Identification of Cardiovascular Disease Risk Factors Among Diabetes Patients using ontological Data Mining Techniques

Abdelaziz A. Abdelhamid[*1], Marwa M. Eid[2], Mostafa Abotaleb[3], S. K. Towfek[4]

[1] Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
[2] Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 11152, Egypt
[3] Department of System Programming, South Ural State University, 454080 Chelyabinsk, Russia
[4]Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA
Emails: abdelaziz@cis.asu.edu.eg; mmm@ieee.org; abotalebmostafa@bk.ru; sktowfek@jcsis.org

Abstract

Diabetes patients face a severe health cost from cardiovascular disease (CVD). Recognising the risk factors for CVD in this group of people is critical for developing effective preventative and management measures. In this study, we use an ontological data mining approach, LightGBM, to analyze a dataset of diabetes patients and investigate the risk variables that contribute to CVD. The association between diabetes and CVD is investigated, emphasising the increased risk that diabetes patients confront. We look into the demographics, health behaviors, and physiological indicators that influence the emergence of heart disease in this population. We use LightGBM to find complicated relationships and trends within the dataset, allowing us to identify critical risk variables. Our research contributes to the field by offering a thorough examination of the diabetes-CVD link and applying an advanced machine-learning technique for information extraction. The results have implications for specific interventions, risk evaluation models, and personalised therapy approaches aimed at reducing the effect of CVD in diabetics.

Keywords: Cardiovascular disease; Diabetes, Risk causes; Ontological data mining; Knowledge representation; Data-driven techniques; Semantic reasoning; Health data analysis.

1. Introduction

Diabetes mellitus is a chronic metabolic condition that is characterised by high blood glucose levels due to insulin insufficiency or insulin resistance. These elevated blood glucose levels can be caused by either insulin resistance or insulin insufficiency. It has been linked to some different consequences, one of the most serious and life-threatening of which is cardiovascular disease (also known as CVD), which affects millions of individuals all over the world [1]. The term CVD refers to a group of disorders that can have an adverse effect on the heart and blood vessels. These conditions include heart failure, stroke, peripheral artery disease, and coronary artery disease. People who have diabetes

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

have a significantly increased risk of acquiring CVD as compared to people who do not have diabetes [2-3].

There is a strong and nuanced correlation between diabetes and CVD. Atherosclerosis, characterised by the formation of fatty plaques within the artery walls, is exacerbated by the elevated blood glucose levels characteristic of diabetes [4]. The risk of cardiovascular events like heart attacks and strokes is raised when arteries become narrowed and hardened due to atherosclerosis. The risk of CVD is already elevated by diabetes, and it is commonly accompanied by other risk factors such as hypertension, dyslipidemia, and obesity. Insulin resistance, in which cells do not properly respond to insulin, is strongly linked to the development of type 2 diabetes. Diabetes and cardiovascular problems are two conditions in which insulin resistance is thought to have a major role [5].

For these reasons, it is critical to screen diabetic patients for CVD risk factors. To begin, healthcare providers are better able to stratify patients based on risk when they have a firm grasp on the risk variables linked with CVD in diabetic patients. Patients at high risk of developing CVD and its complications can benefit from this risk classification by receiving targeted interventions and therapies [6]. Second, being able to take preventative actions and make changes to one's way of life when CVD risk factors are first recognised is a powerful tool. Diabetes patients can greatly reduce their risk of cardiovascular events by making adjustments to their lifestyles, such as eating healthier, being more physically active, and giving up smoking. The most effective use of healthcare resources is achieved when treatment plans are individualised for each patient based on their unique risk profiles. It paves the way for individualised care plans to meet the specific requirements of each patient, which may improve health and well-being [7-8].

Ontological data mining is an innovative technique that integrates the principles of data mining and ontology to extract knowledge and discover hidden relationships from complex datasets. In the context of our study, an ontology represents a structured and formalized representation of relevant concepts and their relationships in the domain of diabetes and CVD [9-10]. By leveraging ontological data mining, researchers can process vast amounts of heterogeneous data, such as electronic health records, clinical databases, and biomedical literature, to uncover meaningful patterns and associations.

This work provides several substantial advances in the field of diabetes-related CVD research. For starters, it presents a thorough examination of the association between diabetes and CVD, emphasising the elevated risk that diabetes patients confront. Our study intends to help the development of tailored preventive measures and interventions by identifying and understanding the risk factors linked with CVD in this specific population. Furthermore, to analyse and extract knowledge from complicated datasets, this study applies an ontological data mining technique, especially LightGBM. We can use LightGBM to find hidden patterns, correlations, and prediction features that would not be seen using typical analytical methods.

2. Related Work

Several studies have been conducted to explore the intricate relationship between diabetes and CVD and to identify the risk factors contributing to this deleterious association. These investigations have illuminated critical insights into the pathophysiological mechanisms underlying the heightened CVD risk in individuals with diabetes. Moreover, previous research has emphasized the significance of identifying and understanding these risk factors to implement targeted interventions, enhance patient care, and alleviate the burden of cardiovascular complications in diabetic populations. Qrenawi and Al Sarraj [1] conducted a study using ontological data mining techniques to identify CVD risk factors among diabetes patients. Their work laid the foundation for exploring the application of ontological data mining in this specific domain. More, Kuo et al. [2] presented a medical case study where domain ontology was used to drive data mining. This approach demonstrated the potential of leveraging domain knowledge through ontologies to enhance data mining outcomes. In addition, Abbas et al [3] focused on data mining and ontology-based techniques in healthcare management. Although not specifically targeting CVD risk causes among diabetes patients, their research highlighted the potential of data mining and ontologies in healthcare settings. Kavakiotis et al. [4] conducted a comprehensive review of machine learning and data mining methods in diabetes research. While not explicitly utilizing ontological data mining, their work provided insights into various machine-learning approaches and their applications in diabetes-related studies. Mehmood et al. [5] explored the use of deep convolutional neural networks for predicting heart disease. Although not directly related to ontological data mining, this study demonstrates the broader range of machine learning techniques employed in CVD research.

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

Kumar et al. [6] discussed big medical data mining and processing in e-healthcare, providing a broader perspective on data mining techniques in healthcare domains. This work emphasized the importance of efficient data processing and analysis in healthcare applications. Tiwari and Singh [7] proposed a significant attribute selection and classification approach for predicting diabetes disease. Although not focused on CVD risk causes, their work presented insights into feature selection and classification techniques that can be relevant in identifying risk factors. Further, Karystianis et al. [8] investigated the use of local lexicalized rules to identify heart disease risk factors in clinical notes. Their study showcased a different approach, highlighting the importance of textual analysis and rule-based methods for risk factor detection.

While the preceding research has provided significant insights into the relationship between heart disease and diabetes, it is clear from the discussion of these connected studies that ontological data mining approaches should be used to detect risk causes among diabetes patients. This motivates the current study and highlights the promise of ontological data mining for overcoming the shortcomings of prior research into the root causes of cardiovascular illness in people with diabetes. Ontological data mining provides a distinct advantage in the full analysis of complicated interactions and the discovery of hidden patterns within heterogeneous healthcare data by merging domain-specific knowledge represented through ontologies with advanced data mining methods. To better understand risk variables and their interplay in the context of CVD and diabetes, this method allows for the incorporation of structured, unstructured, and cognitively supplemented data sources.

3. Methodology

Ontological data mining techniques force the supremacy of machine learning algorithms to abstract knowledge and patterns from composite datasets, allowing us to uncover valuable insights and relationships between variables. One such powerful algorithm used in this study is LightGBM (Light Gradient Boosting Machine). LightGBM is a gradient-boosting framework that has obtained acceptance due to its efficiency, scalability, and capability to handle large-scale datasets with high-dimensional features. It is specifically well-suited for ontological data mining tasks, including the identification of CVD risk factors among diabetes patients [11-14].

LightGBM employs gradient boosting, a technique that iteratively builds an ensemble of weak models to minimize a specific loss function. Let's assume we have a training dataset consisting of N samples and K features. Each sample is represented by a feature vector X and a corresponding target variable y. LightGBM utilizes an objective function that defines the loss to be minimized during training.

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \quad (1)$$

The objective function takes the current model's predictions, $\hat{y}_i^{(t)}$, and the true labels, y_i, as inputs and calculates a measure of the prediction error. LightGBM uses decision trees as the weak learners within the boosting framework. A decision tree is a hierarchical structure of nodes and branches, where each internal node represents a feature, and each leaf node represents a prediction. All nodes of the decision tree are traversed through the accumulation of n samples.

$$Obj^{(t)} \cong \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_{(t)}) \quad (2)$$

LightGBM employs a unique approach called "leaf-wise" tree growth, which differs from the traditional depth-wise growth strategy. In leaf-wise growth, instead of growing the tree level by level, LightGBM selects the leaf with the maximum gradient to expand, resulting in a more balanced and efficient tree structure [15].

LightGBM uses the gradient-based optimization technique to calculate the gradients and Hessians, H_j, of the loss function with respect to the predicted values.

$$\frac{\partial}{\partial w_j}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] = \sum_{i\in I_j} g_i + \left(\sum_{i\in I_j} h_i + \lambda\right) w_j \quad (3)$$

where,

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

$$w_j = -\frac{G_j}{H_j+\lambda}, \quad (4)$$

and

$$G_j = \sum_{i\in I_j} g_i\,, H_j = \sum_{i\in I_j} h_i \quad (5)$$

The model is then able to update its parameters in the direction that best reduces the loss function with the help of these gradients. Its efficient implementation makes it appropriate for ontological data mining tasks since it can handle large-scale datasets with high-dimensional features.

4. Results and Analysis

In our experiment, we evaluated the performance of our proposed model on a publicly available dataset that focuses on exploring the risk factors associated with CVD in adults. The dataset aims to investigate how various demographic factors, health behaviors, and biological markers contribute to the development of heart disease. The data is composed of many variables, which are summarized as follows.

- Age: The age of the participant is recorded as an integer value.
- Gender: The gender of the participant, categorized as either male or female.
- Height: The height of the participant, measured in centimeters (integer value).
- Weight: The weight of the participant, measured in kilograms (integer value).
- Ap_hi: Systolic blood pressure reading taken from the patient, recorded as an integer.
- Ap_lo: Diastolic blood pressure reading taken from the patient, recorded as an integer.
- Cholesterol: The total cholesterol level of the participant, measured in mg/dL on a scale of 0 to 5+ units (integer value). Each unit represents an increase or decrease of 20 mg/dL.
- Gluc: The glucose level of the participant, measured in mmol/L on a scale of 0 to 16+ units (integer value). Each unit represents an increase or decrease of 1 mmol/L.
- Smoke: Indicates whether the participant is a smoker or not (binary value: 0 = No, 1 = Yes).
- Alco: Indicates whether the participant consumes alcohol or not (binary value: 0 = No, 1 = Yes).
- Active: Indicates whether the participant is physically active or not (binary value: 0 = No, 1 = Yes).
- Cardio: Indicates whether the participant suffers from CVD s (binary value: 0 = No, 1 = Yes).

To evaluate the performance of our model, we utilized this dataset, leveraging the information provided by the different variables to predict the likelihood of developing CVD. By training our model on this data and employing appropriate evaluation metrics, we aimed to assess the effectiveness of our approach in identifying and understanding the risk factors associated with CVD in the given population. In Table 1, we provide some descriptive analysis of the dataset.

Table 1: Descriptive statistics for the case study for CVD risk factors

	count	mean	std	min	25%	50%	75%	max
age	70000	19468.87	2467.252	10798	17664	19703	21327	23713
gender	70000	1.349571	0.476838	1	1	1	2	2
height	70000	164.3592	8.210126	55	159	165	170	250
weight	70000	74.20569	14.39576	10	65	72	82	200
ap_hi	70000	128.8173	154.0114	-150	120	120	140	16020
ap_lo	70000	96.63041	188.4725	-70	80	80	90	11000
cholesterol	70000	1.366871	0.68025	1	1	1	2	3
gluc	70000	1.226457	0.57227	1	1	1	1	3
smoke	70000	0.088129	0.283484	0	0	0	0	1
alco	70000	0.053771	0.225568	0	0	0	0	1
active	70000	0.803729	0.397179	0	1	1	1	1
cardio	70000	0.4997	0.500003	0	0	0	1	1

In Figure 1, we present a bar chart that visualizes the distribution of variables within the dataset related to the risk factors for cardiovascular heart disease. This visualization enables to obtain a comprehensive identification of the frequency or occurrence of different variables, enabling us to

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

identify patterns and trends in the data. Figure 1 illustrates the distribution of each variable as individual bars, with the x-axis representing the variable categories and the y-axis denoting the frequency or count of each category.

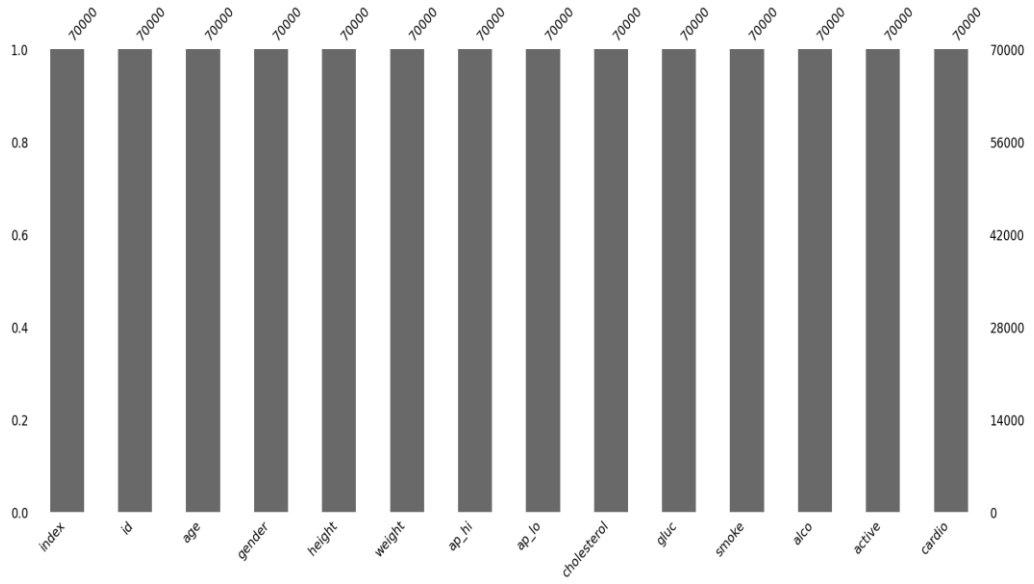

Figure1: Variable distributional analysis in our case study for CVD risk factors

Table 2 presents the Pearson correlation results, which showcase the strength and direction of the linear relationship between pairs of variables in the dataset. The Pearson correlation coefficient measures the degree of linear association between two variables, ranging from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship between the variables.

Table 2: The results of Pearson correlation analysis on the features of our case study.

	age	gende	heigh	weigh	ap_hi	ap_lo	chole	gluc	smok	alco	active	cardi
age	1	-0.02291	-0.08151	0.053561	0.020854	0.01762	0.154012	0.098388	-0.04765	-0.02976	-0.01	0.237985
gender	-0.02291	1	0.499033	0.155406	0.006005	0.015254	-0.03582	-0.02049	0.338135	0.170966	0.005866	0.008109
height	-0.08151	0.499033	1	0.290968	0.005488	0.00615	-0.05023	-0.0186	0.187989	0.094419	-0.00657	-0.01082
weight	0.053561	0.155406	0.290968	1	0.030702	0.04371	0.141768	0.106857	0.06778	0.067113	-0.01687	0.18166

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

ap_hi	0.020854	0.006005	0.005488	0.030702	1	0.016086	0.023778	0.011841	-0.00092	0.001408	-3.3E-05	0.054475
ap_lo	0.01762	0.015254	0.00615	0.04371	0.016086	1	0.024019	0.010806	0.005186	0.010601	0.00478	0.065719
cholesterol	0.154012	-0.03582	-0.05023	0.141768	0.023778	0.024019	1	0.451578	0.010354	0.03576	0.009911	0.221147
gluc	0.098388	-0.02049	-0.0186	0.106857	0.011841	0.010806	0.451578	1	-0.00476	0.011246	-0.00677	0.089307
smoke	-0.04765	0.338135	0.187989	0.06778	-0.00092	0.005186	0.010354	-0.00476	1	0.340094	0.025858	-0.01549
alco	-0.02976	0.170966	0.094419	0.067113	0.001408	0.010601	0.03576	0.011246	0.340094	1	0.025476	-0.00733
active	-0.01	0.005866	-0.00657	-0.01687	-3.3E-05	0.00478	0.009911	-0.00677	0.025858	0.025476	1	-0.03565
cardio	0.237985	0.008109	-0.01082	0.18166	0.054475	0.065719	0.221147	0.089307	-0.01549	-0.00733	-0.03565	1

In Figure 2, we present the distributional characteristics of each variable in the dataset using probability plots and density plots.

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

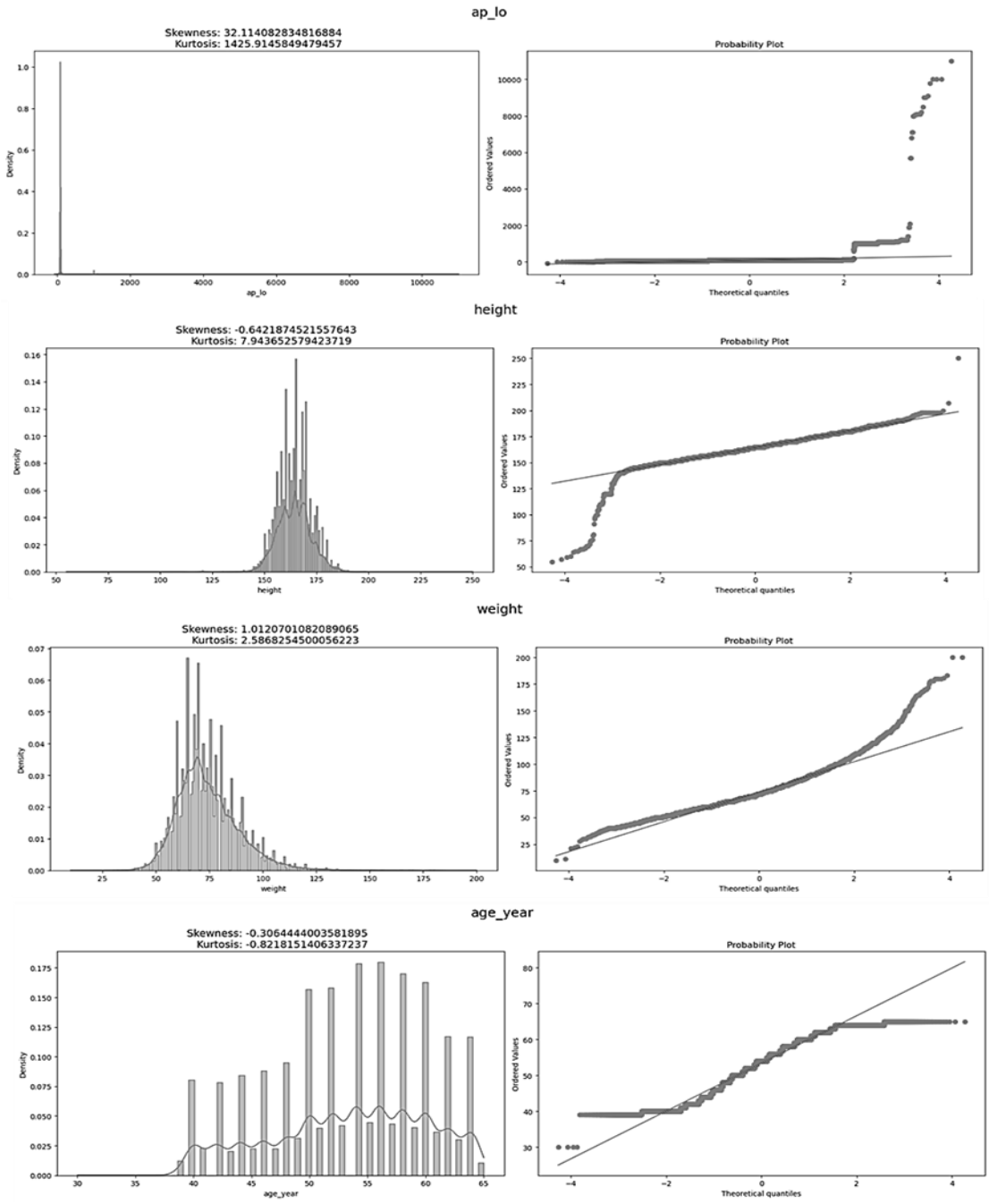

Figure 2: visualization of probability and density plot for each variable in our study

These visualizations provide a comprehensive understanding of the distributional properties of the variables, shedding light on their underlying statistical characteristics. The probability plots, also known as quantile-quantile (Q-Q) plots, are displayed for each variable. The density plots, on the other hand, depict the estimated probability density function of each variable. The x-axis represents the values of the variable, and the y-axis represents the corresponding probability density. The density plots provide insights into the shape, symmetry, and skewness of the variable's distribution. In addition, we present the results of our study by comparing the performance of our proposed model with other ML algorithms commonly used in the domain of cardiovascular disease risk prediction. To evaluate the models, we employ precision-recall curves, which provide insights into the trade-off between precision and recall for different classification thresholds. Figure 3 displays the precision-recall curves for each model, allowing us to assess their performance. A higher precision indicates a lower rate of false positives, while a higher recall suggests a lower rate of false negatives. As noted, we can determine which model strikes the best balance between these two metrics.

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

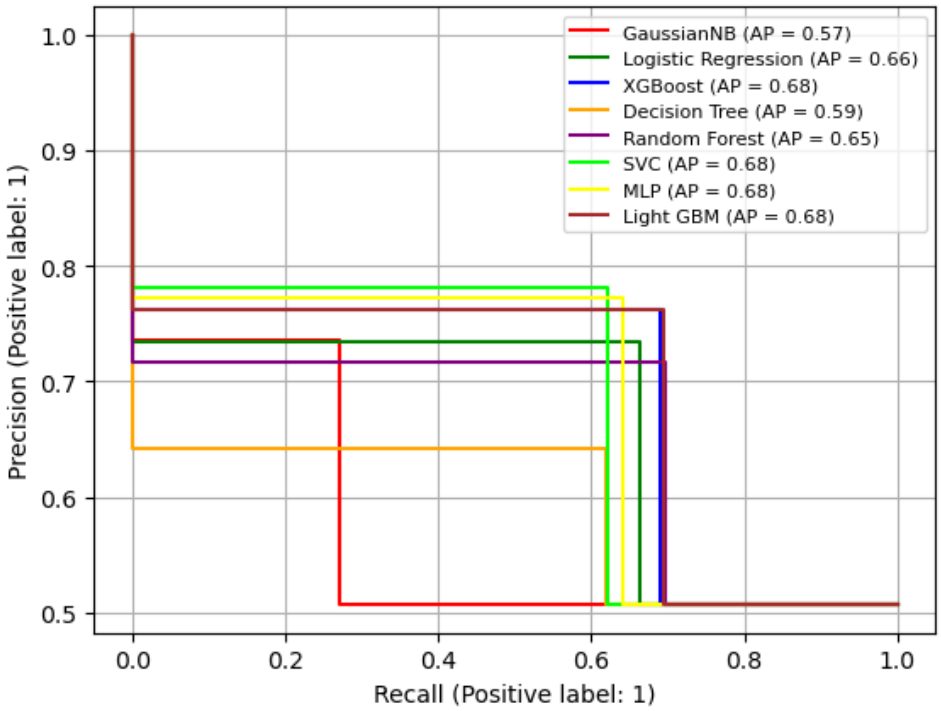

Figure 3: Visualization of Precision-recall curves for different classifiers.

6. Conclusion

This study aimed to identify cardiovascular disease (CVD) risk factors among diabetes patients using an ontological data mining method, specifically LightGBM. Through a inclusive analysis of the relationship between diabetes and CVD, we highlighted the heightened risk faced by individuals with diabetes. Our study contributes to the field by giving valuable understandings into the complicated interactions between risk factors and CVD in this specific population. By employing LightGBM, we were able to extract knowledge and uncover hidden patterns within the dataset. The utilization of this advanced ML algorithm enabled us to identify key risk factors associated with CVD among diabetes patients. The precision-recall curves demonstrated the superior performance of our proposed model, indicating its ability to accurately predict individuals at risk of developing CVD.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Qrenawi M I, Al Sarraj W, Identification of cardiovascular diseases risk factors among diabetes patients using ontological data mining techniques. In 2018 International Conference on Promising Electronic Technologies (ICPET), 129-134, 2018.
[2] M. Saber, Efficient phase recovery system, Indonesian Journal of Electrical Engineering and Computer Science (lJEECS), 5(1), 123-129, 2017.
[3] Mahmoud H, Abbas E, Fathy I, Data mining and ontology-based techniques in healthcare management. International Journal of Intelligent Engineering Informatics, 6(6), 509-526, 2018.
[4] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I, Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116, 2017.

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023

[5] Mehmood A, Iqbal M, Mehmood Z, Irtaza A, Nawaz M, Nazir T, Masood M, Prediction of heart disease using deep convolutional neural networks. Arabian Journal for Science and Engineering, 46(4), 3409-3422, 2021.
[6] Mohamed Saber, A novel design and Implementation of FBMC transceiver for low power applications, Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 8(1), 83-93, 2020.
[7] Tiwari P, Singh V, Diabetes disease prediction using significant attribute selection and classification approach. In Journal of Physics: Conference Series, 1714(1), p. 012013, 2021.
[8] Eid Marwa M, Fawaz Alassery, Abdelhameed Ibrahim, and Mohamed Saber, Metaheuristic optimization algorithm for signals classification of electroencephalography channels. Computers, Materials & Continua, 71(3), 4627-4641, 2022.
[9] Alharbi AH et al., Diagnosis of Monkeypox Disease Using Transfer Learning and Binary Advanced Dipper Throated Optimization Algorithm. Biomimetics, 8(3),313, 2023.
[10] Ali F, El-Sappagh S, Islam S R, Kwak D, Ali A, Imran M, Kwak, K S, A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63, 208-222, 2020.
[11] M. M. E. Bahy, S. A. Ward, M. Badawi and R. Morsi, "Particle-initiated negative corona in co-axial cylindrical configuration. Annual Report Conference on Electrical Insulation and Dielectric Phenomena, Montreal, QC, Canada, 343-348, 2012.
[12] E. M. Shaalan, S. M. Ghania and S. A. Ward, Analysis of electric field inside HV substations using charge simulation method in three dimensional. Annual Report Conference on Electrical Insulation and Dielectic Phenomena, West Lafayette, IN, USA,1-5, 2010.
[13] Mohamed A. Abouelatta, et al. , Measurement and assessment of corona current density for HVDC bundle conductors by FDM integrated with full multigrid technique. Electric Power Systems Research, 199, 2021.
[14] Amin Samy, Sayed A. Ward, Mahmud N Ali, Conventional Ratio and Artificial Intelligence (AI) Diagnostic methods for DGA in Electrical Transformers. International Electruical Engineering Journal, 6, 2096-2102, 2015.
[15] El-Kenawy, El-Sayed M., Marwa Eid, and Alshimaa H. Ismail, A New Model for Measuring Customer Utility Trust in Online Auctions. International Journal of Computer Applications, 975, 8887, 2020.

Doi: https://doi.org/10.54216/JAIM.040205
Received: November 12, 2022 Revised: May 01, 2023 Accepted: July 12, 2023