# UCS1603

# Introduction to Machine Learning

Mini Project

PCOS Prediction using ML

**Kiruthika J 185001078**

**Likhitha Verma A 185001084**

**Madhumitha S 185001086**

**AIM:**

**Polycystic Ovary Syndrome (PCOS)** is a medical condition which causes hormonal disorder in women in their childbearing years. The hormonal imbalance leads to a delayed or even absent menstrual cycle. Women with PCOS majorly suffer from excessive weight gain, facial hair growth, acne, hair loss, skin darkening and irregular periods leading to infertility in rare cases. The existing methodologies and treatments are insufficient for early-stage detection and prediction.

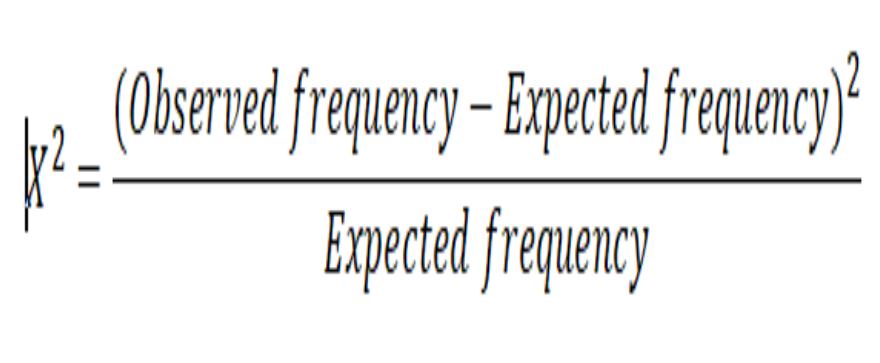
To deal with this problem, we propose a system which can help in early detection and prediction of PCOS treatment from an optimal and minimal set of parameters.

**PROPOSED METHOD:**

The dataset consisted of 41 features, and in order to select the dominant 30 features two methods were employed and accuracies compared. The two methods include:

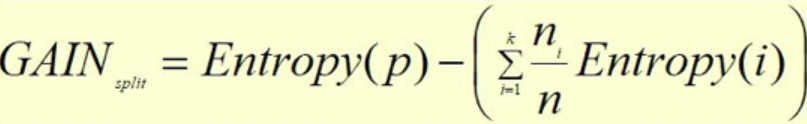
1. **Chi Square test:**

Chi-square test is used for categorical features in a dataset. We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. It determines if the association between two categorical variables of the sample would reflect their real association in the population.



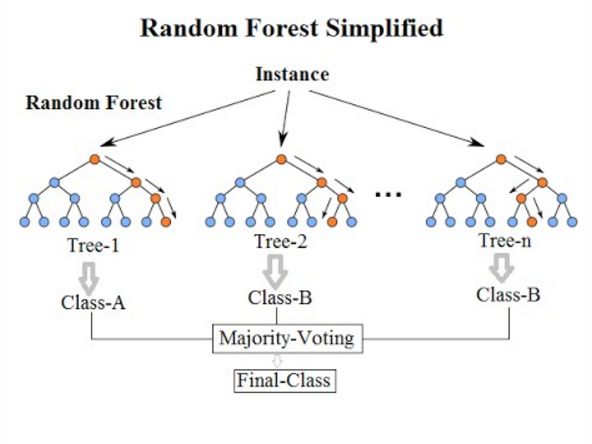
1. **Information Gain**:

Information gain can also be used for feature selection, by evaluating the gain of each variable in the context of the target variable. In this slightly different usage, the calculation is referred to as mutual information between the two random variables.

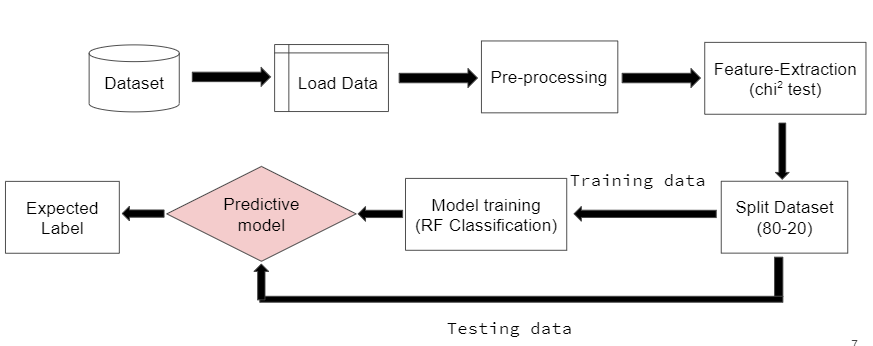
****

**Random Forest Classification:**

Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.

****

**Architecture:**

****

**Stages of Architecture:**

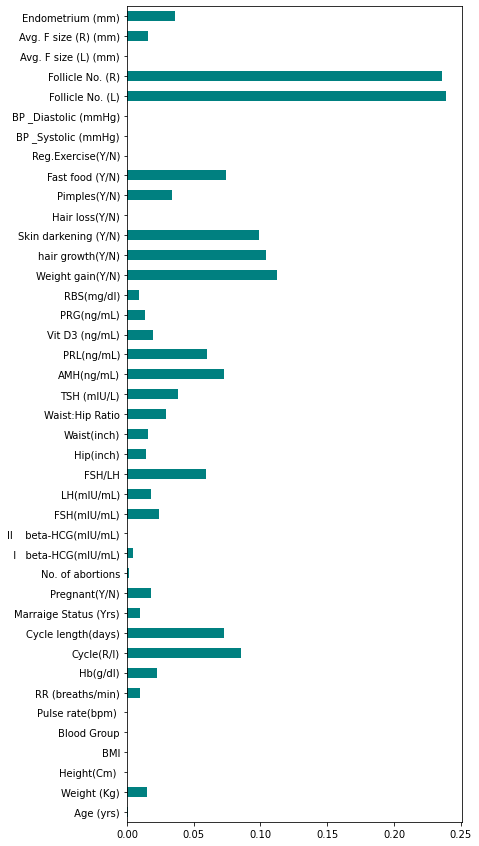
* **Dataset:** The dataset consists of 41 columns representing the features used for Early prediction of PCOS, with 531 row values.
* **Load data:** This phase of the architecture explains the way in which the data is read as a CSV file using the Pandas library in python.
* **Pre-Processing:** The null valued attributes are removed, certain column fields like Patient File No, Sl.No etc which are not required for the prediction purposes are removed. StandardScaler() numerical transform is applied on all other features after converting it to numeric type.
* **Feature Extraction:** Here two feature extraction methodologies are used and the results compared. They are Chi square test and Information Gain methodologies to select the top 30 best features.
* **Split dataset:** The dataset is split in the ratio 80:20 for training and testing.
* **Model training:** The model is trained using the training dataset and K fold cross validation is used to determine the mean accuracy score. Then the ROC graph is also plotted for visualising the performance of the model in predicting whether the patient has PCOS or not.

**FEATURE SELECTION METHODS:**

**Chi square test-features selected:**

|  |  |  |
| --- | --- | --- |
|  | **Feature** | **Score** |
|  | **PRL(ng/mL)** | **9600.594045** |
|  | **No. of abortions** | **6899.358709** |
|  | **FSH(mIU/mL)** | **2572.753526** |
|  | **II beta-HCG(mIU/mL)** | **1592.273459** |
|  | **I beta-HCG(mIU/mL)** | **1012.628577** |
|  | **Follicle No. (L)** | **673.143812** |
|  | **BP \_Diastolic (mmHg)** | **564.595226** |
|  | **TSH (mIU/L)** | **221.815736** |
|  | **LH(mIU/mL)** | **96.235867** |
|  | **hair growth(Y/N)** | **85.664994** |
|  | **Weight gain(Y/N)** | **84.038102** |
|  | **RBS(mg/dl)** | **65.013528** |
|  | **Age (yrs)** | **50.858288** |
|  | **Pimples(Y/N)** | **37.437317** |
|  | **Hb(g/dl)** | **27.793800** |
|  | **Vit D3 (ng/mL)** | **25.008276** |
|  | **Hair loss(Y/N)** | **23.562113** |
|  | **Cycle length(days)** | **19.710939** |
|  | **Height(Cm)** | **15.105580** |
|  | **Skin darkening (Y/N)** | **8.910647** |
|  | **Cycle(R/I)** | **8.230296** |
|  | **Follicle No. (R)** | **7.460844** |
|  | **FSH/LH** | **5.426396** |
|  | **Hip(inch)** | **5.219221** |
|  | **PRG(ng/mL)** | **4.779813** |
|  | **Avg. F size (L) (mm)** | **3.352904** |
|  | **Avg. F size (R) (mm)** | **3.144839** |
|  | **Pregnant(Y/N)** | **2.824165** |
|  | **Fast food (Y/N)** | **1.856357** |
|  | **Blood Group** | **1.235629** |

**Information gain -features selected :**

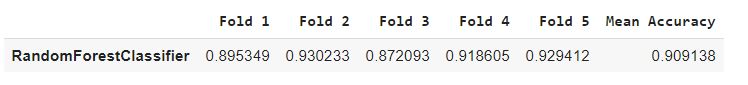
****

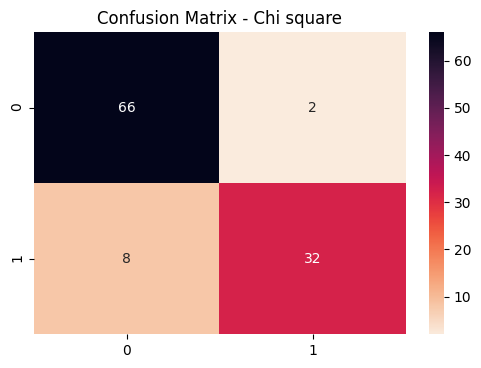
**RESULT:**

PCOS prediction was performed using Random Forest Classification. A comparison was made between two methods of feature selection: Chi Square Test and Information Gain.

**Chi Square Test**

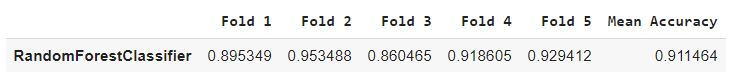
Accuracy

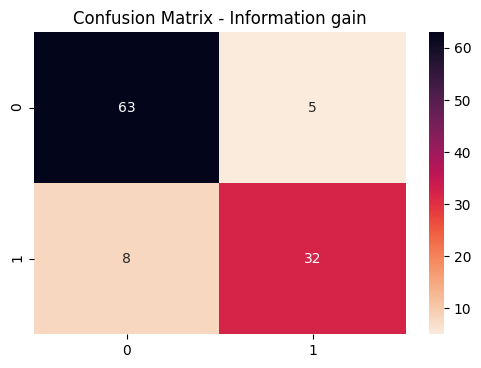
****

****

**Information Gain**

Accuracy





**INFERENCE:**

Using physical and medical parameters the percentage of risk that exists for any woman of getting affected by PCOS is predicted using Random Forest Classifier. Two feature selection methods have been used and compared here to decide upon the main factors that influence the risk of the disease. Chi square test and Information Gain for both these methods the margin of accuracy difference is very minimal.