Concepts and Technologies of AI

Final Portfolio Project
An End - to - End Machine Learning Report
on Classification Task

Name: Monalika Tamang
Student ID: 2408566
Group: L5CG8

Table of Contents

Abstract

Purpose: The purpose of the report is to show the dignosis analysis of Polycystic ovary syndrome(PCOS) of top 75 countries using classification techniques.

Approach: The classification of this dataset is done following the steps: Conducting Exploratory Data Analysis(EDA), building model from scratch (sigmoid function), building two primary models (Decision tree and random forest), optimizing hyper-parameter, feature selection and rebuilding final model.

Key Results: The performance of all the models were evaluated based on the four metrices. The evaluation of first model Decision Tree demonstrated: Accuarcy: 80.48%, Precision:89%, Recall: 89% and F1-Score: 0.89. Similarly, the evaluation of second model Random Forest demonstrated: Accuracy: 89.40%, Precision: 89%, Recall:92% and F1-Score: 0.94. The final model demonstrated: Accuracy: 89.40%, Precision: 79.93%, Recall:89.40% and F1-Score: 0.84.

Conclusion: The performance of Random Forest was the best among all the other models. The key insight includes the analysis of PCOS diagnosis with the factors present in dataset.

1. Introduction:

1.1 Problem Statement:
The main aim of this task is to predict whether diagnosis of PCOS can be predicted based on the factors present on the dataset.

1.2 Dataset:
The datset used for the classificationt task is PCOS Prediction Dataset(Top 75 Countries) obtained from Kaggle. The datsaet aims to answer the diagnosis of PCOS of the top 75 countries present in the dataset based on health factors such as BMI, menstrual regularity, family history of PCOS, etc. and environmental factors such as lifestyle, urabn/rural area, etc. The dataset aligns with SDG 3: Good Health and Wellbeing.

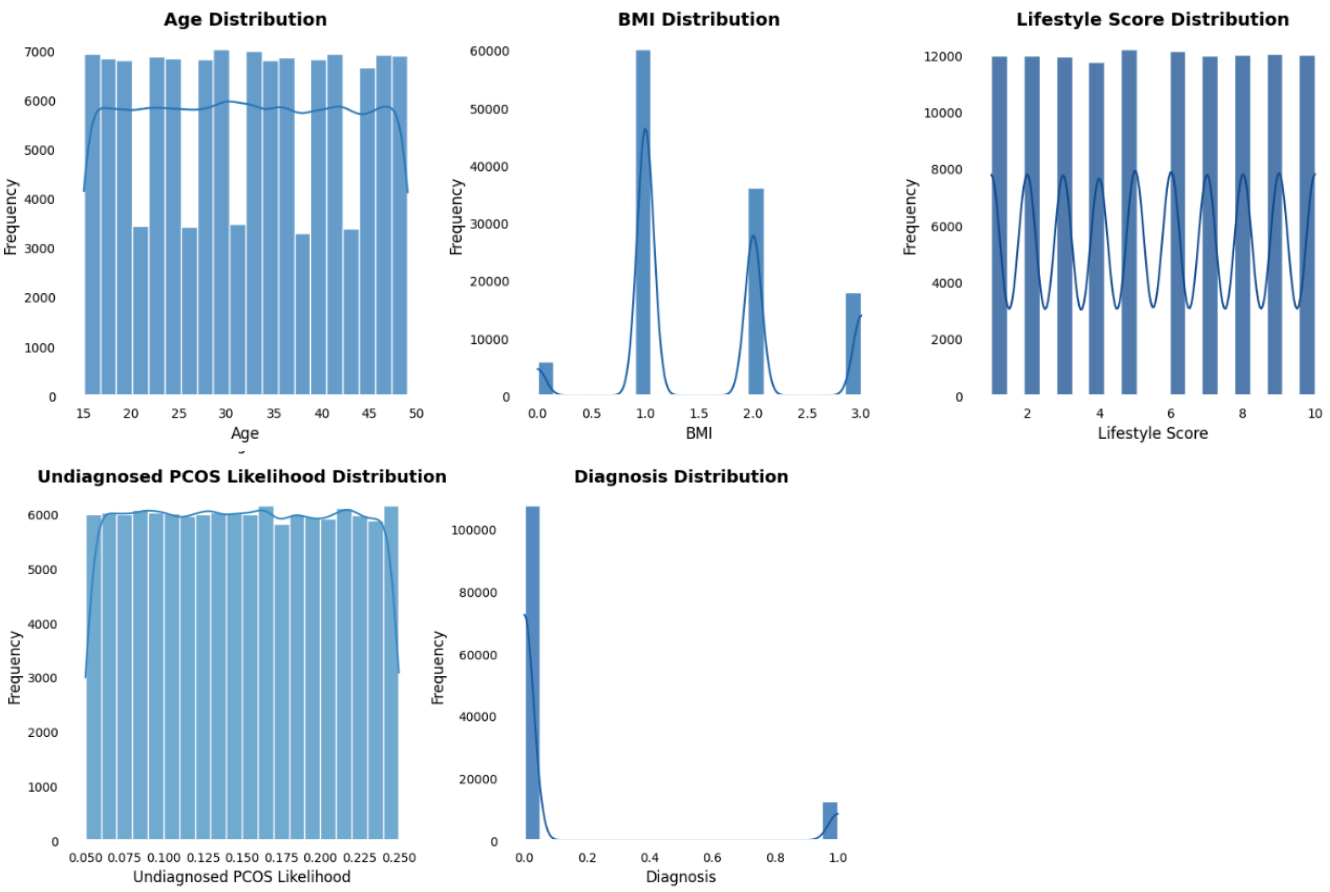
1.3 Objective:
The main objective of this task is to perform classification task by building models and predict the target variable diagnosis of PCOS on various factors of dataset.

2. Methodolgy:

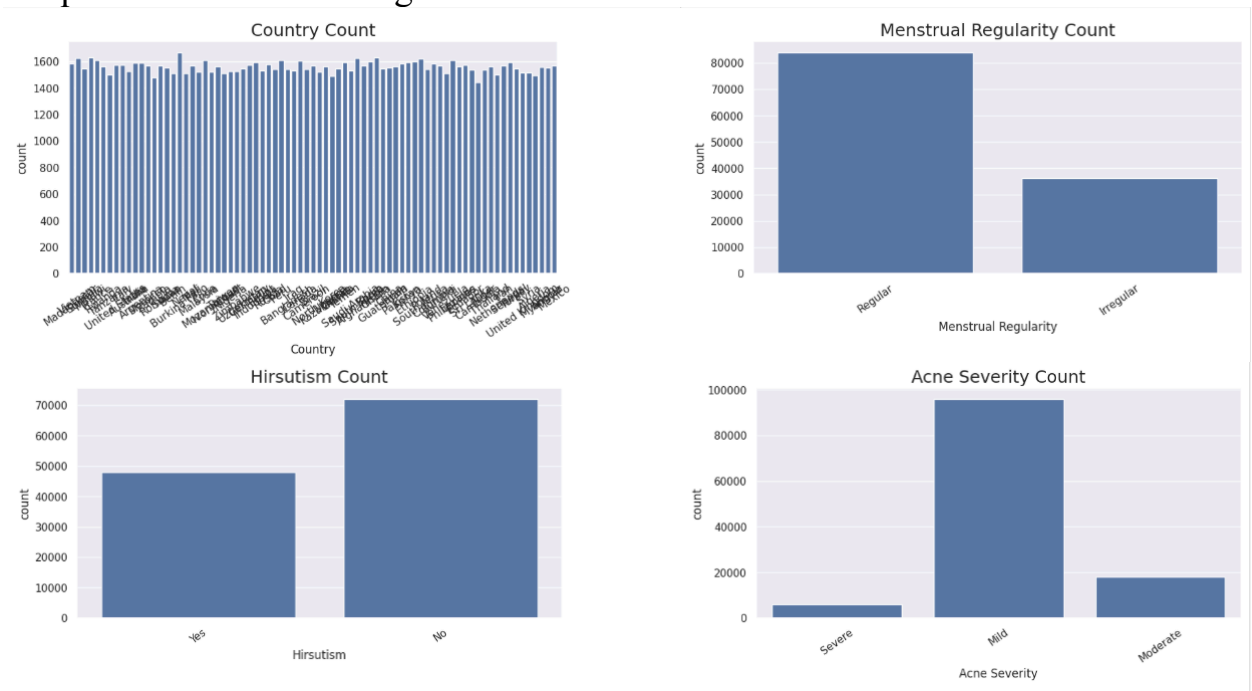
2.1 Data processing:
The first step is to clean the data by handling missing values and handling numerical and categorial columns. The missing categorial values are filled with mode.

2.2 Exploratory Data Analysis (EDA):
After the data was cleaned, it was then visualized in the form of histogram, heatmap, boxplot, scatterplot, etc. which helps to analyze and explore data much effectively.

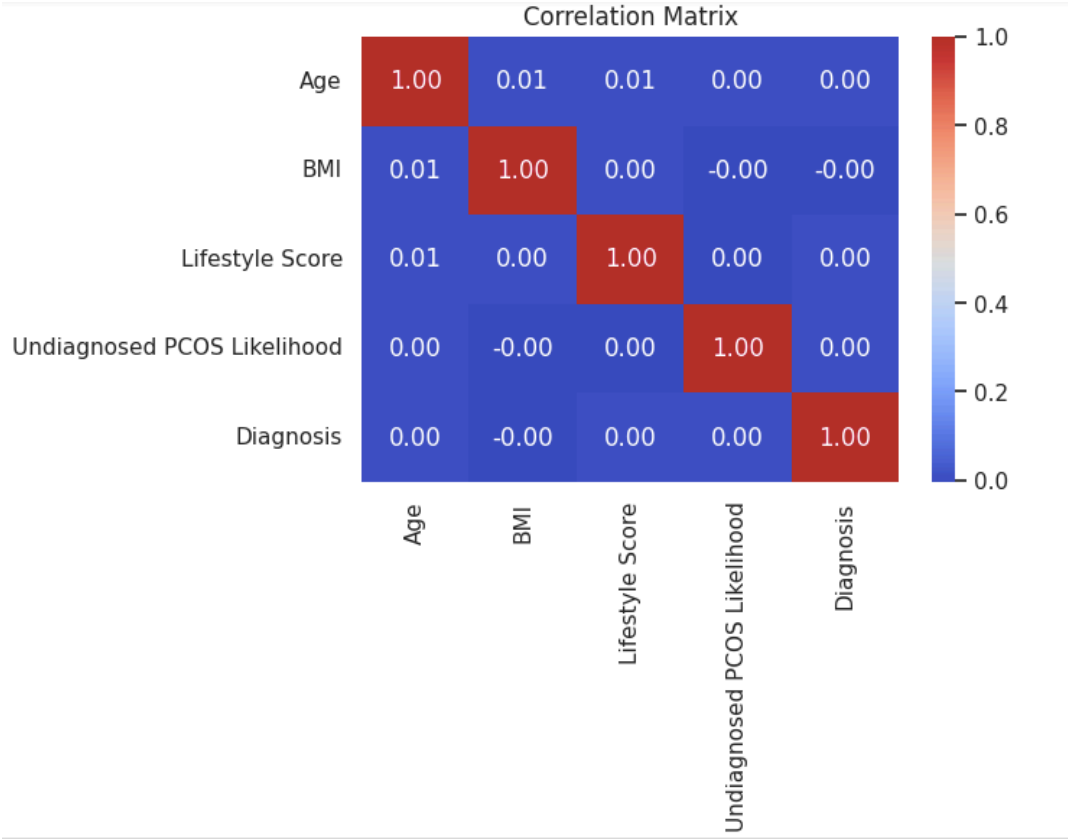
The histograms of numerical columns was plotted which demonstrates the distribution between each numerical columns and their frequencies.



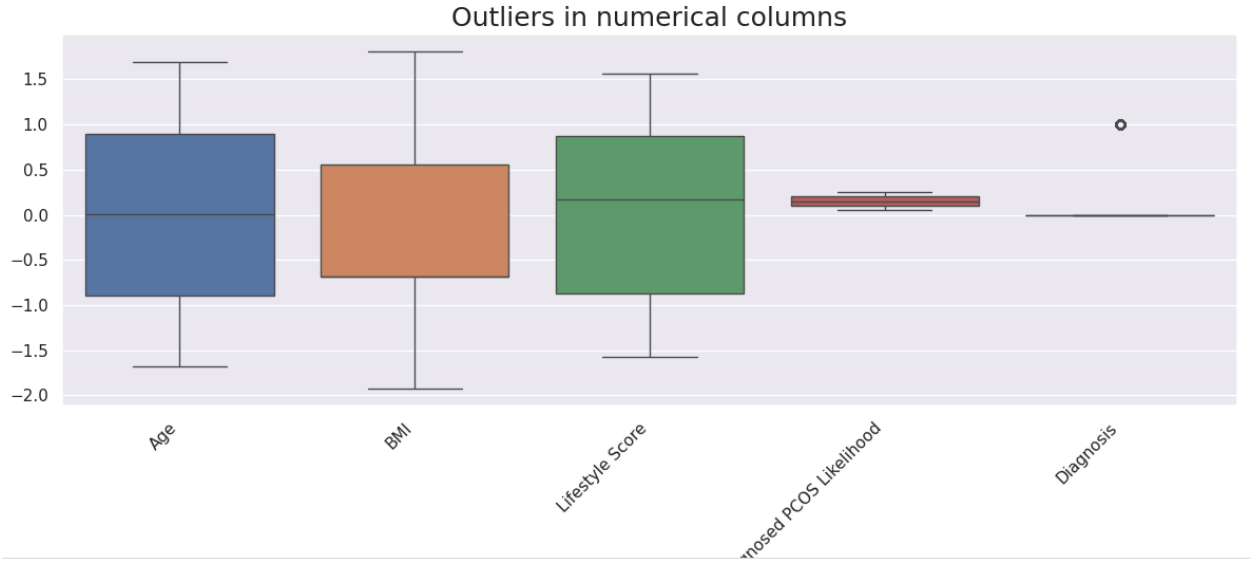
Also, several count plots of categorial columns were plotted to display the count of unique values in each categorial variable.



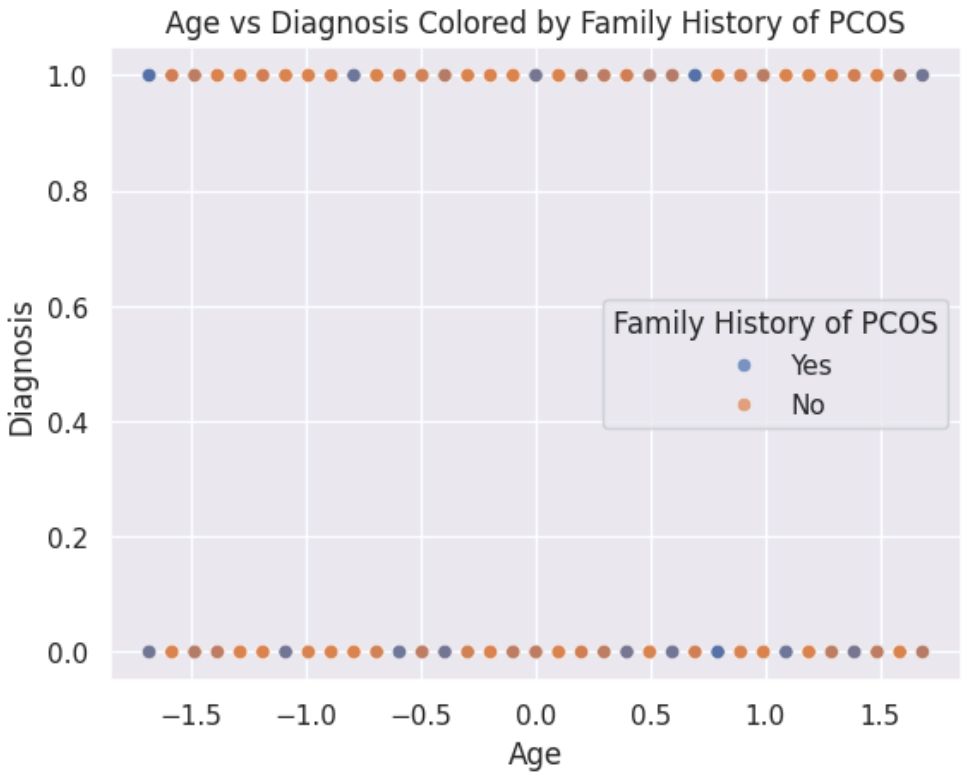
Similarly, a heatmap of correlation matrix is plotted to demonstrate the correlation coefficients between numerical variables of the dataset.



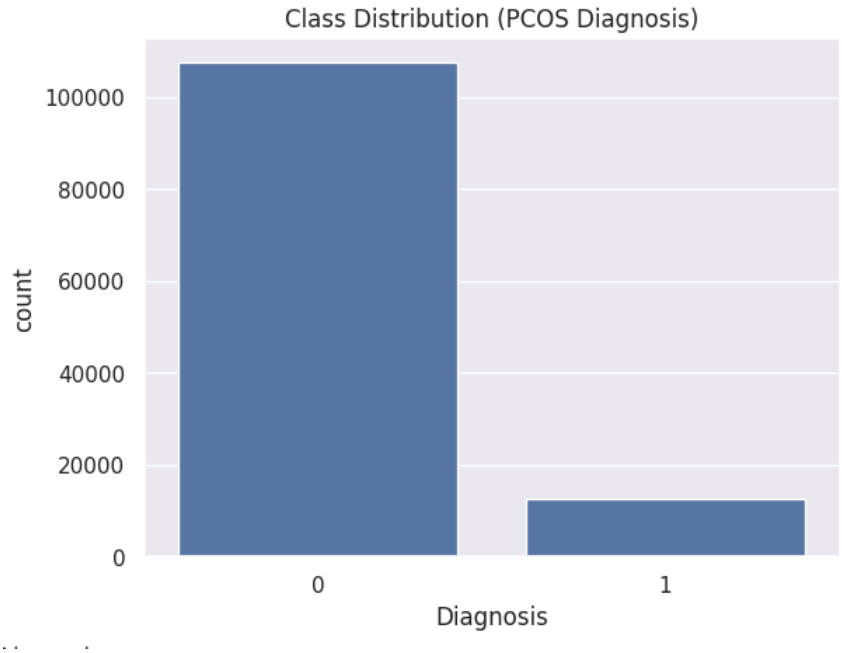
To detect outliers in numerical , boxplot of each numerical column is plotted. From the plot, it is concluded that ‘Diagnosis’ column has an outlier.



Likewise, a scatterplot was plotted to visualize the realtion between age and diagnosis by family history.



The main objective of this task is to predict diagnosis(yes/no) for PCOS. So, diagnosis is the key feature of this task.



The key insight of EDA includes the visualization of features of dataset and the key feature of dataset.

2.3 Model Building:

The models built in this task are:

1. Logistic Regression: The model is built from sratch which is then trained on sliptted train and test datas.
2. Decision Tree Classifier: The another is built with the help of sklearn and trained on training and test datas.
3. Random Forest Classifier: The last model is also built with the help of sklearn and similarly trained.

2.4 Model Evaluation:

The model evaluation is evaluated using four metrices: Accuracy, Precision, Recall and F1-Score which are imported from sklearn. The metrices are evaluated by all the models respectively. Based on the Accuracy of two models (Decision Tree Classifier and Random Forest Classifier), performance of each model is evaluated.

2.5 Hyper-parameter Optimization:

The hyper-parameter optimization of each models were conducted using GridSearchCV imported from sklearn.

- The best hyperparameters of Decision tree are: {'criterion': 'gini', 'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5}and the Cross-Validation Accuracy is: 0.8952
- The best hyperparameters for Random Forest are: {'bootstrap': True, 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 10} and the Cross-Validation Accuracy is: 0.8953

2.6 Feature Selection:

For the feature selection, SelectFromModel imported from sklearn was used to find the best 3 features. The features were selected using their indices.

- The feature indices of Decision Tree are: [5 95 96].
- The feature indices of Random Forest are:[0 2 3].

3. Conclusion

3.1 Key findings:

The model's performance used for classification tasks are evaluated by the four metrices: Accuracy, Precision, Recall and F1 Score.

- The evluation of first model Decision Tree demonstrated: Accuarcy: 80.48%, Precision:89%, Recall: 89% and F1-Score: 0.89
- The evaluation of second model Random Forest demonstrated: Accuracy: 89.40%, Precision: 89%, Recall:92% and F1-Score: 0.94

From the accuracy, it is concluded that Random Forest model performed better than Decision Tree model. Although, both the model's evaluation performance was overall good.

3.2 Final Model:

Finally, the final model was created with Random Forest as its evaluation was better using best hyperparameters and selected features. The evaluation of final model using four metrices are: Accuracy:89.40%, Precision:79.93%,

Recall:84.40% and F1-Score:84.40. The accuracy of both final model and the best model is same but there is difference in other three metrices.

3.3 Challenges:

There were many challenges while completing this task. Finding a good dataset was one of the first challenge and handling missing values, encoding categorial values and training the datas were some challenges that were encountered along the task. Similarly, to find hyper-parameters and selecting features were one of the main challenges.

3.4 Future Work:

The perfomance of models could improve more in future works. For this utilization of much more advanced classification algorithms could be very helpful. Also, the use of many other feature selection techniques and using it to build final model could definietly help to improve the model for future work.

4. Discussion

4.1 Model's Performance:

The two models Decision Tree and Random Forest performed overally good based on the four matrices. However, Random Forest Performed slightly better with the accuracy of 89.40% while the accuracy of Decision Trees was 80.48%. Similarly, the final model performed same as the Random Forest with accuracy 89.40% but increase in other metrices.

4.2 Impact of Hyperparameter Tuning and Feature Selection:

The process of hyperparameter tuning and feature selection helped to know the best hyperparameter and best features of each model which resulted in building a final model with slight improvement in other three metrices ascept the accuracy.

4.3 Interpretation of Results:

The task concludes that the diagnosis of PCOS can be predicted on the basis of

various health, environment and lifestyle factors. The diagnosed patients should get treatment for improvement of their health.

4.4 Limitations:

The dataset contained many categorial values which frequently led to over-fitting and challenges in training models resulting in affecting model's ability to generalize well.

4.5 Suggestion for Future Research:

For future task and reasearch, exploration and analysing more algorithms of classification techniques could help to improve model and perform better.