Brisa Salazar
Kenny Gonzalez
Collins Kariuki
12/5/2023

Final Paper

Introduction

The motivation for our project was the high numbers of type 2 and prediabetes in teenagers and young adults. For context, prediabetes and type 2 diabetes are some of the most prevalent diseases for young adults. However, they are also one of the most preventable diseases as they are based on a person's lifestyle.[1] Therefore, we wanted to implement a classifier that would help us predict whether someone is non-diabetic or either prediabetic or type 2 diabetes. Furthermore, from these predictions, we wanted to examine which features are important in determining diabetes or not. These features ranged from lifestyle (income, education, etc) to general health (physical health, BMI, etc).

To achieve our goals, we wanted to pick a classifier that had good binary classification, was able to handle both binary and non-binary features, and was efficient and fast since we ran over 70 thousand examples with it. Then, we wanted to preprocess the data in an accurate way that did not disturb the integrity of the data. Thus, the only preprocessing we did was on the labels. The data came with labels 0 and 1, and our classifier worked with -1 and 1, so we changed it so that it would fit our purposes. Finally, we chose to use the gradient descent classifier, and we experimented with accuracies, features, and statistical tests to come to our conclusions.

Experimental Setup

We decided to use a dataset found on Kaggle, containing survey responses from the 2015 Behavioral Risk Factor Surveillance System, a survey sent out annually by the Center for Disease Control (CDC).[2] The dataset has a total of 21 features that are a mix of binary and non-binary. The features included: high blood pressure, high cholesterol, cholesterol check, BMI, smoker, stroke, heart disease or attack, physical health, diet with veggies, diet with fruits, heavy alcohol consumption, healthcare coverage, no visits to the doctor because of cost, general health, mental health, physical health, difficulty walking, sex, age, education, and income. The dataset contained two labels, making it a binary dataset, allowing us to use the Gradient Descent classifier. The dataset contained 70,692 examples and had an equal number of -1.0 and 1.0 labels.

To determine what feature(s) are the most significant in determining prediabetes or Type 2 diabetes, we decided to create 22 different models, each one with one feature removed. For

[1] https://www.cdc.gov/diabetes/basics/diabetes.html
[2] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

example, we had our Control Model which contained all 21 features, Model 0 which had feature 0 removed, Model 1 which had Feature 1 removed, and so on. In doing so, we could then compare the accuracies of the different models and see which feature's removal affected the accuracy the most. We created a 10-fold cross-validation for each model, then used t-tests between the control model and the models with the lowest accuracies to determine which model(s) were the most significant.

Results.

In summary, we compiled a table featuring 23 models, including a control model, each with 10 accuracies per cross-validation run. The corresponding Sheets link is here. Upon examination, models 3 and 13 exhibited lower-than-normal accuracies, prompting us to conduct paired t-tests against the control model to assess their statistical significance in affecting the base accuracy. Utilizing Excel's TTEST function yielded the following results:

Key:

Model 3 - ran on the dataset less the BMI feature.

Model 13 - ran on the dataset less the GenHlth (General Health) feature.

Control - ran on the dataset with all the features present.

MODELS	P-VALUE	EVALUATION
Control vs Model 3	0.0001619802116	Model 3 is statistically significant.
Control vs Model 13	0.00007015590996	Model 13 is statistically significant
Model 3 vs Model 13	0.3129240653	It does not matter which one you take out, the difference is insignificant.

The Excel TTEST function calculates the probability of a t-test, indicating whether two samples likely originate from populations with the same mean. We assessed statistical significance using a standard threshold of 0.01; if the p-value is below 0.01, we consider the sample statistically significant, and vice versa.

Conclusions.

The findings suggest that BMI and General Health are crucial factors in predicting (pre)diabetes. Exploring additional project directions could involve reconsidering our approach to model selection and conducting paired t-tests on all models compared to the control.

When considering the application of ML in healthcare, ethical considerations arise. Achieving success in clinical implementation hinges on the explainability of ML models. This informed our model choice, favoring gradient descent for its robustness and explainability over more accurate but less interpretable deep learning models. In clinical contexts, the balance between explainability and accuracy is crucial, as convincing clinicians to adopt ML for diagnoses relies on sound explanations.[3]

[3]Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey (Rasheed et al., 2022)