

# Predictive machine learning applying cross industry standard process for data mining for the diagnosis of diabetes mellitus type 2

Victor Garcia-Rios<sup>1</sup>, Marieta Marres-Salhuana<sup>1</sup>, Fernando Sierra-Liñan<sup>2</sup>,  
Michael Cabanillas-Carbonell<sup>3</sup>

<sup>1</sup>Facultad de Ingeniería y Arquitectura, Universidad Autónoma del Perú, Lima, Perú

<sup>2</sup>Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú

<sup>3</sup>Vicerrectorado de Investigación, Universidad Privada Norbert Wiener, Lima, Perú

## Article Info

### Article history:

Received Jul 15, 2022

Revised Jan 20, 2023

Accepted Jan 30, 2023

### Keywords:

Diagnosis

Machine learning

Prediction

Random forest

Type 2 diabetes mellitus

## ABSTRACT

Currently, type 2 diabetes mellitus is one of the world's most prevalent diseases and has claimed millions of people's lives. The present research aims to know the impact of the use of machine learning in the diagnostic process of type 2 diabetes mellitus and to offer a tool that facilitates the diagnosis of the disease quickly and easily. Different machine learning models were designed and compared, being random forest was the algorithm that generated the model with the best performance (90.43% accuracy), which was integrated into a web platform, working with the PIMA dataset, which was validated by specialists from the Peruvian League for the Fight against Diabetes organization. The result was a decrease of (A) 88.28% in the information collection time, (B) 99.99% in the diagnosis time, (C) 44.42% in the diagnosis cost, and (D) 100% in the level of difficulty, concluding that the application of machine learning can significantly optimize the diagnostic process of type 2 diabetes mellitus.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Michael Cabanillas-Carbonell

Universidad Privada Norbert Wiener

Lima, Perú

Email: mcabanillas@ieee.org

## 1. INTRODUCTION

In recent years, diabetes mellitus has increased its prevalence on the world scene, information from the World Diabetes Atlas, which is periodically published by the International Diabetes Federation, shows that by the year 2045 diabetes mellitus is projected to increase by up to 143% on the African continent and 55% in South America [1]. In addition, the publication mentions that there are currently approximately 537 million people around the world who suffer from diabetes, of which 352 million are in the active phase, within a range of 20 to 64 years of age. The problematic situation of the study is based on the increasing prevalence of type 2 diabetes mellitus in the world scenario, at present, due to COVID-19, people suffering from it are the most likely to develop a critical picture of this disease, which is why early diagnosis is so important [2]. Type 2 diabetes is caused by varying degrees of insulin resistance, altered insulin secretion, increased glucose production, and various genetic metabolic defects in insulin action [3]. According to the pan american health organization (PAHO) [4], it has been identified that approximately 90% to 95% of all cases suffer from type 2 diabetes [5], i.e., of the three main types of diabetes mellitus, type 2 is the most common, significantly affecting the health status of those who suffer from this disease, causing various disabilities, even death, and representing a high economic and social cost.

In recent years, humanity has been immersed in health problems, especially in resource-poor environments, and the situation is aggravated by the limited capacity of the health system to provide health care [6]. This is why it is important to develop and implement technologies such as machine learning models that serve as tools for doctors and patients, through preventive medicine that can help diagnose patients early and provide them with health advice. In this sense, implementing the use of machine learning facilitates the identification of patterns and predictions through analysis techniques such as statistics and empirical data [7]–[9]. Particularly in Perú, different population-based health studies indicate that, in recent years, as in the rest of the world, the rate of cases of type 2 diabetes mellitus has been increasing considerably, as evidenced in the latest study published by the National Institute of Statistics and Informatics (INEI) [10] on this disease in 2020. The research aims to determine to what extent the use of predictive machine learning, applying the cross-industry standard process for data mining (CRISP-DM) methodology, impacts the diagnostic process of type 2 diabetes mellitus.

## 2. LITERATURE REVIEW

According to the research [11], the use of mobile applications for the prediction and monitoring of diabetes proves to be more efficient and practical than conventional methods, providing users with early identification and early diagnosis of the disease to prevent the development of future complications [12]. The development of the application required the collection of various data from users to obtain more accurate results [13]. Finally, users were satisfied with the compliance and design of the application. The research [13] was developed using machine learning algorithms, employing 2 of these to evaluate the effectiveness and overall accuracy of the prediction, in order to identify and diagnose the type of diabetes, minimize the risk of death and generate the improvement of the patient's health. Finally, it was shown in this work that the results were highly accurate and it was possible to predict diabetes with less time in the process. The article [14] states that diabetes is the most common and dangerous disease that can lead to additional problems such as heart attacks, strokes, blindness, nerve damage, kidney failure, and blood vessel disease. Predictive analytics in healthcare is mainly used to determine patients who have early stages of diabetes, asthma, and heart disease, among other critical lifelong diseases. The proposed method using K-Means and random forest provides greater accuracy in predicting type 2 diabetes, obtaining the most effective model the random forest model, with which 80% accuracy was obtained, having as dataset the Pima Indian Diabetes. The article indicates that the fuzzy system and the deep learning method could also be used to improve the proposed method. In the investigation [15], dataset processing is performed to detect and diagnose diabetes mellitus, focusing on the use of machine learning algorithms. The investigation of a hybrid model where a coyote optimization algorithm (COA) and least squares support vector machine (LS-SVM) was proposed, where an average accuracy of 98.811% was obtained outperforming the other algorithms. The use of methods with a metaheuristic approach helps to solve complex optimization problems to determine fuzzy parameters as shown in research [16].

The articles [17], [18] optimization for machine learning feature selection is performed by employing logistic regression, decision trees, and random forest machine learning algorithms. The results showed additional advantages to the machine learning models when tested on a real data set, allowing optimization for machine learning feature selection, which serves greater utility in problems involving real data set cases. The investigation [19] preprocessing was performed by training the model to maintain the same characteristics of the images intended for the main processing. A convolutional neural network was used for the processing of 88,700 retinal fundus images, thus classifying the level/state of the disease, and determining whether a person suffers from diabetes mellitus. A sensitivity of 81.12%, a specificity of 89.16%, and an accuracy of 84.16% were achieved. The article [20] focuses on tuberculosis disease; in this case, the project was developed in Peru and was called eRx. Medical professionals or trained technical staff used smartphones as devices to capture images of chest X-rays of patients in health centers, after capture, these X-ray images were transmitted through the application developed based on artificial intelligence methods, convolutional neural networks algorithm was applied, demonstrated through the appreciation of medical specialists that artificial intelligence tools can optimize the process of diagnosis of tuberculosis.

The scientific article [21] it is mentioned that the process of disease diagnosis is complex and needs to be treated with automated techniques such as artificial intelligence to support decision-making and generate greater certainty. The authors performed a comparison of artificial intelligence techniques, two algorithms were evaluated, neural networks and Bayesian networks, the latter being the one that generated the best performance. Finally, they used certain and also applied the Osgood scale, which provides a positive evaluation of the results of the research conducted, showing the possibility of using artificial intelligence techniques for the diagnosis of diseases from data. This scientific article [22] from Western India mentioned that the incidence of lateral malleolar fractures (LMF) is increasing, and current classification systems have poor prognostic value in assessing the stability of these fractures, so they implemented artificial intelligence into the existing diagnostic procedure.

A semi-automated artificial intelligence diagnostic approach was developed that can significantly improve diagnostic understanding to aid in a more accurate diagnosis of LMF, allowing for resource savings.

### 3. METHODOLOGY

CRISP-DM refers to the proven method for guiding or providing a framework for data mining projects. It is a methodology that includes descriptions of the normal phases of a project description of the tasks required in each phase, and an explanation of the relationships between the tasks. CRISP-DM as a model provides an overview of the life cycle of a data mining project, has six flexible phases, and can be easily customized [23]. In the present study, all phases of the methodology will be applied, to the results of the research paper.

#### 3.1. Understanding the business

Determine business objectives, for this, the League is required to diagnose type 2 diabetes in the people who are screened for the disease in the campaigns they carry out, as well as:

- Reliably detect diabetes in people and act in a timely manner.
- Promote healthy lifestyles.
- Identify the risk of having diabetes.
- Conduct dialectological research.
- Provide counseling to the person with diabetes.
- Prevent the death of people with diabetes.
- Offer facilities to low-income people for the diagnosis and follow-up of diabetes.

Assessment of the current situation, to carry out the data mining project we had a database on GitHub that provided patient information on certain risk factors related to diabetes mellitus type 2, the factors were consulted and validated by volunteers, a medical staff of the organization of the Peruvian League for the Fight against diabetes so it gives greater input to this study.

Determine the objectives of data mining.

- Comparison of predictive models for the classification of potential patients with type 2 diabetes.
- Determination of the best predictive model for the classification of potential type 2 diabetes patients based on machine learning measurement metrics.
- Determination of patient profile characteristics with respect to presenting features and symptoms.
- Identification of relevant information using statistical graphs.
- Diagnostic prediction of patients based on certain measurements included in the dataset to determine whether or not they have diabetes.

#### 3.2. Understanding the data

Collect initial data, the dataset used for the development of the project has been extracted from the globally recognized intensive care unit (ICU) repository, this dataset is known as "PIMA", and contains information on people in India with positive and negative diagnoses of diabetes. Description of data from a set of 768 observations (row) and 10 variables (columns), as detailed in Table 1. The target variable is "outcome" and it was possible to visualize that there are 500 negative cases and 268 positive cases. Verify data quality, using the R programming RStudio as a tool to load the data set and analyze it, missing data representing up to 10% of some variables were found, as shown in Figure 1.

#### 3.3. Data preparation

Data selection was performed for the study by selecting all the columns that made up the database, where each variable was validated as a risk factor for type 2 diabetes mellitus by expert endocrinologists from the Peruvian League against diabetes, with the exception of column "n" because it refers to the patient's identification number, which is not an important variable for the analysis, as well as the column "insulin every 2 hours" because it is a data that required the intervention of laboratory analysis, which did not allow a non-invasive collection of the symptoms. Clear data, after identifying the missing data, the R package "DMwR" was used to perform the central tendency imputation technique, which allowed us not to lose observations from the data set, see Figure 2. Structuring the data, for the preparation of the data used by the models, coding, and data balancing techniques were applied.

The coding technique required the R package "car", which was applied to the variable "outcome", reassigning the values it had for 0 (negative) and 1 (positive) in order to numerically represent each prediction target and improve the analysis of the data set. The data balancing technique required the R package "ROSE" and was applied because the data have a higher number of records of persons with negative diagnoses, which could represent a problem of unbalanced classification when generating the models. Specifically, the

oversampling technique, also known as SMOTE, shown in Figure 3, was applied, which generates fairly accurate records from existing minority records [24]. The original data had 500 negative and 268 positive records, but after the over-sampling, it had 394 negative and 374 positive records.

Table 1. Dataset variables

Field	Type	Description
N	int	It is the identifier of each patient.
Number of pregnancies	int	A number of times the study person has been pregnant.
Plasma glucose concentration	int	It indicates the amount of glucose in the blood. When a person has eaten food, the normal values are less than 140 mg/dl, and when the results are between 140 to 190 they are indicative of diabetes.
Diastolic blood pressure	int	The amount of pressure in your arteries between heartbeats.
Skin fold thickness	int	It is a frequently used procedure, in combination with the body mass index (BMI), to estimate body fat. Measuring skinfolds allows for assessing the fat deposits in the human body. According to medicine the normal thickness: ♂ 12 mm; ♀ 23 mm.
Insulin every 2 hours	int	It is an insulin test that consists of testing before administering glucose and 2 hours later. The reason why these tests are performed is to see your glucose response curve.
Body mass index (BMI)	int	It is a method used to estimate the amount of body fat that a person has, and therefore determine if the weight is within the normal range, or on the contrary, if the person is overweight or thin.
Diabetes pedigree function		A function that rates the likelihood of diabetes based on family history.
Age	int	Age of patients in years
Result	categorical	If positive or negative for the diagnosis of diabetes.

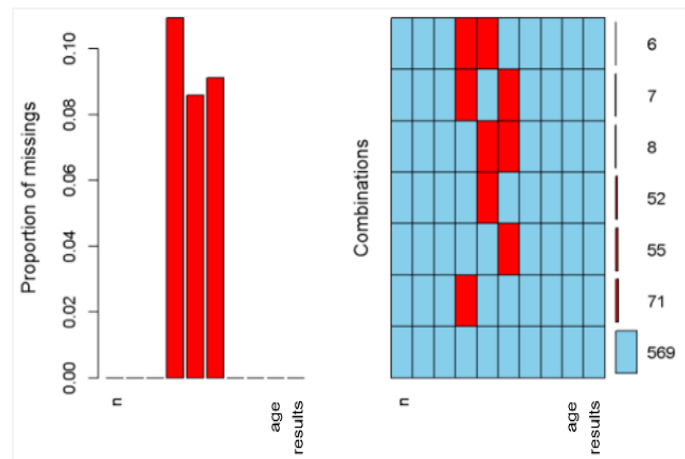


Figure 1. Datos missing

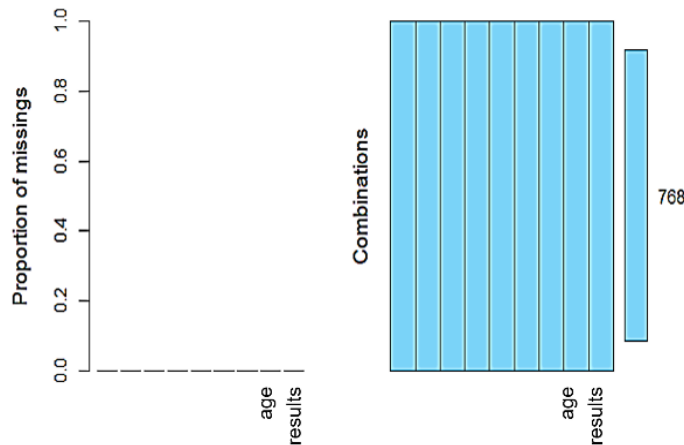


Figure 2. Input data

```

library(ROSE)
set.seed(123)
resultado = datosnew2[,9]
diabetes_1 = cbind(datosnew2,resultado)

muestra = ovun.sample(resultado ~., data = diabetes_1, method="both")$data
table = rbind(table(resultado), table(muestra$resultado))
rownames(table) <- c("data_original", "data_muestreada")
table

```

Figure 3. SMOTE application

### 3.4. Modeling

Three models were generated and compared in order to work with the one that obtained the best performance and was best adapted to the research objective: support vector machine (SVM) [25], artificial neural network (ANN), and [26] random forest [27]. Generate the test plan, before proceeding to the generation of the models, it was necessary to perform the design verification of the model results. The "caret" package was used to generate the partitions, shows in Figure 4 training data are composed of 70% of the total data, which were used to train the model and were obtained by random sampling. The test data made up of 30% of the total data were used to test the results of the model and were obtained by completely random sampling.

#### 3.4.1. Model building

Model 1, the support vector machine was generated by incorporating the R package "e1071". Before building the model, with the training data frame, a test or tuning script was run to obtain the best values for the gamma and cost parameters. The model was built with the "svm()" function of the "e1071" package, indicating within the function as the first parameter the variable to be predicted as a function of the other variables, as the second parameter the training data frame, as third parameter the gamma value and as fourth parameter the cost. From the sample data frame, the resulting classification variable was converted into a factor and defined as seed 123 so that the results would not be altered, shows in Figure 5.

Model 2, ANN as a first step, the "nnet" package was called in order to make use of the function that allowed us to generate the ANN model, a seed "123" was also defined so as not to alter the results obtained, the result classification variable is converted to a factor, then within the parameters of the "nnet" function, it was necessary to define the size of the network, i.e. the number of neurons with which it will work, the maximum allowed a number of weights (MaxNWts) were also defined, finally, the switch for the trace optimization was defined as FALSE, the modification can be seen in Figure 6. The Peruvian League against diabetes is an organization made up of health professionals and volunteers who seek to promote a healthy lifestyle in the Peruvian population and raise awareness of the prevention and early identification of diabetes in Peru.

```

set.seed(123)
particion = createDataPartition(datosnew2$resultado, p = 0.7, list = FALSE, times = 1)
train = datosnew2[particion,]
test = datosnew2[-particion,]

```

Figure 4. Data partition script

```

set.seed(123)
modelo.svm <- svm(resultado ~ ., data = datosnew2[training.sv, ], cost=1, gamma=1)
pred_sv <- predict(modelo.svm, datosnew2[-training.sv,], type = "class")

```

Figure 5. SVM model script

```

set.seed(123)
modelo.ann = nnet(resultado ~., data=train, size=15, MaxNWts=84581, trace=FALSE, method="class")
predicciones_red = predict(modelo.ann, test, type = "class")
predicciones_red = as.factor(predicciones_red)

```

Figure 6. ANN model script

Model 3, random forest, the "randomForest" package was used to generate this model. As shown in Figure 7, starting from the sample data frame, the resulting classification variable is converted into a factor, to define the seed "123" and thus does not alter the results. Then, the model was built with the function "randomForest()", indicating as the first parameter the training data frame obviating the result class variable, in the second parameter only the class of the training data frame is taken, the third parameter the number of trees that are most useful to us and in the fourth parameter it is indicated if all the outputs of the modeling are retained.

```
set.seed(123)
model.rf <- randomForest(x = datosnew2[training.rf, 1:8],
                        y = datosnew2[training.rf, 9],
                        ntree = 600,
                        keep.forest = TRUE)
pred_rf <- predict(model.rf, datosnew2[-training.rf,], type = "class")
```

Figure 7. Random forest model script

### 3.5. Evaluate the model

In order to carry out the evaluation of the models, we took into account metrics that helped us to evaluate the performance of each one, which are: accuracy, sensitivity, specificity, and specificity. Table 2 shows the performance of each classification method: SVM model with 86.09% accuracy, which tells us the number of times the model is correct when applied to the data, as in its other metrics, obtained 90.18% effectiveness in identifying positive cases and 82.20% negative cases. ANN model, with an accuracy of 71.30%, a specificity of 66.95%, and a sensitivity of 75.89%, is the model with the lowest percentage of sensitivity and specificity compared to the others, which means that it has the lowest proportion of hits in the identification of positive and negative cases. Finally, the random forest model is the one that generated the best results in the tests, culminating in the study with 90.43% accuracy, 89.83% specificity, and 91.07% sensitivity in the three metrics with the highest percentage of all, making it the model with the best performance in the identification of positive cases.

Table 2. Comparison of metrics between models

Model	Accuracy %	Sensitivity %	Specificity %
SVM	86.09%	90.18%	82.20%
ANN	71.30%	75.89%	66.95%
RF	90.43%	91.07%	89.83%

### 3.6. Evaluate the model

Figure 8 explains how the integration of a model developed in R language with Laravel and VueJS was done, having as its purpose to show an intuitive form where the required data will be entered to be sent through an application programming interface (API), which is consumed by VueJS, The model, which has been developed in R and through the language and its libraries, an endpoint was enabled which is located on a Linux server from where the result is sent to the Laravel project and it is here where the registration is made to a database that was raised thanks to an amazon web services (AWS). Steps to integrate the model with laravel and VueJS, creation of the API in R, after generating the prediction model following the CRISP-DM methodology, it was saved as a reporting data source (RDS) file, which is a native R file type that allowed storing the model as an object to later load it into the API. The next step was the creation of the API, the R package called Plumber was used, this package allows exposing of the prediction model as a service. In the coding, the libraries needed to build the API were declared (Plumber, randomForest, and caret), and a Cross-origin resource sharing (CORS) filter function was defined, since, by default, the API endpoint prohibits "cross-domain" requests, for this, the function forwards the correct headers to the API and allows to receive the request. Then the parameters to receive from the API were defined with a brief description of each one, the method (Post) by which these values are received and the endpoint of the API (/diabetes). Finally, a function was defined that receives all the parameters of the prediction request, loads the RDS file with the prediction model, generates a data frame with the received values that is loaded to the model to perform the prediction and assembles the output to return the response in JSON format.

Hosting the API with amazon EC2 for this stage of the implementation, Amazon Web Services cloud computing services were used, starting with the creation of an EC2 instance with the characteristics shown in Table 3. Using the FileZilla application, a connection is made to the instance, where a folder was created to

upload the RDS file with the model and the API file. To get the API up on the instance, the latest version of the Docker Engine package was installed, which allowed coding a custom Docker image with the dependencies and configurations required for the API.

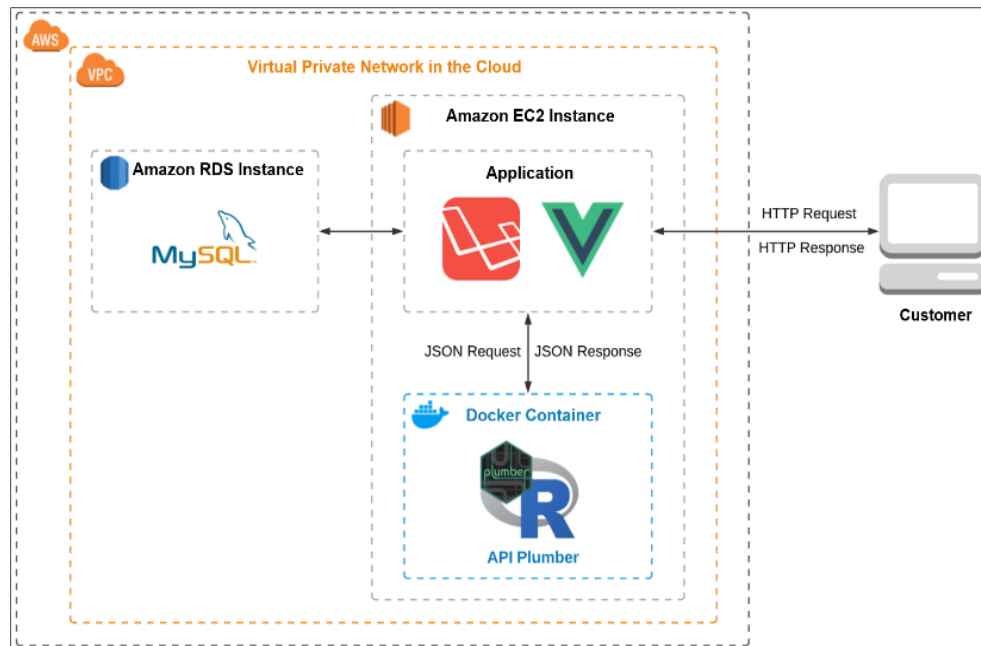


Figure 8. Software architecture model

Table 3. Aws ec2 instance characteristics

Characteristic	Detail
Image by Amazon Machine (AMI)	Amazon Linux 2 AMI
Type of Instance	T2.micro (1 vCPU de 2.5 GHz, 1GiB memory)
Storage	SSD de 8 GiB

Once the Dockerfile was coded, the image was run in a new container and by as-sociating port 8,000 of the container with port 8,000 of the local machine, the API endpoint was generated. To enable consumption of the API, in the AWS console, port 8,000 was enabled in the security group assigned to the EC2 instance. Finally, the API was tested in the browser with the swagger user interface (UI) tool integrated into the Plumber package, as shown in Figure 9. Creating MySQL Database with Amazon RDS service, in the Amazon RDS service interface, a database instance was created with the following characteristics, shows in Table 4. Finally, to manage the database created, the MySQL Workbench tool was used to create a connection to the database using the credentials generated by the Amazon service.

Web platform development, first created a working repository in Github for the project version control, we used the Laravel framework and VueJS, we decided to work with these languages because they allow a dynamic interaction and greater efficiency in the responses and performance in the web application and server communication. The web project has been coded in Visual Studio Code, using the Laravel directory structure, under the commonly used model-view-controller software design pattern with which we implement the UI, data, and control logic. Implementation of the prediction web platform, the platform for the medical personnel that carries out the campaigns of discarding and prevention of type 2 diabetes of the Peruvian League of the fight against diabetes where it presents two important sections, one where the prediction is made through the form and the second shows graphs on the predictions made. In the first instance, we have the prediction form where the fields that must be completed to obtain the result of the prediction are found, and the fields were validated using the "validate" package of VueJS. In addition to the fields with the necessary factors, we have considered fields such as ID, first and last names, and cell phone, which are data required to validate the identity of the patients, as well as the cell phone, gender, and district where the patient resides in order to make graphs that will allow an overview of the predicted cases and make decisions for the organization, shows in Figure 10.

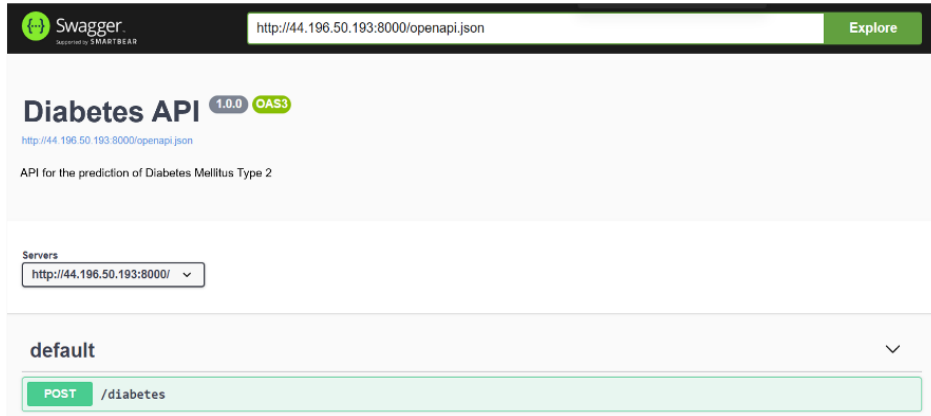


Figure 9. Testing the API in swagger UI

Table 4. MySQL instance characteristics in amazon rds

Characteristic	Detail
Database engine	MySQL v8.0.23
Instance class	db.t2.micro (1 vCPU, 1GiB memoria)
Storage	SSD de 20 GiB

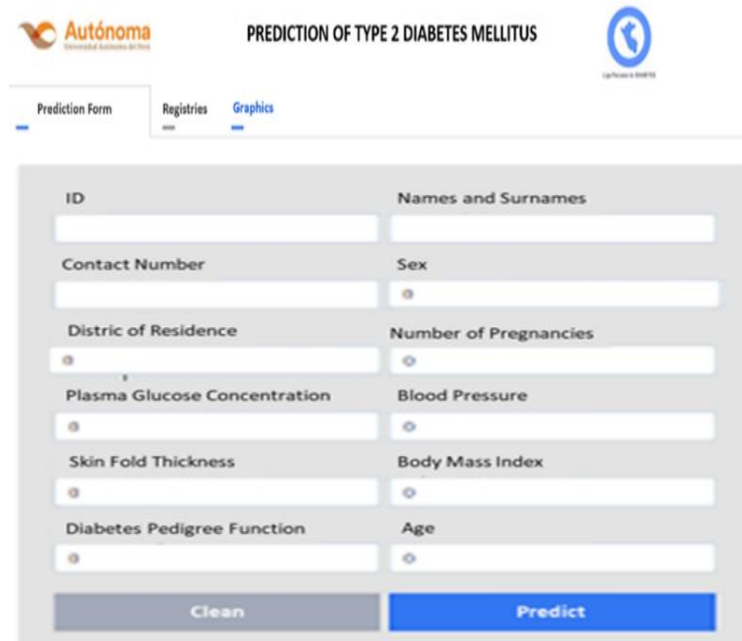


Figure 10. Prediction web form

The following records tab shows in real time the records entered, that is, the information of the patients and the data based on the determined symptoms. They are listed from the last record entered to the first, also, it has a search engine by identity document to facilitate the location of a specific record, as shown in Figure 11. A button is shown in Figure 12 with the option to view the detail of all the data related to the selected case and this view has the option to be exported to a PDF format to perform any necessary transaction such as sharing the information with the patient. The Figure 13 tab is shown dynamically and in real-time statistical graphs based on the records that have been carried out for the prediction, providing the organization with a reporting option where we can identify the total number of people registered, positive cases, negative cases, as well as cases by gender and district, allowing the identification of specific groups or sectors where the necessary measures can be focused and strategies for the prevention of type 2 diabetes can be applied.



#	Names and Surnames	ID	Results	Date of Registration	Detail
30	ETA MAGALLANES BENITES	34315869	POSITIVE	34315869	View
29	LILIA MUÑOZ ANGULO	65823751	POSITIVE	65823751	View
28	JORGE VALDIVIA FLORES	31542265	POSITIVE	31542265	View
27	ELIZABETH ARROYO RIVERA	21956583	POSITIVE	21956583	View
26	RUTH ASCUE POMAQUITA	29678545	POSITIVE	29678545	View
25	CLAUDIA TOLEDO QUIROGA	69858773	POSITIVE	69858773	View
24	MATILDA CACHAY SALAZAR	24565859	POSITIVE	24565859	View
23	MERCEDES YANQUI AQUJE	25699884	POSITIVE	25699884	View
22	RONALDO CANCIA ACHANTE	21564548	POSITIVE	21564548	View
21	MIRYAM MEJIA CHAUCO	54785698	POSITIVE	54785698	View

Figure 11. Record listing section

**Informe**

Names and Surnames	ID	District
ETA MAGALLANES BENITES	18200960	CHORRILLOS
Contact Number	Gender	Age
914923521	FEMENINO	25
Number of pregnancies	Plasma Glucose Concentration	Blood Pressure
3	193	70
Skin Fold Thickness	Body Mass Index	Diabetes Pedigree Function
31	34.90	0.24
Registration Date	Results	
17/11/2021	POSITIVE	

Close Export PDF

Figure 12. Detail view

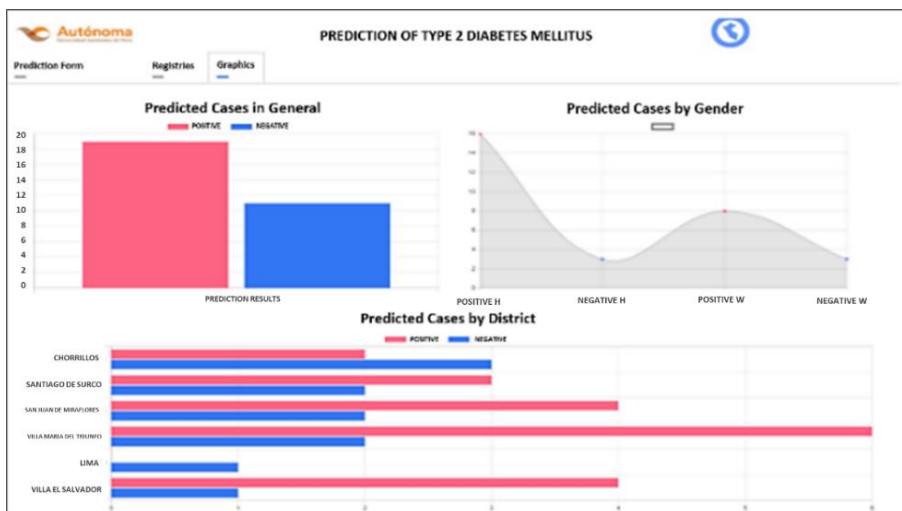


Figure 13. Web reports section

**4. RESULTS**

The objective of the research is to determine to what extent the use of predictive machine learning, applying the CRISP-DM methodology, impacts the diagnostic process of diabetes Mellitus type 2 in Lima-Perú, considering four indicators or also called key performance indicators (KPI's): i) time to collect information, ii) time of diagnosis, iii) cost of diagnosis, and iv) difficulty level. For this purpose, we opted for a type of applied research and a pre-experimental design, with a random sample of 30 patients from the organization Liga Peruana de Lucha Contra la Diabetes, and we measured the defined KPIs in two contexts, a pretest without the experimental condition and posttest with the experimental condition present. The results obtained for each KPI in both contexts were analyzed and interpreted at a descriptive and inferential level to finally respond to the initially stated objective.

**4.1. KPI1: data collection time**

In the inferential analysis, as shown in Figure 14 the results of the first indicator; Figure 14(a), the significance level of the pretest is 0.007 and of the posttest is 0.075. Where one of the cases (Pretest) is less than 0.05, then it is stated that the data do not have a normal distribution. In the descriptive analysis, according to the results shown, see Figure 14(b), for the information gathering time indicator (KPI1), in the pretest a mean value of 8701.83 was obtained and for the posttest, it was 1019.90. With these results, it can be observed that there was a decrease of 88.28%. Since KPI1 does not have a normal distribution, the Wilcoxon hypothesis test was performed as shown in Figure 15, which yielded a significance level equal to 0.000, which is less than 0.05, the threshold value to see if the research hypothesis is accepted. It is accepted that the use of predictive Machine Learning, applying CRISP-DM, has an impact on the time to collect information for the diagnosis of type 2 diabetes mellitus.

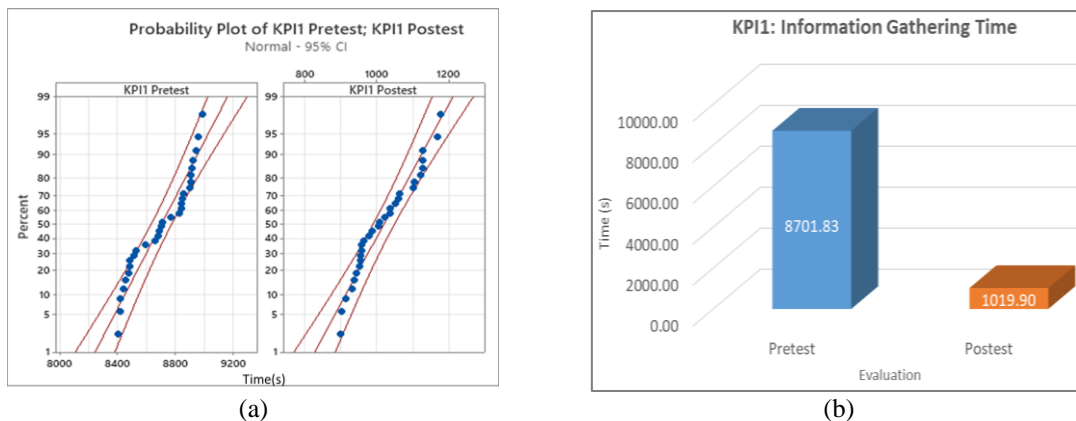


Figure 14. Results of the first indicator in (a) normality graph of KPI1 using Minitab 20.3 and (b) pretest and posttest histogram of KPI1 data collection time

Test statistics	
	Final Information Gathering Time
	Initial Information Gathering Time
Z	-4,782 <sup>b</sup>
Asymptotic sig. (bilateral)	,000

a. Wilcoxon signed-rank test  
b. It is based on positive ranges

Figure 15. Wilcoxon test report for KPI1 using statistical package for social sciences (SPSS)

**4.2. KPI2: diagnostic time**

In the inferential analysis, as shown in Figure 16 the results of the second indicator; Figure 16(a) the significance level of the pretest is 0.114 and of the posttest is 0.459. Where in both cases (Pretest and Posttest) their significance is greater than 0.05, then it is affirmed that the data have a normal distribution. In this case, since the significance value is less than 0.05, it is accepted that the use of predictive Machine Learning, applying CRISP-DM, has an impact on the time of diagnosis of type 2 diabetes mellitus. In the descriptive

analysis, according to the results shown in Figure 16(b), for the diagnostic time indicator (KPI2), in the pretest a mean value of 4424.2 was obtained and for the posttest, it was 0.353. With these results, it can be observed that there was a decrease of 99.99%. Since the KPI2 has a normal distribution, the T-student hypothesis test for related samples was performed as shown in Figure 17, which yielded a significance level equal to 0.000, which is less than 0.05, the threshold value to see if the research hypothesis is accepted. In this case, since the significance value is less than 0.05, it is accepted that the use of predictive Machine Learning, applying CRISP-DM, has an impact on the time of diagnosis of type 2 diabetes mellitus.

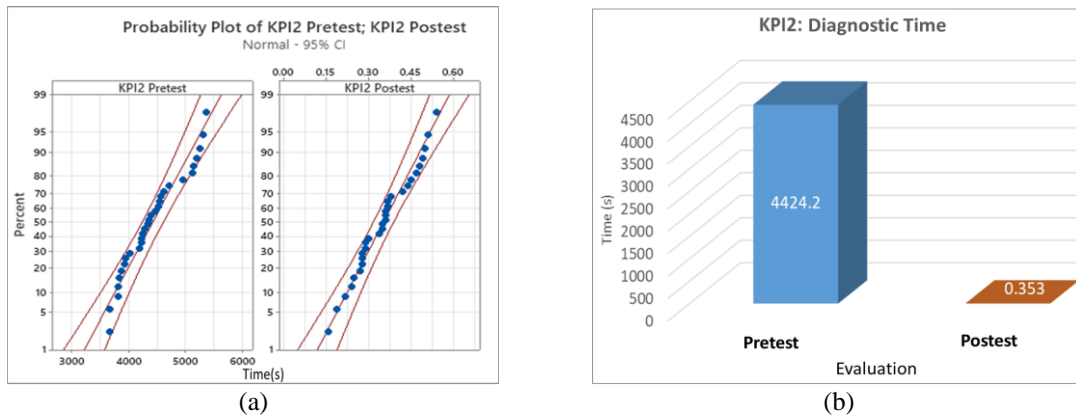


Figure 16. Results of the second indicator in (a) normality graph using Minitab 20.3 and (b) pretest and posttest histogram of KPI2 diagnostic time

Paired sample testing								
Differences matched								
	Mean	Standard Dev.	Mean Standard Error	95% confidence interval of the difference		t	df	Sig. (bilateral)
				Lower	Upper			
Pair 1: Time to Initial Diagnosis - Time to Final Diagnosis	4423.84700	519.29486	94.80984	4229.93911	4617.75489	46.660	29	,000

Figure 17. T-student T-test report for related samples of KPI2 using SPSS

**4.3. KPI3: cost of diagnosis**

The descriptive statistics for KPI3 Cost of diagnosis are shown in Figure 18. In the inferential analysis, as shown in Figure 18(a), the significance level of the pretest is <0.005, and of the posttest is <0.005. Where in both cases (Pretest and Post-test) their significance is less than 0.05, then it is affirmed that the data do not have a normal distribution. In this case, since the significance value is less than 0.05, it is accepted that the use of predictive machine learning, applying CRISP-DM, has an impact on the cost of diagnosing type 2 diabetes mellitus. In the descriptive analysis, according to the results shown in Figure 18(b), for the indicator cost of diagnosis (KPI3), in the pretest the mean value was S/12 and for the posttest it was S/6.67. With these results, it can be seen that there was a decrease of 44.42%. Since KPI3 does not have a normal distribution, the Wilcoxon hypothesis test was performed as shown in Figure 19, which yielded a significance level equal to 0.002, which is less than 0.05, the threshold value to see if the research hypothesis is accepted. In this case, since the significance value is less than 0.05, it is accepted that the use of predictive machine learning, applying CRISP-DM, has an impact on the cost of diagnosing type 2 diabetes mellitus.

**4.4. KPI4: difficulty level**

The instrument used to measure this KPI was a Likert scale questionnaire. According to the results shown in Figure 20, for the level of difficulty indicator (KPI4), favorable results are shown in the Post-Test, increasing the responses in "very easy" and "easy", and decreasing in "very difficult", "difficult" and "normal". In this case, since the significance value is less than 0.05, it is accepted that the use of predictive Machine Learning, applying CRISP-DM, has an impact on the level of difficulty in the diagnosis of type 2 diabetes mellitus.

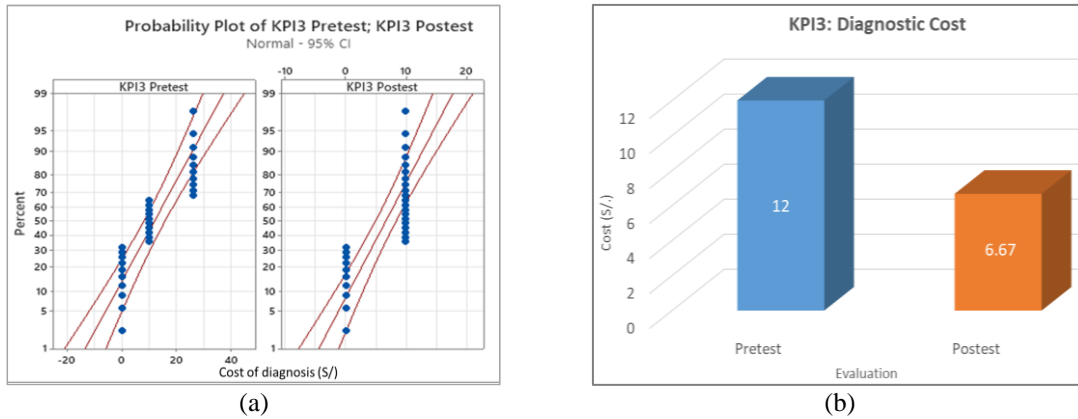


Figure 18. Results of the third indicator in (a) normality graph using Minitab 20.3 and (b) histogram of pretest and posttest of cost of diagnosis

**Test Statistics**

	Final Diagnostic Cost
	Initial Diagnostic Cost
Z	-3,162 <sup>b</sup>
Asymptotic sig. (bilateral)	,002

a. Wilcoxon signed-rank test  
 b. It is based on positive ranges

Figure 19. Wilcoxon test report for KPI3 using SPSS

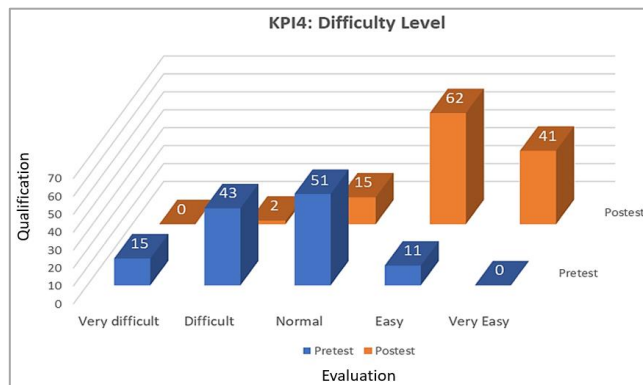


Figure 20. Histogram of pretest and posttest of the fourth KPI4 indicator

### 5. DISCUSSION

From the results obtained in the present research work, it can be seen in Figure 14, that for the information-gathering time indicator, in the pre-test a value in the mean of 8701.83 seconds was obtained and for the post-test, it was 1019.90 seconds; a decrease of 88.28% was observed. These results are consistent with research [20] in which he states that the application of artificial intelligence makes it possible to overcome some of the common complexities surrounding diseases such as tuberculosis, mainly the high workloads of health personnel, the limitations of human resources for the care and collection of signs and symptoms, often generate problems in urban health. As shown in Figure 16, for the diagnostic cost indicator, the pre-test showed a mean value of 12 soles and for the post-test, it was 6.67 soles; a decrease of 44.42% was observed. These results coincide with the research of [22] in which he indicates that through the support of artificial intelligence, the current diagnostic performance of fractures was improved through more comprehensive and complete use of existing clinical data, achieving data proofing and potentially reducing unnecessary tests, saving resources and allowing significant cost reductions. The diagnostic difficulty level indicator is visualized in Figure 20 where the pre-test 63.33% of respondents consider that the process is difficult and for the post-test, 100% of

respondents consider that after the implementation of the solution, the process is easy. These results coincide with the research of [21] where the application of an algorithm would facilitate the diagnosis of diseases from the data due to its ability to model the processes of medical reasoning, the values obtained in the study are evidence of the favorable evaluation of AI techniques to decrease the difficulty of disease diagnosis and increase satisfaction.

## 6. CONCLUSIONS

In order to carry out the present study, data from the ICU repository was used, which is mentioned and recommended by numerous studies; the factors and data were validated by the Peruvian League Against Diabetes organization. The CRISP-DM methodology was used to manage the data, deploy all its phases, and facilitate the development and application of the models described herein, which were developed in R language. First, a comparison of the classification models was made, and through the performance evaluated by metrics such as accuracy, it was determined which of them provided the most accurate and efficient predictions, and the best model was random forest, which obtained 90.43% accuracy. Subsequently, this model was implemented in a web platform that was developed with the laravel and vuejs frameworks, in addition to libraries such as R Plumber, which allowed the interaction and the necessary tests to be carried out with the League's volunteers to determine the impact of the proposed solution. The graphs show that there is evidence of improvements in the indicators determined in the study. We believe that it is important to continue with similar studies to allow mining to generate improvements in the diagnosis of various diseases and support the medical field.




## REFERENCES

- [1] I. D. Federation, "IDF diabetes atlas, 10th ed.," 2021.
- [2] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 261–268, 2019, doi: 10.14569/ijacsa.2019.0100637.
- [3] I. Gnanadass, "Prediction of gestational diabetes by machine learning algorithms," *IEEE Potentials*, vol. 39, no. 6, pp. 32–37, 2020, doi: 10.1109/MPOT.2020.3015190.
- [4] "Diabetes - OPS/OMS | Organización Panamericana de la Salud," 2022, [Online]. Available: <https://www.paho.org/es/temas/diabetes>.
- [5] P. B. K. Chowdary and R. U. Kumar, "Diabetes classification using an expert neuro-fuzzy feature extraction model," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, pp. 368–374, 2021, doi: 10.14569/IJACSA.2021.0120842.
- [6] WHO, "Recommendations for people living with NCDs, caregivers, family members and the public," *World Health Organization*, no. April, pp. 1–6, 2020, [Online]. Available: <https://apps.who.int/iris/handle/10665/331473>.
- [7] A. Mísbah and A. Ettlbi, "Towards machine learning models as a key mean to train and optimize multi-view web services proxy security layer," *International Journal of Recent Contributions from Engineering, Science & IT (IJES)*, vol. 6, no. 4, p. 65, 2018, doi: 10.3991/ijes.v6i4.9883.
- [8] O. Iparaguire-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente, and M. Cabanillas-Carbonell, "Sentiment analysis of tweets using unsupervised learning techniques and the k-means algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 571–578, 2022, doi: 10.14569/IJACSA.2022.0130669.
- [9] S. Afrin et al., "Supervised machine learning based liver disease prediction approach with LASSO feature selection," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3369–3376, 2021, doi: 10.11591/eei.v10i6.3242.
- [10] Instituto Nacional de Estadística e Informática, "Perú: Enfermedades no transmisibles y transmisibles," *Instituto Nacional de Estadística e Informática*, p. 196, 2021, [Online]. Available: [https://www.inei.gov.pe/media/MenuRecursivo/publicaciones\\_digitales/Est/Lib1657/libro.pdf](https://www.inei.gov.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1657/libro.pdf).
- [11] D. K. Yadav, C. Azad, K. Bala, P. K. Sharma, and S. Kumar, "Genetic algorithm and Naïve Bayes-based (GANB) diabetes mellitus prediction system," *Lecture Notes in Electrical Engineering*, vol. 887, pp. 561–572, 2023, doi: 10.1007/978-981-19-1906-0\_47.
- [12] A. Samanta, A. Saha, S. C. Satapathy, S. L. Fernandes, and Y. D. Zhang, "Automated detection of diabetic retinopathy using convolutional neural networks on a small dataset," *Pattern Recognition Letters*, vol. 135, pp. 293–298, 2020, doi: 10.1016/j.patrec.2020.04.026.
- [13] A. S. Alanazi and M. A. Mezher, "Using machine learning algorithms for prediction of diabetes mellitus," *2020 International Conference on Computing and Information Technology, ICCIT 2020*, 2020, doi: 10.1109/ICCIT-144147971.2020.9213708.
- [14] M. S. Geetha Devasena, R. Kingsy Grace, and G. Gopu, "PDD: Predictive diabetes diagnosis using datamining algorithms," *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*, 2020, doi: 10.1109/ICCCI48352.2020.9104108.
- [15] B. S. Bahnam and S. A. Dawwod, "A proposed model for diabetes mellitus classification using coyote optimization algorithm and least squares support vector machine," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 1164–1174, 2022, doi: 10.11591/ijai.v11.i3.pp1164-1174.
- [16] N. A. M. Aseri et al., "Comparison of meta-heuristic algorithms for fuzzy modelling of COVID-19 illness' severity classification," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, pp. 50–64, 2022, doi: 10.11591/ijai.v11.i1.pp50-64.
- [17] Y. Ye et al., "Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: A retrospective cohort study," *Journal of Diabetes Research*, vol. 2020, 2020, doi: 10.1155/2020/4168340.
- [18] M. T. Le, M. T. Vo, N. T. Pham, and S. V. T. Dao, "Predicting heart failure using a wrapper-based feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, pp. 1530–1539, 2021, doi: 10.11591/ijeecs.v21.i3.pp1530-1539.
- [19] O. Khaled, M. ElSahhar, M. A. El-Dine, Y. Talaat, Y. M. I. Hassan, and A. Hamdy, "Automatic classification of preliminary diabetic retinopathy stages using CNN," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, pp. 713–721, 2021, doi: 10.14569/IJACSA.2021.0120289.
- [20] W. H. Curioso and M. J. Brunette, "Inteligencia artificial e innovación para optimizar el proceso de diagnóstico de la tuberculosis," *Revista Peruana de Medicina Experimental y Salud Pública*, vol. 37, no. 3, pp. 554–8, Sep. 2020, doi: 10.17843/rpmesp.2020.373.5585.




- [21] N. González, V. Estrada, and A. Febles, "Estudio y selección de las técnicas de Inteligencia Artificial para el diagnóstico de enfermedades," *Revista de Ciencias Médicas de Pinar del Río*, vol. 22, no. 3, pp. 534–544, 2018, [Online]. Available: [http://scielo.sld.cu/scielo.php?script=sci\\_arttext&pid=S1561-31942018000300014&lng=es&nrm=iso&tlng=pt](http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-31942018000300014&lng=es&nrm=iso&tlng=pt).
- [22] J. E. Karakowski, "Artificial intelligence support for more accurate diagnosis of lateral malleolar fractures," *2017 IEEE MIT Undergraduate Research Technology Conference, URTC 2017*, vol. 2018-January, pp. 1–4, 2018, doi: 10.1109/URTC.2017.8284171.
- [23] I. Corporation, "Manual CRISP-DM de IBM SPSS modeler," *IBM Corporation*, p. 280, 2012, [Online]. Available: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/UsersGuide.pdf>.
- [24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [25] R. S. Raj, D. S. Sanjay, M. Kusuma, and S. Sampath, "Comparison of support vector machine and Naïve Bayes classifiers for predicting diabetes," *1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering, ICATIECE 2019*, pp. 41–45, 2019, doi: 10.1109/ICATIECE45860.2019.9063792.
- [26] T. Mahboob Alam *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, 2019, doi: 10.1016/j.imu.2019.100204.
- [27] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm," *2018 21st International Conference of Computer and Information Technology, ICCIT 2018*, 2019, doi: 10.1109/ICCITECHN.2018.8631968.

## BIOGRAPHIES OF AUTHORS






**Victor Garcia-Rios**    Systems Engineer from the Universidad Autónoma del Perú, Development Analyst, Web Master at IEEE Peru Section, Certified in Scrum fundamentals. Author of scientific articles indexed in IEEE Xplore, Scopus and WoS. He can be contacted at email: [vgarciar@autonoma.edu.pe](mailto:vgarciar@autonoma.edu.pe)






**Marieta Marres-Salhuana**    Systems Engineer from the Universidad Autónoma del Perú, Development Analyst at Falabella, Web Master at IEEE Peru Section, Certified in Scrum fundamentals. Author of scientific articles indexed in IEEE Xplore, Scopus and WoS. E-mail: [mmarres@autonoma.edu.pe](mailto:mmarres@autonoma.edu.pe).



**Fernando Sierra-Liñan**    Mg. Fernando Sierra-Liñan has a Bachelor's degree in Education, specializing in Science and Technology at USIL, a Master's degree in Edumatics and University Teaching at UTP, a Bachelor's degree in Systems Engineering and Computer Science at UTP, with a technical specialty in Computer Science and Computer Science. He is currently working as a researcher and thesis advisor in the faculty of Computer Engineering and Systems at the Universidad Privada del Norte, Lima-Peru. He has 20 years of teaching experience. His areas of interest are programming, database and data analysis. E-mail: [fernando.sierra@upn.edu.pe](mailto:fernando.sierra@upn.edu.pe), [pfsierra.D02052@gmail.com](mailto:pfsierra.D02052@gmail.com).



**Michael Cabanillas-Carbonell**    Engineer and Master in Systems Engineering from the National University of Callao - Peru, PhD candidate in Systems Engineering and Telecommunications at the Polytechnic University of Madrid. President of the chapter of the Education Society IEEE-Peru. Conference Chair of the Engineering International Research Conference IEEE Peru EIRCON. Research Professor at Norbert Wiener University, Professor at Universidad Privada del Norte, Universidad Autónoma del Perú. Advisor and Jury of Engineering Thesis in different universities in Peru. International lecturer in Spain, United Kingdom, South Africa, Romania, Argentina, Chile, China. Specialization in Software Development, Artificial Intelligence, Machine Learning, Business Intelligence, Augmented Reality. Reviewer IEEE Peru and author of more than 50 scientific articles indexed in IEEE Xplore and Scopus. He can be contacted at [mcabanillas@ieee.org](mailto:mcabanillas@ieee.org)