"Sistema para análisis de comportamiento de consumo de la población diabética ecuatoriana aplicando técnicas de minería de datos"

Fabricio Echeverría[1], Alfredo Cáceres Zambrano[2], Mario Iturralde Orellana[3], David Perugachi Rojas[4].

[1] Director de Tópico, Magíster en Sistema de información Gerencial, 2006, Profesor de ESPOL; e-mail: pechever@uniplex.com.ec
[2] Ingeniero en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: acaceres@espol.edu.ec
[3] Ingeniero en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: miturral@tecnoquimicas.com.ec
[4] Ingeniero en Computación Especialización Sistemas de Información, 2006; e-mail: dperugachi@onlycontrol.com

Resumen

El presente proyecto está basado en el desarrollo de un "sistema para el análisis de comportamiento de consumo de la población diabética ecuatoriana (SACC)" el cual tiene como objetivo primordial mostrar cuales son las características demográficas de los diabéticos en el Ecuador; permitiendo describir cuales son los patrones de consumo de productos especializados que ayudan al paciente al tratamiento de la diabetes y mejora de los síntomas, dando como resultado el ofrecer a los distribuidores de dichos productos potenciales oportunidades de mejora en la promoción y distribución de los mismos.

Además el presente trabajo estima la tendencia de una persona a desarrollar diabetes dependiendo de factores de comportamiento, consumo, hereditarios y de cuidado de su salud.

Para el proceso de extracción del conocimiento, el SACC tiene como base el algoritmo de Naive Bayes para el modelo predictivo y el algoritmo de reglas de asociación para el modelo descriptivo. Para construir los modelos se utilizaron las características demográficas de la población, así como también información de ventas del primer semestre del presente año.

Palabras Claves: *Minería Datos, Modelo Predictivo, Predicción, Naive Bayes, Reglas de asociación, Diabetes.*

Abstract

The present project is based on the development of a "system for the analysis of behavior of the Ecuadorian diabetic population's consumption (SACC)", which has as fundamental objective to show which are the demographic characteristics of the diabetic people in Ecuador. This system allows to describe which are the patterns of consumption of specialized products that help the patient with the treatment of the diabetes and the improvement of the symptoms, giving the distributors of these products a tool to find out the potential opportunities in the promotion and distribution of the same ones.

The present work also estimates the tendency of a person to develop diabetes depending on behavior factors, consumption, hereditary patterns and care of its health.

For the process of the knowledge extraction, SACC has as basis the Naive Bayes's algorithm for the predictive model and the association rule's algorithm for the descriptive model. To build the models the population's demographic characteristics were used, as well as the sales of the first semester of the present year.

Introducción

La Diabetes Mellitus es un grupo de trastornos metabólicos de carácter crónico caracterizados por un elemento común, la hiperglucemia, que contribuye al desarrollo de complicaciones macro vasculares, micro vasculares y neuropáticas, lo que la sitúa como una de las principales causas de mortalidad de las sociedades desarrolladas o en vías de desarrollo. La Organización Mundial para la Salud relaciona el aumento de la diabetes con el crecimiento y envejecimiento de la población, el incremento de la obesidad, hábitos erróneos de la alimentación y modos de vida sedentarios.

Nuestro objetivo es aplicar minería de datos sobre una problemática de nivel social; para tener como resultado una herramienta que permita conocer los perfiles del consumidor de productos para diabéticos; así como también la tendencias que tiene una persona a sufrir esta enfermedad debido a su estilo de vida, factores de salud y características étnicas.

1. Definición del Problema y aplicación

Se identificaron las dos problemáticas más relevantes con respecto al análisis de una enfermedad, en este caso, la diabetes: El análisis de los factores de riesgo e incidencia de la enfermedad tomando en cuenta la población urbana del Ecuador y el análisis de los patrones de consumo de los productos para el tratamiento de la enfermedad tomando en cuenta la información de la base de datos transaccional de una empresa distribuidora de estos productos.

El análisis de factores de riesgo tuvo como tarea inicial la recolección de información de la población mediante una encuesta en línea, la cual sirvió como base para la construcción del modelo predictivo. De esta manera un usuario puede acceder al sitio www.diabetesecuador.com y realizar un test y de forma muy sencilla conocer la probabilidad estimada de desarrollar la enfermedad en un futuro basado en parámetros como: edad, sexo, peso, hábitos alimenticios entre otros.

En el análisis de patrones de consumo la fuente de información son las ventas de una empresa farmacéutica en donde se detallan los valores por artículo en una determinada fecha. Estos datos son los parámetros de entrada del modelo descriptivo utilizando reglas de asociación.

Luego de aplicar el modelo descriptivo el usuario, que en este caso corresponde a un distribuidor, puede concluir qué artículos generan mayores ventas acompañados de otros, herramienta útil a la hora de promocionarlos.

Hay que tener presente que para que el proceso de extracción del conocimiento tenga éxito los datos deben pasar por un proceso de discretización a fin de garantizar la calidad de los mismos. Además se deben tener conocimientos básicos sobre la enfermedad para poder interpretar el resultado obtenido de los modelos de minería de datos

2. Algoritmo predictivo: Red Bayesiana aplicada al análisis de factores de riesgo.

El modelo predictivo se enfoca en generar modelos que descubren relaciones ocultas y complejas a partir de diversas operaciones. Las tareas asociadas al análisis de tipo predictivo son: las redes neuronales, árboles de decisión, redes bayesianas, modelos de regresión, entre otros.

El SACC implementa el modelo predictivo mediante la aplicación del algoritmo de Naive Bayes el cual es una técnica de minería de datos que utiliza conocimiento previo para generar un modelo predictivo que estima la probabilidad de una hipótesis, lo que permite calcular las propensiones sobre cualquiera de las variables que existen en el modelo. Cada nuevo individuo que se evalúa sirve como aprendizaje para la red bayesiana mejorando el porcentaje de precisión.

Por ejemplo se puede calcular cual es la probabilidad para la clase sexo (hombre, mujer) dadas las siguientes premisas: padece diabetes tipo 2, reside en Guayaquil, su edad esta en el rango de adulto mayor y trabaja. Este análisis le puede indicar al usuario hacia que grupo objetivo debería enfocar una campaña publicitaria para personas con el perfil descrito en la premisas, si es mas conveniente realizarla enfocada hacia hombres o mujeres.

Con una red bayesiana, el SACC adquiere las siguientes características:

- Cada ejemplo de entrenamiento afecta a la probabilidad de las hipótesis. Esto es más efectivo que descartar directamente las hipótesis incompatibles.

- Se puede incluir conocimiento a priori: probabilidad de cada hipótesis; y la distribución de probabilidades de los ejemplos.

- Es sencillo asociar un porcentaje de confianza a las predicciones, y combinar predicciones en base a su confianza.

Dificultades del método:

Necesidad de un volumen considerable de información dado que este método considera el conocimiento preexistente para estimar la probabilidad de un nuevo individuo. Mientras mas datos, más exactitud.

Se necesita en la mayor parte de las veces normalizar los datos, es decir, eliminar redundancias entre los datos, establecer variables discretas así como establecer un mismo dominio para los datos de entrada.

3. Algoritmo descriptivo: Reglas de Asociación aplicado al análisis de comportamiento de consumo.

Las reglas de asociación son una manera muy popular de expresar patrones de una base de datos. Estás surgen inicialmente para afrontar el análisis de las cestas de compra de los comercios con el fin de conocer el comportamiento de los compradores a la hora de tomar la decisión de que producto comprar.

En este proyecto las reglas de asociación se generan a partir de las compras realizadas por los consumidores de productos especializados para la diabetes. Para ello se discretizan los datos adaptándolos al modelo de minería de datos. El resultado es una serie de reglas que describen la tendencia de compra del cliente o de venta del artículo. De esta manera obtenemos reglas de la forma:

Si[art1]^[art2]…=> [art3]^[art4]…

Por ejemplo se presenta como resultado de la ejecución del algoritmo que: los reactivos para la medición personal del nivel de glucosa están altamente relacionados con el consumo de medicamentos de tipo inyectables. Esto permitiría realizar una combinación de estos productos (a un menor precio), para una comercialización más efectiva.

Se deben tener en cuenta ciertos parámetros adicionales que ayudan a la exactitud y a la convergencia del modelo, los cuales son: la cobertura y la confianza.

Cobertura: se define como el número de instancias que la regla predice correctamente. (También se utiliza el porcentaje). *Confianza:* mide el porcentaje de veces que la regla se cumple cuando se puede aplicar.

Algunas características de las reglas de asociación son:

- Las reglas de asociación están siempre definidas sobre atributos binarios.
- La generación de los atributos binarios para la aplicación de las reglas de asociación no es complicada inclusive en grandes bases de datos.
- El usuario puede determinar la confianza y cobertura de las reglas de asociación

Dificultades del método:

- El problema es que tal algoritmo eventualmente puede dar información que no es relevante.
- El algoritmo tiende a consumir altos niveles de recursos cuando no se define buenos parámetros de confianza y cobertura.

4. Resultados y Conclusiones

En las pruebas realizadas se pudo observar que el algoritmo de reglas de asociación es un descriptor muy preciso y útil a la hora de extraer conocimiento de un conjunto de datos que están relacionados entre si.

Se necesita una población extensa para que el algoritmo de redes bayaesianas sea preciso, ya que con cada individuo se entrena y se perfecciona.

Los parámetros de cobertura y confianza son vitales a la hora de aplicar las reglas de asociación ya que estos inciden en la convergencia y exactitud del método. Se aconseja tener una confianza alta, así se garantiza que la regla sea efectiva. Hay que tener presente que la cobertura siempre es menor que la confianza.

Las reglas de asociación encontradas describen correctamente el comportamiento del consumidor, esto ayudara a las empresas a ofertar acertadamente lo que el consumidor realmente quiere.

5. Agradecimientos

En primer lugar agradecemos a Dios por guiarnos en nuestro camino, por ser quien nos ilumina siempre y porque nos acompaña en cada situación y está presente en nuestras decisiones.

A nuestros padres por ser nuestros pilares en nuestra educación y desarrollo personal, gracias a ellos que siempre han confiado en cada uno de nosotros y están en cada uno de los momentos que más los necesitamos.

Y a cada uno de nuestros maestros que han puesto un granito de arena para nuestro crecimiento como grandes profesionales ya que gracias a sus enseñanzas estamos listos para triunfar en el mundo.

Referencias

a) Referencias de Internet

[1]http://www.latino-bi.com/servicios/business-intelligence/datamining.htm
[2]
[3]

b) Referencias de libros

- HERNÁNDEZ, J.; RAMÍREZ, M.J. y FERRI, C. Introducción a la Minería de Datos, Prentice Hall, España, primera edición, 2004.
- HAIR, J.F.; ANDERSON, R.E.; TATHAM, R.L. y BLACK, W.C. Análisis multivariante, Prentice Hall, Madrid, 1999