Applied Mathematics and Computational Intelligence

In Press, Accepted Manuscript – Note to users

POLYCYSTIC OVARIAN SYNDROME (PCOS) CLASSIFICATION AND FEATURE SELECTION BY MACHINE LEARNING TECHNIQUES

[1]Satish C.R Nandipati, [2]Chew XinYing*, [3]Khaw Khai Wah

[1,2]School of Computer Sciences, 11800, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[3]School of Management, 11800, Universiti Sains Malaysia, Pulau Pinang, Malaysia

ABSTRACT

The one of the most common endocrine system disorders which affects about 5 to 10 % of the adolescent women is Polycystic ovarian syndrome (PCOS). The symptoms include failure to ovulate and infertility, cardiovascular diseases, type 2 diabetes, etc. The detection of PCOS can be done through biochemical, clinical and ultrasonography methods. It is known that early diagnosis and treatment could reduce the chance of PCOS. Hence, it is necessary to know which classification model and selected feature play a role in the prediction of disease, which is the objective of this study with Python-Scikit learn package and RapidMiner. Despite different tools used, the highest accuracy is shown by random forest (93.12%, RapidMiner) with complete dataset. On the other hand, KNN and SVM show similar accuracy performances (90.83%, RapidMiner) with 10 selected features. The average performances of 10 and 24 selected features show insignificance and significance with the combined dataset, indicating these features could be used and cannot be used for the prediction of PCOS respectively. A comparison of both tools and their performances shows that the RapidMiner performs better than Python. However, it depends on the performance of the classification model which in turn dependent on the nature of the dataset and techniques used.

Keywords: Classification, Feature Selection, PCOS, Python-Scikit learn package, RapidMiner

1. INTRODUCTION

Polycystic ovarian syndrome (PCOS) being one of the most common endocrine system disorders which affects about 5 to 10 % of the women [1]. PCOS manifests during adolescence and is formed as a result of hormonal disturbances. Peripherally inside the ovary, fluid-filled sacs are present which are called Follicles or cysts. A polycystic ovary (PCO) can be characterized by twelve or more follicles with a diameter of 2-9 mm [2]. PCOS affects both health and the quality of women's life. The symptoms include cardiovascular diseases, failure to ovulate and infertility, late menopause, type 2 diabetes, acne, baldness, hair loss, hirsutism, obesity, anxiety, depression, and stress. Globally, the prevalence is said to be in a range of 2.2-26% [3]. Based on the community study in the United Kingdom (UK), it is found that the South Asian population shows a prevalence of 52% when compared with a Caucasian population (22%) [2]. The early diagnosis and treatment can be used to control based on the symptoms and by the prevention of long term problems. PCOS can be detected through ultrasonography by a doctor by counting the number 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 PCOS [4]. Another approach for 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 PCOS [3].

*E-mail of corresponding author: xinying@usm.my

In recent years, machine learning (ML) classification and feature selection algorithms have been used by researchers and clinicians for the prediction of diseases as a non-invasive method [5],[6]. PCOS datasets which consist of heterogeneous attributes related to biochemical, clinical, medical history, symptoms of the patients and ultrasound images are used to build predictive models [2]. To the best of the author's knowledge, the seven classification algorithms and five feature selection methods for PCOS dataset, and performance comparison of Python and RapidMiner tools have not been documented [1], [7], [8]. In this paper, our objective is to know the better performance machine learning classification algorithm, to know which features play a role in the prediction of PCOS disease, and to compare the performances of Python and RapidMiner tools by using the PCOS dataset.

2. LITERATURE REVIEW

The various machine learning classification methods have been used for the detection of various diseases such as breast cancer, heart and ovarian, etc. [9]. Some of the classification algorithms used for the prediction of PCOS dataset are reviewed below:

The dataset which consists of 541 women through the survey from doctor consultations and clinical examinations were used in this study. The SPSS V 22.0 is used to extract 8 features from a total of 23 features which consists of both clinical and metabolic parameters. The Spyder Python IDE was used to evaluate the PCOS dataset using six classification algorithms such as Classification and Regression Trees (CART), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), logistic regression, Random Forest Classifier and Naïve Bayes. The highest accuracy (89.02%) is shown by random forest was addressed by Denny et al., [7]. Similarly, another survey study that is based on their lifestyle and food intake habits, with 119 women between the ages of 18-22 and 18 attributes were used to evaluate the PCOS. The algorithms used were Artificial Neural Network backpropagation, Bayesian Network, and C5.0 Decision Tree using IBM SPSS Modeler 16.0 tool. The highest accuracy is shown by the Naïve Bayes algorithm (97.65%), followed by the ANN backpropagation algorithm (96.27%) and 96.24% accuracy of the C5.0 decision tree algorithm was addressed by Vikas et al., [1].

The study conducted by Anuradha et al., [8] with the dataset of 84 instances and 13 attributes was used to detect PCOS by three machine learning classification methods such as Artificial Neural Network, K- nearest Neighbor and Linear Regression using Python. The most important symptoms as shown by this study are acne, irregular periods, LH, sonography, and weight. The highest accuracy is shown by linear regression (100%), followed by ANN with 94%.

The study by Meena et al., [6] presented the dataset of 31 attributes. The feature selection methods used in this study are IGSE (Information Gain Subset Evaluation Technique using Ranker Search Method) and NFRSE (Neural Fuzzy Rough Set Using Genetic Search Algorithm). A total of 17 and 7 selected features were extracted using the above feature selection methods respectively. The classification techniques used are decision tree (ID3 and J48 algorithm). The root mean squared error of NFRSE ID3 is less than the ID3-IGSE, indicating NFRSE showed the best performance.

The study by Deshpande and Wakankar [10] presented with the biochemical and clinical parameters such as hormonal levels, body mass index (BMI), menstrual cycle length, along with imaging parameters such as several follicles were used to study the PCOS. The classification algorithm used in this study is the Support Vector Machine algorithm (SVM). The results showed an accuracy of 95%.

The study by Vijayalakshmi and UmaMaheswari [11] presented a dataset consisting of 575 patients with 154 fertile women and 421 infertile women were used to study infertility in women. WEKA was used for applying various data mining techniques like Classification (J48 and random forest) and Subset evaluation, Associative rule mining, Clustering Statistical analysis. The J48 pruned tree showed an accuracy of 96%. A total of 7 selected features by Subset evaluation are Age, BMI, Diabetic, FSH, LH, TB and TSH.

3. DATA SOURCES AND ATTRIBUTES DESCRIPTION

The PCOS dataset used in this study is retrieved from Kaggle [12]. Similar to PCOS dataset, from the previous studies it is clear that researchers and clinicians are using different disease datasets to study machine learning classification methods [9]. The PCOS original dataset consists of 541 instances with 42 attributes in which one attribute as patient file number (not taken into consideration for data analysis). Finally, the total number of 41 attributes includes 40 as input attributes, and PCOS as a class label [Positive (Yes) and Negative (No)]. The dataset shows an imbalance nature of class labels (i.e., 364 instances of class label = 0 and 177 instances of class label =1) and missing values. The attributes are categorized as continuous, nominal and ordinal. Table 1 shows the attributes description.

Table 1. Attributes description and their units

No	Attributes	No	Attributes
1	Patient File number	22	Thyroid-Stimulating Hormone : TSH (mIU/L)
2	PCOS (class label)	23	Anti-Müllerian Hormone : AMH(ng/mL)
3	Age (Yrs)	24	Prolactin: PRL(ng/mL)
4	Weight (Kg)	25	Vit D3 (ng/mL)
5	Height(Cm)	26	Progesterone : PRG(ng/mL)
6	BMI : body mass index	27	BP _Systolic (mmHg)
7	Blood Group	28	Random Blood Sugar : RBS(mg/dl)
8	Pulse rate(bpm)	29	Weight gain(Y/N)
9	RR (breaths/min)	30	hair growth(Y/N)
10	Hemoglobin : Hb(g/dl)	31	Skin darkening (Y/N)
11	Menstrual Cycle : Cycle(R/I)	32	Hair loss(Y/N)
12	Cycle length(days)	33	Pimples(Y/N)
13	Marriage Status (Yrs)	34	Fast food (Y/N)
14	Pregnant(Y/N)	35	Reg.Exercise(Y/N)
15	No. of aborptions	36	BP _Systolic (mmHg)
16	Follicle stimulating hormone: FSH(mIU/mL)	37	BP _Diastolic (mmHg)
17	LH(mIU/mL)	38	Follicle No. (R)
18	FSH/LH	39	Follicle No. (L)
19	Hip(inch)	40	Avg. F size (L) (mm)
20	Waist(inch)	41	Avg. F size (R) (mm)

21	Waist: Hip Ratio	42	Endometrium (mm)

4. METHODOLOGY

4.1 Data Preprocessing

The mode value is used to replace the missing values present in the dataset. Later, the Synthetic Minority Oversampling Technique (SMOTE, k = 5, default parameter) with the oversampling method was used to make the balance nature of the dataset (i.e. 728 from 541 instances). Since the dataset consists of different measuring units, the variable values are rescaled with the data normalization method. The normalization methods used in two machine learning tools are Python [preprocessing.normalize(X)] and RapidMiner [normalize operator].

4.2 Data Analysis

The dataset which consists of 728 instances with 41 attributes is considered for a model building. The default values parameter settings available in two analytical tools Python-based open-source Scikit learn version 0.21 and RapidMiner studio version 9.5 was used for this study. Before classification model building, the pre-processing steps (normalization, SMOTE oversampling and data split 70–30% as a training and testing data) respectively has been carried out in both machine learning tools. A total of seven machine learning (ML) techniques were used to evaluate the performance of the classifiers, followed by five feature selection methods from Python (Spyder as IDE) - Scikit learn package and RapidMiner was performed on aforementioned datasets respectively.

In Python (Spyder as IDE), the classification performance is carried out at set seed=123. The five classification models used are KNeighborsClassifier (KNN), SVC (for SVM), Random Forest (RF), Gaussian Naive Bayes (for NB), and multilayer perceptron Classifier (for NN, solver='sgd', hidden_layer_sizes= (10, 10), activation='relu'). The two ensemble classifiers used are Bagging Classifier and GBOOST (GradientBoostingClassifier, for Boosting). The feature selections methods (FS) such as correlation matrix (library “pandas” and “seaborn”), recursive feature elimination method with Logistic Regression model (RFE-LR), Rank Feature Importance method with Extra Classifier Tree model (RFI-ECT) and SelectKBest (score_func=chi2, k=5) were used [13].

In RapidMiner, the five machine learning operators such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB) and Auto Multilayer Perceptron (for Neural network), and two ensemble classifiers such as Bagging (method = decision tree) and Adaboost (method = gradient boosting trees) were evaluated on the training and testing data. The feature selection methods included in this study are correlation matrix, forward selection (method = Naïve Bayes) and backward elimination (method = decision tree) with cross-validation operator [14].

The performance of the model on test data is calculated by accuracy, macro average precision, and recall. The overall average scores of all classifier's accuracy, precision, and recall are used for the comparison of Python and RapidMiner performances. The Figure 1 shows the flow diagram of data preprocessing and analysis (the overall work process).

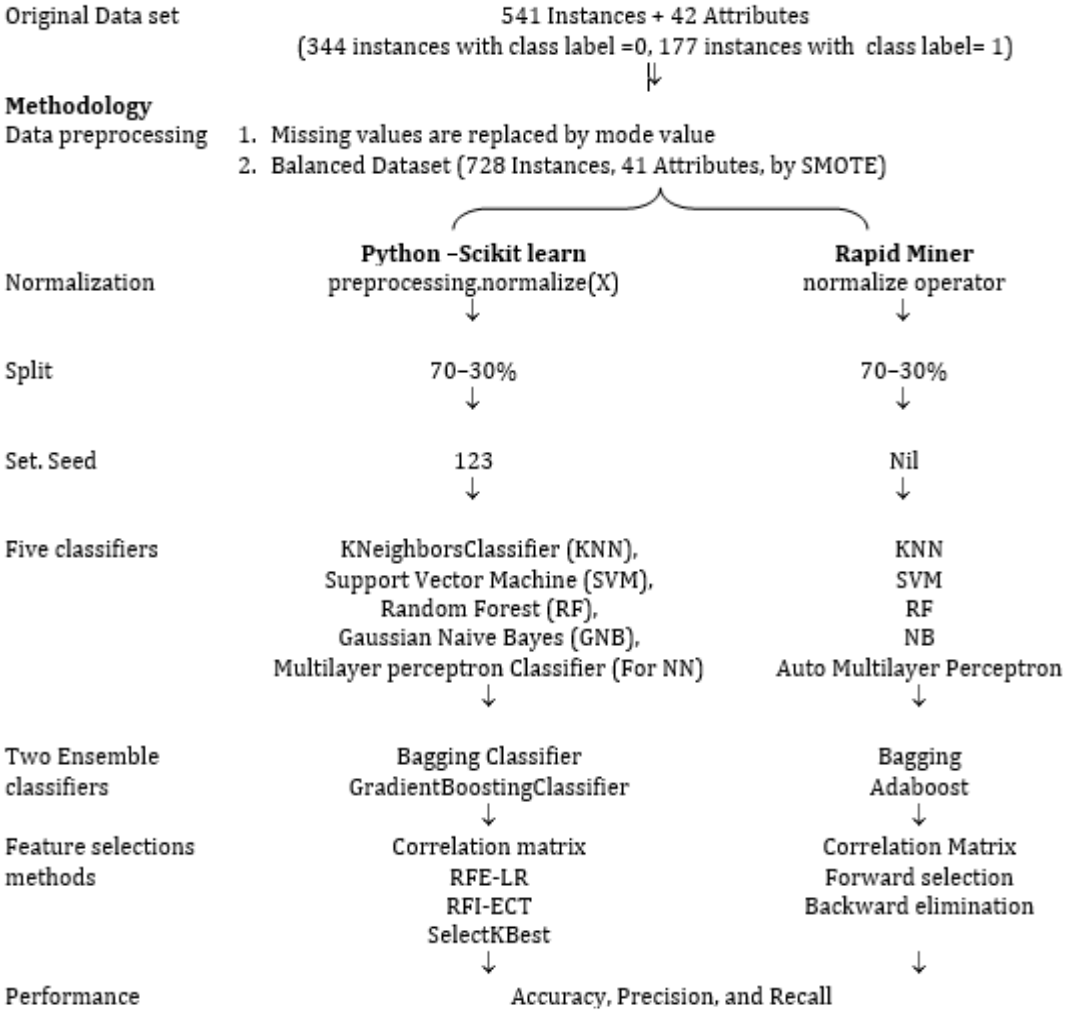

Figure 1. Flow diagram of data preprocessing and analysis

5. RESULTS

5.1 Performance Comparison of Python and RapidMiner ML Techniques on Complete Dataset

To the best of author knowledge, fewer machine learning techniques are used to study the classification model as mentioned in the literature review. Thus, no studies were addressed on this oversampled dataset with the seven machine learning techniques. To understand and compare which classification model and tool have a better performance on the dataset both Python and RapidMiner tools are selected. The performance of each classification algorithm is measured based on accuracy.

In Python-ML techniques, the naïve bayes and adaboost show a similar performance of the classification algorithms with highest accuracy (87.72%), followed by random forest (82.27%). Whereas, in RapidMiner the similar performance of the classification algorithms for highest accuracy, precision and recall are shown by random forest (93.12%), followed by naïve bayes (90.83%) (Table 2). In comparison to both Python and RapidMiner, the 'RapidMiner' shows the highest average accuracies (81.32%) and the same goes for precision (88.46%) and recall (81.55%) respectively. Thus, better performance is shown by the Rapid Miner tool than Python (refer to Table 2).

Table 2 The Performance Comparisons of Python and RapidMiner ML-Techniques with 40 features (attributes).

	Python			**RapidMiner**		
Algorithms	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KNN	75.45	75	75	88.07	88.07	88.12
SVM	82.27	83	82	90.37	90.36	90.45
Random Forest	85	87	85	93.12	93.12	93.12
Naïve Bayes	87.72	88	88	90.83	90.83	90.83
Neural Network	50	25	50	50	100	50
Bagging	80.90	81	81	84.86	84.86	84.93
Adaboost	87.72	89	87	72.02	72.01	73.46
Average	78.43	75.42	78.28	81.32	88.46	81.55

5.2 Feature Selection

Several studies show the subset of relevant features can be useful for better model building. Thus in this study, correlation analysis has been performed to know the highly correlated attributes. The correlation score of 0.59 and above is taken as a positive strong correlation [15]. The scores of the highly correlated features are found to be similar both in Python and RapidMiner (Table 3).

Table 3 Correlated features in Python and RapidMiner, selected and removed attributes based on Python feature selected methods

Python	**RapidMiner**	**Selected Attributes (7)**
Age/ Marriage Status = 0.7	Age/ Marriage Status = 0.66	Marriage Status
BMI/ Hip = 0.6	BMI/ Hip = 0.59	BMI
BMI/Waist = 0.6	BMI/Waist = 0.60	Follicle R
Follicle L/ Follicle R = 0.8	Follicle L/ Follicle R = 0.80	PCOS
Follicle L/PCOS = 0.6	Follicle L/ PCOS = 0.60	FSH
Follicle R/PCOS = 0.6	Follicle R/PCOS = 0.64	Hip
FSH&LH/ FSH = 1.0	FSH&LH/ FSH = 0.97	Weight
Hip/Waist = 0.9	Hip/Waist = 0.87	**Removed Attributes (4)**

Weight / BMI = 0.9	Weight / BMI = 0.90	Age
Weight / Waist = 0.6	Weight / Waist = 0.64	Follicle L
Weight/ Hip = 0.6	Weight/ Hip = 0.63	FSH&LH
		Waist

Since the strongly correlated features show an effect on model performance. Thus, the relevant features among a set of strongly correlated features are selected based on the ranking orders by the Python feature selection methods (refer to data analysis, Table 4). Thus, the 6 correlated attributes are selected (Marriage Status, BMI, Follicle R, PCOS, FSH, Hip, and Weight), and 4 attributes are removed (Age, Follicle L, FSH&LH, Waist) respectively (Table 3).

Table 4 Different types of Feature selections (FS) methods and their selected features from Python and RapidMiner

FS methods	Selected features
RFE-LR	hair growth, PCOS, Skin darkening, Weight gain, Fast food, Pimples, Pregnant, FollicleR, Hair loss, aborptions, PRG, Reg.Exercise, Cycle length, AvgFSizeL, BMI, Weight, Height, Hip, Marriage Status, Cycle, Pulse rate, RR, Hb, TSH, Blood group, BPSystolic, Age, Follicle L, AvgFsizeR, Endometrium, FSH/LH, FSH, AMH, RBS, LH, WaistHip Ratio, BPDiastolic, PRL, Waist, Vit D3
RFI-ECT	0.546 PCOS, 0.065 FollicleR, 0.056 Skin darkening, 0.049 hair growth, 0.047 FollicleL, 0.045 Weight gain, 0.029 Cycle, 0.028 Fast food, 0.015 Pimples, 0.007 Cycle length, 0.006 Hip, 0.005AMH, 0.005 AvgFsizeR, 0.005 Marriage status, 0.005 Reg.Exercise, 0.005 Age, 0.005 LH, 0.005 Waist Hip ratio,0.005 Weight, 0.005 BMI, 0.004 AvgFSizeL, 0.004 Hair loss, 0.004 FSH, 0.004 Height, 0.004 FSH/LH, 0.003 Waist, 0.003 PRL, 0.003 Endometrium, 0.003 Hb, 0.003 PRG, 0.003 TSH, 0.003 Vit D3, 0.003 RR, 0.003 RBS, 0.003 BPSystolic, 0.003 Pulse Rate, 0.002 BPDiastolic, 0.002 Pregnant, 0.002 abortions, 0.002 Blood group
SelectKBest/Chi2	9477.65 Vit D3, 2558.47 LH, 1601.15 FSH, 672.78 FollicleR, 573.65 FollicleL, 364 PCOS, 230.76 AMH, 96.85 FSH/LH, 84.87 Skin darkening, 84.85 hair growth, 65.55 Weight gain, 49.47 Weight, 37.08 Fast food, 27.68 Cycle, 24.64 PRG, 22.59 Pimples , 19.48 Marriage Status, 14.55 BMI, 14.28 Age, 8.85 Hair loss, 8.09 AvgFSizeL, 7.75 Cycle length, 5.89 Hip, 5.59 Waist, 4.46 RBS, 3.67 AvgFsizeR, 3.40 Endometrium, 2.93 abortions, 1.74 Reg.Exercise, 1.22 Pulse rate, 0.59 Height, 0.32 BPDiastolic, 0.28Hb, 0.26 TSH, 0.25 Pregnant, 0.18 Blood Group, 0.13 PRL, 0.11 RR, 0.02 BPSystolic, 0.00 WaistHip Ratio
Forward	PCOS, Follicle R. Fast food, Hair growth, Follicle L, Skin Darkening, BPDiastolic
Backward	PCOS, Age, weight , height, BMI, blood group, pulse rate , RR, HB, cycle, cycle length, marriage status, pregnant, abortions, FSH, LH, FSH/LH, Hip, waist, waist hip ratio, TSH, AMH, PRL,Vit D3, PRG, RBS, Weight gain, hair growth, skin darkening, hair loss, pimples, fast food, reg exercise , bpsystolic, follicle L, follicle R, avgfsize L, Endometrium (expect Avg.F.size (R), BP_Diastolic)

Though the Python FS methods were able to select relevant features from strongly correlation features but did not show similar order of features topology. Finally, a total of 10 selected features is based on correlation analysis and forward selection features. Similarly 24 selected features consist of the features based on the backward elimination. In both cases (i.e. 10 and 24

selected features) repeated features from correlation analysis, forward and backward elimination methods are removed (Table 5) to make unique dataset.

Table 5 The 10 and 24 common selected features

10 selected features	Marriage Status, BMI, Follicle R, FSH, Hip, Weight, Fast food, Hair growth, Skin Darkening, BPDiastolic
24 selected features	Height, blood group, pulse rate , RR, HB, cycle, cycle length, pregnant, abortions, LH, waist hip ratio, TSH, AMH, PRL,Vit D3, PRG, RBS, Weight gain, hair loss, pimples, reg exercise, bpsystolic, avgfsize L, Endometrium

5.3 Performance on 10 selected features

Initially, 10 selected features were taken to build a model. In 'Python', the highest accuracy is shown by SVM (84.54%) with similar performances in precision and recall (85%), followed by AdaBoost (84.9%) followed by random forest (83.63%), with similar performances in precision and recall (85% and 84%) respectively. Whereas in 'RapidMiner' the KNN and SVM show similar performances with the highest accuracy (90.83%) and recall (90.83%), with an exemption for recall (KNN 90.83% and 91.33 % SVM). In comparison to both Python and RapidMiner, the 'RapidMiner' shows the highest average accuracies (85.97%) and the same goes for precision (86.93%) and recall (85.97%) respectively. Thus, better performance is shown by the RapidMiner tool than Python (Table 6).

Table 6 Performance Comparisons of Python and RapidMiner ML-Techniques with 10 selected Features

	Python			**RapidMiner**		
Algorithms	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KNN	72.27	73	72	90.83	90.83	90.83
SVM	84.54	85	85	90.83	91.33	90.82
Random Forest	83.63	85	84	89.91	90.4	89.90
Naïve Bayes	63.18	75	63	86.24	86.54	86.24
Neural Network	78.63	79	79	79.82	84.45	79.82
Bagging	80.90	82	81	80.28	80.4	80.27
Adaboost	84.09	85	84	83.94	84.6	83.94
Average	78.17	80.57	78.28	85.97	86.93	85.97

5.4 Performance on 24 selected features

In 'Python' the highest accuracy, precision, and recall with a similar performance of the classifiers is shown by random forest and adaboost (78.18%, 79%, and 78%) respectively, followed by bagging (75.54%, 76% and 75). Whereas 'RapidMiner' KNN shows the highest accuracy (84.86%), followed by SVM (84.40%). In comparison to both Python and RapidMiner,

the 'RapidMiner' shows the highest average accuracies (73.52%) and the same goes for precision (74.04%) and recall (80.66%) respectively. Thus, better performance is shown by the RapidMiner tool than Python (Table 7).

Table 7 Performance Comparisons of Python and RapidMiner ML-Techniques with 24 selected Features

	Python			RapidMiner		
Algorithms	Accuracy	Precision	Recall	Accuracy	Precision	Recall
KNN	63.63	64	64	84.86	84.93	84.86
SVM	73.63	74	74	84.40	84.82	84.40
Random Forest	78.18	79	78	74.77	75.12	74.77
Naïve Bayes	67.72	77	68	76.61	78.94	76.61
Neural Network	51.36	52	51	50	50	100
Bagging	75.54	76	75	66.97	67.12	66.97
Adaboost	78.18	79	78	77.06	77.39	77.06
Average	69.74	71.57	69.71	73.52	74.04	80.66

5.5 Performance comparison on complete and selected features, and tools

Figure 2 shows the average accuracies comparison between complete (40 features) and selected features (10 and 24 features). In Python, The 10 selected features comparison with 40 features shows an insignificance performance for accuracy, precision and recall with a range of ±0–5.15%. Similarly, the 10 features show a significant difference in performance for 24 selected features for accuracy, precision and recall with a range of 8.43–9% respectively. On the other hand in RapidMiner, the 10 selected features comparison with 40 features show an insignificance performance with a range of ±1.55–4.65%. Similarly, 10 selected features comparison with 24 selected features shows a significant difference in performances with a range of 5.31–12.89% respectively. Thus, indicating the 10 selected features can be useful to build a better model instead of 40 and 24 features respectively. The high evaluation performances in RapidMiner indicate it as a better tool.

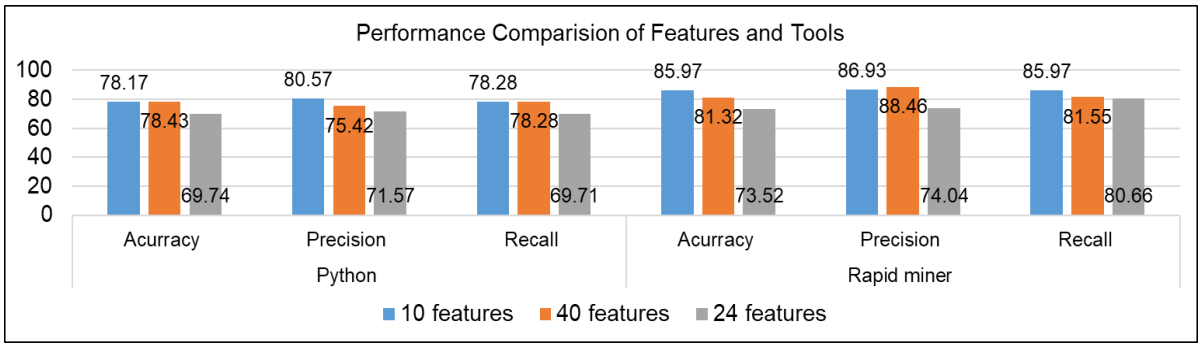

Figure 2. The average performances comparison of 10, 40 and 24 features

6. DISCUSSION

In this study, the seven classification and five feature selection algorithms implemented in Python-Scikit learn package and RapidMiner has been used to evaluate the PCOS. The complete dataset (40 features) and selected features (10 and 24 features) was used to study the classification model. In a complete dataset, the highest accuracy is shown by random forest (93.12%, RapidMiner). The results are in agreement with an insignificance difference (89.02%, random forest, and 95% SVM) of previous studies [7], [10]. On the other hand, the 10 selected features accuracy results (90.83% KNN and SVM) shows an insignificance performance with the previous studies where random forest show and accuracy of 89.02% and J48 (96% accuracy) respectively, where some of the selected features in this study were used previously [7], [11]. Despite different tools used the average performances of 10 selected feature shows similar results with the combined dataset, indicating these features could be used for the prediction of PCOS [7], [11]. On the other hand, the 24 selected features showed a significant difference in their performances with both complete and 10 selected features thus indicating these features cannot be used for model building. In the comparison of both tools, the results showed that RapidMiner performs better than Python. However, these rules could not be applied since the nature of the dataset varies, and where python tool was used which showed good accuracy performance [7], [1]. This scenario could be possible because of the different algorithm performances on datasets since the nature of the dataset does play an important role in the performance of the classification model.

7. CONCLUSION

In this study, the evaluation of the PCOS dataset has been carried out with seven classification and 5 feature selection methods. Regardless of the different tools used, RapidMiner showed 93.12% accuracy with the complete dataset (random forest) and 90.83% for 10 selected features (KNN) respectively. The average performances of 10 selected features show similar results with a complete dataset. The insignificance performance differences between and 10 selected features and complete attributes show these selected features can be useful to build a better model for the PCOS dataset. The performances of each classifier and average performances show that Rapid Miner can be used as an alternative machine learning tool. However, this cannot be a general rule since the performances depend on the nature of the dataset, sampling, and preprocessing steps.

ACKNOWLDGEMENT
This work was supported by the Kementerian Pendidikan Malaysia, Fundamental Research Grant Scheme [Grant Number 203/PKOMP/6711797].

REFERENCES

[1] Vikas, B., Anuhya, B. S., Chilla, M., & Sarangi, S., A Critical Study of Polycystic Ovarian Syndrome (PCOS) Classification Techniques. International Journal of Computational Engineering & Management **21**, 4 (2018) 1-7.

[2] Saravanan, A., & Sathiamoorthy, S., Detection of Polycystic Ovarian Syndrome: A Literature Survey. Asian Journal of Engineering and Applied Technology **7,** 2 (2018) 46-51.

[3] Krishnaveni, V., A Roadmap to a Clinical Prediction Model with Computational Intelligence for PCOS. International Journal of Management, Technology and Engineering **9**, 2 (2019) 177-185

[4] Dewi, R.M., & Wisesty, U.N., Classification of polycystic ovary based on ultrasound images using competitive neural network. Journal of Physics: Conference Series **971,** 1 (2018) 012005, March 2018.

[5] Reddy, N.S.C., Nee, S.S., Min, L.Z., &Ying, C.X., Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. International Journal of Innovative Computing **9,** 1 (2019) 39-46

[6] Meena, K., Manimekalai, M., & Rethinavalli, S., A Novel Framework for Filtering the PCOS Attributes using Data Mining Techniques. International Journal of Engineering Research & Technology **4**, 1 (2015) 702-706.

[7] Denny, A., Raj, A., Ashok, A., Ram, C. M., & George, R., i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques. TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, (2019) 673-678.

[8] Anuradha, D.T., & Priyanka R. L., Genetic Clustering for Polycystic Ovary Syndrome Detection in Women of Reproductive Age. International Journal of Engineering and Advanced Technology **9**, 3 (2020) 1359-1361.

[9] Wang, J., Jain, S., Chen, D., Song, W., Hu, C.T., & Su, Y.H., Development and Evaluation of Novel Statistical Methods in Urine Biomarker-Based Hepatocellular Carcinoma Screening. Sci Rep **8**, 1 (2018) 3799.

[10] Deshpande, S.S., & Wakankar, A., Automated detection of Polycystic Ovarian Syndrome using follicle recognition. IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram (2014) 1341-1346.

[11] Vijayalakshmi, N., & UmaMaheswari, M., JCSMC, Vol. 5, Issue. 8, August 2016, pg.5 – 9 Data Mining to Elicit Predominant Factors Causing Infertility in Women. International Journal of Computer Science and Mobile Computing **5**, 8 (2016) 5-9.

[12] https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[13] Scikit-learn, "Scikit-learn: Machine Learning in Python," 2016.

[14] Mierswa, I., & Klinkenberg, R., RapidMiner Studio (9.1) [Data science, machine learning, predictive analytics]. (2018) Retrieved from https://rapidminer.com/.

[15] Ray, C., and Ray, A., Intrapartum cardiotocography and its correlation with umbilical cord blood pH in term pregnancies: a prospective study. International Journal of Reproduction, Contraception, Obstetrics and Gynecology **6**, (2017) 2745-2752.