OPEN ACCESS

Received: 11/04/25
Accepted: 16/07/25
Online: 26/07/25

ORIGINAL ARTICLE

Untangling the complexity of diabetes risk: a Bayesian approach to learning causal structures

Ney Michel Lituma Villamar[1,a]

[1] Universidad de Guayaquil, Guayaquil, Ecuador.
[a] Master of Applied Artificial Intelligence.

Keywords:
diabetes mellitus; Bayesian networks; artificial intelligence; body mass index; hypertension; glycated hemoglobin A; algorithms; risk factors; prognosis; early diagnosis (source: MeSH-NLM).

ABSTRACT

Objective. To evaluate the performance and interpretability of Bayesian network classifiers for the early detection of diabetes. **Methods.** A model validation study of machine learning applied to healthcare was conducted, focusing on performance assessment and explainability of algorithms on a categorical and preprocessed dataset. Specifically, the following classifiers were trained and applied: Naive Bayes, Tree Augmented Naive–Chow-Liu (TAN–Chow-Liu), Tree Augmented Naive–Hill Climbing with Super Parents (TAN–HCSP), Fast Super-Parent Search with Joint Mutual Information (FSSJ), and the K-Dependence Bayesian Classifier (KDB). Models were tested on 100,000 preprocessed records (filtered by causal relevance and variable discretization) using *bnlearn* and *bnclassify*. Data were partitioned 75/25 (training/testing), and accuracy, sensitivity, specificity, and F1 score were estimated. In addition, the learned structures were analyzed against clinical evidence. **Results.** All models achieved accuracy ≥ 0.95 and F1 score > 0.94. FSSJ showed the best performance (accuracy 0.97; specificity 1.00), while Naive Bayes and KDB achieved comparable metrics with lower computational cost. The learned networks reproduced known associations among body mass index (BMI), hypertension, HbA1c, and glucose, and identified indirect chains (e.g., age influencing BMI, BMI influencing glucose, and glucose influencing diabetes), reinforcing their clinical plausibility. **Conclusions.** Bayesian networks provide transparent, high-quality predictions for diabetes risk. Basic architectures can perform on par with more complex variants when preprocessing is rigorous. The causal pathways highlight modifiable factors (overweight, elevated blood pressure) as priority targets for preventive interventions.

Desenredando la complejidad del riesgo de diabetes: un enfoque bayesiano para el aprendizaje de estructuras causales

Palabras clave:
diabetes mellitus; redes bayesianas; inteligencia artificial; índice de masa corporal; hipertensión; hemoglobina a glucosilada; algoritmos; factores de riesgo; pronóstico; diagnóstico precoz (fuente: DeCs-BIREME).

RESUMEN

Objetivo. Evaluar el rendimiento e interpretabilidad de clasificadores de redes bayesianas para la detección temprana de diabetes. **Métodos.** Se realizó un estudio de validación de modelos de aprendizaje automático (*machine learning*) aplicado al campo de la salud, enfocado en la evaluación de rendimiento y explicabilidad de algoritmos sobre un conjunto de datos categóricos y preprocesado. Específicamente, fueron entrenados y aplicados: Naive Bayes, Tree Augmented Naive-Chow-Liu (TAN–Chow-Liu), Tree Augmented Naive-Hill Climbing with Super Parents (TAN–HCSP), Fast Super-Parent Search with Joint Mutual Information (FSSJ) y K-Dependence Bayesian Classifier (KDB), sobre 100 000 registros preprocesados (filtrados por su relevancia causal y discretización de variables) utilizando *bnlearn* y *bnclassify*. La partición fue 75/25 (entrenamiento/prueba) y fueron estimadas exactitud, sensibilidad, especificidad y F1; además, fueron analizadas las estructuras aprendidas frente a la evidencia clínica. **Resultados.** Todos los modelos alcanzaron exactitud ≥ 0,95 y F1 > 0,94. El FSSJ mostró el mejor desempeño (exactitud 0,97; especificidad 1,00), mientras que Naive Bayes y KDB lograron métricas similares con menor costo computacional. Las redes aprendidas reprodujeron asociaciones conocidas entre el índice de masa corporal (IMC), hipertensión, HbA1c y glucosa, e identificaron cadenas indirectas (por ejemplo, la edad influye en el IMC; este, a su vez, influye en la glucosa y finalmente en la diabetes), reforzando su plausibilidad clínica. **Conclusiones.** Las redes bayesianas proporcionan predicciones transparentes y de alta calidad para el riesgo de diabetes. Las arquitecturas básicas pueden igualar a variantes más complejas cuando el preprocesamiento es riguroso. Las rutas causales resaltan factores modificables (sobrepeso, presión arterial elevada) como objetivos prioritarios para intervenciones preventivas.

Cite as: Lituma-Villamar NM. Untangling the complexity of diabetes risk: a Bayesian approach to learning causal structures. Rev Peru Cienc Salud. 2025;7(3):226-33. doi: https://doi.org/10.37711/rpcs.2025.7.3.12

Correspondence:
Ney Michel Lituma Villamar
mlituma@hotmail.com

INTRODUCTION

Early detection of diabetes is critical for reducing morbidity and mortality. Globally, more than 422 million people live with the disease, and 1.5 million deaths are attributed to it each year, with a disproportionate burden in low- and middle-income countries [1]. In 2021, the International Diabetes Federation (IDF) [2] estimated that 537 million adults (10.5 % of the global population) had diabetes, generating USD 966 billion in healthcare expenditures, with projections reaching 643 million by 2030 and 783 million by 2045, surpassing USD 1.054 trillion in costs.

Given these concerns, therapeutic personalization requires strong causal hypotheses, as data-driven causal discovery can support clinical decision-making [3]. Various machine learning algorithms exceed 85% accuracy in diabetes prediction [4] and detect subtle patterns in rare diseases [5]. In this context, Bayesian networks (BN) stand out for their ability to manage uncertainty and provide interpretability [6,8], along with their flexibility for categorical and continuous variables, incorporation of expert knowledge, and inference with missing data [9,11]. Previous evidence in diabetes has reported positive and negative predictive values of 69.6% and 79.9% in women [12], as well as dependencies between relevant features using Bayesian approaches [13]. A persistent challenge is class imbalance, which reduces classifier performance [14]. To mitigate this issue, techniques such as Synthetic Minority Oversampling Technique (SMOTE) [15], undersampling [16], hybrid strategies [17], weighted feature selection with Random Forest and XGBoost [18], and cost-sensitive learning with dimensionality reduction [19] have been applied. In BN, the combination of feature detection and multiple resampling has facilitated the identification of risk factors [20], and it is crucial to distinguish a generative BN from a BN classifier optimized for predictive accuracy [21]. Recent studies have expanded the framework: BN integration for type 2 diabetes (T2D) and coronary heart disease (CHD) [22], the use of Markov blankets in prediabetic populations [23], and the exploration of biomarker interactions through Tabu search and bootstrap techniques [24].

Despite methodological advances, several gaps remain: (a) head-to-head comparisons of multiple Bayesian classifiers under standardized preprocessing in large, categorical, imbalanced datasets are lacking; (b) it is not clearly delineated when basic architectures—such as Naive Bayes or the K-Dependence Bayesian Classifier (KDB)—match or outperform more complex variants, such as Tree-Augmented Naive Bayes (TAN) in its Chow–Liu, Hill Climbing with Statistical Perturbation (HCSP), or Forward Sequential Selection and Joining (FSSJ) forms, in terms of clinical utility; and (c) it is necessary to verify that the learned structures are clinically plausible and highlight modifiable factors for intervention. Addressing these gaps is relevant for decision-support systems and for prioritizing prevention and risk-management strategies.

The objective of this study was to compare the performance and interpretability of Bayesian classifiers—Naive Bayes, Tree Augmented Naive–Chow–Liu (TAN–Chow–Liu), Tree Augmented Naive–Hill Climbing with Super Parents (TAN–HCSP), Fast Super-Parent Search with Joint Mutual Information (FSSJ), and the K-Dependence Bayesian Classifier (KDB)—applied to a dataset of 100,000 diabetes predictors [25], following preprocessing (discretization and causal-relevance filtering) and imbalance-handling strategies. Accuracy, sensitivity, specificity, and F1-scores were evaluated, and the learned structures were contrasted with existing clinical literature.

METHODS

A machine learning model-validation study was conducted in the health domain, focused on evaluating the performance and explainability of algorithms applied to a categorical, preprocessed dataset. The Diabetes Health Indicators Dataset, a public dataset available on Kaggle with 100,000 records and 16 diabetes predictors [25], derived from the Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System (BRFSS) surveys, was used. The choice of this dataset is justified by its large sample size—over 250,000 original records—which provides the statistical power necessary for training and validating machine learning models. Its accessibility and structure make it a valuable resource for exploratory research on the feasibility of artificial intelligence (AI) algorithms in identifying risk factors in large populations, facilitating reproducibility and cross-study comparisons in the scientific community [26].

During preprocessing, no missing values were detected; the five race variables were merged into a single categorical feature, while age and BMI were discretized. Year and region were excluded due to low

Table 1. Performance metrics using different libraries

Metrics using the *bnclassify* library				
Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**F1-score**
Naive Bayes	0.965	0.975	0.987	0.981
TAN–CL	0.969	0.970	0.997	0.983
TAN–HCSP	0.965	0.974	0.987	0.981
FSSJ	0.970	0.969	1	0.984
KDB	0.965	0.974	0.987	0.981
Metrics using the *bnlearn* library				
Algorithm	**Accuracy**	**Sensitivity**	**Specificity**	**F1-score**
Naive Bayes	0.963	0.987	0.987	0.980
TAN–CL	0.780	0.945	0.806	0.870
TAN–HCSP	0.953	0.951	1	0.975
FSSJ	0.971	0.969	1	0.984
KDB	0.965	0.975	0.987	0.981

causal relevance. All fields were converted to factors in R, and the dataset was shuffled and split 75/25 (training/test), with reproducibility ensured. Naive Bayes, TAN–Chow–Liu, TAN–HCSP, FSSJ, and KDB were trained using *bnclassify* (Laplace smoothing) and subsequently refined with *bnlearn*. Conceptually, Naive Bayes assumed conditional independence; TAN incorporated a dependency tree; TAN–HCSP added hierarchical constraints; FSSJ performed sequential forward feature selection; and KDB combined k-parent structures with dynamic Bayesian reasoning for temporally drifting data. Accuracy, precision, recall, and F1-scores above 0.90 were obtained for all classifiers; TAN–Chow–Liu provided the clearest dependency map, and FSSJ exhibited the highest computational efficiency. Learned structures were inspected using gRain and visualized with Rgraphviz, confirming clinically plausible links between key predictors and diabetes.

RESULTS

Using the *bnlearn* package, three directed acyclic graphs (DAGs) were generated, making the learned Bayesian structures explicit (see Figures 1, 2, and 3). These visualizations shifted the focus from raw predictive accuracy toward the underlying probabilistic dependencies, allowing an assessment of how each algorithm encodes clinical knowledge (see Table 1).

In Figure 1, the star-shaped topology of Naive Bayes is shown: each predictor is modeled as a direct child of diabetes, with no lateral links among variables. An additional layer is displayed in which age channels risk toward hypertension, heart disease, and BMI, highlighting age as the pathway through which diabetes "explains" subsequent morbidity. Although

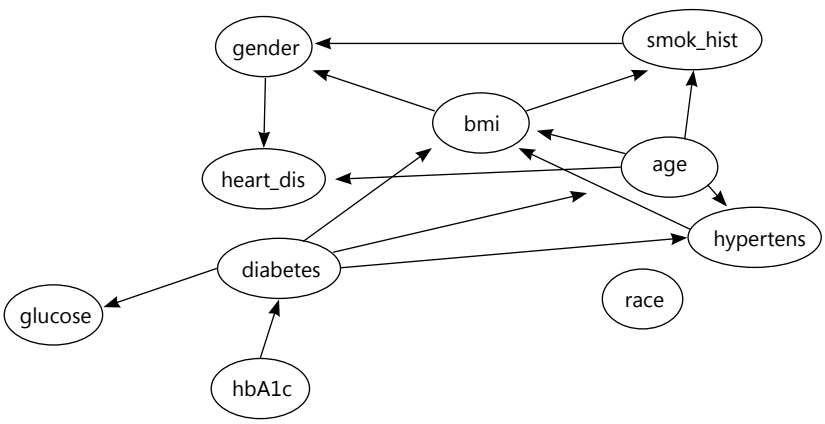

Figure 1. Bayesian model graph of Naive Bayes

 https://doi.org/10.37711/rpcs.2025.7.3.12

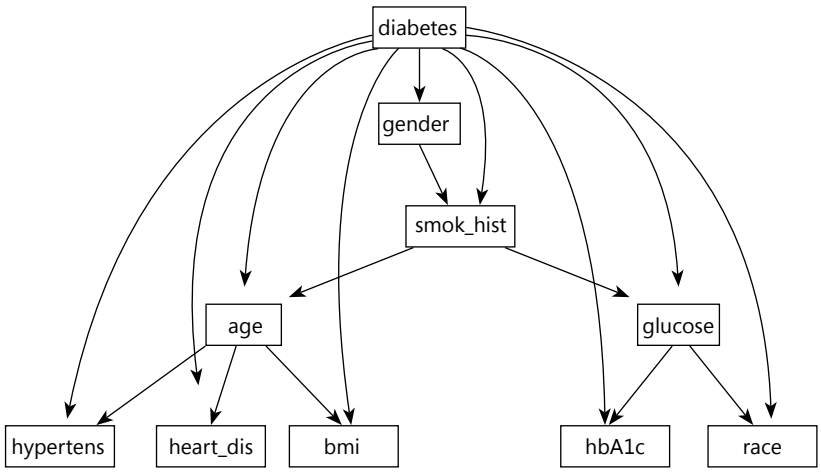

Figure 2. Bayesian model graph of the Chow–Liu structure

the independence assumption simplifies computation, it also flattens real-world correlations; for instance, glucose and glycosylated hemoglobin (HbA1c) do not interact with each other, despite their well-established physiological relationship.

With the TAN–Chow–Liu learner, a tree structure was generated that placed BMI at the root, from which three epidemiological trajectories radiated: smoking history, BMI and sex—capturing sex-specific patterns of consumption, age, BMI, and hypertension, reflecting the increase in blood pressure mediated by weight and aging; BMI toward heart disease; and finally diabetes, representing a canonical cardiometabolic cascade. Sustained glycemia (HbA1c) was modeled as a direct input to diabetes, while glucose and race/ethnicity remained isolated, suggesting that in this cohort chronic exposure constitutes a stronger signal than a single glucose measurement, and that ethnic variability may be negligible (see Figure 2).

Figure 3 was generated by the TAN–HCSP algorithm, enriching the tree without losing interpretability. HbA1c was modeled as the "super-parent" of diabetes, which in turn connected to age, and through age, to BMI and smoking history, eventually reaching sex and heart disease. The metabolic circuit was closed with the BMI loop associated with hypertension, which in turn connects to diabetes; glucose was also directly connected to diabetes. This architecture captured multi-step chains: how age influences BMI, BMI influences glucose, and glucose ultimately influences diabetes—balancing simplicity with deeper epidemiological detail.

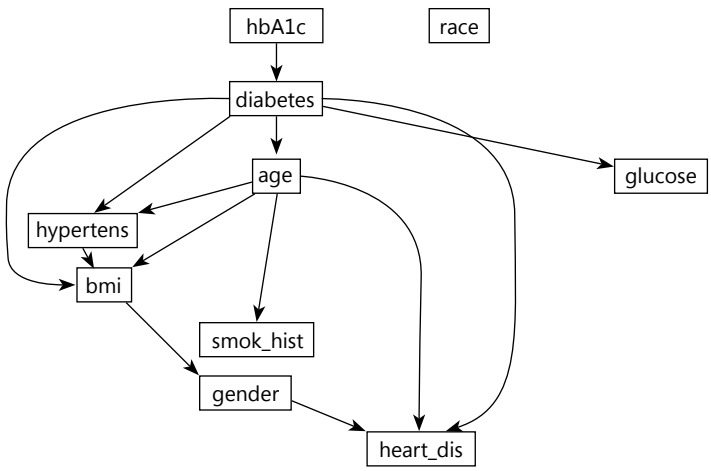

Figure 3. Bayesian model graph of TAN–HCSP (Hill-Climbing)

DISCUSSION

The risk factors identified by our model—such as high BMI, physical inactivity, and hypertension—are consistent with the clinical practice guidelines of the American Diabetes Association (ADA) and the International Diabetes Federation (IDF), thereby validating the model's ability to recognize clinically relevant patterns [27,28]. However, unlike the fixed risk thresholds used in guidelines, the AI model offers a more granular perspective by uncovering complex, non-linear interactions among variables, which could complement and personalize risk stratification in the future [29]. It is essential to interpret these findings as associations rather than direct causality, given the cross-sectional design of the study.

A direct trajectory was observed from elevated levels of glycated hemoglobin (HbA1c) to the presence of diabetes, supporting the role of chronic hyperglycemia as a central axis of progression from intermediate states to overt disease, as well as its value as both a diagnostic and prognostic marker in clinical practice [23,30,31]. Moreover, even in the absence of a formal diagnosis, high HbA1c values have been associated with an increased incidence of medium-term cardiovascular events, reinforcing its clinical relevance beyond the diagnostic threshold [31].

Conversely, the pathway connecting BMI, hypertension, and heart disease identified by the network was consistent with evidence linking adiposity to increased risk of hypertension and, subsequently, cardiovascular disease [32,33]. From a mechanistic standpoint, obesity has been associated with insulin resistance, endothelial dysfunction, and chronic inflammation—elements that promote the development of hypertension and atherosclerosis. Furthermore, hypertension has been reported to act as a major mediator of the cardiovascular consequences of obesity [33]. Accordingly, weight control has been considered a key strategy for interrupting this causal chain and reducing the burden of cardiovascular disease [32,33].

In comparison with recent literature, this study showed agreement with Bayesian models applied in prediabetes, in which HbA1c and BMI emerged as the most influential factors in progression toward type 2 diabetes and in risk stratification [23]. Likewise, the coexistence of glycemic abnormalities and cardiometabolic factors was associated with a meaningful increase in cardiovascular risk, underscoring the need to assess risk comprehensively rather than in isolation [34,35]. When considered jointly, the inferred trajectories suggest that early clinical intervention targeting glycemic control and body weight may translate into both metabolic and cardiovascular benefits [31,33,35].

A practical contribution of this approach lies in the ability of Bayesian networks to represent and communicate clinical interdependencies among variables, as well as to answer counterfactual questions ("what would happen if…?") that are useful in decision-making. This allowed visualization of direct and indirect pathways (e.g., BMI → hypertension → heart disease) and provided a basis for prioritizing interventions according to their potential impact on individual risk [30].

Among the limitations of this study is the risk of population bias, as the BRFSS-derived dataset primarily reflects the demographic profile of the United States and may underrepresent ethnic groups with different risk patterns [36]. This bias compromises the external validity of the model, since an algorithm trained on a specific population may exhibit poor and non-generalizable performance in other demographic contexts, potentially exacerbating existing health disparities—an issue widely documented in AI and health research [37,38]. Therefore, before considering clinical implementation, future research must prospectively validate the model in local and diverse cohorts using more inclusive datasets and fairness-aware machine learning methodologies to ensure that its benefits are universal and not restricted to the original study population [39].

Another fundamental aspect to consider is that the integration of AI tools into diabetes prediction entails important ethical and regulatory implications that require a proactive approach. Chief among these is the "black box" problem inherent to complex algorithms, which compromises both patient autonomy and clinician accountability, highlighting the imperative to develop explainable AI (XAI) systems that enable transparent, evidence-based medical decision-making [40]. From a regulatory perspective, agencies such as the FDA and EMA are developing frameworks for the approval of software as a medical device (SaMD), which require rigorous clinical validation, continuous post-deployment monitoring, and privacy protocols in accordance with regulations such as HIPAA and GDPR [41,42]. Finally, it is imperative to ensure equity in access to these technologies to avoid widening digital and health disparities, guaranteeing that diabetes risk-stratification algorithms become tools for universally accessible preventive

interventions rather than sources of discrimination in healthcare or insurance settings [39].

As an integrated roadmap for improvement, the following actions are proposed: a) enrichment of variables and model structure: incorporate explicit connections for sex and race/ethnicity due to their impact on diabetic and cardiovascular risk; expand the characterization of smoking (active/passive) given its indirect effects through vascular health; add dietary habits, physical activity, and family history (currently absent or insufficiently granular); extend biomarker panels (lipid profile, renal function, inflammatory markers) to capture additional pathophysiological pathways; and include genetic and environmental factors (e.g., polygenic scores, exposure to pollutants) for a more comprehensive assessment of risk; b) rigorous design and validation: conduct multicenter longitudinal studies to ensure temporal precedence and external validity; perform subgroup validation (age, sex, ethnicity, comorbidities) to audit model fairness; report calibration metrics (Brier score, calibration curves) and clinical usefulness (decision curves) in addition to discrimination; and address imbalance using sampling strategies and cost-sensitive learning, quantifying their impact on sensitivity, specificity, and equity; c) interventions and translation: design and test interventions targeting indirect chains (e.g., BMI → hypertension → heart disease), prioritizing modifiable factors (BMI, blood pressure); use counterfactual inference to estimate the impact of 5-10 % weight loss or antihypertensive intensification on future diabetes and cardiovascular risk; and integrate dynamic data (continuous glucose monitoring, home blood pressure, wearables) and temporal models (DBN/KDB) to capture drift and trajectories; d) implementation and governance: progressively integrate the models into electronic health records (EHRs) with explainable outputs (causal pathways, variable contributions), ethical safeguards, and periodic audits; establish update cycles (retraining and threshold review) with monitoring of performance, fairness, and safety; and define protocols for use (who/when/how), supported by educational materials to ensure safe and effective adoption.

Conclusions

Despite their structural differences, all classifiers exceeded an accuracy of 0.96 after cross-validation. FSSJ achieved the highest performance, with approximately 0.97 accuracy and perfect specificity, while Naive Bayes, TAN–HCSP, and KDB demonstrated consistently strong performance. TAN–CL was more sensitive to data idiosyncrasies, although it enabled the identification of new causal pathways. These results confirm that when relationships among variables are relatively simple, basic Bayesian architectures can match the performance of more elaborate variants. Even so, structure-learning tools such as *bnlearn* remain indispensable for extracting clinically interpretable connections, contributing insights that can guide targeted prevention strategies beyond what scalar performance metrics can convey.

Furthermore, the findings suggest that the models are useful for cardiometabolic risk stratification and personalized decision-making. The causal representation obtained would allow the prioritization of interventions on modifiable factors (e.g., weight control and blood pressure management) and support risk communication in terms that are understandable to both patients and clinical teams.

In this sense, the study shows that increasing architectural complexity in Bayesian classifiers does not necessarily translate into proportional predictive improvements, given that all evaluated models achieved comparable performance. The findings highlight the importance of balancing predictive accuracy with clinical interpretability, positioning these methodologies as valuable tools for precision medicine. The ability to generate causally interpretable representations facilitates both patient stratification and the development of personalized preventive strategies, representing a meaningful advance toward optimizing cardiovascular health outcomes through interventions directed at modifiable risk factors.

BIBLIOGRAPHIC REFERENCES

1. World Health Organization. Global Diabetes Report [Internet]. Geneva: WHO; September 10, 2024 [cited 2025 Jul 9]. Available from: https://iris.who.int/bitstream/handle/10665/254649/9789243565255-spa.pdf
2. Hossain J, Al-Mamun, Islam R. Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Sci Rep. [Internet]. 2024 Mar 22 [cited 2025 Jul 9];7(3):e2004. doi: 10.1002/hsr2.2004
3. Bronstein M, Meyer-Kalos P, Vinogradov S, Kummerfeld E. Causal Discovery Analysis: A Promising Tool for Precision Medicine. Psychiatr Ann. [Internet]. 2024 [cited 2025 Jul 9];54(4):e119-e124. https://doi.org/10.3928/00485713-20240308-01
4. Montero Rodríguez JC de J, Roshan Biswal R, Sánchez de la Cruz E. Algoritmos de aprendizaje automático de vanguardia para el diagnóstico de enfermedades. Res Comput Sci. [Internet]. 2019 [cited 2025 Jul 9];148(7):455-68. Available from: https://rcs.cic.ipn.mx/2019_148_7/Algoritmos%20de%20aprendizaje%20automatico%20de%20vanguardia%20para%20el%20diagnostico%20de%20enfermedades.pdf

5. Gómez Ruiz I. Diseño e implementación de modelos de lenguaje para información genómica asociada a enfermedades raras mediante inferencia gramatical [Internet]. Valencia: Universitat Politècnica de València; 2024 [cited 2025 Jul 9]. Available from: https://riunet.upv.es/server/api/core/bitstreams/e752670f-3702-4eee-846b-c16237a5f925/content

6. Darwiche A. Modeling and Reasoning with Bayesian Networks [Internet]. Cambridge: Cambridge University Press; 2009 [cited 2025 Jul 9]. Available from: https://books.google.co.ve/books?id=7AjXGltje7YC&printsec=frontcover#v=onepage&q&f=false

7. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn Comput. [Internet]. 2024 [cited 2025 Jul 9];16:45-74. doi: 10.1007/s12559-023-10179-8

8. Lucas PJ, Van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artif Intell Med. [Internet]. 2004 [cited 2025 Jul 9];30(3):201-14. doi: 10.1016/j.artmed.2003.11.001

9. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. [Internet]. Cambridge: MIT Press; 2009 [cited 2025 Jul 9]. Available from: https://www.researchgate.net/publication/220690050_Probabilistic_Graphical_Models_Principles_and_Techniques

10. Pearl J. Causality: Models, Reasoning and Inference. [Internet]. 2nd ed. Cambridge: Cambridge University Press; 2009 [cited 2025 Jul 9]. Available from: https://dl.acm.org/doi/book/10.5555/1642718

11. Suo X, Huang X, Zhong L, Luo Q, Ding L, Xue F. Development and Validation of a Bayesian Network□Based Model for Predicting Coronary Heart Disease Risk From Electronic Health Records. J Am Heart Assoc. [Internet]. 2024 Jun 2 [cited 2025 Jul 9];13(1):e029400. doi: 10.1161/JAHA.123.029400

12. Coaquira-Flores EE, Torres-Cruz F, Condori-Quispe SJ, Tisnado-Puma JC, Melgarejo-Bolivar RP, Herrera-Urtiaga AP, et al. Predicción de diabetes en mujeres mediante un modelo probabilístico basado en redes bayesianas. Científica Digit. [Internet]. 2023 Apr 29 [cited 2025 Jul 9];16:185-201. doi: 10.37885/230412748

13. Bressan GM, Flamia de Azevedo BC, Molina de Souza R. Métodos de classificação automática para predição do perfil clínico de pacientes portadores do diabetes mellitus. Braz J Biometrics. [Internet]. 2020 Jun 29 [cited 2025 Jul 9];38(2):257-73. https://doi.org/10.28951/rbb.v38i2.445

14. Ndjaboue R, Ngueta G, Rochefort-Brihay C, Delorme S, Guay D, Ivers N, et al. Prediction models of diabetes complications: a scoping review. J Epidemiol Community Health [Internet]. 2022 Jun 30 [cited 2025 Jul 9];76(10):896-904. doi: 10.1136/jech-2021-217793

15. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman C, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS One [Internet]. 2017 [cited 2025 Jul 9];12(7):e0179805. doi: 10.1371/journal.pone.0179805

16. Nejatian S, Parvin H, Faraji E. Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification. Neurocomputing [Internet]. 2018 Feb 7 [cited 2025 Jul 9];276:55-66. https://doi.org/10.1016/j.neucom.2017.06.082

17. Praveenkumar KS. Un enfoque híbrido de analítica de big data para predecir diabetes tipo II usando H-SMOTE tree. Adv Nanotechnol Mater Sci Eng Innov. [Internet]. 2024 [cited 2025 Jul 9];20(S2):606-624. https://doi.org/10.62441/nano-ntp.vi.494

18. Xu Z, Wang Z. A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. In: 11th Int Conf Adv Comput Intelligence (ICACI); Guilin, China; 2019 Jun 7–9. [Internet]. Guilin: IEEE; 2019:278–283. [cited 2025 Jul 9]. doi:10.1109/ICACI.2019.8778622

19. Pes B, Lai G. Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study. PeerJ Comput Sci. [Internet]. 2021 [cited 2025 Jul 9];7:e832. doi: 10.7717/peerj-cs.832

20. Wang X, Ren J, Ren H, Song W, Qiao Y, Zhao Y, et al. Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta. Sci Rep. [Internet]. 2023 [cited 2025 Jul 9];13:12718. doi: 10.1038/s41598-023-40036-5

21. Parrales-Bravo F, Caicedo-Quiroz R, Rodríguez-Larraburu E, Barzola-Monteses J. ACME: A Classification Model for Explaining the Risk of Preeclampsia Based on Bayesian Network Classifiers and a Non-Redundant Feature Selection Approach. Informatics [Internet]. 2024 [cited 2025 Jul 9];11(2):31. https://doi.org/10.3390/informatics11020031

22. Kong D, Chen R, Chen Y, Zhao L, Huang R, Luo L, et al. Bayesian network analysis of factors influencing type 2 diabetes, coronary heart disease, and their comorbidities. BMC Public Health. [Internet]. 2024 May 8 [cited 2025 Jul 9];24:1267. doi: 10.1186/s12889-024-18737-x

23. Fuster-Parra P, Yañez AM, López-González A, Aguiló A, Bennasar-Veny M. Identifying risk factors of developing type 2 diabetes from an adult population with initial prediabetes using a Bayesian network. Front Public Health. [Internet]. 2023 [cited 2025 Jul 9];10:1035025. doi: 10.3389/fpubh.2022.1035025

24. Sun Y, Lei J, Kosmas P. Exploring Biomarker Relationships in Both Type 1 and Type 2 Diabetes Mellitus Through a Bayesian Network Analysis Approach. arXiv [Preprint]. 2024; arXiv:2406.17090. https://doi.org/10.48550/arXiv.2406.17090

25. Choksi P. Comprehensive clinical diabetes dataset (100k rows) [dataset on the Internet]. Kaggle; 2024 [cited 2025 Jul 9]. Available from: https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset

26. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. [Internet]. 2019 [cited 2025 Jul 9];380(14):1347-58. doi: 10.1056/NEJMra1814259

27. ElSayed N, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes-2023. Diabetes Care [Internet]. 2023 [cited 2025 Jul 9];46(Suppl 1):S19-S40. doi: 10.2337/dc23-S002

28. International Diabetes Federation. IDF Diabetes Atlas [Internet]. 10th ed. Brussels: International Diabetes Federation; 2021. Available from: https://diabetesatlas.org/

29. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. [Internet]. 2018 [cited 2025 Jul 9];319(13):1317-1318. doi: 10.1001/jama.2017.18391

30. Zhang J, Zhang Z, Zhang K, Ge X, Sun R, Zhai X. Early detection of type 2 diabetes risk: limitations of current diagnostic criteria. Front Endocrinol (Lausanne). [Internet]. 2023 [cited 2025 Jul 9];14:1260623. doi: 10.3389/fendo.2023.1260623

31. Butalia S, Chu LM, Dover DC, Lau D, Yeung RO, Eurich DT, et al. Association Between Hemoglobin A1c and Development of Cardiovascular Disease in Canadian Men and Women Without Diabetes at Baseline: A Population-Based Study of 608 474 Adults. J Am Heart Assoc. [Internet]. 2024 [cited 2025 Jul 9];13(9):e031095. doi: 10.1161/JAHA.123.031095

32. Lin H, Xiao N, Lin S, Liu M, Liu GG. Associations of hypertension, diabetes and heart disease risk with body mass index in older Chinese adults: a population-based cohort study. BMJ Open [Internet]. 2024 [cited 2025 Jul 9];14(7):e083443. doi: 10.1136/bmjopen-2023-083443

33. Volpe M, Gallo G. Obesity and cardiovascular disease: An executive document on pathophysiological and clinical links promoted by the Italian Society of Cardiovascular Prevention (SIPREC). Front Cardiovasc Med. [Internet]. 2023 [cited 2025 Jul 9];10:1136340. doi: 10.3389/fcvm.2023.1136340

34. Ahmad A, Lim LL, Morieri ML, Tam CH, Cheng F, Chikowore T, et al. Precision prognostics for cardiovascular disease in Type 2 diabetes: a systematic review and meta-analysis. Commun Med (Lond). [Internet]. 2024 [cited 2025 Jul 9];4(1):11. doi: 10.1038/s43856-023-00429-z

35. Bruemmer D, Singh A. Cardiometabolic Risk: Shifting the Paradigm Toward Comprehensive Assessment. JACC Adv. [Internet]. 2023 [cited 2025 Jul 9];2(18):100868. doi: 10.1016/j.jacadv.2024.100867

36. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. [Internet]. 2019 [cited 2025 Jul 9];366(6464):447-453. doi: 10.1126/science.aax2342

37. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ Digit Med. [Internet]. 2020 [cited 2025 Jul 9];3:81. doi: 10.1038/s41746-020-0288-5

38. Chen IY, Johansson FD, Sontag D. Why Is My Classifier Discriminatory? Adv Neural Inf Process Syst. [Internet]. 2018 [cited 2025 Jul 9];31. doi: 10.48550/arXiv.1805.12002

39. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. [Internet]. 2019 [cited 2025 Jul 9];25(9):1337-340. doi:110.1038/s41591-019-0548-6

40. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health [Internet]. 2021 [cited 2025 Jul 9];3(11):e745-e750. doi: 10.1016/S2589-7500(21)00208-9

41. Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device [Internet]. United States: FDA; 2024 [cited 2025 Jul 9]. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

42. European Commission. Proposal for a Regulation on a European Health Data Space [Internet]. Brussels: COM(2022) 197 final; 2022 [cited 2025 Jul 9]. Available from: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space

Funding sources

This research was self-funded.

Conflict of interest statement

The author declares no conflicts of interest.

Authorship contribution

Conceptualization, methodology, formal analysis, research, resources, writing—original draft, writing—revision and editing, and visualization.