MACHINE LEARNING METHODS FOR PREDICTION OF DIABETES PREVALENCE AMONG THE WORLD'S COUNTRIES USING AGGREGATE NUTRITIONAL AND MEDICAL DATA

A THESIS

Presented to the Department of Mathematics and Statistics

California State University, Long Beach

In Partial Fulfillment

of the Requirements for the Degree

Masters of Science in Applied Statistics

Committee Members:

Hojin Moon Ph.D (Chair)

Yong Hee Kim-Park Ph.D

Xiyue Liao Ph.D

By Chioma A. Nwuzi

B.A. 2016 University of California, Santa Barbara

August 2021

ABSTRACT

Type 2 Diabetes Mellitus is known as a chronic disease characterized by comorbidities such as high blood sugar, high blood pressure and obesity. In reality, the significance of these factors varies from country to country based on the access to medical exams, access to nutrient dense foods, and clean living conditions. The goal of this research is to illuminate the effects that certain nutritional outcomes and health measures have on the prevalence of type 2 diabetes with regard to all regions in the world. In order to understand whether the characteristics that are commonly associated with the disease hold a significant relationship with the extent to which the population has diabetes across various countries and regions. This research will show that there are nutritional characteristics, government policies, and health trends that represent a strong relationship to Type 2 Diabetes Mellitus. Backwards regression, stepwise regression, random forests, gradient boosting methods, ridge and lasso regression are employed as variable selection and model selection devices along with model averaging to determine which factors hold across all the countries as significant predictors of diabetes prevalence.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

List of Tables

List of Figures

INTRODUCTION

Type 2 Diabetes Mellitus is the 6th most deadly disease worldwide according to the World Health Organization. There are numerous factors within a country's social and economic individualities that lead to nutritional and health differences between countries. Because of these very individualised characteristics, it is difficult to have a high degree of specificity on the most indicative factors for predicting Type 2 Diabetes Mellitus that apply universally across the world's countries. It is imperative to be able to identify those factors that can be universally known as the highest indicators of a country's prevalence of Type 2 Diabetes Mellitus to ensure the applicability of policy and advisories regarding the mitigation of the disease is as widespread as possible.

It is believed that there are numerous nutrition based indicators of the likelihood of a person contracting Type 2 Diabetes Mellitus in their lifetime. Those factors are typically country specific and have high variables from country to country. The data set is taken from the global nutrition report which is a yearly aggregate report that aims to reduce the amount of malnutrition and nutrition based diseases worldwide. There are 178 observations that represent 178 countries and the outcome is diabetes prevalence in those countries. Type 2 Diabetes Mellitus is typically thought of as a preventable disease if nutritional interventions and policy to maintain high value nutrition and healthcare are implemented within a country.

When dealing with a dataset that contains a higher number of dependent variables compared to the observations such as the data in this study problems around model fitting, interpretability, and dimensionality need to be addressed. In order to address the problem of interpretability and high dimensionality when it comes to the applicability of results, supervised methods such as a backward and stepwise regression models, ensemble methods, and penalized

regression models are applied to the data set containing the aggregate nutritional, health, and government policy data of the countries included in the World Nutrition Report in order to predict the prevalence of Type 2 Diabetes Mellitus and to identify important health or nutritional parameters surrounding this disease, and choose a specific model that performed best when handled the nature of a dataset with a dependent variable and observation balance such as the dataset used in this study. The unsupervised methods are used as a way to narrow the number of nutritional factors that have the highest contributing power and assign weight to these variables to allow for higher visibility and specificity in understanding the way different dependent variables of the report factors affect worldwide diabetes prevalence. These methods in conjunction will allow us to develop a baseline knowledge of how the observations vary amongst each other and how the importance of factors differ among the observations. Having this baseline will add to the predictive interpretation of the significance factors that lead to the prevalence of diabetes among the world's countries.

The ultimate goal of this research is to illuminate the effects that certain nutritional, health, and regulatory outcomes have on the prevalence of Type 2 Diabetes Mellitus across the world. It is also to better understand whether the significance of certain variables such as obesity, childhood growth, population etc. to Type 2 Diabetes Mellitus prevalence that have been observed in regionally specific studies holds worldwide.

TYPE 2 DIABETES MELLITUS BACKGROUND

Type 2 Diabetes Mellitus is a chronic disease characterized by the body's inability to utilize and regulate blood sugar. When someone has this disease, the pancreas does not produce the amount of insulin that is needed for the body. Insulin is the hormone that regulates blood sugar in the body. Along with the issue of insulin production in the body is the issue of the cells not being able to respond to the insulin that the body is able to produce. This specific condition is referred to as insulin resistance. Type 2 Diabetes Mellitus is a disease that is most commonly associated with adults as it was previously referred to as “adult onset diabetes”. But post 2012, the United States is seeing a 5% increase in the rate of children between the ages of 10 and 19 with Type 2 Diabetes Mellitus yearly compared to a 2% increase yearly between the years 2002 and 2012. The IDF Diabetes Atlas Ninth edition reports that there are approximately 463 million people aged 20-79 years who are currently living with diabetes around the world. It is estimated that by 2045 this number will be 700 million.

The rhetoric around the risk factors of diabetes are usually centered around body size, weight, physical activity, and nutrition. Although the risk factors to diabetes are typically focused on health indicators such as weight and physical activity, there are nutritional and other health indicators that are also thought to be risk. The literature typically focuses on such factors and has little emphasis on government policy or the regional differences such as population density or healthcare practices and how these factors could also influence Type 2 Diabetes Mellitus. Factors such as high blood sugar as well as high blood pressure are both factors that are associated with Type 2 Diabetes Mellitus This study seeks to illuminate the factors that act as risk factors to the increase in diabetes prevalence within a country’s population.

LITERATURE REVIEW

The study *Predictive models for diabetes mellitus using machine learning techniques* (Lai, Huang, Keshavjee, Guergachi, Gao 2019) uses patient data from adult Candian patients. The data used in this study was at the patient level and not at the aggregate country level like the data employed in this study. The machine learning methods of this study were a Logistic regression model as well as a Gradient Boosting model in order to determine the probability that a patient, based on their medical data, will develop Type 2 Diabetes Mellitus. In the case of this study, the dependent variable was whether or not the patient had Type 2 Diabetes Mellitus, or did not have Diabetes Mellitus. The method used to understand the predictability of the model was the ARO- area under the ROC curve (receiver operating characteristic curve), sensitivity, and specificity. Sensitivity is the measure of true positives over true positives and false negatives and specificity is the measure of the true negatives over the total of true negatives and false positives.These are all methods of understanding how well a classification model is able to perform in correctly classifying the data. The study found that the Gradient Boosting Model did slightly better than the Logistic Regression Model with an AROC of 84.7% for the GBM model compared to an ARCO of 84% with the Logistic Regression model. The specificity and sensitivity was 71.6% for the GBM model respectively, and 73.4% and 82.3% respectively for the Logistic Regression model. The resulting variables that had the highest significance in the were fasting blood glucose, high density lipoprotein, body mass index, triglycerides, and age, with fasting blood glucose being the most important variable.

The research question of this study centers around the globally holistic factors and how these factors influence diabetes prevalence. Research into diabetes prevalence around the world is typically centered around a single region or a single country. The Research Article *Prevalence*

of Risk Factors for Type 2 Diabetes Mellitus Mellitus in Vietnam: A Systematic Review (Nguyen, 2017) is a research study examining Type 2 Diabetes Mellitus along with the variables that have the highest contributing risk factors to the disease. This study focusing on Vietnam found that the most significant factors in diabetes Type 2 are old age, urban residence, body fat, lifestyle activity, and hypertension. This study did not employ machine learning or statistical methods to come to the conclusions but instead reviewed 10 studies and synthesized the results from the studies. This methodology differs from the methodology in this study not only in the granularity of the data as this study reviewed data that was patient-specific, as opposed to the data being used in this story which uses aggregate data on the country level.

A similar study titled *High prevalence of Diabetes in An Urban Population in South India* (A Ramachandran, 1988) is similar to the previous study as it highlights the risk factors that were identified in South India based on a survey conducted in a township in India. In this study, one of the prominent findings was that the prevalence of diabetes within this town in India was higher among those within the study whose income was greater than the mean. This corroborates the findings that were identified in the exploratory data analysis for the study in this paper that showed that there was an increasing relationship between diabetes prevalence and a country's recorded GDP. This also adds evidence to how diabetes differs from other preventable diseases as the risk factors on a global scale are not as associated with those living below the median income levels as other diseases such as lung cancer and heart disease.

In regards to studies centered around diabetes mellitus that utilized machine learning, *Early Detection of Type 2 Diabetes Mellitus Using machine Learning-Based Prediction Models* (Kopitar 2020) compares machine learning prediction models that are commonly used for diabetes prediction to see which model is best for predicting undiagnosed Type 2 Diabetes

Mellitus. This study found that using the boosting algorithm XGBoost available through the Python programming language performed the best and produced the lowest root mean squared error when compared to other regression methods. Boosting algorithms are a very common method employed in machine learning in the medical field. The results of this study showed that observed hyperglycemia was the most significant variable in 75% of the models tested, along with age and cholesterol. Although it is not outwardly stated in the paper, the guidelines for patient usage of data that was upheld for this paper came from the University of Maribor which is located in Slovenia. Assuming that all of the patients in the study are of Slovenian nationality, and as some of these variables are present in the dataset of the study of this current paper, it will be interesting to see if the same variables are of high significance. It will also lend an interesting note on how not only observations affect the variables that can be concluded to be the most important, but also how different in methodology impacts the variables that are considered of the highest predictive power.

DATA DESCRIPTION

The data used in this study is a combination of nutritional, health, and policy data for 178 countries and the diabetes prevalence from those countries. The nutritional data comes from the 2015 Global Nutrition Report and is a dataset that contains aggregate nutritional and health markers for most countries in the world. The report also contains variables around government and social policies that could affect health and nutritional outcomes .The Global Nutrition Report Data is made up by combining data from UNICEF and from The World Health Organization. This report includes data around specific nutritional deficiency such as iodine and salt consumption, to categorical variables around policy on hypertension and diabetes, blood glucose averages across the country, etc. The Global Nutritional index also includes data around the state of the country as a whole such as poverty indexes, stunting indexes, country GDP, population densities, etc.

The goal of the global nutrition report is to combat malnutrition. The report was created following the Nutrition for Growth Initiative Summit in 2013. The Global Nutrition Report is one of the leading assessments of a country's nutrition profile as well as global nutrition. Although the report and the aggregate data was created and compiled with the goal in mind to rid the world of malnutrition (Global Nutrition Report 2015), the data is a robust source of the state of a country's health profile.The data on the diabetes prevalence of each of the countries comes from the International Diabetes Federation and is the prevalence measured as the percentage of the population with diabetes. Independent of the data from the International Diabetes Federation, there are 235 variables in the data coming from the 2015 Global Nutrition Report.

NUMERIC VARIABLES

The dependent variable is the Type 2 Diabetes Mellitus prevalence in 178 countries/independent states that are a part of the 2015 Global Nutrition Report. In the original dataset from the 2015 Global Nutrition Report there are 201 numeric variables and 34 categorical variables. The nature of the Global Nutrition Report is aggregate as there are variables in the dataset that come from prior years of the Global Nutrition Report. This will evidently lead to high levels of multicollinearity of the variables. The numeric variables include population variables, which include the countries total population, the over 65 population, and the under 5 population, mortality rates of those under 5 years old and more, as well as numeric variables around poverty rates.. There are 15 variables regarding the poverty rates within the country based on two different indexes: the below $1.25 poverty rate and the below $2 a day poverty rate. The dataset also contains accompanying years for these variables. Since the collection of these stats are not uniform across the countries, this limits how the time based variables can be used and eliminates a longitudinal study as a point of analysis in the data. There are variables on the GDP of the country at different points of time, as well as variables around the rate of stunting and the rate of wasting in the country. Stunting is the World Health Organization's categorization of a child whose height is lower than the average for their age and two standard deviations or more away from the WHO's Child Growth Standards Median (Concern 2019). Wasting is defined as a loss of body weight in relation to height, mainly used in response to growing children (World Health Organization 2014). The dataset contains data on the current and the required annual rate of reduction for the rate of overweight children under the age of 5. There are variables that are centered around breastfeeding in the Global Nutrition Report including variables on the percentage of the population that is exclusively breastfeeding, the percentage of the population

that is early breastfeeding, and the availability of breastfeeding alternatives. There are numerous variables around the state of the populations of a country's levels with deficiencies with essential nutrients such as iodine, and vitamin C. As well as more common health indicators for Type 2 Diabetes Mellitus such as the percentage of the population with elevated Blood Pressure, and the percentage of the population with elevated Blood Glucose. These variables are split by gender and then also contain a variable in the dataset that aggregates the variables by gender. Most likely leading to multicollinearity within these categories of variables. Other variables include food supplementation and fortification, availability of produce, population density of health works, water and living space sanitation, water coverage, female education rates, and the percentage of government spending that goes into the different social sectors of the country.

CATEGORICAL VARIABLES

The majority of the categorical variables in the dataset are ordered meaning that there are levels to the variables that represent an increasing or decreasing hierarchy. The categorical variables that are not ordered in this dataset are the variables describing the area where the country resides. This includes the continent, region, and subregion. To keep the definitions of these variables standardized, the definitions for each of these are formatted to match the continent, region, and subregions as defined by the U.N.

The rest of the variables in the dataset are ordinal variables. An ordered categorical variable is a variable that has a clear and natural ordering to the variables. The ordinal variables in the dataset are variables that describe the stage the country is in when it comes to implementing policies to remain on progress with global health and nutrition goals. These variables include the stunting progress, wasting progress, and other variables whose levels range

around how close the country is to being on target with the WHO target in various health indicators such as diabetes policy and hypertension policy.

EXPLORATORY DATA ANALYSIS

Studies into the risk factors of Type 2 Diabetes Mellitus are usually done within certain regional constraints. These studies are either done within country boundaries or within a slightly increased or decreased regional boundary. Because of the specificity in region of these studies, the question on whether the results of these studies can be applied to regions beyond that of these studies arises . One of the questions when going into the exploratory data analysis for this study is to understand if there is a significant difference between the average diabetes prevalence of the continent distribution. Figure 1 shows the distribution of the average diabetes prevalence of the major regions.

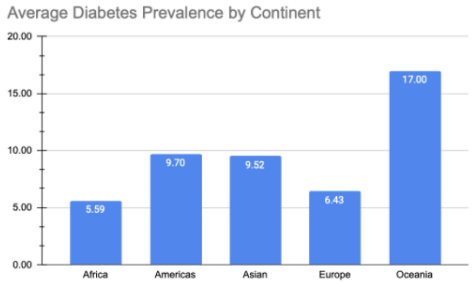

Figure 1. A bar graph of the average diabetes prevalence by the 5 inhabited continental regions as defined by the United Nations.

In order to understand whether the differences between the average diabetes prevalence of the continents is significant, the Wilcoxon Signed Rank Test is used. The Wilcoxon signed rank test is a nonparametric test of data. The null hypothesis of the Wilcoxon Signed Rank Test in the case of the diabetes prevalence of the continental regions is that the mean of diabetes prevalence for one of the regions is significantly different from that of another continental region. This test is run by calculated each of the paired differences for the different regions i.e $d_i = x_i - y_i$ where x_i and y_i are two different pairs of observations. The absolute value of the result of the difference d_i is then order ranked. Each rank of d_i is then labelled with a sign based

on the original sign of d_i.The test statistics $W$is calculated based on the sign of the ranked d_i. W^+ is the sum of the positive ranked $d_i$values and $W^-$is the sum of the negative signed d_i values. The test statistic$W$is then calculates as the minimum of W^- and W^+.

$W = \sum_{j=1}^{P} [sign(x_{2,j} - x_{1,j}) * R_j$ where $W$is the test statistic, sign represents the sign function, P is the sample size of non equal pairs x_1 and x_2, $x_{1,j}$, $x_{2,j}$ are the corresponded ranked couplets, and R_i is the rank. This test statistic is then checked against the critical values of the two sided Wilcoxon Signed Rank sum test and the p-value is evaluated for significance.

Continental Region 1	Continental Region 2	Wilcoxon Test P-Value
Asia	Europe	.00007739
Asia	Africa	.0000003195
Asia	Americas	.4806
Asia	Oceania	.001244
Africa	Europe	.007552
Americas	Europe	.0000006744
Oceania	Europe	.000007223
Africa	Americas	.00000007613
Africa	Oceania	.000002349
Americas	Oceania	.001117

Table 2. Table containing the continental regional pairs that were tested using the Wilcoxon Rank test and the corresponding p-values.

The results of the Wilcoxon signed rank test show that all regions have a significant difference between the means of their diabetes prevalence with the exception of Asia and the Americas.

The Chi-Squared test between two variables and the Fisher Test were also used as a means of understanding the categorical variables in the dataset. Diabetes policy and hypertension policy both represent the availability and stage of implementation of guidelines, protocols, and standards for the management of hypertension as well as the management of diabetes. These are both variables that involve the influence of the government or social entities in ensuring that the population has the resources to mitigate the rise of hypertension and diabetes. The contingency table below shows the count of observations by hypertension policies and diabetes policies. In order to be able to assess the contingency of these two ordered variables, either the Chi-Squared test or the Fisher's test can be employed. The Chi-Squared test and the Fisher test are both tests used to understand the association between categorical variables. The Fisher test is used in place of the Chi Square test most often for contingency tables of 2x2 or when there are uneven values in the cells of the contingency table.

The Fisher Exact test is done by calculating the row and column sums of the contingency table. Let $X\ and\ Y$represent two categorical variables within the data and let i and j represent the number of observed states for both variables. $i\ \times\ j$represents the matrix of observations $p_{m\,n}$ represents the the count of observations where $m\ =\ x$ and $n\ =\ y$. The

$Total\ Sum\ of\ matrix\ =\ \sum_{x} R_{x} = \sum_{y} C_{y}$ The conditional probability of producing the $i\ \times\ j$ matric given any particular row and column total is defined as

$$\frac{(R_1!R_2!R_3!\ldots R_i!)(C_1!C_2!C_3!\ldots C_j!)}{Total\ Sum\ of\ Matrix!\prod_{m,n} p_{mn}!}$$

This is a multivariate form of the hypergeometric distribution and it is used to find all possible combinations of the matrix for all possible row and column sums. This is used to calculate the conditional probability in which the sum of the probabilities is 1.

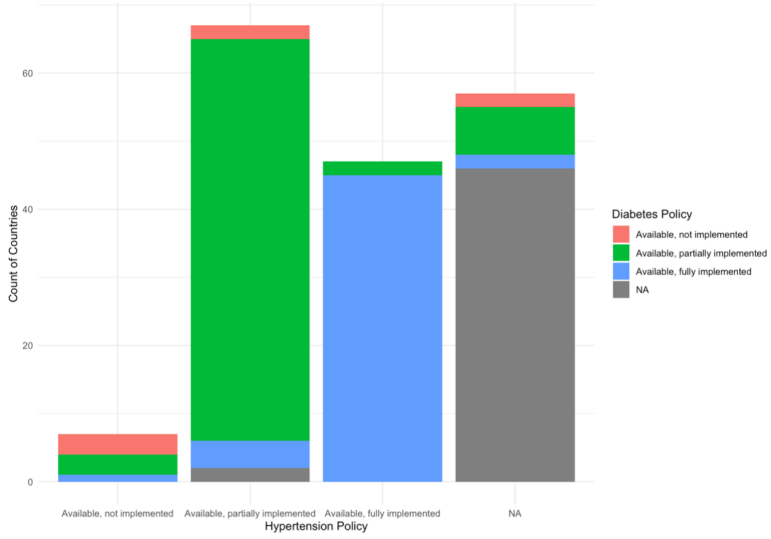

Figure 2. The bar graph shows the count of countries that fall into the three tiers of diabetes policy and hypertension policy. The "NA" signifies missing values for the variable.

		Diabetes Policy		
		Available, Not Implemented	Available, Partial Implemented	Available, Fully Implemented
Hypertension Policy	Available, Not Implemented	3	3	1
	Available, Partially Implemented	2	59	4
	Available, Full Implemented	0	2	45

Table 2. Contingency table showing the number of countries that are within each category for hypertension and diabetes policy implementation.

The Fisher test is more utilised than the Chi-Squared test when the sample size is relatively small. Table 2 and Figure 2 show that there is stark unevenness in the cells of the contingency tables. For the following variables, hypertension policy and diabetes policy, the Fisher Test is used. The result of the Fisher's Exact Test on the variables for hypertension policy and diabetes policy is significant at the 5% significance levels (with p-value less than 2.2e -16). This shows that there does exist a statistically significant relationship between the two variables.

Another test used to understand the significance of categorical variables is the Chi-Squared Test. Table 3 and Figure 3 show the contingency table between the two categorical variables stunting progress and wasting progress. These are two variables associated with the health of the population by how on course the country's population is with having the number of children who are below the WHO level of development as well and how of course the countries population is on par with the people who reside in the population being within an average weight range of their height. These are two variables that intuitively have relation, but because one variable focuses on children and the other is for the total population, the relation between the two variables will be checked using the Chi -Squared Test.

The Chi square test uses a contingency table the same way that the Fisher Test uses. The null hypothesis for the test is that stunting progress and wasting progress are two independent variables while the alternative hypothesis is that these two variables are dependent. This test is conducted by calculated the Chi Test Statistic $\chi^2 = \sum(A - E)^2/E$ where the test statistic is equal to the summation of the actual observed frequency within the contingency table and the expected frequency of the contingency under the null hypothesis that the two variables are

independent. The expected frequency is calculated by $E = R + C / n$ where R is the row total, C is the column total, and n is the sample size.

	Wasting Progress		
Stunting Progress		Off Course	On Course
	Off Course- No Progress	9	9
	Off Course- Some Progress	31	23
	On Course- Good Progress	12	21

Table 3. Contingency Table counting the number of countries that fall into each category of stunting progress and wasting progress.

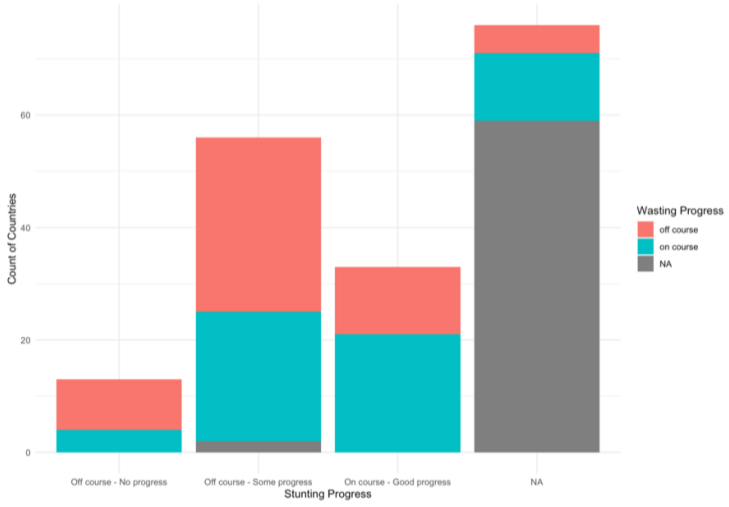

Figure 3. Stacked bar graph displaying the number of countries that fit into each of the categories for wasting progress and stunting progress.

The result of the Chi Squared test of Independent for stunting progress shows that the association is not significant. The p-value of the test is .06682 which is not significant at the 5% significance level. By this result we can conclude that the two categorical variables are independent. Additional tests of the relationship between the categorical variables are shown in the appendix.

MISSING VALUE IMPUTATION

Due to the lack of data collection resources in countries in the Global South or Global East, the threshold for removing observations based on the number of missing variables is too liberal to allow for these data collection discrepancies. The threshold for removing observations or removing variables based on missing data is typically between 25-30%. Although this threshold is an acceptable threshold to maintain data integrity for most machine learning algorithms, it disproportionately excludes observations where the missing values are due to differences in collection methods that are not available to certain demographics, and in the case of this study, disproportionately disadvantages certain countries. Of the 33 countries that represent the highest number of missing variables, only 6 of the countries are in the Global West, including Canada, France, Belgium, and Ireland. The remainder of the countries with the highest number of missing variables are all from the Global South and the Global East. Therefore utilizing the industry standard of removing observations that experience more than 25% of missing variables would disproportionately bias this study towards countries that are in the West, thus skewing the results of the study away from applicability to countries in Africa, Asia, and Eastern Europe.

One of the goals of this study was to not penalize countries because of the data collection limitations of the nation. For that reason, instead of the traditional 25% removal threshold for a variable or an observation, a threshold of 50% was used. The original number of variables in the dataset, excluding the identifying variable of country and the dependent variable of diabetes

prevalence is 233. Of the full dataset there were only 14 variables that had no missing values and about 80 variables with less than 10% missing values. To make a mathematically founded threshold for the number of missing variables, the number of variables that would be eliminated at 10% intervals above 30% were all evaluated. The goal is to find a percentage that does not penalize variables that may have high predictability with the dependent variable diabetes prevalence, in order to reduce data loss, but to also make sure the goal of feature reduction is still being met. Using the 30% threshold, 111 variables (47% of the total variables) would be eliminated. Using this threshold would result in the removal of all of the variables surrounding poverty rates, which exploratory data analysis showed had a negative relationship with diabetes prevalence rates. This is also a metric that is higher for countries in the global south and the global east when compared to the West, therefore this will result in a negative impact on the observations from those areas if this variable proved to be a significant predictive variable for diabetes prevalence. Variables around weight and hypertension were also among the 30% missing group, therefore the conclusion is that this threshold is too conservative for the study. At the 40% threshold, 74 variables (31% of the total variables) would be eliminated. This threshold still includes all of the poverty metrics so it is still a bit too conservative for this study. At the 50% threshold, 49 variables (20% of total variables) would be removed; this threshold does not lead to the elimination of the poverty indexes, the population indexes, and other variables that represent important regional health markers. After evaluating the 60% threshold, it was found to be too lenient on the number of variables that would be best to be removed during this first attempt at feature reductions, therefore the 50% threshold is the threshold for variable removal used in this study.

After removing variables based on the 50% threshold, 137 variables were remaining. Reducing the number of variables is critical before continuing to apply machine learning algorithms since there are algorithms that produce models that are prone to overfitting if the number of observations is less than the number of variables. Some of those methods are applied later in the study.

Along with the threshold for missing variables, there is additional covariate removal that is done to the dataset before obtaining the final dataset that will need to be imputed. These variables do not contain any additional value (as they are redundant across all observations) like the source of the data or the year the data was obtained. After the removal of the covariates in this group or that presented missing variables that were above the threshold, the rest of the missing variables still needed to be imputed before any further analysis or modeling could be performed on the dataset. Imputation is the method of replacing a value that is either missing from the data set or unusable for any reason. Within the dataset, there are two categories of missing variables that need to be imputed. There are 115 numeric variables and 16 factor variables. The factor variables include variables that have 2 or more levels. Because of this mixed nature of the dataset two separate methods for imputation are necessary. One method for numerical imputation, and another method for categorical imputation.For the reproducibility of the analysis, the method chosen for imputation of both groups is the method that takes the least amount of computation time.

NUMERICAL IMPUTATION

The method used for imputation of the numerical variables was the method of predictive mean matching (PMM). This method uses the observed value of an existing observation with a

similar predicted value as the observation with a missing variable and imputes that observed value for the imputation. PMM is a method in which the missing data point is taken as the dependent variable of the dataset and the other variables are used as the independent variables in a random forest model (Wilson 2020). The predicted values of the missing variable are then calculated.The one observed variable from the five predicted values that are closest to the prediction for the missing value is then imputed for the missing variable for that observation.

Let Ybe the variable that is missing for a given observation i. For a set of non-missing variables$n$, fit a linear regression of the form

$Y_i = B_0 + B_1X_{i1} + B_2X_{i2} + \ldots + B_{n-1}X_{i,n-1} + \epsilon_i$where $B_i$represents the coefficients of all of the non missing variables X_n and $B_o$is the intercept of the model. When Y_i has been calculated for all of the observations (regardless of if the observation was missing the variable i). One value is then randomly selected from the five observed values that are closest to the predicted value of Y_i for the observations.

This imputation method is demonstrated for the country of Andorra and for the variable that describes the rate of female education in the country (rate_female_ED). Figure 4 shows how this value would be imputed using the predictive mean matching method. A random forest prediction model using the remaining variables in the dataset as the independent variables, and the rate of female education as the dependent variable for the model is used. The predictions for all the observations, whether or not they were missing for this variable, are then fit to the model. A value is then randomly drawn from five observed values in the dataset that are closests to the predicted values of the observation that originally had a missing. In this case, the predicted value for the rate of female education for Andora is 86.675. The five countries that contain observed values that are close to this predicted value are the United Arab Emirates, Algeria, Cuba,

Bulgaria, and Peru with the values 87.638, 71.467, 88.567, 89.052, 90.189 respectively. The value that is then chosen to be imputed for Adorra is randomly selected from this pool. Therefore Bulgaria's observed value of 89.05255 is the imputation value for the rate of female education.

	Predicted Value	Actual Value
United Arab Emirates	85.69376	87.63841
Algeria	85.92144	71.46703
Andorra	86.67590	*NA*
Cuba	87.77414	88.56739
Bulgaria	89.05202	89.05255
Peru	89.24595	90.18964

	Predicted Value	Actual Value
United Arab Emirates	85.69376	87.63841
Algeria	85.92144	71.46703
Andorra	86.67590	*89.05255*
Cuba	87.77414	88.56739
Bulgaria	89.05202	89.05255
Peru	89.24595	90.18964

Figure 4. The first chart shows the 5 random forest predicted values for the missing covariate for the rate of female education and the actual values from the dataset who's predicted values is closest to the predicted value Andorra. The second chart shows that the observed value of Bulgaria was used to impute for the missing value of Andorra.

CATEGORICAL IMPUTATION

The method of imputation used for the categorical variables is a Classification and Regression Tree (CART) method of imputation. CART algorithms are tree-based algorithms in which the model creates splits in the dependent variable at points in which the sample data may be split into more homogeneous subsamples. This process is repeated in these subsamples to create a tree structure. This method of imputation is particularly good at handling outliers as well as good at handling cases of multicollinearity.

Decision trees are a nonparametric supervised learning method that can be used for the purpose of classification as well as the purpose of regression, hence the name CART. For the purposes of classification, the CART decision trees use a Gini Impurity when determining the optimal splits of the decision tree. The Gini Impurity is a measure of the probability that a randomly chosen variable is misclassified. In the event of Mtotal classes where the probability of randomly selecting a datapoint that is from class $n$is $p(n)$ then the Gini Index is given by

$G = \sum_{n=1}^{M} p(n) \times (1 - p(n))$ and the Gini Impurity is given as $Gini\ Impurity = 1 - G$.

The steps towards the imputation of a missing variable using this method is to fit the regression tree by splitting the dataset into several subsamples that classify the observations in the subsamples but the independent variables (Raileanu 2004). The best split is found by calculating the weighted sum of the Gini Impurity for all possible branch nodes. This process is then repeated until no more splits can be made and/or there is no more information gain that can be earned with the introduction of additional splits. For the missing variables needed for imputation, the imputed value that is used is a random selection from the terminal node in which the observation with the missing variable falls into and is randomly chosen from the observed values of the observations that fall into the same terminal node.

Using the example of the missing stunting progress variable for the United Kingdom a CART decision tree is fit and split at points that create more homogeneous subgroups. There will be very few observations in each node as the tree splits further and further from the initial node. After going through splits in the tree from GDP, to anemia rate, to cholesterol levels, and all the other covariates in the tree, if the tree can no longer be split after the variable for stunting index, then an observation from the observations that ended in the terminal node that also includes the observation that is missing a value for stunting progress prevalence would be randomly selected. The randomly selected observation would have to have a value for stunting progress, and not also be missing. That value from the randomly selected observation is then inputted as the data point for the United Kingdom. This process is then repeated for all missing data points for all of the covariates. This method is a favored method for the imputation of numerical variables but does take a lot of processing time due to the number of numerical variables in the dataset that

require imputation, but using this method for categorical imputation leads to a smaller processing time.

Country	Stunting Progress
United Kingdom	*NA*
Uruguay	Off Course- Some Progress
Uzbekistan	On Course- Good Progress
Vanuatu	Off Course- No Progress
Venezuela	Off Course- Some Progress
Yemen	Off Course- Some Progress
Zambia	Off Course- Some Progress

Table 4a. The first table shows the observations and the United Kingdom missing a value for the variable.

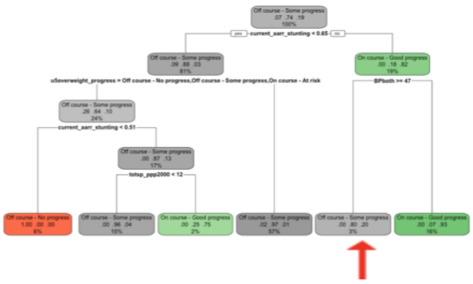

Figure 6a. Shows a subset of the decision tree used to find the terminal node in which the observation the United Kingdom falls into.

Country	Stunting Progress
United Kingdom	*Off Course- Some Progress*
Uruguay	Off Course- Some Progress
Uzbekistan	On Course- Good Progress
Vanuatu	Off Course- No Progress
Venezuela	Off Course- Some Progress
Yemen	Off Course- Some Progress
Zambia	Off Course- Some Progress

Country	Stunting Progress
Nicaragua	Off Course- Some Progress

Figure 6b. Shows the randomly selected observation from the terminal node is Nicaragua. The observed value for this variable for Nicaragua is then imputed as the value for the United Kingdom's "stunting_progress".

CORRELATION ANALYSIS

Correlation analysis is a critical part of feature reduction when using data for means of prediction. This is especially true when there are a large number of variables. In correlation analysis, the relationship between all numerical variables in the data set is calculated to produce a correlation coefficient. The Pearson correlation coefficient is a statistical measure of the strength of the relationship or association between two variables in the dataset. The range of the coefficient is between -1 and 1. A correlation coefficient between (0, 1] represents a positive correlation and a correlation coefficient between [-1, 0) represents a negative correlation. The closer the correlation coefficient is to either -1 or 1 shows that there is either a perfect positive correlation between the two variables, or there is a perfect negative correlation between the two variables. The pearson correlation coefficient is calculated as

$c = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$ with $c$being the Pearson Correlation Coefficient, $x_i$is a variable sample from the data, $y_i$is another variable sample, $\bar{x}$is the mean of the values from the variable x, and $\bar{y}$ is the mean of the values from the variable y.

Correlation analysis is an important portion of exploratory data analysis since it is a way to perform variable reduction before modelling. Two variables that are perfectly correlated could be harmful to keep in a dataset because it can lead to overfitting of the data if two variables have a similar impact on the observations in the dataset. The inclusion of two perfectly correlated variables could also lead to an inflated impact of the variables within the regression model. This is why correlation analysis can be a very important tool when looking at feature reduction.

Figure 7 demonstrates the necessity for the correlation tests of the data. Because of the number of variables that are present in the dataset, the color coded correlation plot is hard to interpret, meaning that the correlation statistic will be calculated for all variables that meet the assumptions under the Pearson's correlation coefficient.

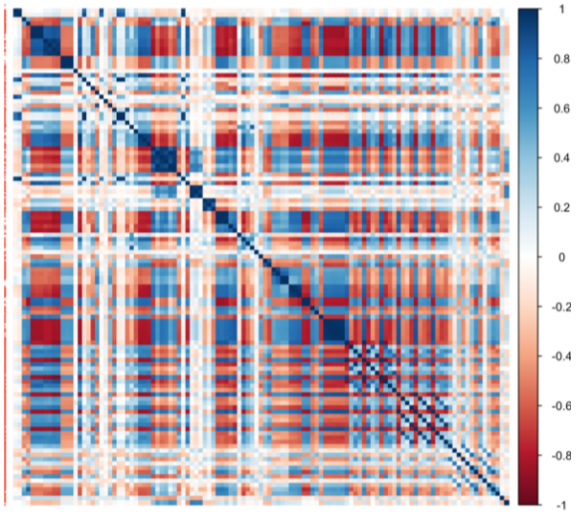

Figure 7. A color coded correlation plot of the variables. Due to the high number of covariate, the chart is to visualize the extent to which there are highly positively correlated variables (dark blue) and highly negatively correlated variables (dark red).

The threshold of how close a correlation coefficient should be to either one or negative-one before the removal of one of the highly correlated variables isn't standardized and depends on numerous factors including the number of variables that the researcher would like to reduce the total number of features by. In this study, the threshold of the removal of features based on the correlation coefficient was .95. After that, the variables were evaluated to ensure that the value of the variables is redundant to each other and therefore the dataset would not suffer from the removal of one of the variables.

Due to the aggregate nature of the data and the multiple sources, some variables would provide the same information to the models as other variables, this dataset is a very good candidate for further variable reduction using the Pearson's Correlation Test to ensure that the correlation coefficient for the variables was significant. In this test the null hypothesis or H_o is that the true correlation coefficient is equal to zero. The alternative hypothesis or H_1 is that the actual correlation coefficient is not equal to zero.

	Under 5 Population	Women Anemica	Under 5 Mortality 2012	Under 5 Mortality 2011	GDP_4
Total Population 2015	0.96996789	0.92345468	0.008798058	0.009876511	-0.07061225
Under 5 Population	1.00000000	0.96745761	0.081542895	0.084584178	-0.11483508
Under Five Mortality 2009	0.08768731	0.05724683	0.984587293	0.998741455	-0.52001210
Under 5 Mortality 2012	0.08154289	0.05302901	1.000000000	0.985491044	-0.51479634
GDP_1	-0.10972266	-0.08180646	-0.520801951	-0.522853345	0.99660757

Table 5. Shows a subset of the variables and the corresponding correlation coefficient. The total population and the under 5 population, and all of the variables for the under 5 mortality rate have correlation coefficients over .95.

Using the .95 or greater correlation coefficient threshold, 56 variables were removed from the dataset due to the high correlations. After the results from the exploratory measures and the feature reduction measures the final dataset that will be used when implementing the modelling techniques contains 13 categorical variables 10 of which are ordered factor variables and 3 of which are unordered. And 59 numerical variables including the dependent variable of diabetes prevalence.

METHODS

The research goals of this study are to be able to find the covariates that are most important when explaining diabetes prevalence across the world's countries. The secondary goal of the study is to choose a model that best performs with the dataset.

One of the initial methods employed in this study is the use of an unsupervised learning method of K-Means clustering. One of the limitations of other studies into diabetes prevalence and the risk factors to diabetes is the regional focus of the study. Intuitively there are many variations between the social, economic, and political norms of a country that could lead to differences in how diabetes is related to the population. Beyond the regional differences the k-means clustering method will allow similarities and differences between the observations that

go beyond regional to be explained. This will allow the models to be evaluated not only on how they are able to explain the dataset as a whole, but also how the models are able to perform within each of the unique clusters that represents a different grouping of the observations.

The supervised learning methods used in this study are 6 different methods that vary in their ability to handle data frames with large numbers of features, and in their feature selection capabilities. These are then split into three categories of stepwise regression methods, ensemble regression methods, and penalized regression methods. This first of methods is stepwise selection and backward selection. The ensemble regression methods included random forests- an ensemble method that utilizes bagging and decision trees and gradient boosted models- an ensemble method that uses a set of boosted weak learners to increase the predictive power of the model. The penalized regression methods include ridge regression- a multi regression method that is especially good at handling multicollinear data, particularly for cases in which the number of covariates is greater than or close to the the number of observations and lasso regression, another penalizing form of shrinkage regression model that typically produces more sparse model.

The evaluation of these models will illuminate what are the variables that the models found to be the most important, and which model was able to produce the smallest prediction error rate.Model averaging methods are also evaluated to understand if combining the models would produce a better predicted error results than the original models alone.

K-MEANS CLUSTERING

The use of unsupervised methods is often utilized to understand the relationships between the observations and how much of the differences in the observations can be explained by the

variations of the covariates. K-Means Clustering is a common unsupervised algorithm used for clustering. In K-Means Clustering the variables, several k centroids are identified, and the variables are grouped around the centroids based on their distance from that centroid and based on the variables that are represented by that centroid. An observation is then assigned to the centroid that it is closest to. The goal of K-Means Clustering is to minimize the variation between points within the cluster as well as minimize the distance between the data points and the centroid of the cluster.

The centroids are determined through a repetitive process in which the K- Means algorithm assigns data points to a cluster based on the initial iterations of clusters, and then afterwards chooses a new centroid that better represents the center of the cluster as more observations are added (Rahman 2013).

In order to be able to place the data points into clusters, a proximity measure (a measure of how similar or dissimilar two data points are) is defined. K-Means clustering utilises the Euclidean distance. The euclidean distance is given by $d(x, y) = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2}$ where $d(x, y)$is the distance measures between the points $x$and y.

The objective function in K-Means Clustering is the function that measures how well the clusters are able to explain the dataset. The objective function in the K-Means Clustering algorithm is the sum of squared error or the SSE. It is the sum of the squared distance between the centroid and all of the members a part of that centroid. The SSE is then defined as

$SSE = \sum_{j=1}^{k} \sum_{i=1}^{l} ||x_i^j - c_j||^2$ where k is the number of clusters l is the number of observations, x_i is observation i, c_j is the centroid for the jth cluster.

The first step of K means clustering is the choice of the most optimal number of clusters k. The methods employed in this study to choose the value of $k$are a combination of the elbow plot and the average proportion of non-overlap (APN). The first measure is the elbow plot. The elbow plot shows the total within sum of squares by the potential number of clusters k. The total within sum of squares is calculated as $\sum_{m=1}^{M} \sum_{x_i \in C_m} (x_i - \overline{x_m})^2$ where $M$is the number of clusters, x_i is each observation of the dataset, and $\overline{x_m}$is the mean of the cluster C_m.

The elbow method uses the elbow plot to understand at what number of clusters the total sum of squares reaches a minimal point without seeing diminishing returns from the increase in clusters. This is because the total within the sum of squares is 0 when the number of clusters matches the number of observations. For K-Means Clustering, it is ideal to reach a point where the total within the sum of squares is close to 0, without an excess in the number of clusters. The elbow method uses the point where the plot starts to elbow as the point in which we have the optimal number of clusters. Figure 8 shows that there is a concave in the plot, but there is no clear elbow for the dataset. Because of this, the elbow plot is used to define a range for the optimal number of clusters in conjunction with the measures of the APN.

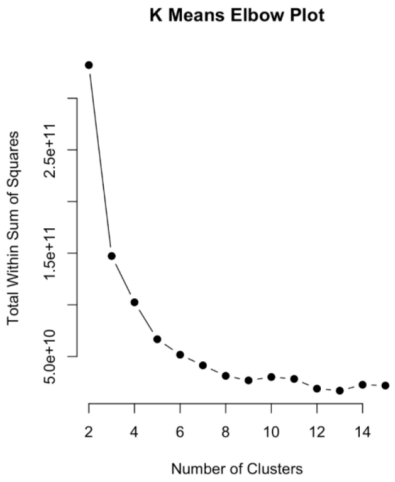

Figure 8. K-Means Clustering Elbow Plot

The average proportion of Non-Overlap or the APN is another measure used to determine the optimal number of clusters in K means clustering. It is a measure of stability and is especially useful when using data that has the potential for high correlation. The APN is used to determine the optimal number of clusters by evaluating the average proportion of the observations in the data that are not within the same cluster based on the entire data set, then clustering on the entirety of the data set with individual columns removed from the dataset. The APN is calculated as $APN(k) = \frac{1}{MR}\sum_{i=1}^{R}\sum_{l=i}^{M}(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n(C^{i,0})})$ where $C^{i,0}$ is the cluster that contains the i^{th} observation based on the entirety of the dataset and $C^{i,j}$ is the cluster that contains the i^{th} observation with the column l removed. K is the total number of clusters. The range of the APN is between 0 and 1 and the optimal cluster will have the minimum APN when compared to those of other clusters (Brock 2008).

The elbow plot began to elbow around k = 5 clusters and therefore that is where the evaluation of the APN begins. Starting at 5, the APN values for each of the clusters are shown in figure 8.

Number of Clusters	APN	Number of Clusters	APN
5	.0118	13	.0158
6	.0066	14	.0152
7	.0041	15	.0178
8	.0095	16	.0175
9	.0147	17	.0159
10	.0099	18	.0176
11	.0140	19	.0162
12	.0137	20	.0153

Table 6: APN of cluster sizes between 5 and 20.

The number of clusters with the lowest APN is 7 clusters. This is in range with the area where the elbow plot showed a slight elbow shape, and thus is used as the final number of clusters for the analysis.

The result of fitting the K-Means clustering algorithm with the 7 clusters is 98.9% of the variation in the dataset being able to be explained by the clusters. Although the amount of variation that can be explained by the clusters is high, the range of the number of observations within each cluster is between 2 in the cluster with the smallest number of countries to 96, being the cluster with the largest number of clusters. Cluster 1 contains only the countries China and India. These two countries are most similar and deviate greatly from the rest of the countries when it comes to population, which could have influence on the other nutritional and health indicators that are correlated to high population levels, leading these two countries to be in the same cluster. For example, looking at the quintile distribution of the total population in Table 7 shows the extent to which the clusters vary just by total population

Cluster Number	25th Percentile	50th Percentile	75th Percentile
1	1312189.344	1341988.438	1371787.531
2	1353.215	5460.590	10151.945
3	26622.020	31649.744	41013.823
4	183523.438	188144.047	203657.203
5	612.290	4034.045	10084.483
6	1898.728	2966.975	4092.267
7	66796.285	81019.156	99657.254

Table 7. Shows the Cluster number, and 25th, 50th, and 75th percentile for the variable for total population.

		Continents				
		Africa	Americas	Asia	Europe	Oceania
Clusters	1	0	0	2	0	0
	2	2	2	6	19	2
	3	12	4	8	4	0
	4	1	1	3	0	0
	5	34	26	13	12	11
	6	0	0	3	1	0
	7	2	1	5	4	0

Table 8. The table for the continental region distribution among the 7 clusters.

Table 8 shows the distribution of the clusters by region. Cluster 2 has 31 countries with the majority of the countries coming from Europe and Asia. Cluster 3 contains 28 countries with a more even spread from the other five regions, although a majority of the countries are within Africa. Cluster 4 is another cluster with a small number of countries as this cluster includes only three countries from Asia (Bangladesh, Indonesia, and Pakistan), one country from Africa (Nigeria), and one country from the Americas (Brazil). These countries in terms of population are not at the scale of India and China, but outside of those two countries and the United States,

the 5 countries in this cluster make up numbers 5-9 of the countries with the highest populations in the world (Population Reference Bureau 2015). The clusters are showing the large role that population plays in terms of variation among the clusters. Cluster 5 is the largest cluster containing 96 countries. 35% of the countries within this cluster are from Africa and 27% of the countries are from the Americas. The rest of the countries are split relatively evenly between Asia, Europe and Oceania with 13,12, and 11 countries respectively. Cluster 6 is another small cluster containing only the countries of Kuwait, Luxembourg, Qatar, and Singapore. Cluster 7 has 12 countries mainly split between Europe and Asia. Although a large number of clusters are split between 3 of the clusters, it is imperative to ensure the training set of the data is representative of the countries from each cluster.

STRATIFIED SAMPLING

Before beginning the modeling portion of the analysis, the training and testing scheme for the models are allocated. The study uses stratified random sampling to ensure that all of the clusters are included in the training and the testing set of the data. Stratified random sampling begins by first separating the data into strata, or data layers. In this case the clusters form the strata. This stratified sampling scheme will take a sample of 75% of the observations in each of the clusters and leave 25% of the cluster for the testing set. Since each of the clusters do not contain the same number of observations, this allows the number of observations in the training and the test set to waver between clusters, but the proportion of the observations in the training and test set to be relatively similar. This comes from each of the clusters and allows for both the training RMSE and the testing RMSE to evaluate the accuracy of the model. Although cluster 1

only contains 2 countries, one observation will be left in the testing data set and the other observation will be selected to be in the training set.

STEPWISE REGRESSION

Stepwise regression methods refer to multiple linear regression methods in which covariates are either incrementally added or removed from the model to produce the optimal accuracy level for the model and the existing covariates. The accuracy measures of the model are routinely assessed until the model with the covariates that achieve this optimal accuracy level are retained. In stepwise regression, depending on whether utilizing stepwise selection or backwards selection (the two methods employed in this analysis) at each recursive step, the model will either remove or add a covariate depending on how this step affects the adjusted R^2 of the models. The R^2 is a value that shows what percentage of the dependent value can be explained by the regression model. The adjusted R^2 value is penalized with the addition of variables especially with variables that are not significant. Therefore the adjusted R^2 value is a better measure of a model's ability to explain the response variable using the independent variables in multiple linear regression. The R^2value is calculated as $1 - \frac{SSE}{SST}$ where the SSE is the sum of squares error calculated as $\sum_{i=1}^{n} (y_i - \widehat{y})^2$ and the $SST$is the sum of squares total calculates as $\sum_{i=1}^{n} (y_i - \overline{y})^2$ where $y_i$the is value of the sample and $\overline{y}$is the mean value of the sample. The adjusted $R^2$is $1 - (1 - R^2)\frac{df_{total}}{df_{error}}$where the df_{total} is the total degrees of freedom $n - 1$where

n is the number of observations and df_{error} is the error degrees of freedom $n - k - 1$where $k$is the number of predictors.

Stepwise selection is a method that combines both the forwards and the backward methods of adding and eliminating variables to obtain the best model. This is generally a successful regression method as it allows the variable importance and the accuracy of the entire model to be evaluated continuously based on additions and subtractions of variables as opposed to just one or the other. The variables are added or removed based on significance measures such as the probability value or the p-value of the variable, or the improvement the addition of the variables has on the adjusted R^2 of the model. The Akaike Information Criterion or the AIC is another measure that is used to determine how well the model is able to fit to the training data. The AIC is evaluated at every step of the process of stepwise selection methods. The AIC is determined using the maximum likelihood estimate of the model as well as the number of independent variables in the model. $AIC = 2(k) \times 2ln(\widehat{L})$where $k$is the number of independent variables in the model and $\widehat{L}$is the maximum likelihood value for the model. Since the number of independent variables is adjusted at every step of stepwise regression, there is a new value for each model until the optimal or minimum value is reached. It is also important to understand stepwise regression is very sensitive to dimensionality, especially where there is a small number of observations in comparison to the number of variables.

The primary metric used to understand the accuracy of the prediction values of the models is the root mean squared error, or the RMSE. The RMSE is a method for understanding the model accuracy that is utilized to compare how well the model was able to fit to the training data and how well the model is able to predict for the test data. The RMSE is calculated as

$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$ Evaluating both the train and the test RMSE give insight into the model fit and prediction accuracy of the model.

STEPWISE REGRESSION: MODEL RESULTS

	R^2	Train RMSE	Test RMSE	AIC
Stepwise Selection	.8701	1.207	5.951	413.665

Table 9. The model statistics for stepwise regression model

Table 9 shows the results of stepwise selection. The adjusted R^2 value of .8701 suggesting that this model is well fit to the dataset, considering the number of variables is close to half of the number of observations of the data. The AIC of this model is 413.665. The AIC of the model is evaluated differently from the R^2 and $adjusted\ R^2$ values as the AIC is best evaluated comparatively to the AIC of other models to understand relative to other methods, which one produces the best fit for the data. The training RMSE of the model is low at 1.207 suggesting that the fit for the training data is doing well with predicting the diabetes prevalence for the observations that were fit in the training data. The testing RMSE is 5.951 which is much higher than the training RMSE. This suggests that although the adjusted R^2 of the model is high and the training RMSE suggests that the prediction error of the training dataset is low, the high value of the testing RMSE suggests that the model is overfit to the training data.

Evaluating the RMSE of the training and the test data across the clusters is helping in understanding where the model is over parametrizing the data. Table 10 shows the training RMSE split by the clusters and Table 11 shows the testing RMSE split by the clusters.

Clusters	Training RMSE
1	0.2566264
2	1.6060195
3	1.2116300
4	1.1167462
5	1.0875930
6	0.9878296
7	1.0585923

Table 10. Stepwise selection training RMSE by the 7 clusters.

Clusters	Testing RMSE
1	7.065779
2	5.108778
3	5.924006
4	6.096839
5	5.423074
6	5.694152
7	11.990259

Table 11. Stepwise selection testing RMSE by the 7 clusters.

The training and the testing RMSE across the clusters shows that the model performed similarly on the training set for all of the clusters. The range of the training RMSE of all of the clusters is .256 and 1.60. Since this initial modelling portion is for the sake of variable selection, iut is important that the model is not weighing more heavily for any cluster. The small range of the training RMSE shows that this is true. The testing RMSE has a wider range across the clusters and the range is more than five times the range of the training RMSE. Therefore although this model is performing well on the training set and will provide useful information regarding the important variables, the wider range of the testing RMSE is indicative that the model is not the best for new data.

Variable Name	Coefficient	Absolute t-statistic	P-Value
subregionUNMicronesia	18.25729239	5.770524407	2.3925E-07
urbanpop	-0.088258667	5.715129096	2.9721E-07
water_unimprov2015	-0.213331109	5.420287563	9.3335E-07
value_UNICEF_ANC_4	0.110199164	5.418633757	9.3931E-07
BGboth	0.575818443	5.416449303	9.4724E-07
GDP_5	9.89067E-05	5.183742194	2.3047E-06
subregionUNEastern Africa	10.66659115	4.736474249	1.2221E-05
subregionUNEastern Asia	15.48064713	4.711631901	1.3384E-05
tothealth_ppp2010	-0.361997223	4.263456863	6.6508E-05
Code_breastfeeding.Q	-6.775681004	4.254852475	6.8537E-05
Class_IodineNutrition.L	-4.334436638	4.044548399	0.00014156
subregionUNMelanesia	10.98214489	4.023637067	0.000152
hypertension_policy.L	-2.072753881	3.651804448	0.00052112
subregionUNMiddle Africa	8.716124197	3.532557572	0.00076303
subregionUNCentral America	8.222212391	3.421351785	0.00108189
LBW	0.217399293	3.280622417	0.00166749
subregionUNSouth America	7.08977928	3.183055036	0.00223663
hypertension_policy.Q	1.262292573	3.108645143	0.00278812
u5overweight_progress.Q	-1.476799927	3.097484768	0.00288105
totag_ppp2010	-0.275623299	2.776243534	0.00717568
subregionUNCaribbean	6.083382077	2.775448287	0.00719135
gender_inequality_index	7.203516059	2.662030512	0.0097775
Code_breastfeeding^6	-3.384598489	2.626142451	0.01075758
subregionUNWestern Africa	6.183364284	2.618509665	0.01097722
subregionUNNorthern Africa	6.685941018	2.609740404	0.01123459
Code_breastfeeding^5	3.957143996	2.607050848	0.01131461
water_piped2015	0.051886293	2.395124682	0.01950543
subregionUNSouthern Asia	5.573920381	2.322450632	0.02335099
WHAtarget_WRAanaemia1.L	-1.93591014	2.247672766	0.02799672
subregionUNCentral Asia	5.18516914	2.241304955	0.02842769
fruitandveg_gram2011	0.002567209	2.036927229	0.04573386

Table 12: The table showing the significant variables/variable levels used in the stepwise regression model.

Table 11 shows the final variables and the levels of the ordered variables that are used in the final model. In order to assess the variable importance of for the stepwise selection, the absolute value of the t-statistic for each of the variables and each of the levels of the are ranked. The p-value, the magnitude of the coefficients and the t-statistic ranking are all methods to understand how the model interpreted the importance of the variables used. The t test statistic is $t_{stat} = \beta_i / SE$ where β_i is the coefficient of the variable and $SE$is the standard error. Table 11 contains the coefficients for the variables in the first column, the absolute t statistics for the variables in the second column, and the p-values are shown in the third column. The results of the model showed that there are 17 significant variables as a result of the stepwise regression

method. The most significant variable is the variable describing the rate of change of the overweight under five population. This is a metric that compares the previous year rate of children under the age of five that are overweight with how that metric has either increased or decreased the following year. It is known that weight concerns are a risk factor to diabetes, but this metric emphasises that risk for the population that is under five years old. The next variable of high significance is the variable for the average annual rate of reduction for stunting. Stunting is a measure of how the child population of a region is doing developmentally compared to the world standard. These first two significant variables are both variables that emphasize the child population of a region, and also show that childhood factors have very high importance when it comes to diabetes prevalence. The variable for subregion is also highly significant for specific regions. The stepwise regression model shows that the regions as categorized by the UN as Melanesia, Caribbean, Polynesia, and Northern Europe are all very significant. Examining the coefficients of these four regions shows that Melanesia, Caribbean, and Polynesia all have high positive coefficients within the range of 4.2 and 9.9. This indicates that these regions present an increase in diabetes prevalence relative to the other regions. Northern Europe is one of the few regions with a negative coefficient, meaning that this region presents a decrease in diabetes prevalence relative to the other regions. Other variables that were significant to this model were GDP, hypertension policy, gender inequality index, breastfeeding code, urban population. The full list is shown in Table 11.

BACKWARDS REGRESSION

Backward elimination is a form of stepwise regression in which all of the covariates of the data are fitted in the model at once. This formula that contains all the covariates of the model is what we are calling the fully saturated or full model:

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where n is the total number of covariates in the dataset and β_i is the coefficient of the variable x_i. This model is called the fully saturated model since the formula is completely full with all of the possible predictors that could influence the dependent variable. During each recursive step of backward elimination, the covariates that have the least significant influence on the model are removed based on their p-values evaluating the null hypothesis that $B_0 = 0$ against the alternative that $B_i = 0$. The removal of these variables results in an improvement in the adjusted R^2 of the model. This process continues until all of the variables retained in the model have a significance level above a predefined threshold. The threshold held for the p-values of the variables in the stepwise models used in this analysis is the .05 significance level.

BACKWARDS REGRESSION: MODEL RESULTS

	R^2	Train RMSE	Test RMSE	AIC
Backwards Selection	.8891	0.8185621	8.960152	164.5369

Table 13. The model statistics for stepwise regression model

The results of backwards selection showed a higher $adjusted\ R^2$ value for the model to be .8891, and $adjusted\ R^2$ value higher than that of stepwise selection. This higher $adjusted\ R^2$ may be due to the higher number of variables that are in the backwards selection model than the

number of variables in the stepwise selection model. In further comparison the AIC of the backwards regression is 164.5369. This is an AIC that is lower than the AIC for the stepwise regression that is adding further indication that the backward selection model is a good fit to the data. However when looking at the training and the Test RMSE shows a difference in the model's prediction ability. The train RMSE is very low at .818 while the test RMSE is very high at 8.960. This is a very testing RMSE and signifies that the testing data was not able to be fit to the dataset as well as the training set, signally overfitting of the training data.

Clusters	Training RMSE
1	0.1836124
2	0.9872797
3	0.6934035
4	0.7896142
5	0.8544768
6	0.5470062
7	0.3106204

Table 14. Backwards selection training RMSE by the 7 clusters.

Clusters	Testing RMSE
1	6.505197
2	7.636887
3	7.474230
4	2.937232
5	8.818825
6	4.093466
7	20.083623

Table 15. Backwards selection testing RMSE by the 7 clusters.

Table 14 and Table 15 show the difference in the training RMSE of the backwards selection model by the 7 clusters and the testing RMSE by the 7 clusters. The results of the testing RMSE shows that the model had very inflated values for the testing RMSE of cluster 7. This is a similar pattern to what occurred using the stepwise method, although for backwards selection the inflation is about twice the magnitude of that rmse of the same cluster in the results of stepwise regression.

Variable Names	Coefficient	Absolute t-statistic	P-Value
water_unimprov2015	-0.304177597	5.451683697	4.08131E-06
BGboth	0.83310837	5.101433577	1.18059E-05
value_UNICEF_skilled_att	-0.16532193	4.965387886	1.78101E-05
toteducation_ppp2010	-0.310651621	4.16899099	0.000190947
valuenurse	-0.40291428	4.071008249	0.000254097
tothealth_ppp2010	-0.451371033	4.039630623	0.000278338
cholesterol_BOTH	0.284946546	3.934968745	0.000376666
stunting_progress.L	2.888377993	3.792977983	0.000565679
Code_breastfeeding.C	-9.567281794	3.71878904	0.000698296
subregionUNEastern Asia	17.95076763	3.699552157	0.000737328
Code_breastfeeding^7	-6.147942515	3.632736168	0.000889995
diabetes_policy.L	-4.735447032	3.607470141	0.000955343
subregionUNMicronesia	16.55487154	3.509460305	0.001255476
Code_breastfeeding^9	-5.343064079	3.47969363	0.001363358
gender_inequality_index	10.91298394	3.256053475	0.002510939
subregionUNEastern Africa	11.17414159	3.25446632	0.002521701
subregionUNMelanesia	12.79523389	3.249962048	0.002552482
ORS	-0.074205129	3.115170557	0.003657939
hypertension_policy.Q	2.249982514	3.114411065	0.003665296
water_surface2015	-0.194095971	3.113477424	0.003674359
water_piped2015	0.094547415	3.076457922	0.004051343
Code_breastfeeding^5	7.558618455	3.07259429	0.004092732
prev_wasting	-0.374277896	2.827959622	0.007698587
rate_stuntingtrend4	-0.194624249	2.733218599	0.009768281
totalpop2015	-6.33583E-06	2.71990981	0.010097388
prevalence_vita	-0.13385007	2.671667722	0.01137857
Fortification.Q	-2.789978544	2.66513443	0.011563216
Class_IodineNutrition.L	-3.825764633	2.580036033	0.014235217
Class_IodineNutrition^4	1.99807331	2.53551767	0.015849361
Code_breastfeeding^8	-5.4821053	2.515416864	0.016631816
WHAtarget_WRAanaemia1.L	-2.527317184	2.42276937	0.02071553
reqd_aarr_stunting	-1.2699611	2.389327643	0.022400694
prev_u5overweight	-0.281964139	2.357943533	0.024094251
fruitandveg_gram2011	0.00406654	2.278711265	0.028896226
rate_stuntingtrend5	-0.120932077	2.239675557	0.031564624
early_BF	0.057266775	2.21732548	0.033189791
Fortification.L	-1.308179269	2.200873233	0.034433413
rate_stuntingtrend3	0.153823442	2.188276156	0.035413521
ob_bothsexes	-0.195486997	2.144358666	0.039027405
diabetes_policy.Q	1.467629035	2.090042042	0.043946285
value_ebf_current	-0.059466272	2.079672547	0.044945113

Table 16. The coefficients, absolute t-statistic, and p-value of the significant variables/variable levels in the backwards selection model.

The variable that is the most significant by the p-value and the absolute t-statistic is the variable for the coverage of unimproved water. Unimproved water is defined as a drinking water source that is not able to protect the water from contamination (Sustainable Sanitation, Water Management & Agriculture 2020). It is the lowest category in terms of safe drinking water. This variable is just above the blood glucose levels of the population as the most important variable.

Drinking water safety is not a variable that is colloquially associated with Type 2 Diabetes Mellitus and may be a proxy for other variables that directly impact the propensity of contracting the illness. Other factors such as food access, and allocation of government spending towards social improvement ventures. The variable that is third on the list in terms of variable importance is the variable that describes the percentage of births that had a skilled attendant present. Similarly to the variable around safe drinking water, unlike blood glucose levels, this is not a variable that is commonly known to be associated with Type 2 Diabetes. This may signify that there are risk factors that can be seen at birth that can be mitigated with the presence of a skilled birth worker, or it could be a proxy for other factors related to the nation that have a more direct influence over the risk of Type 2 Diabetes Mellitus. Some of the other variables that were deemed important by the backwards regression model as the variables are the government expenditure on education, the density of nurses or midwives per 1000 people, and the government expenditure on health. Besides the blood glucose levels, all of the top five variables are around the availability of health workers in the country and social programs related to the state of the resources. This group of variables are shown by the backwards regression model to be the most important for prediction of Type 2 Diabetes Mellitus prevalence.

RANDOM FORESTS

Random Forest is an ensemble supervised learning method that utilizes a technique called bagging. The term "random forest" is called as such because it is made up of a series of decision trees that make up the forest. A decision tree is a supervised method of prediction where the dataset is split at crucial nodes in a way that would make the separate splits at each of the nodes more homogenous subsets. The splits into more homogenous subsets are the algorithm's pursuit of information gain. The decision tree is organized based on the amount of information the algorithm can have at each of the splits. The split in which the algorithm derives the largest amount of information gain is the first node and the amount of information gain trickles down until the terminal node. The terminal node is the final node in which no further splits can be made, or no more information is gained from further splits. The rule of the split is defined as

$$\theta(x, c) = (n_1/n)\times 1/n_i \times \sum_{x_i \le c} (Y_i - \overline{Y}_l)^2 + n_r/n \times 1/n_r \times \sum_{x_i \le c} (Y_i - \overline{Y}_r)^2$$

for the root node $x \le c$ and $x > c$ where l and r are the left and the right split of the node. Where the process of training multiple trees with replacement occurs is to combine trees. This is the ensemble method in which individual classifiers can be combined to make a stronger learner. In the case of the random forests, the individual decision trees make up what are called "weak learners". The random forests use the series of "weak learners" on samples of the dataset and use the information gained from the prediction of this tree to influence the overall prediction of the model. This is where the technique of "bagging" is instrumental to the construction of random forests as this algorithm is based on the idea that combining a series of smaller algorithms or "weak learners" will increase the overall result of the algorithm (Nature Methods, Altman 2017).

Random forests differ from the stepwise methods previously explored as they are less sensitive to data imbalances between the number of predictor variables and the number of

observations. Because the ensemble bagging method can be done with very few variables and observations, that is not a penalty that needs to be considered when creating our training and test sets. The random forest algorithm relies heavily on the mean squared error to determine the next steps in creating the trees that make the forest. The MSE is defined ast

$MSE = \frac{1}{N}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$ where $N$is the number of data points, $\widehat{y_i}$is the predicted value and $y_i$is the observed value, and i represents the datapoint.

The number of trees that make up the random forest is chosen in a way that minimizes the error of the model when fitting the number of trees. When creating the algorithm, there is also the task of not wanting to over-saturate the forest with a very high number of trees even when the extra number of trees is not providing any reduction of error.

RANDOM FORESTS: MODEL RESULTS

	Pseudo R^2	Train RMSE	Test RMSE
Random Forest	.5244	1.381924	2.653056

Table 17. The model statistics for the random forest model.

Differently from the stepwise methods, a $Pseudo\ R^2$value is used for random forests. The $pseudo\ R^2$ is calculated as$\frac{1-MSE}{Var(y)}$. This value shows the percentage of variance explained by the random forest model. The pseudo R^2 for this model is .5244. Comparatively to the stepwise regression models this is a much lower value for R^2. Despite the low$pseudo\ R^2$for the random forest model, the RMSE for the training dataset is 1.381 and the RMSE for the testing

Clusters	Training RMSE
1	0.8758745
2	0.9141516
3	1.3568293
4	0.9401842
5	1.5533973
6	1.1339462
7	1.2649245

Table 18. Backwards selection training RMSE by the 7 clusters.

Clusters	Testing RMSE
1	0.579715
2	1.876098
3	1.831983
4	0.389296
5	3.171361
6	2.060554
7	2.419882

Table 19. Random Forest testing RMSE by the 7 clusters.

dataset is 2.653. Of the previous models, this train RMSE and the test RMSE have smaller range between the two values than the previous models. The range of the testing RMSE of the tree is also smaller than the range of the test RMSE of the stepwise models. This shows that no cluster is disproportionately penalized in the random forest.

There are multiple ways to evaluate the variable importance of a random forest. One of the ways is the minimal depth of the variable. The minimal depth is a measure of the predictive power of a variable in a tree within a random forest. This is a method in evaluating the variable importance of a random forest as it is the measure of the depth of the node relative to the root of the tree. If the minimal depth for the variable is low, then there are a lot of observations of the data that were divided into groups based on this variable and it is closest to the root node of the tree. This is a measure of the variable importance that isn't based on the error rates of the variables like the other variable importance measures, but it is instead based on the shape and the recursive steps of the trees defined in the random forest.

Variable Name	Mean Minimal Depth	P value
subregionUN	1.87E+00	0.00E+00
ob_bothsexes	4.42E+00	5.92E-125
BGboth	5.05E+00	7.32E-97
prevalence_vita	6.99E+00	8.05E-34
value_gini	7.99E+00	5.58E-19
fruitandveg_gram2011	7.99E+00	7.38E-18
number_u5overweight	8.04E+00	1.90E-17
WRAanaemia_NUMBER	7.75E+00	2.61E-17
urbanpop	8.14E+00	2.49E-15
reqd_aarr_stunting	8.13E+00	3.12E-13
BPboth	7.99E+00	7.04E-13
UnmetneedTotal	8.08E+00	3.48E-12
value_undernourishment2000	7.76E+00	1.64E-11
totalpop2015	8.11E+00	1.30E-09
toteducation_ppp2010	8.31E+00	6.38E-09
LBW	7.94E+00	7.96E-09
early_BF	8.11E+00	1.54E-08
cholesterol_BOTH	8.25E+00	8.37E-08
over65pop	8.05E+00	2.82E-07
early_childbearing	8.42E+00	1.15E-05
EBF_trend_2	8.52E+00	7.01E-05
current_aarr_u5overweight	8.72E+00	4.22E-03
totsp_ppp2010	8.83E+00	4.73E-03

Table 20. Significant variables in order of smallest p-values of the random forest model.

Table 20 also shows the p-values for the significant variables that were used in the random forest model. Similarly to the stepwise selection and the backwards selection model, subregion is a very important variable to this model. The next two variables that are shown to be significant to the random forest models are the percentage of the population with elevated blood glucose levels, and the percentage of the population that is considered obese. Elevated blood sugar levels and obesity are two variables that are commonly known to be associated with Type 2 Diabetes Mellitus so it makes sense that these two variables will have strong variable importance for this model. The fourth most important variable is the prevalence of vitamin A deficiency in preschool aged children. The random forest model prioritizes health factors that are more representative of how the population is doing in terms of health markers instead of government policy government intervention in creating social and economic intervention programs to maintain positive health markers, as is shown in the other models.

GRADIENT BOOSTED MODELS

Another ensemble method is gradient boosted models. This method uses decision trees that are recursively fit to the dataset in a sequential additive form such that there is information gain after each iteration of the model. This allows for the next model in the iteration when combined with the information gained from the previous model to have minimal prediction error. This makes this method similar to random forests in that the system of "weak learners" (the individual decision trees) makes up the interactive learners at each step. But the difference lies in how the trees are combined. The loss function for a Gradient Boosted Algorithm is the mean squared error defined as $1/n \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$ where y_i is the observed value and $\widehat{y_i}$ is the predicted value. The gradient descent of the algorithm allows the algorithm to continue to gain information to enhance the predictions of the model and continuously minimize the MSE of the model. As opposed to a random forest that fits the series of trees independently of the other trees and then averages the predictions. In Gradient Boosted Models, the final model is constructed sequentially, building on the previous iterations. This method is repeated until the mean squared error of the model is minimized. The way to illustrate the method in which the gradient boosted model is able to learn through the repetition is to understand the prediction errors. If the errors for given data points are larger through the first iteration within a gradient boosted model, the model can now learn from this information, and use it as a point of focus in the next iteration of the next run of the model. Using this information to focus on gaining strong learners from the information of previous iterations.

The process of Gradient Boosted Trees was developed by Jerome H. Friedman (Friedman 2001). He proposed the following method for Gradient boosted Tree Models with the

differentiable loss function is defined as $L(y,\ H(x))$. There exists a training data set {($x_i,\ y_i)\}_{i=1}^{n}$ with Miterations. The gradient boosting model is initialized as $H_o(x) = argmin_\delta \sum_{i=1}^{n} L(y_i,\ \gamma)$ where γis a constant. For the instance of m = 1 for M iterations, $r_{im} = -[\frac{\partial L(y_i, H(x_i))}{\partial F(x_i)}]_{H(x)=H_{m-1}(x)}\ for\ i = 1,\ldots,n$. The base learner $f_m(x)$ is then train on the pseudo residuals: $\{(x_i,\ r_{im})\}_{i=1}^{n}$. The constant $\gamma_m$is derived by solving $\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(y_i,\ H_{m-1}(x_i) + \gamma f_m(x_i))$ The model is then updated with the constant γ_m as $H_m(x) = H_{m-1}(x) + \gamma_m f_m(x)$

In order to fit the optimal Gradient Boosted Model (GBM) the hyperparameters of the model are necessary to be tuned towards values that will produce the best model. The hyperparameters that are necessary for tuning include the number of trees in the model. The shrinkage of the model is also a hyperparameter to be tuned that is in charge of the learning rate of the model. Typically the smaller the learning rate of the model, the more optimized the model will be. The range for the shrinkage of the model is $0 < l < 1$. The interaction depth of the model is a hyperparameter that determines the total number of splits that are necessary to perform in the model from a single node. In order for these hyperparameters to be determined a max model is iterated through to determine where the model reaches a maximum R^2value. In this case, the GBM model was initially fit on 10000 trees, with a shrinkage of .001, and an interaction depth of 1. After iterating onto these parameters, the optimal hyperparameters for a model of this data results in 150 trees, a shrinkage rate of .1, and an interaction depth of 1.

GRADIENT BOOSTED MODELS: MODEL RESULTS

	R^2	Train RMSE	Test RMSE
Gradient Boosted Model	.7693	2.390929	2.88907

Table 21. The model statistics for the Gradient Boosted Model

Table 21 shows the model performance statistics of the GBM model using the designated optimal hyperparameters. The R^2 value of the GBM model is .7693. The training RMSE of the GBM models is 2.390. The R^2 suggest a moderately well fit model for the data. The difference between the training RMSE and the test RMSE is smaller than in previous models which suggests a better fit for data outside of the training set.

Clusters	Training RMSE
1	0.5563094
2	1.6350604
3	2.3235822
4	2.2233353
5	2.6611661
6	2.7386807
7	2.0450811

Table 22. Gradient Boosted Model training RMSE by the 7 clusters.

Clusters	Testing RMSE
1	0.6387933
2	1.9177054
3	2.1303849
4	2.0530280
5	3.4624484
6	2.8344739
7	1.0848069

Table 23. Gradient Boosted Model testing RMSE by the 7 clusters.

Comparing the RMSE of the training data among the clusters with that of the testing data among the clusters shows that the RMSE of the testing set is a lot closer to the RMSE of the training set. This suggests a model that is a good fit for both the training and the testing data. The ensemble models performed better than the stepwise regression methods in applicability to the testing set when comparing the RMSE of the clusters.

Variable Name	Overall
BGboth	100
ob_bothsexes	71.18014512
totalpop2015	32.72025778
Fortification.Q	29.91568439
prevalence_vita	29.87221877
LBW	29.38181744
BPboth	17.31335501
san_shared2015	13.98307193
poverty2_4	12.9552599
current_aarr_stunting	11.48853742
WRAanaemia_NUMBER	11.09907798
u5mr2011	9.916040859
early_BF	9.388513358
number_sev_wasting	9.328169484
urbanpop	8.369681311
toteducation_ppp2010	7.976486247
Code_breastfeeding^7	7.658044514
value_undernourishment2000	6.065072156
EBF_trend_2	6.050795879
value_ebf_current	5.203697679
prev_sev_wasting	5.104928305
cholesterol_BOTH	4.015115716
value_gini	3.883131034
fruitandveg_gram2011	3.761086892
water_surface2015	3.557377869
reqd_aarr_stunting	3.497980532
UnmetneedTotal	3.314127409
tothealth_ppp2010	3.180008685
u5overweight_progress.C	3.101116975
rate_femaleED_5	2.9671258
number_u5overweight	2.625330787
GDP_5	2.391156938
Class_IodineNutrition.L	1.59175642
early_childbearing	1.447347722
prev_wasting	1.428129209
san_othunimprov2015	1.422055442
undernutritionrank_1to126	1.334823732
Class_IodineNutrition.C	1.302266905
overnutritionrank_1to116	1.301028758
prev_u5overweight	1.273911603
Code_breastfeeding.Q	1.021487653
water_othimprov2015	0.996678495

Table 24. The variables and variable levels used in the gradient boosted model along with the corresponding

Table 24 shows the variables that were used in the GBM model in order by variable importance. The variable importance measure of the GBM is determined using the relative influence of each variable in the model. The relative influence for each of the variables is determined by analyzing how the variable was used as a split during the building of the trees (Breiman 2001). This measure, along with the improvement of the squared error across all the trees changed as a result of this variable, are used to determine the relative influence of the data. The combination of the number of times the variable is selected for splitting, the weighing of the

squared improvement to the GBM model with the variable split then averaged over the total trees in the boosted model is how the relative importance is derived.

The two most important variables in the GBM model were the variables for the percentage of the adult population with elevated blood glucose levels and the percentage of the adult population that is obese. It is common in the literature surrounding Type 2 Diabetes Mellitus for obesity and elevated blood glucose to be risk factors for the illness, so the results of this model are on par with the existing literature on the disease. Both of these variables were also high on the list of important variables for the random forest. The total population in the country is also a significant variable along with Fortification. Fortification is a variable that has appeared in the other models as well as total population. Similarly for the other ensemble method, the variables that are important to the Gradient Boosting Model are those that tell of how well the country is doing in regards to health markers.

RIDGE REGRESSION

Ridge regression is a linear regression method that is especially utilized in cases where either there is suspected multicollinearity in the data, or the number of predictors in the dataset is greater than the number of observations (NCSS 2021). In ridge regression, the coefficients are not determined using an ordinary least squares estimation method as is typical for other linear regression methods, but instead uses a ridge estimator that provides estimates that are higher in terms of bias and lower in terms of variance. The cost function in regression is the ability of the model to use the predictors in the model to understand the dependent variables. Ridge regression uses a method called shrinkage which adjusts and shrinks the coefficients of the model to address problems such as multicollinearity or problems that come from the number of predictors being larger than the number of observations. Using this ridge estimator instead of the ordinary least squares estimator leads to a reduction in the RMSE, which is one of the metrics of model accuracy utilized in this study. The ridge estimation adds the squared magnitude of each of the coefficients in the model as a penalty to the loss function for the regression model. This process is called L2 regularization. The third term is the L2 regularization term.

$$L_{ridge} = \sum_{i=1}^{n} (y_i - \sum_{j=i}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = ||y - X\beta||^2 + \lambda||B||^2$$

This allows the regularization coefficient βridge regression estimate to be $\beta_{ridge} = (X'X + \lambda I)^{-1}(X'Y)$ for the identity matrix I. This allows the Bias of the ridge regularization coefficient to be $Bias(\beta_{ridge}) = -\lambda(X^{-1}X + \lambda I)^{-1}\beta$which illustrates how the bias increases as the ridge regression parameter λ increases. The variance under the ridge regression regularization coefficient is defined as

$Var(\beta_{ridge}) = \sigma^2(X'X + \lambda I)X'X(X'X + \lambda I)^{-1}$ which shows that the variance decreases as the ridge regression parameter λdecreases.

Standardization is a part of regression that varies depending on the regression method chosen. In ridge regression and other shrinkage regression methods in particular standardization is important in ensuring that the shrinkage penalty applied to all of the coefficients does not unfairly penalize some coefficients and fail to penalize others. Since the penalized terms are made up of the sum of squares of all the coefficients, coefficients that are attached to covariates that are not of the same scale will impact the other covariates differently. Therefore, for use within the ridge regression model and then the subsequent lasso model, all of the covariates are scaled to have a variance of 1. K-Fold cross-validation randomly segments the data into k different segments or folds that are of similar size. One of the segments is used as a "holdout" set, and then the remaining segments (k-1 segments) are fit and tested on the holdout set. This process is repeated k times and each time using a different holdout set. The choice of k is arbitrary and ideally between 5 to 10 folds. The more folds chosen, the lower the bias and the higher the variance. The fewer folds chosen the higher the bias and the lower the variance.

After performing the k-fold cross validation the optimal lambda is needed for the L2 regularization term. The optimal lambda is found by selecting the value of the lambda that has the minimum MSE after running K-Fold validation. The value of lambda cannot have too much weight since that will lead to the L2 regularization term being heavily weighted as well and could lead ridge regression to under-fitting the model. On the opposite side, if the lambda is zero, then the L2 Regularization term is also zero, and without this term penalizing the coefficients, the model is then a regular ordinary least squares term. Running the k-fold cross-validation algorithm yielded an optimal lambda value of 0.6069935.

RIDGE REGRESSION:MODEL RESULTS

	R^2	Train RMSE	Test RMSE
Ridge Regression	.7693	2.19743	3.084565

Table 25. The model statistics for the Ridge Regression model.

The R-squared value of the ridge regression model is .769 Which is a relatively high R-squared value for a model. The training RMSE is 2.197 and the test RMSE of the model is 3.084. One of the breakout interesting facts about ridge regression is that ridge regression is not a linear regression model that is used for the sake of feature reduction.

Clusters	Training RMSE
1	0.3506435
2	1.7522617
3	2.6411589
4	1.8470956
5	2.2304747
6	2.3781197
7	2.0486892

Table 26. Ridge Regression training RMSE by the 7 clusters.

Clusters	Testing RMSE
1	1.865288
2	2.372861
3	2.383440
4	2.534459
5	3.565849
6	3.258759
7	1.384905

Table 27.Ridge Regression testing RMSE by the 7 clusters.

The range of the testing RMSE of the ridge model is very tight in comparison to the other models and does not deviate too far from the training RMSE of all the clusters.The influence of all variables within the training dataset are used as a part of the final model, with the difference being the degrees of influence of the variables. Ridge regression is not a model that is used for variable selection, since all of the variables are used in the final ridge model. The use of ridge regression can help understand how the ranking of these variables in the model compare to the

other models, and then eventually reintroduce the ridge regression model in the model selection portion. The process for evaluating the variable importance for the variables in the ridge regression model is by using the absolute value of the coefficients of the model. All of the features when running a ridge regression model remain in the model, but the varying degrees of the magnitude of the variables on the dependent variable represent how important the variable was to the model.

Variable Name	Coefficient
current_aarr_stunting	3.80107291
gender_inequality_index	1.7559961
hypertension_policy	0.77797817
stunting_progress	0.77353733
BGboth	0.56939321
diabetes_policy	0.50101861
WHAtarget_WRAanaemia1	0.47178496
Fortification	0.43364963
current_aarr_u5overweight	0.36206309
regionUN	0.2594516
wasting_progress	0.23929132
valuephysician	0.16810716
LBW	0.15304641
prev_sev_wasting	0.14491506
Class_IodineNutrition	0.14233959
subregionUN	0.12165445
valuenurse	0.11659044
reqd_aarr_stunting	0.11643842
tothealth_ppp2010	0.11641981
prev_wasting	0.09244628
number_sev_wasting	0.09077429
continent	0.08876246
totag_ppp2010	0.07449564
Code_breastfeeding	0.07177065
BPboth	0.06572289
reqd_aarr_u5overweight	0.06120694
WRAanaemia_RATE	0.05620874
water_unimprov2015	0.05398747
over65pop	0.05094491
value_gini	0.04844283

Table 28a. The first 31 variables used in the Ridge Regression model in descending order of the coefficients.

Variable Name	Coefficient
value_undernourishment2000	0.04611183
ob_bothsexes	0.04470564
ORS	0.03699874
early_BF	0.0359241
u5overweight_progress	0.0354669
DTP3	0.03542471
value_nonstaples2009	0.03302311
prevalence_vita	0.03242518
cholesterol_BOTH	0.03107284
toteducation_ppp2010	0.03074198
early_childbearing	0.0206031
san_opendef2015	0.01873228
EBF_trend_2	0.01793086
urbanpop	0.01648851
rate_stuntingtrend3	0.0164271
rate_stuntingtrend4	0.01521509
rate_stuntingtrend5	0.01380229
value_UNICEF_skilled_att	0.01306809
RTF_level1	0.01185373
water_othimprov2015	0.01048008
poverty2_4	0.01038175
totsp_ppp2010	0.01030402
value_ebf_current	0.0092888
undernutritionrank_1to126	0.00846217
UnmetneedTotal	0.00823063
prev_u5overweight	0.00670675
rate_femaleED_5	0.00648655
san_improved2015	0.00600198
continued_BF	0.00433942
san_othunimprov2015	0.00413653
value_UNICEF_ANC_4	0.0038926
u5mr2011	0.00378374
san_shared2015	0.00223529
fruitandveg_gram2011	0.00184666
overnutritionrank_1to116	0.00158967
water_piped2015	0.00133166
water_surface2015	0.00117663
number_u5overweight	0.00039419
GDP_5	3.7317E-05
WRAanaemia_NUMBER	3.7436E-06
totalpop2015	2.4481E-07

Table 28b. The variables used in the Ridge Regression model in descending order of the coefficients.

Tables 28a and 28b show the ranked order of the variables used in the ridge regression model by magnitude. The most important variable to the ridge regression model is the current annual rate of reduction in childhood stunting. Variables around childhood stunting have appeared in all of the models. This indicates that this metric is important when understanding diabetes prevalence on a grand scale. The second most important variable to the ridge regression model is the gender inequality index. This is the first time this variable has appeared as highly significant among the models. Two variables that also appeared significant in other models are

the women of reproductive age that are anemic, the annual average rate of reduction in under 5 overweight population, hypertension policy, diabetes policy, and Fortification. Other variables such as blood glucose levels are high in importance in the ridge regression models, but other variables commonly associated with Type 2 Diabetes Mellitus such as obesity are not ranked within the top 30 important variables in the Ridge Regression model

LASSO REGRESSION

The difference between the LASSO or Least Absolute Shrinkage and Selection Operator method and typical methods of linear regression lies in the cost function. This is typically expressed as the difference between the actual values and the predicted values. The cost function of the LASSO method, which is a modification of the Ridge method differs from both linear regression and ridge regression as it has an added penalty to the cost function of the regression model. The cost function for lasso regression includes the addition of a term denoting the amount of shrinkage multiplied by the summation of the absolute value of the magnitude of the coefficients. This term is referred to as the L1-norm term:

$$L_{LASSO} = \sum_{i=1}^{n} (y_i - \sum_{j=i}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j^2|$$

In LASSO Regression, the value of lambda is the penalty coefficient Lambda is equal to zero is the result if all of the features given to the lasso function are utilized in the model, therefore making the model have the same cost function as a linear regression model. In the case where lambda is equal to infinity, then none of the features are utilized in the model, meaning as the closer lambda gets to infinity, the fewer features are utilized until there are no features used in the final model.

One of the main advantages of LASSO Regression, when compared to Ridge Regression, is that it is specially tuned to be better to work with models with a higher number of independent variables and tuning the model to utilize a smaller number of variables. Similarly to Ridge Regression, the most optimal value for lambda is determined first when beginning to model using LASSO. That value for the optimal lambda is then used to train the model. The optimal lambda is the value of the lambda that will give the most regularized model such that the error of

the model is within one standard error of the minimum. In this study, the optimal value of lambda is 0.1769418

LASSO REGRESSION: MODEL RESULTS

	R^2	Train RMSE	Test RMSE
LASSO Regression	.7131	2.54713	3.14998

Table 29. The model statistics for the LASSO Regression model.

The R^2 value of the LASSO model is .71315 which suggests a good model for the data. The RMSE of the training data of the LASSO model was 2.547 and the test RMSE of the LASSO model was 3.14 . The difference between the RMSE of the training data and the test data is relatively low, with the testing RMSE being just double that of the training RMSE.

Clusters	Training RMSE
1	0.5987938
2	1.9656651
3	2.7811156
4	2.0324489
5	2.6349149
6	3.5291180
7	2.6528456

Table 30. LASSO training RMSE by the 7 clusters.

Clusters	Testing RMSE
1	0.1987977
2	2.7948373
3	2.2673580
4	1.0271966
5	3.6803408
6	2.6941756
7	1.3637687

Table 31. LASSO Model testing RMSE by the 7 clusters.

Table 30 and Table 31show the training RMSE by the clusters and the testing RMSE of the clusters respectively. The range of the RMSE of the testing set among the clusters is within a similar range. Interestingly enough, for some clusters it actually showed that the testing set had a better RMSE than the training set. This shows that although the training RMSE is larger when compared to the other other models, no singular cluster is disproportionately penalized by the lasso model.

Variable Name	Coefficients
current_aarr_stunting	1.06824056
BGboth	0.70615031
Fortification	0.31042692
hypertension_policy	0.27243349
LBW	0.17879129
prev_sev_wasting	0.17293819
stunting_progress	0.16622401
tothealth_ppp2010	0.11324229
subregionUN	0.1091124
valuenurse	0.07212831
valuephysician	0.05974097
prevalence_vita	0.05357513
BPboth	0.05182189
ob_bothsexes	0.04931972
water_unimprov2015	0.04804693
WRAanaemia_RATE	0.03495542
prev_wasting	0.02867591
value_gini	0.02700388
early_BF	0.0225432
EBF_trend_2	0.01726662
ORS	0.01474664
value_undernourishment2000	0.0139656
cholesterol_BOTH	0.01223065
DTP3	0.00913462
over65pop	0.00466371
value_ebf_current	0.00137378
regionUN	0.00055138
fruitandveg_gram2011	0.00049564
undernutritionrank_1to126	0.00040575
GDP_5	2.3866E-05

Table 32. All of the variables used in the LASSO Regression model

Similar to the Ridge Regression model, the most important variable to this model is the current annual rate of reduction in childhood stunting. The second most important variable of this model is the percentage of the population with elevated blood glucose levels. This model is similar to the other models as it shows that the most important variables are a combination of variables that are already associated with Type 2 Diabetes Mellitus and other variables such as Fortification stunting, and subregion that are not commonly known as Type 2 Diabetes risk factors.

RESULTS

VARIABLE IMPORTANCE

Outside of the important variables that are specific to each of the individual models, in order to assess overall variable importance, it is important to understand the frequency of how the variables appear among all of the models. With the exception of the ridge regression model (since the ridge model uses all of the variables with various levels of penalization) the number of times each of the variables appears in the models is provided below.

Variable Name	Model Occurrences
BGboth	5
fruitandveg_gram2011	5
cholesterol_BOTH	4
early_BF	4
LBW	4
ob_bothsexes	4
prev_u5overweight	4
prevalence_vita	4
reqd_aarr_stunting	4
totalpop2015	4
toteducation_ppp2010	4
urbanpop	4
subregionUN	4
BPboth	3
Class_IodineNutrition	3
Code_breastfeeding	3
early_childbearing	3
EBF_trend_2	3
Fortification	3
number_u5overweight	3
tothealth_ppp2010	3
UnmetneedTotal	3
value_gini	3
value_undernourishment2000	3

Table 33a: The first 25 variables and the number of model occurrences.

Variable Name	Model Occurrences
current_aarr_u5overweight	2
GDP_5	2
gender_inequality_index	2
hypertension_policy	2
over65pop	2
prev_wasting	2
totsp_ppp2010	2
u5overweight_progress	2
value_ebf_current	2
value_UNICEF_skilled_att	2
water_piped2015	2
water_surface2015	2
water_unimprov2015	2
current_aarr_stunting	1
diabetes_policy	1
number_sev_wasting	1
ORS	1
overnutritionrank_1to116	1
poverty2_4	1
prev_sev_wasting	1
rate_femaleED_5	1
rate_stuntingtrend3	1
rate_stuntingtrend4	1
rate_stuntingtrend5	1

Table 33b. The second 25 variables and the number of model occurrences.

Variable Name	Model Occurrences
san_opendef2015	1
san_othunimprov2015	1
san_shared2015	1
stunting_progress	1
totag_ppp2010	1
u5mr2011	1
undernutritionrank_1to126	1
value_UNICEF_ANC_4	1
valuenurse	1
water_othimprov2015	1
WHAtarget_WRAaemia1	1
WRAaemia_NUMBER	0
continent	0
continued_BF	0
DTP3	0
regionUN	0
reqd_aarr_u5overweight	0
RTF_level1	0
san_improved2015	0
value_nonstaples2009	0
valuephysician	0
wasting_progress	0
WRAaemia_RATE	0

Table 33c. The last 25 variables and the number of model occurrences.

There are only two variables that appear in every single model. These variables are the amount of the adult population with elevated blood glucose levels, and the availability of the fruit and vegetables in grams in the country. Both of these variables have been associated with Type 2 Diabetes Mellitus in previous studies, so it is expected that these variables would be commonly seen as significant to the models. The next group are the variables that appeared in at least 4 of the 5 models. These variables include the percentage of the adult population with

raised cholesterol, the percentage of the under five population that is overweight, the low birth rate, and the subregion. In this group of variables in which they appeared in at least four of the five variables there are 11 variables. The next group of variables are those variables that were found in at least three of the models. This group of variables includes the percentage of the adult population with elevated blood pressure, fortification of food, the percentage of the children under 5 that are overweight, births to mothers under 18, etc. There are also a total of 11 variables in this group. These are all the variables that appeared in at least half of the models that were utilized, by this metric.

The variables that are within the significant threshold of appearing in at least half of the models can be grouped in ways to offer more intuitive knowledge on the subject and its influence on diabetes prevalence. For example, there are the variables that can be categorized as negative health indicators. These are the variables that show the percentage of the population that have abnormal levels of common health factors such as elevated blood pressure, blood glucose levels, cholesterol, adult obesity, vitamin A deficiency in school aged children ,childhood overweight percentage. Another grouping of the variables are population variables. This grouping includes variables explaining the total population of the country and the urban population. A category that is not commonly thought of high importance for Type 2 Diabetes Mellitus risk is the variables around childbearing mothers. These variables include the percentage of early initiation of breastfeeding to newborn babies, low birth rates, stunting, early childbearning mothers, trends in exclusive breastfeeding, and the percentage of unmet needs for family planning. Another category is one that involves more government intervention such as the country's classification of iodine nutrition, the country policy on the fortification of foods, Natural Implementation of the International Code of Marketing of Breast Milk Substitutes, and the gini index.

It is also interesting to look at the variables that were not utilized by any of the models. One interesting fact is that region was not used in any of the models, whiler the variable for subregion was used in four out of the 5 models. Similarly the variable for continent does not appear in any of the models. The percentage of the population that continued breastfeeding at the 1 year mark. The extent of the constitutional right to food in the country, the rate at which children under 1 are vaccinated by the trio of vaccinations known as the dtp3 (diphtheria,whooping cough, and tetanus), and the population density of healthcare workers per 1000 people within the population.

MODEL SELECTION

Model selection is the secondary portion of this study to decide which modelling method performed best for the data. The final dataset used in modelling is unconventional in that the number of covariates is about half the number of observations. For a lot of modelling techniques, this ratio presents a problem and can lead to issues such as bias and overfitting. Therefore understanding which model was able to best fit this data will show what model produced the best results for the unique dataset and can be utilized for future studies using similar datasets in which the observations represent countries.

In order to evaluate the best model, the test set that was set aside when stratified sampling is used. When going through variable selection, the significant variables to the models that are trained on the training set are used to draw conclusions on the variable importance. For model selection, the model that is the best is the model that is able to fit a testing set (a set of observations of the data that is not used in the training of the model), and produce the best accuracy metric for prediction. In this case the testing RMSE will be compared across the models.

The testing RMSE of the 5 models that were run is given in figure Table 34 shows the breakdown of the testing RMSE of all of the models by cluster. The cluster analysis is an important part of the model selection because it is ideal to make sure that no one cluster is

	Stepwise Selection	Backwards Selection	Random Forest	Gradient Boosted Model	Ridge Regression	LASSO Regression
1	7.065779	6.505197	0.579715	0.6387933	1.865288	0.1987977
2	5.108778	7.636887	1.876098	1.9177054	2.372861	2.7948373
3	5.924006	7.47423	1.831983	2.1303849	2.38344	2.7811156
4	6.096839	2.937232	0.389296	2.053028	2.534459	1.0271966
5	5.423074	8.818825	3.171361	3.4624484	3.565849	3.6803408
6	5.694152	4.093466	2.060554	2.8344739	3.258759	2.6941756
7	11.990259	20.083623	2.419882	1.0848069	1.384905	1.3637687

Table 34. The testing RMSE for all 6 models across all 7 clusters.

drastically misrepresented by the model. For instance, the range of clusters 1 and 2 are 6.8814 and 17.1463 respectively. The stepwise selection model both have much larger ranges compared to the other four models, and both disproportionately represent cluster 7 and cluster 5. The range of the random forest testing RMSE is 2.7820 and the rande of the Gradient Boosted Model is 2.8236. These ranges are drastically smaller than the ranges of the clusters for the stepwise methods, suggesting that no clusters are as heavily favored by the model using these methods. Ridge Regression testing RMSE range is 2.18094 and LASSO Regression test RMSE has a range of 3.4815. Ridge Regression has the smallest range among the clusters followed by the Random Forest model. These two models are good candidates for the best model for the data.

The RMSE of these models will be used as the baseline for comparison when applying other methods of model selection. Evaluating the testing RMSE among the existing models shows that the random forest model performed the best in terms of the testing RMSE. It has the smallest RMSE compared to the 5 other models, with the GBM model being a close second. This shows that the ensemble methods both worked better than the other methods for this dataset. We can also see that among the clusters, the results of the stepwise model and the backwards regression models both showed bad predictive power compared to the other clusters for that model. Looking at the random forest model, this is not shown to be the case.

Step Regression	Backwards Regression	Random Forest	Gradient Boosted Models	Ridge Regression	LASSO Regression
5.951202	8.960152	2.653056	2.88907	3.084565	3.149981

Table 35. Testing RMSE for the six models.

In order to be thorough in the model selection process, it is important to evaluate whether the models would produce more optimized accuracy metrics if combined. The process of

combining models and model predictions in order to produce an optimized model that takes in the weights of better performing models into account is called Model Averaging. Model averaging is a way to utilise the benefits and the individual predictive power of multiple models to make predictions. The model averaging methods are ensemble models that contain more than one model and averages the predictions from the models included in the ensemble model. Using the testing RMSE as the metric of interest means that the model can either be predictive above or below the actual value. Therefore for different data points the model can either be giving an overestimate or an underestimate. Applying model averaging techniques across all of the models can minimize the amount of overestimating and underestimating that is done by both models and creates a flatter prediction line.

In this study, the metric of interest is the RMSE for understanding how well the final model averaging methods did against the original models. The initial assessments of the models showed that the stepwise and backwards selection methods did not perform favoribel with the dataset. And after initial assessment of model averaging methods including these models showed that the inclusion produced subpar results. Therefore only four of the original six models, the random forest model, gradient boosting model, ridge regression model, and lasso regression models are used as part of the model averaging.. Both the stepwise regression and backwards regression were key models in terms of variable selection but performed drastically lower in terms of RMSE when compared to the other models. For these reasons, the backwards and stepwise models are not utilised in the model selection portion.

There are three weighted methods of model averaging employed in this study. The first method is an equal weighted averaging method. In this method, each of the models have equal

weight in the final ensemble models. The arithmetic mean of the model predictions is used making it one of the simplest model averaging methods to implement.

$m_j = 1/C$where m_j is the j^{th}model and C is the total number of models used in the averaging scheme. This equation represents the simplistic model averaging scheme in which the weights of all the models are equal.

The equal weights model averaging method takes the arithmetic mean of the predictions from the four models, then divides by the total number of models, in this case four.

$$m_j = (random.forest_{predicted} + gradient.boosting_{predicted} + ridge.regression_{predicted} + LASSO.regression_{predicted})/4$$

The second model averaging method is a fit-based model averaging. In this method, the models are weighted based on a performance metric that is used to evaluate the individual models. In the case of this study, the root mean squared error (RMSE) is used to evaluate the fit based weights. To implement this method, first the training RMSE is calculated. These are going to be the weights of the models. Since this study uses the RMSE, the reciprocal of these weights is used, since a smaller RMSE is indicative of a better fit and therefore should have additional weight. The last step is to use the weights and obtain the proportion of the sum of weights that each model represents. These proportions are then multiplied by all of the predicted values of the model and summed across all of the predicted values in order to make the new predictions.

The first step is to establish the weights. For this the training RMSE from the original models are used to train the fit based averaged model. The reciprocals ensure that the models that are getting the highest rate are those that have the lowest root means squared errors. Therefore the final predicted values are calculated as

$N_l = \sum_{j,k=1}^{4} (m_j * p_k)\ where\ k = j\ and\ l = 1,\ldots, 132$ where 132 is the number of observations in the training set and $N_l$represents the predicted value for each observation, $m_j$is the model predicted diabetes prevalence and $p_k$is the model proportion and determined by the reciprocal of the RMSE divided by 4. $k = j = 1, 2, 3, 4$represents each of the models and their corresponding proportions (1: random forest, 2: gradient boosted model, 3 ridge regression, 4, lasso regression).

The final model averaging method is called model based combination averaging. The general idea for this method is to use the model outputs as inputs for a new regression model fit against the actuals. The predicted values of all of the variables are used as the dependent variables in a linear regression model for the original dependent variable for the model (based on the training set). Using the predicted values as inputs and the final model is

$combination.averaged_{model =} - 0.25896204 + 1.31998361m_1 - 0.48754646m_2 + 0.31786327m_2$

$- 0.29511742m_4 + 0.11416986m_1m_2 + 0.13652314m_1m_3 - 0.23836036m_2m_3 + - 0.1$

$+ 0.16228503m_2m_4 + 0.03314519m_3m_4 + - 0.00026001m_1m_2m_3 + - 0.00317821m_1m$

$0.00228532m_1m_3m_4 + 0.00242642m_2m_3m_4 - 0.00007503m_1m_2m_3m_4$ *adjust notation

*table must be typed- no jpeg

Country	Random Forest Prediction	GBM Predictions	Ridge Regression Predictions	LASSO Predictions	Diabetes Prevalence Actuals	Equal Weights Predictions	Fit Based Predictions	Model Based Predictions
Albania	9.563085	8.540498	9.3465717	8.921586	10.08	9.092935	9.172004	10.073333
Algeria	7.65029	7.719428	8.4282812	7.403422	6.73	7.800355	7.794063	7.381171
Andorra	8.973862	8.987624	9.5603853	8.438932	7.97	8.990201	9.005355	8.985873
Angola	4.653017	4.722405	5.3894038	5.301529	3.94	5.016589	4.964011	3.982701
Antigua and Barbuda	12.602612	14.284953	12.0967875	11.256096	13.17	12.560112	12.57487	12.882416
Armenia	7.881113	8.735575	7.3447029	8.428751	7.11	8.097535	8.04611	7.394302
Australia	6.907351	7.135154	8.0295656	8.174952	5.07	7.561756	7.462064	6.250914
Austria	6.264114	5.642522	5.4389327	4.82657	6.35	5.543035	5.661024	6.40834
Azerbaijan	8.226846	9.486572	8.9274463	9.134335	7.11	8.9438	8.830994	7.480511
Bahamas	13.052016	15.137538	12.1226963	12.418945	13.17	13.182799	13.152948	13.486175
Belarus	5.943494	5.389507	7.3407746	8.677463	5.18	6.83781	6.686133	5.169592
Belgium	5.537085	6.223373	5.2521531	5.58719	4.29	5.64995	5.626072	4.818517
Bhutan	9.430078	9.398917	8.6312482	8.535621	9.75	8.998966	9.064305	9.681637
Bolivia	6.962656	6.236447	6.9474715	7.34688	6.89	6.873364	6.882337	6.654017
Bosnia and Herzegovina	9.354523	7.342757	6.9705354	7.552334	10.08	7.805037	8.030682	10.04124

Table 36. The chart shows the predictions of the random forest model, the gbm model, the ridge regression model, the LASSO model, the actuals for diabetes prevalence, and the predictions for all three of the model averaging methods: equal weight predictions, fit based predictions, and model based predictions

The final ensemble averaged models are then fit to the testing set to understand the testing RMSE of each of the new models. Table 36 shows a subset of the diabetes prevalence predictions for each of the models used in model averaging, the actual diabetes prevalence in column 5, and the prediction results for all three of the model averaging methods in columns 6-8.

Random Forest	Gradient Boosted Model	Ridge Regression Model	LASSO Regression Model	Equal Weights RMSE	Fit Based RMSE	Model Comparative RMSE
2.653056	2.88907	3.084565	3.149981	2.799276	2.758534	2.692041

Table 37. The testing RMSE of the four models used in model averaging and the three model averaging methods.

Comparing the testing RMSE of all the models shows that among the original 4 models the random forest has the lowest testing RMSE at 2.653. The best RMSE of the model averaging methods is the model combining method which yields a RMSE of 2.692041. Table 37 shows that the model averaging RMSE performed similarly to the RMSE of the random forest model and

the gradient boosted model. But the RMSE of the random forest still yielded the best results and is therefore the best model to fit the data.

The results of model selection parameters show that the Random Forest was the best model fit for the data as it has the smallest RMSE of all of the models and the second lowest RMSE of the training set. The random forest also performed well across all of the clusters without over emphasizing any particular cluster.

CONCLUSION

. The CDC projects that the number of Americans living with Type 2 Diabetes Mellitus is set to triple by the year 2050 (CDC 2019). This statistic, like many of the previous research studies done on the risk factors of diabetes prevalence focuses on the increases in the number of people developing the disease in the West, and therefore focusing these risk factors on variables that are western centric. The goal of this study was to see whether these risk factors as determined in either small regional cohorts held true for regions outside of the West and if other variables that are traditionally not examined in the West held significance. The use of six different machine learning models to find a best model fit and then the evaluation of the variable importance of the predictors used in this model showed that this hypothesis held. Some risk factors such as obesity, blood glucose, blood pressure remained to be significant variables in predicting diabetes prevalence, mirroring what is already found in the literature. While other variables that are traditionally not included in diabetes research like population, vitamin A deficiency, low birth rates, breastfeeding policy ,and the rates of stunted children were also shown to be highly significant when it comes to diabetes prevalence. Some of the variables identified by the models such as the variable around raised blood glucose levels have been commonly associated with diabetes prevalence. But other variables such as vitamin A deficiency in children, stunting rates, the amount of medical care available to pregnant mothers, and low birth rates in a country are variables that are specific to the child population and reproductive health. The emphasis on these variables sheds light on the significance of ensuring the health of people from pregnancy through infancy is essential in mitigating the increase in diabetes prevalence rates. It is projected that by the year 2045, 10.9% of the world's population will have contracted Type 2 Diabetes Mellitus (Saeedi 2019). The mitigation of the continued rise in Type

2 Diabetes Mellitus is a global issue, and having results that illuminate the different ways these risk factors influence the broader world can help programs and initiatives focus on reducing this increasing trend.

STUDY LIMITATIONS

The clustering approach of this study is meant to ensure that the data is representative of the different ways regional, cultural, social and economic differences between the observations have influence on health indicators. These social indicators, specifically when it comes to population measures, were especially stark when the K-Means clustering algorithm grouped high population countries like India and China into a single group. After evaluating the penalized regression models, it was shown that the models were able to predict with the training and the test data set considerably well with the exception of these high populations categories. This result of the study could illuminate the penalized regression measures of Ridge and LASSO regression are a better fit for the data if these population anomalies were not included in the test and training set. This relates back to the original goal of the study because it shows that for certain regions of the world that are especially anomalous of these social and health indicators, studies into risk factors are best done on a smaller regional scale than through aggregate studies like the ones employed in this dataset.

Since there are a finite number of countries in the world, the sample size for this study is stagnant as well as relatively small. In order to see how the results of this study hold, a similar method may be employed but with equal or stratified sampling of patients and their individual health and social markets from most if not all of the countries in the world. This could provide a

way to see if the aggregate data holds on the individual level and how significant the social and economic markers are on the individual probability of developing Type 2 Diabetes Mellitus.

References

An Approach for Selecting Optimal Initial Centroids to Enhance the Performance of K-means. (2013). Conference: International Conference on Advances in Computer and Information Technology. https://www.researchgate.net/publication/331429949_An_Approach_for_Selecting_Optimal_Initial_Centroids_to_Enhance_the_Performance_of_K-means

Breiman, L. (2001) Statistical modeling: the two cultures. *Statistical Science*, 16, 199–215

Brock, V. Pihur, S. Datta, and S. Datta. clValid: An R package for cluster validation. Journal of Statistical Software, 25(4), March 2008. URL http:// www.jstatsoft.org/v25/i04.

CDC Online Newsroom - Press Release - Number of Americans with Diabetes Projected to Double or Triple by 2050. (2010). Number of Americans with Diabetes Projected to Double or Triple by 2050. https://www.cdc.gov/media/pressrel/2010/r101022.html

Concern Worldwide US. (2019). Stunting: What it is and what it means | Concern Worldwide U.S. Concern Worldwide. https://www.concernusa.org/story/what-is-stunting/

Cran MiceRanger. (2021, May 12). Cran R. https://www.who.int/nutrition/topics/globaltargets_wasting_policybrief.pdf

Diabetes Facts & figures. (2020). International Diabetes Federation. https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

Diabetes Increases in Children and Teens. (2017, September 8). NIH News in Health. https://newsinhealth.nih.gov/2017/06/diabetes-increases-children-teens

Ensemble methods: bagging and random forests. (2017, September 29). Nature Methods. https://www.nature.com/articles/nmeth.4438?error=cookies_not_supported&code=888184c9-a8e2-4084-98f0-f5a2811f8f79

Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. (2019, November 1). ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0168822719312306

Global Nutrition Report. (2015). Global Nutrition Report. https://globalnutritionreport.org/about/

https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf. (2020). NCSS Statistics. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf

Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine.." Ann. Statist. 29 (5)1189 - 1232, October 2001. https://doi.org/10.1214/aos/1013203451

Kopitar, L. (2020, July 20). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports. https://www.nature.com/articles/s41598-020-68771-z?error=cookies_not_supported&code=b4e351bf-de4d-44fd-87d4-684c02890db9

Lai, H. (2019, October 15). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders. https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6

Nguyen, C. T. (2015). Prevalence of and Risk Factors for Type 2 Diabetes Mellitus in Vietnam: A Systematic Review. PubMed National Library of Medicine. https://pubmed.ncbi.nlm.nih.gov/26187848/

Population by Country (2021) - Worldometer. (2019). Worldometer. https://www.worldometers.info/world-population/population-by-country/

Ramachandran, A. (1988, September 3). High prevalence of diabetes in an urban population in south India. PubMed. https://pubmed.ncbi.nlm.nih.gov/3139221/

Theoretical Comparison between the Gini Index and Information Gain Criteria. (2004). Annals of Mathematics And Artificial Intelligence. https://www.researchgate.net/publication/226475848_Theoretical_Comparison_between_the_Gini_Index_and_Information_Gain_Criteria

Type 2 Diabetes. (2020). International Diabetes Federation. https://www.idf.org/aboutdiabetes/type-2-diabetes.html

Unimproved Drinking Water Sources. (2020). SSWM - Find Tools for Sustainable Sanitation and Water Management! https://sswm.info/content/unimproved-drinking-water-sources