

# Un enfoque de Machine Learning y razonamiento probabilístico en el pronóstico de factores de riesgo de la diabetes

## A Machine Learning approach and probabilistic reasoning in the prognosis of risk factors for diabetes

Jonathan P. Cárdenas-Ruperti<sup>1</sup>, Francisco S. Durán-Pérez<sup>2</sup>, Luis A. Padilla-Álvarez<sup>3</sup>

### RESUMEN

Este trabajo investigativo propone el diseño de un modelo computacional de clasificación a través del uso de técnicas de Machine Learning, hemos optado por trabajar con el aprendizaje automático apoyado en un modelo probabilístico que permita evaluar los factores de riesgo de la enfermedad de la diabetes tipo 2. Este modelo está basado en el método de regresión logística, que según a los factores de riesgo de la enfermedad, mostrará como resultado el porcentaje que tiene el paciente de contraer la misma. Se aplicó la metodología de investigación bibliográfica que aporte con el conocimiento necesario para la realización de este proyecto en el cual se realizaron pruebas para verificar el comportamiento de cada una de las variables utilizadas en el modelo probabilístico, el cual brindará resultados eficientes y en un corto periodo de tiempo siendo así una herramienta de apoyo en la toma de decisiones a los expertos y aportando con el diagnóstico oportuno para prevenir la enfermedad.

**Palabras clave:** Diabetes, Machine Learning, Razonamiento Probabilístico, Regresión Logística.

### ABSTRACT

This research work proposes the design of a classification computational model through the use of Machine Learning techniques, we have chosen to work with machine learning supported by a probabilistic model that allows evaluating the risk factors of type 2 diabetes disease. This model is based on the logistic regression method, which, according to the risk factors of the disease, will show as a result the percentage that the patient has of contracting it. The bibliographic research methodology was applied that provides the necessary knowledge to carry out this project in which tests were carried out to verify the behavior of each of the variables used in the probabilistic model, which will provide efficient results in a short time. period thus being a tool to support experts in decision-making and providing the appropriate diagnosis to prevent the disease.

**Keywords:** Diabetes, Machine Learning, Probabilistic Reasoning, Logistic Regression.

**Fecha de recepción:** Febrero 8, 2020.

**Fecha de aceptación:** Marzo 25, 2020.

### Introducción

Es una de las mayores emergencias mundiales de salud del siglo XXI. Cada año más y más personas viven con esta condición, que puede desencadenar complicaciones a lo largo de la vida. Además de los 415 millones de adultos que actualmente tienen diabetes, hay 318 millones de adultos con tolerancia a la glucosa alterada, lo que les sitúa en un alto riesgo de desarrollar la enfermedad en el futuro.

El proceso de su tratamiento es costoso. Las medidas y soluciones establecidas por los gobiernos y suscritas por las más altas autoridades en diabetes de todo el mundo, no han tenido a la fecha

el esperado efecto beneficioso, pues este mal no disminuye su frecuencia, intensidad y gravedad, como consecuencia de sus particulares características epidemiológicas. Sin embargo, se tienen la seguridad y la esperanza de que con el esfuerzo y las acciones continuas en todas las naciones, la humanidad se verá libre de las enfermedades crónicas no transmisibles, en un futuro que es de esperar que no tarde demasiado.

La prevención de esta enfermedad exige la adopción de una perspectiva que abarque todo el ciclo vital. En la fase más temprana, cuando se forman los hábitos alimentarios y de actividad física y cuando la regulación del equilibrio energético se puede programar para el futuro a largo plazo, existe un periodo crítico

<sup>1</sup> Ingeniero en Sistemas e Informática, Máster en Gerencia de Sistemas y Tecnologías de la Información, Docente Investigador de la Facultad de Ingenierías en la Universidad Técnica "Luis Vargas Torres" de Esmeraldas, Ecuador. E-mail: jonathan.cardenas.ruperti@utelvt.edu.ec.

<sup>2</sup> Ingeniero Comercial Mención Administración de la Productividad, Magister en Administración de Empresas con Mención en Gerencia de la Calidad y Productividad, Desarrollo Organizacional, Administración de Procesos y Sistemas de Gestión de la Calidad. 8 años de experiencia en Gestión de Calidad Hospitalaria y Seguridad del Paciente en Hospitales Públicos y Privados, Consultoría y Capacitación en Planificación Estratégica y Desarrollo de Proyectos de Factibilidad para la creación de

Empresas e Institutos de Educación Superior y Escuelas Profesionales de Conducción. E-mail: fsduranp@gmail.com.

<sup>3</sup> Tnlgo. En Sistemas, Lic. En Informática Educativa, Ing. En Sistemas Informáticos, Magister en Ciberseguridad, Docente Investigador de la Facultad de Ingenierías en la Universidad Técnica "Luis Vargas Torres" de Esmeraldas, Ecuador. E-mail: luis.padilla@utelvt.edu.ec.

**Como citar:** Cárdenas-Ruperti J, Durán-Pérez, F. & Padilla-Álvarez L. (2020). *Ecuadorian Science Journal*. 4(1), 62-67.  
DOI: <https://doi.org/10.46480/esj.4.1.95>

en el que se puede intervenir con el fin de contrarrestar el riesgo de padecer obesidad y diabetes de tipo 2 en años posteriores. Mellitus, D. (2005).

El aprendizaje automático está siendo utilizado en varias áreas de la medicina, muchas de estas aportaciones han obtenido resultados muy favorables. En este proyecto implementamos esos conocimientos mediante el uso de técnicas de machine Learning a través de un modelo computacional diseñaremos un sitio web dirigido al experto del área, el cual podrá ayudarse mediante el uso de esta herramienta desarrollada en Python la cual podrá ser utilizada en el pronóstico de la enfermedad. Baranwal, A., Bagwe, B. R., & Vanitha, M. (2020).

## Trabajos Relacionados

Ya hace algunos años se han venido implementando el uso de sistemas expertos en el área de la salud, los cuales han sido de gran ayuda y han aportado de manera favorable, como algunos ejemplos que nos ayudaron en la realización de este proyecto tenemos:

Modelo estadístico predictivo para el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II. López Fernández, R. et al, (2016).

Prevalencia de la Diabetes Mellitus tipo 2 y factores asociados en la ciudad de Maracaibo, Venezuela. Cordero, L. C. A, et al. (2017)

Prevalencia de Diabetes por diagnóstico Médico previo en México. Rivas-Acuña, V., et al, (2011).

Una aplicación del Modelo de Regresión Logística en la predicción del rendimiento estudiantil. Domínguez Alonso, E., & Aldana Padilla, D. (2001).

## Materiales y métodos

El desarrollo de la investigación específica el uso de la siguiente metodología que proporcionaron las herramientas necesarias para realizar una evaluación adecuada a la información recopilada.

### Inteligencia Artificial

La Inteligencia Artificial es una de las ramas de las ciencias de la computación que más interés ha despertado en la actualidad, debido a su enorme campo de aplicación. La búsqueda de mecanismos que nos ayuden a comprender la inteligencia y realizar modelos y simulaciones de estos, es algo que ha motivado a muchos científicos a elegir esta área de investigación De Mitri, M. J. (2019).

### Machine Learning

El aprendizaje automático como un conjunto de métodos que pueden detectar automáticamente patrones en los datos, y luego utiliza los patrones descubiertos para predecir futuros datos, o para realizar otros tipos de toma de decisiones bajo incertidumbre Alonso Castrillejo, S. (2020).

### Aprendizaje Supervisado

El aprendizaje supervisado es el más común utilizado entre los dos, incluye algoritmos tales como regresión lineal y logístico, clasificación de clases múltiples y máquinas de vectores de soporte. El aprendizaje supervisado se llama así porque el

desarrollador actúa como una guía para enseñar al algoritmo las conclusiones a las que debe llegar, es decir la salida del algoritmo ya es conocida. Es similar a la forma en que un niño podría aprender de un maestro Aguirre Ascona, Y. D. (2019).

### Modelo de Regresión

Es un problema de predecir una etiqueta de valor real a menudo llamada objetivo dado un Ejemplo sin etiqueta. Estimación de la valoración del precio de la vivienda según las características de la casa, como el área, El número de habitaciones, la ubicación, etc. es un famoso ejemplo de regresión Aguirre Ascona, Y. D. (2019).

### Modelo de Regresión Logística

Cuando se quiere evaluar la relación entre una variable que suscita especial interés (variable dependiente que suele denominarse  $Y$ ) respecto a un conjunto de variables (variables independientes, que se denominan  $X_1, X_2, \dots, X_n$ ) las pruebas de contraste de hipótesis mostradas hasta ahora no nos aportan suficiente información sobre la relación en conjunto de todas ellas, dado que los contrastes de hipótesis que conocemos hasta ahora se basan en probar relaciones bivariantes (2 variables), en las que no se tiene en cuenta la posibilidad de que haya otras variables de interés y en las que el sentido de la relación es bidireccional. Es entonces cuando resulta adecuado y conveniente la aplicación de los modelos de regresión. Los modelos de regresión permiten evaluar la relación entre una variable dependiente respecto a otras variables en conjunto independientes. Abreu, M. N. S., Siqueira, A. L., & Caiaffa, W. T. (2009).

### Regresión Logística Binaria

Se usa cuando la variable dependiente es una variable binaria, es decir, de solo dos categorías, también conocidas como dummy o dicotómica. Por ejemplo: 1 – Sí, 0 – No

### Sobre la bondad del modelo:

1. Significación de chi-cuadrado del modelo en la prueba omnibus: Si la significación es menor de 0,05 indica que el modelo ayuda a explicar el evento, es decir, las variables independientes explican la variable dependiente.

2. R-cuadrado de Cox y Snell, y R-cuadrado de Nagelkerke: Indica la parte de la varianza dependiente explicada por el modelo. Hay dos R-cuadrados en la regresión logística, y ambas son válidas. Se acostumbra a decir que la parte de la variable dependiente explicada por el modelo oscila entre la R-cuadrado de Cox y Snell y la R-cuadrado de Nagelkerke. Cuanto más alto es la R-cuadrado más explicativo es el modelo, es decir, las variables independientes explican la variable dependiente.

3. Porcentaje global correctamente clasificado: Este porcentaje indica el número de casos que el modelo es capaz de predecir correctamente. Me explico. En base a la ecuación de regresión y los datos observados, se realiza una predicción del valor de la variable dependiente (valor pronosticado). Esta predicción se compara con el valor observado. Si acierta, el caso es correctamente clasificado. Si no acierta, el caso no es correctamente clasificado. Cuantos más casos clasifica correctamente (es decir coincide el valor pronosticado con el valor observado) mejor es el modelo, más explicativo, por tanto, las variables independientes son buenas predictoras del evento o variable dependiente. Si es modelo clasifica correctamente más del 50% de los casos, el modelo se acepta. Si no, punto final y a volver a empezar, y seleccionaríamos nuevas variables independientes.

Los siguientes pasos son para evaluar la relación de cada variable independiente con la variable dependiente

**Sobre la relación de las variables independientes con la variable dependiente:**

- 4. Significación de b: si es menor de 0,05 esa variable independiente explica la variable dependiente
- 5. Signo de b: indica la dirección de la relación. Por ejemplo, a más nivel educativo mayor probabilidad que suceda el evento.
- 6. Exp (b): indica la fortaleza de la relación. Cuanto más alejada de 1 está más fuerte es la relación. Para comparar los exponentes de b entre sí, aquellos que son menores a 1 deben transformarse en su inverso o recíproco, es decir, debemos dividir 1 entre el exponente de b (pero solo cuando sean menores a 1). Cevallos-Torres, L., & Botto-Tobar, M. (2019).

**Caso de Uso**

En el presente caso de estudio se va a determinar cómo inciden los factores de riesgo en una persona para que esta sea propensa a contraer la enfermedad de la diabetes. Este trabajo se centró en aquellas personas que tienden a padecer la diabetes tipo 2, del cual se tomó una muestra en una base de datos con 768 registros, los cuales fueron objetos de estudio.

Para conocer el comportamiento de la variable dependiente que en nuestro caso es Clase, se procedió con la utilización del modelo de regresión logística binaria para evaluar su comportamiento, mediante las variables independientes se denotara el modelo logístico de la siguiente manera:

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \epsilon_i$$

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

En donde denotamos que p es la probabilidad de que exista una situación de interés, para nuestro caso cual es la probabilidad de que una persona adquiera la enfermedad de la diabetes tipo 2,  $b_0 + b_1 + b_2 + \dots + b_k$ , son coeficientes  $\beta$  para una mayor comprensión de estos coeficientes es necesario tomar en cuenta una definición muy importante como lo es la variable dependiente clase en nuestro estudio. Este concepto se define como la probabilidad de que un evento suceda o de que no suceda, este proceso también es conocido como ODDS ratios lo cual no muestra cuanto es modificada una variable.

$$p = \frac{eb_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k}{1 + eb_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k}$$

**Clasificación y predicción de los sujetos según los factores de incidencia**

Para poder realizar el pronóstico a una persona que puede llegar a padecer la diabetes tipo 2, se realizó el estudio mediante un análisis de la variable dependiente Clase mediante los factores de incidencia de la enfermedad que son nuestras variables independientes, se procedió a realizar el método de regresión logística binaria con el fin de encontrar las relaciones que tienen estas variables y conocer cuáles eran las más relevantes en nuestro

estudio, y a partir de ese estudio realizado conocer que tan alta o que tan baja es la probabilidad de que una persona mediante esos factores de riesgo pueda contraer la enfermedad.

Para este análisis se tomó en cuenta todas las variables y se analizó el comportamiento de cada una de ellas en relación a la variable dicotómica, de ese análisis se procedió a descartar algunas variables que no eran significativas, analizando su relación y comportamiento con la variable dependiente.

**Análisis de Correlación de Pearson**

Utilizaremos esta herramienta de análisis estadístico para analizar el comportamiento de nuestra variable dependiente con respecto a cada una de nuestras variables independientes:

**Tabla 1:** Correlación de Pearson

	Embarazo	Glucosa	Presión	Piel	Insulina	Imc	Genealogía	Edad
Embarazo	1	,129**	,141**	-,082*	-,074*	0,02	-,034	,544**
Glucosa	,129**	1	,153**	0,057	,331**	,221**	,137**	,264**
Presión	,141**	,153**	1	,207**	,089*	,282**	0,041	,240**
Piel	-,082*	0,057	,207**	1	,437**	,393**	,184**	-,114**
Insulina	-,074*	,331**	,089*	,437**	1	,198**	,185**	-0,042
Imc	0,018	,221**	,282**	,393**	,198**	1	,141**	0,036
Genealogía	-,034	,137**	0,041	,184**	,185**	,141**	1	0,034
Edad	,544**	,264**	,240**	-,114**	-0,042	0,04	0,034	1

En el cuadro 1 es la medida de la asociación lineal entre dos variables. Los valores del coeficiente de correlación van de -1 a 1. El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la fuerza. Los valores mayores indican que la relación es más estrecha.

**Curva ROC**

La línea de puntos del gráfico 1, cuanto más se acerque la curva al extremo superior izquierdo de la cuadrícula la prueba tendrá más capacidad de probabilidad de tener la enfermedad. Si nos damos cuenta en el gráfico 32, podemos observar que existe un área que se forma bajo la curva ROC, a lo que entiendo como la probabilidad de clasificar correctamente si una persona tiene o no la diabetes, seleccionados al azar, los valores AUC bajo la curva ROC van entre 0.5 hasta un máximo que es 1.

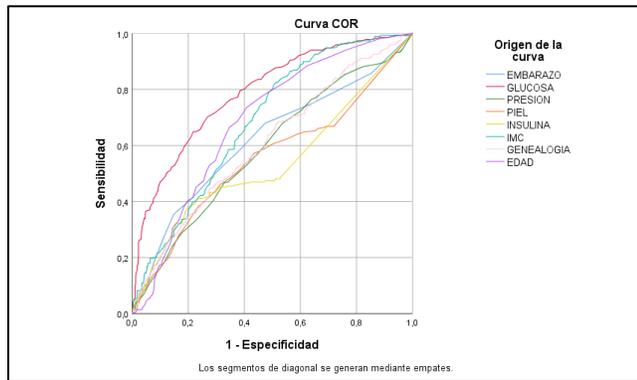


Figura 1. Curva ROC de las variables de estudio

Mediante el análisis de la Curva de Roc en la figura 1 podemos observar que la variable Glucosa, es la variable más significativa ya que se encuentra sobre el ajuste y es la variable que mayor probabilidad tiene a llegar a 1, seguida de la variable Edad que se encuentra en segunda posición y también cuenta con un grado de relevancia menor sobre la variable dependiente, de acuerdo a este análisis la variable IMC también juega un factor muy importante ya que está en el tercer lugar de las variables que tienen a tener mayor influencia, luego ya viene las demás variables las cuales van perdiendo relevancia porque se alejan del área donde tienen a llegar a 1, en este grupo se encuentra la variable piel, presión, insulina, genealogía, embarazo, presión y piel tienen, como mínimo, un empate entre el grupo de las variable que más alejadas se encuentran para ser relevantes. El estado real positivo es 1.

### Área bajo la curva

Tabla 2: Áreas bajo la curva ROC

VARIABLES	Área
EMBARAZO	0,62
GLUCOSA	0,788
PRESION	0,586
PIEL	0,554
INSULINA	0,538
IMC	0,688
GENEALOGIA	0,606
EDAD	0,687

En Tabla 2 Las variables de resultado de prueba: EMBARAZO, GLUCOSA, PRESIÓN, PIEL, INSULINA, IMC, GENEALOGÍA, EDAD tienen, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas

### Análisis multivariado de los datos

Para realizar nuestro estudio en las variables se optó por el modelo de regresión logística para desarrollar el modelo predictivo basado en los factores de riesgo de la enfermedad de la diabetes tipo 2. En este caso su precisión fue evaluada a través de la curva del ROC y la calibración mediante el método de Hosmer y Lemeshow, esta prueba parte de la idea de que si el ajuste es bueno, un valor alto de la probabilidad predicha (p) se asociará con el resultado 1 de la variable binomial dependiente, mientras

que un valor bajo de p (próximo a cero) corresponderá con el resultado Y=0.

### Tabla de Clasificación

También llamada matriz de confusión nos ayuda a evaluar el ajuste de la RL, se lo usa como indicador de bondad de ajuste.

Tabla 3. Matriz de confusión

Tabla de clasificación					
Paso 3	CLASE	0	439	61	87,8
		1	113	155	57,8
Porcentaje global					77,3
a. El valor de corte es ,500					

En la tabla 3 la estimación global es de 77,3 % del ajuste del modelo de regresión logística, y se dio en el paso tres.

### Etapa de validación

Para la etapa de validación se utilizará el método hacia atrás WALD, donde se introducen en el modelo todas las variables y se van suprimiendo si cumplen una serie de condiciones definidas a priori hasta que no se pueden eliminar más, es decir ninguna variable cumpla la condición impuesta.

### Evaluación del Modelo REGRESIÓN LOGÍSTICA

Para la validación estadística se realizaron varias pruebas al modelo:

### Prueba Ómnibus

Esta prueba es útil para analizar la significancia conjunta de las componentes del modelo, donde:

$$H0: \beta_0 = \beta_1 = \beta_2 = \beta_j = \dots = \beta_k = 0$$

$$H1: \exists j \in \{0, 1, 2, \dots, k\} \text{ tal que } \beta_j \neq 0$$

Tabla 4. Pruebas ómnibus de coeficientes de modelo: resultados

		Chi-cuadrado	G.L	Sig.
Paso 1	Paso	270,039	8	0,000
	Bloque	270,039	8	0,000
	Modelo	270,039	8	0,000
Paso 2ª	Paso	-0,008	1	0,929
	Bloque	270,03	7	0,000
	Modelo	270,03	7	0,000
Paso 3ª	Paso	-2,008	1	0,156
	Bloque	268,022	6	0,000
	Modelo	268,022	6	0,000

a. Un valor negativo de chi-cuadrados indica que el valor de chi-cuadrados ha disminuido del paso anterior.

En Tabla 4. para la prueba ómnibus el estadístico de Chi-cuadrado, evalúa la hipótesis nula de que los coeficientes  $\beta_j$  de todos los términos (menos la constante) incluidos en el modelo son cero. Puesto que el p-value en todos los casos es menor a 0.05, podemos afirmar que existe la suficiente evidencia estadística para rechazar la hipótesis nula, es decir la tabla de Chi-cuadrado que evalúa la hipótesis nula nos indica que existe algún  $\beta_j$  diferente de cero.

**Presentación de los Resultados.**

El estadístico de prueba es. Donde  $\sigma^2 \beta_j$  es el error del parámetro  $\beta_j$ , es el estadístico de Wald  $W$ , este sigue una distribución Chi-cuadrado con 1 grado de libertad y significancia  $\alpha = 0.05$ .

$$w = \frac{\beta_j^2}{\sigma^2(\beta_j)}$$

**Contraste de hipótesis**

Recordemos que el presente trabajo sobre el uso de machine Learning y razonamiento probabilístico para determinar los factores que inciden en la diabetes se ha planteado como hipótesis alternativa.

Ha. La probabilidad de tener la enfermedad de la Diabetes se puede predecir mediante los factores de riesgo como son embarazo, glucosa, presión, piel, insulina, IMC, genealogía y edad.

Esta prueba selecciona  $\beta_j$  como el parámetro de interés para contrastar la siguiente hipótesis:

**$H_0: \beta_j = 0$                        $H_a: \beta_j \neq 0$**

**Desarrollo de la prueba de hipótesis**

En el análisis de la prueba de hipótesis, lo haremos mediante el desarrollo de supuesto, en la regresión múltiple nos permite probar un modelo de predicción sobre la variable dependiente la cual se enfoca principalmente en la validez interna.

Para validar nuestra hipótesis alternativa lo haremos con los supuestos analizando sus pruebas de independencia de errores y no multicolinealidad.

**Tabla 5.** Prueba de independencia de errores

R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
.551a	.303	.296	.400	1.982

En tabla 5 el puntaje de la prueba de Durbin-Watson indica que hay independencia de errores de 1.982, lo cual indica que las fuentes que producen los errores no son las mismas es decir que los errores son de condiciones diferentes. Si el valor esta entre 1 y 3 aceptamos el supuesto.

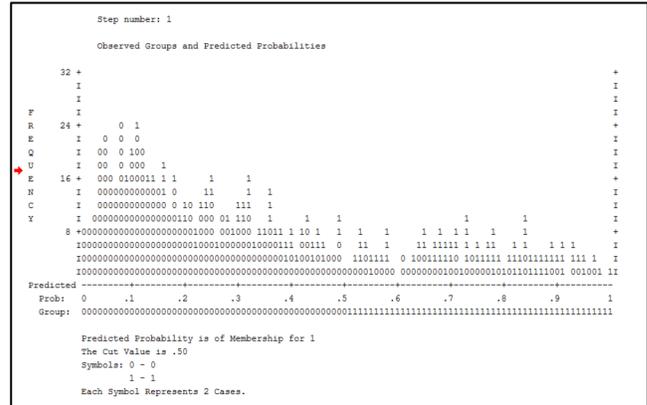
**Tabla 6.** Coeficientes

	B	D. error	Beta	t	sig.	Tolerancia	VIF
<b>(Constante)</b>	<b>-.854</b>	<b>.085</b>		-9.989	.000		
Embarazo	.021	.005	.145	4.014	.000	.699	1.431
Glucosa	.006	.001	.397	11.493	.000	.770	1.299
Presión	-.002	.001	-.095	-2.873	.004	.846	1.182
Piel	.000	.001	.005	.139	.890	.663	1.507
Insulina	.000	.000	-.044	-1.205	.229	.701	1.428
IMC	.013	.002	.219	6.344	.000	.771	1.297
Genealogia	.147	.045	.102	3.268	.001	.937	1.067
Edad	.003	.002	.065	1.693	.091	.630	1.588

en Tabla 6. el factor de varianza inflada VIF. nos indica que se cumple el supuesto de no multicolinealidad donde sus valores están entre 1.067 y 1.588. Se cumple con el supuesto cuando ningún valor está por encima de 10 y en conjunto sus valores están cercanos a 1

**Resultado del contraste de hipótesis**

En ambos casos de independencia de errores y no multicolinealidad se cumple con los supuesto por lo tanto la hipótesis de nuestra investigación es válida.



**Figura 2.** Probabilidades pronosticadas.

En el gráfico 11 podemos notar que si la probabilidad es menor a 0.5 el me lo asigna a la categoría 0 y los 1 que sale en la tabla de especificación están mal asignadas en el modelo, y cuando es más de 0.5 nos asigna a la categoría 1 igual que lo anterior los 0 aquí están mal asignados pero son pocos en el modelo por lo tanto está bien la probabilidad que existe.

**Conclusiones**

El actual artículo presenta un Modelo probabilístico que permite analizar los factores de incidencia en el desarrollo de la enfermedad de la diabetes tipo 2, el mismo que se implementó usando conocimientos basados con el uso de técnicas de machine Learning, mediante información que fue almacenada en una base de datos, se pudo obtener el conocimiento sobre la existencia de la enfermedad, aplicando el método de Regresión logística para determinar su peso porcentual, que fue usado en el modelo creado en SPSS mostrando el resultado del análisis de los factores de incidencia, finalmente se aplicó el modelo

**Agradecimientos**

Con mucha emoción agradezco a Dios primeramente, ya que sin la venia de él no se pudiera avanzar en los proyectos que me propongo en la vida y a mis padres que siempre han estado ahí con su apoyo incondicional hacia mi desarrollo profesional

**Referencias Bibliográficas**

Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes care*, 28(S37), S5-S10.  
 Baranwal, A., Bagwe, B. R., & Vanitha, M. (2020). Machine Learning in Python: Diabetes Prediction Using Machine Learning. In *Handbook of Research on Applications and Implementations of Machine Learning Techniques* (pp. 128-154). IGI Global.

- López Fernández, R., Yanes Seijo, R., Suárez Surí, P., Avello Martínez, R., Gutiérrez Escobar, M., & Alvarado Flores, R. (2016). Modelo estadístico predictivo para el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II. *Medisur*, 14(1), 42-52.
- Cordero, L. C. A., Vázquez, M. A., Cordero, G., Álvarez, R., Añez, R. J., Rojas, J., & Bermúdez, V. (2017). Prevalencia de la diabetes mellitus tipo 2 y sus factores de riesgo en individuos adultos de la ciudad de Cuenca-Ecuador. *Avances en biomedicina*, 6(1), 10-21.
- Rivas-Acuña, V., García-Barjau, H., Cruz-León, A., Morales-Ramón, F., Enríquez-Martínez, R. M., & Román-Alvarez, J. (2011). Prevalencia de ansiedad y depresión en las personas con diabetes mellitus tipo 2. *Salud en Tabasco*, 17(1-2), 30-35.
- Domínguez Alonso, E., & Aldana Padilla, D. (2001). Regresión logística: Un ejemplo de su uso en Endocrinología. *Revista Cubana de Endocrinología*, 12(1), 0-0.
- De Mitri, M. J. (2019). Predicción de marcadores cardíacos en pacientes diabéticos e hipertensos medicados por medio de inteligencia artificial (Doctoral dissertation, Universidad Católica de Córdoba).
- Alonso Castrillejo, S. (2020). Aplicación de algoritmos de machine learning en la predicción de la diabetes mellitus tipo II.
- Aguirre Ascona, Y. D. (2019). Métodos de aprendizaje supervisado para la predicción de diabetes: una revisión sistemática de la literatura.
- Granados, R. M. (2016). Modelos de regresión lineal múltiple. Granada, España: Departamento de Economía Aplicada, Universidad de Granada.
- Abreu, M. N. S., Siqueira, A. L., & Caiaffa, W. T. (2009). Regresión logística ordinal en estudios epidemiológicos. *Revista de Saúde Pública*, 43(1), 183-194.
- Cevallos-Torres, L., & Botto-Tobar, M. (2019). Case study: Project-based learning to evaluate probability distributions in medical area. In *Problem-Based Learning: A Didactic Strategy in the Teaching of System Simulation* (pp. 111-122). Springer, Cham.