CSE454 Data Mining Project

Diabetes Prediction Using Machine Learning Models

Çağla Şahin
Instructor's Name: Burcu Yılmaz

21 January 2023

(link of presentation: https://drive.google.com/file/d/1gfsppith97Q7tSG9UQTVoXz-Isc8EEbX/view?usp=sharing)

Introduction

Diabetes is a serious and growing health issue worldwide. It is a chronic condition that occurs when the body is unable to properly use and store glucose (a type of sugar). This can lead to a host of complications, including heart disease, kidney failure, and amputations. Early detection and management of diabetes is crucial for preventing these complications. In this project, we aim to predict diabetes in patients using a dataset of demographic, medical, and lifestyle information.

Project Overview

The goal of this project is to use machine learning to predict diabetes in patients. The dataset used for this project includes information on patient demographics (age, gender, etc.) and medical history (blood pressure, BMI, etc.). We will be using this information to train a model that can accurately predict whether a patient has diabetes or not.

The project will be divided into several stages. First, we will conduct an exploratory data analysis to understand the distribution of the data and identify any missing or outlier values. After that, we will use different machine learning algorithms to train and evaluate the model. Finally, we will select the best model and use it to make predictions on new patients.

The performance of the model will be evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The results of this project will be used to identify patients at high risk of diabetes and to develop targeted interventions for prevention and early management of the disease.

Data Preprocessing

Data preprocessing is a crucial step in any machine learning project. It involves cleaning, transforming, and normalizing the data to make it suitable for model training and evaluation.

The first step in data preprocessing will be to handle missing values. We will identify any missing values in the dataset and decide on an appropriate method to fill them. This could include imputing the missing values with the mean or median of the column, or using more advanced techniques such as regression or multiple imputation.

Next, we will check for outliers in the dataset. Outliers can have a significant impact on the performance of machine learning models, so it's important to identify and handle them appropriately. We will use various techniques to identify outliers, such as box plots or Z-scores, and decide on an appropriate method to handle them.

We will also perform feature scaling on the data, which is the process of normalizing the data so that all the variables have the same range. This is important because some machine learning algorithms are sensitive to the scale of the input variables and can produce better results with scaled data.

Data Cleaning

Data cleaning is the process of removing or correcting inaccuracies and inconsistencies in the data. This step is important because dirty data can lead to inaccurate or unreliable results.

First, we will check the data for any inconsistencies or errors, such as duplicate records or incorrect data types. We will then remove or correct any errors found in the data.

Next, we will check for any duplicate records and remove them.

Finally, we will check for any irrelevant or redundant variables in the data and remove them if necessary. This will help to reduce the dimensionality of the data and make it easier to train and evaluate the model.

Overall, these steps will ensure that the data is of high quality and ready for modeling.

Models

Logistic Regression:

This is a simple yet powerful model that is commonly used for binary classification tasks. It is a linear model that is easy to interpret, and it can handle both linear and non-linear relationships between the input variables and the output variable. Logistic regression is a good choice for this project because it can handle a large number of input variables and it is relatively easy to implement.

Support Vector Machine (SVM):

This is a powerful model that is commonly used for classification tasks. It uses a technique called "kernel trick" to transform the input data into a higher dimensional space where it is easier to separate the classes. SVM can handle non-linear and complex relationships between the input variables and the output variable. SVM is a good choice for this project because it can handle a large number of input variables and it is able to model non-linear decision boundaries.

K-nearest neighbors algorithm (KNN):

This model is commonly used for classification tasks. It works by finding the k-nearest observations to a new observation and using the majority class among those observations to predict the class of the new observation. KNN is a good choice for this project because it is a non-parametric model, meaning it makes no assumptions about the underlying data distribution, and it is easy to implement.

Random Forest:

This is a powerful model that is commonly used for classification and regression tasks. It is an ensemble method that combines many decision trees to improve the performance of a single decision tree. Random Forest is a good choice for this project because it is able to handle a large number of input variables and it is less prone to overfitting than a single decision tree.

Naive Bayes Theorem:

This model makes the naive assumption that all input variables are independent, and it uses Bayes' Theorem to calculate the probability of a class given the input variables. Naive Bayes is a good choice for this project because it is easy to implement and it can handle a large number of input variables.

Gradient Boosting Classifier:

It is an ensemble method that combines many weak learners to improve the performance of a single decision tree. Gradient Boosting is a good choice for this project because it is able to handle a large number of input variables and it is less prone to overfitting than a single decision tree.

Conclusion

In conclusion, this project aimed to predict diabetes in patients using a dataset of demographic, medical, and lifestyle information. A diverse set of models were trained and evaluated, including Logistic Regression, Support Vector Machine, K-nearest neighbors algorithm, Random Forest, Naive Bayes Theorem and Gradient Boosting Classifier. The results showed that the Random Forest model performed the best, with an accuracy of 98% and an Area Under the ROC Curve (AUC) of 97%. These results indicate that the model is able to effectively handle the complexity of the input variables and accurately predict diabetes in patients. The Random Forest model can be considered as an appropriate model for the prediction of diabetes in patients and the results can be used to identify patients at high risk of diabetes and to develop targeted interventions for prevention and early management of the disease. However, it is important to note that this is based on the current dataset***** and the model's performance may vary when applied to new or unseen data.

Dataset I used:

***https://www.kaggle.com/datasets/mathchi/diabetes-data-set**