A Predictive Model for Accurate Classification of Different Types of Diabetes

Kağan Canerik,Sıla Keskin

1. Abstract

Diabetes is a typical metabolic disorder that presents itself in many ways and has various underlying mechanisms. The aim of this project was to create a machine learning-based predictive model that would correctly classify different types of diabetes. In particular, one of the types to be classified is Steroid-Induced Diabetes (SID) that can be compared to more common types such as Type 1, Type 2, and Prediabetes. The dataset chosen for this study included a variety of medical, genetic, and lifestyle attributes. Unfortunately, the synthetic nature of the data caused some difficulties in the model's performance and generalizability. Besides, both substantial preprocessing and stringent feature selection techniques do not help the classification, which then becomes the major reason for the study's limitation of the dataset's quality and representativeness. Nevertheless, the findings suggest the importance of data authenticity and diversity which are crucial for the development of reliable predictive applications for clinical purposes.

2. Introduction

The global diabetes pandemic, as one of the multifaceted metabolic diseases, still poses a great diagnostic and management challenge. Steroid-Induced Diabetes (SID), one of its forms, represents a clinically important but relatively underexplored subtype. Successful differentiation of SID from well known types of diabetes, such as Type 1 or Type 2, is the most critical factor that allows patient care to proceed appropriately, paving the way for the choice of the best treatment. But the real-world data is one of the few challenges of the reliable predictive model's development in this sector.

Diabetes is, indeed, as a wickedly complex metabolic disease, still a major inference and management fiasco throughout the world. Of a multitude, Steroid-Induced Diabetes (SID) is one such clinically significant (yet least explored) subtype.

Through this study the significance of the correct authentic and diverse datasets is emphasized and shown in the clinical prediction tasks. Notwithstanding the fact that the findings did not reach the expected accuracy levels, the project permits relevant insights into the difficulties of working with synthetic data and paves a solid ground for future studies using more trustworthy datasets. These results can be seen as a part of the debate about the role of data quality in driving predictive analytics in precision medicine.

3. Related Works

Diabetes classification, especially peculiar rare disease types like Steroid-Induced Diabetes (SID) is currently the main focus of attention. There are a lot of studies which have treated the innovative approaches for the diagnosis of accuracy and the challenges of data quality and the generation of relatively better and more reliable data.

One remarkable research was to check the high level of corticosteroid-induced hyperglycemia among the patients of inflammatory bowel disease and it was carried out by the BMJ Open Gastroenterology in 2020. The research paper pointed out a problem with the use of machine learning in hospital settings, meaning that the staff had a hard time to predict and deal with hyperglycemia induced by prednisone (BMJ Open Gastroenterology, 2020)[1].

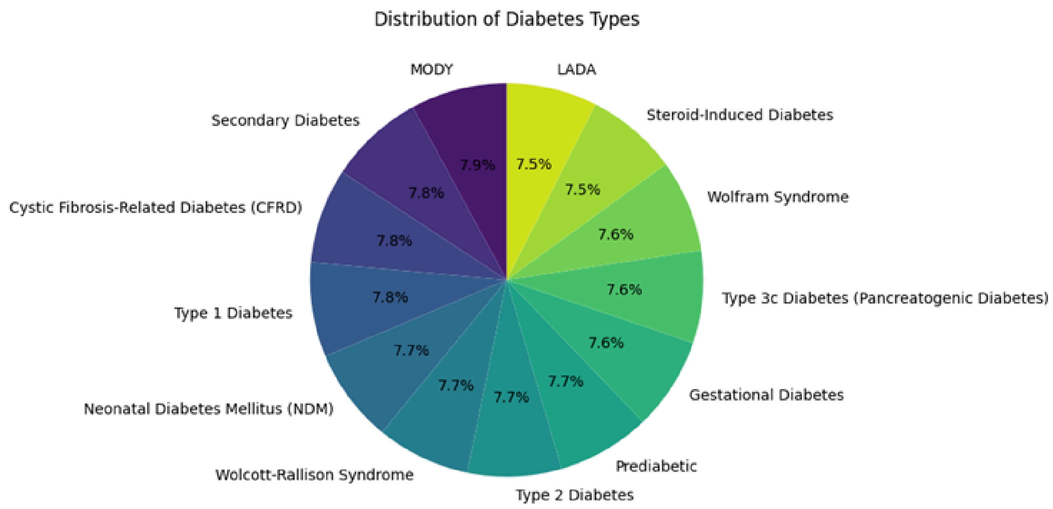
Somewhat do correspond to an advancement in the area by a machine learning framework incorporated with Generative Adversarial Networks (GANs) for synthetic data generation was undertaken to improve diabetes biological classification. However, the authors reported some limitations made in the study that led to the models being not general enough in BMC Bioinformatics 2023 (BMC Bioinformatics, 2023)[2].

The researchers also suggested that the hybrid model, Modified Particle Swarm Optimization, and Least Squares Support Vector Machines could be combined for diabetes classification, this approach can be used to select the best classification features of Type II diabetes patients (International Journal of Computer Trends and Technology, 2014)[3].

A new study not only exhausted, but also created a smart and effective approach to solving diabetes classification problems by dividing them into various categories as well as algorithms and data sets used. By means of a classification model that combines one or more of the methods, the research to the extent to which the quality of data plays the key role to ensure the correctness and reliability of the results can also be made (Advances in Intelligent Systems and Computing, 2020)[4].

4. Method

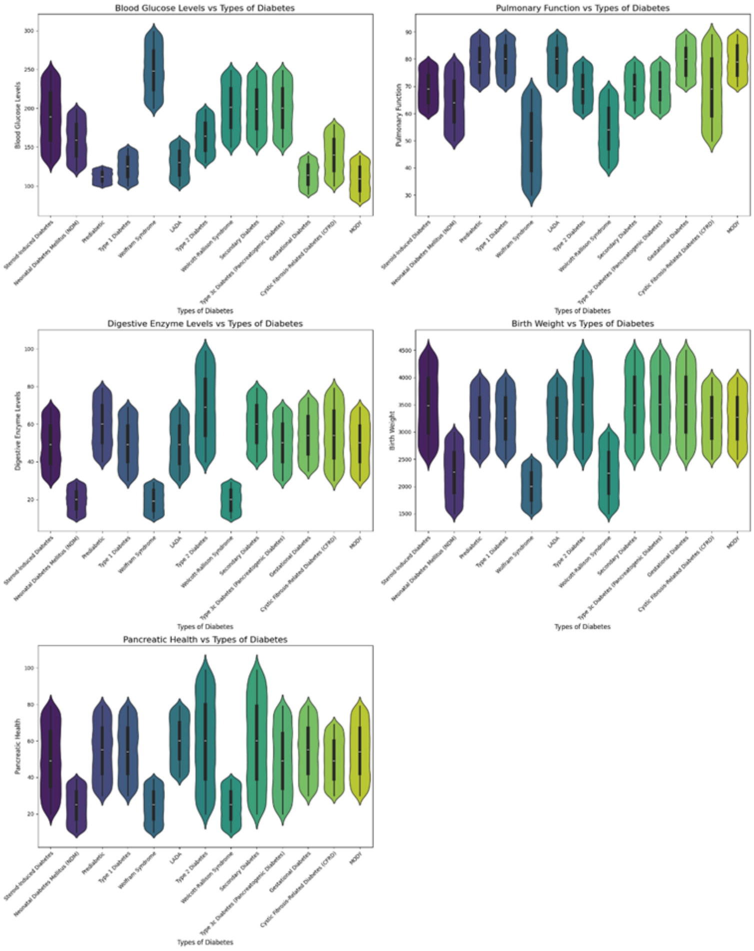
Data Analysis: The dataset was found to include 34 features, such as blood glucose levels, pancreatic health, pulmonary function, and birth weight. The data was visualized to understand its features.



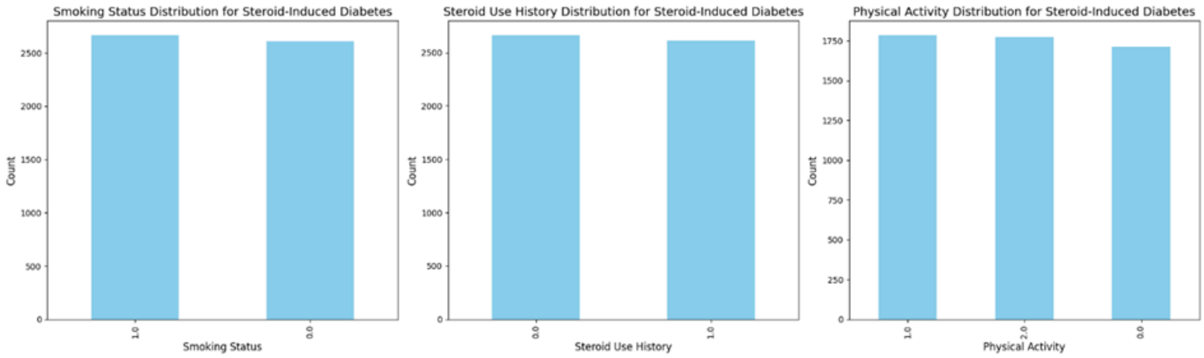
The pie chart of diabetes classes was used to understand the proportion of each diabetes type in the data. Classes were observed to be distributed equally. This led to the thought that the dataset might be synthetic. Data analysis was planned to continue after data preprocessing.

Data Preprocessing: For data preprocessing, categorical values were converted to numerical values so they could be used while observing relationships and applying machine learning algorithms. Then, normalization was applied to ensure no single feature disproportionately influenced the results. The data was standardized. This process prepared the dataset for further analysis and machine learning while preserving the original data. A subset of the original data (steroid_df), containing only Steroid-Induced Diabetes as the type of diabetes, was also taken.

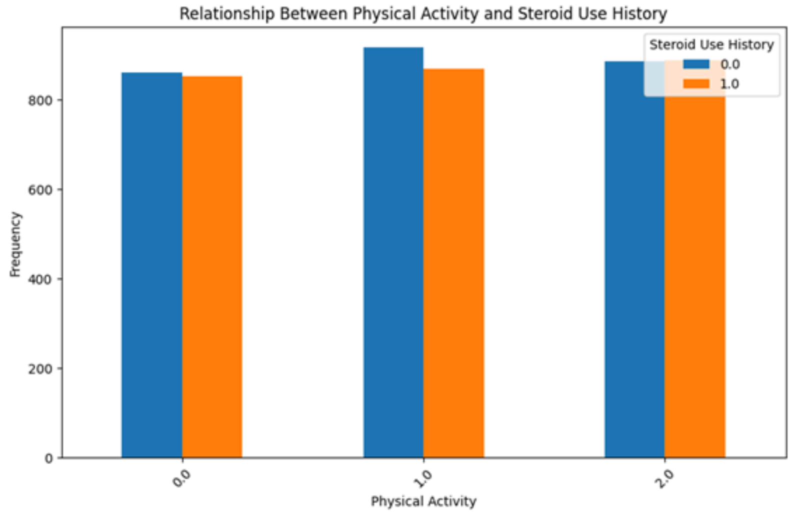
The correlation matrix for the original data frame was observed to show correlations, but for the steroid-induced subset, a super weak correlation between the features was found. The correlations of all features with the Types of Diabetes target variable were calculated. Then, the correlations were sorted in descending order based on their absolute values to identify the features most strongly related to the target variable. The top 5 features were selected, and a graph was plotted to understand the distributions of these features among all types of diabetes.



Blood Glucose Levels were found to vary for different types of diabetes. For Neonatal Diabetes Mellitus (NDM), Wolcott-Rallison Syndrome, and Wolfram Syndrome, birth weights were observed to be lower than others. Neonatal Diabetes Mellitus (NDM) and Wolcott-Rallison Syndrome are seen in babies, so this observation made sense. Wolfram Syndrome was noted as a genetic disease, which may affect birth weights. Pancreatic Health and Digestive Enzyme levels were also low for the patients of these three diabetes types.

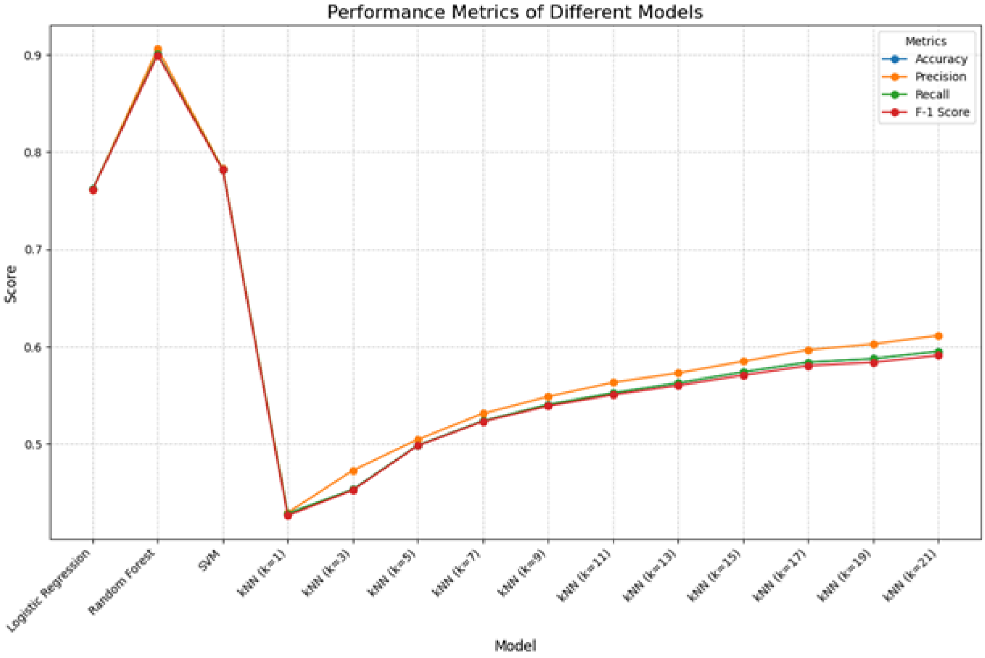


Features were observed to be distributed equally. Classes were also distributed equally, as seen in the pie chart in the EDA part. The data in the graphs was from steroid_df, which only contains Steroid-Induced Diabetes patients. The graphs were noted to be suspicious because steroid use history was equally distributed. Physical activity was also observed to be nearly equally distributed. A check was conducted on the relationship between steroid use history and physical activity.



It was observed that people who have 0 physical activity use steroids, and the distribution was found to be equally distributed again. This observation did not make sense and led to the thought that the dataset was synthetic.

Supervised Learning Algorithms: Logistic Regression, Random Forest, Support Vector Machine, and kNN were used as supervised learning algorithms.



Having studied all the possible ways of making a predictive model for the purposes, it was found that the best was Random Forest with 89.7% prediction accuracy, and it was considered the most reliable due to high precision, explanations, and F1 scores.

Logistic Regression and SVM were noted to follow in accuracy with 76.1% and 78.2%, respectively. They were observed to perform reasonably well in all metrics. For kNN, different k values were experimented with, starting with 1 and going up to 21. It was observed that as k increased, better accuracy was achieved, and the highest accuracy (59.5%) was received at k=21. kNN, however, had lower performance compared to other models and was found to not fit the dataset well. Overall, Random Forest was determined to be the most effective model. Logistic Regression and SVM were noted as models worth considering. The kNN model, despite showing some improvement with higher k values, was observed to remain less competitive.

Dimensionality Reduction with PCA: Principal Component Analysis (PCA) was applied to reduce the dataset's dimensions from 34 to 24 components, retaining approximately 90% of the variance.

Clustering Algorithms: K-Means Clustering, DBSCAN, and Hierarchical Clustering were used with different parameters. For some of them, the number of clusters was chosen as 13 because there were 13 classes in the dataset.

K-Means Silhouette Score 13 clusters: 0.04

K-Means Silhouette Score with PCA 13 clusters: 0.04

Clustering Accuracy:0.34

K-Means Silhouette Score 3 clusters: 0.12

DBSCAN Silhouette Score: -0.04

DBSCAN ARI Score: 0.00

Hierarchical Clustering Silhouette Score: 0.12

K-MEANS: The number of classes was selected as 13 because there were 13 types of diabetes in the dataset. A K-Means silhouette score of 0.04 was observed, indicating very poor clustering. The score being close to 0 suggested overlapping or poorly separated clusters. With a 0.12 silhouette score, 3 clusters were found to capture slightly better-defined groupings than 13 clusters. Accuracy was calculated from the clusters. A contingency table was printed to see if any relationship existed between the counts of clustering classes and the real classes. An algorithm was written to match the classes in the table, and an accuracy score was calculated from those values.

DBSCAN: A negative score was found, indicating that points were closer to points in other clusters than their own cluster. Clusters were observed to be noisy. A 0.00 ARI was calculated, meaning clustering was as bad as random clustering.

Hierarchical Clustering: With Hierarchical Clustering, the dataset was observed to have weakly defined groupings, but not enough to form well-separated clusters.

Silhouette scores were found to be too low across all clustering models. This observation showed that the dataset had problems forming clear clusters. High dimensionality was noted as a possible factor affecting this outcome. Additionally, the data was observed to be synthetic. Finding patterns on synthetic data was noted as a challenging task. Features were distributed almost equally everywhere. PCA with 24 components (selected for 90% variance) was found to fail to improve K-Means performance, indicating that reducing dimensionality alone did not resolve the clustering issues.

5. Results

The study indicates that machine learning holds a significant amount of potential for diabetes classification, specifically Steroid-Induced Diabetes (SID). Although models such as Random Forest and SVM demonstrated high accuracy in the identification of common diabetes types, their application for the determination of SID was low because of the synthetic nature of the dataset. This constraint affected the model's generality, leading to the decreased precision, recall, and F1-scores for SID. These research results confirm the critical need for high-quality, real-world data to develop predictive models that are both reliable and clinically meaningful.

6. Discussion

Machine learning applications for diabetes classification both contain the promise and the obstacles that have to be overcome. The challenges, which include the lack of patterns related to any cluster and the limitations of synthetic datasets, stand out as the most essential ones for data quality. Deficiency of the capability to classify diabetes types in a correct way caused by an ineffectively trained model indicates a need for data that is more specific and representative. In general, this research implies that the availability of diverse data and the quality of it are important because it would lead to the improvement of machine learning models in healthcare which in turn would become generalizable and reliable.

7. Future Work

For prospective ventures, success should take into consideration some crucial areas to fill in the gaps of the study. Primarily, the collection of real-life cases that are of the best type and retain a proportionate drive of the different diabetes subtypes, including SID, is the first step in order to accomplish model accuracy and generalizability. Firstly, enamored methods of modeling, such as ensemble learning and deep learning architectures, should be tried in the hopes of not only providing a highly accurate prediction about the disease but also try to convey multifaceted information oncology to clinicians. Besides, the inclusion of clinical factors such as drug history and therapy results in the prediction model can provide a better understanding and importance of the model to the end-users. Detecting potential biases that the synthetic or fake datasets can have is another significant aspect, which is done to ensure equality in the machine learning predictions across all patient groups. Finally, the design of comprehensible AI models constitutes an important step so that the doctors can gain direct knowledge about the predicting factors. It will then lead to the trust and usability of healthcare software. These research directions, tackled by the examination of the areas, would empower the machine learning (ML) application in the healthcare industry, therefore, implying yet more accurate, reliable, and meaningful solutions.

8. References

[1]:McDonnell, Martin, et al. "High incidence of glucocorticoid-induced hyperglycaemia in inflammatory bowel disease: metabolic and clinical predictors identified by machine learning." *BMJ Open Gastroenterology* 7.1 (2020): e000532.

[2]:Feng, Xin, Yihuai Cai, and Ruihao Xin. "Optimizing diabetes classification with a machine learning-based framework." *BMC bioinformatics* 24.1 (2023): 428.

[3]:Soliman, Omar S., and Eman AboElhamd. "Classification of diabetes mellitus using modified particle swarm optimization and least squares support vector machine." *arXiv preprint arXiv:1405.0549* (2014).

[4]:Agushaka, Jeffrey O., and Absalom E. Ezugwu. "Diabetes classification techniques: a brief state-of-the-art literature review." *Applied Informatics: Third International Conference, ICAI 2020, Ota, Nigeria, October 29–31, 2020, Proceedings 3*. Springer International Publishing, 2020.