ST. TERESA’S COLLEGE (AUTONOMOUS), ERNAKULAM

AFFILIATED TO MAHATMA GANDHI UNIVERSITY

MACHINE LEARNING IN EARLY DETECTION OF POSTPARTUM DEPRESSION

PROJECT REPORT

In partial fulfilment of the requirements for the award of the degree of

BACHELOR OF SCIENCE IN
COMPUTER APPLICATIONS [TRIPLE MAIN]

Submitted By
HANNA SHAJEE
III BSc Computer Applications [Triple Main]
Register No: SB21CA012

Under the Guidance of
Ms. Megha George

DEPARTMENT OF COMPUTER APPLICATIONS
2021-2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

AFFILIATED TO MAHATMA GANDHI UNIVERSITY

MACHINE LEARNING IN EARLY DETECTION OF POSTPARTUM DEPRESSION

PROJECT REPORT

In partial fulfilment of the requirements for the award of the degree of

BACHELOR OF SCIENCE IN
COMPUTER APPLICATIONS [TRIPLE MAIN]

Submitted By
HANNA SHAJEE
III BSc Computer Applications [Triple Main]
Register No: SB21CA012

Under the Guidance of
Ms. Megha George

DEPARTMENT OF COMPUTER APPLICATIONS
2021-2024

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

AFFILIATED TO MAHATMA GANDHI UNIVERSITY

ST. TERESA'S COLLEGE
(AUTONOMOUS)
ESTD 1925

CERTIFICATE

This is to certify that the project report entitled **"Machine learning in early detection of postpartum depression"** is a bonafide record of the work done by **Hanna Shajee** (Register no: **SB21CA012)** during the year 2023-2024 and submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Applications (Triple Main) under Mahatma Gandhi University, Kottayam.

ST. TERESA'S COLLEGE
DEPT. OF COMPUTER
ERNAKULAM

Head of Department

Meghs
22/3/2024.

Internal Examiner

External Examiner

Date: 21 - 03 - 2024

DECLARATION

I, **HANNA SHAJEE** (Register no: SB21CA012), B.Sc. Computer Applications [Triple Main] final year student of St. Teresa's College (Autonomous), Ernakulum, hereby declare that the project submitted named **"MACHINE LEARNING IN EARLY DETECTION OF POSTPARTUM DEPRESSION"** for the Bachelor's Degree in Computer Applications [Triple Main] is my original work. I further declare that the said work has not previously been submitted to any other university or academic body.

PLACE: Ernakulam

DATE: 21 - 03 - 2024

Hanna

HANNA SHAJEE

ACKNOWLEDGMENT

First and foremost, I thank God Almighty for his blessings. I take this opportunity to express my gratitude to all those who helped me in completing this project successfully. I wish to express my sincere gratitude to the **Provincial Superior and Manager Rev. Dr. Sr. Vinitha CSST, Director Rev. Sr Emeline CSST and Principal Dr. Alphonsa Vijaya Joseph** for providing all the facilities.

I express my sincere gratitude towards the **Head of the department Ms. Remya C J** for the support. I deeply express my sincere thanks to my guide **Ms. Megha George** for her proper guidance and support throughout the project work. I thank all other faculty of the department for their exemplary guidance and appreciation throughout the project.

My deepest appreciation goes to all my friends and relatives for their support. I convey my hearty thanks to my parents and my sister who encouraged me and provided all the opportunities for the completion of this project.

HANNA SHAJEE

ABSTRACT

Postpartum depression (PPD) is a significant mood disorder that impacts the mental health of women following childbirth. Understanding and predicting Postpartum Depression is crucial for providing timely support and intervention to affected individuals. The goal of the study was to identify the most effective machine learning model for predicting postpartum depression. This involved a comprehensive exploration of various machine learning algorithms to determine which one yielded the most accurate and reliable predictions. The study revealed that Random Forest emerged as the best performing model for predicting postpartum depression. The discovery of Random Forest as the top- performing model signifies its potential as a valuable tool for predicting postpartum depression. Overall, our study underscores the importance of employing advanced machine learning techniques in mental health research, particularly for predicting postpartum depression. The identification of Random Forest as the best performing model opens up new avenues for improving the early detection and management of Postpartum Depression, ultimately enhancing the well-being of mothers and their infants during the postpartum period.

CONTENTS

1. INTRODUCTION

1.1 About Project

Postpartum depression, a prevalent medical concern among women following childbirth, can significantly disrupt maternal caregiving and bonding processes with their newborns. Additionally, it poses potential risks to the child's developmental trajectory and safety. Given these challenges, the timely identification of postpartum depression is paramount to safeguarding the health and well-being of both mother and child. To address this imperative, the project endeavors to utilize machine learning algorithms to predict postpartum depression accurately. Through this predictive approach, the project aims to contribute to early identification of postpartum depression, enhancing the overall health outcomes and quality of life for mothers and their infants during the postpartum period.

1.2 Objective of Project

The primary aim of the project was to determine the most effective machine learning algorithm for predicting postpartum depression. This involved a thorough exploration and comparison of various algorithms to identify the one with the highest predictive accuracy and reliability. Subsequently, the project focused on leveraging the selected top-performing algorithm to develop a predictive model tailored specifically for detecting postpartum depression. By achieving these objectives, the project aimed to enhance understanding and intervention strategies for postpartum depression using advanced machine learning techniques.

2. LITERATURE REVIEW

The study [1] evaluated five data-driven feature selection methods, including recursive feature elimination, information gain, Relief, stepwise generalized linear modeling, and a bagging-based selection-by-filter method. Nine machine learning algorithms were utilized, such as k-nearest neighbor, support vector machine, random forest, and logistic regression. The primary performance metric used was the area under the ROC curve (AUC), along with sensitivity, specificity, accuracy, precision, and F1 score. Maternal characteristics like age, education, marital status, depression history, smoking behavior, and more were analyzed for their association with postpartum depression. The study highlighted important features like exposure to stress during pregnancy, depression before pregnancy, breastfeeding duration, income, maternal education, dental hygiene before pregnancy, and the baby's gender as significant predictors of postpartum depression. Random forest achieved the highest AUC in predicting postpartum depression cases. The study emphasized the importance of feature selection methods and the performance of machine learning models in predicting postpartum depression risk factors.

The study [2] explores the association of various risk factors with postpartum depression (PPD) symptoms and employs machine learning models to predict PPD. It highlights the potential of early identification and treatment of PPD in pregnant women, emphasizing the impact on mental health and well-being during the postpartum period. The study's findings offer valuable insights into PPD prediction and the potential of machine learning techniques to improve early intervention for at-risk pregnant women, ultimately contributing to better maternal and infant health outcomes.

The paper [3] explores the use of machine learning (ML) models for predicting postpartum depression (PPD) based on data from a cohort of 508 women. The study compares different ML models and identifies key predictive factors for PPD, such as psychological resilience, third-trimester depression, and income level. It acknowledges limitations, including potential selection bias and a small sample size, and suggests further validation of the predictive models. The findings suggest the efficacy of the random forest-based filter feature selection method and the suitability of the support vector machine algorithm for predicting PPD in small sample sizes. Overall, the paper provides valuable insights into the application of ML in predicting PPD and its potential for improving healthcare outcomes for pregnant women.

The study [4] presents a novel approach utilizing machine learning to identify the risk factors and prevalence of postpartum depression (PPD) in Bangladesh. By collecting and processing data from 150 perinatal women, the study implemented various machine learning models, with the Random Forest model emerging as the best performer. The study's findings, including the identification of top risk factors for PPD and a reported prevalence of 66.7% in Bangladesh, mark a significant contribution to the field as the first of its kind to apply a machine learning approach to this issue. The implications of this work could lead to more targeted and effective interventions for PPD, ultimately improving maternal mental health in the region.

The study [5] focuses on using machine learning methods to predict postpartum depression in women. It stands out for its inclusion of a wide range of variables, including resilience and personality factors, not typically considered in similar studies. The research, conducted on a large dataset, aims to develop a screening tool for identifying high-risk women at discharge from the delivery ward. While the study offers valuable insights and potential for preventive interventions, it acknowledges limitations such as a non-representative sample. Overall, the study contributes to the field by exploring new predictive factors and emphasizing the importance of early identification and support for women at risk of postpartum depression.

The literature review in the [6] explores the application of machine learning (ML) methods for predicting postpartum depression (PPD). It highlights the advancements in ML algorithms for analyzing large datasets and improving early detection of PPD. The review acknowledges challenges in assessing ML model performance due to data variability and calls for future systematic reviews to evaluate specific ML techniques in predicting PPD. While noting research gaps and limitations, such as the lack of quality assessment for included studies and the exclusion of non-peer-reviewed literature, the review concludes that ML has the potential to significantly enhance early detection of PPD, emphasizing the need for further research to fully leverage ML techniques in maternal mental health care.

The study [7] focuses on a machine learning-based PPD prediction model using electronic health record (EHR) data from a nationwide longitudinal cohort of 214,359 births. The model, which utilized sociodemographic, clinical, and obstetric features demonstrated the potential to significantly improve the identification of women at risk of developing PPD. The study highlighted the model's ability to identify PPDs at a rate more than three times higher than the overall set at a 90th percentile risk threshold, underscoring the potential of machine learning-based models to augment existing symptom-based screening practices by identifying high-risk populations in need of preventive intervention before the onset of PPD.

The literature review of [8] provides a comprehensive overview of the use of modern machine learning (ML) approaches in identifying predictors of postpartum depression (PPD) as a means to refine patient screening and lower its impact. Through a search of PubMed and Embase, 11 relevant studies were identified, with the support vector machine being the most commonly used algorithm. The studies consistently demonstrated the feasibility of predicting PPD, particularly through variables related to sociodemographic and clinical aspects, with limited incorporation of biological variables. However, the review highlights the current scarcity of literature in this area, with diverse approaches reducing the generalizability of results. Nevertheless, it concludes that the identification of PPD risk populations using ML techniques is promising, emphasizing the need for further research to integrate such approaches into clinical practice.

3. METHODOLOGY

3.1 Data Collection

The dataset was available from Kaggle. The dataset contains 1503 records collected from a medical hospital using Google forms. The dataset includes 10 attributes out of which nine attributes were used for analysis and one was the target attribute. The target attribute, Feeling Anxious was taken as the predictor of Postpartum Depression.

3.2 Data Preprocessing

Data preprocessing is the first step and is crucial for creating machine learning model. It is a process of preparing the raw data and making it suitable for a machine learning model. The dataset which was available from Kaggle does not contain any missing data. The categorical data was then encoded into numbers using a function integer_encode. The dataset was then split into training set and testing set. The 20% of actual data was split into Test set and 80% of actual data was split into Training set. Training set was fed to five different Machine Learning algorithms to teach them to make predictions of postpartum depression. Test set is used to test the trained model and is used for predictions.

3.3 Model Selection

The chosen models encompass a diverse array of machine learning techniques, including Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM). Each of these models offers unique strengths and capabilities, contributing to a comprehensive approach in exploring and analyzing the dataset for predicting postpartum depression.

3.3.1 Random Forest

Random forest is a widely-used machine learning algorithm which combines the output of multiple decision trees to reach a single result. The algorithm’s strength lies in its ability to handle complex datasets and mitigate overfitting, making it a valuable tool for various predictive tasks in machine learning.

3.3.2 Decision tree

A Decision Tree is a non-parametric supervised learning algorithm for classification and regression tasks. It has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits.

3.3.3 Naïve Bayes

Naïve Bayes classifier is one of the simple and most effective Classification algorithm which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

3.3.4 K-NN

The K-NN algorithm is a supervised machine learning model. That means it predicts a target variable using one or multiple independent variables. When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.

3.3.5 Support Vector Machine

Support Vector Machines (SVMs) are predictive models commonly employed in both classification and regression tasks. They are renowned for their capability to identify the optimal hyperplane that maximizes the separation margin between distinct classes within the feature space. By adjusting a regularization parameter, SVMs strike a balance between maximizing this margin and minimizing classification errors, rendering them adaptable and versatile for various prediction tasks.

3.4 Model Training and Evaluation

The five distinct machine learning algorithms were trained using the designated training set. Following training, the performance of each model was meticulously assessed through a comprehensive evaluation process. This evaluation encompassed the calculation of key metrics such as Accuracy, ROC AUC score, classification report, and confusion matrix.

3.4.1 Accuracy

Accuracy is a common metric used to evaluate the performance of a model, particularly in classification tasks. Accuracy represents the proportion of correctly classified instances out of the total instances in the dataset.

Mathematically, accuracy can be expressed as:

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{Total Predictions}}$$

3.4.2 ROC AUC Score

ROC AUC (Receiver Operating Characteristic Area Under the Curve) is another important metric used in evaluating the performance of classification models, particularly binary classifiers. It measures the area under the ROC curve, which is a plot of the true positive rate (Sensitivity) against the false positive rate (1 - Specificity) for different threshold values.

The ROC AUC score ranges from 0 to 1, where:

- A score of 1 indicates perfect classification performance.
- A score of 0.5 indicates performance equivalent to random guessing.

3.4.3 Classification Report

A classification report is a summary of various evaluation metrics for each class in a classification problem. Typically, it includes metrics such as precision, recall, F1-score, and support for each class.

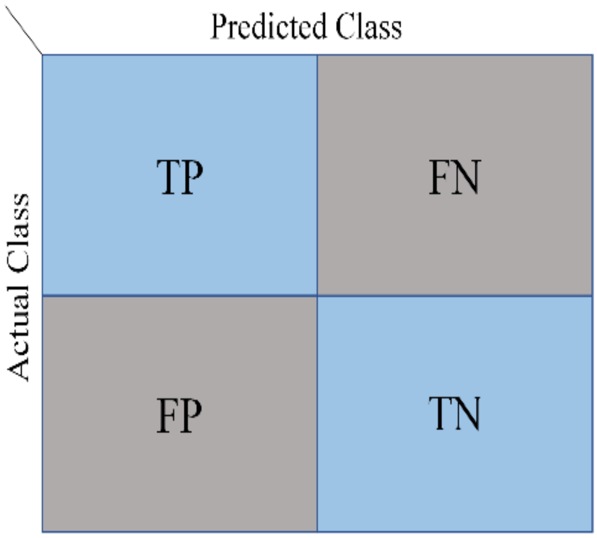
3.4.4 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows visualization of the performance of a classification algorithm by comparing actual values with predicted values.

The confusion matrix consists of four main components:

1. True Positive (TP): The number of observations correctly predicted as positive by the model.
2. True Negative (TN): The number of observations correctly predicted as negative by the model.

3. False Positive (FP): Also known as Type I error, it is the number of observations incorrectly predicted as positive by the model when they are actually negative.
4. False Negative (FN): Also known as Type II error, it is the number of observations incorrectly predicted as negative by the model when they are actually positive.



4. SOURCE CODE

```
df=pd.read_csv('data.csv')
df.head()
```

	Timestamp	Age	Feeling sad or Tearful	Irritable towards baby & partner	Trouble sleeping at night	Problems concentrating or making decision	Overeating or loss of appetite	Feeling anxious	Feeling of guilt	Problems of bonding with baby	Suicide attempt
0	6/14/2022 20:02	35-40	Yes	Yes	Two or more days a week	Yes	Yes	Yes	No	Yes	Yes
1	6/14/2022 20:03	40-45	Yes	No	No	Yes	Yes	No	Yes	Yes	No
2	6/14/2022 20:04	35-40	Yes	No	Yes	Yes	Yes	Yes	No	Sometimes	No
3	6/14/2022 20:05	35-40	Yes	Yes	Yes	Yes	No	Yes	Maybe	No	No
4	6/14/2022 20:06	40-45	Yes	No	Two or more days a week	Yes	No	Yes	No	Yes	No

```
#Calculating accuracy of Support Vector Machine
acc_svm = accuracy_score(y_test,ypred)*100
print("Accuracy Score:",acc_svm)
```

```
Accuracy Score: 91.36212624584718
```

```
print(classification_report(y_test,ypred))
```

```
              precision    recall  f1-score   support

           0       0.93      0.94      0.93       198
           1       0.89      0.85      0.87       103

    accuracy                           0.91       301
   macro avg       0.91      0.90      0.90       301
weighted avg       0.91      0.91      0.91       301
```

Accuracy score and Classification report of Support Vector Machine

```
#Calculating accuracy of Random Forest
acc_rf=accuracy_score(y_test,pred)*100
print("Accuracy Score:",acc_rf)
```

```
Accuracy Score: 99.00332225913621
```

```
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       1.00      0.98      0.99       198
           1       0.97      1.00      0.99       103

    accuracy                           0.99       301
   macro avg       0.99      0.99      0.99       301
weighted avg       0.99      0.99      0.99       301
```

Accuracy score and Classification report of Random Forest

```
#Calculating accuracy of Decision Tree
acc_dt=accuracy_score(y_test,predi)*100
print("Accuracy Score:",acc_dt)
```

```
Accuracy Score: 96.3455149501661
```

```
print(classification_report(y_test,predi))
```

```
              precision    recall  f1-score   support

           0       0.98      0.96      0.97       198
           1       0.93      0.96      0.95       103

    accuracy                           0.96       301
   macro avg       0.96      0.96      0.96       301
weighted avg       0.96      0.96      0.96       301
```

Accuracy score and Classification report
Of Decision Tree

```
#Calculating accuracy score of K-NN
acc_knn=accuracy_score(predic,y_test)*100
print("Accuracy Score:",acc_knn)
```

```
Accuracy Score: 90.69767441860465
```

```
print(classification_report(y_test,predic))
```

```
              precision    recall  f1-score   support

           0       0.93      0.93      0.93       198
           1       0.86      0.86      0.86       103

    accuracy                           0.91       301
   macro avg       0.90      0.90      0.90       301
weighted avg       0.91      0.91      0.91       301
```

Accuracy score and Classification report
of K-NN

```
#Calculating accuracy score of Naive Bayes
acc_nb=accuracy_score(predictt,y_test)*100
print("Accuracy Score:",acc_nb)
```

```
Accuracy Score: 79.734219269103
```

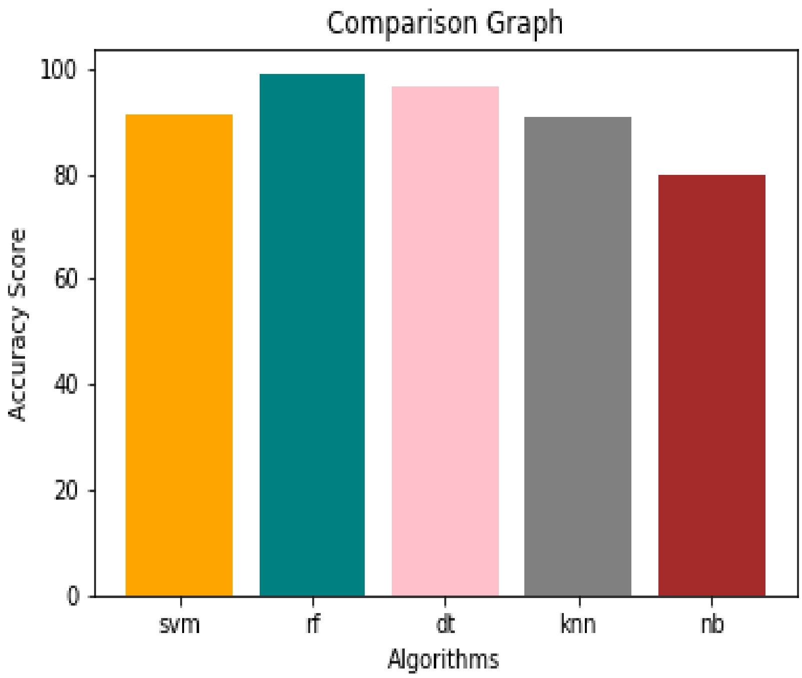
```
print(classification_report(y_test,predictt))
```

```
              precision    recall  f1-score   support

           0       0.80      0.93      0.86       198
           1       0.80      0.54      0.65       103

    accuracy                           0.80       301
   macro avg       0.80      0.74      0.75       301
weighted avg       0.80      0.80      0.79       301
```

Accuracy score and Classification report
of Naïve Bayes

5. RESULTS

Performance of the models were evaluated on the basis of metrics such as Accuracy, Precision, Recall, F1 score and ROC AUC score. The models selected were Random Forest, Decision Tree, Naïve Bayes, K-NN and Support Vector Machine. Random forest achieved the highest Accuracy of 99% followed by Decision Tree with 96%. Naïve Bayes have the least Accuracy score of 79.73%.



Comparison graph showing accuracy score of five algorithms

The area under the receiver–operating-characteristic curve (AUC) of Random forest was highest with a value of 1.0. Decision Tree achieved AUC value of 0.99 making it second best performance model. Naïve Bayes had the least AUC value of 0.82.

```
print("AUC-ROC Score SVM:", auc_svm)
print("AUC-ROC Score Random Forest:", auc_rf)
print("AUC-ROC Score Decision Tree:", auc_dt)
print("AUC-ROC Score K-NN:", auc_knn)
print("AUC-ROC Score Naive Bayes:", auc_nb)
```

```
AUC-ROC Score SVM: 0.967035175879397
AUC-ROC Score Random Forest: 1.0
AUC-ROC Score Decision Tree: 0.99
AUC-ROC Score K-NN: 0.9646733668341708
AUC-ROC Score Naive Bayes: 0.829849246231155 9
```

AUC – ROC score output of five algorithms

```
#Calculating accuracy of Random Forest
acc_rf=accuracy_score(y_test,pred)*100
print("Accuracy Score:",acc_rf)
```

```
Accuracy Score: 99.00332225913621
```

```
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       1.00      0.98      0.99       198
           1       0.97      1.00      0.99       103

    accuracy                           0.99       301
   macro avg       0.99      0.99      0.99       301
weighted avg       0.99      0.99      0.99       301
```

```
confusion_matrix(y_test,pred)
```

```
array([[195,   3],
       [  0, 103]], dtype=int64)
```

Output of accuracy, classification report and confusion matrix of random forest algorithm

In the depicted results, the Accuracy score, Classification report, and Confusion matrix pertain to the Random Forest model. Random Forest exhibits the highest Accuracy and ROC AUC score among the evaluated models. Moreover, it demonstrates the highest Sensitivity at 100% and the highest Specificity at 97%, showcasing its superior performance. Following Random Forest, the Decision Tree model achieves a Sensitivity of 97.9% and a Specificity of 93.3%. Conversely, Naive Bayes records the lowest Sensitivity at 79.6% and a Specificity of 80%.

The performance of the Random Forest model was further assessed by testing it with new data sets that were not included in the initial training or testing phases. The Random Forest model made predictions indicating that the patient is feeling anxious, suggesting the presence of postpartum depression. This determination was based on the predictive attribute of "Feeling Anxious," which was considered as an indicator of postpartum depression.

	Age	Feeling sad or Tearful	Irritable towards baby & partner	Trouble sleeping at night	Problems concentrating or making decision	Overeating or loss of appetite	Feeling of guilt	Problems of bonding with baby	Suicide attempt
0	0	1	1	1	1	1	0	2	1

```
import pickle
```

```
scaler=pickle.load(open('std_scaler.sav', 'rb'))
clf=pickle.load(open('model_rf.sav', 'rb'))
```

```
data = scaler.transform(df)
```

```
val = clf.predict(data)
```

```
if val[0] == 0:
    print('patient is feeling anxious')
else:
    print('patient is not feeling anxious')
```

```
patient is feeling anxious
```

Prediction of postpartum depression using random forest model

6. CONCLUSION

In our study, we conducted a thorough assessment of the performance of five distinct Machine Learning algorithms, each representing different methodologies and approaches to predictive modelling. These algorithms included Random Forest, Decision Tree, Naïve Bayes, K-NN, and Support Vector Machine. Our evaluation process involved the use of several key metrics to gauge the effectiveness of each algorithm in predicting postpartum depression. These metrics encompassed a range of criteria, including Accuracy, Precision, Recall, F1 score, and ROC AUC score.

Upon analyzing the results, we found that Random Forest emerged as the standout performer, demonstrating exceptional performance across multiple metrics. Notably, it achieved the highest accuracy score of 99% and the highest ROC AUC score of 1.0. This indicates that Random Forest excelled in both accurately classifying instances and distinguishing between positive and negative cases of postpartum depression.

Given its impressive performance, Random Forest model was further used to predict Postpartum Depression. By leveraging the predictive capabilities of this model, we aimed to identify individuals who may be at risk of experiencing postpartum depression, thereby facilitating early intervention and support.

Thus the findings presented suggests that Machine Learning approaches might constitute a valid choice for the identification of patients at risk of Postpartum Depression.

7. REFERENCES

[1] Shin, Dayeon, et al. "Machine learning-based predictive modeling of postpartum depression." Journal of Clinical Medicine 9.9 (2020): 2899.

[2] Gopalakrishnan, Abinaya, et al. "Predicting Women with Postpartum Depression Symptoms Using Machine Learning Techniques." Mathematics 10.23 (2022): 4570.

[3] Zhang, Weina, et al. "Machine learning models for the prediction of postpartumdepression: application and comparison based on a cohort study." JMIR medicalinformatics 8.4 (2020): e15516.

[4] Raisa, Jasiya Fairiz, M. Shamim Kaiser, and Mufti Mahmud. "A machine learning approach for early detection of postpartum depression in Bangladesh." International Conference on Brain Informatics. Cham: Springer International Publishing, 2022.

[5] Andersson, Sam, et al. "Predicting women with depressive symptoms postpartum with machine learning methods." Scientific reports 11.1 (2021): 7877.

[6] Saqib, Kiran, Amber Fozia Khan, and Zahid Ahmad Butt. "Machine learning methods for predicting postpartum depression: scoping review." JMIR mental health 8.11 (2021): e29838.

[7] Hochman, Eldar, et al. "Development and validation of a machine learning-based postpartum depression prediction model: A nationwide cohort study." Depression and anxiety 38.4 (2021): 400-411.

[8] Cellini, Paolo, et al. "Machine learning in the prediction of postpartum depression: A review." Journal of Affective Disorders 309 (2022): 350- 357.