group 20-25.
- Approximately 90% of women having PCOD/PCOS consumed fast food, oily/spicy food.
- Age affects the occurrence of PCOD or PCOS, while it does not affect the types ofperiods.
- Marital status affects occurrence of PCOD/PCOS.
- Every 22 of 100 women face problem of PCOD/PCOS.
- Random forest model gives highest (86%) accuracy of classification, whereas logistic regression gives 81% accuracy, the AUC is 0.91 which indicates that discrimination isfair.
- From principal component analysis, 12 number of components from 27 component are extracted which covers maximum variation within dataset.

Limitations of the Study

- Tools applied to the data can change their performance if we change the data
- We develop model only with available variables but if add other important variables then we expect that our models give better results

Suggestion to control PCOD/PCOS



Suggesstion to control PCOD/PCOS

Healthy Food

Exercise

Sleep Management

Stress Management

Weigth Management

Doctor Advice

REFERENCES

- Thomas, N., Kavitha, A.(2020) Prediction of Polycystic Ovarian Syndromewith Clinical Dataset using a Novel Hybrid Data Mining Classification Technique. IJARET 11(11),1872-1881.

- Vikas, B., Anuhya, B. S., Bhargav, K. S., Sarangi, S., & Chilla, M. (2018)Application of the apriori algorithm for prediction of Polycystic Ovarian Syndrome (PCOS) Springer 12(3), 934-944.

- Sinthia Gino, Poovizhi and Khilar Rashmita(2021) Analysis on PolycysticOvarian Syndrome and Comparative Study of Different Machine LearningAlgorithms Springer, 191-196.

- Shamik Tiwari et.al.(2021) SPOSDS: A smart Polycystic Ovary Syndromediagnostic system using machine learning Elsevier, 203,117592

- Danaei Mehr, H., Polat, H. (2022) Diagnosis of polycystic ovary syndromethrough different machine learning and feature selection techniques Healthand Technology, Springer, 12(1), 137-150.

- Silva, I. S., Ferreira, C. N., Costa, L. B. X., Soter, M. O., Carvalho, ´ L. M. L., de C. Albuquerque, J., ... Gomes, K. B. (2022) Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes usingmachine-learning models. Journal of Endocrinological Investigation, Springer, 1-9.

QUESTIONNAIRE

Project Title: Analysis of the Prevalence and Health Impact of PCOS or PCOD Among Women.

1. Age ____________

2. Weight ____________

3. Height ____________

5. Blood Group ____________

6. Haemoglobin ____________

7. Marital Status

□ Married □ Unmarried

8. Places of residence

□ Rural

□ Urban

9. What do you do?

□ Students

□ Housewife

□ Social Worker / Office Worker

□ Wages

10. Are you a vegetarian or non-vegetarian?

□ Vegetarian

□ non-vegetarian

□ Both

11. Which of the following foods do you consume the most?

□ Fast food

□ Balanced diet

□ Oily / Spicy Foods

12. What kind of exercise do you do?

□ Walking

□ Yoga

□ No

13. How many hours do you sleep?

□ 4-6 hrs

□ 6-8 hrs

□ 8-10 hrs

14. Are you addicted to any kind of addiction?

□ Smoking

□ Wine

□ Masher

□ No

15. What year did you start periods? ____________

16. Are periods irregular or regular?

□ Irregular

□ Regular

17. If regular, how many days later it happens?

□ 21-25 Days

□ 26-30 Days

□ 31-35 days

18. If irregular, how many days later does it occur?

□ 40-80 Days

□ 80-120 Days

□ 120-160 Days

□ 160-200 Days

19. Days of flow

□ 2-3 Days

□ 4-5 Days

□ 6-7 Days

20. Do you have trouble with periods?

□ No

□ Back pain

□ Stomach pain

□ Leg pain

21. Do you take any pills to avoid trouble in periods?

□ No

□ Ibuprofen

□ Meftal –spas

□ Cyclopsam

22. Did you know PCOD/PCOS?

□ Yes

□ No

23. Do you have PCOD/PCOS disease?

□ Yes

□ No

24. Has Sonography been done for PCOD/PCOS?

□ Yes

□ No

25. What made you feel PCOD/PCOS?

□ Hormonal Imbalance

□ Eating Disorder

□ Stress

□ Other

26. Changes in you due to PCOD/PCOS?

□ Facial /Body Hair Growth

□ Irregular Periods

□ Weight loss

□ Weight gain

27. What were the effects of PCOD/PCOS on your body?

□ Acene / Pimples

□ Body Rashes

□ Dark neck

□ Other

□ No

28. What diseases were experienced in PCOD/PCOS?

□ Thyroid

□ Cholesterol

□ Diabetes

□ Irregular Period

□ No

29. What treatment did you take for PCOD/PCOS?

□ Homeopathic

□ Allopathic

30. What would you do to reduce PCOD/PCOS?

□ Proper diet

□ Yoga/exercise

□ Doctor's advice

31. Did PCOD/PCOS cause difficulty in getting pregnant?

□ Yes

□ No

APPENDIX

- **File import**

```
  import numpy as np
  import pandas as pd
  import matplotlib.pyplot as plt
Data=pd.read_csv("D:/Shubham/Divya2.csv")
Data
```

- **Data Preprocessing**

```
Data.isnull()
Data.isnull().sum()
Data.info()
Data.describe()
```

- **Split Data into train and test**

```
  from sklearn.model_selection import train_test_split
  from sklearn.metrics import confusion_matrix,accuracy_score,
  mean_squared_error,classification_report
  from sklearn import metrics
  Training_Data=Data.drop(["Do you have PCOD/PCOS disease?    "],axis=1)
Testing_Data=Data["Do you have PCOD/PCOS disease?  "]
  X_train,X_test, y_train,y_test=train_test_split(Training_Data,Testing_Data,test_size=0.2)
```

- **Import SMOTE technic for over_sampling**

```
from imblearn.over_sampling import SMOTEprint(X_train.shape,y_train.shape)
print(X_test.shape,y_test.shape)
print("before oversampling,count of lables'0':{}".format(sum(y_train==0)))print("before oversampling,count
of lables'1':{}".format(sum(y_train==1)))sm=SMOTE(random_state=30)
X_train_res,y_train_res=sm.fit_resample(X_train,y_train.ravel())
print("After oversampling,count of lables'0':{}".format(sum(y_train_res==0)))print("After
oversampling,count of lables'1':{}".format(sum(y_train_res==1))
```

Classification Techniques

❖ KNN

```
from sklearn.neighbors import KNeighborsClassifierk=KNeighborsClassifier(n_neighbors=5)
k.fit(X_train_res,y_train_res)
pred=k.predict(X_test) pred
accuracy=metrics.accuracy_score(y_test,pred) print("Äccuracy from KNN=",accuracy)

print(classification_report(y_test,pred)) print(confusion_matrix(y_test,pred))
```

❖ Naive bayes

```
from sklearn.naive_bayes import GaussianNBn=GaussianNB()

n.fit(X_train_res,y_train_res)class_pred=n.predict(X_test)class_pred

accuracy1=metrics.accuracy_score(y_test,class_pred)print("Accuracy from Naive bayes=",accuracy1)

print(classification_report(y_test,class_pred))print(confusion_matrix(y_test,class_pred))
```

❖ logistics Regression

```
from sklearn.linear_model import LogisticRegressionlogistic=LogisticRegression()
logistic.fit(X_train_res,y_train_res) pred3=logistic.predict(X_test)
pred3 accuracy2=metrics.accuracy_score(y_test,pred3) print("Accuracy of logistics
Regression=",accuracy2)

print(classification_report(y_test,pred3)) print(confusion_matrix(y_test,pred3))
```

❖ Random Forest

```
from sklearn.ensemble import RandomForestClassifierr=RandomForestClassifier(n_estimators=10)
r.fit(X_train_res,y_train_res)

pred5=r.predict(X_test)pred5
accuracy5=metrics.accuracy_score(y_test,pred5)print("Accuracy of Random Forest=",accuracy5)

print(classification_report(y_test,pred5)) print(confusion_matrix(y_test,pred5))
```

```
pred5=r.predict(X_test)pred5
accuracy5=metrics.accuracy_score(y_test,pred5)print("Accuracy of Random Forest=",accuracy5)

print(classification_report(y_test,pred5)) print(confusion_matrix(y_test,pred5))
```

❖ Decision Tree

```
from sklearn.tree import DecisionTreeClassifierD=DecisionTreeClassifier(max_depth=5)
D=D.fit(X_train_res,y_train_res) pred4=D.predict(X_test)
pred4 accuracy4=metrics.accuracy_score(y_test,pred)print("Accuracy of Decision Tree=",accuracy4)

print(classification_report(y_test,pred4)) print(confusion_matrix(y_test,pred4))

from sklearn import treeT=tree.export_text(D) print(T)

fir=plt.figure(figsize=(25,18)) T1=tree.plot_tree(D)
```

❖ Support Vector Machine

```
from sklearn.svm import SVC svm = SVC() svm.fit(X_train_res, y_train_res)pred6 =
svm.predict(X_test) pred6
print(confusion_matrix(y_test,pred6))

a6 = metrics.accuracy_score(y_test,pred6) print("Accuracy from Support Vector Machine = ", a6)

print(classification_report(y_test, pred6))
```

❖ Receiver Operating Characteristic (ROC) Curves

```
import numpy as np
import matplotlib.pyplot as plt

from sklearn.datasets import make_classification from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler from sklearn.neighbors import
KNeighborsClassifier from sklearn.linear_model import LogisticRegressionfrom sklearn.naive_bayes
import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_curve, auc
```

- **Generate some example data**

```
X, y = make_classification(n_samples=1000, n_features=20, n_classes=2, random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

- **Standardize the features**

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

- **Initialize all classifiers**

```
classifiers = {
    "KNN": KNeighborsClassifier(),
    "Logistic Regression": LogisticRegression(),
    "Naive Bayes": GaussianNB(),
    "Random Forest": RandomForestClassifier(),
    "SVM": SVC(probability=True),
    "Decision Tree": DecisionTreeClassifier()
}
```

- **Train all classifiers and compute ROC curves**

```
plt.figure(figsize=(10, 8))
for name, clf in classifiers.items():
    clf.fit(X_train, y_train)
    y_score = clf.predict_proba(X_test)[:, 1]fpr,
    tpr, _ = roc_curve(y_test, y_score) roc_auc
    = auc(fpr, tpr)
    plt.plot(fpr, tpr, label=f'{name} (AUC = {roc_auc:.2f})')
```

- **Plot ROC curve for each classifier**

```
plt.plot([0, 1], [0, 1], linestyle='--', color='k', label='Random Guessing')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curves')
plt.legend(loc='lower right')
plt.show()
```

- **Null Hypothesis (H0): There is no significant association between marital status and theresponse variable (whether individuals answered "Yes" or "No").**

- **Alternative Hypothesis (H1): There is a significant association between marital status andfrom** scipy.stats **import** chi2_contingency
- **Define the observed frequencies**

```
observed = [[35, 56], [71, 250]]
```

- **Perform chi-square test**

```
chi2, p, dof, expected = chi2_contingency(observed)
```

- **Print the results**

```
print("Chi-square statistic:", chi2)
print("p-value:", p) print("Degrees
of freedom:", dof)
print("Expected frequencies:", expected)
```

- **H0:There is no association between age group and the response variable (regular or irregular).**
- **H1:There is association between age group and the response variable (regular or irregular).**

from scipy.stats **import** chi2_contingency

- **Define the observed frequencies**

```
observed = [
   [116, 22],
   [150, 38],
   [47, 7],
   [11, 4],
   [7, 2],
   [6, 0],
   [2, 0]
]
```

- **Perform chi-square test**

```
chi2, p, dof, expected = chi2_contingency(observed)
```

- **Print the results**

```
print("Chi-square statistic:", chi2)
print("p-value:", p) print("Degrees
of freedom:", dof)print("Expected
frequencies:") for row in expected:
   print(row)
```

- **H0:There is no association between age group and the response variable (Yes or No)**
- **H1:There is association between age group and the response variable (Yes or No)**

```
from scipy.stats import chi2_contingency
```

- **Define the observed frequencies**

```
observed = [
   [19, 119],
   [47, 141],
   [17, 37],
   [5, 10],
   [3, 6],
   [0, 6],
   [0, 2]
]
```

- **Perform chi-square test**

```
chi2, p, dof, expected = chi2_contingency(observed)
```

- **Print the results**

```
print("Chi-square statistic:", chi2)
print("p-value:", p) print("Degrees
of freedom:", dof)print("Expected
frequencies:") for row in expected:
   print(row)
```

- **Factor Analysis BY using minitab**

Minitab ➡ Stat ➡ Multivariate ➡ Factor Analysis