XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

Application of Graph Theory in the Analysis of Risk Factors for Type 2 Diabetes

Grillo, E. da S. T.[1], Borges, P. A. S.[2], Dias, T. M. R.[1], Silva, A. M.[1], Silva, M. P.[1]

[1]*Computing Dept. (DECOM), CEFET-MG*
Rua Álvares de Azevedo St., 400, Bela Vista, 35503-822, Divinópolis - MG, Brazil
edgrillocefet@gmail.com, thiagomagela@cefetmg.br, alisson@cefetmg.br, michel@cefetmg.br
[2]*Postgraduate Electrical Engineering, Federal University of Uberlândia*
Avenida João Naves de Ávila, 2121, 38408-100, Uberlândia - MG, Brazil
paulo.engmeca@hotmail.com

Abstract. Type 2 diabetes represents one of the greatest challenges in global public health, due to its complex relationship with metabolic, behavioral, and demographic factors. Traditional statistical methods have shown limitations in analyzing the interdependencies among these risk factors, hindering the identification of relevant patterns for the prevention and management of the disease. This study proposes the application of graph theory as a computational tool to model the connections among clinical variables such as hypertension, obesity, high cholesterol, mental health, and mobility. Using real-world data and correlation analysis techniques, a weighted network of interactions was constructed, allowing a structured and intuitive visualization of these relationships. The results reveal clusters of strongly connected variables that underscore the central role of obesity and hypertension in the type 2 diabetes risk network, as well as relevant associations with psychosocial and functional factors.

Keywords: Type 2 diabetes, graph theory, risk factors, hypertension, obesity.

1 Introduction

Type 2 diabetes mellitus (T2DM) is a chronic, progressive, and multifactorial metabolic syndrome, accounting for up to 95% of diabetes cases worldwide and associated with severe complications such as cardiovascular disease, nephropathy, and neuropathy [1, 2]. It is mainly characterized by insulin resistance and impaired hormone secretion by pancreatic β cells, being influenced by genetic factors, aging, obesity, physical inactivity, dyslipidemia, and inadequate dietary patterns. In addition, demographic aspects such as age, gender, and ethnicity also affect its prevalence. Despite the importance of the topic, understanding the interactions among the multiple risk factors involved still represents a challenge, as traditional statistical analysis methods are not always effective in revealing the complexity and interdependence between clinical, behavioral, and socioeconomic variables [3].

In this context, graph theory emerges as an approach to model complex systems, allowing variables to be represented as nodes and their interrelationships as edges weighted by association metrics, such as correlation. This structure facilitates the identification of hidden patterns and relationships that would go unnoticed in conventional analyses [3–5]. Rather than focusing exclusively on T2DM, this study aims to explore the use of graph modeling as a tool to represent, visualize, and interpret the connections between health-related variables, using T2DM as a practical case to demonstrate the potential of this approach.

Based on public data provided in [6], a correlation matrix was developed between clinical, behavioral, and demographic variables, which was then used to generate a weighted graph. This graphical representation allows not only the identification of variables with higher centrality in the network but also an understanding of how modifiable factors influence multiple aspects of the system under analysis.

The structure of this article has been organized to contextualize the problem, justify the methodological choice, and systematically present the adopted procedures and the results obtained. Initially, a theoretical discussion is conducted on the limitations of traditional statistical approaches in identifying complex interrelationships among multiple variables, especially in contexts with high multicollinearity, such as that of factors associated with T2DM. Next, the foundations and application of graph theory as an alternative tool are presented, highlighting its ability to visually and structurally represent multivariate systems. The modeling section details the process of constructing the correlation matrix, the criteria for defining weights and edges, and the transformation of this matrix

into a weighted graph. Subsequently, the results are explored using topological metrics (such as degree, centrality, and clustering) to identify relevant patterns, key variables, and possible groupings of factors with similar behaviors. Finally, the implications of the findings, their limitations, and potential future applications of the approach in similar contexts are discussed.

This study aims to contribute to the field by adopting graph theory as an innovative approach to represent and analyze multivariate systems. By emphasizing the value of structural visualization and quantification of relationships between variables, it is hoped that this work will not only deepen the understanding of the studied context but also encourage the use of graphs as a complementary tool in population health studies and intervention planning.

2 Literature Review

The study of risk factors associated with type 2 diabetes mellitus (T2DM) has attracted increasing interest due to the complexity of interactions among metabolic, behavioral, and demographic variables. Several studies use statistical and computational methods to explore these relationships, ranging from traditional regression analyses to more advanced techniques such as machine learning and graph modeling.

Graph theory provides a solid mathematical foundation for representing and analyzing systems composed of multiple interconnected entities. In practice, a graph is formed by a set of vertices (or nodes) and a set of edges that connect pairs of vertices; these edges may be directed or undirected, and weighted or unweighted. This structure is widely used to model complex networks in fields such as biology, social networks, and engineering. In the context of this study, an undirected weighted graph was employed, in which each node represents a variable from the dataset (such as BMI, age, family history, cholesterol, among others), and each edge represents the correlation between these variables, weighted by the Pearson coefficient. This representation enables the identification of central variables in the network, evaluation of connectivity levels, and exploration of natural groupings based on the strength of associations, providing a more holistic perspective on factors related to T2DM.

In [1], the authors investigated the maintenance of glycemic levels in T2DM patients, correlating these indicators with lifestyle habits such as diet and physical activity. Although the study identified relevant patterns, its approach was limited to isolated analysis of variables, without considering broader interdependencies among the examined factors. Complementarily, the work of Malta et al. [3] analyzed factors related to diabetes in Brazilian adults using data from the National Health Survey (PNS). In that study, the authors observed statistical associations between T2DM, hypertension, obesity, and physical inactivity; however, the absence of a visual modeling of the connections hindered the identification of more complex patterns.

Other studies have also addressed specific aspects of T2DM. For example, Chen et al. [7] investigated the relationship between carotenoid levels and the risk of metabolic syndrome, providing valuable biochemical insights. However, these studies focused on limited subgroups of variables, without considering the systemic interrelation that characterizes T2DM. Similarly, Araujo et al. [4] applied Pearson correlation matrices in operational studies but did not advance toward transforming these matrices into graphical representations, which could offer a more intuitive and detailed view of variable interdependencies.

This approach complements studies such as that by Lopes et al. [5], who introduced exploratory data techniques using the Python language, and by Varges et al. [1], who emphasized the importance of glycemic control in T2DM management. By integrating advanced computational methods with intuitive graphical representations, this study stands out by offering a comprehensive analysis of the interactions among variables related to type 2 diabetes. Such contributions reinforce the relevance of robust analytical tools in addressing the multifactorial challenges associated with this condition.

3 Methodology

The flowchart shown in Figure 1 describes the methodology applied for constructing a graph based on the correlation analysis between variables. The process begins with the calculation of the correlation matrix, followed by the verification of valid correlations between variable pairs. When the correlation is considered significant, edges and nodes representing that relationship are drawn. Otherwise, the process moves on to generating the graph layout, which, once defined, allows for the visual display of the final structure. This approach enables an intuitive representation of the most relevant interactions among the analyzed factors.

In this study, the methodology was organized into three main stages: tools used, data preparation, and graph modeling. Each of these stages is detailed below.

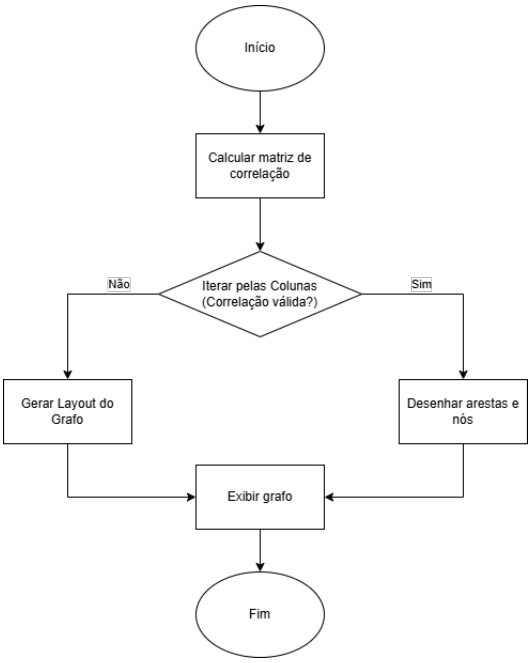


Figure 1. Flowchart of the graph construction methodology based on correlation analysis between variables.

3.1 Data Pre-processing

Before implementing the model, it is essential to ensure that the environment used is equivalent to the one in which data processing and analysis were performed. This ensures that procedures are reproducible and reliable in future studies. The implemented *dataset* is publicly available on the Kaggle repository and archived on **Zenodo** for future reference.

- **Data reading:** using the `pandas` library, converting the CSV file into a `DataFrame` for tabular manipulation.
- **Identification and removal of missing values:** to avoid inconsistencies in correlation calculations.
- Transformation of categorical variables into numerical ones using the *Label Encoding* method.
- Normalization of continuous variables to the range $[0, 1]$ using the *Min-Max Scaling* technique.

These steps are crucial to ensure data integrity and comparability, providing a solid foundation for graph construction and analysis.

3.2 Graph Modeling

The construction of the overall graph was based on a correlation matrix calculated using the `pandas` library. To quantify the strength and direction of the linear relationship between variables, the Pearson correlation coefficient was used. The mathematical formula applied for this calculation is presented below:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{1}$$

where X_i and Y_i represent the individual values of variables X and Y, $\bar{X}$ and $\bar{Y}$ denote their means, and n is the total number of observations.

The resulting correlation matrix was converted into an undirected and weighted graph. In this graph, the nodes represented the variables in the dataset, while the edges connected variables based on their correlation values. The edge weights were defined by the magnitude of the correlation coefficients.

The graph was visualized using the `networkx` library, applying the *Spring Layout* algorithm, which arranges the nodes in a way that maximizes structural clarity and the aesthetic of the graphical representation.

4 Results

The data analysis revealed striking patterns in the distribution of type 2 diabetes and its associations with risk factors. Figure 2 shows the distribution of cases by age group, highlighting a progressive increase in the number of cases starting at age 40, with a peak between 65 and 69 years. This behavior is consistent with previous epidemiological studies that indicate a positive correlation between aging and the prevalence of chronic metabolic conditions such as type 2 diabetes.

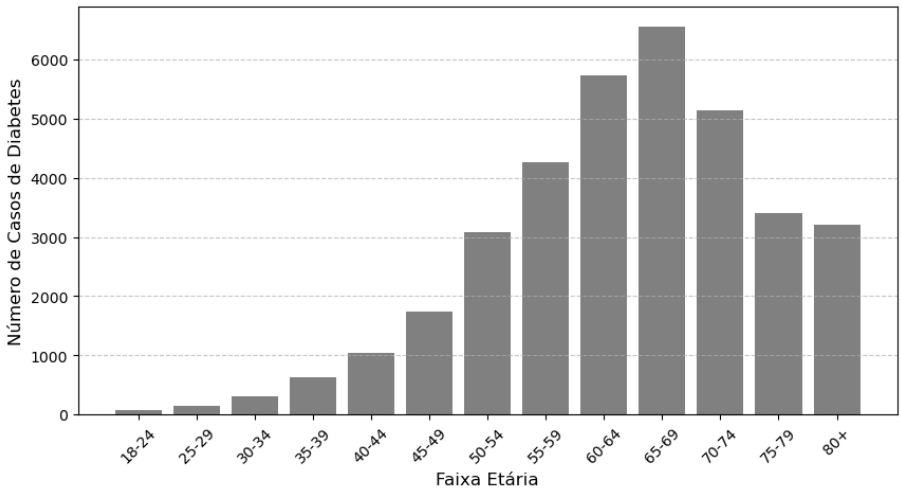


Figure 2. Distribution of diabetes cases by age group. The data show a progressive increase in the number of cases starting at age 40, with a peak between 65 and 69 years.

Additionally, the gender-based analysis revealed an interesting pattern, as shown in Figure 3. Although the absolute number of female patients is higher, the proportion of confirmed diabetes cases (code 1.0) between genders is relatively balanced. This finding suggests that, regardless of gender, factors such as aging and metabolic conditions play a central role in the prevalence of type 2 diabetes.

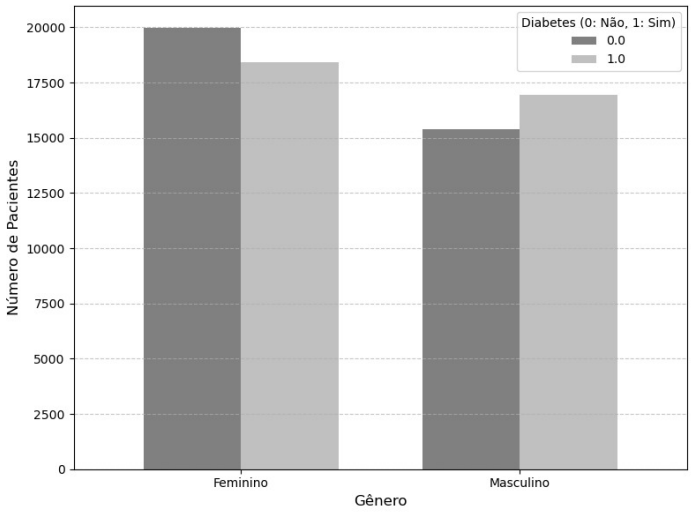


Figure 3. Distribution of diabetes cases by gender. The chart compares the number of male and female patients, distinguishing between individuals diagnosed with diabetes (1) and those without the condition (0).

Figure 4 presents the overall correlation graph, in which each node represents a variable from the dataset, while the edges indicate the strength of the correlation between them. This graph illustrates the complex network of interrelationships among diabetes, comorbidities, and behavioral factors, providing a broad and integrated view of the connections that influence metabolic health.

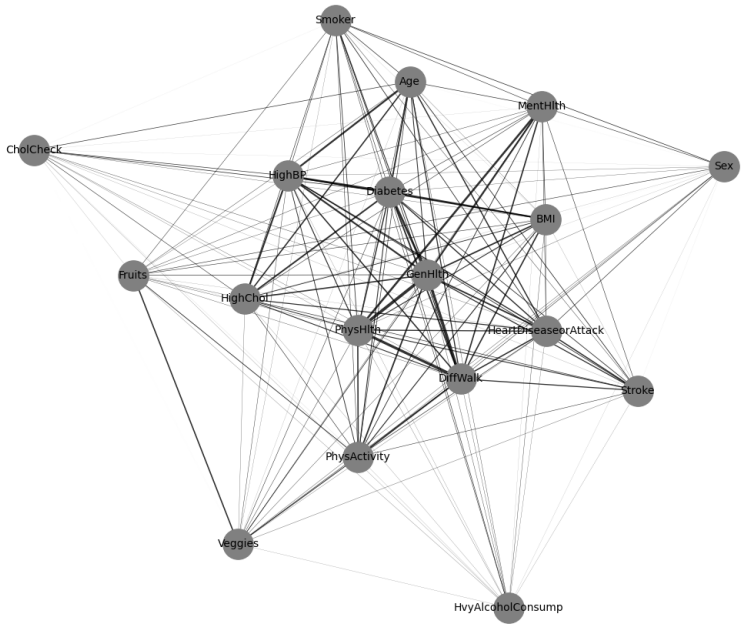


Figure 4. Overall correlation graph showing relationships among variables related to type 2 diabetes. Thicker edges indicate stronger correlations.

To simplify interpretation, a filtered graph was generated (Figure 5), retaining only the most significant connections. This graph highlights key variables such as hypertension (HighBP), high cholesterol (HighChol), body mass index (BMI), and history of cardiovascular disease (HeartDiseaseorAttack), reinforcing the importance of these variables in the context of diabetes.

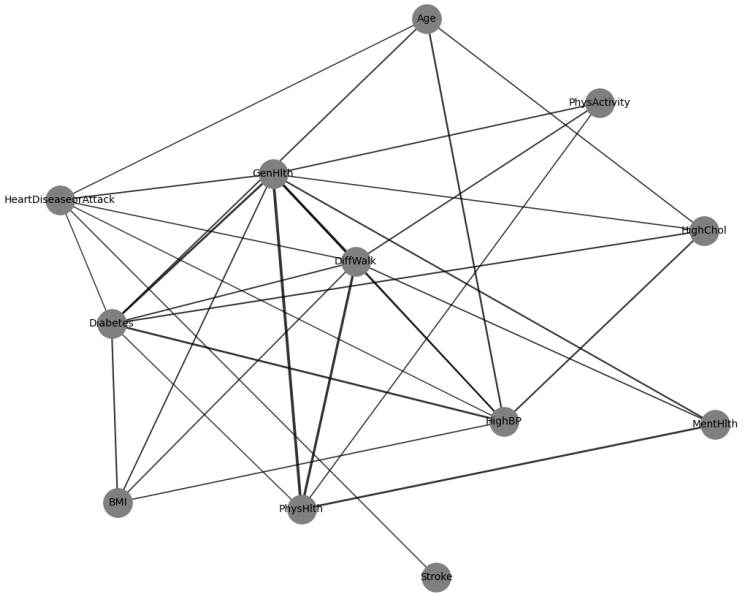


Figure 5. Filtered correlation graph showing only the strongest associations between diabetes and other health-related variables.

The results presented provide a comprehensive overview of the interrelationships among risk factors associated with type 2 diabetes, highlighting how demographic, behavioral, and metabolic variables are interconnected. The age group analysis (Figure 2) showed that advancing age is closely associated with increased prevalence of type 2 diabetes, especially in individuals over 40 years old. This pattern supports epidemiological studies that attribute this phenomenon to factors such as increased insulin resistance, reduced pancreatic β-cell function, and greater accumulation of visceral fat, all of which contribute to the development of the disease [8].

The gender-based analysis (Figure 3) revealed that although the absolute number of women diagnosed with diabetes is higher, the proportional prevalence between men and women is balanced. This suggests that metabolic factors such as insulin resistance and obesity play a more predominant role in the development of the disease than gender-related differences. However, previous studies indicate that women may face additional challenges in managing diabetes, such as higher rates of comorbidities and differences in treatment response, which warrants more detailed investigation.

The generated graphs (Figures 4 and 5) highlighted the most influential variables and their interrelationships in the context of type 2 diabetes. The overall graph revealed a wide network of connections among clinical, behavioral, and demographic variables, emphasizing, for instance, the strong correlation between hypertension (HighBP) and high cholesterol (HighChol) with diabetes. These results support studies that associate these metabolic conditions with underlying mechanisms such as chronic inflammation and insulin resistance, which are key determinants in the development of type 2 diabetes.

On the other hand, the filtered graph simplified the analysis by highlighting more significant connections, such as the association between obesity (BMI) and diabetes. Another important connection identified was between reduced physical activity (PhysActivity) and difficulty walking (DiffWalk). These findings reinforce the importance of lifestyle-based interventions, including regular exercise programs and dietary changes, which can substantially reduce the risk of diabetes.

Another noteworthy finding was the connection between mental health (*MentHlth*) and type 2 diabetes. Recent studies indicate that disorders such as anxiety and depression are associated with an increased risk of diabetes, whether due to hormonal changes that affect glucose metabolism or due to difficulty maintaining healthy habits. These data emphasize the need to integrate psychological support into diabetes management strategies, developing multidisciplinary approaches to address both the emotional and physical aspects of the disease [1].

5 Conclusions

This study highlighted the potential of graph theory as a powerful analytical tool for identifying risk factors associated with type 2 diabetes, offering an innovative and visual approach to exploring complex patterns of interdependence between variables. Based on the generated correlation matrix, it was possible to model a network that revealed significant connections among diabetes, metabolic and cardiovascular comorbidities, as well as behavioral factors.

The results obtained identified hypertension, high cholesterol, obesity, and history of cardiovascular disease as central variables within the correlation network, corroborating findings widely reported in the scientific literature. Additionally, the inclusion of variables such as mental health and difficulty walking provided a broader and more multidimensional perspective, highlighting the psychological and functional impact that type 2 diabetes has on patients' quality of life. These findings reinforce the need for preventive and multidisciplinary interventions that combine metabolic control, psychological support, and the promotion of regular physical activity.

The presented methodology demonstrated clear advantages over traditional statistical approaches by offering a structured and intuitive representation of the relationships among variables. The graphical visualization of the connections allows for a clearer understanding of complex interrelationships, facilitating the communication of results and the formulation of evidence-based strategies. This approach has the potential to be expanded and applied in other public health contexts, contributing to the development of more effective and integrated policies for managing chronic conditions such as type 2 diabetes and its comorbidities.

As future perspectives, it is recommended to integrate predictive techniques based on machine learning with network analysis, aiming to forecast the progression of diabetes in different populations and contexts. This integration would allow for the identification of emerging patterns and anticipation of trends, enhancing the effectiveness of preventive interventions. Furthermore, the inclusion of longitudinal data could deepen the understanding of the temporal dynamics among risk factors, enabling a more detailed analysis of the disease's evolution over time.

By applying this approach, the goal is to contribute to more efficient and personalized management of type 2 diabetes, promoting improvements in patients' health and quality of life. Additionally, the data from this study can support the formulation of more integrated public health strategies focused on both prevention and disease management, ensuring a positive impact across different populations and socioeconomic contexts.

Acknowledgements. The authors thank CEFET-MG and UFU for the institutional support.

Authorship statement. This article uses the CRediT (Contributor Roles Taxonomy) system to specify individual author contributions:

- Borges, P. A. S.[2]: Investigation, Data curation, Validation, Writing – review & editing.
- Dias, T. M. R.[1]: Methodology, Software, Formal analysis, Visualization, Writing – original draft.
- Silva, A. M.[1]: Conceptualization, Supervision, Project administration, Funding acquisition.
- Silva, M. P.[1]: Investigation, Validation, Review & Editing, Writing – review & editing.

References

[1] J. Varges, T. Santos, and M. Barros. Avaliação da manutenção de níveis glicêmicos em pacientes portadores de diabetes tipo 2, correlacionando com seus hábitos de vida. *Research, Society and Development*, vol. 10, pp. e38710716857, 2021.

[2] dos V. C. Santos. Diabetes mellitus tipo 2 - aspectos epidemiológicos, fisiopatológicos e manejo terapêutico. *Brazilian Journal of Development*, vol. 9, n. 3, pp. 9737–9749, 2023.

[3] D. Malta, R. Bernal, A. Sá, T. Silva, B. Iser, B. Duncan, and others. Diabetes autorreferido e fatores associados na população adulta brasileira: pesquisa nacional de saúde, 2019. *Ciência & Saúde Coletiva*, vol. 27, n. 7, pp. 2643–2653, 2022.

[4] J. V. G. A. Araujo, dos M. Santos, and C. F. S. Gomes. Desenvolvimento de um código em python para geração de matrizes de correlação de pearson com laços a partir de 'n' variáveis tomadas duas a duas. In *Simpósio de Pesquisa Operacional e Logística da Marinha*, Rio de Janeiro, RJ, Brasil, 2019.

[5] G. Lopes, A. Almeida, A. Delbem, and C. Toledo. Introdução à análise exploratória de dados com python, 2019.

[6] da E. Silva Torres Grillo. Diabetes_data. `https://zenodo.org/records/14806872`. Zenodo, 2025.

[7] M. Chen, S. Cai, Q. Jia, Y. Suo, Y. Tang, Y. Shi, X. Zhu, and H. Zhang. Inverse relationship between serum carotenoid levels and cardiovascular-kidney-metabolic syndrome among the general adult population. *Journal of Diabetes*, vol. 17, pp. e70046, 2025.

[8] K. C. P. McLellan, S. M. Barbalho, M. Cattalini, and A. C. Lerário. Diabetes mellitus do tipo 2, síndrome metabólica e modificação no estilo de vida. *Revista De Nutrição*, vol. 20, n. 5, pp. 515–524, 2007.