Diet Recommendation for Poly Cystic Ovarian Syndrome of Indian Patients Using Multi-Attribute and Multi-Labeling Classifier

Santhi Selvaraj[1*], S. Selva Nidhyananthan[2], R. Vanmathi[3], M. Ramya[4]

[1*,3,4]Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.

[2]Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.

[1*]E-mail: santhi31@gmail.com

Abstract

Polycystic Ovarian Syndrome (PCOS) is a poorly understood, under-diagnosed, and under-treated condition, with prevalence ranging from 2.2 percent to 26 percent worldwide. A prevalent endocrine disorder in women of childbearing age, PCOS is a syndrome that results in the development of ovarian cysts and may ultimately cause infertility. Oily skin, darker acne scars, weight gain, hypertension, and irregular menstrual cycles are a few of the prevalent symptoms. According to these symptoms, we have taken the PCOS dataset from Kaggle which contains 541 records and 43 attributes including patient number and target class. This target class contains the labels as 1 for PCOS affected and 0 for normal women. This data was collected from 10 nearest hospital from Kerala in India. Among 43 features, we selected 21 essential features based on gynecologist suggestions and recommending a proper diet. This work focuses on the prevention of both PCOS patients and normal women by recommending diet through the construction of rule set. We transform the existing target label into a new multi-label target class that includes the dietary class by using this rule set. This system recommends a PCOS diet based on clinical data of the patients using machine learning techniques such as K-Nearest Neighbor, Decision Tree, Random Forest classifier, and Multi-Layer Perceptron. Based on a variety of evaluation indicators, the findings are analyzed, and the performance of the algorithms is validated. This type of analysis is useful for early prevention and safe recovery of PCOS patients by recommending nutrition diets.

Keywords: PolyCystic Ovarian Syndrome - Support Vector Machine - K-Nearest Neighbor - Decision Tree - Random Forest - Multi-Layer Perceptron - Diet recommendation.

DOI: 10.47750/pnr.2022.13.S03.255

INTRODUCTION

A recent study has revealed about 19% of the women in North India, 26% from the East, 20% from the West, and 18% from south India suffer from this syndrome. A study conducted in AIIMS illustrates twenty to twenty-five percentage of Indian women of reproduction age are affected for PCOS disease [1].

Nowadays, most women are facing one of the major problems are PolyCystic Ovarian Syndrome (PCOS) in worldwide [2]. It is a hormonal condition that affects women of reproductive age from 15 to 30 years. Women with PCOS includes changes in their androgen hormone levels, irregular menstrual periods and skin related problems. This particular form of PCOS stops the ovaries from releasing eggs on a regular basis and converts them into numerous little fluid or follicle collections. In addition to weight loss, early detection and treatment of polycystic ovary syndrome may lower the risk of long-term consequences such type 2 diabetes, heart disease, cervical cancer, and endometrial cancer. Short-term complications of PCOS [3] includes infertility, sleep apnea, depression, anxiety, eating disorders, miscarriage or premature birth, high blood pressure and blood sugar, gestational diabetes, abnormal uterine bleeding.

Vikas B, et al. [4] addressed the disorder of PCOS which leads to dangerous illnesses like Type 2 Diabetes, Cardiovascular and anxiety disorders which can cause depression independent of BMI. In this study, different signs of PCOS were investigated and discovered, including hirsutism, acne, male pattern baldness, irregular menstrual cycles, oily skin, darkening acne scars, and a failure to respond to ovulation induction. They also listed various treatments of PCOS as Metformin therapy, Progesterone-based oral contraceptives, etc. The algorithms suggested by this work concentrates on records of patients, analyzing the data, and possible solutions of treatment were obtained for diagnosing PCOS. This disorder that is growing more prevalent because women's eating habits and lifestyles changed. [5].

PCOS is typically diagnosed with a pelvic exam to examine the reproductive organs, blood tests to determine hormone levels, and ultrasounds to examine the ovaries, endometrial layer, and uterus [6]. Consequently, the treatment approach benefited from an early diagnosis of PCOS.

Recently, a number of machine learning and deep learning techniques have produced hopeful outcomes in medical diagnostics. [7]. Machine learning algorithms are helpful to improve the prediction and clustered levels of medical data by repeatedly doing the process through their experience. It builds model based on training data and send the test data for making the better decisions in medical field.

The diagnosis of PCOS disease is how much important like that the prevention of the disease also important. This PCOS will be cured by not only taking the medicine in addition with that correctly follow the diets and exercises. In this regard, we propose this work for recommending the PCOS diets for affected women in Kaggle dataset according to doctor's suggestions by using machine learning algorithms.

Related Works

Some works employed various models to analyze different concerns about PCOS. We summarize them below regarding the techniques employed. Palvi Soni et al. [8] described various data mining techniques for predicting PCOS. Naive Bayes Classifier, Decision Tree and Support Vector Machine were used for classifying and predicting the PCOS from extracted features of the data. This work also used for association rule mining for finding relationship between symptoms, clustering for group the similar kind of symptoms and outlier detection for finding unwanted clusters.

Clinical and metabolic factors serve as an early marker for PCOS, and Amsy Denny et al. [9] suggested a strategy for the earlier prediction of this condition. The Kaggle PCOS data sets required for their system, which consists of 541 women with 43 clinical parameters. Effective eight potential features are selected out of 43 features by using SPSS (Statistical Package for the Social Sciences) and Principal Component Analysis (PCA). Classification of PCOS was done by using different machine learning techniques such as Logistic Regression, K-Nearest neighbor, Naive Bayes classifier, Support Vector Machine, Random Forest Classifier, Classification and Regression Trees (CART). Random Forest Classifier was the most appropriate and accurate method compare to other classifiers for PCOS prediction with an accuracy of 89.02%.

Vedpathak, et al. [10] detected whether a woman was suffering from PCOS or not by using five different machine learning classifiers like Gaussian Naive Bayes, Logistic Regression, SVM, Random Forest and K-Nearest Neighbor. The same Kaggle PCOS dataset was taken for this work within that the top features were recognized using the Chi-Square method and converted into feature vectors. Random Forest Classifier achieved the highest accuracy and the most reliable compared to other classifiers.

Neetha Thomas et al. [11], predicted the PCOS with a clinical data obtained from various clinics throughout the Thodupzha Municipality. This work used hybrid classification algorithm for combining both Bayesian classifier and Neural Network classifier. These two classifiers predicted values have cross analyzed and produced a superior outturn. This system was one of the hybrid structures, which was used for determining the likelihoods of PCOS and producing the best result. PCOS prediction performance parameters like Accuracy, Precision, Recall, F-measure, and specificity of the system have been evaluated. The Navies Bayes algorithm predicted output with the least accuracy compared to Neural Networks.

Bharati, et al. [12], predicted the PCOS for Kaggle data using machine learning algorithms. This system was detected the 177 women have PCOS among 541 women based on 43 attributes. Initially, select the essential attributes for PCOS prediction by applying univariate attribute selection algorithm. Follicle-stimulating hormone (FSH) and Luteinizing hormone (LH) ratio attribute was selected this method by computing the ranking of the attributes. The training and testing data were then separated from the dataset using holdout and cross-validation techniques. Logistic Regression, Random Forest, Hybrid Random Forest and Logistic Regression (RFLR) and Gradient Boosting classifiers were applied for detection of PCOS from Kaggle dataset. RFLR parades the best accuracy for testing data of 91.01% and 90% of recall using 40-fold cross-validation of significant attributes.

Malik Mubasher Hassan et al. [13] diagnosed PCOS disease using Naive Bayes, Logistic Regression, SVM, Random Forest, Classification and Regression Trees according to clinical data of the patients. The result was evaluated, and the algorithms' effectiveness was confirmed using the Kappa Coefficient, accuracy, precision, recall, and F-statistics. The Random Forest algorithm showed the best accuracy of 96% according to the validation measures.

Preethi Chauhan et al. [14] proposed a mobile application based PCOS prediction system using Machine Learning techniques. The dataset was formed by collecting review from application and it was cleaned using Python with NLP tools. The features and its importance were calculated by applying the Gini coefficient. Various machine learning algorithms were used to classify PCOS and Decision Tree Classifier achieved the best accurate model. The PCOS was predicted at earlier stage by using developed mobile application.

Bhat et al. [15], detected the PCOS by using various machine learning classifiers and compare the results. Data was collected from the Kaggle repository, which was preprocessed and selected the features using feature selection mechanism for detecting PCOS at earlier. Next, data splitting and non-sampling were done by using the SMOTE (Synthetic Minority Over-sampling Techniques) function. Following this, they built the model and applied the machine learning algorithms which were evaluated based on resultant matrices.

Homay Danaei Mehr et al. [16], proposed a system for diagnosing the PCOS by using ensemble classifiers for Kaggle dataset. The performance of various classifiers like Ensemble Random Forest, Decision Tree, and Multi-Layer

Perceptron (MLP) was examined using the dataset with all attributes. Ensemble Random Forest classifier performed best compared to other classifiers with an accuracy of 98.89%.

Subrato Pijush Dutta et al. [17] built a prediction model using Synthetic Minority Oversampling Technique (SMOTE) with machine learning algorithms, which industrializes the PCOS detection at earlier. The classification of PCOS was done by eliminating the missing values in datasets, and the strongest attributes were selected using PCA. Then the model was learnt and its efficiency was found in terms of classification accuracy, Training Time, F1-score, Recall (Sensitivity), Precision & Area under the ROC. SMOTE based Logistic Regression model achieved all the mentioned evaluation metrics.

Wanyun Cui et al. [18] developed the open rule induction model which was worked by combining knowledge-based rule induction and language model (LM)-based rule generation. Without checking annotated rules, this system was mined open rules from LMs automatically. This model was unsupervised and optimized the language model by continuously giving the training. They tested the quantity and quality of the inducted and generated open rules and these rules outperformed compared to the manual annotated rules and also identify the errors in LMs.

Dhatri Ganda et al. [19] surveyed different multi-label classification algorithms based on label ranking. They applied various multi-labeling algorithms into two groups such as problem transformation and algorithm adaptation. They suggested different problem transformation methods such as Binary Relevance, Label Powerset, Random K-Label Sets, Calibrated Label Ranking, Ranking by Pair wise Comparison, Pruned Sets (PS)/ Pruned Problem Transformation. They used the following multi-labeling methods for algorithm adaptation such as KNN, Back Propagation, Multi-Class Multi-Labeling Association, Tree Based Boosting and Hierarchical methods. They evaluated the algorithms using different measures like Accuracy, Exact match ratio, Precision, Average Precision, F1-measure, Hamming Loss and Ranking Loss.

Eyke H¨ullermeier et al. [20] suggested the multi-label classification (MLC) rule generation and rule learning. This included the rules' potential interpretability, their flexibility in modelling label dependencies, and their ease in easily adapting a predictor to various loss functions. They presented the rule-based MLC modular framework and discussed the challenges and opportunities of multi-label rule learning.

PROPOSED WORK

The above all related works only focused on diagnosing the PCOS based on clinical symptoms. According to the Kaggle dataset and other resources, the researchers suggested only prediction and detection of PCOS disease using Machine Learning classifiers. But in this proposed work helps to recommend the diet by diagnosing the PCOS patients and normal women who have slight symptoms from the Kaggle dataset. The proposed system comprises modules to explore PCOS data and to recommend a diet according to symptoms. Figure 1 shows the overall scheme of Diet Recommendation for PCOS symptoms from Kaggle dataset.

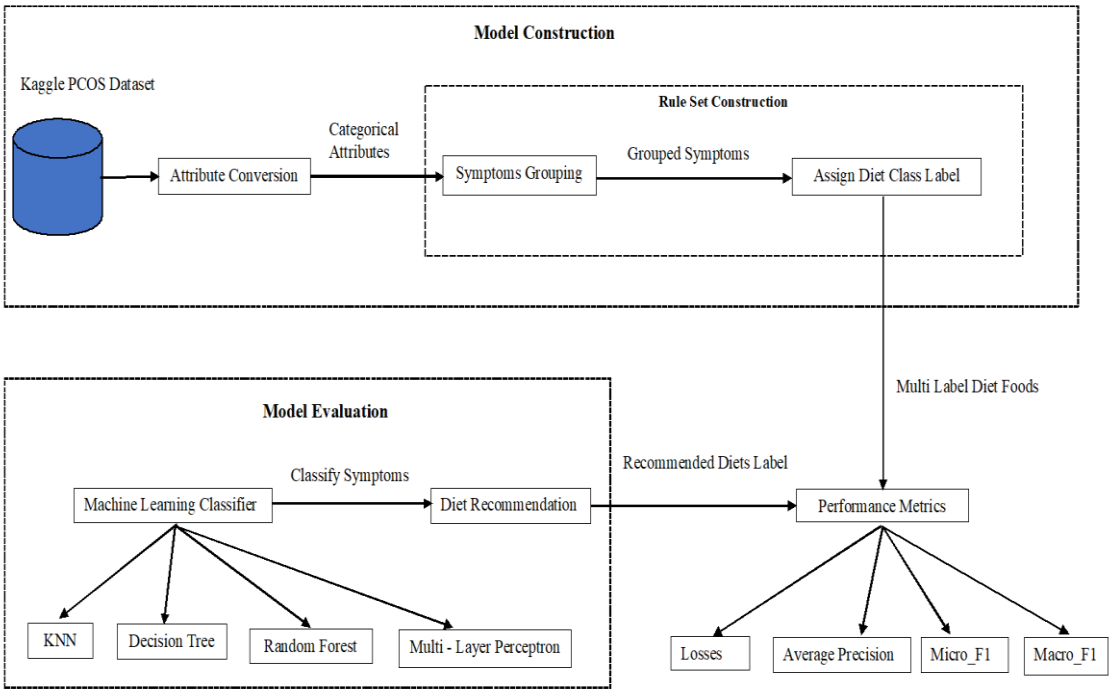

Figure 1. The System Design of Diet Recommendation for PCOS Patients

1. Model Construction

In this phase, the Kaggle dataset is preprocessed and assigned diet class labels by constructing the rule set induction.

Dataset Description

The dataset of this work is PCOS without fertility dataset from Kaggle repository. It contains 541 instances and 43 attributes including patient number and target class. This dataset was collected from different hospitals from Kerala in India. This dataset contains target class labels as 1 for PCOS affected women and 0 for normal women. Table 1 shows the various attributes and description of PCOS Kaggle dataset. Among 43 attributes, 13 attributes are categorical value and 30 attributes are numerical value.

Table 1: Attributes and Description of Kaggle PCOS Dataset

#	Attribute Name	Description	Numerical / Categorical
1.	Patient File Number	Patient ID number.	Categorical
2.	PCOS(Y/N) (Target Class Label)	It shows whether the patient has Poly-cystic ovary syndrome (PCOS) or not.	Categorical
3.	Age (in Yrs)	Age of the patient in Years.	Numerical
4.	Weight (in Kg)	Weight of the patient in Kilograms.	Numerical
5.	Height (in Cm)	Height of the patient in Centimeters.	Numerical
6.	BMI	Body Mass Index of the patient.	Numerical
7.	Blood Group	Blood Group of the patient.	Categorical
8.	Pulse Rate	Pulse rate of the patient in beats per minute (bpm).	Numerical
9.	RR	Number of breaths per minute.	Numerical
10.	HB (in g/dl)	Hemoglobin concentration of the patient	Numerical
11.	Cycle (R/I)	Menstrual cycle of the patient.	Categorical
12.	Cycle Length (in days)	Number of days of the menstrual cycle of the patient.	Numerical
13.	Marriage status	Marital status of the patient	Categorical
14.	Pregnant (Y/N)	Whether the patient is pregnant or not.	Categorical
15.	No. of Abortions	Number of abortions carried out by the patient	Numerical
16.	I beta-HCG (in mIU/mL)	Beta-human chorionic gonadotropin hormone level of the patient.	Numerical
17.	II beta-HCG (in mIU/mL)	II beta-human chorionic gonadotropin hormone level of the patient.	Numerical
18.	FSH (in mIU/mL)	Follicle Stimulating Hormone level of the patient.	Numerical
19.	LH (in mIU/mL)	Luteinizing Hormone level of the patient.	Numerical
20.	FSH/LH	Follicle Stimulating Hormone and Luteinizing Hormone ratio in the patients' blood.	Numerical
21.	Hip (in inch)	Hip size of the patient.	Numerical
22.	Waist (in inch)	Waist size of the patient.	Numerical
23.	Waist : Hip Ratio	Waist to Hip ratio of the patient.	Numerical
24.	TSH (in mIU/mL)	Thyroid Stimulating Hormone level of the patient.	Numerical
25.	AMH (in ng/mL)	Anti-Mullerian Hormone level of the patient.	Numerical
26.	PRL (in ng/mL)	Pro Lactin Hormone level of the patient.	Numerical
27.	Vit D3 (in ng/mL)	Vitamin D3 level of the patient.	Numerical
28.	PRG (in ng/mL)	Progesterone hormone level of the patient.	Numerical
29.	RBS (in mg/dl)	Random Blood Test in of the patient.	Numerical
30.	Weight Gain (Y/N)	Whether the patient has a Weight gain or not.	Categorical
31.	Hair Growth (Y/N)	Whether the patient has abnormal hair growth.	Categorical
32.	Skin Darkening (Y/N)	Whether the patient had skin darkening.	Categorical
33.	Hair Loss (Y/N)	Whether the patient has hair loss or not.	Categorical
34.	Pimples (Y/N)	Whether the patient has pimples or not.	Categorical
35.	Fast Food (Y/N)	Whether the patient eats fast food or not.	Categorical
36.	Regular Exercise (Y/N)	Whether the patient does regular exercise	Categorical
37.	BP Systolic (in mmHg)	Blood Pressure of the patient during systolic cycle of the patient.	Numerical
38.	BP Diastolic (in mmHg)	Blood Pressure of the patient during diastolic cycle of the patient.	Numerical
39.	Follicle no. In Left Ovary	Number of follicles in the left ovary of the patient.	Numerical
40.	Follicle no. In Right Ovary	Number of follicles in the right ovary of the patient.	Numerical
41.	Avg. Follicle Size (Left) (in mm)	Average follicle size in left ovary of the patient.	Numerical
42.	Avg. Follicle Size (Right) (in mm)	Average follicle size in right ovary of the patient.	Numerical
43.	Endometrium (in mm)	Endometrium size of the patient.	Numerical

Attribute Conversion

In our diet recommendation system doesn't need all 43 features including target label. So, we reduce 43 attributes into 21 essential attributes and also consider the target class. Categorical attributes are maintained as the same format and numerical attributes have changed into categorical form for construction of rule set induction in easy manner. Table 2 shows the chosen attributes conversion format according to normal range of each attribute.

Table 2: Attributes Conversion based on Standard Range

#	Selected Attribute Name	Old Form	New Form	Range
1.	Age	Numerical	Numerical	From 20 to 48 in Years
2.	BMI	Numerical	Categorical	BMI>=32 - 1 BMI<32 - 0
3.	Cycle (R/I)	Categorical	Categorical	R – Regular I - Irregular
4.	Cycle Length (in days)	Numerical	Categorical	3-5 days – Regular Otherwise - Irregular
5.	Pregnant (Y/N)	Categorical	Categorical	Y – Pregnant N – Not Pregnant
6.	No. of Abortions	Numerical	Categorical	Abortions>0 - 1 No Abortions - 0
7.	FSH/LH	Numerical	Categorical	Range [1.3:7] – Normal Otherwise – Abnormal
8.	TSH (in mIU/mL)	Numerical	Categorical	Range [0.2:4.7] – Normal Otherwise – Abnormal
9.	AMH (in ng/mL)	Numerical	Categorical	Range [0.7:4] – Normal Otherwise – Abnormal
10.	PRL (in ng/mL)	Numerical	Categorical	Range [0:20] – Normal Otherwise – Abnormal
11.	PRG (in ng/mL)	Numerical	Categorical	Range [5:20]– Normal Otherwise – Abnormal
12.	Weight Gain (Y/N)	Categorical	Categorical	Y – Weight Gain N – Normal Weight
13.	Hair Growth (Y/N)	Categorical	Categorical	Y – Hair Growth in Face N – No Hair Growth
14.	Skin Darkening (Y/N)	Categorical	Categorical	Y – Dark Skin N – Normal Skin
15.	Hair Loss (Y/N)	Categorical	Categorical	Y – Hair Loss N – No Hair Loss
16.	Pimples (Y/N)	Categorical	Categorical	Y – Pimples N – Normal or No pimples
17.	Follicle no. In Left Ovary	Numerical	Categorical	From 1 to 10 – Normal >10 - Abnormal
18.	Follicle no. In Right Ovary	Numerical	Categorical	From 1 to 10 – Normal >10 - Abnormal
19.	Avg. Follicle Size (Left) (in mm)	Numerical	Categorical	From 2 to 17 – Normal >17 - Abnormal
20.	Avg. Follicle Size (Right) (in mm)	Numerical	Categorical	From 2 to 17 – Normal >17 - Abnormal
21.	Endometrium (in mm)	Numerical	Categorical	From 2 to 11 – Normal >11 - Abnormal

Rule Set Induction

Rule Induction is one of the supervised machine learning techniques and it creates if -then rules based on attributes of the dataset. The standard form of the rule is represented as "IF Conditions THEN Decisions". Generally, the rules are known as expressions and it is written as:

EXP.1: IF ((Attribute-1, Value-1) AND (Attribute-2, Value-2) … AND (Attribute-k, Value-k)) THEN (Decisions, Value)

This expression emphasizes that multiple number of attributes and its values are used to take the decisions and

return its actions. These returned decisions are not only single value or class, it may be binary class or multi class or multi-labeling or numeric values. The decision is a dependent variable since it depends on the features but attributes are independent variable meanwhile one attribute doesn't depend on the other attributes.

The standard dataset was handled in CSV file which contains the records as a row and attributes as a column. The attribute values are either numerical or categorical and we convert the numerical information into categorical form for performing effective classification. The typical rule set depends on all attributes of the dataset so the missing values of the dataset need to be filled based on mean, mode and median value of the data. There are no missing values in PCOS Kaggle data set so we effectively dispense a diet class label by constructing the ruleset.

This diet ruleset is derived based on gynecologist suggestions from Lakshmi Hospital & Fertility Centre, Sivakasi [21]. The diet ruleset induction module contributes two processes in Kaggle dataset:

1. Group the relevant symptoms
2. Diet Label Assignment

A. Symptoms Grouping

This module helps to reduce the above 21 chosen attributes (Table 2) into eight groups of symptoms and its respective rule is written in Algorithm 1. This attribute reduction and symptom grouping is used to write an effective diet ruleset for both PCOS affected people and normal women. Table 3 shows the attributes combination and name of the symptom groups for PCOS dataset.

Algorithm 1: Rule Set for Symptoms Grouping
Input: DS – PCOS Kaggle Dataset, A - Selected Attributes, V - values
Output: SG - Symptom Group Name

```
Function symptomsGrouping(DS,A,V)
BEGIN
        SG = ""
        for each record in DS
          for each A,V
            IF BMI>=32 AND WeightGain==YES THEN SG ← "Obesity"
            IF CycleLength>5 || CycleLength<=2 AND CYCLE==irregular THEN SG ← "Menstrual"
            IF Pregnant==No AND No. of Abortions>0 THEN SG ←  "Infertility"
            IF FSH/LH != [1.3:7] AND TSH!= [0.2:4.7] AND AMH!= [0.7:4] AND
                PRL!=[0:20] AND PRG!=[5:20] THEN SG ← "Hormone Abnormality"
            IF HairGrowth==YES THEN SG ← "Skin Hair Growth
            IF SkinDarkening==YES AND Pimples==YES THEN SG ← "Skin Problems"
            IF Hair Loss==YES THEN SG ← "Baldness"
            IF Follicleno.L&R>10 AND Avg.FollicleSizeL&R>17 AND Endometrium>11 THEN
                SG←"Ovulation Failure"
          end for
        end for
        return SG
END Function
```

Table 3: Symptom Grouping for PCOS Kaggle Dataset

Sl. No.	Attributes Combination	Symptom Group Name
1.	Body Mass Index (BMI), Weight Gain	Obesity
2.	Cycle, Cycle Length	Menstrual Issues
3.	Pregnant, No. of Abortions	Infertility
4.	FSH/LH, TSH, AMH, PRL, PRG	Hormone Abnormality
5.	Hair Growth	Skin Hair Growth
6.	Skin Darkening, Pimples	Skin Problems
7.	Hair Loss	Baldness
8.	Follicle no. In Left Ovary and Right Ovary, Avg. Follicle Size in Left and Right, Endometrium	Ovulation Problem

B. Assign Diet Class Label

According to symptom grouping and gynecologist suggestions, we give a diet food for all 541 patients and its Rule set is written in Algorithm 2. Table 4 shows the symptom name, diet food and its class label for PCOS patients in dataset.

Algorithm 2: Rule Set for Diet Class Label Assignment
Input: DS – PCOS Dataset, SG – Symptom Group
Output: DF – Diet Food, DL - Diet Label
Function diet Assignment(DS,SG)
BEGIN
DF = "", DL=0
for each record in DS
for each SG
IF SG == Obesity THEN DF ← "SeaFood", DL←1
IF SG == Menstrual THEN DF ← " Cinnamon ", DL←2
IF SG == Infertility THEN DF ← "Avacados", DL←3
IF SG == Hormone Abnormality THEN DF ← "Pumpkin Seeds", DL←4
IF SG == Skin Hair Growth THEN DF ← " Garlic Juice", DL←5
IF SG == Skin Problems THEN DF ← "Lemon", DL←6
IF SG == Baldness THEN DF ← "Omega-3s food", DL←7
IF SG == Ovulation Failure THEN DF ← " Spinach", DL←8
end for
end for
return DF, DL
END Function

Table 4: Symptom Grouping, Diet Food and its Class Label

Sl. No.	Symptom Group Name	Diet Food (DF)	Diet Label (DL)
1.	Obesity	Sea Food	1
2.	Menstrual Issues	Cinnamon	2
3.	Infertility	Avocados	3
4.	Hormone Abnormality	Pumpkin Seeds	4
5.	Skin Hair Growth	Garlic Juice	5
6.	Skin Problems	Lemon	6
7.	Baldness	Omega-3s food	7
8.	Ovulation Problem	Spinach	8

Applying Algorithm 1 and Algorithm 2 for all instances in PCOS Kaggle dataset and get the new multi-label target class as a diet food. Table 5 shows some instances from the dataset after applying symptom grouping and Diet assignment.

Table 5: PCOS Kaggle Dataset with New Diet Label

Patient No.	Symptoms	Existing PCOS Labels	New Diet Labels
1	Hormone, Ovulation Problem	0 (No)	4, 8
2	Ovulation Problem	0 (No)	8
3	Hair Growth, Skin Problems	0 (No)	5, 6
4	Hormone, Baldness, Skin Problems, Ovulation	1 (Yes)	4,7,6,8
5	Hormone, Obesity, Skin Problem,	1 (Yes)	4,1,6
6	Obesity, Menstrual Issues, Hormone, Skin Hair Growth, Baldness, Skin Problem, Ovulation, Infertility	1 (Yes)	1,2,4,5,7,6,8,3

2. Model Evaluation

The machine learning classifiers algorithms are used to evaluate binary class, multi-class and multi-label classifiers and compare the performance of the model. The target label of modified dataset consists of multiple diet labels based on symptoms grouping. In this work, rule induction based new diet class labels are evaluated by using various multi-label machine learning classifiers like KNN, Decision Tree, Random Forest classifier and Multi-Layer Perceptron. Each classification algorithm performs two operations like classifies the symptoms as per symptom grouping and then recommends the diet. Finally, the classifier generates

suggested diet labels and compares them to multi-label diet foods that were produced using rule-based induction.

3. Performance Metrics

This recommendation system employs multi-label classification, hence the performance metrics Ranking Loss, Hamming Loss, Label Ranking Average Precision (LRAP), Micro_F1 and Macro_F1 are utilized to assess classification issues. These measures are calculated by constructing the Confusion Matrix.

Confusion Matrix

Confusion matrix is a table-based approach with two dimensions such as Actual classes in column and Predicted classes in rows. Here, actual class as Rule Set Induction diet labels and predictive class as Machine Learning classifier diet labels which was taken from model evaluation. Both classes have Positives (P) and Negatives (N) and it has four measures like:

a. True Positives (TP) – Both Rule set (Actual) diet labels and Machine Learning Classifier (Predicted) diet labels are one.
b. True Negatives (TN) − Both Rule set (Actual) diet labels and Machine Learning Classifier (Predicted) diet labels are zero.
c. False Positives (FP) − Rule set (Actual) diet labels are zero but Machine Learning Classifier (Predicted) diet labels are one.
d. False Negatives (FN) − Rule set (Actual) diet labels are one but Machine Learning Classifier (Predicted) diet labels are zero.

Ranking Loss

Ranking Loss (RL) means the number of inaccurate labels divided by the number of correct labels and zero is the optimal value for ranking loss. Consider the matrix with ground truth labels are given in Equation 1.

$$x \in \{0,1\}^{n_{instances} * n_{labels}} \quad (1)$$

The score for each label is represented by $\hat{g}$ and it is written in Equation 2.

$$\hat{g} \in \{R\}^{n_{instances} * n_{labels}} \quad (2)$$

Ranking loss is calculated in Equation 3:

$$RL(x,\hat{g}) = \frac{1}{n_{instances}} * \sum_{i=0}^{n_{instances}-1} \frac{1}{||x_i||_0 * (n_{labels} - ||x_i||_0)} |\{(k,l): \hat{g}_{ik} \leq \hat{g}_{il}\ ;\ x_{ik} = 1, x_{il} = 0\}| \quad (3)$$

$Where$ $||x_i||_0$ means number of non-zero elements in the set and |.| means the number of elements in the vector.

Label Ranking Average Precision (LRAP)

Instead of using precision and recall, it measures the prediction model's average precision. Here, each example's label rating is calculated, and the result must be greater than one. Consider the matrix with ground truth labels are given in Equation 1 and the score for each label is represented by $\hat{g}$ and it is given in Equation 2.

LRAP is calculated in Equation 4:

$$LRAP(x,\hat{g}) = \frac{1}{n_{instances}} * \sum_{i=0}^{n_{instances}-1} \frac{1}{||x_i||_0} \sum_{j:x_{ij}=1} \frac{|L_{ij}|}{rank_{ij}} \quad (4)$$

Where,

$$L_{ij} = \{k : x_{ik} = 1, \hat{g}_{ik} \geq \hat{g}_{ij}\} \quad (5)$$

$$rank_{ij} = |\{k : \hat{g}_{ik} \geq \hat{g}_{ij}\}| \quad (6)$$

Hamming Loss

Hamming Loss (HL) is the division of wrong labels with total number of labels. The hamming loss penalizes only the individual labels in multi-label classification. The prediction error and missing error are considered along with the total number of classes and examples in the Hamming loss.

$$HL = \frac{1}{nL} \sum_{i=1}^{n} \sum_{j=1}^{L} I\ (x_i^j \neq \hat{x}_i^j) \quad (7)$$

Where n is the total number of instances, L is the total number of labels, x_i^j is the actual label, $\hat{x}_i^j$ is the predicted label and I is the indicator function. Practically, the learning algorithms perform better when the hamming loss is reduced.

Micro_F1

It is one of the global averages F1 score and it is calculated by counting the sum of True Positives, False Positives and False Negatives.

$$Micro\_F1 = \frac{TP}{TP + \frac{1}{2}(FP+FN)} \quad (8)$$

Micro-averaging basically calculates the percentage of correctly classified observations among all observations.

Micro_F1 gives more weight to the common labels and this value can be high even if the model is carrying out very poorly on an erratic label.

Macro_F1

Macro_F1 score is calculated by taking arithmetic mean of all the per-class F1 scores and it is used to evaluate the quality of multiple binary labels. Each label is given equal weight, and the result is 1 for the best performance and 0 for the worst performance.

$$Macro\_F1 = \frac{1}{N} \sum_{i=0}^{N} F1 - Score_i \quad (9)$$

Where i is the label index and N is the number of labels. Macro averaging is more preferred in case of imbalanced labels since it gives equal weights to each label and doesn't encourage the number of samples of each label.

Experimental Result

In this section, we discuss the details of experimental evaluation and its various metrics for our Kaggle PCOS

dataset. This dataset contains 541 instances and comprises PCOS's pre-existing binary class labels (1 indicating PCOS and 0 indicating No PCOS), and then we developed a new multi-label diet based on the patients' symptoms. It is done by symptom grouping and diet label assignment steps using rule set induction.

In symptom grouping, we combine the twenty-one attributes into eight symptom groups as per Algorithm 1 and Table 3. From this result, we first display how much patients affected for single or multiple symptoms or no symptoms. Figure 2 describes the number of patients affected for number of symptoms group in Kaggle dataset.

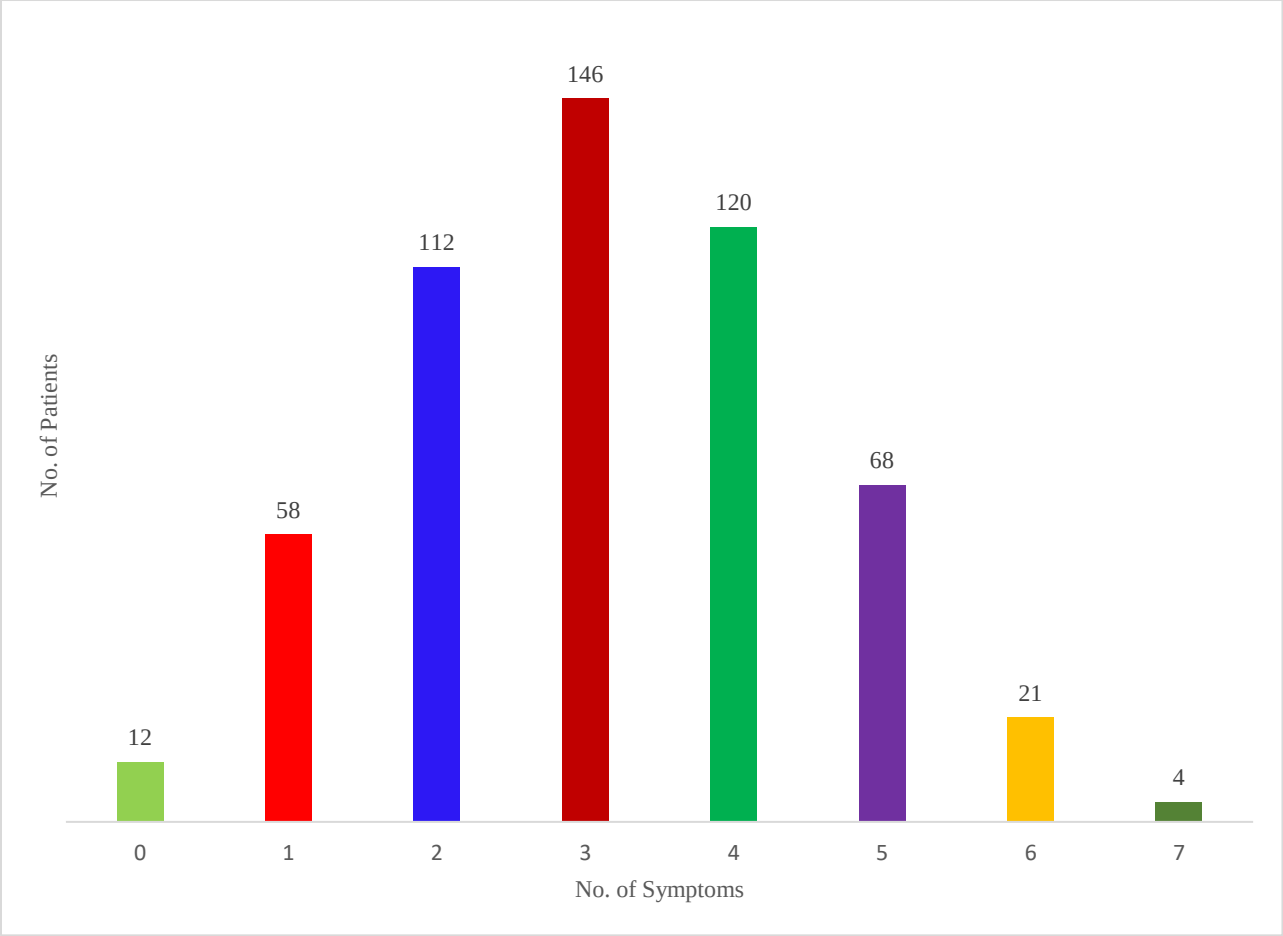

Figure 2. No. of Symptoms Vs No. of Patients

Figure 2 suggests among 541 patients, 12 patients have no symptoms, nearly 170 patients affected less than three symptoms, 146 patients mostly affected three symptoms, nearly 210 patients have affected more than three symptoms and 4 patients have highly affected all symptoms.

According to this symptom grouping next we assign the multiple diet labels using rule set induction and it has been evaluated by using various machine learning algorithms. We evaluate the rule sets for each symptom group by constructing the confusion matrix with actual labels and predicted labels. The following figures describe the confusion matrix for KNN, DT, RFC and MLP.

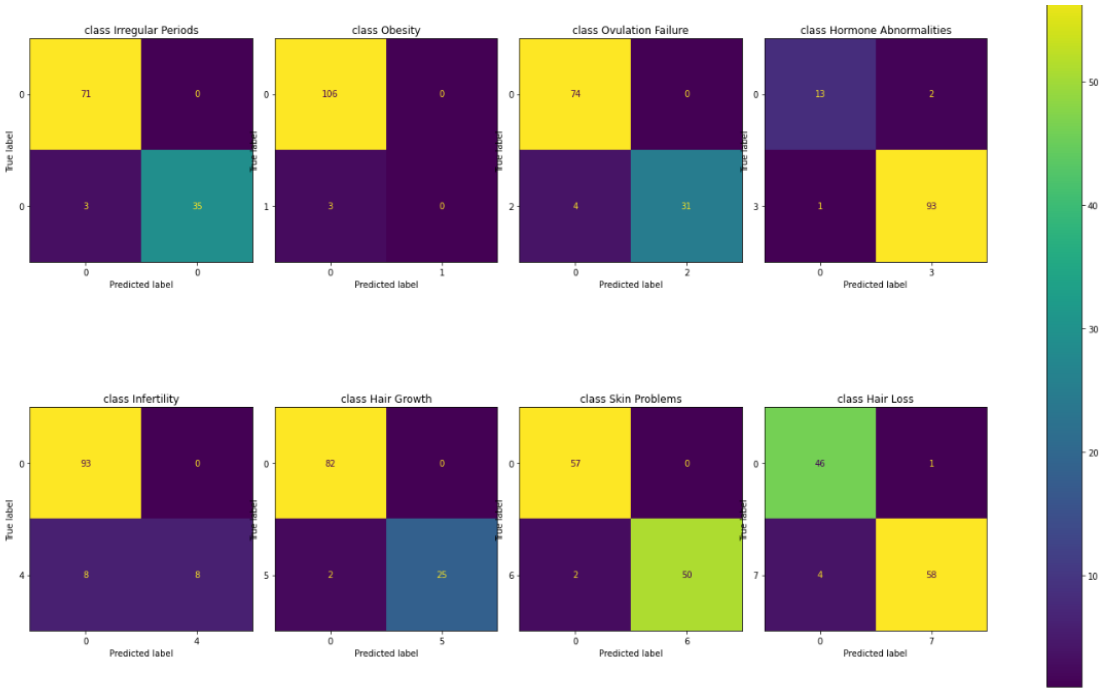

Figure 3(a): KNN Confusion Matrix

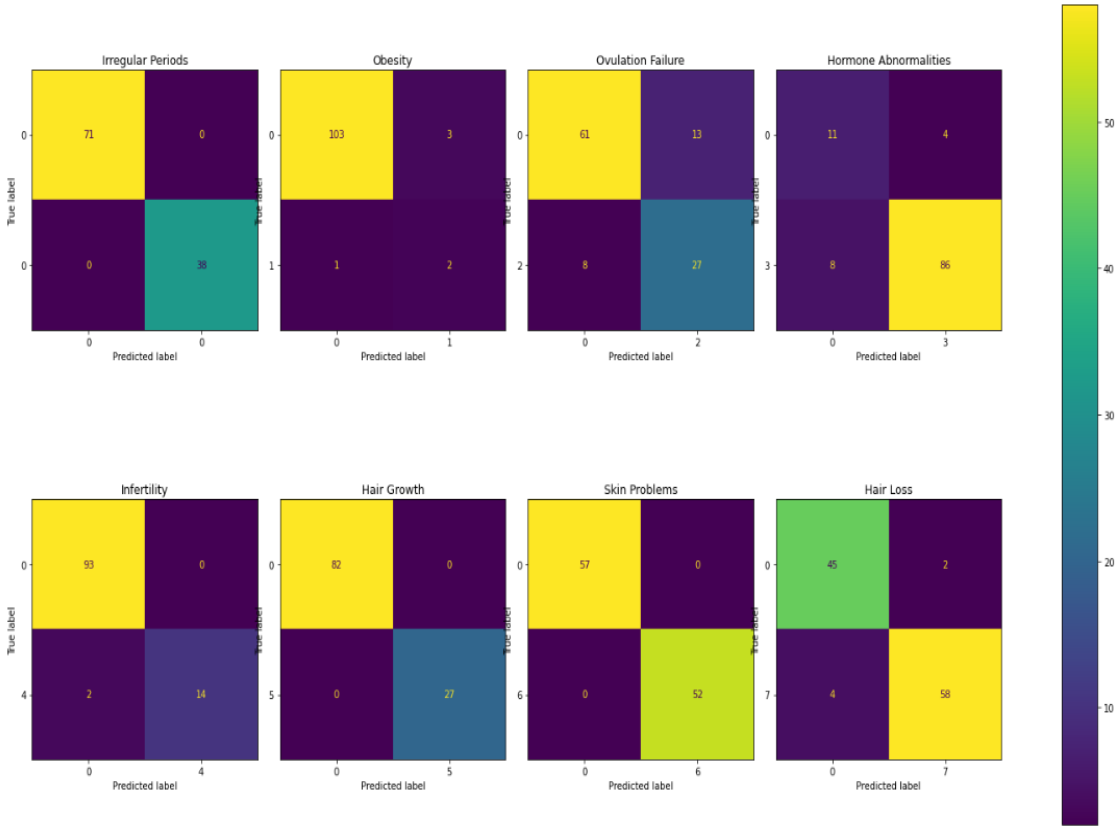

Figure 3(b): DT Confusion Matrix

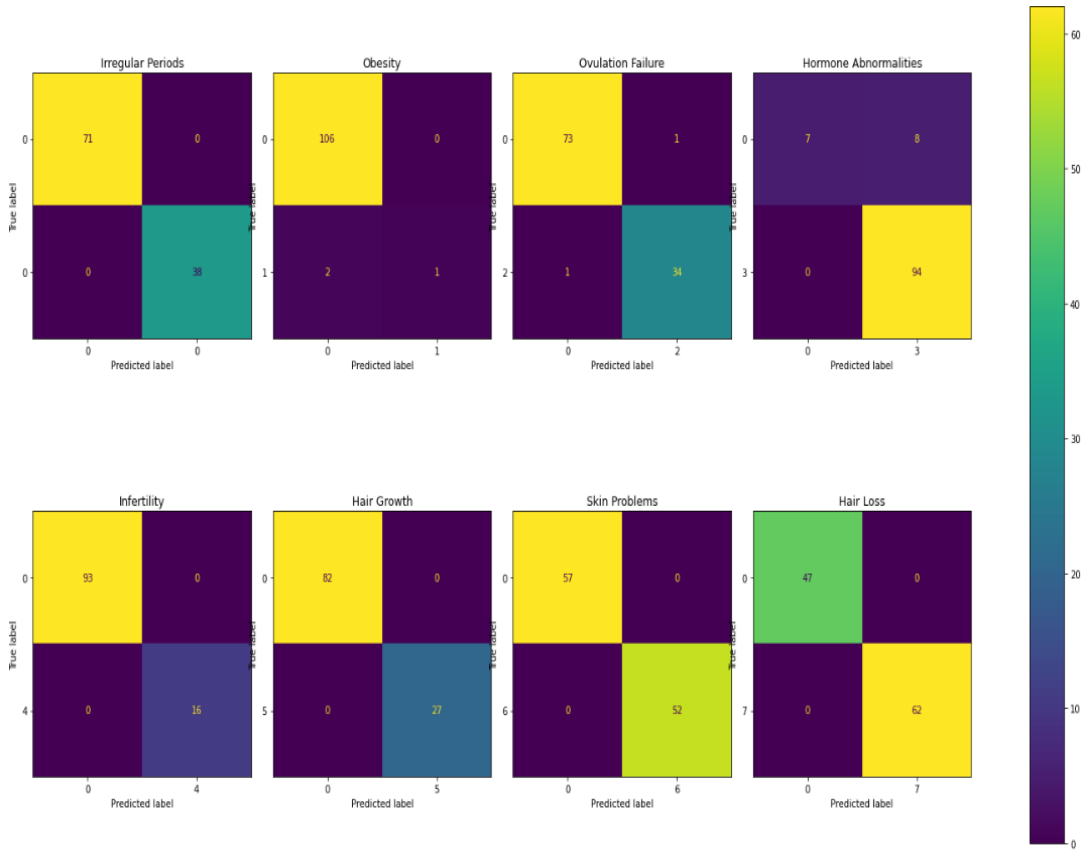

Figure 3(c): RF Confusion Matrix

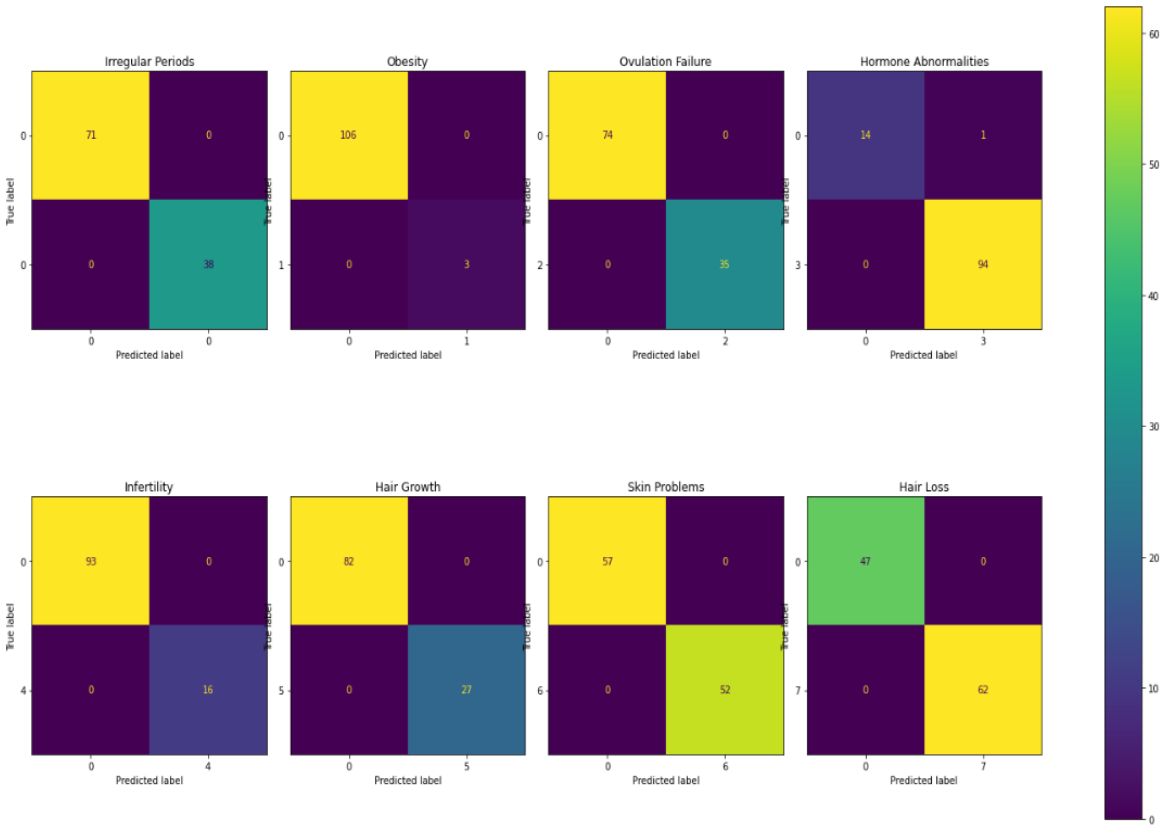

Figure 3(d): MLP Confusion Matrix

From highest to lowest True Positive in the dataset, these statistics suggested that the following order of symptom groups were present: obesity, infertility, hair growth, ovulation failure, irregular periods, skin problems, hair loss, and hormonal irregularities. The symptom groups from highest to lowest True Negative rate happened in opposite order at the same time. False Positive and False Negative rates are very low or nonexistent. The confusion matrix comes to the conclusion that the rule set induction model provides the best multi-diet labels and it should be properly assessed using various multi-label classifier models. After construction of confusion matrix some other important measures also assessed such as Ranking and Hamming Losses, Average Precision and F1 Score. These measures for different algorithms are given in Figure 4.

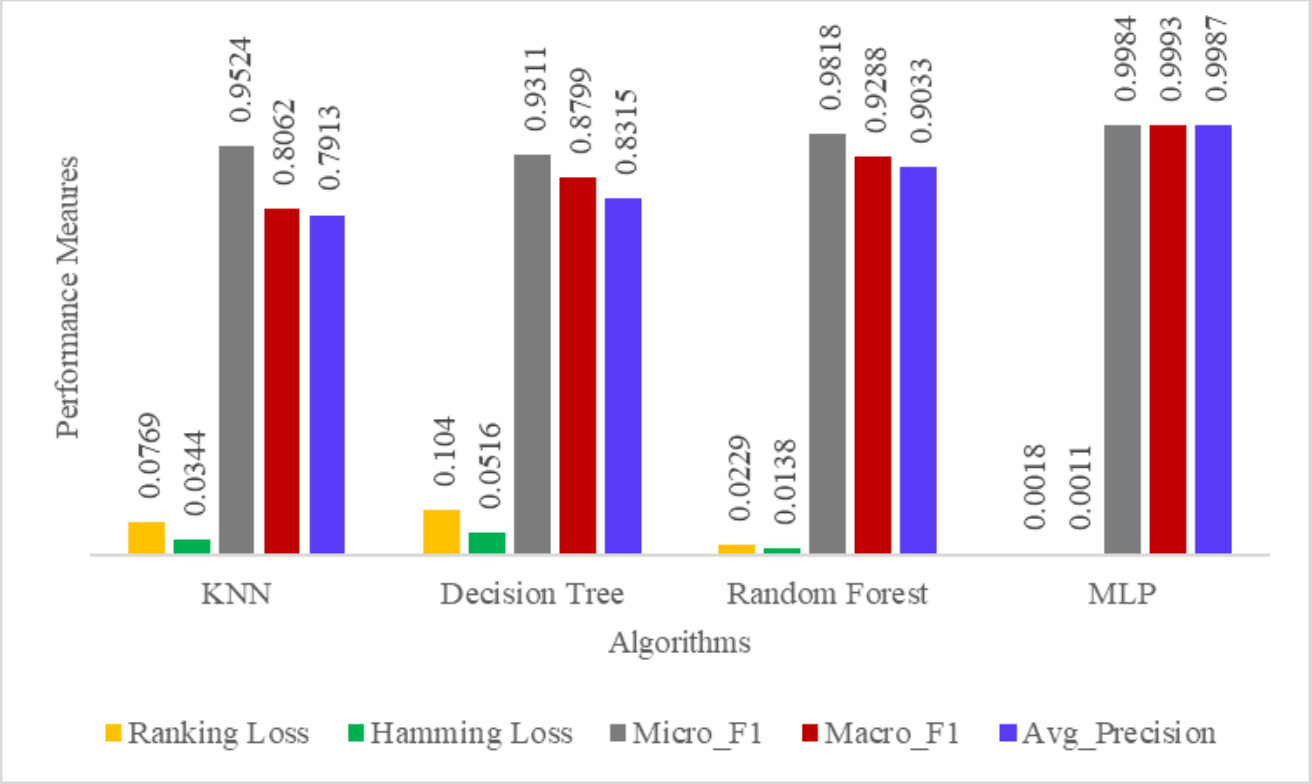

Figure 4: ML Classifiers and its Performance

From this analysis, we decided the following results:

a. Ranking Loss: DT > KNN > RF > MLP
b. Hamming Loss: DT > KNN > RF > MLP
c. Micro_F1: MLP > RF> KNN > DT
d. Macro_F1: MLP > RF > DT > KNN
e. Avg_Precision: MLP > RF > DT > KNN

Random Forest algorithm outperform the decision tree algorithm because random forest is combination of multiple decision tree. KNN performs better than Decision tree based on ranking loss, hamming loss and average precision but micro_f1 and macro_f1, decision tree performs well. Based on this result, it is concluded that both algorithms produce average performance for multi label classification. MLP produces maximum value for micro f1, macro f1 and average precision and also minimum value for ranking loss and hamming loss. So, Multilayer Perceptron has given the best result among the four chosen algorithm.

Conclusion

This proposed system analyzed PCOS disease in the Kaggle dataset based on symptoms like hormonal abnormality, obesity, irregular periods, infertility, skin hair growth, hair loss, skin problems, and ovulation problem. This system is used to assign the new multi diet labels according to these symptoms and existing PCOS class labels. The F1-Score, precision, and losses are used to evaluate the different multi-attribute with multi-label classification system methods. The designed diet rules are most appropriate for PCOS disease where the F1-score and precision are raised and losses are lowered. And also, this proposed work reduces the severity of the disease by analyzing the disease with symptoms in advance and also it overcomes the effects by taking a recommended diet. In future, the system can be extended to collect the data from PCOS patients post on social media and the diet will be recommended for those patients.

REFERENCES

PCOS AIIMS Study (2017) Why PCOS is on the rise among Indian women. https://www.dailyo.in/variety/polycystic-ovarysyndrome-womens-health/story/1/16785.html. Accessed 21 April 2017.

Jeshica Bulsara, Priyanshi Patel, Arun Soni, Sanjeev Acharya (2021) A review: Brief insight Into Polycystic Ovarian Syndrome. Journal of Endocrine and Metabolic Science 3.

PCOS The Hindu Mumbai Study (2019) One in five Indian women suffers from PCOS. https://www.thehindu.com/sci-tech/health/one-in-five-indian-women-suffers-from-pcos/article29513588.ece Accessed 26 September 2019.

Vikas B, Sipra Sarangi, Manaswini Chilla, K. Santosh Bhargav, BS Anuhya (2017) A Literature Review On The Rising Phenomenon PCOS. International Journal of Advances in Engineering and Technology. 10: 216-224.

The Telegraph Study (2022) PCOS: A lifestyle disorder. https://www.telegraphindia.com/health/pcos-a-lifestyle-disorder/cid/1677239 (2022) Accessed 15 April 2022.

St. Clair Health: Polycystic ovary syndrome (PCOS) (2020) https://www.stclair.org/services/mayo-clinic-health-information/diseases-and-conditions/CON-20314571/ Accessed 3 October 2020.

Emids (2019) A Practical Application of Machine Learning in Medicine. https://www.macadamian.com/learn/a-practical-application-of-machine-learning-in-medicine/ Accessed 1 February, 2019.

Palvi Soni and Sheveta Vashisht (2018) Exploration on Polycystic Ovarian Syndrome and Data Mining Techniques. Proceedings of the International Conference on Communication and Electronics Systems.

Amsy Denny, Anita Raj, Ashi Ashok, C Maneesh Ram, Remya George (2019) i-HOPE: Detection and Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques. TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 673-678.

Shreyas Vedpathak, & Vaidehi Sunil Thakre (2020) Pcocare: PCOS Detection and Prediction Using Machine Learning Algorithms. Bioscience Biotechnology Research Communications, 240-244.

Neetha Thomas and Dr.A. Kavitha (2020) Prediction of Polycystic Ovarian Syndrome with Clinical Dataset Using a Novel Hybrid Data Mining Classification Technique. International Journal of Advanced Research in Engineering and Technology, 11: 1872-1881.

S. Bharati, P. Podder, and M. R. Hossain Mondal (2020) Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms. IEEE Region 10 Symposium (TENSYMP). 1486-1489.

Malik Mubasher Hassan & Tabasum Mirza (2020) Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. International Journal of Computer Applications, 175: 42-53.

Preeti Chauhan, Pooja Patil, Neha Rane, Pooja Raundale, Harshil Kanakia (2021) Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS. International Conference on Communication information and Computing Technology (ICCICT), 1-7.

Bhat, Shakoor Ahmad (2021) Detection of Polycystic Ovary Syndrome Using Machine Learning Algorithms. Masters Thesis, Dublin. National College of Ireland.

Homay Danaei Mehr, Huseyin Polat (2022) Diagnosis of polycystic ovary syndrome through different machine learning and feature selection technique. Health and Technology, 1.

Pijush Dutta, Shobhandeb Paul, Madhurima Majumder (2021) An Efficient SMOTE Based Machine Learning classification for Prediction and Detection of PCOS. Research Square.

Wanyun Cui, Xingran Chen (2021) Open Rule Induction. 35th Conference on Neural Information Processing System.

Dhatri Ganda, Rachana Buch (2018) A Survey on Multi Label Classification. STM Journals. 5: 19-23.

Eyke H¨ullermeier, Johannes F¨urnkranz, Eneldo Loza Mencia, Vu-Linh Nguyen, and Michael Rapp (2020) Rule-based Multi-label Classification: Challenges and Opportunities. Rules and Reasoning.

Lakshmi Hospital (2021) [online] Available at: https://lakshmifertilitycentre.com/hormone-analysis/#/