Assignment for PMIM202

Swansea University
Prifysgol Abertawe
Medical School
Ysgol Feddygaeth

Module number:	PMIM202
Module name:	Health Data Modelling
Title of assignment:	A1 – Research Report + Summary
Student ID number:	2311233
Word count:	2500+500
Declaration:	I understand the following conditions which apply throughout this course: 1. I confirm that I am the sole author of this work. 2. I understand that proof reading by a third party is discouraged, but if used, records should be available as per guidelines. 3. I understand the need for academic integrity and that all my submitted work will adhere to its principles. 4. I understand that the teaching team will take measures to deter, detect and report any academic misconduct. 5. I agree to my work being submitted to the TurnItIn academic database. 6. I understand the importance of assignment deadlines and the need to seek help in good time where personal circumstances interrupt my work.
Please copy and paste this declaration onto the front of the submission.	

Report Summary for Exclusive Breastfeeding (EBF): Focus cohorts to improve EBF rates using significant factors affecting "Intention to breastfeed"

Previous studies suggest that, physical, mental, social, economic factors and COVID directly influence the intention to breastfeed. Intention to breastfeed is found to be directly associated to exclusive breast feeding. Women with the intention to exclusive breastfeed are more likely to exclusively breastfeed for the recommended initial 6 months. This study revolves around the idea of improving the exclusive breastfeeding rates by focusing on those group of people who have no intention of breastfeeding. Encouraging them to choose the intention to breastfeed might increase exclusive breastfeeding rates. Exclusive breastfeeding is important because a child develops more immunity from breastmilk stated by other research studies. Such development has long term positive effect on individuals' life. Not limited to individual, but society, health services, finance system and environment as well. In this study, the analysis was performed to understand if the factors considered as important in other studies are also turning up to be important on the data surveyed on Welsh region for the period of 2020 – 2022. Further, the analysis intends to check if these important factors affect the duration of exclusive breastfeeding in the postnatal phase. Also, the analysis was carried out to find which group of women has the highest odds of not having the intention to exclusive breastfeeding, that is they plan to prefer alternate methods like bottle feeding or mixed feeding. The limitations on understanding of the factors are limited to dataset and the conclusions associated with similar studies.

Following are the suggestions to create an action plan based on the identified factors,

1. **Weight Management** programmes could be developed for maintaining healthy weight plans. Women with overweight BMI has highest odds of not having the intention to breastfeeding.
2. **Quit Smoking** awareness programmes to curb smoking habits could prove helpful in longer duration of breastfeeding. Women who smoke greater 4 cigarettes / day or week has highest odds of losing the intention to breastfeeding.
3. **Family** awareness programmes to educate the family members to encourage pregnant women in the family to breastfeed.
4. **Update to date training** programmes to midwives and consultant led care should be provided as majority of the women select either midwives or consultant or combination of both as support during pregnancy phase. Negative experience or inadequate awareness shared by the care providers could lead to highest odds of losing the intention to breastfeeding.
5. **Mental Wellbeing** sessions to women who experience restlessness and anxiety emotions on a every day and several days basis could increase the odds of intention to breastfeeding.
6. **Work-Life** balance sessions could help working pregnant women better manage stress from work and household.
7. **Couple Therapy** could be provided to have positive effect on intention, initiation, and continuation of exclusive breastfeeding as married and living with partner women has the highest odds of not breastfeeding.

Exclusive Breastfeeding (EBF): Focus cohorts to improve EBF rates using significant factors affecting “Intention to breastfeed”

ABSTRACT:

Background
EBF is deemed mandatory at least for the first 6 months of infancy by WHO and national governmental bodies. It affects not only at individual levels but also at social, economic, governmental, and environmental levels. Research studies state “Intention to Breastfeed” is a strong predictor for EBF.

Aim
This study aims to understand the relevance of precedent factors associated to “Intention to Breastfeed”, report new ones. Identify factors affecting duration of breastfeeding through investigation on the 2020-2022 breastfeeding survey. Create focus cohorts to target groups where special care is needed and thereby increase EBF rates.

Method
Data processing and feature engineering were performed to clean and encode the data. To identify the significant variables, data was modelled using regression analysis with logistic regression and decision tree. Secondary analysis, ROC analysis for model confirmation and model diagnostics to exclude Influential – Leverage – Outlier samples to improve model quality. Post hoc analysis to create focus cohorts and determine the Odds Ratio (OR) of the outcome. Survival analysis using Kaplan Meier Estimate and Cox regression to identify breastfeeding duration and factors affecting the duration were conducted.

Result
The identified factors affecting intention to breastfeed are **weight / BMI**; Odds Ratio (OR: 3.09; 1.43), **smoking status** (5.73; 2.13), **other members of family** (4.01; 1.80; 1.21; 1.12), **type of maternity care** (4.37; 1.37; 0.70), **restlessness** (2.34; 1.70; 1.08; 0.87), **anxiety** (1.05; 1.08; 0.98; 0.87), **working status** (1.09; 1.01) and **relationship status** (3.34; 1.26; 1.08; 0.85) The OR is calculated as Intention to NOT EBF against Intention to EBF. Multiple OR corresponds to multiple focus cohorts made from multiple levels of factors. Survival analysis concludes that probability **of EBF for > 5 months is 50%** and **relationship status** is a factor that affect duration of breastfeeding.

Conclusion
The factors identified in results are relevant and in agreement with previously conducted studies. No new factor was found. The focus cohorts suggest which group of women could be targeted to improve EBF rates based on IV. This can be achieved by designing and updating programmes to better support and improve the experience during pregnancy, aid in prenatal and postnatal preparations. It is required to build programmes to encourage with idealistic expectations as oppose to unrealistic ones.

INTRODUCTION:

The World Health Organization (WHO), and Welsh Government recommends nurturing infants with EBF, especially in the first 6 months, mandatory for complete nourishment of infant and lifelong optimal health (1). Earlier studies associate increase in rates of EBF with increase in chance of prevention of an illness throughout the period of an individual’s life and reduced cost incurred at the health services. Evidence shows that benefits of EBF are not only limited to physical, mental health but also to financial and environmental benefits (2). It is important to consider significant investments to be placed for services, support and awareness programmes, training programmes and schemes to attain higher EBF rates which is likely to provide significant short and long term returns (3).

Wales infant breastfeeding survey conducted in the year 2010 reported that there has been increase in the EBF rate in the region (4). However, the survey dates to a decade. There are recent studies which suggest that EBF is discontinued or substituted with other means of feeding. The alternative methods

include bottle feeding and mixed feeding. For the period 2019-2021 during the COVID pandemic period there has been an increase in rates of EBF choices for Welsh region. In addition, “Intention to EBF” is established as the most likely predictor for EBF for the duration of 6 months. Studies suggests it is directly associated with social and economic factors (5). This study aims to retest precedented factors on the 2020-22 breastfeeding survey, report new factors and identify factors affecting the duration of the breastfeeding postnatal and identify cohorts. The factors could be used to create or update support and awareness programmes. These programmes could provide encouragement to choose breastfeeding as an option during the pregnancy phase (6). The factors could be used to update or create new guidelines for maternity care service providers to help implement better policies and training for midwives and other consultants.

This study uses statistical approaches including the regression analysis, secondary analysis, post hoc analysis, and survival analysis to detect the significance of the factors associated with the intention to EBF.

METHOD:

Dataset

The flat file consists of 753 rows and 73 columns. “*How are you planning to feed your baby*?” column served as the Dependent Variable (DV). The dataset consisted of participants from the Welsh region. Figure 1, represent the demographic distribution of the participants in the dataset. Majority of the cohort has British nationality. The information was collected in the form of online questionnaire (7). The flat file consisted of questions could be categorized into sections as personal information, mood, health and well-being, pregnancy, work, covid, period of stress, children, home. WIMD is an additional column that indicates the region deprivation statistic (8).

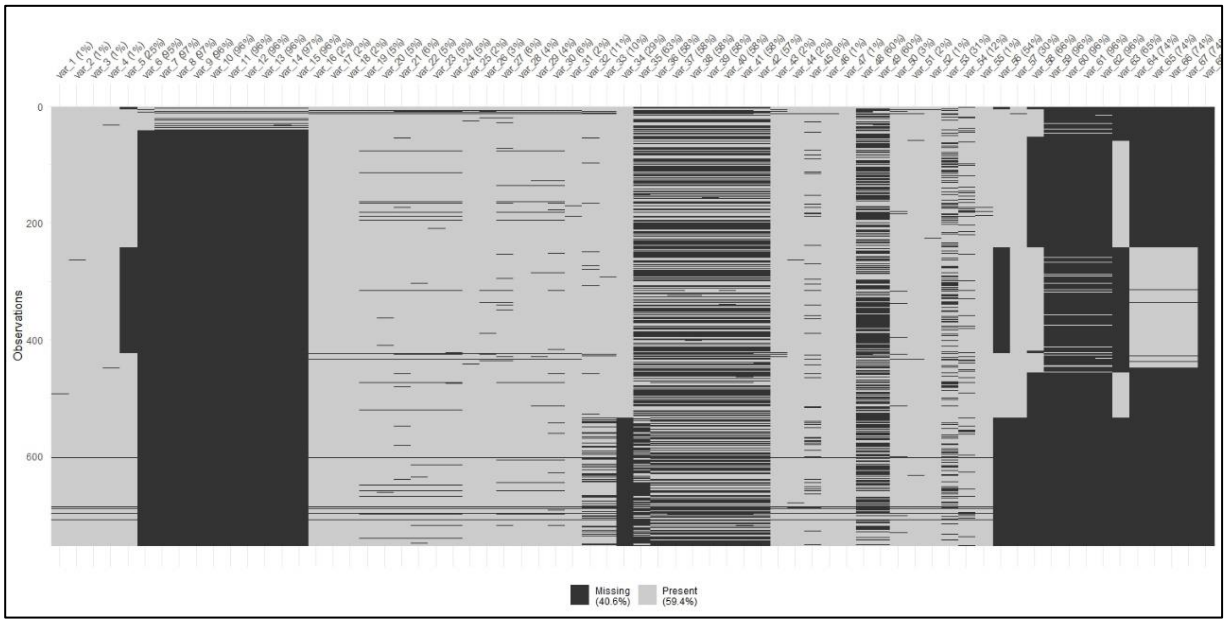
Figure 1. Demography of the dataset

Labels	Nationality										
	British	Welsh	European	British & Other	Indian	Irish	American	New Zealand	Greek	Brazilian	Filipino
Total Percent	88.80%	6.45%	2.28%	0.81%	0.27%	0.27%	0.13%	0.13%	0.13%	0.13%	0.13%
Ethinicity											
White	85.90%	6.45%	2.28%	0.81%	--	0.27%	0.13%	0.13%	0.13%	0.13%	--
Mixed	2.15%	--	--	--	--	--	--	--	--	--	--
Asian	0.54%	--	--	--	0.27%	--	--	--	--	--	0.13%
Chinese	0.13%	--	--	--	--	--	--	--	--	--	--
Others	0.08%	--	--	--	--	--	--	--	--	--	--
Employed											
Yes	61.70%	4.57%	1.48%	0.54%	0.27%	--	0.13%	0.13%	0.13%	0.13%	0.13%
No	27%	1.88%	0.80%	0.27%	--	0.27%	--	--	--	--	--
Others	0.1%	--	--	--	--	--	--	--	--	--	--
Education											
Higher Ed.	58.90%	4.97%	1.75%	0.81%	0.27%	0.27%	0.13%	0.13%	0.13%	0.13%	0.13%
Further Ed.	7.66%	0.81%	0.13%	--	--	--	--	--	--	--	--
Secondary Ed.	21%	0.67%	0.27%	--	--	--	--	--	--	--	--
Primary Ed.	--	--	--	--	--	--	--	--	--	--	--
NEET	0.40%	--	--	--	--	--	--	--	--	--	--
Others	0.84%	--	0.13%	--	--	--	--	--	--	--	--
Income											
< 10K	5.30%	0.27%	--	--	--	--	--	--	--	--	--
10 to < 50K	45.73%	3.36%	1.61%	0.53%	--	--	0.13%	0.13%	0.13%	--	--
50K to < 100K	32.80%	2.25%	0.67%	0.27%	0.27%	0.27%	--	--	--	0.13%	0.13%
100K to < 200K	0.13%	--	--	--	--	--	--	--	--	--	--
>=200K	0.13%	--	--	--	--	--	--	--	--	--	--
Others	4.71%	0.57%	--	0.01%	--	--	--	--	--	--	--
Intention											
BreastFeed	43.20%	2.69%	1.48%	0.54%	0.27%	0.27%	0.13%	--	--	0.13%	--
NotBreastFeed	45.60%	3.76%	0.80%	0.27%	--	--	--	0.13%	0.13%	--	0.13%

Data Preprocessing

The preliminary steps of data preprocessing consisted of manual inspection, Feature Engineering (FE), Imputation, Visualization, and Validation using internet research. Manual inspection focused to fix data errors and explore the values of the columns. There were data errors which were identified through inspection of histogram visualization of the column. The errors were manually fixed.

Figure 2. Missing data in the dataset before encoding



After manual inspection, 68 columns were selected including the DV. Figure 2. shows the overall missing data in the dataset. For DV, approximately 1% of the data were NA values. The 1% of the data were removed considering no scope of imputation for the DV. Table 1. shows number of columns as per the percentage of missingness. Out of 68 columns, 32 columns have NA values greater than 40% and 30 columns have NA values less than 10%.

Table 1. Missingness of data as per columns

Percentage Range	<10%	10 - 20%	20 - 30%	30 - 40%	>40%
Number of columns	30	2	3	1	32

Studies on data exclusion states that with no scope of imputation, the data should be excluded when missing data crosses the upper limit threshold of 40%, and if the missingness could not be determined, that is, if it's of the type MCAR or MAR (9–11). Post data exclusion 744 rows and 36 columns remained.

FE steps were performed. Columns were categorized into ordinal, nominal, text, and numeric statistical data types. The ordinal and nominal variables were manually inspected and assigned levels based on logic and information about the variable from internet research. The nominal variables were encoded and converted into dummy variables, to incorporate reference encoding and converted them to factors (12,13). The ordinal variables were left unchanged in-order to preserve the order level information in the data after the values were converted from string to numbers. Text data type columns were encoded using vector representations. They were aggregated and transformed using TF-IDF algorithm (14). TF-IDF values represents each word in the sentences and the aggregated product represents the sentence. High cardinality variables increases complexity in model interpretation and data analysis (15). "*What is your occupation*?" is one such column. The unique levels in the column are more than 50% of the sample data, the levels of the column were reduced by using Target Encoding. The column was grouped into 21st percentile of the median of estimated coefficients of a model including this column and the DV, where 21

is the square root of column cardinality (16–18). A BMI variable was derived using the height and weight column.

Figure 3. Missing data after encoding

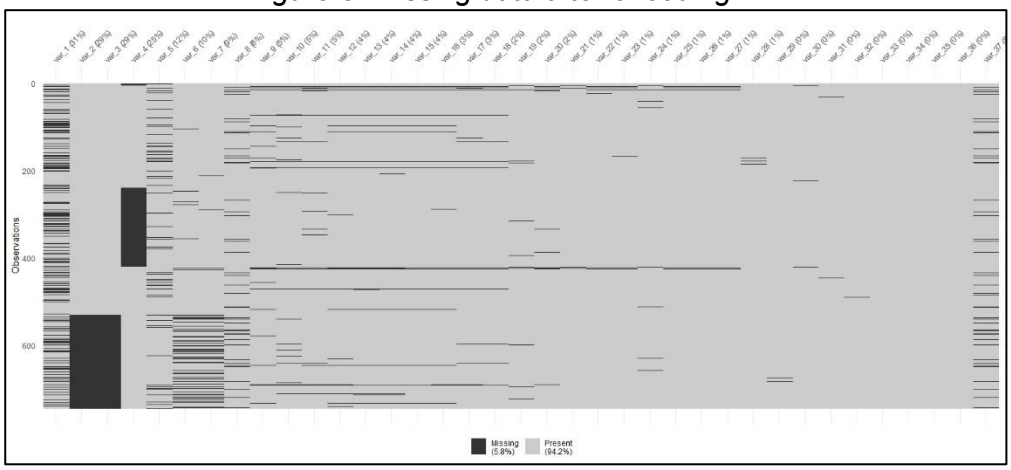


Figure 3. represents the missing data after data encoding. Imputation was performed using the Multiple Imputation using Chained Equations (MICE) technique (19). Outlier treatment process was carried out using winsorization process (20). There are 744 samples and 36 Independent Variable (IV) and 1 DV.

Data Modelling

The DV is binary with the proportion of 49% “Intention to Exclusive Breastfeed” and 51% “Intention to Exclusive Not Breastfeed”. The “Intention to Exclusive Not Breastfeed” samples were encoded as 1 (positive outcome) to find linear combinations of IV to determine those with higher probability of not EBF.

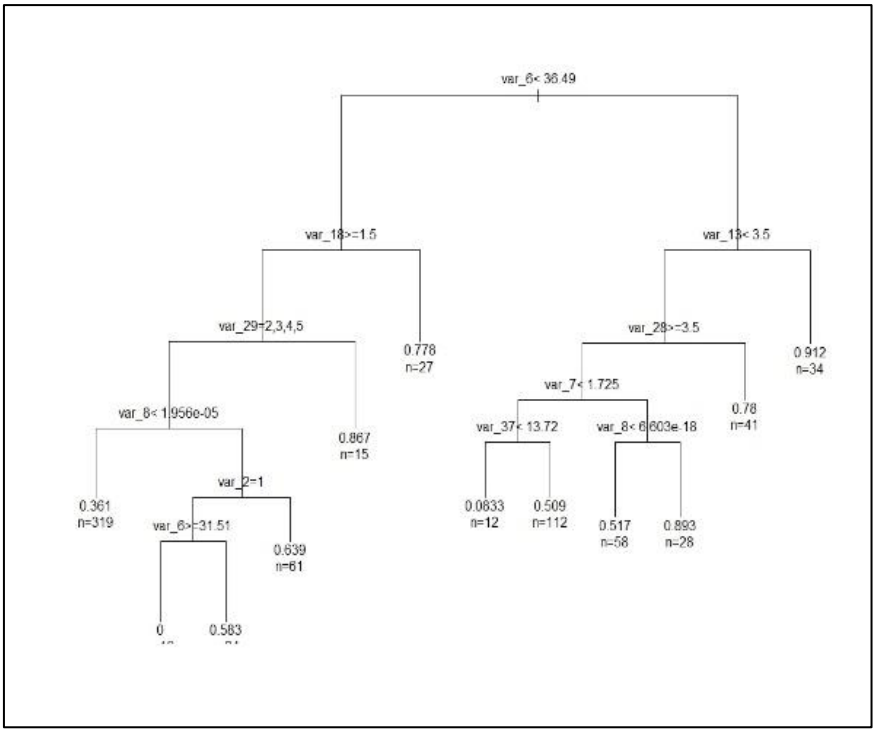
First, a Stepwise Multiple Logistic Regression (MLR) was modelled (21). The scope of the step algorithm was from intercept (null model) to a model with all IV. By including “*What is your occupation*?” an IV encoded using the DV, has a possibility of target leakage. Thus, the model was checked with and without this IV. Model2 produced the lowest AIC value. With k = 5 folds, k-fold cross validation (22). A Decision Tree (DT) model was estimated on the dataset. Table 2 shows the most significant variables extracted from each model. The variables from DT mentioned in the Table 2 are in the descending order of its variable importance.

Table 2. Significant variable list for Logistic Regression Model

Model Name	Variables	Occupation
MLR 1	**var_6, var_25, var_29, var_12, var_5, var_13**	Without
MLR 2	**var_19, var_6, var_21, var_13**	With
DT	var_6, var_25, var_31, var_29, var_30	Without

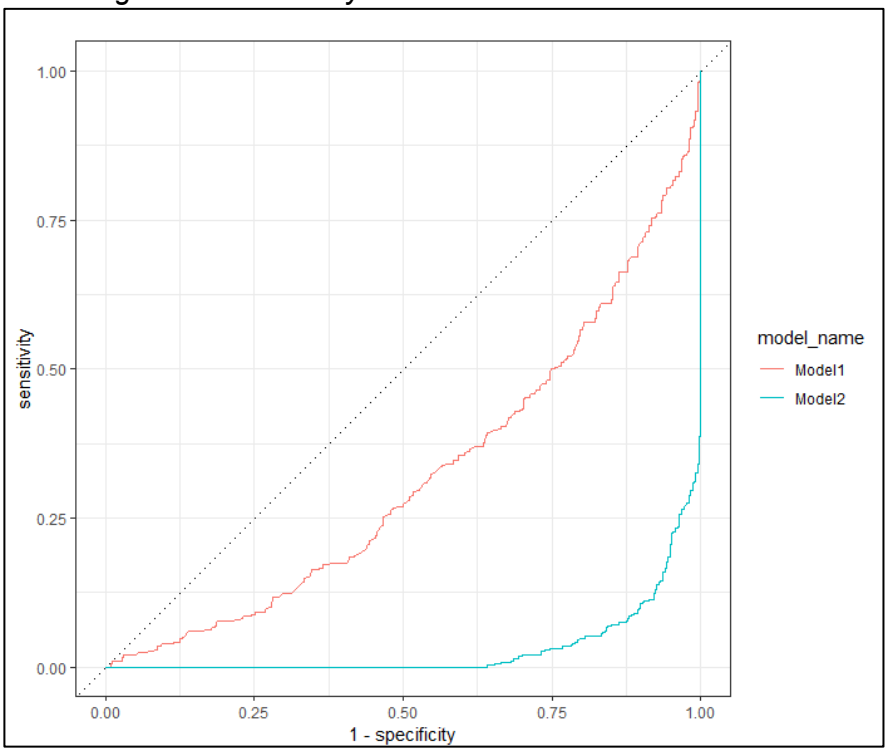
Estimates of MLR models, colour coded as green has positive association on the DV and the ones colour coded as orange has negative association on the DV. As shown in Figure 4, the decision rules include some of the variables that are found in the MLR models.

Figure 4. Decision Rules from Decision tree model



Second, secondary analysis was conducted with the significant IV from Table 2. A ROC Analysis was performed to choose from the two MLR models. It was found that the Model2 had a high False Positive Rate (FPR), see Figure 5 (23). Thus, Model1 is favoured.

Figure 5. ROC Analysis for Model 1 & Model 2



Further, in MLR Model1, adding interaction terms led to increase in multicollinearity. The significance of an IV which were used as interaction term were observed to increase (24). Thus, square & log transformations were used instead of interaction terms. The use of the transformations on the variables did not benefit as the linearity of model could not be established with IV being categorical predictors without increasing complexity. So, next step was to perform model diagnostics to understand the model quality. Cook's

Distance, Studentized tests, Bonferroni p-value, hat-values were iteratively examined to remove and reduce the Influential samples, Leverage samples, possible Outliers in the dataset that affected model prediction (25). Figure 6.1 gives an outline of diagnostics plots. Figure 6.2 shows marginal plots of how Model1 fits for significant IV.

Figure 6.1. Model Diagnostics Plot

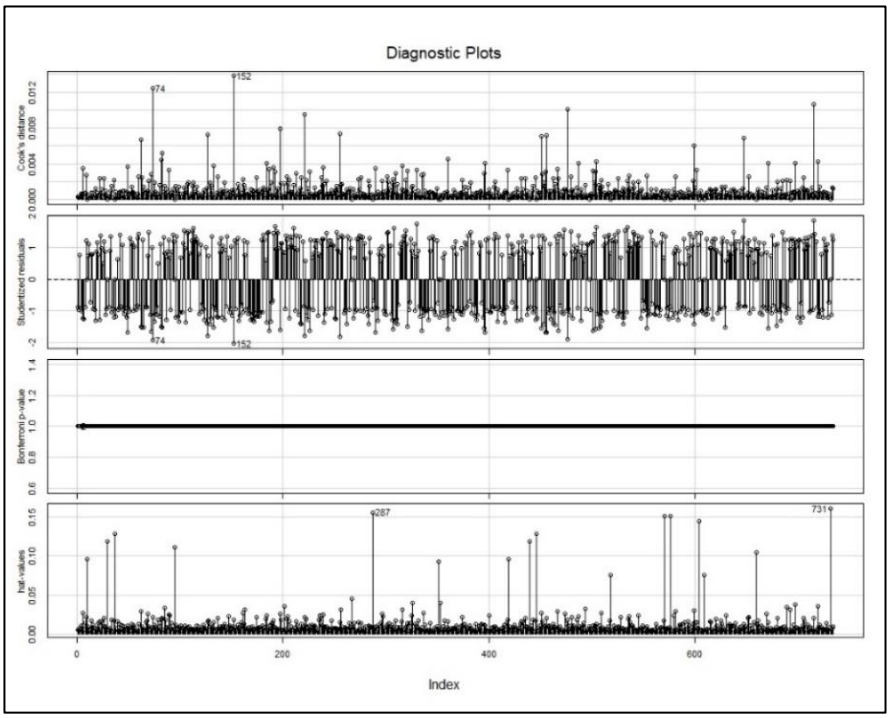
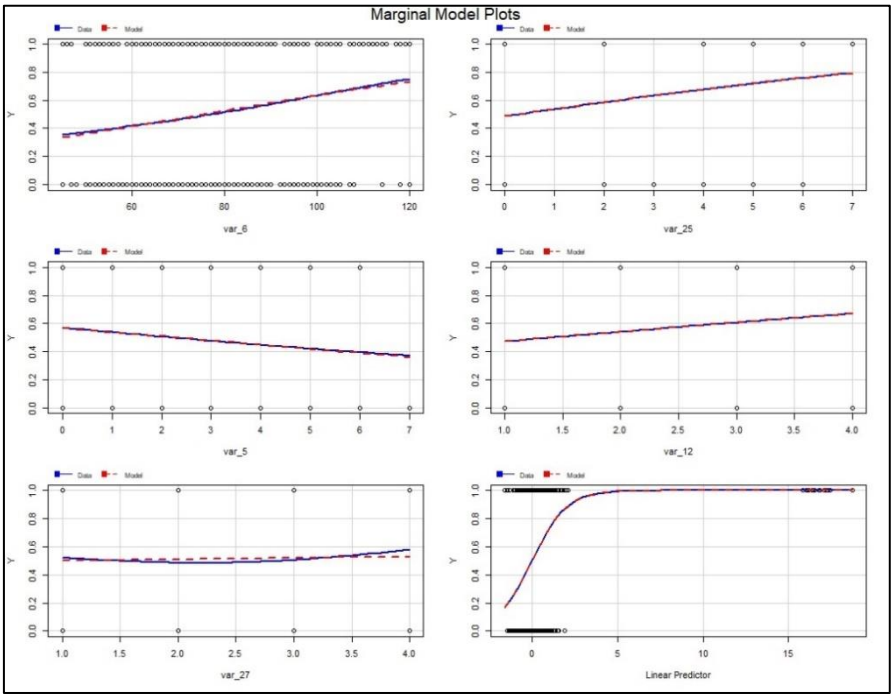


Figure 6.2. Model Diagnostics Marginal Plot



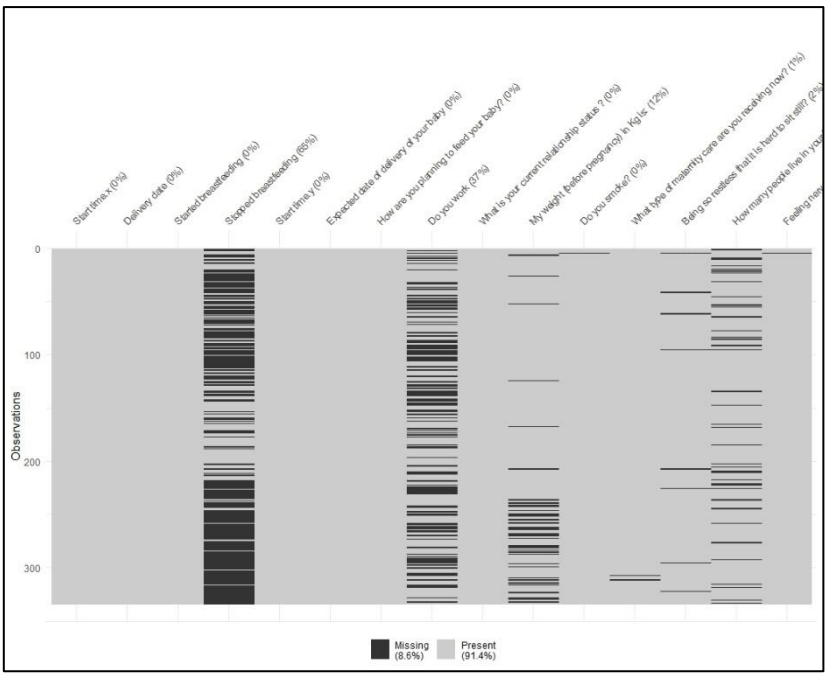
Post Hoc Analysis

Post hoc analysis was performed to create Focus Cohorts. Focus Cohorts are subset of dataset sliced using a contingency table. Odds Ratio (OR) were calculated based on these contingency tables. The OR is the ratio of odds of Intention to NOT EBF to the odds of Intention to EBF. Table 4 in the result section, outlines the details of focus cohorts.

Survival Analysis

The survival analysis was conducted by joining the “Born in Wales” and “Follow Up” flat files. 50% of the missing data with respect to the “Started Breastfeeding” column were dropped. Figure 7 shows the missing data after data exclusion. FE was carried out to encode the respective columns like that of data modelling. Status column was created using the “Stopped Breastfeeding” column. The status = 1 indicated available values and status = 0 as censored. Time column was created by subtracting the start and stop breastfeeding time and then converted them to months. Survival object was created and Kaplan-Meier Survival Analysis, Cox PHR was conducted with the significant IV. Relationship, working hours and working status IV were added for Cox PHR.

Figure 7. Missing data in dataset for Survival Analysis



RESULT:

Table 3 represents the model performance trained on K fold cross validation. Due to limitations of the dataset, the specificity of the model learning represents moderate performance.

Table 3. Model Metrics

Model Name	Accuracy K fold = 5	Kappa K fold = 5	Specificity	Sensitivity	95% CI
MLR Model1	0.6224	0.246	0.656	0.60	(0.59, 0.66)

Table 4 exhibits the odds ratio of the selected IV in the post hoc analysis. OR with higher than 1 means that women with “Intention to NOT EBF” are higher in the cohort. The table describes top 4 OR in the respective cohorts. Focus cohort column describes the cohort subcategory or levels with respect to the variable and the percentage of the size of the cohort with respect to the whole dataset.

Table 4. Odds ratio on selected variables

Variable	Not Exclusive to Exclusive Odds	Focus Cohort (size %)
var_6: Weight before pregnancy	3.09, 1.43	Weight category (in kg) > 100 (12%) 82 to 100 (20%)
var_25: Do you smoke	5.73, 2.13	Smoke < 5 cig/day (3.3%), Smoke > 5 cig/ week (2.8%)
var_5: How many people live in your home (not including you)?	4.01, 1.80, 1.21, 1.12	Family members >= 6 (0.5%), 5 (1.4%), 1 (37%) and 2 (36%)
var_29: What type of maternity care are you receiving now?	4.37, 1.37, 0.70	Midwife & Consultant (1.6%), Midwife only (50%), Consultant only (45%)
var_12: Being so restless that it is hard to sit still?	2.34, 1.70, 1.08, 0.87	Every day (5%), More than half days (8%), Several days (25%), Not all days (62%)
var_27: Feeling nervous, anxious or on edge?	1.05, 1.08, 0.98, 0.87	Every day (9 %), Several days (42%), More than half days (13%), Not all days (36%)
var_30: Are you currently working?	1.09, 1.01	No (30%) Yes (70%)
var_31: What is your current relationship status?	3.34, 1.26, 1.08, 0.85	Single (5.2%), Dating (3.3%), Living with partner (42%), Married (46%)

Figure 8. Predicted probabilities of Logistic Regression

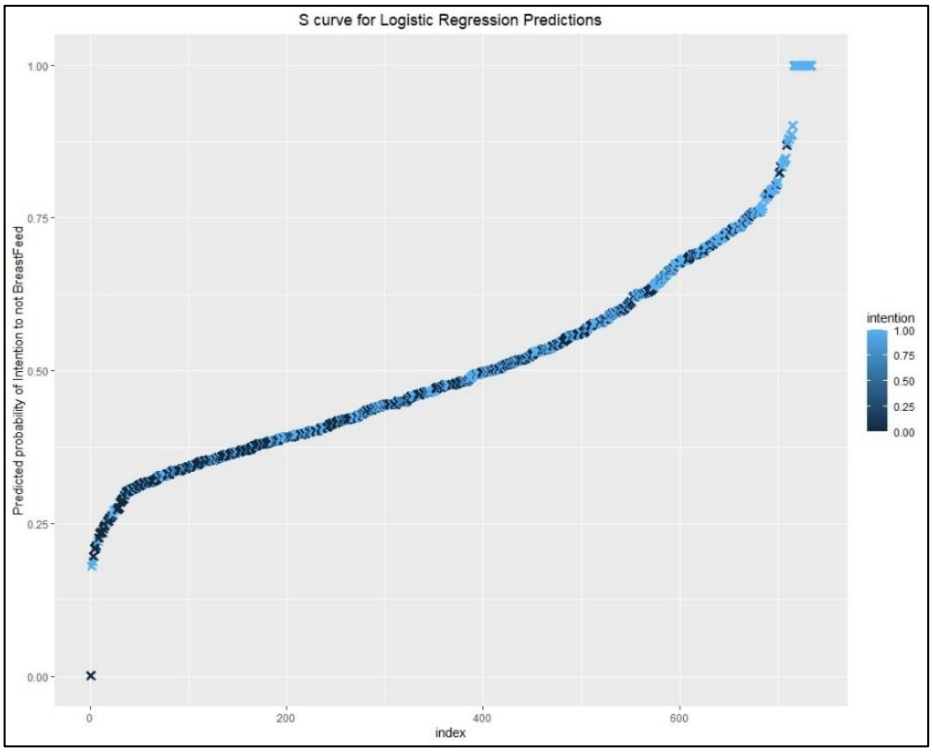


Figure 8 provides the overall idea of the distribution of predicted probabilities generated from the model. Figure 9 shows that the breastfeeding for pregnant women from the provided samples, has a 50% probability to exclusively breastfeed at least 4-5 months of duration.

Figure 9. Kaplan Meier Survival Analysis

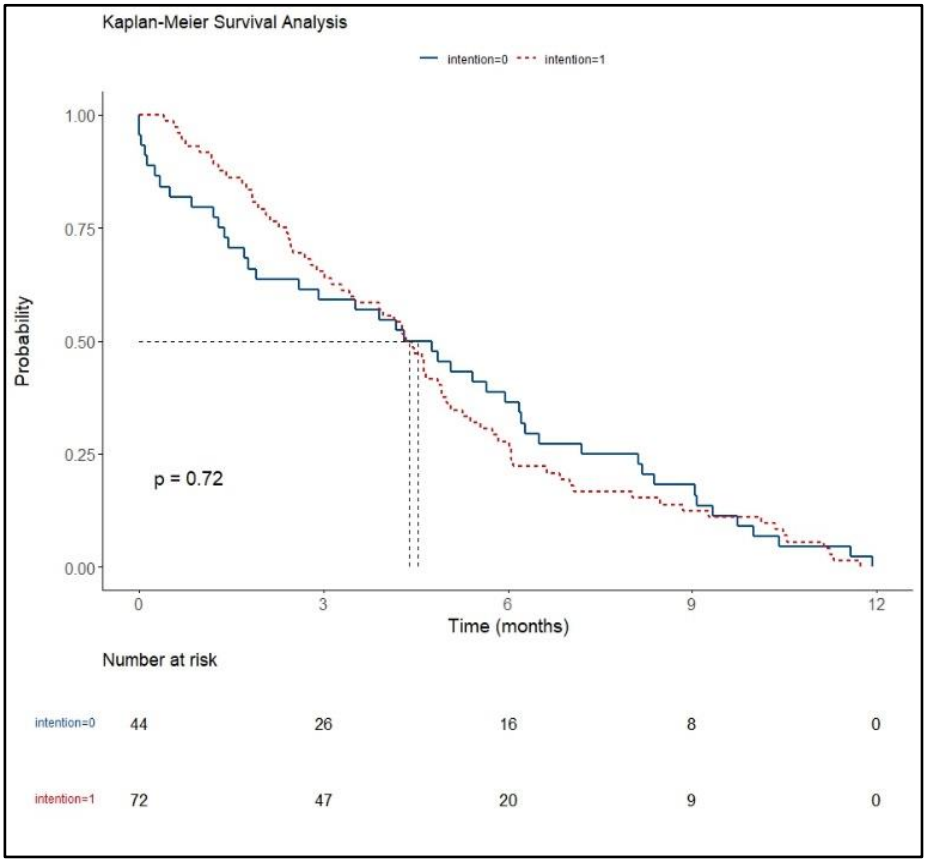
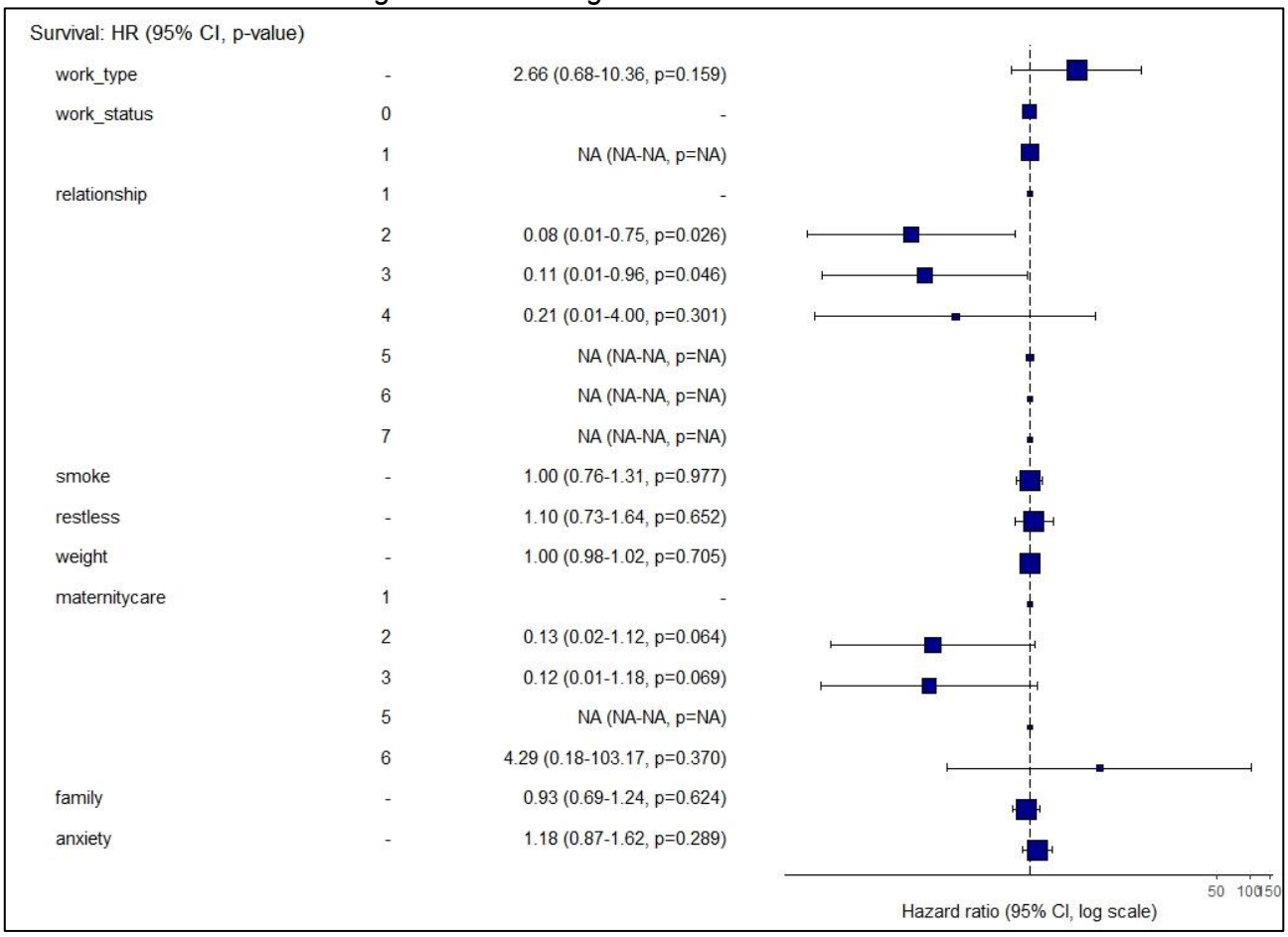


Figure 10. shows that in the cox regression, “Living with partner” encoded as 2 and “Married” as 3, type of relationship has a significant hazard / risk on EBF. Work status and work type are not resulted as significant hazard.

Figure 10. Cox Regression Hazard Ratio Plot



DISCUSSION:

The study was aimed to investigate the factors that are associated with intention to breastfeed and the factors that affect the duration of EBF.

The results show that

1. Data Modelling suggests IV specific to physical, mental, and social health and well-being have influence on “Intention to EBF,” no new factor was found.
2. Post hoc analysis through focus cohorts suggests the levels of the variables that has higher odds of NOT EBF to EBF. Table 4 shows cohort of women to be targeted for special care to increase EBF rates.
3. In Survival Analysis Kaplan Meier states that the probability of EBF for >5 months is 50% and Cox PHR states that relationship status affects the duration of EBF.
4. Covid-19 effects on the “Intention to EBF” cannot be determined as the investigation is limited by the dataset.

Due to data limitations, model has moderate performance. Adjusted models that fit the data more accurately could be the scope of future experiments, with the caution to not overfit the model on the given dataset.

The results support that,

1. **Weight / BMI** Special programmes could be developed for maintaining healthy weight plans. Women with weight > 82 kg has highest odds of not breastfeeding, it is consistent with other research studies (26,27).

2. **Smoking** studies show it affects duration of breastfeeding despite the intention to breastfeed (28). Programmes to curb smoking habits could prove helpful to increase duration of breastfeeding (29). Women who smoke >= 4 cigarettes / day or week has highest odds.
3. **People living with,** have direct a direct influence on EBF. Educating and including family members would positively encourage EBF (30). Women living with family members has highest odds.
4. **Type of maternity care** selected by women during pregnancy plays an important role as per studies (31). Midwives are more preferred as they are more trained in providing better support than others. Women selecting midwives, consultants or both should be considered as negative experience led to highest odds of not breastfeeding.
5. **Restlessness & Anxiety** could reduce EBF (32). It is essential to avoid negative emotions, a prevalent predictor for Nursing Aversion (33). Women who experience these emotions every day and several days has the highest odds of not breastfeeding.
6. **Working status** are not conducive to reduced EBF as per studies (34).
7. **Relationship Status** has positive effect on intention, initiation, and continuation of EBF (35). Married and living with partner women has the highest odds of not breastfeeding.

There are studies that shows that unrealistic expectations were set by the professional practitioners for the pre-natal, post-natal, antenatal preparations (36). So, during preparation of the guidelines and policy making that encompasses above mentioned factors, preparation and support should be provided as per idealistic expectations.

REFERENCES:

1. Exclusive breastfeeding for six months best for babies everywhere [Internet]. [cited 2024 Jan 3]. Available from: https://www.who.int/news/item/15-01-2011-exclusive-breastfeeding-for-six-months-best-for-babies-everywhere

2. Jones K. BCUHB Infant Feeding Strategic Plan 2019.

3. Preventing disease and saving resources: the potential contribution of increasing breastfeeding rates in the UK.

4. McAndrew F, Thompson J, Fellows L, Large A, Speed M, Renfrew MJ. Infant Feeding Survey 2010: Summary. 2012;

5. Santacruz-Salas E, Segura-Fragoso A, Cobo-Cuenca AI, Carmona-Torres JM, Pozuelo-Carrascosa DP, Laredo-Aguilera JA. Factors Associated with the Abandonment of Exclusive Breastfeeding before Three Months. Children. 2020 Dec 16;7(12):298.

6. Jones H, Seaborne M, Mhereeg M, James M, Kennedy N, Bandyopadhyay A, et al. Intention to breastfeed and association with subsequent breastfeeding duration, a linked population level routine data study - The Born in Wales cohort 2018-2021 [Internet]. medRxiv; 2022 [cited 2024 Jan 8]. p. 2022.12.13.22283407. Available from: https://www.medrxiv.org/content/10.1101/2022.12.13.22283407v1

7. Born In Wales - Information Sheet for Mums [Internet]. NCPHWR. [cited 2024 Jan 3]. Available from: https://ncphwr.org.uk/healthy-families-information-sheet/

8. Welsh Index of Multiple Deprivation [Internet]. [cited 2024 Feb 4]. Available from: https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation

9. Mack C, Su Z, Westreich D. Types of Missing Data. In: Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [Internet] [Internet]. Agency for Healthcare Research and Quality (US); 2018 [cited 2024 Jan 19]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK493614/

10. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019 Jun 1;110:63–73.

11. Dong Y, Peng CYJ. Principled missing data methods for researchers. SpringerPlus. 2013 May 14;2(1):222.

12. Polissar L, Diehr P. Regression analysis in health services research: the use of dummy variables. Med Care. 1982 Sep;20(9):959–66.

13. Grotenhuis MT, Thijs P. Dummy variables and their interactions in regression analysis: examples from research on body mass index.

14. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Doc. 2004 Oct;60(5):503–20.

15. Cerda P, Varoquaux G. Encoding high-cardinality string categorical variables. IEEE Trans Knowl Data Eng. 2022 Mar 1;34(3):1164–76.

16. Bruce P, Bruce A, Gedeck P. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python [Internet]. O'Reilly Media; 2020. Available from: https://books.google.co.uk/books?id=F2bcDwAAQBAJ

17. Zumel N, Mount J. vtreat: a data.frame Processor for Predictive Modeling [Internet]. arXiv; 2019 [cited 2024 Jan 23]. Available from: http://arxiv.org/abs/1611.09477

18. Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explor Newsl. 2001 Jul;3(1):27–32.

19. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011 Feb 24;20(1):40–9.

20. Aguinis H, Gottfredson RK, Joo H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organ Res Methods. 2013 Apr 1;16(2):270–301.

21. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis [Internet]. Springer New York; 2013. Available from: https://books.google.co.uk/books?id=7D0mBQAAQBAJ

22. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques [Internet]. Elsevier Science; 2011. Available from: https://books.google.co.uk/books?id=bDtLM8CODsQC

23. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Casp J Intern Med. 2013;4(2):627–35.

24. Cortina JM. Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. J Manag. 1993 Dec 1;19(4):915–22.

25. Zhang Z. Residuals and regression diagnostics: focusing on logistic regression. Ann Transl Med. 2016 May;4(10):195.

26. Turcksin R, Bel S, Galjaard S, Devlieger R. Maternal obesity and breastfeeding intention, initiation, intensity and duration: a systematic review. Matern Child Nutr. 2012 Aug 20;10(2):166–83.

27. Chen CN, Yu HC, Chou AK. Association between Maternal Pre-pregnancy Body Mass Index and Breastfeeding Duration in Taiwan: A Population-Based Cohort Study. Nutrients. 2020 Aug 7;12(8):2361.

28. Ariz U, Gutierrez-De-Terán-Moreno G, Fernández-Atutxa A, Montero-Matía R, Mulas-Martín MJ, Benito-Fernández E, et al. Despite intention to breastfeed, smoking during pregnancy is associated with shorter breastfeeding duration. J Neonatal Nurs. 2023 Apr 1;29(2):334–40.

29. Donath S, Amir L, Team the AS. The relationship between maternal smoking and breastfeeding duration after adjustment for maternal infant feeding intention. Acta Paediatr. 2004;93(11):1514–8.

30. Lok KYW, Bai DL, Tarrant M. Family members' infant feeding preferences, maternal breastfeeding exposures and exclusive breastfeeding intentions. Midwifery. 2017 Oct;53:49–54.

31. Balyakina E, Fulda KG, Franks SF, Cardarelli KM, Hinkle K. Association Between Healthcare Provider Type and Intent to Breastfeed Among Expectant Mothers. Matern Child Health J. 2016 May 1;20(5):993–1000.

32. Shao S, Yan S, Zhu P, Hao J, Zhu B, Tao F. Persistent Pregnancy-Related Anxiety Reduces Breastfeeding Exclusiveness and Duration: A Prospective Cohort Study. Breastfeed Med. 2022 Jul 1;17(7):577–83.

33. Yate ZM. A Qualitative Study on Negative Emotions Triggered by Breastfeeding; Describing the Phenomenon of Breastfeeding/Nursing Aversion and Agitation in Breastfeeding Mothers. Iran J Nurs Midwifery Res. 2017;22(6):449–54.

34. Tsai SY. Impact of a Breastfeeding-Friendly Workplace on an Employed Mother's Intention to Continue Breastfeeding After Returning to Work. Breastfeed Med. 2013 Apr;8(2):210–6.

35. Gibson-Davis C, Brooks-Gunn J. The Association of Couples' Relationship Status and Quality With Breastfeeding Initiation. Princet Univ Woodrow Wilson Sch Public Int Aff Cent Res Child Wellbeing Work Pap. 2007 Dec 1;69.

36. Fox R, McMullen S, Newburn M. UK women's experiences of breastfeeding and additional breastfeeding support: a qualitative study of Baby Café services. BMC Pregnancy Childbirth. 2015 Jul 7;15:147.