END PROJECT REPORT ON
"PCOS DETECTION"
(Polycystic Ovary Syndrome)

Submitted for end project of Introduction to Artificial Intelligence and Machine Learning by:

Sakshi S Jangi-23ECB0B51

Supervised By:

Sri. Ravi Kishore Kodali

Associate Professor

Department of Electronics and Communication Engineering National Institute of Technology, Warangal

November 13 ,2024

ABSTRACT:

PCOS is the most common endocrine disorder in women of childbearing age, and it may be presented by irregular menstrual cycles, hyperandrogenism, and polycystic ovaries. Early detection and treatment of the syndrome are crucial to manage the syndrome and reduce all the long-term health risks. Therefore, in the proposed work, a comprehensive dataset will be utilized for developing a machine learning model to predict the possibility of PCOS. The platform that follows KNIME Analytics is used for data preprocessing, model building, and evaluation. The workflow includes PCA, SMOTE, and several algorithms, such as Decision Trees, Random Forest, Gradient Boosted Trees. Its model performance will be evaluated based on accuracy.

INTRODUCTION:

PCOS is a hormonal disorder affecting almost 10% of women worldwide. it is the one of the most cause of infertility and is also associated with obesity, insulin, type 2 diabetes and PCOS is one of the most prevalent disorders globally, yet it is often underdiagnosed, with its heterogeneity of symptoms and lack of defined standard diagnostic criteria. Machine learning can be a great tool in helping to recognize, even predict PCOS diagnosis by analysing complex data and possibly pointing to patterns not readily apparent through diagnostic means alone. The project develops a predictive model for PCOS using the KNIME Analytics Platform. This study aimed at constructing a reliable model able to predict the likelihood of PCOS among women based on clinical and lifestyle data

ABOUT THE DATSET:

Source: I obtained the dataset for this project from an open source platform,

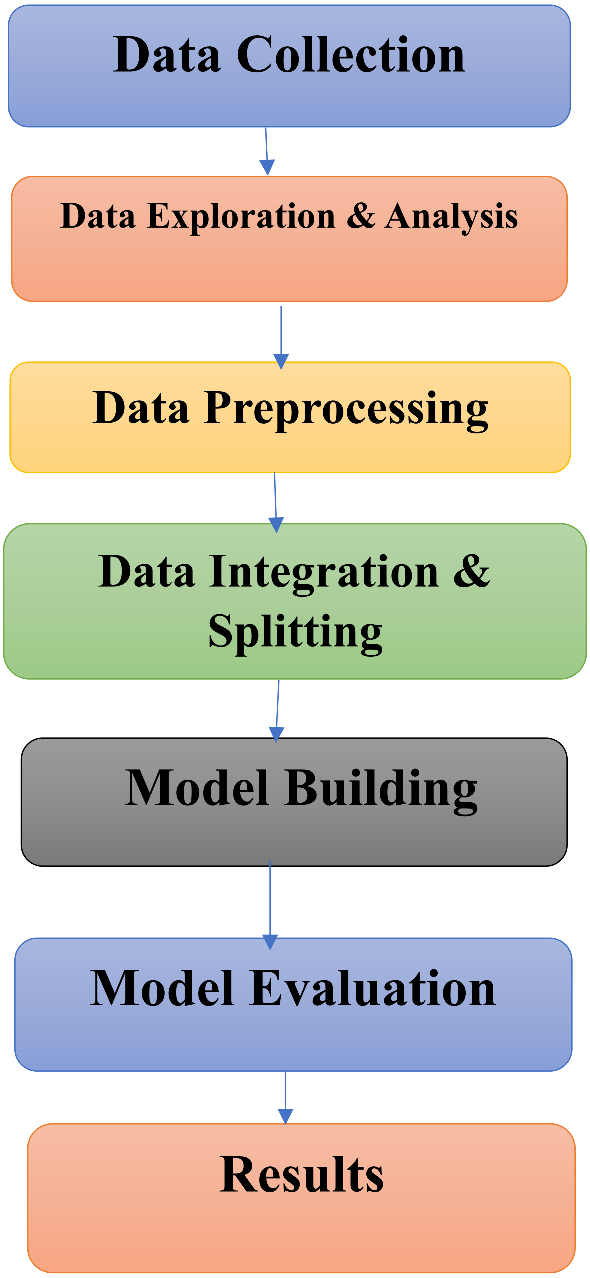
github.com

https://github.com/PCOS-Prediction/Machine-Learning

EXPLANATION:

This KNIME PCOS prediction workflow has been designed in order to analyze clinical and lifestyle data as regards the possibility of a woman having Polycystic Ovary Syndrome. The import of the given data sets is done within the workflow using the CSV Reader nodes. Then, they feed into the Concatenate node. Pre-processing diminishes missing values, irrelevant features, and duplicate values for the data set. PCA node decreases the process dimensionality, reducing the complexity of a set of data with no loss in relevant information. SMOTE doesn't fit to address imbalances in dataset but overlays synthetic samples onto the minority class. The Normalizer node scales features keeping them in common ranges; this helps improve model performance on a set. The Partitioning node splits the dataset into training and testing sets. Three algorithms were trained on the data-Decision Tree, Random Forest, and Gradient Boosted Tree-whose strengths lie in handling complex patterns. Predictor nodes provide predictions for the outcomes of the test set. To grade these predictions, scorer nodes are combined with accuracy, precision, recall, and the F1 score metrics. The ROC Curve nodes plot the Receiver Operating Characteristic curve, providing the AUC score with the capability to differentiate between cases of PCOS and non-PCOS. The Table View and Concatenate nodes gather results from all models for comparison and thus provide an overview of performance metrics. This workflow makes use of data preprocessing, model training, and assessment in the case of early PCOS diagnosis prediction and provides key predictive features.

WORKFLOW:



Data Collection

Data Exploration & Analysis

Data Preprocessing

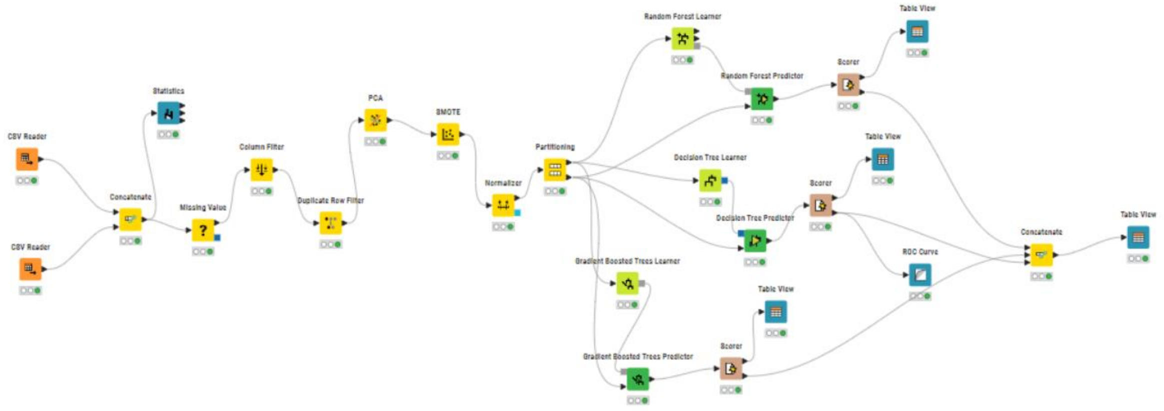
Data Integration & Splitting

Model Building

Model Evaluation

Results

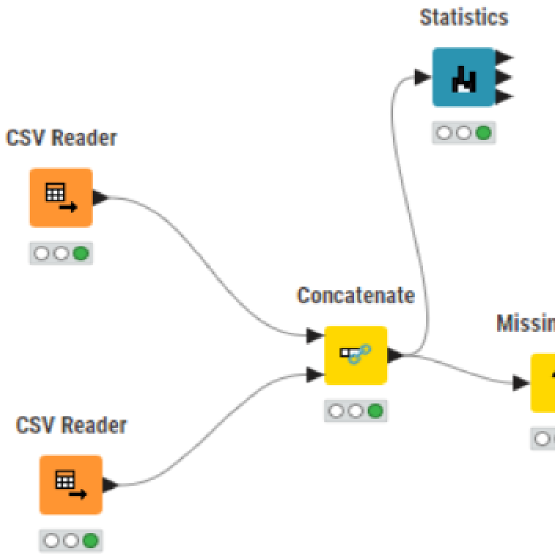
MY KNIME WORKFLOW:



WORKFLOW EXPLANATION:

Data Input:

- CSV Reader Node
- CONCATENATE

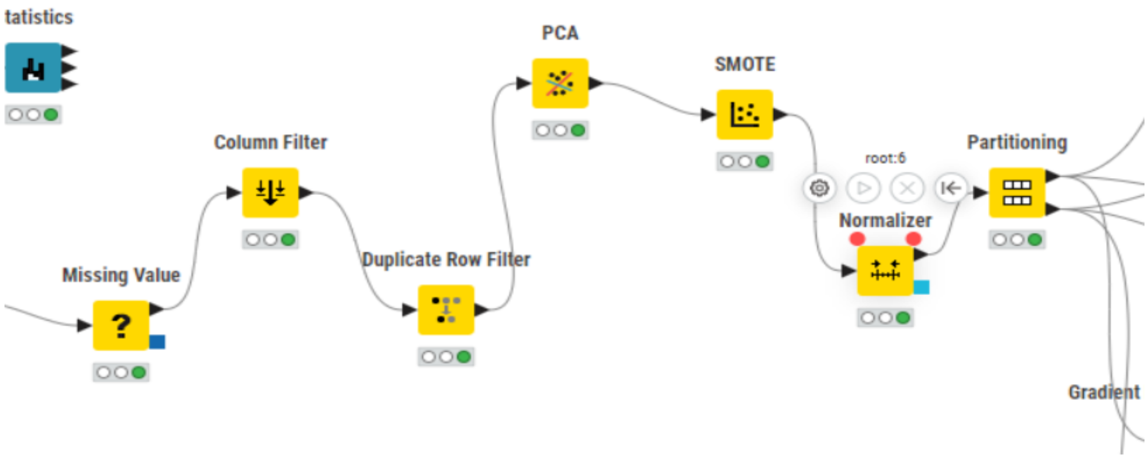


Data Preprocessing:

- **Missing Value Node**
- **Column Filter Node**
- **Duplicate Row Filter Node**
- **PCA (Principal Component Analysis)**
- **SMOTE Node**
- **Normalizer**

Data Partitioning:

- Partitioning

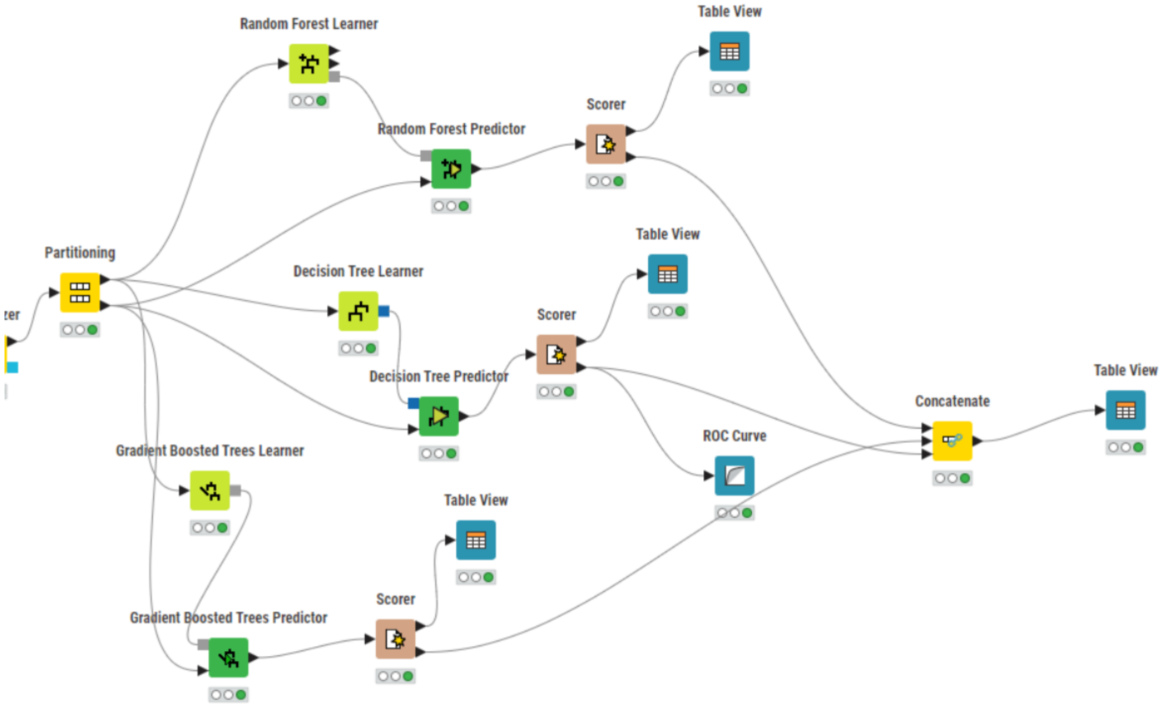


Model Building:

- **Gradient Boosted Trees**
- **Decision Tree**
- **Random Forest Learners:**

Prediction and Evaluation:

- **Predictor Nodes (Random Forest, Decision Tree, Gradient Boosted Tree)**
- **Scorer Nodes**



RESULTS:

ACCURACY OF RANDOMN FOREST PREDICTOR AND GRADIENT BOOSTED TREE PREDICTOR IS 98.587% AND THAT OF DECISION TREE PREDICTOR IS 97.8%.

The reason for the accuracy difference is that ensemble models(Random Forest and Gradient Boosted Tree) have the advantage that they combine multiple trees, capture more complex patterns, and reduce overfitting for better performance.

CONFUSION MATRIX FOR RANDOM FOREST AND GRADIENT BOOSTED MODEL:

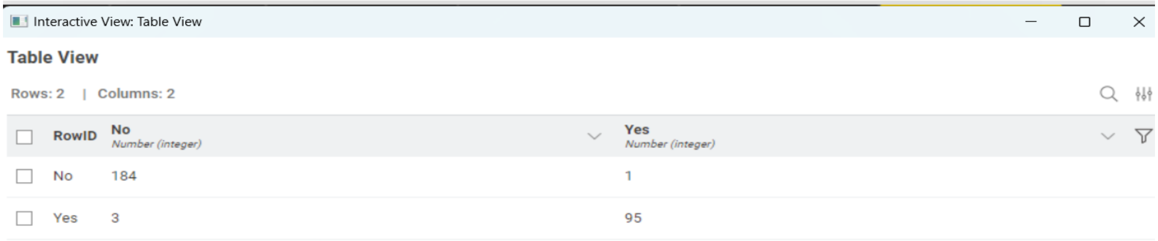


Table View

Rows: 2 | Columns: 2

RowID	No Number (integer)	Yes Number (integer)
No	184	1
Yes	3	95

CONFUSION MATRIX FOR DECISION TREE MODEL:

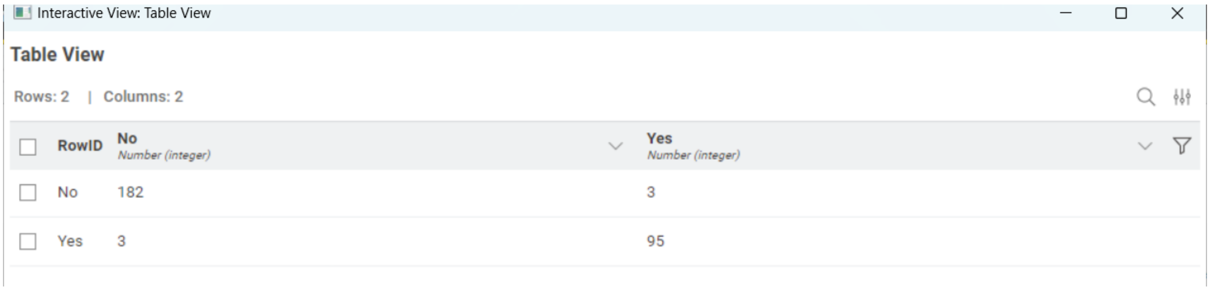


Table View

Rows: 2 | Columns: 2

□	RowID	No *Number (integer)*	Yes *Number (integer)*
□	No	182	3
□	Yes	3	95

CONCLUSION:

This KNIME workflow promises to predict PCOS with clinical and lifestyle data. The best result was achieved using Gradient Boosted Tree and Random forest.

REFRENCE:

- **Decision Tree – KNIME Community Hub**
- **Training a Random Forest – KNIME Community Hub**
- https://github.com/PCOS-Prediction/Machine-Learning