Project Outline

Data Mining Project - Diabetes Risk Prediction

Team 11 - Support Vector Superstars
Philipp Gänz, Salome Heckenthaler, Patricia Paskuda,
Benedikt Prisett, Matthias Fast

October 13, 2024

Submitted to
Data and Web Science Group
Prof. Dr. Sven Hertling
University of Mannheim

Motivation. Diabetes is one of the most common chronic diseases in the United States, affecting millions and imposing a significant financial burden on the economy, with annual costs approaching $400 billion. It occurs when the body either fails to produce enough insulin or cannot use it effectively, leading to high blood sugar levels due to the inability to absorb glucose from the bloodstream. Since glucose is the primary energy source for the brain and muscles, its proper regulation is essential for overall health. Uncontrolled high blood sugar can lead to severe complications, including heart disease, kidney failure, and vision loss, ultimately reducing both quality of life and life expectancy. While diabetes can have a genetic basis, it can also develop in individuals without a family history.

In 2018, the Centers for Disease Control and Prevention (CDC) reported that over 34 million Americans had been diagnosed with diabetes, with an additional 88 million at risk due to prediabetes. These numbers are expected to rise. Given the increasing prevalence of diabetes and its serious health consequences, early detection is crucial for managing the disease.

This project aims to develop an accurate predictive model for identifying individuals at risk of diabetes using the `Diabetes Health Indicators` dataset from Kaggle. Our goal for this project is to enhance early diagnosis and to identify crucial features of diabetes patients, helping to guide prevention strategies and reduce its societal impact.

Dataset. The data this project deals with was originally provided by the CDC, a U.S. government agency, and was obtained by an annually conducted telephone survey called "Behavioral Risk Factor Surveillance System". For the purpose of this project, we make use of an already preprocessed dataset by Alex Teboul which is based on the original dataset of CDC.

The version of the dataset we are going to use is called "diabetes_binary_health_indicators_BRFSS2015" and only includes clean data (due to being already preprocessed). However, it is still unbalanced as it contains 218,334 samples (86.1%) with a binary target variable called "Diabetes_binary" of 0.0, meaning no diabetes in that case, and only 35,346 samples (13.9%) with "Diabetes_binary" being 1.0, standing for either prediabetes or diabetes.

Besides, the in total 253,680 provided samples, the dataset includes 21 features which are available to us in order to use for predicting the binary target. Most of the feature names are self-explanatory like "Smoker", "Stroke", "Physical Activity" etc. However, some people might not be familiar with the feature "BMI" (Body Mass Index) which is a value used to measure the body fat percentage by using the weight and height of a person. Out of all features are 16 of categorical type and most of them (14) are binary. "BMI", in this dataset, is a discrete variable with integer values ranging from 12 to 98, even though, all samples are displayed with one decimal point which is zero in every case. The other four continuous features are "MentHlth", "PhysHlth", "Age" and "Income", although the feature "Age" is divided into 13 bins making it somewhat categorical.

In the following, all of the 21 features and the target variable are listed without any further explanation, as we are going to provide details on them in our final project

report, if necessary: *HighBP* (= high blood pressure), *HighCol* (= high cholesterol), *CholCheck* (= conducted cholesterol check), *BMI*, *Smoker*, *Stroke*, *HeartDiseaseorAttack* (= heart disease or attack), *PhysActivity* (= physical activity), *Fruits*, *Veggies*, *HvyAlcoholConsump* (= heavy alcohol consumption), *AnyHealthcare* (= any kind of health care coverage), *NoDocbcCost* (= need of doctor but could not afford), *GenHlth* (= general health), *MentHlth* (= mental health), *PhysHlth* (= physical health), *DiffWalk* (= difficulty walking), *Sex*, *Age*, *Education*, *Income* and **Diabetes_binary** (target variable). Nonetheless, their respective detailed description can already be accessed on kaggle.com.

Methodology. To gain an initial understanding and overview about the data at hand, we will generate summary statistics for each feature, including mean, median, and standard deviation. This analysis will help us understand the data types present (categorical versus continuous) and identify any anomalies or inconsistencies. Next, we will assess the dataset for missing values, analyzing their distribution and potential impact on model performance. Understanding the extent of missing data is vital for deciding on appropriate handling methods, such as imputation or removal. A correlation analysis will follow, where we calculate the correlation matrix to uncover relationships between features and the target variable. Visualizing these correlations through heatmaps will allow us to identify which features have the strongest associations with diabetes diagnosis, guiding our feature selection process.

In the visualization phase, we will conduct univariate analysis using histograms and boxplots for numerical features to observe their distributions and pinpoint any outliers. For categorical variables, bar plots will illustrate frequency distributions, providing a clear picture of each category's prevalence. Moving to bivariate analysis, scatter plots will help us visualize relationships between numerical features and the outcome variable, while grouped boxplots will compare the distributions of numerical features across different categories of the target variable. Pair plots will further enhance our understanding of the interactions among features, revealing patterns that could inform our model.

Following exploration and visualization, we will proceed to data preprocessing. Since the dataset has already been preprocessed, this step will probably only need to be carried out to a small extent. Nonetheless, we will address missing values through strategies tailored to their distribution and importance. Feature scaling will be implemented to normalize or standardize numerical features, ensuring uniformity across the dataset, especially for algorithms sensitive to feature scales. As the categorical features are already encoded, we do not need to process them any further. Apart from that, we will perform feature selection using techniques like Principal Component Analysis (PCA), identifying the most relevant features for our classifier.

Building upon the first data exploration and preprocessing, several models will be implemented in an iterative process to perform the task of diabetes risk prediction. While the exact scope and models to be explored might change slightly throughout the project, a preliminary overview of the most applicable techniques can be provided.

Logistic regression is widely established for binary classification problems and will therefore be one of the algorithms used. It models the probability of a class outcome

using a linear combination of input features. Another technique suitable for classification tasks are decision trees, which split data into branches based on feature values to make decisions. One issue that is, however, often encountered with decision trees is overfitting. To address this, ensemble methods will be applied. These include random forests, which combine multiple decision trees to reduce variance and improve accuracy; gradient boosting, which builds trees sequentially to minimize errors; and AdaBoost, which adjusts the weight of misclassified instances to focus on difficult cases. Support Vector Machines aim to find the optimal hyperplane that best separates the classes in the data and work well for high-dimensional data and complex decision boundaries. These characteristics make them applicable to the problem at hand. Additionally, the methods Naïve Bayes and K-Nearest Neighbors can be explored. Lastly, we plan to explore more complex deep learning approaches by training a deep neural network, which can capture very complex patterns in the data.

A train, validate, and test split will be performed, which should pose no issue given the large number of data samples. For hyperparameter selection, which is needed for models such as random forests and support vector machines, we will implement techniques such as grid search or randomized search to optimize model performance, as well as cross-validation to ensure robust selection of hyperparameters. To address the class imbalance of the target variable, we will look into best practices such as oversampling, undersampling, or synthetic data generation to improve the model's performance. Another aspect of concern is feature selection and engineering. While this is closely related to our first data exploration, we will include these steps in our iterative process of model implementation and evaluation.

Applying these techniques, several different machine learning models will be trained, tested, and compared. This will allow us to find the most suitable techniques that can perform accurate diabetes risk prediction.

Evaluation. For evaluating our trained models we will mainly rely on the *confusion matrix* by computing the *precision* and *recall* scores, in order to estimate our models' generalization performances. In addition, the *F1-score*, combining both the precision and recall score into one value, will be highly important, since our dataset is imbalanced. Moreover, the *specificity* (true negative rate) provides also an interesting metric to look at. Plotting the receiver operating characteristic (ROC) curve will also be insightful as it visualizes true positive and false positive rate in relation to a model's confidence. Furthermore, a learning curve can show how the accuracy changes with a growing training set size, indicating which size is actually significant e.g., for similar classification problems predicting another disease.

For each model we train, the above-mentioned values will serve as metrics for comparing the various models. Additionally, we will contrast each model with our baselines, the majority class (assigns the most likely label to each new data point) and a random classifier (assumes the distribution of our training data and predicts the positive class with the proportion of the positive samples in our training data)