A yellow and blue logo with a cross and a lamp

AI-generated content may be incorrect.

INGENIERÍA EN INTELIGENCIA ARTIFICIAL

**DOCUMENTO DE PROYECTO FINAL EN LA MATERIA DE ADQUISICION, ANALISIS Y PROCESAMIENTO DE DATOS**

**MODELOS PREDICTIVO PARA LA DETECCION DE DIABETES BASADO EN DATOS CLINICOS DE PACIENTES DIAGNOSTICADOS Y NO DIAGNOSTICADOS CON LA ENFERMEDAD**

**Docente: Marin Salazar Carmen Rosa**

**Estudiante: José Alfredo Zambrana Cruz**

Santa Cruz - Bolivia

Junio, 2025

**Resumen ejecutivo**

El proyecto tuvo como propósito desarrollar un modelo predictivo para la detección de diabetes, mediante el análisis de variables clínicas extraídas de un conjunto de datos disponible en la plataforma Kaggle (ver Anexos). El objetivo principal fue identificar los factores con mayor influencia en el diagnóstico de esta enfermedad. Para ello, se empleó un enfoque basado en inteligencia artificial que permitió evaluar la relevancia de cada variable en relación con la presencia o ausencia de diabetes.

El problema abordado se centró en la limitada implementación de herramientas de inteligencia artificial en el ámbito clínico boliviano, a pesar de su creciente adopción en otros países. Esta investigación buscó evidenciar el potencial de dichas herramientas, mediante la aplicación de un modelo predictivo que pudiera aportar precisión y utilidad al proceso diagnóstico.

Los resultados obtenidos mostraron que las variables con mayor influencia en el diagnóstico fueron el nivel de glucosa en sangre, el nivel de hemoglobina glicosilada (HbA1c) y la edad. El modelo alcanzó una precisión del 88 %, lo que indicó una capacidad significativa para predecir la presencia de diabetes con base en los datos disponibles.

El análisis permitió demostrar que la inteligencia artificial puede constituirse en una herramienta complementaria de valor para el diagnóstico clínico. Su implementación futura podría optimizar la identificación temprana de enfermedades como la diabetes, fortaleciendo así la toma de decisiones médicas basadas en datos.

**Definición del Problema**

El diagnóstico temprano de la diabetes representa un desafío constante en el ámbito clínico. A pesar de los avances tecnológicos, el uso de herramientas de inteligencia artificial (IA) en contextos de atención médica continúa siendo escaso en diversos países en desarrollo, entre ellos Bolivia. Este proyecto busca evaluar la viabilidad de implementar un modelo predictivo de IA que permita detectar la presencia de diabetes, a partir de datos clínicos de pacientes diagnosticados y no diagnosticados. Asimismo, se pretende identificar las variables con mayor influencia en el diagnóstico automatizado, y determinar cuál de los modelos evaluados proporciona el mejor desempeño en este contexto.

**Preguntas de investigación**

* ¿Qué tan eficiente es la inteligencia artificial en la detección de la diabetes?
* ¿Qué variables influyen en la predicción del diagnóstico según el modelo?
* ¿Qué modelo predictivo presenta el mejor rendimiento con el conjunto de datos empleado?

**Alcance y limitaciones**  
El proyecto se centró en la aplicación y evaluación de modelos de inteligencia artificial entrenados exclusivamente con los datos proporcionados por un conjunto disponible en la plataforma Kaggle. La metodología incluyó la comparación de distintos algoritmos de clasificación (Decision Tree, Logistic Regression, Random Forest, Support Vector Machine y Naive Bayes) para determinar cuál ofrecía mejores resultados predictivos.

* El alcance metodológico se limitó a la selección, entrenamiento y evaluación de modelos de clasificación supervisada. No se incorporaron datos clínicos externos ni se realizaron validaciones en entornos hospitalarios reales.
* El análisis de datos se restringió a las variables contenidas en el dataset original, sin adición de nuevas fuentes. Por tanto, los resultados y conclusiones solo son válidos en el marco de este conjunto de datos y no pretenden generalizarse a toda la población.

**Descripción de los Datos**

La fuente de datos utilizada para el desarrollo del modelo predictivo fue la plataforma Kaggle, específicamente un conjunto de datos relacionado con características clínicas de pacientes diagnosticados y no diagnosticados con diabetes.

El dataset estuvo conformado por **100.000 registros (filas)** y **9 variables (columnas)**. Estas variables incluyeron información demográfica, antecedentes médicos, historial de tabaquismo y valores clínicos relevantes para el diagnóstico de la enfermedad.

**Diccionario de variables utilizadas:**

| **Variable** | | **Descripción** | |
| --- | --- | --- | --- |
| gender | | Sexo biológico del paciente (Male, Female, Other) | |
| age | | Edad del paciente en años | |
| hypertension | | Presencia de hipertensión (1: Sí, 0: No) | |
| heart\_disease | | Presencia de enfermedad cardíaca (1: Sí, 0: No) | |
| smoking\_history | | Historial de consumo de tabaco (never, former, current, etc.) | |
| bmi | | Índice de masa corporal (Body Mass Index) | |
| HbA1c\_level | | Nivel de hemoglobina glicosilada | |
| blood\_glucose\_level | | Nivel de glucosa en sangre | |
| diabetes | Diagnóstico de diabetes (1: Sí, 0: No) | |

**Muestra del dataset:**

A continuación, se presenta una muestra representativa del conjunto de datos

A screenshot of a computer screen

AI-generated content may be incorrect.

Esta muestra ilustra la estructura de los datos disponibles y los valores esperados para cada una de las variables consideradas en el análisis.

**Columnas y sus valores:**

**Gender:**  
A screen shot of a computer

AI-generated content may be incorrect.

**Age:**

A screenshot of a computer

AI-generated content may be incorrect.

**Hypertension:**

A screenshot of a computer

AI-generated content may be incorrect.

**heart\_disease:**

A screenshot of a computer

AI-generated content may be incorrect.**smoking\_history:**

A screenshot of a computer

AI-generated content may be incorrect.

**Bmi:**

A screenshot of a computer

AI-generated content may be incorrect.

**HbA1c\_level:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**blood\_glucose\_level:**

A screenshot of a computer

AI-generated content may be incorrect.

**diabetes:**

A screenshot of a computer

AI-generated content may be incorrect.

**Adquisición de Datos**

Los datos fueron obtenidos mediante carga local, a partir de un archivo en formato CSV descargado desde la plataforma Kaggle. Esta fuente proporcionó un conjunto de datos estructurado, listo para su procesamiento y análisis.

**Herramientas utilizadas:**

* **Pandas:** Utilizado para la lectura del archivo CSV, así como para la limpieza, transformación y preparación de los datos antes del entrenamiento de los modelos.
* **Matplotlib:** Empleado en la elaboración de gráficos exploratorios como diagramas de barras, diagramas de cajas y curvas ROC, con el fin de facilitar el análisis visual de las variables.
* **Seaborn:** Aplicado en la representación gráfica de matrices de correlación, lo que permitió identificar relaciones entre variables.
* **Scikit-learn (sklearn):** Herramienta principal para el entrenamiento de los modelos predictivos. También se utilizó para el cálculo de métricas de evaluación (precisión, sensibilidad, exactitud, puntuación F1), visualización de árboles de decisión, generación de matrices de confusión y curvas ROC. Adicionalmente, fue empleada en el proceso de *feature engineering*, que incluyó la normalización de variables numéricas, la codificación de variables categóricas y la selección de atributos relevantes para mejorar el desempeño de los modelos.

**Problemas encontrados y soluciones:**

Uno de los principales problemas se presentó en la variable smoking\_history, la cual contenía seis categorías: never, former, current, ever, no\_info y not current. La categoría no\_info representaba aproximadamente el 36 % de los registros de la columna. Debido a esta proporción, aplicar un procedimiento de eliminación directa mediante dropna() habría reducido significativamente el tamaño del conjunto de datos, comprometiendo la validez del análisis.

Por último smoking\_history no es una variable comúnmente tomada en cuenta al momento de realizar un analisis de diabetes, es decir no es tan determinante en el diagnostico como si lo son otras. Por estas razones, se decidió eliminar esta variable del conjunto de datos para preservar la integridad del dataset y evitar introducir ruido en el proceso de entrenamiento.

**Preprocesamiento y Limpieza**

El conjunto de datos fue sometido a una revisión exhaustiva para garantizar su calidad antes del entrenamiento de los modelos predictivos. Las etapas principales del preprocesamiento incluyeron el manejo de valores nulos, la normalización de variables numéricas, la codificación de variables categóricas y la aplicación de transformaciones adicionales justificadas.

**Manejo de valores nulos**

Inicialmente, se verificó la presencia de valores nulos explícitos mediante el conteo estándar de celdas vacías. No se detectaron valores nulos en ninguna columna. Sin embargo, se realizó una inspección adicional en las variables bmi, HbA1c\_level, blood\_glucose\_level y age, ya que en estas columnas un valor de cero puede interpretarse como ausente o anómalo. El conteo de ceros en dichas columnas resultó ser cero, lo que permitió concluir que el conjunto de datos se encontraba completamente limpio para su uso.

**Normalización de variables numéricas**

Se aplicó la técnica de **MinMaxScaler** para escalar las variables numéricas continuas. Esta decisión se fundamentó en la necesidad de homogeneizar las escalas de entrada, especialmente para algoritmos sensibles a la magnitud de los datos. Las variables normalizadas fueron:

* age: por su amplio rango (0–80+).
* bmi: por su dispersión significativa (17–35).
* HbA1c\_level: a pesar de su rango estrecho, se incluyó para evitar sesgos.
* blood\_glucose\_level: por su alta variabilidad (80–158).

Las variables categóricas o binarias (gender, hypertension, heart\_disease, diabetes) no fueron normalizadas, ya que no lo requerían por su naturaleza.

**Codificación de variables categóricas**

Se aplicaron dos transformaciones para categorizar variables continuas:

* age\_group: se creó una nueva variable que agrupa la edad en rangos de 10 años.
* glucose\_group: se categorizó el nivel de glucosa en sangre en intervalos de 25 unidades, hasta un máximo de 400.

Estas variables facilitaron el análisis exploratorio y la evaluación del comportamiento del modelo en diferentes segmentos de datos.

Adicionalmente, la variable gender fue transformada a formato numérico: se asignó el valor **0** a Female y **1** a Male.

**Filtrado de outliers**

Se realizó un análisis estadístico y visual (boxplots) para identificar valores atípicos. No se encontraron outliers relevantes en las variables numéricas, por lo que no se aplicaron filtros ni eliminaciones en esta etapa.

**Transformaciones aplicadas y justificación**

Las transformaciones se enfocaron en:

* Escalar variables numéricas para asegurar convergencia y estabilidad en los modelos.
* Codificar correctamente las variables categóricas y crear rangos útiles para análisis segmentado.
* Preparar los datos en un formato compatible con algoritmos de aprendizaje supervisado, sin pérdida significativa de información.

**Análisis Exploratorio de Datos (EDA)**

**Estadísticas descriptivas**

Se analizaron las variables numéricas continuas (age, bmi, HbA1c\_level, blood\_glucose\_level) mediante gráficos de cajas y bigotes para evaluar su distribución, identificar valores extremos y estimar la dispersión de los datos.

* Edad (age): La distribución fue simétrica, con una mediana cercana a los 45 años. No se detectaron valores atípicos.

A diagram of a graph

AI-generated content may be incorrect.

* Índice de masa corporal (bmi): La variable mostró una gran cantidad de valores atípicos por encima del rango esperado (superior a 40). Estos datos fueron conservados, ya que reflejan variaciones clínicas reales.

A diagram of a diagram

AI-generated content may be incorrect.

* Nivel de hemoglobina glicosilada (HbA1c\_level): Presentó una distribución concentrada entre 4.0 y 6.5, con algunos valores atípicos moderados por encima de 8.

A diagram of a graph

AI-generated content may be incorrect.

* Nivel de glucosa en sangre (blood\_glucose\_level): Mostró una dispersión amplia, con valores atípicos en el extremo superior que reflejan hiperglucemias propias de estados diabéticos.

A diagram of a level of glucose

AI-generated content may be incorrect.

**Visualizaciones clave**

* Distribución del diagnóstico por género: Se construyó una gráfica de barras que muestra el porcentaje de pacientes diagnosticados y no diagnosticados con diabetes según el género. La proporción de casos con diabetes fue similar entre hombres y mujeres, mientras que la categoría "Other" no presentó incidencia significativa.

A graph of diabetes

AI-generated content may be incorrect.

* Relación entre edad y diabetes: Se elaboró un gráfico que representa el porcentaje de pacientes con diabetes en función de rangos de edad de 10 años. La incidencia de la enfermedad se incrementó progresivamente a partir de los 40 años, alcanzando su punto máximo entre los 70 y 80 años.

A graph of diabetes

AI-generated content may be incorrect.

* Porcentaje de diabetes según niveles de glucosa: Se categorizó el nivel de glucosa en sangre en intervalos de 25 unidades. Se observó una marcada concentración de casos con diabetes en rangos superiores a los 225 mg/dL, con un 100 % de incidencia en dichos valores.

A graph with green bars

AI-generated content may be incorrect.

* Distribución de HbA1c: Se representó gráficamente el porcentaje de pacientes con diabetes según los rangos de nivel de hemoglobina glicosilada (HbA1c). Se evidenció una clara relación entre valores elevados de HbA1c y la probabilidad de diagnóstico de diabetes.

A graph with orange bars

AI-generated content may be incorrect.

* Matriz de correlación: Se generó una matriz de correlación entre las variables numéricas. Las variables blood\_glucose\_level y HbA1c\_level presentaron los coeficientes más altos de correlación con la variable objetivo diabetes (0.42 y 0.40 respectivamente), seguidas por age (0.26).

A screenshot of a computer

AI-generated content may be incorrect.

**Principales patrones observados**

**El análisis exploratorio reveló tres patrones relevantes:**

1. La edad, el nivel de glucosa en sangre y la hemoglobina glicosilada influyeron significativamente en la probabilidad de desarrollar diabetes.
2. La mayoría de los casos diagnosticados se concentraron en grupos etarios mayores y en pacientes con desbalances metabólicos evidentes.
3. Variables demográficas como el género no mostraron influencia directa significativa en el diagnóstico.

**Feature Engineering**

Durante el desarrollo del modelo predictivo se implementaron distintas estrategias de *feature engineering* con el objetivo de mejorar la representación de los datos, facilitar el entrenamiento de los modelos y optimizar su rendimiento.

**Nuevas variables creadas**

* **age\_group**: Se generó una variable categórica que agrupa la edad de los pacientes en rangos de 10 años. Esta transformación permitió realizar un análisis segmentado de la distribución de la diabetes según grupos etarios.
* **glucose\_group**: Se creó una variable categórica que clasifica los niveles de glucosa en sangre en intervalos de 25 unidades, hasta un máximo de 400 mg/dL. Esta variable fue útil en el análisis exploratorio para identificar umbrales críticos de riesgo.

Ambas variables fueron utilizadas con fines descriptivos y visuales en el análisis exploratorio de datos (EDA), pero no fueron incluidas en el entrenamiento de los modelos debido a su naturaleza redundante con respecto a las variables continuas originales.

**Transformaciones aplicadas**

* Se aplicó la técnica de **MinMaxScaler** para normalizar las variables numéricas continuas: age, bmi, HbA1c\_level y blood\_glucose\_level. Esta técnica transformó los valores al rango [0, 1], lo cual fue particularmente útil para modelos sensibles a la escala como regresión logística, máquinas de soporte vectorial (SVC).
* La variable gender fue transformada a formato numérico. Se asignó el valor **0** a Female y **1** a Male, con el objetivo de integrarla en los modelos que requieren entradas numéricas.

**Selección y eliminación de atributos irrelevantes**

* La columna smoking\_history fue eliminada del conjunto de datos. Esta decisión se fundamentó en que:
  + Presentaba un **36 % de valores con la categoría no\_info**, lo que impedía una imputación confiable.
  + Su correlación con la variable objetivo fue baja.
  + Su eliminación permitió conservar el tamaño del conjunto de datos sin introducir ruido o incertidumbre innecesaria en el modelo.

**Modelado y Evaluación**

**Tipo de modelos utilizados**

Se evaluaron cinco algoritmos de clasificación binaria supervisada:

* **Logistic Regression**
* **Decision Tree**
* **Random Forest**
* **Support Vector Machine (SVC)**
* **Naive Bayes (GaussianNB)**

Todos los modelos fueron entrenados con class\_weight='balanced' para mitigar el desbalance en la variable objetivo (diabetes)

**División de los datos**

Se seleccionaron las siguientes variables predictoras:  
['gender', 'age', 'hypertension', 'heart\_disease', 'bmi', 'HbA1c\_level', 'blood\_glucose\_level']  
La variable objetivo fue diabetes. El conjunto de datos se dividió en un 80 % para entrenamiento y un 20 % para prueba, empleando random\_state=369 para garantizar reproducibilidad.

**Métricas de evaluación**

Se utilizaron las siguientes métricas estándar:

* **Accuracy (exactitud)**: proporción de predicciones correctas.
* **Precision**: proporción de verdaderos positivos respecto a los positivos predichos.
* **Recall (sensibilidad)**: proporción de verdaderos positivos respecto a los positivos reales.
* **F1-score**: media armónica entre precisión y recall.
* **Matriz de confusión**: para identificar errores por clase.

**MODELOS**

**Modelo de Regresión Logística**

**Resultados**

* **Accuracy**: 88.66 %
* El modelo mostró una **alta precisión para la clase 0 (sin diabetes)** y una **recall notable para la clase 1 (con diabetes)**, aunque con menor precisión.

A white background with black text

AI-generated content may be incorrect.

A screenshot of a phone

AI-generated content may be incorrect.

**Modelo Árbol de Decisión**

Se evaluó el rendimiento del modelo variando el parámetro max\_depth entre 2 y 40. Se observó una mejora progresiva hasta estabilizarse alrededor de 7, punto en el que se obtuvo un buen equilibrio entre sobreajuste y generalización.

A blue screen with white text

AI-generated content may be incorrect.

Tambien se analizó el árbol en otros casos como por ejemplo max\_depth = 15 y 2

A screenshot of a computer

AI-generated content may be incorrect.

Al final se optó por un max\_depth = 7. El proyecto sigue un enfoque científico por lo que disminuir la cantidad de Falsos Negativos (FN) es crucial.

**Resultados del Árbol de Decisión con max\_depth = 7:**

* **Accuracy**: 88.85 %
* Mejoró el F1-score de la clase positiva (diabetes = 1) respecto a la regresión logística.

A white background with black text

AI-generated content may be incorrect.

A screenshot of a phone

AI-generated content may be incorrect.

**Modelo SVC y kernel selection**

Se evaluaron cuatro kernels: linear, poly, rbf y sigmoid:

A line graph with different lines

AI-generated content may be incorrect.  
Las mejores opciones eran; linear, poly y rbf. Se opto por linear por su rapidez de procesamiento:

A close-up of a white background

AI-generated content may be incorrect.

El modelo tuvo una accuracy del: **0.8852%**

**Naive Bayes (GaussianNB)**

Este modelo simple logró un **accuracy del 90.16 %**, con buen desempeño general, aunque con menor recall en comparación con Random Forest y Decision Tree.

A screenshot of a computer

AI-generated content may be incorrect.

**Comparacion de modelos y resultados**

A screenshot of a graph

AI-generated content may be incorrect.

En este caso solo escogeremos Logistic Regression, Decision Tree y Random Forest.

Dejaremos de lado SVC y Naive Bayes por que contienen muchos FN y desproporcionado FP:

A blue screen with white text

AI-generated content may be incorrect.**Naive Bayes SVC (linear)**

A blue screen with white text

AI-generated content may be incorrect.

Extraemos las accuracy de los modelos Logistic Regression, Decision Tree y Random Forest.

A screenshot of a computer screen

AI-generated content may be incorrect.

**Analisis Individual**

**Logistic Regression:**

A screenshot of a computer program

AI-generated content may be incorrect.

**Decision Tree:**

A screenshot of a computer program

AI-generated content may be incorrect.

**RandomForest:**

A screenshot of a computer program

AI-generated content may be incorrect.

Descartamos Logistic regresión, pq tiene muchos FN (213) y FP(2055), por lo que tenemos dos contendientes RandomForest o DecisionTree:

RandomForest cuenta con un acuraccy de 96% pero tiene 502 FN.

Por otro lado, tenemos DecisionTree, este tiene un acuraccy del 88% pero solo 148 FN lo cual es bastante bueno.

**Como queremos reducir los FN, nuestro modelo ganador es DecisionTree.**

**Matriz de confusión Decisión Tree Classifier:**

A graph with numbers and a number on it

AI-generated content may be incorrect.

**Classification Report:**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Curva ROC:**

A graph with a red line

AI-generated content may be incorrect.

**Análisis de la Curva ROC**

La curva ROC presenta una trayectoria que se eleva de forma pronunciada hacia la esquina superior izquierda del gráfico y luego se estabiliza cerca de un valor de sensibilidad (TPR) igual a 1. Esta forma sugiere que el modelo posee una alta capacidad para discriminar entre clases positivas y negativas.

En el gráfico, la curva ROC se mantiene claramente por encima de la diagonal representada por la línea roja. Esta línea corresponde al rendimiento esperado de un clasificador aleatorio. La posición de la curva respecto a esta referencia indica que el modelo supera con amplitud el desempeño de una clasificación sin inteligencia.

Aunque el valor exacto del área bajo la curva (AUC) no se encuentra representado numéricamente en la gráfica, la forma de la curva permite inferir que el AUC se aproxima a 0.97 o 0.98. Este valor se interpreta como un excelente nivel de rendimiento. Un AUC igual a 1.0 describe un modelo perfecto, mientras que un valor superior a 0.9 se considera altamente satisfactorio.

**Conclusión:**  
El modelo evaluado mediante esta curva ROC muestra una capacidad significativa para distinguir entre las clases del problema. Se caracteriza por una alta sensibilidad, una baja tasa de falsos positivos y un comportamiento general que refleja un desempeño predictivo sobresaliente.

**Variables influyentes para el modelo:**

A graph with blue bars

AI-generated content may be incorrect.