PCOS Detection

Using Machine Learning Algorithms

Data Warehousing and Data Mining Lab

Team Members:

Name	Registration Number
Sanjeev Kushal Pendekanti	200911160
Riddhi Rajendra Dayma	200911212
Anoushka Kondaskar	200911188

Abstract:

This research paper aims to predict the likelihood of PCOS based on specific factors, using a machine learning and data mining model. Among women of reproductive age, polycystic ovarian syndrome (PCOS) is a prevalent endocrine condition. Early diagnosis and efficient treatment can prevent long-term health complications.

To efficiently predict the likelihood of PCOS in females, the data was pre-processed and cleaned followed by employing five different classification algorithms, including Logistic Regression, Naive Bayes Classification, KNN, and Random Forest Classification and Support Vector Machine. The models were trained using data on variables such as follicle number (R), follicle number (L), skin darkening, hair growth, weight gain, regularity of menstrual cycle, and fast food consumption. The outcome of the model was either categorized as "yes" or "no" depending on whether the patient had PCOS or not.

The findings indicate that the Support Vector Machine algorithm had the highest level of PCOS prediction accuracy, outperforming the other three algorithms. According to the analysis of the significance of each feature in predicting the result, follicle number (R) was the most crucial feature, followed by Follicle No. (L) and Weight gain (Y/N).

This model has the potential to assist medical professionals in identifying patients who may be at risk for PCOS and providing timely interventions. The project highlights the value of feature selection and algorithm selection in creating precise predictive models, underscoring the potential of machine learning and data mining in healthcare.

Introduction:

Polycystic Ovarian Syndrome is a common hormonal disorder in women that occurs in up to 13% of women that are in the reproductive-aged category. A variety of clinical symptoms, including irregular menstruation, hyperandrogenism, infertility, and metabolic abnormalities, are caused by the complex interaction of ovarian, hormonal, metabolic, and metabolic dysfunctions that define it.

It is usually also accompanied by extremely high levels of androgen hormone. This is usually present in small amounts in women since it is a male sex hormone. These women have high insulin resistance causing high levels of insulin in the body causing the presence of high androgen levels. Another problem with high levels of insulin is the addition of a bigger appetite that results in the sudden gain of weight.

The diagnosis of PCOS is sometimes difficult since it necessitates the presence of several clinical and laboratory criteria, despite its high frequency and clinical relevance. The cause of PCOS is not exactly known but as of now there is a high probability that it is genetic in nature and is passed down from generations.

Machine learning and predictive modeling approaches have had recently had promising results in the prediction of the presence of PCOS based on various factors such as age, height, weight, follicle size etc. as well as lifestyle based factors such as the fast food consumption. The early detection of this disorder can help young women manage the disorder in a timely manner and so that can prevent further complications. We aim in this paper; to do exactly this and give a more certain prediction for the presence of PCOS based on these factors.

Literature Review:

In [1]:

This paper is addressing the problem of diagnosing PCOS through various blood tests and pelvic ultrasound. The model employs several classifiers, including K-Nearest Neighbor (KNN), Naive Bayes, Support Vector Machine (SVM), Classification Tree and Logistic Regression. According to the results, the KNN classifier excels in terms of sensitivity, while the Linear Discriminant classifier excels in terms of precision.

In [2]:

The purpose of the paper is to increase the diagnosis rates for PCOS, which are low due to poor awareness among women. The model evaluates algorithm performance using metrics such as accuracy, F-statistics, recall, and precision utilizing network-obtained data. The study found that logistic regression, out of all the techniques, has a 90% accuracy rate.

In [3]:

The approach for predicting PCOS in this paper is based on artificial neural networks. Standard scalar methods like oversampling minority data, undersampling majority data, Synthetic Minority Over Sampling Technique, and ensemble are used to organize the data.The paper concluded that the ANN model is able to diagnose PCOS with a precision of 84%.

In [4]:

The intersection between typical PCOS characteristics and known risk factors for severe COVID-19 is evaluated in this paper. It provides a comprehensive analysis of parameters like Hyperandrogenism, Cardio-metabolic comorbidity, hyper-inflammation and low vitamin D levels as correlating factors. The results show the strong correlation between the two and the authors propose measures to control the potential risks.

In [5]:

In order to diagnose polycystic ovary syndrome (PCOS) in obese and non-obese women, the LH: FSH ratio is examined in this study.Computerized documents were retrieved in order to acquire the demographic, clinical, and laboratory data. The BMI and LH/FSH ratio Spearman correlation was calculated, and the study groups were contrasted using the t-test. According to the findings, a reduced body mass index was not associated with a higher LH/FSH ratio.

In [6]:

This model uses various metrics to study the prediction of PCOS to help with early diagnosis. Employed algorithms include CatBoost, RF, SVM, Logistic Regression, Bernoulli Naive Bayes, Decision Tree and K-Nearest Neighbor. With Pearson's correlation feature extraction technique and 38 variables, CatBoost and RFC perform the best, with rate of accuracy being of 93.90% and 92.68%, respectively.

In [7]:

PCOS must be identified as it increases the risk of complex long-term issues. In this study, a hybrid classifier has been made by combining SVM, RF and XGboosting techniques. The hybrid model has an accuracy rate of 93.8% in detecting PCOS in its early stages.

In [8]:

The occurrence of PCOS is determined using all of the ailments that could be its side consequences using decision tree, gradient boosting, random forest, logistic regression, K-nearest neighbor, hybrid RFLR, and SVM algorithms to predict it. The gradient boosting algorithm has the highest accuracy of 98.9%.

In [9]:

The actual causes of PCOS are still unknown, despite the fact that it is largely accepted as a lifestyle disorder. This study uses the following algorithms to classify PCOS- Random Forest, SVM, decision tree, logistic regression, linear discriminant analysis, KNN, XGboost, etc. RF has the highest accuracy rate of 92.4% at a correlation threshold of 0.8.

In [10]:

The primary focus of this study is the diagnosis of PCOS in females using the LH:FSH ratio for detection. The study uses gradient boosting, RF, hybrid random forest and logistic regression to test accuracy. RFLR is found to have the highest testing accuracy of 91.01% and recall value of 90%.

In [11]:

In this paper machine learning methods and approaches are used to analyze medical images of PCOS. The basis for this is that ultrasound images and other advanced medical imaging of ovaries and vaginal area can easily detect patients suffering from PCOS. The automated detection of PCOS being analyzed is through methods like segmentation and classification. Conclusion that segmentation is used to focus on specific areas of the ultrasound images and carry out more extensive analysis on them whereas classification is employed to locate the follicles visible in the picture.

In [12]:

The prediction of PCOS using advanced machine learning algorithms on data from 541 patients. This is done based on an optimized version of the chi square mechanism. The machine learning algorithms used for this are Random Forest, Linear Regression, Stochastic gradient descent, Support Vector Machine, K nearest neighbor classifier, MLP classifier, Logistic regression, Gaussian NB and Gradient Boosting Classifier. Conclusion that Gaussian NB performed the best with an accuracy of 100% and time computation of 0.002.

In [13]:

The self-prediction of PCOS using machine learning algorithms for both potential patients and clinical providers using the data of 541 women in Kerala. CatBoost method is used for classification. The results being 81-82.5% accurate without invasive predictor variables for self-prediction and 87.5-90.1% accurate with and without invasive predictor variables.

In [14]:

They have used a RNA-sequence based dataset and applied Random Forest classifier and artificial neural network. The RF classifier found key genes with a very high level of accuracy of more than 75% as compared to the RNA that was not biomarked.

In [15]:

They have used ultrasound for detection of infertile patients and employed the use of SVM, KNN and Logistic Regression for classification. A hybrid approach is used to detect PCOS and the accuracy was found to be 98%.

In [16]:

They have used a hybrid approach of ML algorithms to detect PCOS. The results showed that the Linear Support Vector Machine performed the best among all the algorithms applied in terms of accuracy, recall and precision with an accuracy rate of 93.665%.

In [17]:

The researchers have used various ML algorithms like Naive Bayes, logistic regression, KNN, RFC and SVM. They found prediction of PCOS to be most accurate when the Random Forest Classifier was applied (89.02%).

In [18]:

The researchers use the Logistic Regression and Bayesian classifiers to detect the selected features. The overall accuracy of the Bayesian and logistic regression algorithms are found to be 93.93% and 91.04% respectively.

In [19]:

A hybrid algorithm is used that combines SVM, RF and XGBoosting algorithms. This new classifier is compared with results from other commonly applied ML classifiers but is found to be superior with an accuracy of 93.8%.

In [20]:

This research shows the use of decision tree algorithms to diagnose PCOS in women. Using the RF classifier, an efficient decision tree is derived that provides a cautionary advisory to women regarding development of PCOS.

In [21]:

In this case the research follows the pursuit to check how well suited the Random Forest classifier is for gene selection methods. It was found that the Boruta algorithm from the Random Fern classifier found the most potent and important genes.

In [22]:

An attempt to develop an algorithm to detect PCOs is made. Different classifiers like SVM, KNN and decision tree are applied and it is found that SVM performs the best out of these three classifiers.

In [23]:

This research studies the differences between the variations of the RF classifier to check if the acceptance-rejection variation will perform better. It is found that the AR version of random forest performs marginally better.

In [24]:

This research combines a CNN with the XGBoost classifier for image classification of the PCOS dataset to accurately predict the presence of PCOS. This combined algorithm is found to have an accuracy of 99.89%.

In [25]:

This study goes into the depth of the impact of feature selection algorithms. It is found that the accuracy and validity of the classification increases when various feature selection and feature agglomeration techniques are applied.

Methodology:

Initially data obtained from the dataset was visualized to provide a better insight into the algorithms and machine learning models to be used for the analysis.

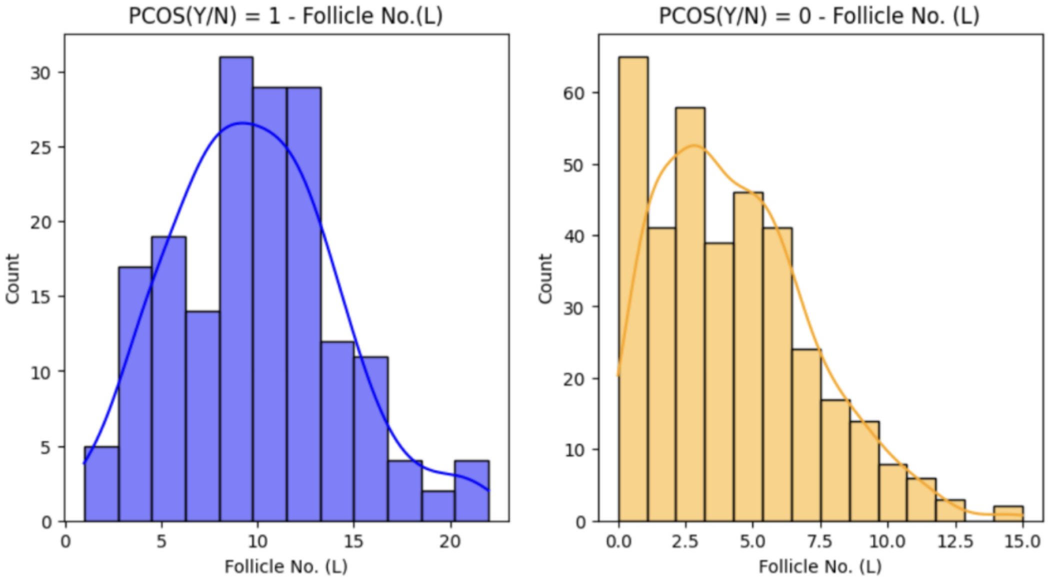


Figure 1. The above figure shows a histogram plotted between the attribute Follicle No. (L) and PCOS (Y/N)

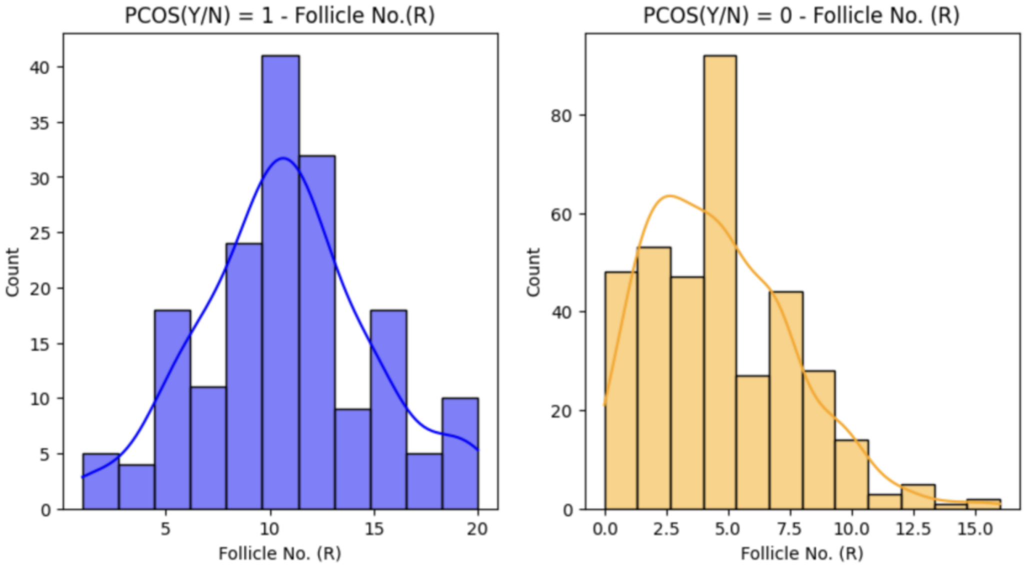


Figure 2. The above figure shows a histogram plotted between the attribute Follicle No. (R) and PCOS (Y/N)

From looking at the graphs it was concluded that the data was normally distributed. This resulted into Naïve Bayes being a potentially suitable model of choice for the prediction. To ensure accurate results in this study, the dataset underwent a comprehensive cleaning process. Patient file number and unidentified were omitted from the attributes list since they might have skewed the analysis. Additionally, the respective median values were used to impute missing data. A correlation analysis was performed to exclude features with low correlation ratios in order to further refine the dataset.

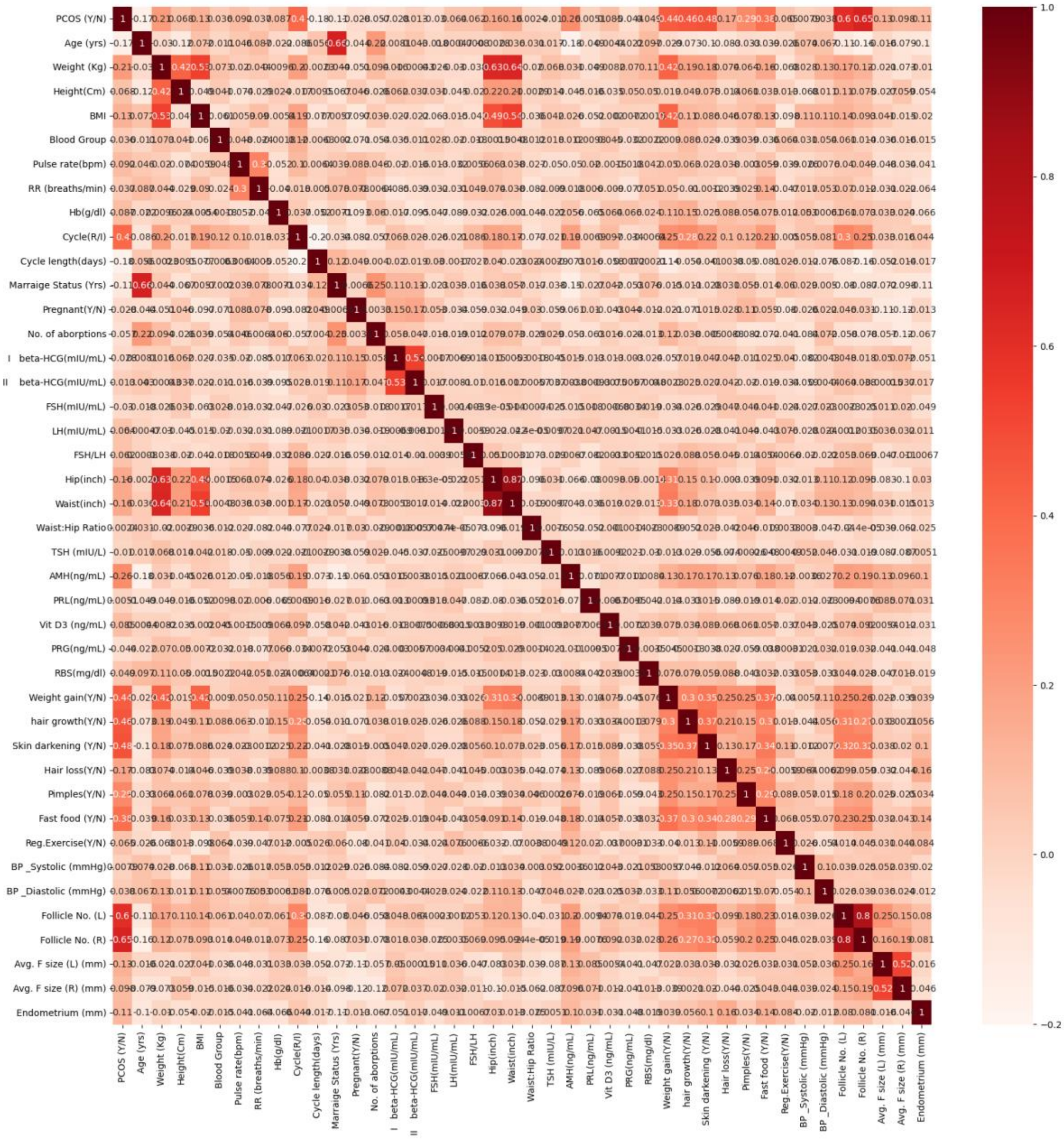


Figure 3. The above figure shows a correlation heatmap between all the attributes.

For the project, the data was split into a train-test split ratio of 0.3 i.e., 30% data was used as test sample. The training feature was denoted as "X_train" and consisted of attributes based on which the target variable "y_train" could be predicted. "y_train" consisted of the diagnosis of PCOS (Y/N), while "X_train" comprised predictor variables.

Based on the correlation matrix in Figure 3, a subset of highly correlated attributes namely - ‘Follicle No (L)’, ‘Follicle No (R)’, ‘Fast food’ , ‘Pimples’, ‘Cycle Regularity’, ‘Skin Darkening’, ‘Hair Growth’ and ‘Weight Gain’ were selected.

Initially, three algorithms were utilized to predict the prevalence of PCOS in females based on various features, such as hormones and physical characteristics.

After cleaning the data, two additional algorithms, KNN and SVM, were applied to improve the accuracy of the predictive models. These algorithms were chosen based on their proven track record in classification tasks and their ability to handle high-dimensional datasets with complex relationships between features.

The discussion and analysis of the two algorithms we have used are given below:

1. Logistic Regression:

A statistical approach for binary classification issues is logistic regression. It models the likelihood of a binary response variable—say, yes or no as a function of one or more predictor variables. In logistic regression, a logistic function is used to model the response variable as a function of the predictor variables. The logistic function converts any input value to a value between 0 and 1, which can be understood as the likelihood that, given the predictor variables, the response variable will be in the positive class.

The logistic regression model can be represented as:

$$P(Y=1|X)=\frac{\mathrm{e}^{(\beta_o+\beta_1 x)}}{\mathrm{e}^{(\beta_o+\beta_1 x)}+1}$$

Where:

- y is the binary response variable
- X is a vector of predictor variables
- β0 is the intercept term (constant)
- β1, β2, ..., βp are the coefficients of the predictor variables
- exp() is the exponential function

A logistic regression model is fitted to the training data and then the array of predictor variables (X) is used to estimate the coefficients of the model. The coefficients help represent the relationship between each predictor variable and the log-odds of the outcome variable (Y).The log-odds can then be converted to probabilities between 0 and 1 using logistic functions.

The results obtained from the logistic regression model can then be evaluated using metrics like F1 score, accuracy , precision etc. which give us a measure of how well the model can predict the presence of PCOS based on the predictor variables.

2. Naive Bayes Classification

By computing the likelihood of each class given the input data, the Naive Bayes method determines the class with the highest probability. This is accomplished by first determining the antecedent(prior) probability of each class, or the likelihood that each class would exist even if the input data had not been considered, and then establishing the conditional probability of each class in its inclusion of the input data.

- It goes on the premise that each trait is independent. For instance, a car's color, "Yellow," has nothing to do with its type or origin.
- It accords equal weight to all of the features. For instance, the result could be accurately predicted using only the Color and Origin information. Because of this, each component is equally crucial and essential to the outcome.
- Equation for Naive Bayes:

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

On applying this algorithm to our dataset we were able to achieve a prediction accuracy score of 81.59%.

However, in our case, it is not possible that all parameters associated with the tests are independent of each other. Thus, some other algorithm is more likely to give a better score for the prediction model.

3. Random Forest Classifier:

For classification and regression applications, Random Forest is a strong and adaptable machine learning method. It is an ensemble method that integrates various decision trees to produce a model that is more reliable and accurate. When using the Random Forest technique, numerous decision trees are built during the training phase, and their predictions are then combined during the testing phase.

- Random Forest chooses a selection of samples from the training set at random in order to train each decision tree. Bootstrapping is a method that reduces overfitting by building different and independent trees.
- Build decision trees: Next, the Random Forest method builds numerous decision trees using different subsets of the features and samples. A different subset of the training samples and a random subset of the characteristics are used to build each tree.
- Predictions in aggregate: Each decision tree forecasts the test data throughout the testing phase. By combining all of the decision trees' predictions, a class label is assigned to the test data. This aggregate can be done by getting the majority vote of the predictions or by averaging the predicted values.

4. KNN Classification:

A straightforward yet efficient non-parametric approach used for both classification and regression applications is K-Nearest Neighbors (KNN). It operates by locating the K data points in the training set that are closest to an unknown data point, and then classifying the unknown point according to the majority class of its neighbors. The user can significantly affect the algorithm's performance by selecting K's value. KNN is well-liked for being straightforward and easy to understand, although it can be computationally expensive for big datasets.

The steps are as follows:

- Load the training data: The labelled training data must be loaded first in KNN. The training set consists of a set of features (also known as predictor variables) and a corresponding class label for each data point.
- Decide on the value of K: The next stage is to decide on the value of K, which establishes how many nearest neighbors will be taken into account when categorizing a new data point.
- Determine the distance: The distance between a new data point and each data point in the training set is determined in order to identify the K nearest neighbors for that point. The Euclidean distance is the most often used distance measure.
- The K data points with the closest distances to the new point are chosen as its nearest neighbors after the distances have been calculated.
- Decide on the class label: The new data point's class label is finally decided by a majority vote among its K nearest neighbors' class labels. In other words, the new point is given the class label with the highest frequency among its K neighbors.

5. Support Vector machine:

Support Vector Machine (SVM) is employed for classification and regression analysis. It seeks to identify the ideal dividing line between the various groups of data points. Both linearly separable and non-linearly separable data can be handled by SVM. SVM uses a kernel function to transform non-linearly separable data into a higher-dimensional space. The hyperplane that maximizes the margin while minimizing the classification error is then discovered using the SVM method by solving an optimization issue. SVM may classify new data points by identifying which side of the hyperplane they fall on once the hyperplane has been established.

- Pick the kernel function: To move the data into a higher-dimensional space where it can be more easily segregated, the SVM algorithm requires a kernel function. The type of data used and how complicated the decision boundary is determining which kernel function should be used.
- Define the hyperplane: The SVM method determines the hyperplane that maximally divides the various classes of data points after the data is translated into a higher-dimensional space. The data points nearest to the hyperplane, define the hyperplane.
- Margin optimization: The SVM method seeks to maximize the margin, which is the separation between the nearest data points in each class and the hyperplane.
- Identifying which side of the hyperplane new data points land on allows the SVM algorithm to classify them once the hyperplane has been defined. The decision function's sign determines the class label that is applied to the test data.

Results:

Precision: It is the percentage of true positives (positive examples that the model correctly classified) among all positive predictions.

$$Precision\ score = \frac{TP}{TP\ +\ FP}$$

Recall: Out of all real positive examples, it is the proportion of true positives.

$$Recall\ score = \frac{TP}{TP\ +\ FN}$$

F1-score: It is harmonic mean for recall and precision. False positives and false negatives are both taken into account.

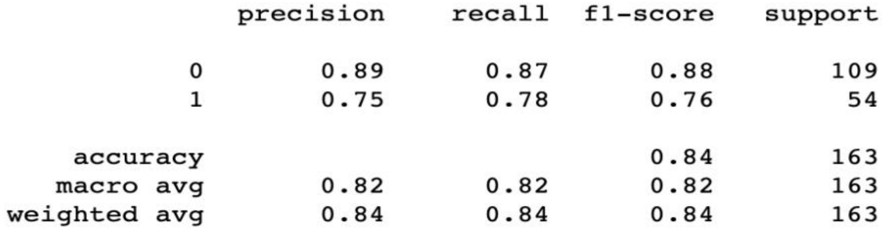
$$F1 - score = \frac{2\ *\ (Precision\ *\ Recall)}{(Precision\ +\ Recall)}$$

Accuracy: It is the percentage of the model's predictions that were accurate overall.

$$Accuracy\ score = \frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP\ +\ FN}$$

Support: It refers to the frequency or percentage of times that a particular itemset or pattern appears in a given dataset.

1. **Logistic Regression**



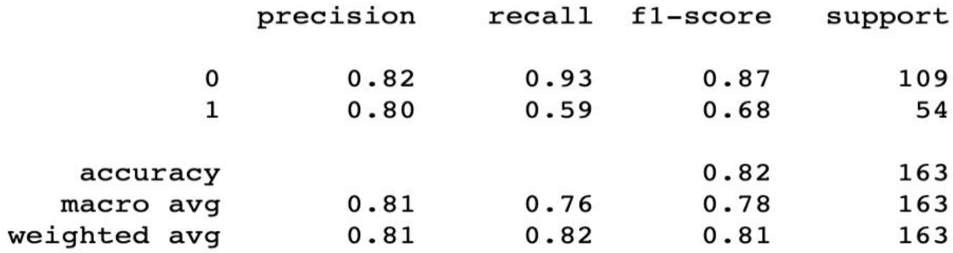
```
              precision    recall  f1-score   support

           0       0.89      0.87      0.88       109
           1       0.75      0.78      0.76        54

    accuracy                           0.84       163
   macro avg       0.82      0.82      0.82       163
weighted avg       0.84      0.84      0.84       163
```

In this instance, the accuracy is 0.84, indicating that 84% of the test set examples had their class (yes/no for PCOS) correctly predicted by the model.

2. **Naive Bayes Classification**



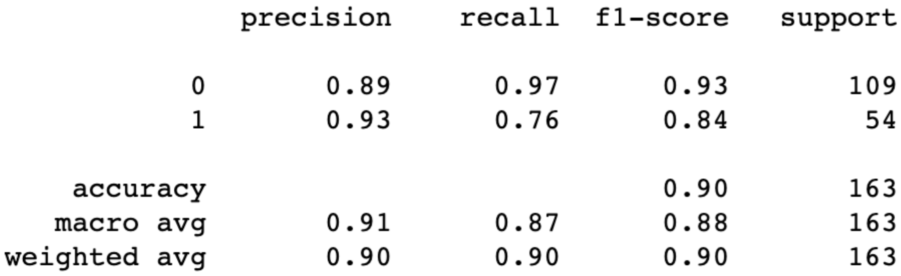
```
              precision    recall  f1-score   support

           0       0.82      0.93      0.87       109
           1       0.80      0.59      0.68        54

    accuracy                           0.82       163
   macro avg       0.81      0.76      0.78       163
weighted avg       0.81      0.82      0.81       163
```

In this instance, the accuracy is 0.82, indicating that 82% of the test set examples had their class (yes/no for PCOS) correctly predicted by the model.

3. **Random Forest Classification**



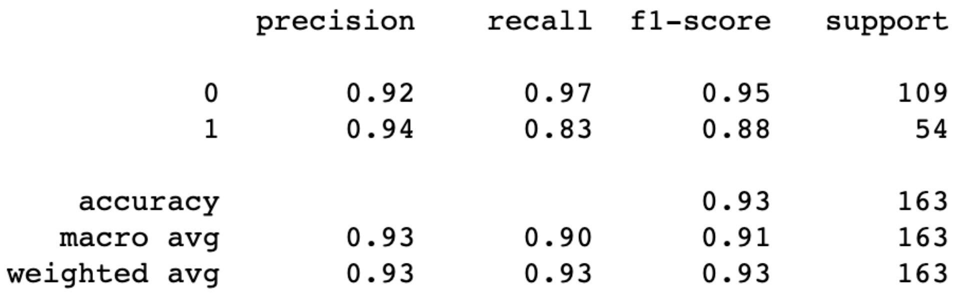
```
              precision    recall  f1-score   support

           0       0.89      0.97      0.93       109
           1       0.93      0.76      0.84        54

    accuracy                           0.90       163
   macro avg       0.91      0.87      0.88       163
weighted avg       0.90      0.90      0.90       163
```

A 90.02% accuracy rate for the dataset was shown by the random forest model. The random forest model's high accuracy highlights both how well it can handle challenging classification problems and how it can take advantage of ensemble learning to provide reliable predictions.

4. KNN Classification



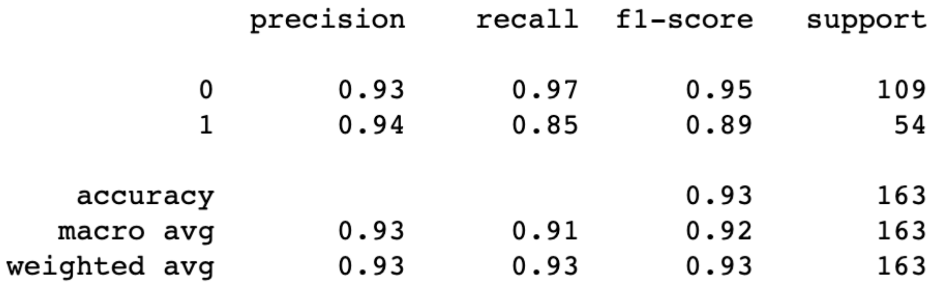
```
              precision    recall  f1-score   support

           0       0.92      0.97      0.95       109
           1       0.94      0.83      0.88        54

    accuracy                           0.93       163
   macro avg       0.93      0.90      0.91       163
weighted avg       0.93      0.93      0.93       163
```

The dataset's accuracy for the K-nearest neighbours (KNN) model was 92.63%. This demonstrates the KNN algorithm's usefulness, especially in conditions when the data displays intricate decision boundaries.

5. Support Vector Machine



```
              precision    recall  f1-score   support

           0       0.93      0.97      0.95       109
           1       0.94      0.85      0.89        54

    accuracy                           0.93       163
   macro avg       0.93      0.91      0.92       163
weighted avg       0.93      0.93      0.93       163
```

The SVM model proved to be the best option for predicting the prevalence of PCOS in females, achieving a remarkable accuracy of 93.25% on the dataset. The SVM algorithm is well-suited for classification tasks in high-dimensional spaces, making it an ideal choice for this study.

Conclusion:

This study compared five different machine learning classifiers, including Logistic Regression, Naive Bayes Algorithm, Random Forest Classification, K-Means, and Support Vector Machine (SVM), for the detection and identification of Polycystic Ovarian Syndrome (PCOS) in menstruating females. The results demonstrated that SVM achieved the highest accuracy rate of 93.25%.SVM was chosen as one of the classifiers due to its ability to handle datasets with a large number of features and those with a clear margin of separation between classes. The dataset used in this study consisted of binary predictions (yes/no) based on a large number of attributes, with a clear margin of classes between PCOS-positive and PCOS-negative samples. Therefore, SVM is a promising tool for PCOS diagnosis.

References:

[1] D. Hdaib, N. Almajali, H. Alquran, W. A. Mustafa, W. Al-Azzawi and A. Alkhayyat, "Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms," 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, 2022, pp. 532-536, doi: 10.1109/IICETA54559.2022.9888677.

[2] S. Dhinakaran, C. Thangavel, S. S and H. V S, "PCOS Perception analysis prediction using Machine learning algorithms," 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), MANGALORE, India, 2022, pp. 260-265, doi: 10.1109/ICRAIE56454.2022.10054279.

[3] P. Chitra, M. Sumathi, K. Srilatha, F. V. Jayasudha and S. Amudha, "Review of Artificial Intelligent based Algorithm for Prediction of Polycystic Ovary Syndrome(PCOS) from Blood Samples," 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2022, pp. 1172-1176, doi: 10.1109/ICIRCA54612.2022.9985699

[4] Kyrou, I., Karteris, E., Robbins, T. et al. Polycystic ovary syndrome (PCOS) and COVID-19: an overlooked female patient population at potentially higher risk during the COVID-19 pandemic. BMC Med 18, 220 (2020). https://doi.org/10.1186/s12916-020-01697-5

[5] Saadia, Z. (2020). Follicle Stimulating Hormone (LH: FSH) Ratio in Polycystic Ovary Syndrome (PCOS) - Obese vs. Non- Obese Women. Medical Archives, 74(4), 289-293. https://doi.org/10.5455/medarh.2020.74.289-293

[6] Y. Rathod et al., "Predictive Analysis of Polycystic Ovarian Syndrome using CatBoost Algorithm," 2022 IEEE Region 10 Symposium (TENSYMP), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/TENSYMP54529.2022.9864439.

[7] S. R. Swamy and N. P. K S, "Hybrid Machine Learning Model for Early Discovery and Prediction of Polycystic Ovary Syndrome," 2022 Second International Conference on Advanced Technologies in Intelligent Control,

Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2022, pp. 1-8, doi: 10.1109/ICATIECE56365.2022.10047488.

[8] Aggarwal, S., & Pandey, K. (2023). Early identification of PCOS with commonly known diseases: Obesity, diabetes, high blood pressure and heart disease using machine learning techniques. Expert Systems with Applications, 217, 119532. https://doi.org/10.1016/j.eswa.2023.119532

[9] Tiwari, S., Kane, L., Koundal, D., Jain, A., Alhudhaif, A., Polat, K., Zaguia, A., Alenezi, F., & Althubiti, S. A. (2022). SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. Expert Systems with Applications, 203, 117592. https://doi.org/10.1016/j.eswa.2022.117592

[10] S. Bharati, P. Podder and M. R. Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1486-1489, doi: 10.1109/TENSYMP50017.2020.9230932.

[11] N. Jan, A. Makhdoomi, P. Handa and N. Goel, "Machine learning approaches in medical image analysis of PCOS," 2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS), Bhubaneswar, India, 2022, pp. 48-52, doi: 10.1109/MLCSS57186.2022.00017.

[12] S. Nasim, M. S. Almutairi, K. Munir, A. Raza and F. Younas, "A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics," in IEEE Access, vol. 10, pp. 97610-97624, 2022, doi: 10.1109/ACCESS.2022.3205587.

[13] Zigarelli A, Jia Z, Lee H, Machine-Aided Self-diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study JMIR Form Res 2022;6(3):e29967 URL: https://formative.jmir.org/2022/3/e29967 DOI: 10.2196/29967

[14] Ning-Ning Xie, Fang-Fang Wang, Jue Zhou, Chang Liu, Fan Qu, "Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network", BioMed Research

International, vol. 2020, Article ID 2613091, 13 pages, 2020. https://doi.org/10.1155/2020/2613091

[15] J. Madhumitha, M. Kalaiyarasi and S. S. Ram, "Automated Polycystic Ovarian Syndrome Identification with Follicle Recognition," 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, 2021, pp. 98-102, doi: 10.1109/ICSPC51351.2021.9451720.

[16] Y. A. Abu Adla, D. G. Raydan, M. -Z. J. Charaf, R. A. Saad, J. Nasreddine and M. O. Diab, "Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques," 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME), Werdanyeh, Lebanon, 2021, pp. 208-212, doi: 10.1109/ICABME53305.2021.9604905.

[17] A. Denny, A. Raj, A. Ashok, C. M. Ram and R. George, "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, pp. 673-678, doi: 10.1109/TENCON.2019.8929674.

[18] P. Mehrotra, J. Chatterjee, C. Chakraborty, B. Ghoshdastidar and S. Ghoshdastidar, "Automated screening of Polycystic Ovary Syndrome using machine learning techniques," 2011 Annual IEEE India Conference, Hyderabad, India, 2011, pp. 1-5, doi: 10.1109/INDCON.2011.6139331.

[19] S. R. Swamy and N. P. K S, "Hybrid Machine Learning Model for Early Discovery and Prediction of Polycystic Ovary Syndrome," 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2022, pp. 1-8, doi: 10.1109/ICATIECE56365.2022.10047488.

[20] A. S. Prapty and T. T. Shitu, "An Efficient Decision Tree Establishment and Performance Analysis with Different Machine Learning Approaches on Polycystic Ovary Syndrome," 2020 23rd International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh, 2020, pp. 1-5, doi: 10.1109/ICCIT51783.2020.9392666.

[21] Kursa, M.B. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics 15, 8 (2014). https://doi.org/10.1186/1471-2105-15-8

[22] Zhang, XZ., Pang, YL., Wang, X. et al. Computational characterization and identification of human polycystic ovary syndrome genes. Sci Rep 8, 12949 (2018). https://doi.org/10.1038/s41598-018-31110-4

[23] Calhoun, P., Hallett, M.J., Su, X. et al. Random forest with acceptance–rejection trees. Comput Stat 35, 983–999 (2020). https://doi.org/10.1007/s00180-019-00929-4

[24] Suha, S.A., Islam, M.N. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Sci Rep 12, 17123 (2022). https://doi.org/10.1038/s41598-022-21724-0

[25] Kondo, M., Bezemer, CP., Kamei, Y. et al. The impact of feature reduction techniques on defect prediction models. Empir Software Eng 24, 1925–1963 (2019). https://doi.org/10.1007/s10664-018-9679-5