Predictive Modelling for Early Detection of Postpartum Depression

Praveen Yeduresi, *E.Mail ID:-py23736@essex.ac.uk and Reg.No:-2321428*

Abstract—Postpartum depression (PPD) is a type of depression that many women get affected after having a baby. According to the NHS, it's a common problem, affecting more than 1 in every 10 women within a year of giving birth. The early prediction of postpartum depression can help identify women at risk and can help many women receive adequate care. In this paper, the potential of Data science techniques was used to analyze the data, extract insights from the data, and develop a predictive model that leverages the selected attributes to estimate the likelihood of postpartum depression. Firstly, the data set has been loaded and explored to check the data properties and respective data challenges. The second stage comprises data preprocessing and data cleaning to solve the challenges in the data. Third, conducted exploratory data analysis (EDA) aimed to contribute valuable insights for early detection and finally implemented a predictive modeling approach which includes training and evaluating various machine learning classification algorithms namely SVM, Random Forest, KNN, XGBoost, Neural Networks, Decision trees and logistic regression. Among all Random Forest and XGBoost models turned to have decent results with F-score of 0.99 for feeling anxious class and 0.97 for Not feeling anxious class. Thus, these models can be incorporated in healthcare to estimate the likelihood of postpartum depression to achieve the objective of identifying women at risk.

Index Terms—Postpartum depression, Data Preprocessing, Exploratory Data Analysis, Machine Learning

1 Introduction

Postpartum depression (PPD) is a health condition that women encounter shortly after giving birth. It can be minor or major depending on the symptoms and it is one of the more common depressions diagnosed in new parents. According to the World Health Organization. Around 13 percentage of women worldwide suffer from postpartum depression. They claim, this illness causes anxiety and, in many cases, even prompts women to commit suicide. It is significantly greater in developing nations i.e., 19.8 percentage. It indicates a major treat to the mother's health [2]. There are no underlying biological variables that could be identified as the cause of PPD and also ongoing rapid changes in food habits, sleep habits, and many hormonal changes are becoming common in women after childbirth. As a result, it is becoming a challenge to diagnose PPD in the early stages [3].

There has been wide research in the field of postpartum depression to develop a predictive model that leverages the available data to estimate the likelihood of PPD. Due to the rapid growth of digitalization, a lot of information during pregnancy and in the post-partum period has been collected and recorded by medical agencies. Several studies and research have determined Data science techniques especially Machine learning models could be a useful approach to accurately classify and interpret crucial complicated data, circumventing the constraints of traditional statistical viewpoints [4]. Furthermore, studies have underlined the significance of early detection and intervention can identify women at risk and can help many women to seek health care earlier before complications arise.

This paper aims to unearth reasons behind PPD using data analytics and build a model that predicts the likelihood of PPD from an online questionnaire administrated data using machine learning. The motivation for this research is to decrease the number of women who suffer from PPD by early detecting and make lives better by leveraging the potential of data science, especially machine learning techniques. This research contributes to the medical field directly by helping doctors to know the technical and logical reasons for the occurrence of PPD and helping them to identify the women who may undergo PPD without any manual interventions.

2 Literature Review

In the healthcare sector, research has been ongoing into the early detection of PPD using data science techniques. For instance, a study [5] established a complete data science framework that includes data preprocessing, data imputation and evaluation of wide range of Machine learning (ML) supervised classification algorithms to predict pregnant women at risk of PPD using basic population-based prospective cohort study in Sweden. Among all the models, extremely randomized random forests performed well. Also, this study demonstrated positive results of using ML and highlighted the capacity of machine learning algorithms in detecting PPD.

The advancements in the machine learning algorithms always bring great value in detecting PPD. For instance, an estimation of PPD from electronic health records using ML has been carried out [6]. According to their study, by applying a gradient tree boosting machine learning algorithm on a data cohort of over 265,544 women who gave first birth, achieved a fair prediction of PDP with a performance of AUC varied from 0.72 to 0.74. Data includes socio-demographic and medical variables. This has provided great value to existing screening tools. Based on these studies, utilizing the power of machine learning algorithms

like gradient tree boosting can be used to develop predictive models to estimate the likelihood of PDP in the early stages.

Another study conducted by [7] used Machine learning capacity to predict women with PPD. Their framework comprises four stages. Firstly, Data has been collected from the Edinburgh Postnatal Depression Scale (EPDS) questionnaire survey. The second phase of data management included data preprocessing and attribute selection. Next to it a multistage assessment where actual training and evaluation of the model has been taken place and finally a prediction phase where predictions of women with PPD have occurred. This study trained a wide range of supervised classification models and conducted a full comparative study to select the best model. The extreme random forest model offered the most efficient predictions. They claimed using this model on given data pregnant women who are likely to be affected by PPD could be identified and taken care of.

Overall previous works focusing on developing predictive modelling to estimate the likelihood of PPD have demonstrated good potential for usage of data science techniques. Based on all these studies, this paper also follows a similar data science framework which is explained in later sections.

3 METHODOLOGY

This section explains the framework used by this paper to estimate the likelihood of postpartum depression. It comprises data description, data exploration, data preparation, exploratory data analysis and finally modeling stages. The technology stack used for the overall project is Python.

3.1 Data Description

A team has collected a dataset from a medical hospital that includes 1503 individual records for this research. The information was obtained using a survey distributed via an online application. It contains ten variables. Nine of these features act as predictors(or) independent variables, and one feature, "Feeling Anxious," has been identified to be the target variable (or) dependent variable.

The features and unique categories in the data are shown in Table 1

Feature	Unique Categories
Age	25-30, 30-35, 35-40, 40-45 and 45-50
Feeling sad or Tearful	Yes , No and Sometimes
Irritable towards Baby and Partner	Yes , No and Sometimes
Trouble sleeping at night	Yes , No and Two or more days a week
Problems concentrating or making decision	Yes , No and Often
Overeating or loss of appetite	Yes , No and Not at all
Feeling of guilt	Yes , No and May be
Problems of bonding with baby	Yes , No and Sometimes
Suicide attempt	Unique article Yes , No and Not interested to say
Feeling anxious	Yes and No

TABLE 1
Data unique categories

Up on checking the above table, data contains age and behavioral attributes of individuals to estimate the likelihood of PPD.

3.2 Data Exploration

The main objective of the data exploration stage is to gain initial insights from the data, understand data characteristics, and investigate challenges in the data. Firstly, it found that the entire dataset contains only categorical variables, henceforth it is not suitable to use traditional descriptive statistics like mean, median, standard deviation, and quartiles. Alternatively, this paper examines the frequency distribution and description of each category within the variables, and this gives a few insights into each variable category' such as count, top variable, and respective frequencies of the categories. For instance, the target variable 'Feeling anxious' has two categories 'Yes' and 'No', and yes is the most frequently occurring category with 980 times in overall 1503 data points. Upon understanding the data properties and objective, it is classified as a supervised binary classification problem. As this paper aims to run wide range of classification algorithms to find the best model to estimate the likelihood of PPD, there is a necessity for converting categorical data to numerical using respective encoding techniques.

Secondly, by looking at the count of each variable, variables named 'Irritable towards baby and partner', 'Problems concentrating or making decisions', and 'Feeling of guilt' had some missing values. In other words, data for those variables are missing, indicating that the dataset is incomplete.

Henceforth the findings of this stage conclude data had few challenges of missing values and categorical variables and treating missing values and encoding categorical data with respective methods are essential to be performed in later stages. This study will use this stage findings in the stages of data preparation and modeling efforts.

3.3 Data Preparation

During the data exploration stage, this paper identified missing values in three variables. This section of data preparation aimed to treat missing values in three affected variables with mode imputation. Given the categorical data, Mode imputation will act as a simple technique to deal with missing values. It involves replacing the missing values in the data with the non-missing value that occurs the most frequently. The outcome of this stage will be complete data with no missing values and data will be ready for analysis.

3.4 Exploratory Data Analysis

Exploratory data analysis is a crucial step, and it aims to perform univariant and bivariant analysis on the data to extract insights that explain patterns inside data, unearth reasons for individuals feeling anxious, and check which models and evaluation metrics can be used in modeling section and to take data decisions. It is important to make clear that all insights presented in this section are related to data used for this paper and insights may not be correct for different datasets used to estimate the likelihood of PPD.

The univariant analysis includes analyzing the distribution of each variable in the dataset individually to understand its distribution. On observation of the Age variable, data implies most individuals are in the 40-45 age group and fewer individuals are in the 25-30 age group. From the distribution of feeling sad (or) tearful variable, this paper found an equal portion of individuals reported both. By observing irritable towards baby and partner variable, this paper infers that, many individuals reported irritation and that irritation is every time (or) sometimes. When investigating the trouble sleeping at night variable, it is evident that a majority number of individuals reported having problems sleeping two (or) more days a week. By looking into the problems concentrating or making decisions, a significant majority of individuals have not encountered such problems. The pattern from overeating (or) loss of appetite variable infers Most individuals indicated problems that might be every time (or) sometimes. Observing the guilt feeling column, we can say that minority respondents reported as guilty. Many individuals reported they do not have problems bonding with the baby and the majority have no intention of committing suicide.The target variable, ”Feeling anxious,” was found to be unbalanced during the analysis, with most of the records reporting feeling anxious. It is essential to address this class imbalance during model training (or) choosing respective evaluation metrics such as precision, recall, and F-score as these are sustainable to class imbalances.

The bivariant analysis includes analyzing two variables in the dataset at a time to understand their patterns and understand their potential associations. According to the analysis, individuals aged above 30 are reporting feeling anxious comparatively than those age below 30. Individuals who reported irritability towards baby and partner, problems bonding with baby, and individuals who are having trouble sleeping at night more often are feeling highly anxious.Conversely, Individuals who reported feeling sad or tearful are not experiencing much anxious. Also, according to the data, individuals who reported overeating (or) losing appetite, guilt, and attempting suicide does not experience any anxiety.

In conclusion, the Exploratory data analysis stage gives valuable findings into distributions and relationships between variables inside the dataset. It revealed target variable ‘feeling anxious is imbalanced. Also, unearth the reasons such as individuals who reported irritability towards baby and partner, problems bonding with baby, having trouble sleeping at night more likely to feel anxious. Understanding these patterns is important for choosing models and metrics for further stages of the research, including modeling.

3.5 Modelling

The modeling stage is the crucial stage that comprises generation of train and test datasets, encoding categorical data and finally training and evaluation of the predictive models.

3.5.1 Data preparation for modelling

The dataset was separated into independent (which are predictors) and dependent (which is target) as part of data preparation for modeling. These independent and dependent variables are further divided into 80 percent of the training set and 20 percent of the testing set. This is often used to train selected machine learning models and evaluate their performance. Also, this paper wants a wide range of binary classification machine learning algorithms to find the best model to estimate the likelihood of PPD. Henceforth there is a necessity of converting categorical variables to numerical representations. Category encoding techniques were used to do this conversion

3.5.2 Model Training using Logistic regression

This step involves training suitable machine learning models using prepared training data. A wide range of supervised binary classification algorithms have been chosen and each algorithm has been trained on the prepared training data.

The first model chosen for training is Logistic regression. It is a widely used Machine learning algorithm for classification problems.The objective of this model is to predict the probability of feeling anxious based on given predictor variables.logistic regression model uses the sigmoid function to find the likelihood of feeling anxiety, which was modeled as a function of the predictor variables. Using the sigmoid function, any real number can be mapped into a range between 0 and 1, which indicates the likelihood of feeling anxiety.

The prepared training data was used to train the logistic regression model where Maximum likelihood estimation was used to estimate the model’s coefficients in the model. The logistic loss function, which calculates the error between the actual binary labels and the predicted probabilities, was minimized during the model’s training process. During prediction phase, model calculates the probability of felling anxious and assigns the label accordingly.

3.5.3 Model Training using Decision trees

Decision Trees are referred to as Classification and Regression techniques (CART) techniques. This is a non-parametric supervised learning method used for both classification and regression. Decision trees are a popular choice due to their interpretability and simplicity. The objective is to create a model that predicts the probability of feeling anxious by learning simple decision rules inferred from the given data features. This model mimics human decision-making by dividing the feature space into a structure like a tree made up of nodes and branches that indicate decisions and potential results of the decisions made.

The prepared training data was used to train the decision tree model. The first decision node, located at the base of the decision tree, divides the dataset into two or more subgroups according to the value of a chosen characteristic. After that, each subset is handled recursively, with decision nodes further dividing the data until a predetermined endpoint is reached. The expected class label is stored in the tree’s terminal nodes, sometimes referred to as leaf nodes.

3.5.4 Model Training using Random Forest

Random Forest is an ensemble machine learning technique that combines the simplicity of decision trees and randomness to provide flexibility and resolve the problem of overfitting. Random Forest is also another CART technique and

uses decision trees as base learners, then applying bagging and column sampling on top of it.

The prepared training data was used to train the Random Forest model. Once we pass training data, the Model will create a forest of decision trees, those trees are trained independently on different random data subsets of the features and finally, those individual decision tree predictions will be aggregated using majority voting to obtain the likelihood of feeling anxious. In addition to the predictions, random forest is also used to find the important features using model's feature scores. According to the scores, top 3 features contributing to models' prediction are 'Feeling of guilt', 'irritable towards baby and partner' and, 'problems concentrating or making decision'. From the section of data exploratory analysis, individuals who reported irritability towards baby and partner are experiencing more anxiety.

3.5.5 Model Training using Support vector classifier

Support vector classifier (SVC) also called support vector machine (SVM) is another machine learning algorithm used for classification problems. It is a very popular, simple, and elegant supervised technique that creates a hyperplane to separate the classes in the feature space. The SVM classifier aims to find the optimal hyperplane (or) maximized margin that separates data points of different classes.

The prepared training data was used to train the SVM. During the training process, the algorithm learns to classify instances of different classes by identifying the optimum hyperplane aiming to maximize the distance between them. After being trained, the SVM model uses the optimal hyperplane to classify incoming instances into the proper feeling anxiety category.

3.5.6 Model Training using K-Nearest Neighbors

K-Nearest Neighbors (KNN) is an instance-based Machine learning algorithm and easy to implement as a result it became one of the popular choices for classification tasks. KNN doesn't require a training phase because it keeps all training instances in memory. The aim is to classify new instances of anxiety by using the majority class among its k nearest neighbors in the feature space. The prepared training data was used to fit the KNN. During the training phase, the algorithm keeps all instances in memory along with the accompanying class labels. During the prediction phase, for each incoming instance, KNN calculates the distance in the feature space between each new instance and all training instances. It then selects k nearest neighbors based on the computed distances and the majority class among these neighbors is then assigned as the predicted label for the new instance.

3.5.7 Model Training using XGBoost

Extreme Gradient Boosting (XGboost) is a machine learning algorithm known for its high performance that can be used for both Classification and Regression tasks. It is an ensemble learning algorithm that uses gradient-boosted Trees, row sampling, and column sampling techniques. A set of decision trees is sequentially built, and their predictions are aggregated to produce a final prediction. It is popularly used in many machine learning applications to deliver state-of-the-art results with minimal efforts.

The prepared training data was used to train the XGBoost. During the training phase, it sequentially builds a set of decision trees to minimize the loss function. It uses ensemble techniques and boosting techniques to enhance the performance of weak learners. The errors of each decision tree will be corrected by successor decision trees to improve predictive accuracy. During the prediction phase, the model evaluates each decision tree's prediction and combines them to produce the final prediction (or) likelihood for feelings of anxiety.

3.5.8 Model Training using Neural networks

Neural Networks are inspired by the function of the human brain which can be used for both classification and regression tasks. They consist of nodes, layers, weights, and activation functions that are responsible for processing input data, performing operations to learn complex data patterns, and generating output predictions. Neural networks with multiple hidden layers are referred to as Deep neural networks.

The prepared training data was used to fit the Neural network. During the training phase, it undergoes backward propagation where the model tries to learn data patterns within the data by adjusting model parameters iteratively to minimize a defined loss function. During the prediction phase, the model evaluates the incoming data and based on associated weights it produces the final prediction label for feelings of anxiety.

4 Evaluation and Results

This step involves evaluating the performance of the trained model using prepared testing data and discussing the results obtained by this paper. The reliability of the trained model will be known by evaluating its performance using prepared testing data. Henceforth Once the Model training phase has been completed, this paper evaluated each model's performance using evaluation metrics such as precision, recall, and F-score. Revising from the data exploration section, the target variable of the 'Feeling anxious' class is imbalanced. As a result, the performance metrics of precision, recall, and F-score have been chosen to evaluate the performance of the model. Precision refers to the measurement of the proportion of true positive predictions among all the model's positive predictions, and Recall refers to the measurement of the proportion of true positive predictions among all actual positive predictions. It is something referred as sensitivity and finally score is single metric to measure the balance between precision and recall, which refers to the harmonic mean of Precision and Recall.

The results of each model are tabulated in Table 2

From table 2, evaluating the model performance, this paper inferred that the Random Forest and XGBoost models performed well and achieved the highest F-score of 0.99 for feeling anxious class and 0.97 for Not feeling anxious class. These states Random Forest and XGBoost models are able to understand the patterns inside data and can effectively predict the likelihood of feeling anxious. The precision of 0.99 infers that the random forest correctly identified 99 percentage of the positive predicted feeling anxious instances out of all positively classified instances, similarly with the

	Yes(Feeling Anxious)			No(Not Feeling Anxious)		
Model	Precision	Recall	F-Score	Precision	Recall	F-Score
Logistic regression	0.84	0.84	0.84	0.66	0.66	0.66
Decision trees	0.85	0.88	0.87	0.72	0.67	0.70
Random Forest	0.99	0.98	0.99	0.96	0.98	0.97
Support vector classifier	0.91	0.88	0.89	0.76	0.8	0.78
K-Nearest Neighbors	0.93	0.91	0.92	0.82	0.85	0.84
XGBoost	1	0.97	0.99	0.94	1	0.97
Neural networks	0.96	0.91	0.93	0.82	0.92	0.87

TABLE 2
Performance Metrics

recall of 0.98, the random forest identified 92 percentage of all actual instances of feeling anxious in the data.

Also, it is evident that the Neural networks also achieved F-score of 0.92 for feeling anxious class and 0.87 for Not feeling anxious class, indicating its capacity to provide decent predictions.

The models logistic regression and decision trees comparably lower performance with F-scores of 0.66 and 0.7 for Not feeling anxious class. This means these models not able to understand the patterns and predict the likelihood of feeling anxious from the given data.

Overall, the outcome of performance metrics shows how well different machine learning models can predict the likelihood of feeling anxiety using the given dataset. As the best performer, the Random Forest and XGBoost models demonstrated its effectiveness in achieving predictions.

5 DISCUSSION

This section highlights the main insights from the exploration stage and discusses about the expectations to achieve from the described methods.

Firstly, this paper performed exploratory data analysis using univariant and bivariant analysis. The main aim of this stage is to extract valuable insights from the data, explain data patterns, and unearth the reasons for individuals feeling anxious.

The univariate analysis in the EDA section showed that most individuals are in the 40-45 age group, an equal portion of individuals reported sadness. Many individuals reported irritation towards the baby, problems sleeping at night, problems with overeating and finally no intention of committing suicide. A significant majority of individuals reported they have not encountered concentration problem, no problems bonding with the baby, no feeling of guilty. Similarly, bivariate analysis in the EDA section showed that the anxiety may be indicated in individuals aged above 30, in individuals who reported irritability towards baby and partner,problems bonding with baby, and finally having trouble more often.

Secondly, based on the data and objective, the probability of feeling anxious is classified as a supervised binary classification problem. The described methodology aims to leverage Machine learning algorithms by training and evaluating a range of classification algorithms to find the best model to predict PPD. This range comprises logistic regression, decision trees, random forest, SVM, KNN, XGBOOST, and neural networks. As the best performer, the Random Forest and XGBoost models demonstrated its effectiveness in achieving predictions.

overall, this paper expects the analysis and predictions will help medical field by helping doctors to know the technical and logical reasons for the occurrence of PPD and helping them in identifying the individuals in the postpartum period at risk and decrease the number of women who suffer from PPD. Furthermore, the methodology used, and insights gathered from this study can guide research efforts meant to enhance the mental health of many individuals in the postpartum period.

6 CONCLUSION

In conclusion, this paper shows the potential of Data science techniques to analyze the data, extract insights from the data, and develop a predictive model using machine learning that leverages the selected attributes to estimate the likelihood of postpartum depression. The methodology used in this paper begins with an exploratory data analysis that extracts powerful insights, explains patterns inside data, and infers potential factors for individuals feeling anxious and then proceeds with training and evaluation of various machine learning algorithms in predicting anxiety feelings among individuals in the postpartum period. The outcome of performance metrics shows how well different machine learning models can predict the likelihood of feeling anxiety using the given dataset. The ultimate objective is to have a positive impact on health among individuals in the postpartum period by identifying individuals at risk and help many individuals to seek health care earlier before complications arise.

REFERENCES

[1] K. Saqib, A. F. Khan, and Z. A. Butt, "Machine learning methods for predicting postpartum depression: scoping review," *JMIR mental health*, vol. 8, no. 11, p. e29838, 2021.
[2] S. Brummelte and L. A. Galea, "Postpartum depression: Etiology, treatment and consequences for maternal care," *Hormones and behavior*, vol. 77, pp. 153–166, 2016.
[3] T. Pearlstein, M. Howard, A. Salisbury, and C. Zlotnick, "Postpartum depression," *American journal of obstetrics and gynecology*, vol. 200, no. 4, pp. 357–364, 2009.
[4] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual review of clinical psychology*, vol. 14, pp. 91–118, 2018.
[5] S. Andersson, D. R. Bathula, S. I. Iliadis, M. Walter, and A. Skalkidou, "Predicting women with depressive symptoms postpartum with machine learning methods," *Scientific reports*, vol. 11, no. 1, p. 7877, 2021.
[6] G. Amit, I. Girshovitz, K. Marcus, Y. Zhang, J. Pathak, V. Bar, and P. Akiva, "Estimation of postpartum depression risk from electronic health records using machine learning," *BMC Pregnancy and Childbirth*, vol. 21, pp. 1–10, 2021.
[7] A. Gopalakrishnan, R. Venkataraman, R. Gururajan, X. Zhou, and G. Zhu, "Predicting women with postpartum depression symptoms using machine learning techniques," *Mathematics*, vol. 10, no. 23, p. 4570, 2022.