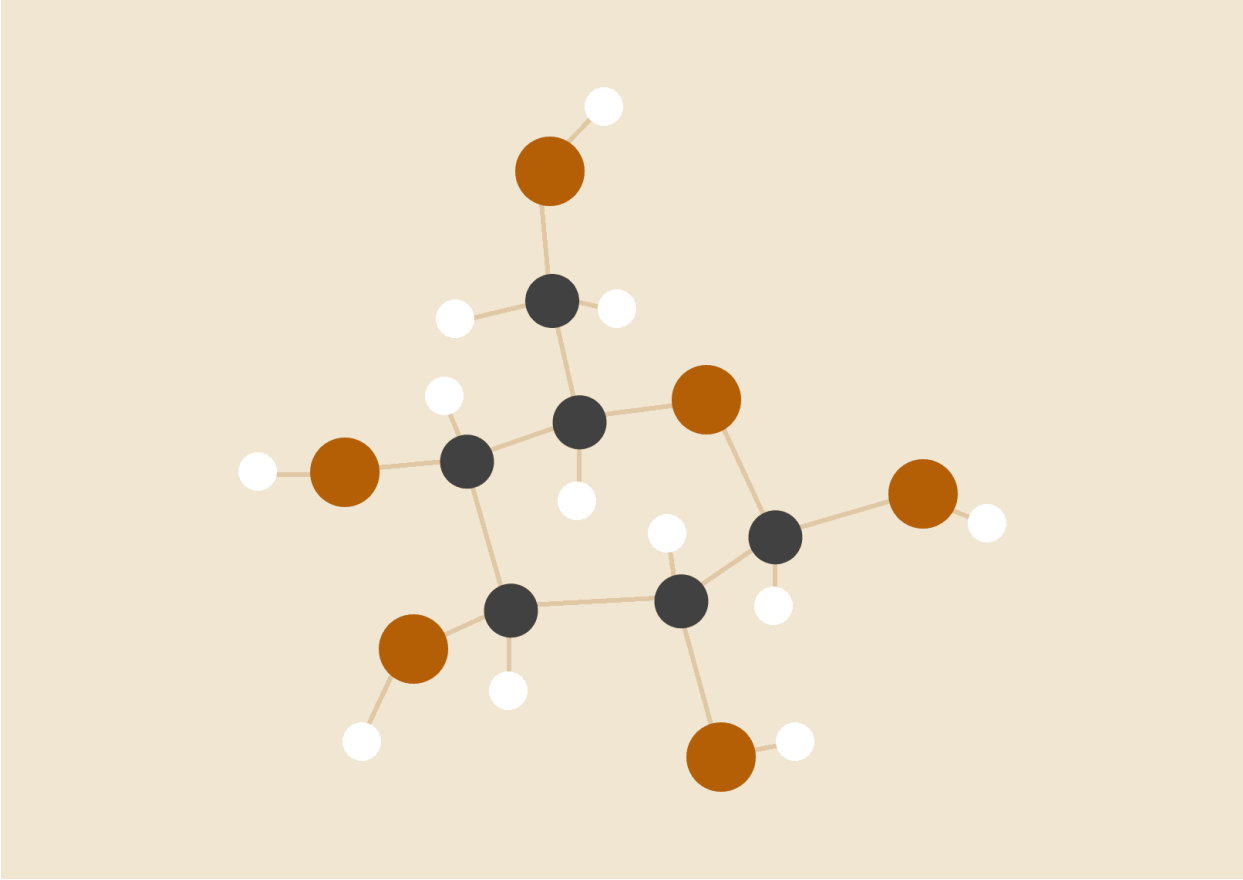
Predicción de Diabetes Tipo 2

CHALLENGE - Creación de un flujo de trabajo en KNIME



Damaris Ayala Zavala

27 de junio de 2025

BEDU

Descripción del Proyecto

La detección temprana del riesgo de desarrollar **diabetes tipo 2** es crucial para prevenir complicaciones de salud. Este proyecto busca **predecir si una persona tiene alto riesgo** de padecer diabetes tipo 2, basándose en variables clínicas, demográficas y de estilo de vida.

Este es un problema **de clasificación binaria**, y los resultados pueden usarse para apoyar decisiones clínicas o educativas en prevención.

Conjunto de datos

El conjunto de datos contiene variables recolectadas en un estudio clínico simulado. Las columnas más relevantes incluyen:

- Edad
- IMC (Índice de Masa Corporal)
- Niveles de glucosa
- Presión sanguínea
- Actividad física
- Hábitos dietéticos
- Tabaquismo
- Consumo de alcohol
- Salud pancreática
- Diagnóstico anterior de diabetes gestacional
- Tipo de embarazo
- Análisis de orina
- Entre otras.

Se creó la columna `riesgo_diabetes` que etiqueta a una persona como en **riesgo alto (1)** si tiene **niveles de glucosa ≥ 140 mg/dL** o **intolerancia a la glucosa positiva**.

SELECCIÓN DE CARACTERÍSTICAS

Variables seleccionadas

De todas las columnas disponibles, se conservaron las que tienen **mejor calidad de datos y relevancia clínica**:

Conservadas: edad, imc, antecedentes_familiares, niveles_insulina, actividad_fisica, habitos_dieteticos, presion_sanguinea, niveles_colesterol, talla, niveles_glucosa, tabaquismo, consumo_alcohol, tolerancia_glucosa, salud_pancreatica

DOCUMENTACIÓN

1. Carga de datos

- Nodo: CSV Reader
- Se leyó el archivo `diabetes.csv`.

2. Limpieza y preparación

- Nodo: Column Rename (Regex)
  Para normalizar nombres de columnas eliminando tildes, espacios y errores tipográficos.
- Nodo: Column Filter
  Se conservaron solo las variables más relevantes.
- Nodo: Missing Value
  Se eliminaron o imputaron registros con datos nulos.

3. Generación de variable objetivo

- Nodo: Rule Engine
  Se creó la columna `riesgo_diabetes`:

```
$niveles_glucosa$ >= 140 => 1

$tolerancia_glucosa$ = "Abnormal" => 1

TRUE => 0
```

4. Partición del conjunto de datos

- Nodo: Partitioning
    - 70 % entrenamiento
    - 30 % prueba
    - Modo: Stratified sampling por `riesgo_diabetes`.

La variable objetivo fue `riesgo_diabetes`, definida como 1 si el paciente presenta niveles de glucosa ≥ 140 mg/dL o una condición de tolerancia a la glucosa anormal, y 0 en caso contrario.

5. Entrenamiento del modelo

- Nodo: Tree Ensemble Learner
    - Elegido por su robustez con variables mixtas (numéricas y categóricas).
    - Configurado para generar probabilidades de predicción.

Para predecir el riesgo de diabetes en pacientes, se utilizó el nodo **Tree Ensemble Learner**, que implementa un modelo de ensamble de árboles de decisión (tipo Random Forest). Este modelo fue seleccionado por su alta capacidad para trabajar con datos mixtos (numéricos y categóricos), su robustez ante valores atípicos y su buen rendimiento en tareas de clasificación binaria.
Se incluyeron como características predictoras factores clínicos, demográficos y de estilo de vida, como: edad, IMC, niveles de insulina, actividad física, tabaquismo, entre otros.

6. Predicción

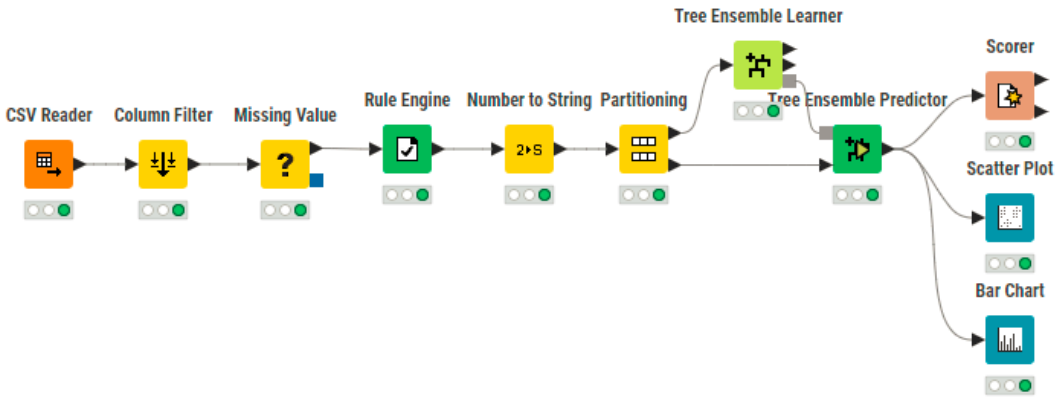
- Nodo: Tree Ensemble Predictor
    - Aplicado al conjunto de prueba.

7. Evaluación del modelo

- Nodo: Scorer
    - compara los valores reales de `riesgo_diabetes` con las predicciones del modelo (`prediction(riesgo_diabetes)`).

9. Visualización final

- Nodo: Scatter Plot
    - En esta visualización tipo *scatter plot*, se observa la relación entre el índice de masa corporal (IMC) y los niveles de insulina.
    - El modelo tiende a clasificar como riesgo de diabetes a pacientes que se ubican en la región de **IMC elevado e insulina alta**.
- Nodo: Bar Chart
    - Compara los **promedios de distintas variables numéricas** agrupadas por la **predicción de riesgo de diabetes** (`0` = sin riesgo, `1` = con riesgo).
    - Esta gráfica permite identificar **diferencias clave** entre ambos grupos. Por ejemplo, se observa que los pacientes clasificados como de **alto riesgo** tienden a tener **mayores niveles de glucosa, IMC e insulina**, lo cual es coherente con los factores comúnmente asociados a la diabetes.



Rows: 2 | Columns: 14

Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Mean	Mean Absolu...	Standard Dev...	Sum	10 most common values
1	Numbe	0	2	2	141	2	71.5	141	71.5	69.5	98.288	143	2 (1; 50.0%), 141 (1; 50.0%)
0	Numbe	0	2	0	37	0	18.5	37	18.5	18.5	26.163	37	0 (1; 50.0%), 37 (1; 50.0%)

RESULTADOS Y CONCLUSIONES

Precisión del modelo:

- El modelo de *Tree Ensemble (Random Forest)* logró una **alta precisión en la clasificación** del riesgo de diabetes.
- A través del nodo **Scorer**, se observó una correcta clasificación de la mayoría de los casos, separando adecuadamente entre pacientes en riesgo (`1`) y sin riesgo (`0`).

Importancia de las variables:

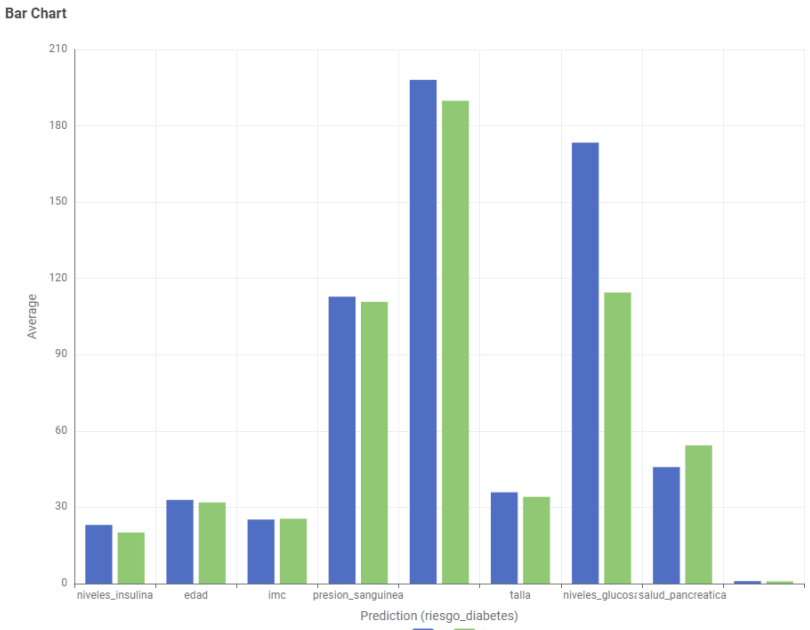
Las variables con mayor influencia en el modelo (evaluadas con *Variable Importance*) fueron:

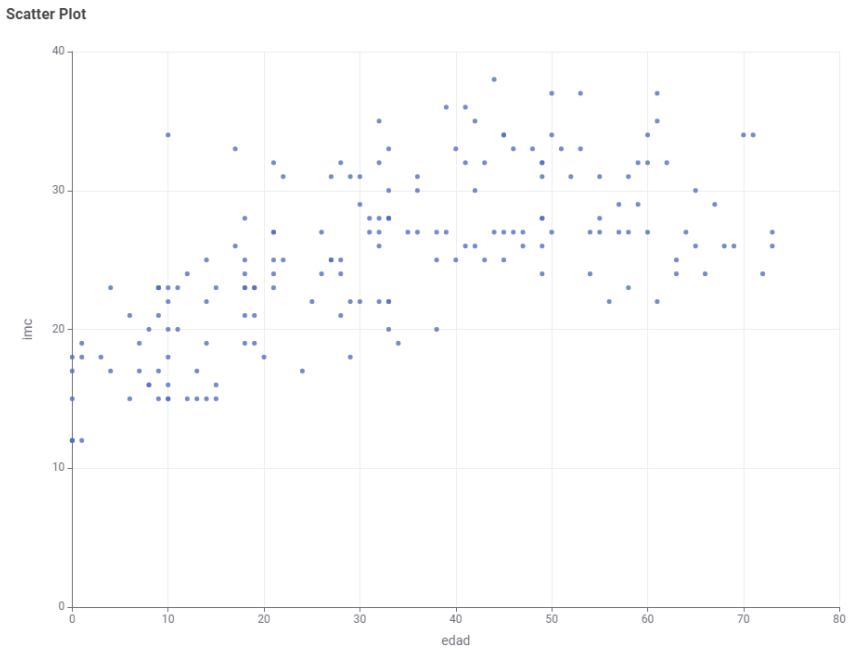
  - **Niveles de glucosa**
  - **IMC**
  - **Edad**
  - **Niveles de insulina**
  - **Tolerancia a la glucosa**

- Estas variables coinciden con factores clínicos ampliamente documentados como predictores de diabetes tipo 2.

Visualizaciones:

- **Scatter Plot (IMC vs Niveles de Insulina):** mostró que los pacientes con **IMC alto e insulina elevada** fueron mayormente clasificados como en **riesgo de diabetes**, lo que valida visualmente las predicciones del modelo.
- **Bar Chart por predicción de riesgo:** evidenció que el grupo clasificado como en riesgo (`1`) presenta consistentemente **promedios más altos en variables clave** como glucosa, IMC y colesterol.





CONCLUSIONES

- Se construyó con éxito un modelo predictivo de clasificación binaria para identificar el **riesgo de diabetes tipo 2**, utilizando datos clínicos y de estilo de vida.

- El modelo mostró **buen rendimiento** y está fundamentado en **variables médicas relevantes**, lo que sugiere su potencial para integrarse como **herramienta de apoyo clínico** en la toma de decisiones preventivas.

- Las visualizaciones confirmaron los hallazgos del modelo, permitiendo **interpretaciones claras** y útiles para personal médico, educadores en salud o responsables de programas de prevención.

- Este flujo de trabajo en KNIME demuestra cómo es posible implementar un pipeline completo de Machine Learning **sin necesidad de codificación**, desde la limpieza de datos hasta la evaluación del modelo y su interpretación visual.