BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Diabetes prediction using machine learning

– ITSG report –

Author
Păpară Ana-Maria
Software Engineering
group 248

2022-2023

Contents

Chapter 1

Introduction

1.1 What? Why? How?

Diabetes is a chronic health condition affecting millions worldwide, often leading to severe complications if undiagnosed or poorly managed. Early detection is crucial for effective treatment and better patient outcomes. Traditional diagnostic methods rely on laboratory tests, which can be time-consuming and resource-intensive. With the use of Artificial Intelligence (AI), I intend to develop an application that can predict diabetes onset using readily available clinical measurements, thereby aiding healthcare providers in early diagnosis and intervention.

The scientific problem addressed in this report is the early and accurate prediction of diabetes based on specific diagnostic measurements, a task that is often challenging due to the complex nature of diabetes risk factors and their interactions. Detecting diabetes risk early can enable timely intervention and help prevent or mitigate these health issues, improving patient outcomes and reducing healthcare costs.

By building a user-friendly application, I aim to deliver a practical tool for healthcare providers and patients to support early diagnosis and informed decision-making.

1.2 Paper structure

This work is organized into five chapters as follows:

- **1:Introduction** Provides an overview of the research problem, its significance, and the objectives of this study.
- **2:State of the art and related work** Reviews relevant literature on diabetes prediction models.
- **3:Approach** Details the proposed algorithm.
- **4:Application** Describes in-detail the steps of each project phase.
- **5:Results** Presents the results from testing and evaluating the model

1.3 Original contributions

This research paper contributes to the advancement of both theoretical and practical aspects of AI-based predictive models, specifically tailored for diabetes prediction. The contributions of this work are threefold:

- **Intelligent algorithm development** The primary contribution is the design and implementation of an intelligent machine learning algorithm that can predict diabetes based on patient diagnostic measurements.
- **User-friendly application**: The second contribution is the development of an intuitive, easy-to-use application that enables healthcare providers and patients to input diagnostic data and receive real-time predictions.
- **Comprehensive model evaluation and comparison**: The third contribution is an in-depth analysis of different models performances, including detailed evaluation metrics and comparison on different dataset sizes.

1.4 Scientific problem

The core problem addressed in this study is the early prediction of diabetes using available patients diagnostic data. By training the model on historical patients data, the algorithm can learn the characteristics of individuals at risk for diabetes, thereby generating rapid and reliable predictions when provided with new input data. This approach can significantly reduce the need for costly and time-consuming diagnostic procedures, allowing healthcare providers to prioritize high-risk patients for further testing and intervention.

Formally, the problem can be defined as a binary classification task:

- **Inputs** (X): A set of diagnostic measurements for each individual, including features such as age, blood glucose levels, mass index, insulin levels, and blood pressure.
- **Output** (Y): A binary prediction indicating the likelihood of diabetes: positive (1) if the individual is at high risk, or negative (0) if they are at low risk.

Chapter 2

State of the art and related work

Diabetes prediction through data analysis has gained significant attention due to the increasing potential of machine learning and artificial intelligence. This chapter reviews some previous articles that focus on diabetes risk predictions using AI approaches and techniques.

2.1 Diabetes prediction using supervised machine learning [3]

- **Dataset**: Pima Indians Diabetes Database[5]
- **Purpose**: Comparison between the KNN and Naive Bayes algorithms to see which algorithm suit the best for diabetes prediction based on several health attributes.

2.2 Diabetes Prediction using Machine Learning Algorithms [4]

- **Dataset**: Pima Indians Diabetes Database[5] & Diabetes Dataset[2]
- **Purpose**: Improve the accuracy of prediction of diabetes by including some external factors along with regular ones and by using various ML algorithms.

2.3 A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods [6]

- **Dataset**: Pima Indians Diabetes Database[5]
- **Purpose**: To compare several classifiers and feature selection techniques to more accurately predict diabetes.

Chapter 3

Investigated approach

3.1 Overview

My approach is to compare the performance of multiple machine learning models (existing ones-Logistic Regression, Random Forest Classifier, Gaussian Naive Bayes, SVC and also new created ones). The goal is to evaluate these models based on accuracy, precision, recall, F1 score, and computational efficiency to determine the most effective approach for predicting diabetes risk using clinical data.

3.2 Algorithm

Algorithm 1 Diabetes prediction

INPUT: Patient diagnostic data
OUTPUT: Best performing model

```
Preprocess data (normalize, handle missing values)
Split data into training and test
for each model in {Logistic Regression, Random Forest, Gaussian Naive Bayes, SVC} do
    Train model on training data
    Test model on test data
    Evaluate model (accuracy, precision, recall, F1 score)
end for
Compare results across models
```

3.3 Project walkthrough

- **Step 1**: Search for different size datasets.
- **Step 2**: Apply algorithm 1 for each dataset.
- **Step 3**: Find improvements.
- **Step 4**: Draw conclusions and discuss.

Chapter 4

Application

This study compare the performance of multiple machine learning models to predict diabetes risk based on diagnostic features. The primary criteria for evaluating the models are accuracy, precision, recall, F1 score, and computational efficiency. These metrics are chosen to ensure the model's performance is robust in predicting diabetes while also being interpretable and efficient for practical use.

4.1 Experiment 1 - small data

4.1.1 Overview

In the first experiment, I evaluated multiple machine learning models (Logistic Regression, Random Forest, Gaussian Naive Bayes, and Support Vector Classifier) on their ability to predict diabetes risk based on the PIMA Indians Diabetes dataset. The objective was to determine the best-performing model by comparing the results of different evaluation metrics.

4.1.2 Data

The PIMA Indians Diabetes[5] is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset consists of 768 instances and 8 attributes. All patients here are females at least 21 years and the attributes are medical health measurements as follows:

- *Number of pregnancies*
- *Glucose level*
- *Blood pressure*
- *Skin thickness*
- *Insulin level*
- *Body mass index*

- *Diabetes pedigree function*
- *Age*
- *Outcome (patient has diabetes or not)*

The dataset was split into 80% training data and 20% testing data.

4.1.3 Results

The models were evaluated based on accuracy on test data. The results showed that:

- **Random Forest** and **KNN** achieved the highest accuracy.
- **SVC** performed reasonably well but had slightly lower accuracy compared to Random Forest.
- **Logistic Regression** and **Gaussian Naive Bayes** had the lowest performance in terms of accuracy.

Algorithm	Accuracy
Logistic Regression	0.69
Random Forest Classifier	0.74
Gaussian Naive Bayes	0.69
Support Vector Classification	0.71
KNN	0.74

Table 4.1: Models accuracies on PIMA dataset

These results suggest that Random Forest is the most reliable model for predicting diabetes risk based on this dataset, although further improvements can be made with tuning and additional data.

4.2 Experiment 2 - real data

4.2.1 Overview

In the second experiment, we scaled up the analysis by using a larger dataset, consisting of over 445,000 records with other types of features. The goal was to test the models' ability to handle larger datasets and determine if they maintain their predictive accuracy when faced with more data.

4.2.2 Data

The dataset was obtained by filtering the 2022 BRFSS Survey Data [1] using python and pandas library. The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

My filtered dataset version consists of 444,046 instances (after pre-processing) and 15 attributes. The features were selcted from a total range of over 300 attributes in order to be relevant in the prediction of diabetes. After multiple filterings, the attributes in the dataset are:

- *sex*: respondent's gender
- *avg_sleep_time*: average sleep time (hours)
- *heart_attack_ever*: if ever suffered a heart attack
- *heart_disease*: if ever had a serious heart disease
- *asthma_ever*: if ever suffered of asthma
- *any_cancer*: if ever had any type of cancer
- *pulmonary_disease*: if ever had chronic pulmonary disease
- *depressive_disorder*: if ever had a depressive disorder
- *kidney_disease*: if ever had kidney disease not including kidney stones
- *has_diabetes*: if ever had diabetes
- *weight*: respondent's weight (in kilograms)
- *height*: respondent's height (in centimeters)
- *is_deaf*: deaf or serious difficulty hearing
- *is_blind*: blind or serious difficulty seeing
- *any_walk_difficulty*: serious difficulty walking or climbing stairs

By filtering, it were excluded all attributes that:

- **cannot have any impact on patient diabetes status**
  - respondent's residence
  - respondent's source of health insurance
  - respondent's educations
  - respondent's screening history
  - respondent's number of removed teeth
- **are limited on a short period of time and so their impact decreases**
  - physical activity during the past 30 days
  - alcohol consumption during the past 30 days

- confusion or memory loss during the past 12 months

- **could have impact on diabetes but have many missing values**
  - smoke cigarettes every day, some days, or not at all
  - ever had an H.P.V. vaccination
  - respondent's race
  - life satisfaction
  - stress

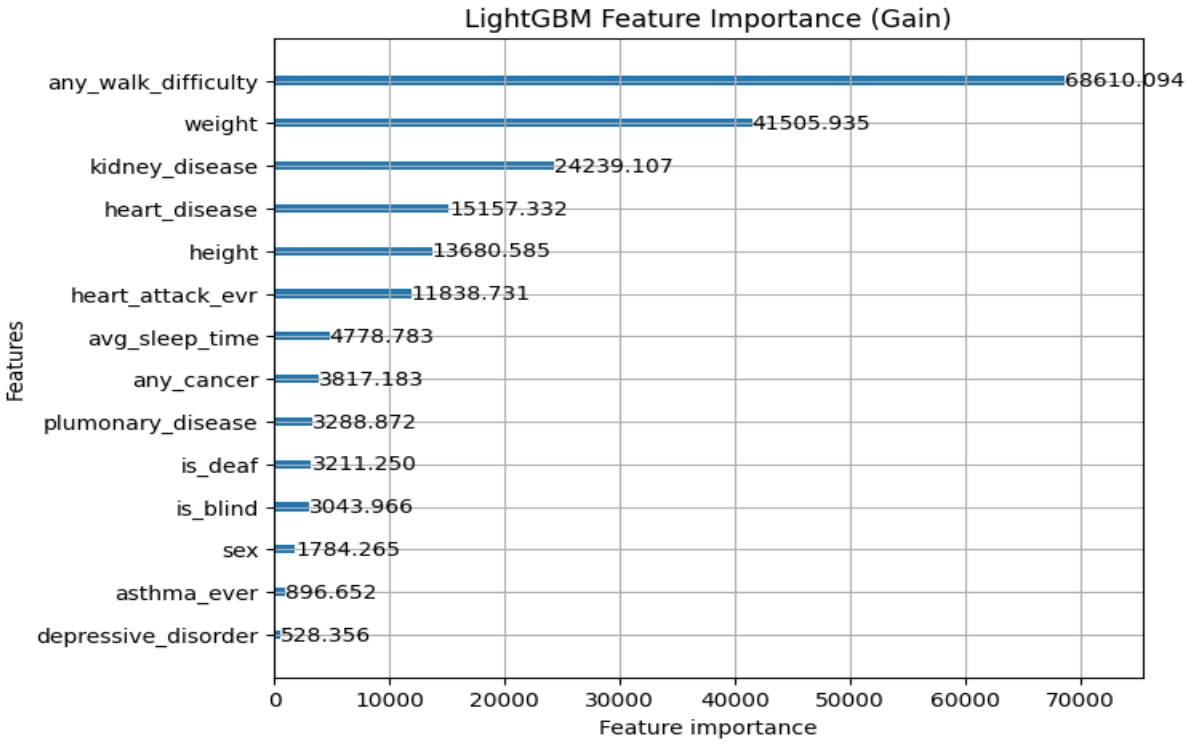
4.2.3 Results

Algorithm	Accuracy	Precision	Recall	F1	run time
Random Forest Classifier	0.84	0.8	0.84	0.81	1m 23s
Categorical Naive Bayes	0.85	0.81	0.85	0.82	0.9s
LightGBM	0.865	0.831	0.865	0.81	3.7s
Decision Tree Classifier	0.81	0.79	0.81	0.8	5.2s

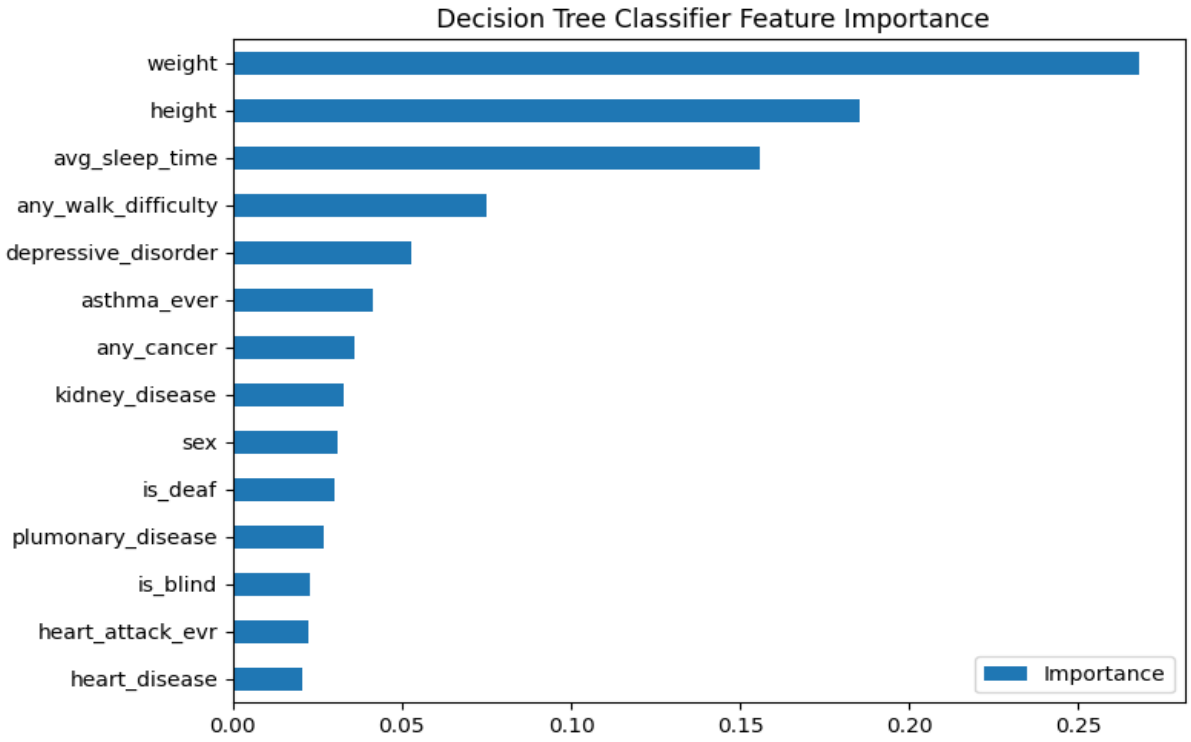
Table 4.2: Models accuracies on BRFFS dataset

Among the models, **LightGBM** achieved the highest accuracy at 0.865, with strong precision (0.831) and recall (0.865), while maintaining an efficient runtime of 3.7 seconds. This performance makes LightGBM particularly appealing for high-accuracy applications on large datasets, especially when speed is crucial. **Categorical Naive Bayes**, the fastest model at 0.9 seconds, demonstrated competitive performance with an accuracy of 0.85, making it a reasonable choice for scenarios prioritizing speed over marginal accuracy gains.

4.2.4 Feature importance



The above plot represents the feature importance for LightGBM model, which had the best accuracy on test data. As it can be seen, the existence of *any walk difficulty* and the *weight* of the patient emerges as the most influential predictors for this model. On the other hand, features like *depressive disorder*, *asthma* diagnosis and the *sex* of the patient have lower importance, implying a lesser impact on the predicted result. This plot aids in understanding model behavior and guiding feature selection.



4.2.5 Model improvements

4.2.5.1 Class weights

One possible obstacle for the current dataset is that the classes are not well balanced. *61 158* of the responders have diabetes, while the rest of *382 887* does not have. In the below table there are reported newly computed accuracy values for the models that support ***class_ weight*** parameter.

Algorithm	Accuracy	Precision	Recall	F1	run time
Random Forest Classifier	0.80	0.78	0.80	079	1m 18s
LightGBM	0.7	0.83	0.7	0.74	3s
Decision Tree Classifier	0.74	0.78	0.74	0.76	2.8s

Table 4.3: Models accuracies on *balanced* BRFFS dataset

Looks like the ***class_ weight*** parameter does not bring any improvement on the accuracy values: LightGB has the same accuracy, but RF and DT classifiers values decreased. At the same time, there can be seen real improvements for the run time where in 2 out of 3 cases the values halved.

4.2.5.2 Feature selection

For feature selection, I will use ***SelectKBest*** class from *sklearn* Python package which selects features according to the k highest scores. From the total of 14 parameters, will be selected the best 5 and then the best 10 features.

- Best 5 features ('weight' 'height' 'is_deaf' 'is_blind' 'any_walk_difficulty')

Algorithm	Accuracy	Precision	Recall	F1	run time
Random Forest Classifier	0.85	0.79	0.85	0.81	49s
LightGBM	0.862	0.82	0.86	0.8	2.4s
Decision Tree Classifier	0.84	0.79	0.84	0.81	0.9s

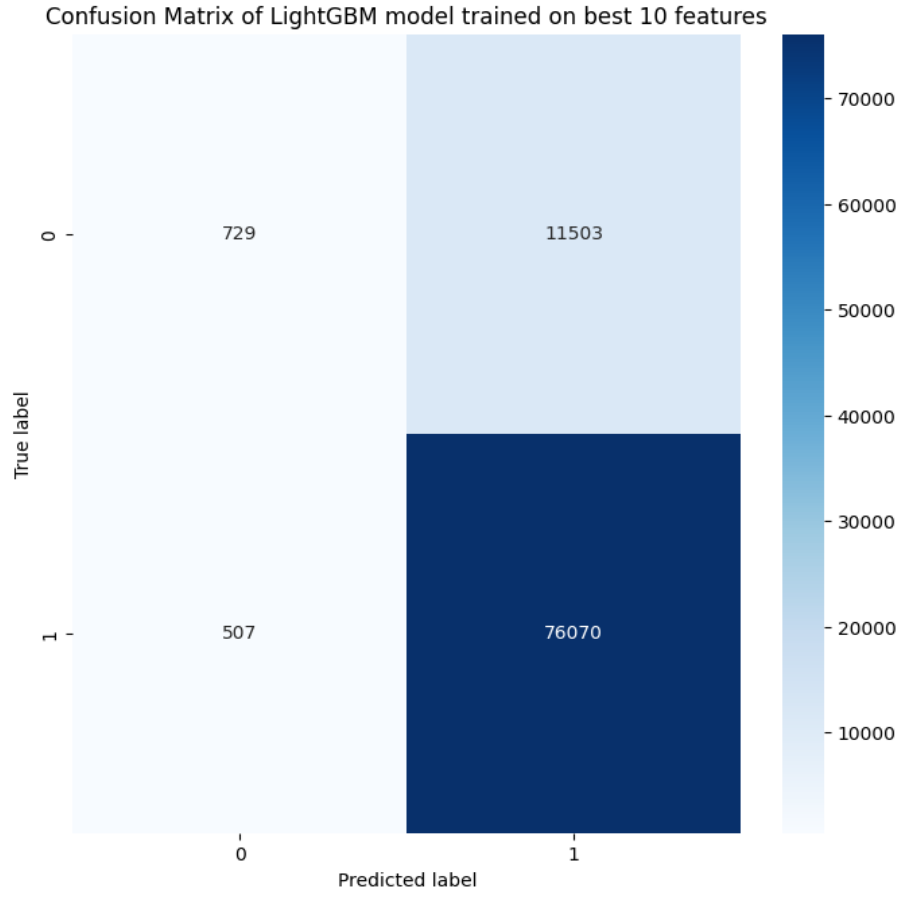
Table 4.4: Models accuracies after selection of best **5** features

- Best 10 features ('heart_attack_evr' 'asthma_ever' 'any_cancer' 'plumonary_disease' 'kidney_disease' 'weight' 'height' 'is_deaf' 'is_blind' 'any_walk_difficulty')

Algorithm	Accuracy	Precision	Recall	F1	run time
Random Forest Classifier	0.84	0.80	0.84	0.81	1m 9s
LightGBM	0.864	0.83	0.86	0.81	2.9s
Decision Tree Classifier	0.83	0.80	0.83	0.81	1.6s

Table 4.5: Models accuracies after selection of best **10** features

From the results, looks like both Random Forest and Decision Tree achieve their best ever accuracy on the 5 feature selection. Increasing the number of features to 10 or even the initial 14 have a negative impact on these two models. On the other hand, with the Light Gradient Boosting Classifier cannot be reached its initial 0.865 accuracy value by feature selection.


Confusion Matrix of LightGBM model trained on best 10 features
729
11503
507
76070
True label
Predicted label
0
1
0
1
70000
60000
50000
40000
30000
20000
10000

Chapter 5

Social impact

By predicting the risk of diabetes based on key health indicators, the project could enable early intervention, reducing the long-term health costs associated with untreated diabetes. It has the potential to allow access to health assessments for everyone, particularly in areas with limited access to healthcare professionals. Moreover, such predictive models can empower individuals by providing them with personal health information that could lead to lifestyle changes, ultimately promoting healthier societies. However, the project also presents challenges related to data privacy and the equitable distribution of health benefits.
All in all, the social impact that the application can be summarized as follows:

- **Early detection of diabetes**
- **Make health assessments more accessible**
- **Educate people regarding lifestyle**

5.1 Clinical use

The tool should not replace healthcare professionals but serve as a supportive tool to guide decision-making. The ethical responsibility lies in how the model's predictions are interpreted and used-decisions should be made in collaboration with medical professionals, rather than relying solely on the algorithm's output.

5.2 Ethical challenges

However, the project also presents challenges related to data privacy. Accessing diabetes patients data to train models presents a significant challenge, particularly due to regulations like the GDPR. These laws protect individuals' privacy and restrict how personal health data can be collected, stored, and used. As a result, obtaining large, diverse datasets for training machine learning models becomes difficult, limiting the ability to develop accurate predictive models.

Bibliography

[1] 2022 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2022.html.

[2] Diabetes dataset. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset.

[3] Muhammad Exell Febrian, Fransiskus Xaverius Ferdinan, Gustian Paul Sendani, Kristien Margi Suryanigrum, and Rezki Yunanda. Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216:21–30, 2023. 7th International Conference on Computer Science and Computational Intelligence 2022.

[4] Aishwarya Mujumdar and V Vaidehi. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165:292–299, 2019. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

[5] UCI Machine Learning repository. Pima indians diabetes database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[6] Gupta M Sampada GC Saxena R, Sharma SK. A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. *Computational intelligence and neuroscience*, 2022:3820360, 2022.