Demographic Data-driven Deprivation Index for Predicting Chronic Diseases

Olugbenga Iyiola
Department of Computer Science
The University of Texas at El Paso
El Paso, TX 79968
otiyiola@miners.utep.edu

Monika Akbar
Department of Computer Science
The University of Texas at El Paso
El Paso, TX 79968
makbar@utep.edu

Abstract—Researchers have worked on modeling and predicting the likelihood of developing chronic diseases, such as diabetes and high blood pressure, using medical data (e.g., heart-rate, blood sugar). However, many of these diseases demonstrate strong links with demographics and socio-economic status (e.g., race, gender, income). It is also less time-consuming to retrieve demographic and socio-economic data, some of which are publicly available through the U.S. Census Bureau, than to conduct medical tests. Hence, such data can give a quicker estimate of the susceptibility of a person to a chronic disease.

In this work, we study the effect of using medical vs. demographic data for modeling and predicting two chronic diseases: diabetes and high blood pressure. We proposed an updated deprivation index to build disease models that consider demographic data. Our results indicate demographic data are as good or better indicators for predicting chronic diseases.

Index Terms—Chronic Diseases, Deprivation Index, Disease Model, Predicting Diseases.

I. INTRODUCTION

There are many studies on disease prognosis using machine learning algorithms [1]–[3]. These diseases can be broadly categorized into two groups: chronic and infectious. Significant progress has been made on accurately detecting many chronic diseases using machine learning algorithms. As a consequence, it is now possible to detect some of these diseases earlier than before, thus significantly reducing their mortality rates [4].

Studies mostly rely on diagnostic test results (or, patients’ vital information) for determining the likelihood of developing chronic diseases [5]. We observe that in such studies, demographic data have been used less, although such data hold much promise as they contain valuable information about common trends seen across communities. Accordingly, a few researchers have developed deprivation indices integrating different types of demographic data [6]. These indices are useful in measuring health outcomes in a geographical area. Such indices mostly use generic demographic attributes such as unemployment rate, household size, and the number of vehicles [7], [8]. However, for communities that demonstrate different traits - such as, communities on the border of a country - these indices fall behind in accurately capturing the vulnerability towards certain diseases. In this paper, we extend an existing deprivation index, Townsend [9], to include community-specific demographics data, such as, citizenship information and the preferred choice of language. We retrieved these data from the U.S. Census Bureau [10]. The data is available at multiple levels of abstraction (e.g., zip code, census tract). We used the zip code level (i.e., the ZCTA*) census data. In particular, we investigated if a person’s area of residence, along with the proposed deprivation index, is a better indicator for predicting his/her vulnerability towards chronic diseases. Overall, our results demonstrate that the index significantly improves the performance of the algorithms tested in predicting the likelihood of someone developing a chronic disease.

The rest of the paper is organized as follows: Section II presents some of the relevant work in this research area; Section III gives a detailed description of the methodology adopted in this work; Section IV describes all the steps taken to process the datasets; Section V explains the steps taken to construct the new deprivation index proposed in this paper; Section VI summarizes the results from the selected machine learning algorithms; Section VII concludes the research suggesting possible future directions and ways to improve on the work done in this study.

II. RELATED WORK

Deprivation index is commonly used to measure levels of disadvantage across small areas [9]. The index is helpful in assessing the health outcomes as well as addressing health inequities for a geographical area. Researchers at the University of South Carolina developed a census-based small-area socio-economic deprivation index optimized to predict chronic disease burden among Medicaid recipients in South Carolina, a largely impoverished southern state where more than one in five residents are enrolled in the Medicaid system [7]. In developing the index named Palmetto Small-Area Deprivation Index (Palmetto SADI), they used U.S. Census Bureau population and housing data at the ZIP Code Tabulation Area (ZCTA) level. They assessed two variables in each of five distinct socio-economic domains:

*ZIP Code Tabulation Areas (ZCTAs) are generalized areal representations of United States Postal Service ZIP Code service areas.
education, income, employment, social fragmentation, and material deprivation. A logistic regression analysis was then used to compare the ability of Palmetto SADI to predict chronic disease with other similar deprivation indices.

Our approach improves this by considering other relevant factors such as immigration and language barrier, that significantly affect the living conditions of people living in border towns, like El Paso, where most of the people in our datasets reside. We also use a methodology that addresses missing values, imbalanced datasets, and ensures there is no percentage estimation bias, i.e., where worst areas are those with few observations, when measuring deprivation [11]–[14].

III. PROBLEM DESCRIPTION AND METHODOLOGY

In this section, we discuss the methodology employed to study the effects of the proposed demographic data-driven deprivation index for predicting two chronic diseases: diabetes and high blood pressure.

A. Problem Description

The city of El Paso shares a border with Mexico. The city is the second-largest bi-national urban city along the US-Mexico border. The city hosts people from the neighboring city of Juárez from Mexico who come to El Paso for several reasons including, education, employment, healthcare, and residence. Around 70% of the residents of El Paso speak Spanish. Many residents have close ties with their families on the other side of the border. This is a bilingual and bi-cultural city.

According to the Center for Medicare and Medicaid, it is estimated that about 32.7% of the adult population in El Paso, Texas are currently living with diabetes while about 31.6% in the city are living with high blood pressure [15], which result in a negative economic consequence for the county. This research aims to study whether the prevalent socio-economic conditions in this border town have any connections to these chronic diseases.

Let us assume a set of the two chronic diseases, \( CD = \{CD_d, CD_{hbp}\} \), where \( CD_d \) is diabetes and \( CD_{hbp} \) is high blood pressure. Also, let \( Y \) be a two-dimensional dataset, where every row represents a vector \( v \) with information of respondents, such that \( v = \{x_{demo}, x_{health}\} \), where \( x_{demo} \) are the demographic features (e.g., income, race) and \( x_{health} \) are the health features (e.g., vaccination status, blood pressure). Lastly, assume for every zip code in a reference area (e.g., county, state, country), we have a deprivation index, \( DI_{zip} \). Our goal is to develop a new deprivation index that incorporates demographic data of a border city, such as El Paso, so that existing machine learning algorithms can be used to accurately predict occurrences of \( CD_d \) and \( CD_{hbp} \). The contributions of this research are hinged on the fact that it provides answers to the following two questions.

1. What is the effect of the proposed deprivation index in predicting the occurrence of two chronic diseases: diabetes and high blood pressure?
2. What is the effect of using demographic data for predicting chronic diseases compared to health-related data?

In order to answer these questions, we use the following two datasets [17]:

- Atlas X: Used to estimate accuracy and performance of the algorithms when only demographic information of users is available.
- Atlas Y: Used to compare the accuracy and performance of the algorithms on both demographic and health-related data.

B. Approach

The proposed approach in this study for the prediction of diabetes and heart blood pressure using demographic data optimizes the predictive performance of selected machine learning algorithms by efficiently replacing the missing values in the dataset, employing a combination of feature engineering procedures, and introducing a deprivation index based on the respondents’ zip codes. Specifically, the procedure consists of the following steps:

- Gain a better understanding of the dataset through exploratory data analyses.
- Efficient replacement of the missing values in the dataset.
- Selection of the best features for the algorithms and balancing the dataset.
- Construction of the proposed Novel deprivation index.
- Conduct comparative analyses on the effects of the proposed Novel deprivation index on detecting the two chronic diseases: diabetes and high blood pressure.
- Study the effects of demographics vs. health-related data on detecting patients’ susceptibilities to chronic diseases.

IV. DATASET PROCESSING

The first dataset used in this research was named Atlas X, and it contains de-identified structured text information of 21,810 people, mainly residents of El Paso. It has 75 features which include demographic information such as age, income, education, race, address, insurance information, and management data on cancer, flu, diabetes, and high blood pressure. The data was accessed from a secure database, ATLAS, which stores historical and current regional health information from existing databases and community-based sources to supplement information from Medicaid, low-income, and uninsured populations not captured through traditional electronic medical records. These databases include the U.S. Census tract, Medicare and Medicaid (CMS), Texas Health and Human Services Commission (HHSC), Centers for Disease Control and Prevention (CDC), American Community Survey, and data collected through the Medicaid Waiver 1115 Program [17].

Atlas X was used to train our model for both diabetes and high blood pressure, given that only the demographic information of the respondents is known. One of the challenges of this dataset was the large number of missing values. Figure I shows the missing values in the dataset, where the yellow lines represent the missing values in the features (name of the features on the X-axis). To address the missing values, we convert all the cells with blank and unknown values to
We also downloaded the second dataset, named Atlas Y, from the ATLAS database. It contains de-identified demographic and health-related information of 1372 people also mainly living in El Paso. It has 30 features which include age, income, race, Body Mass Index (BMI), systolic and diastolic blood pressure, blood test, mammography, etc. The percentage of missing values in this data was less than 2%.

A. Partitioning the Datasets

Before handling the missing values in both datasets, we split each of them into train and test sets. The training set is a set of examples used to fit the parameters, while the test set is used to provide an unbiased evaluation of how the final model(s) fit on the training dataset. We adopted this approach to prevent any form of data leakage that might arise as a result of imputation. We divided the data into training and test sets in the ratio of 80:20. Then, to achieve a near-optimal use of the available data, we partitioned the training set using the cross-validation technique. We repeated this ten different times while ensuring that the training set has the larger ratio in each iteration, until all data in the training subset have served as validation data once. For every iteration, we used different parameter combinations until we achieved the best accuracies for the models.

B. Handling the Missing Values in the Datasets

To handle the missing values in each of the features in Atlas X, we first employed the use of mean and mode imputation for the continuous and categorical features in them respectively [22]. Figures 2(A) and 2(B) show how the kernel density plots differ after imputation. The kernel density estimation (KDE) is a non-parametric way to estimate the probability density function (PDF), i.e., the probability distribution of the variables. We see a significant difference after the imputation with a larger concentration in the variable modes.

Hence, to get better results, we replaced the missing continuous and discrete values with linear regression iterative imputation and the categorical values with K nearest neighbor (KNN) imputation. Using this method proved to be more efficient, with minimal KDE difference after imputation. This same technique was also used to replace the missing values in the Atlas Y dataset.

C. Encoding the Categorical Features

We used one-hot encoding for the nominal features – race, employed, and marital status – to prevent the curse of dimensionality and optimize time complexity. This was
done by creating a new dummy feature for each unique value in them. Then, we used ordinal encoding for the ordinal categorical features, with age group, health status, income, and education, having cardinalities of 8, 7, 14, and 11, respectively. All the remaining features were encoded using integer encoding, by mapping each unique label in them to an integer. We assigned 0 for No and 1 for Yes for all binomial features, which constituted a larger part of the dataset.

D. Selecting the Best Features for Modeling

Feature selection was carried out on both datasets to reduce their dimensionalities. This process removed the redundant features and also helped to enhance the predictive performance of the models. We used LASSO (Least Absolute Shrinkage and Selection Operator) [19], which performs L1 regularization by adding penalty equivalent to the absolute value of the magnitude of coefficients. LASSO helped us minimize the prediction error by putting a constraint on the sum of the absolute values of the model features, where the sum has to be less than a fixed value or upper limit. The method applied a shrinking process and penalized the coefficients of the regression variables shrinking some of them to zero and retaining only those features with non-zero coefficients after its completion. The LASSO estimate is computed from the solution to the l1 optimization problem as

\[
\text{minimize} \left( \frac{||Y - X\beta||_2^2}{n} \right) \quad \text{subject to } ||\beta||_1 < t
\]

where t is the upper bound for the sum of the coefficients, Y is the target variable, X is the explanatory variable, and \( \beta \) is the regression coefficient. This optimization problem is equivalent to the parameter estimation

\[
\hat{\beta}(\lambda) = \arg\min_{\beta} \left( \frac{||Y - X\beta||_2^2}{n} + \lambda ||\beta||_1 \right)
\]

where \( ||Y - X\beta||_2^2 = \sum_{i=0}^{n} (Y_i - (X\beta))_i^2 \), and \( ||\beta||_1 = \sum_{j=1}^{k} |\beta_j| \), where k is the number of explanatory variables, and \( \lambda \geq 0 \) is used to determine the strength of the penalty. A high \( \lambda \) value results in a greater shrinkage.

E. Fixing Imbalance in the Datasets

In both Atlas X and Y datasets, we noticed imbalance with respect to the two target variables: diabetes and high blood pressure. An imbalanced data is one where the number of observations per class is not equally distributed, which results in bias in the algorithm’s classification. In Atlas X, as depicted in Figures 3(A) and 3(B), we noticed that for the target variables with 0 (No) and 1 (Yes) values, we have a significantly higher amount of No for both diseases. Hence to avoid the accuracy paradox, i.e., a situation where the model accuracy measures and reflects the underlying dominant class distribution, it was essential that we employ a method to remove the imbalance in the dataset. We used the Synthetic Minority Over-sampling Technique (SMOTE) technique [23] to add copies of instances from the under-represented class to obtain a balanced dataset.

F. Data Standardization

We then standardized all the features using the z-score, which re-scaled the features to be normally distributed with the mean value (\( \mu \)) of zero and standard deviation (\( \sigma \)). This step was applied to both datasets. The standard z-scores were calculated as

\[
z = \frac{x - \mu}{\sigma}
\]

G. Hyperparameter Tuning

Lastly, we carried out hyperparameter tuning to sample possible model architecture candidates from the space of possible hyperparameter values using a combination of grid and random search techniques.

V. THE PROPOSED NOVEL DEPRIVATION INDEX FOR PREDICTING CHRONIC DISEASES

As mentioned earlier, deprivation indices are being built and used to measure the level of disadvantage experienced by a population. In our proposed deprivation index – called Novel deprivation index – we included eight domains to measure physical and socioeconomic deprivation in a border town, such as El Paso, using knowledge from past literature.
TABLE I: Domains and Domain-Specific Indicators included in the Proposed Novel Deprivation Index

<table>
<thead>
<tr>
<th>Domains with corresponding indicators</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Households receiving food stamps</td>
<td>0</td>
<td>100</td>
<td>11.43</td>
<td>9.96</td>
</tr>
<tr>
<td>%Below poverty level in the past 12 months</td>
<td>0</td>
<td>100</td>
<td>14.45</td>
<td>11.10</td>
</tr>
<tr>
<td>Employment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Unemployment rate for population 16 years and over</td>
<td>0</td>
<td>51.9</td>
<td>5.03</td>
<td>4.78</td>
</tr>
<tr>
<td>%Employed with no health insurance coverage</td>
<td>0</td>
<td>100</td>
<td>14.67</td>
<td>9.68</td>
</tr>
<tr>
<td>Disability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Total civilian population with a disability</td>
<td>0</td>
<td>100</td>
<td>14.77</td>
<td>8.48</td>
</tr>
<tr>
<td>%Women who had a birth in the past 12 months</td>
<td>0</td>
<td>100</td>
<td>5.63</td>
<td>7.01</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%25 years old and over with less than 9th grade</td>
<td>0</td>
<td>100</td>
<td>8.19</td>
<td>8.68</td>
</tr>
<tr>
<td>%between 18–24 years not enrolled in school</td>
<td>0</td>
<td>100</td>
<td>54.11</td>
<td>26.03</td>
</tr>
<tr>
<td>Immigration/Language Barrier</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Limited English-speaking households</td>
<td>0</td>
<td>100</td>
<td>5.91</td>
<td>9.39</td>
</tr>
<tr>
<td>%Non-citizens foreign-born population</td>
<td>0</td>
<td>100</td>
<td>6.93</td>
<td>8.00</td>
</tr>
<tr>
<td>Housing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Households with more than one occupant per room</td>
<td>0</td>
<td>72</td>
<td>4.13</td>
<td>4.79</td>
</tr>
<tr>
<td>Material</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Households with no vehicle</td>
<td>0</td>
<td>100</td>
<td>4.80</td>
<td>5.91</td>
</tr>
<tr>
<td>%Households that are renter occupied</td>
<td>0</td>
<td>100</td>
<td>29.59</td>
<td>19.63</td>
</tr>
<tr>
<td>Technology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Households with a computer but no internet subscription</td>
<td>0</td>
<td>100</td>
<td>10.89</td>
<td>10.20</td>
</tr>
<tr>
<td>%Households with no computer</td>
<td>0</td>
<td>100</td>
<td>9.07</td>
<td>9.90</td>
</tr>
</tbody>
</table>

Min: Texas Minimum Value; Max: Texas Maximum Value; Mean: Texas Mean Value; Stdev: Texas Standard Deviation;

The proposed Novel index captures domains like immigration and technology that have scarcely been considered in previous studies. Table I lists those domains. Each one of the eight domains has two domain indicators. For example, we have two domain indicators for the domain Income: (1) %Households receiving food stamps, and (2) %Below poverty level in the past 12 months. We list the minimum, maximum, mean, and standard deviation for each domain indicator across all the 1975 ZCTAs studied. For example, for the domain indicator %Households receiving food stamps under the Income domain, a Min value of zero indicates the zip codes where no household received any food stamp. A Max value of 100 in the same domain indicator indicates zip codes where every household received a food stamp.

We ensured that the domain indicators were current, relevant, statistically robust, measurable, and quantifiable. The data for these indicators can be accessed by users at minimal or zero cost, relatively easy to update periodically, non-overlapping for each domain, and available at a small area level for most parts of Texas. Note that, even though we limited the coverage of this study to Texas, the same methodology can be extended to other geographical areas for which such data is available. Each of these indicators directly measures a major aspect of the dimension of deprivation, as stated in [20].

Researchers proposed several different deprivation indices. In constructing the proposed deprivation index, we adopted a variation of indices of multiple deprivations (IMD) similar to that of Cantalejo et al. [20], with adjustments in the way domain indicators were combined. The IMD methodology represents one of the most comprehensive and complete way of performing the operation. The data for all indicators were retrieved from the US Census Bureau files [10]. There were a total of 1975 ZCTAs across Texas in 2019 – the timeline for which we carried out the experiments. After carefully choosing the indicators for each domain, by ensuring they meet up with all the conditions earlier mentioned, a shrinkage estimation, i.e., Empirical Bayes estimation, was carried out on all indicators. This was done to improve the inference for each indicator so that fluctuations would not result in a situation where the most or least deprived ZCTAs would be those with very few observations.

For example, using percentage estimation, we could assume erroneously that a ZCTA with 250 households with no vehicle, out of 500 total households (50%), is more deprived than another ZCTA with 5000 households with no vehicle, out of a 20000 total number of households (25%). To address this, we performed shrinkage estimation, for each indicator in Table I, computing the mean ($\tilde{x}_i$) and variance $\tau$ across all ZCTAs in the state of Texas.

Next, for each domain, we calculate Bartlett’s Sphere Test and Kaiser-Meyer-Olkin (KMO) values [24]. Bartlett’s Sphere Test verifies the null hypothesis that affirms that the correlation matrix is not significantly different from the identity matrix, while the KMO value gives us the proportion of variance in the variables that are probably caused by some underlying factors. A KMO value close to 1.0 generally means that factor analysis is suitable for the data, and a KMO value less than 0.50 means it is not suitable [21], [20].

We used these scores to check the suitability of using
factor analysis to combine our variables and the results. In our case, the scores ranged between 0.001–2.000 and 0.49–0.50, respectively, showing that standardizing the indicators using the z-scores would give a better output. Indicators were then combined to form each domain score. We decided to give all the domains equal weight since studies showed that all of them are critical in assessing the deprivation in a city like El Paso. Lastly, the standardized domain scores were combined using equal domain weights to form an overall ZCTA Index of Multiple Deprivation (IMD).

Figure 4 shows the resultant Texas map showing the proposed deprivation index. The darker the areas in the map, the more deprived those regions are. Noticeably, parts of the US-Mexico border area have darker colors indicating that the population living in these border regions are deprived according to the domains considered in the Novel deprivation index.

VI. RESULTS AND DISCUSSION

In this section, we present the results of experiments that study the effects of the proposed Novel deprivation index in classifying a person at risk of diabetes and high blood pressure (HBP). The performance of the algorithms was measured using the following four metrics:

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
Precision = \frac{TP}{TP + FP}
\]

\[
Recall = \frac{TP}{TP + FN}
\]

\[
F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}
\]

where \(TP\) is the number of True Positives, \(FP\) is the number of False Positives, \(TN\) is the number of True Negatives, and \(FN\) is the number of False Negatives in the classifications.

In order to identify the likelihood of developing a chronic disease, we used several classification algorithms. The algorithms used for this study include: logistic regression (LR), decision tree (DT), Random Forest (RF), Gaussian Naïve Bayes (GNB), Multilayer Perceptron Neural Network (MLPNN), K Nearest Neighbor (KNN), and Support Vector Machine (SVM). These algorithms vary in the way they classify entities. While some use statistical methods (e.g., LR, GNB), others use a tree-based approach (e.g., DT, RF), or are motivated by the workings of neural networks (e.g., MLPNN). By including a variety of these algorithms, we intend to study if the algorithms have any impact on classification and, if they do, to what extent.

A. Using Novel Deprivation Index to Predict Chronic Diseases.

We seek to answer the following question in this section: What is the effect of the Novel deprivation index in predicting the occurrence of two chronic diseases: diabetes and high blood pressure? The experiments of this section used the Atlas X dataset. We carried out several experiments by training the machine learning models to make the predictions.

The best features selected by LASSO in these models for predicting diabetes were age calculated, city, language, other language, government assistance, uninsured healthcare, primary healthcare, HBP, family history, race (white), and health.

The best features selected for high blood pressure (HBP) were age calculated, city, ethnic background, government assistance (or not), form of insurance, uninsured healthcare, primary healthcare, ER visit, hospitalized, BP treatment, family blood pressure, race(white), marital status (married or civil union), income, and health status. We will discuss different experiments we conducted to validate the use of the proposed deprivation index in the following sections.

Table II presents the results of the experiments for both diseases. In this table, for each of the diseases, we list the four metrics: Accuracy, Precision, Recall, and F1 Score, for three variations – no index, Townsend index, and the proposed Novel index. The bold numbers indicate the best performance observed in any given metric. For example, for diabetes, SVM achieves the highest accuracy of 0.98 for all three variations of the deprivation index across all four metrics.

1) Data with No Index Included: In this experiment, we trained the machine learning models with the Atlas X dataset. Using randomized search over 1000 iterations, hyperparameter tuning was done for all the algorithms to optimize their predictive accuracies. The data did not contain any deprivation index, and only the selected features mentioned earlier were used to train the models. The first column in each performance metric shows the results of predictions with no index included. For diabetes, as shown in Table II, the best accuracy for this model was from SVM with an accuracy of 98% while the best accuracy for HBP was 88% from RF and SVM.

2) Data with Townsend Index Included: Here, the Townsend Index was added as a feature to Atlas X by matching the respondent’s zip code with the corresponding index. LASSO selected the index along with the earlier mentioned...
TABLE II: Experimental Results Comparing the proposed Novel Index with No Index and Townsend Index

<table>
<thead>
<tr>
<th>Disease</th>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No Index</td>
<td>Townsend</td>
<td>Novel</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>LR</td>
<td>0.74</td>
<td>0.74</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.54</td>
<td>0.65</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.58</td>
<td>0.59</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GNB</td>
<td>0.74</td>
<td>0.75</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MLPNN</td>
<td>0.76</td>
<td>0.77</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.82</td>
<td>0.82</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>HBP</td>
<td>LR</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.87</td>
<td>0.84</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GNB</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MLPNN</td>
<td>0.87</td>
<td>0.87</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.86</td>
<td>0.87</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td></td>
</tr>
</tbody>
</table>

ones as the best features for the models. Same procedures as in section VI-A1 were carried out to find the best combination of parameters for each of the classifiers from the available space. For diabetes, as shown in Table II, the best accuracy for this model was SVM with 98%, and the best for HBP was also 88% from RF and SVM.

3) Data with the Proposed Novel Index Included: Similar to section VI-A2, the proposed index was added as a feature to Atlas X. The best features for the model selected by LASSO were those mentioned earlier, along with the added Novel index. We also tuned hyperparameters for all the classifiers as before. For diabetes, the best accuracy for this model was SVM with 98%, and the best for HBP was 88% from DT, RF, MLPNN, and SVM.

In Table II, for diabetes, we see an increase in the predictive accuracies of all the models as a result of the added proposed Novel index to the data features. Compared to a commonly used index like Townsend, it is apparent that the Novel index performs better in enhancing the accuracy of predicting those who are susceptible to diabetes in El Paso. We obtained better accuracy in LR, DT, GNB, MLPNN, and KNN by adding the proposed Novel index than adding the Townsend index. The best F1 performance for diabetes prediction was by SVM with a score of 98%. We also observed an improvement in the F1 scores of DT, RF, GNB, and KNN for diabetes prediction as a result of the addition of the Novel index.

For HBP, the Novel index boosted the accuracy of DT from 87% to 88%, KNN from 86% to 87%, and resulted in a fairly stable performance for most of the other algorithms. The Novel index also performed fairly better than the Townsend index. We observed an increment in the F1 scores of DT, RF, GNB, MLPNN, KNN, and SVM as a result of adding the Novel index. Next, we plotted the ROC-AUC graphs for diabetes to help us separate the signal from the noise, i.e., we wanted to make sure that the models were not just randomly guessing.

In Figure 5, we observe that the models correctly predict a good number of those that are truly positive to diabetes. We also see an improvement in the performance of LR, SVM, KNN, DT, MLPNN, and GNB as a result of the Novel index.

Fig. 5: ROC-AUC Curves for Diabetes (top-left) With No Index, (top-right) With Townsend Index, (bottom) With the Novel Index

Notable is the increment in the AUC for DT from 0.39 to 0.61. We also measured the performance of the models for high blood pressure. Figure 6 shows that all the models with the proposed Novel index overall have better performance. Note that the proposed index performed better than the Townsend index for all the algorithms except SVM.

B. Using Demographic Data to Predict Chronic Disease

We seek to answer the following question in this section: What is the effect of using demographic data for predicting chronic diseases compared to health-related data? To answer this question, the machine learning models were trained with the Atlas Y dataset. Hyperparameters for all classifiers were tuned to get optimal results. The following section will discuss
TABLE III: Experimental Results Comparing Demographic Data to Health Data

<table>
<thead>
<tr>
<th>Disease</th>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>LR</td>
<td>0.77 0.81 0.78</td>
<td>0.77 0.81 0.78</td>
<td>0.77 0.80 0.78</td>
<td>0.77 0.80 0.78</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.83 0.91 0.88</td>
<td>0.83 0.91 0.88</td>
<td>0.83 0.91 0.89</td>
<td>0.83 0.91 0.86</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.85 0.85 0.89</td>
<td>0.87 0.89 0.90</td>
<td>0.85 0.85 0.89</td>
<td>0.85 0.84 0.89</td>
</tr>
<tr>
<td></td>
<td>GNB</td>
<td>0.74 0.61 0.71</td>
<td>0.75 0.70 0.75</td>
<td>0.74 0.61 0.71</td>
<td>0.74 0.56 0.70</td>
</tr>
<tr>
<td></td>
<td>MLPNN</td>
<td>0.88 0.82 0.88</td>
<td>0.88 0.82 0.89</td>
<td>0.88 0.82 0.88</td>
<td>0.88 0.82 0.88</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.95 0.96 0.94</td>
<td>0.95 0.97 0.95</td>
<td>0.95 0.96 0.94</td>
<td>0.95 0.96 0.94</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td><strong>0.99 0.96 0.96</strong></td>
<td><strong>0.99 0.96 0.96</strong></td>
<td><strong>0.99 0.96 0.96</strong></td>
<td><strong>0.99 0.96 0.96</strong></td>
</tr>
<tr>
<td>HBP</td>
<td>LR</td>
<td>0.83 0.80 0.83</td>
<td>0.83 0.80 0.83</td>
<td>0.83 0.80 0.83</td>
<td>0.83 0.80 0.83</td>
</tr>
<tr>
<td></td>
<td>DT</td>
<td>0.85 0.82 0.84</td>
<td>0.85 0.84 0.85</td>
<td>0.85 0.82 0.85</td>
<td>0.85 0.82 0.85</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.89 0.86 0.86</td>
<td>0.90 0.88 0.88</td>
<td>0.89 0.86 0.86</td>
<td>0.89 0.86 0.86</td>
</tr>
<tr>
<td></td>
<td>GNB</td>
<td>0.80 0.68 0.81</td>
<td>0.80 0.71 0.81</td>
<td>0.80 0.68 0.81</td>
<td>0.80 0.67 0.81</td>
</tr>
<tr>
<td></td>
<td>MLPNN</td>
<td>0.88 0.78 0.86</td>
<td>0.88 0.78 0.86</td>
<td>0.88 0.78 0.86</td>
<td>0.88 0.78 0.86</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.93 0.92 0.88</td>
<td>0.94 0.93 0.90</td>
<td>0.93 0.92 0.88</td>
<td>0.93 0.92 0.88</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td><strong>0.96 0.95 0.92</strong></td>
<td><strong>0.96 0.96 0.92</strong></td>
<td><strong>0.96 0.95 0.91</strong></td>
<td><strong>0.96 0.95 0.91</strong></td>
</tr>
</tbody>
</table>

Fig. 6: ROC-AUC Curves for High Blood Pressure (top-left) With No Index, (top-right) With Townsend Index, (bottom) With the Novel Index

the results of experiments, as shown in Table III, that address the aforementioned question.

1) Using Health plus Demographic Data (H+D): Firstly, we used the entire Atlas Y data, i.e., containing both health and demographic features, to train the model. The best features selected by LASSO in this model for predicting diabetes were age, household size, systolic BP, dysstolic BP, BMI, cholesterol, glucose, pulse, CE4, CO4, address, city, state, medical check-up, emergency room, admitted to hospital, treatment for diabetes, mammography, colonoscopy, immediate family diagnosed, tobacco, flu vaccine, HBP, gender (male), race (Hispanic/Latino), marital status (divorced), language (Spanish), and education.

For HBP, the best features selected were age, systolic BP, dysstolic BP, BMI, cholesterol, glucose, weight, BR4, CE4, CO4, state, medical check-up, emergency room, admitted to hospital, treatment for diabetes, treatment for hypertension, mammography, colonoscopy, immediate family diagnosed, which vaccines, diabetes, marital status (married/civil union), marital status (single), language (Spanish), government assistance, income, and education. Table III the first column of each performance metric, shows the results of experiments when using both health and demographic data (i.e., H+D). We observe that the best algorithm for predicting diabetes and HBP was SVM, with an accuracy of 99% for diabetes and 96% for HBP.

2) Using Health Data: Then, we separated Atlas Y into two parts; demographic data and health-related data. Next, we used only the health-related features to train the model. The best features selected by LASSO for predicting diabetes were systolic BP, dysstolic BP, BMI, cholesterol, glucose, weight, pulse, PMC, BR4, CE4, CO4, dental check-up, emergency room, admitted to hospital, treatment for diabetes, treatment for hypertension, mammography, colonoscopy, immediate family diagnosed, flu vaccine, which vaccines, and HBP.

The best features selected for HBP were systolic BP, dysstolic BP, BMI, cholesterol, glucose, weight, BR4, CE4, CO4, medical check-up, emergency room, admitted to hospital, treatment for diabetes, treatment for hypertension, mammography, fecal occult blood test, colonoscopy, immediate family diagnosed, tobacco, flu vaccine, which vaccines, and diabetes. The best accuracy for predicting diabetes was 96% from both KNN and SVM, while the best accuracy for HBP (i.e., hypertension) was 95% from SVM.
3) Using Demographic Data: Lastly, we used only the demographic data from Atlas Y to train our model. This was done to answer the second question. The best features selected by LASSO for predicting diabetes were age, address, city, state, zip code, household size, gender (male), race (Hispanic/Latino), employment status (homemaker), marital status (divorced), marital status (separated), marital status (single), language (Spanish), government assistance, and education. For HBP, the best features selected were age, address, city, zip code, employment status (part-time), marital status (married/civil union), marital status (single), language (Spanish), government assistance, income, and education. The best classifier for both diabetes and HBP using demographic data only, was SVM with an accuracy of 96% and 92%, respectively.

In summary, we tested the algorithms on the entire dataset (H+D), health-related data only, and demographic data only. Table III shows that overall, demographic data are as effective as health data in predicting the two chronic diseases, especially high blood pressure. For example, we achieved an accuracy of 84% for the decision tree (i.e., DT) algorithm using demographic data, compared to 82% using health data. We also achieved the same best accuracy of 96% from SVM for diabetes using only health data and only demographic data.

Aside from SVM, we achieved considerable accuracies for RF and MLPNN for diabetes prediction by using only demographic data. Accuracies of LR, DT, KNN, and SVM from using demographic data were also close to the ones of using health data only for diabetes prediction. For HBP prediction, LR and GNB achieved their best accuracies by using demographic data only. We also got close accuracies from DT, RF, and SVM by using either health or demographic data only for HBP prediction. In addition, the best F1 scores from LR, DT, and GNB were by using only demographic data for HBP prediction.

We also compared the performance of using the demographic versus the health data. In Figure 7, we see that RF and GNB perform optimally using the demographic data alone. We also see that for all the other models, demographic data still help in achieving good predictions for diabetes. Lastly, in Figure 8, we observe that LR, MLPNN, and GNB achieved higher AUC using demographic data than using health data only.

Our experiments captured all the questions raised earlier in section III-A. We observed that the accuracies and performances were higher for some of the algorithms when the deprivation index was included as a feature in the data. It is also noteworthy that demographic data alone can help to predict chronic diseases in a border town like El Paso, when health data are not readily available.

VII. CONCLUSION

The aim of this work was to study the effects of using demographic data for the prediction of two chronic diseases: diabetes and high blood pressure, instead of relying solely on the health-related data (e.g., blood pressure, heart rate). We also want to understand how a deprivation index can enhance the result of such predictions in border cities, using El Paso, Texas, as a case study. Our approach builds on the knowledge
from previous studies and extends the Townsend index to include demographic data to optimize the performance of some machine learning algorithms. Experimental results show that we do not have to collect health-related data to accurately predict patients that are susceptible to the diseases. Patients' demographic data are sufficient for a good prediction. We compared our approach using two datasets and different types of machine learning algorithms. Results indicate demographic data is an equally strong indicator for predicting chronic diseases at a border city. Such data performs equally well as health data for predicting chronic diseases such as diabetes and high blood pressure.

Further work that can be done to improve the framework introduced in this study includes using nationwide census tract data, which would produce a finer level of deprivation measurement. In the future, we plan to test the proposed index across other border towns.

REFERENCES


