Optimizing Diabetes Risk Prediction: Comparative Analysis of Machine Learning Models and Feature Importance

Ana-Maria Păpară[0009−0009−7443−0170]

Faculty of Mathematics and Computer Science, Babeş-Bolyai University
ana.maria.papara@gmail.com

Abstract. Diabetes is known to be one of the most common diseases in the world today. A significant alternative to traditional diagnostic methods that can be time and resource-consuming is an ML-based approach. With the use of historical health lifestyle and diagnostic data, artificial intelligence can predict someone's risk of developing diabetes. This study aims to explore the potential of ML models for the prediction of diabetes based on different health and lifestyle factors. Furthermore, using a feature importance analysis, the objective is to determine the factors that affect the most diabetes condition. By evaluating multiple methods, configurations and datasets, current study aims to identify the most effective approach for predicting diabetes risk. For training and prediction, the following 4 models were utilized: Random Forest Classifier (RFC), LightGBM, Decision Tree (DTC) and Naive Bayes. The performance of the models was evaluated on the basis of accuracy, precision, recall and F1-score. The study involved testing these models on two datasets: a smaller one that contains medical attributes (PIMA dataset) and another much larger that incorporates mostly lifestyle factors (derived from the Behavioral Risk Factor Surveillance System Survey from 2022). The results show that the LightGBM and Random Forest models outperform the other evaluated classifiers. LightGBM achieved the highest accuracy of 86% in the larger dataset, highlighting the effectiveness of incorporating lifestyle factors. Furthermore, feature importance analysis revealed key predictors, such as weight, height and mobility difficulties to be those that impact diabetes diagnostic the most.

Keywords: Diabetes · ML · Feature importance · Prediction.

1 Introduction

Diabetes is a chronic health condition that affects millions of people around the world, often leading to severe complications if not diagnosed or poorly managed. Early detection is crucial for effective treatment and better patient outcomes. Traditional diagnostic methods are based on laboratory tests, which can be time-consuming and resource-intensive. Using Artificial Intelligence (AI), this study aims to develop an intelligent system that can predict the onset of diabetes using readily available clinical measurements, thus helping healthcare providers in early diagnosis and intervention.

The core problem addressed in this study is early and accurate prediction of diabetes using available patient diagnostic data, 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 problems, improving patient outcomes and reducing healthcare costs. By training a Machine Learning (ML) model on historical patient data, the algorithm can learn the characteristics of people at risk of 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.

Most studies addressing diabetes prediction [9], [7], [10], [13] rely on small datasets, such as PIMA [3], which are limited both in the number of attributes and the number of samples. In addition, these studies often overlook the significance of data augmentation techniques for training datasets and the methods for attribute analysis and selection. Consequently, this research aims to evaluate the performance of ML models trained on larger datasets, to investigate the importance of various attributes and to answer the following research questions:

RQ1: How do different machine learning models (RFC, Gaussian Naive Bayes, DTC, LGBM) compare in terms of accuracy, precision, recall, F1 score and run time for diabetes prediction?

RQ2: What features are the most influential in predicting diabetes according to the Decision Tree Classifier and LightGBM models?

RQ3: How does the performance of machine learning models change when trained on oversampled (using SMOTE and ADASYN) or class-weighted datasets compared to the original?

RQ4: How does the selection of features (all features, SelectKBest and model-based feature importance) impact the performance of machine learning models in the prediction of diabetes?

The main contribution of the current research is an **in-depth analysis and comparison** of different models' performances, including detailed evaluation metrics and comparison on different dataset sizes. The second contribution is the **novelty of the dataset**. This study introduces a new data set derived from a large US archive that has not been used before in the prediction of diabetes. Although most of the previous studies have relied on well-known public datasets, the current approach provides novel insights and uncovers new patterns. Another original contribution concerns the **feature importance analysis**. The current study analyzes the importance of the features involved in identifying the key factors that influence diabetes. This contributes to the existing literature by offering new information on diagnostic measurements that can significantly impact the risk of developing diabetes in patients.

This work is organized into five sections as follows: section 1 provides an overview of the research problem, its significance, the objectives and contributions of this study, section 2 reviews relevant studies from the literature on diabetes prediction using artificial intelligence and ML models, section 3 details the proposed approach and the steps followed. Here, the results of the experiments and their interpretation are also depicted. In the end, section 4 summarizes the important results and conclusions of the article and highlights the strengths and weaknesses of the approach of this study.

2 Literature review

Diabetes prediction through data analysis has gained significant attention due to the increasing potential of ML and AI. This section reviews some previous articles that focused on diabetes risk predictions using AI approaches and techniques.

In [9], the Pima Indians Diabetes Database (PIMA) was used to train the ML model. A detailed description of the PIMA dataset can be found in subsection 3.2. The main objective of this research was to compare various ML algorithms to see which one suits the best for the prediction of diabetes based on several health attributes. The researchers compared the accuracies of two different ML algorithms: **K-Nearest Neighbor**, which is based on supervised learning techniques and **Naive**

Bayes, a probabilistic classifier based on the Bayesian theorem. Both models were trained on eight health-related attributes to predict the presence of diabetes as a binary outcome. Eight experiments were conducted that differed by the split factor for the test and training data. The percentage of training data ranges from 80 to 10, in each experiment decreasing by 10. To evaluate the models, researchers used accuracy, precision and recall and then compared the results for the models. The best results were obtained in the first experiment that had a split of 80-20 for training and test. In this case, the KNN algorithm produced an accuracy of 77.92%, while Naive Bayes had an accuracy of 78.57%. Regarding the other experiments, most of the time, the Naive Bayes obtained better results than the KNN, having an average accuracy of 76.07%, while the KNN only 73.33%.

The authors of [7] also used the PIMA dataset, but removing patients with missing values. Remaining with only 392 observations, the features in the data set are: number of pregnancies, glucose level, diastolic measure of blood pressure, triceps (value measured in millimeters that estimate body fat), insulin, body mass index, age, history of diabetes in family and a binary variable for diabetes test. The main goal of the study was to identify the critical health aspects that impact diabetes in a patient. In addition to this, the researchers attempted to improve the accuracy of the prediction by training various ML algorithms and comparing the results. Three studies with three different algorithms for binary classification have been conducted: Logistic Regression (LogReg), Support Vector Machines (SVM) and Random Forest Classifier (RFC). The data set was divided into train and test in a ratio of 67% and 33%. Because the data set was imbalanced, the researchers chose to evaluate the models using only F1 and recall metrics. RFC was the most ideal algorithm for predictions with an accuracy of around 79.1% and having the highest true positive and true negative values. The feature importance analysis performed on RFC highlighted that **glucose** has the highest importance with 0.21, followed by **insulin** with 0.19 and **age** with 0.13 score.

A study [10] from 2022 uses a private data set collected using offline and online forms between 2019 and 2021 in India. The form was designed with the help of medical field specialists and has been distributed to cover as many areas as possible. The resulting data set consists of 1939 records and 11 features. In addition to the features from PIMA dataset, this one includes more lifestyle parameters such as smoking and drinking habits, along with patient weight, height and variables related to urination (number of times of passing urine in a day/night), thirst (number of times of drinking water in a day/night) and fatigue (if the subject feels fatigued or not). This study aimed to develop a model based on ML and ensemble learning techniques to predict diabetes using the collected data. To enhance the accuracy of the predictions, the researchers combined different ensemble techniques and ML models. Using the Bagging method, they analyzed Bagged Decision Tree (bagged DT), RFC and Extra Tree algorithms. AdaBoost and Stochastic Gradient Boosting have been used through the Boosting method. Then, finally, LogReg, SVM and DTC have been used via Voting method. Performance of the models was evaluated using metrics like training accuracy, testing accuracy, miss classification rate and running time. Among all models, the Bagged DT algorithm achieved the highest testing accuracy rate of 99.14% followed by Stochastic Gradient Boosting with 98.45% and RFC with 93.64%. In terms of runtime, the boost models performed the best, both running in 0.03 seconds. A statistical feature importance analysis showed that all features, except *sex* significantly contribute to outcome prediction, with **urination**, **weight** and **thirst** being the most influential.

Other study [13] performed in China in 2024 uses a data set from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), which includes 30,691 records and 14 key variables such as age, gender, BMI, high blood pressure, high cholesterol, smoking, stroke, exercise, fruit and vegetable consumption, alcohol use, difficulty walking, education and income. The primary goal of

this research is to analyze the factors that contribute to diabetes and determine their statistical significance. The study employs a Binary LogReg model to evaluate the relationship between the selected variables and the presence of diabetes. This statistical approach allows the identification of significant predictors by examining how each factor influences the likelihood of developing diabetes. The author found that the factors that have the most significant impact on diabetes risk are: **stroke** (74.33%), **walk difficulty** (73.64%) and **high blood pressure** (67.08%). In contrast, exercise, fruit consumption, alcohol intake and education are the features with the lowest impact on diagnosis.

3 Research design and findings

3.1 Overview

The current study aims to evaluate and compare the performance of multiple ML models, including Random Forest Classifier (RFC), Naive Bayes, Decision Tree Classifier (DTC) and LightGBM (LGBM) [11]. Each model is trained and tested on a structured data set containing diagnostic and lifestyle characteristics, in an attempt to assess its effectiveness in predicting diabetes (following the pipline presented in Figure 1). To ensure comprehensive performance, the models are evaluated based on the following metrics: **accuracy** (overall correctness of the predictions), **precision** (proportion of correctly predicted positive cases), **recall** (model's ability to detect actual positive cases), **F1** score and **run time**.

The first experiment evaluates the following ML models: RFC, Gaussian Naive Bayes, LGBM and DTC on their ability to predict diabetes risk based on the PIMA Indians Diabetes dataset, which consists of 768 records. The objective was to determine the best performing model comparing them using evaluation metrics. In the second experiment, a much larger size data set is used to determine if the models maintain their predictive accuracy. The new data set contains over 445,000 records, but it involves a completely different set of attributes compared to PIMA. Dealing now with more complex data, the experiments were structured according to different key dimensions: **number of features** (all 15 features, features selected by **SelectKBest** before model building, features identified based on their importance in the predictive model), **data variations** (original data set, data oversampled using SMOTE, data oversampled using ADASYN), **training approach** (standard training, training with class weights).

Understanding the impact of individual characteristics on model prediction is important for clinical relevance. This study employs a feature importance analysis in two selected models: DTC and LGBM. These models have built-in methods to provide scores for each independent feature based on how much the feature contributes to the target outcome. By integrating feature selection with model evaluation, the objective is to determine the optimal combination of ML techniques for an accurate and efficient prediction of the risk of diabetes.

3.2 Data description

PIMA dataset. The PIMA Indians Diabetes Dataset [3] was collected by the US National Institute of Diabetes and Digestive and Kidney Diseases from a population of 768 women over 21 years of age, of Pima Indian heritage, in the late 1990s. The outcome tested was diabetes, 258 tested positive and 500 negative. Therefore, there is one target variable (dependent) and eight other attributes that describe the health measurements presented in Table 1.

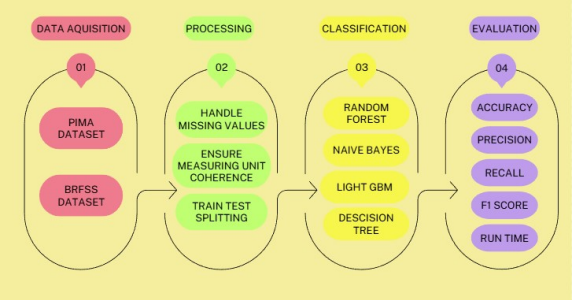


Fig. 1: Algorithm steps details

BRFSS dataset. This data set was obtained manually by filtering and processing the 2022 BRFSS Survey Data [1]. The **Behavioral Risk Factor Surveillance System (BRFSS)** is a large-scale ongoing health survey conducted by the Centers for Disease Control and Prevention (CDC) in the United States. Established in 1984, it collects self-reported data on health-related risk behaviors, chronic diseases and preventive health practices from adults through telephone interviews. BRFSS covers topics such as **smoking**, **alcohol use**, **physical activity**, **diet** or **chronic conditions**, making it one of the most comprehensive sources of public health data. The data set is widely used for epidemiological research, policymaking and public health interventions. To prepare data from the 2022 BRFSS survey for the prediction of diabetes, the data set was first filtered to retain only the most relevant characteristics. The survey results comprise 445,132 records and include 326 variables. Given the large number of variables, a careful selection process was applied to keep the attributes that are the most likely to impact the risk of diabetes. This selection was based on previous research, focusing on factors that may be relevant in the prediction of diabetes. Filtering and transformation were performed using Python and the Pandas library. The raw data set, initially in an XPT transport format, was processed into a clean CSV file that contained only the selected attributes. Unnecessary columns were dropped and categorical variables were standardized for consistency. The final version of the data set consists of 444,046 instances (after preprocessing) and 15 attributes that are detailed in Table 1. In addition to the information provided in Table 1, all categorical variables contain a third value: "7" that encloses "don't know / not sure" and "refuse" answers.

The filtering process involves excluding all attributes from the original BRFSS survey data that: **cannot have any impact on patient diabetes status** (respondent's residence, source of health insurance, educations, screening history, 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 too many missing values** (smoking habits, ever had an H.P.V. vaccination, respondent's race, life satisfaction, stress). In the next step, on the just obtained dataset, the following changes were applied: (1) missing values for the categorical variables were marked as "don't know / not sure" answers, (2) missing values for the numerical variables were replaced by the average value of the existing values, (3) refused answers were also marked as "don't know / not sure" in order to have fewer labels, (4) for the output variable (has diabetes) were stored only two values: *positive* and *negative* as prediabetes, borderline and pregnancy diabetes were treated as no diabetes and the missing values, refused and "don't know" answers were dropped, (5) weight values were converted to kilograms and (6) height values were converted to centimeters.

Reference			[9]	[7]	[10]	[13]	new
Dataset			PIMA	PIMA	private	BRFSS	BRFSS
#samples			768	392	1939	30691	445132
#features			8	8	10	14	14
pregnancies	numeric	[0, 17]	✓	✓			
glucose	numeric	[0, 199]	✓	✓			
blood pressure	numeric	[0, 122]	✓	✓			
skin thickness	numeric	[0, 99]	✓	✓			
insulin	numeric	[0, 864]	✓	✓			
BMI	numeric	[0, 67.1]	✓	✓		✓	
age	numeric	[21, 81]	✓	✓	✓	✓	
history of diabetes	numeric	[0.08, 2.42]	✓	✓	✓		
gender	categorical	male / female			✓	✓	✓
smoking	categorical	Yes / No			✓	✓	
drinking	categorical	Yes / No			✓	✓	
weight	numeric	[15, 160]			✓		✓
height	numeric	[60, 250]			✓		✓
urination	numeric	[2, 15]			✓		
thirst	numeric	[1, 15]			✓		
fatigue	categorical	Yes / No			✓		
high blood pressure	categorical	Yes / No				✓	
high cholesterol	categorical	Yes / No				✓	
stroke	categorical	Yes / No				✓	
exercise	categorical	Yes / No				✓	
fruit consumption	categorical	Yes / No				✓	
vegetable consumption	categorical	Yes / No				✓	
difficulty walking	categorical	Yes / No				✓	✓
education	categorical	1/2/3/4/5/6				✓	
income	categorical	1/2/3/4/5/6/7/8				✓	
sleep time	numeric	[1, 24]					✓
heart attack ever	categorical	Yes / No					✓
heart disease	categorical	Yes / No					✓
asthma	categorical	Yes / No					✓
any cancer	categorical	Yes / No					✓
pulmonary disease	categorical	Yes / No					✓
depressive disorder	categorical	Yes / No					✓
kidney disease	categorical	Yes / No					✓
is deaf	categorical	Yes / No					✓
is blind	categorical	Yes / No					✓
Data - output			0 / 1	0 / 1	0 / 1	0 / 1	0 / 1
ML performance (KNN)			77.92				
ML performance (Naive Bayes)			78.57				85.00
ML performance (LogReg)				79.10		72.80	
ML performance (SVM)				79.70	89.51		
ML performance (RFC)				84.00	93.64		84.00
ML performance (Stochastic GB)					98.45		
ML performance (Bagged DT)					99.14		
ML performance (DTC)							81.00
ML performance (LGBM)							86.50
Feature importance analysis				✓	✓	✓	✓

Table 1: Comparison of datasets and results between literature studies and the current one

3.3 Numerical experiments

Experiment 1 - PIMA dataset On the PIMA data set, the models were evaluated based on accuracy, precision, recall, F1 score on the test data. The runtime was also used to compare the models. The runtime is measured in milliseconds and includes only the model training time. The evaluation of ML models on the PIMA dataset is presented in Table 2 and reveals that the RFC achieved the highest accuracy (74%), along with strong precision, recall and F1 scores, making it the best performing model in terms of predictive power. Additionally, it maintained a low run-time of 359 milliseconds (0.4 seconds), balancing both efficiency and effectiveness. LGBM followed closely with an accuracy of 72%, although it had the longest run time of 515 milliseconds (4.6 seconds), making it computationally more expensive. The Gaussian Naive Bayes model demonstrated a slightly lower accuracy of 69%, but performed comparably in precision and had a great run time performing the training in only 15 milliseconds. Overall, the RFC stands out as the most effective model for the prediction of diabetes on this data set, although further improvements can be made with parameter tuning and additional data.

Algorithm	Accuracy	Precision	Recall	F1	train time (ms)
Random Forest Classifier	0.74	0.73	0.73	0.73	359
Gaussian Naive Bayes	0.69	0.70	0.69	0.69	15
LightGBM	0.72	0.72	0.72	0.72	515
Decision Tree Classifier	0.66	0.65	0.66	0.65	15

Table 2: Model evaluation metrics on PIMA dataset

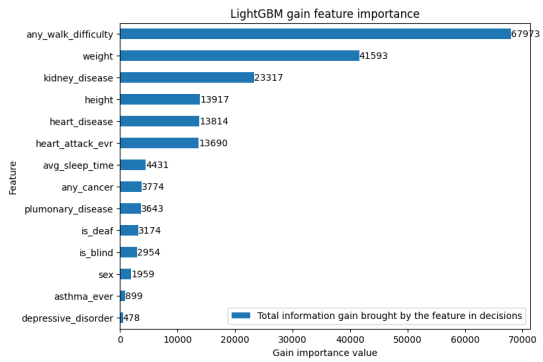
Experiment 2 - BRFSS dataset. The results obtained on the BRFSS dataset depicted in Table 3 (column 3) indicate a noticeable improvement in models performance compared to the PIMA dataset, probably due to the larger dataset size and richer feature set. LGBM emerged as the best performing model, achieving the highest accuracy of 86%, making it the most reliable for the identification of diabetes cases. Interestingly, the Categorical Naive Bayes closely followed with an accuracy of 85%, demonstrating that a probabilistic approach can be highly effective when the dataset contains categorical features. Despite its strong performance, the RFC, which has an accuracy of 84%, took significantly longer to run, suggesting that while it is a robust model, it may not be the most efficient choice for large-scale applications. Unlike in the PIMA dataset experiment, where DTC struggled, here the DTC performed reasonably well with 81% accuracy, suggesting that the dataset's structure allows for better boundary learning. Another interesting observation is that LGBM, despite its superior performance, ran significantly faster (3.7 seconds) than RFC, highlighting the advantage of gradient boosting techniques in balancing both speed and accuracy. The differences in model rankings between the two datasets reinforce the idea that model performance is highly data dependent, highlighting the importance of dataset-specific tuning rather than a one-size-fits-all approach.

Experiment 3 - Feature importance analysis. Feature analysis is an important step in results analysis because it helps determine which variables have the most influence on model predictions. Understanding the impact of each feature allows for easier understanding of the model, improved performance and potential feature selection. In addition, analysis of feature importance has clinical relevance because it helps identify which factors contribute the most to the risk of disease. In this study, a feature importance analysis was conducted using RFC, LGBM and DTC. Because Naive Bayes models simplify the computation and make the algorithm more efficient by assuming feature independence, it is known to be not suitable for feature importance analysis.

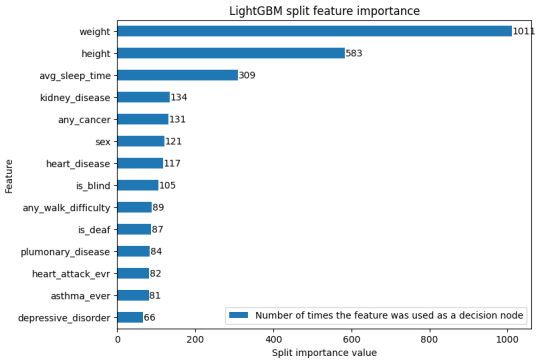
LGBM provides importance scores based on how often a feature is used for splitting and how much it improves the ability of the model to predict the correct results. Two primary metrics [2] are used: *split importance*, which counts the number of times a feature is used as a decision node and *gain importance*, which measures the total gain of information provided by the feature across all splits. In other words, *gain* value measures how much better the split makes predictions compared to before the split, while *split* value measures how frequently the feature was chosen as the splitting criterion. However, a high split count does not necessarily mean that the feature is highly influential; it might just be frequently used. So, gain-based importance is generally more reliable, as it reflects both frequency and impact on model performance. On the other hand, the **DTC** assigns importance scores based on how effectively each feature reduces uncertainty (impurity) in the dataset when making splits. The features that contribute the most to distinguishing between diabetic and non-diabetic individuals receive higher scores, while those with minimal impact are ranked lower.

Similarly to DTC, the **RFC** determines the importance of the features by measuring how much each contributes to reducing the impurity in all trees in the ensemble. Since RFC aggregates results from multiple decision trees, its feature importance scores are more stable and reliable than those from a single decision tree.

The feature importance analysis on **LGBM** presented in Figure 2a and Figure 2b reveal which features are the most commonly used to divide the data (split values) and also which features are most valuable to improve the accuracy of the model (gain values). In terms of gain importance, *any_ walk_ difficulty*, *weight*, *kidney_ disease* and *height* play a key role in improving the model accuracy. This suggests that mobility limitations are strongly correlated with the risk of diabetes. On the other hand, *weight*, *height* and *sleep time* are highlighted as the most frequently used features in the decision-making process, having the highest split importance values. However, **weight** and **height** are consistently important in both types of importance, while **depressive disorder** and **asthma ever** appear to be less influential in both split and gain importance.



(a) LGBM GAIN feature importance

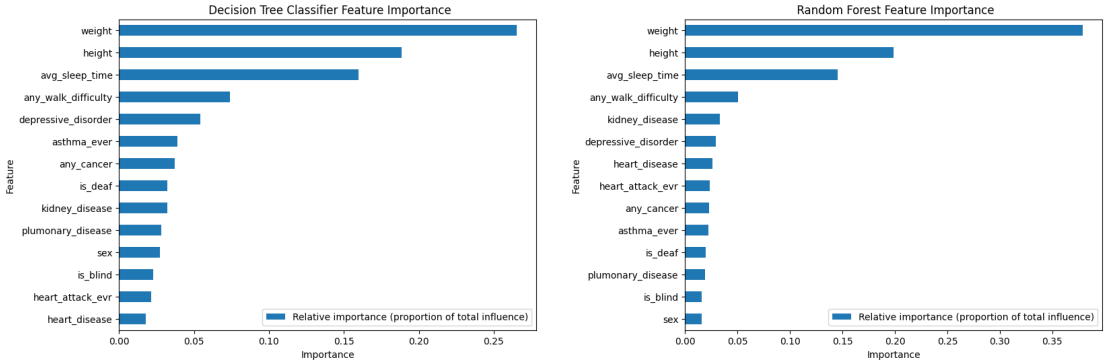


(b) LGBM SPLIT feature importance

Fig. 2: Feature importance - GAIN and SPLIT

On the other hand, the **DTC** produced a different ranking of important features, exposed in Figure 3a. The most influential attributes according to this model were *weight*, *height*, *avg_ sleep_ time*, *any_ walk_ difficulty* and *depressive_ disorder*. Unlike LGBM, which emphasized chronic conditions, DTC placed greater importance on average sleep time and depressive disorder, indicating that lifestyle and mental health factors had a stronger influence on the model output. Weight and height remained among the top features in both models, confirming their relevance in the assessment of diabetes risk.

Feature importance scores from **RFC**, highlighted in Figure 3b, are very similar to the DTC scores, at least in terms of the first part of the ranking. Although the first 4 most influential features remain the same in these 2 models, it can be observed that they have different scores with *weight* counting more in Random Forest. These similarities are due to the fact that both models use impurity reduction to determine splits, so the most important features according to DTC will also tend to be important in the RFC. It is also worth mentioning that in the RFC feature ranking, the *kidney_ disease* takes a higher place, defending *depressive_ disorder*, a fact that is also met in LGBM ranking.



(a) DTC feature importance

(b) RFC feature importance

Fig. 3: Feature importance

The key differences between the four different approaches for feature importance analysis are the way each model learns and evaluates the contributions of the feature. LGBM, as a boosting algorithm, builds multiple weak learners sequentially, refining feature importance by prioritizing those that consistently reduce error across iterations. In contrast, DTC and RFC rely on building individual trees or ensembles of trees, where feature importance is based on how often features are used in the reduction of impurity. The differences in value ranges are because in RFC and DTC (both imported from scikit-learn library) the importance values are, by default, normalized so that they all sum to one. Regarding the results, it can be observed that features like **weight** and **height** are consistently ranked as the most important in all models, indicating their strong influence on the prediction of diabetes. There is a strong correlation between split importance from LGBM, DTC and RFC feature importance. All three approaches are rating *weight*, *height* and *sleep time* as the features with the highest impact on the output variable. However, it is interesting to see how *depressive disorder* is placed in the top 6 in both DTC and RFC and, at the same time, has the lowest value in the split importance of LGBM.

The findings of this feature importance analysis can be compared with those of similar studies in the literature. The study [13] uses a dataset derived from the 2015 BRFSS survey, but there are notable differences in the feature filtering process compared to the current study. These differences arise due to variations in data filtering, leading to different sets of features being used in each study. As shown in **??** the only common features of both datasets are *gender* and *difficulty walking*. In study [13] *walking difficulty* was identified as the second most important feature, followed by *stroke*. In contrast, in the current study, LGBM gain-based feature importance (Figure 2a) ranked *walking difficulty* as the most influential factor.

Another relevant study [10], although based on a different dataset, includes three common features with the current research: *gender*, *weight* and *height* (as can be observed in Table 1). In that study, weight was assigned an importance of 23%, being the second most important feature after urination. Height had a minimal influence of 0.03% and gender was found to be completely insignificant (0% importance). In comparison, if the LGBM gain-based feature importance values (Figure 2a) were normalized to sum 1, weight would be similarly influential at 21.26%, while height played a much larger role at 7.11%. Additionally, while gender was still relatively irrelevant in both

studies, it held slightly higher weight in the current analysis (1.00%). These variations suggest that although some feature rankings remain consistent across studies, differences in dataset composition, feature engineering and model selection can significantly affect the perceived importance of individual variables.

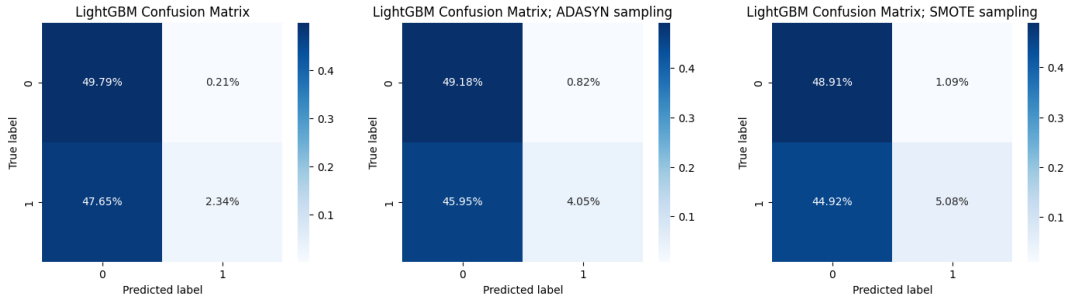
Experiment 4 - Class weights One of the challenges in training ML models on the BRFSS dataset is the imbalance between the two classes. Out of 444.045 total respondents, only 61.158 have diabetes, while the remaining 382.887 do not. This imbalance can lead to biased models that favor the majority class, reducing the model's ability to predict diabetes. To avoid this problem, the **class weighting** was applied. Class weighting is a technique that adjusts the model's learning process by assigning higher importance to underrepresented classes. [5].

This ensures that the contribution of each class is balanced in the model learning process. The adjusted weights penalize misclassifications of the minority class more heavily, helping the model pay more attention to the underrepresented class. In Table 3 (column 3) are described new computed evaluation metrics for the models that support *class_ weight* parameter. After applying class weights to the models that support them (RFC, DTC and LGBM), results showed that this approach did not significantly improve accuracy. RFC maintained nearly the same accuracy of 80%, while both LGBM and DTC saw a slight decrease in performance. However, an interesting observation is the significant reduction in terms of run-time. The training time was almost halved for LGBM and DTC. This suggests that class weighting may introduce regularization effects, reducing overfitting and allowing models to converge faster.

Experiment 5 - Resampling Considering the fact that applying class weights did not improve the performance of the models, another feasible solution is **resampling**. This method involves adjusting the balance between minority and majority classes through up-sampling or down-sampling. In the case of an imbalanced dataset, **oversampling** involves generating synthetic records for the minority class while **undersampling** entails removing rows from the majority class. Because undersampling can discard valuable data from the majority class, leading to a loss of potentially important information, it is best to choose oversampling. **SMOTE (Synthetic Minority Oversampling Technique)** [8] is a popular oversampling technique that works by identifying the k-nearest neighbors in the feature space for each minority class instance and then creating synthetic examples starting from the original instance and using the difference between the instance and its neighbors. **ADASYN (Adaptive Synthetic Sampling)** [12] is an improved version of SMOTE. What it brings over SMOTE is that it adds a random value in the computation process so instead of all the samples being linearly correlated to the parent they have a little more variance.

In order to ensure a consistent comparison, a single test set was used to evaluate the models trained on different data. Firstly, a subset of 20,000 samples (10,000 positive and 10,000 negative) were held back for testing and the rest of 424,045 were collected in training set. Afterwards, a LGBM model was trained separately on the obtained training set (without any sampling), on the training set oversampled using ADASYN and on the training set oversampled using SMOTE. In the end, the models were evaluated on the balanced test set. The performance of the models was compared both based on confusion matrices (see Figure 4) and based on accuracy, precision, recall and F1 score as shown in Table 3 columns 4 and 5.

It can be observed in the matrices that there is a small improvement from the original model to the models trained on the oversampled data, especially in terms of true positives: increasing from



(a) Original data (b) ADASYN data (c) SMOTE data

Fig. 4: Confusion matrices for the LGB models trained on original, ADASYN and SMOTE data; tested on same test set.

2.34% to 4.05% in ADASYN and to 5.08% in SMOTE. Despite this improvement appearance, the measurements presented in Table 3 show a significant decrease in the five evaluation metrics.

Experiment 6 - Feature selection Feature selection can be a crucial step in ML aiming to reduce dimensionality, improve model accuracy and enhance model efficiency [6]. In this study, SelectKBest is used for feature selection along with a manual selection based on feature importance scores computed by DTC. **SelectKBest** is a univariate feature selection method that selects the top **k** features based on statistical scoring. SelectKBest is a filter-based feature selection method; the feature selection process is carried out independently of any specific ML algorithm [4]. Instead, it relies on statistical measures to score and rank features. This technique helps to retain only the most relevant features, eliminating those that have a minimal impact on the prediction of diabetes.

The feature selection process in the current study followed a structured pipeline: (1) **Dataset Preparation**: The data set used was the BRFSS after preprocessing steps. (2) **Feature Selection**: (a) SelectKBest was applied to identify the most predictive features based on statistical scores. (b) Separately, a DTC was used to calculate the importance of the feature and the first 5 features with the highest importance were selected. (3) **Model Training & Evaluation**: Three selected feature subsets (first 5 from SelectKBest, first 10 from SelectKBest, first 5 from Decision Tree Classifier) were used to train RFC, LGBM and DTC. Afterwards, the models were evaluated based on accuracy, precision, recall and F1 score.

SelectKBest was used in two experiments: the first time with the parameter *k=5*, results provided in Table 3 (column 6) and the second time with *k=10*, results provided in Table 3 (column 7). Furthermore, an experiment (results in Table 3, column 8) was conducted using only the subset of the first five features with highest importance computed by the DTC. So, the models were trained again only on the following features:

1. **Best 5 features computed with SelectKBest**: *any_walk_difficulty*, *is_blind*, *is_deaf*, *height*, *weight*
2. **Best 10 features computed with SelectKBest**: *any_walk_difficulty*, *is_blind*, *is_deaf*, *height*, *weight*, *kidney_disease*, *plumonary_disease*, *any_cancer*, *asthma_ever*, *heart_attack_ever*
3. **Best 5 features computed by Decision Tree Classifier**: *weight*, *height*, *avg_sleep_time*, *any_walk_difficulty*, *depressive_disorder*

While both feature selection techniques (SelectKBest and best features from DT) identified *weight*, *height* and *any_walk_difficulty* as key predictors, the variation in the remaining features indicates that different algorithms focus on different aspects of the data.

The results indicate that RFC and DTC achieved their highest accuracy when trained on the top five features selected by SelectKBest. However, increasing the number of features to 10 or using the complete set of 14 features (Table 3, column 2) led to a slight decline in performance for these models. However, LGBM's accuracy remained stable across different feature sets, demonstrating its robustness. Although it cannot reach its initial accuracy of 86.5%, LGBM maintains a high accuracy value of 86.2%.

Algo	BRFSS																																			
	Simple						Weights						Resample ADASYN						Resample SMOTE						5 SKB						10 SKB						5 DTC					
	acc	pr	rec	F1	run		acc	pr	rec	F1	run		acc	pr	rec	F1	run		acc	pr	rec	F1	run		acc	pr	rec	F1	run		acc	pr	rec	F1	run		acc	pr	rec	F1	run	
RFC	84	80	84	81	83.0s		80	78	80	79	78.0s		54	61	54	46	295s		54	61	54	46	315s		85	79	85	81	49.0s		84	80	84	81	69.0s		84	79	84	81	?	
Bayes	85	81	85	82	0.9s		-						65	66	65	65	1.1s		65	66	65	65	1.8s		-						-						-					
LGBM	86	83	86	81	3.7s		70	83	70	74	3.0s		53	67	53	41	8.5s		53	67	53	43	9.6s		86	82	86	80	2.4s		86	83	86	81	2.9s		86	81	86	79	?	
DTC	81	79	81	80	5.2s		74	78	74	76	2.8s		56	61	56	50	11.9s		55	60	55	49	11.8s		84	79	84	81	0.9s		83	80	83	81	1.6s		83	79	83	80	?	

Table 3: Models' performance in various scenarios. Each cell contains separate columns for accuracy, precision, recall, F1-score and training time.

4 Conclusions and future work

This study explored the efficiency of various ML algorithms for prediction of the risk of diabetes using two distinct data sets with different setups. Through extensive experimentation, the impact of class weighing, feature selection, and oversampling on model performance is analyzed. Moreover, a feature importance analysis carried out on two of the considered models. The results highlight that models such as LightGBM and Random Forest consistently outperform traditional classifiers in terms of accuracy and robustness. Additionally, feature importance analysis provided valuable insights about the most influential health attributes: *weight*, *height* and mobility issues (*any_walk_difficulty*) are the features that influence the most the prediction of diabetes.

Despite these promising results, some limitations remain. One key limitation is the potential for bias in the data set due to self-reported health information, which can introduce inaccuracies and inconsistencies. Furthermore, while the models performed well on the training data, their correctness to real-world clinical settings remains to be validated. Future work can focus on incorporating more advanced deep learning techniques and also validating the models on datasets obtained from healthcare institutions, which contain only medically verified diagnoses.

Ultimately, this research contributes to the development of more reliable diabetes prediction models, helping healthcare professionals make data-driven decisions and improving early detection strategies.

5 Acknowledgment

I would like to express my gratitude to Mrs. Laura Dioşan from Babeş-Bolyai University for the continuous guidance and assistance throughout this research.

References

1. 2022 BRFSS Survey Data and Documentation. https://www.cdc.gov/brfss/annual_data/annual_2022.html, last accessed: 2025-04-08
2. LightGBM Parameters webpage. https://lightgbm.readthedocs.io/en/stable/Parameters.html, last accessed: 2025-04-08
3. Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database, last accessed: 2025-04-08
4. Scikit-learn SelectKBest Homepage. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html, last accessed: 2025-04-08
5. Bakirarar, B., ELHAN, A.: Class weighting technique to deal with imbalanced class problem in machine learning: Methodological research. Turkiye Klinikleri Journal of Biostatistics **15**, 19–29 (01 2023). https://doi.org/10.5336/biostatic.2022-93961
6. Barbieri, M.C., Grisci, B.I., Dorn, M.: Analysis and comparison of feature selection methods towards performance and stability. Expert Systems with Applications **249**, 123667 (2024). https://doi.org/https://doi.org/10.1016/j.eswa.2024.123667, https://www.sciencedirect.com/science/article/pii/S0957417424005335
7. Dutta, D., Paul, D., Ghosh, P.: Analysing feature importances for diabetes prediction using machine learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) pp. 924–928 (2018). https://doi.org/10.1109/IEMCON.2018.8614871
8. Elreedy, D., Atiya, A.F.: A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. Information Sciences **505**, 32–64 (2019). https://doi.org/https://doi.org/10.1016/j.ins.2019.07.070, https://www.sciencedirect.com/science/article/pii/S0020025519306838
9. Febrian, M.E., Ferdinan, F.X., Sendani, G.P., Suryanigrum, K.M., Yunanda, R.: Diabetes prediction using supervised machine learning. Procedia Computer Science **216**, 21–30 (2023). https://doi.org/https://doi.org/10.1016/j.procs.2022.12.107, https://www.sciencedirect.com/science/article/pii/S1877050922021858, 7th International Conference on Computer Science and Computational Intelligence 2022
10. Ganie, S.M., Malik, M.B.: An ensemble machine learning approach for predicting type-ii diabetes mellitus based on lifestyle indicators. Healthcare Analytics **2**, 100092 (2022). https://doi.org/https://doi.org/10.1016/j.health.2022.100092, https://www.sciencedirect.com/science/article/pii/S2772442522000399
11. Geron, A.: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc., 2nd edn. (2019)
12. He, H., Bai, Y., Garcia, E., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the International Joint Conference on Neural Networks pp. 1322 – 1328 (07 2008). https://doi.org/10.1109/IJCNN.2008.4633969
13. Xu, P.: Research on the influence factors that possibly lead to diabetes. Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (2024), https://api.semanticscholar.org/CorpusID:274300013