A Hybrid Approach to Identifying Key Factors in Environmental Health Studies

Shi Dong[1], Zlatan Feric[1], Xiangyu Li[1], Sheikh Mokhlesur Rahman[3], Guangyu Li[1], Chieh Wu[1],
April Z. Gu[2], Jennifer Dy[1], David Kaeli[1], John Meeker[4], Ingrid Y. Padilla[5],
Jose Cordero[6], Carmen Velez Vega[7], Zaira Rosario[7], Akram Alshawabkeh[1]
[1]Northeastern University [2]Cornell University [3]Bangladesh University of Engineering and Technology
[4]University of Michigan [5]University of Puerto Rico at Mayaguez [6]University of Georgia
[7]University of Puerto Rico Medical Campus

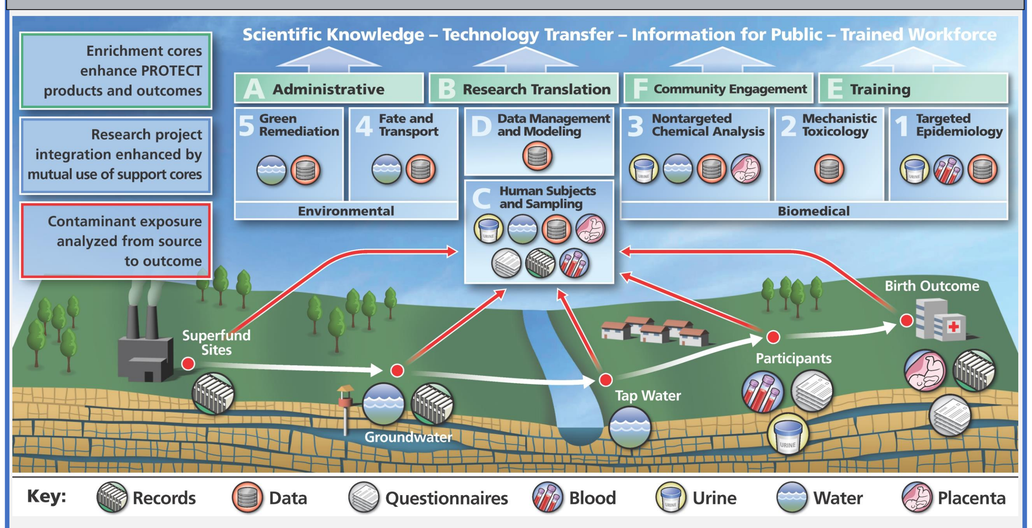
INTRODUCTION

Data analytics frameworks have become a key tool in discovery in public health and environmental science research, building on recent advances in machine learning algorithms. We leverage these frameworks in the NIEHS P42 PROTECT Center.

In this poster, we present:

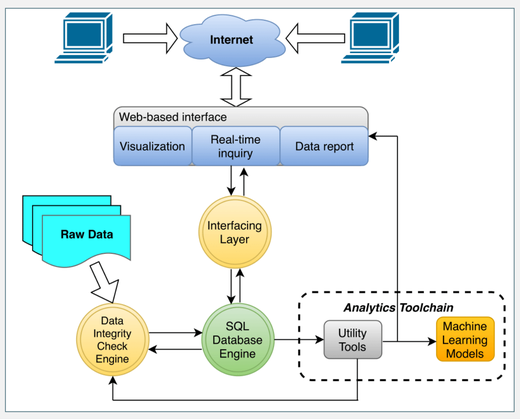
- Our use of Linear Correlation, Normalized Mutual Information, Logistic Regression, and Decision Trees to identify dominant factors/features potentially responsible for the high rate of premature births in Puerto Rico.
- We discuss our customized end-to-end analytics toolchain which performs preprocessing of all PROTECT data.
- We identify top-ranked features produced by our model as potential key contributors of high preterm birth rates in Puerto Rico, as well as the model performance across selected algorithms.

PROTECT



- Human subjects information - medical history, reproductive health records, product use data surveys and birth outcomes
- Biological samples - blood, urine, hair and placental samples
- Environmental samples and measurements - soil samples, well and tap water samples, historical Environmental Protection Agency (EPA) data, soil samples and superfund site data

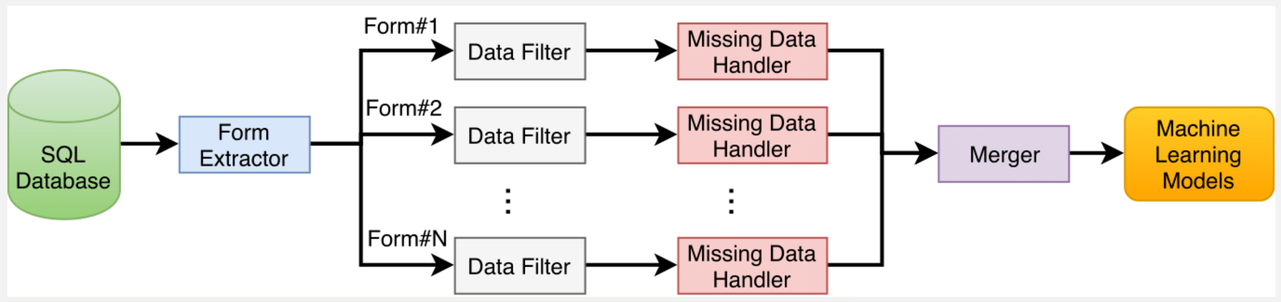
PROTECT Database System



- A web-based interface for data visualization, data query handling, and data report generating
- A data cleaning and a database engine to ensure the data integrity and manage the data, respectively
- A set of utility tools for multiple purposes, including workflow management, data statistics and visualization, and data processing using machine learning algorithm

Feature Name	Description
SVHEALTH	*The current health level during the second visit.*
MR1PULSER	*Whether or not a pulse is reported.*
PESTSTOREV2	*Whether or not a pesticide is currently stored in the house, during the second visit.*
DATEPRENATCAREM	*The month of the pregnancy in which prenatal care began.*
FVBPDIAS	*Diastolic Blood Pressure*
ULTRAGESTAGEWFV	*Ultrasound estimated gestational age, during the first visit.*
MR1FHY	*Fundal Height*
FVHEARBUZZ	*Whether the mother reports hearing a buzz.*
SLEEPCHNGNEGPOS	*The rate of impact when changing sleep pattern*
CSECTION	*Is there a c-section performed on at least one of the previous pregnancies?*
HOMTIME	*How long the mother lived in the current house?*
PAINPILLTYP	*The type of pain killer used most often.*
HOMSURRAREA	*The type of the surrounding environment.*
BUGSPRY	*How often has the mother used insect repellents, creams or wipes?*
HOMCHILDREN	*The number of children that are living in the current house.*
ED	*The mother's education level.*
FVBPSYS	*The mother's systolic blood pressure.*
SHVCREAM	*How often the patient used shaving cream.*
FVSCHL	*The type of school the mother is currently attending during the first visit.*
WTPREPREG	*The weight in pounds before conceiving.*

Data Preprocessing Toolchain



Similarity-based Missing Data Handling Algorithm

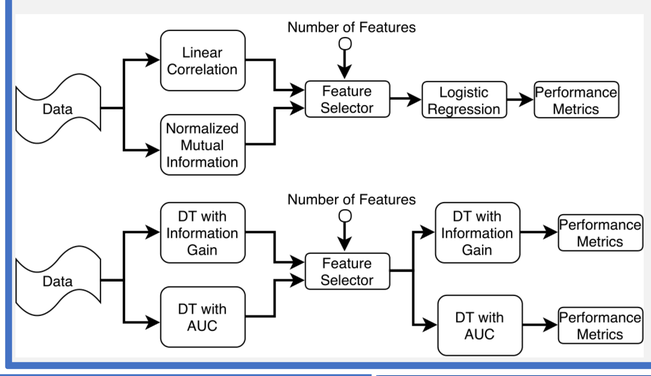
1. Transform categorical data entries to use a one-hot encoding format
2. Normalize the numerical data using sample means and variances
3. Calculate a similarity matrix sample-by-sample, based on equation on the right side
4. Fill in missing values using the value of the same feature in its first nearest neighbor

$$S_{ij} = \frac{\vec{x_i} \cdot \vec{y_j}}{N}$$

S_{ij} stands for the similarity between a sample vector x_i and y_j. The · denotes a dot product operation, and N denotes the number of features involved in the dot product operation

Data Analytics

Number of samples	681
Number of positive samples	64
Number of negative samples	617
Number of features	911



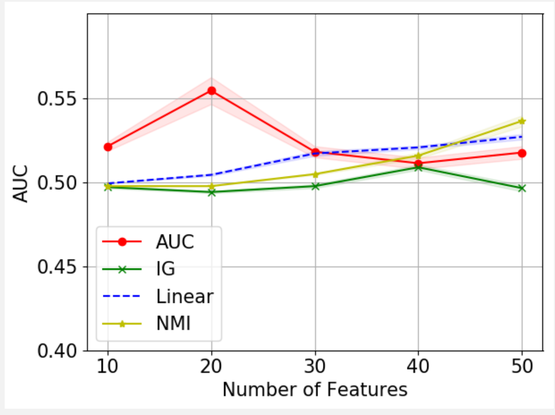
Machine Learning Algorithms

- Feature ranking
  - Linear Correlation, Normalized Mutual Information, and Decision Tree
- Performance evaluation
  - Logistic Regression and Decision Tree
- Decision Tree with Information Gain
- Decision Tree with AUC

$$Info\_Gain(D, SD) = Entropy(D) - \sum_{i=1}^{n} \frac{Num(SD_i)}{Num(D)} Entropy(SD_i)$$

$$if LPR_1 > LPR_2 \quad AUCSplit = \frac{p_1 n + p n_2}{2pn}$$
$$else \quad AUCSplit = \frac{p_2 n + p n_1}{2pn}$$

Results



- From the figure above, we can see that the Decision Tree using the AUC (criterion) performs the best when using 20 features.
- Even though the absolute AUC is not very high, the relative difference is still significant.
- We select the 20 top-ranked features using the Decision Tree with AUC criteria as the final feature set, as listed in the left table.

Conclusion

- We present a framework used to analyze the complex data in PROTECT to identify factors contributing to the high rate of preterm birth in Puerto Rico
- We present a multi-stage approach to data preprocessing
- We describe hybrid methods for feature selection and performance evaluation

Reference

[1] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification (2Nd Edition). Wiley-Interscience, 2000.
[2] X. Li, L. Yu, D. Kaeli, Y. Yao, P. Wang, R. Giese, and A. Alshawabkeh, "Big Data Analysis on Puerto Rico Testsite for Exploring Contamination Threats," ALLDATA, 2015.
[3] P. Flach, Machine Learning: The Art and Science of Algorithms That Make Sense of Data. New York, NY, USA: Cambridge University Press, 2012.
[4] K. Delaney, Inside Microsoft SQL Server 2000. Redmond, WA, USA: Microsoft Press, 2000.
[5] EarthSoft, "EQuIS Professional," http://www.earthsoft.com/products/professional/, 2017 (accessed October 1, 2017).
[6] ——, "EQuIS Enterprise," http://www.earthsoft.com/products/enterprise/, 2017 (accessed October 1, 2017).
[7] Wen Jiang, "PyPyODBC 1.3.5," https://pypi.python.org/pypi/pypyodbc, 2017 (accessed October 1, 2017).
[8] C. Ferri, P. A. Flach, and J. Hernandez-Orallo, "Learning decision trees using the area under the roc curve," in Proceedings of the Nineteenth International Conference on Machine Learning, ser. ICML '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 139–146.

Acknowledgment

The work presented in this paper is supported in part by NIEHS P42 Program award P42ES017198, and NSF awards OAC-1559894 and IIS-1546428.