Feature Selection for Gestational Diabetes Mellitus Prediction using XAI based AutoML Approach

Alia MAALOUL[1*], Meriam JEMEL[2] and Nadia BEN AZZOUNA[3]

[1,2] LR11ES03 SMART Lab, ISG Tunis, Le Bardo,
[3] ESSECT, Tunis, Tunisia, Université de Tunis
[*] Maaloul.alia@gmail.com

Abstract. Predicting Gestational Diabetes Mellitus (GDM) is crucial for pregnant women to enable regular monitoring of their blood sugar levels and adherence to a healthy diet. Early intervention can significantly lower the risk of developing this condition. To assess this risk, Machine Learning (ML) and Deep Learning techniques are employed. However, traditional ML models often face challenges in accurately predicting GDM risk due to the complex processing required to optimize their hyperparameters for the best performance. This study presents a feature selection for GDM prediction using AutoML-XAI techniques (Automatic Machine Learning – eXplainable Artificial Intelligence techniques) approach, which aims to automatically predict GDM risk as accurate as possible while providing meaningful interpretations of the predictive results used in feature selection. The AutoML models generated utilize a Kaggle dataset and several combinations of features selected based on their scores of importance determinated with XAI techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). The proposed approach of autoML and features selection with XAI techniques leads to the creation of a precise, efficient, and easily interpretable model which surpasses other machine learning models in predicting GDM risk without the need for human intervention. The scores of importance of features are involved in the feature selection process and multiple AutoML models are generated and assessed, with the optimal AutoML model being established automatically.

Keywords: IA, GDM,XAI, SHAP, AutoML.

1 Introduction

Gestational Diabetes Mellitus (GDM) is a form of diabetes that arises during pregnancy, affecting approximately 2-10% of pregnant women. This condition is marked by elevated blood sugar levels, which can pose risks to both the mother and her baby. Women with gestational diabetes face an increased likelihood of developing type 2 diabetes in the future, while their infants may experience complications such as macrosomia (excessive birth weight) and hypoglycemia (low blood sugar) [1]. Consequently, early prediction and diagnosis of gestational diabetes are essential for effec-

© The Author(s) 2025
S. Krichen et al. (eds.), *Proceedings of International Conference on Decision Aid and Artificial Intelligence (ICODAI 2024)*, Atlantis Highlights in Intelligent Systems 12,
https://doi.org/10.2991/978-94-6463-654-3_10

tive management and minimizing associated risks. Screening methods typically include a glucose challenge test or an oral glucose tolerance test, with heightened monitoring recommended for women at greater risk, such as those with a family history of diabetes or obesity. By effectively predicting and managing gestational diabetes, women can lower the risk of complications for themselves and their children [1]. As Artificial Intelligence (AI) techniques have advanced, a variety of increasingly sophisticated and effective prediction models have emerged in recent years [2]. While certain machine learning models, like decision trees and linear regression, are straightforward and offer reasonable accuracy that can be enhanced, others achieve high predictive accuracy but are challenging to optimize [2]. In fact, the optimization of ML models consists on Hyperparameter's optimization which involves selecting the optimal values for these parameters in order to obtain the best performance of the model. This process can be complex and time-consuming, as there are usually many possible combinations of hyperparameters to consider. The increasing complexity of data and the need for more accurate models has led to the emergence of Automated Machine Learning (AutoML) as a rapidly growing field. AutoML seeks to automate the process of implementing high-performance ML models and hyperparameters optimization on a given dataset, thereby reducing the need for manual intervention and improving the efficiency of model development and deployment [3]. Thus, AutoML offers a pipeline for the process of implementation of performant ML models based on key steps which include the Data Preprocessing phase during which AutoML can automate steps such as data cleaning and handling missing values. The next phase is the features and models selection phase which allows selecting important features and testing multiple algorithms machine learning to determine which one works best for a given dataset. Then the phase of optimizing and adjusting the models hyperparameters to improve their performance. After, AutoML can evaluate the performance of models using cross-validation techniques and other metrics. Finally, some AutoML tools include features to help interpret model results, which is essential for understanding how decisions are made. Furthermore, employing feature selection techniques can enhance the performance of predictive models. These techniques encompass embedded, filter, and wrapper methods; however, they come with certain drawbacks, such as increased computational costs and sensitivity to hyperparameters. Thus, the application of Explainable Artificial Intelligence (XAI) as an innovative method for feature selection was proposed; this approach seeks to pinpoint the key features that play a significant role in prediction, while also offering insights into the impact of these features on the model's outcomes [17].

Consequently, integrating explainable AI into feature selection process is crucial for predicting gestational diabetes mellitus (GDM), as it helps to clarify the importance of each feature in the system's decision-making process and boosts its effectiveness.

In this research work, the AutoML platform used is based on some explanation techniques used through the proposed process of features selection based on the score of importance of each feature. These scores indicate the percentage of the contribution of each feature in the prediction [4]. In this research work, the SHAP technique was used to determinate these scores of importance for explanation through the proposed process of features selection.

The aim of this paper is to propose a hybrid strategy of AutoML and XAI techniques to generate and evaluate several ML models in order to determinate the best accurate model based on a proposed feature selection method. This method proposes to eliminate features contributing to unfairness based on the feature attribution explanations of the model's predictions.

This approach of feature selection permits to examine the variable importance scores for a machine learning model. After that the process of generating and evaluating AutoML models is restarted after removing insignificant features based on their scores of importance.

The rest of this paper is structured as follows: section II presents current research trends and related works to AutoML and features selection methods, section III describes the process of the proposed strategy and exposes experimental setup and obtained results and finally the conclusion and some perspectives.

2 Litterature review

This section provides an overview of various related works that focus on features selection methods and highlights some important aspects of AutoML.

2.1 Features selection methods

Types of features selection methods

It has been demonstrated that feature selection is highly effective for handling high-dimensional data and improving learning efficiency in both theoretical studies and practical applications [9]. Through this method, a subset is extracted from the original feature set based on specific criteria, and selecting the pertinent features of the dataset are isolated. Consequently, redundant and irrelevant data features are eliminated, thus streamlining the data processing process.

As a result, this process can enhance learning accuracy, decrease learning duration, and streamline learning outcomes [10]. Some researchers in GDM prediction [11] choose to work closely with clinical experts and gynecologists to manually select features and reduce the number of variables, while others use automated techniques for feature selection. There are three main techniques for feature selection: filter methods, wrapper methods, and embedded methods [4]. Filter methods evaluate the relevance of features independently of the learning algorithm while wrapper methods evaluate the performance of features by directly using them with a specific machine learning algorithm [4]. Embedded methods incorporate feature selection as part of the model training process itself [4].

XAI features selection methods

In addition, new methods of features selection based on eXplainable Artificial Intelligence (XAI) are proposed such as Layer-wise Relevance Propagation (LRP), Local Interpretable Model-agnostic Explanation (LIME), and Shapley Additive exPlanations (SHAP) [4]. In fact, numerous studies have explored the application of Explainable Artificial Intelligence (XAI) as an innovative method for feature selection. Feature selection approach using XAI can help to improve the accuracy and interpretability of the prediction models. Additionally, this approach offers transparency during the process of selection of important features [12].

Especially, in our context, implementing XAI-based feature selection for GDM prediction has the potential to enhance both the accuracy and interpretability of machine learning models [17].

2.2 Automated Machine Learning (AutoML)

In the realm of artificial intelligence, Machine Learning and Deep Learning models have gained significant importance and applicability across various industries, effectively tackling complex AI challenges [5].

However, developing these sophisticated models typically requires a labor-intensive, trial-and-error approach that is manually executed by domain experts, necessitating considerable resources and time. To address these issues, the AutoML paradigm has emerged as a promising solution aimed at streamlining and enhancing the machine learning process [5]. Nonetheless, the interpretation of AutoML varies among different sectors within the scientific community. For instance, The researchers in [6] suggest that AutoML is primarily intended to reduce the reliance on data scientists, empowering domain experts to create machine learning applications without needing extensive ML expertise.

The authors in [7] consider AutoML as a seamless integration of automation and machine learning. This perspective highlights the automated construction of a machine learning pipeline, which operates within a defined computational budget. Thus, a robust AutoML system showcases the dynamic combination of various techniques, resulting in an intuitive, end-to-end machine learning pipeline. Numerous AI-focused companies, such as Google, Microsoft Azure, Amazon, H2O.ai, and RapidMiner, have created and made available systems like Cloud AutoML [7].

Considering AutoML and its role in Healthcare, AutoML is designed to streamline the selection of optimal machine learning algorithms, hyperparameter tuning, data pre-processing management and feature selection. This automation significantly decreases the time and expertise required to create effective models, making it particularly appealing in healthcare, where swift and precise decision-making is vital. As a result, AutoML has become a crucial resource in the medical sector for identifying risk factors, forecasting disease progression, and informing treatment plans. Several studies have demonstrated remarkable success in various medical applications, including diabetes diagnosis and prediction [5]. In the realm of diabetes prediction, AI-driven predictive models have played a key role in identifying early indicators of the disease and aiding clinicians in making informed, data-driven choices. These models utilize extensive data from diverse sources, including electronic health records, genomics, and wearable technology, to deliver accurate predictions regarding diabetes risk and onset [5] but these works present some limits such as Data Quality and Quantity since AutoML systems require high-quality, well-structured data and also AutoML processes can be computationally intensive, requiring significant processing power and time, especially for large datasets or complex models.

This is in the context of GDM prediction, we propose in this study an AutoML feature selection approach based on XAI which involves generating GDM predictive models that provide the best accuracy and interpretability in their decision-making process and we are considered among the pioneers who have addressed this topic. In our context, the proposed approach aims to generate ML models based on the most relevant features or variables that contribute to the prediction of GDM and also provides insights into how these features influence the model's predictions.

3 feature selection for GDM prediction using XAI based AutoML approach

This section describes the proposed feature selection for GDM prediction using XAI based AutoML approach in this research work by describing some experiments and their results in order to improve the performance of the predictive models generated by the autoML platform used in our study in the context of prediction of GDM.

3.1 XAI with SHAP

XAI is a rapidly growing field that focuses on developing AI systems that can provide explanations for their decisions and actions. In recent years, there has been a significant increase in research and literature on XAI, as the need for transparency and accountability in AI systems becomes more apparent.
In fact, researchers have proposed various methods for making AI systems more explainable, including model-agnostic techniques, post-hoc explanations, and interpretable models [13].
These approaches based on XAI use techniques such as SHAP (SHapley Additive exPlanations) values [4] or LIME (Local Interpretable Model-agnostic Explanations) [4] to explain the importance of each feature in the model's predictions. These techniques provide a clear understanding of how each feature impacts the model's output, allowing for the identification of key predictors of diabetes. SHAP is a powerful XAI technique used in this research work to interpret the used predictive machine learning model. The SHAP interprets machine learning models based on optimal Shapley values which are calculated through coalitional game theory. In this theory, two or more players or factors are involved in planning a strategy to achieve a desired outcome or payoff. The same concept is used to know how to distribute the "payout" among the features fairly. The SHAP explanation works on both global and local levels. On the local level, it shows each feature's contribution to predicting output on a particular

data point, however the global-level explanation shows the importance of each feature to the model [14]. SHAP is used in various domains, including healthcare, finance, and natural language processing, where understanding model decisions is crucial for trust, accountability, and compliance with regulations [15].

The AutoML platform used in this study which is RapidMiner, is based on the use of SHAP to determinate the scores of importance of features in the dataset used.

3.2 Proposed feature selection for GDM prediction using XAI based AutoML approach

The feature selection for GDM prediction using XAI based AutoML approach proposed in this research work aims to identify the most accurate GDM predictive model. This process is based on generation and evaluation of several AutoML models. The remove of irrelevant features from the dataset through some iterations can improve the model's performance and its hyperparameters and reduce also overfitting. By selecting only the most important features and regenering the autoML models, these models become more accurate and interpretable.

The process, the experiments and the results of our proposed approach are exposed in this subsection.

Proposed process of feature selection

The process of our proposed feature selection for GDM prediction using XAI based AutoML approach is presented in Figure. 1:

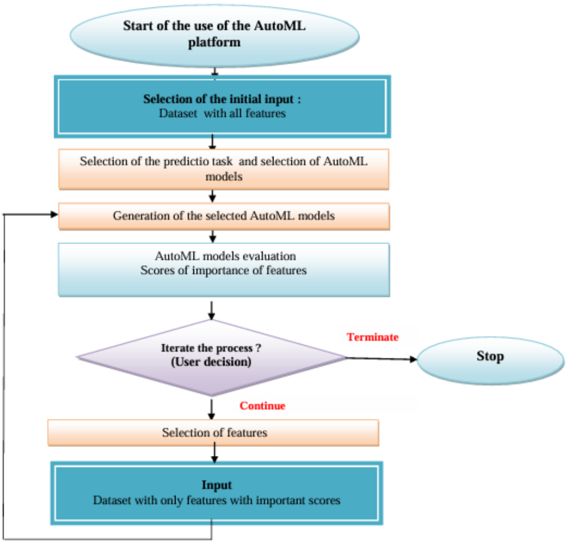

Fig. 1. Flowchart of the process of the proposed feature selection for GDM prediction using XAI based on autoML approach

As illustrated in Figure 1, regarding the process of the proposed feature selection approach for GDM prediction, Initially**,** this process started by launching the AutoML platform and the first step after that consists on selecting the dataset with all features.

In second time, a new project of AutoML was created, the task of prediction and the features of the dataset to use were selected. Then, the models to generate and some others hyperparameters such as max runtimes and the number of models were specified. After that, the specified AutoML models were generated with all their metrics of performance and the scores of importance for the used features based on SHAP value. In the next iterations, we removed the features having null contributions or scores of importance with absolute values less than the fixed threshold (0.20) in order to readjust another time the AutoML models and improve their performance. Finally, this process can be stopped when all features present significant scores of importance above the fixed threshold. This process is guided by the user decision.

Experimental setup and results

This section deals with the description of the used AutoML platform, the exploited dataset and the presentation of adopted AutoML models in our approach. Additionally, the experimental results will be exposed and discussed.

Dataset description

The experimental results are based on a diabetes prediction dataset for GDM sourced from Kaggle[1]. This dataset comprises 15 clinical features from 3,525 female patients, including attributes such as age, number of pregnancies, HDL (High-density lipoprotein) levels, family history, and BMI (Body Mass Index), among others. The output is represented as '1' for a positive diagnosis (diabetes detected) and '0' for a negative diagnosis (no diabetes detected).

AutoML platform description

RapidMiner is a comprehensive data science platform that provides tools for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is designed to facilitate the entire data science workflow, making it accessible to both data scientists and business [5]. It's a powerful tool which offers a visual interface that allows users to build and deploy machine learning models without extensive programming knowledge. Users can drag and drop various operators to create data workflows. In addition, RapidMiner includes AutoML capabilities that automate the process of features selection, models selection, hyperparameter tuning, and evaluation.

1 [1]https://www.kaggle.com/datasets/sumathisanthosh/gestational-diabetes-mellitus-gdm-dataset

This AutoML platform helps users quickly identify the best models for their data without deep expertise in machine learning. Overall, RapidMiner supports XAI which is increasingly important in the context of AutoML platforms. These techniques help users understand, trust, and effectively utilize machine learning models. The most key XAI techniques that can be integrated with RapidMiner is SHAP.

This technique is used to assess the importance of different features in a model. This helps users understand which variables are driving predictions. By incorporating this XAI technique, RapidMiner enhances the usability and trustworthiness of its AutoML capabilities, enabling users to make informed decisions based on their machine learning models [5].

Adopted ML Models

The used AutoML system in this study proposes six machine learning models deemed most promising and pperformant, including the following:

1. Generalized Linear Model (GLM): a flexible extension of standard linear regression that accommodates response variables with error distribution models beyond the normal distribution [5].
2. Decision Tree: an intuitive model that forecasts results by dividing the dataset into smaller groups according to feature values, similar to a tree-like structure [5].
3. Gradient Boosted Trees: An ensemble learning technique that builds a strong model by sequentially refining a collection of weak learners through the application of the gradient descent algorithm [5].
4. Support Vector Machine (SVM): A boundary-based model that identifies the optimal hyperplane for clearly distinguishing data points in high-dimensional space. Support Vector Machines (SVMs) determine the hyperplane that maximizes the margin between different classes [5].
5. Deep Learning: a neural network model modeled like the human brain, which is highly effective at identifying complex patterns within extensive datasets [5].
6. Random Forest: An ensemble learning technique that trains multiple Decision Tree classifiers on different subsets of the dataset and employs averaging to enhance predictive accuracy while mitigating overfitting [5].

Experimental results and discussion

After the selection of the kaggle GDM dataset with 15 features initially and the specification of the predictive models to use, we proceed to analyse the results for GDM prediction in depth through iterations of the process of generation of the autoML models.

The metrics of performance proposed by the AutoML platform RapidMiner and used in this study are as follow:

- Relative error lenient or accuracy : evaluates the accuracy of predictions by taking into account an acceptable margin of error [16].
- Runtimes : is a measure used to represent the execution time for each specified model [16].

At the first iteration, the AutoML project is initiated with all the features and the AutoML models are selected for a task of GDM prediction as illustrated in Figure 2.

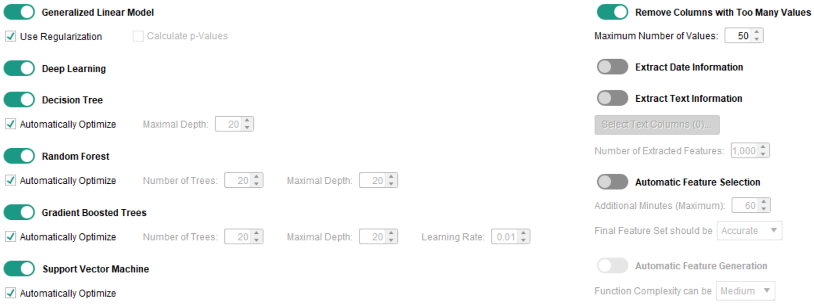

Fig. 2. AutoML selected models

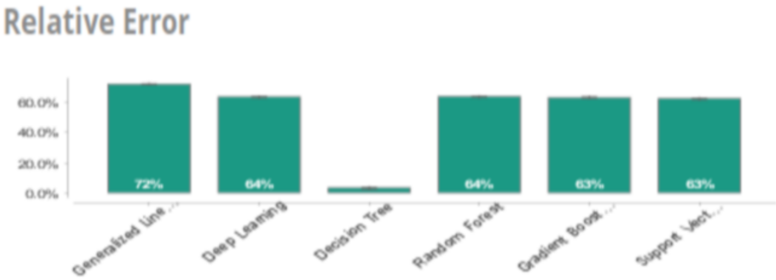

Fig. 3.. Comparison of the generated AutoML models based on Accuracy

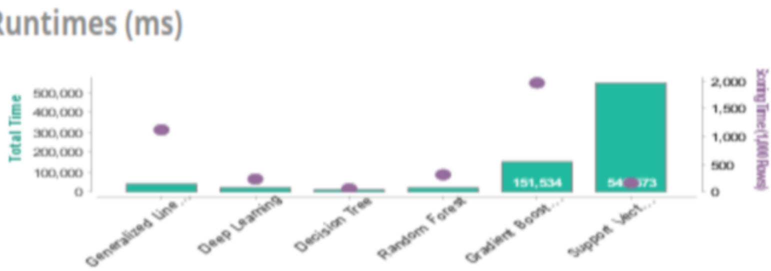

Fig. 4. Comparison of the generated AutoML models based on Runtimes

After the launch of the process of generation of the selected AutoML models, the comparison between these models based on runtimes and Accuracy (relative error) are displayed. These results are illustrated in Figures 3 and 4. The optimal generated AutoML model is the Generalized Linear Model which presents the metrics presented in Table 1 (Column first iteration).

For this first iteration of the process of the proposed approach, the output includes the features and their scores of importance based on SHAP value as illustrated in Figure 5 which demonstrates that the features named HDL (High Density Lipoprotein), Number of pregnancy, Large child or Birth Default and Sys BP have insignificant contribution since their scores of importance based on SHAP value are null. Also, the features Family History, sedentary lifestyle, Age, unexplained parental loss and Gestation in previous Pregnancy have absolute values of their weights less than the fixed threshold (0.2). Thus, all these features must be removed in the Next iteration.

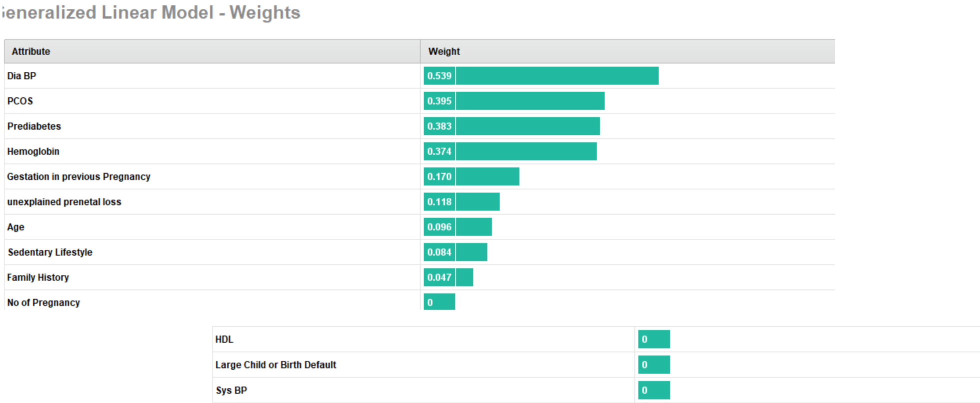

Fig. 5. Features weights of Generalized Linear Model (First iteration)

At the second iteration, we proceed to remove these features and the AutoML project is initiated with the features selected after the first iteration having significant scores of importance for the prediction and we select another time the same AutoML models for a task of GDM prediction as shown in Figure 6. The comparison between the generated predictive models for this second iteration based on runtimes and accuracy (relative error) are illustrated in Figures 7 and 8.

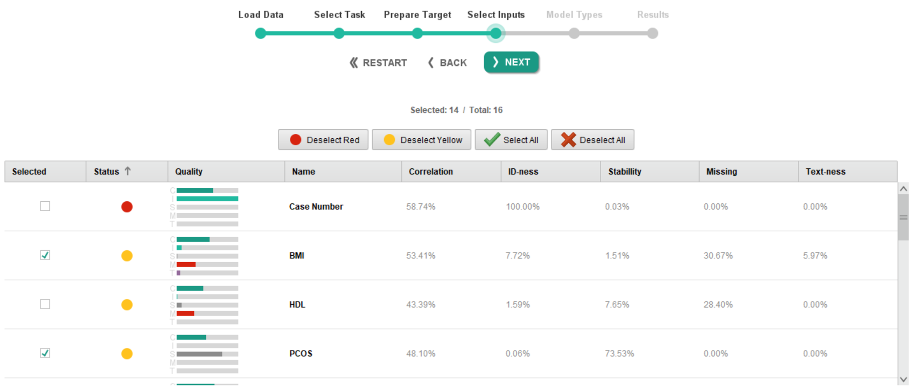

Fig. 6. Interface of features selection

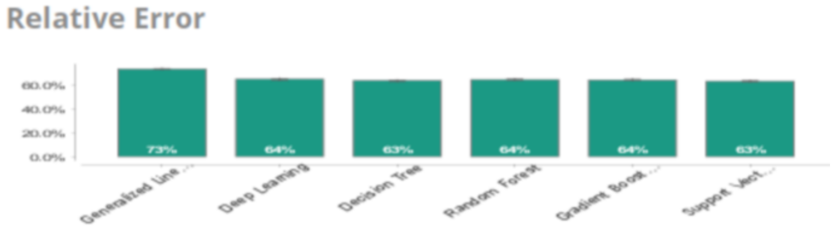

Fig. 7. Comparison of AutoML models based on Accuracy

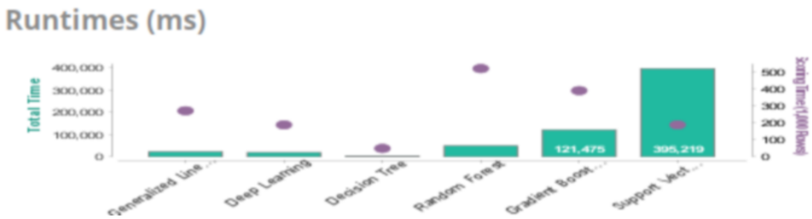

Fig. 8. Comparison of AutoML models based on Runtimes

Table 1 (Column Second iteration) presents the performance of the optimal model which is another time the Gnerelized Linear Model.The results through these two iterations are illustrated in what follows (Table I).

Table 1. Generalized Linear Model performance (First and Second iteration)

Metrics	First iteration (Without features selection)	Second iteration (With features selection)
Relative_error_lenient or accuracy	72.43% +/- 0.25%	72.92% +/- 0.34%
Runtimes	6 s	4 s

We notice the optimization of the performance of this model in terms of accuracy and runtimes after the selection of important features and the elimination of insignificant features in the second iteration relatively to our approach.

In addition, for the second iteration of our proposed approach, we have also as output the weights of the selected features (Figure 9) having all absolute values of weights above the fixed threshold (0.2). The process is stopped after this second iteration.

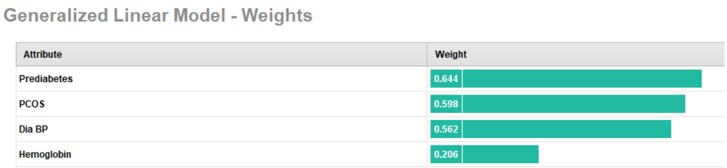

Fig. 9. Features weights of Generalized Linear Model (Second iteration)

4 Conclusion

This paper describes our study dealing with the experimentation of a feature selection for GDM prediction using XAI based on AutoML approach. The principal contribution in this research work is to demonstrate that the proposed feature selection approach leads to optimize the performance of the generated AutoML models. The explanation technique SHAP is utilized for interpretation and the most important features are determinated and used. The proposed approach was adopted and tested at the level of six AutoML models in the context of GDM prediction through an iterative process which resulted in the improvement of the performance of these models. The Generalized linear model achieves the highest prediction performance. The future scope is to study other approaches of feature selection based on AutoML models and XAI techniques and other AutoML systems and datasets will also be explored.

References

1. Kurt, B., Gürlek, B., Keskin, S. et al. Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. Med Biol Eng Comput 61, 1649–1660 (2023). https://doi.org/10.1007/s11517-023-02800-7

2. Cubillos, G., Monckeberg, M., Plaza, A. et al. Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. BMC Pregnancy Childbirth 23, 469 (2023). https://doi.org/10.1186/s12884-023-05766-4

3. El-Rashidy, N., ElSayed, N.E., El-Ghamry, A. et al. RETRACTED ARTICLE: Prediction of gestational diabetes based on explainable deep learning and fog computing. Soft Comput 26, 11435–11450 (2022). https://doi.org/10.1007/s00500-022-07420-1

4. R. Kumari and H. Garg, "Interpretation and Analysis of Machine Learning Models for Brain Stroke Prediction," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-6, doi: 10.1109/ISCON57294.2023.10112188

5. Zhuhadar, L.P.; Lytras, M.D. The Application of AutoML Techniques in Diabetes Diagnosis: Current Approaches, Performance, and Future Directions. Sustainability 2023, 15, 13484. https://doi.org/10.3390/su151813484

6. Popova Zhuhadar, L. AutoML Workflow. Available online:https://commons.wikimedia.org/wiki/File:AutoML_diagram.png (accessed on 30 March 2023)
   a. .

7. Zöller, M.-A.; Huber, M.F. Benchmark and survey of automated machine learning frameworks. J. Artif. Intell. Res. 2021, 70, 409–472. [CrossRef]

8. Yao, Q.; Wang, M.; Chen, Y.; Dai, W.; Li, Y.-F.; Tu, W.-W.; Yang, Q.; Yu, Y. Taking human out of learning applications: A survey on automated machine learning. arXiv 2018, arXiv:1810.13306.

9. H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer Science & Business Media, 2012.

10. Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, H. Liu, Advancing Feature Selection Research, ASU Feature Selection Repository (2010) 1–28

11. Du Y, Rafferty AR, McAuliffe FM, Wei L, Mooney C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Sci Rep. 2022 Jan 21;12(1):1170. doi: 10.1038/s41598-022-05112-2.

12. Ye Y, Xiong Y, Zhou Q, Wu J, Li X, Xiao X. Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study. J Diabetes Res. 2020 Jun 12;2020:4168340. doi: 10.1155/2020/4168340.

13. Yang, Z., Wang, Z., Huang, C., Yao, X. (2023). An Explainable Feature Selection Approach for Fair Machine Learning. In: Iliadis, L., Papaleonidas, A., Angelov, P., Jayne, C. (eds) Artificial Neural Networks and Machine Learning – ICANN 2023. ICANN 2023. Lecture Notes in Computer Science, vol 14261. Springer, Cham. https://doi.org/10.1007/978-3-031-44198-1_7

14. M. S. Islam, M. Minul Alam, A. Ahamed and S. I. Ali Meerza, "Prediction of Diabetes at Early Stage using Interpretable Machine Learning," SoutheastCon 2023, Orlando, FL, USA, 2023, pp. 261-265, doi: 10.1109/SoutheastCon51012.2023.10115152.

15. S. Shahane, S. T C, B. R, A. Kodipalli, T. Rao and R. B R, "Prediction of Diabetes using Machine Learning Algorithms and Correlation with Pregnancy," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-7, doi: 10.1109/CONIT59222.2023.10205558.

16. Bochao Sun, Wenjun Cui, Gaoyang Liu, Biao Zhou, Weijian Zhao,A hybrid strategy of AutoML and SHAP for automated and explainable concrete strength prediction, Case Studies in Construction Materials,Volume 19,2023

17. A. MAALOUL, M. JEMEL and N. BEN AZZOUNA, XAI based feature selection for gestational diabetes Mellitus prediction, CODIT Conference,from 1 to 4 Jully 2024 Malte

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.