**Post Partum Depression**

Vishwasree Samala

Indiana University-Purdue University, Indianapolis, IN 46202, USA.

**Abstract**: Postpartum Depression (PPD) affects approximately 10-20% of new mothers and can develop within days or weeks after giving birth. Hormonal changes, sleep deprivation, life changes, and medical complications are several factors that contribute to the high risk of PPD. This project aims to analyze the factors that influence postpartum depression and predict PPD by using anxiety as an indicator of depression. Additionally, we aimed to identify the most important factors that affect PPD. We analyzed data collected from patient surveys in hospital settings, which were uploaded to Kaggle.com. Based on the data, we initially hypothesized that factors such as age, eating habits, and sleep do not have an influence on PPD. To verify our hypothesis, we conducted hypothesis testing and built machine learning models like SVM, Random Forest, Gradient Boosting, and Logistic Regression, to predict PPD. Our analysis showed that our initial hypothesis was not accurate; in fact, factors such as age, eating habits, and sleep do have an impact on anxiety and depression. The SVM model performed best in predicting anxiety, achieving an accuracy and precision of 96%. We were also able to identify the most important features influencing anxiety, which allowed us to answer our research questions.

**Keywords**: Postpartum Depression (PPD), Feeling Anxious, Suicide Attempt, Age, Overeating or Loss of Appetite, Trouble sleeping at night, Data Visualisation, Univariate, Bivariate Analysis, SQL, Hyperparameters, SVM, Logistic Regression, Gradient Boosting, KNN, Random Forest Classifier, Feature Importance.

# Project Scope

# Introduction

Many women are undergoing a phase called postpartum depression. According to estimates, 13% to 19% of mothers have postpartum depression, a significant psychological condition after giving birth to a child [1] (Pope & Mazmanian, 2016). The mother, along with the child, is affected by this condition. PPD can cause feelings of anxiety, sadness, irritability, and exhaustion, which will interfere with the bond between baby and mother and may last several months or longer. It is clear from the results of one of the articles that “Postpartum depression will occur in one out of every three pregnant women who experience psychological suffering with a sense of social isolation.” [4] (Nielsen et al., 2000) Women are unable to care for their babies and breastfeed them due to postpartum depression. “One in seven women was diagnosed with depression after pregnancy and received treatment for it.” [5] (Dietz et al., 2007) “However, more than 50 percent of these women experienced recurring depressive symptoms.” [5] (Dietz et al., 2007) While postpartum depression is a well-known condition, there is still much that is not understood about its causes, risk factors, and effective treatments. Delaying actions in women at greater risk of PPD until symptoms appear reduces the probability of depression in women only slightly, whereas treating with appropriately focused treatment prior to the emergence of symptoms significantly reduces the possibility of depression relapse. [3] (Zhang et al., 2021) [2](Cohen et al., 2006)

# 1.2 Aim

Our aim is to help healthcare professionals gain a better understanding of postpartum depression, identify potential risk factors, create improved screening and evaluation methods, and develop more specialized intervention plans. By examining the risk factors, we should be able to predict whether a woman is experiencing postpartum depression or not.

# Research Questions

* + 1. Do factors like age, trouble sleeping at night, overeating, or loss of appetite significantly influence Post-Partum Depression?
    2. Is anxiety a good predictor of Post-Partum Depression?

Based on the dataset and the research questions, our hypothesis is:

**Null Hypothesis:** The variables provided in the dataset like ‘age’, ‘Trouble sleeping at night’,

‘Overeating, or loss of appetite’ does not have a significant impact on postpartum depression.

**Alternate Hypothesis:** The variables provided in the data set like ‘age’, ‘Trouble sleeping at night’,

and ‘Overeating or loss of appetite’ have an impact on postpartum depression.

# Purpose

Postpartum depression has deleterious effects on both the mother and the child. For some people, postpartum depression may last for a few weeks or perhaps a year. In this project, we look at the symptoms of postpartum depression like irritability, trouble sleeping, anorexia, anxiety, guilt feelings, etc., and identify the most prevalent symptoms among women who have postpartum depression.

**1.4.1** Our project purpose is to investigate and examine PPD risk factors, pick out significant risk factors from this group, and rely on these significant risk factors for PPD diagnosis and prognosis.

**1.4.2** The purpose is to increase awareness of PPD and the advantages of early intervention.

# Methodology

# For this project, we followed the below methodology.

# 2.1 Type of Study: We used Quantitative Analysis methods on the data extracted. All columns in the dataset are categorical, and thus we applied quantitative analysis methods by treating them as nominal data. We gained insights into the relationship between the target variable and each column by plotting them with their respective frequency counts. This approach allowed us to visualize the distribution of each categorical variable and how they relate to the target variable.

# 2.2 Data collection: We used a sample dataset that was posted on Kaggle on postpartum depression.

# <https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression>

# 2.3 Storage and extraction: The data provided is in CSV format. The data is loaded into the MySQL database for analysis.

# 2.4 Data description: The data set contains 11 columns and 1503 rows. We identified statistically significant factors in the data that influenced postpartum depression. We used this data set to predict postpartum depression.

# 

# 2.5 Data Mining: After having the data, we did Data Cleaning, then Data Analyzing, and finally Data Visualization.

# 2.6 Statistical analysis: We performed univariate analysis on all of the selected variables, bivariate analysis to understand the relationship between the variables, and co-relation checks on the variables.

# 2.7 Model building: We used Random Forest Classifier, SVM, KNN, Gradient Boosting, and Logistic Regression.

# 

# 2.8 Evaluating Results: We evaluated the results after performing model development and analysis.

# 2.9 Tools used: MySQL, Python, different packages like matplotlib, and seaborn for

# data visualizations.

# 2.10 Contents:

# 

# 2.10.1 Data Collection

# 2.10.2 Data Storage and Extraction

# 2.10.3 Data Import

# 2.10.4 Data Cleaning

# 2.10.5 Data Analysis

# 2.10.6 Data Visualizations

# 2.10.7 Model Development and Analysis

# 2.10.8 Evaluating results

# 2.10.9 Hypothesis testing

# 2.10.10 Summary of our findings

# 2.10.11 Appendix

# 2.10.12 References

# 

# 2.11 Project Challenges

# We encountered several obstacles while working on the project. Firstly, we had difficulty

# creating a table in the SQL database. Additionally, we were unable to identify the primary

# key, so we loaded the data without it. Due to the observed impact of each characteristic in

# the dataset on anxiety levels, it was challenging to formulate a precise research question.

# Eventually, we decided to focus on only three attributes: age, trouble sleeping at night,

# and overeating or loss of appetite. We assumed that these factors would not affect the

# anxiety, which allowed us to narrow our research question and proceed with the analysis.

# During the data analysis phase, we imported the data into Python and discovered null values.

# We discussed various methods for identifying these null values and ultimately decided to

# use the bfill method. Since the data set was collected in a hospital environment, we felt

# that all participants would respond similarly to survey questions given the environment

# they were in.

# We also faced some hurdles in selecting the best parameters to verify the model’s

# performance. Ultimately, we agreed upon using accuracy and precision. We chose precision

# as we felt it was important to minimize false negatives. It is crucial not to start treatment

# on healthy individuals, so we wanted to ensure that the model would correctly identify

# those who were actually experiencing anxiety and depression.

# 

# 3. Data Collection

The dataset was taken from the Kaggle on postpartum depression.

<https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression>.

The target attribute selected for our analysis is "Feeling Anxious," which is used as a predictor for postpartum depression.

# 4. Data Extraction and Storage

# 4.1 Data extraction

Our first step in this project was to extract the data from the provided Kaggle dataset, which

was in CSV format. We carefully reviewed the dataset to understand how the values in each

column were recorded, and to identify any missing values. In cases where we found null

values, we discussed potential solutions to address this issue.

Our ultimate goal in this project is to assist healthcare professionals in gaining a deeper

understanding of postpartum depression, including identifying potential risk factors,

developing more effective screening and evaluation methods, and creating more

specialized intervention plans.

All of the attributes in the dataset share the same variable type, which is categorical.

However, they differ in their data types. "Timestamp" has a data type of date time,

"Age group" has a data type of integer, and the remaining attributes have a data type

of string.

Now let's examine the descriptions of each attribute that were provided in the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Data Type** | **Variable Type** | **Description** |
| Timestamp | Date Time | Categorical | Timestamp of survey |
| Age group | Integer | Categorical | The age group of people surveyed. |
| Feeling sad or tearful | String | Categorical | Feeling sad or Tearful |
| Irritable towards baby and partner | String | Categorical | Whether the mother is irritable towards the baby & partner or not. |
| Trouble sleeping at night | String | Categorical | Finding out if there are any sleeping troubles tonight or not. |
| Problems Concentrating or making decisions | String | Categorical | To find out whether there are problems concentrating or making a decision. |
| Loss of Appetite | String | Categorical | Overeating or loss of appetite |
| Anxiety | String | Categorical | Feeling anxious or not |
| Guilt | String | Categorical | Whether there is a feeling of guilt or not. |
| Problems of bonding with baby | String | Categorical | To find out whether there are problems with bonding with the baby or not. |
| Suicide Attempt | String | Categorical | Suicide attempt by the person is surveyed. |

**Fig 3:** Table showing attributes and their description

# 4.2: Data import

# We began the process of importing the CSV file into the I501jolburnsSpring23grp\_01\_db

# database in phpMyAdmin.

# Graphical user interface Description automatically generated with medium confidence

# Fig 4: Structure of postpartum\_1 in PhpMyAdmin

# 4.3 Data Storage:

# Graphical user interface, text, application Description automatically generated

# Fig 5: Data stored in the database

# Graphical user interface, text, application, chat or text message Description automatically generated

# Fig 6: Database and table postpartum\_1

# 4.4: Data importing from SQL to Jupyter

# This is how we imported the data from SQL to Jupyter python notebook by connecting to

# the database.

# 

# Text, application Description automatically generated

# Fig 7: Importing data from SQL to Jupyter

# 

# Table Description automatically generated

# Fig 8: Showing the top few rows and columns from the data set.

# Once a successful connection to the database has been established, the next step is to

# write code to fetch all the columns from the database and examine their data types. Since

# the column names in the database may differ from those in the dataset, we provided

# suitable codes in the appendix to facilitate the renaming process.

# To rename the columns, we opted to use Python code instead of performing the operation

# in phpMyAdmin. Once the renaming is completed, we need to check whether the data has

# been cleaned properly. To accomplish this, we wrote a code to check the number of

# null values in each column.

# 4.5 Data Cleaning

# For the analysis to be accurate and reliable, the data cleaning phase is essential. We found numerous null values in our dataset that need to be handled with caution.

# To address this issue, we began by reviewing the data in Excel, examining each row to identify any missing values. We discovered that certain columns, such as "Irritable towards baby and partner," "Problems concentrating or making decisions," and "Feeling of guilt," contained null values that needed to be handled with care. We decided to handle null values

# in Python so that the process would be repeatable. In Python, we decided to treat null values by replacing them with the values from the next row using the bfill method. This decision was made based on our assumption that the data set was collected in a hospital environment and that all participants would respond similarly to the survey questions given the environment they were in.

# Table Description automatically generated with medium confidence

# Fig 9: Data showing Null values

# Text Description automatically generated

# Fig 10: Count of null values from the dataset

# Based on the above image, it appears that there are null values in the dataset that need to be handled. We have included the relevant code to address this issue in the image below.

# 

**Fig 11:** Data cleaning code

After addressing the null values, we can verify that our code has successfully removed any

remaining null values from the dataset. By reviewing the null values count, we can confirm

that there are no longer any null values in the dataset.

Text

Description automatically generated

**Fig 12:** Count of null values after the bfill method

We made a minor change to the dataset by renaming the column names with

underscores to improve readability and consistency. The new column names

reflect the same information as the original column names but are easier to read

and understand.

Text, scatter chart

Description automatically generated

Text

Description automatically generated

**Fig 13:** Changed column names

# 4.6 Statistical Analysis:

# 4.6.1 Univariate and Bivariate Analysis

# In this section, we will present the results of our statistical analysis and highlight some interesting observations that we have made. To begin with, we conducted univariate analysis on the categorical columns of the dataset. Since frequency counts are the most effective way to understand categorical data, we first present the frequency counts for each column. Subsequently, we will conduct a bivariate analysis and then explore various visualization techniques to gain further insights.

# It's important to note that we have excluded the timestamp attribute from our analysis as it does not provide any significant information beyond the time of form submission.

# During our initial analysis, we identified certain attributes that were distinct from others, and we will discuss our findings from these attributes in detail.

A picture containing text, screenshot, font

Description automatically generated

**Fig 14:** Frequency count code for the age column

To begin with, we created a list of all the columns in the dataset and wrote code to

display the frequency count for the Age column. This enabled us to gain a better

understanding of the distribution of ages among the survey participants.

A picture containing text, font, receipt, screenshot

Description automatically generated

# Fig. 15: Frequency values of the age column

# While the frequency counts provided us with a basic understanding of the

# distribution of ages among the survey participants, visualizations can help us gain more

# insights. We used libraries such as Seaborn and Matplotlib to create visualizations and

# analyzed the patterns in the data. By doing so, we were able to derive more meaningful

# insights from the data.

# Chart Description automatically generated

# Fig 16: Histplot for Age column with the target attribute

# We can derive some useful insights from the histogram visualization of the Age column

# with respect to the Feeling Anxious attribute. The age group is shown on the X-axis, while

# each group's frequency count is shown on the Y-axis. The blue bars represent individuals

# who reported feeling anxious, while the orange bars indicate that they did not. We observe

# that the age groups 30-35, 35-40, and 40-45 have the highest rates of anxiety positivity,

# while the other two age groups have the lowest. The age group 25-30 appears to be the

# safest, with fewer positive and negative cases compared to all other groups.

# The previous analysis on the age column provided valuable insights, so let's perform a similar

# analysis on another interesting column, "Suicide attempt." We will explore the frequency

# counts and create visualizations to understand the distribution of the data and any potential

# relationships with the target attribute.

# A screenshot of a computer code Description automatically generated with medium confidence

# Fig. 17: Frequency count code and values for the column Suicide Attempt

# The above figure shows that there are three values in the "suicide attempt" column

# and their respective count values are also displayed.

# Chart, bar chart Description automatically generated

# Fig 18: Hist plot for Suicide Attempt Column with target attribute

# The above histplot shows the count of each value in the suicide attempt column, plotted against

# the Feeling Anxious column. The X-axis represents the three possible values in the suicide

# attempt column and the Y-axis shows their respective counts.

# We understand from the graph that people who did not attempt to commit suicide are

# feeling more anxious than the people who tried to commit suicide. This seems contradictory

# to the premise of feeling anxiety.

# The pattern observed in the histogram is consistent even for those who do not report feeling

# anxious, which is a strange behavior for this column.

# In this section, we will discuss two more attribute plots. One of these columns is the

# "Problems of bonding with baby or partner" column.

# Chart, bar chart Description automatically generated

# Fig 19: Hist plot for problems of bonding with the baby or partner column with target attribute.

# In the above plot, we can see that woman who reported having problems with bonding with

# their baby sometimes has the highest rate of feeling anxiety.

# Similarly, for women who reported never having problems with bonding with their baby, the ratio of feeling anxious and not feeling anxious is almost equal.

# Here, we will explain one more attribute plot, which is related to overeating or loss of appetite. First, let's take a look at the plot and then we will discuss the insights provided by it.

# Chart, bar chart Description automatically generated

# Fig 20: Histplot for Overeating or loss of appetite column with target attribute.

# The X-axis of the plot represents the Overeating or Loss of Appetite column and its values

# are Yes, No, and Not at all. The Y-axis displays the frequency count of each category of overeating.

# We plotted this with respect to the target variable Feeling Anxious. This column is particularly

# interesting because according to the data, women who do not overeat have the highest

# anxiety rate compared to those who do. Furthermore, we can infer that overeating does

# not affect the anxiety rate and may actually decrease it to a significant extent.

# Now we will see insights into the column Feeling sad or tearful column.

# A screenshot of a computer code Description automatically generated with medium confidence

# Fig 21: Frequency code and count for feeling sad or tearful column

# Now we will see the visualization of feeling sad or tearful attribute.

# A screenshot of a graph Description automatically generated with low confidence

# Fig 22: Histplot for feeling sad or tearful column with target attribute.

# The data suggests that when women are not feeling sad, they are more likely to experience

# anxiety, which can be confusing to accept as there is a narrow margin between the two

# emotions. Additionally, it appears that both not feeling sad and feeling anxious contribute

# significantly to the overall emotional state of women.

# Now we will see insights into the column irritable towards baby and partner.

# A screenshot of a computer code Description automatically generated with low confidence

# Fig 23: Frequency count code and values for irritable towards baby & partner

# Now we will see visualizations for this irritable towards baby & partner attribute.

# A screenshot of a graph Description automatically generated with medium confidence

# Fig 24: Histplot for irritable towards baby & partner with target attribute.

# As expected, the data shows that when women become irritated, they often experience a

# significant amount of anxiety. This is in contrast to the fact that when people are not feeling

# anxious, there is an equal ratio of women who are either getting irritated or not.

# Now we will see trouble sleeping at night column.

# A screenshot of a computer code Description automatically generated with low confidence

# Fig 25: Frequency count code and values for Trouble sleeping at night

# Now we will see visualizations for this trouble sleeping at night.

# A screenshot of a graph Description automatically generated with medium confidence

# Fig 26: Histplot for trouble sleeping at night

# The data indicates that for women who experience trouble sleeping for more than two days,

# there is a significant correlation with anxiety. This correlation appears to decrease linearly

# for both women without sleeping problems and those with sleeping problems. Furthermore,

# women who do not experience anxiety are generally those without any kind of sleeping

# problem at night.

# Now we will see problems concentrating or making decisions column.

# A screenshot of a computer code Description automatically generated with low confidence

# Fig 27: Frequency count code and values for problems concentrating or making decisions.

# Now Let’s see the visualizations for this problem concentrating or making decisions.

# A picture containing text, screenshot, diagram, parallel Description automatically generated

# Fig 28: Histplot showing problem concentrating or making decisions with feeling anxious.

# The data suggests that the anxiety rate is similar for both "yes" and "no" responses in the

# column related to problem concentrating or decision-making. However, women who often

# report having problems concentrating show the highest rates of anxiety. Conversely, women

# who do not report any issues with concentration or decision-making tend to have the lowest

# anxiety rates.

# Now we will see the insights into the feeling guilty column.

# A screenshot of a computer code Description automatically generated with medium confidence

# Fig 29: Frequency count code and values of feeling guilty column

# A screenshot of a computer screen Description automatically generated with low confidence

# Fig 30: Histplot of feeling guilt with feeling anxious

# The data suggests a contradictory pattern where women who report not feeling guilty

# exhibit higher rates of anxiety, while those who report feeling guilty exhibit the lowest

# anxious rates. In contrast, when the columns are reversed, women who believed they might

# be guilty fall in between the other two columns.

# Before proceeding with further analysis, let us first examine the frequency count of the

# target attribute, Feeling Anxious. We will check if it is biased towards any particular value

# or if it is balanced. The following image is a count plot using seaborn for the Feeling Anxious

# column. From the plot, we can observe that more women are feeling anxious and the graph is

# slightly indicating that the outputs are skewed towards Yes compared to No.

# Chart, bar chart Description automatically generated

# Fig 31: Count plot for the target column Feeling Anxious

# After analysis, we will convert all categorical values into numerical format because the expected

# format for the model is in numbers. We used Label Encoder for this purpose.

# We chose to use a LabelEncoder instead of OneHotEncoder for this purpose because we have

# multiple categorical variables, and using a OneHot Encoder could result in the risk of cardinality

# with multiple columns.

# Text Description automatically generated

# Fig 32: Label Encoder Code

# We should also convert the data in the target attribute to numerical format and store it separately.

# Once the conversion is done, we can take a look at the data to see how it looks. After converting

# the categorical data to numerical values, we removed the original columns and will be using the

# newly formed columns for our analysis.

# Table Description automatically generated

# Fig 33: Transformed Data

# Let us explore another way to analyze these attributes, apart from using a hist plot.

# 4.6.2 Correlation Map:

# The correlation factor has been used to produce a heatmap.

# Text Description automatically generated with medium confidence

# Fig 34: Code for generating Heat Map

# A picture containing chart Description automatically generated

# Fig 35: Generated Heat Map

# Although we can create a heatmap by converting all variables into discrete ones, it may not

# provide us with valuable insights.

# The heatmap above shows how closely all the variables are correlated, with deeper colours

# denoting a greater degree of correlation and lighter colours denoting a lesser correlation.

# Overeating or loss of appetite and feeling sad or tearful are strongly correlated with values of

# 0.36. Problems bonding with baby and partner and feeling anxious are strongly correlated with

# values of 0.23. Feeling guilty and feeling anxious has a strong negative correlation with a value

# of -0.27.

# 4.6.3 Training and Testing Data

# We Stored the data into independent (x) and target (y) variables.

# 

# Fig 36: Code showing for independent and target variables

# As stated in our Aim, we would like to predict postpartum depression by using the

# variable ‘Feeling\_Anxious’ as the target variable which is a good indicator of depression.

# We created a train-test split of 70-30 on the data set to evaluate our machine-learning models.

# A picture containing text, font, screenshot, line Description automatically generated

# Fig 37: Test train Split code

# As seen above, there are only 9 columns in the dataset as we dropped the timestamp column

# because it would not have a significant impact on the model building.

# 4.7 Building Models

# 4.7.1 Random Forest Classifier

# The first step after preparing the data is to import the necessary packages. Then, we will plot

# the ROC curve and classification report for each model.

# Text Description automatically generated

# Fig 38: Random Forest Code

# Chart, line chart Description automatically generated

# Fig 39: Result of Random Forest Classifier

# Before discussing the results, let's take a look at the code where we implemented hyperparameter

# tuning. This process is important as it helps us find the best suitable model for the given input

# data. The classification report shows that we achieved an accuracy of about 91%, with a descent

# F1-scores, precision, and AUC values of 0.89. We can observe that it shows a good true positivity

# rate, and after a threshold of 0.18, the false positives are zero, whereas, before 0.18, there are

# false positives that vary linearly.

# 4.7.2: Support Vector Machine (SVM)

# After preparing the data, we need to import the necessary packages and plot the ROC curve

# and classification report for the SVM model. As mentioned earlier, we will also use

# hyperparameter tuning, and the tuned parameters will differ for each model.

# Text Description automatically generated

# Fig 40: Code for SVM

# Graphical user interface, chart, line chart Description automatically generated

# Fig 41: Results of SVM

# The classification report shows an achieved accuracy of about 96%, as well as decent

# f1-scores and precision. From the above AUC, we can understand that at 0.05, we achieve

# a TPR of 1. The main important thing is that after 0.2, the false positives are 0, which suggests

# that it is performing decently.

# 4.7.3: Logistic Regression

# Once we have the data ready, the first step is to import the necessary packages. We will then

# proceed to plot the ROC curve and classification report for the logistic regression model. As

# mentioned earlier, we will also be using hyperparameter tuning for this. The tuned parameters

# will differ for each model.

# Text Description automatically generated

# Fig 42: Code for Logistic Regression model

# Table Description automatically generated

# Fig 43: Classification report of Logistic Regression Model

# The ROC curve analysis for the logistic regression model shows that its performance

# is not satisfactory with the given input data.

# Chart, line chart Description automatically generated

# Fig 44: ROC curve for Logistic Regression Model

# The logistic regression model's performance seems to be suboptimal, as it requires a

# longer time to achieve the positivity rate, and the false positivity rate is also high.

# 4.7.4: Gradient Boosting

# After preparing the data, the first step is to import the necessary packages, and then we can

# plot the ROC curve and classification report for the Gradient Boosting model. As mentioned

# earlier, we will use hyperparameter tuning for this model, and the tuned parameters will

# differ from each model.

# Text Description automatically generated

# Fig 45: Code for Gradient Boosting

# Before discussing the results, let's take a look at the code where we implemented the

# hyperparameter tuning. Hyperparameter tuning is an important process as it helps to fit the

# best suitable model for the given input data.

# Chart, line chart Description automatically generated

# Fig 46: Results of Gradient Boosting

# The classification report shows that the accuracy achieved is around 73%, with poor f1-scores

# and precision. The AUC value is 0.63, which suggests a bad true positivity rate. We can observe

# that the false positivity rate is linearly increasing up to the threshold of 0.6, indicating that it is

# performing poorly when compared to the previous models.

# 4.7.5: K Nearest Neighbor (KNN)

# After preparing the data, we need to import the necessary packages and plot the ROC curve

# and classification report for the KNN model.

# Text Description automatically generated

# Fig 47: Code for KNN

# Chart, line chart Description automatically generated

# Fig 48: Results of KNN

# The classification report shows that an accuracy of about 92% was achieved, along with good

# f1-scores and precision. The AUC value is 0.90, indicating a good true positivity rate. The

# false positivity rate increases linearly until the threshold of 0.16, suggesting that the model is

# performing decently compared to the previous models.

# 5. Evaluating Results

# As observed from all the models, the SVM is performing very well with an accuracy of 96%

# and good F1-scores 0.95, and a precision score of 0.96.

# We used the Scikit-learn library to extract feature importance and understand the important

# features in the SVM model. The feature importance of the SVM model is shown below:

# A screenshot of a computer code Description automatically generated with medium confidence

# Fig 49: Feature importance code for SVM

# Here, we are using permutation importance from the sklearn library. It should be noted that the

# feature importances are not normalized, and since we have categorical values, normalization is

# not necessary. The output simply means that a higher value indicates greater importance. Based

# on the metrics, we have selected SVM as the final model, as it is performing better than the other

# models.

# A screenshot of a computer code Description automatically generated with low confidence

# Fig 50: Feature importance of each column

# A picture containing text, screenshot, font, diagram Description automatically generated

# Fig 51: Feature importance graph

# The graph shows that feeling of guilt is the most important feature in predicting anxiety

# followed by problem concentrating or making decisions. The remaining features have similar

# importance in the prediction of anxiety. Age has the least important feature in prediction.

# 6. Hypothesis

# The null hypothesis is that parameters such as age, trouble sleeping at night, and overeating or

# loss of appetite do not have any significant effect or impact on postpartum depression, whereas

# the alternative hypothesis would state the opposite, indicating that these parameters do have a

# significant impact on postpartum depression. Instead of the Fischer test in this instance, the

# Chi-square test is used to get the p-value. The reason for choosing the chi-square test over the

# Fischer test is that Chi–Square is used for large data samples, whereas Fischer’s test gives

# appropriate results for only a small sample size.

# The code and results for the hypothesis testing can be seen in the images below.

# 

# Fig 52: Hypothesis Testing

# 

# Fig 53: Hypothesis Results

# We can see that in this case, the p-values are below the significance level, indicating that

# the null hypothesis should be rejected.

# 7. Summary of the findings

# After following our approach of data cleaning, analysis, and visualization, we gained

# several insights related to our two research questions. Here are some of our significant

# findings for each research question, based on our methodology.

# 7.1 Findings for Research Question 1

# 1. Do factors like age, trouble sleeping at night, overeating, or loss of appetite have a

# significant influence on Post-Partum Depression?

# Yes, factors such as age, trouble sleeping at night, overeating, or loss of appetite have a

# significant impact on postpartum depression.

# The following is the evidence for this question:

# 

# 

# 

# Fig 54: Evidence for research question 1

# 7.2 Findings for Research Question 2

# 2. Is anxiety a good predictor of Post-Partum Depression?

# The machine learning models showed that, for the given data, we were able to predict

# anxiety with an accuracy of 96% and precision of 0.96, indicating that the model is a

# good predictor of anxiety and postpartum depression.

# Additionally, we used permutation importance to identify the feature importance, which

# provided insight into each feature's effect on anxiety and depression.

# The following is the evidence for this question:

# A picture containing text, screenshot, font, diagram Description automatically generated

# Fig 55: Feature Importance

# 8. Limitations

# The data collected in this study was limited to one hospital in a specific region, which raises

# concerns about the representativeness and bias of the sample. Further testing of the model in

# different settings is necessary before generalizing the results.

# Although the data was collected from multiple participants, it was only over a two-day period in

# 2022. To conduct a more comprehensive analysis, it is recommended to collect data at different

# stages of postpartum for at least one year.

# Given that the data was collected in 2022, it is possible that the COVID-19 pandemic may have

# influenced the responses in a confounding manner, which was not accounted for during data

# collection.

# Our knowledge of machine learning and Python was limited while testing and modeling our

# hypothesis, which may have impacted the accuracy and robustness of our results.

# 9. Appendix – Python Code Snippets

# A screenshot of a computer program Description automatically generated with low confidence

# 

# A screenshot of a computer code Description automatically generated with low confidence

# A screenshot of a computer program Description automatically generated with medium confidence

# 

# A screenshot of a computer code Description automatically generated with low confidence

# 

# A picture containing screenshot, text, square, diagram Description automatically generated

# A screen shot of a computer Description automatically generated with medium confidence

# 

# A screenshot of a computer Description automatically generated with medium confidence

# A screenshot of a computer Description automatically generated with medium confidence

# A screenshot of a computer Description automatically generated with low confidence

# 

# 

# A screenshot of a computer program Description automatically generated with low confidence

# A screenshot of a computer program Description automatically generated with low confidence

# A screenshot of a computer program Description automatically generated with low confidence

# A screenshot of a computer program Description automatically generated with medium confidence

# A screenshot of a computer program Description automatically generated with medium confidence

# A screenshot of a computer program Description automatically generated with medium confidence

# A screenshot of a computer code Description automatically generated with medium confidence

# A screenshot of a computer code Description automatically generated with low confidence

# References

Post Partum Depression Data Set Link -

<https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression>

1. Pope, C. J., & Mazmanian, D. (2016). Breastfeeding and Postpartum Depression: An Overview and Methodological Recommendations for Future Research. *Depression Research and Treatment*, *2016*, 1–9. <https://doi.org/10.1155/2016/4765310>

2. Cohen, L. S., Altshuler, L. L., Harlow, B. L., Nonacs, R., Newport, D. J., Viguera, A. C., Suri, R., Burt, V. K., Hendrick, V., Reminick, A. M., Loughead, A., Vitonis, A. F., & Stowe, Z. N. (2006, June). Relapse of Major Depression During Pregnancy in Women Who Maintain or Discontinue Antidepressant Treatment. *Obstetrical & Gynecological Survey*, *61*(6), 368–370. <https://doi.org/10.1097/01.ogx.0000219497.90073.85>

3. Zhang, Y., Wang, S., Hermann, A., Joly, R., & Pathak, J. (2021, January). Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. *Journal of Affective Disorders*, *279*, 1–8. <https://doi.org/10.1016/j.jad.2020.09.113>

4. Nielsen, D., Videbech, P., Hedegaard, M., Dalby, J., & Secher, N. H. (2000). Postpartum depression: identification of women at risk. Bjog: An International Journal of Obstetrics and Gynaecology, 107(10), 1210–1217. <https://doi.org/10.1111/j.1471-0528.2000.tb11609.x>

5. Dietz, P. M., Williams, S. B., Callaghan, W. M., Bachman, D. J., Whitlock, E. P., & Hornbrook, M. C. (2007, October). Clinically Identified Maternal Depression Before, During, and After Pregnancies Ending in Live Births. American Journal of Psychiatry, 164(10), 1515–1520. <https://doi.org/10.1176/appi.ajp.2007.06111893>