A COMPARISON OF MACHINE LEARNING APPROACHES AND TRADITIONAL STATISTICAL ANALYSIS METHODS USING SYNTHETIC ELECTRONIC HEALTH RECORDS DATA

APRIL 20, 2024

Submitted by:
Lori Krammer
lorikrammer@gwu.edu

Raja Mazumder
The George Washington University
mazumder@gwu.edu

Faculty Advisor
Scott Quinlan
squinlan@gwu.edu

In Partial Fulfillment of the Requirements
For the Masters of Public Health Degree
Department of Epidemiology and Biostatistics
The George Washington University
School of Public Health and Health Services

Abstract

Background/Objectives

Traditional statistical methods (TSM) are widely used in epidemiologic research. Novel machine learning (ML) approaches have the potential to expand clinical understanding and further facilitate complex statistical analysis. This project aims to compare the accuracy, strengths, and limitations of traditional statistical methods and machine learning approaches for predictive modeling of clinical intervention outcomes in a synthetic electronic health records (EHR) dataset based on United States veterans.

Methods

The data was cleaned and harmonized, descriptive statistics were generated and assessed, imputation was performed, and multiple models were trained and validated. The machine learning models included support vector machine, decision tree classifier, boosted trees, and random forest. These were compared against logistic regression, a well-known traditional statistical algorithm for data analysis.

Results

Models performed comparatively, with subtle differences in accuracy, precision, and recall. The random forest and boosted tree models performed marginally better than logistic regression and decision tree classifier models. Findings for this project indicate that both ML and TSM models perform similarly, with regard to the size and scope of the dataset used.

Discussion

This work expands on existing knowledge regarding the use of ML in epidemiologic and public health research. Despite their limitations, predictive models have significant potential to improve clinical decision-making in the context of precision medicine. The closeness of model performance may not be present when large datasets are used.

Introduction

Epidemiologic research primarily relies on traditional statistical methods (TSM) to adequately assess causal relationships between variables of interest.[1] In recent years, epidemiologists have begun to incorporate an application of artificial intelligence (AI), known as machine learning (ML), into their analyses.[2] With the new age of ML and AI, these sophisticated algorithmic models hold great potential for improving not only our understanding of causal relationships, but also common practices in epidemiologic research. Machine learning algorithms, as the name implies, are designed to continuously learn and improve based on the data provided to them.[3,4] Several types of ML exist, perhaps the most common of which are supervised and unsupervised learning. Supervised learning models are given a training dataset with clear parameters. The ML techniques used in the scope of this project are supervised learning models. Unsupervised learning is a type of ML where clear parameters are not provided, and the algorithm identifies associations and naturally occurring patterns. Reinforcement and Deep Learning are other types of Machine Learning.[3]

ML techniques have recently been employed to predict mortality outcomes related to peripheral artery disease, non-metastatic prostate cancer, and heart failure with preserved ejection fraction.[5-7] Additionally, various types of ML algorithms—Including support vector machine, decision tree, and random forest—have been employed at various stages of the drug discovery process and have shown promise in improving current clinical trial methodology.[4] The use of ML in other sectors of public health research and practice, such as policy and surveillance, is widespread.[8,9] To increase understanding of the commonalities, differences, and potential of both statistical approaches, we will conduct a comparative analysis of multiple ML model types and binary logistic regression.

Prediabetes, a condition characterized by impaired glucose levels, is of growing global concern.[10] It is commonly known as a precursor to type 2 diabetes mellitus (T2DM). Despite substantial research into T2DM, additional research into pre-diabetic conditions is needed to understand the multifactorial etiology of the disease and its progression.[11] Though a consensus has not been reached, diet-based interventions for the management of diabetes have been explored.[12,13] This project aims to utilize data from patients with prediabetes, which can expand upon existing research into associations between diet interventions and the pre-diabetic state.

Specific Aim 1: To compare machine learning approaches and traditional statistical analysis methods using synthetic EHR data for the prediction of a five percent minimum reduction in weight following a 'dietary counseling and surveillance' intervention for individuals at risk of progression to diabetes from prediabetes. Binary logistic regression and random forest, decision tree classifier, support vector machine, and boosted trees machine learning models will be trained on the synthetic dataset and pre-specified outcome metrics will be compared. The models aim to accurately predict a 5% reduction in weight or, conversely, a weight gain or reduction of less than 5%, for individuals who underwent a 'dietary counseling and surveillance' intervention. It is hypothesized that ML approaches will outperform logistic regression in the primary metrics used for comparison - accuracy and the area under the receiver operating characteristic (AUROC) curve.

Specific Aim 2: To produce a functional and highly performing model for single-patient predictions of a five percent minimum reduction in weight following a 'dietary counseling and surveillance intervention. The highest-performing model is planned to be integrated into the PredictMod platform for clinical use. Accuracy will be measured by the fewest numbers of false positive and false negative predictions when the model is tested, as illustrated by a confusion matrix.

Methods

A comparison of ML and TSM was conducted using synthetic data made available from the MDClone platform (www.mdclone.com). This data is representative of real veteran patient data gathered from Veterans Affairs (VA) medical centers across the United States.[14] The synthetic dataset is comprised of individuals at least 18 years old with prediabetes and a body mass index (BMI) of greater than or equal to 30 who underwent a ‘dietary counseling and surveillance’ intervention. The outcome of interest is responder status (responder(R)/non-responder (NR)) which is defined as a 5% reduction in weight within the 6-month timeframe following the initiation of the ‘dietary counseling and surveillance’ intervention. This intervention is associated with ICD-10 code Z71.3.

The MDClone dataset consists of 19,902 observations and 27 variables, which are listed in Table A1 of the appendix. A query to generate the synthetic data was created in MDClone and the data was extracted from the platform and then cleaned in preparation for analysis. For many of the variables in MDClone, data points are not readily available when building the query using the associated Logical Observation Identifiers Names and Codes (LOINC) values. For example, a LOINC search for Hemoglobin A1c (HbA1c) would use the code 4548-4, but many HbA1c levels are associated with the keyword “HbA1c” or “Hemoglobin A1c” instead. So, to obtain a large quantity of high-quality observations for each patient, an additional variable based on an internal keyword search was generated. The data points for both columns were merged during cleaning, taking the keyword search as a priority.

Each instance of synthetic data created in MDClone is accompanied by documentation of the comparative statistics to the real MDClone patient dataset. Prior to data cleaning, the comparative statistics were examined to confirm the synthetic data is representative of the real patient data.

Statistical Analysis Software (SAS, Cary, NC) version 9.4 and Visual Studio Code version 1.87 were used for data cleaning, analysis, and model training. This process included renaming the extracted variables according to the existing PredictMod data dictionary and mapping properties. Text string values, such as race and ethnicity, were replaced with categorical integer values and converted to binary columns via one-hot encoding. The response status was calculated by comparing the weight at the pre-intervention and six-month follow-up time intervals.

Unused variables were removed from the dataset. The clean dataset is maintained as a comma separated values (CSV) document within the PredictMod team's private SharePoint site.

Once the data has been cleaned and prepared for model training, it was used to train both the ML and TSM models. The model training process was as follows:

Descriptive statistics were generated, including means and standard deviations for continuous variables as well as frequencies and percentages for categorical variables. The distribution of each variable was evaluated. The descriptive statistics were compared against the ranges expected from each data property as described in the PredictMod data dictionary. The ranges in the data dictionary are based on LabCorp values and the VA's CIPHER tool. This was done to ensure data quality and to obtain an accurate understanding of the cohort characteristics. For some variables, extreme or unrealistic values were identified as erroneously queued from the MDClone database. This was not unexpected and likely due to unassociated keywords or mistakes when the data was entered into the EHR system. To ameliorate this, values outside of the expected ranges determined in the data dictionary for calcium, cholesterol, carbon dioxide, creatinine, high-density lipoprotein, heart rate, platelet, low-density lipoprotein, protein, and sodium were dropped prior to imputation. The descriptive statistics for the dataset with the dropped outliers (dataset v2.0) are described in Table 1.

The descriptive statistics were also stratified based on response to the intervention. For dataset v2.0, these are described in Table 2. Stratified descriptive statistics for the dataset with erroneous outliers (dataset v1.0) are available in the Appendix (Table A2). Missing values were resolved via nearest neighbor (KNN) imputation. Multiple imputation was initially done, but likely due to the complexity of the data, it generated impossible, even negative values for certain variables. Nearest neighbor imputation was recommended by members of the PredictMod team and it performed well. This method of imputation replaces missing values with predictions based on neighboring rows with existing values for the missing cell. Descriptive statistics (of the full variables and stratified by status) were compared to verify that the imputation did not generate impossible values. Descriptive statistics for the dataset that included imputed values (dataset v3.0) are described in Table 3.

Model fitting was initially conducted using logistic regression, via a stepwise method at both 5% and 1% thresholds to enter/stay in the model. These fitted variables were compared to feature selection conducted using two machine learning algorithms: decision tree classifier and random forest. A summary table of model fitting techniques is shown in Table 4. Both of the decision tree classifier and random forest techniques included variables identified as multicollinear, so the logistic regression selections were used to train the fitted models.

A logistic regression (LOGR) model was compared with the following machine learning algorithms: decision tree classifier (DTC), random forest (RF), boosted tree (XGB), and support vector machine (SVM). Decision trees are an early ML method and are known for their user-friendly, tree-based classification approach.[15,16] Random forest is an ensemble learning method that creates multiple decision trees from random subsets of the data.[17] Support vector machines are skilled at classifying data using hyperplanes designated by mathematically theorized margins.[18] These ML model types have been found to be high performing and suited for prediction modeling.[7,15,19-21] Given that logistic regression is a well-known and widely used predictive modeling approach for epidemiological research, it is a good fit for comparison against ML models.[22-26] Direct and stepwise model fitting techniques were employed to ensure the highest-quality models were used in the comparison. Model training was first done for the tree-based models (DTC, RF, XGB) on the dataset with dropped outliers (v2.0). These models can handle missing data and can potentially identify important predictors that would otherwise be undetected. Additional models for all algorithms were trained on the imputed (direct fit) dataset (v3.0), and then again as threshold model fitting techniques of 5% (0.05 fit) and 1% (0.01 fit) were employed. Logistic regression assumptions of no perfect collinearity and linearity were tested prior to model training.[27] Multiplicative interaction analysis was conducted to identify potential interactions. The interactions explored were selected based on residual regression plots from the logistic regression models that indicated non-linear or complex relationships.

The models were assessed to ensure high-quality discrimination and calibration. Discrimination was determined by reviewing the C-Statistic and the AUROC curve, which have been used to successfully identify the quality of logistic regression model discrimination.[26,28,29] Calibration was assessed using the Hosmer-Lemeshow statistic and Brier score. These techniques for reviewing calibration are widely used in biostatistics research.[30,31] Prediction accuracy, and AUROC curves were the primary metrics used to compare the ML and TSM models. Confusion matrices for each model were also considered in the comparison. In addition to the model training, hyperparameter tuning techniques such as random search, grid search, cross validation, and class weights were employed to the best performing model (v2.0, v3.0 (direct), 1% fit, or 5% fit) for each algorithm before comparison.

The ML and TSM models were compared across several metrics. Accuracy indicates how many outcomes were correctly predicted. The area under the receiving operator curve (AUROC) is indicative of how well the model differentiates between responder/non-responder status. Training and testing scores indicate the accuracy of the model on the training portion of the data (75%) and the testing portion (25%). Precision is the proportion of responders that were correctly predicted out of all predicted responders. Recall, otherwise known as sensitivity, indicates the model's ability to correctly predict responders out of all actual responders. The f1-score is a metric used to balance precision and recall. Higher f1-scores are desirable when assessing model performance. Confusion matrices are another metric used to assess model accuracy. They categorize the number of predicted and actual outcomes in the training dataset into true positive (lower right), false positive (upper right), true negative (upper left), and false negative (lower left) sections.

Given the synthetic nature of the dataset, institutional review board (IRB) review was not required as part of this project. Nevertheless, the proposed project was submitted to the Student Oversight Portal for review and they determined IRB review was not necessary.

Results

Weight and race were identified as potential confounders based on a comparison of the variable distribution when stratified by responder/non-responder status. With regard to model fitting, the 5% threshold fit included weight, hematocrit, race, heart rate, blood urea nitrogen, diastolic blood pressure, chloride, carbon dioxide, sodium, cholesterol, low-density lipoprotein, and potassium. The 1% threshold fit included weight, hematocrit, race, heart rate, blood urea nitrogen, diastolic blood pressure, chloride, and carbon dioxide. Model fit diagnostics did not indicate any high-leverage outliers. Model fit statistics are based on the logistic regression model. Based on Cook's D statistics, there were several influential observations; this was likely due to the number of observations and the number of variables included in the dataset. The model fit histogram and Q-Q plot indicated a normal residual distribution. Residual regression plots for blood urea nitrogen, chloride, creatinine, fasting glucose, HbA1c, sodium, and triglycerides showed signs of heteroscedasticity. BMI, weight, and height showed significant signs of multicollinearity (variance inflation factors of 17.18, 27.98, and 12.78, respectively), which was expected. Chloride and sodium showed moderate signs of multicollinearity (variance inflation factors of 6.73 and 6.00, respectively). Multiplicative interaction analysis revealed significant interactions between weight and chloride (p-value 0.001) and weight and blood urea nitrogen (p-value 0.023). The Hosmer-Lemeshow Goodness-of-Fit test was conducted for the logistic regression model and was not significant (p-value 0.45), indicative of sufficient calibration. The brier score of 0.17 further substantiated this claim.

All models performed comparatively. RF had the among the highest test and train scores at 96% and 79%. SVM had the lowest with 69% and 70%. The LOGR model had the highest

AUROC at 56%. In terms of accuracy, the RF and XGB models performed the best with 79%. SVM performed the worst at 70%. The RF model performed the best in precision and recall with scores of 73% and 79%, respectively. XGB and LOGR both had precision scores of 71%, with XGB having a slightly higher recall (79%) compared to LOGR (78%). For f1-score, XGB and LOGR performed best at 71%, DTC and RF had 70%, and SVM had 69%. While many of these scores were quite similar, the RF and XGB models outperformed the other models in 71% of the performance metrics. A summary of the model comparison is shown in Table 5. Confusion matrices for each of the models are shown in Figure 1. The model with the highest number of true positives was SVM with 187, as compared to 4 from the RF model. The model with the highest number of false positives was SVM with 617, as compared to 4 from the RF model. This indicates that the SVM model had the highest number of both correctly and incorrectly identified responders. With regard to false negatives, the RF model performed the worst and incorrectly identified 1052 non-responders, compared to SVM's 869 incorrectly identified non-responders. On the other hand, the RF model correctly identified the most non-responders, with 3916 correct predictions. The SVM model performed poorly, correctly identifying only 3303 non-responders. Due to the nature of these interventions and the overall positive benefits associated with dietary counseling, the model's ability to correctly identify non-responders (true negatives) holds slightly more weight than the ability to identify responders (true positives). Health risks associated with withholding dietary counseling have greater potential to be damaging than those associated with recommending an improved diet. These results indicate that, with regard to the MDClone dataset in question, both ML and TSM models perform similarly when predicting intervention outcomes.

Discussion

This comparison reports on methods used for predictive modeling in an epidemiological context. The scope is a descriptive effort to obtain baseline characteristics of both ML and TSM model accuracy and understand the benefits and drawbacks of the algorithms used, as well as a functional model for single-patient predictions. The use of ML in epidemiologic research can be a highly efficient and effective tool, as long as the appropriate algorithms are used, and their parameters are well-understood. These types of models have significant potential to improve public health research as well as clinical decision-making in the realm of precision medicine.

This project incorporates the use of high-quality data prepared for model training through an established data ingestion pipeline, as well as a critical analysis of several ML and TSM models. While this analysis is not exhaustive, it is inclusive of highly relevant algorithms used for predictive modeling in the context of public health research. Quality control measures imposed on the data used and models generated strengthened the validity of the results.

Limitations for this project include the generalizability of results outside of veteran patients comprising the MDClone database, the broad interpretations of the 'dietary counseling and surveillance' intervention of interest, and limitations associated with the algorithms used. The tree-based models (DTC, RF, and XGB) are prone to overfitting and there was strong evidence of this prior to model tuning. Unlike the tree-based models, SVM and LOGR models do not handle missing data well. SVM models are computationally more exhaustive than tree-based classifiers, and hyperparameter tuning of the model did not yield improved results. While both ML and TSM models performed adequately on the existing dataset, this balance may shift when scaling up to larger datasets. The following future directions for this project are recommended: utilizing a generative adversarial network to expand on the existing synthetic data and compare model performance on a larger scale, as well as a further analysis of the features included and potential biological pathways present. Additional hyperparameter tuning may occur prior to model integration in the PredictMod tool.

Table 1. Descriptive statistics for synthetic dataset, from the MDClone database, n=19,902.

Variable	Unit	N	Mean (Std.)	Freq. (%)
Age	yrs	19,902	57.90 (12.35)	
Diastolic BP	mm[Hg]	19,238	78.42 (9.56)	
Systolic BP	mm[Hg]	19,238	130.06 (14.86)	
BMI	kg/m^2	19,902	36.89 (5.98)	
BUN	mg/dL	17,193	15.95 (7.42)	
Calcium	mg/dL	16,281	9.31 (0.49)	
Chloride	mmol/L	17,259	103.68 (3.31)	
CO2	mmol/L	15,582	26.21 (2.62)	
Creatinine	mg/dL	16,471	1.04 (0.35)	
Fasting Glucose	mg/dL	157	115.27 (33.51)	
HbA1C	%	16,329	6.21 (0.81)	
HDL	mg/dL	14,078	43.84 (12.05)	
Heart Rate	bpm	19,110	77.86 (13.89)	
Height	in	19,902	68.73 (3.60)	
Hematocrit	%	15,904	43.00 (4.36)	
LDL	mg/dL	13,903	107.70 (37.03)	
Platelet Count	10*3/uL	15,336	247.80 (62.67)	
Potassium	mmol/L	17,263	4.15 (0.41)	
Protein	g/dL	14,266	7.30 (0.54)	
Sodium	mmol/L	17,234	139.31 (2.64)	
Total Cholesterol	mg/dL	14,971	177.77 (42.64)	
Triglycerides	mg/dL	14,911	160.98 (114.34)	
Weight	lbs	19,902	247.09 (48.02)	
Sex		19,902		
Male				15,665 (78.71)
Female				4,232 (21.26)
Unknown				5 (0.03)
Race		19,902		
American Indian or Alaska Native				80 (0.40)
Asian				107 (0.54)
Black or African American				5,613 (28.20)
Native Hawaiian or Pacific Islander				90 (0.45)
White				10,406 (52.29)
Unknown				3,606 (18.12)
Ethnicity		19,902		
Hispanic or Latino				1,644 (8.26)
Not Hispanic or Latino				18,258 (91.74)
Smoking Status		19,902		
Yes				9,336 (46.91)
No				10,566 (53.09)
Status		19,902		
Responder				4,297 (21.59)
Non-Responder				15,605 (78.41)

Abbreviations: N=Sample Size, Std=Standard Deviation, Freq=Frequency, %=Percentage, BP=Blood Pressure, BUN=Blood Urea Nitrogen, CO2=Carbon Dioxide, HbA1C=Hemoglobin A1C, HDL=High-Density Lipoprotein, LDL=Low-Density Lipoprotein, yrs=Years, mm[Hg]=millimeters of mercury, kg/m^2=Kilograms per meter squared, mg/dL=milligrams per deciliter, mmol/L=millimoles per liter, 10*3/uL=thousands per microliter of blood, g/dL=grams per deciliter , lbs=Pounds

Table 2. Descriptive statistics for dataset v2.0 (dropped outliers) stratified by status. Based on synthetic data from the MDClone database, n=19,902.

Variable	Unit	Responder		Non-Responder	
		Mean (Std.)	Freq. (%)	Mean (Std.)	Freq. (%)
Age	yrs	58.03 (12.61)		57.86 (12.28)	
Diastolic BP	mm[Hg]	77.98 (9.73)		78.54 (9.51)	
Systolic BP	mm[Hg]	129.79 (14.77)		130.13 (14.89)	
BMI	kg/m^2	37.44 (6.49)		36.74 (5.83)	
BUN	mg/dL	16.43 (7.63)		15.81 (7.35)	
Calcium	mg/dL	9.30 (0.50)		9.31 (0.49)	
Chloride	mmol/L	103.53 (3.45)		103.73 (3.27)	
CO2	mmol/L	26.10 (2.67)		26.25 (2.60)	
Creatinine	mg/dL	1.06 (0.41)		1.04 (0.33)	
Fasting Glucose	mg/dL	108.57 (22.85)		116.92 (35.52)	
HbA1C	%	6.22 (0.84)		6.20 (0.81)	
HDL	mg/dL	43.59 (12.66)		43.91 (11.89)	
Heart Rate	bpm	78.64 (14.24)		77.64 (13.78)	
Height	in	68.84 (3.64)		68.70 (3.59)	
Hematocrit	%	42.68 (4.62)		43.10 (4.28)	
LDL	mg/dL	106.41 (38.11)		108.06 (36.72)	
Platelet Count	10*3/uL	248.86 (64.27)		247.50 (62.21)	
Potassium	mmol/L	4.17 (0.42)		4.15 (0.41)	
Protein	g/dL	7.29 (0.55)		7.31 (0.54)	
Sodium	mmol/L	139.25 (2.77)		139.32 (2.60)	
Total Cholesterol	mg/dL	175.46 (42.86)		178.40 (42.56)	
Triglycerides	mg/dL	162.08 (110.61)		160.68 (115.34)	
Weight	lbs	251.92 (53.85)		245.76 (46.20)	
Sex					
Male			3,425 (79.71)		12,240 (78.44)
Female			872 (20.29)		3,360 (21.53)
Unknown			0 (0.00)		5 (0.03)
Race					
American Indian or Alaska Native			17 (0.40)		17 (0.40)
Asian			19 (0.44)		88 (0.56)
Black or African American			1,070 (24.90)		1,070 (24.90)
Native Hawaiian or Pacific Islander			13 (0.30)		13 (0.30)
White			2,393 (55.69)		8,013 (51.35)
Unknown			785 (18.27)		2,821 (18.08)
Ethnicity					
Hispanic or Latino			350 (8.15)		350 (8.15)
Not Hispanic or Latino			3,947 (91.85)		3,947 (91.85)
Smoking Status					
Yes			2,013 (46.85)		7,323 (46.93)
No			2,284 (53.15)		8,282 (53.07)

Abbreviations: N=Sample Size, Std=Standard Deviation, Freq=Frequency, %=Percentage, BP=Blood Pressure, BUN=Blood Urea Nitrogen, CO2=Carbon Dioxide, HbA1C=Hemoglobin A1C, HDL=High-Density Lipoprotein, LDL=Low-Density Lipoprotein, yrs=Years, mm[Hg]=millimeters of mercury, kg/m^2=Kilograms per meter squared, mg/dL=milligrams per deciliter, mmol/L=millimoles per liter, 10*3/uL=thousands per microliter of blood, g/dL=grams per deciliter, lbs=Pounds

Table 3. Descriptive statistics for dataset v3.0 (imputed dataset) stratified by status. Based on synthetic data from the MDClone database, n=19,902.

Variable	Unit	Responder		Non-Responder	
		Mean (Std.)	Freq. (%)	Mean (Std.)	Freq. (%)
Age	yrs	58.03 (12.61)		57.86 (12.28)	
Diastolic BP	mm[Hg]	78.02 (9.62)		78.56 (9.38)	
Systolic BP	mm[Hg]	129.79 (14.58)		130.14 (14.70)	
BMI	kg/m^2	37.44 (6.49)		36.74 (5.83)	
BUN	mg/dL	16.34 (7.39)		15.78 (7.02)	
Calcium	mg/dL	9.31 (0.47)		9.31 (0.45)	
Chloride	mmol/L	103.56 (3.27)		103.75 (3.09)	
CO2	mmol/L	26.16 (2.45)		26.28 (2.37)	
Creatinine	mg/dL	1.05 (0.38)		1.04 (0.30)	
Fasting Glucose	mg/dL	115.66 (14.84)		114.59 (14.48)	
HbA1C	%	6.21 (0.77)		6.18 (0.74)	
HDL	mg/dL	43.60 (11.10)		43.93 (10.59)	
Heart Rate	bpm	78.60 (14.03)		77.67 (13.57)	
Height	in	68.84 (3.64)		68.70 (3.59)	
Hematocrit	%	42.84 (4.31)		43.15 (3.95)	
LDL	mg/dL	107.64 (33.82)		108.91 (32.77)	
Platelet Count	10*3/uL	247.82 (58.83)		246.88 (56.57)	
Potassium	mmol/L	4.17 (0.40)		4.15 (0.39)	
Protein	g/dL	7.29 (0.50)		7.31 (0.48)	
Sodium	mmol/L	139.26 (2.62)		139.32 (2.45)	
Total Cholesterol	mg/dL	176.66 (38.64)		179.12 (38.60)	
Triglycerides	mg/dL	161.24 (98.22)		160.93 (102.95)	
Weight	lbs	251.92 (53.85)		245.76 (46.20)	
Sex					
Male			3,425 (79.71)		12,240 (78.44)
Female			872 (20.29)		3,360 (21.53)
Unknown			0 (0.00)		5 (0.03)
Race					
American Indian or Alaska Native			17 (0.40)		63 (0.40)
Asian			19 (0.44)		88 (0.56)
Black or African American			1,070 (24.90)		4,543 (29.11)
Native Hawaiian or Pacific Islander			13 (0.30)		77 (0.49)
White			2,393 (55.69)		8,013 (51.35)
Unknown			785 (18.27)		2,821 (18.08)
Ethnicity					
Hispanic or Latino			350 (8.15)		350 (8.15)
Not Hispanic or Latino			3,947 (91.85)		3,947 (91.85)
Smoking Status					
Yes			2,013 (46.85)		7,323 (46.93)
No			2,284 (53.15)		8,282 (53.07)

Abbreviations: N=Sample Size, Std=Standard Deviation, Freq=Frequency, %=Percentage, BP=Blood Pressure, BUN=Blood Urea Nitrogen, CO2=Carbon Dioxide, HbA1C=Hemoglobin A1C, HDL=High-Density Lipoprotein, LDL=Low-Density Lipoprotein, yrs=Years, mm[Hg]=millimeters of mercury, kg/m^2=Kilograms per meter squared, mg/dL=milligrams per deciliter, mmol/L=millimoles per liter, 10*3/uL=thousands per microliter of blood, g/dL=grams per deciliter, lbs=Pounds

Table 4. Summary of model fitting and feature selection techniques from the synthetic MDClone dataset. Both techniques were done with 5% and 1% thresholds.

LOGR Model Fitting		**DTC Model Fitting**		**RF Model Fitting**	
5%	**1%**	**5%**	**1%**	**5%**	**1%**
Weight	Weight	Age	Age	Age	Age
Hematocrit	Hematocrit	Heart Rate	BMI	Heart Rate	BMI
Race	Race	BMI	Weight	BMI	Weight
Heart Rate	Heart Rate	Weight	Platelet	Weight	Platelet
BUN	BUN	Systolic BP	Fast. Glucose	Systolic BP	Fast. Glucose
Diastolic BP	Diastolic BP	Platelet	LDL	Platelet	LDL
Chloride	Chloride	Fast. Glucose	Hematocrit	Fast. Glucose	Hematocrit
CO2	CO2	LDL	Triglycerides	LDL	Triglycerides
Sodium		Hematocrit		HDL	
T. Cholesterol		Creatinine		Hematocrit	
LDL		T. Cholesterol		T. Cholesterol	
Potassium		Triglycerides		Triglycerides	

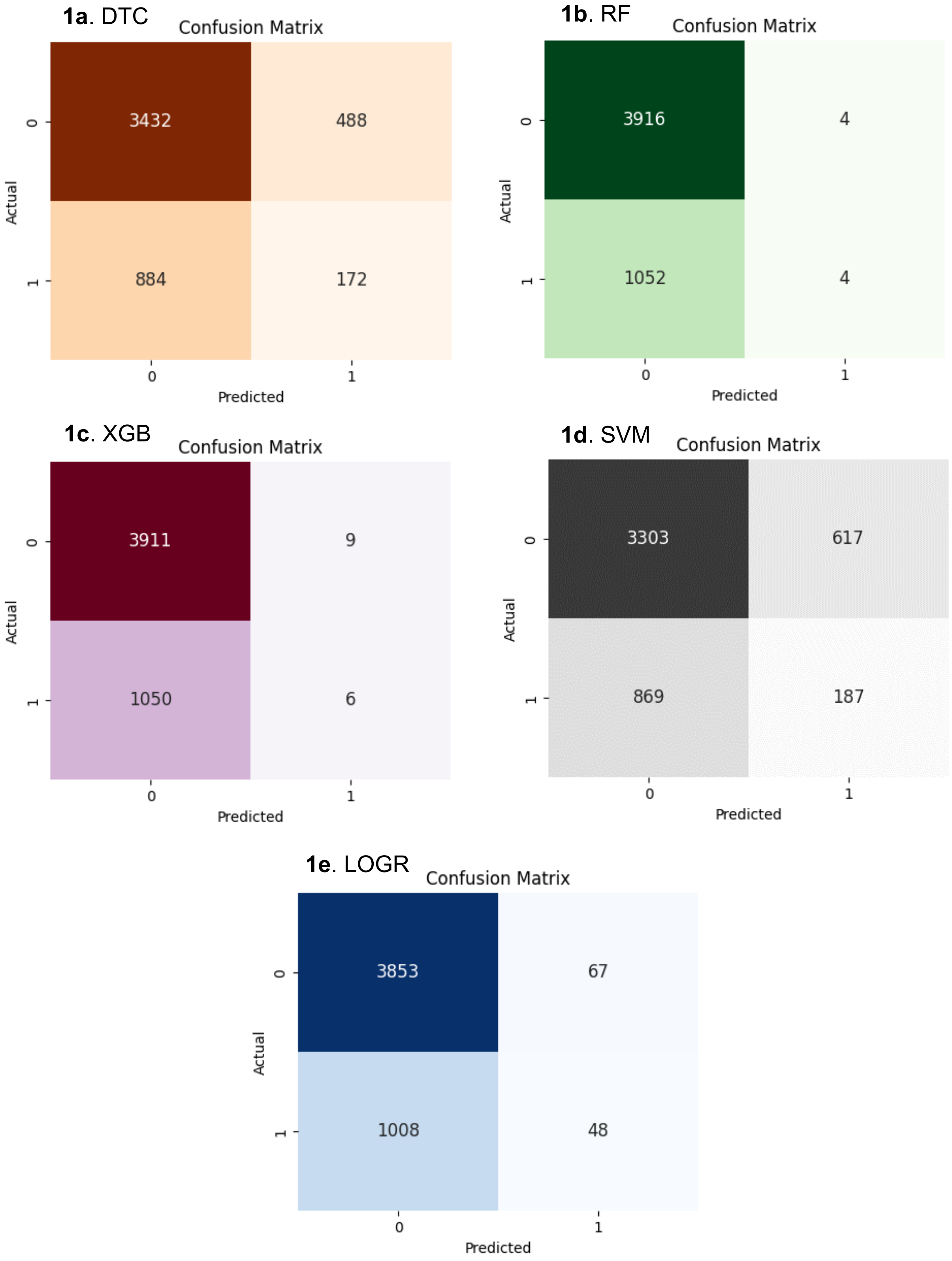
Abbreviations: LOGR=Logistic Regression, DTC=Decision Tree Classifier, RF=Random Forest, %=Percentage, BUN=Blood Urea Nitrogen, BP=Blood Pressure, CO2=Carbon Dioxide, T=Total, LDL=Low-Density Lipoprotein, BMI=Body Mass Index, Fast=Fasting

Table 5. Summary of model performance metrics. Precision, recall, and f1-score are based on weighted averages, which account for the class imbalance in status. Models were trained on synthetic data from the MDClone database.

Metric	**DTC (direct fit)**	**RF (5% fit)**	**XGB (5% fit)**	**SVM (direct fit)**	**LOGR (5% fit)**
Training score	0.76	0.96	0.79	0.69	0.78
Testing score	0.72	0.79	0.79	0.70	0.78
AUROC	0.53	0.53	0.55	0.51	0.56
Accuracy	0.72	0.79	0.79	0.70	0.78
Precision	0.68	0.73	0.71	0.67	0.71
Recall	0.72	0.79	0.79	0.70	0.78
f1-score	0.70	0.70	0.71	0.69	0.71

Abbreviations: DTC=Decision Tree Classifier, RF=Random Forest, XGB=Boosted Trees, SVM=Support Vector Machine, LOGR=Logistic Regression, AUROC=Area Under the Receiving Operator Characteristic Curve, %=percent

Figure 1. Confusion matrices for each model. Figure 1a: DTC model; Figure 1b: RF model; Figure 1c: XGB model; 1d: SVM model; 1e: LOGR model.



1a. DTC

1b. RF

1c. XGB

1d. SVM

1e. LOGR

References

1. Matranga D, Bono F, Maniscalco L. Statistical Advances in Epidemiology and Public Health. Int J Environ Res Public Health 2021;18(7). DOI: 10.3390/ijerph18073549.
2. Hamilton AJ, Strauss AT, Martinez DA, et al. Machine learning and artificial intelligence: applications in healthcare epidemiology. Antimicrob Steward Healthc Epidemiol 2021;1(1):e28. DOI: 10.1017/ash.2021.192.
3. Jayatilake S, Ganegoda GU. Involvement of Machine Learning Tools in Healthcare Decision Making. J Healthc Eng 2021;2021:6679512. DOI: 10.1155/2021/6679512.
4. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 2021;25(3):1315-1360. DOI: 10.1007/s11030-021-10217-3.
5. Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. J Vasc Surg 2016;64(5):1515-1522 e3. DOI: 10.1016/j.jvs.2016.04.026.
6. Lee C, Light A, Alaa A, Thurtle D, van der Schaar M, Gnanapragasam VJ. Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database. Lancet Digit Health 2021;3(3):e158-e165. DOI: 10.1016/S2589-7500(20)30314-9.
7. Angraal S, Mortazavi BJ, Gupta A, et al. Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction. JACC Heart Fail 2020;8(1):12-21. DOI: 10.1016/j.jchf.2019.06.013.
8. Ramezani M, Takian A, Bakhtiari A, Rabiee HR, Ghazanfari S, Mostafavi H. The application of artificial intelligence in health policy: a scoping review. BMC Health Serv Res 2023;23(1):1416. DOI: 10.1186/s12913-023-10462-2.
9. Wu WT, Li YJ, Feng AZ, et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. Mil Med Res 2021;8(1):44. DOI: 10.1186/s40779-021-00338-z.
10. Rooney MR, Fang M, Ogurtsova K, et al. Global Prevalence of Prediabetes. Diabetes Care 2023;46(7):1388-1394. DOI: 10.2337/dc22-2376.
11. Khan RMM, Chua ZJY, Tan JC, Yang Y, Liao Z, Zhao Y. From Pre-Diabetes to Diabetes: Diagnosis, Treatments and Translational Research. Medicina (Kaunas) 2019;55(9). DOI: 10.3390/medicina55090546.
12. Gardner CD, Landry MJ, Perelman D, et al. Effect of a ketogenic diet versus Mediterranean diet on glycated hemoglobin in individuals with prediabetes and type 2 diabetes mellitus: The interventional Keto-Med randomized crossover trial. Am J Clin Nutr 2022;116(3):640-652. DOI: 10.1093/ajcn/nqac154.
13. Toi PL, Anothaisintawee T, Chaikledkaew U, Briones JR, Reutrakul S, Thakkinstian A. Preventive Role of Diet Interventions and Dietary Factors in Type 2 Diabetes Mellitus: An Umbrella Review. Nutrients 2020;12(9). DOI: 10.3390/nu12092722.
14. Zautke K. MDClone partners with VHA Innovation Ecosystem to provide better, smarter, Faster Healthcare to U.S. veterans. (https://www.mdclone.com/news-press/articles/mdclone-partners-with-vha-innovation-ecosystem-to-provide-better-smarter-faster-healthcare-to-us-veterans).
15. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17. DOI: 10.1016/j.csbj.2014.11.005.
16. Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL. Review of Medical Decision Support and Machine-Learning Methods. Vet Pathol 2019;56(4):512-525. DOI: 10.1177/0300985819829524.

17. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh A. A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes. J Res Health Sci 2018;18(2):e00412. (https://www.ncbi.nlm.nih.gov/pubmed/29784893).
18. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24(12):1565-7. DOI: 10.1038/nbt1206-1565.
19. Austin AM, Ramkumar N, Gladders B, et al. Using a cohort study of diabetes and peripheral artery disease to compare logistic regression and machine learning via random forest modeling. BMC Med Res Methodol 2022;22(1):300. DOI: 10.1186/s12874-022-01774-8.
20. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015;27(2):130-5. DOI: 10.11919/j.issn.1002-0829.215044.
21. Silva GFS, Fagundes TP, Teixeira BC, Chiavegatto Filho ADP. Machine Learning for Hypertension Prediction: a Systematic Review. Curr Hypertens Rep 2022;24(11):523-533. DOI: 10.1007/s11906-022-01212-6.
22. Ogasawara K, Hiraga H, Sasaki Y, et al. A Logistic Regression Model for Predicting the Risk of Subsequent Surgery among Patients with Newly Diagnosed Crohn's Disease Using a Brute Force Method. Diagnostics (Basel) 2023;13(23). DOI: 10.3390/diagnostics13233587.
23. Zargar N, Khosravi K, Zadsirjan S, et al. The association of endodontic prognostic factors with the presence of periapical lesion, its volume, and bone characteristics in endodontically treated molars: a cross-sectional study. BMC Oral Health 2024;24(1):28. DOI: 10.1186/s12903-023-03818-x.
24. Nilsgard TL, Oiestad BE, Randsborg PH, Aroen A, Straume-Naesheim TM. Association between single leg hop tests and patient reported outcome measures and patellar instability in patients with recurrent patellar dislocations. BMJ Open Sport Exerc Med 2023;9(4):e001760. DOI: 10.1136/bmjsem-2023-001760.
25. Kasper G, Momen M, Sorice KA, et al. Effect of neighborhood and individual-level socioeconomic factors on breast cancer screening adherence in a multi-ethnic study. BMC Public Health 2024;24(1):63. DOI: 10.1186/s12889-023-17252-9.
26. Lim WXS, Seah XFV, Thoon KC, Han Z. Comparison of Vancomycin Trough-Based and 24-Hour Area Under the Curve Over Minimum Inhibitory Concentration (AUC/MIC)-Based Therapeutic Drug Monitoring in Pediatric Patients. J Pediatr Pharmacol Ther 2023;28(5):430-438. DOI: 10.5863/1551-6776-28.5.430.
27. Harris JK. Primer on binary logistic regression. Fam Med Community Health 2021;9(Suppl 1). DOI: 10.1136/fmch-2021-001290.
28. Meurer WJ, Tolles J. Logistic Regression Diagnostics: Understanding How Well a Model Predicts Outcomes. JAMA 2017;317(10):1068-1069. DOI: 10.1001/jama.2016.20441.
29. Bellavia A, Rotem RS, Dickerson AS, Hansen J, Gredal O, Weisskopf MG. The use of Logic regression in epidemiologic studies to investigate multiple binary exposures: an example of occupation history and amyotrophic lateral sclerosis. Epidemiol Methods 2020;9(1). DOI: 10.1515/em-2019-0032.
30. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Stat Med 2013;32(1):67-80. DOI: 10.1002/sim.5525.
31. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128-38. DOI: 10.1097/EDE.0b013e3181c30fb2.

Appendix

Table A1. Variables included in the MDClone dataset. All variables, with the exception of weight, were measured at the pre-intervention time interval. Weight was provided for the pre-intervention and six-month follow-up windows. Responder status was calculated during data cleaning based on the pre-intervention and six-month follow-up weight measurements.

	Property	Definition
1	Age	Age
2	Blood Pressure (diastolic)	Diastolic blood pressure
3	Blood Pressure (systolic)	Systolic blood pressure
4	BMI	Body mass index (BMI) [Ratio]
5	BUN (blood urea nitrogen)	Urea nitrogen [Mass/volume] in Blood
6	Calcium	Calcium [Mass/volume] in Blood
7	Chloride	Chloride [Moles/volume] in Blood
8	CO2 (carbon dioxide)	Carbon dioxide, total [Moles/volume] in Blood
9	Creatinine	Creatinine [Mass/volume] in Blood
10	Ethnicity	Ethnicity (Hispanic or Latino/Not Hispanic or Latino)
11	Fasting Glucose	Fasting glucose [Mass/volume] in Serum or Plasma
12	HbA1C (Hemoglobin A1C)	Hemoglobin A1c/Hemoglobin.total in Blood
13	HDL (high-density lipoprotein)	High Density Lipoprotein Cholesterol
14	Heart Rate	Heart rate
15	Height	Body height
16	Hematocrit	Hematocrit [Volume Fraction] of Blood by Automated count
17	LDL (low-density lipoprotein)	Cholesterol in LDL [Mass/volume] in Serum or Plasma
18	Platelet Count	Platelets [#/volume] in Blood by Automated count
19	Potassium	Potassium [Moles/volume] in Blood
20	Protein	Protein [Mass/volume] in Serum or Plasma
21	Race	Race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Pacific Islander, White)
22	Sex	Sex (Male/Female)
23	Smoking Status	Tobacco smoking status (Yes/No)
24	Sodium	Sodium [Moles/volume] in Blood
25	Total Cholesterol	Cholesterol [Mass/volume] in Serum or Plasma
26	Triglycerides	Triglyceride [Mass/volume] in Serum or Plasma
27	Weight	Body weight
28	Status	Responder Status (R/NR)

Table A2. Descriptive statistics for dataset v1.0 (initial dataset) stratified by status, n=19,902.

Variable	Unit	Responder		Non-Responder	
		Mean (Std.)	Freq. (%)	Mean (Std.)	Freq. (%)
Age	yrs	58.03 (12.61)		57.86 (12.28)	
Diastolic BP	mm[Hg]	77.98 (9.73)		78.54 (9.51)	
Systolic BP	mm[Hg]	129.79 (14.77)		130.13 (14.89)	
BMI	kg/m^2	37.44 (6.49)		36.74 (5.83)	
BUN	mg/dL	16.43 (7.63)		15.81 (7.35)	
Calcium	mg/dL	9.30 (0.50)		9.31 (0.49)	
Chloride	mmol/L	103.53 (3.45)		103.73 (3.27)	
CO2	mmol/L	26.10 (2.67)		26.25 (2.60)	
Creatinine	mg/dL	1.06 (0.41)		1.04 (0.33)	
Fasting Glucose	mg/dL	108.57 (22.85)		116.92 (35.52)	
HbA1C	%	6.22 (0.84)		6.20 (0.81)	
HDL	mg/dL	43.59 (12.66)		43.91 (11.89)	
Heart Rate	bpm	78.64 (14.24)		77.64 (13.78)	
Height	in	68.84 (3.64)		68.70 (3.59)	
Hematocrit	%	42.68 (4.62)		43.10 (4.28)	
LDL	mg/dL	106.41 (38.11)		108.06 (36.72)	
Platelet Count	10*3/uL	248.86 (64.27)		247.50 (62.21)	
Potassium	mmol/L	4.17 (0.42)		4.15 (0.41)	
Protein	g/dL	7.29 (0.55)		7.31 (0.54)	
Sodium	mmol/L	139.25 (2.77)		139.32 (2.60)	
Total Cholesterol	mg/dL	175.46 (42.86)		178.40 (42.56)	
Triglycerides	mg/dL	162.08		160.68 (115.34)	
Weight	lbs	251.92 (53.85)		245.76 (46.20)	
Sex					
Male			3,425 (79.71)		12,240
Female			872 (20.29)		3,360 (21.53)
Unknown			0 (0.00)		5 (0.03)
Race					
American Indian or Alaska Native			17 (0.40)		63 (0.40)
Asian			19 (0.44)		88 (0.56)
Black or African American			1,070 (24.90)		4,543 (29.11)
Native Hawaiian or Pacific Islander			13 (0.30)		77 (0.49)
White			2,393 (55.69)		8,013 (51.35)
Unknown			785 (18.27)		2,821 (18.08)
Ethnicity					
Hispanic or Latino			350 (8.15)		1,294 (8.29)
Not Hispanic or Latino			3,947 (91.85)		14,311
Smoking Status					
Yes			2,013 (46.85)		7,323 (46.93)
No			2,284 (53.15)		8,282 (53.07)

Abbreviations: N=Sample Size, Std=Standard Deviation, Freq=Frequency, %=Percentage, BP=Blood Pressure, BUN=Blood Urea Nitrogen, CO2=Carbon Dioxide, HbA1C=Hemoglobin A1C, HDL=High-Density Lipoprotein, LDL=Low-Density Lipoprotein, yrs=Years, mm[Hg]=millimeters of mercury, kg/m^2=Kilograms per meter squared, mg/dL=milligrams per deciliter, mmol/L=millimoles per liter, 10*3/uL=thousands per microliter of blood, g/dL=grams per deciliter, lbs=Pounds