








Modeling the Association Between Prenatal Exposure to Mercury and Neurodevelopment of Children

Stefan Popov^{1,2} , Janja Snoj Tratnik³ , Martin Breskvar² ,
Darja Mazej³, Milena Horvat^{1,3} , and Sašo Džeroski^{1,2} 

¹ Jožef Stefan International Postgraduate School, Jamova cesta 39,
Ljubljana, Slovenia

{stefan.popov,milena.horvat,saso.dzeroski}@ijs.si

² Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39,
Ljubljana, Slovenia

martin.breskvar@ijs.si

³ Department of Environmental Sciences, Jožef Stefan Institute, Jamova cesta 39,
Ljubljana, Slovenia

{janja.tratnik,darja.mazej}@ijs.si

Abstract. This work presents an application of machine learning methods in the area of environmental epidemiology. We have used lifestyle and exposure data from 769 mother-child pairs from Slovenia and Croatia to predict the neurodevelopment of the children, expressed through five Bayley-III test scores. We have applied single- and multi-target (semi-)supervised predictive methods to build models capable of predicting the Bayley-III scores. Additionally, we have used feature ranking methods to estimate the importance of individual lifestyle and mercury exposure attributes on the Bayley-III test scores. The learned models offer useful insights into the effect of prenatal mercury exposure on the neural development of children.

Keywords: Machine learning · Multi-target regression · Environmental epidemiology · Feature ranking

1 Introduction

Mercury (Hg) is known to have adverse impacts on human health [5]. The general population is mainly exposed to mercury in two ways: (1) through the diet - mostly by fish consumption (methyl Hg) and (2) through dental amalgam fillings (Hg^o vapour). Prenatal or early postnatal exposure to methyl Hg can cause neurodevelopmental disorders in children. A recent study [11] investigates the association between prenatal exposure to mercury and neurodevelopment of children, taking into account gene data (apolipoprotein E-*ApoE*). For their purpose they have surveyed mother-child pairs from the central region in Slovenia and from Rijeka, a city on the Croatian coast in the northern Adriatic, and have collected data on their lifestyle and Hg exposure. The neurodevelopment of some

children at 18 months of age has been assessed with the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III) Test [2]. This test helps to identify children with delay in development and assesses their development in different domains.

This study focuses on the data set provided by the PHIME study [11]. Our goal is to train machine learning models that will be able to predict the Bayley-III scores from the lifestyle and exposure data.

2 Data

The PHIME project [11], is part of a larger longitudinal birth cohort study set in the Mediterranean area. The project was designed to investigate the association between prenatal mercury exposure from fish consumption during pregnancy and neuropsychological development of children as well as to investigate the co-exposure to other potentially neurotoxic elements and their role in biological response of the children exposed to Hg in the prenatal period.

The PHIME project started with the recruitment of women in their last trimester of pregnancy or at child birth. The collected data consists of 540 mother-child pairs from Slovenia, and 229 from Croatia. At birth, cord blood and maternal scalp hair were sampled for determination of trace elements concentrations. Mothers filled out a brief inclusion questionnaire, including general information about health, dietary habits and socio-economic status. Six to eight weeks later, breast milk was collected by mothers. Mothers were also required to fill out a detailed questionnaire regarding their health, life-style and dietary habits, socio-economic status, residential and occupational history. The children were followed up at 18 months of age for assessment of their neuropsychological performance using Bayley Scales of Infant and Toddler Development, Third Edition Test (Bayley, 2006), administering cognitive, language and motor (fine and gross) scores:

1. Cognition composite score (CCS)
2. Language composite score (LCS)
3. Motor composite score (MCS)
4. Fine Motor scaled score (FMSS)
5. Gross Motor scaled score (GMSS)

At the time of testing, another (supplementary) questionnaire was filled out by the mothers, including the type of feeding from birth onwards, and behavioural features. Table 1, taken from the workbook on Bayley-III scores, summarizes the information about their values.

The data set is rather incomplete as there are a lot of missing data. For example, the blood and urine features are available only for the Croatian population, which makes up for less than 30% of our data set. For 331 mother-child pairs there are no Bayley-III scores available. In total, the data set consists of 769 mother-child pairs which are described with 82 descriptive attributes (lifestyle and exposure data). Our goal is to predict the values of 5 target attributes (Bayley scores).

Table 1. Descriptive classification of the Bayley target scores

Composite score or equivalent	Class
130 and above	Very superior
120–129	Superior
110–119	High average
90–109	Average
80–89	Low average
70–79	Borderline
69 and below	Extremely low

3 Machine Learning Methods

The availability, dimensionality and type of the target variables that we are trying to predict (Bayley-III scores) determine the machine learning task. Given that there are multiple numerical (integer) target variables, and the fact that not all instances (mother-child pairs) have known values for them, the task at hand is semi-supervised multi-target regression (MTR).

Generally, MTR problems can be approached in two ways: locally or globally. When a local approach is used, one predictive model is learned for each target attribute. Alternatively, when using a global approach to MTR, a single model is learned that is able to predict values for all targets simultaneously. The difference between the two approaches is in the way how the target space is interpreted by the algorithm. When using the local approach, no potential relations between the target attributes can be exploited as the algorithm only focuses on one target attribute. With global approaches, the potential relations between the target attributes are taken into account and can, in some cases, lead to better predictive performance. In cases, when interpretable model types are used, such as decision trees and decision rules, the global approach yields a single interpretable model, as opposed to several interpretable models resulting from the local approach to MTR. It can be challenging for the domain experts to combine local models into an overall interpretation. In this work, we apply both local and global approaches and compare their predictive performance.

Simple models (models with low complexity in terms of how they interpret the input space) often exhibit low predictive power. It is a standard practice to combine many such models into ensembles, which is a known way of improving the predictive performance. An ensemble model combines predictions of the individual models within the ensemble to produce the final prediction. Such ensemble models are also often used for feature ranking.

A feature ranking is a list of all descriptive attributes (inputs), ordered according to their ranking scores. The idea is to determine which descriptive attributes carry the most discriminating information w.r.t. the target attribute(s), i.e., the higher the ranking score of a given descriptive attribute, the higher its importance. In settings with high-dimensional descriptive spaces,

it is often beneficial to reduce the number of descriptive attributes before learning the predictive models. Obviously, removing highly important attributes will result in poor predictive performance. Hence, by using a feature ranking algorithm, one can determine the importance of all attributes and make an educated cut-off (ultimately considering only the attributes with high importance scores). The ranking of attributes can also be used to validate interpretable predictive models, i.e., if an attribute appears high in a decision tree and also has a high feature ranking score, one can be confident that that attribute is quite important w.r.t. discriminating the values of the target attribute(s).

In this study, we have used machine learning methods that produce interpretable predictive models as well as ensembles thereof. Additionally, we have also used a feature ranking method to produce a ranking of descriptive attributes. The used methods are briefly described below. All methods are implemented within the CLUS¹ software. The first type of interpretable models that we used are predictive clustering trees (PCTs). In particular, we have built MTR trees [12], such as the ones shown in Figs. 1 and 2. PCTs are based on the predictive clustering paradigm [3], which generalizes decision trees and parametrizes them to support multitude of structured output prediction tasks, one of which is MTR. Decision trees can also be seen as a hierarchical clustering, where the structure of the decision tree mirrors the clustering hierarchy. Each node represents a cluster that can be described by the tests that appear in the tree. Each node holds a test and if we combine all the tests from the root node to the selected node, we get the description of the cluster at the selected node. A prediction with a PCT is made in the same way as with a standard decision tree.

The importance scores were calculated by using the feature ranking method for MTR [8]. This method is based on ensembles of MTR trees [6] and calculates the Genie3 importance score, based on Random forests (RFs) of 100 PCTs for MTR. The importance scores and the corresponding ranking denote the relative importance of each attribute for predicting all targets, jointly and separately. Highly ranked attributes contain the most discriminative information w.r.t. the target(s) of choice.

Our data set contains missing values for many of the target attributes, i.e., not all mother-child pairs have known values for the Bayley-III scores. The standard PCT top-down induction algorithm does not support such cases. Therefore, we have also used semi-supervised PCTs (SSL-PCTs), an extension to the standard PCT induction algorithm, where both, labeled and unlabeled instances are used for calculating the heuristic score of candidate splits during model learning [7].

Both, PCTs and SSL-PCT were also used in the ensemble setting. In particular, we have used the RF algorithm to build our ensembles of (SSL-)PCTs. The RF algorithm builds an ensemble of many decision trees in order to lift the predictive performance over that of individual PCTs in the ensemble. The RF ensembles with PCTs and SSL-PCTs are denoted as RF-PCTs and RF-SSL-PCTs, respectively.

¹ CLUS software is available for download at <http://source.ijs.si/ktclus/clus-public>.

The second type of interpretable models we used are predictive clustering rules (PCRs) [16]. PCRs are multi-target decision rules, capable of modeling MTR problems. The PCR algorithm implements the standard sequential covering algorithm for rule discovery. In each step, the standard covering algorithm generates a single rule and removes data instances from the data set which are covered by that rule. A data instance is *covered* by a rule if it satisfies its condition clause. The algorithm continues to generate rules until there are no more instances left in our data set. A rule is added to the rule set if the predictive performance of the rule set with the new rule is better than without it. When making predictions, the discovered rules can be used in one of two ways: ordered or unordered. When rules are ordered (such models are often called decision lists), only one rule can be triggered. The order of the rules is determined by the algorithm. If none of the rules are triggered, the default rule is applied. The triggered rule gives the final prediction. This explicitly gives higher importance to those rules that have a higher weight, which can affect the interpretation of the predictions. When using unordered rules, several rules can be triggered, i.e., the instance, for which the predictions are being produced, can, depending on the rule conditions, trigger more than one rule. In those cases, predictions are combined into the final prediction (similar to what is done with tree ensembles).

4 Related Work

The predictive clustering framework has been successfully applied to many diverse problems in the domain of life and medical sciences. Here we name a few. [15] have applied predictive clustering methods to reveal the relationship between fungi and different salt concentrations. Their study has revealed new interesting properties about halophilic fungi and has expanded the knowledge of possible life performance under diverse and extreme environmental conditions. [4] have utilized the clustering aspect of PCTs and have discovered interesting clusters of patients with Alzheimer’s disease that share biological features. The clusters have discovered both gender specific differences and several biological features that can relate to the progression of the disease. [14] have used PCTs to identify subgroups of patients with Parkinson’s disease that would react positively or negatively to medication modification. Their findings will assist physicians that make the therapy modifications for a given patient by narrowing down the number of possible scenarios.

In the recent works by [1, 11, 13] multiple linear regression has been applied to evaluate possible relationship between Hg exposure in prenatal life and 5 neurodevelopmental scores of children at 18 months of age. The model adjusted for potential confounders (mother’s age, child’s sex, birth weight, education of the mother, smoking during pregnancy and concentration of selenium and lead in cord blood) revealed that doubling the Hg concentration on cord blood would result in 0.33 points lower fine motor score. Similar decrement was observed for Slovenian and Croatian populations in the meta-analysis done by [1]. On the other hand, doubling the Hg concentration in cord blood of Apoe $\epsilon 4$ carriers

would decrease the cognitive score for 5.4 points [11]. The observed changes were small on an individual level, but were statistically significant and relevant on a global (population) scale.

To the best of our knowledge, there is no publication related to the application of machine learning methods to the problem of associating prenatal and early postnatal exposome with the neural development of children. Given the geographic specificity of the problem, and its potential to generalise to the entire human population, we consider this publication to be very relevant in the field of environmental epidemiology.

5 Experimental Setup

PCT and SSL-PCT models were built with a variance reduction heuristic and M5Multi pruning method [10]. Same setup was used for both single- and multi-target variants. In the standard PCT top-down induction (TDI) algorithm, the variance reduction heuristic is calculated based on the values of the target variables. Our data set contains some instances where values of the target variables are not known. Therefore, the standard PCT TDI algorithm needs to be instantiated with a different variance reduction heuristic function. In particular, SSL-PCTs introduce the w parameter, used to control the contribution of target and descriptive attributes variances towards the overall variance in the currently observed instances. This parameter is data set sensitive and must be optimized [7] for each data set individually. Therefore, we optimize it by using 5-fold internal cross-validation to select one of the candidate values which range between 0.1 and 1.0 with a step of 0.1.

To build the random forest ensemble models (for prediction and feature ranking) we used 100 individual (SSL-)PCTs as base learners. Each (SSL-)PCT was allowed to grow without limiting the number of instances in the leaf nodes, i.e., no pre-pruning was applied, and had only a subset of $\text{sqrt}|D|$ random attributes available when learning, where $|D|$ is the number of descriptive attributes. The final prediction of the ensemble is obtained by taking the predictions of the individual (SSL-)PCTs and calculating their arithmetic mean.

Ordered PCRs were learned by using the standard covering algorithm, adding additional rules only if they improve the predictive performance of the model. Unordered PCRs were learned by using the weighted covering algorithm, where the only difference from the former algorithm is that we do not immediately remove instances that are covered by a new rule, but rather decrease their weight inversely proportional to the error that the new rule makes when predicting their target values. For both rule-based models we used multiplicative dispersion search heuristic and added rules to the resulting rule set if and only if they cover at least 45 instances. Unordered PCRs were obtained by setting the weight controlling the amount by which weights of covered instances are reduced within the error weighted covering algorithm, to 0.5 and the instance’s weight threshold to 0.1 (if an instance’s weight falls below this value, it is removed from the learning set).

We calculated root relative squared errors (RRSEs) to evaluate the predictive performance of the generated models. RRSE is relative to what it would have been if we had just predicted the average value for each score. Thus, the relative squared error takes the total squared error from our model and normalizes it by dividing it with the total squared error of a model that simply predicts the average. In general, we want the RRSEs to be lower than one and as close to zero as possible. The formula for calculating RRSE for the target attribute t is:

$$\text{RRSE}_t = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y}_i)^2}}, \quad (1)$$

where N is the number of data points, y_i is the true target value of instance i , \hat{y}_i is the predicted value for the target and \bar{y}_i is the arithmetic mean, calculated over the target values within the training set. The average RRSE over T target attributes is then calculated as:

$$\text{aRRSE} = \frac{1}{T} \sum_{t=1}^T \text{RRSE}_t. \quad (2)$$

We used 10-fold cross-validation to estimate the RRSEs of our models. Table 4 contains the obtained RRSEs values.

6 Results

The algorithms for building PCTs and SSL-PCTs yield models that can easily be interpreted. The produced PCT and SSL-PCT models are shown in Fig. 1 and Fig. 2, respectively. The PCT model identified the child’s gender, the concentration of methyl Hg in the mother’s blood and the mother’s age as the most relevant attributes. The semi-supervised PCT model identified the concentration of methyl Hg in the cord blood and the number of pregnancies as most relevant attributes.

The PCR model illustrated in Table 2 consists of an ordered list of 9 rules. Each data instance that we are trying to predict is tested against the condition clause in the rules in the specified order. Prediction is done by the first rule that has its condition clause satisfied (i.e. the rules are ordered). If there exists no such rule, then the prediction from the default rule is applied.

Similarly, in Table 3 we illustrate an unordered PCR model. There, the collection of rules can be seen as a set rather than a list. An instance is tested against each rule and a prediction is obtained by averaging the predictions from all individual rules that had their condition satisfied by the instance. If the instance fails to satisfy any condition, then the prediction from the default rule is taken as final.

Table 4 summarizes the values of root relative squared errors (RRSE) for each method per target score.

The random forest of PCTs with Genie3 feature ranking method outputs a list of attributes ordered by their importance scores. Each score is calculated

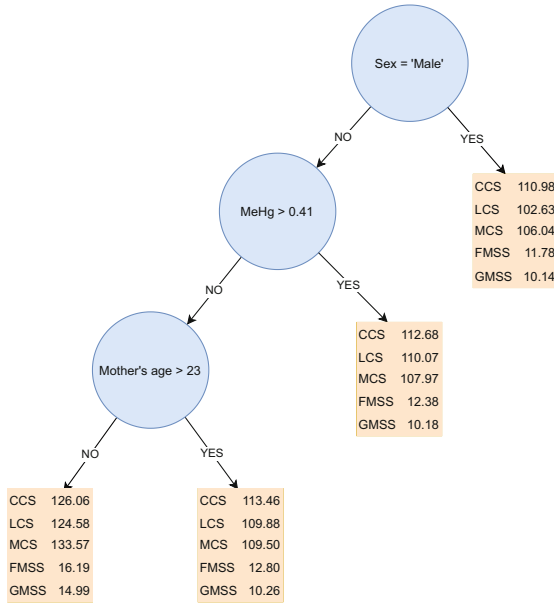


Fig. 1. A predictive clustering tree for MTR predicting the values of the five Bayley-III scores simultaneously.

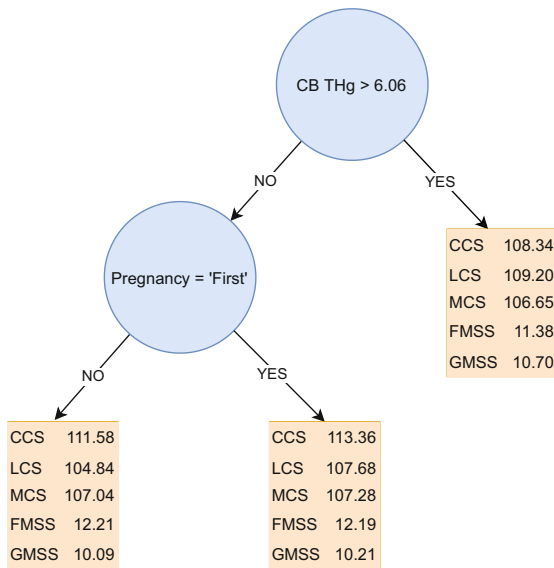


Fig. 2. A semi-supervised predictive clustering tree for MTR predicting the values of the five Bayley-III scores simultaneously.

Table 2. A list of PCRs for MTR predicting the five Bayley-III scores simultaneously. The rule conditions are given in the second column. The predictions are in columns CCS, LCS, MCS, FMSS and GMSS.

#	Rule conditions	CCS	LCS	MCS	FMSS	GMSS
1	CB_Zn \leq 1507.795	117.06	106.73	107.50	12.32	10.04
2	CB_Serum_Ca $>$ 2.89 AND CB_Serum_Mg $>$ 0.71	113.26	106.15	108.69	12.69	10.08
3	M_milk_Mn \leq 1.599	111.19	106.89	106.89	12.10	10.06
4	CB_As \leq 0.466	114.34	105.32	106.76	12.36	9.78
5	GA $>$ 40 AND CB_Se \leq 112.38	112.17	105.34	106.13	12.17	9.78
6	M_milk_Zn \leq 2539.085	108.04	104.83	106.56	12.00	10.13
7	no_amalgams $>$ 2 AND M_hair_THg $>$ 26	114.02	107.43	107.02	12.21	10.06
8	frozenfish = 3 AND gest_age = 2 AND BMI $>$ 18.9069	106.41	105.39	107.26	11.60	10.71
9	CB_Serum_Ca $>$ 2.51 AND BMI \leq 33.91	115.65	112.34	110.84	12.32	11.23
10	Default	103.33	101.91	102.79	10.87	10.00

Table 3. A set of PCRs for MTR predicting the five Bayley-III scores simultaneously. The rule conditions are given in the second column. The predictions are in columns CCS, LCS, MCS, FMSS and GMSS.

#	Rule conditions	CCS	LCS	MCS	FMSS	GMSS
1	CB_Zn \leq 1507.795	117.06	106.73	107.50	12.32	10.04
2	CB_Serum_Ca $>$ 2.91 AND CB_Serum_FeIII \leq 44.10	114.12	107.12	107.74	12.65	9.81
3	M_milk_Mn \leq 1.397 AND M_milk_Se $>$ 8.344747	110.27	106.48	106.87	12.14	10.04
4	Default	111.70	106.38	107.19	12.09	10.21

as the average decrease in impurity while inducting the PCTs. The greater the decrease, the more important an attribute is. Table 5 gives a snapshot of the list, showing only the top and bottom three attributes as well as some other attributes in between that appear in the (SSL-)PCT models.

The ranking in Table 5 identifies the concentration of total Hg in the mother’s hair as the most important attribute for determining the Bayley-III scores, the concentration of Cu (copper) and As (arsenic) in the cord blood as the second and third most important attribute, and continues down to the information about whether the mother is smoking currently, the concentration of methyl Hg

Table 4. The RRSEs for each multi-target model the an overall average RRSE (aRRSE) across all five scores. Bolded numbers denote the best-performing model for a given score. Underlined number denotes the best overall-performing model.

Method	Cognition	Language	Motor	Fine motor	Gross motor	Overall
PCT	1.0094	0.9961	1.0158	1.0071	1.0147	1.0086
SSL-PCT	1.0094	1.0130	1.0096	1.0070	1.0016	1.0081
Ordered PCR _s	0.9836	1.0168	1.0117	1.0095	1.0136	1.0070
Unordered PCR _s	0.9942	1.0048	1.0019	1.0029	0.9944	0.99968
RF-PCT _s	0.9855	0.9928	1.0118	0.9888	1.0098	<u>0.9977</u>
RF-SSL-PCT _s	0.9868	1.0072	1.0113	0.9931	1.0051	1.0007

Table 5. Feature ranking of the top ten/bottom three attributes and of those that appear in the (SSL-)PCT models.

Rank	Attribute name	Importance score
1	M_hair_THg	806.3
2	CB_Cu	786.8
3	CB_As	769.3
4	mothers_age	768.4
5	CB_Pb	762.5
6	CB_MeHg	697.5
7	freshfish	680.6
8	BMI	672.6
9	CB_Se	665.4
10	CB_THg	663.9
11	child_sex	605.3
41	M_blood_MeHg	123
54	firstpregnancy	90.8
...
80	smokingcurrently	9.7
81	M_milk_MeHg	5.9
82	numberofcigarettesperday	1.9

in her milk and the number of cigarettes she smokes per day as the three least important factors. Mother’s age, child’s sex and the concentration of total Hg in the cord blood can also be considered very important, because they appear in the tree models, and rank very high in the rankings.

The above described results are in line with the main outcomes of the Slovenian and Croatian birth cohort study PHIME, which tested the existence of association between exposure to methyl Hg in prenatal or early life and neurodevelopmental performance of children. There have been reports [9, 11, 13] on

significantly negative association between total or methyl Hg in cord blood or maternal hair (both indicate exposure to Hg in prenatal period) and fine motor scores of Bayley-III assessment, as well as cognitive scores, although only in a sub-population of carriers of apolipoprotein epsilon 4 gene variant [11, 13]. Prenatal exposure to Hg was confirmed as a significant predictor for cognitive and fine motor scores regardless of the genotype by the RF-PCT+Genie3 and SSL-PCT methods, ranking at the first position in the former method, and being the root in the tree model in the latter. However, some additional predictors were revealed in the present study, namely copper (Cu) and arsenic (As) concentrations in cord blood, the first known for its pro-inflammatory effects and the second for potential neurotoxicity, similarly as Hg. Both Hg and As share the source of exposure which might explain the observed significance. First pregnancy also came out as an important attribute, which is yet to be explained (Table 6).

Table 6. The RRSEs for each single-target model and the overall average RRSE (aRRSE) across all five scores. Bolded numbers denote the best-performing model for a given score. Underlined number denotes the best overall-performing model.

Method	Cognition	Language	Motor	Fine motor	Gross motor	Overall
PCT	1.0157	0.9745	1.0067	1.0226	1.0212	1.0081
SSL-PCT	1.0221	1.0092	1.0040	1.0051	1.0000	1.0080
Ordered PCRs	1.0147	1.0364	1.0066	1.0134	1.0099	1.0162
Unordered PCRs	0.9993	1.0113	1.0004	1.0083	1.0013	1.0041
RF-PCTs	0.9921	0.9895	1.0165	0.9934	1.0158	1.0014
RF-SSL-PCTs	0.9909	0.9892	1.0099	0.9896	1.0112	<u>0.9981</u>

In our particular case, local and global approaches exhibit similar predictive performance. Local approach outperforms the global only for the best language and motor score, and, for other scores, including the overall one, the global approach is marginally better. SSL-PCTs perform slightly better than fully supervised PCTs.

7 Conclusion

In this paper, we have applied machine learning methods to model the associations between exposure to mercury in the environment and neural development of children. In this multi-target regression problem our target attributes represent Bayley-III scores. The problem was modeled with PCTs, RF of PCTs and semi-supervised variants of them, as well as with PCRs. We have also produced a ranked list of attributes, according to their importance when used for predicting the target attributes.

All methods generate models with comparable predictive performance but the best performing model was generated with the RF-PCTs method. Given the

specific nature of the problem, an observation can be made that a global approach is better than a local one, because it generalizes better and thus captures only the high-level relationships between the features, and does not succumb to the noise introduced by the missing data and limited number of instances. The random forest of PCTs model marginally outperformed the baseline model (simply predicting the average value) for the targets Cognitive score, Language score and Fine motor (scaled). Similarly, the random forest of SSL-PCTs outperformed the baseline model for the targets Cognitive score and Fine motor (scaled). PCRs and PCTs were able to predict one target (Language and Cognition score, respectively) better than the simple baseline model. Other models were not able to outperform the baseline model, predicting arithmetic mean for individual targets, calculated on the training data. We believe that the poor predictive performance of generated models can be attributed to high sparsity of the data set. Obtaining more labeled data should result in better performing models.

Despite this rather low predictive power of models, the results obtained are in line with the main findings of previous work on the PHIME data set. Some additional predictors were revealed, providing valuable insight into the environmental epidemiology aspects of chronic low-level exposures and will be further evaluated by using an expanded version of the data set. Application of machine learning methods is particularly valuable in studies like PHIME, where a large number of attributes is used to make a prediction within a rather narrow range of values.

Acknowledgements. SP would like to acknowledge the financial support in the form of a scholarship of the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia. SD, MB and SP would like to acknowledge the grant number P2-0103 (the research programme Knowledge Technologies). All authors acknowledge the NEURODYS project (J7-9400, Neuropsychological dysfunctions caused by low level exposure to selected environmental pollutants in susceptible population) for financial support of the overall work. JST, DM and MH also acknowledge the EU funded 6th FP project PHIME for providing the data used in this work.

References

1. Barbone, F., et al.: Prenatal mercury exposure and child neurodevelopment outcomes at 18 months: results from the Mediterranean Phime cohort. *Int. J. Hyg. Environ. Health* **222**(1), 9–21 (2019). <https://doi.org/10.1016/j.ijheh.2018.07.011>
2. Logsdon, A.: Bayley Scales of Infant and Toddler Development. 3rd edn (2008). http://images.pearsonclinical.com/images/pdf/bayley-iii_webinar.pdf, last accessed 07.09.2020
3. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 55–63 (1998)
4. Breskvar, M., Zenko, B., Džeroski, S.: Relating biological and clinical features of Alzheimer’s patients with predictive clustering trees. In: International Multi-Conference Information Society (2015)

5. Kim, K.H., Kabir, E., Jahan, S.A.: A review on the distribution of hg in the environment and its human health impacts. *J. Hazard. Mater.* **306**, 376–385 (2016). <https://doi.org/10.1016/j.jhazmat.2015.11.031>
6. Kocev, D., Vens, C., Struyf, J., Džeroski, S.: Tree ensembles for predicting structured outputs. *Pattern Recogn.* **46**(3), 817–833 (2013)
7. Levatić, J., Kocev, D., Ceci, M., Džeroski, S.: Semi-supervised trees for multi-target regression. *Inf. Sci.* **450**, 109–127 (2018). <https://doi.org/10.1016/j.ins.2018.03.033>
8. Petković, M., Kocev, D., Džeroski, S.: Feature ranking for multi-target regression. *Mach. Learn.* **109**(6), 1179–1204 (2019). <https://doi.org/10.1007/s10994-019-05829-8>
9. Prpic, I., et al.: Prenatal exposure to low-level methyl mercury alters the child's fine motor skills at the age of 18 months. *Environ. Res.* **152** (2016). <https://doi.org/10.1016/j.envres.2016.10.011>
10. Quinlan, J.R.: Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence, pp. 343–348. World Scientific (1992)
11. Snoj Tratnik, J., et al.: Prenatal mercury exposure, neurodevelopment and apolipoprotein e genetic polymorphism. *Environ. Res.* **152**, 375–385 (2017). <https://doi.org/10.1016/j.envres.2016.08.035>
12. Struyf, J., Džeroski, S.: Constraint based induction of multi-objective regression trees. In: International Workshop on Knowledge Discovery in Inductive Databases, pp. 222–233. Springer, Berlin (2005). <https://doi.org/10.1007/978-3-540-75549-4>
13. Trdin, A., et al.: Mercury speciation in prenatal exposure in Slovenian and Croatian population - Phime study. *Environ. Res.* **177**, 108627 (2019). <https://doi.org/10.1016/j.envres.2019.108627>, <https://www.sciencedirect.com/science/article/pii/S0013935119304244>
14. Valmarska, A., Miljkovic, D., Konitsiotis, S., Gatsios, D., Lavrac, N., Robnik-Sikonja, M.: Combining multitask learning and short time series analysis in Parkinson's disease patients stratification. In: Conference on Artificial Medicine in Europe, pp. 116–125, May 2017. https://doi.org/10.1007/978-3-319-59758-4_13
15. Zajc, J., et al.: Chaophilic or chaotolerant fungi: a new category of extremophiles? *Front. Microbiol.* **5** (2014)
16. Ženko, B.: Learning predictive clustering rules. Ph.D. thesis, University of Ljubljana (2007), <http://eprints.fri.uni-lj.si/709/1/zenko%2Dphd%2Dthesis.pdf>