Using Undersampling with Ensemble Learning to Identify Factors Contributing to Preterm Birth

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

PROTECT
Puerto Rico Testsite for Exploring Contamination Threats

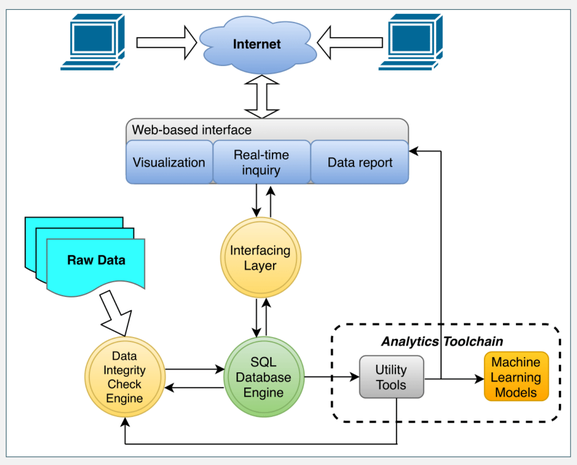
INTRODUCTION

Preterm birth has been identified as a major cause of serious birth defects and even infant deaths. Given the availability of environmental health records for preterm birth, data-driven research can begin to identify potential factors contributing to premature birth outcomes.

In this poster, we present:

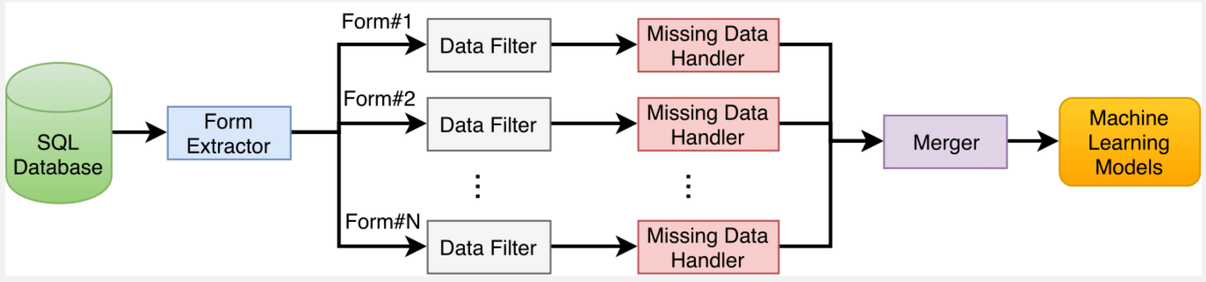
- a customized data preprocessing toolchain equipped with similarity-based missing data handling capabilities
- an ***undersampling ensemble learning*** model for both feature selection and performance evaluation
- analysis results using multiple ensemble feature selection methods to handle missing data

PROTECT Database System



- A web-based framework for data management
- A data cleaning ensures the integrity of the data
- The system supports workflow management, data analysis and visualization

Data Preprocessing Toolchain



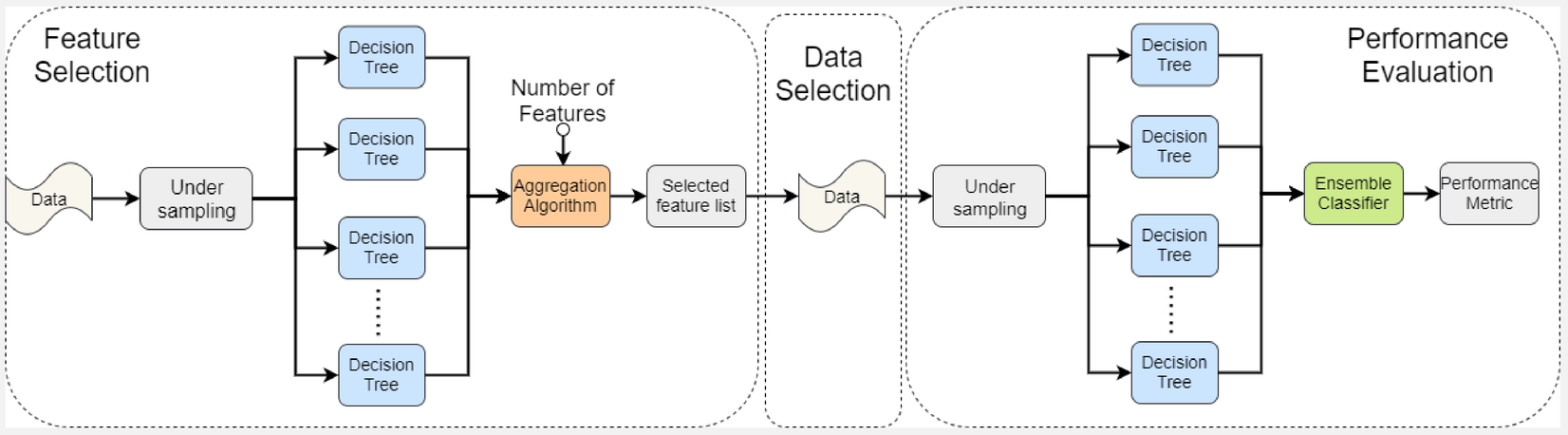
Similarity-based Missing Data Handling Algorithm

1. Transforms categorical data entries to one-hot encoding format
2. Normalizes numerical data using mean and variance
3. Calculates a similarity matrix sample-by-sample using the equation
4. Fills in missing values using nearest neighbor of the same feature

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

S_{ij} stands for the similarity between a sample vector x_i and y_j. The $\cdot$ denotes dot product operation and N the number of features

Undersampling Ensemble Learning Model



Description

- Undersampling
  - Given dataset P of minority class (positive) and dataset N with majority class (negative)
    - $|P| \ll |N|$
  - Randomly select samples from N to construct a series of subset $\overline{N_i}$ so that $|P| = |\overline{N_i}|$
  - Combine P and all $\overline{N_i}$ to generate a series of undersampling training instances $\{P \cup \overline{N_1}, P \cup \overline{N_2}, \ldots, P \cup \overline{N_n}\}$
- Ensemble Feature Selection
  - Complete Linear Aggregation (CLA)
  - Weighted Mean Aggregation (WMA)
  - Feature Occupancy Frequency (OFA)
  - Classification Accuracy Based Aggregation (CAA)

Missing Data Rate and Accuracy Based Aggregation (MAA)

$$\frac{Accuracy}{(Missing\_Rate + \alpha)^{\beta}}$$

- Features with low missing data rate will experience less data variance after missing data handling
- Adjust the accuracy based on the magnitude of data variance represented by the missing data rate

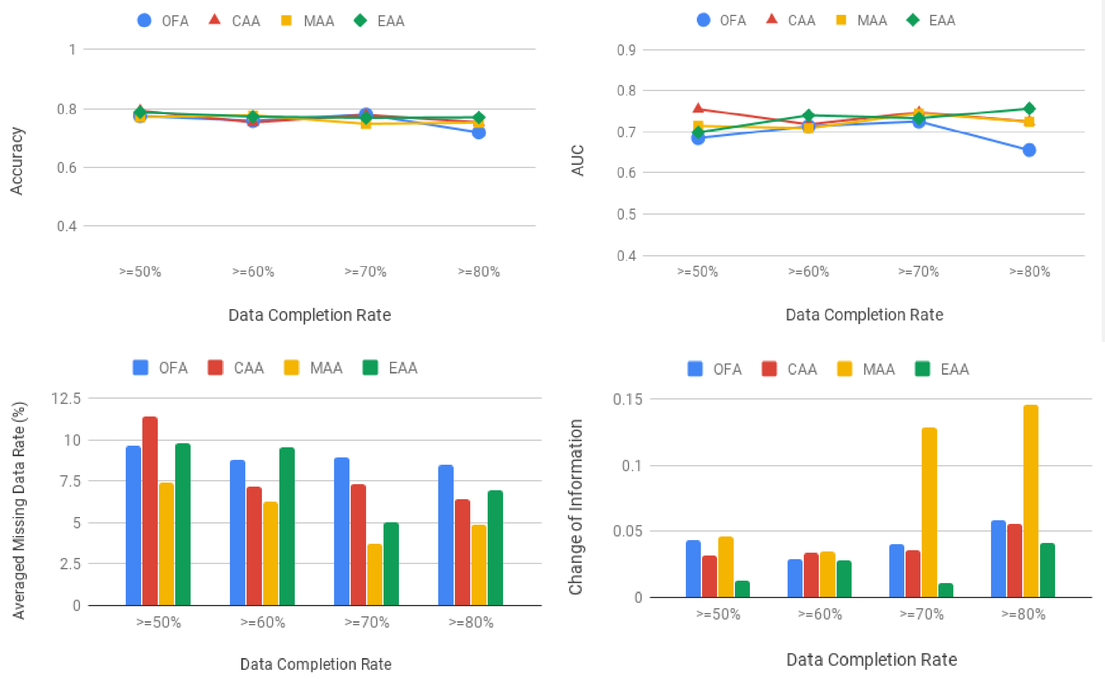
Entropy and Accuracy Based Aggregation (EAA)

$$\frac{Accuracy}{(\Delta Entropy + \alpha)^{\beta}}$$

- A small entropy difference preserves the information entropy, leading to small data variance
- Adjust the accuracy based on the magnitude of data variance represented by the missing data rate

Results

	CLA	WMA	OFA	CAA
Accuracy	0.42	0.69	0.72	0.75
AUC	0.5	0.65	0.66	0.72



- The proposed MAA and EAA targets selecting features with less data variance, focusing on different aspects
- The two methods have 50% different features, meaning that EAA is able to select important features potentially ignored by MAA
- The EAA can serve as a complement of MAA when selecting features for dataset

Conclusions

- We present an undersampling ensemble model for selecting key factors contributing to high preterm birth rates in northern Puerto Rico
- We found that the proposed model, equipped with CAA, can achieve a 44% improvement in AUC as compared to previous studies
- We also propose two novel feature selection methods, MAA and EAA, limiting the variance introduced by missing data values

Acknowledgments

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