



## **Universidad Autónoma Metropolitana**

Departamento de Ciencias Naturales e Ingeniería

Licenciatura en Ingeniería en Computación

Datos a Gran Escala (BIG DATA)

Proyecto: Predicción de Diagnóstico de Diabetes.

Integrantes:

García Núñez Rodrigo - 2203025158  
Cortes Lopez Alan Yair - 2203066542  
Chávez Flores Alejandro - 2203024955  
Ortega García Rodolfo André - 2203066275

## Introducción

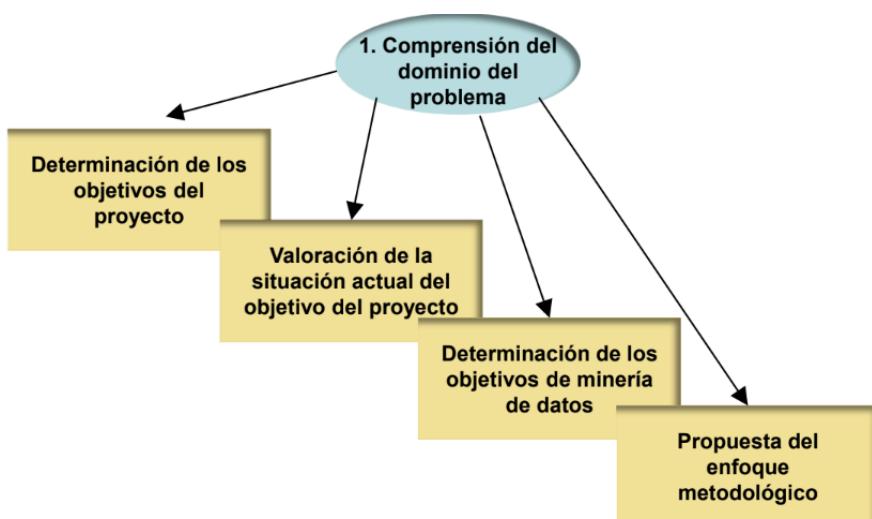
La diabetes es una enfermedad crónica que se caracteriza por niveles elevados de glucosa en la sangre debido a la incapacidad del cuerpo para producir o utilizar adecuadamente la insulina, una hormona que regula el azúcar en la sangre. En México, la diabetes es un problema de salud pública significativo y ha alcanzado proporciones alarmantes en las últimas décadas.

Según datos de la Secretaría de Salud de México, la prevalencia de la diabetes ha aumentado de manera constante en el país. Se estima que alrededor del 10% de la población mexicana vive con diabetes, y el número de casos sigue en aumento. Además, la diabetes tipo 2, que está fuertemente relacionada con el estilo de vida y la dieta, es la forma más común de diabetes en México.

Factores como la dieta rica en calorías y carbohidratos, la falta de actividad física, la obesidad y la predisposición genética contribuyen al aumento de la diabetes en la población mexicana. Además, el acceso desigual a la atención médica y la falta de conciencia sobre la prevención y el manejo de la diabetes son desafíos adicionales.

Si bien existen diferentes tipos de diabetes, la diabetes tipo II es la forma más común y su prevalencia varía según la edad, la educación, los ingresos, la ubicación, la raza y otros determinantes sociales de la salud. Gran parte de la carga de la enfermedad recae también en las personas de nivel socioeconómico más bajo.

## Fase 1: Comprensión del dominio del problema.



## Determinación del objetivo del proyecto.

En este proyecto, pretendemos desentrañar la relación existente entre el estilo de vida de las personas y factores médicos respecto a la aparición de Diabetes, en base a datos obtenidos por el CDC (Centro de Control y prevención de Enfermedades) en Estados Unidos.

Pretendemos predecir con un cierto nivel de exactitud si una persona sufrirá de un infarto relacionado a su diagnóstico de Diabetes, respecto a su estilo de vida y a sus datos médicos.

Asimismo, como meta adicional, pretendemos categorizar el diagnóstico de Diabetes respecto a los atributos contenidos en el DataSet.

## Valoración de los objetivos del proyecto

- ¿Se comprende de forma clara el problema que se intenta abordar?

Sí, se comprende. *En este proyecto pretendemos determinar varios modelos predictivos y categóricos que nos permitan predecir y categorizar el diagnóstico de Diabetes de las personas en base a datos que representen su estilo de vida y datos médicos.*

*Para comprender el problema que se quiere atender, es necesario tener en cuenta algunos conceptos sobre estilo de vida y salud presentes en el DataSet. Por lo siguiente a se puede utilizar el siguiente diccionario.*

## Diccionario.

Término	Definición
Diabetes	Grupo de enfermedades que afecta la forma en que el cuerpo utiliza la glucosa en la sangre.
Glucosa	Importante fuente de energía para las células que forman los músculos y tejidos. También es la principal fuente de combustible del cerebro.

Presión Sanguínea. - BP	Tensión ejercida por la sangre que circula por las paredes de los vasos sanguíneos.
Colesterol - Chol	Sustancia cerosa similar a la grasa implicada en la producción de hormonas, vitamina D y sustancias digestivas. Un exceso de colesterol en la sangre puede combinarse con otras sustancias y pegarse en las paredes de los vasos sanguíneos.
Índice de masa corporal - BMI	Es el peso de una persona en kilogramos dividido por el cuadrado de su estatura en metros. Se utiliza como método de evaluación para la categoría de peso: bajo peso, peso saludable, sobrepeso, y obesidad.
Accidente Cerebrovascular - Stroke	Ocurre cuando el flujo de sangre que debe llegar al cerebro se detiene por varios segundos o bien, cuando existe un derrame de sangre en el cerebro o alrededor del mismo.
Ataque cardiaco - HDA	Ocurre cuando el flujo de sangre a una parte del músculo cardíaco se bloquea.
Actividad Física	Cualquier actividad o ejercicio que involucra un gasto de energía y fenómenos a nivel corporal, psíquico y emocional en la persona que lo realiza.
Consumo de alcohol importante.	Consumo no saludable de alcohol que pone en riesgo la salud o seguridad. Se considera cuando un hombre toma 5 o más bebidas en 2 horas o si una mujer toma 4 bebidas en 2 horas.
HealthCare	Cualquier tipo de asistencia de salud, como seguros, planes, etc.

Salud Mental	Engloba el estado de equilibrio y bienestar de nuestro ser emocional, psicológico y social.
Salud Física	Estado de bienestar donde el cuerpo funciona de manera óptima. Se toma en cuenta la ausencia o presencia de enfermedades y heridas que tengan alguna repercusión en el funcionamiento fisiológico del organismo.
Sexo	Categoría taxonómica que clasifica entre seres vivos con aparatos reproductores femeninos o masculinos.

- ¿Existen datos disponibles para efectuar el análisis?

Sí, se tiene un DataSet lo suficientemente grande. En él se encuentran los datos suficientes que describen el estilo de vida y estado de salud de las personas participantes en el estudio. En el DataSet se encuentran 253680 tuplas, de las cuales sólo se utilizarán 25000 tuplas escogidas de manera aleatoria y sin repetición, y 21 atributos que serán de ayuda para predecir si una persona desarrollará Diabetes respecto a su estilo de vida.

- ¿Cuál es la fuente de esos datos y de qué tipo son?

La fuente de los datos son un grupo de personas en Estados Unidos, de los cuales el CDC recopiló sus datos referentes al estilo de vida y antecedentes médicos junto con sus diagnósticos de diabetes. La descripción de los datos y la tabla que los contiene pueden consultarse a través de la siguiente liga: [CDC Diabetes Health Indicators - UCI Machine Learning Repository](https://www.cdc.gov/diabetes/pdfs/info/basics/diabetes_health_indicators_UCI_Machine_Learning_Repository.pdf).

- ¿Se dispone de recursos humanos y tecnológicos para desarrollar el proyecto?

Sí, se cuenta con un equipo de trabajo para el desarrollo de este proyecto. Asimismo, se cuenta con el software necesario para realizar las diferentes actividades objetivo para la realización del proyecto. El software que se usará es IBM SPSS MODELER.

- ¿Se han identificado factores de riesgo que afecten el desarrollo del proyecto?

Sí, de los que se pueden destacar:

- *Falta de conocimiento sobre temas relacionados a la medicina. Particularmente, temas referentes al campo de estudio de la Diabetes.*
- *Falta de conocimiento de la herramienta IBM SPSS MODELER.*
- *Tiempo limitado para llevar a cabo las actividades objetivo en la realización de este proyecto.*
- *Cierto nivel de incompatibilidad en los horarios de los involucrados en este proyecto.*

## Determinación de los objetivos de minería de datos

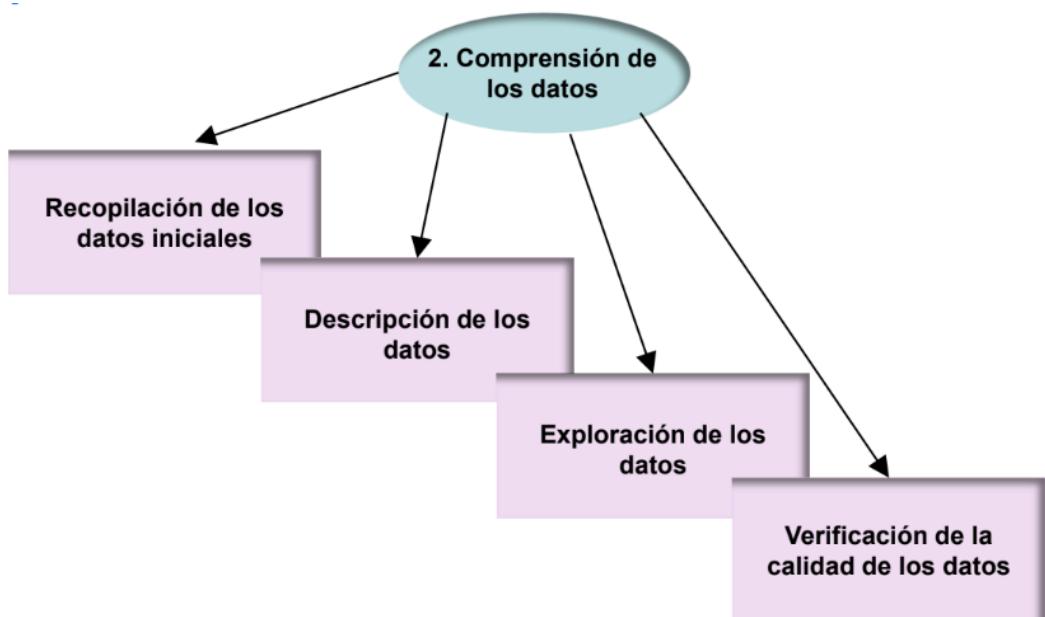
- Predicción/pronóstico: Construir varios modelos predictivos supervisados que permitan predecir la aparición de ciertos problemas de salud (como eventos cardiacos y/o cerebrovasculares) respecto a diagnóstico de Diabetes de una persona, a su estilo de vida y datos médicos. Por ejemplo :
  - Reconocer patrones de los distintos hábitos que a una persona le puedan provocar el desarrollo de esta enfermedad.
  - Construir un modelo de predicción de la enfermedad a través de los registros que se tienen en el dataset.
  - Clasificación de los pacientes si tienen o no tienen diabetes e incluso si son propensos a tener dicha enfermedad.
  - Predicción si es propenso o no a tener la enfermedad.
- Clasificación:
  - Construir modelos de clasificación que permitan clasificar el estado del diagnóstico de diabetes de un paciente (sin presencia de diabetes, prediabetes o presencia de diabetes)
  - Construir modelos de clasificación que permitan clasificar si un paciente tuvo un infarto respecto a su diagnóstico de Diabetes y datos que representan su estilo de vida.

## Propuesta de enfoque metodológico

Fase	Tiempo	Recursos Humanos y Tecnológicos	Riesgos Atribuibles
Comprensión del dominio del problema.	1 semana	-Equipo de desarrollo	Términos que tal vez no puedan ser entendidos debido a que los integrantes del equipo no cuentan con el conocimiento de Medicina.
Comprensión de los datos.	2 semanas	-Equipo de desarrollo  -Hojas de cálculo (Excel)  -IBM SPSS MODELER	Algunos de los atributos requieren un cierto nivel de conocimiento de medicina, lo que dificulta la comprensión de los datos por parte del equipo de desarrollo.
Preparación de los datos.	3 semanas	-Equipo de desarrollo  -IBM SPSS MODELER	Poca experiencia con el software IBM SPSS MODELER, lo que puede entorpecer un poco la fase de preparación de los datos.
Modelado	2 semanas	-Equipo de desarrollo  -IBM SPSS MODELER	Poca experiencia con el software IBM SPSS MODELER y una gran curva de aprendizaje en el área de la minería de datos, lo que puede entorpecer el modelado.
Evaluación	2 semanas	-Equipo de desarrollo	No se han identificado riesgos.

		-IBM SPSS MODELER	
Presentación	1 semana	-Equipo de desarrollo -IBM SPSS MODELER	No se han identificado riesgos.

## Fase 2: Comprensión de los datos.



### Recopilación de los datos iniciales

Se tiene una tabla compuesta por 21 atributos y 253680 registros. Datos obtenidos el 25/9/2023. Los datos corresponden a un DataSet público y descargado desde el siguiente enlace: [CDC Diabetes Health Indicators - UCI Machine Learning Repository](#).

Con esto podemos asegurar que los datos mostrados en el DataSet son confiables debido a que fue obtenido a través del url ya mencionado.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDisease	PhysActivity	Fruits	Veggies	HvyAlcoholC	AnyHealthcar	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk
2	0	1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15
3	0	0	0	0	25	1	0	0	1	0	0	0	1	3	0	0	0
4	0	1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30
5	0	1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0
6	0	1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0
7	0	1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2
8	0	1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14
9	0	1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0
10	2	1	1	1	30	1	0	1	0	1	1	0	1	0	5	30	30
11	0	0	0	1	24	0	0	0	0	0	1	0	1	0	2	0	0
12	2	0	0	1	25	1	0	0	1	1	1	0	1	0	3	0	0
13	0	1	1	1	34	1	0	0	0	1	1	0	1	0	3	0	30
14	0	0	0	1	26	1	0	0	0	1	0	1	0	3	0	15	
15	2	1	1	1	28	0	0	0	0	0	1	0	1	0	4	0	0
16	0	0	1	1	33	1	1	0	1	0	1	0	1	1	4	30	28
17	0	1	0	1	33	0	0	0	1	0	0	0	1	0	2	5	0
18	0	1	1	1	21	0	0	0	1	1	1	0	1	0	3	0	0
19	2	0	0	1	23	1	0	0	1	0	0	0	1	0	2	0	0
20	0	0	0	0	23	0	0	0	0	1	0	1	0	2	15	0	
21	0	0	1	1	28	0	0	0	0	0	1	1	0	2	10	0	
22	0	1	1	1	22	0	1	1	0	1	0	0	1	0	3	30	0
23	0	1	1	1	38	1	0	0	0	1	1	0	1	0	5	15	30
24	0	0	0	1	28	1	0	0	0	0	1	0	1	0	3	0	7
25	2	1	0	1	27	0	0	0	1	1	1	0	1	0	1	0	0
26	0	1	1	1	28	1	0	0	0	1	1	0	1	0	3	6	0
27	0	0	0	1	32	0	0	0	1	1	1	0	1	0	2	0	0
28	2	1	1	1	37	1	1	0	0	1	0	0	1	0	5	0	0
29	2	1	1	1	28	1	0	1	0	0	1	0	1	0	4	0	0

Imagen. Captura de pantalla del DataSet de Diabetes Health Indicators.

## Descripción de los datos

Dato (Atributo)	Tipo de Dato	Descripción	Entrada/Salida
ID	Entero	Identificación del paciente	-
Diabetes012	Entero	0 = no diabetes, 1 = pre-diabetes, 2= diabetes (este es el objetivo)	Salida
HighBP	Binario	0 = no presión arterial alta, 1 = presión arterial alta	Entrada
HighChol	Binario	0 = no colesterol alto, 1 = colesterol alto	Entrada
CholCheck	Binario	0 = no chequeos de colesterol en 5 años, 1 = chequeos de colesterol en 5 años	Entrada
BMI	Entero	Índice de masa corporal	Entrada
Smoker	Binario	¿Has fumado al menos 100 cigarrillos a lo largo de tu vida? [Nota: 5 cajetillas = 100 cigarrillos] 0 = no, 1 = sí	Entrada

Stroke	Binario	¿Alguna vez te dijeron que tuviste un derrame cerebral? 0 = no, 1 = sí	Entrada
HeartDisease orAttack	Binario	Enfermedad coronaria o ataque al corazón. 0 = no, 1 = sí	Entrada/Salida
PhysActivity	Binario	Actividad física en los últimos 30 días, excluyendo el trabajo. 0 = no, 1 = sí	Entrada
Fruits	Binario	Consumir frutas 1 o más veces al día. 0 = no, 1 = sí	Entrada
Veggies	Binario	Consumir verduras 1 o más veces al día. 0 = no, 1 = sí	Entrada
HvyAlcoholConsump	Binario	Consumo elevado de alcohol (hombres adultos que consumen más de 14 bebidas por semana y mujeres adultas que consumen más de 7 bebidas por semana). 0 = no, 1 = sí	Entrada
AnyHealthcare	Binario	¿Tienes algún tipo de cobertura de atención médica, incluido seguro médico, planes prepagos como HMO, etc.? 0 = no, 1 = sí	Entrada

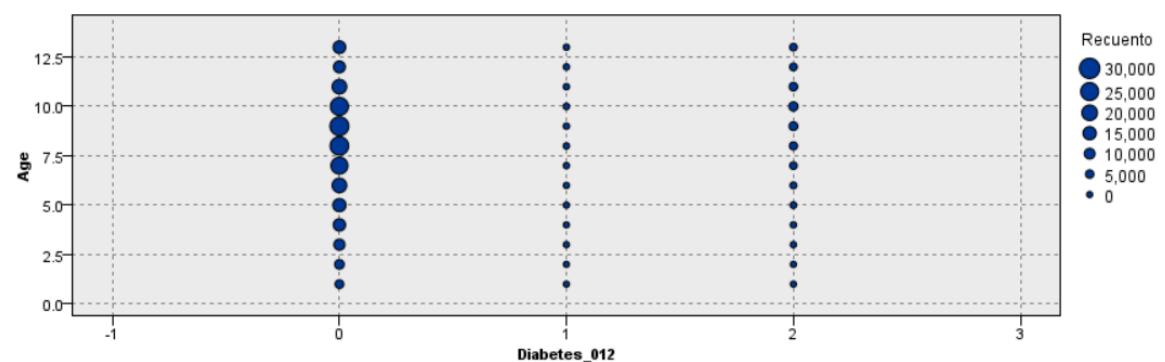
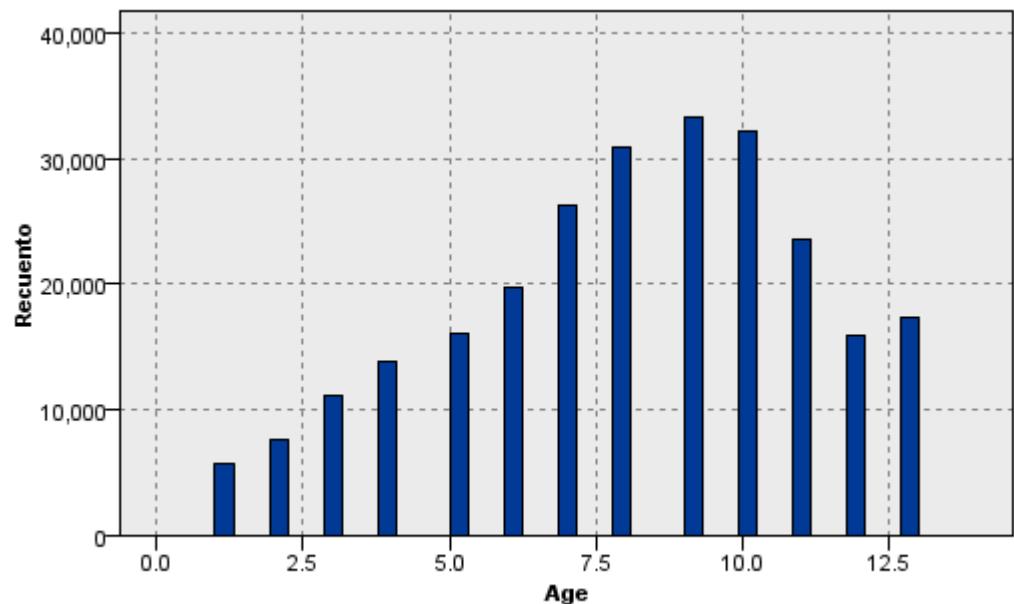
NoDocbcCost	Binario	¿Hubo un momento en los últimos 12 meses en que necesitaste ver a un médico pero no pudiste debido al costo? 0 = no, 1 = sí	Entrada
GenHlth	Entero	¿Dirías que en general tu salud es: escala 1-5? 1 = excelente, 2 = muy buena, 3 = buena, 4 = regular, 5 = mala	Entrada
MentHlth	Entero	Ahora, pensando en tu salud mental, que incluye estrés, depresión y problemas emocionales, ¿por cuántos días durante los últimos 30 días no fue buena tu salud mental? Escala 1-30 días	Entrada
PhysHlth	Entero	Ahora, pensando en tu salud física, que incluye enfermedades físicas y lesiones, ¿por cuántos días durante los últimos 30 días no fue buena tu salud física? Escala 1-30 días	Entrada
DiffWalk	Binario	¿Tienes dificultades serias para caminar o subir escaleras? 0 = no, 1 = sí	Entrada
Sex	Binario	0 = mujer, 1 = hombre	Entrada

Age	Entero	Categoría de edad en 13 niveles (_AGEG5YR, ver libro de códigos). 1 = 18-24, 9 = 60-64, 13 = 80 o más	Entrada
Education	Entero	Nivel educativo (EDUCA, ver libro de códigos). Escala 1-6: 1 = nunca asistió a la escuela o solo kindergarten, 2 = grados 1 al 8 (primaria), 3 = grados 9 al 11 (algo de secundaria), 4 = grado 12 o GED (graduado de secundaria), 5 = universidad 1 a 3 años (alguna universidad o escuela técnica), 6 = universidad 4 años o más (graduado universitario)	Entrada
Income	Entero	Escala de ingresos (INCOME2, ver libro de códigos). Escala 1-8: 1 = menos de \$10,000, 5 = menos de \$35,000, 8 = \$75,000 o más	Entrada

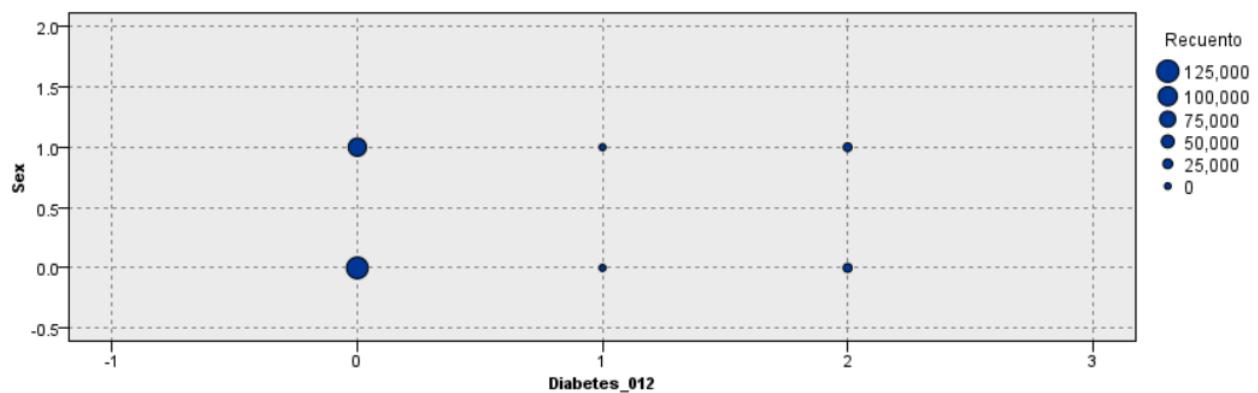
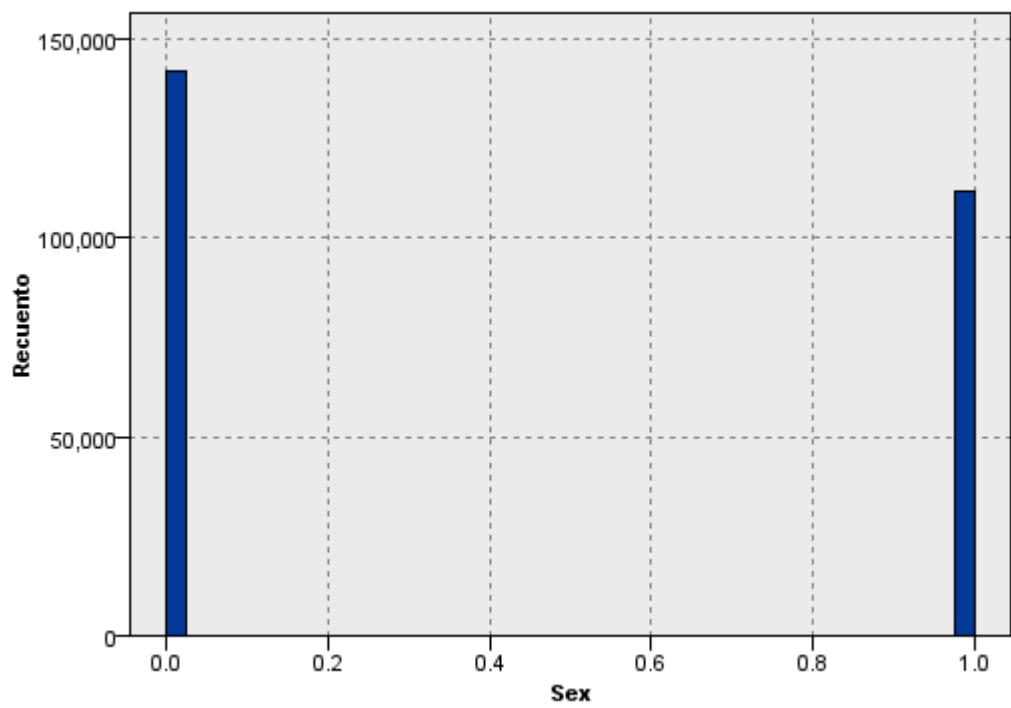
## Exploración de los datos

A continuación se modelaron todos los datos del dataset, con el fin de que se tener de una manera más visual que datos son los que nos ayudarán a lograr los objetivos de nuestra minería de datos.

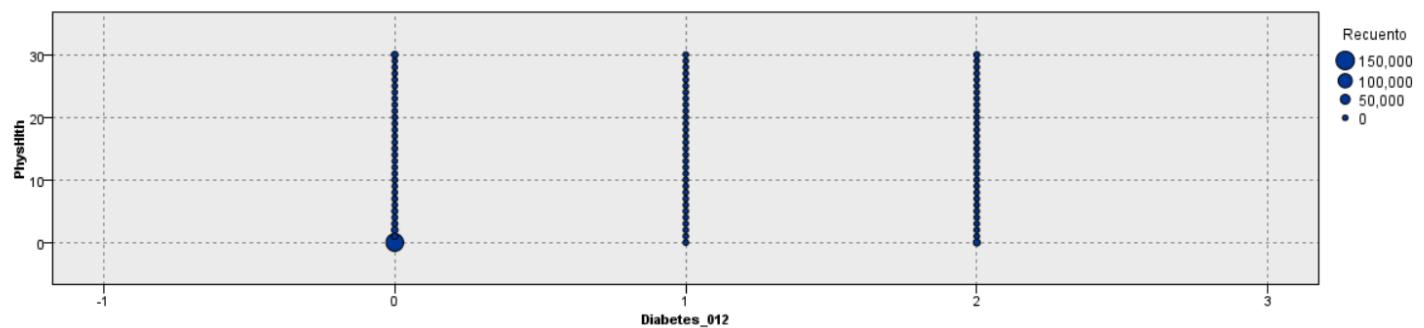
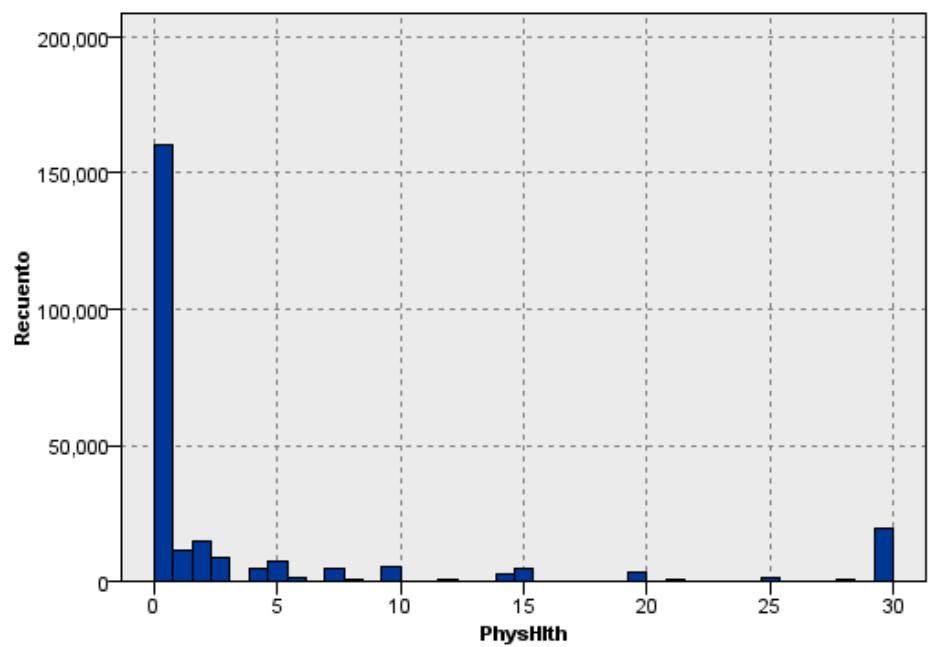
**Edad:** 1 = 18-24, 2 = 25-29, 3 = 30-34, 4 = 35-39, 5 = 40-44, 6 = 45-49, 7 = 50-54, 8 = 55-60, 9 = 60-64, 10 = 65-69, 11 = 70-74, 12 = 75 -80, 13 = 80 o más



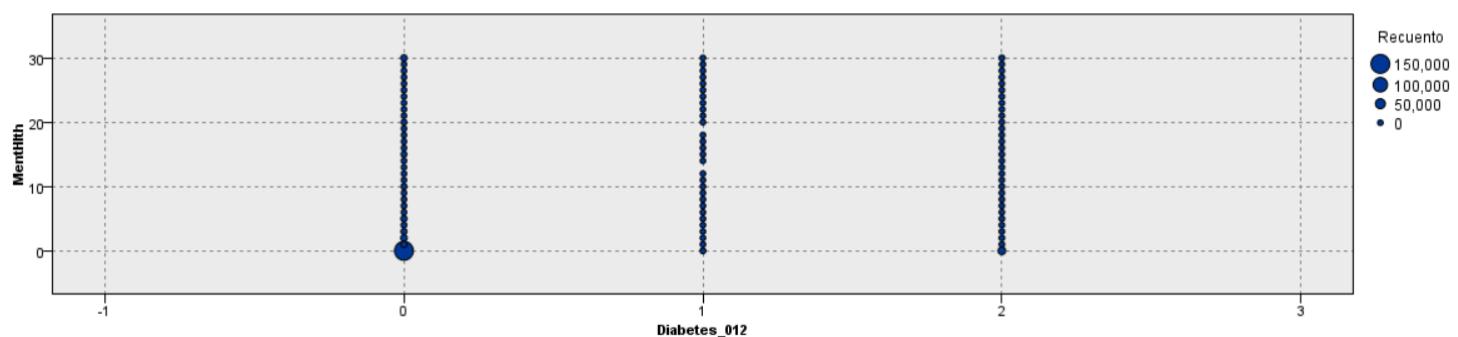
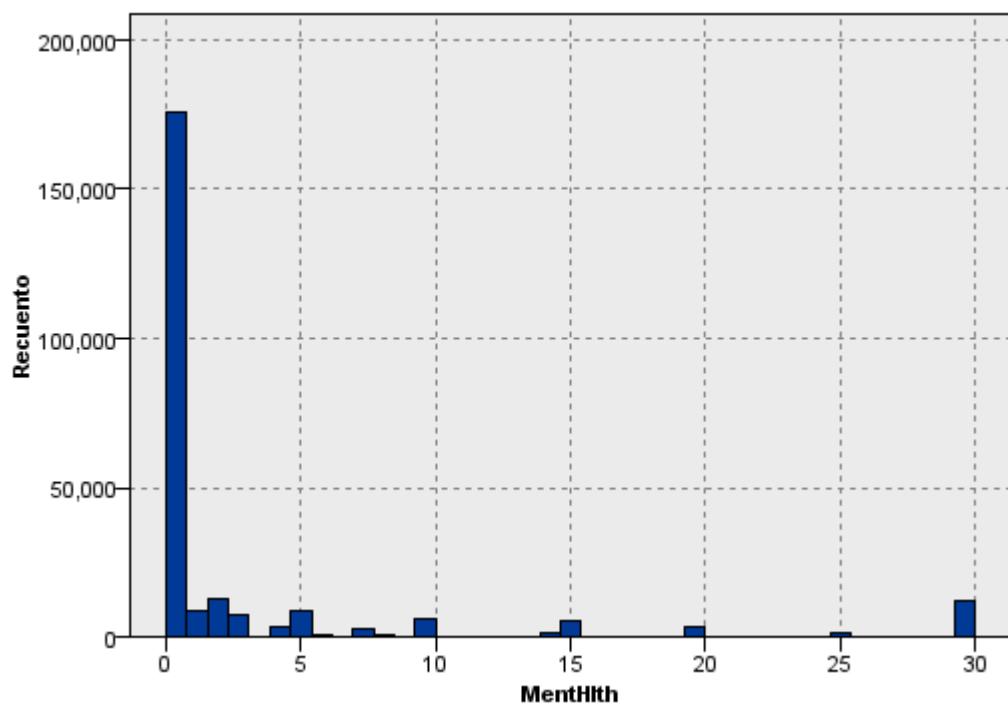
**Sexo (Genero); 0 = mujer, 1 = hombre**



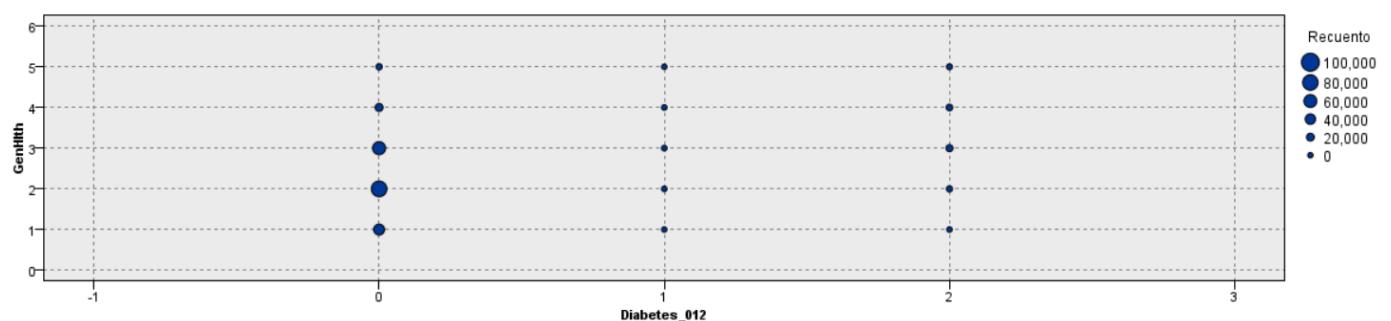
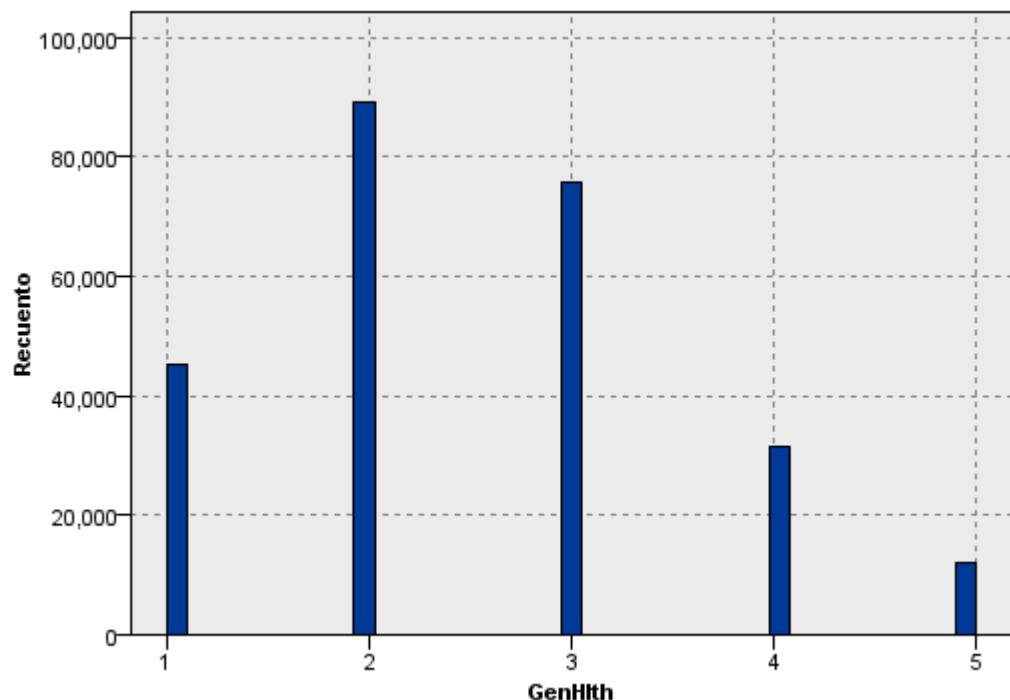
**Salud Física (¿por cuántos días durante los últimos 30 días no fue buena tu salud física?); 1-30**



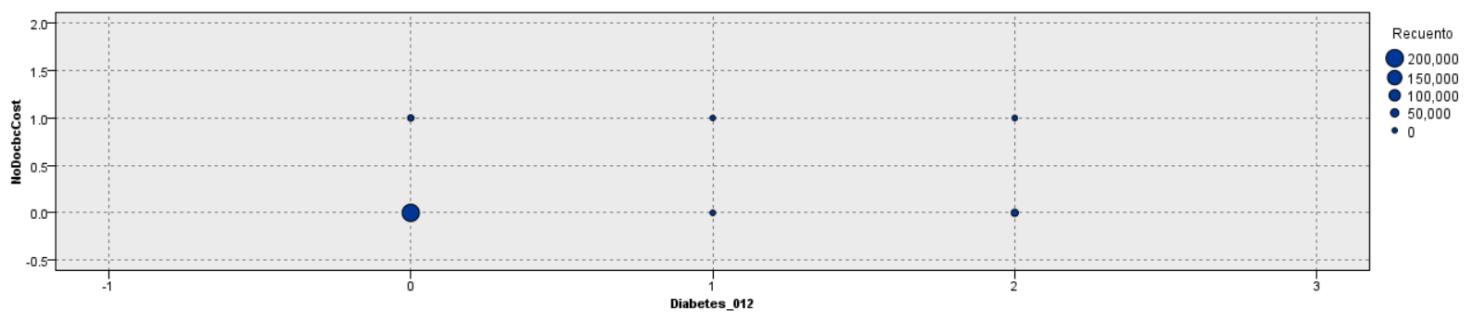
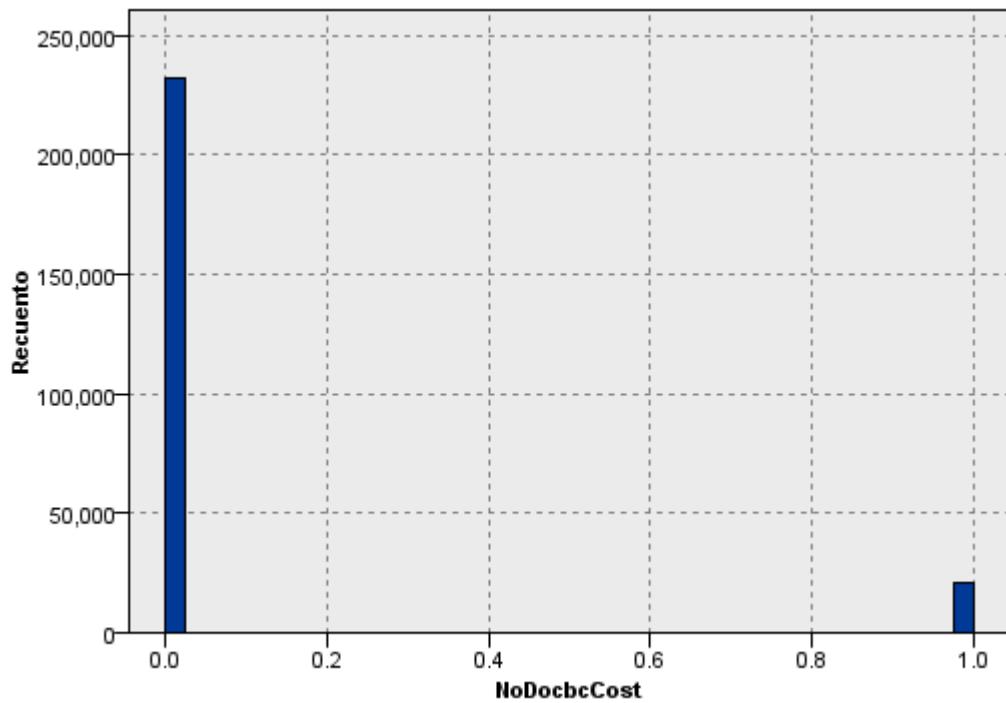
**Salud Mental (¿por cuántos días durante los últimos 30 días no fue buena tu salud mental?); 1-30**



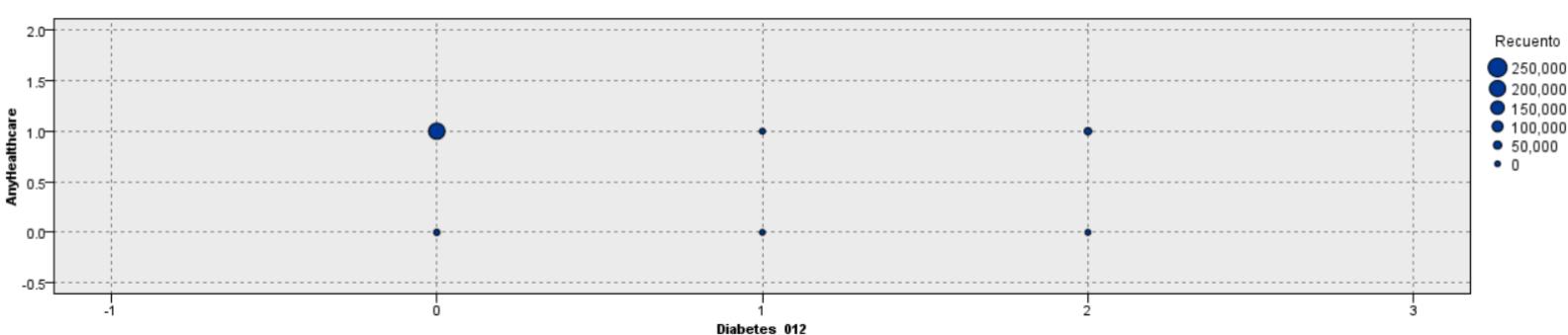
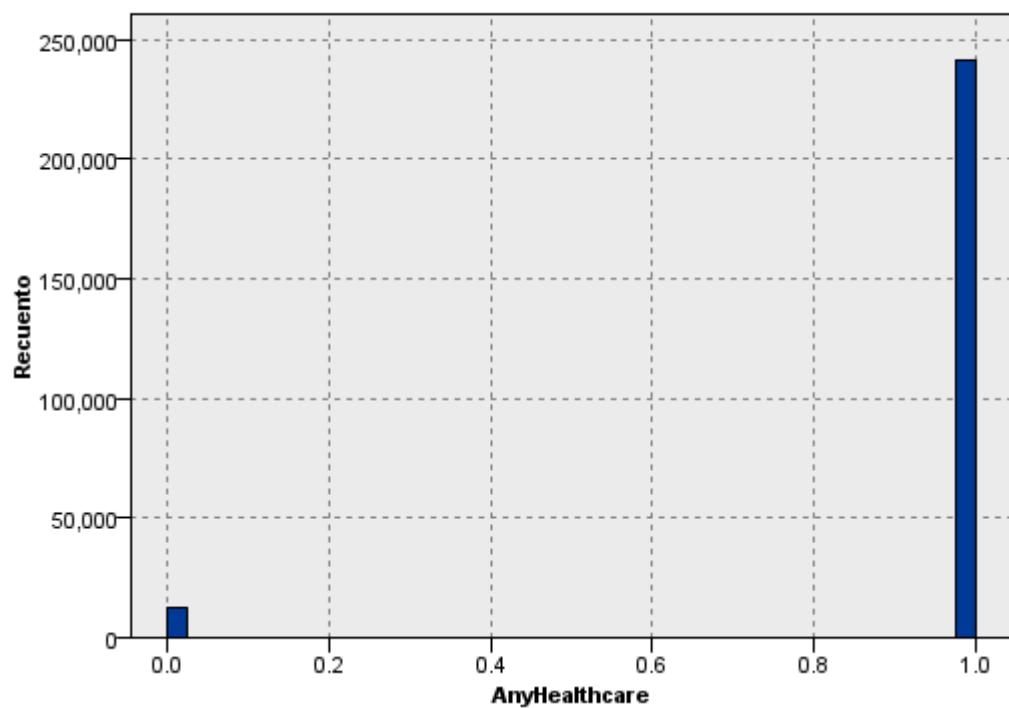
**Salud en general; 1 = excelente, 2 = muy buena, 3 = buena, 4 = regular, 5 = mala**



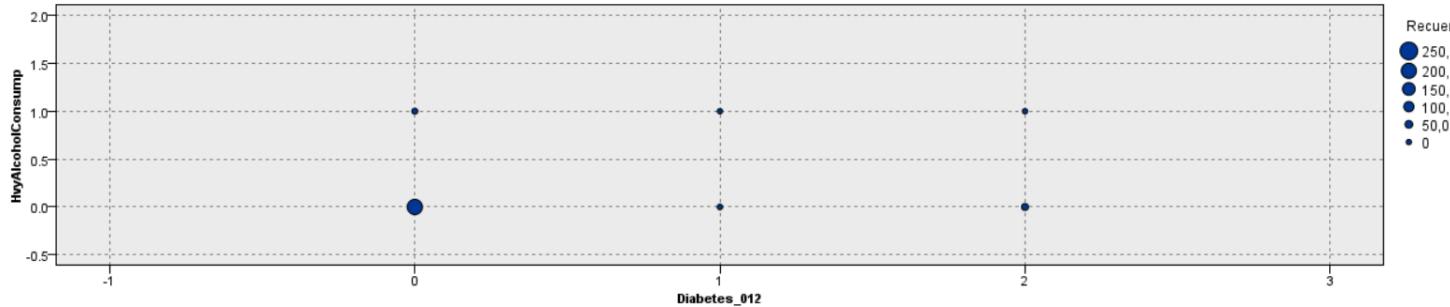
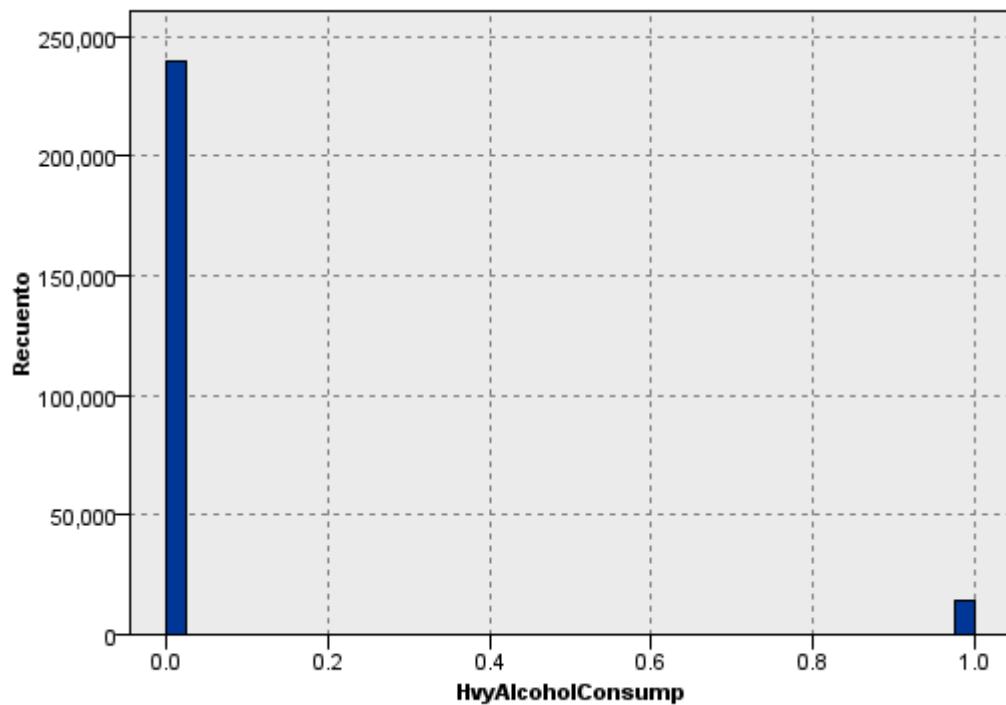
**No obtuvo atención médica en los últimos 12 meses debido a no poder cubrir el costo; 0 = no, 1 = sí**



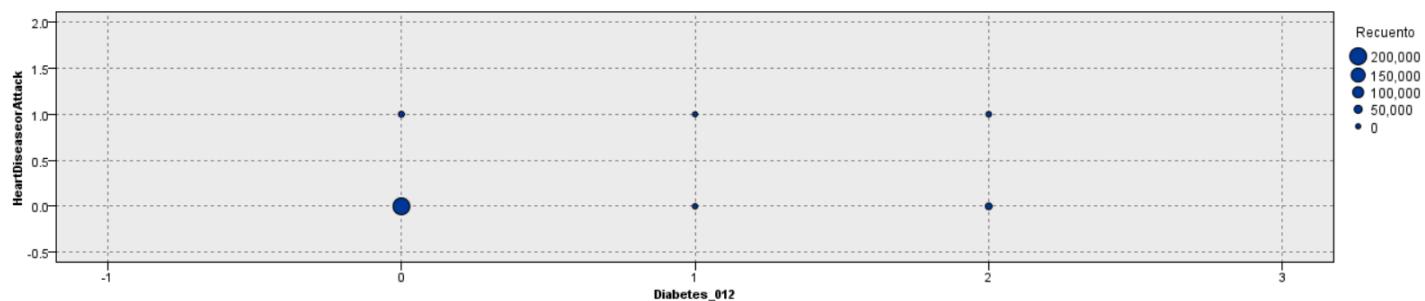
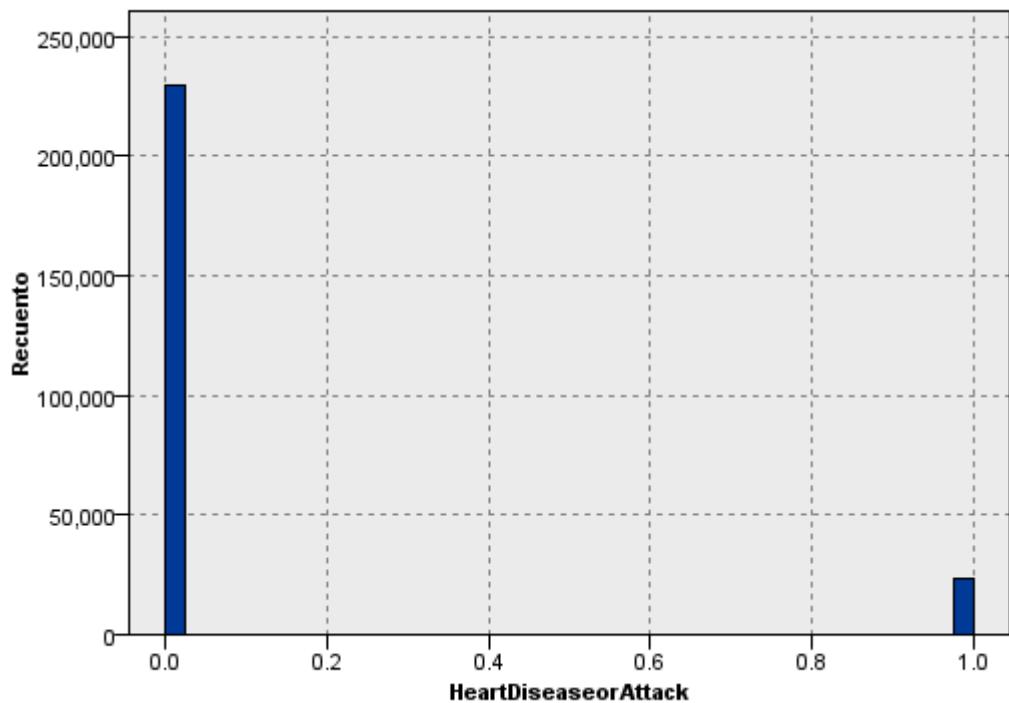
**Afiliación a Servicios de salud; 0 = no, 1 = sí**



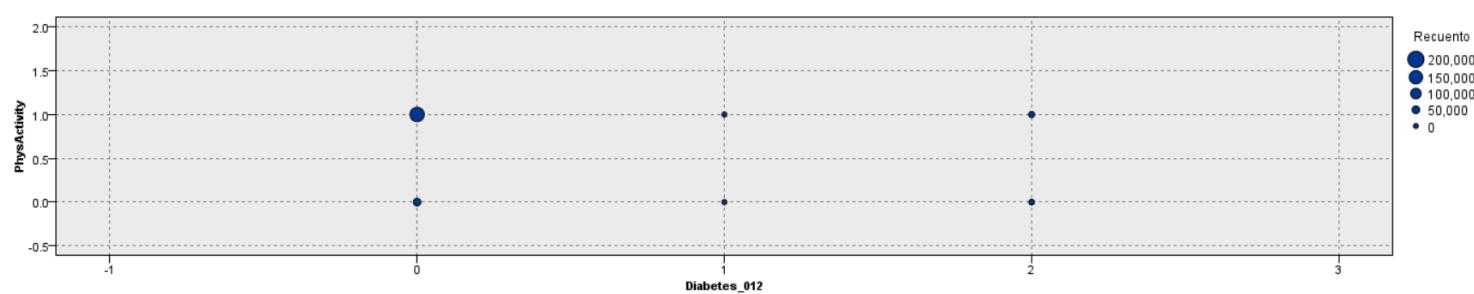
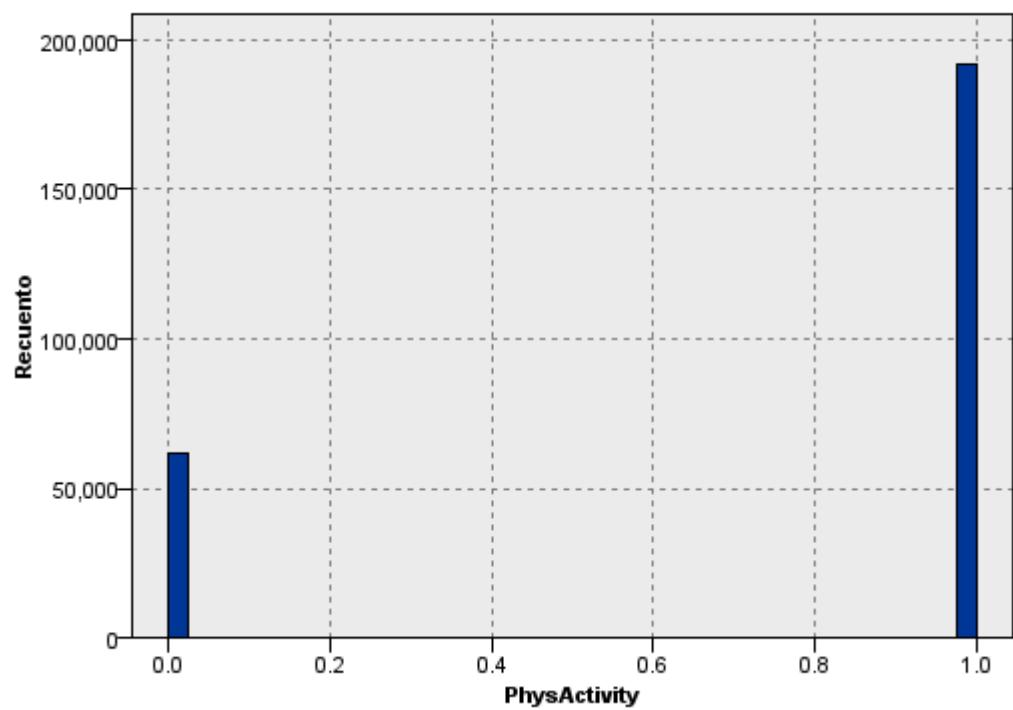
**Consumo Importante de Alcohol; 0 = no, 1 = sí**



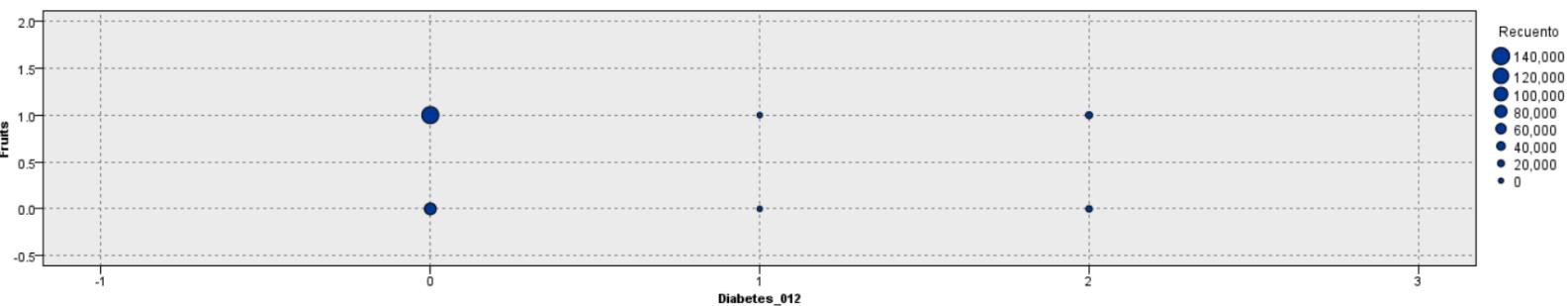
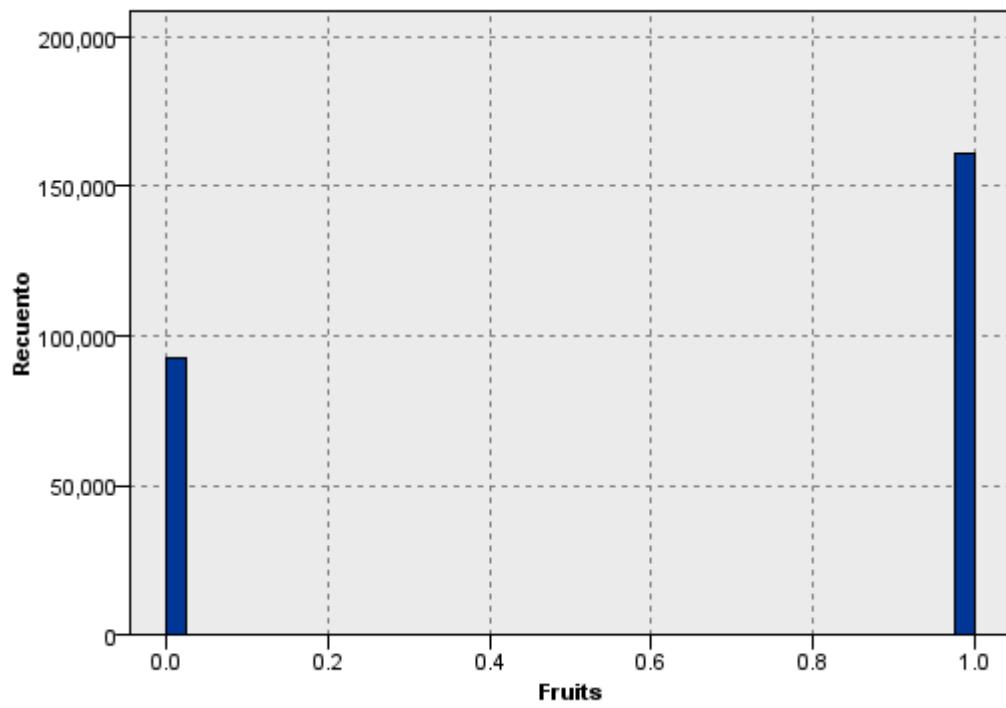
**Ataque Cardíaco; 0 = no, 1 = sí**



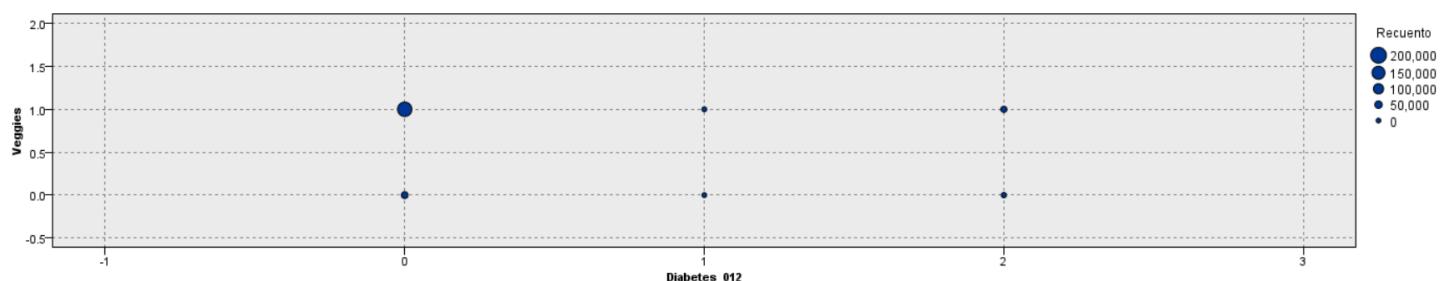
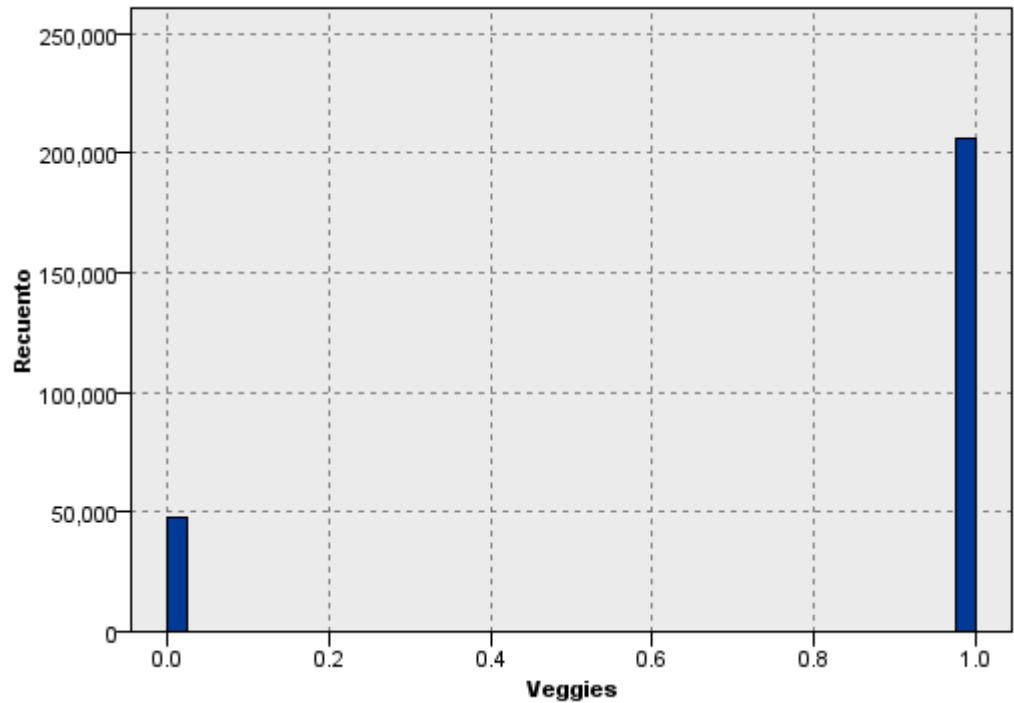
**Actividad Física en los últimos 30 días; 0 = no, 1 = sí**



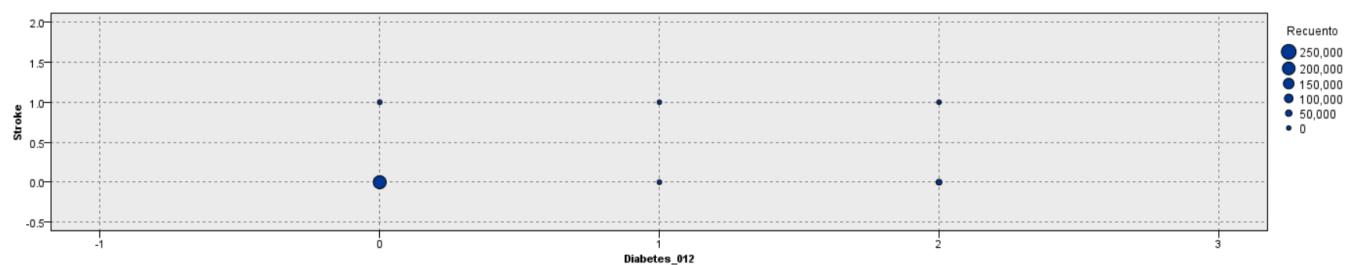
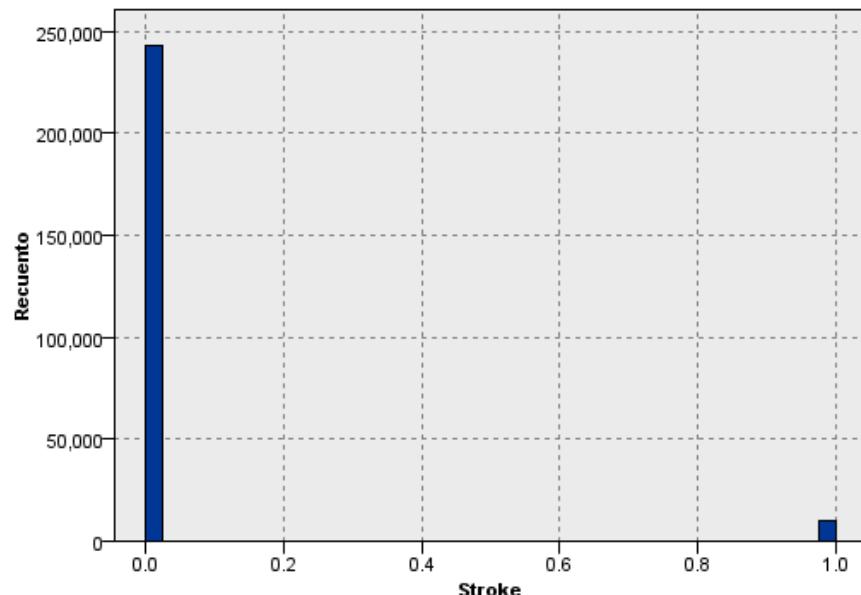
**Consumo de Frutas; 0 = no, 1 = sí**



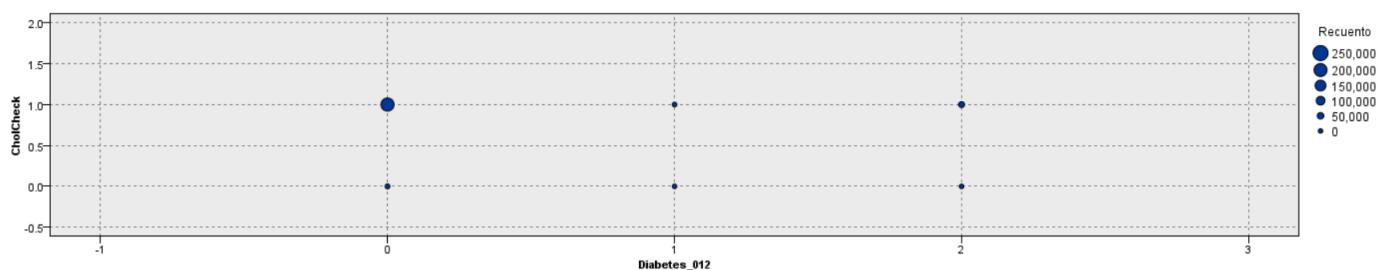
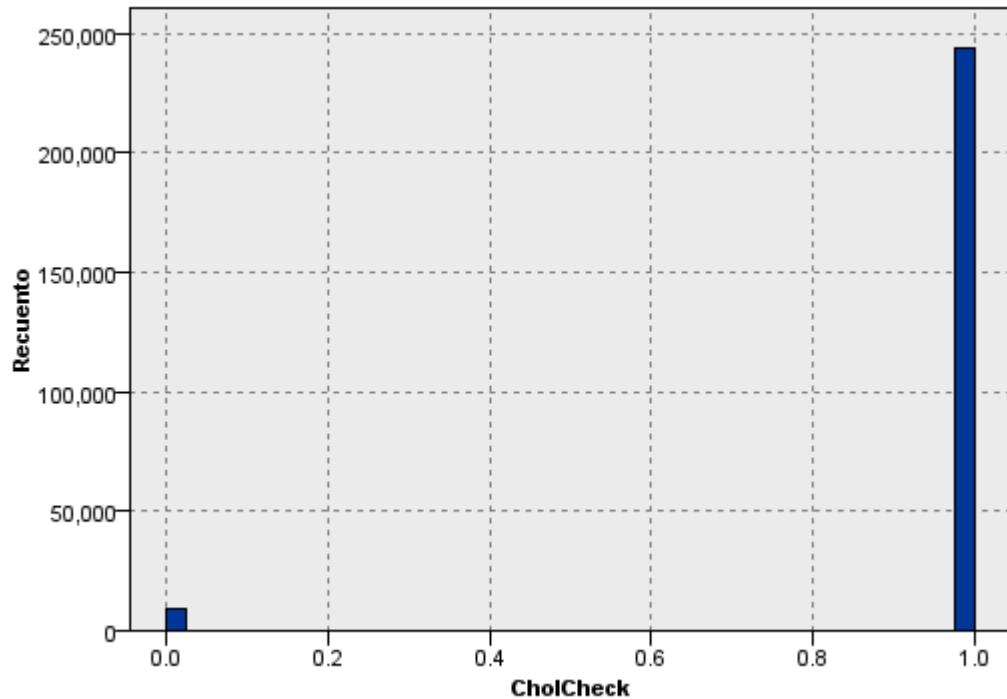
**Consumo de vegetales; 0 = no, 1 = sí**



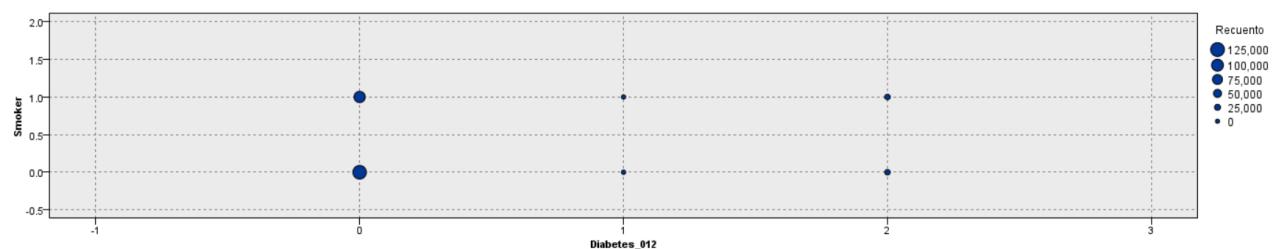
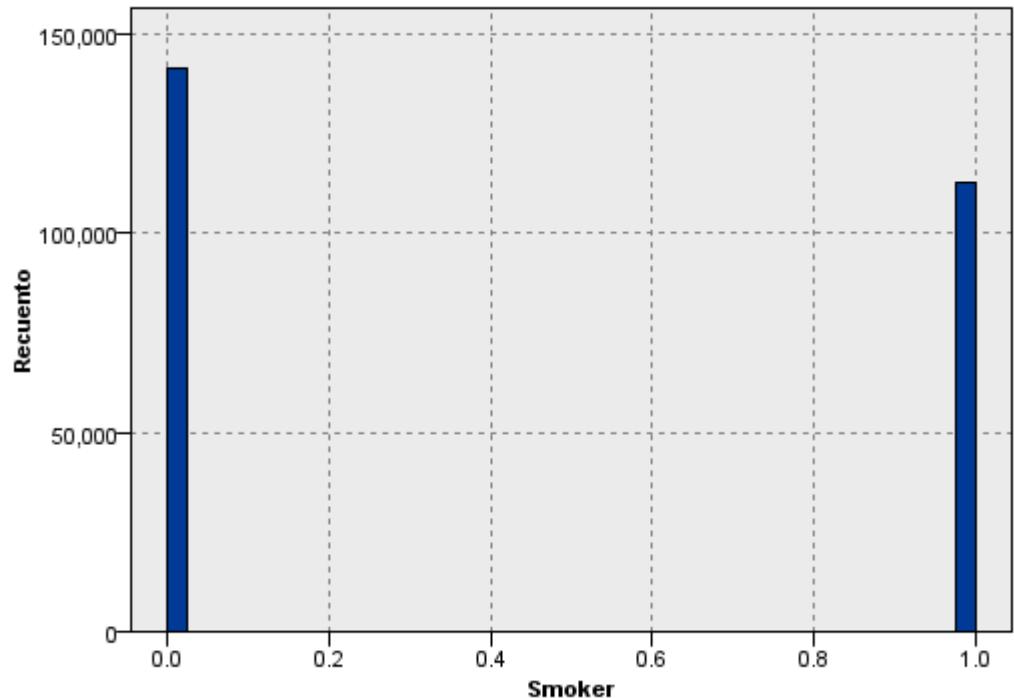
**Accidente Cerebrovascular; 0 = no, 1 = sí**



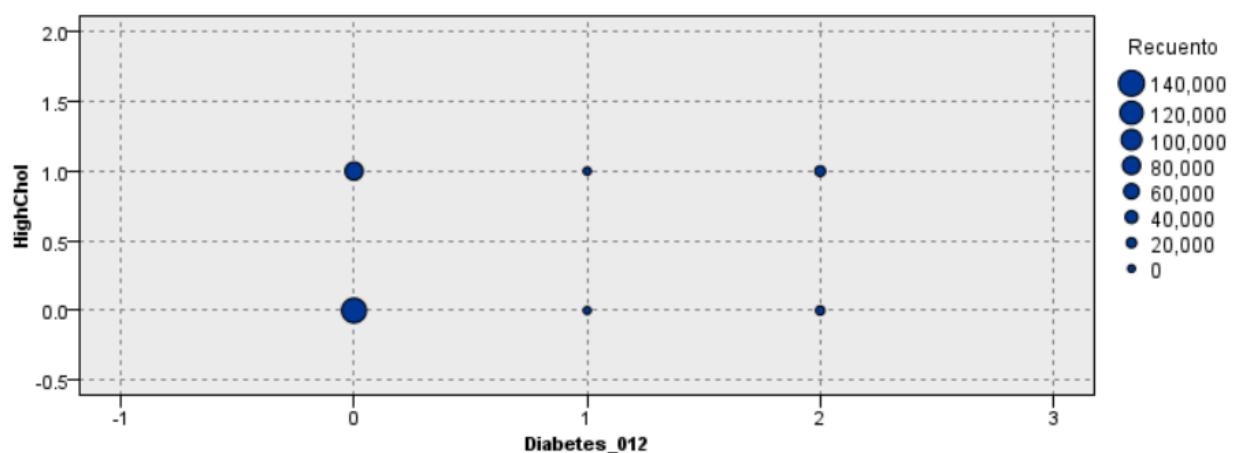
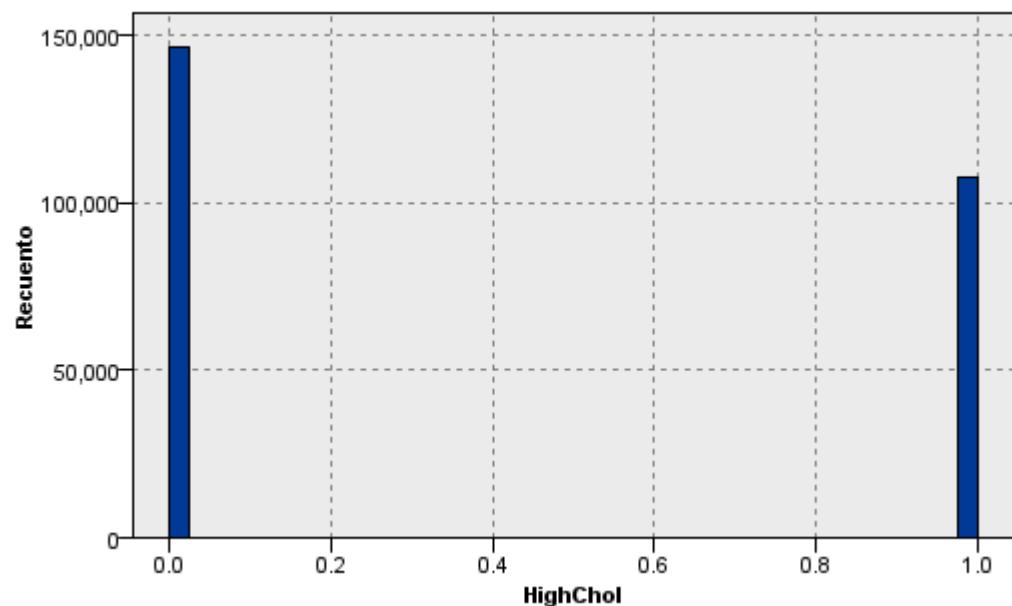
**Chequeo de colesterol;** 0 = no hubo chequeos de colesterol en 5 años, 1 = sí hubo chequeos de colesterol en 5 años.



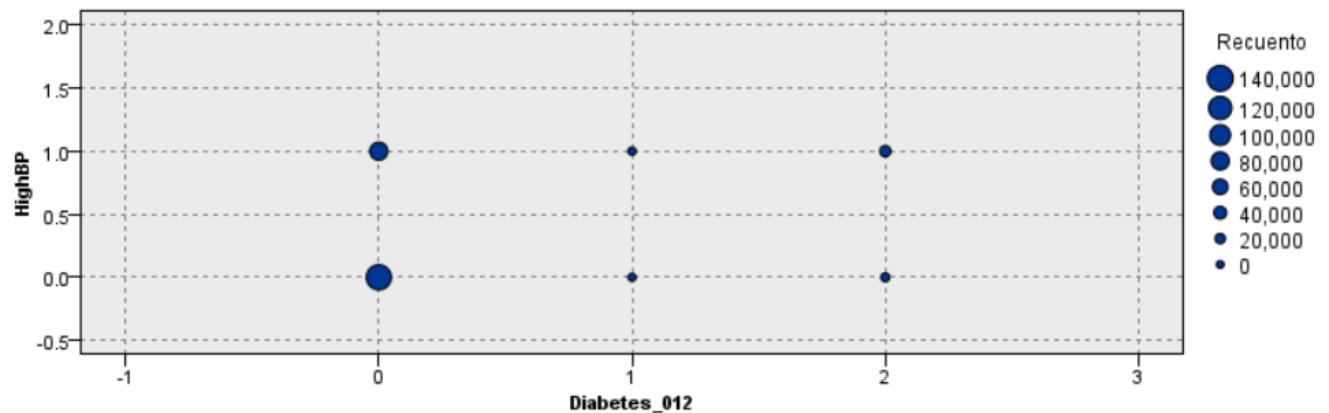
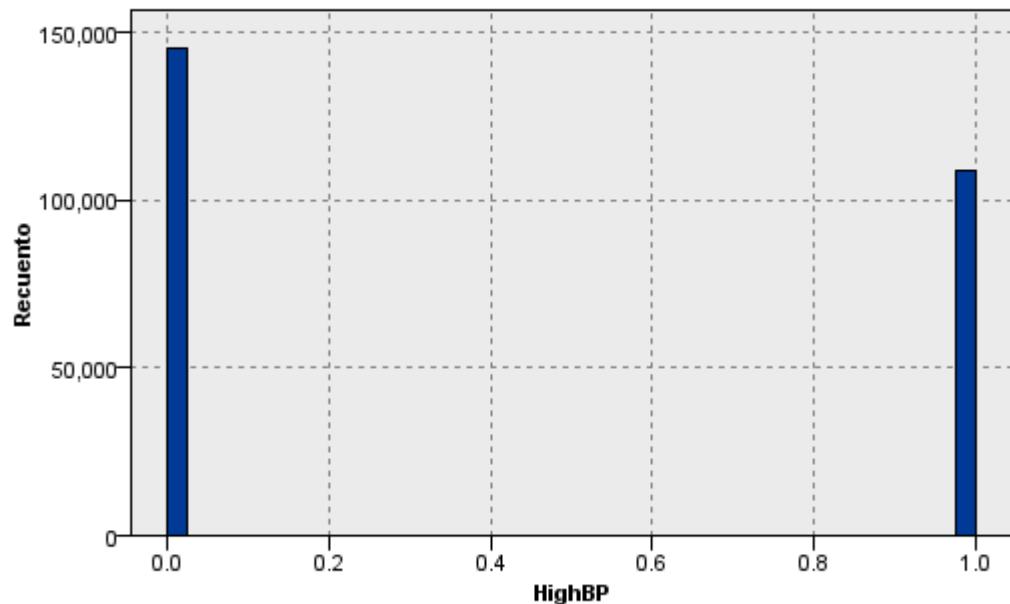
**Fumador Sí/No; 0 = no, 1 = sí**



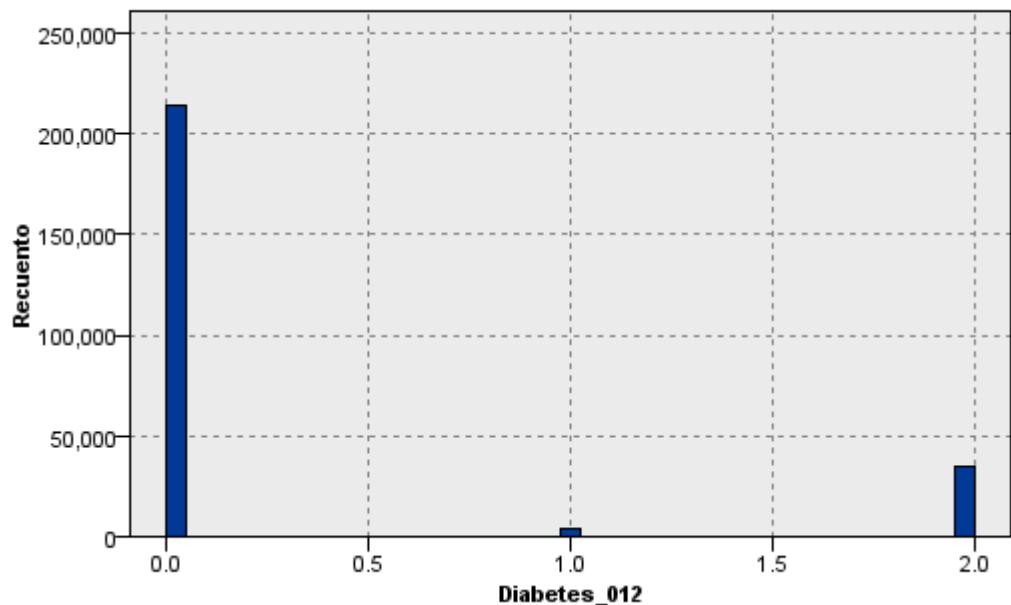
**Colesterol Alto;** 0 = no colesterol alto, 1 = colesterol alto



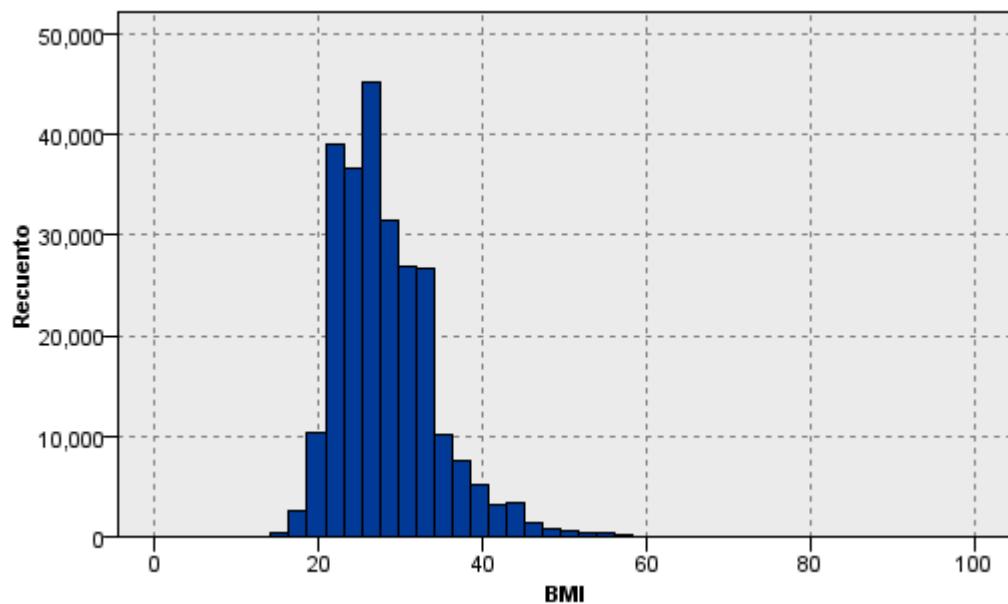
**Presión Sanguínea Alta;** 0 = no presión arterial alta, 1 = presión arterial alta

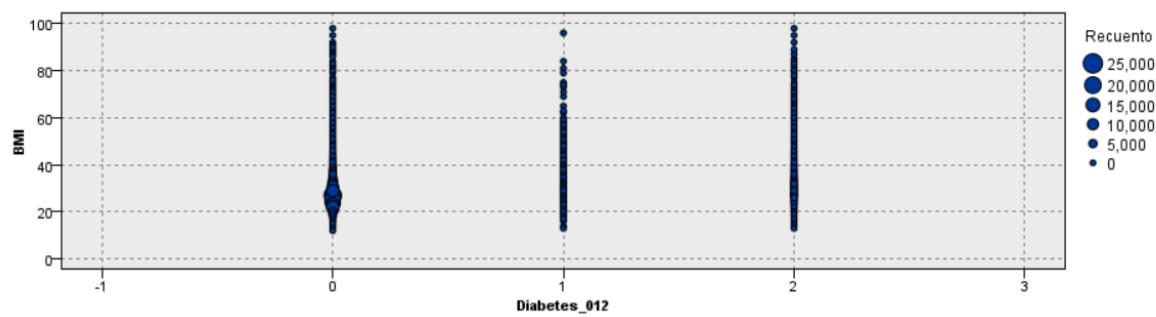


**Presencia de Diabetes o Prediabetes; 0= No Diabetes, 1 = prediabetes, 2 = Diabetes**

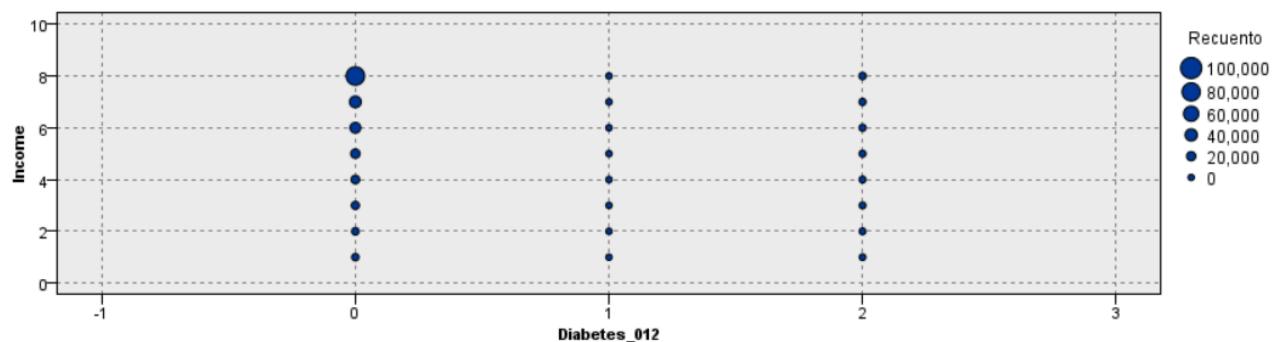
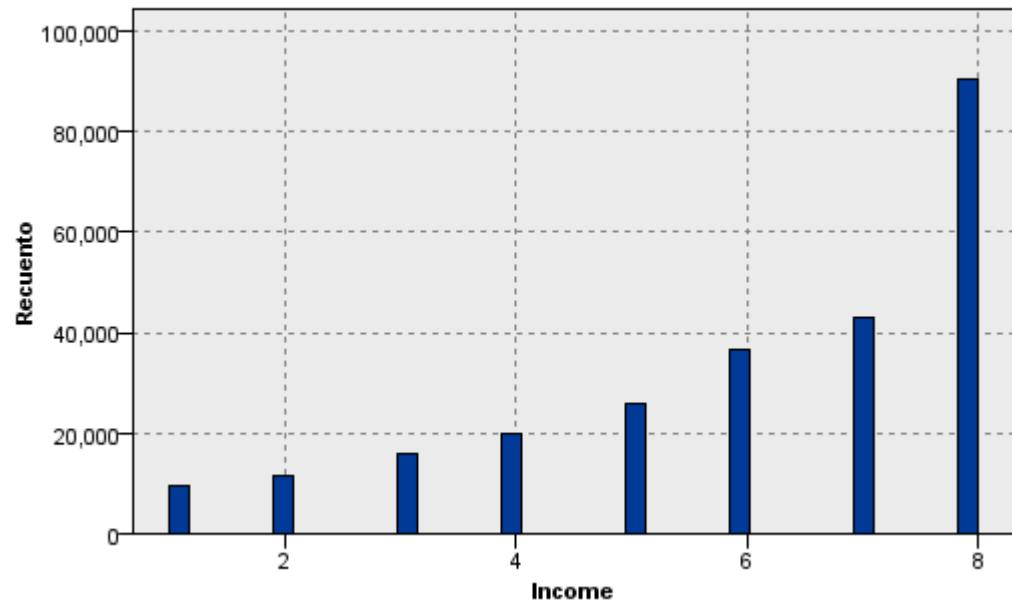


### Índice de Masa Corporal

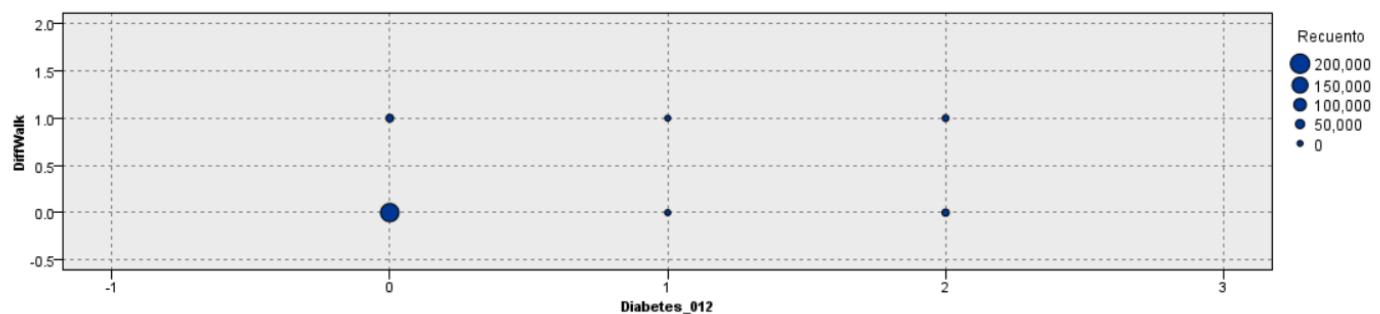
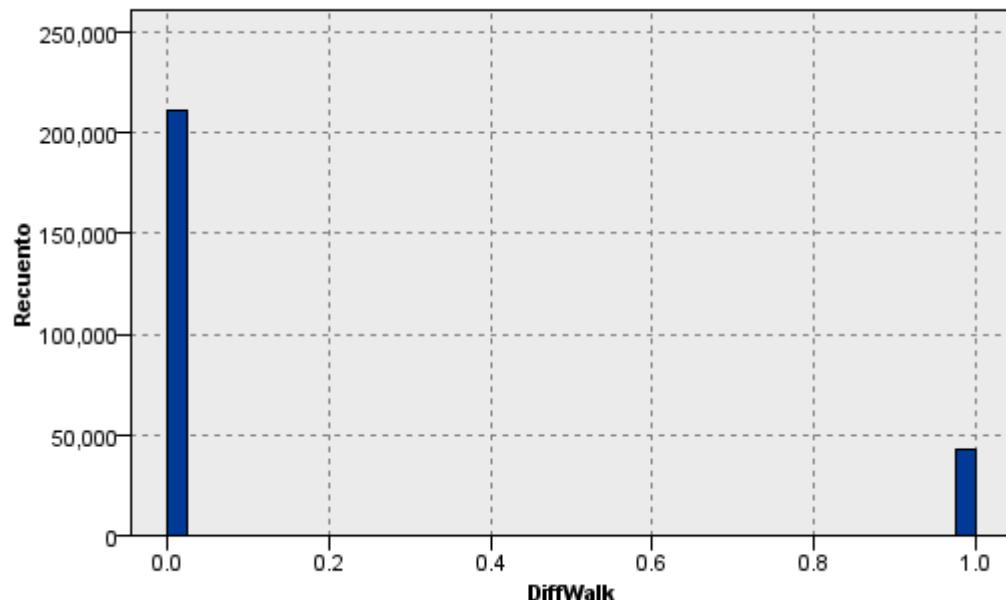




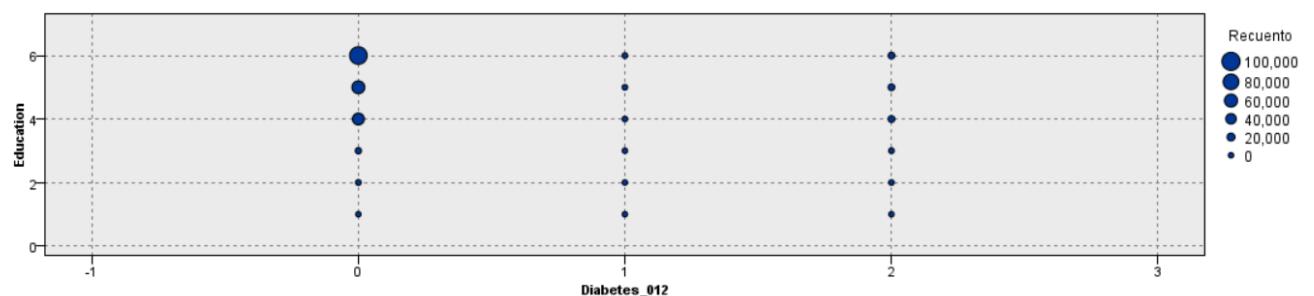
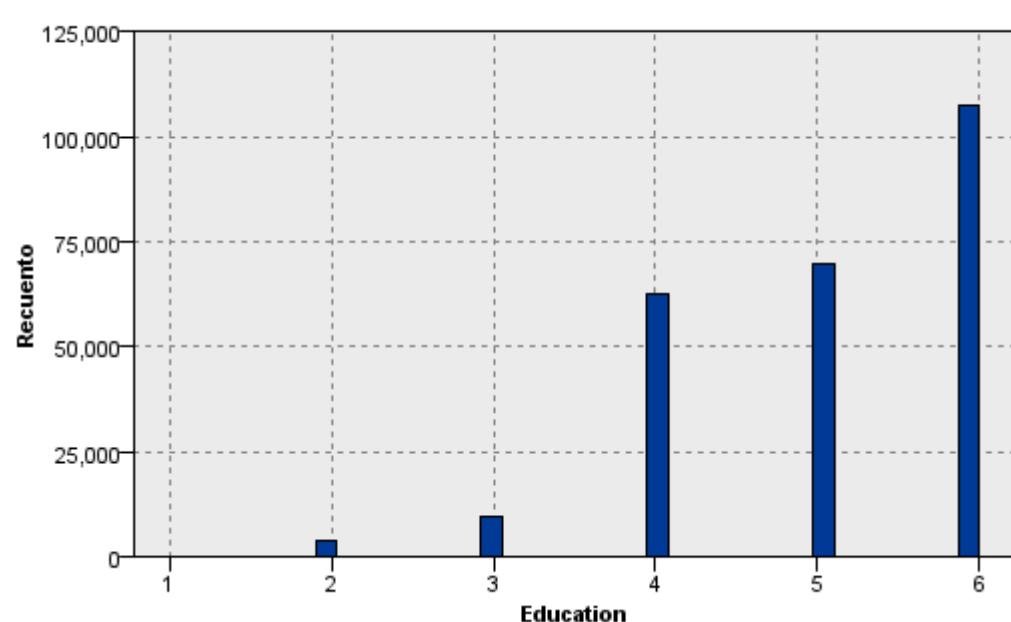
**Ingreso;** 1 = menos de \$10,000, 5 = menos de \$35,000, 8 = \$75,000 o más

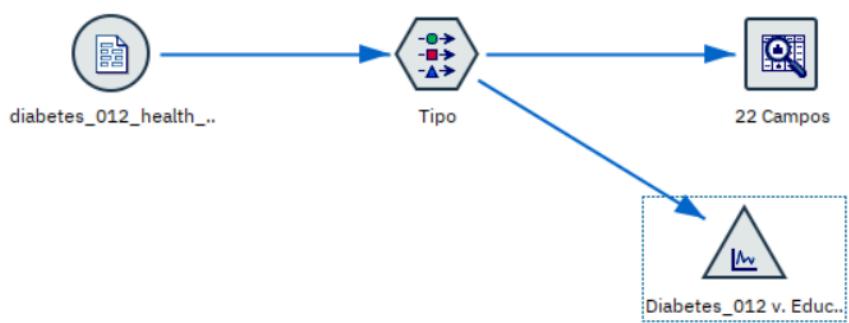


**Dificultad para caminar; 0 = no, 1 = sí**



**Nivel de Educación:** 1 = nunca asistió a la escuela o solo kindergarten, 2 = grados 1 al 8 (primaria), 3 = grados 9 al 11 (algo de secundaria), 4 = grado 12 o GED (graduado de secundaria), 5 = universidad 1 a 3 años (alguna universidad o escuela técnica), 6 = universidad 4 años o más (graduado universitario)





## Verificación de la calidad de los datos

Llegamos a la conclusión de que el DataSet no contiene errores en sus datos. Con error nos referimos a registros de datos nulos, cadenas vacías, espacios en blanco o valor en blanco, por lo que no se tendrán que hacer la sustitución de datos erróneos.

Campo	Medida	Valores atípicos	Extremos	Acción	Imputar perdidos	Método	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio en blanco	Valor en blanco
Diabetes_012	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
HighBP	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
HighChol	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
CholCheck	Continuo	0	9470 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
BMI	Continuo	2193	770 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Smoker	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Stroke	Continuo	10292	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
HeartDisease...	Continuo	23893	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
PhysActivity	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Fruit	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Veggies	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
HyalcoholC...	Continuo	14256	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
AnyHealthcare	Continuo	12417	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
NoDocbcCost	Continuo	21354	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
GenHlth	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
MentHlth	Continuo	12697	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
PhysHlth	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
DiffWalk	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Sex	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Age	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Education	Continuo	4217	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0
Income	Continuo	0	0 Ninguno	Nunca	Fijo	100	253680	0	0	0	0	0

Imagen. Tabla donde se muestran las columnas que componen al DataSet, así como su estatus sobre la presencia de errores en sus tuplas.

**Volumen:**

El conjunto de datos del CDC se presenta como una vasta reserva de información, compuesto por 253,680 tuplas y 21 atributos. Este volumen brinda una base para realizar análisis en profundidad y desarrollar modelos predictivos precisos. La amplitud del conjunto de datos permite abordar la complejidad de los factores relacionados con la diabetes y ofrece la capacidad de extraer patrones y tendencias significativas.

**Veracidad:**

La fuente confiable del CDC asegura la veracidad y confiabilidad de los datos. La reputación internacionalmente reconocida del CDC respalda la calidad de la información proporcionada. Además, al explorar las variables del conjunto de datos, se observa una codificación clara y coherente. Este nivel de precisión contribuye a la confiabilidad de los resultados derivados del conjunto de datos y garantiza que las conclusiones extraídas sean representativas y válidas.

**Variedad:**

El conjunto de datos del CDC abarca variables de distintos tipos, desde binarias hasta enteras y categóricas. Esta diversidad permite un enfoque completo al análisis, brindando una comprensión más rica de los factores relacionados con la diabetes. Desde información demográfica hasta hábitos de salud, la variedad en las variables enriquece la complejidad del conjunto de datos.

**Valor:**

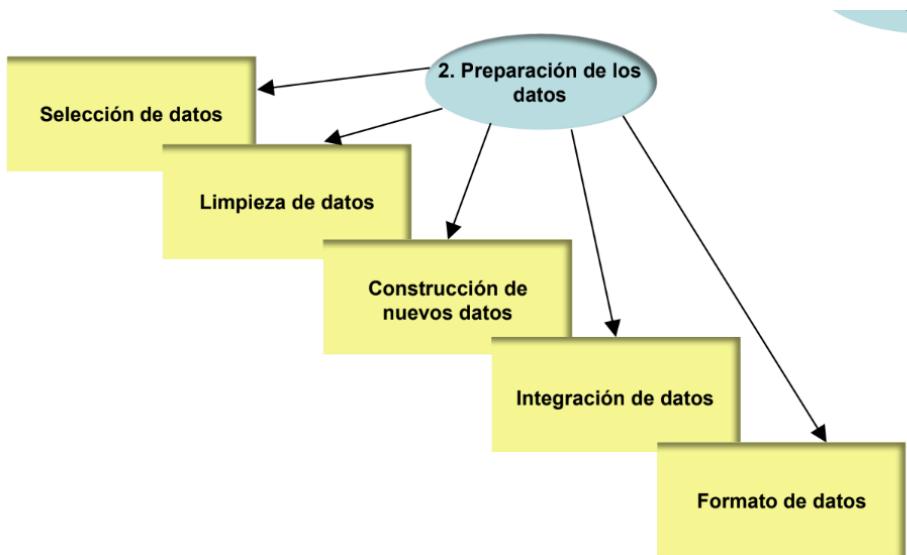
En este sentido, el conjunto de datos del CDC se destaca al incluir variables clave relacionadas con hábitos alimenticios, actividad física, antecedentes médicos y más. Estos datos fundamentales son esenciales para comprender y predecir la aparición de la diabetes en función del estilo de vida. La riqueza de información presente en el conjunto de datos respalda su valor en el contexto del proyecto de minería de datos.

**Velocidad:**

El formato CSV del conjunto de datos del CDC, que contiene información de 253,680 respuestas, facilita la carga rápida en diversas plataformas y herramientas analíticas. La eficacia en la manipulación de datos contribuye a la velocidad en la generación de insights(patrones de análisis) y resultados.

Link hacia el dataSet: [CDC Diabetes Health Indicators - UCI Machine Learning Repository](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)

## Preparación de los Datos



### Selección de los Datos

Debido a que el DataSet es muy grande, hemos dividido el DataSet en dos. Uno será una muestra aleatoria y sin repetición del 10% del DataSet original (25 mil registros), mientras que el otro será el 90% restante. Principalmente se trabajará con la muestra del 10% del DataSet, mientras que el 90% se quedará con reserva.

Para cumplir con los objetivos de minería de datos, hemos considerado que el campo “Education” de nuestro DataSet original podría ocasionar ruido a la hora de generar los modelos. Es por ello que se ha decidido no tomar en cuenta los datos contenidos en ese campo en aras de mejorar la bondad de los modelos clasificatorios y predictivos.

## Limpieza de los datos.

Como ya se determinó en la fase “Comprensión de los datos”, no será necesario corregir problemas relacionados con datos perdidos, vacíos o errores en los datos. Realizamos un análisis manual para corroborar lo que argumentamos. Sin embargo, para poder estar aún más seguros debido a que en nuestro DataSet se encuentran 253680 tuplas, la plataforma IBM SPSS Modeler nos permite hacerlo de manera automatizada. A continuación mostraremos que se realizaron para verificar que se realizó de manera adecuada.

Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Educatio
0.000	1.000	1.000	1.000 40...	1.000 0.000	0.000	0.000 0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	5.000	18.000	15.000	1.000	0... 9.0...	4.0(	
0.000	0.000	0.000	0.000 25...	1.000 0.000	0.000	0.000 0.000	0.000	1.000	0.000	0.000	0.000	1.000	3.000	0.000	0.000	0.000	0... 7.0...	6.0(		
0.000	1.000	1.000	1.000 28...	0.000 0.000	0.000	0.000 1.000	0.000	1.000	1.000	0.000	1.000	1.000	5.000	30.000	30.000	1.000	0... 9.0...	4.0(		
0.000	1.000	0.000	1.000 27...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	0.000	0.000	0.000	0... 11...	3.0(		
0.000	1.000	1.000	1.000 24...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	3.000	3.000	0.000	0... 11...	5.0(		
0.000	1.000	1.000	1.000 25...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	0.000	14.000	0.000	0... 9.0...	6.0(		
0.000	1.000	1.000	1.000 25...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	0.000	0.000	0.000	1.000 0... 11...	4.0(		
2.000	1.000	1.000	1.000 30...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	5.000	30.000	30.000	1.000	0... 9.0...	5.0(		
0.000	0.000	0.000	1.000 24...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	0.000	0.000	0.000	1... 8.0...	4.0(		
2.000	0.000	0.000	1.000 25...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	0.000	0.000	0.000	1... 13...	6.0(		
0.000	1.000	1.000	1.000 34...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	0.000	30.000	1.000	0... 10...	5.0(		
0.000	0.000	0.000	1.000 26...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	0.000	15.000	0.000	0... 7.0...	5.0(		
2.000	1.000	1.000	1.000 28...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	4.000	0.000	0.000	1.000	0... 11...	4.0(		
0.000	0.000	1.000	1.000 33...	1.000 1.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	1.000	4.000	30.000	28.000	0.000	0... 4.0...	6.0(		
0.000	1.000	0.000	1.000 33...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	5.000	0.000	0.000	0... 6.0...	6.0(		
0.000	1.000	1.000	1.000 21...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	0.000	0.000	0.000	0... 10...	4.0(		
2.000	0.000	0.000	1.000 23...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	0.000	0.000	0.000	1... 7.0...	5.0(		
0.000	0.000	0.000	1.000 23...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	15.000	0.000	0.000	0... 2.0...	6.0(		
0.000	0.000	0.000	1.000 28...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	10.000	0.000	0.000	1... 4.0...	6.0(		
0.000	1.000	1.000	1.000 22...	0.000 1.000	1.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	30.000	0.000	1.000	0... 12...	4.0(		
0.000	1.000	1.000	1.000 38...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	5.000	15.000	30.000	1.000	0... 13...	2.0(		
0.000	0.000	0.000	1.000 28...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	0.000	7.000	0.000	1... 5.0...	5.0(		
2.000	1.000	0.000	1.000 27...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0... 13...	5.0(		
0.000	1.000	1.000	1.000 28...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	3.000	6.000	0.000	1.000	0... 9.0...	4.0(		
0.000	0.000	0.000	1.000 32...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	0.000	0.000	0.000	0... 5.0...	6.0(		
2.000	1.000	1.000	1.000 37...	1.000 1.000	1.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	5.000	0.000	0.000	1.000	1... 10...	6.0(		
2.000	1.000	1.000	1.000 28...	1.000 0.000	1.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	4.000	0.000	0.000	1.000	1... 12...	2.0(		
2.000	1.000	1.000	1.000 27...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	4.000	20.000	20.000	1.000	0... 8.0...	4.0(		
0.000	1.000	1.000	1.000 31...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	1... 12...	6.0(		
2.000	1.000	1.000	1.000 34...	1.000 1.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	4.000	0.000	7.000	1.000	0... 9.0...	5.0(		
0.000	1.000	0.000	1.000 33...	1.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	1.000	1... 13...	3.0(		
0.000	0.000	0.000	1.000 23...	0.000 0.000	0.000	0.000 1.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	0.000	0.000	0.000	0... 6.0...	4.0(		

Imagen 1. DataSet Original antes de realizar cualquier cambio.

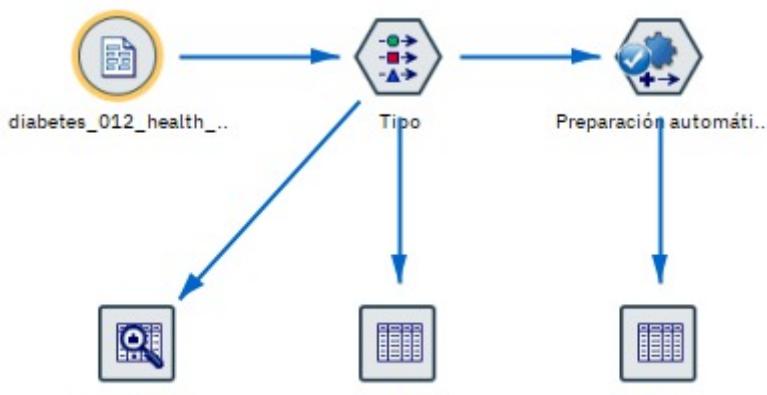


Imagen 2. DataSet Representado en la plataforma IBM SPSS Modeler

#### Ajustar el tipo y mejorar la calidad de los datos

Desti...

- Ajustar el tipo de campos numéricos (ordinales y continuos)
- Reordenar campos nominales para tener la categoría en primer lugar y la mayor en último lugar
- Sustituir valores extremos en campos continuos (recomendado para campos de entrada si se utilizarán en una escala común)
- Campos continuos: sustituir valores perdidos por la media
- Campos nominales: sustituir valores perdidos por el modo
- Campos ordinales: sustituir valores perdidos por la mediana

Número máximo de valores de campos ordinales:

Número mínimo de valores de campos ordinales:

Valor de corte atípico:  (desviaciones estándar)

Método de sustitución de valores atípicos:  Sustituir con valor de corte  Eliminar valor

#### Transformar campo continuo

Poner todos los campos de entrada continuos en una escala común (muy recomendado si se ejecutará la característica de constru...

Método de cambio de escala:  Media final:  Desviación estándar final:

Imagen 3. Configuración que se le realizó al DataSet en la plataforma IBM SPSS Modeler para poder hacer la corrección de manera automatizada.

Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Educati
0.000	1.000	1.000	1.000	40....	1.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	5.000	18.000	15.000	1.000	0... 9.0...	4.0(	
0.000	0.000	0.000	0.000	25....	1.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	3.000	0.000	0.000	0.000	0... 7.0...	6.0(		
0.000	1.000	1.000	1.000	28....	0.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	1.000	30.000	30.000	1.000	0... 9.0...	4.0(		
0.000	1.000	0.000	0.000	27....	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	2.000	0.000	0.000	0... 11....	3.0(		
0.000	1.000	1.000	1.000	24....	0.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	2.000	3.000	0.000	0... 11....	5.0(		
0.000	1.000	1.000	1.000	25....	0.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	2.000	0.000	2.000	0.000	1... 10....	6.0(	
0.000	1.000	0.000	0.000	30....	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	3.000	0.000	14.000	0.000	0... 9.0...	6.0(		
0.000	1.000	1.000	1.000	25....	1.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	3.000	0.000	0.000	1.000	0... 11....	4.0(		
2.000	1.000	1.000	1.000	30....	1.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	5.000	30.000	30.000	1.000	0... 9.0...	5.0(		
0.000	0.000	0.000	0.000	24....	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	2.000	0.000	0.000	1... 8.0...	4.0(		
2.000	0.000	0.000	0.000	25....	1.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	3.000	0.000	0.000	0.000	1... 13....	6.0(		
0.000	1.000	1.000	1.000	34....	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	3.000	0.000	30.000	1.000	0... 10....	5.0(	
0.000	0.000	0.000	0.000	26....	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	3.000	0.000	15.000	1.000	0... 7.0...	5.0(	
2.000	1.000	1.000	1.000	28....	0.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	4.000	0.000	0.000	1.000	0... 11....	4.0(	
0.000	0.000	1.000	1.000	33....	1.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	1.000	4.000	30.000	28.000	0.000	0... 4.0...	6.0(	
0.000	1.000	0.000	0.000	33....	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	2.000	5.000	0.000	0.000	0... 6.0...	6.0(	
0.000	1.000	1.000	1.000	21....	0.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	3.000	0.000	0.000	0.000	0... 10....	4.0(		
2.000	0.000	0.000	0.000	23....	1.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	2.000	0.000	0.000	1... 7.0...	5.0(		
0.000	0.000	0.000	0.000	23....	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	15.000	0.000	0.000	0... 2.0...	6.0(	
0.000	0.000	1.000	1.000	28....	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	2.000	10.000	0.000	0.000	1... 4.0...	6.0(	
0.000	1.000	1.000	1.000	22....	0.000	0.000	1.000	0.000	1.000	0.000	0.000	1.000	0.000	3.000	30.000	0.000	1.000	0... 12....	4.0(	
0.000	1.000	1.000	1.000	38....	1.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	0.000	5.000	15.000	30.000	1.000	0... 13....	2.0(	
0.000	0.000	1.000	1.000	28....	1.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	0.000	2.000	5.000	0.000	0.000	0... 6.0...	6.0(	
2.000	1.000	0.000	0.000	27....	1.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0... 13....	5.0(	
0.000	1.000	1.000	1.000	31....	1.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	7.000	0.000	0.000	1... 12....	6.0(
2.000	1.000	1.000	1.000	34....	1.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	4.000	0.000	7.000	1.000	0... 9.0...	5.0(	
0.000	1.000	0.000	0.000	33....	1.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	1.000	0... 13....	3.0(	
0.000	0.000	0.000	0.000	23....	0.000	0.000	0.000	1.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	1.000	0... 6.0...	4.0(	

Imagen 4. DataSet después de haber aplicado la preparación automática de datos

De esta etapa podemos concluir que no hubo necesidad de hacer limpieza de datos debido a que no se encontraron datos que estuvieran dañados.

Nuestro dataset se caracteriza por su notable completitud, ya que, desde su origen, no presentaba ningún vacío de información ni datos atípicos. Cada atributo contenía información numérica o categórica, y el único campo de datos continuos, referente al índice de masa corporal, nos mostró una integridad impecable sin impurezas detectables. Es por eso que en nuestra parte de limpieza de datos se limitó a una muy estricta verificación para confirmar esta integridad, eliminando así la necesidad de correcciones significativas. Este sólido fundamento inicial ha facilitado enormemente la calidad y confiabilidad de nuestro conjunto de datos para cualquier análisis en el futuro.

## Balanceo de datos.

Antes de comenzar con el balance, hay que recordar que se tomó el 10% de los registros del DataSet Original, es decir, con 25,369 registros.

- *Balanceo Diabetes\_012:*

Originalmente, el campo Diabetes\_012 tiene un desbalance de ocurrencias, pues como se puede observar en la siguiente tabla, la categoría de no presentar Diabetes (representado con el número 0) ocupaba un poco más de 84% de las ocurrencias.

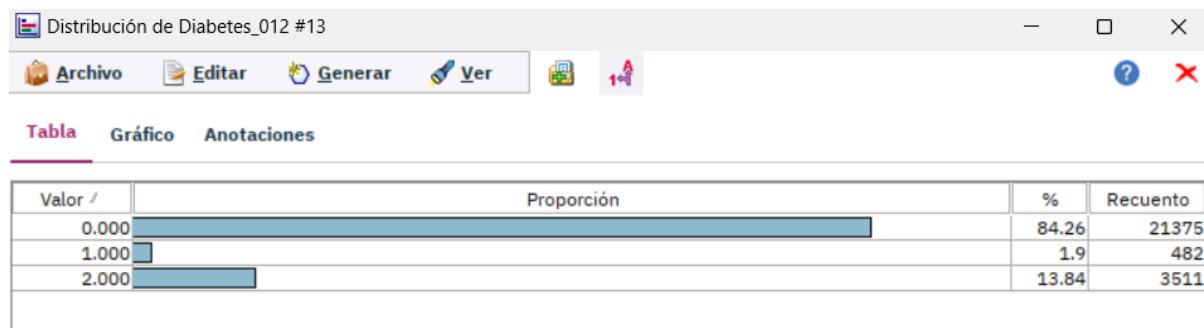


Imagen 5. Recuento de datos Desbalanceado

Para mitigar esto, utilizamos un script de Python con la biblioteca “pandas”, con el que agregamos registros del 90% restante del DataSet original.

```
import pandas as pd

df_original = pd.read_csv('DataSet90restante.csv')

df_muestra = pd.read_csv('muestra10.csv')
```

```

condicion1 = (df_original['Diabetes_012'] == 1)

condicion2 = (df_original['Diabetes_012'] == 2)

condiciones = condicion1 | condicion2

df_condicion1=df_original[condicion1].sample(frac=1).reset_index(drop=True)

df_condicion2=df_original[condicion2].sample(frac=.55).reset_index(drop=True)

df_condicion = pd.concat([df_condicion1, df_condicion2],
ignore_index=True)

df_balanceado = pd.concat([df_muestra, df_condicion],
ignore_index=True)

df_balanceado.to_csv('muestraBalanceadaD012.csv', index=False)

```

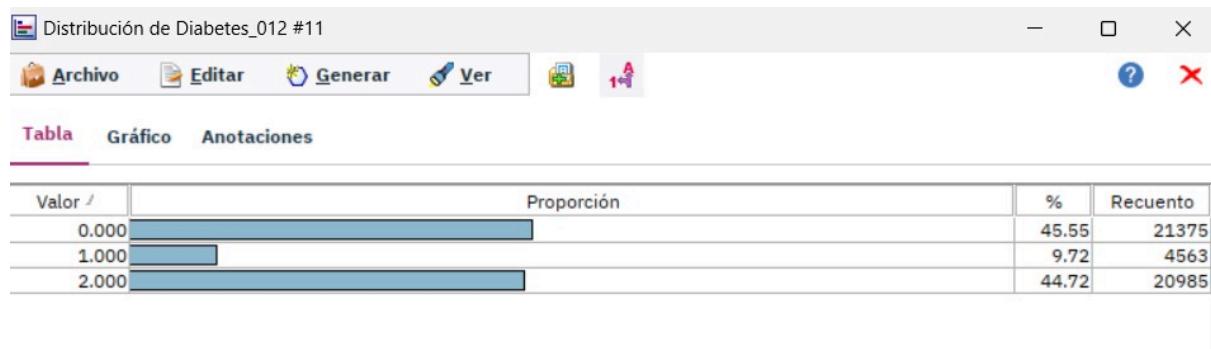
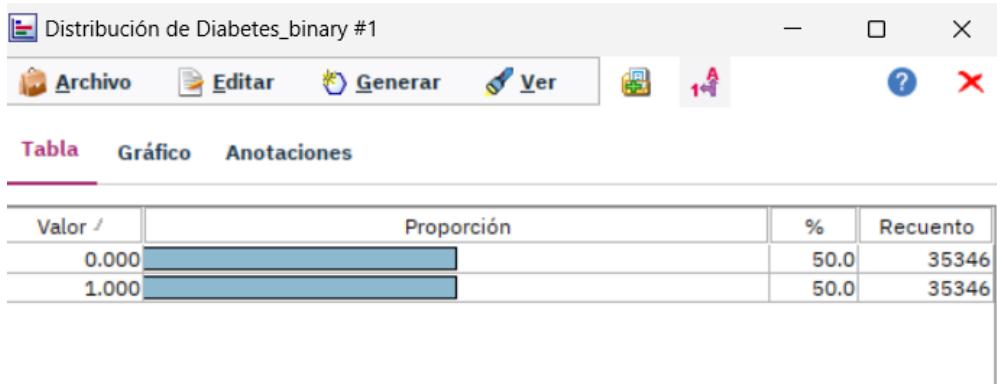


Imagen 6. Recuento de datos con una primera instancia de balanceo.

Con el fin de trabajar con un DataSet más balanceado, los creadores de este DataSet ya habían hecho un versión del dataSet balanceada y que cambia al campo Diabetes\_012 en el que se indica si un paciente tiene diabetes (2), prediabetes (1) o ninguno (0), a un campo binario en el que la presencia de diabetes o prediabetes se representa con 1, mientras que la ausencia de estas afecciones se representa con 0. Sin embargo, sólo presentamos este DataSet como alternativa, más no como una solución final al desbalance. A continuación presentamos la distribución de este DataSet:



#### - Balanceo HeartDiseaseorAttack

Así mismo, para el campo de HeartDiseaseorAttack, tenemos un desbalance tal que la ocurrencia de no presentar afecciones cardiacas abarca un poco más del 90% de las ocurrencias.



Imagen 7. Recuento de datos Desbalanceado

Para corregir esto, también utilizamos un script de Python con la biblioteca “pandas”, con el que agregamos registros aleatoriamente y sin repetición del 90% restante del DataSet original.

```
import pandas as pd

df_original = pd.read_csv('DataSet90restante.csv')

df_muestra = pd.read_csv('muestra10.csv')

condicion1 = (df_original['HeartDiseaseorAttack'] == 1)

df_condicion=df_original[condicion1].sample(frac=.95).reset_index(drop=True)
```

```

df_balanceado= pd.concat([df_muestra, df_condicion], ignore_index=True)

df_balanceado.to_csv('muestraBalanceadaInfarto.csv', index=False)

```

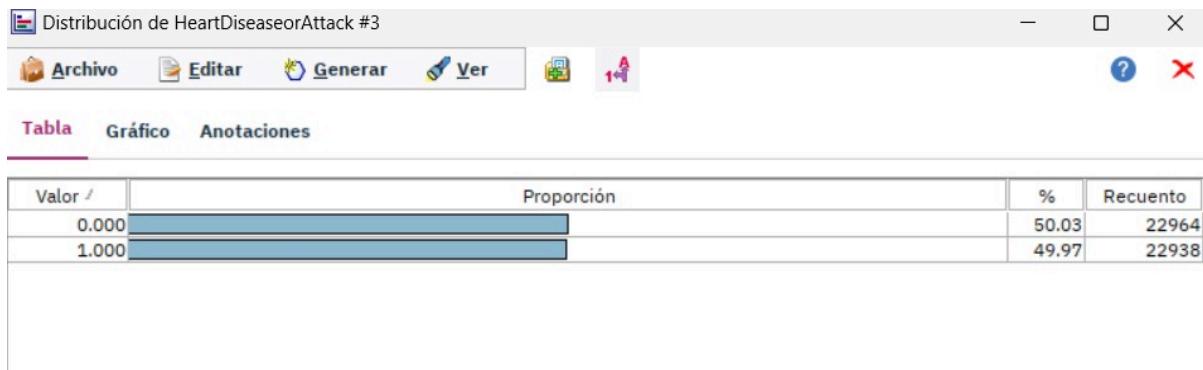


Imagen 8. Recuento de datos Balanceado.

## Construcción de nuevos datos

No consideramos necesaria la integración de nuevos campos, pues creemos que con los campos dados en el DataSet original es suficiente para cumplir con los objetivos de minería de datos propuestos en este trabajo.

### *Suficiencia de Atributos Actuales:*

- Los 21 atributos actuales en el DataSet abarcan una amplia gama de información relacionada con la salud y el estilo de vida.
- Cada atributo, desde la presión arterial hasta el índice de masa corporal (BMI), ofrece una visión integral que puede contribuir a predecir la diabetes.

### *Integridad de los Datos Existentes:*

- El análisis de calidad de datos no reveló errores ni valores faltantes en las 253,680 tuplas del conjunto actual.
- La información presente se considera completa y confiable para respaldar los objetivos del proyecto.

### *Complejidad Potencial de Nuevos Campos:*

- La introducción de nuevos campos podría aumentar la complejidad del análisis sin proporcionar beneficios sustanciales.

- Dada la diversidad de atributos existentes, la adición de más variables podría no ser necesaria para cumplir con los objetivos.

#### *Optimización de Recursos:*

- Al no construir nuevos datos, se optimizan recursos en términos de tiempo y esfuerzo, centrándose en la utilidad de los datos actuales.
- Se evitan posibles complicaciones innecesarias al mantener un enfoque claro en los factores ya identificados como esenciales.

#### *Claridad en el análisis*

- La decisión se toma para ser eficientes, enfocándose en la información existente y evitando introducir complejidades que podrían no agregar valor significativo.
- Se busca mantener la claridad en el análisis, asegurando que los recursos se utilicen de manera efectiva para lograr los objetivos del proyecto.

Además de los puntos mencionados, es importante destacar que la decisión de no construir nuevos datos se alinea con la filosofía de maximizar la utilidad de los recursos disponibles. Al evitar la introducción de campos adicionales, se busca simplificar el proceso de análisis y modelado, reduciendo la complejidad innecesaria.

Al mantener la concentración en los datos actuales, se reduce el riesgo de posibles errores introducidos al construir nuevos campos. Esta aproximación busca equilibrar la eficiencia con la integridad de los resultados finales, asegurando que el análisis sea claro, preciso y alineado con los objetivos del proyecto.

#### **Integración de datos**

Decidimos integrar registros del 90% restante del DataSet original para balancear las ocurrencias de las clases que protagonizan nuestros objetivos de minería de datos. Esto para evitar perder registros al utilizar el nodo “equilibrado” proporcionado por la herramienta IBM SPSS Modeler.

Para el objetivo de clasificación de Diabetes\_012, una vez equilibrado con el script de python, nos hemos quedado con 46,923 registros y con los 21 predictores originales. Por otro lado, usando el DataSet equilibrado proporcionado por los creadores del DataSet, nos quedamos con 70, 692 registros con los 21 predictores originales.

Para el objetivo de clasificación de HeartDiseaseorAttack, una vez equilibrado con el script de python, nos hemos quedado con 45,902 registros con los predictores originales.

## **Formato de datos**

A continuación, presentamos el formato de los datos que se utilizaran en la fase de Modelado.

### **1. Datos Ordinales:**

- 1.1. Diabetes\_012
- 1.2. GenHlth
- 1.3. MentHlth
- 1.4. PhysHlth
- 1.5. Age
- 1.6. Education
- 1.7. Income

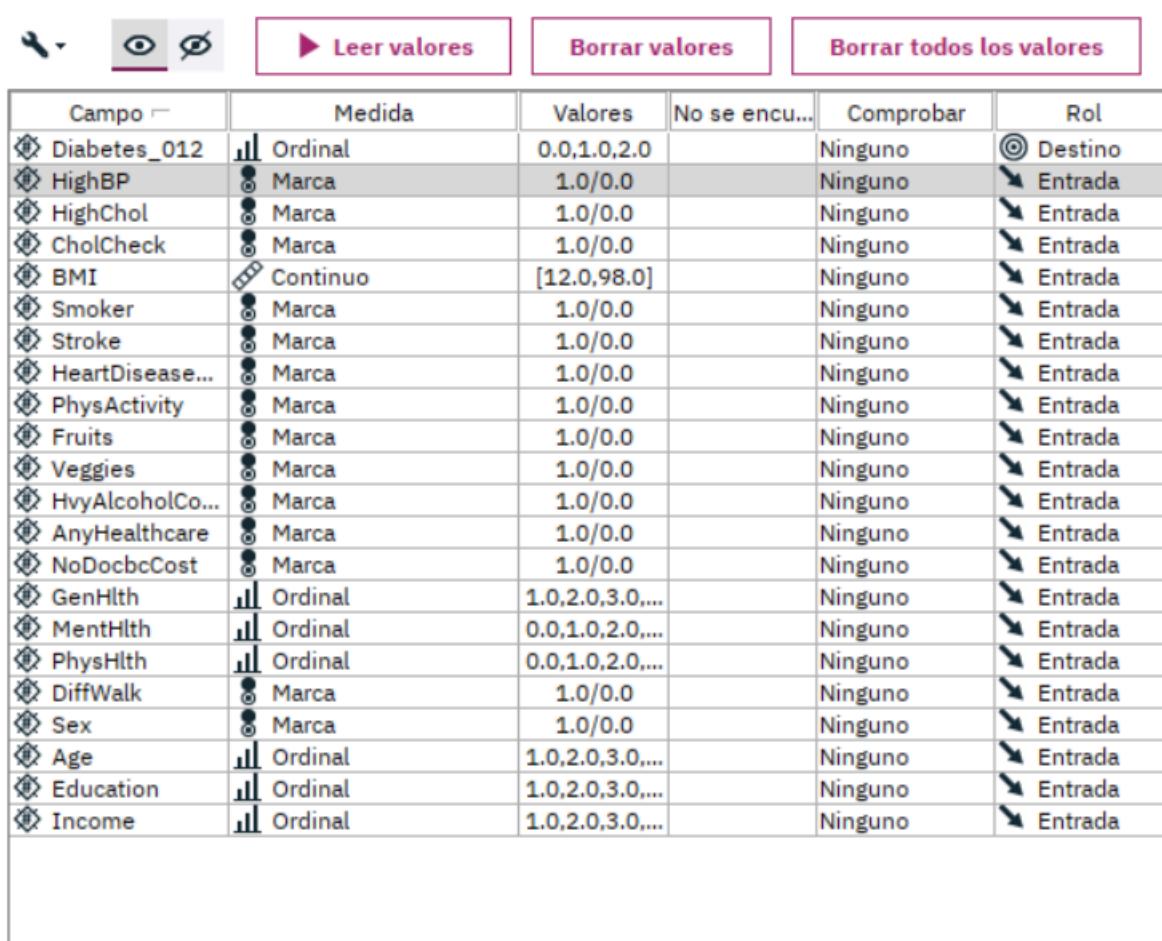
### **2. Datos Categóricos:**

- 2.1. HighBP
- 2.2. HighChol
- 2.3. CholCheck
- 2.4. Smoker
- 2.5. Stroke
- 2.6. HearthDiseaseOrAttack
- 2.7. PhysActivity
- 2.8. Fruits
- 2.9. Veggies
- 2.10. HvyAlcoholConsump
- 2.11. AnyHealthCare
- 2.12. NoDocbcCost
- 2.13. DiffWalk
- 2.14. Sex

### **3. Datos Continuos:**

- 3.1. BMI

A continuación, presentamos la imagen de la tabla obtenida de la herramienta IBM SPSS Modeler en la que cada campo adquiere su formato apropiado para la fase de Modelado:



The screenshot shows a table of field properties in the IBM SPSS Modeler interface. The table has columns for Campo, Medida, Valores, No se encu..., Comprobar, and Rol. The rows list various fields with their corresponding properties. The 'Medida' column includes icons for Ordinal (bar chart), Marca (radio buttons), and Continuo (thermometer). The 'Valores' column contains specific value ranges or counts. The 'Comprobar' column shows 'Ninguno' for most fields. The 'Rol' column indicates roles: 'Destino' for Diabetes\_012, and 'Entrada' for all other fields.

Campo	Medida	Valores	No se encu...	Comprobar	Rol
Diabetes_012	Ordinal	0.0,1,0,2,0		Ninguno	Destino
HighBP	Marca	1,0/0,0		Ninguno	Entrada
HighChol	Marca	1,0/0,0		Ninguno	Entrada
CholCheck	Marca	1,0/0,0		Ninguno	Entrada
BMI	Continuo	[12,0,98,0]		Ninguno	Entrada
Smoker	Marca	1,0/0,0		Ninguno	Entrada
Stroke	Marca	1,0/0,0		Ninguno	Entrada
HeartDisease...	Marca	1,0/0,0		Ninguno	Entrada
PhysActivity	Marca	1,0/0,0		Ninguno	Entrada
Fruits	Marca	1,0/0,0		Ninguno	Entrada
Veggies	Marca	1,0/0,0		Ninguno	Entrada
HvyAlcoholCo...	Marca	1,0/0,0		Ninguno	Entrada
AnyHealthcare	Marca	1,0/0,0		Ninguno	Entrada
NoDocbcCost	Marca	1,0/0,0		Ninguno	Entrada
GenHlth	Ordinal	1,0,2,0,3,0,...		Ninguno	Entrada
MentHlth	Ordinal	0,0,1,0,2,0,...		Ninguno	Entrada
PhysHlth	Ordinal	0,0,1,0,2,0,...		Ninguno	Entrada
DiffWalk	Marca	1,0/0,0		Ninguno	Entrada
Sex	Marca	1,0/0,0		Ninguno	Entrada
Age	Ordinal	1,0,2,0,3,0,...		Ninguno	Entrada
Education	Ordinal	1,0,2,0,3,0,...		Ninguno	Entrada
Income	Ordinal	1,0,2,0,3,0,...		Ninguno	Entrada

Modelado.

- Modelado para la clasificación de diabetes.

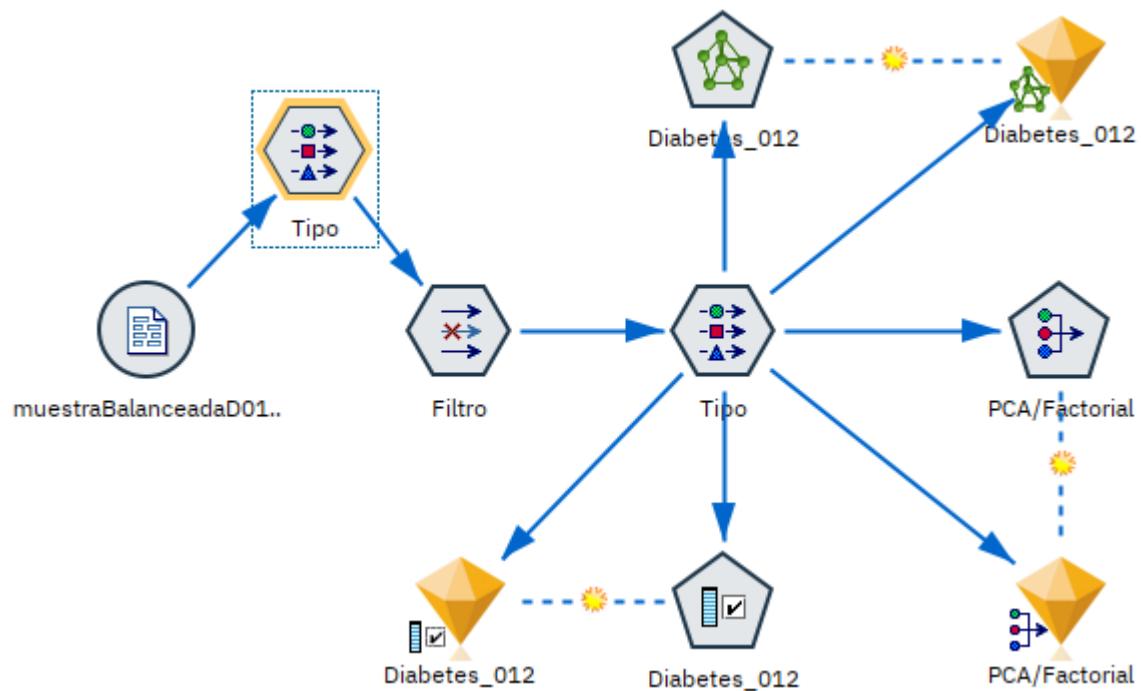


Imagen : modelado.

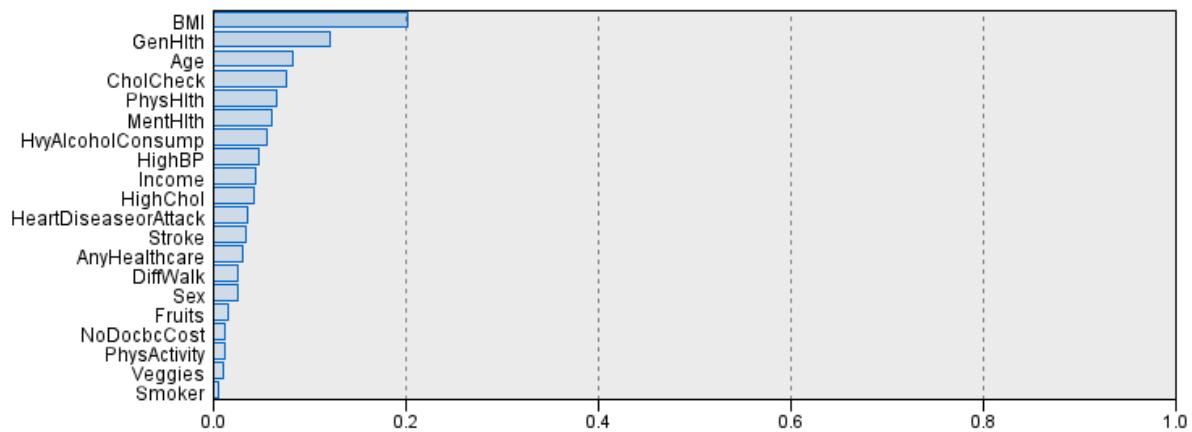


Imagen : importancia de los predictores usando la red neuronal.

	Clasificación /	Campo	Medida	Importancia	Valor
✓	1	GenHlth	Ordinal	Importante	1.0
✓	2	HighBP	Marca	Importante	1.0
✓	3	Age	Ordinal	Importante	1.0
✓	4	HighChol	Marca	Importante	1.0
✓	5	DiffWalk	Marca	Importante	1.0
✓	6	Income	Ordinal	Importante	1.0
✓	7	PhysHlth	Ordinal	Importante	1.0
✓	8	BMI	Continuo	Importante	1.0
✓	9	HeartDiseaseorAtt...	Marca	Importante	1.0
✓	10	PhysActivity	Marca	Importante	1.0
✓	11	MentHlth	Ordinal	Importante	1.0
✓	12	Smoker	Marca	Importante	1.0
✓	13	Veggies	Marca	Importante	1.0
✓	14	Fruits	Marca	Importante	1.0
✓	15	Sex	Marca	Importante	1.0

Campos seleccionados: 15    Total de campos disponibles: 20

★ > 0.95    + <= 0.95    □ < 0.9

5 Campos representados

	Campo	Medida	Motivo
✓	Stroke	Marca	Categoría única demasiado grande
✓	NoDocbc...	Marca	Categoría única demasiado grande
✓	HvyAlcoh...	Marca	Categoría única demasiado grande
✓	CholCheck	Marca	Categoría única demasiado grande
✓	AnyHealt...	Marca	Categoría única demasiado grande

Imagen : importancia de los predictores usando la selección de características.

Comunalidades		
	Inicial	Extracción
HighBP	1,000	,514
HighChol	1,000	,361
CholCheck	1,000	,198
BMI	1,000	,311
Smoker	1,000	,421
Stroke	1,000	,214
HeartDiseaseorAttack	1,000	,399
PhysActivity	1,000	,305
Fruits	1,000	,555
Veggies	1,000	,459
HvyAlcoholConsump	1,000	,226
AnyHealthcare	1,000	,621
NoDocbcCost	1,000	,494
GenHlth	1,000	,614
MentHlth	1,000	,477
PhysHlth	1,000	,620
DiffWalk	1,000	,549
Sex	1,000	,454
Age	1,000	,542
Income	1,000	,428

Método de extracción: análisis de componentes principales.

Imagen : importancia de los predictores usando la PCA.

Comparando los resultados de los tres modelos anteriores llegamos a la conclusión de que los 5 campos más importantes son: "BMI", "GenHlth", "Age", "PhysHlth", "MentHlth".

Y los 5 campos con menor impacto como predictores son: "stroke", "HvyAlcoholConsump", "NoDocbcCost", "sex".

- Modelado para la clasificación de ataque o enfermedad cardiaca.

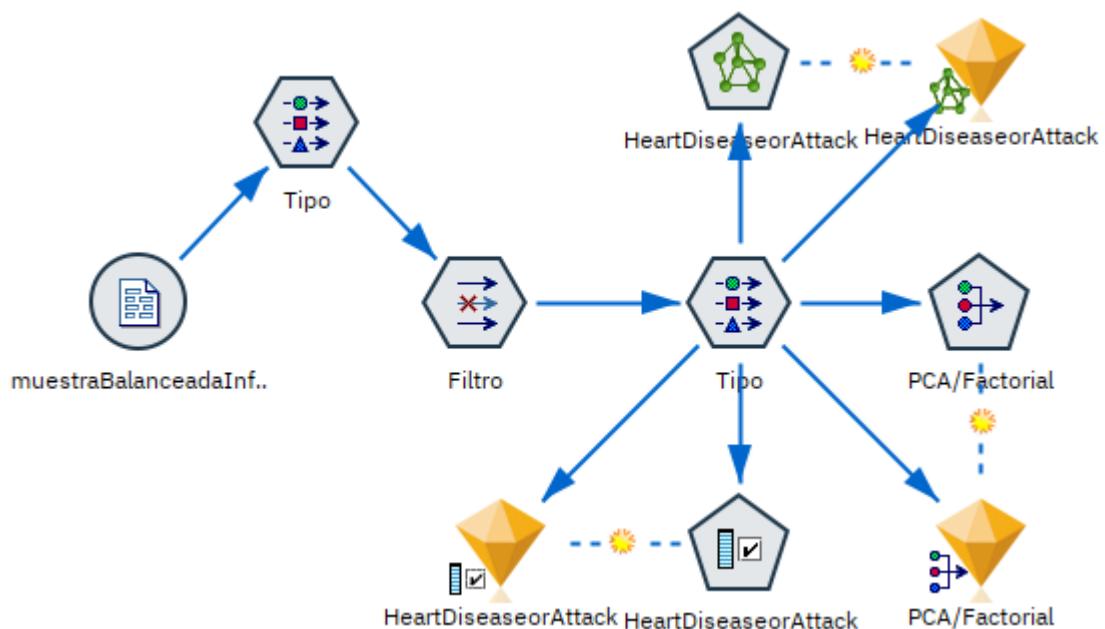


Imagen : modelado.

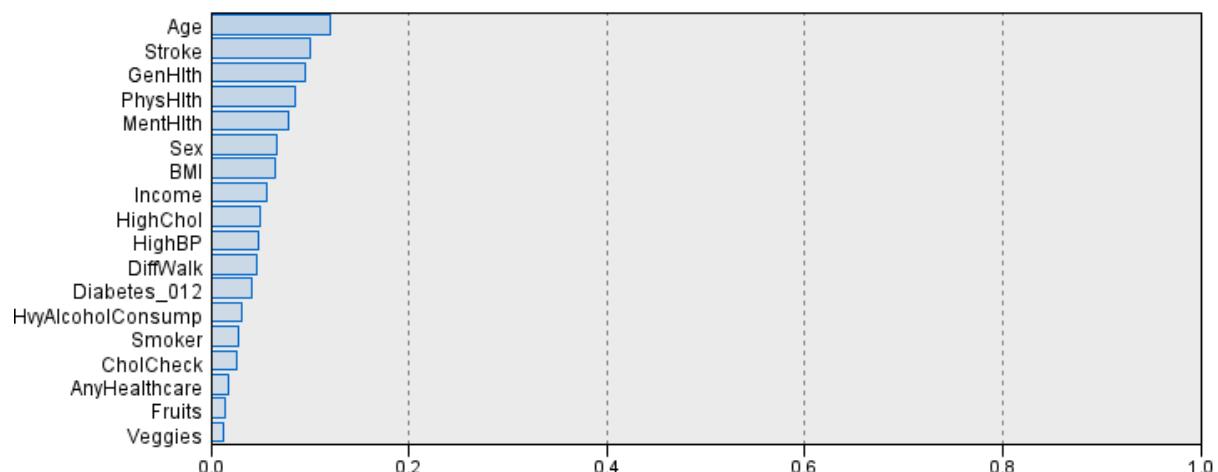


Imagen : importancia de los predictores usando la red neuronal.

	Clasificación /	Campo	Medida	Importancia	Valor
✓		1 GenHlth	Ordinal	Importante	1.0
✓		2 Age	Ordinal	Importante	1.0
✓		3 HighBP	Marca	Importante	1.0
✓		4 HighChol	Marca	Importante	1.0
✓		5 DiffWalk	Marca	Importante	1.0
✓		6 PhysHlth	Ordinal	Importante	1.0
✓		7 Diabetes_0...	Ordinal	Importante	1.0
✓		8 Income	Ordinal	Importante	1.0
✓		9 Smoker	Marca	Importante	1.0
✓		10 Sex	Marca	Importante	1.0
✓		11 PhysActivity	Marca	Importante	1.0
✓		12 MentHlth	Ordinal	Importante	1.0
✓		13 BMI	Continuo	Importante	1.0
✓		14 Veggies	Marca	Importante	1.0
✓		15 Fruits	Marca	Importante	1.0

Campos seleccionados: 15    Total de campos disponibles: 20

★ > 0.95    + <= 0.95    □ < 0.9

5 Campos representados

	Campo	Medida	Motivo
□	Stroke	Marca	Categoría única demasiado grande
□	NoDocbc...	Marca	Categoría única demasiado grande
□	HvyAlcoh...	Marca	Categoría única demasiado grande
□	CholCheck	Marca	Categoría única demasiado grande
□	AnyHealt...	Marca	Categoría única demasiado grande

Imagen : importancia de los predictores usando la selección de características.

<b>Comunalidades</b>		
	Inicial	Extracción
Diabetes_012	1,000	,435
HighBP	1,000	,501
HighChol	1,000	,390
CholCheck	1,000	,172
BMI	1,000	,612
Smoker	1,000	,461
Stroke	1,000	,204
PhysActivity	1,000	,328
Fruits	1,000	,586
Veggies	1,000	,543
HvyAlcoholConsump	1,000	,244
AnyHealthcare	1,000	,439
NoDocbcCost	1,000	,451
GenHlth	1,000	,632
MentHlth	1,000	,385
PhysHlth	1,000	,546
DiffWalk	1,000	,556
Sex	1,000	,471
Age	1,000	,582
Income	1,000	,423

Imagen : importancia de los predictores usando la PCA.

Comparando los resultados de los tres modelos anteriores llegamos a la conclusión de que los 5 campos más importantes son: "BMI", "GenHlth", "PhysHlth", "HighBP", "age".

Y los 2 campos con menor impacto como predictores son: "CholCheck", "HvyAlcoholConsump".

## Modelado.

### 1. Selección de los modelos

- Red Neuronal MLP
- Red Neuronal LSVM
- Red Neuronal SVM
- Red Neuronal Bayesiana
- Algoritmo KNN
- Algoritmo de Decisión C5.0
- Algoritmo de Decisión CRT
- Algoritmo de Decisión CHAID
- Regresión Logística
- Árboles Aleatorios

#### **Red Neuronal MLP (Perceptrón Multicapa):**

Definición: Una red neuronal artificial con al menos tres capas: una capa de entrada, una o más capas ocultas y una capa de salida. Cada capa está compuesta por nodos que utilizan funciones de activación no lineales.

Tipo de uso: Clasificación y regresión en problemas complejos, como reconocimiento de imágenes, procesamiento de lenguaje natural y predicción de series temporales.

#### **Red Neuronal LSVM (Máquina de Soporte Vectorial con Regresión Lineal):**

Definición: Una variante de las máquinas de vectores de soporte que se utiliza para problemas de regresión lineal, donde se busca encontrar la línea que mejor se ajusta a los datos.

Tipo de uso: Predicción de valores numéricos en problemas de regresión lineal.

#### **Red Neuronal SVM (Máquina de Soporte Vectorial):**

Definición: Un algoritmo de aprendizaje supervisado que se utiliza tanto para clasificación como para regresión. Busca encontrar el hiperplano que mejor separa las clases en el espacio de características.

Tipo de uso: Clasificación y regresión en problemas donde las clases o los valores a predecir son linealmente separables.

### **Red Neuronal Bayesiana:**

Definición: Un modelo que combina la teoría de la probabilidad bayesiana con las redes neuronales, permitiendo la incorporación de incertidumbre en los parámetros del modelo.

Tipo de uso: Clasificación y regresión, especialmente en problemas donde se necesita modelar la incertidumbre de los datos y parámetros.

### **Algoritmo KNN (K Vecinos más Cercanos):**

Definición: Un método de clasificación o regresión que asigna a un punto el valor o la clase de la mayoría de sus k vecinos más cercanos en el espacio de características.

Tipo de uso: Clasificación y regresión en problemas donde la estructura local de los datos es importante y no se conocen a priori las distribuciones de las clases.

### **Algoritmo de Decisión C5.0:**

Definición: Un algoritmo de aprendizaje de árboles de decisión que utiliza una serie de reglas para dividir el conjunto de datos en subconjuntos homogéneos en términos de la variable objetivo.

Tipo de uso: Clasificación y regresión en problemas donde se pueden representar las decisiones como reglas condicionales.

### **Algoritmo de Decisión CRT (Árbol de Regresión de Clasificación):**

Definición: Una variante de los árboles de decisión que se utiliza específicamente para problemas de clasificación, donde cada nodo interno representa una regla de decisión.

Tipo de uso: Clasificación en problemas donde se busca una representación interpretable de las decisiones.

### **Algoritmo de Decisión CHAID:**

Definición: Un algoritmo de construcción de árboles de decisión que utiliza análisis de varianza para determinar las divisiones óptimas en cada nodo del árbol.

Tipo de uso: Clasificación en problemas donde se necesitan árboles de decisión jerárquicos y se desea tener en cuenta la interacción entre las variables predictoras.

### **Regresión Logística:**

Definición: Un modelo estadístico que se utiliza para modelar la probabilidad de que una variable dependiente binaria tenga éxito en función de una o más variables independientes.

Tipo de uso: Clasificación en problemas binarios donde se requiere una probabilidad como salida.

### **Árboles Aleatorios:**

Definición: Un conjunto de árboles de decisión construidos de forma aleatoria, donde cada árbol vota por la clase más popular.

Tipo de uso: Clasificación y regresión en problemas donde se busca reducir la varianza y mejorar la precisión predictiva en comparación con un solo árbol de decisión.

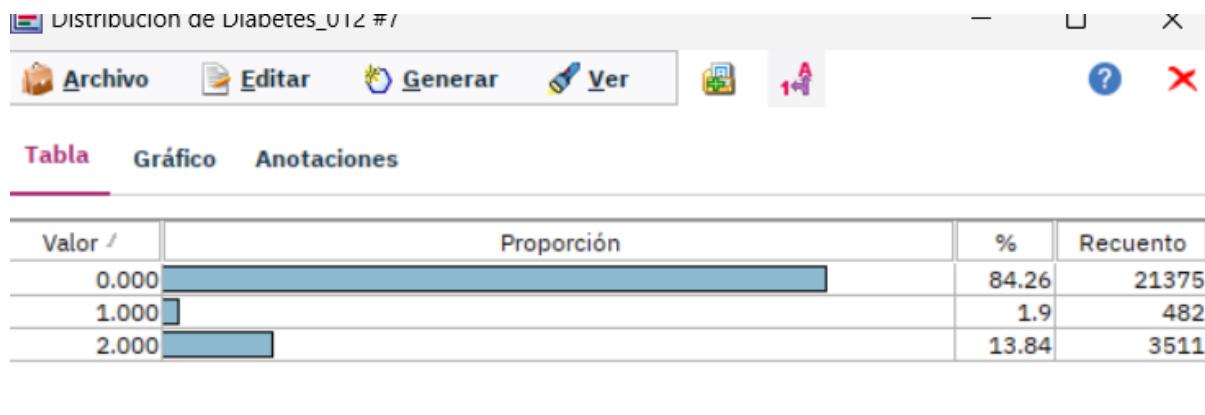
### **Métodos de comprobación.**

Dado a que se usarán modelos supervisados, se considera a la bondad como criterio para determinar la calidad de los modelos generados.

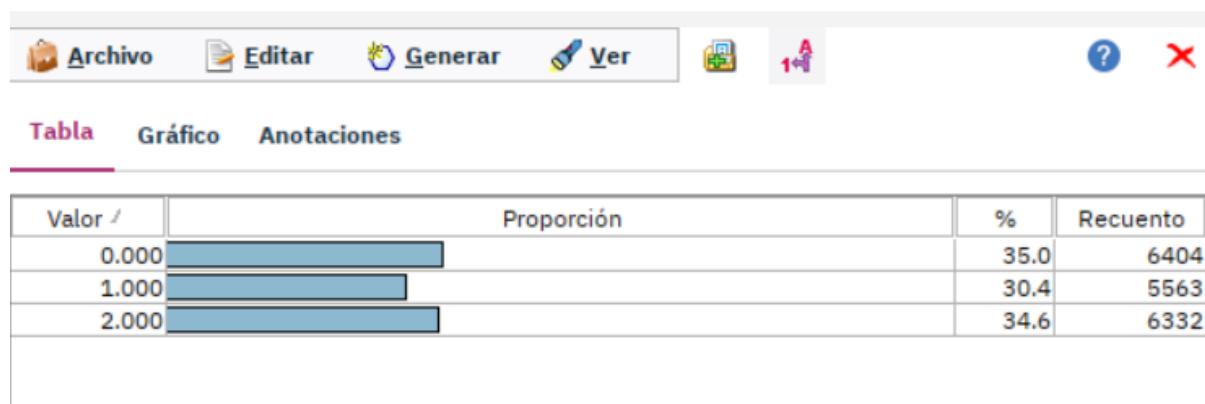
Así mismo, se utilizará el DataSet restante para poner a prueba la clasificación de los modelos, tal que comprobaremos si estos siguen las tendencias de ocurrencias en los atributos objetivo.

# Objetivo de Minería de Datos - Clasificación de Diabetes\_012

## Ocurrencias del atributo destino en DataSet desbalanceado



## Ocurrencias de la clase destino en DataSet Balanceado



## Modelado desbalanceado

Obtención de matriz de confusión en base a la evaluación de cada modelo.

MLP

	Negativo1	Negativo2	Positivo
Negativo1	21076	0	296
Negativo2	455	0	27
Positivo	3072	0	438

LSVM

	Negativo1	Negativo2	Positivo
Negativo1	211300	0	245
Negativo2	450	0	32
Positivo	3092	0	419

SVM

	Negativo1	Negativo2	Positivo
Negativo1	21160	2	213
Negativo2	91	385	6
Positivo	850	4	2657

Red Bayesiana

	Negativo1	Negativo2	Positivo
Negativo1	19769	3	1603
Negativo2	347	3	132
Positivo	2152	1	1358

### C5.0

	Negativo1	Negativo2	Positivo
Negativo1	21000	0	375
Negativo2	428	0	54
Positivo	2896	0	615

### CRT

	Negativo1	Negativo2	Positivo
Negativo1	20963	0	412
Negativo2	432	0	50
Positivo	2964	0	547

### Logistica

	Negativo1	Negativo2	Positivo
Negativo1	20903	0	472
Negativo2	434	0	48
Positivo	2869	0	642

### Algoritmo KNN

	Negativo1	Negativo2	Positivo
Negativo1	21146	5	224
Negativo2	394	16	72
Positivo	2630	10	871

## CHAID

	Negativo1	Negativo2	Positivo
Negativo1	21199	0	176
Negativo2	470	0	12
Positivo	3290	0	221

## Arboles Aleatorios

	Negativo1	Negativo2	Positivo
Negativo1	15581	1556	4238
Negativo2	71	335	76
Positivo	549	311	2651

## Métricas Desbalanceada

Se ocuparon las métricas de exactitud, precisión, sensibilidad y especificidad de cada uno de los modelos para poder evaluar su eficiencia.

Modelo	Exactitud(bonda d)	Precisió n	Sensibili dad	Especifici dad
Red Neuronal (MLP)	0.848210062	0.575558 48	0.110466 58	0.9849058 37
Red Neuronal(LSVM)	0.982281547	0.602011 49	0.105781 37	0.9986907 84
Red Neuronal (SVM)	0.954036582	0.923852 57	0.737236 4	0.9899375 11

Red Neuronal Bayesiana	0.832939136	0.439055 93	0.351722 35	0.9193285 91
Algoritmo KNN	0.868535162	0.746358 18	0.222762 15	0.9862056 11
Algoritmo de decisión C5.0	0.852057711	0.589080 46	0.156131	0.9799804
Arbol de decision CRT	0.847918638	0.542120 91	0.138726 86	0.9784364 06
Árbol de decisión CHAID	0.844370861	0.540342 3	0.055513 69	0.9912096 13
Árboles aleatorios	0.731906339	0.380617 37	0.515959 52	0.7867523 48
Regresión Logística	0.849298329	0.552495 7	0.162737 64	0.9757270 22

## Modelado sin selección de Atributos (balanceado)

Obtención de matriz de confusión en base a la evaluación de cada modelo.

### LSVM

	Negativo1	Negativo2	Positivo
Negativo1	4322	706	1376
Negativo2	1615	1322	2626
Positivo	1152	1001	4179

### Arboles aleatorios

	Negativo1	Negativo2	Positivo
Negativo1	4397	894	1113
Negativo2	1053	2767	1743
Positivo	942	1096	4294

### Red Bayesiana

	Negativo1	Negativo2	Positivo
Negativo1	4304	728	1372
Negativo2	1724	1272	2567
Positivo	1221	873	4238

### C5.0

	Negativo1	Negativo2	Positivo
Negativo1	4624	810	970
Negativo2	960	3017	1586
Positivo	762	1070	4500

CRT

	Negativo1	Negativo2	Positivo
Negativo1	4612	281	1511
Negativo2	2120	522	2921
Positivo	1539	466	4327

Logistica

	Negativo1	Negativo2	Positivo
Negativo1	4379	699	1326
Negativo2	1656	1281	2626
Positivo	1190	978	4164

MLP

	Negativo1	Negativo2	Positivo
Negativo1	4322	624	1458
Negativo2	1699	867	2997
Positivo	1146	724	4462

SVM

	Negativo1	Negativo2	Positivo
Negativo1	5821	223	360
Negativo2	386	4812	365
Positivo	353	250	5729

## Algoritmo KNN

	Negativo1	Negativo2	Positivo
Negativo1	5064	652	688
Negativo2	1030	3162	1371
Positivo	1134	733	4465

## CHAID

	Negativo1	Negativo2	Positivo
Negativo1	3963	810	1631
Negativo2	1476	1111	2976
Positivo	1014	872	4446

## Métricas sin selección de Atributos (balanceado)

Se ocuparon las métricas de exactitud, precisión, sensibilidad y especificidad de cada uno de los modelos para poder evaluar su eficiencia.

Modelo	Exactitud(bondad)	Precisión	Sensibilidad	Especificidad
Red Neuronal (MLP)	0.518486485	0.50217969	0.56046042	0.9991537
Red Neuronal(LSVM)	0.532050932	0.50681133	0.48455509	0.574872701
Red Neuronal (SVM)	0.853981092	0.85121913	0.76177901	0.913724784
Red Neuronal Bayesiana	0.531832341	0.51253448	0.48689608	0.573243726
Algoritmo KNN	0.694081644	0.68304593	0.55970989	0.798388662

Algoritmo de decisión C5.0	0.641237226	0.59267648	0.58596666	0.684533671
Arbol de decision CRT	0.517022788	0.49400617	0.49547693	0.536692452
Árbol de decisión CHAID	0.520247008	0.49110792	0.51589696	0.524119409
Árboles aleatorios	0.621454724	0.59958275	0.51573155	0.710362173
Regresión Logística	0.532761353	0.50924799	0.48092393	0.57976451

- **Diabetes balanceado sin stroke, HvyAlcoholConsump**

Se ocuparon las métricas de exactitud, precisión, sensibilidad y especificidad de cada uno de los modelos para poder evaluar su eficiencia.

Modelo	Exactitud(bondad)	Precisión	Sensibilidad	Especificidad
Red Neuronal (MLP)	0.523088693	0.48504644	0.56343372	0.488596047
Red Neuronal(LSVM)	0.534728674	0.50744462	0.48535371	0.578865659
Red Neuronal (SVM)	0.884583857	0.88017362	0.80917771	0.931483779
Red Neuronal Bayesiana	0.534510083	0.51438588	0.48595249	0.579063188
Algoritmo KNN	0.662221979	0.65682826	0.52174441	0.777093474

Algoritmo de decisión C5.0	0.651620307	0.6183268	0.56300763	0.722402438
Arbol de decision CRT	0.517022788	0.49400617	0.49547693	0.536692452
Árbol de decisión CHAID	0.520247008	0.4911 0792	0.515896 96	0.5241 19409
Árboles aleatorios	0.624187114	0.6047 5914	0.500901 33	0.7269 99399
Regresión Logística	0.535275152	0.5104 0392	0.480747 06	0.5844 15584

- **Diabetes balanceado sin stroke, heavyAlcoholCons, NoDcbCost**

Se ocuparon las métricas de exactitud, precisión, sensibilidad y especificidad de cada uno de los modelos para poder evaluar su eficiencia.

Modelo	Exactitud(bondad)	Precisión	Sensibilidad	Especificidad
Red Neuronal (MLP)	0.523689819	0.48480448	0.55490918	0.497063588
Red Neuronal(LSVM)	0.533854309	0.50739449	0.48769213	0.575264358

Red Neuronal (SVM)	0.878408656	0.87346875	0.7999148	0.92751177
Red Neuronal Bayesiana	0.534127548	0.51540785	0.4857631	0.578737262
Algoritmo KNN	0.692660801	0.68095819	0.55815408	0.797049403
Algoritmo de decisión C5.0	0.644297503	0.60280068	0.5700495	0.703004208
Arbol de decision CRT	0.517022788	0.49400617	0.49547693	0.536692452
Árbol de decisión CHAID	0.520247008	0.49110792	0.51589696	0.524119409
Árboles aleatorios	0.632548227	0.6068424	0.52473327	0.721022784
Regresión Logística	0.534728674	0.51048269	0.48215519	0.582232394

## Modelado con Selección de Atributos

Obtención de matriz de confusión en base a la evaluación de cada modelo con un DataSet balanceado y quitando los atributos de menor impacto.

- Objetivo Diabetes(Balanceado sin stroke, hvyalcohol, NoDocbcCost, sex)

LSVM

	Negativo1	Negativo2	Positivo
Negativo1	4325	664	1415
Negativo2	1680	1207	2676
Positivo	1173	955	4204

Arboles aleatorios

	Negativo1	Negativo2	Positivo
Negativo1	4404	903	1097
Negativo2	1124	2657	1782
Positivo	936	1085	4311

Red Bayesiana

	Negativo1	Negativo2	Positivo
Negativo1	4264	718	1422
Negativo2	1726	1195	2642
Positivo	1230	829	4273

C5.0

	Negativo1	Negativo2	Positivo
Negativo1	4462	754	1188
Negativo2	952	2562	2049
Positivo	787	835	4710

CRT

	Negativo1	Negativo2	Positivo
Negativo1	4612	281	1511
Negativo2	2120	522	2921
Positivo	1539	466	4327

Logistica

	Negativo1	Negativo2	Positivo
Negativo1	4374	662	1368
Negativo2	1708	1190	2665
Positivo	1209	938	4185

MLP

	Negativo1	Negativo2	Positivo
Negativo1	4329	390	1685
Negativo2	1705	5523306	2997
Positivo	1161	448	4723

SVM

	Negativo1	Negativo2	Positivo
Negativo1	5606	321	477
Negativo2	542	4540	481
Positivo	482	369	5481

### Algoritmo KNN

	Negativo1	Negativo2	Positivo
Negativo1	5092	622	690
Negativo2	1043	3133	1387
Positivo	1125	731	4476

### CHAID

	Negativo1	Negativo2	Positivo
Negativo1	3963	810	1631
Negativo2	1476	1111	2976
Positivo	1014	872	4446

### Métricas con selección de Atributos

- Objetivo Diabetes(Balanceado sin stroke, hvyalcohol, NoDocbcCost, sex)

Se ocuparon las métricas de exactitud, precisión, sensibilidad y especificidad de cada uno de los modelos para poder evaluar su eficiencia.

Modelo	Exactitud(bonda d)	Precisió n	Sensibili dad	Especifici dad
Red Neuronal (MLP)	0.527405869	0.50039251	0.51554015	0.538054749

Red Neuronal(LSVM)	0.53680529	0.510817 75	0.482953 89	0.585113
Red Neuronal (SVM)	0.894147221	0.887666 56	0.825385 39	0.9361683 39
Red Neuronal Bayesiana	0.53631346	0.518282 99	0.482468 12	0.5860220 7
Algoritmo KNN	0.693535166	0.684396 08	0.557149 99	0.7998055 42
Algoritmo de decisión C5.0	0.663478879	0.637755 1	0.555418 42	0.7493380 41
Arbol de decision CRT	0.517022788	0.494006 17	0.495476 93	0.5366924 52
Arbol de decisión CHAID	0.520247008	0.491107 92	0.515896 96	0.5241194 09
Árboles aleatorios	0.626154435	0.600559 44	0.518661 67	0.7149700 6
Regresión Logística	0.536859938	0.513060 62	0.479336 94	0.5888472 74

## Mejores Modelos

- Red neuronal SVM - 89%

- Algoritmo KNN - 69%
- Algoritmo C5.0 - 66%

## Propuesta - Fusión de las clase prediabetes (1) y Diabetes (2)

Dada la baja cantidad de personas prediabeticas y los resultados no tan buenos en la mayoría de los modelos al utilizar una muestra balanceada, proponemos integrar la clase de prediabetes (1) con la clase de Diabetes (2), dando como resultado el atributo Diabetes\_binary, el cual representa con uno si la persona padece de diabetes o prediabetes con un 1, caso contrario se representa con un 0.

### Ocurrencias del atributo destino (Diabetes\_binary) en DataSetBalanceado



The screenshot shows a software interface with a menu bar at the top containing 'Archivo', 'Editar', 'Generar', 'Ver', and icons for 'Nuevo', 'Abrir', 'Guardar', 'Imprimir', and 'Ayuda'. Below the menu is a toolbar with buttons for 'Tabla', 'Gráfico', and 'Anotaciones'. The main area displays a table with the following data:

Valor	Proporción	%	Recuento
0.000		50.09	13455
1.000		49.91	13408

### Modelado sin selección de Atributos

- 1. Red neuronal MLP

#### Clasificación para Diabetes\_binary

Porcentaje correcto global = 74.7%

**Clasificación para Diabetes\_binary**

Porcentaje correcto global = 74.7%

		Previsto		Porcentaje de filas
		0.000	1.000	
Observado	0.000	9520	3935	100.00
	1.000	2860	10548	0.00

- 2. Red neuronal LSVM

- 
- Resultados para el campo de resultado Diabetes\_binary
    - Comparando \$L-Diabetes\_binary con Diabetes\_binary

Correctos	20,117	74.89 %
Erróneos	6,746	25.11 %
Total	26,863	
    - Matriz de coincidencias para \$L-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,688	3,767
1.000000	2,979	10,429
    - Evaluación del rendimiento

0.000000	0.423
1.000000	0.387
  
  - 3. Red neuronal SVM

---
  - Resultados para el campo de resultado Diabetes\_binary
    - Comparando \$S-Diabetes\_binary con Diabetes\_binary

Correctos	23,737	88.36 %
Erróneos	3,126	11.64 %
Total	26,863	
    - Matriz de coincidencias para \$S-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	11,823	1,632
1.000000	1,494	11,914
    - Evaluación del rendimiento

0.000000	0.572
1.000000	0.567

---

- 4. Red Bayesiana

■ Resultados para el campo de resultado Diabetes\_binary

■ Comparando \$B-Diabetes\_binary con Diabetes\_binary

Correctos	19,960	74.3 %
Erróneos	6,903	25.7 %
Total	26,863	

■ Matriz de coincidencias para \$B-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,572	3,883
1.000000	3,020	10,388

■ Evaluación del rendimiento

0.000000	0.417
1.000000	0.377

## - 5. Algoritmo KNN

- Contraer todo + Desplegar todo

■ Resultados para el campo de resultado Diabetes\_binary

■ Comparando \$KNN-Diabetes\_binary con Diabetes\_binary

Correctos	21,382	79.6 %
Erróneos	5,481	20.4 %
Total	26,863	

■ Matriz de coincidencias para \$KNN-Diabetes\_binary (las filas muestra

	0.000000	1.000000
0.000000	10,390	3,065
1.000000	2,416	10,992

■ Evaluación del rendimiento

0.000000	0.482
1.000000	0.449

## - 6. Algoritmo De Decisión C5.0

- Contraer todo

+ Desplegar todo

■ Resultados para el campo de resultado Diabetes\_binary

■ Comparando \$C-Diabetes\_binary con Diabetes\_binary

Correctos	20,407	75.97 %
Erróneos	6,456	24.03 %
Total	26,863	

■ Matriz de coincidencias para \$C-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,834	3,621
1.000000	2,835	10,573

■ Evaluación del rendimiento

0.000000	0.438
1.000000	0.4

## - 7. Algoritmo de Decisión CRT

- Contraer todo

+ Desplegar todo

■ Resultados para el campo de resultado Diabetes\_binary

■ Comparando \$R-Diabetes\_binary con Diabetes\_binary

Correctos	19,674	73.24 %
Erróneos	7,189	26.76 %
Total	26,863	

■ Matriz de coincidencias para \$R-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,368	4,087
1.000000	3,102	10,306

■ Evaluación del rendimiento

0.000000	0.405
1.000000	0.361

## - 8. Algoritmo de Decisión CHAID

- Resultados para el campo de resultado Diabetes\_binary

  - Comparando \$R-Diabetes\_binary con Diabetes\_binary

Correctos	19,751	73.52 %
Erróneos	7,112	26.48 %
Total	26,863	

  - Matriz de coincidencias para \$R-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,481	3,974
1.000000	3,138	10,270

  - Evaluación del rendimiento

0.000000	0.405
1.000000	0.368

- 9. Regresión Logística

Contraer todo

Desplegar todo

  - Resultados para el campo de resultado Diabetes\_binary

    - Comparando \$L-Diabetes\_binary con Diabetes\_binary

Correctos	20,122	74.91 %
Erróneos	6,741	25.09 %
Total	26,863	

    - Matriz de coincidencias para \$L-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,760	3,695
1.000000	3,046	10,362

    - Evaluación del rendimiento

0.000000	0.42
1.000000	0.39

- 10. Árboles Aleatorios

- Contraer todo

+ Desplegar todo

■ Resultados para el campo de resultado Diabetes\_binary

■ Comparando \$R-Diabetes\_binary con Diabetes\_binary

Correctos	20,853	77.63 %
Erróneos	6,010	22.37 %
Total	26,863	

■ Matriz de coincidencias para \$R-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,739	3,716
1.000000	2,294	11,114

■ Evaluación del rendimiento

0.000000	0.48
1.000000	0.406

## Métricas Sin Selección de Atributos

Modelo	Exactitud ()	Precisión	Sensibilidad	Especificidad
Red Neuronal MLP	0.7470498 46	0.7283021 47	0.7866945 11	0.707543664
Red Neuronal LSVM	0.7488739 16	0.7346435 62	0.7778192 12	0.720029729
Red Neuronal SVM	0.8836317 61	0.8795216 3	0.8885739 86	0.8787068
Red Neuronal Bayesiana	0.7430294 46	0.7279097 47	0.7747613 37	0.711408398

Algoritmo KNN	0.7960832 5	0.7819591 66	0.8200537 15	0.772203642
---------------	----------------	-----------------	-----------------	-------------

Algoritmo de Decisión C5.0	0.7596694 34	0.7448922 08	0.7885590 69	0.730880713
-------------------------------	-----------------	-----------------	-----------------	-------------

Algoritmo de Decisión CRT	0.7323828 31	0.7160425 21	0.7686455 85	0.696246748
------------------------------	-----------------	-----------------	-----------------	-------------

Algoritmo de Decisión CHAID	0.7352492 28	0.7210053 36	0.7659606 21	0.704645113
--------------------------------	-----------------	-----------------	-----------------	-------------

Regresión Logística	0.7490600 45	0.7371416 38	0.7728221 96	0.725380899
---------------------	-----------------	-----------------	-----------------	-------------

Árboles Aleatorios	0.7762721 96	0.7494268 37	0.8289081 15	0.723820141
--------------------	-----------------	-----------------	-----------------	-------------

---

### Tabla de Valores Matrices de Confusión

---

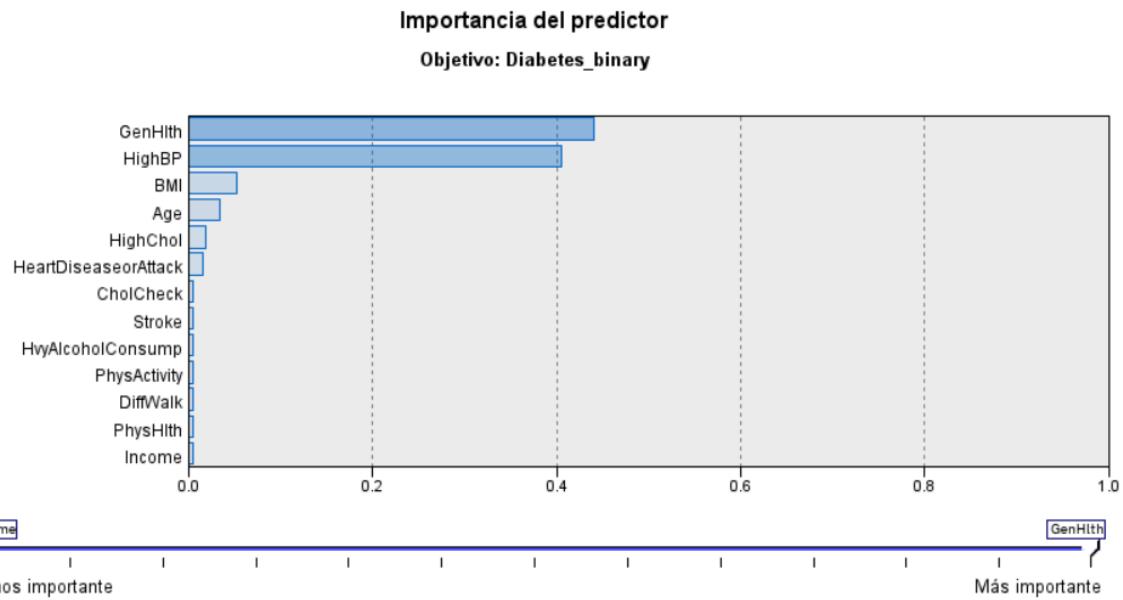
Modelo	TN	FP	FN	TP
RED MLP	9520	3935	2860	10548
LSVM	9688	3767	2979	10429

SVM	11823	1632	1494	11914
Bayesiana	9572	3883	3020	10388
KNN	10390	3065	2412	10992
C5.0	9834	3621	2835	10573
CRT	9368	4087	3102	10306
CHAID	9481	3974	3138	10270
REG LOGISTIC	9760	3695	3046	10362
Árboles Aleatorios	9739	3716	2294	11114

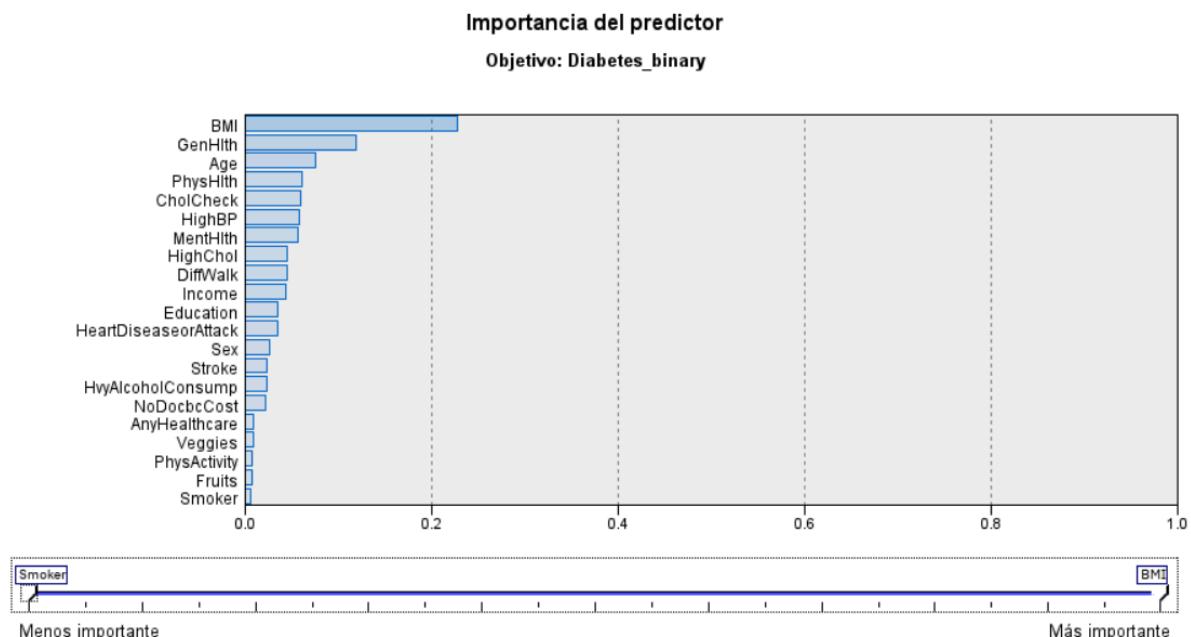
---

### Importancia de los Predictores por Modelo

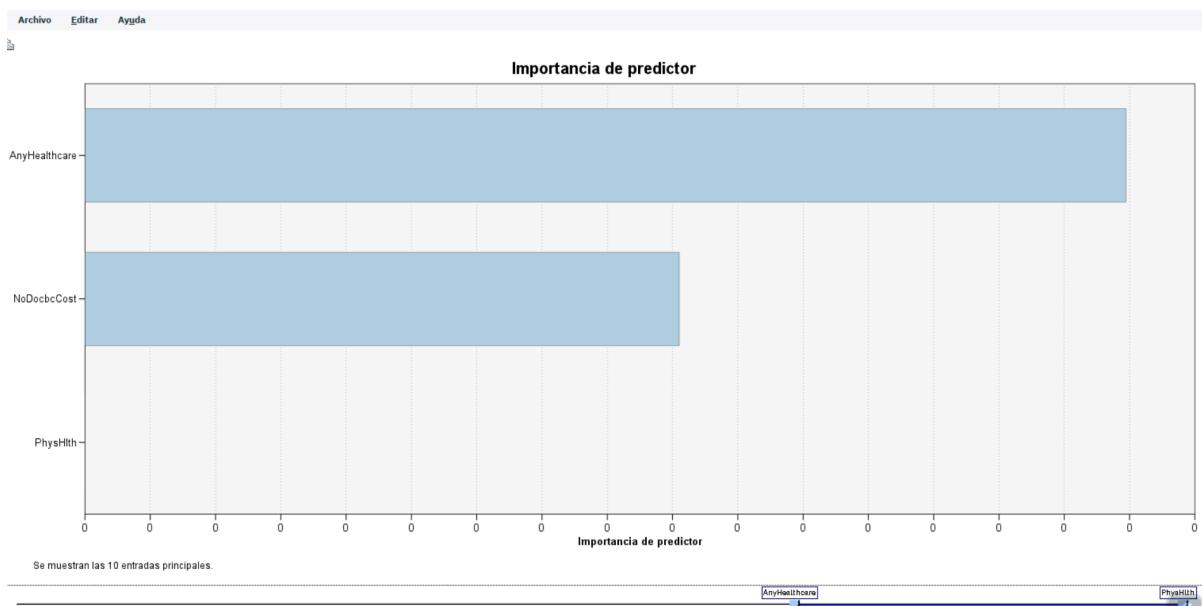
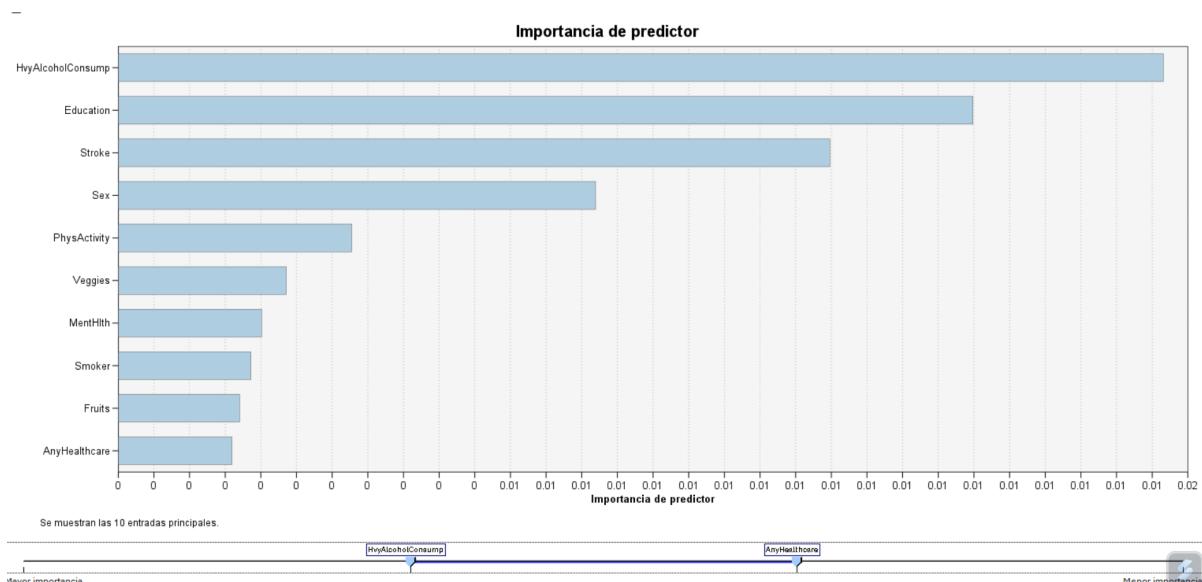
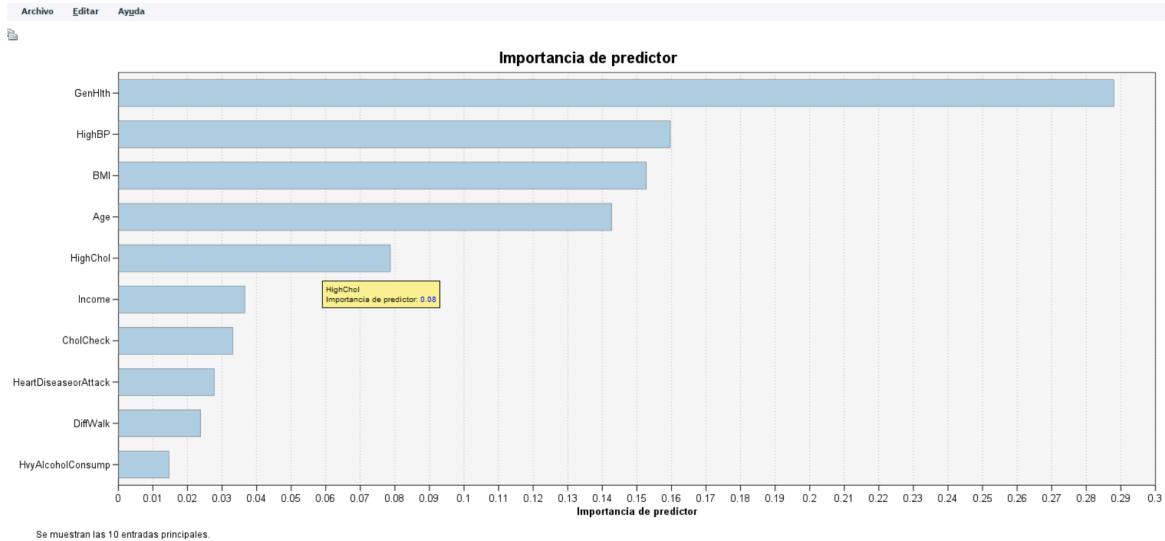
- Algoritmo de Decisión CRT



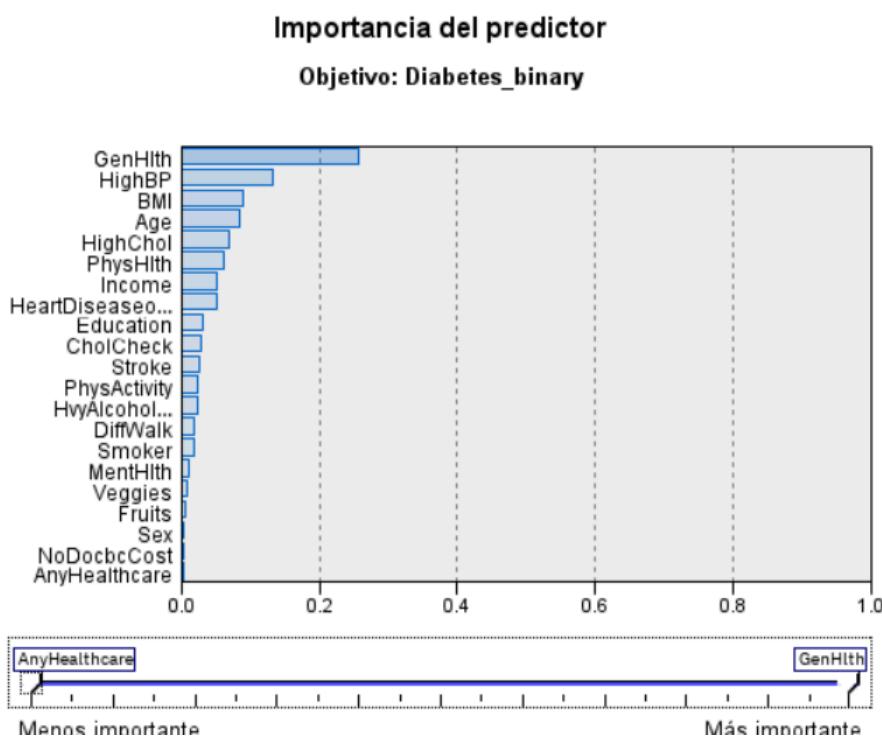
### - Red Neuronal MLP



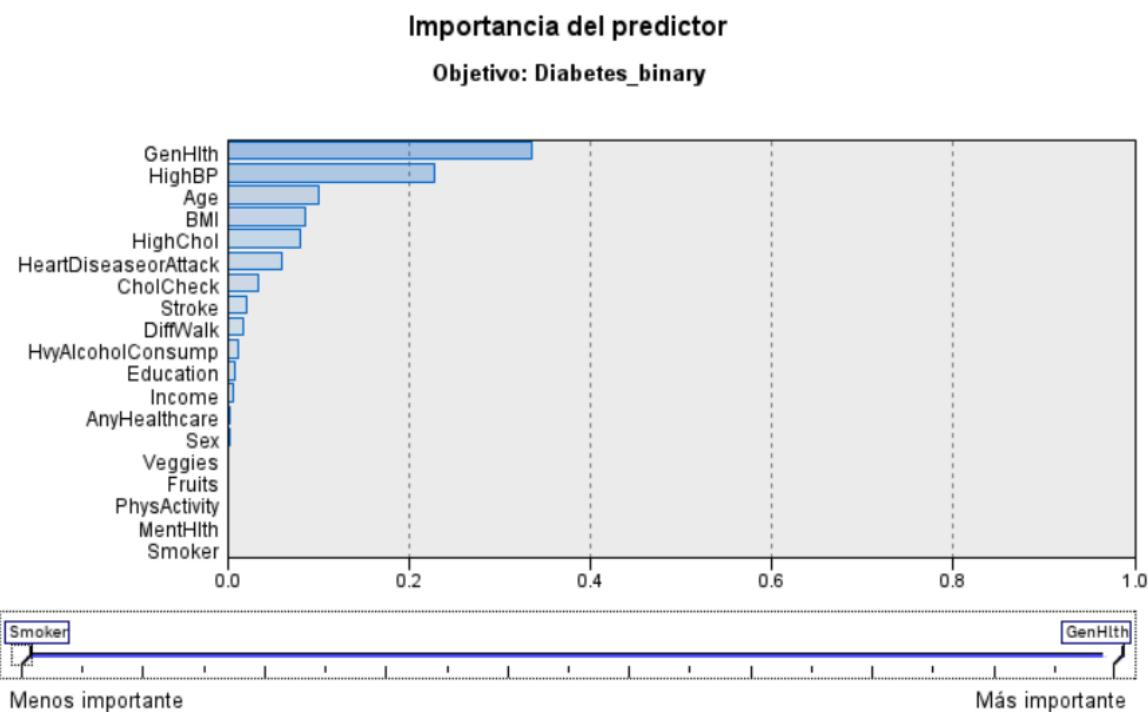
### - Red Neuronal LSVM



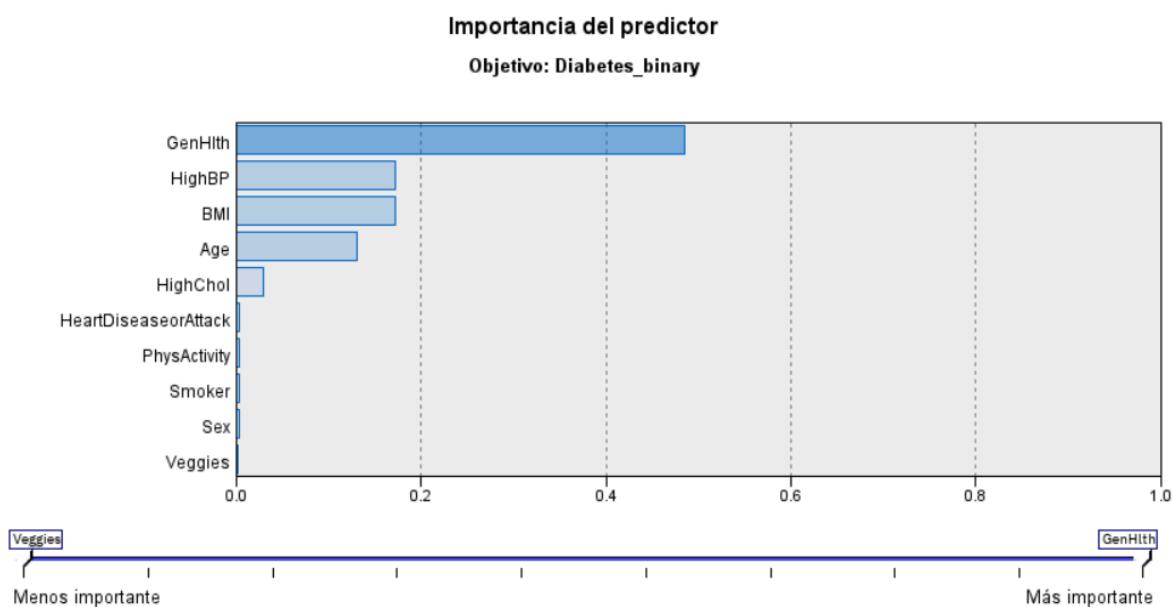
### - Red Bayesiana



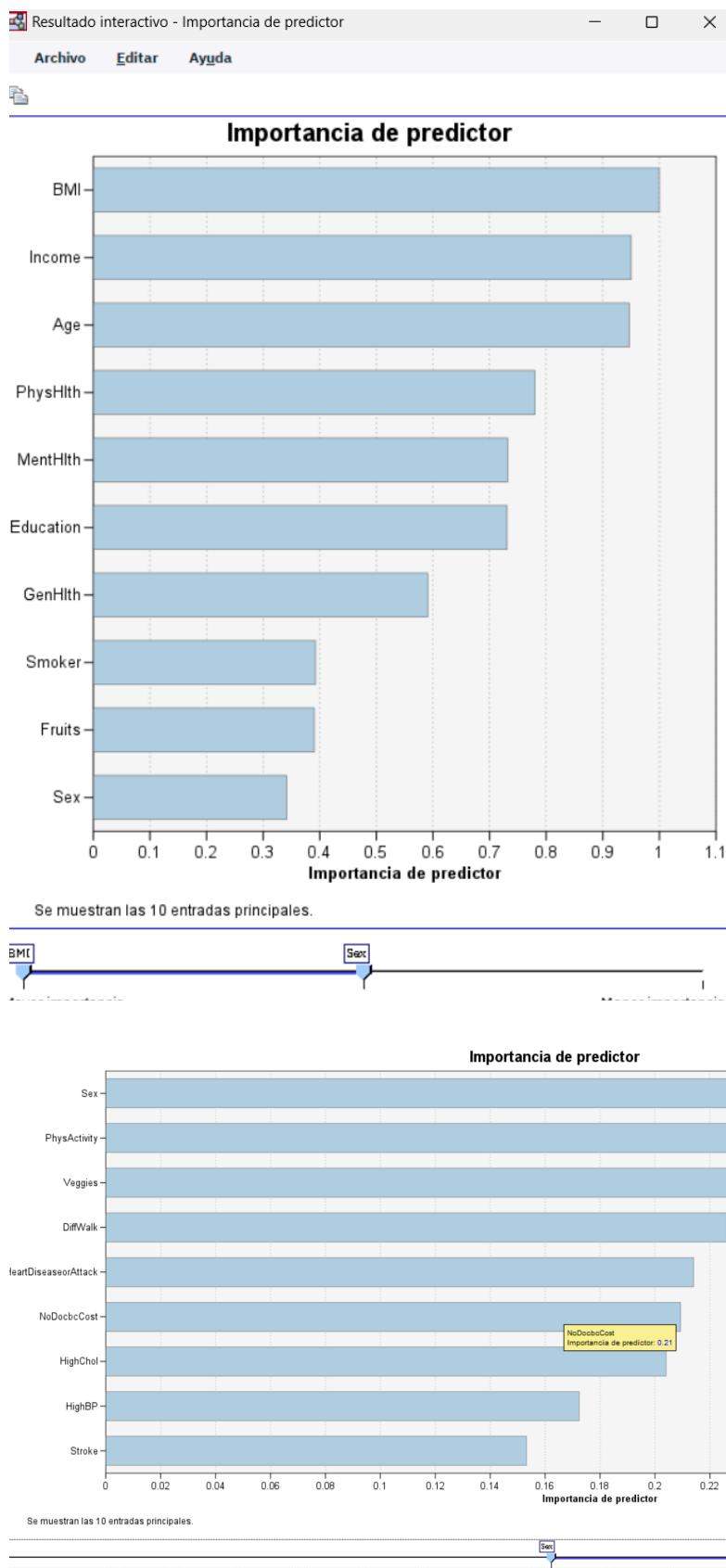
#### - Algoritmo de Decisión C5.0



#### - Algoritmo de Decisión CHAID



## - Árboles Aleatorios



## Modelado con selección de Atributos.

Para este modelado, se excluyeron los atributos *stroke*, *hvylcohol*, *NoDocbcCost* y *sex*

### - 1. Red neuronal MLP

#### Clasificación para Diabetes\_binary

Porcentaje correcto global = 74.8%

Observado	Previsto	
	0.000	1.000
0.000	9435	4020
1.000	2744	10664

Porcentaje de filas

- 100.00
- 80.00
- 60.00
- 40.00
- 20.00
- 0.00

### - 2. Red neuronal LSVM

Análisis de [Diabetes\_binary] #7

Archivo Editar

Análisis Anotaciones

– Contrair todo + Desplegar todo

Resultados para el campo de resultado Diabetes\_binary

Comparando \$L-Diabetes\_binary con Diabetes\_binary

Correctos	20,206	75.22 %
Erróneos	6,657	24.78 %
Total	26,863	

Matriz de coincidencias para \$L-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,706	3,749
1.000000	2,908	10,500

Evaluación del rendimiento

0.000000	0.429
1.000000	0.39

### - 3. Red neuronal SVM

The screenshot shows the Orange data mining software interface. The title bar reads "Análisis de [Diabetes\_binary] #8". The menu bar includes "Archivo" and "Editar". The main window has tabs "Análisis" and "Anotaciones", with "Análisis" selected. Below the tabs are buttons "+ Contraer todo" and "+ Desplegar todo". The main content area displays the results for the "Diabetes\_binary" field:

- Resultados para el campo de resultado Diabetes\_binary
  - Comparando \$S-Diabetes\_binary con Diabetes\_binary

Correctos	24,489	91.16 %
Erróneos	2,374	8.84 %
Total	26,863	
  - Matriz de coincidencias para \$S-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	12,227	1,228
1.000000	1,146	12,262
  - Evaluación del rendimiento

0.000000	0.602
1.000000	0.599

### - 4. Red Bayesiana

The screenshot shows the Orange data mining software interface. The title bar reads "Análisis de [Diabetes\_binary] #11". The menu bar includes "Archivo" and "Editar". The main window has tabs "Análisis" and "Anotaciones", with "Análisis" selected. Below the tabs are buttons "+ Contraer todo" and "+ Desplegar todo". The main content area displays the results for the "Diabetes\_binary" field:

- Resultados para el campo de resultado Diabetes\_binary
  - Comparando \$B-Diabetes\_binary con Diabetes\_binary

Correctos	19,975	74.36 %
Erróneos	6,888	25.64 %
Total	26,863	
  - Matriz de coincidencias para \$B-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,584	3,871
1.000000	3,017	10,391
  - Evaluación del rendimiento

0.000000	0.418
1.000000	0.378

## - 5. Algoritmo KNN

Análisis de [Diabetes\_binary] #6

Archivo Editar

Análisis Anotaciones

– Contraer todo + Desplegar todo

Resultados para el campo de resultado Diabetes\_binary

Comparando \$KNN-Diabetes\_binary con Diabetes\_binary

Correctos	21,470	79.92 %
Erróneos	5,393	20.08 %
Total	26,863	

Matriz de coincidencias para \$KNN-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	10,481	2,974
1.000000	2,419	10,989

Evaluación del rendimiento

0.000000	0.484
1.000000	0.455

## - 6. Algoritmo De Decisión C5.0

Análisis de [Diabetes\_binary] #10

Archivo Editar

Análisis Anotaciones

– Contraer todo + Desplegar todo

Resultados para el campo de resultado Diabetes\_binary

Comparando \$C-Diabetes\_binary con Diabetes\_binary

Correctos	20,465	76.18 %
Erróneos	6,398	23.82 %
Total	26,863	

Matriz de coincidencias para \$C-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,858	3,597
1.000000	2,801	10,607

Evaluación del rendimiento

0.000000	0.441
1.000000	0.403

## - 7. Algoritmo de Decisión CRT

The screenshot shows the RapidMiner interface with the following details:

- Toolbar:** Archivo, Editar, etc.
- Top Bar:** Análisis, Anotaciones, etc.
- Buttons:** Contraer todo, Desplegar todo
- Result Section:**
  - Resultados para el campo de resultado Diabetes\_binary
  - Comparando \$R-Diabetes\_binary con Diabetes\_binary
 

Correctos	19,674	73.24 %
Erróneos	7,189	26.76 %
Total	26,863	
  - Matriz de coincidencias para \$R-Diabetes\_binary (las filas muestran las reales)
 

	0.000000	1.000000
0.000000	9,368	4,087
1.000000	3,102	10,306
  - Evaluación del rendimiento
 

0.000000	0.405
1.000000	0.361

## - 8. Algoritmo de Decisión CHAID

The screenshot shows the RapidMiner interface with the following details:

- Toolbar:** Archivo, Editar, etc.
- Top Bar:** Análisis, Anotaciones, etc.
- Buttons:** Contraer todo, Desplegar todo
- Result Section:**
  - Resultados para el campo de resultado Diabetes\_binary
  - Comparando \$R-Diabetes\_binary con Diabetes\_binary
 

Correctos	19,751	73.52 %
Erróneos	7,112	26.48 %
Total	26,863	
  - Matriz de coincidencias para \$R-Diabetes\_binary (las filas muestran las reales)
 

	0.000000	1.000000
0.000000	9,481	3,974
1.000000	3,138	10,270
  - Evaluación del rendimiento
 

0.000000	0.405
1.000000	0.368

## - 9. Regresión Logística

The screenshot shows the SPSS Modeler interface with the title bar "Análisis de [Diabetes\_binary] #4". The menu bar includes "Archivo", "Editar", "Analizar", "Visualizar", "Anotaciones", and "Ayuda". Below the menu is a toolbar with icons for "Analizar", "Visualizar", and "Anotaciones". The main area is divided into sections: "Analizar" (selected) and "Anotaciones". At the top of the main area are two buttons: "- Contraer todo" and "+ Desplegar todo". The results section contains the following items:

- Resultados para el campo de resultado Diabetes\_binary
  - Comparando \$L-Diabetes\_binary con Diabetes\_binary

Correctos	20,205	75.21 %
Erróneos	6,658	24.79 %
Total	26,863	
  - Matriz de coincidencias para \$L-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,776	3,679
1.000000	2,979	10,429
  - Evaluación del rendimiento

0.000000	0.425
1.000000	0.393

## - 10. Árboles Aleatorios

The screenshot shows the SPSS Modeler interface with the title bar "Análisis de [Diabetes\_binary] #4". The menu bar includes "Archivo", "Editar", "Analizar", "Visualizar", "Anotaciones", and "Ayuda". Below the menu is a toolbar with icons for "Analizar", "Visualizar", and "Anotaciones". The main area is divided into sections: "Analizar" (selected) and "Anotaciones". At the top of the main area are two buttons: "- Contraer todo" and "+ Desplegar todo". The results section contains the following items:

- Resultados para el campo de resultado Diabetes\_binary
  - Comparando \$R-Diabetes\_binary con Diabetes\_binary

Correctos	20,934	77.93 %
Erróneos	5,929	22.07 %
Total	26,863	
  - Matriz de coincidencias para \$R-Diabetes\_binary (las filas muestran las reales)

	0.000000	1.000000
0.000000	9,972	3,483
1.000000	2,446	10,962
  - Evaluación del rendimiento

0.000000	0.472
1.000000	0.419

**Métricas con Selección de características.(sin stroke, hac, cholchek, sex, noDC )**

---

Modelo	Exactitud ()	Precisión	Sensibilidad	Especificidad
Red Neuronal MLP	0.74820384 9	0.7262326 34	0.79534606 2	0.70122631
Red Neuronal LSVM	0.75216857 2	0.7368938 17	0.78311455 8	0.721326098
Red Neuronal SVM	0.91162565 6	0.9089696 07	0.91452864	0.908732813
Red Neuronal Bayesiana	0.74358783 5	0.7285794 42	0.77498508 4	0.71230026
Algoritmo KNN	0.79924059 1	0.7870085 23	0.81958532 2	0.778966927
Algoritmo de Decisión C5.0	0.76305699 3	0.7450420 99	0.79855310 3	0.727684876
Algoritmo de Decisión CRT	0.73238283 1	0.7160425 21	0.76864558 5	0.696246748
Algoritmo de Decisión CHAID	0.73524922 8	0.7210053 36	0.76596062 1	0.704645113
Regresión Logística	0.75214979 7	0.7392259 71	0.77781921 2	0.726570048

Árboles Aleatorios	0.77928749 6	0.7588785 05	0.81757159 9	0.741137124
--------------------	-----------------	-----------------	-----------------	-------------

---

### Mejores Modelos:

- Red Neuronal SVM - 91%
- Algoritmo KNN - 79%
- Árboles Aleatorios - 77%

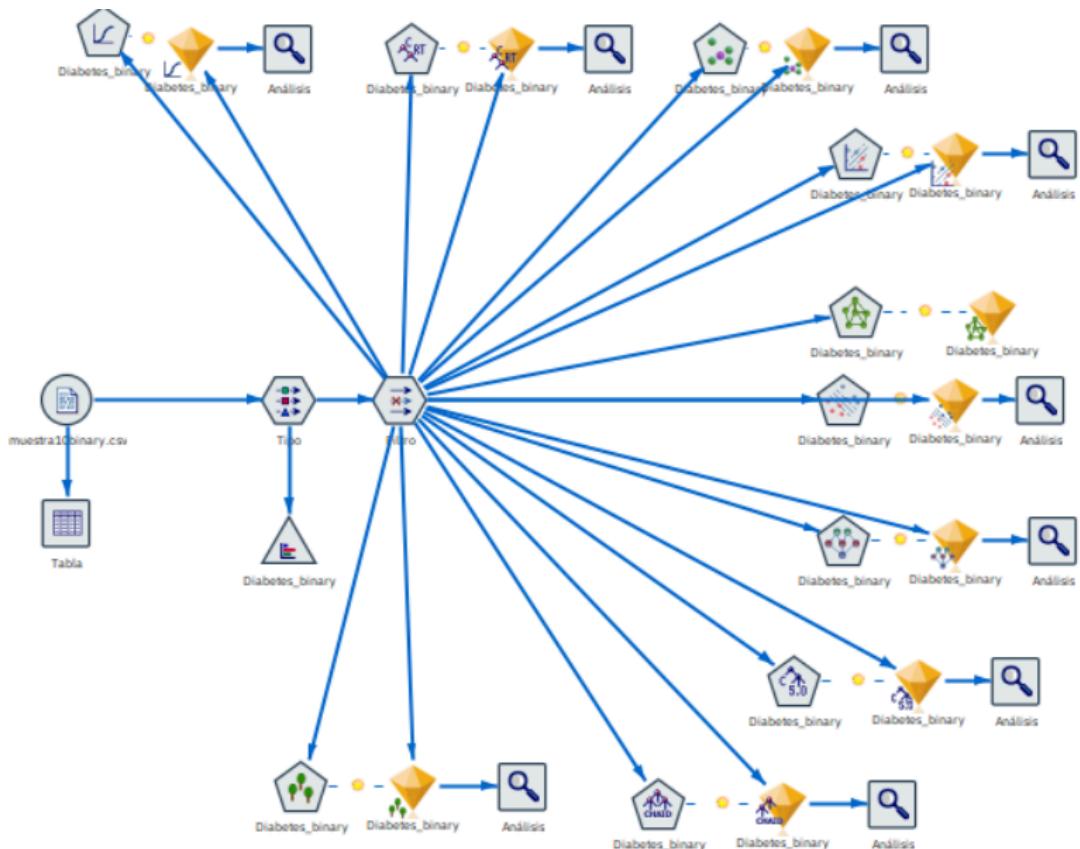
**Tabla de Valores de Matrices de confusión**

Modelo	TN	FP	FN	TP
RED MLP	9435	4020	2744	10664
LSVM	9704	3749	2908	10500
SVM	12227	1228	1146	12262
Bayesiana	9584	3871	3017	10391
KNN	10481	2974	2419	10989
C5.0	9791	3664	2701	10707

CRT	9368	4087	3102	10306
CHAID	9481	3974	3138	10270
Regresión Logística	9776	3679	2979	10429
Árboles Aleatorios	9972	3483	2446	10962

---

### Ruta IBM SPSS MODELER

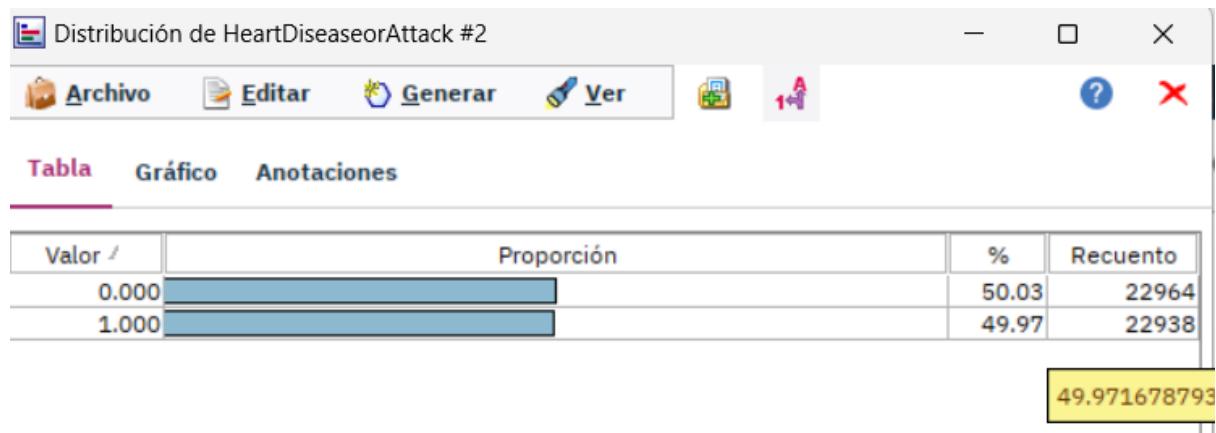


# Objetivo de Minería de Datos - Clasificación de HeartDiseaseorAttack

## Ocurrencias del atributo Destino en DataSet desbalanceado



## Ocurrencias del atributo Destino en DataSet Balanceado



## Modelado desbalanceado

Obtención de matriz de confusión en base a la evaluación de cada modelo.

SVM		
	Negativo	Positivo
Negativo	22846	118
Positivo	545	1859

Red Bayesiana		
	Negativo	Positivo
Negativo	21850	1114
Positivo	1670	734

Logistica		
	Negativo	Positivo
Negativo	22713	251
Positivo	2127	277

MLP		
	Negativo	Positivo
Negativo	22887	74
Positivo	2303	100

LSVM		
	Negativo	Positivo
Negativo	22895	69
Positivo	2266	138

### CHAID

	Negativo	Positivo
Negativo	22964	0
Positivo	2404	0

### C5.0

	Negativo	Positivo
Negativo	22792	172
Positivo	2213	191

### KNN

	Negativo	Positivo
Negativo	22874	90
Positivo	2058	346

### CRT

	Negativo	Positivo
Negativo	22964	0
Positivo	2404	0

QUEST		
	Negativo	Positivo
Negativo	22964	0
Positivo	2404	0

## Métricas (Desbalanceado)

Se ocuparon las métricas de exactitud, precisión, sensibilidad y especificidad de cada uno de los modelos para poder evaluar su eficiencia.

Modelo	Exactitud(bon dad)	Precisión	Sensibili dad	Especifi cidad
Red Neuronal (MLP)	0.771535009	0.7528327 17	0.80813 497	0.73497 648
Red Neuronal(LSVM)	0.907954904	0.6666666 67	0.05740 433	0.99699 53
Red Neuronal (SVM)	0.981354462	0.9565011 82	0.84151 414	0.99599 373
Red Neuronal Bayesiana	0.761796872	0.7454805 73	0.79462 028	0.72901 063

Algoritmo KNN	0.708942159	0.7971641 23	0.83577 47	0.27882 91
Algoritmo de decisión C5.0	0.784061697	0.7622825 39	0.82522 452	0.74294 548
Arbol de decision CRT	0.862725879	0.7408537 88	0.76187 985	0.73380 073
Árbol de decisión CHAID	0.746220208	0.7207383 36	0.80347 022	0.68903 501
Árboles aleatorios	0.743998083	0.7098638 05	0.82483 216	0.66325 553
Regresión Logística	0.773059997	0.7608215 64	0.79614 613	0.75

## Modelado sin selección de Atributos (balanceado)

Obtención de matriz de confusión en base a la evaluación de cada modelo.

SVM		
	Negativo	Positivo
Negativo	17112	5852
Positivo	4571	18367

### Red Bayesiana

	Negativo	Positivo
Negativo	16741	6223
Positivo	4711	18227

### Logistica

	Negativo	Positivo
Negativo	17223	5741
Positivo	4676	18262

### MLP

	Negativo	Positivo
Negativo	16878	6086
Positivo	4401	18537

### LSVM

	Negativo	Positivo
Negativo	22895	69
Positivo	2266	138

### CHAID

	Negativo	Positivo
Negativo	15823	7141
Positivo	4508	18430

### C5.0

	Negativo	Positivo
Negativo	17061	5903
Positivo	4009	18929

### KNN

	Negativo	Positivo
Negativo	1886	4878
Positivo	3767	19171

### CRT

	Negativo	Positivo
Negativo	16851	6113
Positivo	5462	17476

## QUEST

	Negativo	Positivo
Negativo	15231	7733
Positivo	4018	18920

### Métricas sin selección de Atributos (balanceado)

Se ocuparon las métricas de exactitud, precisión, sensibilidad y especificidad de cada uno de los modelos para poder evaluar su eficiencia.

Modelo	Exactitud(bo ndad)	Precisión	Sensibi lidad	Especif icidad
Red Neuronal (MLP)	0.76796218	0.749725 621	0.80408 057	0.73188 469
Red Neuronal(LSVM)	0.77096648	0.758071 175	0.79703 099	0.74481 798
Red Neuronal (SVM)	0.831311583	0.825771 17	0.84203 189	0.82047 896
Red Neuronal Bayesiana	0.761470088	0.744903 379	0.79488 186	0.72809 615

Algoritmo KNN	0.811533267	0.798030 79	0.83390 008	0.78919 178
Algoritmo de decisión C5.0	0.783974554	0.762688 615	0.82413 462	0.74385 995
Arbol de decision CRT	0.862725879	0.740853 788	0.76187 985	0.73380 073
Árbol de decisión CHAID	0.746220208	0.720738 336	0.80347 022	0.68903 501
Árboles aleatorios	0.743998083	0.709863 805	0.82483 216	0.66325 553
Regresión Logística	0.772297503	0.760320 24	0.79492 545	0.74969 518

## Modelado con Selección de Atributos

- HeartAttackorDisease(balanceado sin CholChe)

SVM		
	Negativo	Positivo
Negativo	17112	5852
Positivo	4571	18367

### Red Bayesiana

	Negativo	Positivo
Negativo	16771	6193
Positivo	4752	18186

### Logistica

	Negativo	Positivo
Negativo	17232	5732
Positivo	4691	18247

### MLP

	Negativo	Positivo
Negativo	16858	6106
Positivo	4632	18306

### LSVM

	Negativo	Positivo
Negativo	17129	5835
Positivo	4587	18351

### CHAID

	Negativo	Positivo
Negativo	15823	7141
Positivo	4508	18430

### C5.0

	Negativo	Positivo
Negativo	17020	5944
Positivo	4020	18918

### KNN

	Negativo	Positivo
Negativo	18099	4865
Positivo	3816	19122

### CRT

	Negativo	Positivo
Negativo	16851	6113
Positivo	5462	17476

## QUEST

	Negativo	Positivo
Negativo	15231	7733
Positivo	4018	18920

- HeartAttackorDisease(balanceado sin HAC)

## SVM

	Negativo	Positivo
Negativo	17112	5852
Positivo	4571	18367

## Red Bayesiana

	Negativo	Positivo
Negativo	16720	6244
Positivo	4705	18233

## Logistica

	Negativo	Positivo
Negativo	17216	5748
Positivo	4704	18234

### MLP

	Negativo	Positivo
Negativo	16807	6157
Positivo	4494	18444

### LSVM

	Negativo	Positivo
Negativo	17104	5860
Positivo	4576	18362

### CHAID

	Negativo	Positivo
Negativo	15823	7141
Positivo	4508	18430

### C5.0

	Negativo	Positivo
Negativo	17082	5882
Positivo	4034	18904

KNN		
	Negativo	Positivo
Negativo	18123	4841
Positivo	3810	19128

CRT		
	Negativo	Positivo
Negativo	16851	6113
Positivo	5462	17476

QUEST		
	Negativo	Positivo
Negativo	15231	7733
Positivo	4018	18920

- HeartAttackorDisease(balanceado sin Chol,AI )

SVM		
	Negativo	Positivo
Negativo	17112	5852
Positivo	4571	18367

### Red Bayesiana

	Negativo	Positivo
Negativo	16746	6218
Positivo	4723	18215

### Logistica

	Negativo	Positivo
Negativo	17225	5739
Positivo	4715	18223

### MLP

	Negativo	Positivo
Negativo	16973	5991
Positivo	4551	18387

### LSVM

	Negativo	Positivo
Negativo	17117	5847
Positivo	4607	18331

### CHAID

	Negativo	Positivo
Negativo	15823	7141
Positivo	4508	18430

### C5.0

	Negativo	Positivo
Negativo	17045	5919
Positivo	4078	18860

### KNN

	Negativo	Positivo
Negativo	18136	4828
Positivo	3863	19075

### CRT

	Negativo	Positivo
Negativo	16851	6113
Positivo	5462	17476

QUEST		
	Negativo	Positivo
Negativo	15231	7733
Positivo	4018	18920

## Métricas con selección de Atributos

- HeartAttackorDisease(balanceado sin Chol Check)

Modelo	Exactitud(bondad)	Precisión	Sensibilidad	Especificidad
Red Neuronal (MLP)	0.766066838	0.74987711	0.79806435	0.73410556
Red Neuronal(LSVM)	0.77295107	0.758744728	0.80002616	0.74590664
Red Neuronal (SVM)	0.83412177	0.832730716	0.83988886	0.82825069
Red Neuronal Bayesiana	0.761557231	0.745969892	0.79283285	0.73031702
Algoritmo KNN	0.8108797	0.797181807	0.8336385	0.78814666

Algoritmo de decisión C5.0	0.782928848	0.76092028	0.82474496	0.74116008
Arbol de decision CRT	0.862725879	0.740853788	0.76187985	0.73380073
Árbol de decisión CHAID	0.746220208	0.720738336	0.80347022	0.68903501
Árboles aleatorios	0.743998083	0.709863805	0.82483216	0.66325553
Regresión Logística	0.772929284	0.760957504	0.7954922	0.75039192

- HeartAttackorDisease(balanceado sin Hvy Alcohol Consum)

Modelo	Exactitud(bondad )	Precisión	Sensibilidad	Especificidad
Red Neuronal (MLP)	0.76796218	0.749725621	0.80408057	0.73188469
Red Neuronal LSVM	0.77096648	0.758071175	0.79703099	0.74481798

Red Neuronal (SVM)	0.831311583	0.82577117	0.84203189	0.82047896
Red Neuronal Bayesiana	0.761470088	0.744903379	0.79488186	0.72809615
Algoritmo KNN	0.811533267	0.79803079	0.83390008	0.78919178
Algoritmo de decision C5.0	0.783974554	0.762688615	0.82413462	0.74385995
Arbol de decision CRT	0.862725879	0.740853788	0.76187985	0.73380073
Arbol de decision CHAID	0.746220208	0.720738336	0.80347022	0.68903501
Árboles aleatorios	0.743998083	0.709863805	0.82483216	0.66325553
Regresion Logistica	0.772297503	0.76032024	0.79492545	0.74969518

- HeartAttackorDisease(balanceado sin Hvy Alcohol Consump y Chol Check)

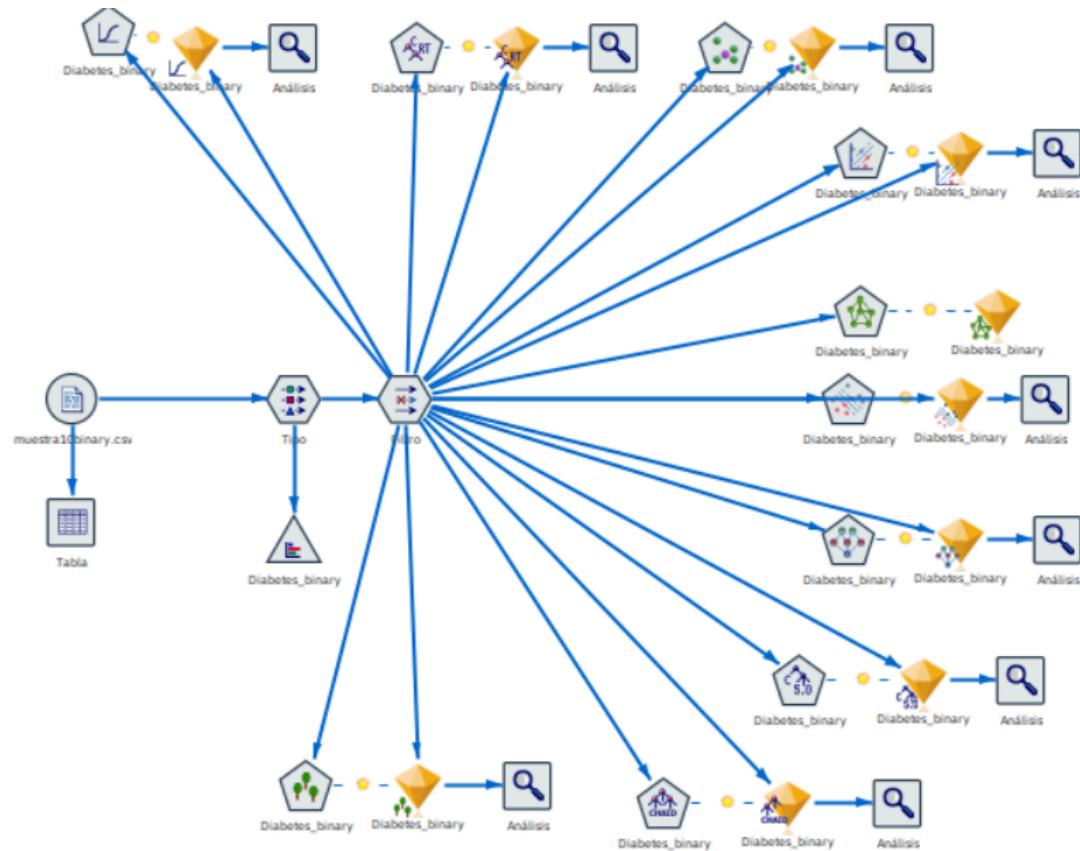
Modelo	Exactitud(bon dad)	Precisión	Sensibilidad	Especificidad
Red Neuronal (MLP)	0.771535009	0.752832717	0.80813497	0.73497648

Red Neuronal(LSVM)	<b>0.907954904</b>	0.6666666 67	0.05740 433	0.99699 53
Red Neuronal (SVM)	<b>0.836757137</b>	0.8332901 89	0.84443 861	0.82898 115
Red Neuronal Bayesiana	<b>0.761796872</b>	0.7454805 73	0.79462 028	0.72901 063
Algoritmo KNN	<b>0.708942159</b>	0.7971641 23	0.83577 47	0.27882 91
Algoritmo de decisión C5.0	<b>0.784061697</b>	0.7622825 39	0.82522 452	0.74294 548
Arbol de decision CRT	0.862725879	0.7408537 88	0.76187 985	0.73380 073
Árbol de decisión CHAID	0.746220208	0.7207383 36	0.80347 022	0.68903 501
Árboles aleatorios	0.743998083	0.7098638 05	0.82483 216	0.66325 553
Regresión Logística	<b>0.773059997</b>	0.7608215 64	0.79614 613	0.75

## Mejores Modelos obtenidos

- CRT - 86%
- SVM - 83%
- Regresión Logística - 77.3%

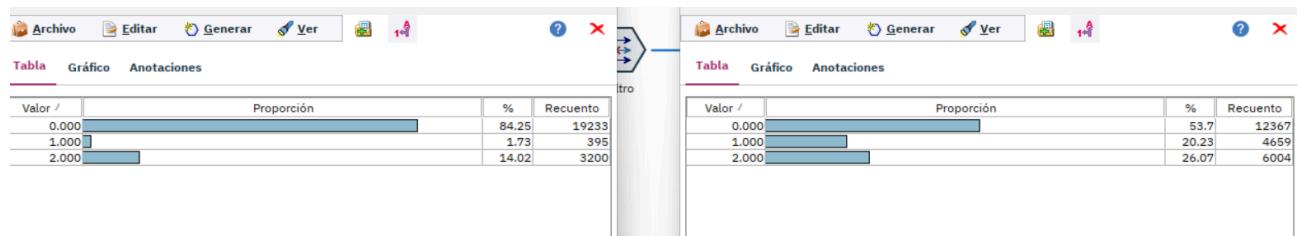
## Ruta IBM SPSS MODELER



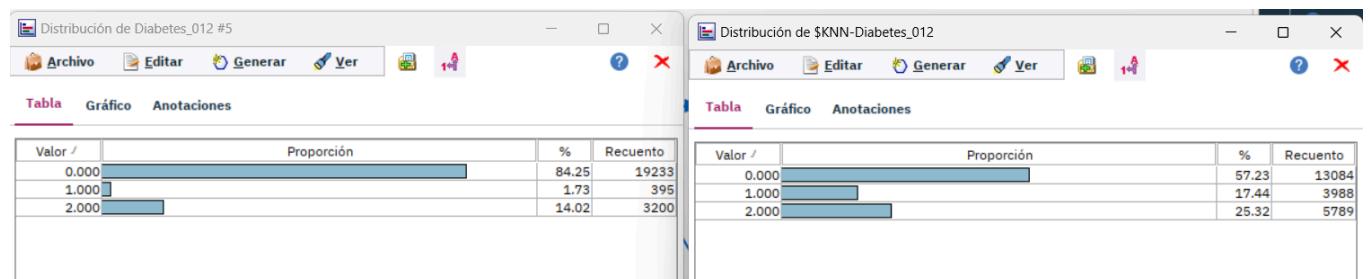
## Prueba de Modelos (Evaluación)

### Prueba de Clasificación Diabetes\_012

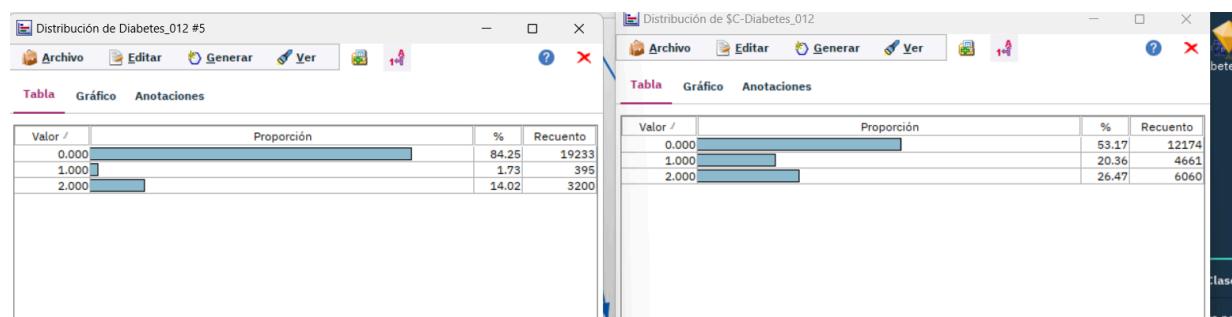
- SVM - 89%



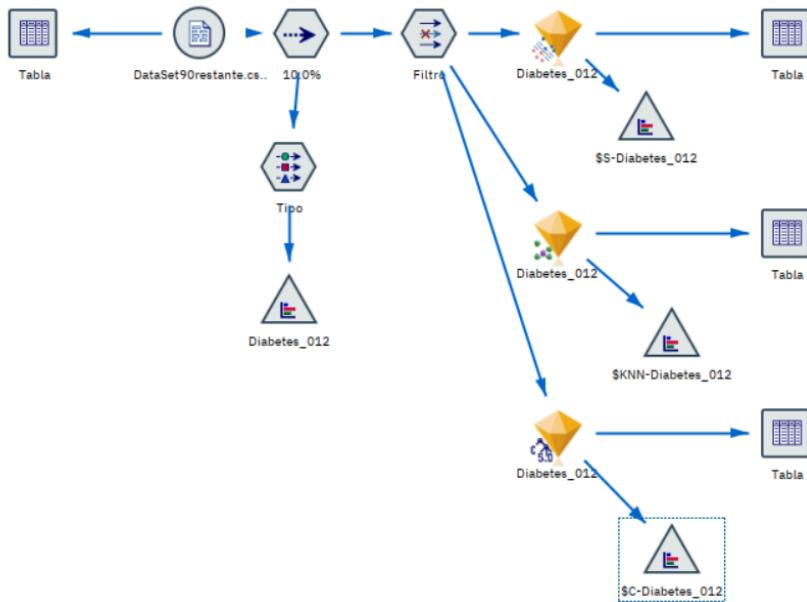
- KNN - 69%



- C5.0 - 66%

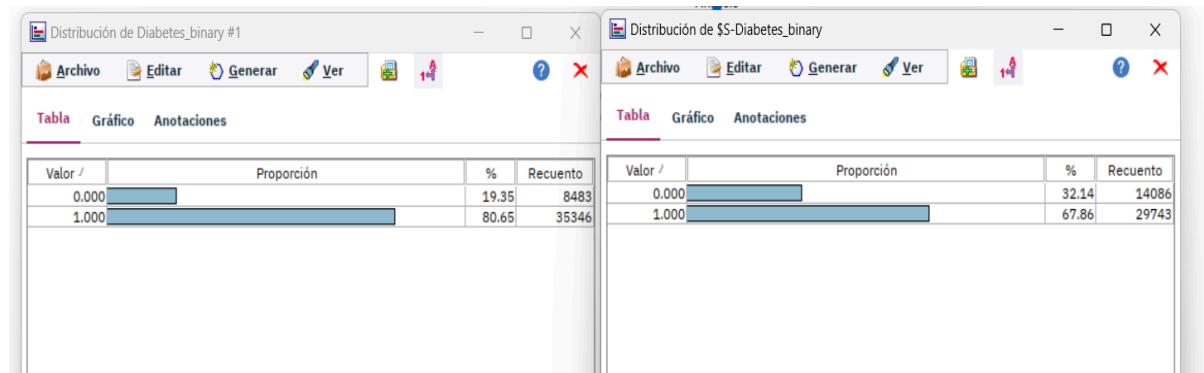


- Ruta IBM



## Prueba de Clasificación Diabetes\_binary

- Red Neuronal SVM - 91%



**- Contraer todo**

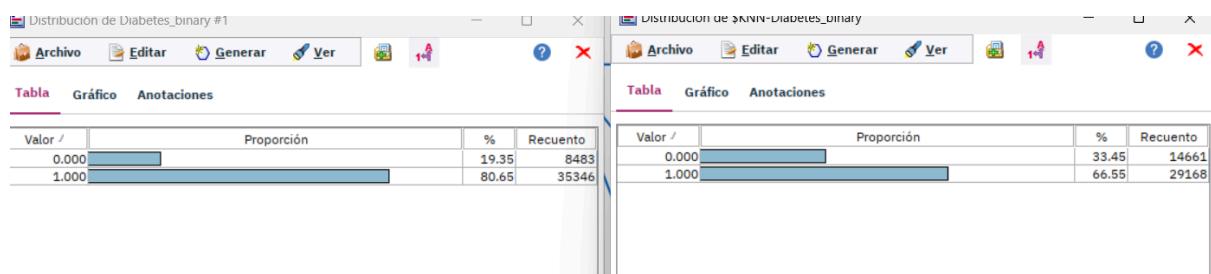
**+ Desplegar todo**

■ Resultados para el campo de resultado Diabetes\_binary

■ Comparando \$S-Diabetes\_binary con Diabetes\_binary

Error mínimo	-1.0
Error máximo	1.0
Error promedio	-0.221
Error absoluto promedio	0.281
Desviación estándar	0.482
Correlación lineal	0.354
Ocurrencias	22,711

- Algoritmo KNN - 79%



**Análisis Anotaciones**

**- Contraer todo**

**+ Desplegar todo**

■ Resultados para el campo de resultado Diabetes\_binary

■ Comparando \$KNN-Diabetes\_binary con Diabetes\_binary

Error mínimo	-1.0
Error máximo	1.0
Error promedio	-0.237
Error absoluto promedio	0.306
Desviación estándar	0.499
Correlación lineal	0.309
Ocurrencias	22,940

- Árboles Aleatorios 77%

The image shows two side-by-side SPSS Statistics windows. Both windows have a title bar with the text 'Distribución de [Dataset]'. The left window is titled 'Distribución de Diabetes\_binary #1' and the right one is 'Distribución de \$R-Diabetes\_binary'. Both windows have tabs for 'Tabla', 'Gráfico', and 'Anotaciones'. Below the tabs is a table with four columns: 'Valor /', 'Proporción', '%', and 'Recuento'. In the first window, the rows are '0.000' (Proporción 19.35, Recuento 8483) and '1.000' (Proporción 80.65, Recuento 35346). In the second window, the rows are '0.000' (Proporción 29.44, Recuento 12902) and '0.000' (Proporción 70.56, Recuento 30927).

**Analysis Anotaciones**

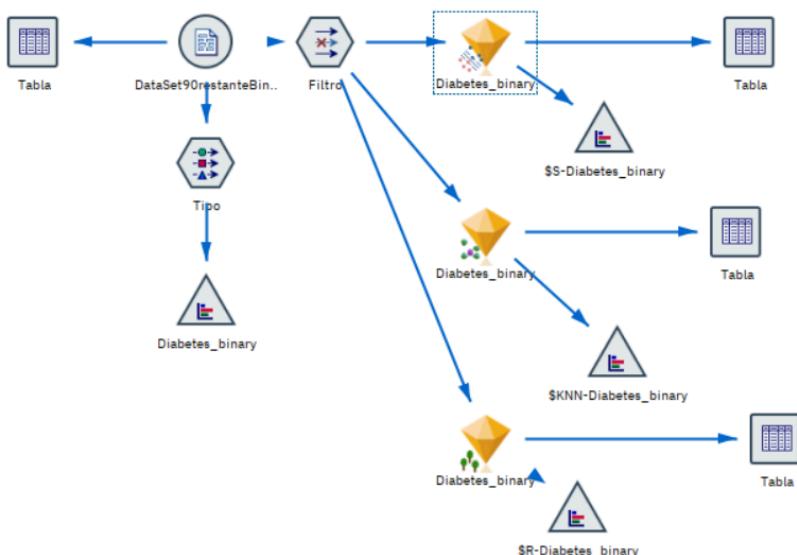
**Contraer todo** **+ Desplegar todo**

Resultados para el campo de resultado Diabetes\_binary

Comparando \$R-Diabetes\_binary con Diabetes\_binary

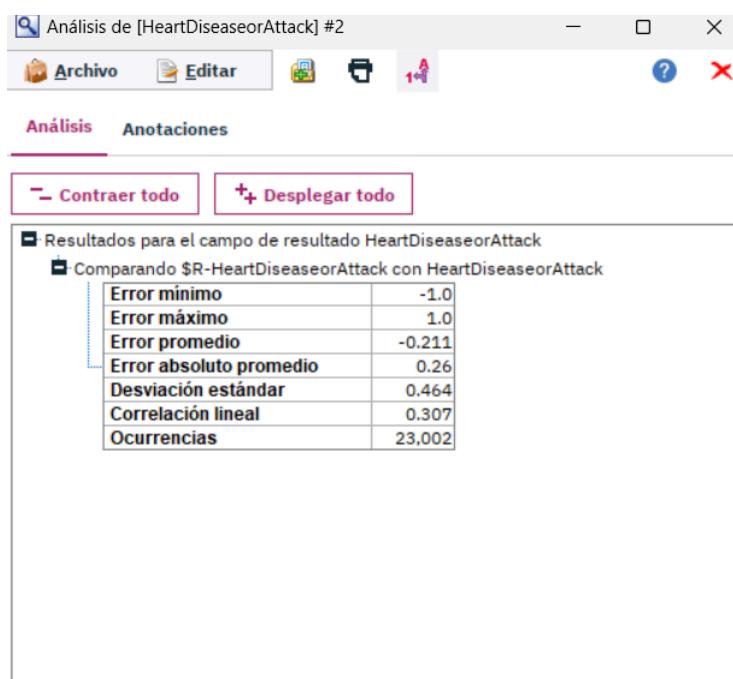
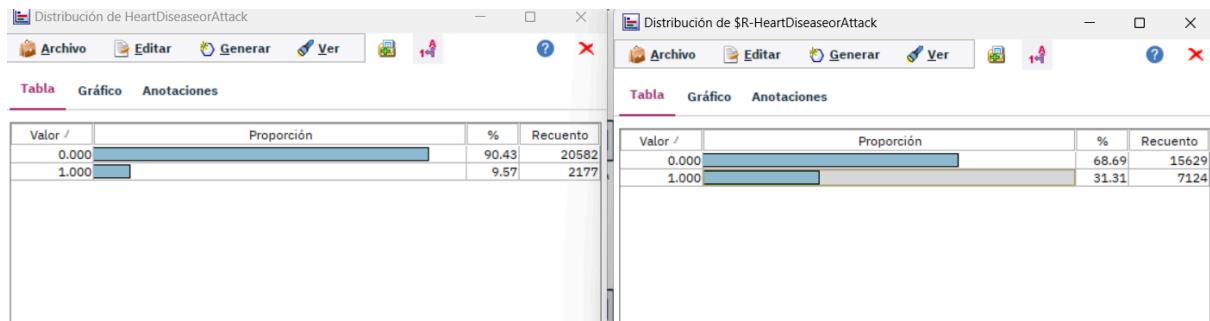
Error mínimo	-0.0
Error máximo	1.0
Error promedio	0.137
Error absoluto promedio	0.137
Desviación estándar	0.344
Correlación lineal	
Ocurrencias	22,901

- Ruta IBM

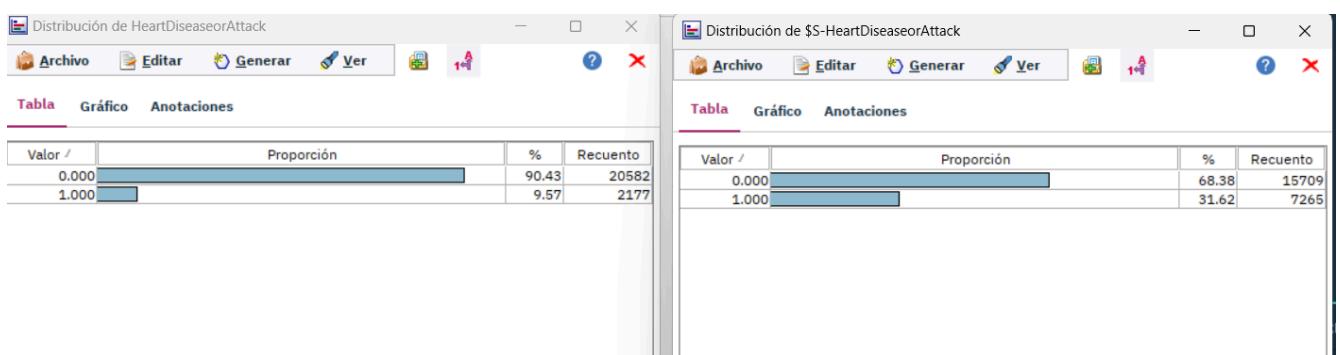


## Prueba de Clasificación HeartAttackorDisease

- CRT - 86%



- SVM - 83%



Archivo Editar

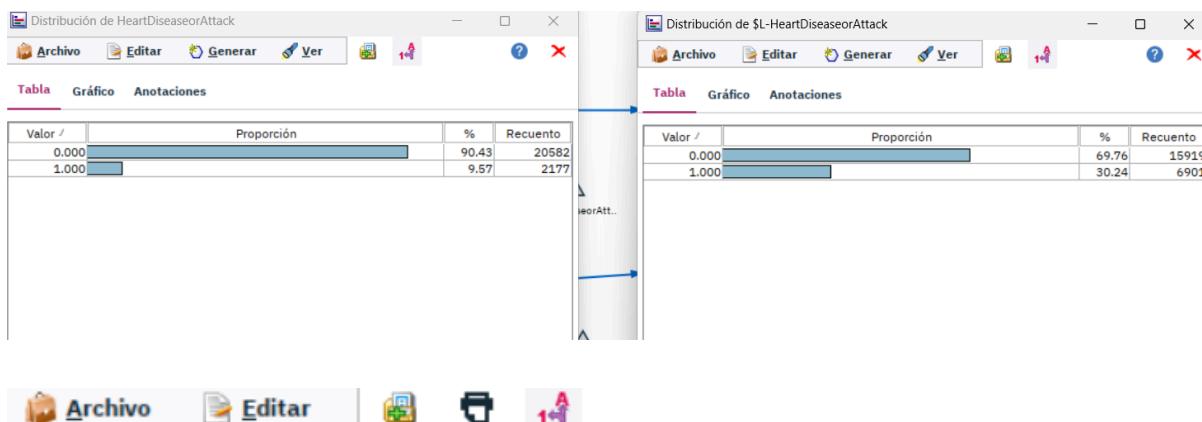
Análisis Anotaciones

**- Contraer todo** **+ Desplegar todo**

- Resultados para el campo de resultado HeartDiseaseorAttack
  - Comparando \$S-HeartDiseaseorAttack con HeartDiseaseorAttack
 

Error mínimo	-1.0
Error máximo	1.0
Error promedio	-0.213
Error absoluto promedio	0.247
Desviación estándar	0.449
Correlación lineal	0.373
Ocurrencias	22,678

- Regresión Logística - 77.3%



Archivo Editar

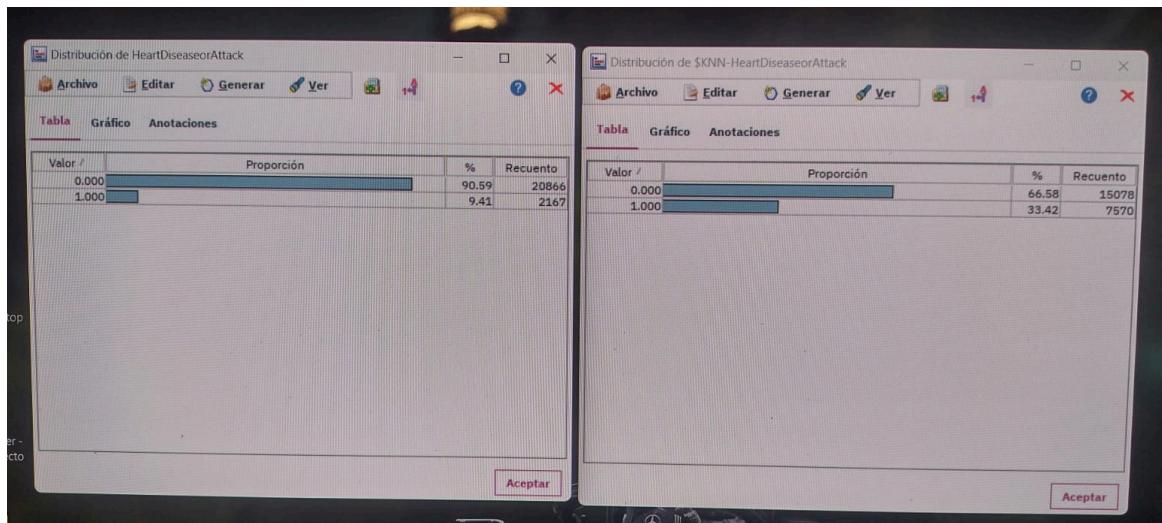
Análisis Anotaciones

**- Contraer todo** **+ Desplegar todo**

- Resultados para el campo de resultado HeartDiseaseorAttack
  - Comparando \$L-HeartDiseaseorAttack con HeartDiseaseorAttack
 

Error mínimo	-1.0
Error máximo	1.0
Error promedio	-0.209
Error absoluto promedio	0.248
Desviación estándar	0.452
Correlación lineal	0.348
Ocurrencias	22,766

- KNN (Sin selección de Atributos) -81%



## Análisis Anotaciones

– Contraer todo

+ Desplegar todo

Resultados para el campo de resultado HeartDiseaseorAttack

Comparando \$KNN-HeartDiseaseorAttack con HeartDiseaseorAttack

Error mínimo	-1.0
Error máximo	1.0
Error promedio	-0.236
Error absoluto promedio	0.27
Desviación estándar	0.463
Correlación lineal	0.344
Ocurrencias	22,743

- Ruta IBM

