The Role of Social Networks and Personal Characteristics in Shaping Fertility Intentions: A Multi-Method Machine Learning Perspective

Ezgi GÜNBATAR
Applied Data Science Master Thesis

Supervisor: Dr. Javier Garcia Bernardo

Utrecht University
July, 2024

Table of Contents

Abstract

This study aims to explore how individual and social network characteristics influence women's desire to have children. For this purpose, individual and network attributes were used for predicting fertility intentions. The data was acquired from the "Social Networks and Fertility Research" at the LISS panel. The sample of the study consists of 738 Dutch women aged 18 to 40, and their individual and social network characteristics. In total, over 18,000 relationships were collected from respondents. First, a Graph Neural Network (GNN) was applied to grasp the effect of the network variables. Since the accuracy of the GNN performed low (with 0.30- 0.40 accuracy) and it was able to produce prediction only for one class the other machine learning methods were used. The methods, Histogram-Based Gradient Boosting Decision Tree (HGBT), Support Vector Machine, and Random Forest were trained and tested with 5-fold cross-validation also Grid Search CV hyperparameter tuning was implemented for reaching the best parameters. HGBT performed the best among all, so this model was used in further steps to describe the relations. It was found that individual characteristics, especially age and family pressure, had a more significant impact compared to network variables. On the other hand, while not as influential as individual attributes, network variables also demonstrated a significant role in explaining fertility intentions. The results showed that the network variables, such as the total number of children in women's networks, the number of people who want to have children, their connections with these individuals, and the frequency of their contact also have an influence.

Keywords: *Social network analysis, predicting fertility preferences, graph neural networks, histogram-based gradient boosting decision tree.*

1. Introduction

Human behavior is complex and not always easy to explain. One of these concepts is fertility behaviors. The reason why researchers want to explain people's fertility behavior is because it affects many areas such as economics, demographics, sociology, and politics. In this context, in order to explain people's fertility behavior, studies have been conducted on individual and environmental factors that may affect this.

1.1.Literature Review

The preferences and actions of others shape people's desires about having children. Individuals do not act in isolation but are embedded in a network of social relations. They exchange information, learn, transmit, negotiate, and challenge social norms through social interactions with their network partners (Keim et al, 2009). Research by Coale and Watkins (1986) and Bongaarts and Watkins (1996) demonstrates that fertility is influenced not only by individuals' characteristics but also by the behaviors of their social circle.

Social interaction is discussed in two distinctions in the literature: Social learning and social influence. “Social learning” involves the exchange and collective evaluation of information and ideas within a network, leading to behavior change by reducing perceived risk and uncertainty. Social influence is the process where individuals conform to the expectations of gatekeepers and promoters of social norms to gain approval and avoid conflict within their social group (Madhavan et al., 2003). To give an example specifically in this field, acquiring social judgments about when is the right time to have children and start a family and learning this implicitly (Kavas & De Jong, 2020). There is also a third mechanism called “social (emotional) contagion” which describes catching an idea or behavior from another person unconsciously in an emotional way. For example, women may experience emotional arousal when they spend time with babies in their social circles, and this positive emotion may trigger their own desire to have a family or baby (Bernardi & Klaerner, 2014).

The study by Richter et al. (2012), indicating that social contagion is related to birth rates, shows that the proportion of people in the network with children under 3 positively affects the respondent's likelihood of giving birth to a second child.

Stulp et al. (2023), also focus on the impact of networks and individual characteristics on fertility preferences. In the paper, a data-driven approach is employed using the data of Dutch women with over 18,000 relationships. LASSO regression is applied to understand how well it can predict five different outcomes (parent pressure, friend pressure, children in the future, happiness related to having children, ideal number of children) relating to fertility preferences, and which variables are most important in explaining these different outcomes. As a result, for all outcome measures,

individual characteristics are the strongest predictors. The most effective (in a negative relation) predictors for 'having children in the future' outcome variables are two individual characteristics: number of children and age. When predicting the likelihood of having children in the future, individual and network variables explained 40% of the variation, but network variables alone had no explanatory effect. However, this does not imply that networks are unimportant.

Using the same dataset as Stulp et al. (2023), this study was also focused on a similar research question. New approaches and models were experimented with to investigate the impact of network variables on fertility intentions.

1.2. The Study & Research Question

In the previous study, due to the lack of significant impact found for network variables, models that could better analyze the network structure were initially considered. First, it was planned to try a Graph Neural Network model that can analyze the network structure as a whole to determine whether there is any information loss when converting network information into structural variables.

The study aims to find an answer to the "Do the individual characteristics and the social network attributes affect women's desire to have children?" question. For this purpose, individual and network characteristics are used for predicting the having children intentions.

After discussing the factors influencing fertility behaviors and tendencies, and examining the approaches and studies related to this topic, the following section introduces the data of the study and provides information about the variables. In the 3rd section, details about the analyses and analytical methods used in the study are provided. Finally, the results of the analysis regarding which factors are relevant for fertility decisions and the conclusions are presented in the 4th section.

2. Data

The data was provided from LISS (Longitudinal Internet Studies for the Social Sciences) panel administered by Centerdata (Tilburg University, The Netherlands). The LISS panel is an online research infrastructure in the Netherlands, representative of the Dutch population. It allows for the collection of new data, downloading of existing data, and conducting innovative experiments. Every year the panel is carried out on ten core surveys that cover a wide range of topics. (LISS,2024)

The survey named "Social Networks and Fertility Research" organized by Stulp et al. (2023) at the LISS panel was used for the study. The sample of the study consists of 738 Dutch women aged 18 to 40 from households where at least one member speaks Dutch, out of a total of 1332 invited to participate. The respondents were asked two kinds of questions. In the first part of the questionnaire, participants answered questions about themselves such as age, education, number of children, partnership status, origin, fertility intentions (outcome variable), etc. In the second part of the questionnaire, they were asked to list 25 individuals, 18 years or older with whom they had contact in the last year. Then respondents gave answers about these people's characteristics like the type of their relationship, sex, frequency of contact, closeness, having children preferences, whether those 25 people had contact with one another, etc. In total, over 18,000 relationships were collected from 738 women.

In the study, each respondent is meant as "ego", and the people in the respondents' network (25 individuals) are meant as "alters". So, attributes that come from the first block of the questionnaire, (information about the respondent herself) are called "ego attributes" and the attributes that come from the second block of the questionnaire (information about the respondent's personal network) are called "alter attributes". Here, the answer to the question "*Do you think you will have (more) children in the future?*" is the outcome of the study is called the "childwish" attribute. More information about the survey questions and answers can be found in the codebook (Stulp, 2020).

2.1.Data Preprocessing & Variables

Before starting to analyze, the pre-processing steps applied in Stulp et al. (2023) were utilized. The R package FertNet was used (Stulp 2023a) (Stulp 2023b) to process the data, igraph and tidygraph packages were used for calculating the network variables. In this context, units with the following characteristics were removed from the data,

- reported fewer than 25 alters or mentioned the same person twice,
- more than 10 missing on alter attributes,
- reported no existing ties between the alters.

The answers containing strings in ego and alter attributes were labeled and converted into categorical variables. In addition, for the missing attributes, mean imputation was applied if the variable was numerical, and mode imputation was applied if it was categorical. In attributes expressing measurements such as degree and density in network variables, missing values were filled with 1, which represents the smallest degree.

The variables used in the study are defined and grouped as follows:

- **Ego variables:** attributes that come from the first block of the questionnaire (information about the respondent herself)

Variable Name	Description / Question
age	"How old are you?"
partner_num	"Do you currently have a partner?" 1 = Yes 0 = No
has_child_num	"Do you have children?" 1 = Yes 0 = No
educ_bin	1= High education level 0 = Low education level
child_num	"How many children do you have?"
relationship_duration_num	"How long are you in a relationship with your partner?"
cohabiting_num	"Do you live together with your partner?" 1= Yes 0= No
cohabitation_form_num	"What kind of a cohabitation form do you have with your partner?" 1= Marriage 2 = Registered partners 3= No formal cohabitation form 0= Other
pressure_f_num	"To what extent do you agree with the following statements: Most of my friends think that I should have (more) children:" 1= Completely disagree 2= Disagree 3= Somewhat disagree 4= Neither agree nor disagree 5= Somewhat agree 6= Agree 7= Completely agree 0= I don't know

pressure_p_num	"My parents/ caretakers think that I should have (more) children:" 1= Completely disagree 2= Disagree 3= Somewhat disagree 4= Neither agree nor disagree 5= Somewhat agree 6= Agree 7= Completely agree 0= I don't know & not applicable
civil_status_num	1 = Married 2 = Divorced 3 = Separated 4 = Never been married 0 = Other
happiness_num	"Which statement best reflects your view when it comes to having children and happiness?" 1 = People without children are much happier than people with children 2= People without children are somewhat happier than people with children 3= People with and without children are much are equally happy 4= People with children are somewhat happier than people without children
urban_num	1= Extremely urban and very urban 2 = Moderately urban and slightly urban 3= Not urban 0 = Other
type_dwelling_num	1 = Self-owned dwelling 2 = Cost-free dwelling 3 = Rental dwelling 0 = Other
origin_num	1= Dutch background 2= First-generation foreign, western background 3= First-generation foreign, non-western background 4= Second-generation foreign, western background 5= Second-generation foreign, non-western background 0 = Other

childwish_num (outcome variable)	Do you think you will have (more) children in the future? 1= Absolutely not 2= Probably not 3= I don't know 4= Probably so 5= Absolutely so

Table 1: Ego variables and definitions.

- **Alter variables:** The attributes come from the second block of the questionnaire. The respondents give answers about their personal network (25 individuals).

Variable name	Description / Question
age_a	Age person
sex_a	Gender person 1= Female 0 = Male
num_child_a	"How many children does this person have?" 1= Expecting first child 1-5 =Number of children 0= I don't know 5= More than 5
child_free_a	"Does this person prefer to remain childless?" 0= Prefers to remain childless 1= Wishes to have children 2= I don't know whether the person wishes to have children
friend_a	"Do you consider this person to be a friend? 1= Yes, is a friend 0= No, is not a friend
help_a	"If you have a child or were to have a child in the future, could you ask this person for help in caring for the child (e.g., as a babysitter)?" 1= Could ask for help in caring for a child 0 = Could not ask for help in caring for a child
childwish_a	"Does this person wish to have children?" 1= Wishes to have children 0= I don't know whether the person wishes to have children
has_child_a	"Does this person have children or are currently expecting a child?" 1= Does have (a) child(ren) or is expecting a child 0= Does not have (a) child(ren) and is not expecting a child

age_child_a	"How old is the youngest child of this person?" 0.5= Between 0 and 6 month & Between 6 and 12 months & Expecting first child 1= Between 1 and 2 years 2= Between 2 and 3 years 3= Between 3 and 4 years 4= Between 4 and 5 years 5= Older than 5 years 0= I don't know
f2f_a	"How often face-to-face contact with the person?" 1= About once a month & Several times a month 2= Several times a week 3= Daily 0= A few times a year or less
non_f2f_a	"How often are you in touch with these people in other ways than face-to-face, for instance by (mobile) phone, post, email, chat, SMS, and other forms of online and offline communication?" 1= About once a month & Several times a month 2= Several times a week 3= Daily 0= A few times a year or less

Table 2: Alter variables and definitions.

- **Network variables:** Except for the Graph Neural Network method, other machine learning models can't handle the node-edge structure and the graph network information itself. To use alter variables in these models, network variables are created.

"Network composition variables" and "Network structure variables" are used from the study by Stulp et al. (2023) as "Network variables" in this study. Briefly, we can group these network variables as follows:

- Tie strength variables (e.g. closeness to the alter, frequency of face-to-face contact)
- Number of particular alter attribute groups (e.g. number of kin in the network, the number of people with children)
- Different combinations of network variables (e.g. closeness measure for friends, average face-to-face contact with alters who want children)
- Structural network variables (density, degree, eigenvector, betweenness, modularity measures of the network)

Network variables that are significant after the models will be defined and explained in detail.

2.2.Ethical Considerations

Participants of the LISS panel underwent a double informed consent procedure. Ethical approval for the social networks and fertility survey within the LISS panel was granted by the ethical committee of sociology at the University of Groningen (Stulp et al., 2023).

3. Methods

In previous studies, it was observed that individual characteristics of women are more important in predicting their fertility intentions, while the network effect was explained using structural variables such as density, closeness, or number of people with specific attributes in the network derived from the attributes of alters, through Lasso Regression. (Stulp et al. 2023). In this study, first a Graph Neural Network (GNN) model was employed to investigate the network effect more deeply.

3.1.Graph Neural Networks

GNNs are effective learning frameworks when the data has a graphical structure, such as social networks, road routes, or molecular bond structures in chemistry. Learning from this kind of data requires an effective representation of their graph structure (Xu et al., 2018). GNNs are advanced and specialized classes of neural networks designed to work with graph-structured data. Different than traditional neural networks, which handle data in structured formats like grids (images) or sequences (text), GNNs can effectively capture the relations of graphs by message passing between the nodes of graphs (Zhou et al., 2020).

In this study, a variant of GNN, Graph Convolutional Network (GCN) was used. In a social network, nodes (vertices) represent individuals, and edges (links) represent the connection between individuals. Unlike traditional neural networks that work on grid-structured data (like images or sequences), GCNs can handle data represented as graphs, which consist of nodes and edges. The convolution operation in GCNs involves aggregating information from a node's neighbors to update its feature representation. (Kipf & Welling, 2016)

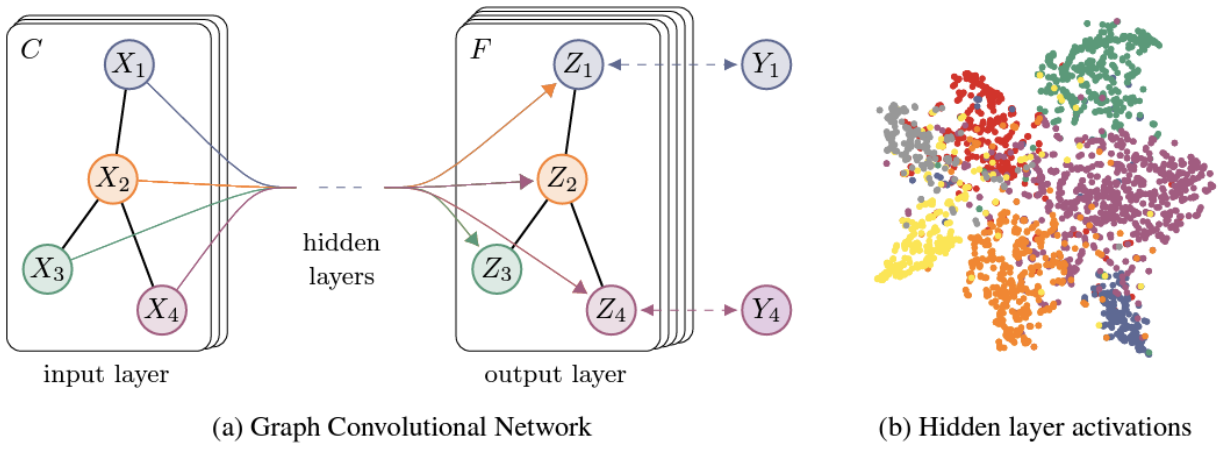

Figure 1: Schematic depiction of multilayer Graph Convolutional Network (GCN) with C input channels and F feature maps in the output layer (Kipf & Welling, 2016).

In the fertility network, as shown in Figure 2, “n25” represents the respondent (ego) and the other 24 individuals are the alters of this respondent. Data includes 706 different graphs (after cleaning the data) for each respondent as in Figure 2. Each ego has connections with her all alters, and some alters may also have connections with each other. All the relationships and features known about the network consist of those between the ego and its alters, there is no information other than whether there are connections between the alters themselves. The network graphs as a whole have no connection among themselves. This means we also don’t have information about whether each ego has a relation with other egos.

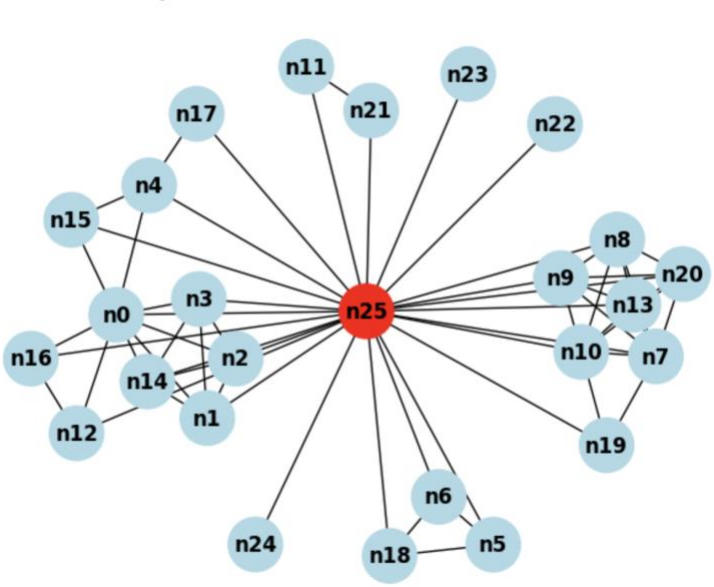

Figure 2: Node-edge representation of a network for one respondent in the “Social Networks and Fertility Research”.

Here, the “alter attributes” are utilized as predictors to determine each ego's "childwish" classification. As our goal is to predict the classification of each sub-network, this constitutes a graph-level prediction task.

The GCN model for this task was built using the PyTorch and PyTorch Geometric libraries in Python. First networks were exported for each ego and alter attributes and edge information were transformed into tensors since these libraries work with tensors. Then different GCN models were built with a combination of alter attributes. Some of these attributes like “happiness” were removed from the model because they reduced the performance.

The GCN model consists of three graph convolutional layers followed by a global mean pooling layer and a fully connected linear layer. The GCN layers were initialized with the number of node features from the dataset and the specified hidden channel size, which was set to 64. The forward method first applied the three convolutional layers to the input features (x) and edge index, each followed by a ReLU activation function. After the convolutional layers, the node embeddings were aggregated using global mean pooling to produce a fixed-size graph representation. Finally, a

dropout layer with a dropout rate of 0.5 was applied for regularization before passing the result through a linear layer to obtain the final class predictions. The use of dropout helps prevent overfitting by randomly setting a fraction of the input units to zero during training.

The accuracy of the GNN Models changed in the range of 0.30-0.40. The performance of the model wasn't as good as expected, when it was checked, it was observed that the models can only predict one class ("Probably so") and make very few predictions for other classes. To solve this problem, group imbalances in "childwish" were tried to be reduced. However, even after attempting methods such as merging groups or excluding the 'I don't know' classes from the data, while the accuracy increased, the model failed to make predictions for other classes. The detailed results reported in the Summary performance metrics can be seen in the "Results" section in Table 3.

For this reason, the effects of individual and network characteristics were examined by using 3 other machine learning models instead of the GNN: Histogram Gradient Boosting Tree, Random Forest, and Support Vector Machine.

While building these models, hyperparameter tuning was made using Grid SearchCV and the aim was to reach the best parameters. 5-fold cross-validation was applied to separate train and test sets. The scikit-learn library was used for building these models in Python.

3.2.Histogram-Based Gradient Boosting DecisionTree

The Gradient Boosting Decision Tree (GBT) is a powerful ensemble machine learning algorithm that uses many week learner decision trees. Each new tree tries to minimize the errors of the previous ones, and this process continues. A major drawback of GBT is that it is slow to train the model, especially in large datasets (Friedman, 2001).

Histogram-Based Gradient Boosting Decision Tree (HGBT) is a variation of GBT that uses histogram-based methods. The histogram-based algorithm groups continuous feature values into discrete bins instead of determining split points from sorted feature values. These bins are used to create feature histograms during training, optimizing memory usage and training speed, constructing histograms largely determines computational complexity. In HGBT, data is divided into specific intervals and histograms are created for each interval. This provides a faster and more efficient process since the model goes through the intervals instead of the entire data set (Ke et al., 2017)

With HGBT 3 kinds of models were built: Ego model, Network Model, and Full Model.

The ego model consists only of ego variables as predictors to assess the extent to which women's individual characteristics define their fertility intentions. The network model relies on network variables as predictors to examine how much network characteristics define their fertility

intentions. In the full model, both types of variables are included to investigate their combined effects. Since this method can't handle network data as a graph, "network variables" are utilized instead of "alter attributes", as detailed in the Data section previously. Additionally, in the "childwish" variable, the 'I don't know' class was removed from the data since it doesn't carry scalable value.

To optimize the HGBT model, GridSearchCV was used in hyperparameter tuning. For the ego, network, and full model different hyperparameters (maximum iteration, maximum depth of each tree, minimum number of the sample leaf, learning rate) showed the best performance. After calculating performance metrics(accuracy, precision, recall, F1-score) for each model and class, feature importances were evaluated as shown in the "Results" section, also detailed outcomes are shown in the Appendix.

3.3.Random Forest

Another machine learning method applied in the study is Random Forest. Random forest is an ensemble learning method that is a combination of multiple decision trees such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. They combine the simplicity of decision trees with the power of ensemble learning to create models that perform well on a variety of tasks (Breiman, 2001).

The same dataset and preprocessing steps were applied with HGBT. Using GridSearch CV, tuning was performed for hyperparameters such as number of trees, maximum depth, minimum sample split, and minimum sample leaf to find the best parameter values. Summary performance metrics can be seen in the "Results" section in Table 3, also detailed outcomes are shown in Tables 10, 11, and 12 in the Appendix.

3.4.Support Vector Machine

Support Vector Machines (SVM) represent a powerful supervised learning technique for classification, regression, and outlier detection with an intuitive model representation. It aims to find the optimal separating "hyperplane" into classes by maximizing the margin between the classes' closest points. These data points that are closest to the hyperplane are called support vectors, and the middle of the margin is the optimal separating hyperplane. Hyperplanes can be linear as well as have nonlinear boundaries, which is possible through the use of different kernel functions (Kecman, 2005).

The same dataset and preprocessing steps were applied with HGBT and Random Forest. Using GridSearch CV, tuning was performed for hyperparameters such as kernel function, gamma, and regularization parameter to find the best parameter values. Summary performance metrics can be

seen in the “Results” section in Table 3, also detailed outcomes are shown in Tables 13, 14, and 15 in the Appendix.

The Python code to produce the results for all methods (GNN, HGBT, Random Forest, and SVM) in the current study can be found in the “github.com/Ezgigunbatar/FertilityStudy” repository.

4. Results

After applying the GNN and observing that this model was not successful in predicting all classes, models HGBT, Random Forest, and SVM were implemented. When using these models, network information was transformed into "network variables" that traditional machine learning methods can process as predictors.

	Ego Model	Network Model	Full Model
HGBT	0.56	0.46	0.53
GNN	-	0.35	-
Random Forest	0.54	0.45	0.51
SVM	0.50	0.45	0.53

Table 3: Accuracy metrics for all models.

Almost all models except GNN, showed similar performance in terms of accuracy as seen in Table 3. Ego model has the best performance with an accuracy range of 0.50-0.54 for all methods. Although the performance of the network models is lower, they still have almost as much explanatory power with around 0.45 accuracy as the ego and full models.

The detailed tables for the best method, HGBT, are as follows in Table 4, Table 5, and Table 6:

Childwish	Precision	Recall	F1 Score	#
1- Absolutely not	0.52	0.55	0.54	89
2- Probably not	0.48	0.33	0.39	98
4- Probably so	0.56	0.75	0.64	227
5- Absolutely so	0.61	0.45	0.52	190

Table 4: Performance metrics for the ego model (HGBT)

For the ego model, out of all the instances that truly belong to the "Probably so" class, the model correctly predicts 75% of them, this measure is 55% for “Absolutely not” class.

Childwish	Precision	Recall	F1 score	#
1- Absolutely not	0.40	0.34	0.37	89
2- Probably not	0.33	0.24	0.28	98
4- Probably so	0.48	0.56	0.52	227
5- Absolutely so	0.49	0.50	0.49	190

Table 5: Performance metrics for the network model (HGBT)

The network model has lower performance scores than the ego model, yet it still has a certain level of explanatory capacity for classes “Probably so” and “Absolutely so" with 0.52 and 0.49 F1 scores.

Childwish	Precision	Recall	F1 score	#
1- Absolutely not	0.59	0.57	0.58	89
2- Probably not	0.37	0.26	0.30	98
4- Probably so	0.53	0.66	0.59	227
5- Absolutely so	0.54	0.48	0.51	190

Table 6: Performance metrics for the full model (HGBT)

As it is seen from the tables, models are good at predicting classes “Probably so” and “Absolutely so” it may happen because of the high number of the group sample. But at the same time even with the low sample size “Absolutely not” class still has reliable prediction performance.

Additionally, to understand which variables have more contribution to models, permutation impartance scores are calculated for each model. These variables have the highest importance:

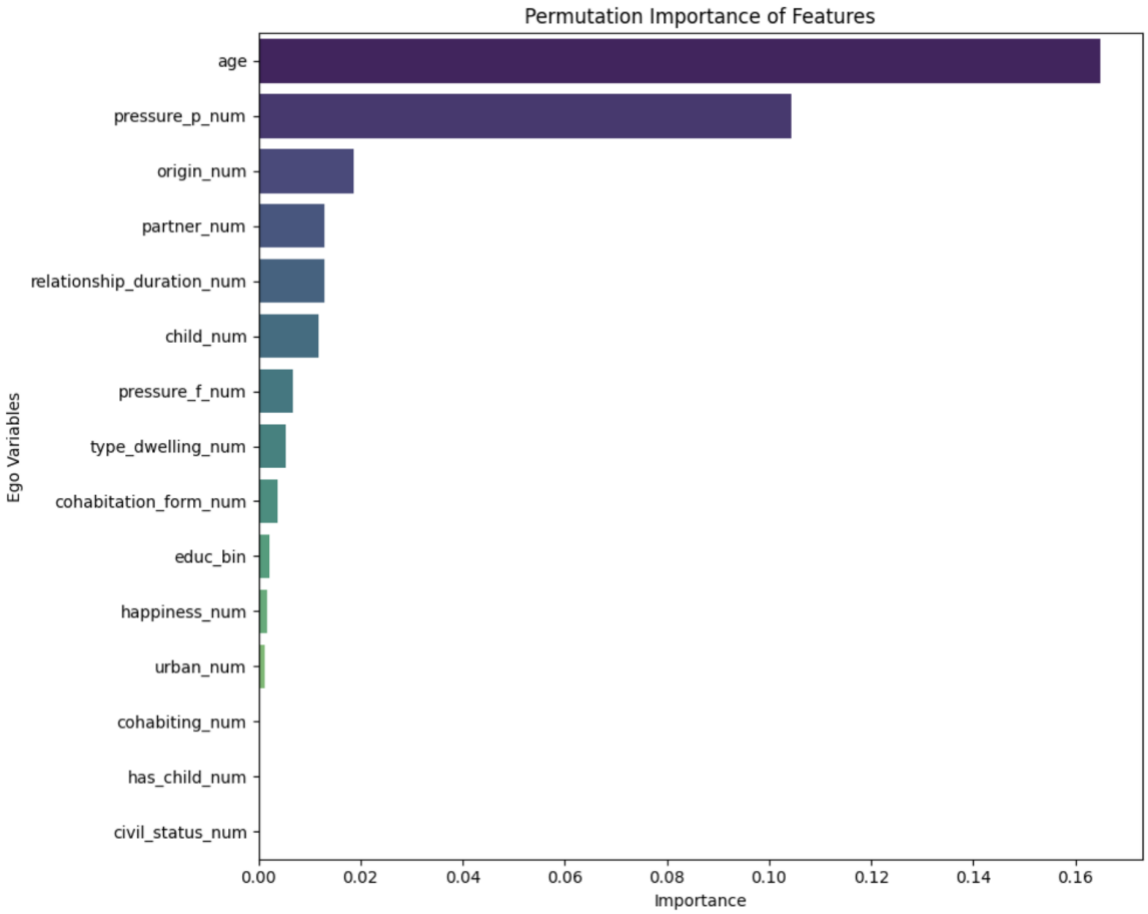

Figure 3: Permutation importance in the ego model

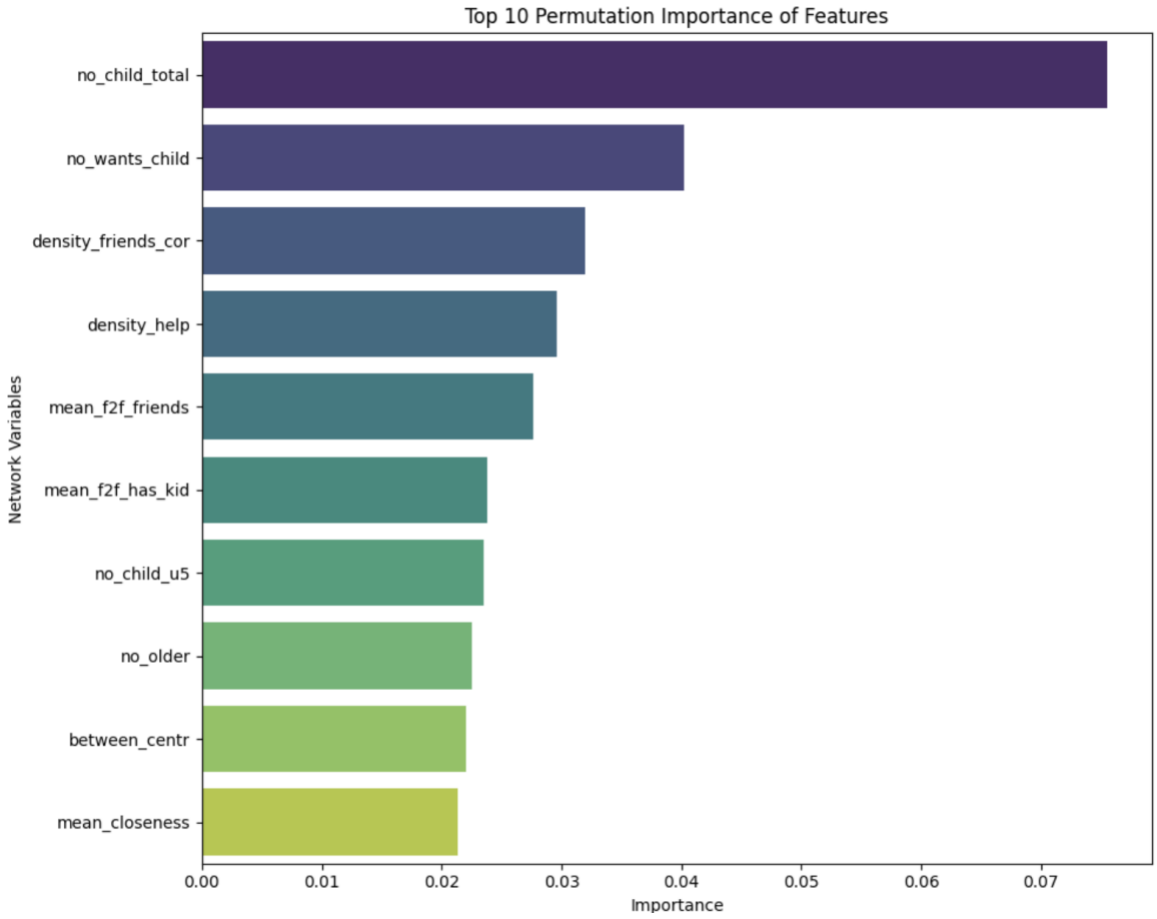

Figure 4: Permutation importance for the top 10 variables in the network model

Since the network model has many variables, the first 10 important variables are listed. Although network variables are not as dominant as ego variables, they still have a certain degree of explanatory power as seen also from the Figure 4. Descriptions of network variables with high importance are as follows:

- no_child_total: Number of total children in the network
- no_wants_child: Number of alters who want children in the network
- density_friends_cor: Density (proportion of ties) within friends
- density_help: Density within who the respondent can ask for help for child
- mean_f2friends: Average face-to-face contact for friends
- mean_f2f_has_kid: Average face-to-face contact for alters with kids
- no_child_u5: Number of children under 5 in the network
- no_older: How many alters are older than the respondent

- between_centr: Betweenness centrality[1] of the network
- mean_closeness: Average for the closeness attribute in the network

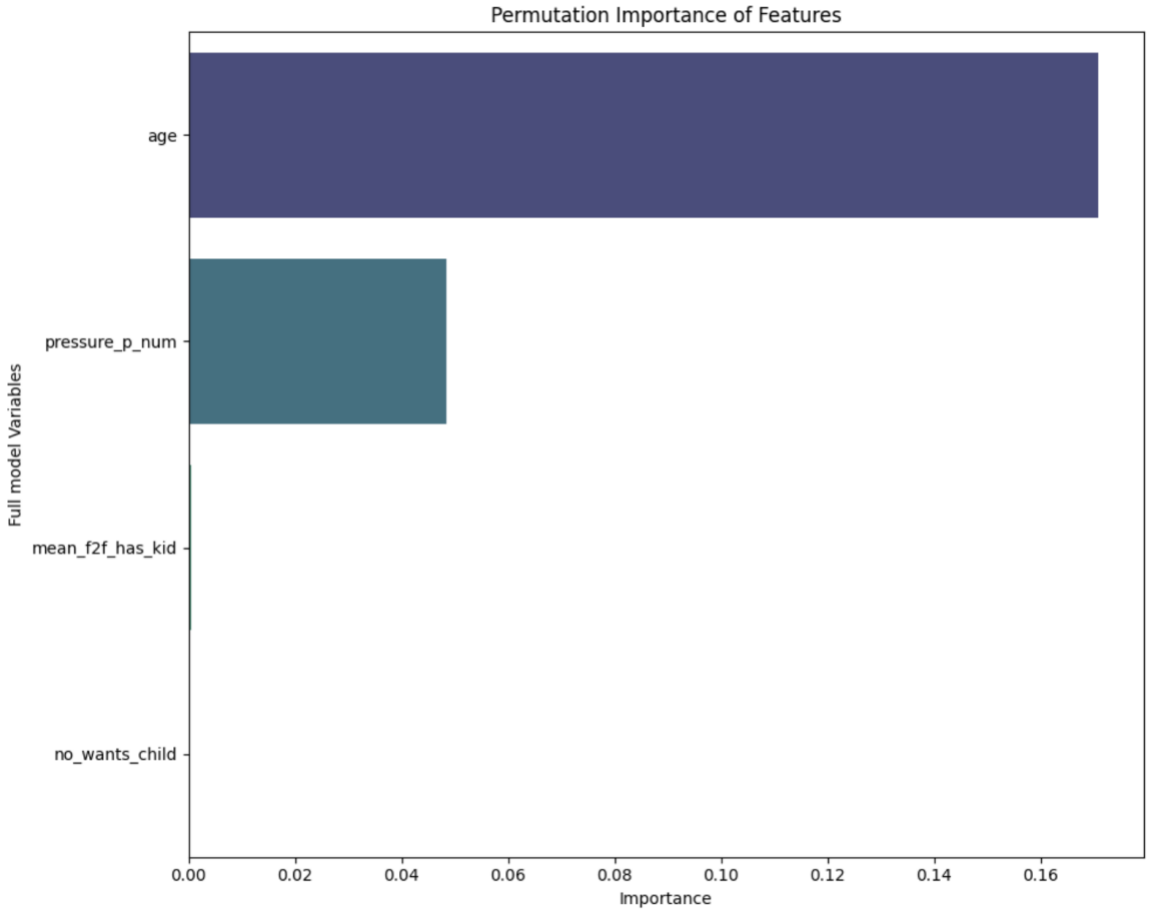

Figure 5: Permutation importance for both ego and network variables in the full model

When we look at the full model, which includes both ego and network variables, it can be seen that in Figure 5 "age" and "parent pressure" have the greatest importance. It can be said that individual characteristics have a greater impact on fertility intentions than network variables.

Detailed permutation importance score tables can be seen in Table 7, 8 and 9 in the Appendix.

Partial dependency plots are shown below to understand which variables have significant effects on which childwish classes and the direction of these effects. These plots show the relationship between a feature and the predicted outcome while marginalizing the values of all other features in the model. Plots for all childwish classes are as follows:

[1]: Betweenness centrality: The determination of a node's centrality is based on the quotient of the number of all shortest paths between nodes in the network that include the regarded node and the number of all shortest paths in the network. Individual is considered to be well connected if he is located on as many shortest paths as possible between pairs of other nodes (Landherr et al., 2010)

- **“Absolutely not”**

Ego variables:

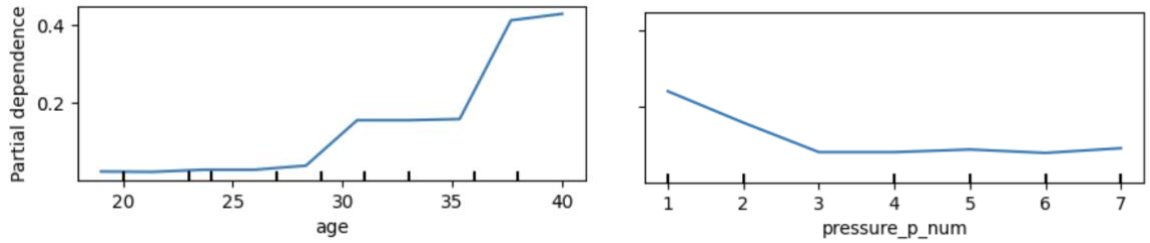

Figure 6: Partial dependency plots (1= “Absolutely not”) for important ego variables

Based on Figure 6 and the initial plot, the probability of saying "Absolutely not" to wanting a child increases by 0.2 units when the age factor increases from around 35 to 38. Additionally, it can be stated that after age 28, the thought of having a child decreases as age increases.

Network variables:

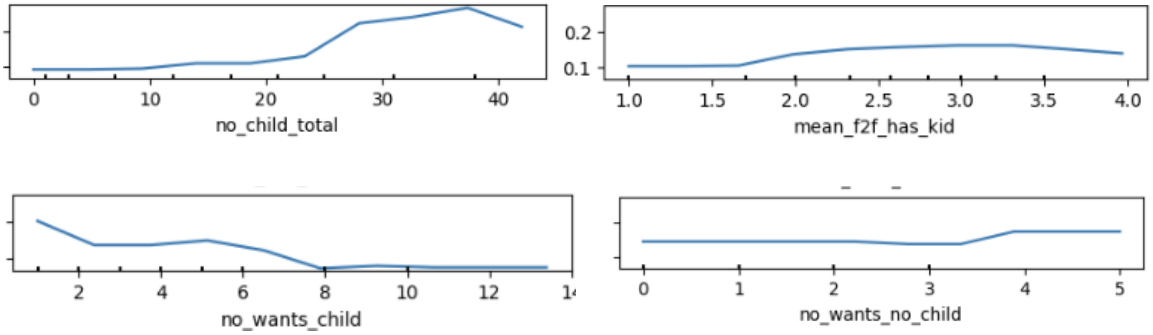

Figure 7: Partial dependency plots (1= “Absolutely not”) for important network variables

Based on Figure 7, as the number of people with children with whom they communicate face to face and the number of children in their social network increases, women are more likely to say "Absolutely not".

At the same time, it is observed that the likelihood of saying “Absolutely not” decreases as the number of people who want children increases, and it decreases as the number of those who do not plan to have children increases. From this, we can conclude that the desire for children in women's social circles influences them in a consistent manner.

- **“Probably not”**

Ego variables:

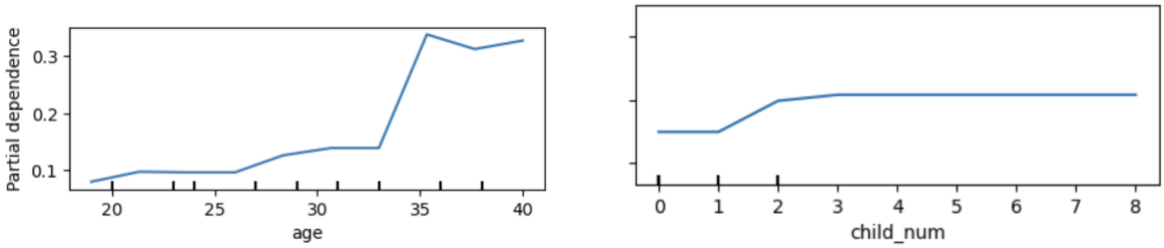

Figure 8: Partial dependency plots (2= “Probably not”) for important ego variables

In Figure 8, similar to the "absolutely not" class, the probability of saying "probably not" to wanting a child also increases as age increases. Women with two children currently have a 0.05 higher probability of saying "probably not" compared to women with one child.

Network variables:

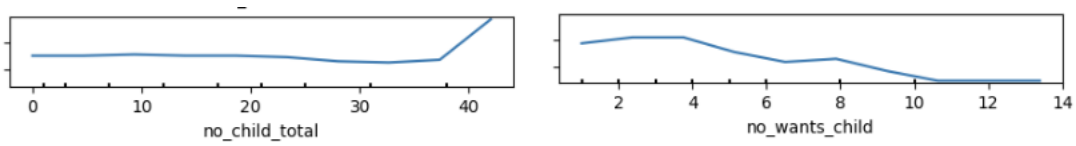

Figure 9: Partial dependency plots (2= “Probably not”) for important network variables

In Figure 9, as the total number of children in a woman's network increases (38 or more), the probability of her saying "probably not" to having children also increases. Additionally, as the number of people in her network who want children increases, this probability decreases.

- **“Probably so”**

Ego variables:

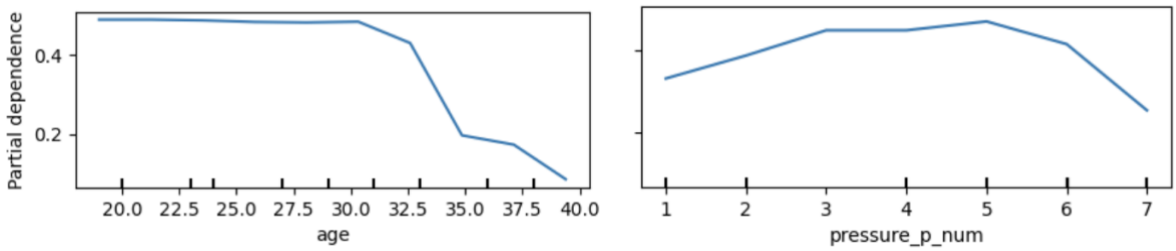

Figure 10: Partial dependency plots (4= “Probably so”) for important ego variables

In Figure 10, following the results from classes “Absolutely not” and “Probably not” when age gets older, the probability of saying “Probably so” decreases. Parent pressure affects women negatively (also in Figure 6), when family pressure is low, there is a higher likelihood of having a positive attitude towards having children, whereas this likelihood decreases as family pressure increases.

Network variables:

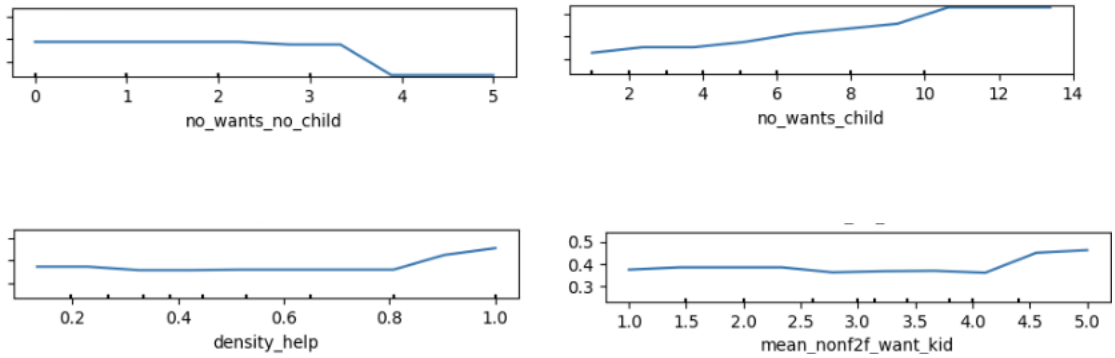

Figure 11: Partial dependency plots (4= “Probably so”) for important network variables

Based on Figure 11, as the number of people in the network who want children increases, saying “probably so” to having a child probability increases. At the same time, as the number of alters expressing a desire for children through non-face-to-face communication increases, and as connections with alters who could assist children grow, this likelihood is observed to rise.

- **“Absolutely so”**

Ego variables:

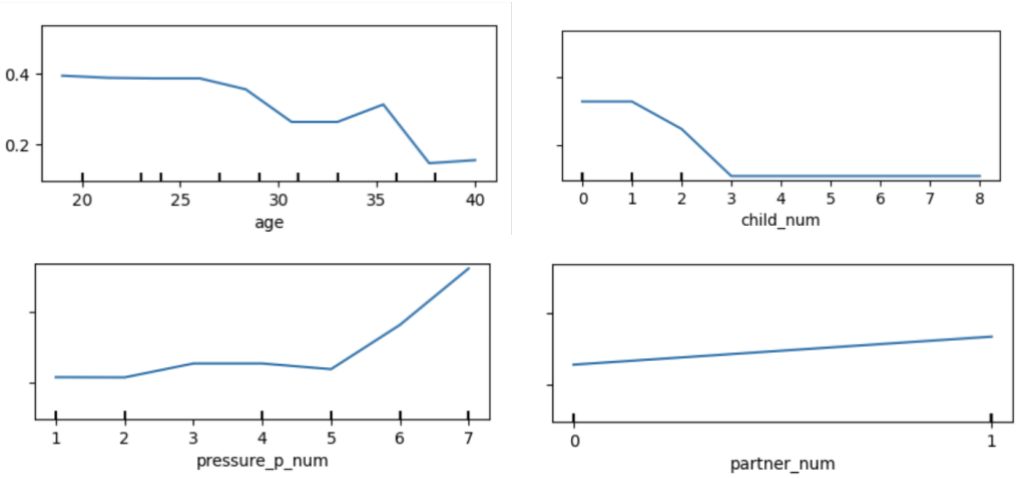

Figure 12: Partial dependency plots (5= “Absolutely so”) for important ego variables

In Figure 12, also when the age gets older and, the number of children increases the probability of saying “absolutely so” decreases. Women with partners are more likely to have a positive attitude towards having children compared to those without partners.

Network variables:

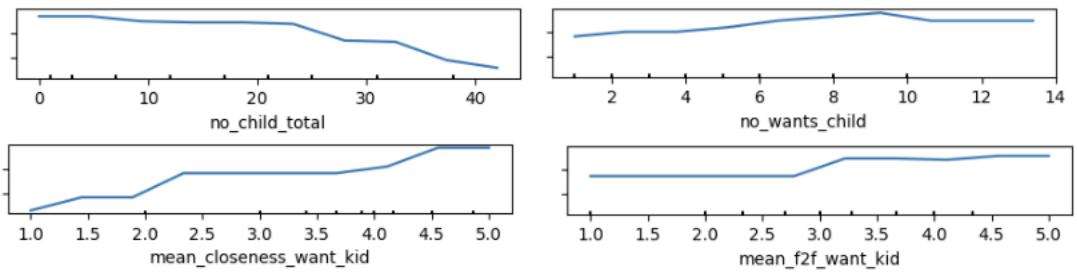

Figure 13: Partial dependency plots (5= “Absolutely so”) for important network variables

In Figure 13, as closeness to alters expressing a desire for children increases, and as face-to-face meetings with these individuals become more frequent, the likelihood of saying “absolutely so” also rises.

Detailed partial dependency plots, for all classes and models in HGBT can be seen in Figures 14-21 in the Appendix.

4.1.Conclusion and Discussion

In conclusion, this study offers a quantitative approach to defining the effects of individual and social network characteristics on fertility decisions. Since we believed that a GNN model would provide better insights into the graph structure of the social network data, initially we experimented with it. However, this method could only predict one class of the outcome variable accurately and struggled with other classes. Therefore, the network structure was transformed into structural variables and was explored with different machine learning methods such as HGBT, SVM, and Random Forest. Among these, HGBT performed the best, so this model was used in subsequent steps to describe the variables further.

In the study, it was found that ego variables (individual characteristics), especially age and family pressure, had a more significant impact compared to network variables. This finding is consistent with the study by Stulp et al. (2023), which also highlighted the importance of ego variables. However, at the same time, it was observed that the model incorporating network variables also had explanatory power for fertility intentions to a certain extent. Factors such as the number of total children, the number of people who want children, having contact with friends with children, the number of older people, and ties with people who can be asked for help for the child in the

network were found to have an impact on fertility intentions. Significant variables were observed in different classes using partial dependency plots.

In general, it can be said from the model's results;

- As age increases, the likelihood of considering having children decreases.
- As the number of people in a woman's network who want children increases, and as her connections with these individuals strengthen, she becomes more inclined to consider having children.

Other variables also have different effects depending on the childwish classes. This is because there may not be sufficient data related to that class for the specific feature being predicted.

Additionally, although we refer to these as "individual characteristics," it is not entirely possible to separate these traits from the influence of the network. It is expected that people form their networks based on their own character traits, and the reverse is also possible. Thus, what we investigate as network effects may be intertwined with individual variables. It might be necessary to approach this subject from a causality perspective. For example, does a person who is considering having children form their social circle based on this intention, or does their tendency to have children increase because the majority of people around them are parents? When it comes to social sciences and human behaviors, it is often not possible to depict the subject with very clear boundaries.

In this study, one factor that could reduce the success of the analyses and lead to bias is the way the respondents answered. Respondents were first asked for information about themselves and then expected to list 25 names and answer subsequent questions on behalf of these individuals. These responses may not be accurate and could also reflect the respondents' personal thoughts.

One of the other limitations of the study, especially for GNN, our framework currently does not naturally support edge features and is limited to undirected graphs. When information is available about the relationships and edge features among alters themselves, an analysis can be conducted across the entire network, not just through each ego. Knowing whether the networks of each respondent are interconnected can also positively influence the analysis results. Although a study conducted with such data would be effective, collecting the data could be costly and challenging in terms of time and resources.

In future studies, research on the relationship between fertility intentions and networks could be conducted from a causality perspective. This two-way approach could provide more insights. Also, it is believed that GNN will work more efficiently when unbiased and more detailed information about the other nodes in the network, not limited to the respondent, is obtained.

In addition, there are many 'I don't know' responses, indicating indecision, in the “childwish” outcome column that was excluded from the study. A large majority of these responses are likely to turn into positive or negative attitudes over time. If a longitudinal study can be conducted in a certain time period, the predictions related to these categories may also provide meaningful results.

References

- Bernardi, L., & Klaerner, A. (2014). Social networks and fertility. *Demographic Research, 30*, 641–670. https://doi.org/10.4054/demres.2014.30.22.
- Bongaarts, J., & Watkins, S. C. (1996). Social interactions and contemporary fertility transitions. Population and Development Review, 22, 639–682.
- Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324
- Coale, A. J., & Watkins, S. C. (1986). The decline of fertility in Europe. Princeton, NJ: Princeton University Press.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451.
- Kavas, S., & De Jong, J. (2020). Exploring the mechanisms through which social ties affect fertility decisions in Turkey. *Journal of Marriage and the Family/Journal of Marriage and Family*, *82*(4), 1250–1269. https://doi.org/10.1111/jomf.12668.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). Lightgbm: a highly efficient gradient-boosting decision tree. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 3149–3157*
- Kecman, V. (2005). Support Vector Machines – An Introduction. In *Studies in fuzziness and soft computing* (pp. 1–47). https://doi.org/10.1007/10984697_1
- Keim, S., Klärner, A., Bernardi, L. (2009). Who is relevant? Exploring fertility relevant social networks. *MPIDR Working Paper. Max Planck Institute for Demographic Research, Rostock, Germany.*
- Kipf, T.N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1609.02907.

- Landherr, A., Friedl, B., & Heidemann, J. (2010). A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, *2*(6), 371–385. https://doi.org/10.1007/s12599-010-0127-3
- Madhavan, S., Adams, A., & Simon, D. (2003). Women's networks and the social world of fertility behavior. International Family Planning Perspectives, 29(2), 58. https://doi.org/10.2307/3181059
- Richter, N., Lois, D., Arra´nz Becker, O., & Kopp, J. (2012). Mechanisms of social network influences on fertility decisions in East and West Germany. In J. Huinink, M. Kreyenfeld, & H. Trappe (Eds.), Special Issue 2012 of Journal of Family Research-Zeitschrift fur Familienforschung (pp. 95–118).
- Stulp G. 2020 Methods and materials of the social networks and fertility survey (Sociale relaties en kinderkeuzes). https://doi.org/10.34894/EZCDOA
- Stulp, G., Top, L., Xu, X., & Sivak, E. (2023). A data-driven approach shows that individuals' characteristics are more important than their networks in predicting fertility preferences. Royal Society Open Science, 10(12). https://doi.org/10.1098/rsos.230988.
- Stulp G. (2023a) Describing the Dutch Social Networks and Fertility Study and how to process it. Demogr. Res. 49, 493–512. (doi:10. 4054/DemRes.2023.49.19)
- Stulp G. (2023b) FertNet: process data from the social networks and fertility survey. R package version 0.1.1. See https://github.com/gertstulp/FertNet
- The Social Networks and Fertility Survey, downloaded from https://dataarchive.lissdata.nl.
- The code used in this thesis can be found at the following address: https://github.com/Ezgigunbatar/FertilityStudy

- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001
- Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018, October 1). How Powerful are Graph Neural Networks? https://arxiv.org/abs/1810.00826

Appendix

Permutation Importance Scores:

	Feature	Importance	Std Dev
0	age	0.164901	0.011944
10	pressure_p_num	0.104470	0.012311
14	origin_num	0.018709	0.006541
1	partner_num	0.012914	0.007094
5	relationship_duration_num	0.012914	0.007285
4	child_num	0.011755	0.004651
9	pressure_f_num	0.006623	0.004188
13	type_dwelling_num	0.005298	0.003210
7	cohabitation_form_num	0.003642	0.002943
3	educ_bin	0.002152	0.005238
8	happiness_num	0.001656	0.001481
12	urban_num	0.001325	0.000993
6	cohabiting_num	0.000166	0.001880
2	has_child_num	0.000000	0.000000
11	civil_status_num	0.000000	0.000000

Table 7: Permutation Importance Scores for Ego model (HGBT)

Permutation Importance Scores:

	Feature	Importance	Std Dev
0	age	0.170861	0.016437
10	pressure_p_num	0.048344	0.005629
27	mean_f2f_has_kid	0.000497	0.000759
69	no wants child	0.000331	0.000662

Table 8: Permutation Importance Scores for the Full model (HGBT)

Permutation Importance Scores:

	Feature	Importance	Std Dev
51	no_child_total	7.549669e-02	0.006378
54	no_wants_child	4.023179e-02	0.010883
79	density_friends_cor	3.195364e-02	0.004744
65	density_help	2.963576e-02	0.006474
11	mean_f2f_friends	2.764901e-02	0.005186
12	mean_f2f_has_kid	2.384106e-02	0.011735
52	no_child_u5	2.350993e-02	0.005329
46	no_older	2.251656e-02	0.007148
71	between_centr	2.201987e-02	0.007976
0	mean_closeness	2.135762e-02	0.007042
56	no_help	2.119205e-02	0.003612
73	avg_betweenness	2.069536e-02	0.005651
23	mean_nonf2f_talk	1.953642e-02	0.004844
19	mean_nonf2f_has_kid	1.903974e-02	0.005889
6	mean_closeness_want_kid	1.870861e-02	0.012588
14	mean_f2f_wants_no_kid	1.821192e-02	0.004967
57	no_talk	1.804636e-02	0.007746
2	mean_nonf2f	1.655629e-02	0.006455
61	density_children	1.556291e-02	0.006024
10	mean_f2f_kin	1.523179e-02	0.005277
13	mean_f2f_want_kid	1.523179e-02	0.007434
7	mean_closeness_wants_no_kid	1.490066e-02	0.005288
74	avg_closeness	1.473510e-02	0.006259
20	mean_nonf2f_want_kid	1.324503e-02	0.004253
84	density_help_cor	1.307947e-02	0.005851
55	no_wants_no_child	1.274834e-02	0.007065
76	cliques	1.192053e-02	0.002543
4	mean_closeness_friends	1.076159e-02	0.003646
3	mean_closeness_kin	1.043046e-02	0.007065
49	no_high_edu	1.043046e-02	0.005238
66	comm_1or2	1.026490e-02	0.007171
82	density_childfree_cor	8.940397e-03	0.004136
18	mean_nonf2f_friends	8.609272e-03	0.004844
1	mean_f2f	8.443709e-03	0.005259
8	mean_closeness_help	8.443709e-03	0.006844
5	mean_closeness_has_kid	7.781457e-03	0.007256
83	density_talk_cor	7.781457e-03	0.004744
16	mean_f2f_talk	7.615894e-03	0.003244
62	density_wantschildren	6.953642e-03	0.002848
58	density	6.291391e-03	0.006186
48	no_kin	4.470199e-03	0.004914
68	modularity	3.807947e-03	0.002966
63	density_childfree	2.814570e-03	0.002227
22	mean_nonf2f_help	1.821192e-03	0.000892
80	density_children_cor	1.324503e-03	0.001443
69	comp_largest	1.324503e-03	0.003124
59	density_kin	1.158940e-03	0.002966
75	avg_eigenv	1.158940e-03	0.004126
15	mean_f2f_help	1.158940e-03	0.003146
64	density_talk	9.933775e-04	0.001325
50	no_has_child	6.622517e-04	0.007375

Table 9: Permutation Importance Scores for the Network model (HGBT)

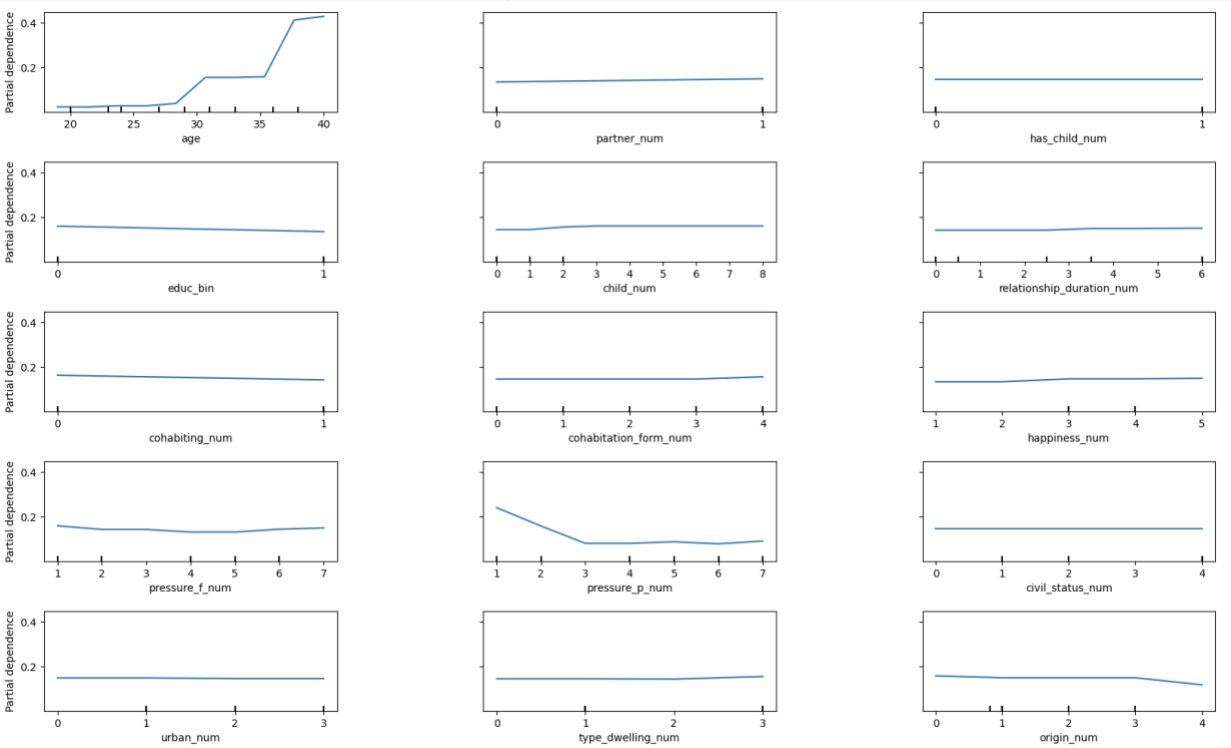

Figure 14: Partial dependency plots (1= “Absolutely not”) for all ego variables (HGBT)

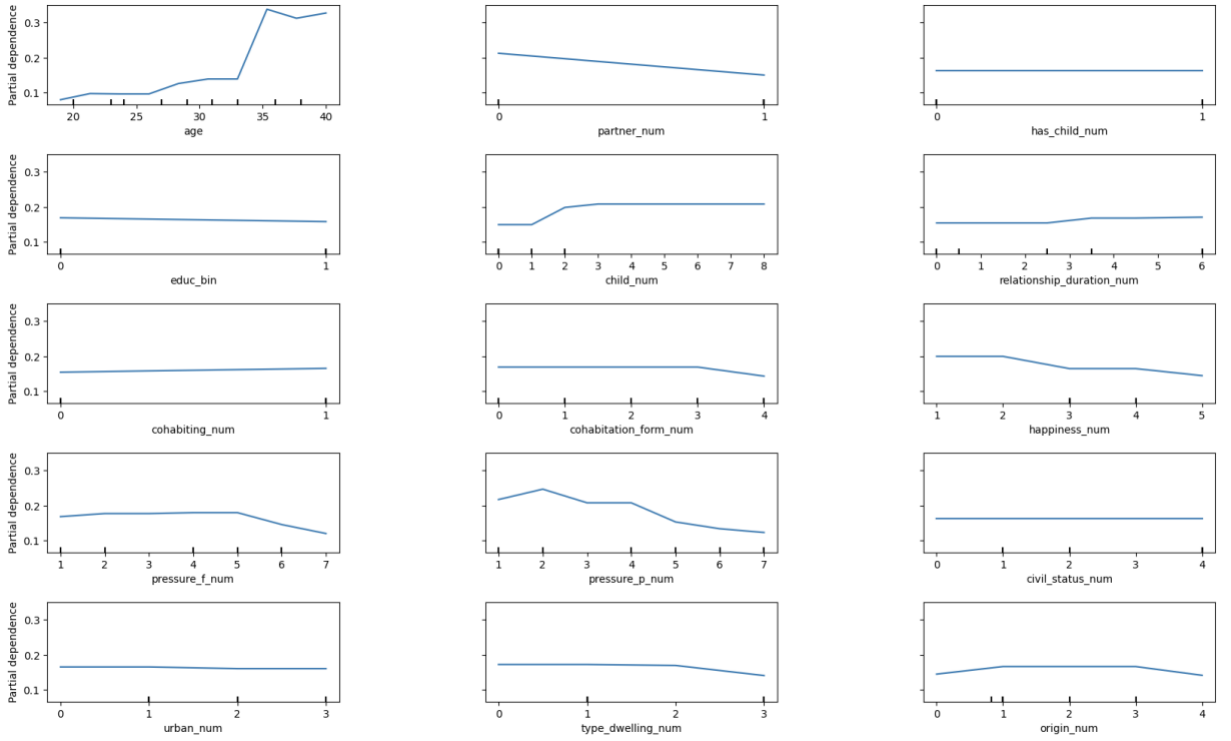

Figure 15: Partial dependency plots (2= “Probably not”) for all ego variables (HGBT)

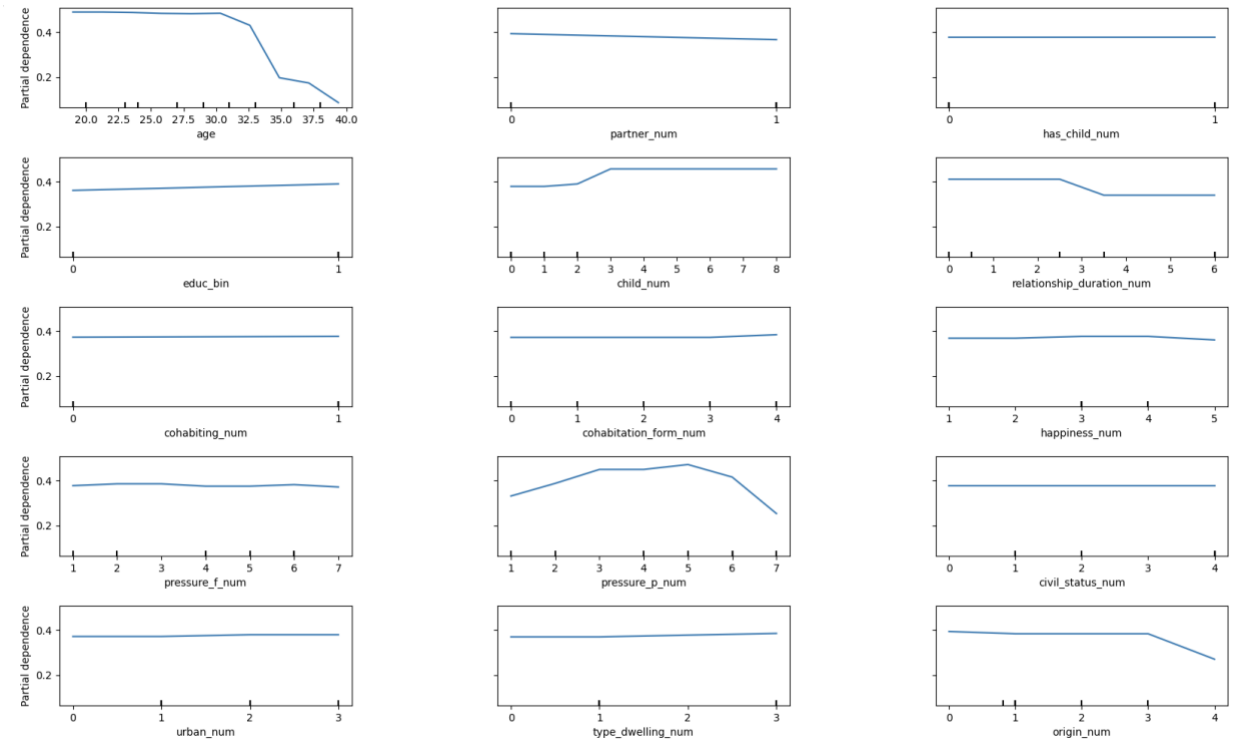

Figure 16: Partial dependency plots (4= “Probably so”) for all ego variables (HGBT)

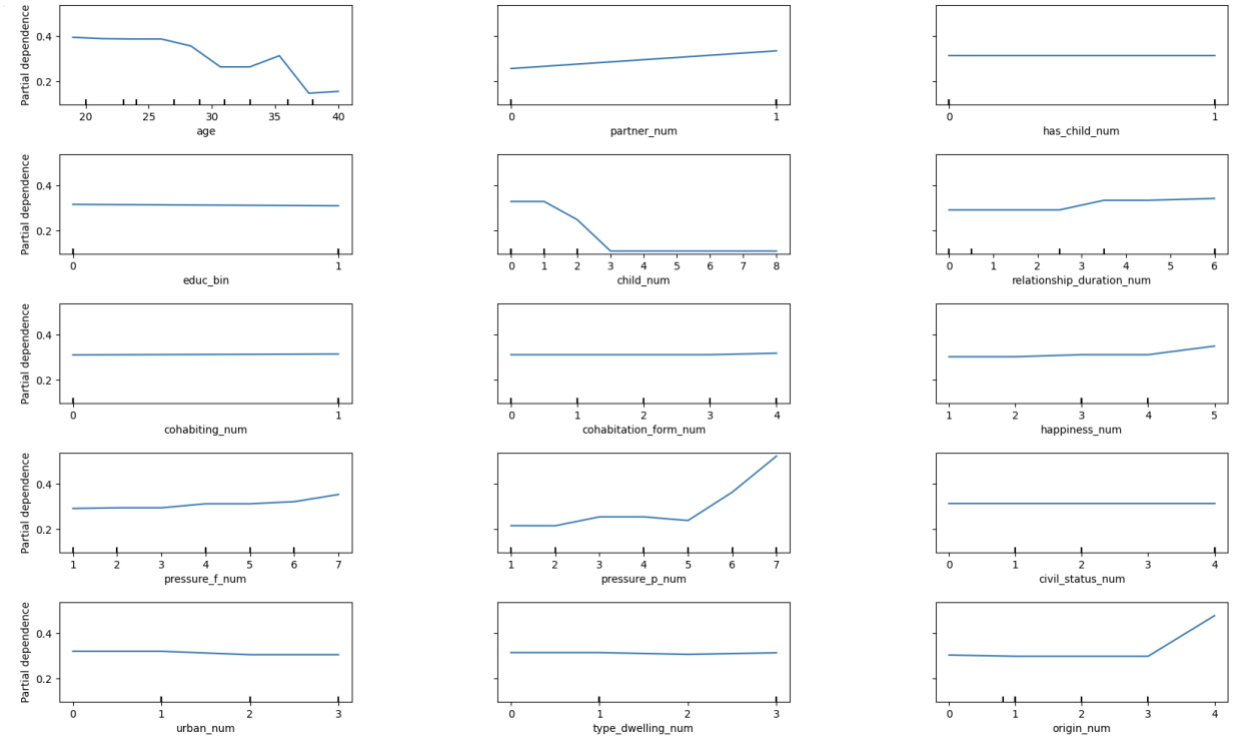

Figure 17: Partial dependency plots (5= “Absolutely so”) for all ego variables (HGBT)

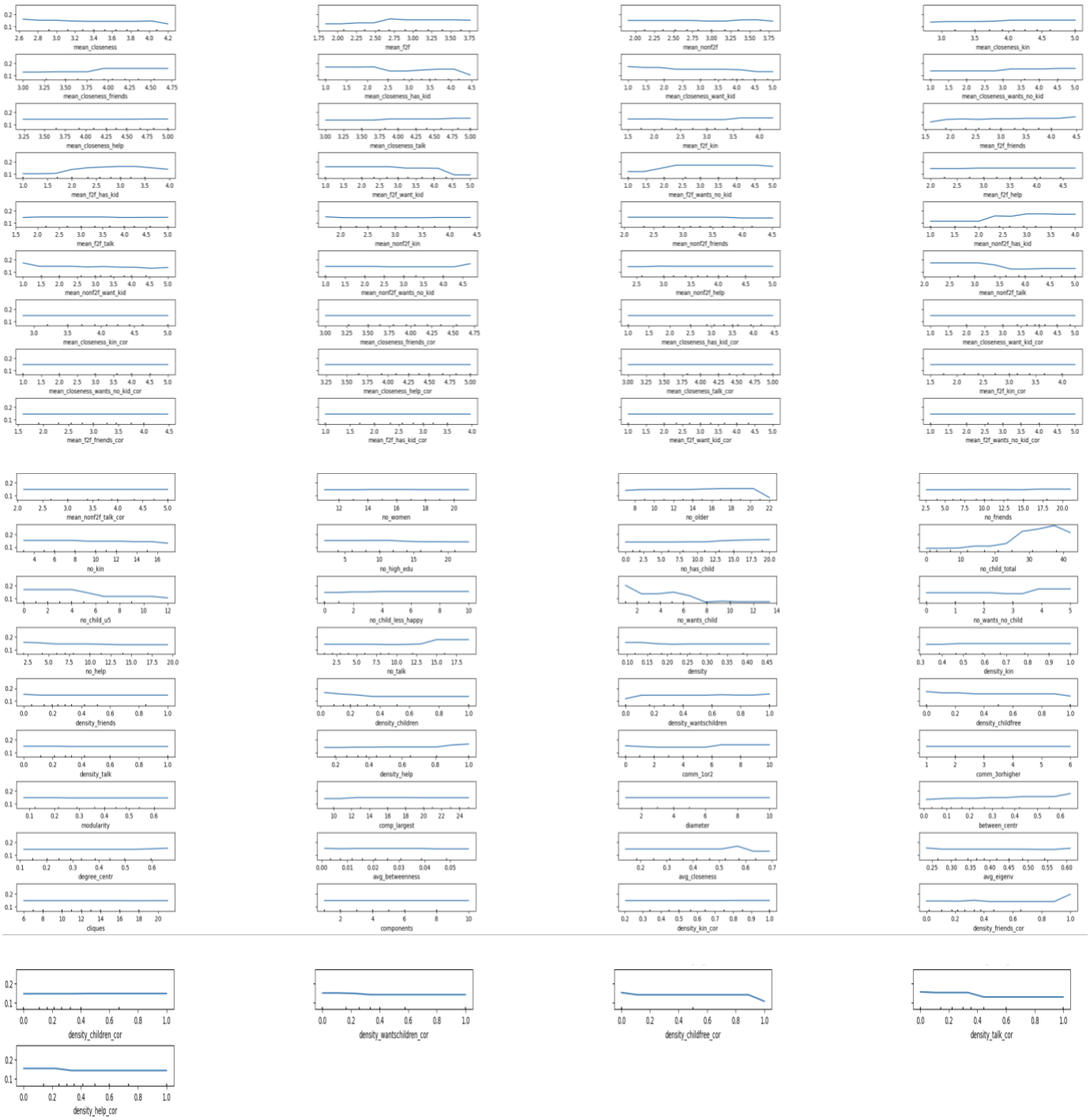

Figure 18: Partial dependency plots (1= “Absolutely not”) for all network variables (HGBT)

Figure 19: Partial dependency plots (2= “Probably not”) for all network variables (HGBT)

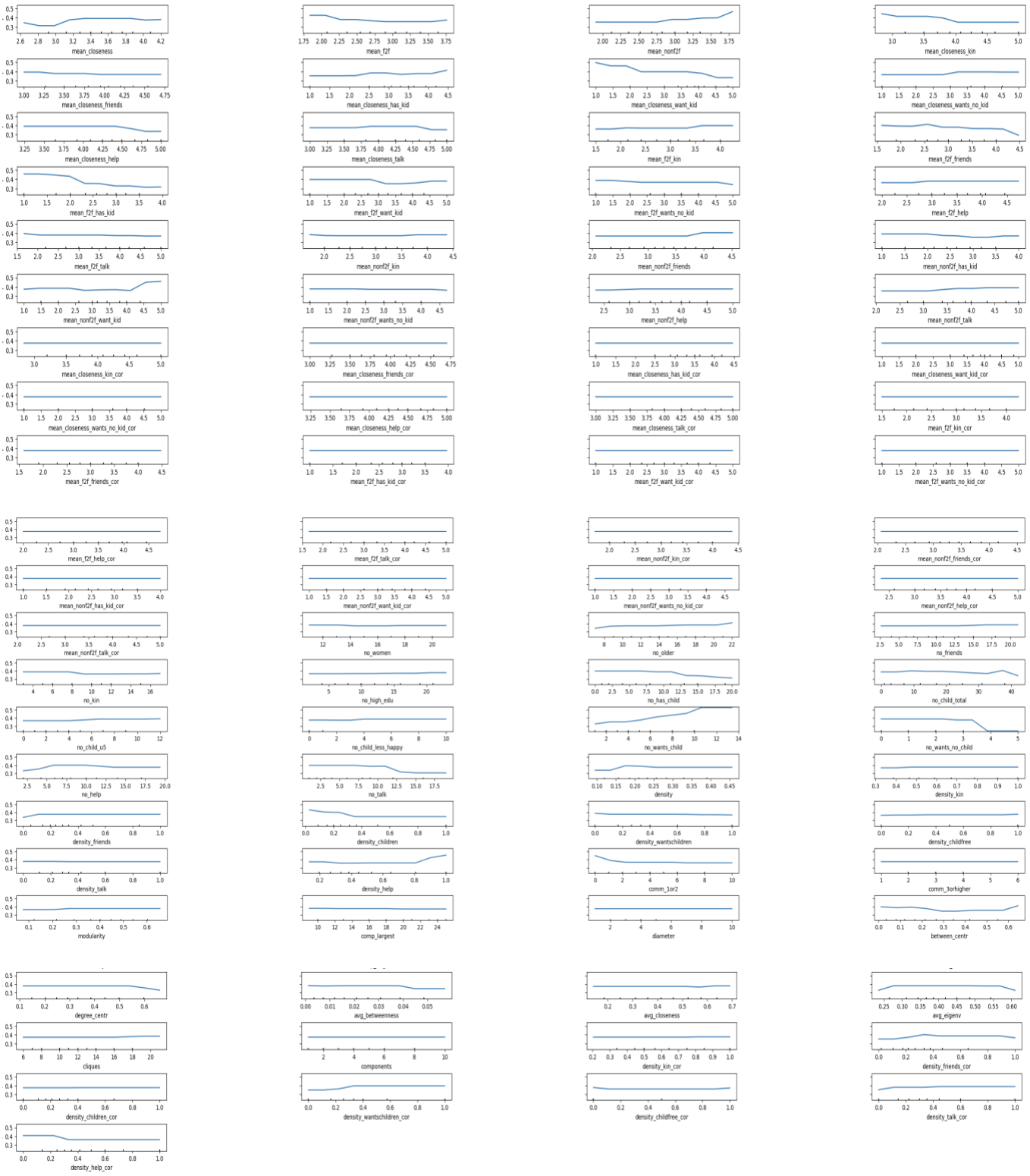

Figure 20: Partial dependency plots (4= “Probably so”) for all network variables (HGBT)

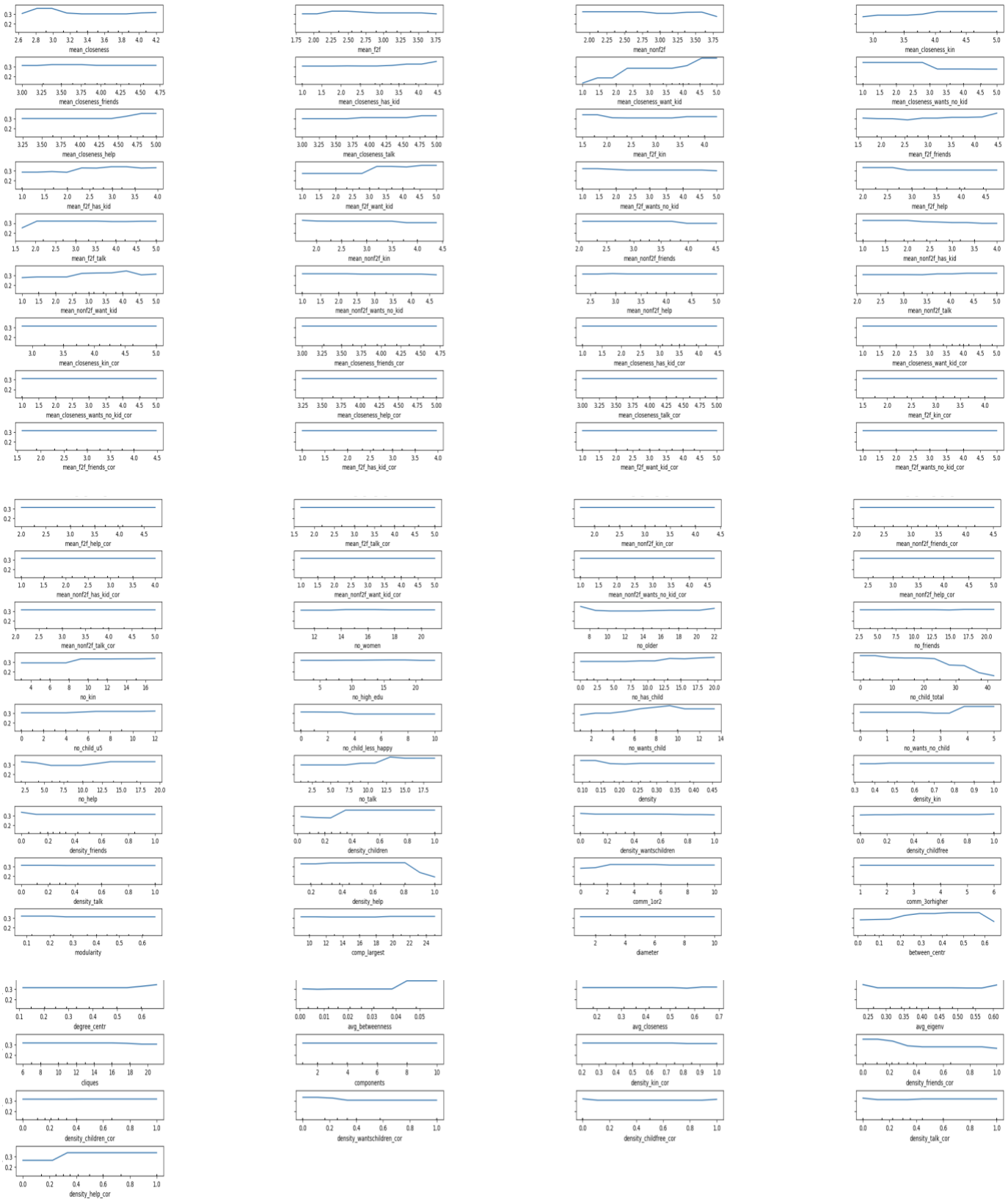

Figure 21: Partial dependency plots (5= “Absolutely so”) for all network variables (HGBT)

Childwish	Precision	Recall	F1 Score	#
1- Absolutely not	0.51	0.56	0.53	89
2- Probably not	0.43	0.30	0.33	98
4- Probably so	0.55	0.67	0.60	227
5- Absolutely so	0.58	0.49	0.53	190

Table 10: Performance metrics for the ego model (Random Forest)

Childwish	Precision	Recall	F1 Score	#
1- Absolutely not	0.49	0.42	0.45	89
2- Probably not	0.41	0.22	0.29	98
4- Probably so	0.46	0.64	0.54	227
5- Absolutely so	0.43	0.36	0.39	190

Table 11: Performance metrics for the network model (Random Forest)

Childwish	Precision	Recall	F1 Score	#
1- Absolutely not	0.52	0.49	0.51	89
2- Probably not	0.40	0.30	0.34	98
4- Probably so	0.51	0.66	0.58	227
5- Absolutely so	0.54	0.44	0.49	190

Table 12: Performance metrics for the full model (Random Forest)

Childwish	Precision	Recall	F1 Score	#
1- Absolutely not	0.50	0.53	0.51	89
2- Probably not	0.42	0.28	0.33	98
4- Probably so	0.52	0.65	0.57	227
5- Absolutely so	0.51	0.43	0.47	190

Table 13: Performance metrics for the ego model (SVM)

Childwish	Precision	Recall	F1 Score	#
1- Absolutely not	0.56	0.26	0.35	89
2- Probably not	0.24	0.08	0.12	98
4- Probably so	0.44	0.66	0.53	227
5- Absolutely so	0.47	0.48	0.48	190

Table 14: Performance metrics for the network model (SVM)

Childwish	Precision	Recall	F1 Score	#
1- Absolutely not	0.50	0.48	0.49	89
2- Probably not	0.29	0.22	0.25	98
4- Probably so	0.54	0.64	0.59	227
5- Absolutely so	0.58	0.53	0.55	190

Table 15: Performance metrics for the full model (SVM)