<u>Online Supplementary Document</u>

Identification of distinct risk-subsets for under five mortality in India using CART model: an evidence from NFHS 4

Appendix S1

Table S1: Definition of the Explanatory variables.

S.No	Factors	Definition
1	**Highest educational level**	Highest education level attended. This is a standardized variable providing level of education in the following categories: No education, Primary, Secondary, and Higher. In some countries the educational system does not fit naturally within this scheme and a different categorization was used for the Final Report. In this case, this variable is constructed as accurately as possible from the country's own scheme and the variable used for the Final Report is included as a country-specific variable.
2	**Type of place of residence**	Type of place of residence where the household resides as either urban or rural.
3	**Covered by health insurance**	Covered by health insurance
4	**Wealth index**	The wealth index is a composite measure of a household's cumulative living standard. The wealth index is calculated using easy-to-collect data on a household’s ownership of selected assets, such as televisions and bicycles; materials used for housing construction; and types of water access and sanitation facilities. Generated with a statistical procedure known as principal components analysis, the wealth index places individual households on a continuous scale of relative wealth. DHS separates all

		interviewed households into five wealth quintiles to compare the influence of wealth on various population, health and nutrition indicators. The wealth index is presented in the DHS Final Reports and survey datasets as a background characteristic.
5	**States**	States
6	**Mothers age at birth**	Mothers age at birth
7	**Religion**	Religion
8	**Caste**	Caste
9	**Breastfeeding**	Whether the respondent is currently breastfeeding a child. This is based on the entries in the maternity history for children born in the last three/five years. If no child was born in the last three/five years, the respondent is assumed not to be breastfeeding. This variable is created by looking for any child which is still being breastfed, and not just whether the last child is being breastfed.
10	**Type of cooking fuel**	Type of cooking fuel
11	**Type of toilet facility**	Type of toilet facility in the household.
12	**Source of drinking water**	Main source of drinking water for members of the household
13	**Preceding Birth Interval**	Preceding birth interval is calculated as the difference in months between the current birth and the previous birth, counting twins as one birth. In the DHS VII recode, B11 is also based on the CDC of date of birth of the children (B18). In previous recodes B11 was based on the CMC date of birth of the children (B3). BASE: All births except the first birth and its twins.
14	**Birth in past five years**	Total number of births in the last five years is defined as all births in the months 0 to 59 prior to the

		month of interview, where month 0 is the month of interview.
15	**Birth order number**	Birth order number gives the order in which the children were born
16	**Birth weight in kilograms**	Reporting of birth weight is based on either a written record or mother's recall
17	**Sex of child**	Sex of child
18	**Child is twin**	Twin code gives an order number for each child of a multiple birth. Code 0 indicates a single birth, code 1-upwards give the number of the child. Twins are ordered in the birth history with the higher twin codes appearing before the lower twin codes. See the example of the birth history structure below.
19	**Number of antenatal visits during pregnancy**	Number of antenatal visits during the pregnancy. Women who did not see anyone for antenatal care during the pregnancy are coded 0. BASE: Last births in the three/five years before the survey.
20	**Delivery by caesarean section**	Whether child was born by caesarean section.
21	**Assistance at delivery**	The type of person who assisted with the delivery of the child (14 variables)
22	**Delivery complications**	This is based on breech complication, labour complication, and bleeding complication. If any of these is present, then it is defined as Yes, otherwise No. BASE: Received postnatal check within 2 months
23	**Place of delivery**	Place of delivery of the child (Categorized into institutional and non-institutional)
24	**Time before postnatal check up**	How long after delivery postnatal check took place BASE: Received postnatal check within 2 months

Appendix S2

Figure S1: Classification tree Model 1.

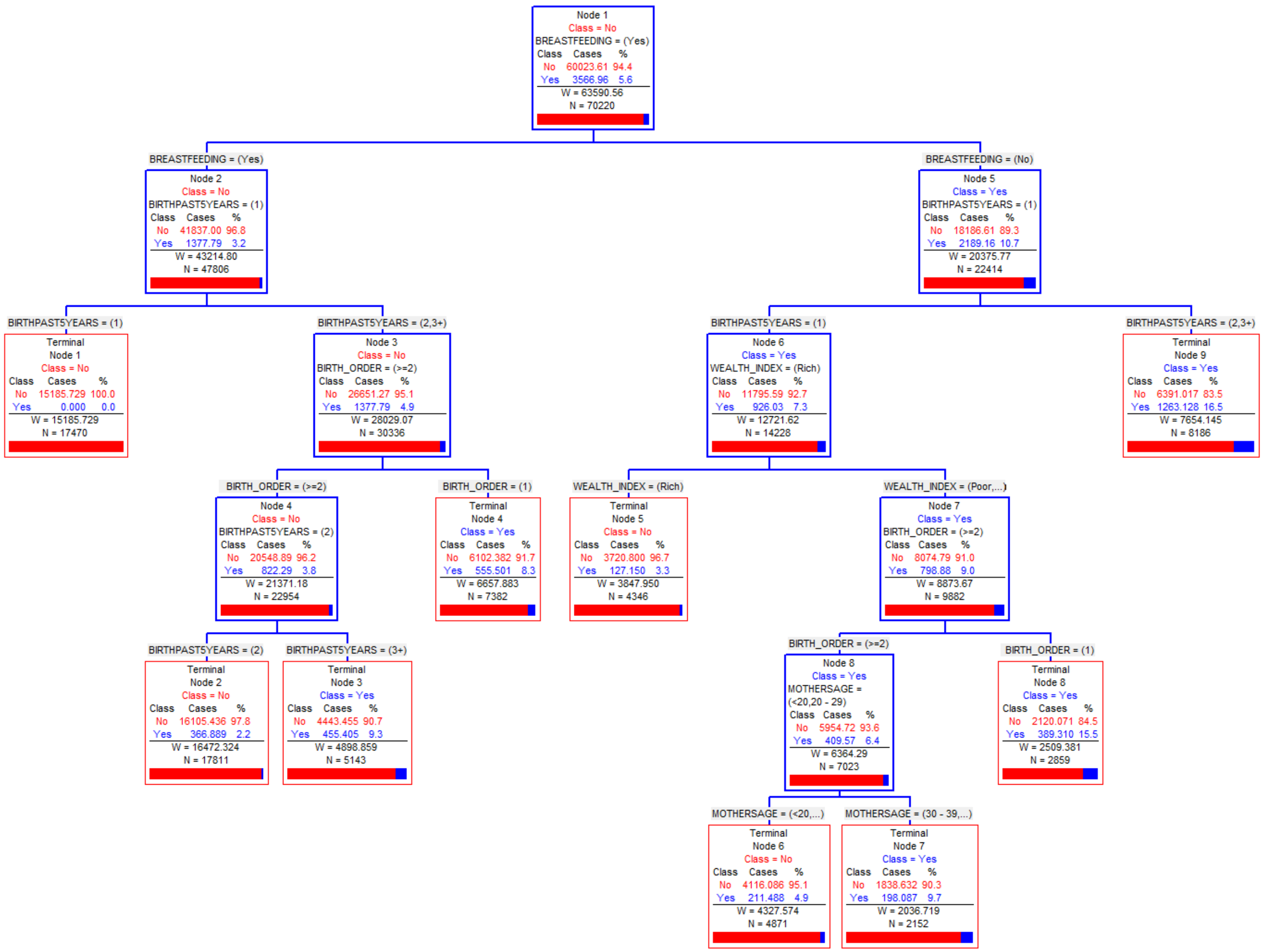

Figure S1 shows the classification tree model-1 using demographic factors, socioeconomic factors, nutritional factor, environmental factors, and maternal and biological factors for classifying children with under-five mortality. The rectangle represents node and terminal nodes. Terminal nodes (no further child node) are mutually exclusive and exhaustive subgroups of the study population.

Figure S2: Classification tree Model 2.

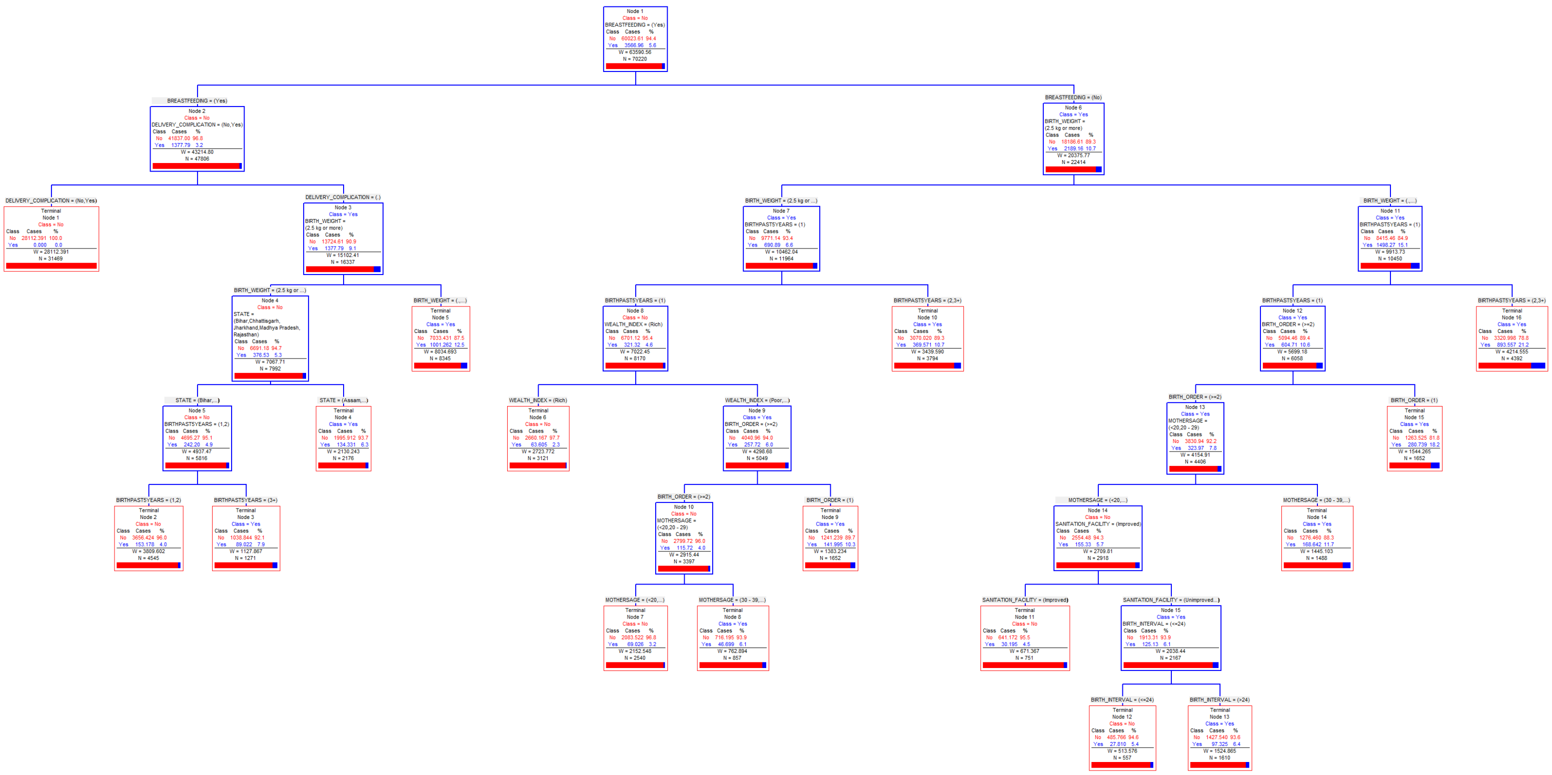

Figure S2 shows the classification tree model-2 using demographic factors, socioeconomic factors, nutritional factor, environmental factors, and maternal and biological factors for classifying children with under-five mortality. The rectangle represents node and terminal nodes. Terminal nodes (no further child node) are mutually exclusive and exhaustive subgroups of the study population.

Appendix S3

The classification tree for model-1 (Table 2) can be represented as following equation:

Tree = 0*I(1) + 0*I(2) + 1*I(3) + 1*I(4) + 0*I(5) + 0*I(6) + 1*I(7) + 1*I(8) + 1*I(9)   (1)

Where ***I*** is the Indicator function defined as:

$$\boldsymbol{I(x) = \begin{Bmatrix} 1 & If\ x = yes \\ 0 & If\ x = no \end{Bmatrix}} \qquad (2)$$

Here *x* is the terminal node conditions which need to be satisfied with above condition. Similarly, the classification tree for model-2 (Table 3), keeping other conditions similar to (1) and (2), can be represented as following equation:

Tree = 0*I(1) + 0*I(2) + 1*I(3) + 1*I(4) + 1*I(5) + 0*I(6) + 0*I(7) + 1*I(8) + 1*I(9) + 1*I(10) + 0*I(11) + 0*I(12) + 1*I(13) + 1*I(14) + 1*I(15) + 1*I(16)   (3)

Appendix S4

Table S2: Root (Node) Competitor Splits for Model-1.

-	**Competitor**	**Split**	**Improvement**	**Imp. ratio**
Main	Breastfeeding	Yes	0.04725	-
1	Birth in past 5 years	1,2	0.02426	0.51344
2	Birth weight	2.5 kg or more	0.01348	0.2852
3	Type of birth	Single	0.01013	0.21445
4	Birth interval	>24	0.00785	0.16614
5	Postnatal check up	4-23 hrs,1-2 days,3+ days	0.00497	0.10523
6	ANC visit	<4visits,atleast 4 visits	0.00459	0.09715
7	Wealth index	Rich	0.00355	0.07522
8	State	Assam, Bihar, Chhattisgarh, Jharkhand, MP, Rajasthan, UP	0.00353	0.0747
9	Mother’s age at birth	20 - 29	0.00351	0.07435
10	Delivery complication	No	0.00323	0.06833
11	Education	Secondary and above	0.00301	0.06375

12	Delivery assistance	Skilled	0.00292	0.0618
13	Sanitation facility	Improved	0.00238	0.05044
14	Place of delivery	Institutional	0.00219	0.04634
15	Cooking fuel	Safe	0.00168	0.03547
16	Residence	Urban	0.00123	0.02599
17	Caste	Other	0.00085	0.01806
18	Birth order	>=2	0.00064	0.01344
19	Source of water	Unimproved	0.00039	0.00821
20	Caesarean	Yes	0.00027	0.00575
21	Gender	Female	0.00025	0.00519
22	Religion	Other	0.00013	0.00272
23	Insurance	Yes	0.0001	0.00202

Appendix S5

Table S3: Root (Node) Competitor Splits for Model-2.

-	Competitor	Split	Improvement	Imp. ratio
Main	Breastfeeding	Yes	0.04725	-
1	Delivery complication	No, Yes	0.03109	0.65797
2	Postnatal check up	<4 hrs,4-23 hrs,1-2 days,3+ days, No check-up	0.02999	0.63472
3	ANC visit	No antenatal visits,<4visits,atleast 4 visits	0.02973	0.62922
4	Birth weight	2.5 kg or more	0.02965	0.62747
5	Birth in past 5 years	1,2	0.02426	0.51344
6	Type of birth	Single	0.01013	0.21445
7	Birth interval	>24	0.00785	0.16614
8	Wealth index	Rich	0.00355	0.07522

9	State	Assam, Bihar, Chhattisgarh, Jharkhand, MP, Rajasthan, UP	0.00353	0.0747
10	Mother's age at birth	20 - 29	0.00351	0.07435
11	Delivery assistance	Skilled	0.00328	0.06934
12	Education	Secondary and above	0.00301	0.06375
13	Place of delivery	Institutional	0.00248	0.05252
14	Sanitation facility	Improved	0.00238	0.05044
15	Cooking fuel	Safe	0.00168	0.03547
16	Residence	Urban	0.00123	0.02599
17	Caste	Other	0.00087	0.01832
18	Birth order	>=2	0.00064	0.01344
19	Source of water	Unimproved	0.00039	0.00821
20	Caesarean	Yes	0.00027	0.00575
21	Gender	Female	0.00025	0.00519
22	Religion	Other	0.00013	0.00272
23	Insurance	Yes	0.0001	0.00202

Appendix S6

Methods

Tree building: Steps of procedure

The Classification tree construction is based on the technique known as binary recursive partitioning. The tree construction process, which we adopted, starting from the root node using Gini diversity index as the splitting rule are given as the following:

- Firstly, the outcome variable, independent variables, splitting criteria, and pruning method were specified in the software with additional criteria such as priors, minimum costs, minimum parent node size and minimum child node size below which node will not split. Priors are set as EQUAL = 0.50 for the two classes of the outcome assuring that no matter how small a class may be relative to the other classes, it will be treated as if it were of equal size. Misclassification costs = 1 were kept as default. Minimum parent node size = 1000 and minimum child node size = 500.

- For model-2 additionally, missing together (MT) approach [Zhang et al. (1996)] was applied by creating missing categorical levels for predictors only. Suppose that we try to split node t by variable x_j and that x_j is missing for a number of subjects. The MT approach forces all these subjects to the same daughter node of node t.
- CART splits the first variable at the best split point with highest split improvement value compared to the same of other best splits of other variables. At each possible split point of a variable the sample splits into two child nodes. Cases with a "yes" and those with "no" response to the question were sent to the left child node and the right node, respectively.
- CART ranks all of the "best" splits on each variable according to the reduction in impurity achieved by each split and selects the variable and its corresponding split point that most reduced impurity of the root or parent node.
- CART then assigns classes to these nodes according to the rule that minimizes misclassification costs
- CART approach to the decision tree construction is based on the foundation that it is impossible to know for sure when to stop growing a decision tree. Steps 2 – 4 are repeatedly applied to each nonterminal child node at each stage recursively.
- CART uses extraordinarily fast algorithms, so it does take much time to grow the initial largest tree.
- The pruning technique was used to get the "right-sized" tree. CART uses two test procedures- tenfold cross validation and a random test sample to select optimal tree with the lowest overall misclassification cost, thus the highest accuracy. Both the test procedures are automated and ensure the optimal tree will accurately classify existing data and predict results.
- For larger dataset as in this study, we separated the data into two parts, the training set (50 %) and testing set (50 %). The tree was grown using only the training set, and the

testing set was used to estimate the error of all possible subtrees that can be built, and the subtree with the lowest error on the testing set was chosen as the decision or classification tree.

- The SPM software by default gives the optimal tree however one of other nearby trees are just as good as the optimal tree, therefore it is suggested that we use a "1 standard error" or 1SE rule to identify these trees. The optimal tree is "better" but it is also twice the size and our measurements are always subject to some statistical uncertainty. Thus, 1SE tree was selected as the final tree model

Appendix S7

Results

The CART decision tree for model-1 and model-2 are represented in Figure S1 and Figure S2, respectively. Both the trees were obtained after applying the three analytic steps: recursive partitioning, pruning and an independent test sample to measure the predictive accuracy of the pruned tree. At any given split in the tree into two descendent groups, the split to the left indicates survival groups, and the split to the right indicates mortality groups. Because the percentage of under-five mortality in the total sample was 5.6%, terminal subsets comprised of more than 5.6% mortality cases were considered as mortality groups. In both the models, breastfeeding was used as the 1st primary splitter variable selected with the highest split improvement among all the predictors considered (See Appendix S4 and Appendix S5), optimally splitting the entire sample involved in it with value "yes" splitting subjects to the left and values "No" splitting subjects to the right with highest reduction in impurity, indicating that breastfeeding was used as most important predictor of mortality among under-five children.

Appendix S8

Discussion

We observed that how specific risk factors, especially modifiable, jointly influence U5M (for example: breastfeeding & birth in past 5 years) and concluded that decision tree is a useful tool for identifying homogeneous subgroups defined by combinations of individual characteristics. Also, we observed important factors responsible for U5M in high focused states of India and found that breastfeeding & number of births in past five years were the two most crucial factors.
By applying CART model based recursive partitioning technique to NFHS-4 data, the performance in terms of correct classification

was found more in the classification rule without considering missing observations as a category.
Now a days, CART is an important recursive partitioning algorithm-based decision tree that gives the foundation of machine learning (ML) techniques and it is the basis for many powerful ML concepts like bagging and boosting, and algorithms like random forest and gradient boosting decision trees. The present study uses the recursive partitioning method which has been used in different types of studies in public health with respect to different outcomes. In this study, CART interaction is implicitly modelled over certain regions of the data i.e. locally so there was no need to add interaction terms or local terms in the model. The risk subgroups identified by classification tree structure could be used to generate hypothesis for future studies or could be examined using data from prospective studies of the same condition. If a classification tree grown with data from one study identified risk subgroups that were confirmed with data from other studies then conclusion regarding the influence of multiple factors to outcome risk would be enriched.
In terms of variable importance to classify U5M, Model-1 identified birth in past 5 years, breastfeeding, birth order, wealth index, mother‘s age at birth. Model-2 additionally identified delivery complications, birth weight, state, sanitation facility, birth interval, caste, education. Variable importance describes the role of a variable in a specific tree. It is natural to expect that the root node splitter will be the most important variable in a CART tree. However, we cannot generalize it for every tree. In our case, breastfeeding which was the root node splitter, turned out to be ranked second in terms of variable importance whereas, births in past 5 years and delivery complication were ranked first as the most important in model-1 and model-2, respectively. Sometimes a variable that splits the tree below the root is most important because it ends up splitting many nodes in the tree and splitting powerfully. The importance score given with variables deals with a variable's ability to perform in a specific tree of a specific size either as a primary splitter or as a surrogate splitter. It utters nothing about the value of the variable in the construction of other trees. For example, a variable that is very important in a ten-node tree might not be important at all in a two-node tree because it exhibits no role in the splitting of the root node (which is the only split in case of a two-node tree). Variables have more chances to play a role in the tree, if a tree is allowed to become larger, and thus to take non-zero importance scores. In case of comparing trees of substantially different sizes, the relative importance rankings of variables can alter dramatically. Thus, the rankings are strictly relative to a given tree structure;

and one should not consider importance scores to specify an absolute information value of a variable.

Major strengths of our study may also be noted. Recursive partitioning is a valuable data exploration method in the study of better understanding of how the socio-economic, demographic, cultural and environmental factors available at household-level, maternal-level, child-level, community-level, child-care program-level influence and affecting under-five mortality. It permits for the detection of higher order interactions within the data locally which would be very difficult to inspect using Generalized Linear Models. CART method has the primary benefit of illustrating the natural interaction and important variable selection related to outcome. The small data set generally adds instability of the classification tree and yielded imprecise measures of associations. Our study tried to avoid this problem by using large data set. This study is first of its kind from India carried out to find the distinct risk subsets based on decision tree. CART based recursive partitioning algorithm may be the best method in such situation.

Appendix S9

Future Research suggested

The combination of factors may be combined with traditional method (Logistic regression) to enhance the prediction accuracy. Ensemble methods (Bagging, or Bootstrap Aggregating, Random Forest Models) can be used to combine several base CART models in order to produce one optimal predictive model.