**Early identification of PCOS with commonly known diseases and obesity**

**risk prediction using machine learning techniques**

*Report submitted to the SASTRA Deemed to University as the requirement for the course*

**INT300 - MINI PROJECT**

*Submitted by*

**Karthik R**

**(Reg. No.: 124015045, B.Tech. Information Technology)**

**Infant Akash S**

**(Reg. No.: 124015034, B.Tech. Information Technology)**

**Haresh S**

**(Reg. No.: 124003103, B.Tech. Information Technology)**

**MAY 2023**



**SCHOOL OF COMPUTING**

**THANJAVUR, TAMIL NADU, INDIA – 613 401**



**SCHOOL OF COMPUTING**

**THANJAVUR – 613 401**

**Bonafide Certificate**

This is to certify that the report titled “ **Early identification of PCOS with commonly**

**known diseases and obesity risk prediction using machine learning techniques**”

submitted as a requirement for the course, INT300 : MINIPROJECT for B.Tech. is a

bonafide record of the work done by **Mr. Karthik R (Reg.No.:124015045, B.Tech.**

**Information Technology), Mr. Infant Akash S (Reg.No.:124015034, B.Tech. Information**

**Technology), Mr. Haresh S (Reg. No.:124003103, B.Tech. Information Technology)**

during the academic year 2022-23 in the School of Computing, under my supervision.

Signature of Project Supervisor :

Name with Affiliation :

Date :

Mini Project Viva voce held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Examiner 1 : Examiner 2 :

# **Acknowledgements**

We would like to thank our Honorable Chancellor Prof. R. Sethuraman for providing us with an opportunity and the necessary infrastructure for carrying out this project as a part of our curriculum.

We would like to thank our Honorable Vice-Chancellor Dr. S. Vaidhyasubramaniam and Dr. S. Swaminathan, Dean, Planning & Development, for the encouragement and strategic

support at every step of our college life.

We extend our sincere thanks to Dr. R. Chandramouli, Registrar, SASTRA Deemed to be University for providing the opportunity to pursue this project.

We extend our heartfelt thanks to Dr. A. Umamakeswari, Dean, School of Computing,

Dr. V. S. Shankar Sriram, Associate Dean, Department of Computer Science and Engineering,

Dr. R. Muthaiah, Associate Dean, Department of Information Technology and Information & Communication Technology and Dr. B.Santhi, Associate Dean, Department of Computer Application.

Our guide Dr. Rajkumar K, Sr. Asst. Professor, School of Computing was the driving force behind this whole idea from the start. His deep insight in the field and invaluable suggestions helped us in making progress throughout our project work. We also thank the project review panel members for their valuable comments and insights which made this project better. We would like to extend our gratitude to all the teaching and non-teaching faculties of the School of Computing who have either directly or indirectly helped us in the completion of the project.

We gratefully acknowledge all the contributions and encouragement from my family and friends resulting in the successful completion of this project. We thank you all for providing us an opportunity to showcase our skills throughout the project.

# 

# **List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | Page No. |
| 1.1 | Workflow of Proposed  Methodologies |  |
| 4.1.1 | Learning curve of random forest |  |
| 4.1.2 | Learning curve of Decision Tree |  |
| 4.1.3 | Learning curve of Gradient Boosting |  |
| 4.1.4 | Learning curve of Hybrid RFLR |  |
| 4.1.5 | Learning curve of Logistic Regression |  |
| 4.1.6 | Learning curve of KNN |  |
| 4.2.1 | Elbow curve of K-means |  |
| 4.2.2 | Scatter plot of K-means |  |
| 4.3.1 | Interface - I |  |
| 4.3.2 | Interface - II |  |

**ABBREVIATIONS:**

|  |  |
| --- | --- |
| ML | MACHINE LEARNING |
| PCOS | POLYCISTIC OVARY SYNDROME |
| AI | ARTIFICIAL INTELLIGENCE |
| BP | BLOOD PRESSURE |
| CP | CHEST PAIN |
| BMI | BODY MASS INDEX |
| FBS | FASTING BLOOD SUGAR |

# **Abstract**

Polycystic Ovary Syndrome (PCOS) is a health disorder that affects around 10 million

women worldwide. Due to a lack of awareness in India, it affects one out of every-five

women. If not diagnosed at an early stage, it leads to various other harmful diseases as well

like diabetes, high blood pressure, obesity, heart disease etc. Thus, PCOS identification at an

early stage is essential. Although several expensive tests are available there is no proper

treatment and people are not aware of the same.

In modern times, obesity has also become a significant threat all over the world. Obesity

means an unnatural or excessive amount of fat that is present in our bodies. People are

constantly moving towards an unhealthy lifestyle, eating excessive junk food, late-night

sleep, and spending a long time sitting down. Adolescents are being affected because of their

unconscious attitudes. It is a medical problem known as a very complex disease. It promotes

the spread of complex illnesses, stroke, heart disease, liver cancer. Consequently, as an

aware multitude of Bangladesh, we have to move forward to prevent this risk of obesity.

The main objective of this project is to find the diseases which can help early identification

of PCOS and also predict the risk of obesity. In a nutshell, we want to answer the question

“Can we identify some commonly known diseases which are taken as an indication of having

PCOS and predict the risk of obesity”.

**KEYWORDS:** PCOS, diabetes, heart disease, obesity, ML

## Table of Contents

**Title PageNo**

[Bonafide Certificate ii](#_Toc17728)

[Acknowledgements iii](#_Toc17729)

[List of Figures iv](#_Toc17730)

[List of Tables iv](#_Toc17731)

[Abbreviations v](#_Toc17732)

[Abstract vi](#_Toc17733)

[1. Summary of the base paper 1](#_Toc17734)

[2 Merits and Demerits of the base paper 5](#_Toc17735)

[3 Source Code 8](#_Toc17736)

[4 Snapshots 15](#_Toc17737)

[5 Conclusion and Future Plans 22](#_Toc17738)

[6 References 23](#_Toc17739)

7 Appendix 24

**CHAPTER 1**

1. **BASE PAPER DETAILS:**

* **Title:** Early identification of PCOS with commonly known diseases: Obesity, diabetes, high blood pressure and heart disease using machine learning techniques
* **Year:** 2023
* **Journal Name:** Expert System with Applications
* **Indexing:** SCI-E
* **Base paper URL:** <https://www.sciencedirect.com/science/article/pii/S0957417423000337#b0135>
* **DOI:** [**https://doi.org/10.1016/j.eswa.2023.119532**](https://doi.org/10.1016/j.eswa.2023.119532)

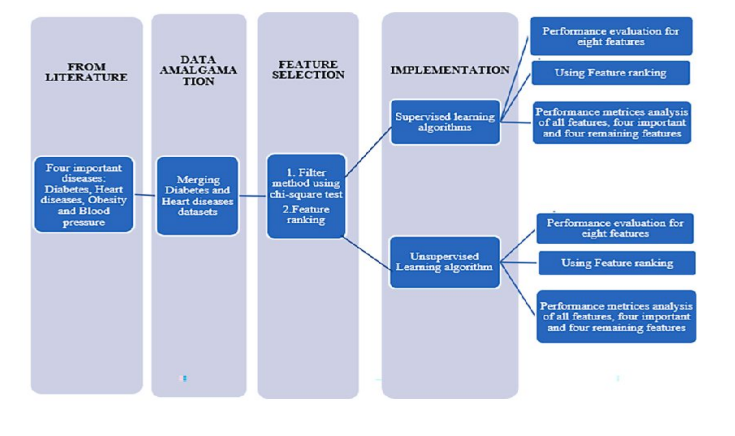
1. **INTRODUCTION:**
   1. **NOVELTY OF THE BASE PAPER:**

* The approach used in the base paper deals only with the important features.
* This results in more accuracy compared to all features.

1. **RESEARCH ADDRESSED:**

* The basic procedure which is used to successfully predict PCOS is:
* To predict PCOS using other major diseases such as heart disease, blood pressure, obesity and diabetes
* The important features are taken and processed in machine learning models and accuracies are obtained
* They produce higher accuracy compared to all features

1. **PROPOSED METHODOLOGIES:**

****

**Fig 1.1 Workflow of Proposed Methodologies**

**CHAPTER 2**

**LITERATURE SURVEY:**

|  |  |  |  |
| --- | --- | --- | --- |
| Title | Work | Method | Results |
| Embedded feature-selection support vector machine for driving pattern recognition | In this work, a more efficient and robust driving pattern recognition technique, extended Support Vector Machine (SVM) with embedded feature selection ability | The recognition results of this extended SVM are compared with results from standard 2-norm SVM and linear 1-norm SVM, using representative driving cycle data to demonstrate the function and superiority of the new technique. | In this work, a Support Vector Machine (SVM) with an embedded feature selection ability is developed |
| Polycystic Ovary Syndrome and Risk of Type 2 Diabetes, Coronary Heart Disease, and Stroke | The fact that PCOS is a syndrome with multiple features complicates efforts to establish causality between PCOS and adverse outcomes because individual features may contribute differentially to outcomes | Conducted a two-sample Mendelian randomization study to investigate the associations of PCOS with type 2 diabetes, coronary heart disease (CHD), and stroke | There is insufficient evidence to support the association between genetically predicted PCOS and diabetes in European women |

|  |  |  |  |
| --- | --- | --- | --- |
| Title  Polycystic ovary syndrome (PCOS) and COVID-19: an overlooked female patient population at potentially higher risk during the COVID-19 pandemic | Work    Emerging data link the risk of severe COVID-19 with certain factors such as hyper-inflammation,low vitamin D levels, and hyperandrogenism, all  have known direct associations with PCOS | Method  Particularly since women with PCOS often receive fragmented care from multiple healthcare services | Results  Despite the immense challenges posed by the COVID-19 outbreak to the healthcare systems in affected countries, attention should be directed to maintain a high standard of care for complex patients such as many women with PCOS |
| An Analysis of PCOS Disease Prediction Model Using Machine Learning Classification Algorithms | A manual feature selection is made over it and  Machine learning algorithms are used to identify the important features to diagnose PCOS. | SVM,logistic Regression,Gradient Boosting and Random forest | There are several types of research progressing in the direction of diagnosing PCOS, but till now, the relevant features are not identified for the same |
| Determining the representative features of PCOS via design experiments | The objective is to find the most important parameters for identifying the PCOS with the Design of experiments (DOE) | Experiment has been carried out for each of the 2k-p potential parameter combinations | The features obtained was not suitable for performing ML Techniques |

|  |  |  |  |
| --- | --- | --- | --- |
| Title | Work | Method | Results |
| Cardiometabolic Risk in PCOS: More than a Reproductive Disorder | The pathogenesis of PCOS metabolic symptoms and the relationship between metabolites and the pathophysiology of PCOS. | PCOS is also associated with metabolic abnormalities, including insulin resistance and β-cell dysfunction | PCOS increases the risk for cardiovascular disease |
| PCOS and diabetes mellitus: from insulin resistance to altered beta pancreatic function, a link in evolution | The assessment of IR may be performed with the hyperinsulinemic clamp (based on the infusion of a specific amount of insulin as a constant level of 100 U/mL for 120 min). | BAT-mediated process,  PPAR-gamma receptor function process | Sedentary lifestyle and over nutrition may be accelerating causes for metabolic PCOS |
| Associations of preconception Body Mass Index in women with PCOS and BMI and blood pressure of their offspring | Women with polycystic ovary syndrome (PCOS) have unfavorable metabolic profiles. Their offspring may be affected by such risks. The objective of the current study was to disclose associations between preconception health of these women and health of their offspring | The offspring follow-up was conducted in the two earlier mentioned age categories. Similar in both groups was a parental questionnaire in which parental and children’s habits and health were evaluated: family history regarding cardiovascular events and diabetes mellitus, anthropometrics, intoxications (smoking, alcohol, and drug use), physical exercise, and chronic diseases | A total of 74 women-child pairs were included: 42 children in the young offspring group (2.5–4 years) and 32 children in the older group (6–8 years). The offspring was female in 49% (*n* = 36) of all cases |

**MERITS AND DEMERITS:**

**Merits:**

1. It has been observed from the literature that if a woman is having PCOS, then there is a high chance that she also has heart disease, obesity, high blood pressure or diabetes. Generally, women go for regular check-ups of these commonly known diseases but they are unaware of PCOS tests. This prediction model does the opposite by using some metrics of these 4 major diseases and uses it to predict PCOS.
2. This prediction model uses only 8 features instead of the original 20 features in its dataset and provides a good accuracy in its prediction.

**Demerits:**

1. The amalgamation process to obtain the final dataset from different datasets is difficult.
2. The amount data used in this model is only around 1000 and much more data will be required to get a more efficient result

**CHAPTER 3**

**SOURCE CODE:**

**# import libraries**

*from sklearn.feature\_selection import SelectKBest*

*from sklearn.feature\_selection import chi2*

*import seaborn as sns*

*import matplotlib. pyplot as plt*

*import numpy as np*

*import pandas as pd*

**# importing the dataset**

*data=pd.read\_csv("/content/finaldataset.csv")*

*x=data.drop('PCOS',axis=1)*

*y=data.PCOS*

*from sklearn.model\_selection import train\_test\_split*

*X\_train, X\_test,y\_train, y\_test = train\_test\_split(x,y,random\_state=104,test\_size=0.25,shuffle=True)*

**# random forest algorithm**

*from sklearn.ensemble import RandomForestClassifier*

*clf = RandomForestClassifier(criterion="entropy")*

*clf.fit(X\_train, y\_train)*

*y\_pred = clf.predict(X\_test\_)*

**# gradient boosting**:

*import numpy as np*

*from sklearn.ensemble import GradientBoostingClassifier*

*gbc = GradientBoostingClassifier(n\_estimators=100, learning\_rate=0.5, max\_depth=1)*

*gbc.fit(X\_train, np.ravel(y\_train, order='C'))*

*y\_pred = gbc.predict(X\_test)*

**# k nearest neighbour**

*from sklearn.neighbors import KNeighborsClassifier*

*from sklearn.model\_selection import train\_test\_split*

*knn = KNeighborsClassifier(n\_neighbors=10)*

*knn.fit(X\_train, y\_train)*

*y\_pred=knn.predict(X\_test)*

*from sklearn.metrics import precision\_score, \recall\_score, confusion\_matrix, classification\_report, \accuracy\_score, f1\_score*

*print('confusion matrix',confusion\_matrix(y\_test,y\_pred))*

*print('Accuracy:', accuracy\_score(y\_test, y\_pred))*

*print('F1 score:', f1\_score(y\_test,y\_pred ))*

*print('Recall:', recall\_score(y\_test, y\_pred))*

*print('Precision:', precision\_score(y\_test,y\_pred))*

**#Logistic regression**

*from sklearn import linear\_model*

*from sklearn.linear\_model import LogisticRegression*

*model=LogisticRegression(max\_iter=3000)*

*model.fit(X\_train,y\_train)*

*y\_pred=model.predict(X\_test)*

**# SVM classifier**

*from sklearn.svm import SVC*

*svm = SVC(kernel='linear')*

*svm.fit(X\_train, y\_train)*

*y\_pred=svm.predict(X\_test)*

**#Decision Tree**

*from sklearn.tree import DecisionTreeClassifier*

*from sklearn import metrics*

*dtree = DecisionTreeClassifier(criterion="entropy")*

*dtree.fit(X\_train, y\_train)*

*y\_pred = dtree.predict(X\_test)*

**# hybrid RFLR**

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=42)*

*rf = RandomForestClassifier(n\_estimators=10, random\_state=42)*

*rf.fit(X\_train, y\_train)*

*rf\_features = rf.apply(X\_train)*

*X\_train\_new = X\_train.copy()*

*X\_train\_new = np.concatenate((X\_train\_new, rf\_features), axis=1)*

*lr = LogisticRegression(random\_state=42)*

*lr.fit(X\_train\_new, y\_train)*

*rf\_features\_test = rf.apply(X\_test)*

*X\_test\_new = X\_test.copy()*

*X\_test\_new = np.concatenate((X\_test\_new, rf\_features\_test), axis=1)*

*y\_pred = lr.predict(X\_test\_new)*

**#KMEANS**

*from sklearn.cluster import KMeans*

*kmeans = KMeans(n\_clusters=2, init ='k-means++', max\_iter=300, n\_init=10,random\_state=0 )*

*y\_kmeans = kmeans.fit\_predict(x)*

*from sklearn.metrics import silhouette\_score*

*print(f'Silhouette Score(n=2): {silhouette\_score(x,y\_kmeans)}')*

*from sklearn.metrics import davies\_bouldin\_score*

*print(f'davies\_bouldin\_score(n=2): {davies\_bouldin\_score(x,y\_kmeans)}')*

**RESULT ANALYSIS:**

**SUPERVISED LEARNING:**

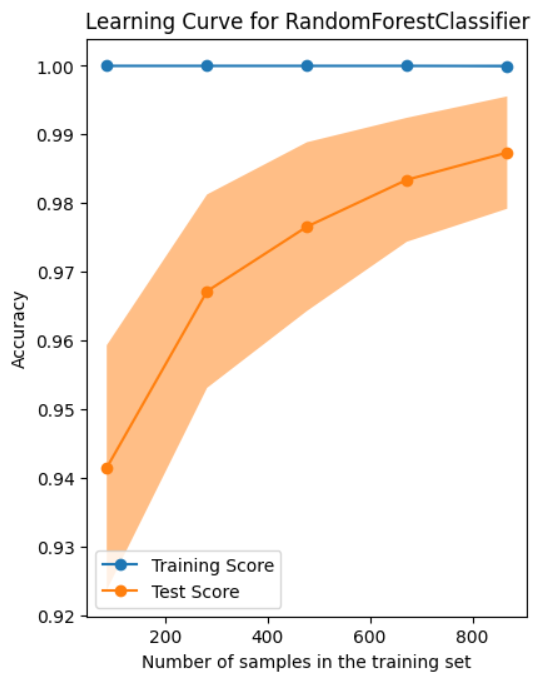
|  |  |  |
| --- | --- | --- |
| Algorithm | Accuracy (IF) | Accuracy (AF) |
| Random Forest | 98.5 | 98.8 |
| Gradient Boosting | 98.5 | 98.8 |
| KNN | 93.75 | 85.8 |
| Logistic Regression | 94.8 | 94.1 |
| SVM | 94.1 | 90.5 |
| Decision Tree | 98.9 | 96.4 |
| Hybrid RFLR | 98.1 | 96.4 |

**UNSUPERVISED LEARNING:**

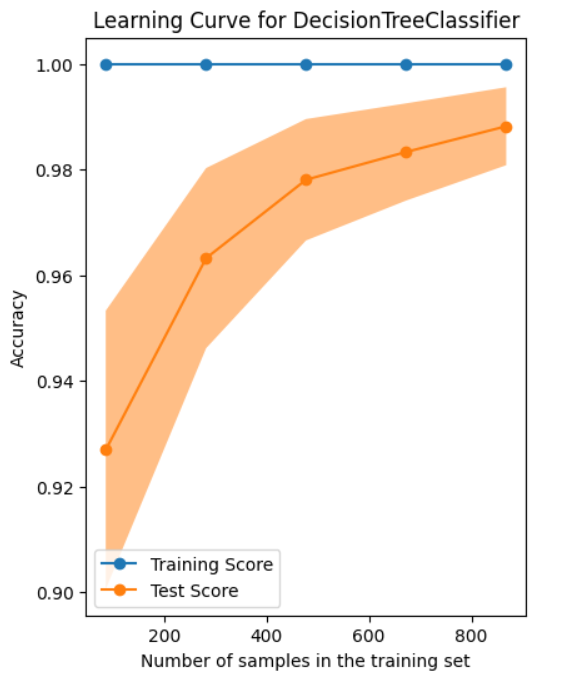
|  |  |  |
| --- | --- | --- |
| Algorithm | Silhouette score | Davies Bouldin score |
| k-means | 68.6 | 37.1 |

**CHAPTER 4**

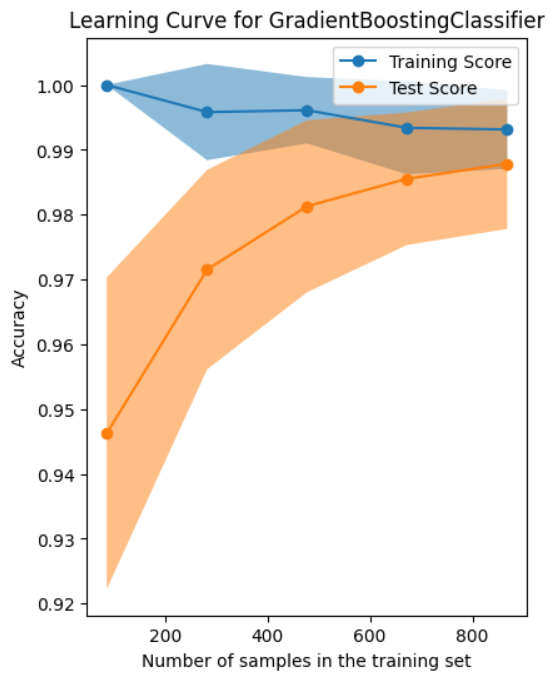
**OUTPUTS OF SUPERVISED LEARNING ALGORITHM**

****

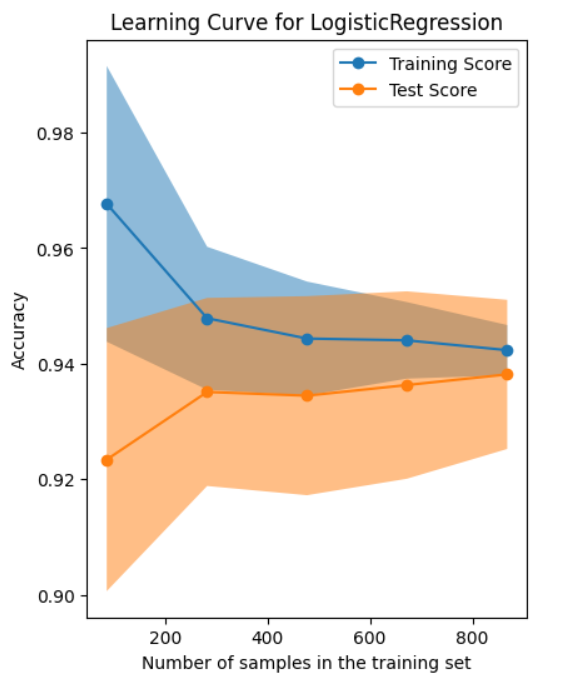
**Fig :4.1.1 learning curve of Random Forest Algorithm**

****

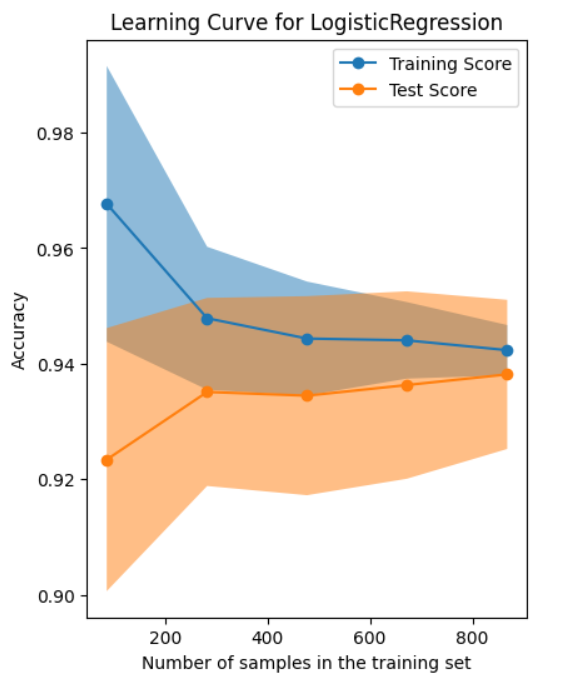
**Fig:4.1.2 learning curve of Decision Tree Algorithm**



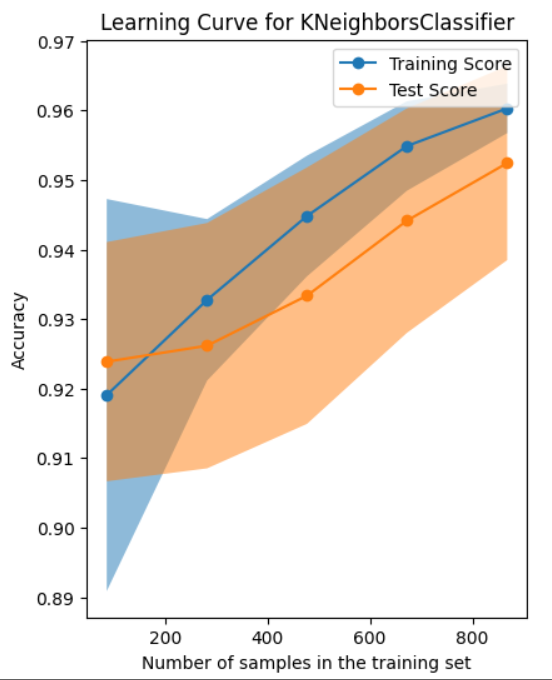
**Fig: 4.1.3 learning curve of Gradient Boosting Algorithm**

****

**Fig: 4.1.4 learning curve of Hybrid RFLR**

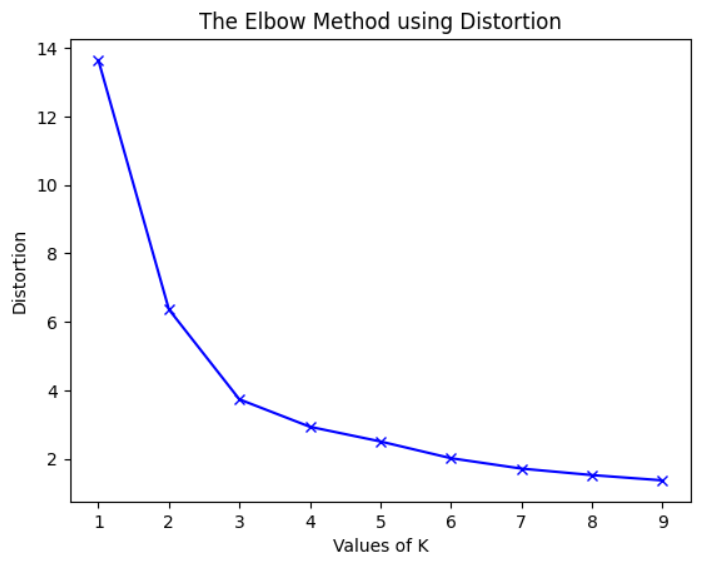
****

**Fig:4.1.5 learning curve of logistic regression**

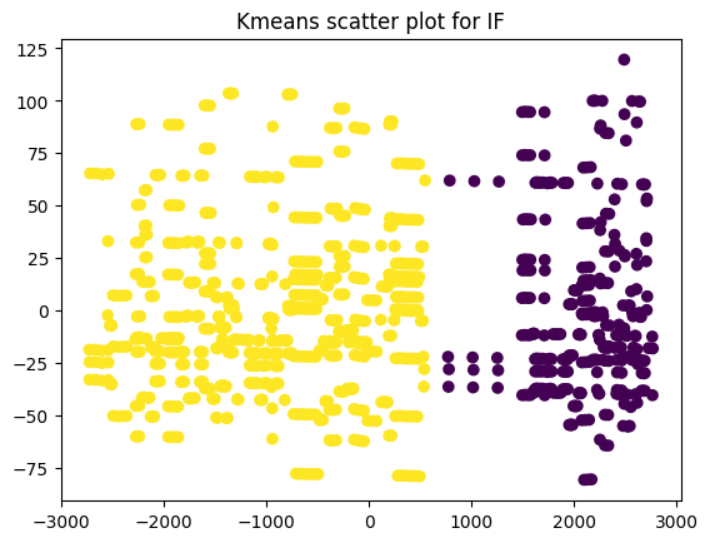
****

**Fig: 4.1.6 learning curve of KNN**

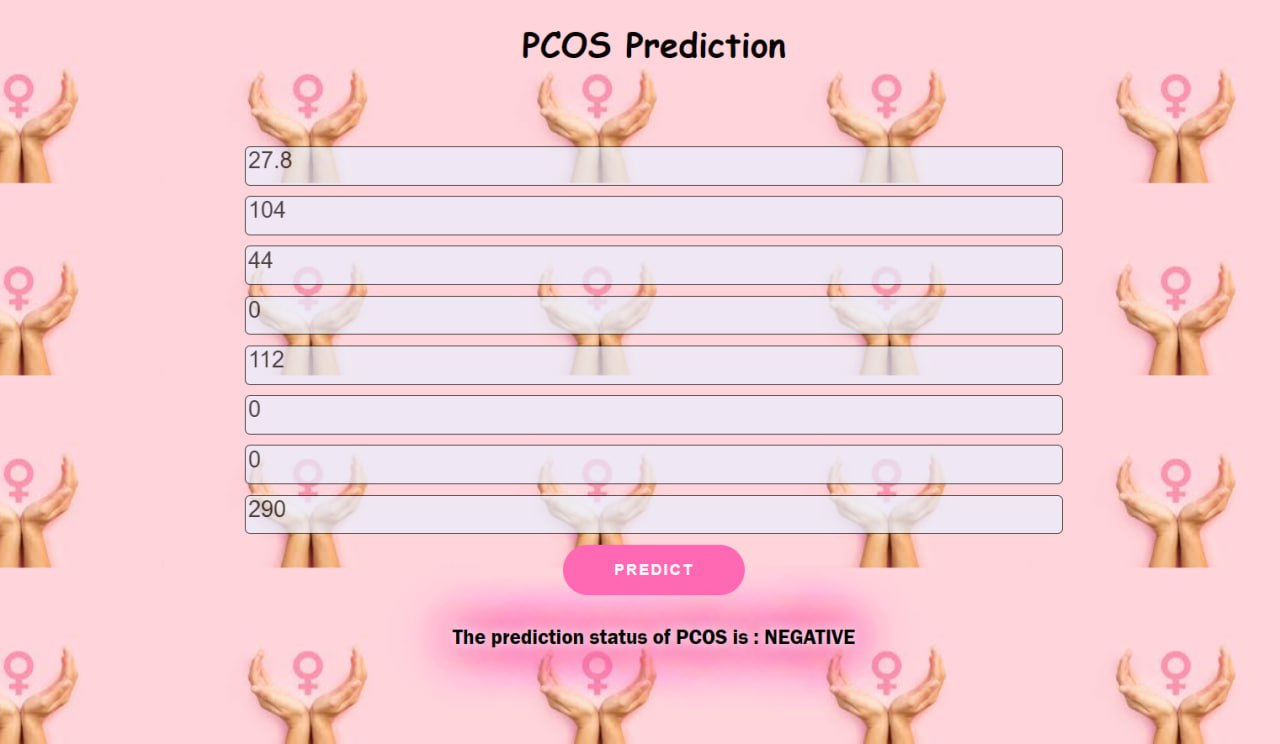
**SNAPSHOTS OF UNSUPERVISED LEARNING ALGORITHMS**

****

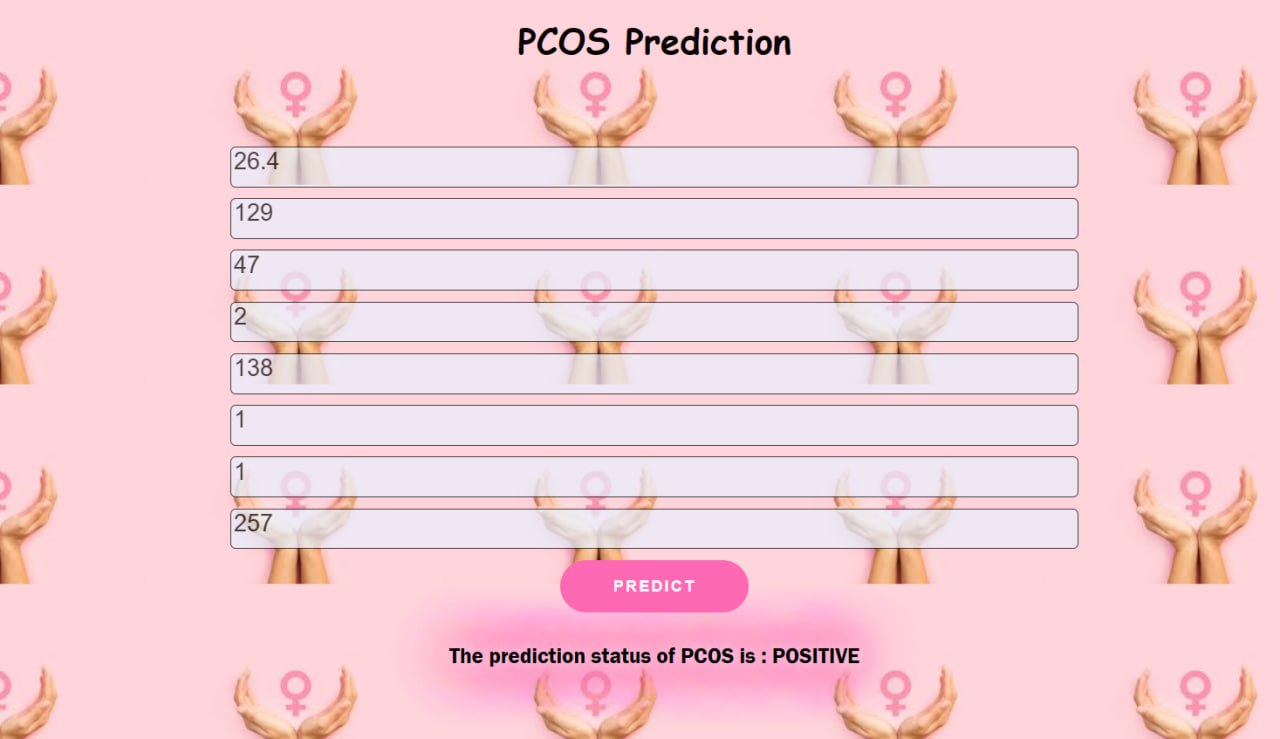
**Fig:4.2.1 Elbow plot for k-means**

****

**Fig: 4.2.2 Scatter plot of k-means**



**4.3.1 Interface snapshot I**



**4.3.2 Interface snapshot II**

**CHAPTER 5**

**CONCLUSION AND FUTURE PLANS**

**1) CONCLUSION:**

* From the performance metrics of the model obtained, it shows that important features provide better accuracies of the algorithms compared to all features such as,
* **RANDOM FOREST (98.5 %)**
* **DECISION TREE (98.9 %)**
* **GRADIENT BOOSTING (98.5 %)**
* **K-NEAREST NEIGHBOR (93.75 %)**
* **LOGISTIC REGRESSION (94.8 %)**
* **SUPPORT VECTOR MACHINE (94.1 %)**
* **HYBRID RFLR (98.1 %)**
* Thus, we conclude that decision tree has the best metric among the supervised algorithms tested for the present dataset followed by gradient boosting algorithm.

**2) FUTURE PLANS:**

* We are planning to publish the model as an awareness website which can be accessed by everyone who are in need of it.
* As per the survey taken and based on the data available in the base paper quoting “Despite affecting millions of women and the serious health results, PCOS is unknown to most people. It is more shocking when studies reveal that around 50 % of the women living with PCOS are going undiagnosed (Escobar-Morreale, 2018)”, we are planning to create a separate awareness website which will hold the details regarding PCOS and also possible treatments and health conditions
* Along with this prediction model and additional feature of indicating obesity at an early age. With respect to the model, we plan to test with additional supervised learning algorithms and check for better accuracy.

**CHAPTER 6**

**REFERENCES:**

* Aggarwal, S., & Pandey, K. (2021). An Analysis of PCOS Disease Prediction Model Using Machine Learning Classification Algorithms. Recent Patent of Engineering, 15(6), 53–63. <https://doi.org/10.2174/1872212115999201224130204>
* Glueck, C. J., & Goldenberg, N. (2019). CHARACTERISTICS OF OBESITY IN POLYCYSTIC OVARY. Metabolism, 92, 108–120. [https://doi.org/10.1016/j. metabol.2018.1100%02](https://doi.org/10.1016/j.%20metabol.2018.1100%02)
* Wilson, M. S., & Metink-Kane, M. M. (2012). Polycystic Ovary Syndrome and Risk for Long-Term Diabetes and Dyslipidemia. Obstet Gynecol, 23(1), 6–13. [https://doi.org/ 10.1097/AOG.0b013e31820209bb.](https://doi.org/%2010.1097/AOG.0b013e31820209bb.)
* <https://www.sciencedirect.com/science/article/pii/S095741742300033>
* <https://www.sciencedirect.com/science/article/pii/S2666518221000401>
* Centers for Disease Control and Prevention. (2020). PCOS (Polycystic Ovary Syndrome) and Diabetes. (n.d.). <https://www.cdc.gov/diabetes/basics/pcos.html>. Accessed February, 2022.
* Chen, W., & Pang, Y. (2021). Metabolic Syndrome and PCOS: Pathogenesis and the Role of Metabolites. Metabolites, 11(12). <https://doi.org/10.3390/metabo11120869>
* Glueck, C. J., & Goldenberg, N. (2019). CHARACTERISTICS OF OBESITY IN POLYCYSTIC OVARY. Metabolism, 92, 108–120[. https://doi.org/10.1016/j. metabol.2018.1100%02](.%20https:/doi.org/10.1016/j.%20metabol.2018.1100%02)