Predicting Polycystic Ovary Syndrome (PCOS): using Machine Learning on Electronic Health Records

Priyanshi Sharma   
*Business Intelligence System Infrastructure (BISI), Algonquin College – Ottawa*

*24W\_CST2101\_010: Business Intelligence Programming*

*Professor Siju Philip*

*Abstract*— This paper explores the utilization of machine learning algorithms to predict the occurrence of Polycystic Ovary Syndrome (PCOS) using electronic health records. The objective is to develop a robust predictive model that aids in early detection, personalized intervention, and optimized management of PCOS, thereby improving patient outcomes. Data preprocessing steps, including data cleaning, transformation, and feature selection, are outlined. The paper also discusses exploratory data analysis, machine learning model training, comparative analysis of model performance, and detailed evaluation of the Random Forest model's effectiveness in predicting PCOS.

# Proposed Solution/Approach

## Polycystic ovary syndrome (PCOS) is a common hormonal disorder affecting women of reproductive age. Early detection and management are crucial for improving health outcomes. Machine learning algorithms offer promise in predicting PCOS using electronic health records, thereby assisting healthcare providers in timely diagnosis and personalized intervention.This Is a Level 2 Heading

### . Early detection and management of PCOS are

### crucial for improving the overall health and quality of life of affected individuals (Anagnostis et al., 2018). In recent years, the use of machine learning algorithms has shown promise in predicting the occurrence of PCOS using electronic health records. An automated system can serve as a valuable tool for doctors, saving them considerable time in examining patients and reducing delays in diagnosing the risk of PCOS (Source: "An automated system will act as an assisted tool for the doctor for saving considerable time in examining the patients and hence reducing the delay in diagnosing the risk of PCOS"). Using machine learning algorithms on electronic health records can help healthcare providers accurately identify and manage PCOS in patients.

### 

### In addition to early detection, machine learning-basedprediction models can also play a significant role in optimizing treatment plans and monitoring the progression of PCOS in patients over time. Thesemodels can adapt to evolving patient data, providing dynamic and adaptive support for healthcare providers in delivering tailored care..

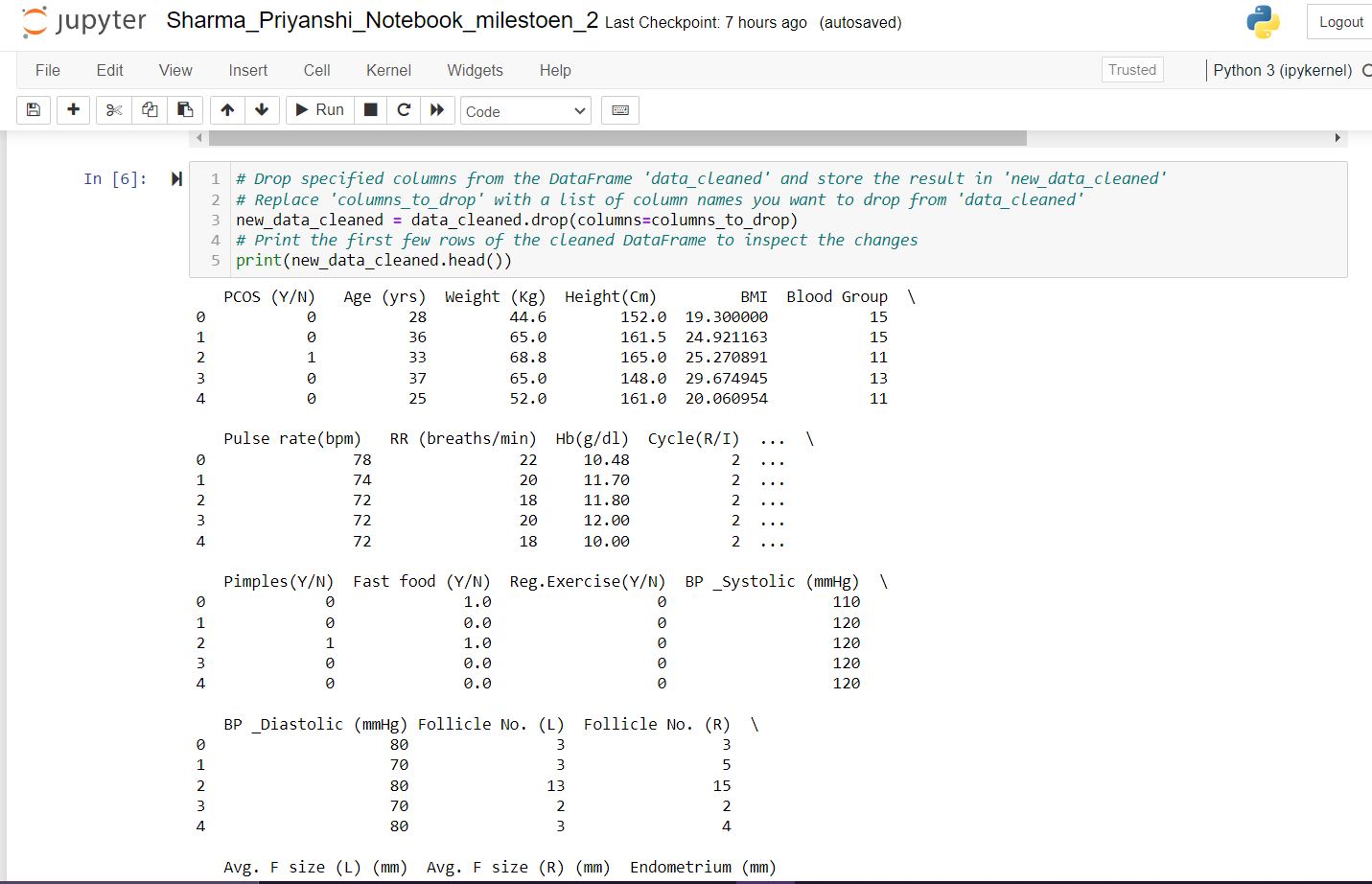
#### **Data Preprocessing Steps**

#### In order to develop an accurate predictive model for PCOS, several data preprocessing steps will be undertaken.

The dataset preprocessing entailed data importation, cleaning, and ensuring suitability for analysis. Initially, pandas was used to import the data into a DataFrame. Missing values were detected with the isnull() method and quantified by the sum() method. Rows with missing data in 'Marriage Status' and 'Fast food (Y/N)' were removed using dropna(). Irrelevant columns 'Sl. No', 'Patient File No.', and 'Unnamed: 44' were also excluded.

Further cleaning standardized values in 'Cycle(R/I)' and corrected 'II beta-HCG(mIU/mL)' from '1.99.' to '1.99' before converting it to a float type. The 'AMH(ng/mL)' column had values marked 'a', regarded as irrelevant, and thus were removed.

The resulting cleaned dataset had 538 entries and 42 columns. To confirm data integrity and correctness, the dataset was inspected thoroughly. Statistical summaries provided insights into categorical variables like 'PCOS (Y/N)', 'Cycle length(days)', and 'BP \_Systolic (mmHg)', visualized using seaborn's countplot. The fully cleaned dataset was then saved, ready for model building and further exploration.



1. This is a figure showing Data Preprocessing Steps

#### The Exploratory Data Analysis for the PCOS dataset involved the following steps:

#### Loaded "PCOS\_clean\_data\_without\_infertility.csv" into a pandas DataFrame named 'eda'; the dataset had a shape of.

#### Verified the absence of missing values in the dataset.

#### Checked for and found no duplicates in the 'Full\_new' sheet.

#### Utilized describe() to obtain statistical information on numerical variables.

#### Distinguished between numerical and categorical features within the dataframe.

#### Transformed categorical values to numeric where needed for analysis.

#### Used box plots to examine the distribution of numerical variables and identify outliers.

#### Created scatter plots to explore the relationships between numerical variables and the 'PCOS (Y/N)' target variable.

#### Analyzed the distribution of categorical variables with respect to 'PCOS (Y/N)' using bar charts.

#### Constructed a correlation matrix to assess feature relationships and their impact on the target variable.

#### Selected key features displaying a correlation coefficient greater than 0.4 with 'PCOS (Y/N)'.

#### Generated a heatmap to visually represent the correlation between the selected features.

#### Investigated relationships between chosen features and the 'PCOS (Y/N)' outcome using scatter plots and regression plots.

#### Illustrated the distribution of follicles in ovaries using scatter plots.

#### Employed swarm plots and boxen plots to observe the distribution of key features based on PCOS status.

#### Pictured the frequency of 'PCOS (Y/N)' values using a pie chart.

#### 

1. This is a figure showing correlation matrix

# *c)* Here are the key steps involved in the modeling process for PCOS prediction using the PCOSClassifier class:

# Data Loading: Load the data from the CSV file 'PCOS\_clean\_data\_without\_infertility.csv'.

* Instantiate PCOSClassifier: Create an instance of the PCOSClassifier class with the loaded data.
* Run All Steps: Execute all key steps including data preprocessing, training, evaluation, and finding the best model.
* Save Best Model: Save the best model obtained from the training process to a file using pickle.
* Plot Model Accuracies: Plot a bar graph showing the accuracies of all trained models.
* Plot Confusion Matrix: Plot the confusion matrix for the best model.
* Plot Precision-Recall Curve: Plot the precision-recall curve for the best model.
* Perform Cross-Validation: Perform k-fold cross-validation with the best model.

# 

1. This is a figure showing rints the accuracy scores of the trained models

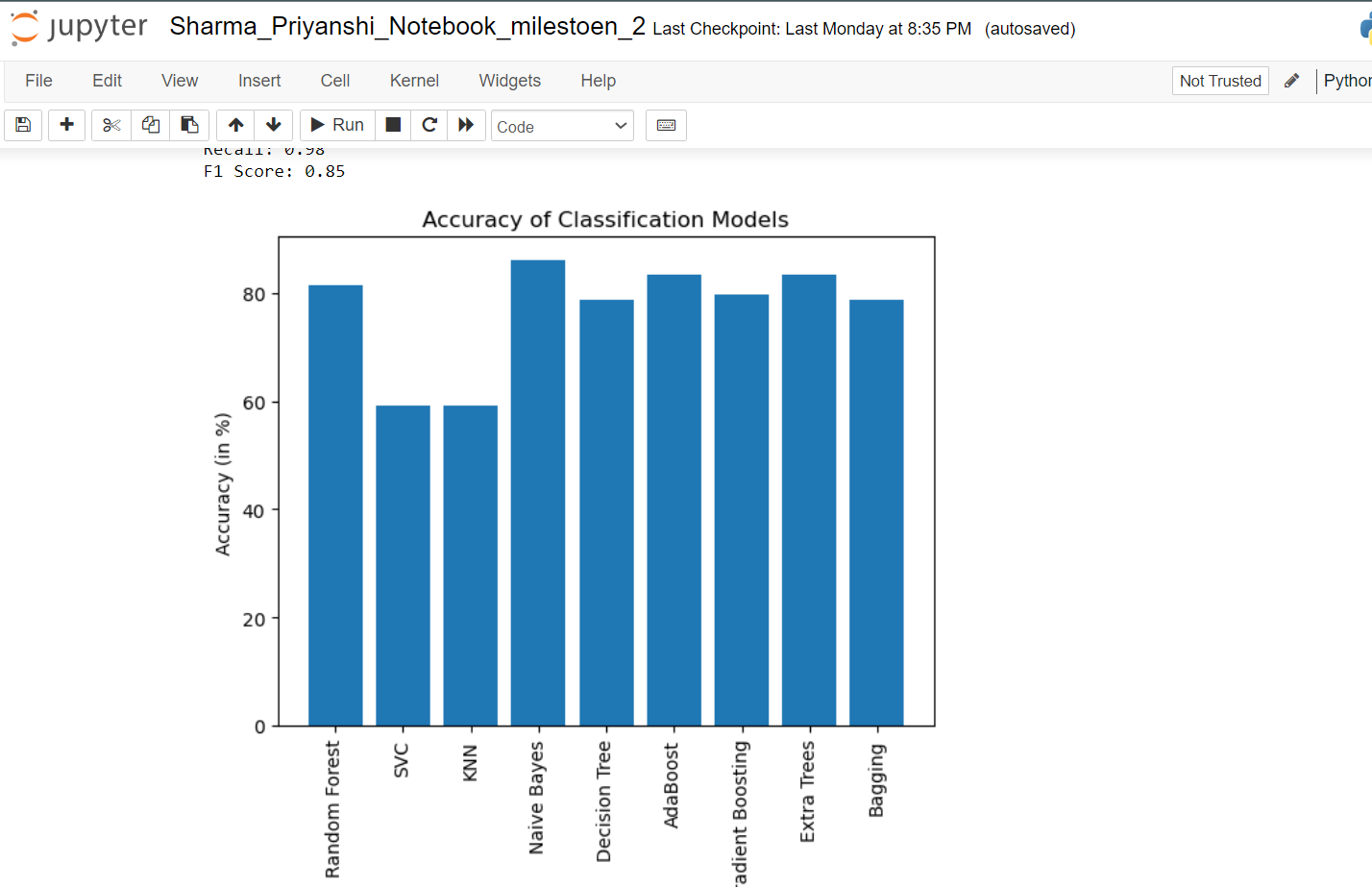


Fig. 4. This is a figure showing Plots the accuracies of the trained models.

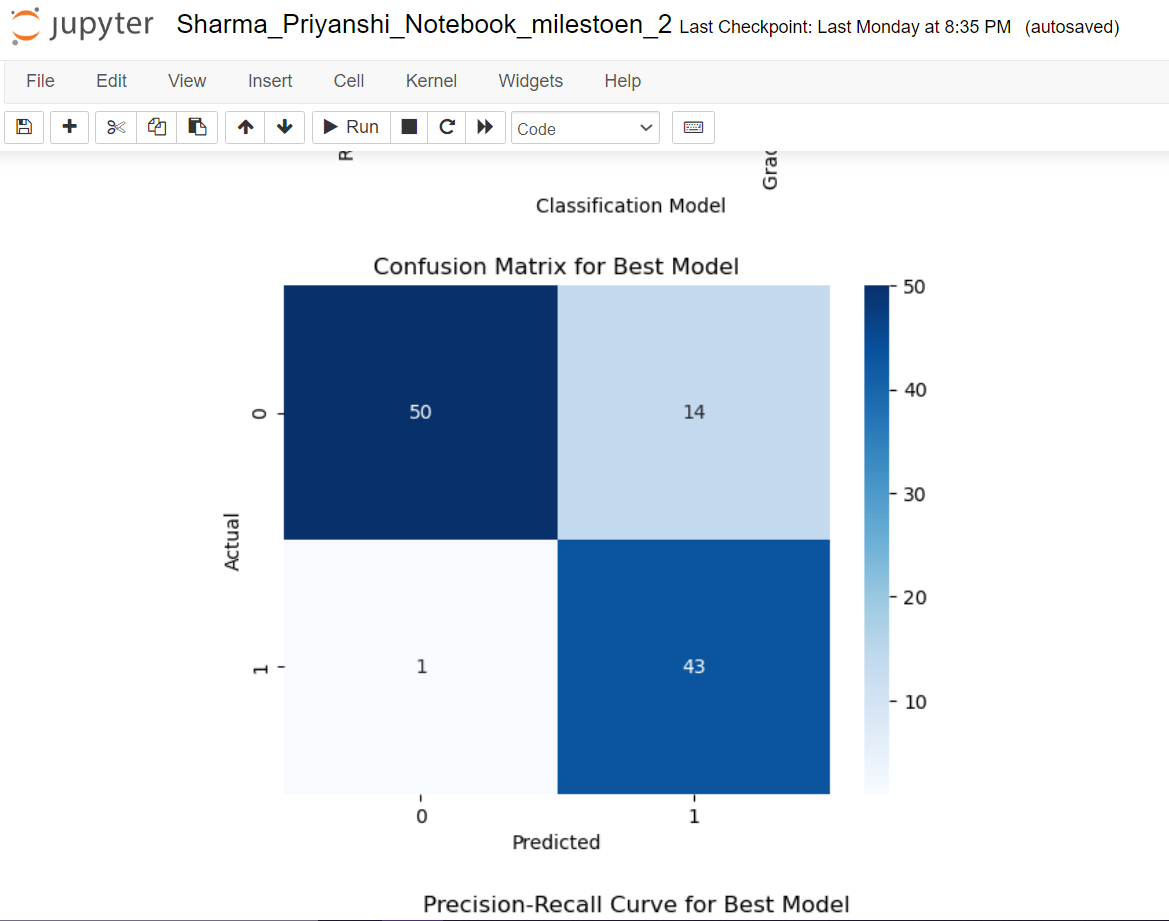


Fig. 5. This is a figure showing Plots the confusion matrix for the best model.

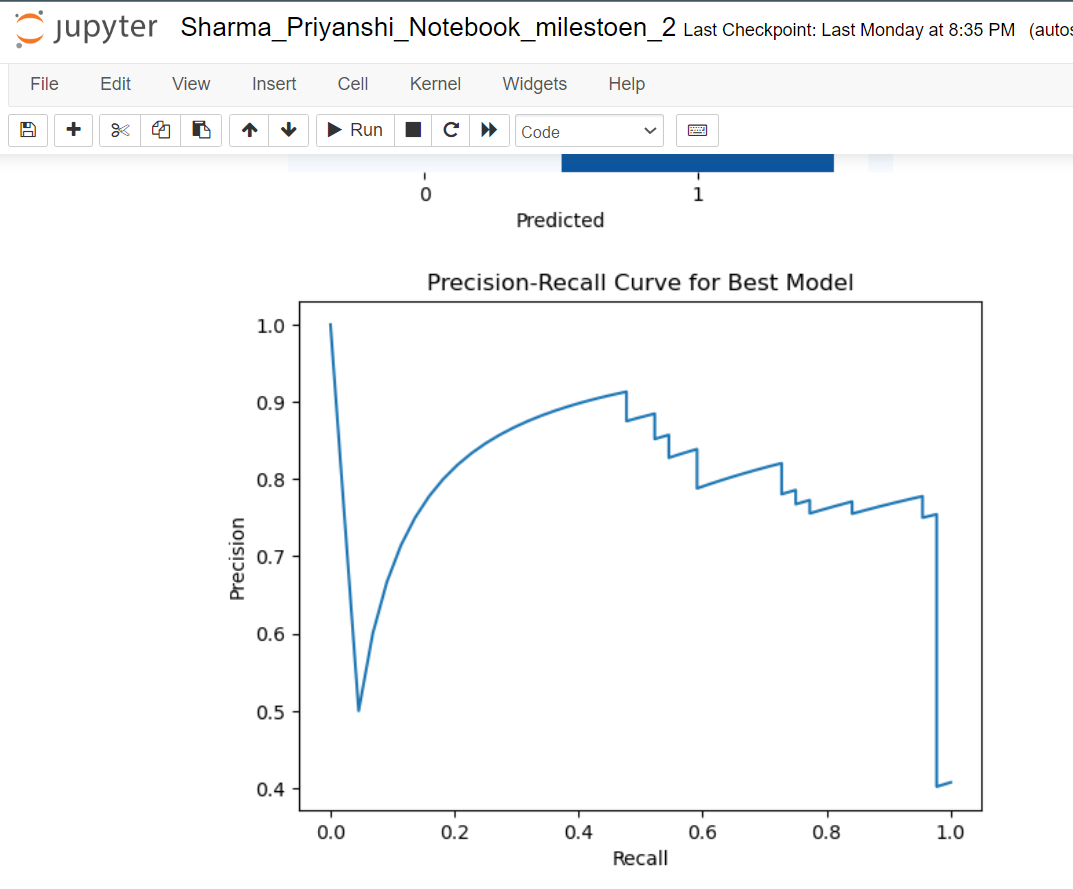


Fig. 6. This is a figure showing Plots the precision-recall curve for the best model.

# Resources Required

#### Dataset

#### PCOS\_data\_without\_infertility.xlsx (from an archive folder)

* PCOS\_infertility.csv

##### References

1. Prasoon Kottarathil. (2021). Polycystic Ovary Syndrome (PCOS)
2. [Dataset].Kaggle. <https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>
3. Mehrotra, P., Chatterjee, J., Chakraborty, C., Ghoshdastidar, B., & Ghoshdastidar, S. (2011). Automated screening of Polycystic Ovary Syndrome using machine learning techniques. <https://doi.org/10.1109/indcon.2011.6139331>