Smarter Predictions, Healthier Lives: Machine Learning in Diabetes

Authors: Ashrit Wajjala, Harshith Varma, Melissa Mercer, FNU Sabha Sultana, Remy Cron

Keywords *diabetes, machine learning, hyperparameter optimization, class imbalance*

Abstract

Diabetes is a growing common public health concern in the United States, driven by largely modifiable lifestyle and dietary factors. This project investigates the application of *machine learning* methods in predicting diabetes status—non-diabetic, pre-diabetic, or diabetic—based on personal health indicators. The dataset used is the CDC Diabetes Health Indicators, which comprises more than 250,000 records and 21 attributes extracted from the Behavioral Risk Factor Surveillance System (BRFSS). Preprocessing of the data for model training involved handling missing values with mean substitution and encoding categorical features using one-hot encoding. The intrinsic class imbalance of the dataset was corrected using *SMOTENC*, an oversampling technique that considers both categorical and numerical features. Z-score scaling was applied to normalize numerical variables, achieving comparability between features. Dimensionality reduction via *Principal Component Analysis (PCA)* was performed, preserving 95% of the variance to enhance computational efficiency and minimize noise. Several classification models—*Logistic Regression, Multilayer Perceptron (MLP), Kernel SVM, and Random Forest*—were trained to address two binary classification problems: (1) distinguishing non-diabetic individuals from those at risk and (2) distinguishing pre-diabetic from diabetic individuals. Performance metrics appropriate for *imbalanced classification* problems, including precision, recall, and F1-score, were used to evaluate models. Findings indicated that nonlinear models performed better than linear models in identifying diabetes progression, although computational cost trade-offs were noted. This research illustrates the potential benefit of *machine learning* in early prediction of diabetes, which can facilitate preventive treatment and better health outcomes.

I. Introduction

Diabetes has been ever increasing as a public health issue within the United States, which is largely influenced by one's lifestyle or dietary habits. This project was presented as our final project to allow us to understand different ways of detecting diabetes predictors or helping diagnose and treat affected patients earlier.

A. Background

Diabetes mellitus is a group of metabolic diseases that cause problems in insulin release results in ongoing hyperglycemia within its affected [1]. In 2012, it was estimated that 1.5 million people died from the disease and in 2016, 9% of adults had some form of diabetes mellitus, causing the World Health Organization to declare this as the 7th leading cause of death by 2030 [1]. For this project, the dataset we are using is "CDC Diabetes Health Indicators Dataset", which contains lifestyle survey information and healthcare statistics about people, including their diabetes status, which is intended to be used to analyze diabetes and the associated risk factors that are present in the United States [2]. The dataset contains 253, 680 instances and 21 attributes, which act as indictors based on data from the Behavioral Risk Factor Surveillance System (BRFSS) (i.e: age, BMI, smoking status, alcohol consumption, etc.) [3]. The target variable used as the diabetes indicator has three possible values: healthy, pre-diabetic, and diabetic.

One of the notable characteristics of our dataset is that it is unbalanced, where most instances in our dataset define healthy individuals, with less instances acting as pre-diabetic, and even less established as diabetic. This means that we will be employing techniques to handle this imbalanced dataset during data preprocessing.

B. Motivation

Our dataset contains data on pre-diabetic individuals, which makes it a valuable resource for diabetes prediction and health trends. Because of its large size and wide number of attributes, this dataset is appropriate for machine learning applications to help build predictive models for early diabetes detection and see how factors influence one's risk of diabetes.

Since diabetes is such a pressing issue globally, it is important that we come up with better methods to detect diabetes, and pre-diabetes within individuals to catch the disease earlier and optimize future health outcomes.

C. Objectives

There have been multiple objectives raised to complete the project, including:

1. *Classification of individuals.*

   Apply methods of Single Vector Machine (SVM), Logistic Regression, Multi-layer perceptron, and random forest to solve the classification problem of assigning individuals from our dataset into one of the three categories defining diabetes. The categories to designate diabetes status are diabetic, pre-diabetic, and not diabetic.

D. Handle imbalanced dataset.

Handle working with an imbalanced dataset by mediating with techniques like oversampling and class weighting.

2. *Identify the significant predictors of diabetes.*

   Use Principal Component Analysis to complete dimensionality reduction to solve for feature importance, allowing us to detect the significant predictors of diabetes.

3. *Train multiple models and establish accuracy metrics.*

   Use metrics specific for class-imbalanced data like precision, recall/sensitivity, specificity, and F1 score to evaluate model prediction performance.

II. Methodology

A. Data Preprocessing

Exploratory Data Analysis (EDA) was used to help understand the initial features of the dataset. A heatmap was created to display the correlation between the dataset features, and a histogram of each feature was created to display the frequency of responses for each feature to detect any redundancies. A pie plot was created to give an idea of the percentage of healthy individuals compared to non-healthy individuals (pre-diabetic/diabetic), showing the class imbalance as many were healthy individuals (non-diabetic).

Implemented mean substitution across all the features to handle the missing values within the data, and separated the target variable, Diabetes_012, to be able to find the value count of each diabetes result, as 0 represents no diabetes, 1 represents pre-diabetic, and 2 represents diabetic.

To handle the imbalanced dataset, the SMOTENC technique was chosen to handle both categorical and numerical features. Categorical features were identified and SMOTENC function was applied to oversample the data and show that pre-diabetic and diabetic data is represented enough. To normalize the numerical features, the categorical and numerical columns were all identified and transformed with a column transformer to normalize numerical features and one-hot encode categorical features.

Cross-validation was completed by defining the dataset size (e.g., 100,000, 150,000 sample size) and number of random train-test splits (e.g., 5) and completed train-test split by randomly sampling and splitting samples into training and test data. Then, each training set was passed through SMOTENC, then normalized, and internal one-hot encoding is applied inside each training loop. One-hot encoding was used to create separate columns for each result of the Diabetes_012 feature, splitting each result (e.g., 0, 1, 2).

PCA is completed after scaling to eliminate noise and keep the components that explain 95% of variance within the data. Preprocessed datasets were saved for model training and further evaluation. For each dataset size and random split, training and test sets were created, which act as input in further analysis.

B. Logistic Regression

Logistic Regression was employed to address the two binary classification tasks:1) distinguishing between non-diabetic individuals from those who are either pre-diabetic or diabetic, and 2) distinguishing between the pre-diabetic and diabetic individuals. In this each classification task was performed across five randomized train-test splits and two dataset sizes: 100,000 and 150,000 samples. Principal Component Analysis (PCA) was applied beforehand, and only the PCA-transformed features were used for training. In the second task, non-diabetic samples were excluded as irrelevant.

For both tasks, a grid search with 3-fold cross validation was used to optimize hyperparameters, including the regularization strength(C), penalty type (L1 or L2), and solver (liblinear), maximizing the weighted F1 score, to improve generalization. A standardized pipeline of feature scaling and logistic regression was constructed using Pipeline for maintaining consistency. Final models were evaluated on independent test sets using accuracy, precision, recall and F1 score. Results were aggregated to report summary statistics across all dataset sizes and splits.

C. Multilayer Perceptron

The multilayer perceptron (MLP) model was used to solve two binary classification tasks: 1) distinguishing between the non-diabetic group and the combined pre-diabetic group and diabetic group, and 2) distinguishing between the prediabetic group and the diabetic group. In both tasks, the MLP models were tested on the 100,000-tuple split #1 only due to the high computational cost of training MLP models. Note that in the task distinguishing between pre-diabetic and diabetic tuples, the model did not receive any non-diabetic data as it was not relevant. Models received the training data outputs of the PCA step and underwent a cross-validation process to select hyperparameters. The models each had two hidden layers, and the number of neurons per hidden layer (N) was selected through a grid search over the values 10, 15, and 25. For each N value, the optimal value of α, the hyperparameter representing the strength of L2 regularization, was selected with a random search over the log-uniform distribution over the interval $[10^{-6}, 1]$ with 3 total samples using 3-fold cross validation. The validation F1 scores were obtained for each (N, α) pair and the highest pair was used to construct a new model over the entire training dataset. This model was then tested on the test set, and evaluation metrics including accuracy, precision, recall, and F1 score were computed.

D. *Kernel SVM*

This analysis evaluated a Kernel SVM (Gaussian/RBF kernel) on two binary diabetes classification tasks: Pre-diabetic vs. Diabetic, and Healthy vs. Diabetic. The input data originated from pre-processed features (scaled, SMOTENC-corrected) which were then reduced using PCA and loaded from a saved data split (e.g., 100k samples/split 1). Subsets for each binary task were created by filtering based on original labels (0/1/2) and remapping them to a binary format. For efficiency, these subsets were randomly down-sampled (stratified) to a maximum of 10,000 points. The SVM model (using default hyperparameters) was evaluated on these sampled subsets via 3-fold cross-validation. Performance was measured using Accuracy, Weighted Precision/Recall/F1-

Score, Sensitivity, Specificity, and Class 1 Precision, calculated from the out-of-fold predictions. These results were saved to separate CSV files for each task

E. *Random Forest*

The *Random Forest (RF)* algorithm was used to perform two *binary classification* tasks: (1) the classification of non-diabetic individuals compared to pre-diabetic or diabetic individuals, and (2) the classification between pre-diabetic and diabetic individuals. Preprocessed datasets of 100,000 and 150,000 instances were used; each divided into 80/20 training and test partitions. *Hyperparameter optimization* was achieved using *RandomizedSearchCV*, evaluating five randomly sampled hyperparameter configurations using *RepeatedStratifiedKFold* cross-validation (2 folds repeated 5 times). Important hyperparameters that were tuned were the number of estimators, the depth of the trees, and the minimum samples per node splits and leaves. For training, to overcome class imbalance, balanced class weights were used. For the first task, labels were binarized to match pre-diabetic and diabetic subjects; however, the second task only included pre-diabetic and diabetic cases. Feature importance scores were calculated from the learned models to identify the most predictive features. Model performance was assessed using *accuracy*, *weighted precision*, *recall, F1 score, confusion matrix, sensitivity,* and *specificity*. The best-performing hyperparameters and evaluation metrics were reported and compared for different dataset sizes and divisions to assess the robustness and generalization of the model.

III. Results

A. Logistic Regression Results

The Logistic Regression models delivered varying results between the two binary tasks when using optimized parameters and PCA features with multiple splits and sizes. LR_Model1 effectively split non-diabetics by reaching an accuracy rate of 81-82% along with an equivalent percentage of correct true and false cases (80-83%). LR_Model2 performed the second classification task for pre-diabetic patients versus diabetics but reached accuracy rates at 78%. The model achieved an acceptable amount of specificity in the range of 81-82 percent but demonstrated reduced sensitivity at 65 percent, which made the correct identification of 'diabetic' patients more challenging.

B. Multilayer Perceptron Results

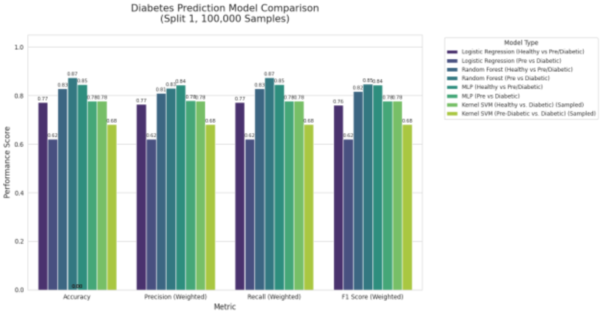
The optimization parameters with PCA transformation resulted in varying performance degrees for the MLP network on its two binary classification tests. The first MLP model produced 84-85% accuracy in identifying healthy from other classes. This model maintained a high success rate (90-91%) to identify non-healthy subjects, while its ability to detect healthy individuals reached only 72-73%. About 77-78% accuracy emerged from the second MLP model when it differentiated Pre-diabetics from Diabetics. The sensitivity and specificity metrics for this task matched each other closely at ~81-82% and ~74-75% respectively; however, the overall performance fell short of the first multi-layer perceptron model.

C. Random Forest Results

Random Forest models that used PCA features and adjusted hyperparameters demonstrated excellent overall accuracy yet proved less effective at detecting the positive outcome during both tasks. The model achieved an accuracy rate of ~87-88% when discriminating between Diabetics and Pre-diabetics, while producing great specificity rates of ~97-98%. However, the model demonstrated very poor sensitivity for detecting Diabetics at ~5-8%. Random Forest models achieved an accuracy of ~92% for identifying Healthy versus Pre-diabetic patients while maintaining very high specificity at ~98%, but produced low sensitivity at ~28-29% toward the pre-diabetic class.

D. Kernel SVM Results

Multiple Kernel SVM models utilizing the RBF kernel with default parameters underwent three-fold cross-validation of the sampled (maximum 10k points) PCA-transformed data. The SVM generated results with moderate performance during differentiation of Pre-diabetics and Diabetics, achieving 68% accuracy/F1 scores and sensitivity/specificity ratios of ~65% and ~71.5% respectively. The model distinguishing Healthy from Diabetic individuals delivered performance around 78% accuracy/F1, as well as sensitivity at approximately 81% and specificityat74%.



IV. Discussion

The logistic regression model consistently performed the worst across both the non-diabetic vs. (pre-diabetic or diabetic) task and the pre-diabetic vs. diabetic task. Note that the logistic regression is a linear model, while each of the other models have a nonlinear component. Thus, it is likely that this discrepancy in the performance is due to both tasks being not linearly separable, limiting the performance of any linear classifier on the data. The logistic regression did take much less time to train than any of the other models, allowing for more extensive testing on different subsets of the training data. Thus, the choice of whether to use a linear model or nonlinear model for these classification tasks and similar ones depends on the tradeoff between performance and computational cost.

The next worst model was the SVM with the gaussian kernel. It performed marginally better than the logistic regression model on non-diabetic vs. (pre-diabetic or diabetic) task and substantially better than the logistic regression on the pre-diabetic vs diabetic task, but still less than the other nonlinear models. SVM with the gaussian kernel, like the nonlinear models, took a significant amount of computational resources to train. Because of this, the choice was made to forgo significant hyperparameter optimization for this model (although hyperparameter optimization was still done for the MLP and the random forest). Thus, the decrease in performance of the SVM with the gaussian kernel relative to the other nonlinear models could be due to the lack in hyperparameter optimization, again highlighting the issue of high computational cost when training complex models.

Most of the models performed significantly worse on the pre-diabetic task vs. diabetic task than on the non-diabetic vs. (pre-diabetic or diabetic) task. There could be a couple of reasons behind this discrepancy. One possibility is that there is greater separation between the non-diabetic class and the other two classes than there is between the pre-diabetic class and the diabetic class. Another possibility deals with the class imbalance in the dataset. Of the three overall classes, the pre-diabetic class had the lowest representation in the original dataset. To deal with this class imbalance, SMOTE was used to create synthetic data for the underrepresented classes until the classes were balanced with synthetic data. Unfortunately, synthetic data is not the same quality as true data, and the less true data available in an underrepresented class (like pre-diabetic), the harder it is for SMOTE to properly model this data and generate quality synthetic data from it. In both tasks, synthetic pre-diabetic data is used to train each of the models. However, in testing, the models made predictions on less pre-diabetic data in the non-diabetic vs. (pre-diabetic or diabetic) task since pre-diabetic data is lumped together with diabetic data, which could explain the higher performance on this task. The notable exception to this trend is the random forest model, which performed about the same or even better on the pre-diabetic vs. diabetic task than the other task; it is unclear why this is the reason, but it could be due to a higher generalization capacity of the random forest model.

IV. Conclusion

This study applied multiple ML models and techniques –logistic regression, multilayer perceptron, kernel SVM and random forest to predict diabetes status using the CDC Diabetes health indicators dataset. Our experiments revealed that non-linear models generally outperform linear models in handling the complex, Imbalanced data of diabetes prediction. While high specificity and accuracy were achieved, sensitivity remains a challenge, particularly in distinguishing pre-diabetic from diabetic cases. The use of SMOTENC to address class imbalance helped improve model training but the limited quality of synthetic data especially for underrepresented classes like pre-diabetes impacted performance.

Future work should explore more diverse datasets, refined hyperparameters tuning and robust methods for handling missing data to further improve predictive performance. Testing on real-world data with varying levels of completeness and using alternative resampling or ensemble techniques could provide more reliable results.

Overall, our findings support the potential of machine learning to enhance early diabetes detection and inform targeted public health interventions, helping identify at-risk individuals sooner and enabling timely preventive measures.

References

[1] M. Karamanou, A. Protogerou, G. Tsoucalas, G. Androutsos, and E. Poulakou-Rebelakou, "Milestones in the history of diabetes mellitus: The main contributors," *World J. Diabetes*, vol. 7, no. 1, pp. 1–7, 2016.

[2] "CDC Diabetes Health Indicators - UCI Machine Learning Repository," *Uci.edu*. [Online]. Available: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators. [Accessed: 24-Apr-2025].

[3] 'Behavioral Risk Factor Surveillance System," *Cdc.gov*, 17-Apr-2025. [Online]. Available: https://www.cdc.gov/brfss/index.html. [Accessed: 24-Apr-2025].