THE CURIOSITY CUP 2024
A Global SAS® Student Competition

Maternal Mental Health and Infant: An analysis of the impact of Maternal Mental Health on Infant Behavior from zenodo.org using SAS Studio and SAS Viya

Dreamland Data

INTRODUCTION

Maternal mental health influences the quality of parent-child interactions, which are vital for the child's social and cognitive development. Early intervention and support for maternal mental health are essential for promoting positive infant outcomes and fostering secure attachment relationships. When mothers experience depression, anxiety, or stress, it affects their ability to provide consistent care and emotional support to their babies. This can lead to disruptions in the mother-infant attachment bond, resulting in infants showing signs of distress like excessive crying, difficulty falling asleep, the hours they sleep, and other factors. This paper describes a systematic approach for analyzing the impact of maternal mental health on infant behavior. The dataset was collected from Swiss National Science Foundation from 410 mothers in postpartum period, and their infants aged between 3-12 months old. So, we used SAS Studio, SAS Visualization and Machine Learning using SAS Viya to answer a valuable question such as What is factor are most related to infant behavior”? What is the relationship between how the baby falls asleep and the duration of sleep at night? Before we make descriptive and predictive analytics, we classify maternal mental health for being normal or abnormal for different maternal mental health scales based on calculated the score for every scale. Finally, that emphasizes how distinct influences of maternal mental health difficulties shape infant`s sleep patterns and their behavior depending on different moderating factors. And from this paper we can gain valuable insights that can inform government interventions aimed at improving families' physical health and social wellbeing.

METHODOLOGY

IMPORTING AND ACCESSING DATASET

The given text describes a code that uses the SAS programming language to import a data file in xlsx format located at a specific file path. A SAS library named CUP is assigned to the directory "CurisityCup2024","validvarname=v7" option is important for specifying the rules that govern the naming of variables within a dataset.

```
libname CUP "/home/u58481968/CurisityCup2024";
option validvarname=v7;
```

Then we import the data from an excel file into the SAS dataset, these steps enable the availability of data for analysis, the integration of diverse datasets, and the maintenance of consistency in data format and structure.

```
PROC IMPORT datafile="/home/u58481968/CurisityCup2024/FINAL_MENTAL_HEALTH.xls"
out=CUP.FINAL_MENTAL_HEALTH dbms=xls
replace;
RUN;
```

EXPLORATION AND PREPARING DATA

The preparation and exploration phases of the data analysis process are two crucial stages. Data exploration assists in comprehending the structure of the dataset and identifying patterns, while data preparation addresses issues such as missing values, outliers, and inconsistencies. These processes ensure data quality, facilitate variable conversions, and enhance the overall dependability and effectiveness of further analysis and model building. These actions basically lay out the foundation for significant discoveries and decisions that are well-informed.

To explore the data and understand it better, the first 100 rows then starting with data cleaning to remove any unnecessary columns. Adjusting the length of the new character columns to prevent data truncation, ensure consistency across datasets, and facilitate compatibility with downstream processes.

convert the Sleep_night_duration format from time to numeric to display the variable with the most suitable dimension, ensuring that it is presented in a readable and understandable manner when printed or viewed. And calculate its value by dividing Sleep_night_duration_bb1 by 3600.

```
format Sleep_night_duration_bb1 BEST12.;

Sleep_night_duration=(Sleep_night_duration_bb1/3600);
```

Then we removed the outliers to increase the robustness of the descriptive statistics. then convert the gestational age from weeks to years to unify the factor of the ages and then format Gestational_age

```
Gestationnal_age=Gestationnal_age/52;

format Gestationnal_age comma10.2;
```

Label the column for enhancing understanding, readability, and collaboration in data analysis, handle Marital_status_edit column if it= "Pacse" then set it to 2, create 4 new columns EPDS_score, HADS_score by sum the answers of questionnaires, HADS_info, EPDS_info by provide the status of the mother depend on EPDS_score and HADS_score.

Then we assume that the CBTS_score, CBTS_info, IBQ_R_VSF_score, and IBQ_R_VSF_info based on previous assumptions, create these 4 new columns, and divide the CBTS_score by 2 to normalize the values. Then convert the numeric columns to character columns based on the metadata. Lastly, exporting the data into an xls file to using it in analysis.

ANALYZING DATA

The power of knowing things that we know that we don`t know is analyzing the data to get information and then processing the information to get an advanced one to explore insights and hidden patterns, behaviors, and information by using different types and techniques of analysis.

The descriptive analysis is answering what questions. So, first, we use the macro language that allows us to write programs that will rewrite themselves, describe the values of each scale, and know what our data values are. The macro function is %analysis (scale) using scale as a macro variable, and it contains multiple procedures for accurate analysis. Using the PROC SQL and selecting the AVG (age) function, we get 30.20 as the average age of all the participants, and most mothers in our data are in relationship. The largest number of participants who fall into each category of marital status and education level is 185, and they are in a relationship with no education or compulsory school.
The range of each scale is calculated by:

```
title "What is the range of &scale score ";
```

```
    PROC SQL;
        SELECT MIN(&scale) AS Min&scale , MAX(&scale) AS Max&scale
      FROM CUP.FINAL_MENTAL_HEALTH;
    QUIT;
```

- EPDS from 0 to 28
- HADS from 0 to 20
- CBTS from 0 to 23.5
- IBQ_R_VSF from 6 to 6

After knowing what we have, we must discover the how questions. and we do that by using associations between the fields to explore more hidden features.
How does each scale vary based on marital status and education level?

```
title "Do &scale vary based on marital status or education level?";
PROC SQL;
      SELECT Marital_status, AVG(&scale) AS Average&scale FROM
CUP.FINAL_MENTAL_HEALTH GROUP BY Marital_status ORDER BY Average&scale DESC;
QUIT;

PROC SQL;
      SELECT Education_level, AVG(&scale) AS Average&scale FROM
CUP.FINAL_MENTAL_HEALTH GROUP BY Education_level ORDER BY Average&scale DESC;
QUIT;
```

The maximum average score in all scales is when the mother is single and has no education level. The second question was How each scale of the maternal scales affects or depends on infant behavior. And how do all scales relate to each other? We did that using PROC CORR, VAR statement for the maternal scales, and WITH statement for the infant behavior scale to find out that the scale most related to the IBQ_R_VSF is the EPDS scale.

```
title "Is there a correlation between EPDS (Edinburgh Postnatal Depression
Scale) score, HADS (Hospital Anxiety and Depression Scale) score, CBTS (City
Birth Truma Scale) score and IBQ_R_VSF (Infant behavior questionaries)
score?";

PROC CORR nosimple DATA=CUP.FINAL_MENTAL_HEALTH;
         VAR EPDS_score HADS_score CBTS_score;
         with IBQ_R_VSF_score;
RUN;
```

So, we use SAS Visual Analytics to make a useful prediction that we deduced valuable insights and hidden patterns that need a lot of analytics. We discover more on What are the characteristics of EPDS_info? Finding out the most related factors to EPDS_info is CBTS_score. Figuring out also the HADS (Hospital Anxiety and Depression Scale) for maternal. We asked, what are the characteristics of HADS_info? What are the groups based on EPDS_score by chance of HADS_info being abnormal?

- HADS info has a 25.67% chance (105 of 409) of being Abnormal
- If EPDS score is between 13 and 20 Gestationnal_age is greater than or equal to .79, then HADS_info has a 100.00% chance (5 out of 5 cases) of being Abnormal.
- If EPDS score is greater than or equal to 20, then HADS_info has a 91.18% chance (31 out of 34 cases) of being Abnormal.
- If EPDS score is between 13 and 20, Sleep_night duration is 7, 8.5, 9. 9.5 or 11, then HADS_info has an 86.36% chance (19 out of 22 cases) of being Abnormal.

The power of automated explanation in SAS Visual Analytics helped us continue to find out more about infant behavior and the infant sleep pattern. And answer the question of What factors are most related to IBQ_R_VSF_info? And the figure (1) illustrates that the night awakening number of babies is most related to infant behavior after that EBDS_score (Edinburgh Postnatal Depression Scale), and we also indicate that if the night awakening number is 6 or 10, then IBQ_R_VSF_info has a 66.67% chance of being abnormal.

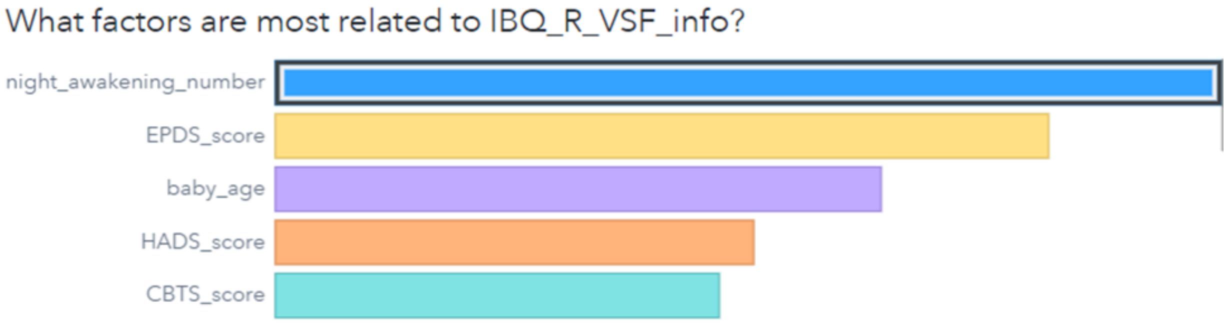


Figure 1: Shows What factors are most related to IBQ_R_VSF_info (Infant Behavior Questionnaire).

Because the infant is our primary concern and objective, and we want the baby to have better health and better sleep, So, we make a relationship between night awakening number and falling asleep? And that is illustrated in Figure 2. To conclude, what is the best position for the baby to sleep comfortably and sleep more at night?

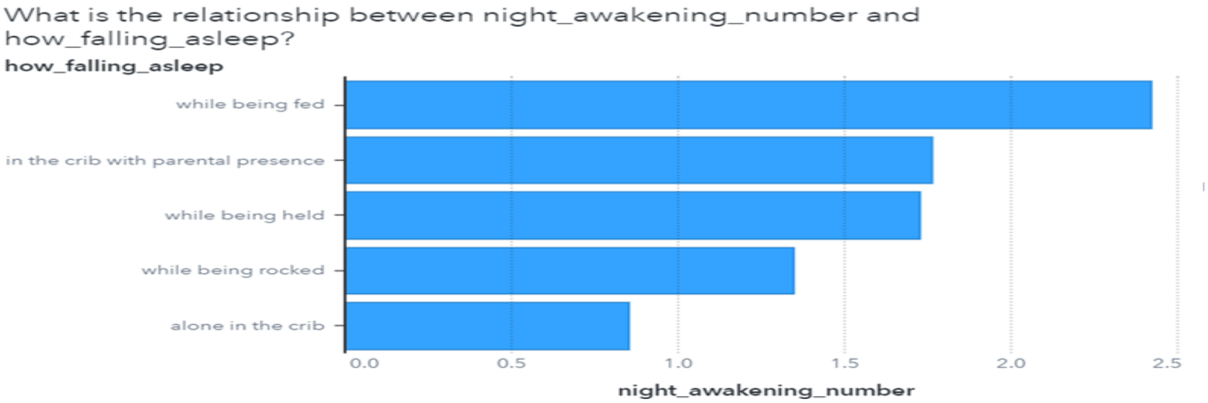


Figure 2: What is the relationship between night_awakening_number and how_falling_asleep?

Furthermore, we discover that the sex of the infant has an impact on maternal mental health, particularly in CBTS and HADS, if the infant is a boy, although girls occur more frequently in the data. So that this proves the validity of the proof.

DATA MODEL

The following section focuses on predicting the impact of different variables on CBTS results by machine learning using SAS Viya®. Our aim was to find the best-suited supervised learning model for our target. We chose CBTS_info as the target with two classes (normal or abnormal), which is suitable for classification models. We also changed the role of CBTS_info, HADS_info, EPDS_info, and IBQ_R_VSF_info to be rejected to prevent overfitting. We constructed four models: Logistic Regression, Neural Network, Support Vector Machines (SVM), and Decision Tree. The Model Comparison node was used to compare the models and determine the champion model. The data was automatically split into 60:30:10 for training, validation, and testing datasets for all the models and chosen

Test data for selection partition. After running our pipeline, SAS Viya determined the logistic regression model as the most effective model (CHAMPION MODEL) with an accuracy of 0.9024 against 0.8780 for SVM, 0.8293 for Decision Tree and 0.8293 for Neural Network In SAS, logistic regression is a statistical method used to model the relationship between a binary. In our case Normal or Abnormal.

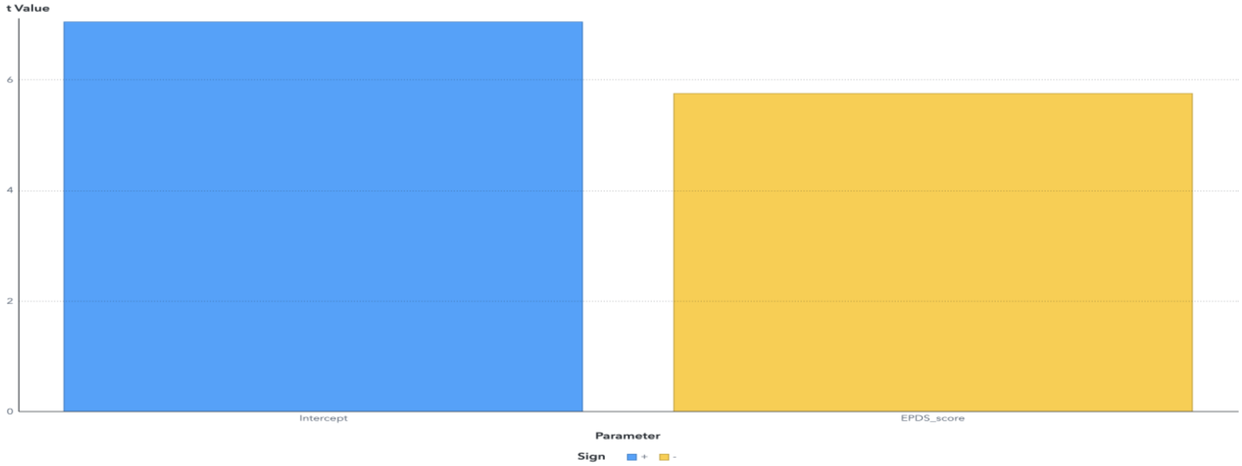


Figure 3: Shows the t value for the two parameters estimate in a logistic regression model (Innercept and EPDS_score).

So, this graph illustrates that the most significant parameter in this graph is the Intercept, which has a t value of 7.0464. In logistic regression, the intercept represents the log-odds of the dependent variable when all predictor variables are zero. The t-value associated with the intercept indicates whether the log-odds are statistically significantly different from zero. We assess the goodness of fit using ROC (receiver operating characteristic) curve for model discrimination. The ROC curve is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate).and the KS Cutoff reference line is drawn at the value of 1-specificity where the greatest difference between sensitivity and 1-specificity is observed for the VALIDATE partition. The KS Cutoff line is drawn at the cutoff value 0.85, where the 1-specificity value is 0.182 and the sensitivity value is 0. 893.Cutoff values range from 0 to 1, inclusive, in increments of 0.05. At each cutoff value, the predicted target classification is determined by whether P_CBTS_infonormal, which is the predicted probability of the event "normal" for the target CBTS_info, is greater than or equal to the cutoff value. When P_CBTS_infonormal is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event.

CONCLUSION

In this work, we found that the EBDS score (Maternal mental health scale) is the most related factor to IBQ_R_VSF score (Infant behavior scale). Also, sleep duration is based on and affected by how the baby falls asleep. After proving that the CBTS score is the most relevant factor for EBDS information, we built a machine learning model to predict the impact of different variables on the CBTS results. In the future, we will work more on building models to predict the other maternal mental health scales against infant behavior and the infant behavior itself based on the maternal mental health status.

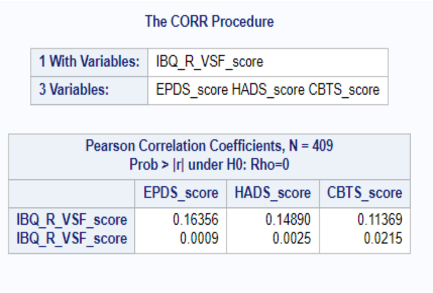
REFERENCES

- HADS (2).pdf
- Edinburgh Postnatal Depression Scale (EPDS) Calculator (perinatology.com)
- City Birth Trauma Scale (all questions and diagnostic subscales) | Right Decisions (scot.nhs.uk)
- The Infant Behavior Questionnaire-Revised: Factor Structure in a Culturally and Sociodemographically Diverse Sample in the United States - PMC (nih.gov)
- https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/procstat/procstat_corr_toc.htm

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Waleed Ead for her supervision and support throughout our work.

APPENDIX



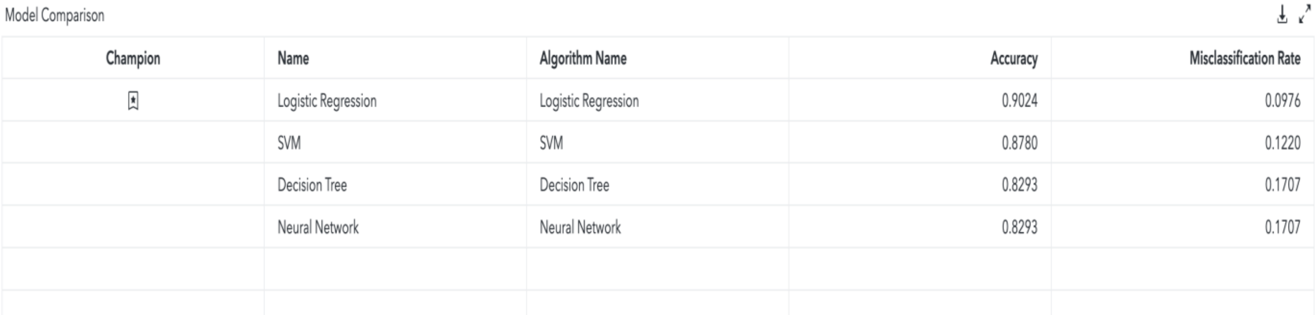
The CORR Procedure

1 With Variables:	IBQ_R_VSF_score
3 Variables:	EPDS_score HADS_score CBTS_score

Pearson Correlation Coefficients, N = 409
Prob > \|r\| under H0: Rho=0

	EPDS_score	HADS_score	CBTS_score
IBQ_R_VSF_score IBQ_R_VSF_score	0.16356 0.0009	0.14890 0.0025	0.11369 0.0215

Table 1: Shows the correlation between EPDS, HADS, CBTS scores and IBQ_R score.



Model Comparison

Champion	Name	Algorithm Name	Accuracy	Misclassification Rate
★	Logistic Regression	Logistic Regression	0.9024	0.0976
	SVM	SVM	0.8780	0.1220
	Decision Tree	Decision Tree	0.8293	0.1707
	Neural Network	Neural Network	0.8293	0.1707

Table 2: Shows that the champion model is Logistic Regression.

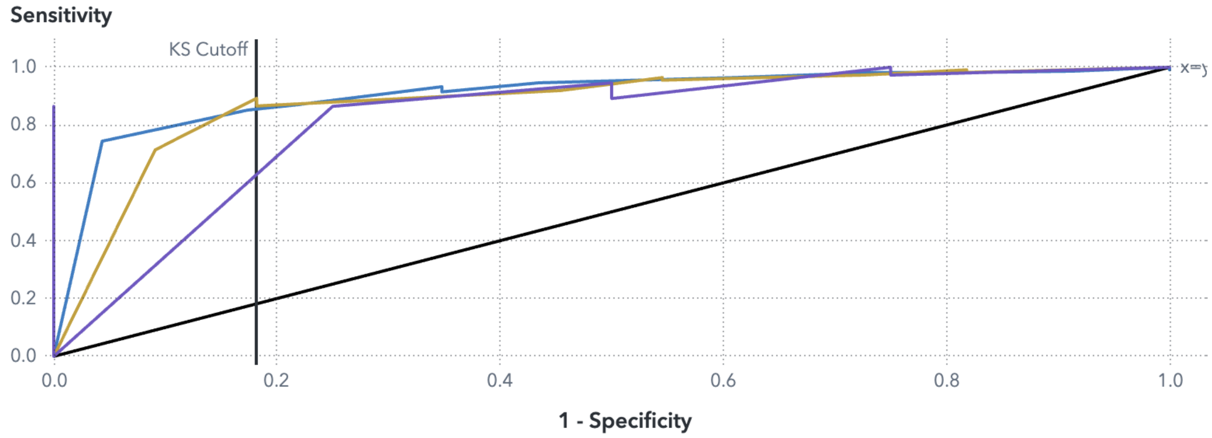


Figure 4: The ROC curve is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate) of the champion model (logistic regression).