Diabetes Detection Using Machine Learning with Feature Engineering

Awais Naeem Neil Roy Nicholas Strohmeyer Ahmet Gunhan Aydin

1. Introduction

Diabetes Mellitus refers to a class of chronic diseases that are characterized by excessive blood sugar. This excess causes chronic issues and can lead to failure of vital organs such as the eyes, nerves, heart, blood vessels and kidneys [1]. Furthermore, diabetes greatly increases the risk of stroke, heart disease and developing various forms of cancer, thus posing a significant health risk to all individuals with the disease [2].

It was estimated that 422 million people worldwide had diabetes in 2014. This number rose from just 108 million in 1980 [3]. In the past 30 years, there has been a rise in type 2 diabetes in particular. In the United States, about 11.3% of the population has diabetes, this figure is 14.7% for adults (over 18). Notably, about 8.5 million of these adults were not aware they had diabetes or did not report it [2].

This last fact, plus the fact that preventive measures can have a significant impact on delaying the early onset of type 2 diabetes in adulthood, underscores the importance of developing predictive systems that are able to detect diabetes in the early stages and diagnose prediabetic individuals. Some studies have suggested predicting the early onset of diabetes using machine learning techniques and one group of authors was able to achieve over 90% prediction accuracy by training/evaluating the Random Forest classifier on a dataset containing 14 risk factors associated with the diabetes [4]. Another study includes a fairly good survey of machine learning methods that have been tried by the community to predict the early onset of diabetes [5].

In our project, we will utilize a health indicators dataset with multiple risk factors and try a variety of leading ML classification models in an attempt to develop a model that can classify an individual as non-diabetic, prediabetic or diabetic. Furthermore, using the results of our models, we will attempt to analyze the risk factors which play a significant role in corresponding predictions. The combination of these tasks not only provide benefit in helping identify diabetes in its early stages but also has the potential to point out possible methods of effective treatment.

2. Dataset

We will be using the Diabetes Health Indicators dataset from Kaggle which contains around 253,680 survey responses to the CDC's BRFSS 2015 [6]. The target variable has three classes namely non-diabetic, prediabetic and diabetic. There exists a class imbalance in the dataset such that non-diabetic, prediabetic and diabetic classes have 213704, 4632 and 35345 samples respectively.

There are a total of 21 feature variables which will constitute the training data for our project. The feature space is a mix of nominal, ordinal and numeric features. Out of 21 input features, 14 are binary categorical features whereas the rest are either continuous or ordinally ranked. Using these feature variables, we will train our models such that they are able to predict whether a person is diabetic, prediabetic or non-diabetic.

Dataset cleaning will involve the removal of duplicate records, imputation/dropping of the missing sample values across the features and modification of feature data types from string to int/floating point values. Dataset preprocessing will involve feature selection by removing highly correlated features containing redundant information, encoding of categorical/nominal features using one-hot encoder, and scaling/normalization to ensure that all the features have zero mean and unit variance.

To separate out the data for training and testing purposes, a stratified train-test split will be formed so that 80% of data ($\sim$200k samples) is designated for training and 20% data ($\sim$50K samples) is reserved to evaluate the model's performance. Since our data has a class imbalance, we will need to impose stratification here so that the splitted data has equal proportions of all the three classes.

3. Methods

To predict the diabetes onset using various risk factors, we have carefully chosen five distinct machine learning algorithms, each with its unique strengths and capabilities. Logistic Regression, our first algorithm, is a linear model suitable for binary and multi-class classification. Its simplicity and interpretability make it an ideal choice for this project. The Support Vector Classifier (SVC) is a powerful algorithm that creates decision boundaries by maximizing the margin between classes, handling complex data distributions effectively. Random Forest Classifier, an ensemble method, combines multiple decision trees to enhance model robustness and provide feature importance analysis, crucial

for understanding the relevance of various features in diabetes detection. Multi-Layer Perceptron (MLP) represents a neural network-based approach, known for its capacity to capture complex, non-linear patterns in the data. Finally, we employ XGBoost [7] an extreme gradient boosting algorithm that excels in predictive performance and is capable of handling large datasets efficiently. Each of these algorithms offers a distinct approach to diabetes detection, and their combination ensures a comprehensive and well-rounded exploration of the problem domain, increasing the likelihood of achieving accurate and robust results.

To ensure the accuracy and robustness of our models, we will adopt the Stratified K-Fold Cross Validation technique, which is particularly valuable when working with imbalanced datasets. By partitioning the data into stratified subsets, we can train and test our models in a way that ensures representation of all classes in each fold, preventing any class from being underrepresented during evaluation. In the quest for optimal model performance, we will delve into hyperparameter tuning through a Grid Search process. This step is vital for finding the most suitable hyperparameters for each algorithm, maximizing their predictive capabilities. Moreover, our project will feature an ablation study employing Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). PCA is a linear dimensionality reduction technique that extracts a set of uncorrelated variables called principal components, focusing on preserving the variance within the data. In contrast, t-SNE is a nonlinear method that emphasizes maintaining pairwise data point similarities. PCA simplifies the data while retaining critical information, while t-SNE reveals complex relationships in high-dimensional data. These dimensionality reduction techniques will enable us to identify a reduced feature space that strikes a balance between preserving the majority of the variance within the data while minimizing the number of features. This is important not only for enhancing model efficiency but also for improving model interpretability.

4. Results

After the models have been trained, we will evaluate them on the test data containing the health indicators to analyze how accurately our models can predict whether an individual has diabetes or is at risk to develop diabetes. Similar experiments have been done using the same BRFSS data, with accuracies reaching up to 82.4% [8]. We are hoping to achieve similar or better results and to gain insight into the pros and cons of the different models and methods that we are testing on the dataset. Furthermore, we will analyze the risk factors or combination of risk factors which are most predictive of diabetes and if we can still achieve high accuracies compared to other state of the art predictors, while using a smaller subset of health indicators.

The design space exploration of the different machine learning models will likely be a major part of our project along with finding ways to effectively deal with the data imbalance present in the dataset. We want to also test the effects of different preprocessing techniques, or the lack of preprocessing on the various ML models as well. Our findings will be useful in understanding not only which risk factors are indicative of diabetes or prediabetes, but also to get an idea of whether or not the BRFSS dataset's survey questions can be used to provide accurate predictions on whether an individual has diabetes or not.

For each model, we will be using metrics such as accuracy, precision, recall, F1 score, ROC, and confusion matrices to compare model performances against each other and get a better understanding of their strengths and weaknesses. Model accuracy is our primary concern, and will be used during the grid search for hyperparameter optimization. Similarly, during the ablation study, these metrics will give us a good idea about the effects of removing certain features and how well certain subsets of features can be used for classification compared to using all of the features.

References

[1] American Diabetes Association. "Diagnosis and classification of diabetes mellitus." Diabetes care 33.Supplement_1 (2010): S62-S69.

[2] Centers for Disease Control and Prevention. Diabetes Basics: What is Diabetes? Accessed November 9, 2021. https://www.cdc.gov/diabetes/basics/diabetes.html

[3] World Health Organization Fact Sheet: https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics 9 (2018): 515.

[5] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.

[6] Kaggle Dataset: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[7] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4, 2015.

[8] Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis. 2019 Sep 19