**Evaluation of Machine Learning techniques for PCOS prediction**

**Supplementary Material Document**

**Dataset-** Dataset description is the process of summarizing and organizing the main characteristics of a data set. It is an important step in the data analysis process because it helps researchers to understand and make sense of the data that they are working with.

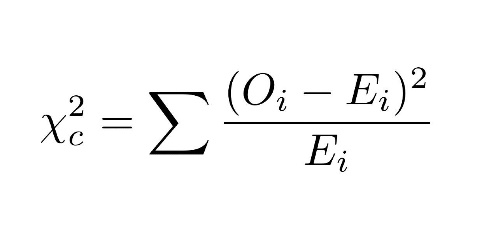
**Algorithms used:**

1. **Logistic Regression Classifier-** The method of estimating the likelihood of a discrete result over a given input variable is known as logistic regression. The most popular types of logistic regression produce a binary result, such as true or false, yes or no, and so on. Using multinomial logistic regression, events with more than two distinct potential outcomes may be modelled. When attempting to establish which category a new sample most closely resembles, classification issues are a good place to employ logistic regression as an analysis technique. Logistic regression is a helpful analytical method since cyber security involves classification difficulties, such as attack detection.
2. **Support Vector Classifier-** Support vector machines in machine learning are supervised learning models with related learning algorithms that examine data used for regression and classification analysis. An SVM training method creates a model that classifies incoming instances into one of two categories based on a collection of training examples that have each been identified as belonging to one of the two categories or the other.
3. **Decision Tree Classifier-** Decision tree is an approach for machine learning that is used for classification and regression applications.  In order to successfully anticipate the outcome of a given input, it operates by building a tree-like model of decisions and their potential outcomes. The method starts by taking into account all of the characteristics/features (sometimes referred to as "attributes") of the input data. The feature that divides the data into the most useful groupings or categories is then chosen. For each split, this procedure is repeated, with the algorithm selecting the characteristic that separates the data most effectively at each stage. The cycle repeats until the tree becomes fully grown or an end criterion is met, whichever comes first.
4. **Random Forest-** A large number of decision trees are built during the training phase of the random forest ensemble learning approach, which is used for classification, regression, and other tasks. The class that the majority of the trees choose is the output of the random forest for classification problems. The mean or average forecast of each individual tree is returned for regression tasks.
5. **K-Nearest Neighbor-** K-nearest neighbor (KNN) is a supervised machine learning algorithm utilized for both classification and regression applications. Based on the properties of the present training data points, KNN is utilized to make predictions on the test data set. Assuming that similar objects exist nearby, this is accomplished by measuring the distance between the test data and training data. The algorithm will have learning data stored, making it better at anticipating and classifying new data points. The KNN algorithm will pick up new data point's characteristics/features as it is input. The current training data points with the same characteristics or attributes will then be placed closer to the new data point.
6. **XG-Boost Classifier-** The gradient boosted trees approach is widely used and well implemented in open-source software called XGBoost. Gradient boosting is a supervised learning process that combines the predictions of a number of weaker, simpler models to attempt to properly predict a target variable. Regression trees serve as the weak learners when utilizing gradient boosting for regression, and each one of them associates each input data point with a leaf that holds a continuous score.

**Cat-Boost Classifier-** Another component of the gradient boosting method for decision trees is the CatBoost algorithm. Gradient Boosting Decision Tree (GBDT) implemented through CatBoost is a well-known and powerful open-source version. GBDT is a supervised learning technique that combines an ensemble of estimates from a variety of weaker and simpler models in an effort to properly forecast a target variable.

**Experimental Procedure**

1. **Data wrangling-** Data wrangling, also known as data munging, is the process of cleaning, preparing, and transforming raw data into a format that is suitable for analysis and machine learning. It is a critical step in the machine learning process because the quality and structure of the data you feed into a model can significantly impact the accuracy and performance of the model.
2. **Variable Description-** Variable description involves explaining the meaning and context of each variable in the data set, as well as the values that the variable can take on.
3. **Feature Selection- -** Feature selection is the process of selecting a subset of features to use in a machine learning model. The goal of feature selection is to select a subset of features that are most relevant and informative, while removing features that are redundant or irrelevant.
   1. **Correlation Matrix-** A correlation matrix is a table that shows the correlations between a set of variables. It is a useful tool for understanding how different variables are related to each other and can be used to identify patterns and trends in data.
   2. **Chi-Square Test-** The chi-square test of independence is a statistical test used to determine if there is a significant association between two categorical variables. The test is based on the chi-square statistic, which measures the difference between the expected frequencies of the variables in the population and the observed frequencies in the sample.

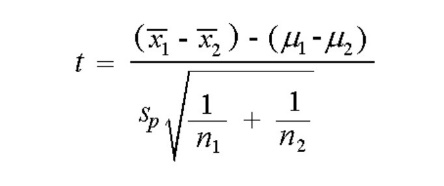


where;

O is the observed frequency in each cell of the contingency table

E is the expected frequency in each cell of the contingency table

* 1. **T-test-**



Where:

x̄1 is the sample mean of group 1

x̄2 is the sample mean of group 2

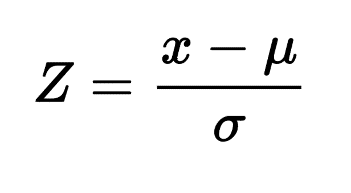
sp is the pooled estimate of variance

n1 is the sample size of group 1

n2 is the sample size of group 2

1. **Exploratory Data Analysis-** Exploratory data analysis (EDA) is an approach to analysing and summarizing a dataset in order to understand the underlying patterns and relationships.
2. **Data Modelling-**

**Standardized Value,**



Where;

x is value of variable

µ is mean of variable

σ is standard deviation of variable

1. **Data building-**
   1. **Model building** in machine learning refers to the process of creating a mathematical model that represents the relationships between different variables in a dataset. The goal of model building is to build a model that can make accurate predictions or decisions based on the data.
   2. **Hyperparameter tuning-** Hyperparameter tuning is the process of choosing the optimal set of hyperparameters for a machine learning model. Hyperparameters are values that are set before training a model, and they can have a big impact on the performance of the model. For example, in a decision tree model, the depth of the tree is a hyperparameter. A tree that is too shallow will not be able to capture the complexity of the data, while a tree that is too deep will overfit the data.

Hyperparameter tuning is used to find the best set of hyperparameters for a particular problem. This can be done manually, but it is often done using some type of automated optimization process, such as grid search or random search.

**Results and Discussions**

**Sensitivity**

Accuracy and Sensitivity are used as important metrics here because higher the accuracy means the chances of correct predictions are higher. Also, it is hoped to miss as few positive cases as possible for medical investigations. Thus, Recall is considered as higher recall values lower the positive cases missed.

1. Accuracy
2. Precision
3. Sensitivity
4. Specificity
5. f1-score
6. Confusion matrix
7. AUC-ROC curve