**STANDARDFORM OUTPUT OR EXPORT[[1]](#footnote-1)**

|  |  |
| --- | --- |
| Name *[name user]* | Hanzhang Ren |
| Telephone number *[telephone number where the user can be reached when there are questions about this output]* | +1 2032857184 |
| Institution *[name institution]* | Rijksuniversiteit Groningen |
| Date *[date on which the output is produced]* | 2024-12-31 |
| Project *[contractnumber + name research]* | 9469 - Predicting Fertility Data Challenge (PreFer) |
| This output consists entirely of documentation *[syntax or text file without research results] If Yes, then you don’t need to fill in the rest of the form* | No |
| This output is an adjusted version of a stopped output. If Yes, then please fill in the date of the stopped output. | No |
| Data *[used datasets]* | KINDOUDERTAB, GBAPERSOONTAB, GBABURGERLIJKESTAATBUS, HOOGSTEOPLTAB, SPOLISBUS, GBAHUISHOUDENSBUS, GBAMIGRATIEGEBEURTENISBUS, SECMBUS, INHATAB, VEHTAB, NABIJHEIDKINDOPVTAB, GBAADRESOBJECTBUS, VSLGWBTAB, GBAVERBINTENISPARTNERBUS, FAMILIENETWERKTAB |
| Purpose *[description of the purpose of the analysis in this output]* | The purpose of this analysis is to examine how sample size and the set of predictor variables influence predictive performance for models predicting fertility. |
| Relation with previous output *[description of the relation with previously submitted output]* | This is similar to a few previous outputs we did in which we examined how sample size and feature sets impact predictive performance. This time, we used different feature sets. |
| Content of the output  *[description of the output, for example: there will be frequencies or crosstabs or regression analysis[[2]](#footnote-2) in this output]* | This output contains:  (1) a folder with files containing code and code alone.  (2) topics\_performance.csv and feature\_importance\_performance.csv contain predictive performance metrics for various feature sets at different sample sizes. The former file contains model performance for various combinations of substantive topic areas like education and family structure. As a baseline, we also report the performance of a model not informed by any feature sets and simply predicting the mean of the training set. We also include a model based on a set of features chosen by a domain expert. feature\_importance\_performance.csv reports the performance of the top 10, 20, 30, and 40 percent of the most informative variables in a particularly well-performing model we made. All performance metrics are reported on evaluation sets of around 200,000 to 400,000 people. We trained our models on random samples of Dutch residents between 18 and 45 years olds, with sample sizes ranging from 1000 to 3.7 million.  (3) feature\_importance.csv reports the percentage of variation in predicted values attributable to each of 138 predictors in a particularly well-performing model trained on 3.7 million people.  (4) missingness.csv is a frequency table that counts the number of individuals with unknown data values for each of 96 variables in random subsets of our data. We also check whether missingness rates differ based on our prediction outcome—whether one had children or not between 2021 and 2023.  (5) metadata.csv includes no information directly derived from CBS microdata. We created the csv ourselves to keep track of which features are in which feature sets. |
| 1 – Does this output contain zeros? | No for the most part. However, missingness.csv indicates that many variables are missing for 0 people. |
| 2 – Does this output contain values < 10 or is it possible to recalculate values < 10? | No for the most part. However, missingness.csv indicates that in some of our randomly selected training sets or test sets, the number of people with missing data falls below 10. This is largely an artifact of the fact that we randomly sampled people into groups. If we add across each column of numbers, either for people who had kids or not, we can see that the number of people with missing data is either zero or at least 10 for all predictors. |
| 3 – Is there any group disclosure or dominance in this output? | No for the most part. However, missingness rates are frequently far below 10 percent. |
| 4 – Does this output contain information on 1 institution/company/household/etc? | No for the most part. However, there is a single 1 in missingness.csv: when we took a roughly 4 percent random sample of 18-45-year-old Dutch people who had kids to create an evaluation set, 1 person in that group happened to have a mother with unknown country of birth. |
| 5 – Are there any other points where this output does not meet the [output guidelines](file:///K:\Utilities\Regels_Richtlijnen\Guidelines%20for%20RA%20Output.pdf) ? | No |
| If you have answered “Yes” to at least one of the above 5 questions, please give a short but clear explanation on why there is no risk of disclosure in your opinion. | We figured that the confidentiality risks related to missingness.csv is extremely low. Without CBS access, it is difficult to recognize whether anyone has missing data for a particular variable. Even if one could somehow recognize such a person, we do not disclose any additional detail about that person. |

1. If this form is not fully or correctly filled in, the output will not be checked. [↑](#footnote-ref-1)
2. In case the output contains results from models, please clearly mention the underlying number of observations. [↑](#footnote-ref-2)