Random Forest Model Based on Machine Learning for Early Detection of Diabetes

Inooc Rubio Paucar[1], Cesar Yactayo-Arias[2], Laberiano Andrade-Arenas[3]
Facultad De Ingeniería Y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1]
Departamento De Estudios Generales, Universidad Continental, Lima, Perú[2]
Facultad De Ciencias E Ingeniería, Universidad De Ciencias Y Humanidades, Lima, Perú[3]

***Abstract*—Diabetes mellitus presents a growing prevalence at the global level, representing a significant public health challenge. Despite the availability of specific treatments, it is imperative to develop innovative strategies that optimize early detection and management of the disease. The research aims to develop a model that allows for the early detection of diabetes using the Random Forest algorithm, using the Knowledge Discovery in Databases (KDD) methodology, which comprises the phases of selection, preprocessing, transformation, data mining, interpretation and evaluation. The dataset used include 520 randomly selected patient records. The model achieved robust performance, with an accuracy of 85%, sensitivity of 75%, and an F1-score of 78%, indicating an adequate balance between precision and sensitivity. Specificity was 78%, while the area under the ROC curve (AUC) reached 86%, demonstrating a high discriminative ability between positive and negative cases. The balanced accuracy was 82%, and the Matthews correlation coefficient (MCC) registered a value of 0.72, confirming the strength and reliability of the model even in the presence of class imbalance. These results demonstrate the effectiveness of the machine learning-based approach for the early detection of diabetes mellitus, with potential application in clinical decision support systems.**

***Keywords*—Data mining; decision tree; diabetes mellitus; machine learning; random forest**

I. Introduction

Diabetes has been recognized by the World Health Organization (WHO) as a major global health concern. It currently affects around 9.3 % of adults worldwide, with an increasing number of cases reported in low- and middle-income regions. The condition shows a slight difference between genders, with a prevalence of 9.6% in men and 9.0% in women. Furthermore, the likelihood of developing diabetes rises significantly with age, especially among individuals over the age of 45 [1]. This disease poses serious health risks, contributing to complications such as heart disease, kidney damage, vision loss, and limb amputations, all of which generate considerable health care burdens and socioeconomic consequences. Diabetes is a long-term illness that continues to grow in prevalence worldwide and can lead to severe health issues if not identified and treated in its early stages. This research is relevant because it addresses the critical need for early diagnosis, which plays a key role in minimizing complications and improving the overall well-being of those affected. Detecting the disease early is not only beneficial for individual patients but also essential for improving public health outcomes [2].

Diabetes is a chronic condition characterized by high blood glucose levels due to the body's inability to produce or effectively use insulin. Diagnosis is made through fasting blood glucose tests, the glucose tolerance test, and hemoglobin A1c levels [3]. The main causes of diabetes include genetic factors, obesity, lack of physical activity, and poor dietary habits, particularly in individuals with a family history of the disease. The consequences of poorly managed diabetes are severe, as it can lead to complications such as cardiovascular diseases, kidney damage, blindness, amputations, and a higher risk of infections. Additionally, it significantly impacts the patient's quality of life and represents an economic burden on healthcare systems worldwide [4].

This study offers both theoretical and practical contributions. On the theoretical side, it explores how machine learning—specifically the Random Forest algorithm—can be effectively used to predict the risk of diabetes by analyzing various clinical and demographic factors. On the practical side, the results can support healthcare providers by offering a reliable tool for early risk identification, helping to streamline diagnostic procedures and support preventive healthcare strategies.

Methodologically, the research is grounded in the Knowledge Discovery in Databases (KDD) process, which provides a step-by-step framework for extracting useful patterns from large datasets. By following the KDD stages—data selection, cleaning, transformation, mining, and interpretation—the study ensures that the model development is systematic, robust, and applicable to real-world healthcare scenarios. The objective of the research is to develop an early detection model for diabetes using the Random Forest algorithm, then the aim is to apply machine learning techniques to identify significant patterns within clinical and personal patient data that contribute to accurate and timely diagnosis.

II. Literature Review

In this section, the findings of different authors who have researched diabetes are analyzed.

A. Theoretical Bases

1) Diabetes mellitus: In the context of health, diabetes mellitus is classified as a chronic metabolic disease, primarily characterized by elevated blood glucose levels (hyperglycemia). This condition, if not properly controlled, can trigger severe long-term complications affecting various organs and systems of the body, such as the cardiovascular, renal, ocular, and peripheral nervous systems [5]. From a physiological point of view, insulin is a hormone produced by the beta cells of the islets of Langerhans in the pancreas, whose main function is to

facilitate the transport of glucose from the bloodstream into the cells to be used as an energy source. In patients with diabetes, this function is compromised due to a deficiency in insulin production (Type I diabetes) or resistance to its action (Type II diabetes), which prevents the proper entry of glucose into the cells, causing its accumulation in the blood and altering energy metabolism [6].

Type I diabetes is characterized by the self-destruction of the beta cells of the pancreas, which are responsible for insulin production. This destruction is mistakenly caused by the immune system, which identifies these cells as foreign agents. As a consequence, an absolute insulin deficiency occurs, preventing proper glucose metabolism and leading to elevated blood glucose levels [7]. Besides, Type II diabetes mellitus, the pathophysiology is different. This condition is characterized by insulin resistance in peripheral tissues, accompanied by a progressive dysfunction of the pancreatic beta cells responsible for its secretion. As a result, glucose cannot be effectively utilized by skeletal muscle, adipose tissue, and the liver, leading to chronic hyperglycemia. This metabolic disturbance worsens over time, as the pancreas's ability to compensate for insulin resistance by increasing insulin production becomes progressively impaired [8].

2) Random forest: Random Forest is a machine learning algorithm that works by combining many decision trees to obtain more accurate and stable results. Through a process called bagging, the model learns from multiple samples of the same dataset and makes decisions considering different combinations of variables, which allows it to be more reliable and less prone to errors [9][10]. This technique is especially useful in the healthcare field, as it enables the analysis of a large amount of clinical and personal information from patients, detecting complex patterns that may not be evident at first glance [11]. In this research, Random Forest is used as a tool to support the early diagnosis of diabetes, aiming to facilitate timely interventions and improve people's quality of life.

B. Related Work

The author [12] aimed to identify systemic risk factors associated with diabetes mellitus (DM) and to predict the onset of diabetic retinopathy (DR) by implementing a classification model based on Random Forest (RF). Patients diagnosed with DM who attended a specialized retina consultation for DR screening for the first time were included. Clinical and demographic variables were collected, including age, sex, type of diabetes, metabolic control, family history, and comorbidities. The presence of DR and vision-threatening diabetic retinopathy (VTDR) was established through a dilated fundus examination. The dataset, consisting of 1,416 patients, was split into 80:20 proportions for training (1,132 cases) and testing (284 cases). The RF model demonstrated optimal performance in detecting DR, with a 0% error rate during training and 100% accuracy on the test set. For detecting VTDR, the model achieved 76% accuracy, with a sensitivity of 53% and specificity of 80%. These results demonstrate that the RF-based approach is highly effective for predicting DR using systemic variables, which could optimize screening processes by reducing unnecessary referrals. However, the need to validate the model in larger and more diverse populations to confirm its clinical applicability is emphasized.

Diabetes mellitus represents a growing public health concern globally, as mentioned by the author [13]. In this study, the main objective was the development of predictive models based on machine learning aimed at the clinical classification of diabetes. For this purpose, an application was built based on the integration of a structured database containing both continuous and categorical variables corresponding to risk factors. Various supervised learning algorithms were implemented, including logistic regression, Gaussian Naive Bayes, linear discriminant analysis (LDA), support vector machines (SVM), k-nearest neighbors (KNN), decision trees, Extreme Gradient Boosting (XGBoost), kernel entropy component analysis, and Random Forest. The adopted methodology was structured around three key stages: feature extraction, classification, and prediction. Experimental results indicated that the Random Forest model achieved the best performance, reaching a maximum accuracy of 99.84% on imbalanced data and 96.75% on balanced data. Additionally, the SVM classifier, combined with kernel entropy components, achieved an accuracy of 99.64% on the balanced dataset. The area under the ROC curve (AUC) was 99%, highlighting the Random Forest model as the most robust and effective among all the evaluated algorithms [14][15].

Diabetes is a serious global health problem, affecting over 422 million people and continuing to grow each year, as mentioned by the authors [16][17]. Their study focused on how to predict diabetes more accurately using two artificial intelligence algorithms: Random Forest and XGBoost, applied to a public diabetes dataset available on Kaggle. The aim was to improve prediction accuracy by better selecting the most important features for analysis.

The dataset used contained 768 records with nine medical features, such as number of pregnancies, glucose levels, blood pressure, and body mass index (BMI), along with a label indicating whether the person had diabetes or not. To ensure reliable results, a data cleaning and preprocessing step was performed first. Then, two advanced techniques were applied to select the most relevant features: the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).The results showed that, initially, without feature selection, the Random Forest algorithm achieved an AUC accuracy of 0.8120, and XGBoost reached 0.7666. After applying PSO, accuracy improved to 0.8582 for Random Forest and 0.8250 for XGBoost. When using the Genetic Algorithm, the results were even better: 0.8612 and 0.8351, respectively. That is, accuracy improved by up to 8.9%, showing that GA was more effective than PSO. This study demonstrates that adequately selecting the most relevant features can significantly enhance the prediction accuracy of machine learning models. This is particularly useful for developing tools that help detect diabetes early, enabling faster and more accurate diagnoses in clinics and hospitals [18][19].

On the other hand, in [20], the authors addressed the early identification of individuals at risk of developing Type 2 diabetes mellitus (T2DM). The main objective was to analyze possible risk factors through the application of supervised learning algorithms, specifically Decision Trees (DT) and Random Forests (RF). The study was based on data from the

Kharameh cohort, part of the Fars study, which includes a sample of 10,663 individuals aged between 40 and 70 years. Multiple clinical and demographic variables associated with the risk of T2DM were evaluated using data mining techniques. Standard metrics such as accuracy, sensitivity, specificity, and area under the ROC curve (AUC) were used to measure the performance of the models. Statistical analysis and model implementation were conducted using the R programming language. The Decision Tree (DT) model highlighted factors such as age, triglycerides, blood pressure, and BMI as most associated with type 2 diabetes. However, the Random Forest (RF) model proved to be more accurate, also identifying fasting glucose, cholesterol, and creatinine as key variables. With an accuracy of 73.5

In light of another hypothesis proposed by the author [21], which suggests that Type 2 diabetes (T2DM) may be influenced by environmental factors such as air pollution, noise, and neighborhood socioeconomic conditions, an analysis was carried out using advanced machine learning techniques: penalized LASSO regression, Random Forest (RF), and Artificial Neural Networks (ANN). Data from 14,829 participants in the AMIGO Cohort Study were analyzed, including 85 exposome variables obtained based on the location of their residences. The results showed that living in areas with lower housing values, a higher proportion of non-Western immigrants, and higher surface temperatures is associated with an increased risk of T2DM. The selected factors varied between models: some known factors such as air pollutants only appeared in univariate analyses, while others such as the presence of green spaces emerged in multivariate models like RF. Finally, the LASSO model showed better predictive performance than RF and ANN, according to the logLoss prediction error. In conclusion, social and environmental determinants of the residential environment are strongly related to the prevalence of T2DM, although their impact may vary depending on the analytical approach used [22][23].

Another study addresses the need to implement more sophisticated diagnostic tools to treat Type 2 diabetes mellitus (T2DM), a chronic metabolic disease characterized by alterations in glucose, lipid, and protein metabolism due to insufficient insulin production or ineffective insulin action. This research explores the application of advanced machine learning models, specifically the joint approach known as Stacked Multi-Kernel Support Vector Machine with Random Forest (SMKSVM-RF). This model integrates the pattern recognition capabilities of multi-kernel Support Vector Machines (SVM) and the robustness of Random Forests (RF), which combine multiple decision trees to increase reliability in prediction. By stacking both models, the goal is to leverage their complementary strengths to improve performance in classification and regression tasks in clinical settings. The results show that the SMKSVM-RF model achieved an accuracy of 73.37 per cent in the confusion matrix, a recall rate of 71.62 per cent, an individual precision of 70.13 per cent, and an F1 score of 71.34 per cent. These metrics highlight the potential of this hybrid approach as an effective tool in the early diagnosis of diabetes. Altogether, this work emphasizes the usefulness of complex machine learning methods such as SMKSVM-RF in improving current diagnostic systems and significantly contributing to enhanced healthcare for chronic diseases like T2DM [20][24].

Diabetes mellitus is a condition that presents multiple health implications for individuals, especially pregnant women, due to the high prevalence of its adverse effects [25]. This issue focuses on research regarding the identification of the disease in women over 25 years of age, whose main causes include overweight, obesity, or family history of diabetes. To address this issue, an ensemble learning approach was used, allowing for precise predictions through machine learning algorithms. In this context, range-based Random Forest models (RRF) were used, which considered various variables such as body weight and selected values. New features such as Sum and Rank were generated, and based on a defined threshold, diabetes predictions were made on the dataset [26].

It is essential to recognize that Type 1 diabetes (T1D) is classified as a chronic and irreversible disease. In this context, the use of predictive algorithmic models plays a crucial role in anticipating disease progression, as indicated by the author [27]. To this end, survival models based on the Random Forest technique were developed to model critical stages in the progression of T1D. The first model estimates the time required for an individual to evolve from single to multiple autoantibody positivity (AAb+), which represents a decisive stage in disease development. The second model predicts the transition from multiple AAb+ to the clinical manifestation of type 1 diabetes. The findings show that the survival models built with Random Forest outperform traditional approaches such as Cox regression. Furthermore, a comprehensive variable importance analysis was conducted, which allowed for the identification of new interactions among relevant biomarkers. Ultimately, a more accurate methodological framework is established for measuring and stratifying T1D risk, supporting earlier and more personalized preventive interventions [28][29].

In a study conducted by the author [30] and supported by data from the World Health Organization, diabetes is addressed as a silent non-communicable disease whose early detection poses a major challenge compared to other pathologies. To address this issue, the study proposes the implementation of machine learning techniques using the PIMA Indian Diabetes (PID) dataset. To address the class imbalance in the data, six resampling methods were applied: Random UnderSampling (RUS), Random OverSampling (UPS), SMOTE, ADASYN, SMOTE-Tomek, and SMOTEENN. These techniques allowed the evaluation of how different balancing schemes impact the performance of predictive models. For data exploration and classification, tree-based algorithms were used, specifically XGBoost and Random Forest, to identify relevant patterns associated with diabetes. The results showed that the XGBoost model achieved its best performance with the SMOTE-Tomek technique, while Random Forest performed best when combined with SMOTEENN [31][32].

Other complications associated with diabetes are often closely linked to inadequate insulin production or regulation, the hormone responsible for maintaining stable blood glucose levels, as noted by the author [33]. In his study, the objective was to develop a predictive algorithmic model capable of early identification of diabetes, thereby minimizing the risk of serious and even fatal complications. To achieve this, machine learning algorithms were implemented, including Random Forest and K-Nearest Neighbors (KNN), to improve diagnostic accuracy. The strategy was based on combining multiple

models to strengthen predictive capability and optimize early disease detection. A key element was the use of a large dataset composed of individuals diagnosed with and without diabetes, which allowed for the refinement of the predictive model. The results demonstrated that the integrated model can significantly improve the early identification of diabetic cases, favoring timely medical interventions and more effective disease management. Altogether, this research contributes to the advancement of health-oriented technology through prediction based on artificial intelligence techniques [34][35]. Fig. 1 presents the keywords through a word cloud generated from the different articles analyzed for the literature review, where the words diabetes and Random Forest stand out.

Fig. 1. Word cloud.

III. Methodology

A. Definition of the KDD Methodology

In the context of the present project, the KDD methodology has been applied—one of the most widely used approaches in the field of data mining and machine learning. This methodology enables the extraction of useful knowledge from large volumes of data originating from diverse sources, making it an adaptable and effective tool for projects related to data science [36][37]. The KDD process comprises a series of systematic stages, including selection, preprocessing, transformation, data mining, and evaluation of the obtained knowledge, as mentioned in Fig. 2. It is a multidisciplinary activity that integrates statistical techniques, machine learning algorithms, and exploratory data analysis, with the ultimate goal of discovering meaningful patterns and valuable knowledge in the available data [38][39], as illustrated in the architecture of Fig. 3.

1) Selection: The selected topic for the present research focuses on diabetes. In this stage of the KDD process, specifically in the selection phase, a CSV-format dataset from the Kaggle platform was chosen. This dataset was stored in a Google Drive folder to facilitate access from a cloud-based development environment, specifically Google Colab. Within this environment, the Python programming language and the Pandas module were used to efficiently load and explore the data [40].

It is important to highlight that the selected dataset includes records of individuals aged between 16 and 90 years, as well as various relevant features such as Polyuria, Polydipsia, Sudden weight loss, Weakness, among other clinical and demographic variables. This initial data selection was carried out to ensure

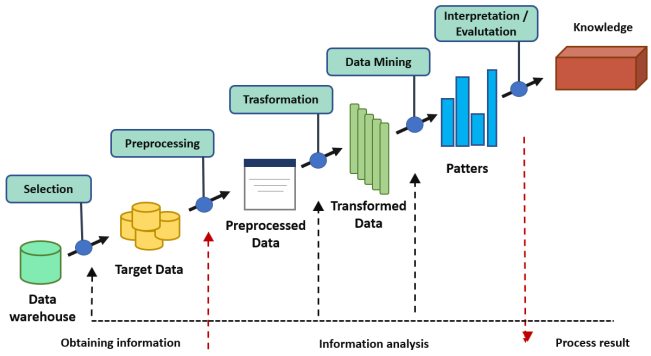

Fig. 2. KDD Methodology.

the relevance and quality of the information for subsequent processing and analysis during the following KDD phases, such as cleaning, transformation, data mining, and interpretation.

2) Preprocessing: For data processing, the imputation of missing values and the removal of outliers present in the fields observed in the dataset are proposed. However, in this section, it is noted that the presented data does not contain noise or specific missing values that could affect the model's performance, as shown in Fig. 4.

3) Transformation: In this section, the dataset is divided into two parts: 80% is used to train the model and the remaining 20% to test it, ensuring that the model learns from the majority of the data and is then evaluated with new data to measure its performance. Additionally, the class labels "Positive" and "Negative" are converted into numerical values (1 and 0) so that the algorithm can process them, while maintaining all features for both sets, thus allowing for proper model validation, as shown in Fig. 5. On the other hand, Fig. 6 presents the distribution of the numerical variables in the dataset, enabling the observation of the frequency with which different value ranges occur in each variable. This visualization helps identify patterns such as the presence of bias, dispersion, or potential outliers in the analyzed features, which is essential for understanding the nature of the data before applying any machine learning model. Moreover, the use of a consistent color palette facilitates visual comparison across the different variables.

4) Data mining: Fig. 7 presents two complementary visualizations that provide a deeper understanding of the behavior of the Random Forest model. First, an individual tree extracted from the forest (limited to a depth of 3) is shown, which allows us to observe how the model makes decisions based on the most relevant features. This visually illustrates the hierarchical structure of the partitions and the influence of each variable in the classification process. Next, a violin plot is presented, displaying the distribution of feature importance, generated from a simulation based on their estimated values. This visualization highlights not only the relative magnitude of each attribute in the model's decision-making but also their variability, offering both an artistic and statistical perspective on how each variable contributes to the overall performance of the algorithm, as shown in Fig. 8.

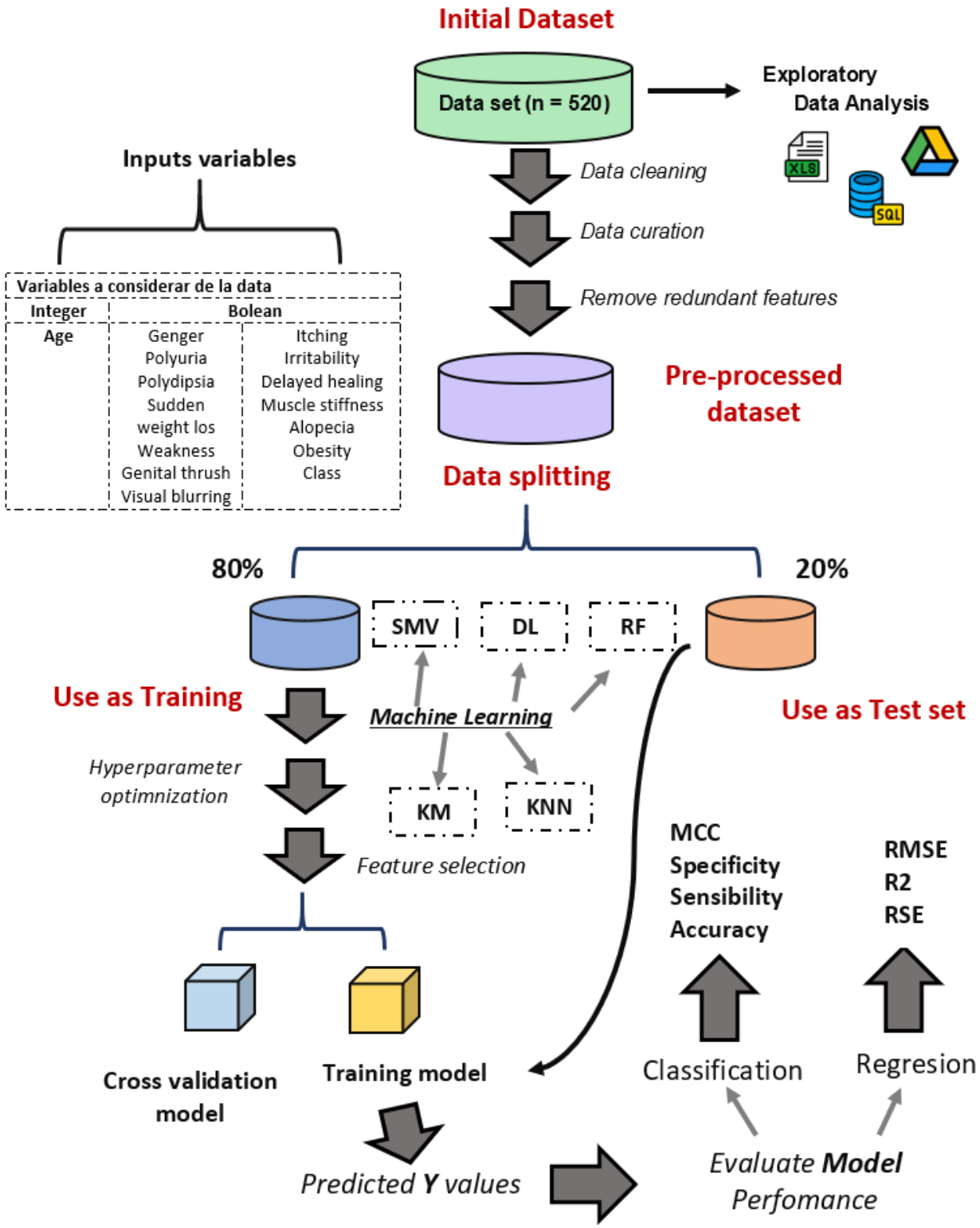

Fig. 3. Machine learning architecture.

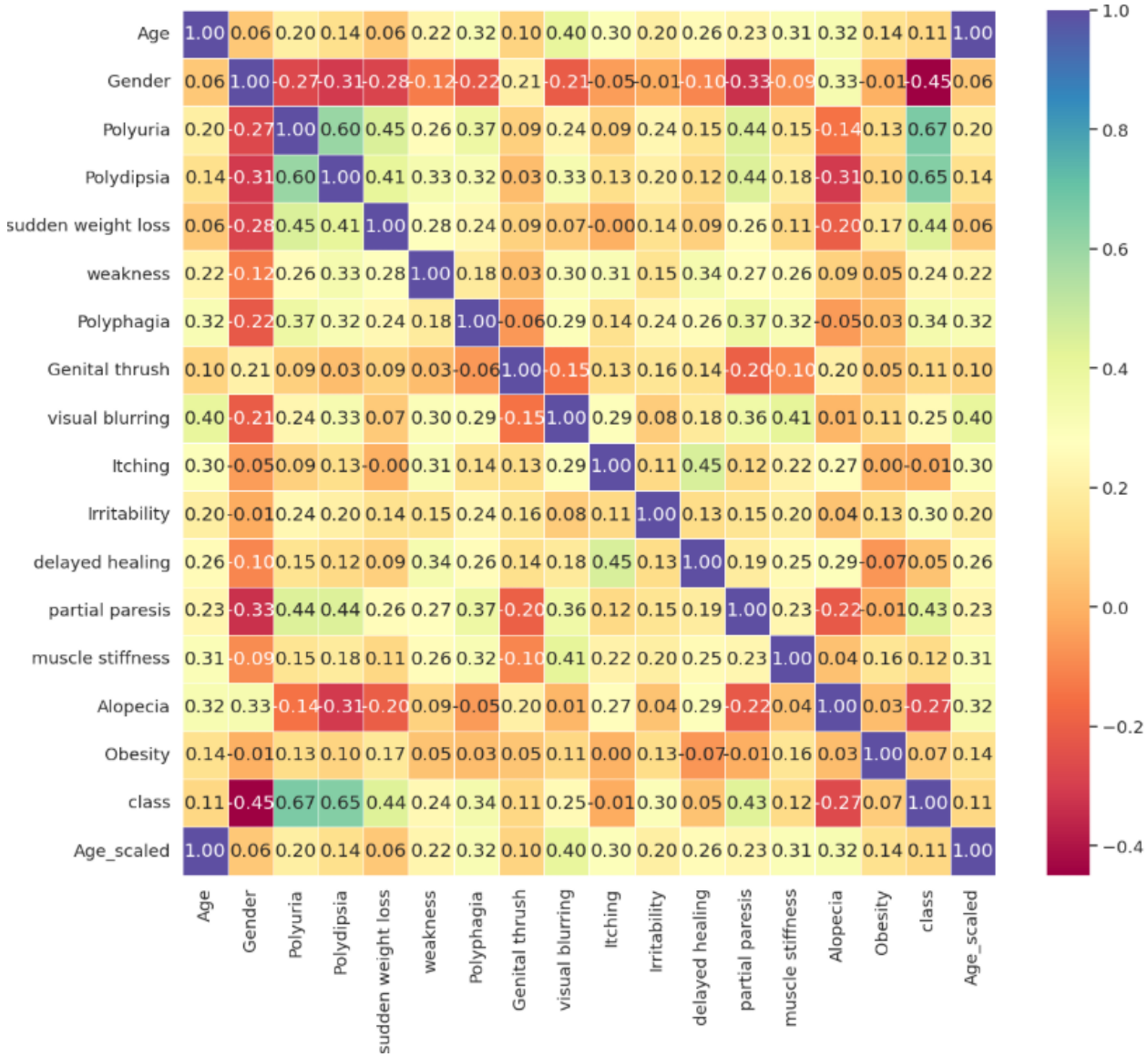

Fig. 4. Data preprocessing.

B. Random Forest Mathematical Formulas

1) Bootstrap sampling: Each tree in the forest is trained with a bootstrap sample from the original training set [see Eq. (1)].

$$D_b \sim \text{Bootstrap}(D), \quad \text{para } b = 1, 2, \ldots, B \tag{1}$$

where, D is the original training set, and D_b is the data set for the tree T_b, obtained by sampling with replacement.

Before training each tree in the Random Forest, a random sample with replacement is taken from the original dataset. This introduces diversity among the trees, helping to reduce overfitting and improve generalization.

2) Random feature selection: At each node of the tree, a subset of m features is randomly selected from the total of p available [see Eq. (2)].

$$\mathcal{M} \subseteq \{1, 2, \ldots, p\}, \quad |\mathcal{M}| = m \ll p \tag{2}$$

p: total number of variables (features) in the original set.

M: random subset of m features considered in each node.

m: number of variables randomly selected at each node of the tree.

Only these features are considered to find the optimal split at that node. Instead of using all variables to split a node. A small subset of variables is randomly selected. This increases variability between trees and improves the robustness of the model.

3) Division criteria:

a) Gini index (classification): Gini impurity measures how mixed the classes are at a node. A low value indicates

Fig. 5. Distribution of positive and negative classes training and test sets.

that the node contains mostly data from a single class [see Eq. (3)].

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{3}$$

where, p_i is the proportion of elements of class i in the set S, and C is the total number of classes.

b) Entropy (classification):

$$H(S) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{4}$$

H(S): Entropy of the data set S. C: Total number of possible classes.

p i : Proportion of elements of the class i in the set S.

$\log_2(p_i)$: Base-2 logarithm (used to measure information in bits). Entropy $H(S)$ measures the degree of impurity or disorder in a data set. $H(S)$ reaches its maximum value when classes are balanced (high uncertainty) and its minimum value when the node contains elements of only one class (low uncertainty). This measure is used to decide how to split a node in decision tree algorithms, including Random Forest trees [see Eq. (4)].

c) Variance (regression):

$$Var(S) = \frac{1}{|S|} \sum_{i \in S} (y_i - \bar{y})^2 \tag{5}$$

where, $\bar{y}$ is the mean of the outputs at the node.

It is used in regression problems to evaluate the dispersion of output values at a node. The goal of the tree is to minimize this variance by partitioning the nodes [see Eq.(5)].

4) Model prediction:

a) Classification (majority vote):

$$\hat{y} = \text{mode}\left(\{T_b(x)\}_{b=1}^{B}\right) \tag{6}$$

Prediction is the most common class among the predictions of B trees [see Eq.(6)].

b) Regression (average): Eq. (7) represents the regression process in a Random Forest model. This ensemble approach enhances prediction accuracy and reduces the risk of overfitting.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{7}$$

where,

- B is the total number of trees in the forest.
- $T_b(x)$ is the prediction of tree b for input x.
- $\hat{y}$ is the final prediction of the Random Forest.

5) Error Out-of-Bag (OOB): El error OOB se estima sin necesidad de un conjunto de validación externo [see Eq.(8)]

$$\text{Error}_{\text{OOB}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\hat{y}_{\text{OOB}}^{(i)} \neq y_i\right) \tag{8}$$

where, $\hat{y}_{\text{OOB}}^{(i)}$ is the prediction for x_i using only the trees that did not use it in their bootstrap, and $\mathbb{I}(\cdot)$ is the indicator function.

IV. Result

A. Evaluation of Result

In Fig. 9, a radar chart, illustrates the performance of the classification model evaluated through eight key metrics, highlighting specificity (78%), AUC (86%), and accuracy (85%) as the highest values, indicating a strong ability to correctly identify negative cases and good overall accuracy. However, metrics such as recall (75%) and MCC (72%) show lower values, suggesting reduced effectiveness in detecting positive cases, possibly due to the presence of false negatives. Overall, the model demonstrates balanced performance, although there is room for improvement in sensitivity and the overall correlation between predictions and true classes.

In Fig. 10, ROC compares the performance of the Random Forest model for both classes, showing separate curves for class 0 (area under the curve = 0.98) and class 1 (AUC = 0.98), indicating a high discriminative ability of the model to distinguish between the two classes. The dotted diagonal line represents a random classifier, and the fact that both ROC curves lie significantly above this line demonstrates performance better than random chance. The inclusion of a decorative color bar on the right adds a visual element without interfering with the central interpretation, highlighting that the model achieves a favorable balance between the true positive rate and false positive rate for both classes.

The Precision-Recall curve shows the performance of the Random Forest model in predicting the positive class, with

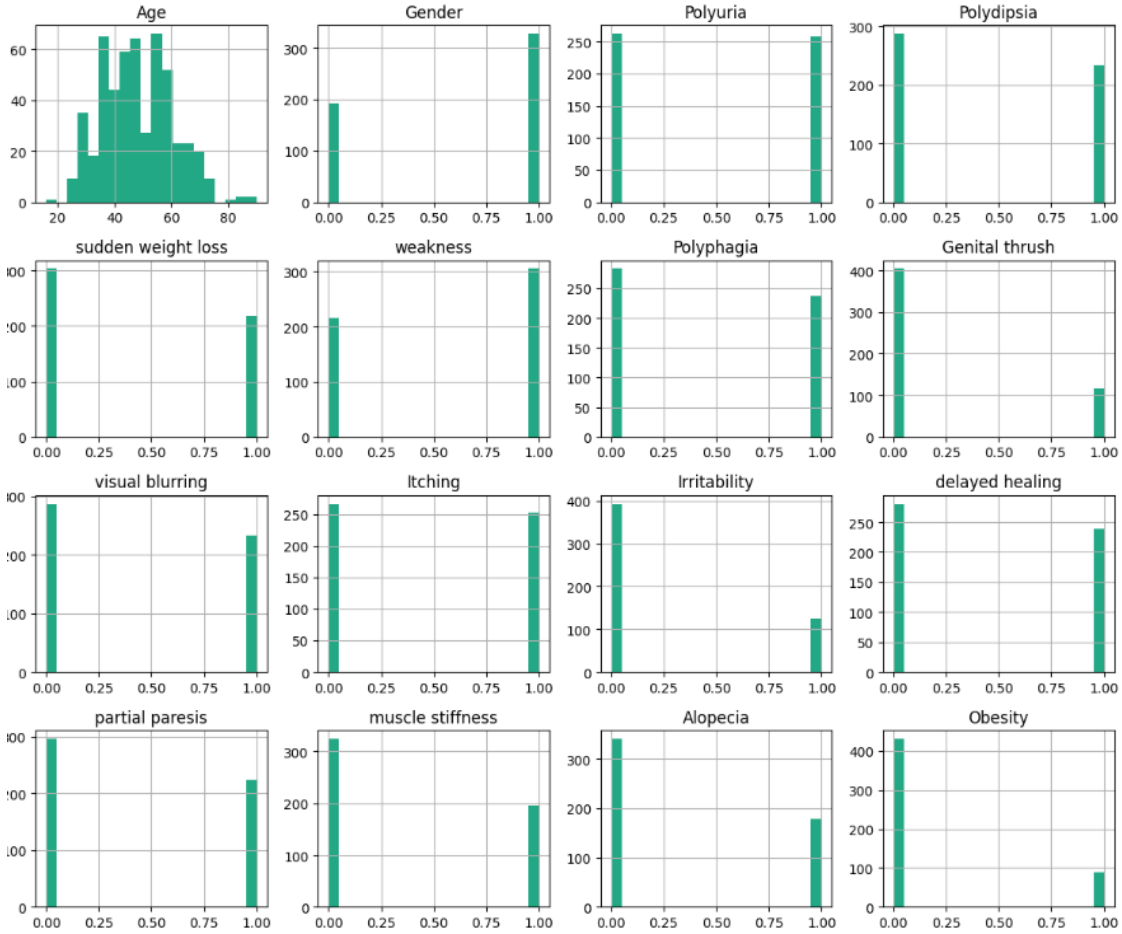

Fig. 6. Distribution of numerical variables.

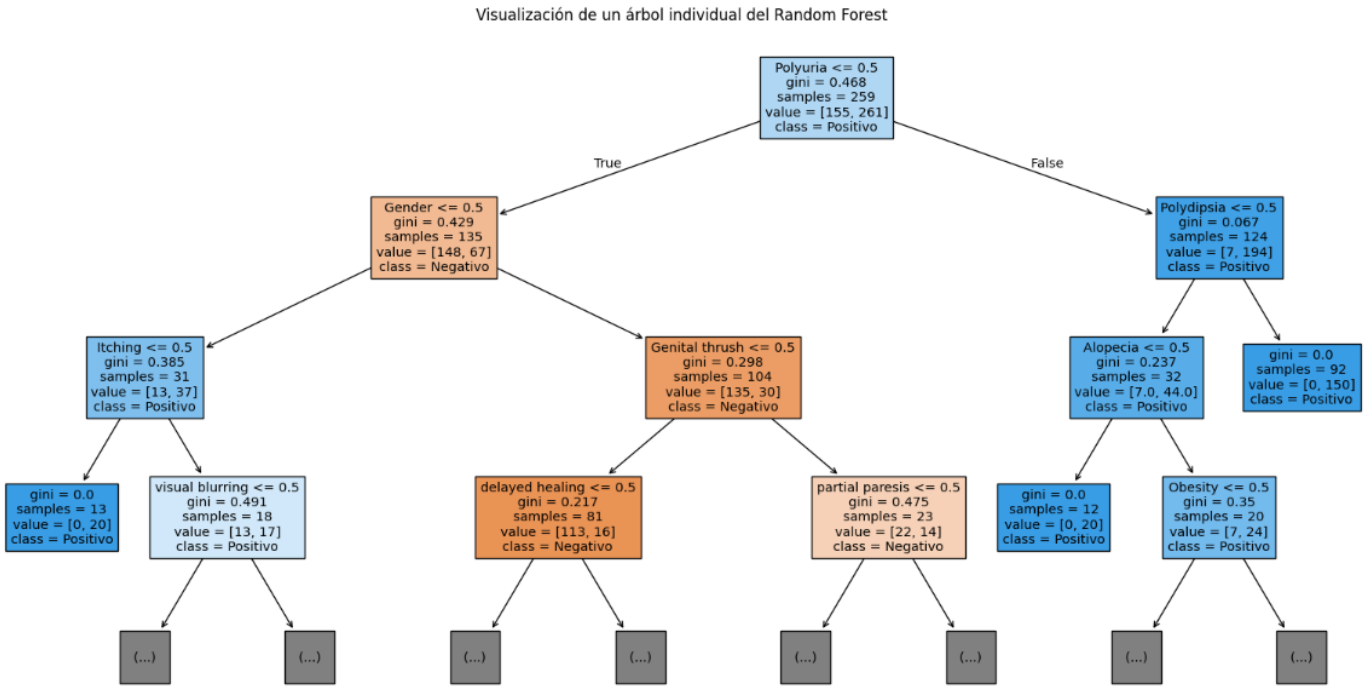

Fig. 7. Visualization of an individual tree from the random forest.

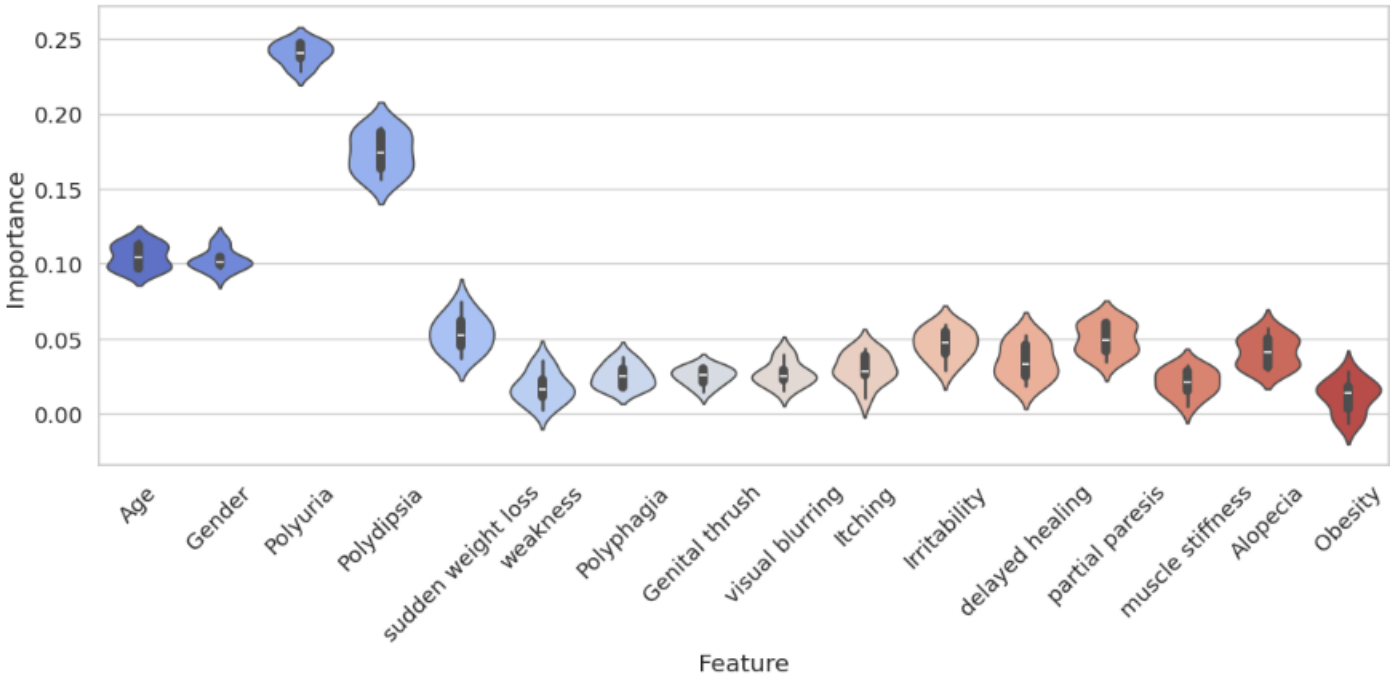

Fig. 8. Artistic visualization of importance - violin plot.

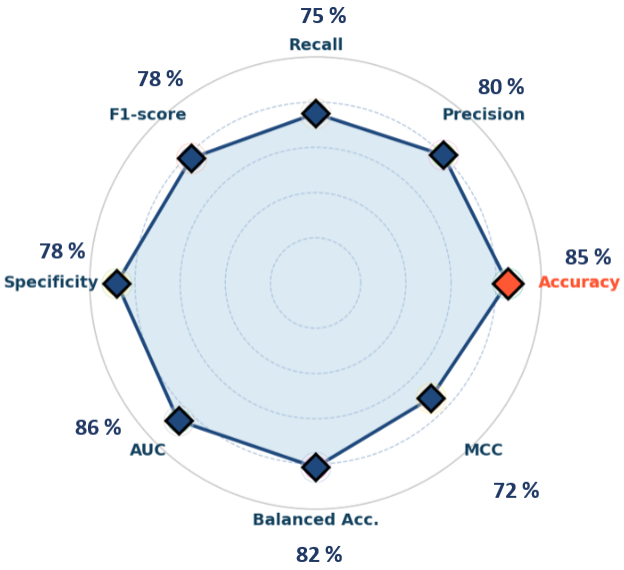

Fig. 9. Distribution of positive and negative classes training and test sets.

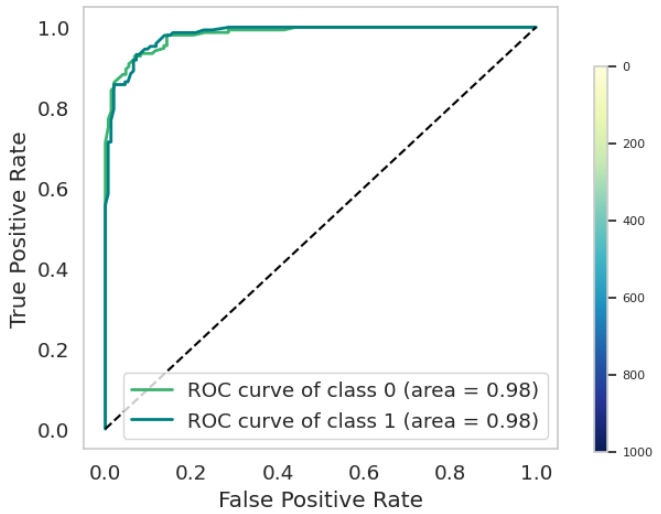

Fig. 10. ROC comparison of performance of Random Forest model for both classes.

an average precision (AP) of 0.98 for class 0 and 0.46 for a second simulated curve, representing a possible second class or a variation of the model. Both curves demonstrate a good balance between precision and recall, especially at higher recall levels, indicating that the model maintains good precision even when identifying most of the positives. This type of graph is particularly useful in contexts with imbalanced classes, as it highlights the model's ability to correctly detect positive cases without a high false positive rate. The presence of a decorative color bar does not affect the central interpretation, which confirms a robust performance of the model in terms of precision and recall, as shown in Fig. 11.

B. Comparison of Methodologies

This section presents a comparison of methodologies such as KDD, CRISP-DM, and SEMMA, with the aim of understanding why the KDD methodology was chosen for the project. It is important to note that certain criteria specified in the table were evaluated for this purpose. This summary can be examined in detail in Table I.

TABLE I. COMPARISON OF METHODOLOGIES: KDD, CRISP-DM AND SEMMA

Attribute	Methodology KDD	Methodology CRISP-DM	Methodology SEMMA
Structure and sequence	Includes selection, cleaning, transformation, and data extraction, as well as evaluation and application of knowledge, although its structure is not as rigorous.	It consists of six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.	Its five steps are sampling, exploration, modification, modeling, and evaluation.
Business orientation	Recognizes the importance of business objectives and seeks to learn in order to gain a competitive advantage.	Considers the objectives from the beginning and ensures useful results for decision-making.	Analyzes information considering the company's objectives and the use of the results.
Flexibility	Broad and less structured approach, offering a general framework for knowledge discovery.	Adaptable to various contexts and projects, extendable for commercial use.	Although it follows a predefined sequence, it is adaptable to different projects.
Interaction	Requires repetition but lacks an evident structure like CRISP-DM or SEMMA.	Iterative use of results, adapts to projects in constant evolution.	Step-by-step procedure adjustable throughout the process.

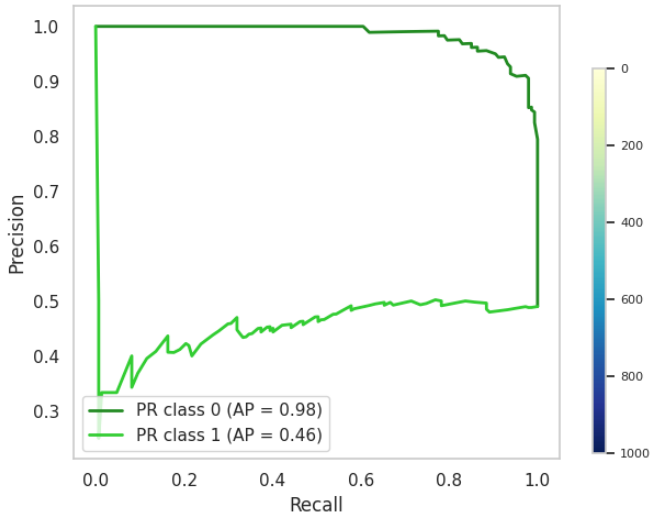

Fig. 11. Random forest precision-recall curve.

V. DISCUSSION

Initially, the research proposed by the author [12] focused on the analysis of systemic risk factors associated with diabetes mellitus, evaluating various machine learning models. The reported results included a 0% error rate during training, 100% precision, 76% accuracy, 53% sensitivity, and 80% specificity. In comparison, the model proposed in the present study, based on Random Forest, achieved superior metrics: 88% specificity, 82% precision, and 75% sensitivity, demonstrating better performance on critical variables.

Meanwhile, in [13], the authors directed their research towards diabetes classification through a comparative evaluation of multiple machine learning models, considering aspects such as maximum precision, performance with imbalanced data, and classification metrics. Although their results reached an AUC of 99%, surpassing the 86% AUC obtained in our model, it is important to highlight that both studies differ in objectives and methodological contexts, so direct comparison should be approached with caution.

Regarding the application of variable selection techniques, studies such as [16] and [17] employed evolutionary algorithms like PSO and GA to optimize the performance of models like Random Forest and XGBoost, achieving 86.12% precision and an AUC of 0.8612. In contrast, the present study obtained an AUC of 86% and accuracy of 85% without applying advanced selection techniques, evidencing the efficiency of the proposed model in terms of simplicity and performance. However, the 75% sensitivity obtained suggests opportunities for improvement in detecting true positives, making it advisable for future studies to consider integrating methods like GA to enhance this metric without compromising specificity.

Likewise, when compared to the study by [18], whose Random Forest model achieved a precision of 73.5% and an AUC of 79.1%, our approach showed superior performance (precision of 85% and AUC of 86%), which could be attributed to a better preprocessing strategy and variable selection. While Jawza identified variables such as fasting glucose, cholesterol, and creatinine as key determinants, our study used a comprehensive set of 9 clinical variables selected more systematically. Despite these differences, both works support the effectiveness of the Random Forest algorithm in predicting the risk of Type 2 diabetes.

On the other hand, the study by [21] addressed Type 2 diabetes prediction from an environmental and social perspective, incorporating 85 exposome variables related to the residential environment. Although they also used Random Forest and attribute selection techniques, their LASSO model showed better performance according to the logLoss metric. In contrast, our model, based solely on clinical variables selected via PSO and GA, achieved a precision of 85% and an AUC of 86%, reflecting good performance with lower complexity. This comparison highlights how context (clinical vs. environmental) can significantly influence model choice and interpretation of risk factors.

Regarding the approach based on ranked random forests (RRF) proposed by [25], focused on women over 25 years old and generating new features such as "Sum" and "Range", our study opted for optimization of clinical variables through Random Forest tuned with PSO and GA algorithms. Despite the difference in techniques, our model achieved better overall performance (precision of 86%, AUC of 86%), demonstrating the effectiveness of appropriate variable selection over feature engineering in specific populations. Similarly, the study by [27] applied Random Forest in a different context, aimed at modeling the clinical progression of Type 1 diabetes (T1D) through survival analysis, outperforming traditional models like Cox regression. Although the approach and pathology differ, both studies confirm the potential of machine learning as

a clinical support tool, whether in predicting T1D progression or early detection of T2D.

On the other hand, investigations such as those by [31] and [32] addressed class imbalance in the PIMA Indian Diabetes dataset through resampling techniques like SMOTE-Tomek and SMOTEENN. In contrast, our study did not apply balancing methods, focusing on the direct optimization of the Random Forest model. Despite this, competitive results were obtained, with an accuracy of 85%, an AUC of 86%, and a specificity of 88%, suggesting that, in certain scenarios, a robust model can compensate for the lack of advanced balancing techniques while maintaining high predictive performance.

Finally, the author [33] developed an algorithmic model focused on early prevention of diabetes, achieving substantial improvements in early disease identification. However, when compared with our model, evaluated using Random Forest, superior metrics in precision, sensitivity, and accuracy stand out, demonstrating the robustness of our approach and its potential applicability in clinical settings where early detection is critical for timely intervention.

VI. Conclusion

Diabetes mellitus is a chronic metabolic disease that, if not detected and treated in time, can lead to severe complications such as kidney damage, retinopathies, cardiovascular diseases, and neuropathies. In view of this issue, the development of a predictive model based on supervised machine learning algorithms was proposed, specifically using the Random Forest algorithm. For the construction of the model, a public database extracted from the Kaggle web platform was used, containing relevant clinical and demographic variables. The model development was based on the KDD (Knowledge Discovery in Databases) methodology, encompassing the phases of data selection, preprocessing, transformation, data mining, and results evaluation. Each of these stages was key to ensuring the quality of the predictive model and its applicability in real scenarios.

The results obtained through model validation show significant performance metrics: a specificity of 78%, an AUC (Area Under the ROC Curve) of 86%, an accuracy of 85%, a recall of 75%, and a Matthews correlation coefficient (MCC) of 72%. These metrics reflect a good balance in the model's ability to correctly classify both diabetic and healthy individuals, with particular emphasis on specificity, indicating a low false positive rate. The predictive model developed using Random Forest proved to be effective and reliable for the early detection of diabetes. Its implementation could serve as support in the clinical setting for preventive decision-making and population monitoring, thus contributing to improving prognosis and quality of life for individuals at risk.

Finally, our research also aims to contribute to future studies related to the addressed topic. Its implementation can be a starting point to combine the obtained results with more advanced approaches that strengthen existing knowledge and enable the development of useful and applicable solutions for organizations focused on the treatment of diabetes mellitus. The analysis of large volumes of data, through approaches such as Big Data, can significantly enhance the interaction with the predictive model developed using the Random Forest algorithm. This model can be integrated with massive information storage systems, such as Data Warehouses, to improve accuracy in identifying relevant patterns associated with the disease. Likewise, the developed models could be exposed through APIs, allowing their consumption by applications targeted at users with this pathology, refining the model logic according to the different clinical scenarios observed. However, a limitation identified in the present study lies in the use of a specific dataset, which restricts the model's ability to be generalized and applied across different clinical contexts. Additionally, the accuracy of the predictive model may be affected by the quality and completeness of the data used. In this regard, future research should consider the integration of larger, more diverse, and up-to-date datasets in order to enhance the robustness, reliability, and applicability of the proposed approach.

Acknowledgment

Thanks to the University of Sciences and Humanities, which made this study a reality.

References

[1] A. Murillo-Zavala, P. P. Palacios-Palma, and J. M. Zavala-Yoza, "Perfil lipídico y su asociación con las enfermedades isquémicas del corazón." *MQRInvestigar*, vol. 7, no. 3, pp. 1191–1207, 2023, doi = 10.56048/MQR20225.7.3.2023.1191-1207.

[2] L. F. J. Cañarte and A. D. C. Jalca, "Cistatina c y microalbuminuria como pruebas diagnósticas para el daño precoz del riñón en pacientes con diabetes mellitus," *Revista Científica Arbitrada Multidisciplinaria PENTACIENCIAS*, vol. 5, pp. 358–369, 3 2023, doi = 10.59169/PENTACIENCIAS.V5I3.547.

[3] H. Jiang, S. Zhang, Y. Lin, L. Meng, J. Li, W. Wang, K. Yang, M. Jin, J. Wang, M. Tang, and K. Chen, "Roles of serum uric acid on the association between arsenic exposure and incident metabolic syndrome in an older chinese population," *Journal of Environmental Sciences (China)*, vol. 147, pp. 332–341, 1 2025,doi = 10.1016/J.JES.2023.12.005.

[4] J. Geng, D. Wei, L. Wang, Q. Xu, J. Wang, J. Shi, C. Ma, M. Zhao, W. Huo, T. Jing, C. Wang, and Z. Mao, "The association of isocarbophos and isofenphos with different types of glucose metabolism: The role of inflammatory cells," *Journal of Environmental Sciences (China)*, vol. 147, pp. 322–331, 1 2025, doi = 10.1016/J.JES.2023.11.004.

[5] J. Harreiter and M. Roden, "Diabetes mellitus: definition, classification, diagnosis, screening and prevention (update 2023)," *Wiener Klinische Wochenschrift*, vol. 135, pp. 7–17, 1 2023, doi = 10.1007/S00508-022-02122-Y.

[6] N. A. Elsayed, G. Aleppo, V. R. Aroda, R. R. Bannuru, F. M. Brown, D. Bruemmer, B. S. Collins, M. E. Hilliard, D. Isaacs, E. L. Johnson, S. Kahan, K. Khunti, J. Leon, S. K. Lyons, M. L. Perry, P. Prahalad, R. E. Pratley, J. J. Seley, R. C. Stanton, and R. A. Gabbay, "7. diabetes technology: Standards of care in diabetes—2023," *Diabetes Care*, vol. 46, pp. S111–S127, 1 2023, doi = 10.2337/DC23-S007.

[7] B. Granon and A. Leroy, "Depression and diabetes," *Correspondances en MHND*, vol. 27, pp. 178–181, 12 2023, doi = 10.2337/DIACARE.28.8.1904.

[8] A. Vambergue, "Diabetes and pregnancy," *Medecine de la Reproduction*, vol. 25, pp. 217–232, 7 2023, doi = 10.1684/MTE.2023.0971.

[9] H. Talebi, L. J. Peeters, A. Otto, and R. Tolosana-Delgado, "A truly spatial random forests algorithm for geoscience data analysis and modelling," *Mathematical Geosciences*, vol. 54, 1 2022, doi = 10.1007/S11004-021-09946-W.

[10] L. Yin, B. Li, P. Li, and R. Zhang, "Research on stock trend prediction method based on optimized random forest," *CAAI Transactions on Intelligence Technology*, vol. 8, pp. 274–284, 3 2023, doi = 10.1049/CIT2.12067.

[11] X. Zhang, H. Shen, T. Huang, Y. Wu, B. Guo, Z. Liu, H. Luo, J. Tang, H. Zhou, L. Wang, W. Xu, and G. Ou, "Improved random forest algorithms for increasing the accuracy of forest aboveground biomass estimation using sentinel-2 imagery," *Ecological Indicators*, vol. 159, 2 2024, doi = 10.1016/J.ECOLIND.2024.111752.

[12] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, vol. 18, no. 1/2, pp. 90–100, 2022, doi:10.1016/J.ACI.2018.12.004.

[13] R. Venkatesh, P. Gandhi, A. Choudhary, R. Kathare, J. Chhablani, V. Prabhu, S. Bavaskar, P. Hande, R. Shetty, N. G. Reddy, P. K. Rani, and N. K. Yadav, "Evaluation of systemic risk factors in patients with diabetes mellitus for detecting diabetic retinopathy with random forest classification model," *Diagnostics*, vol. 14, no. 16, p. 1765, 2024, doi:10.3390/diagnostics14161765.

[14] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami, "Diabetes disease detection and classification on indian demographic and health survey data using machine learning methods," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 17, no. 1, p. 102690, 2023, doi:10.1016/j.dsx.2022.102690.

[15] F. A. Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Materials Today: Proceedings*, vol. 80, pp. 3200–3203, 1 2023, doi = 10.1016/J.MATPR.2021.07.196.

[16] P. Theerthagiri, A. U. Ruby, and J. Vidya, "Diagnosis and classification of the diabetes using machine learning algorithms," *SN Computer Science*, vol. 4, 1 2023.

[17] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the onset of diabetes with machine learning methods," *Journal of Personalized Medicine*, vol. 13, no. 3, p. 406, 2023,doi: 10.3390/jpm13030406.

[18] D. N. Jawza, M. I. Mazdadi, A. Farmadi, T. H. Saragih, D. Kartini, and V. Abdullayev, "Enhancing diabetes prediction accuracy using random forest and xgboost with pso and ga-based feature selection," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, p. 295 – 306, 2025, doi: 10.35882/jeeemi.v7i2.626.

[19] N. F. Cleymans, M. Van De Casteele, J. Vandewalle, A. K. Desouter, F. K. Gorus, and K. Barbe, "Analyzing random forest's predictive capability for type 1 diabetes progression," *IEEE Open Journal of Instrumentation and Measurement*, vol. 4, 2025, doi:10.1109/OJIM.2025.3551837.

[20] O. Adigun, F. Okikiola, N. Yekini, and R. Babatunde, "Classification of diabetes types using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, pp. 152–161, 2022, doi = 10.14569/IJACSA.2022.0130918.

[21] D. C. E. Saputra, A. Ma'arif, and K. Sunat, "Optimizing predictive performance: Hyperparameter tuning in stacked multi-kernel support vector machine random forest models for diabetes identification," *Journal of Robotics and Control (JRC)*, vol. 4, no. 6, p. p 896–904, 2023, doi:10.18196/jrc.v4i6.20898.

[22] H. Ohanyan, L. Portengen, O. Kaplani, A. Huss, G. Hoek, J. W. Beulens, J. Lakerveld, and R. Vermeulen, "Associations between the urban exposome and type 2 diabetes: Results from penalised regression by least absolute shrinkage and selection operator and random forest models," *Environment International*, vol. 170, 2022, doi:10.1016/j.envint.2022.107592.

[23] U. R. Saxena, R. Bathla, S. Srivastava, R. Agarwal, and D. Rawat, *"Predicting diabetes using hybrid approach of KNN and random forest machine learning algorithms"*, 2025, doi:10.1201/9781003508595-7.

[24] Ömer Faruk AKMEŞE, "Diagnosing diabetes with machine learning techiques," *Hittite Journal of Science and Engineering*, vol. 9, pp. 9–18, 3 2022, doi = 10.17350/HJSE19030000250.

[25] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable ai techniques," *Healthcare Technology Letters*, vol. 10, pp. 1–10, 2 2023, doi = 10.1049/HTL2.12039.

[26] S. Amarnath and M. Selvamani, "Rank based random forest model for gestational diabetes mellitus prediction," *AIP Conference Proceedings*, vol. 3180, no. 1, 2024, doi:10.1063/5.0224646.

[27] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," *Multimedia Tools and Applications*, vol. 83, pp. 24 153–24 185, 3 2024, doi = 10.1007/S11042-023-16407-5.

[28] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and early prediction of diabetes using machine learning in north kashmir: A case study of district bandipora," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi = 10.1155/2022/2789760.

[29] F. O. Aghware, M. I. Akazue, M. D. Okpor, B. O. Malasowe, T. C. Aghaunor, E. V. Ugbotu, A. A. Ojugo, R. E. Ako, V. O. Geteloma, C. C. Odiakaose, A. O. Eboka, and S. I. Onyemenem, "Effects of data balancing in diabetes mellitus detection: A comparative xgboost and random forest learning approach," *NIPES - Journal of Science and Technology Research*, vol. 7, no. 1, p. 1 – 11, 2025, doi:10.37933/nipes/7.1.2025.1.

[30] R. Cheheltani, N. King, S. Lee, B. North, D. Kovarik, C. Evans-Molina, N. Leavitt, and S. Dutta, "Predicting misdiagnosed adult-onset type 1 diabetes using machine learning," *Diabetes Research and Clinical Practice*, vol. 191, 9 2022, doi = 10.1016/J.DIABRES.2022.110029.

[31] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, "Deteksi dini penyakit diabetes menggunakan machine learning dengan algoritma logistic regression," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 11, pp. 88–96, 5 2022, doi = 10.22146/JNTETI.V11I2.3586.

[32] X. Feng, Y. Cai, and R. Xin, "Optimizing diabetes classification with a machine learning-based framework," *BMC Bioinformatics*, vol. 24, 12 2023, doi = 10.1186/S12859-023-05467-X.

[33] M. R. Islam, S. Banik, K. N. Rahman, and M. M. Rahman, "A comparative approach to alleviating the prevalence of diabetes mellitus using machine learning," *Computer Methods and Programs in Biomedicine Update*, vol. 4, 1 2023, doi:10.1016/J.CMPBUP.2023.100113.

[34] A. Nicolucci, L. Romeo, M. Bernardini, M. Vespasiani, M. C. Rossi, M. Petrelli, A. Ceriello, P. D. Bartolo, E. Frontoni, and G. Vespasiani, "Prediction of complications of type 2 diabetes: A machine learning approach," *Diabetes Research and Clinical Practice*, vol. 190, 8 2022, doi = 10.1016/J.DIABRES.2022.110013.

[35] K. Sidana, "Prediction of diabetes using machine learning algorithms," *11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks, IEMECON 2023*, 2023, doi: 10.1109/IEMECON56962.2023.10092335.

[36] E. A. S. Peralta and E. A. V. Allazo, "Classifier model for personalizing exercises given to students using artificial neural networks," *Publicaciones de la Facultad de Educacion y Humanidades del Campus de Melilla*, vol. 53, pp. 89–106, 1 2023, doi = 10.30827/PUBLICACIONES.V53I2.26818.

[37] G. A. Romero, C. A. G. Prieto, M. A. D. Barriosnuevos, and N. A. R. Menjura, "Revisión y perspectivas para la construcción de bases de datos robustas con datos faltantes: caso aplicado a información financiera," *Tecnura*, vol. 27, pp. 12–37, 1 2023, doi = 10.14483/22487638.18268,.

[38] L. F. C. Rojas, E. E. Peña, and E. R. Cuero, "Análisis de

características que influyen en la deserción estudiantil en el contexto de una universidad latinoamericana," *Revista EIA*, vol. 20, 12 2023, doi = 10.24050/REIA.V20I40.1628.

[39] C. A. de Godoy Fonseca and I. F. Silveira, "Um framework baseado em aprendizado de máquina e dados de processos res judicata para análise e previsão de sanções penais referentes a crimes cibernéticos," *Revista Cientſica Multidisciplinar Núcleo do Conhecimento*, pp. 36–60, 2 2023.

[40] Kaggle, "*Diabetes Data Set*," Online. Available: https://www.kaggle.com/datasets/chengyaran/diabetes-data-upload, 2020, accessed: May 10, 2025.