Sistema generador de alertas que ayuden a prevenir el embarazo a temprana edad utilizando minería de datos

David Mauricio Moreno Torres

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas y Computación
Bogotá, Colombia
2016

Sistema generador de alertas que ayuden a prevenir el embarazo a temprana edad utilizando minería de datos

David Mauricio Moreno Torres

Trabajo final presentado como requisito para optar al título de:
Magister en Ingeniería de Sistemas y Computación

Director(a):
Ph.D. Luis Fernando Niño

Grupo de Investigación:
Laboratorio de Investigación en Sistemas Inteligentes

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas y Computación
Bogotá, Colombia
2016

Dedicatoria

A Todas aquellas niñas y adolescentes de Bogotá que pudieran tener una mejor planificación de sus vidas, la construcción de sus sueños y las oportunidades de disfrutar sus etapas de niñez y adolescencia sin tener que madurar tan rápido a causa de un embarazo, a todas ellas les dedico este trabajo con la esperanza de que algún día los gobernantes de este país piensen en desarrollar proyectos que realmente estén dirigidos al bienestar de vida de los ciudadano por encima de sus intereses propios.

Agradecimientos

Agradezco a Dios quien es el artífice de mi vida y de los logros que obtengo, porque es Él quien brinda los medios para cumplir mis sueños y mis objetivos.

También, agradezco a todas las personas que contribuyeron al desarrollo de este trabajo de investigación, a aquellas que me motivaron y me dieron buenos consejos.

Al profesor Luis Fernando Niño, PhD, Director del grupo de Investigación, por su paciencia, disposición y tiempo dedicado para guiarme, por sus recomendaciones, por contribuir a mi formación profesional y personal.

Al grupo de investigación porque en las reuniones de grupo sus recomendaciones y comentarios permitieron mejorar el trabajo realizado. En especial a Carlos Manuel Steves quien siempre fue un gran apoyo en sus contribuciones y entusiasmo mostrado en este trabajo.

A la Universidad Nacional de Colombia y a los profesores que guiaron con sus recomendaciones el desarrollo de la investigación y la elaboración del documento final.

A la Doctora Luz Adriana Zuluaga funcionaria de la Secretaría Distrital de Salud, por haber compartido la idea de realizar este trabajo y por haberme guiado en el marco conceptual de la problemática global. Por haberme planteado la necesidad de contar con modelos predictivos que ayuden a la construcción de una mejor sociedad basados en mitigar problemas de salud pública como el que se aborda en este trabajo investigativo.

A mi familia quienes me han apoyado siempre, infinitas gracias.

Resumen

Durante esta investigación se llevó a cabo un proceso conocido como KDD (*Knowledge Discovery in Databases*), el cual involucra un trabajo extenso en minería de datos como uno de los pasos más importantes de este proceso. La investigación se realizó bajos dos objetivos principales: Encontrar patrones de comportamiento relacionados con el embarazo a temprana edad en las poblaciones más vulnerables de Bogotá y generar un modelo predictivo capaz de identificar las adolescentes con mayor riesgo de quedar en embarazo. El desarrollo de este modelo predictivo se basó en la teoría de determinantes de la salud de la OMS (Organización Mundial de la Salud), la cual indica que existen ciertos factores de riesgo asociados al embarazo adolescente determinados por factores sociales, económicos, personales y de entorno. Finalmente, con el propósito de brindar una herramienta tecnológica a los equipos médicos de promoción y prevención del distrito, se desarrolló una solución que involucra una aplicación móvil para geolocalizar a las mujeres menores de 20 años con mayor riesgo usando el modelo predictivo desarrollado.

***Palabras Clave*— Minería de datos, Aprendizaje de Máquina, KDD, Embarazo adolescente**

Abstract

In this work, a process known as KDD (Knowledge Discovery in Databases), which involves extensive work in data mining as one of the most important steps in this process, was carried out. The main goal was to find patterns associated with early pregnancy in the most vulnerable population in Bogotá and to develop a predictive model in order to identify the women with high risk of becoming pregnant during their girlhood, adolescence and part of their youth. The predictive model developed was based on the theory of health determinants of the WHO (World Health Organization). This theory states that teenage pregnancy risk is associated with social, economic, personal and environmental factors. Finally, a software tool (an app) to geolocate teenagers girls with most high risk detected by the predictive model was implemented. Such app could be used by medical professionals from health programs in Bogotá.

***Keywords Data*— Mining, Knowledge Discovery in Databases, kDD, Machine Learning**

Contenido

Abstract IX

Lista de figuras XIV

1. Marco teórico 3

1.1 Descubrimiento de Conocimiento en Bases de Datos – KDD 3

1.2.1 Identificación del Problema y del Dominio de Trabajo 6

1.2.2 Crear el conjunto de datos. 7

1.2.3 Preprocesamiento de los Datos 7

Algunos métodos para el preprocesamiento 8

- ML_T2LA 8

1.2.4 Reducción de datos 9

1.2.5 Formulación de los objetivos. 9

1.2.6 Exploración de análisis, modelo y selección de la hipótesis. 9

1.2.7 Minería de Datos 9

Objetivos de la minería de datos 9

Consultas 10

Modelos de regresión 10

Modelos basados en *Clustering*, Segmentación o Agrupación 11

Modelos de asociación (Summarization) 12

Modelos basados en Clasificación 12

1.2.8 Interpretación de los patrones encontrados 14

1.2.9 Evaluación de los modelos predictivos elaborados con clasificadores 14

Métricas de rendimiento 14

Curvas ROC 16

1.3 Aproximaciones en el campo del cuidado de la salud 16
1.4 Tendencias 19
2 Enfoque de la Investigación 21
2.1 Planteamiento del problema 21
2.2 Hipótesis 23
2.3 Metodología 23
2.4 Enfoque propuesto 23
2.5 Objetivos 24
3 Primera aproximación - búsqueda de patrones y generación del modelo predictivo 27
3.1 Preprocesamiento 27
3.1.1 Reducción de dimensionalidad 28
Reducción por correlaciones entre las variables 32
3.2 Minería de Datos 34
2.1.1 Métodos de asociación 34
A priori y crecimiento de patrones frecuentes (pattern growth) 34
Balanceos de datos con respecto a la clase 36
Búsqueda de asociaciones con las muestras de datos balanceados 36
3.2.1 Métodos de clasificación 38
Árboles de decisión 38
Redes Bayesianas 40
Ascenso de Colina (*Hill Climbing*) 41
Naive Bayes 43
Redes neuronales 44
3.2.2 Métodos de agrupación 50
Agrupamiento basado en distribución 50
METODOS BASADOS EN LA DENSIDAD 54
3.3 Conclusiones de resultados con el primer conjunto de datos 57
4 Segunda Aproximación – Mejorando el modelo predictivo 58
4.1 Pre-procesamiento 61
4.1.1 Balanceo de datos 62

4.1.2 Transformación de datos 63
4.2 Minería de Datos 63
4.2.1 Perceptron Multicapa Utilizando Rprop Como Algoritmo De Optimización. 64
4.2.2 Redes Neuronales Probabilísticas entrenada con algoritmo DDA 67
4.2.3 Naive Bayes 70
4.2.4 Árboles de decisión 71
4.3 Conclusiones segunda aproximación 74
5 Aplicación móvil desarrollada 77
5.1 Descripción general de la aplicación 77
5.2 Arquitectura de la aplicación 79
5.2.1 Casos de Uso 79
5.2.2 Vista Lógica 79
5.2.3 Vista Física 81
5.2.4 Modelo Entidad Relación 82
..... 83
6 Conclusiones y recomendaciones 85
6.1 Conclusiones 85
6.2 Recomendaciones 86

Lista de figuras

Figura 1. Pasos para realizar KDD. 6
Figura 2. Ejemplo ilustrativo de datos sobre embarazo adolescente en mujeres según edad y educación. 6
Figura 3. Método de limpieza con recorrido de pila (Hernández & Stolfo, 1998) 8
Figura 4. Regresión lineal de primer embarazo en mujeres según edad y educación. (Ejemplo ilustrativo) l 11
Figura 5. Segmentación de primer embarazo en mujeres según edad y educación. (Ejemplo ilustrativo) 11
Figura 6. Modelo general para la construcción de un modelo de clasificación. 13
Figura 7. Clasificación de primer embarazo en mujeres según edad y educación. (Ejemplo ilustrativo) 13
Figura 8. Curvas ROC para dos clasificadores no discretos 16
Figura 9. Estadísticas de embarazo adolescente reincidente 22
Figura 10. Datos estadísticos del número de personas dentro de la familia 30
Figura 11. Histograma Nivel Educativo. 31
Figura 12. Plano euclidiano de la variable Educación 32
Figura 13. Diagrama de correlación de variables 33
Figura 14. Correlación entre variables del nivel educativo y la edad 34
Figura 15. Histograma atributo embarazo 35
Figura 16. Esquema del balanceo del atributo embarazo usando la herramienta KNIME. A la derecha se muestra el histograma 36
Figura 17. Modelo desarrollado con árboles de decisión. 38
Figura 18. Primer nivel del árbol de decisión 39
Figura 19. Red Bayesiana generada con algoritmo de búsqueda K2 41
Figura 20. Red Bayesiana generada con algoritmo de entrenamiento Ascenso de colina. 43
Figura 21. Arquitectura de una red neuronal probabilística (PNN). 47
Figura 22. Modelo de aplicación para redes neuronales PNN y RProp Multilayer Algoritm. 47
Figura 23. Ejemplo de probabilidades asignadas para la clasificación dadas por el modelo PNN a cada registro. 48

Figura 24. Ilustración de grupos generados con DBSCAN como algoritmo de agrupamiento 55
Figura 25. Comportamiento del agrupamiento del conjunto de datos con el algoritmo k-NN respecto a registros cuyo atributo Embarazo es 1. 56
Figura 26. Modelo desarrollado para clasificadores con redes neuronales de perceptron multicapa entrenado y probado con métodos de 10 validación cruzada y algoritmo RProp como algoritmo de entrenamiento 65
Figura 27. Modelo desarrollado para clasificadores con redes neuronales PNN entrenado y probado con métodos de validación cruzada y algoritmo DDA como algoritmo de entrenamiento. 67
Figura 28. Modelo desarrollado para clasificadores con Naive Bayes entrenado y probado con métodos de validación cruzada 70
Figura 29. Modelo desarrollado para clasificador con Árboles de decisión entrenado y probado con métodos de validación cruzada 72
Figura 30. Árbol de decisión obtenido por el modelo para el segundo set de datos balanceado 73
Figura 31. Diagrama de Pareto para los modelos predictivos desarrollados 75
Figura 32. Curva ROC para el clasificador PNN con algoritmo de entrenamiento DDA 76
Figura 33. Ilustración general de la aplicación móvil desarrollada (prevengo.app) 78
Figura 34. Diagrama de casos de uso 79
Figura 35. Diagrama de clases 80
Figura 36. Diagrama de secuencias 80
Figura 37. Vistas graficas del aplicativo desarrollado (prevengo.app) 81
Figura 38. Modelo Entidad Relación (prevengo.app) 83

Lista de tablas

Tabla 1. Matriz de confusión para un problema de clasificación binaria 15
Tabla 2. Matriz de confusión para un problema de clasificación SI/NO 15
Tabla 3. Determinantes proximales asociados por la OMS al embarazo adolescente 28
Tabla 4. Determinantes intermedios asociados por la OMS al embarazo adolescente 28
Tabla 5. Determinantes distales asociados por la OMS al embarazo adolescente. 29
Tabla 6. Matriz de confusión resultante de clasificar 220 registros con el modelo de árboles de decisión. 40
Tabla 7. Matriz de resultados obtenidos de la red bayesiana utilizando K2 como algoritmo de búsqueda 41
Tabla 8. . Matriz de confusión obtenida de la red bayesiana utilizando K2 como algoritmo de búsqueda 41
Tabla 9. Matriz de resultados obtenidos de la red bayesiana utilizando Ascenso de Colina como algoritmo de búsqueda. 42
Tabla 10. Matriz de confusión obtenida de la red bayesiana utilizando ascenso de colina como algoritmo de búsqueda. 42
Tabla 11. Matriz de resultados con la aplicación de Naive bayes 43
Tabla 12. Matriz de confusión obtenida con la aplicación de Naive Bayes como método de clasificación para el conjunto de datos 44
Tabla 13. Matriz de resultados con clasificación de redes neuronales entrenado por gradiente descendiente 44
Tabla 14. Resultados de clasificaciones usando RProp MLP 45
Tabla 15. Reglas generadas por el modelo de RNN entrenado con Algoritmo DDA 48
Tabla 16. Resultados de clasificaciones utilizando PNN con algoritmo de entrenamiento DDA 48
Tabla 17. Resultado de Agrupamiento EM de 2 grupos con respecto a los registros de la etiqueta clase 50
Tabla 18. Resultado de Agrupamiento EM de 3 grupos con respecto a los registros de la etiqueta clase 52
Tabla 19. . Resultado de Agrupamiento EM de 4 grupos con respecto a los registros de la etiqueta clase 52

Tabla 20. Resultado de Agrupamiento EM de 5 grupos con respecto a los registros de la etiqueta clase 52
Tabla 21. Resultado de Agrupamiento EM de 6 grupos con respecto a los registros de la etiqueta clase 52
Tabla 22. Resultado de Agrupamiento k-medias de 2 grupos con respecto a los registros de la etiqueta clase 52
Tabla 23. . Resultado de Agrupamiento k-medias de 3 grupos con respecto a los registros de la etiqueta clase 53
Tabla 24. Resultado de Agrupamiento k-medias de 4 grupos con respecto a los registros de la etiqueta clase 53
Tabla 25. Resultado de Agrupamiento k-medias de 5 grupos con respecto a los registros de la etiqueta clase 53
Tabla 26. Resultado de Agrupamiento k-medias de 6 grupos con respecto a los registros de la etiqueta clase 53
Tabla 27. Características descriptivas de los clusters generados con métodos basados en distribución 54
Tabla 28. Resultados obtenidos durante experimentos con k-NN 56
Tabla 29. Asociación de variables con los determinantes sociales de la salud relacionados al riesgo de embarazo a temprana edad 59
Tabla 30. Resultado de entrenamiento de modelo clasificador con red neuronal PNN con técnicas de entrenamiento con el set de datos original y el set de datos balanceado 62
Tabla 31. Resultados modelo predictivo con red neuronal Rprop MPL. Número máximo de iteraciones: 100 65
Tabla 32. Resultado de la clasificación del conjunto de datos con redes neuronales RProp MPL 66
Tabla 33. Resultados modelo predictivo con red neuronal PNN. 67
Tabla 34. Resultados de clasificación del conjunto de datos con los modelos de red neuronal PNN 69
Tabla 35. Las 8 reglas generadas por el modelo PNN con mayor peso para la clasificación . 69
Tabla 36. Resultados modelo predictivo con Naive Bayes. Máximo número de valores nominales permitidos por variable: (2, 3, 5, 7 y 9) 70
Tabla 37. Resultados de clasificación del conjunto de datos con el modelo de desarrollado con árboles de decisión 74
Tabla 38. Diagrama de Pareto para los modelos predictivos desarrollados 75

Introducción

Este trabajo se enfoca en analizar un conjunto de datos referentes a registros de niñas y adolescentes captadas mediante el programa denominado “Salud a su Hogar” de la Secretaría Distrital de Salud (SDS), el cual contiene variables socio-económicas de la población bogotana intervenida durante los últimos 10 años por equipos médicos que realizan trabajos enfocados en la política mundial de Atención Primaria en Salud (APS). El programa de la Secretaría Distrital de Salud opera en 19 localidades de la ciudad enfocado en las poblaciones más vulnerables. La mayoría de las personas caracterizadas en el sistema de información pertenecen a los estratos socioeconómicos más bajos (1, 2 y parte del 3), por lo cual los patrones encontrados y el comportamiento de las variables socioeconómicas no son una representación de toda la ciudad, pero hacen una síntesis de la realidad de las poblaciones más vulnerables. Con los datos obtenidos en el sistema de información que respalda dicho programa de salud, se realizaron dos aproximaciones buscando encontrar nuevo conocimiento acerca de la problemática del embarazo adolescente y la construcción del modelo predictivo más efectivo mediante la aplicación de diferentes técnicas computacionales.

Para la primera aproximación se tomó un total de 209.978 registros de la base de datos del sistema de información referentes a mujeres menores de 21 años al momento de ser caracterizadas por el programa. Dichos registros cuentan con datos poblacionales, socioeconómicos y el registro que indica si se encontraba en estado de gestación durante esta captación. Con este primer conjunto de datos se probaron varias técnicas comunes en minería de datos basadas en asociación de reglas, clasificación y agrupación para la generación del modelo predictivo. También se logró hacer hallazgos importantes respecto a correlación de variables que ayudan a entender mejor la problemática del embarazo adolescente en Bogotá. Sin embargo cuando se utilizaron todos los registros no se lograban generar reglas de asociación con respecto a los embarazos por la gran cantidad de casos negativos que presenta el conjunto de datos. Finalmente los modelos se probaron con 22000 registros. Durante esta etapa se evaluaron los modelos con respecto al 10% de los datos utilizados.

La segunda aproximación se realizó con 5743 registros de mujeres activas en el programa que para finales del año 2015 tenían 21 años y quienes habían sido captadas antes del

2006. Con este segundo enfoque se logró mejorar los modelos predictivos desarrollados durante la primera aproximación, debido a que se tenía el historial de las adolescentes y saber si habían quedado o no en embarazo durante su adolescencia. En esta aproximación se ingresaron menos variables al modelo dejando únicamente las de mayor influencia para la generación de modelos predictivos detectadas durante la primera aproximación y aquellas que tienen una mejor calidad respecto a datos nulos o fuera de rango.

El propósito final de esta investigación fue la creación de una herramienta tecnológica que le permitiera a la Secretaría Distrital de Salud, enfocar sus campañas de promoción y prevención a las mujeres adolescentes con mayor riesgo de quedar en embarazo a temprana edad. Con base en este propósito se desarrolló una aplicación móvil que permite georreferenciar el lugar donde viven las adolescentes y a quienes el modelo puede clasificar en alto riesgo de quedar en embarazo. La información contenida en esta aplicación es muy sensible y debe ser únicamente operada por los equipos médicos dedicados al programa de Atención Primaria en Salud de la red distrital de hospitales con los cuales el gobierno lleva a cabo este programa. Esto podría permitir a la SDS realizar campañas enfocadas a la población más vulnerable a que esto ocurra y teóricamente tener mejores resultados en la disminución de embarazos tempranos.

1. Marco teórico

El Descubrimiento de conocimiento en bases de datos (KDD - *Knowledge Discovery in Databases*) ha sido estudiado intensamente los últimos años a pesar de ser un campo de investigación joven por ser un tema que involucra una gran cantidad de técnicas computacionales de última generación. En las investigaciones relacionadas con este campo se pueden encontrar trabajos realizados con métodos de aprendizaje maquinal desarrollados por medio de técnicas estadísticas. Estas técnicas se han utilizado para extraer las reglas de clasificación de diferentes conjuntos de datos (Zhang, Hu, & Xie, 2015). En muchos escritos se suele mencionar KDD indiscriminadamente con Minería de Datos como si se trataran de lo mismo, en este apartado se hace una introducción al término KDD y a minería de datos, se muestra las principales diferencias y la relación entre los dos términos para definir la metodología y el enfoque que se dio a la investigación en estos campos. Se hace una revisión general del proceso de KDD en donde se van exponiendo trabajos recientes que involucran la profundización de cada etapa o algoritmo aplicado, este capítulo se desarrolla dentro del paradigma que plantea una diferencia entre KDD como proceso general y Minería de Datos como la etapa más importante de este proceso. Se hace una revisión de las aproximaciones investigativas en el campo de la salud y las diferentes aplicaciones de los métodos de clasificación y aprendizaje maquinal relacionando algunos ejemplos de investigaciones recientes que hacen uso de dichas técnicas.

1.1 Descubrimiento de Conocimiento en Bases de Datos – KDD

Por mucho tiempo el término minería de datos ha sido relacionado con las técnicas de inteligencia artificial (AI) y aprendizaje de máquina que permite extraer nueva información útil basándose en la detección de patrones entre los datos almacenados (Körting, Garcia, & Camara, 2013). Por otro lado, el término KDD fue mencionado por primera vez por Piatetsky-Shapiro en 1989 para describir el proceso en el cual se logra extraer de una base de datos información útil hasta ahora desconocida para quien realiza el proceso. Este proceso involucra un conjunto de etapas definidas para el tratamiento de los datos, antes de aplicar las diferentes técnicas de minería de datos en la búsqueda de patrones ocultos en los datos y finalmente hacer el análisis de los patrones encontrados cuyo objeto es brindar nueva información y conocimiento (Thakur & Mahajan, 2015).

Hasta hace unos años minería de datos y KDD eran mencionadas indiscriminadamente para referirse a la extracción de patrones de las variables en una base de datos. Hasta hace poco, se puede decir que menos de 10 años, algunos investigadores han comenzado a utilizar el término KDD para referirse a la extracción de conocimiento proveniente de bases de datos como un macro proceso, mientras que minería de datos es catalogada como el mecanismo de aplicación de algoritmos para extraer patrones presentados por los datos. Se dice entonces que KDD es el proceso para la identificación de patrones válidos, nuevos, útiles y sobretodo comprensibles, que conlleva al descubrimiento de nuevo conocimiento. En cambio, la minería de datos es tan solo una de las etapas de este proceso, considerada por muchos como la más importante (Fayyad & Stolorz, 1997), (Fayyad & Uthurusamy, 1996) (Körting, Garcia, & Camara, 2013).

Esta técnica de extracción de información es utilizada ampliamente en muchas áreas como Economía para descubrir tendencias del mercado, Cuidado de la Salud para realizar prevención o anticipar un diagnóstico, Mercadeo para encontrar patrones entre los tipos de clientes y sus compras predilectas, Seguridad Informática en la detección anormal de una relación lo cual implicaría una alerta de fraude, entre muchos otros ejemplos (Magnisalis, Demetriadis, & Karakostas, 2011).

En la actualidad los sistemas de información comúnmente cuentan con bases de datos demasiado robustas. La capacidad de almacenamiento se ha incrementado en los últimos años de la misma manera que lo han hecho los datos y variables de información en las compañías, entes gubernamentales, universidades y demás instituciones. Esta particularidad hace que el análisis de la información almacenada sea complejo para quienes requieren tomar decisiones prácticas con base a los dato. El uso de herramientas que permitan extraer información de las bases de datos ha tomado fuerza en los campos de investigación como estadística, aprendizaje de máquina, bases de datos entre otras (Huang, Hsu, & Wang, 2012). En el campo de la salud los estudios estadísticos, de minería de datos y de adquisición de información en bases de datos han generado importantes contribuciones para el área del cuidado de la salud. En la cual las oportunidades de construcción de conocimiento basado en la información de prácticas clínicas es de vital importancia para las intervenciones futuras que se puedan realizar a los pacientes (Goodwin, Vandyne, Lin, & Talbert, 2003).

Este capítulo está organizado de la siguiente manera. En la sección 1.2 se presenta una breve explicación del proceso KDD, describiendo cada una de las etapas, luego las técnicas más usadas en la detección de patrones en bases de datos. También se aborda las técnicas más conocidas de extracción de conocimiento sobre bases de datos y se ponen ejemplos de trabajos que profundizan en cada técnica en específico, en algunos casos se hace mención a trabajos recientes que relacionan dichas técnicas con la investigación en el campo del cuidado de la salud. En la sección 1.3 se hace una breve revisión de algunos trabajos relativamente recientes relacionados al área de la salud, se mencionan las técnicas utilizadas en estos trabajos para la extracción de la información y el propósito con el cual el investigador aplicó dichas técnicas. Se muestran ejemplos específicos en estudios relacionados con la detección de patrones y clasificación de datos para diagnósticos

médicos de diferente tipo de enfermedades. La sección 1.4 hace referencia a las tendencias de futuras investigaciones según los expertos en este tema.

Proceso para el descubrimiento de conocimiento en bases de datos

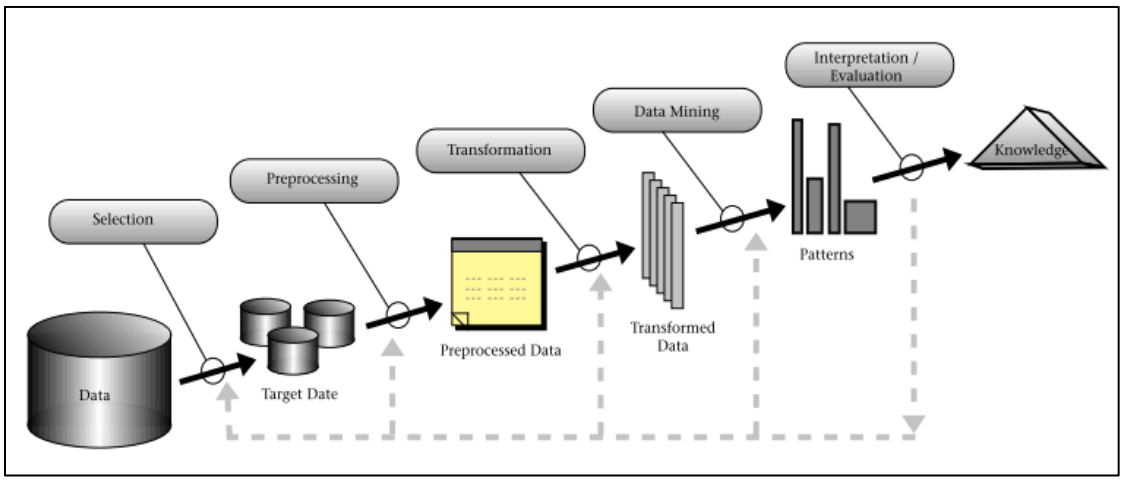
El almacenamiento de datos en los diferentes sistemas de información ha venido incrementándose de manera dramática, alejándose de las posibilidades humanas para extraer información útil de manera eficiente. Por este motivo, es necesario el uso de métodos que ayuden a las personas a interpretar la información almacenada en estas enormes fuentes de datos y poder extraer nuevo conocimiento. El proceso conocido como KDD puede tener algunos pasos como: Selección, reducción de los datos, minería de datos y evaluación de los hallazgos (Verma, 2015).

Los pasos que involucran la extracción de conocimiento de las bases de datos son nueve en general, la secuencia es importante para la obtención de los resultados esperados. En algunos casos puede llegar a ser necesario regresar tras la identificación de alguna oportunidad de mejora en el tratamiento de los datos. Las nueve etapas de KDD son:

1. Identificación del problema y del dominio de trabajo.
2. Crear el conjunto de datos.
3. Pre-procesamiento de los datos.
4. Reducción de datos y proyección.
5. Formulación de los objetivos del PROCESO KDD (Paso 1)
6. Exploración de análisis, modelo y selección de la hipótesis.
7. Minería de datos
8. Interpretación de los patrones encontrados
9. Descubrimiento de nuevo conocimiento (Fayyad & Uthurusamy, 1996), (Köksal, Batmaz, & Testik, 2011).

En la Figura 1 se presenta de manera resumida los pasos o etapas para la obtención de nuevo conocimiento en bases de datos. De aquí en adelante se hace una pequeña introducción a cada uno de estos pasos, los algoritmos y técnicas usadas y algunos trabajos recientes que enfatizan en cada etapa son mencionados o referenciados.

Figura 1. Pasos para realizar KDD.

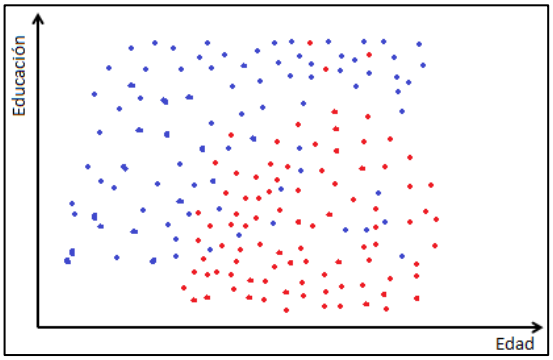


Fuente: Fayyad, U. (1996)

1.2.1 Identificación del Problema y del Dominio de Trabajo

Esta primera etapa consiste en identificar el objetivo para el cual se va a realizar KDD, se debe tener en cuenta que no todos los datos de un dominio serán útiles para el resultado esperado. Se debe definir el subdominio de datos que son relevantes para la aplicación de la técnica (Matheus, Chan, & Piatetsky-Shapiro, 1993). En el 2011 se presentó un trabajo que hace una gran profundización en la selección de dominio para la detección de hipótesis médicas donde los investigadores hacen referencia a la importancia de las reglas de asociación dentro del dominio específico (Then, 2011). Por ejemplo, para el caso de la detección de mujeres embarazadas menores de edad, se deben sesgar solo los registros que cumplan con dichas características y las variables asociadas a esos registros.

Figura 2. Ejemplo ilustrativo de datos sobre embarazo adolescente en mujeres según edad y educación.



1.2.2 Crear el conjunto de datos.

En la figura 2 por ejemplo se tiene a manera de ilustración los casos de embarazo con respecto a los determinantes intermedios y distales[1] de un individuo con respecto a factores personales como la educación y la edad. Para este caso en particular los colores indican la condición verdadera o falsa de un embarazo adolescente. El dominio en este caso estaría representado por registros pertenecientes a mujeres entre los 10 y los 20 años. En el conjunto de datos seleccionado se deben incluir atributos o variables relacionadas a sus condiciones de entorno social, condiciones económicas, de educación entre otras que puedan influir directa o indirectamente a la condición dada por el problema abordado.

1.2.3 Preprocesamiento de los Datos

Uno de los grandes retos cuando se trabaja con bases de datos robustas es precisamente la cantidad de datos que estas contienen, generalmente los datos son introducidos manualmente y suelen tener errores. Esto se conoce como ruido y este es precisamente uno de los casos más frecuentes en los registros clínicos ingresados de manera textual por los profesionales de la salud. Trabajos recientes trabajan en técnicas para el tratamiento de este problema en registros clínicos (Friedlin, Mahoui, Jones, & Jamieson, 2011). Para eliminar los datos inconsistentes es necesario hacer una clasificación de los datos con los cuales se va a trabajar (Chai, Liu, & Ngai, 2013). El preprocesamiento busca mejorar la calidad de los datos y reducir el riesgo de error. Las técnicas básicas para el preprocesamiento son cuatro:

- Limpieza de los datos.
- Transformación de los datos.
- Reducción de Datos
- Hacer los datos de tipo discreto

En cuanto a la limpieza de datos existen investigaciones que sugieren cómo hacer este proceso. En este punto es importante definir los umbrales de los datos que se permitirán y aquellos que no serán objeto de estudio. Encontrar nuevas reglas de asociación tales que el umbral definido permita encontrar reglas realmente útiles y no aquellas en donde las variables no tienen una relación real (Vashishtha, Kumar, & Ratnoo, 2011). La definición de estos umbrales es muy importante, por ejemplo, para encontrar buenas predicciones de diagnósticos médicos, en este sentido existen trabajos relacionados que documentan métodos y técnicas para llevar a cabo este paso. Algunas sugerencias generales son:

[1] La Organización Mundial de la Salud (OMS) define a los determinantes como los factores sociales, políticos, económicos, ambientales y culturales que condicionan el proceso vital, y específicamente el proceso salud-enfermedad de un individuo (Documento Compes Social 147 - Consejo Nacional de Política Económica y Social República de Colombia Departamento Nacional de Planeación)

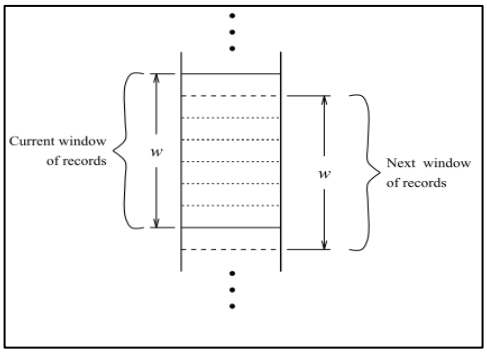
- Generar umbrales flexibles que permitan encontrar alguna relación inusual, dado que las nuevas reglas a menudo apuntan a casos poco frecuentes, es probable que esto se dé cuando el umbral definido esté fuera de las condiciones comunes.
- Los umbrales definidos deben permitir que los algoritmos de minería de datos funcionen correctamente con los datos permitidos.
- Muchas veces generar umbrales demasiado flexibles repercutirá en la generación de reglas sin interés.

Algunos métodos para el preprocesamiento

OAG (Generalización orientada a objetos), es un método propuesto que se basa en el recorrido de los datos haciendo una estructura de árbol, donde cada nivel jerárquico del árbol tiene una generalización. Todos los nodos de un mismo nivel corresponden de manera jerárquica a un grupo determinado por su condición, el ejemplo ciudad – localidad-Territorio- puede ser uno de ellos, entonces los datos se clasifican según este orden jerárquico. También pueden hacerse clasificaciones de la combinación de otros árboles existentes como Nivel educativo alto – medio - bajo. Esta es una manera de transformar los atributos del conjunto de datos (Hilderman, Hamilton, & Cercone, 1999).

La figura 3 ilustra otra técnica propuesta en 1998, la cual consiste en eliminar los datos irregulares en un subconjunto determinado. Los registros se almacenan en forma de pila y se va comparando el valor que ingresa con la ventana de comparación mientras se hace el recorrido de los datos, la pila tiene el mismo número de registros durante todo el recorrido (Hernández & Stolfo, 1998).

Figura 3. Método de limpieza con recorrido de pila (Hernández & Stolfo, 1998)



Fuente: Hernández, M. A., Stolfo, S. J. (1998)

- **ML_T2LA**

El método ML_T2LA es un algoritmo de tipo *a priori* mejorado haciéndolo multinivel, dando mayor relevancia a las reglas que tienen una asociación más fuerte. Este algoritmo sirve

principalmente para encontrar la frecuencia de patrones en los diferentes niveles de la estructura de datos (Han & Fu, 1999).

1.2.4 Reducción de datos

La reducción de los datos es el resultado de los pasos realizados en el preprocesamiento. Para esto se debe tener en cuenta las características principales de los datos para que cumplan con el objetivo de búsqueda durante la aplicación de las técnicas de minería de datos (Verma, 2015).

1.2.5 Formulación de los objetivos.

Este paso consiste en determinar si el objetivo que se desea con KDD y en las futuras aplicaciones de minería de datos será de clasificación, regresión, clustering o de asociación (Fayyad & Uthurusamy, 1996), (Verma, 2015).

1.2.6 Exploración de análisis, modelo y selección de la hipótesis.

Esta parte debe realizarse de manera crítica, aquí se escogen los métodos y algoritmos que se utilizarán más adelante en la minería de datos (Fayyad & Uthurusamy, 1996).

Los métodos y algoritmos seleccionados generalmente suelen ser de tipo predictivo y se tiene un conocimiento previo del posible resultado (Tesauro & Kephart, 2000). También existen trabajos que han planteado modelos híbridos de algoritmos por ejemplo el uso de ANN (Redes Neuronales Artificiales) y Lógica Difusa apodado como Neurofuzzy (Tsoukalas, 1998) o técnicas de algoritmos genéticos y lógica difusa (Herrera, 2008), (Carse & A., 1996). También puede escogerse Otros Algoritmos recientes como el de Optimización de colonia de hormigas (Dorigo, 1991) entre otras técnicas de inteligencia artificial - IA (Yang, 2001).

1.2.7 Minería de Datos

Minería de datos en KDD es quizá la etapa más importante del proceso, en este paso se hace la búsqueda de patrones relacionales con técnicas como agrupación (*clustering*), regresión, clasificación, entre otras (Verma, 2015)

Objetivos de la minería de datos

La minería de datos puede tener como fin diferentes objetivos entre los que se encuentra la confirmación de Hipótesis o el descubrimiento de conocimiento. Tambien se dice que los resultados pueden ser de tipo predictivo o descriptivo (Vashishtha, Kumar, & Ratnoo, 2011).

La minería de datos se basa en formalismos matemáticos que involucran la lógica y la estadística, esta segunda más utilizada en el proceso de KDD que la primera (Fayyad & Stolorz, 1997).

Consultas

En las bases de datos relacionales es muy importante escoger de manera correcta las sentencias. Existen consultas que implican el uso de algoritmos genéticos, los cuales permiten buscar la mejor solución en la búsqueda de patrones o un conjunto de posibles soluciones: esto se puede hacer a través del lenguaje DMQL (*Data Mining Query Language*). Se requiere que estas consultas satisfagan cualquiera de las siguientes características:

1. Las consultas deben buscar tuplas que satisfacen una o más condiciones normalmente especificadas por la cláusula WHERE.
2. Se construye una consulta de agregado que consiste en la extracción de la información estadística que no existe tal como es, sino que tienen que ser deducida de los atributos existentes
3. Deben satisfacer el descubrimiento del conocimiento y los patrones en un sistema de información (Srinivasa, Jagadish, Venugopal, & Patnaik, 2007).

Algunos patrones pueden surgir de relaciones que no son evidentes, pero que puede ser consecuencia natural de lo que sucede en el mundo real (Vashishtha, Kumar, & Ratnoo, 2011).

Se deben definir medidas para los patrones encontrados teniendo en cuenta los siguientes aspectos:

- Eliminar patrones de bajo interés durante el proceso.
- Otorgar un rango de acuerdo con los objetivos.

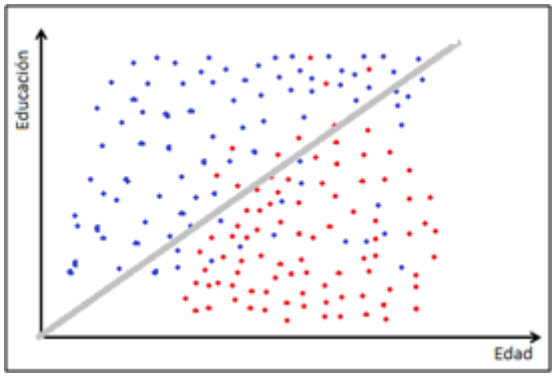
Estas medidas se pueden utilizar para precisar las reglas de interés (Vashishtha, Kumar, & Ratnoo, 2011).

De aquí en adelante se hace una descripción general de las técnicas comúnmente usadas en minería de datos.

Modelos de regresión

Este modelo está basado en la predicción de la existencia de un dato dependiendo de variables asociadas a este dato. Se trata de un método basado en la estadística y en la creación de una fórmula de regresión que separa las zonas donde se encuentran los datos según las variables, para después utilizar la fórmula como predictor de la zona ubicación de un registro determinado dentro del espacio vectorial (Fayyad & Stolorz, 1997).

Figura 4. Regresión lineal de primer embarazo en mujeres según edad y educación. (Ejemplo ilustrativo) l

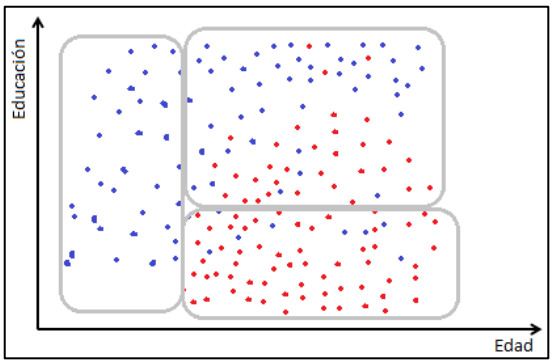


Estos métodos consisten en modelar matemáticamente las relaciones entre variables. Entre las técnicas conocidas se tienen la regresión logística, las redes neuronales y las máquinas de soporte vectorial.

Modelos basados en *Clustering*, Segmentación o Agrupación

Consiste en organizar por grupos los registros que se encuentren cercanos a un punto, o que tienen particularidades comunes en los valores de sus variables en una zona dentro de un espacio vectorial. Se determinan umbrales para la función de admisión al grupo, para esto se debe definir una medida que determina la pertenencia a cada grupo, la cual en muchas ocasiones está definida por una métrica.

Figura 5. Segmentación de primer embarazo en mujeres según edad y educación. (Ejemplo ilustrativo)



Cuando se especifica el número de grupos, estos métodos pueden ser divididos en tres clases: métodos basados en métricas, métodos basados en modelos y métodos basados en particiones (Verma, 2015).

Uno de los algoritmos más conocidos y utilizados para la aplicación de esta técnica es *k-means*, el cual fue propuesto por primera vez en 1955. A pesar de su antigüedad este

algoritmo tiene un gran rendimiento, cuando se requiere realizar agrupamiento como técnica de minería de datos (Jain, 2010).

Modelos de asociación (Summarization)

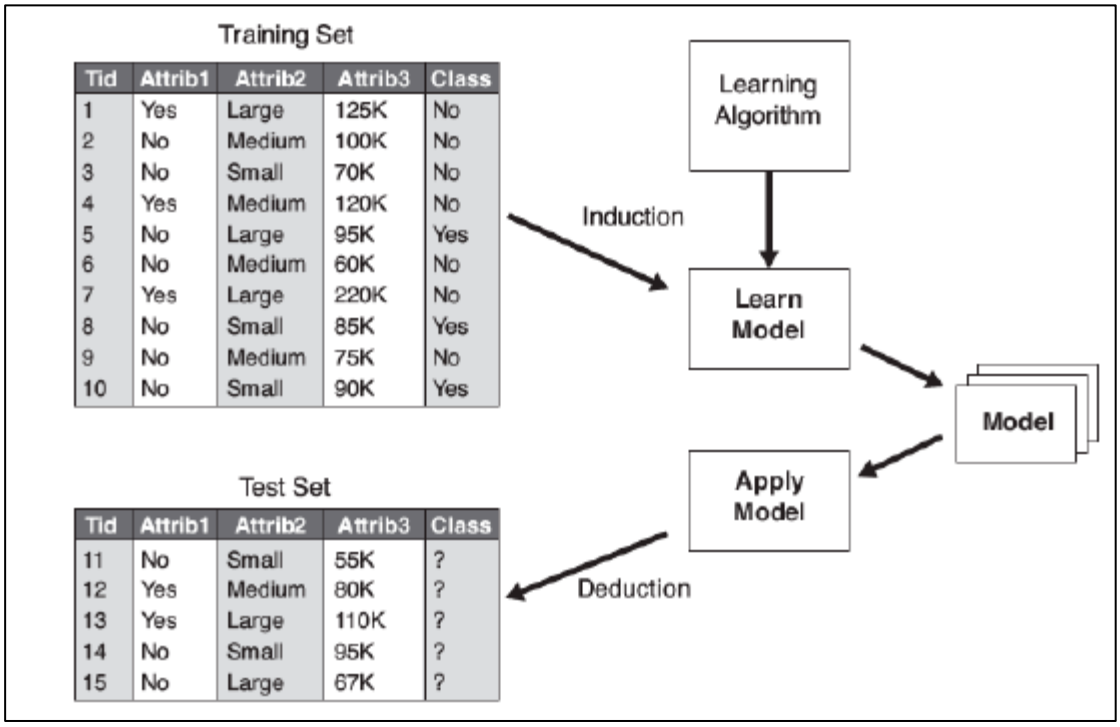
Estos métodos consisten en extraer patrones compactos que describen subconjuntos de los datos (Fayyad & Uthurusamy, 1996). Un método muy conocido para este tipo de técnica es la obtención de reglas de asociación, estableciendo relaciones entre las variables (Fayyad & Stolorz, 1997). Algunas propuestas que involucran estructuras de datos formales plantean el uso de métodos derivados que utilizan asociaciones de reglas multinivel en forma de búsqueda de árbol (Han & Fu, 1999).

En la vida práctica unas variables están relacionadas con otras, esta relación puede ser descubierta cuando se hace minería con estos métodos. Esto podría representar un problema cuando estos descubrimientos se ven traducidos en violaciones a la privacidad de los individuos al realizar búsquedas en bases de datos donde se encuentra almacenada información poblacional (Wang, Maskey, Jafari, & Hong, 2008).

Modelos basados en Clasificación

Estos métodos buscan asociar los datos basados en relaciones de vecindad (Verma, 2015).Una técnica de clasificación es una aproximación sistemática para la construcción de modelos predictivos a partir de un conjunto de datos de entrada. Ejemplos de estas técnicas incluyen clasificaciones como árboles de decisión, clasificadores basados en reglas de asociación, redes neuronales, máquinas de soporte vectorial y clasificadores basados en redes bayesianas. Cada técnica utiliza un algoritmo de aprendizaje para identificar relaciones entre los atributos de entrada y la etiqueta clase que se utiliza para clasificar. Un buen modelo predictivo debería ser capaz de clasificar un registro a partir de sus atributos, incluso con algunos valores diferentes a los ingresados durante su entrenamiento (Tan, Steinbach, & Kumar). La figura 6 ilustra el esquema general para la solución de problemas de clasificación. Se dispone de un conjunto de datos de entrada para el entrenamiento del modelo donde los valores para la variable o etiqueta clase son conocidos. Este conjunto de datos de entrenamiento es utilizado para crear el modelo de clasificación, sometiendo los datos a la búsqueda de patrones relacionales entre los valores que toman sus atributos en cada registro con respecto a la etiqueta clase; esta búsqueda de reglas y patrones de asociación son obtenidas mediante un algoritmo de aprendizaje maquinal. Después se somete al modelo para que clasifique registros de un conjunto de datos que no tienen valores asignados a las clases.

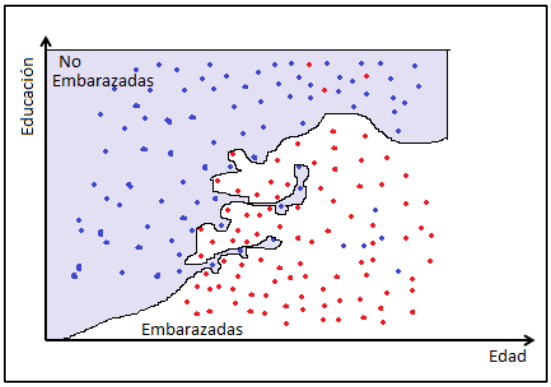
Figura 6. Modelo general para la construcción de un modelo de clasificación.



Fuente: Pal, Mandana, & Pal. (2012)

Alguno de los problemas del mundo real en la administración de salud pública a nivel mundial, es que los datos no se encuentran almacenados en bases de datos estructuradas. La mayoría de la información que se almacena en los sistemas de información de las entidades al cuidado de la salud, es primero puesta en papeles a manuscritos por los profesionales de la salud (Goodwin, Vandyne, Lin, & Talbert, 2003).

Figura 7. Clasificación de primer embarazo en mujeres según edad y educación. (Ejemplo ilustrativo)



1.2.8 Interpretación de los patrones encontrados

La interpretación de los patrones es una etapa muy importante en el proceso de KDD ya que define el estado final de todo el proceso. El inconveniente con la interpretación es que puede presentar rasgos subjetivos según las expectativas del usuario final. Para que la interpretación pueda ser objetiva debe plantearse desde una visión global y partiendo desde el conocimiento previo. En este sentido la definición de los umbrales ha jugado un papel muy importante en todo el proceso (Vashishtha, Kumar, & Ratnoo, 2011).

Otro problema frecuente es la dificultad en la justificación de la validez de las nuevas reglas de asociación. La aceptación de estas reglas por parte del usuario es de vital importancia. En este aspecto se debe tener en cuenta los límites permitidos para determinar si una nueva regla de asociación es apropiada, existen planteamientos para probar la pertinencia de los datos aceptados dentro de los umbrales los cuales están sujetos a la aceptación por parte del usuario (Then, 2011).

1.2.9 Evaluación de los modelos predictivos elaborados con clasificadores

Métricas de rendimiento

Para la evaluación de los modelos de clasificación existen varias métricas que indican que tan eficiente es el clasificador desarrollado. Estas métricas se basan en el conteo de registros clasificados correctamente e incorrectamente según la etiqueta clase. Estos conteos son tabulados en una tabla conocida como Matriz de Confusión. La tabla 1 representa una matriz de confusión para un problema de clasificación binaria. Cada dato f_{ij} en la tabla representa la cantidad de registros de la clase *i* cuya predicción del modelo fue la clase *j*. Por ejemplo, la cantidad de registros mal clasificados como 1 que en realidad eran 0 se ingresan en la cuadrícula f_{01}. Por lo tanto, la cantidad de registros correctamente clasificados corresponden a la suma de f_{11} y f_{00}, mientras que los datos clasificados incorrectamente son ($f_{01} + f_{10}$).

La tabla 2 representa una matriz de confusión para casos típicos de clasificación SI/NO, donde TP (True Positive) representa la cantidad de registros cuya clase es SI y fueron bien clasificados, TN (True Negative) representa la cantidad de registros cuyo valor para la etiqueta clase es NO y fueron bien clasificados, FP (False Positive) corresponde a la cantidad de registros clasificados incorrectamente como SI mientras que FN (False Negativo) son aquellos registros que fueron mal clasificados por que su valor real era SI y el modelo los predijo como NO.

Las matrices de confusión brindan información para comparar modelos y determinar cuál realiza mejores clasificaciones y cual es más conveniente para implementar comparando el rendimiento de cada modelo según el objetivo por el cual fue construido.

Tabla 1. Matriz de confusión para un problema de clasificación binaria

		Predicción	
		Clase = 1	Clase = 0
Clase	Clase = 1	f_{11}	f_{10}
	Clase = 0	f_{01}	f_{00}

Tabla 2. Matriz de confusión para un problema de clasificación SI/NO

		Predicción	
		SI	NO
Clase	SI	TP	FN
	NO	FP	TN

La evaluación de los modelo puede realizarse mediante la aplicación de métricas de rendimiento. Las métricas mayormente utilizadas para la evaluación de este tipo de modelos cuando se requiere la detección de los casos positivos se definen a continuación:

$$Exactidud = \frac{Numero\ de\ Predicciones\ Correctas}{Total\ de\ registros} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad (1)$$

$$Tasa\ de\ Error = \frac{Numero\ de\ Predicciones\ Incorrectas}{Total\ de\ registros} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} \quad (2)$$

$$FP\ rate\ (tasa\ de\ Falsos\ Positivos) = \frac{FP}{N} = \frac{FP}{(FP + VN)} \quad (3)$$

Donde, N es el número real de casos negativos.

$$Sensibilidad = TP\ rate\ (tasa\ de\ Verdaderos\ Positivos) = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (4)$$

Donde, P es el número real de casos positivos.

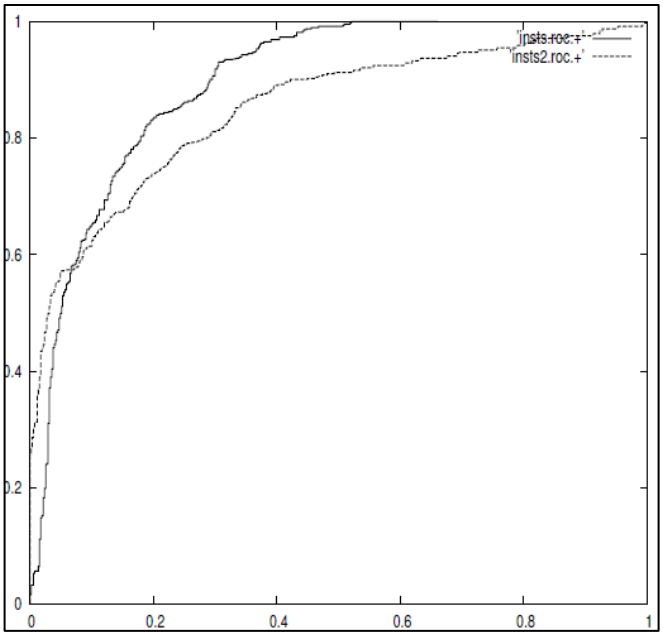
$$Precisión\ (Valor\ Predictivo\ Positivo) = \frac{TP}{TP + FP} \quad (5)$$

$$F\ score = Precisión * Sensibilidad \quad (6)$$

Curvas ROC

Una curva ROC es la representación del desempeño de un modelo de clasificación basado en dos métricas: la tasa de verdaderos positivos en el eje Y y la tasa de falsos positivos en el eje X. Cuando un conjunto de datos es sometido a algún clasificador discreto tal como los árboles de decisión, se genera una matriz de confusión cuyos valores sirven para calcular las dichas métricas. Esto permite graficar un punto dentro del espacio ROC que evalúa el rendimiento general del predictor. Para otro tipo de clasificadores como las redes neuronales la clasificación no es discreta, es decir que cada registro trae implícito un valor que define el grado de pertenencia respecto a los valores de la etiqueta clase y la clasificación de cada registro se realiza con base en un umbral definido. Para este tipo de clasificadores en particular se pueden dibujar todos los puntos que generan los grados de pertenencia de cada registro sobre el espacio ROC. Esta grafica genera una curva que permite calcular métricas de desempeño adicionales a las evaluadas con la matriz de confusión. La métrica más común con esta gráfica es el área bajo la curva (*Area Under Curve*, AUC) que siempre toma valores entre 0.5 y 1, siendo 1 el valor para un clasificador ideal (Tom Fawcett, 2004).

Figura 8. Curvas ROC para dos clasificadores no discretos



Fuente: (Tom Fawcett, 2004)

1.3 Aproximaciones en el campo del cuidado de la salud

Esta sección describe algunas de las técnicas usadas en el campo del cuidado de la salud. Se hace referencia de algunas investigaciones recientes que han recurrido a las técnicas de KDD y aprendizaje de máquina para lograr algún avance en esta área. Sin embargo, se debe

tener en consideración que existe una diferencia entre estos trabajos y el propósito general de esta tesis. Estos dos difieren en el modo de captación de la información y del subconjunto al que pertenecen dentro del área del cuidado de la salud. Aquí se muestran trabajos de investigaciones relacionados a pronósticos médicos, historias clínicas, toma de exámenes y detección de enfermedades. Mientras que la investigación realizada en este caso consiste en el subconjunto de la prevención de enfermedades y la atención primaria en salud como política internacional orientado por la Organización Mundial de la Salud (OMS), del mismo modo los repositorios de datos tienen características diferentes.

Los distintos trabajos realizados involucran diferentes campos investigativos como minería de datos, modelamiento de agentes inteligentes, construcciones de algoritmos con lógica difusa, redes neuronales, así como el aprendizaje de máquina, entre otras que podrían enmarcarse en alguna etapa de KDD. De hecho los algoritmos de aprendizaje maquinal fueron diseñados para el análisis y clasificación de conjuntos de datos médicos, estas técnicas fueron implementadas con el objetivo de predecir resultados clínicos o encontrar patologías asociadas a alguna enfermedad (Kononenko, 2001).

Las investigaciones en este campo suelen hacer uso de las diferentes técnicas aplicadas en las etapas de pre-procesamiento, minería de datos y análisis de la información de KDD (Cismondi, Fialho, & Vieira, 2011), con el fin de obtener información predictiva de los resultados de un examen específico. Algunos trabajos de investigación[2] hacen uso por ejemplo de redes neuronales y lógica difusa para determinar patrones de comportamiento de pacientes y sus reacciones fisiológicas. Las técnicas aplicadas buscan encontrar los patrones de relación entre estas variables y la patología sufrida por el paciente. Para hacer este estudio los autores tomaron un conjunto de datos de pacientes y los resultados de sus exámenes de laboratorio clínico. Los datos del paciente fueron seleccionados junto a las variables fisiológicas asociadas a la unidad de cuidados intensivos y se buscaron relaciones estadísticos entre los datos. Al momento de definir el modelo es importante una correcta selección de los métodos apropiados según la problemática que se quiera abordar, dependiendo del subconjunto de datos. Una buena aproximación para la selección de algún método apropiado para estudios referentes al cuidado de la salud se presenta en un trabajo reciente que compara 5 tipos diferentes de datos clínicos y que recomienda la utilización de los métodos dependiendo las características de los datos (ver Srimani & Koti, 2011).

Uno de los grandes desafíos al realizar minería de datos en sectores al cuidado de la salud es que, en muchos casos, la información no se encuentra almacenada de manera estructurada. Gran parte de esta información ha venido siendo ingresada a los sistemas que se han implementado en centros médicos o en sus departamentos (Kononenko, 2001). Aunque en proyectos de ingeniería generalmente los datos se encuentran almacenados en bases de datos estructuradas y robustas, no necesariamente tiene que ser así para que sea posible descubrir conocimiento en bases de datos. Uno de los problemas más comunes en las organizaciones que se encargan del cuidado de la salud es precisamente que sus registros se encuentran almacenados en herramientas que no son robustas. KDD también

[2] Cismondi, F. 2011 Predicting laboratory testing in intensive care using fuzzy and neural modeling

puede ser aplicada en datos planos o en formatos no estructurales como hojas de Excel; algunas investigaciones[2] tratan este tipo de almacenamiento y aplican KDD sin distinción,' permitiendo demostrar que esta técnica es independiente de la tecnología de almacenamiento. Para estos casos es recomendable utilizar una herramienta que se encuentre disponible en el mercado como SPSS MODELER, WEKA, KNIME o cualquier otra herramienta analítica que se especialice en la generación de modelos predictivos y minería de datos. Estas herramientas pueden ayudar a realizar el proceso de minería de datos cuando los datos no están estructurados. Algunos estudios[3], por ejemplo, han hecho uso de estas herramientas en diferentes campos de la medicina, enfáticamente en la obtención de patrones relacionales de los exámenes médicos y exámenes de sangre (Minnie & Srinivasan, 2011).

Algunos investigadores han realizado esfuerzos para mejorar la interacción entre el computador y las personas, estos trabajos se han venido realizando durante mucho tiempo de manera independiente para desarrollar métodos que pueden ayudar a los usuarios finales a identificar, extraer y entender la información oculta entre los datos, es evidente que se necesita un enfoque interdisciplinario entre el área de la salud y las técnicas computacionales para generar sinergias en estos dos campos. Lo cual podría generar sinergias en la aplicación de estos métodos para conjuntos de datos médicos complejos y débilmente estructurados (Holzinger, 2012).

Para la detección de riesgos patológicos de pacientes se estudian sus comportamientos, historia clínica y resultados previos. Recientemente para la detección de enfermedades de la arteria coronaria se han realizado trabajos que clasifican imágenes de electrocardiogramas con base en registros lingüísticos descritos por los médicos. Para esta clasificación se utilizan técnicas de lógica difusa para determinar el grado de pertenencia entre los diferentes subconjuntos y de este modo predeterminar los riesgos de los pacientes y los cuidados que deben tener sobre esta enfermedad (Pal, Mandana, & Pal, 2012).

La obtención de relaciones entre las variables de las bases de datos suele ser una fuente valiosa para la obtención de conocimiento para un experto médico, trabajos relacionados con un conjunto de datos de la hepatitis crónica es un buen ejemplo del estudio de los usos de KDD (Li, 2012).

Otro ejemplo se presenta en los estudios de datos cardiovasculares. En post-procesamiento, las reglas pueden ser descubiertas mediante la aplicación de filtros basados en la semántica basada en ontología UMLS (Unified Medical Language System) (Sebastian & Then, 2011).

Algunas aproximaciones se realizan utilizando minería de texto debido a que en la mayoría de los repositorios de datos clínicos los registros son ingresados de manera textual por el

[3] Minnie, D. 2011 - Application of Knowledge Discovery in Database to blood cell counter data to improve quality control in clinical pathology

médico. En estos casos la captura de datos y la asociación con patologías presentan una mayor dificultad (Friedlin, Mahoui, Jones, & Jamieson, 2011).

1.4 Tendencias

En los últimos años los datos puestos en la Web han crecido de manera exponencial, la mayoría de estos datos carece de una estructura específica (Yang, Slattery & Ghani, 2002). La mayoría de investigadores coinciden en que las tendencias de investigación de KDD, como técnica y como parte fundamental de minería de datos, están orientadas a satisfacer la necesidad de encontrar información en la gran variedad de texto no estructurado (Ball & Brunner, 2010). La minería de textos y la construcción de consultas que satisfagan la extracción de nuevo conocimiento tienen gran interés en investigación actualmente. Otro campo importante de investigación será el que permita la construcción de modelos de sistemas para la toma de decisiones importantes (Skulimowski, 2011).

2 Enfoque de la Investigación

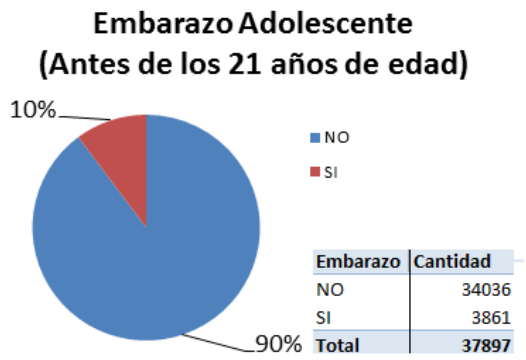
El incremento exponencial que ha sufrido durante los últimos años la recolección de datos en diferentes campos del conocimiento conlleva a la necesidad de crear técnicas computacionales que ayuden al ser humano a extraer información útil para generar nuevo conocimiento de manera ágil (Verma, 2015). Un ejemplo típico de este fenómeno se presenta en el campo de la salud. Es muy común para los especialistas del cuidado de la salud, analizar periódicamente las tendencias de los datos en un sistema de información y el efecto colateral que estos puedan tener.

La minería de datos, el aprendizaje de máquina, el análisis y la estadística, son temas de alto interés de investigación actual, pues permiten encontrar información en fuentes estructuradas como no estructuradas. Durante el desarrollo de los siguientes capítulos se expondrá el proceso KDD aplicado a un conjunto de datos poblacional tomado de la Secretaría Distrital de Salud con el objetivo de encontrar patrones comunes en mujeres con gestación a temprana edad.

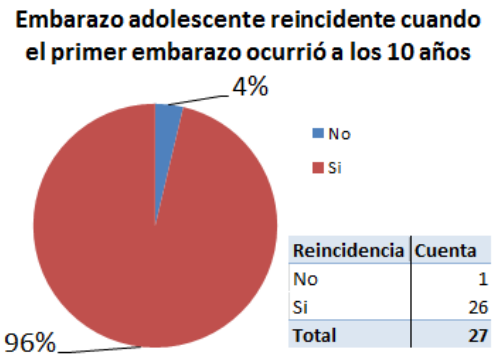
2.1 Planteamiento del problema

En el boletín estadístico entregado por la Secretaría Distrital de Salud (SDS) para los periodos de 2004 a 2007, se caracterizó en el Sistema de Información de Atención Primaria en Salud (APS) a 154 mujeres embarazadas entre los 10 a 14 años de edad y 1.773 entre los 15 a 17 (Secretaría Distrital de Salud, 2007). La cifra de mujeres en estado de gestación ha disminuido en la última década gracias a programas de prevención y educación sexual en la población. Sin embargo, las cifras siguen siendo altas en comparación con otras ciudades del mundo. Aunque el Sistema de Información del programa “Salud a Su Casa” de la SDS capta otras variables de entorno social de la población, no se cuenta con un sistema que indique si existe una relación entre estas variables y la vulnerabilidad de las mujeres a quedar en estado de embarazo a temprana edad.

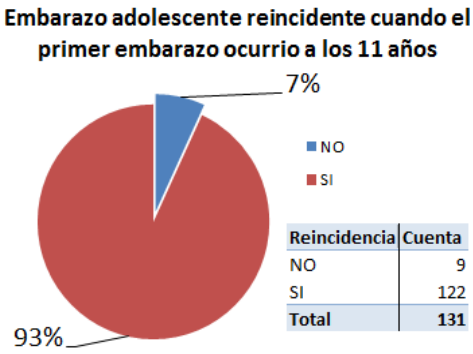
Figura 9. Estadísticas de embarazo adolescente reincidente



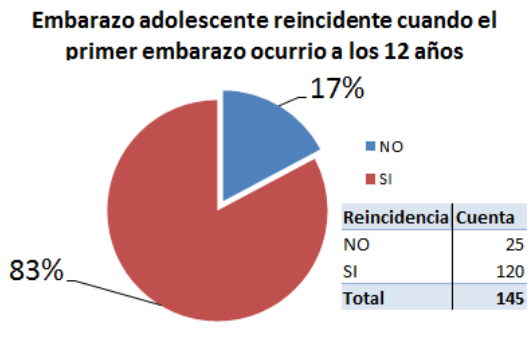
Embarazo	Cantidad
NO	34036
SI	3861
Total	37897



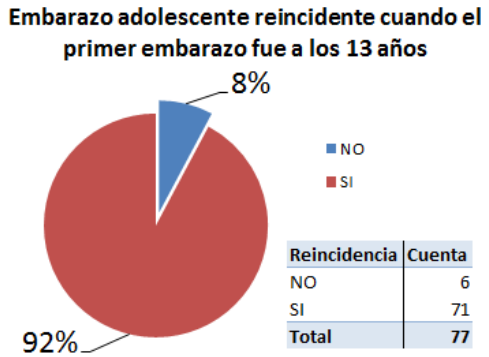
Reincidencia	Cuenta
No	1
Si	26
Total	27



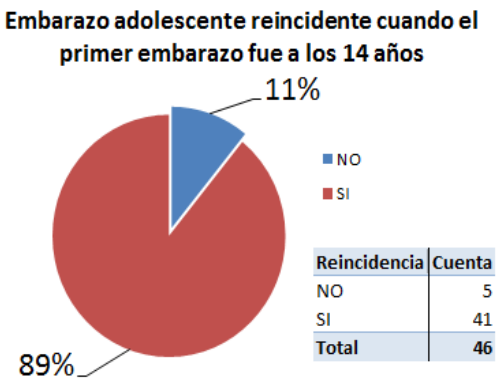
Reincidencia	Cuenta
NO	9
SI	122
Total	131



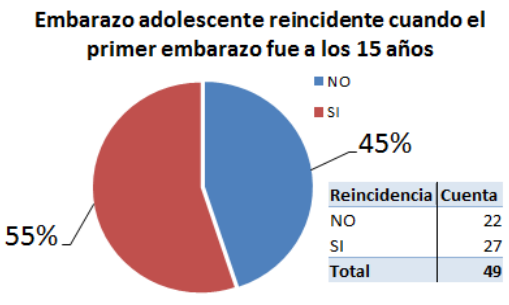
Reincidencia	Cuenta
NO	25
SI	120
Total	145



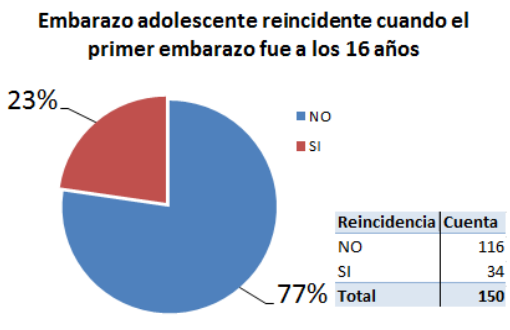
Reincidencia	Cuenta
NO	6
SI	71
Total	77



Reincidencia	Cuenta
NO	5
SI	41
Total	46



Reincidencia	Cuenta
NO	22
SI	27
Total	49



Reincidencia	Cuenta
NO	116
SI	34
Total	150

Con un análisis estadístico básico del conjunto de datos utilizado durante el segundo enfoque descrito en el capítulo 4, se pudo determinar que la mayoría las mujeres que presentaron embarazos a muy temprana edad (Entre los 10 y 15 años), tuvieron al menos otro embarazos antes de cumplir los 21 años. En la figura 9 se hace una ilustración de la situación de este fenómeno. Como se puede observar una de cada diez mujeres atendidas por el programa "Salud a su Casa" tuvo un embarazo antes de cumplir los 21 años. Por otra parte, durante los 12 años del programa se han reportado 27 casos de embarazos de niñas de tan solo 10 años, pero lo más alarmante de esta situación es que tan solo una de ellas no reincidió en otro embarazo antes de los 21 años.

2.2 Hipótesis

Las tendencias de salud de una población están asociadas a diferentes variables del entorno relacionado a los individuos. La OMS (Organización Mundial de la Salud), plantea que la salud de los individuos se ve afectada por los determinantes sociales de la salud, estos determinantes son caracterizados en diferentes enfoques que son propios de cada individuo y que se relacionan con sus creencias, su estilo de vida, su entorno socio económico, entre otros.

2.3 Metodología

Se cuenta con una base de datos del programa de Atención Primaria en Salud también conocido como "Salud a Su Casa", con información recolectada desde el 2004 y enfocada en la población de estrato socioeconómico 1, 2 y 3. Con estas variables se trabajó para generar un modelo capaz de buscar patrones en los datos asociados a los determinantes de salud de los individuos que están relacionados con embarazos adolescentes.

El análisis se realizó desde las perspectivas de determinantes sociales, personales y familiares que están relacionados con el riesgo de niñas y adolescentes de tener un embarazo a temprana edad, estos determinantes ya han sido definidos por la OMS.

2.4 Enfoque propuesto

La base de datos con la que se cuenta, tiene información relacionada con la población, en la que se puede encontrar información acerca de la salud de las personas, de su entorno social y familiar. El análisis se realizó basado en la teoría de los Determinantes Sociales amparada por las políticas de la OMS, la cual refiere al conjunto de variables individuales, sociales y estructurales que relacionadas entre sí, condicionan el proceso vital y explican el conjunto de fenómenos asociados a la salud sexual y reproductiva de adolescentes y

jóvenes. Estos determinantes se pueden clasificar en tres conjuntos basado en la proximidad y relación con el individuo:

Determinantes proximales

Están relacionados con las características individuales en los que se destacan factores biológicos como el desarrollo puberal y factores del comportamiento como el inicio de relaciones sexuales, la nupcialidad o las uniones tempranas, el uso de los métodos de anticoncepción, y el acceso a servicios de salud y educación.

Determinantes intermedios

Se refieren al ambiente escolar, de familia y hogar en el que se encuentre la adolescente. Se consideran las condiciones familiares, la existencia de abuso o violencia, la supervisión y el diálogo entre padres o cuidadores, las normas de funcionamiento del hogar, el cuidado por mantener a la niña o adolescente en el sistema escolar.

Determinantes Distales

Están relacionados con los ingresos, pobreza, la cobertura y acceso oportuno a servicios públicos, la oportunidad de participar en las decisiones públicas y el ejercicio de la democracia, los legados culturales que trascienden las instituciones, las comunidades, el macro entorno social, las normas sociales sobre la sexualidad, la feminidad, la masculinidad, las relaciones de pareja o la participación de los adolescentes en los procesos de decisión e identidad social (DNP-DSS.SS, Ministerio de Protección Social).

2.5 Objetivos

El enfoque propuesto para la búsqueda de patrones y la generación de los modelos predictivos consistió en relacionar las variables existentes en el repositorio de datos con estos determinantes definidos por la OMS como factores del embarazo a temprana edad. Con el conjunto de variables seleccionadas se realizó el proceso de KDD en busca de dos objetivos específicos:

- Hallar patrones relacionales entre las variables con respecto al embarazo adolescente que permitieran un mejor entendimiento de esta problemática en las comunidades atendidas por el programa.

- Generar un modelo predictivo que fuera capaz de identificar menores con alto riesgo según los patrones encontrados. Evaluar los modelos construidos y utilizar el mejor para el desarrollo de una herramienta tecnológica que pudiera ser utilizada por el equipo médico del programa para realizar campañas de prevención enfocadas a las menores con mayor riesgo.

3 Primera aproximación - búsqueda de patrones y generación del modelo predictivo.

Durante los primeros acercamientos al desarrollo del modelo predictivo, se trabajó con un conjunto de datos que pertenecen a registros de mujeres entre los 10 y los 19 años de edad y contenía un total de 209.978 registros y 93 variables que describen condiciones sociales, económicas, culturales y de entornos familiar basados en la primera encuesta que realizan los profesionales de la salud a los ciudadanos que son captados por primera vez. Tomando este conjunto de datos se llevaron a cabo todos los procesos descritos como KDD en búsqueda de nuevo conocimiento respecto a la problemática del embarazo adolescente en Bogotá y buscando desarrollar un modelo predictivo que sirviera de fuente de insumo principal para el desarrollo de una herramienta tecnológica que pueda permitir localizar las adolescentes con mayor riesgo de tener un embarazo a temprana edad. En este capítulo se describe el trabajo realizado con este conjunto de datos, los resultados obtenidos, algunos inconvenientes presentados y las conclusiones a las que se llegaron después de aplicar varias técnicas de minería de datos.

Las variables extraídas del sistema de información “Salud a Su Casa” se relacionan en el anexo A, en donde se describen las asociaciones hechas con los determinantes de la salud de la OMS y los valores que tienen dentro del conjunto de datos.

3.1 Preprocesamiento

Como se describió en el marco conceptual, uno de los mayores inconvenientes cuando se trabaja con registros en área de la salud es precisamente la mala calidad del dato. El sistema de información del cual se extrajo el conjunto de datos no es la excepción a la regla, ya que estos son captados de manera manual por los profesionales de la salud en formularios en papel que después son digitalizados por un grupo de personas al interior de cada hospital. Además las aplicaciones de captura de la información no cuentan con muchas validaciones para impedir el ingreso de datos fuera de rango o datos incoherentes. Por este motivo el proceso de preprocesamiento de los datos fue una de las tareas que llevó mayor tiempo durante esta investigación. Tarea que se dificulto también por el alto

número de variables y el elevado número de registros con problemas de calidad en el conjunto de datos.

3.1.1 Reducción de dimensionalidad

Un primer paso consistió en realizar una clasificación empírica de las 96 variables disponibles en el sistema y asociarlas a los determinantes sociales de la OMS. Del mismo modo, se definieron aquellos determinantes que no pudieron ser tenidos en cuenta por falta de información del repositorio de datos con el cual se realizó la investigación. Uno de los principales inconvenientes en la elaboración de este trabajo precisamente fue el no disponer de información referente de algunos de estos determinantes que tienen un alto impacto con el problema abordado.

Las tablas 3, 4 y 5 muestran los determinantes sociales establecidos por la OMS como factores de alto impacto relacionados a la vulnerabilidad de las mujeres a tener un embarazo a temprana edad. En estas tablas se diferencia con color rojo aquellos determinantes de los cuales no se tuvo variables disponibles dentro del Sistema de Información y de color negro aquellas que pudieron ser asociadas al menos con una variable del repositorio de datos.

Tabla 3. Determinantes proximales asociados por la OMS al embarazo adolescente

Factor	Determinante Social
Factores biológicos	Desarrollo de pubertad
	Edad menarquía
Factores de Comportamiento	Inicio de relaciones Sexuales
	Nupcialidad y/ o uniones tempranas
	uso de los métodos de anticoncepción
	acceso a servicios de salud
	acceso a servicios de educación
	creencias del individuo

Tabla 4. Determinantes intermedios asociados por la OMS al embarazo adolescente

Factor	Determinante Social
Factores	Familia

Factor	Determinante Social
Interpersonales	Redes Sociales
	Hogares con Jefatura femenina
	Historia de embarazo adolecente en la Familia
	Violencia Intrafamiliar
	Abandono
	Falta de Monitoreo
	Comunicación con cuidadores
	Aceptación y apoyo paternal
	lugar que se ocupa en la familia
	Calidad en la educación y sexualidad
	Cohesión con pares
Factores Intrapersonales	Edad
	Nivel educativo
	Imaginarios del amor y la sexualidad
	Manejo del tiempo libre

Tabla 5. Determinantes distales asociados por la OMS al embarazo adolescente.

Factor	Determinante Social
Factores Contextuales	Normas sociales y de genero
	Valores de la sociedad
Factores Estructurales	Nivel de pobreza
	Inequidad social
	Relaciones de dominación y subordinación de género

Con esta primera asociación de las 93 variables a los determinantes de la salud se logró reducir la dimensionalidad de las variables. Al momento de revisar el contenido del conjunto de datos, se encontraron algunas variables que en su mayoría tenían registros

con valores nulos. Al realizar la preclasificación por grupos de variables pertenecientes a cada uno de los determinantes sociales se pudieron descartar dichas variables.

Algunos procesos en la depuración de datos, limpieza y transformación se describen a continuación:

Edad

Esta variable tuvo que ser recalculada con la fecha de nacimiento respecto a la fecha de la captación, debido a una alta tasa de datos perdidos o nulos.

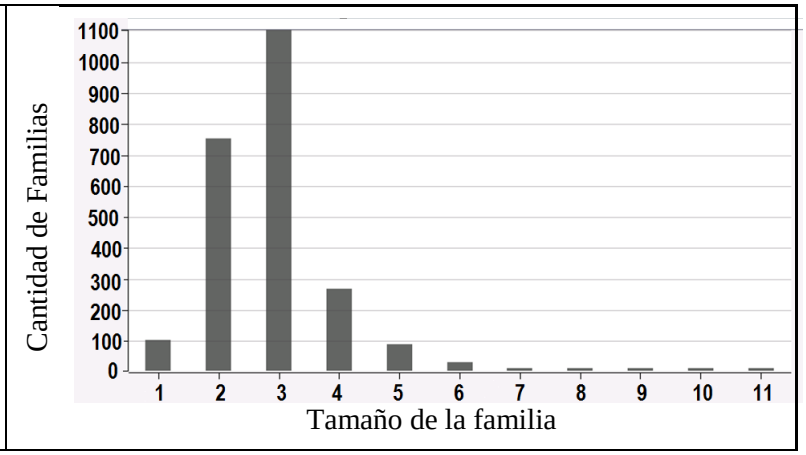
Familia

Para el determinante de familia establecido por la OMS (Zamudio & Rubiano, 1999) se dice que este está asociado con las particularidades del entorno familiar del individuo, en la revisión bibliográfica también se describe que el número de individuos que componen la familia y la posición que ocupa la persona son factores sensibles para el estudio de la vulnerabilidad de tener un embarazo a temprana edad.

En un primer análisis descriptivo se pudo notar que para la variable del tamaño de la familia existían valores fuera de rango, cuyo tamaño máximo llegó a ser de 60 siendo la media tan solo 4,28. Para este caso se eliminaron los registros donde la variable "Tamaño De La Familia" superaba dos desviaciones estándar con respecto a la media.

Figura 10. Datos estadísticos del número de personas dentro de la familia

Valores estadísticos	
Minino	1
Máximo	60
Media	4.288
Desviación estándar	1.721
Varianza	2.963



Nivel Educativo

Existen 7 variables asociadas al nivel educativo, la primera de ellas es de tipo nominal e indica si el individuo es analfabeta, para poder computar esta variable con las de tipo ordinal fue necesario realizar una binarización para los valores NO y SI.

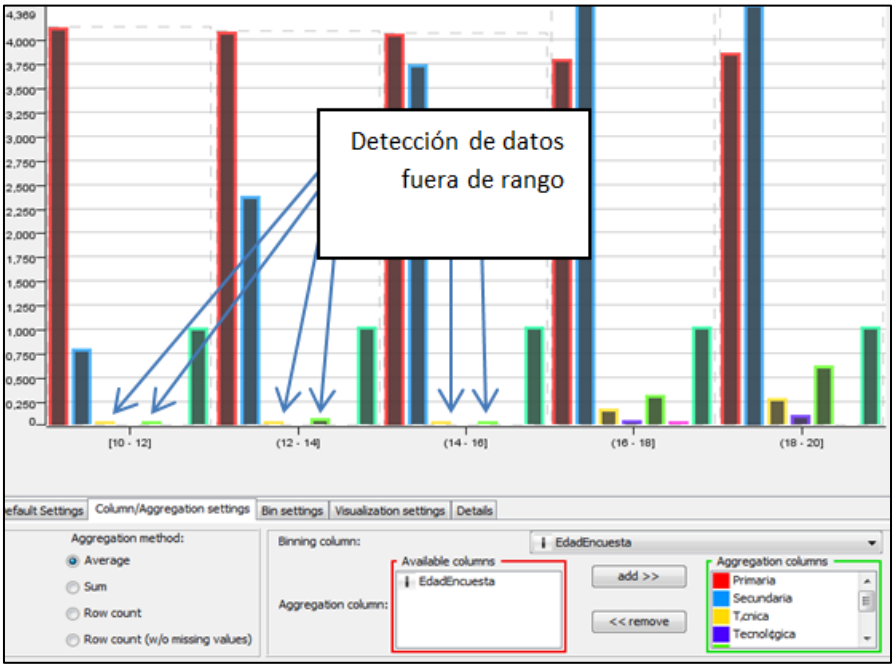
Por otra parte, las variables Primaria, Secundaria, Técnica, Tecnológica, Universitaria y Postgrado son variables ordinales que cuentan con escalas definidas.

En el histograma de los valores promedio de las escalas dadas a cada nivel educativo versus la edad de los individuos se pueden encontrar varios problemas o inconvenientes de fidelidad y confiabilidad de los datos.

Se hace evidente que existe ruido producto de la mala digitación de las personas que ingresaron la información de las encuestas. En la figura 11 se pueden observar los valores de niveles de educación superior en personas caracterizadas con edades muy cortas, ruido que fue considerado en el tratamiento de los datos.

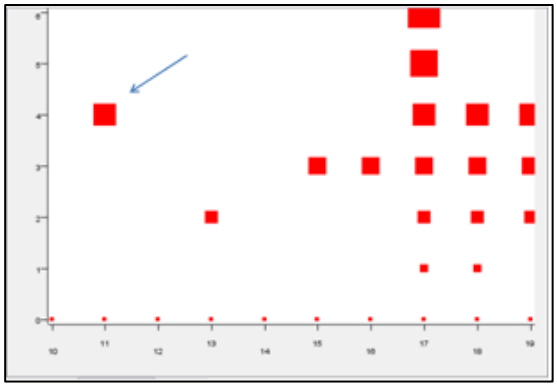
La limpieza de estos datos se pudo realizar mediante técnicas conocidas de agrupamiento y eliminación de los datos que no estén dentro de los rangos comunes. Los datos se pueden representar en un plano euclidiano de *n* dimensiones y existirán vecindades entre las diferentes configuraciones de los datos. Para el conjunto de datos que representa el nivel educativo, las diferentes dimensiones están dadas por los niveles de escolaridad de los individuos, su edad y el estado positivo o negativo de gestación.

Figura 11. Histograma Nivel Educativo.



La Figura 12 es una representación del plano Euclidiano de los datos representado en las siguientes tres dimensiones: grado de escolaridad a nivel técnico y tecnológico de los individuos (Eje Y), La edad (Eje X) y la frecuencia o cantidad de datos con esta particularidad representados por el tamaño de los cuadros.

Figura 12. Plano euclidiano de la variable Educación



Como se puede observar existen muchos registros de niñas con 11 años en niveles altos de secundaria. Estos datos están alejados de las zonas comunes con respecto a los demás registros del conjunto de datos. Para hacer limpieza en el conjunto de datos se recurrió a técnicas de similitud o de distancias, para hacer esto posible fue necesario realizar un cambio de dimensiones a los datos, categorizándolos y haciendo eliminación de registros enteros que cumplan con cierta característica. Se diseñó un modelo para realizar la limpieza que hace uso de simples reglas lógicas IF ELSE para los casos en que un nivel de escolaridad resulta ilógico a una edad directamente en las sentencias SQL que se utilizaron para extraer el conjunto de datos.

Reducción por correlaciones entre las variables

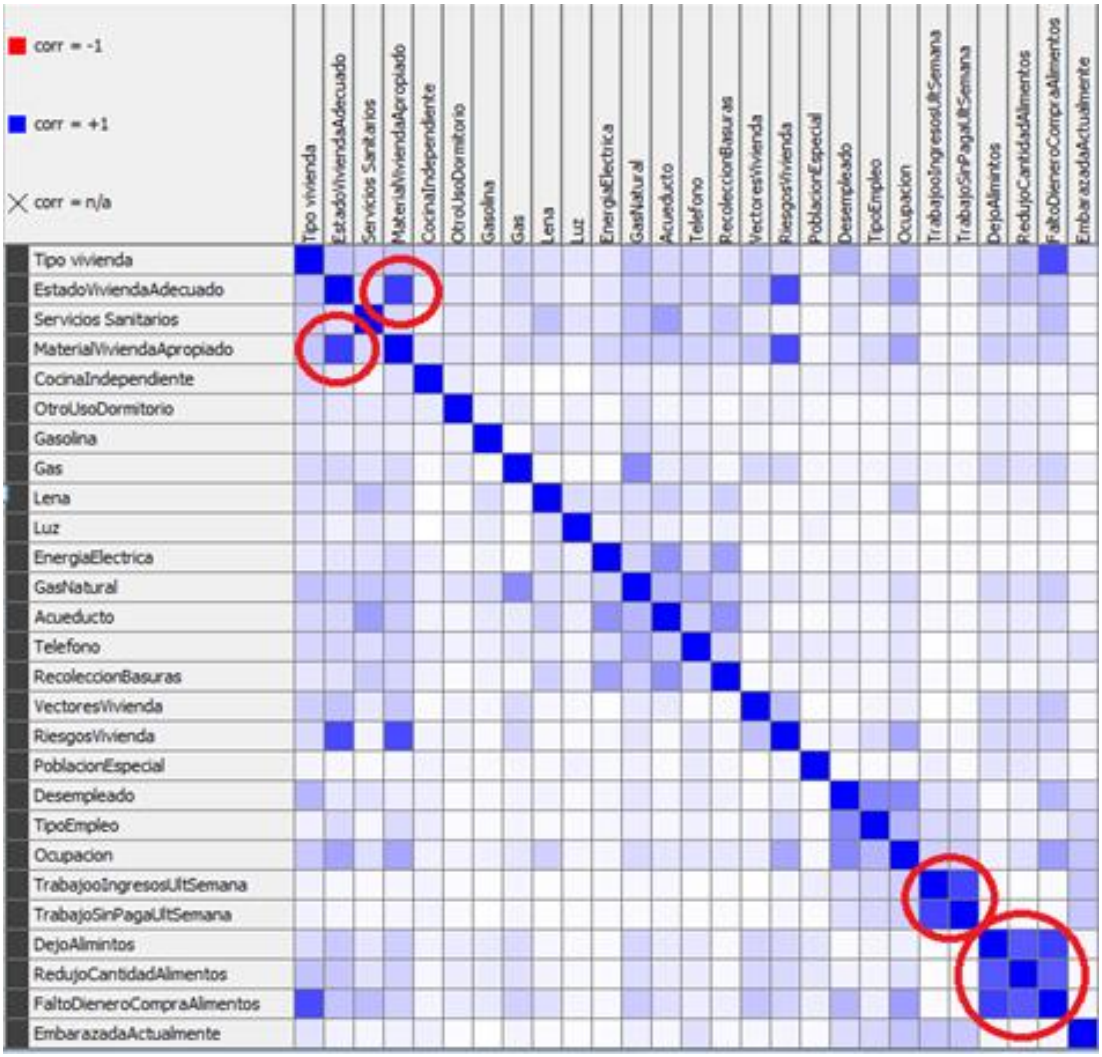
Muchas de las variables que se analizaron tienen relaciones entre ellas y el análisis de todas ocasiona reglas redundantes, entre otras cosas. Las correlaciones más fuertes se pueden observar en las gráficas 13 y 14 con colores azules, siendo los colores oscuros los de mayor correlación, las cuales se listan a continuación:

- EstadoViviendaAdecuado con MaterialViviendaApropiado.
- TrabajoSinPagaUltAsemana y TrabajoIngresoUltimaSemana.
- DejoAlimentos con RedujoCantidadAlimentos, FaltoDineroCompraAlimentos y TipoVivienda.
- Ocupación con TiempoEmpleo y Desempleado
- Gas con GasNatural

- ServiciosSanitarios con Acueducto

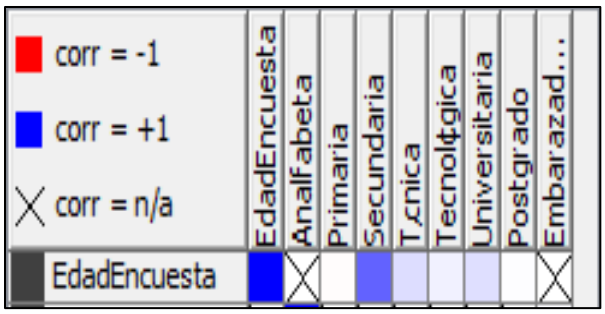
Por lo cual, se dejaron únicamente las variables: Ocupación, EstadoViviendaAdecuado, DejoAlimentos, Gas, TrabajoSinPagaUltimaSemana y ServicioSanitarios.

Figura 13. Diagrama de correlación de variables



En la figura 14 se puede observar que no existe relación entre el nivel de primaria y el aumento en la edad, esto se debe a que el rango de edades objeto de estudio está entre los 10 y los 20 años, rangos de edad en los cuales la mayoría de individuos ha cursado casi la totalidad de sus estudios básicos. Lo mismo sucede con la variable postgrado ya que este nivel educativo no alcanza a ser desarrollado en este rango de edades. Estas dos variables fueron excluidas del conjunto de datos.

Figura 14. Correlación entre variables del nivel educativo y la edad



Otras transformaciones a los datos

Durante algunas aplicaciones de técnicas de minería de datos fue necesario realizar algunas transformaciones de variables, discretizando, binarizando o simplemente realizando transformaciones de valores lógicos a valores numéricos. Esta última se utilizó en todas las técnicas aplicadas y fue aplicada a variables con dos opciones de respuesta, por ejemplo:

Género jefe hogar: JefeFemenino = 0 para jefe de hogar = Hombre y 1 para Jefe de Hogar = Mujer.

Embarazo: Se realizó cambio en el valor de la variable, de “SI” y “NO” a una variable binaria 1 y 0 respectivamente.

3.2 Minería de Datos

2.1.1 Métodos de asociación

A priori y crecimiento de patrones frecuentes (pattern growth)

Con el conjunto de datos ya preprocesados, se establecieron las pautas necesarias para aplicar los siguientes métodos de asociación aplicados al conjunto de datos: a priori y crecimiento de patrones frecuentes (Frequent Pattern growth, FP-growth), para los cuales, fue necesario binarizar los datos. Sin embargo, teniendo en cuenta que se estaba trabajando con algunos atributos ordinales, estos debieron ser primero discretizados para poder binarizarlos. La idea básica era encontrar aquellas reglas que permitan establecer bajo qué condiciones una joven queda en estado de embarazo a temprana edad, este atributo se convierte en la clase requerida para la aplicación de estos métodos.

Algunos inconvenientes presentados

Para la implementación de esta etapa del análisis fue necesario establecer dos valores: el soporte y la confianza, así que se iban modificando los valores para estos parámetros con el fin de ver que reglas se obtenían. Para el soporte de 0.6, confianza de 0.9 y un número de reglas predefinido en 10, no se generó ninguna regla que permitiera establecer bajo qué parámetros se podría presentar un embarazo a temprana edad. Lo mismo sucedió al establecer el valor de soporte en 0.4, la confianza en 0.5 y un número de reglas predefinido en 30. Estos son valores son muy bajos cuando se quieren obtener reglas confiables y todas las reglas generadas daban como resultado (NO embarazo). Algunas de estas reglas se muestran a continuación.

1. Edad=16 ==> Embarazo=0 [confianza:(0.96)]

2. Analfabetismo=0 Edad=16 ==> Embarazo=0 [confianza:(0.96)]

3. RelacionJefe=Hija ==> Embarazo=0 [confianza:(0.96)]

4. RelacionJefe=Hija Analfabetismo=0 ==> Embarazo=0 [confianza:(0.96)]

5. TamañoFlia=4 ==> Embarazo=0 [confianza:(0.96)]

6. Analfabetismo=0 TamañoFlia=4 ==> [Embarazo=0 confianza:(0.96)]

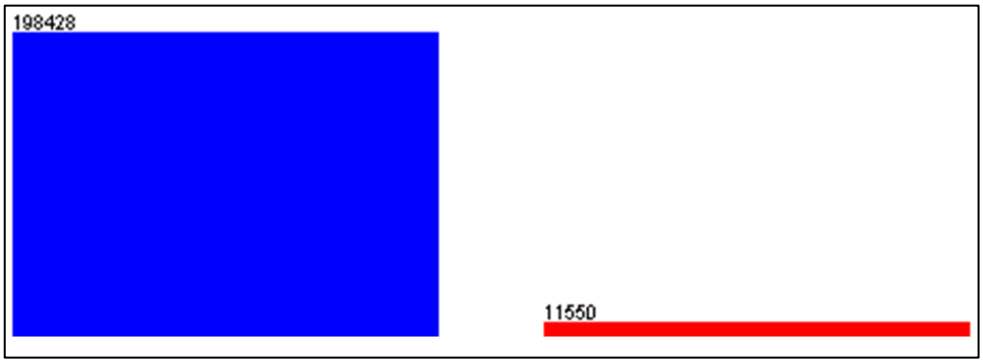
7. JefeFemenino=0 ==> Embarazo=0 [confianza:(0.95)]

8. Analfabetismo=0 ==> Embarazo=0 [confianza:(0.94)]

9. PrimariaD=1 ==> Embarazo=0 [confianza:(0.94)]

10. Analfabetismo=0 PrimariaD=1 ==> Embarazo=0 [confianza:(0.94)]

Figura 15. Histograma atributo embarazo



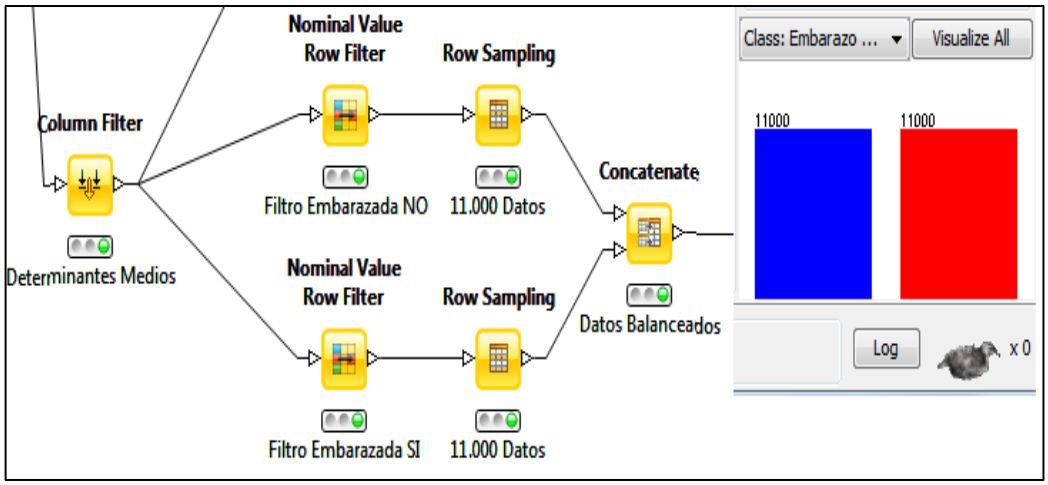
Al observar los datos se evidenció la causa del problema: el desbalance de la clase embarazo como se puede ver en la Figura 15 no permitía la generación de reglas para la condición Embarazo = SI, ya que tan solo el 5% de los registros pertenecían a embarazos adolescentes y para las condiciones de aquellos registros "SI" existían muchos más "NO" por lo tanto los algoritmos generaban este tipo de reglas.

Balanceos de datos con respecto a la clase

Para solucionar el problema del desbalanceo inicialmente se tomó una muestra aleatoria de 11.000 registros de cada uno de los dos valores que toma la clase (Embarazada: Si o No) como se puede ver en la Figura 16. Después se realizaron tomas de datos de muestra haciendo balances desequilibrados entre las dos clases, siempre preservando una mayoría de casos de adolescentes no embarazadas para mantener el conjunto de datos similar a la realidad. Con los diferentes conjuntos de datos se realizaron los entrenamientos de los métodos de clasificación.

Algunos de los resultados que se mostraran de las diferentes técnicas de minería de datos, corresponden a este subconjunto de datos balanceados con a un total de 22.000 registros, las matrices de confusión y la evaluación de las métricas corresponden al 10% de los datos que se utilizaron como prueba. El restante 90% fueron utilizados para el entrenamiento.

Figura 16. Esquema del balanceo del atributo embarazo usando la herramienta KNIME. A la derecha se muestra el histograma.



Búsqueda de asociaciones con las muestras de datos balanceados

Al inicio de la investigación se pretendía encontrar patrones por separado para cada tipo de determinante, para definir cuáles eran las condiciones personales, de entorno y de

sociedad que propiciaban embarazos tempranos. Por lo cual este método fue inicialmente probado con cada conjunto de variables por separado según las agrupaciones por determinante sociales (Proximales, Intermedios y Distales). Para este caso se obtuvieron resultados poco satisfactorios, ya que solo se pudieron encontraron algunas reglas de asociación con niveles de soporte bajos (alrededor del 40%). Adicionalmente, las reglas generadas por el algoritmo no representaban un descubrimiento de conocimiento importante dentro de la investigación, pues las condiciones lógicas para que una mujer fuera vulnerable están presentes en las variables de analfabetismo negativo y primaria positiva. Estas dos características se presentan en la mayoría de los individuos (embarazadas y no embarazadas). No se encontraron reglas satisfactorias en las otras variables analizadas en esta primera parte del estudio, por lo cual se puede decir que las variables con la que se hizo el análisis no son suficientes para crear reglas de interés y de alta confiabilidad. Con el objetivo de encontrar reglas que aporten conocimiento, se incluyeron todas las variables sin importar a cual determinante se habían asociado.

Finalmente, al realizar el ejercicio nuevamente se encontraron reglas de asociación con un valor de confiabilidad más alto y de mayor interés para el primer objetivo basado en encontrar patrones ocultos entre los datos que permitieran entender mejor la problemática abordada. Las reglas obtenidas con confiabilidad 0.7 o superior se listan a continuación:

1. RelacionJefe=Hija JefeFemenino=0 ==> Embarazo=0 confianza:(0.71)

2. RelacionJefe=Conyuge JefeFemenino=0 Poblacion_Especial=Ninguno ==> Embarazo=1 confianza:(0.71)

3. RelacionJefe=Conyuge Analfabetismo=0 JefeFemenino=0 Poblacion_Especial=Ninguno ==> Embarazo=1 confianza:(0.71)

4. RelacionJefe=Conyuge JefeFemenino=0 Servicio Sanitario=Sin Definir Poblacion_Especial=Ninguno ==> Embarazo=1 confianza:(0.71)

5. RelacionJefe=Conyuge Analfabetismo=0 JefeFemenino=0 Servicio Sanitario=Sin Definir Poblacion_Especial=Ninguno ==> Embarazo=1 confianza:0.71)

6. RelacionJefe=Conyuge JefeFemenino=0 Desempleado=No Poblacion_Especial=Ninguno ==> Embarazo=1 conf:(0.71)

7. RelacionJefe=Conyuge Analfabetismo=0 JefeFemenino=0 Desempleado=No Poblacion_Especial=Ninguno ==> Embarazo=1 confianza:(0.7)

8. RelacionJefe=Conyuge JefeFemenino=0 Desempleado=No Servicio Sanitario=Sin Definir Poblacion_Especial=Ninguno ==> Embarazo=1 confianza:(0.7)

9. RelacionJefe=Conyuge JefeFemenino=0 ==> Embarazo=1 confianza:(0.7)

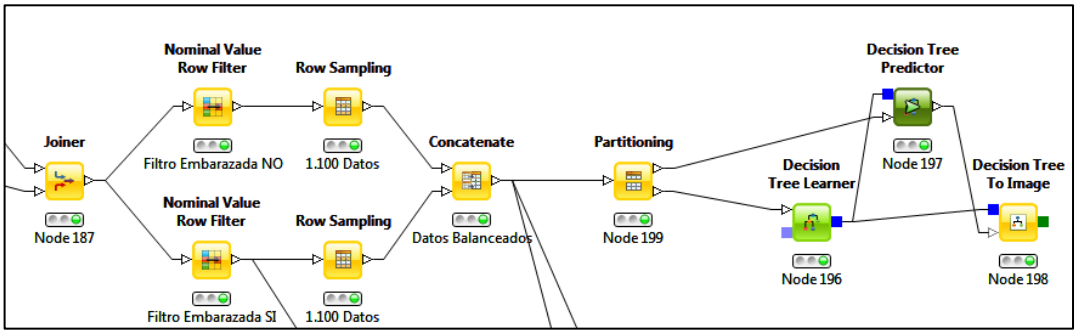
10. RelacionJefe=Conyuge Poblacion_Especial=Ninguno ==> Embarazo=1 confianza:(0.7)

3.2.1 Métodos de clasificación

Árboles de decisión

Con el conjunto de datos balanceado y con un mayor número de variables, donde se incluyeron aquellas relacionadas con los determinantes proximales, intermedios y distales, se construyó un predictor con árboles de decisión (ver Figura 17). Durante la etapa de pre procesamiento de los datos para esta técnica el conjunto de datos se redujo de 209.987 a tan solo 12.114 debido a una alta proporción de datos nulos en las variables socioeconómicas tales como Ocupación, Tipo de empleo y el tipo de vivienda. Este subconjunto de datos presentaba el desbalanceo natural de la clase embarazo con 1.179 registros de casos positivos y 11.497 de casos negativos. En la figura 17 se puede observar al lado izquierdo la unión de los registros correspondientes a los determinantes utilizados, después de haber pasado por la etapa de pre-procesamiento de las variables asociadas a dichos determinantes. Más adelante se realizó el balanceo de la clase realizando un muestreo a fin de conseguir un conjunto de datos balanceado. Finalmente, se realiza la partición de los datos en dos muestras configuradas para tener el 70% de los datos que se utilizaron para el entrenamiento y el 30% para la validación del modelo.

Figura 17. Modelo desarrollado con árboles de decisión.



En la figura 18 se pueden observar los primeros nodos del árbol generado por este predictor. El nodo principal corresponde a la relación Jefe de hogar, el cual indica la alta correlación con el riesgo de embarazo. Existe un número alto de valores positivos para el nodo “cónyuge o compañero permanente”, lo cual confirma la teoría presentada por la

OMS respecto a las uniones maritales a temprana edad. En total, el árbol de decisión obtuvo una profundidad de 5 niveles a pesar de haber sido entrenado con 13 variables. El segundo nivel del árbol fue el número de integrantes que conforman la familia, las variables Ocupación, Escolaridad y Edad generaron los niveles faltantes, Las restantes 8 variables que hicieron parte del entrenamiento y que no generaron nodos indican su baja influencia en la generación de modelos predictivos con esta técnica.

Una vez obtenido estos resultados se realizaron experimentos con un sistema de validación cruzada, como se puede observar en la tabla 6 la cual muestra uno de los resultados obtenidos con este clasificador y este tipo de validación se obtuvieron mejores resultados. En resumen las métricas de rendimiento fueron: Exactitud del 67%, Precisión del 66% y Sensibilidad del 73% con un total de 82 verdaderos positivos.

Una regla de clasificación interesante generada por este método describe que cuando el jefe de hogar está representado por un hermano, el clasificador asigna directamente el valor de embarazo = SI. Esto indica un factor de riesgo específico para la población más vulnerable en Bogotá.

La tabla 6 presenta los resultados de clasificación obtenidos con este método, hasta ahora presenta las mejores métricas de rendimiento teniendo en cuenta que la Sensibilidad está relacionada con el acierto de los casos positivos reales de embarazos.

Figura 18. Primer nivel del árbol de decisión

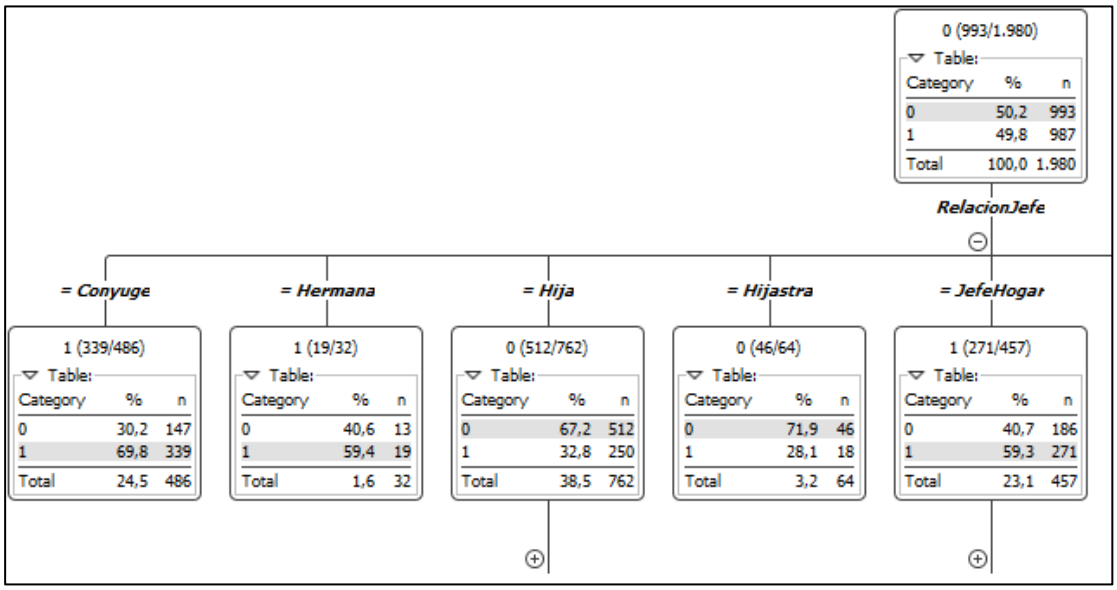


Tabla 6. Matriz de confusión resultante de clasificar 220 registros con el modelo de árboles de decisión.

Partición para entrenamiento 30/70

Total de registros: 660 Correspondientes al 30% de 2.200

		Predicción	
		SI	NO
Clase	SI	232	115
	NO	116	197
Exactitud:			0.65
Precisión:			0.54
Sensibilidad:			0,67

Validación cruzada 10/90

Total de registros: 220 Correspondientes al 10% de 2.200

		Predicción	
		SI	NO
Clase	SI	82	31
	NO	41	66
Exactitud:			0.65
Precisión:			0.54
Sensibilidad:			0,67

Redes Bayesianas

En la aplicación de esta técnica de clasificación se realizaron experimentaciones con dos tipos de algoritmos de entrenamiento. En esta sección se realiza una pequeña explicación de la historia y la lógica que los caracterizan.

K2

Algoritmo que fue desarrollado por Cooper y Herskovits(1992). El cual encuentra el conjunto de padres más probables, utilizando la métrica Bayesiana. Mide la probabilidad de la estructura dados los datos. La heurística de este algoritmo se basa en un ordenamiento topológico que tiene que ser especificado por el usuario (Li, Wang, & Leung, 2012).

En la figura 19 se puede observar la red bayesiana creada con este algoritmo de búsqueda y en las tablas 6 y 7 se muestran los resultados y eficiencia de la misma.

SE observa que el resultado generado con la aplicación de este algoritmo no es bueno, ya que el nodo principal está asociado a todos los demás nodos y no se establece un patrón definido entre los diferentes atributos que definen el conjunto de datos. Además, los valores de las métricas de rendimiento fluctúan en promedio en 0.64 y la sensibilidad para casos positivos fue tan solo de 0.654.

Figura 19. Red Bayesiana generada con algoritmo de búsqueda K2

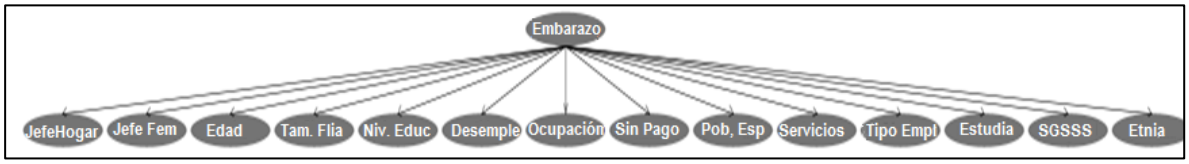


Tabla 7. Matriz de resultados obtenidos de la red bayesiana utilizando K2 como algoritmo de búsqueda

	Tasa VP	Tasa FP	Precisión	Sensibilidad	F-Score	AUC	Clase
Promedio ponderado	0.628	0.346	0.645	0.628	0.636	0.701	0
	0.654	0.372	0.637	0.654	0.645	0.701	1
	0.641	0.359	0.641	0.641	0.641	0.701	

Tabla 8. . Matriz de confusión obtenida de la red bayesiana utilizando K2 como algoritmo de búsqueda

	a	b
a=0	691	409
b=1	381	719
Exactitud:	0.64	

Ascenso de Colina (*Hill Climbing*)

La idea principal del uso de este algoritmo de entrenamiento es encontrar el valor máximo de una función objetivo, mejorándolo paso a paso. Se debe definir un criterio de satisfacción para la búsqueda, comúnmente asociado a un valor esperado o a un número máximo de iteraciones (Gent & Walsh , 1992). Los resultados de aplicar ascenso de colina se muestran en las tablas 9 y 10, como se puede observar los valores para las métricas de rendimiento precisión y sensibilidad varían muy poco respecto a las de las redes bayesianas entrenadas con algoritmo K2. Por otra parte la exactitud es mejor utilizando Ascenso de colina como algoritmo de entrenamiento, sin embargo no deja de ser un clasificador con rendimientos bajos de clasificación con respecto a los obtenidos con los árboles de decisión.

Tabla 9. Matriz de resultados obtenidos de la red bayesiana utilizando Ascenso de Colina como algoritmo de búsqueda.

	Tasas VP	Tasa FP	Precisión	Sensibilidad	F score	AUC	Clase
Promedio ponderado	0.634	0.336	0.653	0.634	0.643	0.699	0
	0.664	0.366	0.644	0.664	0.654	0.699	1
	0.649	0.351	0.649	0.649	0.649	0.699	

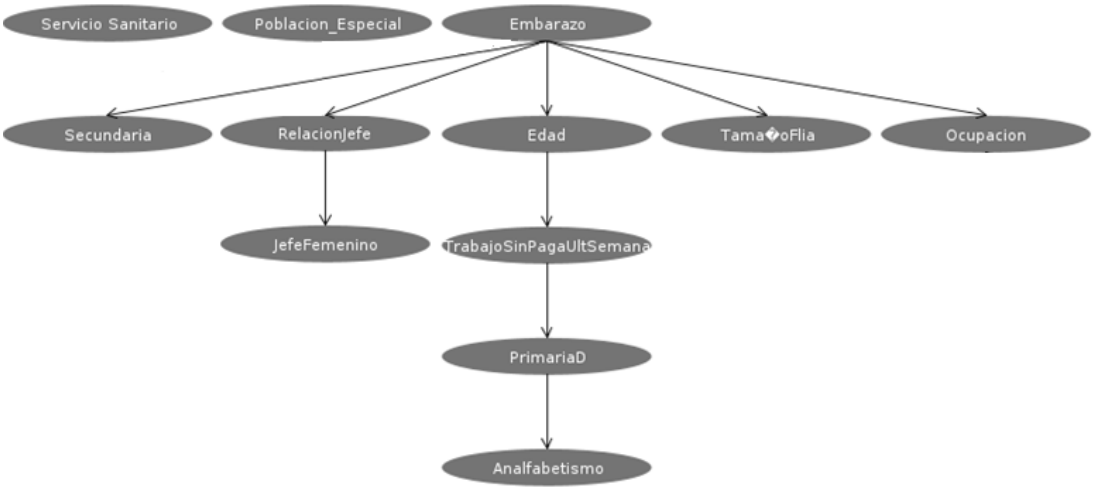
Tabla 10. Matriz de confusión obtenida de la red bayesiana utilizando ascenso de colina como algoritmo de búsqueda.

	A	B
a=0	697	403
b=1	370	730
Exactitud:	0.68	

En la figura 20 se observa la red generada con la aplicación de ascenso de colina como algoritmo de entrenamiento en métodos de redes bayesianas. A diferencia de la red generada con K2, se obtiene una red jerárquica en la cual se presentan patrones relacionales entre los nodos. Cada nodo representa las dimensiones del conjunto de datos. Para el objetivo de esta investigación de encontrar patrones entre las variables que puedan determinar el riesgo de una menor en quedar en estado de gestación partiendo de sus condiciones socioeconómicas, este tipo de redes resultan más útiles que aquellas que no presentan un nivel de ramificación jerárquico.

Con los resultados obtenidos se puede inferir que los servicios sanitarios con los que cuentan las viviendas donde viven las adolescentes y la pertenencia a algún tipo de población especial no son factores influyentes para la caracterización del riesgo. Por otra parte resulta interesante analizar la relación existe entre la Edad y Trabajo sin Paga la Última Semana que a su vez se asocia a las variables que indican un bajo grado de escolaridad (Primaria y Analfabetismo) con respecto al embarazo. Esto brinda una perspectiva de la problemática presentada en las poblaciones de menores recursos económicos con respecto a la baja educación, la explotación infantil y el embarazo como consecuencia de las variables de inequidad social presentadas por la teoría de la OMS como factores fundamentales de riesgo.

Figura 20. Red Bayesiana generada con algoritmo de entrenamiento Ascenso de colina.



Naive Bayes

Este es un método probabilístico que realiza la clasificación suponiendo los datos por medio de las probabilidades de los atributos del conjunto de datos, asumiendo la presencia o ausencia de una característica. Es un método de aprendizaje supervisado ya que requiere de la evaluación de la clasificación con respecto al atributo clase. La probabilidad de que un evento X ocurra dado Y es representado por la ecuación:

$$P(Y|X_1, \dots, X_N) = \frac{P(Y)}{P(X)} \prod_i P(X_i|Y) \quad (7)$$

Durante la aplicación de este método se crean probabilidades para cada atributo con respecto al valor que toma la clase. Después de generado el modelo se realizó validación cruzada de 10 iteraciones (10 *fold cross validation*), los resultados se muestran en las tablas 11 y 12. Al igual que los experimentos realizados con redes bayesianas los valores para las métricas oscilan en valores promedio de 0.66 para la exactitud, la precisión y la sensibilidad. Sin embargo, los resultados muestran que para el conjunto de datos el método de clasificación Naive Bayes tiene un mejor rendimiento y confiabilidad frente a los otros dos métodos, debido a que es la que presenta mayor área bajo la curva ROC.

Tabla 11. Matriz de resultados con la aplicación de Naive bayes

	Tasa VP	Tasa FP	Precisión	Sensibilidad	F-Score	ROC Área	Clase

Promedio ponderado	0.651	0.333	0.662	0.651	0.656	0.706	0
	0.667	0.349	0.657	0.667	0.662	0.706	1
	0.659	0.341	0.659	0.659	0.659	0.706	

Tabla 12. Matriz de confusión obtenida con la aplicación de Naive Bayes como método de clasificación para el conjunto de datos

	A	b
a=0	716	384
b=1	366	734
Exactitud:	0.66	

Redes neuronales

Perceptron multicapa entrenado por gradiente descendiente

Se realizaron algunos experimentos con redes neuronales tipo perceptron multicapa entrenado por gradiente descendiente. La experimentación se realizó variando el número de capas ocultas y la cantidad de neuronas por capa. El mejor clasificador obtenido con esta técnica fue el de 5 capas ocultas con 8 neuronas por capa, la cual obtuvo una exactitud de 0.62 como se puede observar en la tabla 13. El valor de sensibilidad también es bajo, teniendo en cuenta que esta medida mide los casos positivos clasificados con base en los valores positivos reales, este método no representa una buena solución para el modelo predictivo.

Tabla 13. Matriz de resultados con clasificación de redes neuronales entrenado por gradiente descendiente

		Predicción	
		SI	NO
Clase	SI	700	460
	NO	382	658
Exactitud:		0.62	
Precisión:		0.64	
Sensibilidad:		0.60	

Perceptron multicapa utilizando RProp como algoritmo de optimización

RProp es un algoritmo que realiza una adaptación local de los cambios de los pesos de acuerdo con el comportamiento de una función de error, mejorando los inconvenientes

presentados por métodos de gradiente descendiente respecto a la velocidad de entrenamiento (Riedmiller & Braun, 1993).

La configuración de la red neuronal se puede realizar modificando el número de capas y el número de neuronas que componen cada capa. Para el caso de estudio, se evaluaron diferentes configuraciones para la red neuronal, modificando la cantidad de capas ocultas al igual que se realizó con las redes neuronales entrenadas con gradiente descendiente. Para todos los casos se dejó como parámetros de configuración los siguientes valores: 100 épocas de entrenamiento debido a que se comprobó que eran suficientes para lograr la convergencia y con 8 neuronas por cada capa. Para cada aproximación se realizó 10 veces el entrenamiento y la clasificación y se tomó el promedio para determinar los resultados, los cuales se muestran en la tabla 14.

Con base en los resultados obtenidos se puede afirmar que las redes neuronales representan un buen modelo de clasificación para el problema propuesto. Los resultados muestran que la arquitectura que presentó mejor desempeño para estas redes neuronales fue la de tres capas ocultas.

Tabla 14. Resultados de clasificaciones usando RProp MLP

1 capas oculta

		Predicción	
		SI	NO
Clase	SI	773	306
	NO	439	682

Exactitud:	0.66
Precisión:	0.63
Sensibilidad:	0.71

2 capas ocultas

		Predicción	
		SI	NO
Clase	SI	773	306
	NO	430	691

Exactitud:	0.66
Precisión:	0.64
Sensibilidad:	0.72

3 capas ocultas

Predicción	
SI	NO

4 capas ocultas

Predicción	
SI	NO

Clase			
	SI	829	250
	NO	391	730

Exactitud:	0.71
Precisión:	0.68
Sensibilidad:	0.77

Clase			
	SI	771	308
	NO	433	688

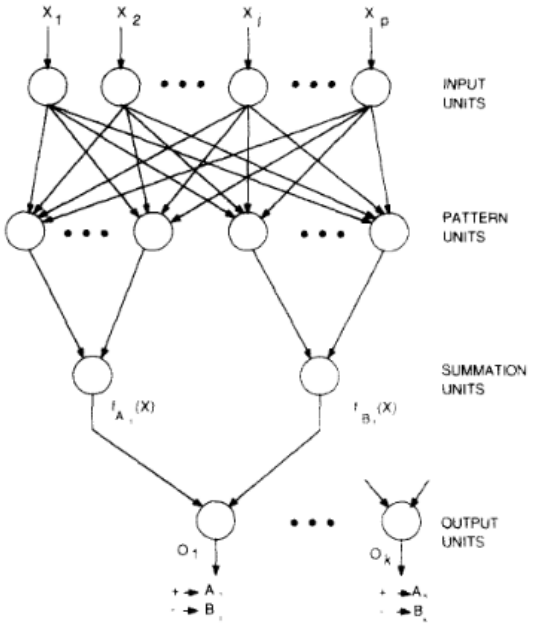
Exactitud:	0.66
Precisión:	0.64
Sensibilidad:	0.72

Redes neuronales probabilísticas (*Probabilistic Neural Network*, PNN)

PNN es una red neuronal principalmente compuesta por una arquitectura de cuatro capas. La primera de ellas es la de entrada la cual cuenta con un número *n* de nodos que representan las dimensiones de los vectores dada por la cantidad de variables que ingresan al sistema, la segunda se conoce como patrón o capa de formación. En cuanto a su arquitectura una PNN es similar a un perceptron multicapa pero difiere en el método en que se entrena. Este tipo de red neuronal no contiene el peso de las neuronas en la capa oculta, en vez de esto está compuesta por vectores de ejemplo, los cuales no necesariamente son modificados durante el aprendizaje. Todos los nodos de la capa de entrada están interconectados con la capa oculta. Una red neuronal PNN tiene una tercera capa que suma los vectores resultantes antes de conectarse a la capa de salida. Sin embargo, no todas las neuronas de la capa oculta se conectan con las neuronas de esta capa. La última capa es la de salida la cual representa las clases en que pueden ser clasificados los registros del conjunto de datos (Spetch, 1990).

Un aspecto interesante de las PNN es que dado un vector de entrada desconocido, las activaciones de los nodos ocultos se calculan y se suman después en la capa de salida. La clase con la activación mayor es la que determina la clase a la que pertenece el registro (Padhy, 2005).

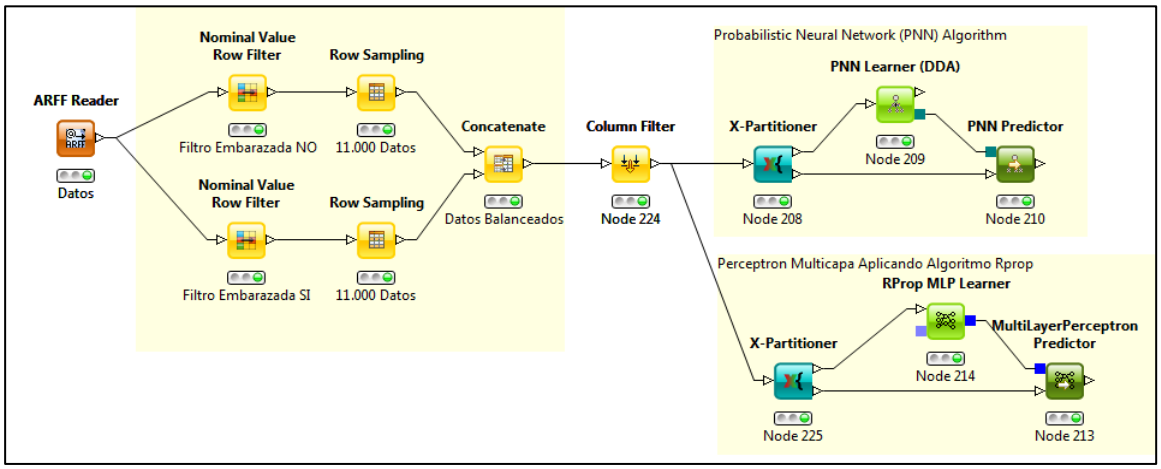
Figura 21. Arquitectura de una red neuronal probabilística (PNN).



Fuente: Donald F. Specht (1990)

La figura 22 muestra el modelo desarrollado con redes neuronales PNN con algoritmo de entrenamiento DDA (*Dynamic Decay Adjustment*). Este algoritmo genera reglas basadas en datos numéricos. Cada regla se define como una función de Gauss de alta dimensión que se ajusta por dos umbrales, ϕ mínimo (ϕMín) y ϕ máximo (ϕMáx). Estos umbrales permiten evitar conflictos con las reglas para las diferentes clases. Cada función de Gauss se define por un vector de centro y una desviación estándar que se ajusta durante el entrenamiento para permitir únicamente los casos donde no hay conflicto entre las reglas generadas (Berthold & Diamond, 21 April 1998).

Figura 22. Modelo de aplicación para redes neuronales PNN y RProp Multilayer Algoritm.



Existen dos aspectos interesantes de usar este tipo de redes neuronales entrenadas con este algoritmo. Por una parte se pueden determinar las reglas aprendidas por el modelo y saber cuáles de estas tiene mayor importancia para la clasificación, ya que el algoritmo genera un peso a cada regla para el proceso de clasificación. Por otra parte, se puede determinar la probabilidad que el modelo le da a cada registro para ser clasificado en cada valor de la etiqueta clase. Para el caso de estudio estas dos particularidades son muy importantes, porque no solo se desarrolla el modelo predictivo, sino que también se puede hacer descubrimiento de conocimiento al hacer análisis de las reglas con mayor peso, las cuales permiten tener un mayor entendimiento de la problemática de embarazo adolescente en Bogotá. La probabilidad dada a cada registro podría ser interpretada como el riesgo de una adolescente de quedar en embarazo según los patrones encontrados en las variables analizadas. La tabla 15 muestra un ejemplo sobre la regla 140 generado por el modelo desarrollado con PNN durante los trabajos realizados con el segundo conjunto de datos que se explicara en el próximo capítulo. En esta tabla se puede observar en la última columna que asigna un mayor peso a esta regla que a las demás para la clasificación. De este modo se puede inferir que el embarazo adolescente está relacionado con los bajos niveles educativos y con los niveles de pobreza asociados a las variables de índice de hacinamiento, riesgos de la vivienda y la falta de dinero para comprar alimentos.

Tabla 15. Reglas generadas por el modelo de RNN entrenado con Algoritmo DDA

Row ID	NivelEducativo	Oficios escolares	Oficios del hogar	Otras actividades	FaltoDineroCompraAlimentos	RiesgosVivienda	IndiceHacinamiento	Embarazo	Weight
Rule_138	0.682	0	0	0	0	1	1.5	SI	1
Rule_139	0.442	0	0	0	1	0	3	SI	1
Rule_140	0.488	0	0	0	1	1	7	SI	5
Rule_141	0.493	0	0	0	1	1	1	SI	1
Rule_142	0.544	0	0	0	1	1	1.5	SI	1

La figura 23 muestra algunos de los registros de predicción y la probabilidad asignada por el modelo, por ejemplo, para el registro 1710 el clasificador asignó una probabilidad a la adolescente de quedar en embarazo del 88.7%.

Figura 23. Ejemplo de probabilidades asignadas para la clasificación dadas por el modelo PNN a cada registro.

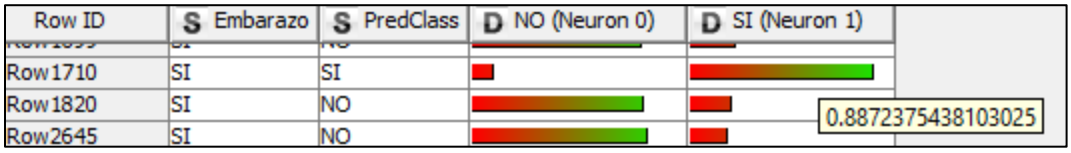


Tabla 16. Resultados de clasificaciones utilizando PNN con algoritmo de entrenamiento DDA

ϕ Mí = 0.2 – ϕ Max = 0.4 **ϕ Mí = 0.2 – ϕ Max = 0.6** **ϕ Mí = 0.2 – ϕ Max = 0.8**

		Predicción	
		SI	NO
Clase	SI	778	322
	NO	435	665

Exactitud: 0.66
Precisión: 0.64
Sensibilidad: 0.71

		Predicción	
		SI	NO
Clase	SI	775	325
	NO	437	663

Exactitud: 0.65
Precisión: 0.64
Sensibilidad: 0.7

		Predicción	
		SI	NO
Clase	SI	782	318
	NO	439	661

Exactitud: 0.66
Precisión: 0.64
Sensibilidad: 0.71

ϕ Mí = 0.4 – ϕ Max = 0.6

ϕ Mí = 2 – ϕ Max = 0.8

ϕ Mí = 0.6 – ϕ Max = 0.8

		Predicción	
		SI	NO
Clase	SI	783	317
	NO	437	663

Exactitud: 0.66
Precisión: 0.64
Sensibilidad: 0.71

		Predicción	
		SI	NO
Clase	SI	779	321
	NO	437	663

Exactitud: 0.66
Precisión: 0.64
Sensibilidad: 0.71

		Predicción	
		SI	NO
Clase	SI	786	314
	NO	440	660

Exactitud: 0.66
Precisión: 0.64
Sensibilidad: 0.71

Para el caso de las redes neuronales probabilísticas, se realizó una fase experimental modificando los valores de separación angular entre ϕMín y ϕMáx. Como se puede observar en las diferentes métricas de cada configuración mostradas en la tabla 16, para este conjunto de datos en particular la modificación de los parámetros del algoritmo de entrenamiento no afecta significativamente el resultado de la clasificación, esto indica que las reglas generadas por la red neuronal PNN entrenada bajo algoritmo DDA no presentan conflictos entre ellas y siempre se generan las mismas reglas de clasificación para este conjunto de datos en particular. En total cada configuración generó un total de 75 reglas de clasificación, de las cuales 36 fueron para clasificaciones negativas y 39 para clasificaciones positivas.

Análisis de resultados obtenidos con los métodos de clasificación.

Después de revisar los resultados obtenidos se puede concluir que las redes neuronales representan un buen modelo de clasificación para el problema propuesto. La variación de las capas ocultas de la red genera un cambio significante en el resultado final. Por otra parte, para el caso de los árboles de decisión el índice de sensibilidad para todos es cercano al 70% al igual que las redes neuronales generadas con PNN y RProp. Aunque los resultados muestran que la arquitectura que presentó mejor desempeño para estas redes fue el de

perceptron multicapa con tres capas ocultas y ocho neuronas por capa, se puede observar que la variación entre un modelo y otro no es significativamente alta. La selección del mejor modelo predictivo se convierte en una tarea que requiere una evaluación más concienzuda debido a que estos modelos son superiores en algunas métricas de rendimiento pero inferiores en otras, ej. Los árboles de decisión presentan una mejor sensibilidad que las redes neuronales, esto implica que será más asertivo en la detección de menores con alto riesgo real, pero tiene un valor inferior en la precisión lo cual lo llevara a predecir más adolescentes sin riesgo real (ver ecuaciones 4 y 5). Por otra parte, los métodos desarrollados con redes bayesianas no fueron tan eficientes como estos últimos, por lo cual no fueron considerados durante los trabajos realizados con el segundo conjunto de datos. Mientras que los clasificadores desarrollados con el método de Naive Bayes permitió encontrar patrones relacionales de interés entre las variables y su desempeño de clasificación tuvo un comportamiento medio comparado con los otros métodos.

3.2.2 Métodos de agrupación

Los métodos de agrupamiento pueden clasificarse en cuatro grupos fundamentales: jerárquicos, de centroide, basados en densidad y basados en distribución. Dependiendo de los métodos en que los algoritmos realicen la agrupación del conjunto de datos (Tan, Steinbach, & Kumar). En este trabajo dos métodos de agrupación fueron utilizados: Agrupamiento basado en distribución y Agrupamiento basado en densidad.

Agrupamiento basado en distribución

En la experimentación con métodos de agrupamiento para el conjunto de datos se utilizaron dos métodos, algoritmo maximización de la esperanza (*Expectation Maximization*, EM) y K-medias (*K-means*). Para cada uno de los cuales se realizaron pruebas cambiando el número de grupos entre 2 y 6. Los resultados obtenidos durante cada aproximación se consignan en las siguientes tablas donde se resalta de color rojizo el grupo en el cual se agrupan más registros con valor 0 para el atributo embarazo con respecto al 1 y de color verde para el caso contrario que representa mayor cantidad de puntos agrupados para embarazos positivos.

Algoritmo maximización de la esperanza (*expectation maximization*, EM)

Tabla 17. Resultado de agrupamiento EM de 2 grupos con respecto a los registros de la etiqueta clase

Grupo	0	1
No Embarazada	394	706

Embarazada	332	768
% respecto al total	54.27%	52.10%

Tabla 18. Resultado de agrupamiento EM de 3 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*
No Embarazada	561	162	377
Embarazada	478	370	252
% respecto al total	53.99	69.55	59.94%

Tabla 19. . Resultado de agrupamiento EM de 4 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*	*3*
No Embarazada	155	220	344	381
Embarazada	363	285	231	221
% respecto al total	70.08%	56.44%	59.83%	63.29%

Tabla 20. Resultado de agrupamiento EM de 5 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*	*3*	*4*
No Embarazada	86	320	153	335	206
Embarazada	96	176	362	202	264
% respecto al total	52.75%	64.52%	70.29%	62.38%	56.17%

Tabla 21. Resultado de agrupamiento EM de 6 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*	*3*	*4*	*5*
No Embarazada	143	177	85	257	162	276
Embarazada	344	213	97	145	164	137
% respecto al total	70.64%	54.62%	53.30%	63.93%	50.31%	66.83%

Algoritmo k-medias (*K-Means*)

Tabla 22. Resultado de agrupamiento k-medias de 2 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*
No Embarazada	666	434
Embarazada	682	418
% respecto al total	50.59%	50.94%

Tabla 23. . Resultado de agrupamiento k-medias de 3 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*
No Embarazada	377	396	327
Embarazada	298	333	469
% respecto al total	55.85%	54.32%	58.92%

Tabla 24. Resultado de agrupamiento k-medias de 4 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*	*3*
No Embarazada	301	380	162	257
Embarazada	195	257	352	296
% respecto al total	60.69%	59.65%	68.48%	53.53%

Tabla 25. Resultado de agrupamiento k-medias de 5 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*	*3*	*4*
No Embarazada	266	364	83	258	129
Embarazada	165	246	114	303	272
% respecto al total	61.72%	59.67%	57.87%	54.01%	67.83%

Tabla 26. Resultado de agrupamiento k-medias de 6 grupos con respecto a los registros de la etiqueta clase

Grupo	*0*	*1*	*2*	*3*	*4*	*5*
No Embarazada	287	150	88	238	124	213
Embarazada	168	141	126	252	246	167
% respecto al total	63.08%	51.55%	58.88%	51.43%	66.49%	56.05%

Cuando se aplicaron estas dos técnicas los algoritmos realizaban agrupaciones dentro del espacio vectorial de conjuntos de datos con características similares, para un número pequeño de grupos no se creaban grupos cuya característica separara casos de embarazo, en las tablas se puede observar que a medida que se aumentaba el número de grupos, comienzan a aparecer grupos capaces de segmentar aquellos registros de adolescentes con y sin embarazo. Para el caso de las dos técnicas utilizadas por agrupamiento basado en distribución, el método EM tuvo mejores resultados que k-medias. El siguiente paso consistió en realizar análisis a las características de los registros agrupados en los grupos que generaron mejor segmentación entre los casos positivos y negativos de embarazo. Las características de las variables cuyo valor permanece igual para todos los registros dentro del espacio vectorial se describen a continuación para cada grupo.

Tabla 27. Características descriptivas de los grupos generados con métodos basados en distribución

Técnica	Grupos	Características
EM	4	RelacionJefe=Otro Pariente, Desempleado=No, Analfabetismo=0, TrabajoSinPagaUltMes = 1,
EM	5	RelacionJefe=Nuera, Analfabetismo=1, Servicio Sanitario=Sin Definir
EM	6	RelacionJefe=Conyuge, JefeFemenino=0
k-*Means*	5	RelacionJefe=Conyuge, JefeFemenino=0
k-*Means*	6	RelacionJefe=Conyuge

Nuevamente se encuentra que las variables asociadas al Jefe de Hogar, género del jefe de la familia, Analfabetismo y trabajó sin pago durante el último mes son patrones de incidencia sobre el embarazo a temprana edad.

En cuanto a los resultados mostrados por este método y el de reglas de asociación se puede inferir también que existe una relación directa con el género del Jefe de Hogar, a diferencia de la teoría presentada por la OMS donde se estipula un mayor índice de riesgo en las adolescentes cuya jefaturas del hogar está representada por una mujer en la población de Bogotá analizada en este estudio sucede exactamente lo contrario, se presenta un patrón de riesgo en jefaturas masculinas.

METODOS BASADOS EN LA DENSIDAD

Algoritmo DBSCAN

Para aplicas este algoritmo es necesario definir un número de vecinos alrededor de cada punto los cuales son usados para calcular la densidad del grupo a medida que se va formando. El algoritmo comienza por la selección de un punto al azar, conocido como punto central. Con base a este punto, comienza a formar un grupo al ir uniendo a todos los puntos con una densidad alcanzable desde el punto central. Si el punto seleccionado para unirse a algún grupo no puede estar relacionado con un punto central, se toma otro punto del conjunto de datos. Este proceso iterativo continúa hasta que todos los puntos han sido procesados. Los puntos que caen fuera de los grupos se denominan puntos de ruido. Los puntos que no son ni puntos ni ruido ni centros se denominan puntos de la frontera. De esta forma DBSCAN construye grupos en donde un grupo puede tener más de un punto

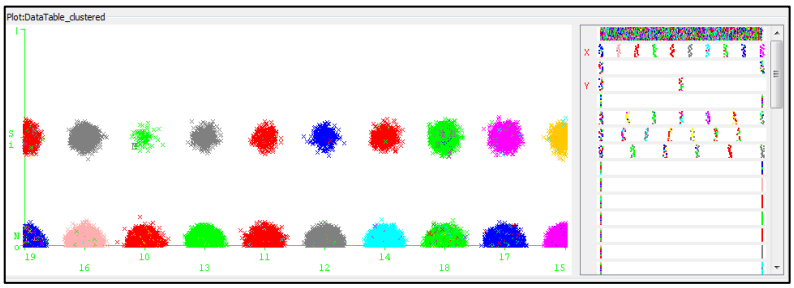
central, al igual que varios puntos bordes y putos ruido (Ester , Kriegel, Sander, & Xu, 1996). Los resultados obtenidos con esta técnica se describen a continuación:

Parametrización: Epsilon: 0.9; minPoints: 6 (Valores para la densidad)

Grupos generados: 112

Puntos que no son agrupados: 156

Figura 24. Ilustración de grupos generados con DBSCAN como algoritmo de agrupamiento



En la figura 24 se pueden observar los grupos generados por este método. El algoritmo no permite un número de grupos definidos por el usuario y puede representar una buena opción si se quiere encontrar algún patrón de agrupamiento desconocido sin una clase especifica. Por otra parte se puede observar grupos definidamente separables. Para el caso de estudio la aplicación de este tipo de técnicas no generó información útil.

Para el caso experimental cuando se asignó el atributo clase al algoritmo, genero un total de 112 grupos, pero no se lograron encontrar características que definieran una segmentación de los patrones buscados que permitieran aportar información valiosa.

Algoritmo k-NN

Este algoritmo es usado generalmente para agrupar conjuntos de datos con un gran número de variables como es el caso de este estudio. Utiliza la regla de los vecinos más cercanos. La regla k-NN ha sido extensamente usada en muchos métodos de clasificación donde existe un conjunto de objetos etiquetados (Pascual & Pla, 2007). En este algoritmo se determina el número de grupos de manera automática. Necesita como parámetro de entrada únicamente la cantidad de vecinos de manera similar que DBSCAN.

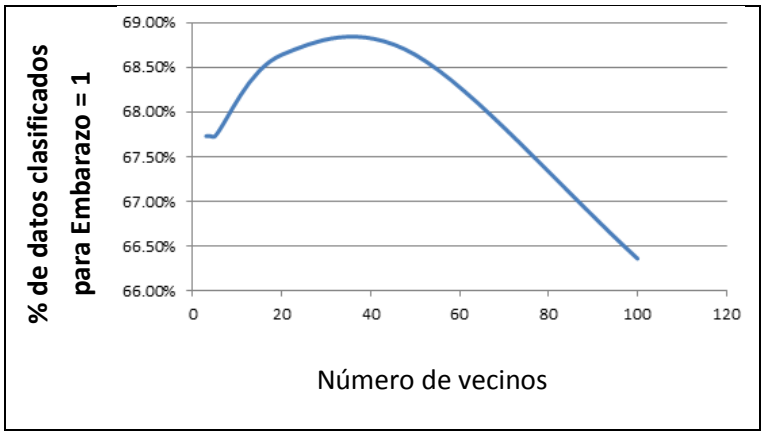
El algoritmo comienza asignando a cada punto a un grupo individual. Se calcula para cada punto sus k vecinos más cercanos. Para cada punto de la base de datos se aplica la regla k-NN y se asigna a un grupo según esta regla, este proceso se repite hasta que ninguno de los objetos cambia de grupo en dos iteraciones sucesivas.

En la aplicación de este método se encontraron buenos resultados con respecto al agrupamiento de registros con características de embarazo = 1, el experimento se repitió varias veces modificando la cantidad de vecinos para cada punto central. La tabla 28 y la figura 25 muestran los resultados obtenidos durante este proceso. En estas se puede observar que se clasificaron correctamente hasta el 68,64% de los registros con valores entre 20 y 50 para el único parámetro del algoritmo (Cantidad de vecinos para un punto central).

Tabla 28. Resultados obtenidos durante experimentos con k-NN

Número de vecinos	% de datos clasificados para Embarazo = 1
3	67.73%
4	67.73%
5	67.73%
20	68.64%
50	68.64%
100	66.36%

Figura 25. Comportamiento del agrupamiento del conjunto de datos con el algoritmo k-NN respecto a registros cuyo atributo Embarazo es 1.



3.3 Conclusiones de resultados con el primer conjunto de datos

En el contexto del caso de estudio, la mejor aproximación para desarrollar el segundo objetivo de crear un modelo predictivo corresponde a los métodos de clasificación, puesto que se cuenta con la clase. Por otra parte los modelos de clasificación pueden perfeccionarse ingresando aquellas variables de mayor incidencia, en este aspecto las técnicas de minería de datos basadas en agrupamiento y las de asociación juegan un papel importante en la detección de patrones relacionales entre las variables. Los modelo de clasificación que presentaron mayor exactitud y precisión fueron: Naive Bayes y las redes neuronales.

Aunque el resultado de la agrupación no está directamente relacionada con la clase definida por un experto, al hacer una evaluación de la clase dentro de los grupos obtenidos en el contexto definido, se puede concluir que entre mayor sea el número de grupo se pueden encontrar los espacios vectoriales donde se encuentran las características para la incidencia de embarazo a temprana edad. Por otra los métodos de agrupamiento por densidad presentan un mejor comportamiento y resultado para el conjunto de datos de este estudio.

Se pudo definir un conjunto de técnicas apropiadas para la búsqueda de patrones en el conjunto de datos a partir de la información que se pueda extraer y otras para la generación de un modelo predictivo. Con este conjunto de técnicas se debe trabajar en el desarrollo de la herramienta tecnológica basada en el mejor modelo predictivo que permita identificar a las menores con mayor índice de riesgo de embarazo.

No todas las técnicas son apropiadas para un conjunto de datos específico; en este trabajo se pudo evidenciar que algunos métodos de minería presentaron mejores resultados para los objetivos planteados.

4 Segunda Aproximación – Mejorando el modelo predictivo

¿Por qué una segunda aproximación?

Los datos considerados inicialmente consistían en la información de la encuesta realizada a las personas por primera vez por parte del programa del gobierno "Salud a su Casa". Esta encuesta es conocida como encuesta de caracterización. Con esta encuesta se obtienen los datos de entorno familiar y social de los individuos. Sin embargo, para el caso de este estudio estos datos presentan la particularidad de niñas que no se encontraban en embarazo al momento de ser encuestadas pero que después de un tiempo quedaron embarazadas estando aun en su edad adolescente, pero no se tuvo en cuenta este factor en el primer conjunto de datos.

El programa Salud a Su Casa no consiste únicamente en hacer encuestas a la población, también realiza intervenciones médicas de atención, prevención y educación en salud. El programa cuenta con un repositorio de datos obtenidos durante los once años que lleva el programa en funcionamiento. El nuevo enfoque consistió en analizar los datos de todas las mujeres con 21 años de edad que cumplieran con las siguientes características:

1. Haber sido caracterizadas antes del 2006, lo cual asegura que se tienen registros desde sus 10 años, edad de la cual se tienen casos reportados de embarazo adolescente en Bogotá por el programa "Salud a su Casa".
2. Estar activas y haber tenido seguimientos durante su momento de captación hasta el momento en el que se extrajo el conjunto de datos del Sistema de Información. Esto permite tener la confianza de conocer si se presentaron embarazos durante la adolescencia.

Características específicas de los nuevos datos

Se obtuvo un total de 37.897 registros de mujeres con 21 años de las cuales el 10% tuvieron un embarazo adolescente (3.861 en total). Como se describió anteriormente lamentablemente la captura de datos tiene muchas falencias en cuanto a la calidad del dato. Una vez realizados los procesos de depuración de registros que permitieran entrada de datos sin ruido y sin datos perdidos al modelo se obtuvo un conjunto de datos total de 5.743 registros, de los cuales 413 pertenecen a mujeres que tuvieron un embarazo durante su adolescencia, es decir el 7.2%.

Nuevas dimensiones de los datos

Gracias al proceso realizado con el primer conjunto de datos se redujo la cantidad de variables teniendo en cuenta aspectos de impacto sobre el modelo y la calidad del dato. Este nuevo enfoque también trajo como inconveniente no contar con la variable Edad, la cual fue determinante durante la primera aproximación, además de ser una variable considerada como influyente para determinar el riesgo de un embarazo temprano según los estudios de la Organización Mundial de la Salud. Se consideró inicialmente tener en cuenta la edad del primer embarazo de los 413 registros de mujeres que presentaron embarazos adolescentes. Sin embargo esto traía consigo otro dilema: ¿Qué valor se debería colocar a los otros 5.330 registros?, si se colocaba 21 años los modelos quedarían mal entrenados, porque todas las edades adolescentes estarían relacionadas con embarazo. Por este motivo se tuvo que prescindir de esta variable. Una nueva variable llamada índice de hacinamiento fue considerada la cual está fuertemente asociada a la pobreza y a la inequidad social en trabajos de investigación que abordan este tema y la cual se calcula mediante la siguiente formula:

$$i_h = \frac{Número\ de\ Dormitorios}{Número\ de\ personas} \tag{8}$$

Generalmente se aceptan los valores:

- Hasta 2.4 - sin hacinamiento.
- De 2.5 a 2.9 - hacinamiento medio.
- Más de 3.0 - hacinamiento crítico.

(Reporte de indicadores ONU-HABITAT en las ciudades de Veracruz periodo 2000-2010)

En total se trabajó con 17 variables las cuales se relacionaron con los determinantes sociales establecidos por la OMS como de alto impacto para la determinación del riesgo de embarazos adolescentes. La tabla 29 presenta las variables que se tuvieron en cuenta y la asociación que se realizó con los determinantes sociales para ser incluidos en la generación del modelo predictivo.

Tabla 29. Asociación de variables con los determinantes sociales de la salud relacionados al riesgo de embarazo a temprana edad

Determinante Social	Factor	Variable OMS	Variable del programa "Salud a su Casa"
Determinantes próximos	Factores biológicos	Desarrollo de pubertad	
		Edad menarquia	

Determinante Social	Factor	Variable OMS	Variable del programa "Salud a su Casa"
	Factores de Comportamiento	Inicio de relaciones Sexuales	
		Nupcialidad y/ o Uniones tempranas	Estado Civil
		Uso de los métodos de anticoncepción	
		Acceso a servicios de salud	Condición SGSSS (Aseguramiento a la Salud)
		Acceso a servicios de educación	Estudiando
		***Creencias del individuo**	Etnia
Determinantes intermedios	Factores Interpersonales	Familia	
		Redes Sociales	
		Hogares con Jefatura femenina	
		Historia de embarazo adolescente en la Familia	Historia Embarazo Adolescente Madre
		Violencia Intrafamiliar	
		Abandono	Existen Personas Sin Cuidador
		Falta de Monitoreo	Mujeres Menores Sin Cuidador
		Comunicación con cuidadores	
		Aceptación y apoyo paternal	
		Lugar que se ocupa en la familia	Relación con el Jefe de Hogar
		Calidad en la educación y sexualidad	

Determinante Social	Factor	Variable OMS	Variable del programa "Salud a su Casa"
		Cohesión con pares	
	Factores Intrapersonales	Edad	
		Nivel educativo	Nivel Educativo
		Imaginarios del amor y la sexualidad	
		Manejo del tiempo libre	Oficios escolares Oficios del hogar Otras actividades
Determinantes Distales	Factores Contextuales	Normas sociales y de género	
		Valores de la sociedad	
	Factores Estructurales	**Nivel de pobreza**	Faltó Dinero para Comprar Alimentos Índice de Hacinamiento Nivel de Hacinamiento
		Inequidad social	Tipo de Vivienda Riesgos de Vivienda
		Relaciones de dominación y subordinación de género	

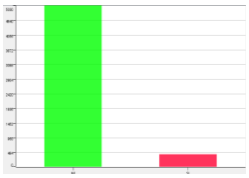
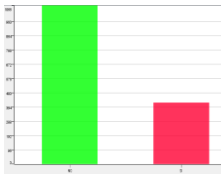
4.1 Pre-procesamiento

Gracias al análisis realizado a las variables durante el desarrollo del primer modelo predictivo, se realizó la depuración en la calidad del dato y se incluyeron únicamente las variables con mayor impacto para la construcción de este segundo modelo. En este caso el pre-procesamiento relacionado a la limpieza y transformación de datos se realizó directamente con las sentencias SQL que se desarrollaron para la generación del repositorio de datos (ver Anexo B).

4.1.1 Balanceo de datos

Como se describió anteriormente, por la naturaleza del conjunto de datos el 93% de los registros pertenecen a mujeres que no quedaron en embarazo. Esto repercute en que el modelo no puede generar reglas de asociación o sectores de clasificación dentro del espacio vectorial para los casos de embarazo adolescente al igual que en los trabajos realizados con el primer conjunto de datos. Por esta razón, se realiza un balanceo de la muestra, disminuyendo la diferencia porcentual entre los dos tipos de registros, la tabla 30 muestra el ajuste que se realiza durante la fase de entrenamiento.

Tabla 30. Resultado de entrenamiento de modelo clasificador con red neuronal PNN con técnicas de entrenamiento con el set de datos original y el set de datos balanceado

Cantidad	 SI NO Si: 413 No: 5.330 Total: 5.743	 . SI NO Si: 413 No: 1.066 Total: 1.479
Matriz de Confusión Modelo Red Neuronal PNN	**Predicción** SI NO **Clase** SI 0 41 NO 0 534 Exactitud: 0.929 Precisión: 0 Sensibilidad: 0	**Predicción** SI NO **Clase** SI 11 36 NO 4 97 Exactitud: 0.729 Precisión: 0.733 Sensibilidad: 0.234

Existen mujeres que pueden tener las variables socioeconómicas de aquellas que tuvieron un embarazo adolescente, pero como tan solo el 7% de los registros pertenece a mujeres que tuvieron algún embarazo antes de llegar a los 20 años pueden existir más registros de mujeres que bajo las mismas condiciones socioeconómicas no quedaron en embarazo. Esto genera que algunos algoritmos de entrenamiento no pueden identificar zonas dentro del espacio vectorial para las clasificaciones de tipo SI o no generan reglas para la clasificación asociada a la variable embarazo. Como se puede observar en la tabla 30 para un clasificador PNN no se generan reglas de clasificación para embarazos positivos con el conjunto de datos completo. Por ello se hace necesario realizar una disminución a la diferencia porcentual entre los valores de la clase para entrenar el modelo y una vez entrenado someterlo a pruebas de clasificación con el conjunto de datos completo. En la

tabla 30 también se puede observar que al realizar este desbalanceo de carga, los algoritmos pueden generar clasificaciones para los casos positivos de embarazo.

4.1.2 Transformación de datos

Al igual que el trabajo realizado con el primer conjunto de datos las variables tuvieron que ser transformadas para la aplicación de algunas técnicas específicas. En general estas modificaciones a los datos son necesarias por la naturaleza de los algoritmos de cada técnica. En particular solo se realizó una transformación diferente respecto al nivel educativo. El conjunto de datos contenía siete variables relacionadas con este determinante. El primero de tipo nominal que indica si la persona es analfabeta o no; para calcular esta variable con las variable ordinal, era necesario binarizar mediante la asignación de valores de cero y uno para sí y no respectivamente. Las demás variables indicaban el nivel más alto alcanzado durante las etapas de primaria, secundaria, técnica o tecnológica, estudios universitarios y de postgrado. Estas seis categorías son variables ordinales que han definido escalas de acuerdo con el nivel máximo que se puede alcanzar en cada uno de ellas. Con base en esto se propuso una ecuación que permitió generar un índice de nivel educativo. Esta ecuación está directamente relacionada con la edad como sigue:

$$EL = \frac{EY}{A} \Rightarrow LE = \frac{6 + \mathrm{P} + \mathrm{S} + 0.5\mathrm{Tec} + 0.5\mathrm{Tlg} + 0.5\mathrm{U}}{A} \quad (9)$$

donde *EL* significa el nivel educativo; *EY* los años de educación, A se refiere a la edad; P, S, Tec, Tlg y U son el nivel máximo en la escuela primaria, secundaria, estudios técnicos, estudios tecnológicos y los estudios universitarios respectivamente.

En consecuencia, se consideró empíricamente que, en promedio, un niño comienza la escuela primaria a la edad de seis años de edad. Se decidió no tener en cuenta los niveles de postgrado en la ecuación, ya que no se aplican a los rangos de edad de este estudio y generalmente pertenecen a valores fuera del rango considerado.

4.2 Minería de Datos

Para este caso se sometió el conjunto de datos a las técnicas de clasificación que presentaron mejores resultados durante la primera aproximación, esta vez se realizaron más experimentos con una mayor variedad de configuraciones a los parámetros de cada técnica buscando ajustar el modelo. Sin embargo, para disminuir los tiempos de procesamiento durante la etapa experimental se redujo la cantidad de datos haciendo selecciones aleatorias del repositorio de datos principal.

Por otra parte, para que el modelo pudiera generar reglas de asociación con respecto a la clase principal que permitiera realizar clasificaciones asociadas a el valor “SI” de la clase, se

desbalanceo el conjunto de datos dejando 72% de registros con embarazo positivo y 38% con embarazo negativo. Los experimentos y resultados se describen en esta sección.

4.2.1 Perceptron Multicapa Utilizando Rprop Como Algoritmo De Optimización.

Para este método en específico se realizaron transformaciones de datos de tipo texto a tipo entero. Ya que por la naturaleza del algoritmo de entrenamiento se crea un espacio vectorial de N dimensiones con vectores aleatorios, donde N pare este caso son las 17 variables que ingresan al modelo.

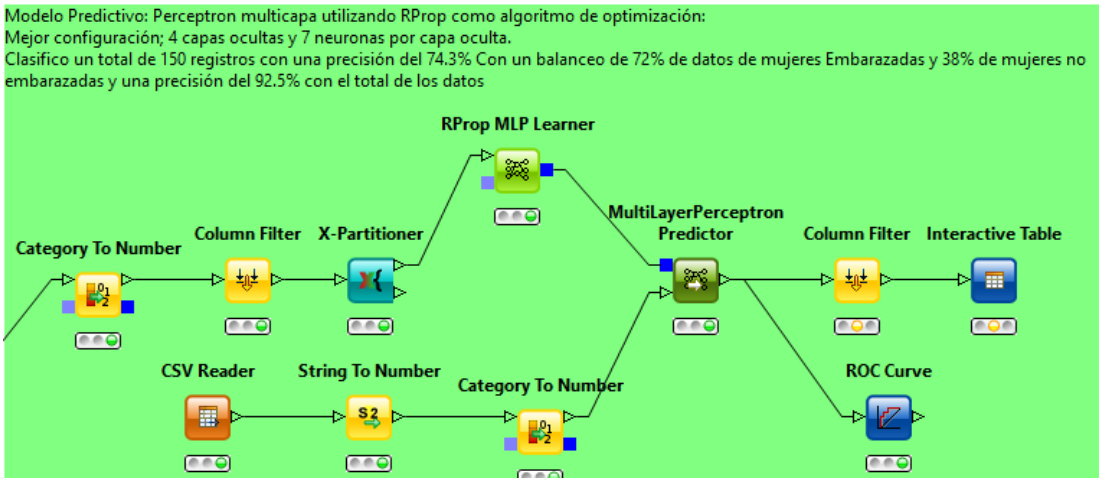
Para los casos donde las variables toman el valor SI y NO se remplazó por 1 y 0 respectivamente. Sin embargo otras transformaciones se realizan dependiendo su naturaleza. Por ejemplo, Estado Civil:

0: Soltero.
1: Unión Libre.
2: Casado.
3: Divorciado.
4: Viudo.

Otras variables cuya naturaleza no permite una distancia entre los valores ej. (Etnia cuyos valores son Ninguno, Afrocolombiano, Indígena y Rom/Gitano), fueron binarizadas con técnicas de frecuencia.

La figura 26 muestra el modelamiento de la red neuronal, el modelo fue entrenado y probado con la técnica validación cruzada utilizando el 90% de los datos para el entrenamiento y el restante 10% para su validación.

Figura 26. Modelo desarrollado para clasificadores con redes neuronales de perceptron multicapa entrenado y probado con métodos de 10 validación cruzada y algoritmo RProp como algoritmo de entrenamiento



La tabla 31 contiene las matrices de confusión y algunas métricas de evaluación del modelo como resultados de los experimentos realizados.

.

Tabla 31. Resultados modelo predictivo con red neuronal Rprop MPL. Número máximo de iteraciones: 100

CAPAS OCULTAS

NÚMERO DE NEURONAS POR CAPA OCULTA

3

		1		2		3		4	
		Predicción		Predicción		Predicción		Predicción	
		SI	NO	SI	NO	SI	NO	SI	NO
Clase	SI	8	35	10	33	11	32	10	33
	NO	5	100	7	98	9	96	9	96
Exactitud:		0.73		0.73		0.72		0.72	
Precisión:		0.61		0.59		0.55		0.53	
Sensibilidad:		0.19		0.23		0.26		0.23	

5

		Predicción		Predicción		Predicción		Predicción	
		SI	NO	SI	NO	SI	NO	SI	NO
Clase	SI	9	34	9	34	8	35	10	33
	NO	7	98	7	98	8	97	6	99
Exactitud:		0.72		0.72		0.71		0.74	

Precisión	0.56	Precisión: 0.56	Precisión: 0.50	Precisión: 0.65
Sensibilidad:	0.21	Sensibilidad: 0.21	Sensibilidad: 0.19	Sensibilidad: 0.23

7

		Predicción	
		SI	NO
Clase	SI	9	34
	NO	7	98

Exactitud: 0.72
Precisión: 0.56
Sensibilidad: 0.21

		Predicción	
		SI	NO
Clase	SI	8	35
	NO	6	99

Exactitud: 0.72
Precisión: 0.57
Sensibilidad: 0.19

		Predicción	
		SI	NO
Clase	SI	8	35
	NO	6	99

Exactitud: 0.72
Precisión: 0.57
Sensibilidad: 0.19

		Predicción	
		SI	NO
Clase	SI	9	34
	NO	4	101

Exactitud: 0.74
Precisión: 0.69
Sensibilidad: 0.21

Una vez se realizó esta aproximación se hizo la evaluación para determinar la mejor configuración del modelo predictivo implementado. Por la importancia de detectar los casos de verdaderos positivos que para el caso de estudio significan casos reales de embarazos adolescentes, las métricas a tener en cuenta son: la exactitud que le da un valor calificativo general al modelo en cuanto a la clasificación, la precisión que evalúa los verdaderos positivos con base en las predicciones positivas y la sensibilidad cuyo valor indica la predicción de casos positivos con base en los casos reales (ver ecuaciones 1, 4 y 5). Como se puede observar las redes neuronales de 4 capas ocultas con 5 y 7 neuronas por capa oculta presentan la mejor exactitud y precisión de clasificación. Sin embargo, sólo detectaron 10 y 9 registros respectivamente de las mujeres que en realidad tuvieron un embarazo durante su adolescencia. Por otra parte la red neuronal de 3 capas ocultas y 3 neuronas por capa presentó una mayor cantidad de verdaderos positivos y una mejor sensibilidad.

Se seleccionaron estos tres modelos como los mejores para el problema abordado y se probó cada modelo con el conjunto total de datos. Los resultados se describen en la tabla 32 donde las mejores evaluaciones al modelo fueron para los médelos de 4 capas ocultas con 7 neuronas por capa oculta en cuanto a la sensibilidad y el de 3 capas ocultas con 4 neuronas por capa en precisión y exactitud.

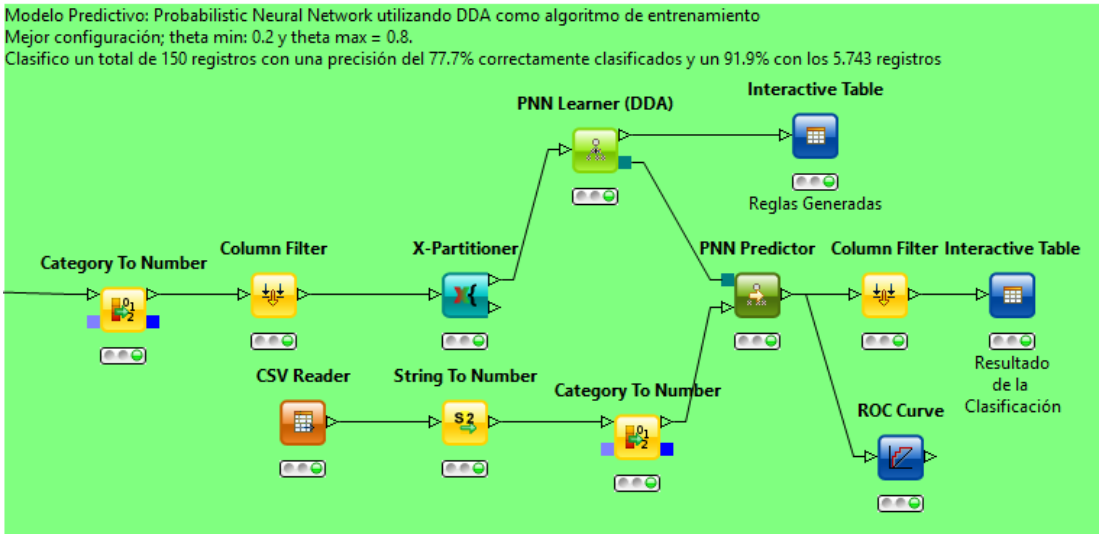
Tabla 32. Resultado de la clasificación del conjunto de datos con redes neuronales RProp MPL

Capas ocultas: 4 **Neuronas por capa oculta: 7**	**Capas ocultas: 4** **Neuronas por capa oculta: 5**	**Capas ocultas: 3** **Neuronas por capa oculta: 3**
Predicción / SI / NO **Clase** SI: 145 / 268	**Predicción** / SI / NO **Clase** SI: 140 / 273 NO: 371 / 4959	**Predicción** / SI / NO **Clase** SI: 123 / 290 NO: 261 / 5069

	NO	401	4929				
			0.88	Exactitud:	0.89	Exactitud:	0.90
		Exactitud:	0.2	Precisión:	0.27	Precisión:	0.32
		Precisión:	0.3	Sensibilidad	0.34	Sensibilidad:	0.30
		Sensibilidad:					

4.2.2 Redes Neuronales Probabilísticas entrenada con algoritmo DDA

Figura 27. Modelo desarrollado para clasificadores con redes neuronales PNN entrenado y probado con métodos de validación cruzada y algoritmo DDA como algoritmo de entrenamiento.



Como se describió en la aplicación de esta técnica cuando se realizaron los modelos para el primer conjunto de datos, el algoritmo de entrenamiento DDA utiliza dos parámetros conocidos como ϕ Mínimo y ϕ Máximo que impiden que las reglas generadas entren en conflicto. Esta vez el conjunto de datos fue sometido a experimentos variando estos dos parámetros y realizando máximo 100 iteraciones antes de terminar con el entrenamiento, los resultados de las pruebas se describen en la tabla 33 y el modelo desarrollado se muestra en la figura 27.

Tabla 33. Resultados modelo predictivo con red neuronal PNN.

		Φ Max							
		0.4		0.6		0.8		1.0	
Φ Min	0.2	Predicción		Predicción		Predicción		Predicción	
		SI	NO	SI	NO	SI	NO	SI	NO

Clase	SI	12	35
	NO	6	95

Exactitud: 0.72
Precisión: 0.66
Sensibilidad: 0.25

Clase	SI	12	35
	NO	5	96

Exactitud: 0.73
Precisión: 0.71
Sensibilidad: 0.25

Clase	SI	12	35
	NO	5	96

Exactitud: 0.73
Precisión: 0.71
Sensibilidad: 0.25

Clase	SI	47	0
	NO	100	1

Exactitud: 0.32
Precisión: 0.32
Sensibilidad: 1.00

0.4

		Predicción	
		SI	NO
Clase	SI	9	38
	NO	4	97

Exactitud: 0.72
Precisión: 0.69
Sensibilidad: 0.19

		Predicción	
		SI	NO
Clase	SI	11	36
	NO	4	97

Exactitud: 0.73
Precisión: 0.73
Sensibilidad: 0.23

		Predicción	
		SI	NO
Clase	SI	11	36
	NO	4	97

Exactitud: 0.73
Precisión: 0.73
Sensibilidad: 0.23

		Predicción	
		SI	NO
Clase	SI	47	0
	NO	99	2

Exactitud: 0.33
Precisión: 0.32
Sensibilidad: 1.00

0.6

		Predicción	
		SI	NO
Clase	SI	3	44
	NO	1	100

Exactitud: 0.70
Precisión: 0.75
Sensibilidad: 0.06

		Predicción	
		SI	NO
Clase	SI	10	37
	NO	4	97

Exactitud: 0.72
Precisión: 0.66
Sensibilidad: 0.21

		Predicción	
		SI	NO
Clase	SI	46	1
	NO	94	7

Exactitud: 0.36
Precisión: 0.32
Sensibilidad: 0.98

0.8

		Predicción	
		SI	NO
Clase	SI	0	47
	NO	0	101

Exactitud: 0.68
Precisión: 0
Sensibilidad: -

		Predicción	
		SI	NO
Clase	SI	42	5
	NO	65	36

Exactitud: 0.53
Precisión: 0.39
Sensibilidad: 0.89

Al igual que se realizó con el modelo de redes neuronales de Perceptron Multicapa, se tomaron las mejores configuraciones halladas para el modelo de red neuronal PNN y se sometió a la clasificación del conjunto total de datos. Para este caso se realizó con las configuraciones ϕMín=0.2 y ϕMáx = 0.6 la cual logró mejores resultados de exactitud y sensibilidad y ϕMín=0.4 y ϕMáx = 0.6 cuyos resultados de precisión fueron los mejores. Los resultados obtenidos con ϕMáx = 0.8 para estas dos configuraciones obtuvieron los mismos resultados que la configuración con ϕMáx = 0.6 como se puede observar en la tabla 33 y el tiempo de procesamiento es mayor para estos últimos; por esta razón no se sometió el

conjunto de datos a clasificaciones con estas configuraciones. Los resultados obtenidos se describen en la tabla 34 donde se evidencia que se obtuvieron mejores resultados con la primera configuración para todas las métricas de desempeño.

Tabla 34. Resultados de clasificación del conjunto de datos con los modelos de red neuronal PNN

ϕMín=0.2 y ϕMáx = 0.6

		Predicción	
		SI	NO
Clase	SI	145	268
	NO	192	5138

Exactitud:	0.92
Precisión:	0.43
Sensibilidad:	0.35

ϕMín=0.4 y ϕMáx = 0.6

		Predicción	
		SI	NO
Clase	SI	77	336
	NO	335	4995

Exactitud:	0.8832
Precisión:	0.187
Sensibilidad	0.186

Como se describió en la sección 2.4.1, a diferencia de otro tipo de redes neuronales estas tienen la particularidad de generar las reglas que utiliza para la clasificación gracias al algoritmo utilizado para su entrenamiento. En total género 751 reglas para la clasificación NO y 336 reglas para la clasificación SI, a cada una de estas les asigna un peso según la importancia de cada una para asignar la clasificación. La tabla 35 muestra las reglas con mayor peso generadas por el modelo. De aquí se pueden validar las apreciaciones realizadas por la Organización Mundial de la Salud, al describir como factores de alto riesgo las variables de jefatura femenina en la familia (valor 1), el nivel de pobreza (representado por el índice de hacinamiento) y el nivel educativo.

Tabla 35. Las 8 reglas generadas por el modelo PNN con mayor peso para la clasificación

GeneroJefeHogar	IndiceHacinamiento	NivelEducativo	Embarazo	Weight
0	1.3	0.573	NO	9
0	1.7	0.534	NO	7
0	2	0.577	NO	6
0	1.7	0.546	NO	6
0	1.3	0.496	NO	5
0	2.5	0.526	NO	5
0	2.5	0.497	NO	5
0	2	0.588	NO	5
1	7	0.488	SI	5

4.2.3 Naive Bayes

El tercer método que obtuvo buenos resultados durante los trabajos realizados con el primer conjunto de datos fueron los modelos predictivos de Naive Bayes. Para esta segunda aproximación se sometió el modelo a varias pruebas modificando la cantidad de atributos ingresados al modelo dependiendo de la cantidad máxima de valores nominales permitidas por atributo. Aquellos que superen esa cantidad máxima no son tenidos en cuenta para la generación del modelo. El modelo genera distribuciones gaussianas para los atributos cuyos valores son continuos como el índice de hacinamiento y nivel educativo. La tabla 36 muestra los resultados obtenidos con la aplicación de este método.

Como se puede observar en la tabla, mientras se iba incrementando el número de variables que se tienen en cuenta para el entrenamiento del modelo, se mejoraba el número de datos bien clasificados y la capacidad del modelo para predecir las adolescentes con embarazo adolescente.

Finalmente, se tomó el modelo que incluyó todas las variables para la etapa de entrenamiento y se sometió al conjunto de datos a la clasificación, obteniendo muy buenos resultados en cuanto a los TP o mujeres con embarazo adolescente bien clasificadas; sin embargo, la precisión del modelo disminuyó.

Figura 28. Modelo desarrollado para clasificadores con Naive Bayes entrenado y probado con métodos de validación cruzada

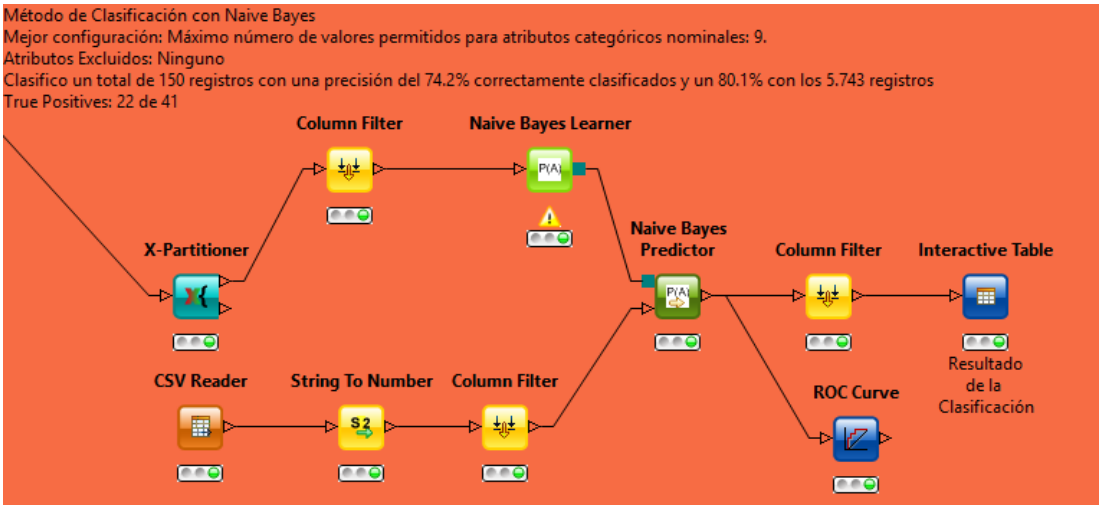


Tabla 36. Resultados modelo predictivo con Naive Bayes. Máximo número de valores nominales permitidos por variable: (2, 3, 5, 7 y 9)

2	3	5
Predicción	Predicción	Predicción

		SI	NO
Clase	SI	17	24
	NO	21	86

Exactitud: 0.70
Precisión: 0.45
Sensibilidad: 0.31

Atributos excluidos por el modelo:

- Tipo Vivienda
- Condición SGSSS
- Relación con el Jefe de Hogar
- Riesgo de la Vivienda
- Estado Civil
- Etnia
- Hacinamiento

		SI	NO
Clase	SI	17	24
	NO	24	83

Exactitud: 0.68
Precisión: 0.41
Sensibilidad: 0.41

Atributos excluidos por el modelo:

- Tipo Vivienda
- Condición SGSSS
- Relación con el Jefe de Hogar
- Etnia

		SI	NO
Clase	SI	19	22
	NO	22	85

Exactitud: 0.70
Precisión: 0.46
Sensibilidad: 0.46

Atributos excluidos por el modelo:

- Tipo Vivienda
- Relación con el Jefe de Hogar

7

9

Todo el conjunto de Datos

		Predicción	
		SI	NO
Clase	SI	20	21
	NO	18	89

Exactitud: 0.74
Precisión: 0.53
Sensibilidad: 0.49

Atributos excluidos por el modelo:

- Relación con el Jefe de Hogar

		Predicción	
		SI	NO
Clase	SI	22	19
	NO	19	88

Exactitud: 0.74
Precisión: 0.54
Sensibilidad: 0.54

Ningún atributo excluido

		Predicción	
		SI	NO
Clase	SI	179	234
	NO	907	4423

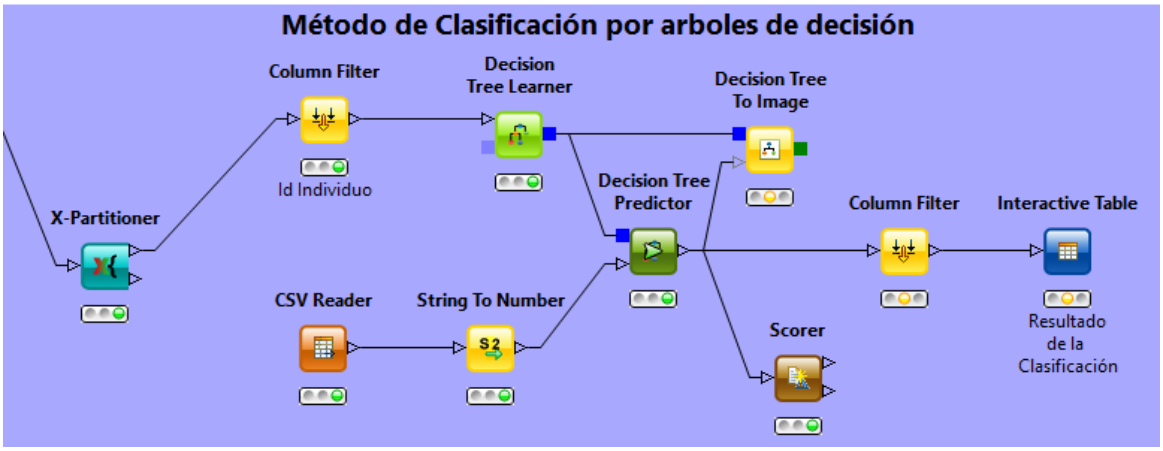
Exactitud: 0.80
Precisión: 0.16
Sensibilidad: 0.43

Máximo número de valores nominales permitidos por variable: 9

4.2.4 Árboles de decisión

La última técnica que presentó buenos resultados durante los trabajos realizados con el primer conjunto de datos fueron los árboles de decisión; esta técnica produjo buenos resultados en cuanto al análisis de la problemática y la generación de reglas de clasificación. Para la segunda aproximación se realizaron nuevamente pruebas a este modelo y los resultados se describen a continuación.

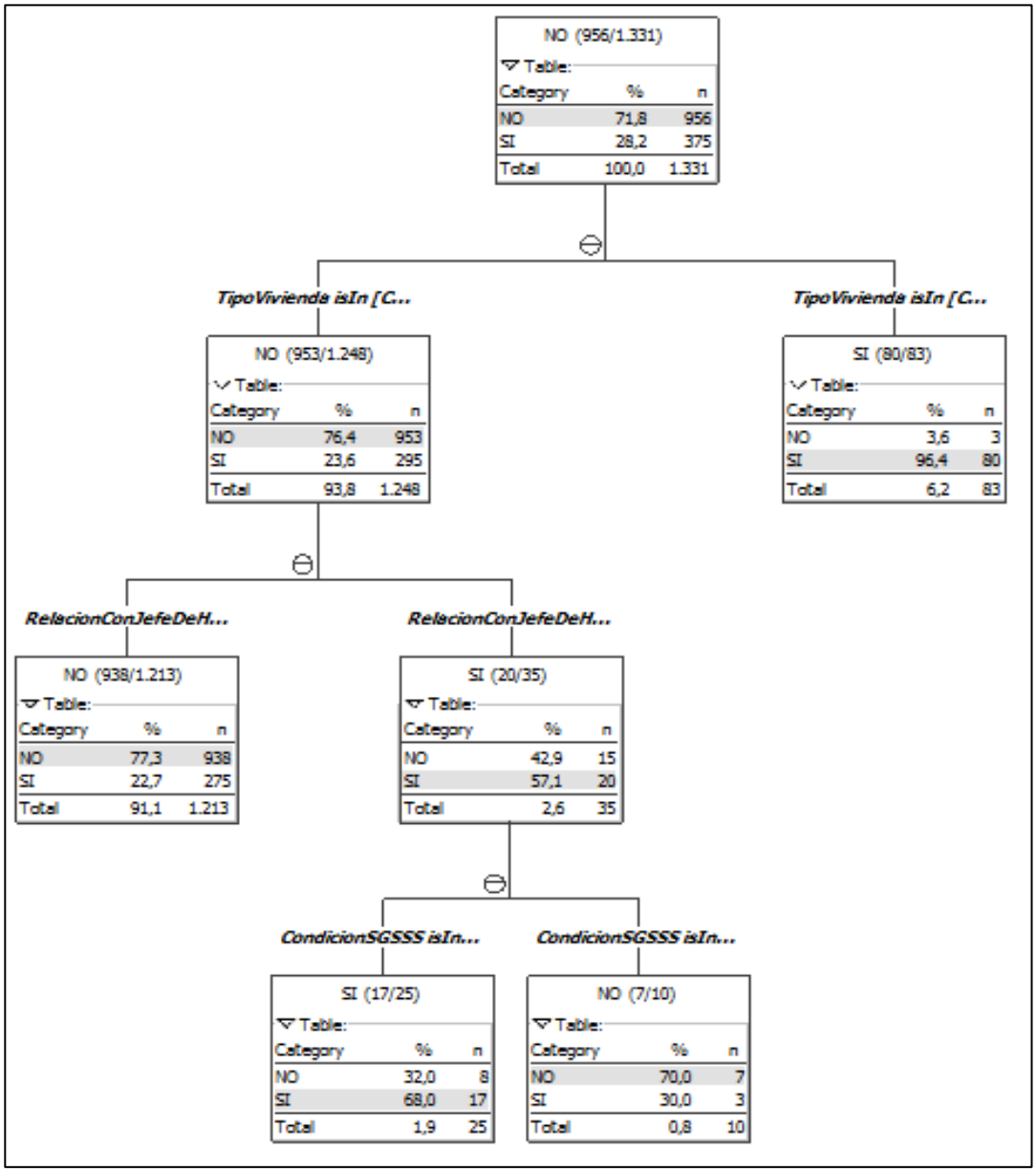
Figura 29. Modelo desarrollado para clasificador con Árboles de decisión entrenado y probado con métodos de validación cruzada



Configuración de parámetros del modelo:

- Columna Clase: Embarazo
- Medida de Calidad: Relación de ganancia (*Gain Ratio*)
- Número mínimo de registros por nodo: 10
- Evitar columnas nominales sin valores: SI
- División binaria de valores nominales: SI
- Máximo número de valores para atributos nominales para ser agrupados: 8 en un nodo

Figura 30. Árbol de decisión obtenido por el modelo para el segundo set de datos balanceado



Reglas inferidas del árbol de decisión:

a. Tipo de Vivienda = (Móvil o Refugio Natural o Carpa) => SI
b. Tipo de Vivienda = (Casa o Apartamento) & Relación con Jefe de Hogar = (Hijastro o Hijo o Nieto o Hermano o No Pariente) => NO
c. Tipo de Vivienda = (Casa o Apartamento) & Relación con Jefe de Hogar = (Otro Pariente o Yerno-Nuera o Jefe de Hogar o Conyugue) & Condición SGSS = (Contributivo o Subsidiado) => SI

d. Tipo de Vivienda = (Casa o Apartamento) & Relación con Jefe de Hogar = (Otro Pariente o Yerno-Nuera o Jefe de Hogar o Conyugue) & Condición SGSS = (No Asegurado o Régimen Especial) => NO

Al igual que con los otros métodos de clasificación, una vez realizado el entrenamiento con el conjunto de datos balanceado, se probó el modelo ingresando todos los registros del conjunto de datos y los resultados se describen con su respectiva matriz de confusión.

Tabla 37. Resultados de clasificación del conjunto de datos con el modelo de desarrollado con árboles de decisión

		Predicción	
		SI	NO
Clase	SI	103	310
	NO	110	5209

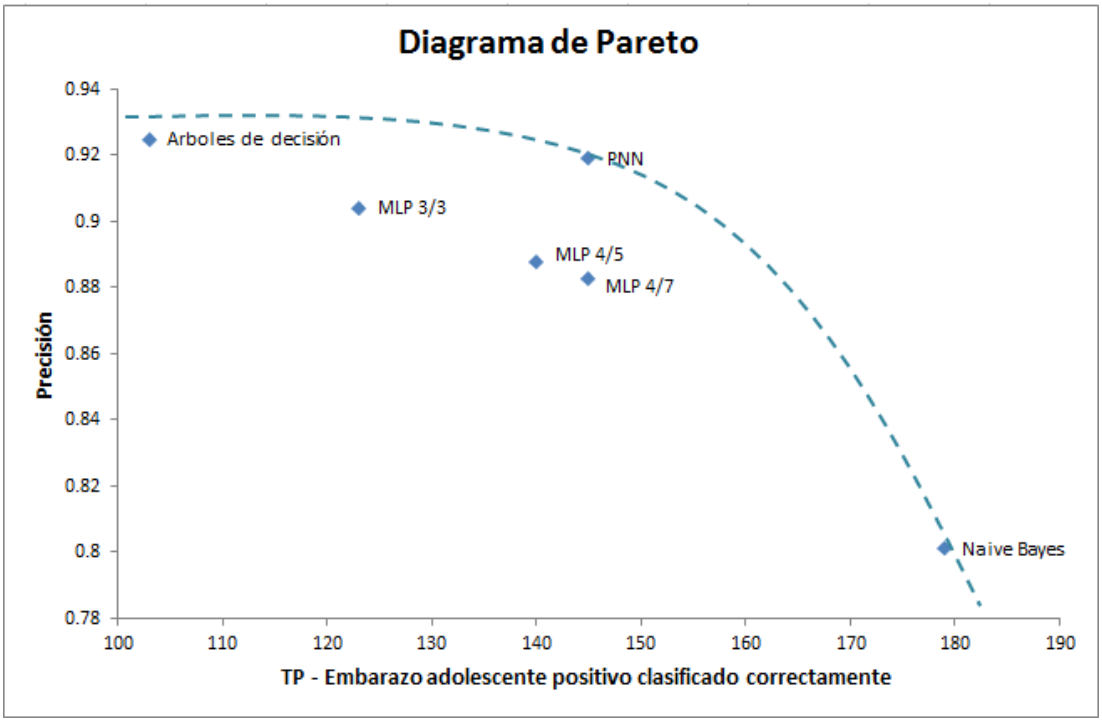
Exactitud:	0.927
Precisión:	0.484
Sensibilidad:	0.249

4.3 Conclusiones segunda aproximación

Todos los modelos mejoraron su rendimiento en cuanto a la clasificación durante la segunda aproximación. Los mejores clasificadores durante la primera aproximación oscilaban en valores entre el 65% y el 70.5% de precisión y exactitud con pocos registros de verdaderos positivos clasificados. El cambio del conjunto de datos produjo mejoras al modelo permitiendo desarrollar modelos que presentaron hasta 74% de exactitud y un número mayor de verdaderos positivos. Esto debido a la preclasificación de las variables de mayor impacto encontradas durante el primer acercamiento, las mejoras en la calidad de los datos basadas en depuraciones y transformaciones realizadas desde las sentencias SQL que generaron el conjunto de datos, y el hecho de contar con los registros que contienen el historial de toda la adolescencia y determinar si hubo un embarazo o no durante ella.

Debido a que el interés principal de desarrollar un modelo predictivo es la identificación de las adolescentes con alto riesgo de quedar en embarazo, la cantidad de registros de embarazo clasificados correctamente (True Positive) resultan de alto interés. Por lo tanto para el caso de estudio, la elección del mejor modelo resulta del análisis conjunto sobre las métricas de rendimiento. Para la elección de los mejores modelos se consideró desarrollar un análisis con un diagrama de Pareto teniendo en cuenta estas dos los datos correctamente clasificados y la exactitud del modelo. La gráfica 31 es la representación del diagrama de Pareto para los resultados obtenidos teniendo en cuenta estos dos aspectos, teniendo en cuenta que se trata de un problema de maximización, ya que se quiere tener una mayor cantidad de registros bien clasificados pero a la vez un número alto de verdaderos positivos del embarazo adolescente, los mejores modelos para el caso de estudio son los clasificadores con los árboles de decisión, Naive Bayes y las redes neuronales probabilísticas PNN.

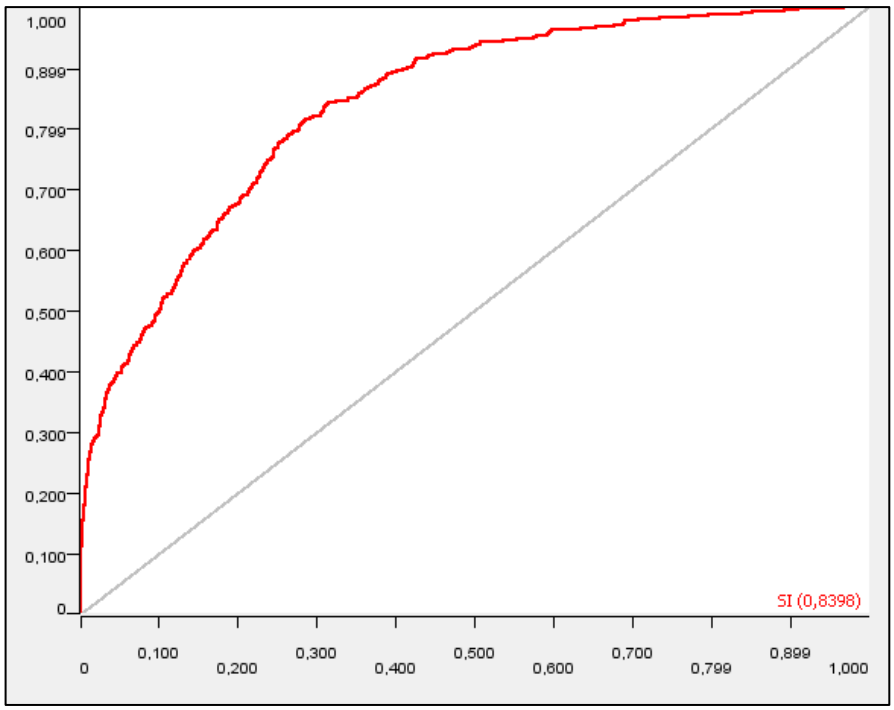
Figura 31. Diagrama de Pareto para los modelos predictivos desarrollados



Con las técnicas implementadas se lograron generar reglas de asociación de la problemática del embarazo adolescente en Bogotá con las variables socioeconómicas como el nivel de pobreza, el nivel educativo, la conformación de la familia entre otras. De este modo se pudo ratificar la teoría de los determinantes sociales en la salud de la OMS relacionados con el riesgo de embarazos adolescentes.

Se puede concluir que para realizar una clasificación del riesgo de embarazo adolescente el mejor modelo predictivo fue el desarrollado con redes neuronales probabilísticas (PNN) utilizando DDA como algoritmo de entrenamiento y configurando sus parámetros en ϕMín=0.2 y ϕMáx = 0.6.

Figura 32. Curva ROC para el clasificador PNN con algoritmo de entrenamiento DDA



La figura 32 ilustra la curva ROC obtenida por el modelo después de la clasificación del conjunto de datos. El área bajo la curva obtenida fue de 0.84, lo cual confirma el buen rendimiento de este clasificador.

5 Aplicación móvil desarrollada

5.1 Descripción general de la aplicación

Con la definición del mejor modelo predictivo para la detección de adolescentes con alto riesgo de quedar en embarazo, se realizó la etapa final de este proyecto. Se tomaron los registros del sistema de Información del programa "Salud a su Casa" de la Secretaria Distrital de Salud. Se filtró el conjunto de datos dejando únicamente aquellas menores de edad que no han tenido embarazos. Este conjunto de datos fue sometido al proceso de clasificación de con el modelo de red neuronal PNN definido como el mejor modelo en el capítulo 4. La clasificación realizada dejo un total de 20.432 adolescentes etiquetadas con alto riesgo. Con este conjunto de datos se extrajeron los datos de ubicación de vivienda y los datos personales para el desarrollo de la aplicación. Cabe resaltar que esta información es muy sensible y no puede ser divulgada abiertamente por las leyes de protección de los datos personales de los individuos. Como se describió en la definición de los objetivos del proyecto la aplicación desarrollada no es para uso comercial ni debe ser puesta en una tienda de aplicaciones, en caso de ser implementada debe ser utilizada únicamente por los equipos médicos que trabajan en campañas de promoción y prevención de la SDS.

La figura 33 es una representación gráfica de la app desarrollada, esta aplicación móvil básicamente tiene tres funcionalidades:

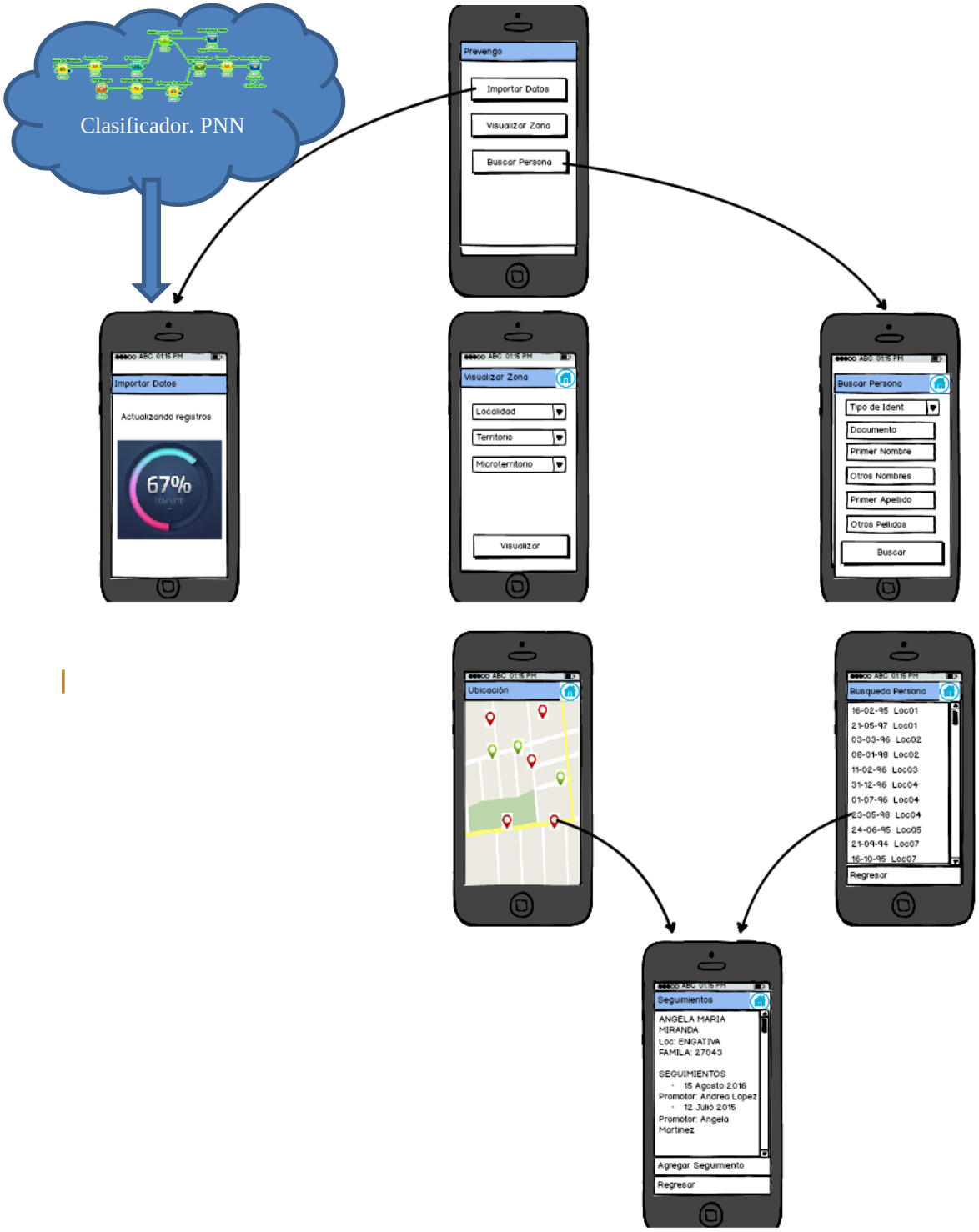
1. Sincronización: Descarga de un servicio web los datos de las adolescentes clasificadas como de alto riesgo.

2. Visualización geográfica: Permite encontrar los datos básicos obtenidos durante la caracterización al programa "Salud a su Casa", en donde se encuentran los datos para su geolocalización. Mediante un mapa se hace la ubicación del hogar de la joven.

3. Seguimiento en promoción: Permite establecer fechas y actividades realizadas para la educación en aspectos de prevención de embarazos adolescentes, que de ser

implementado serian de ayuda para los profesionales médicos en cuanto al estado de los seguimientos.

Figura 33. Ilustración general de la aplicación móvil desarrollada (prevengo.app)

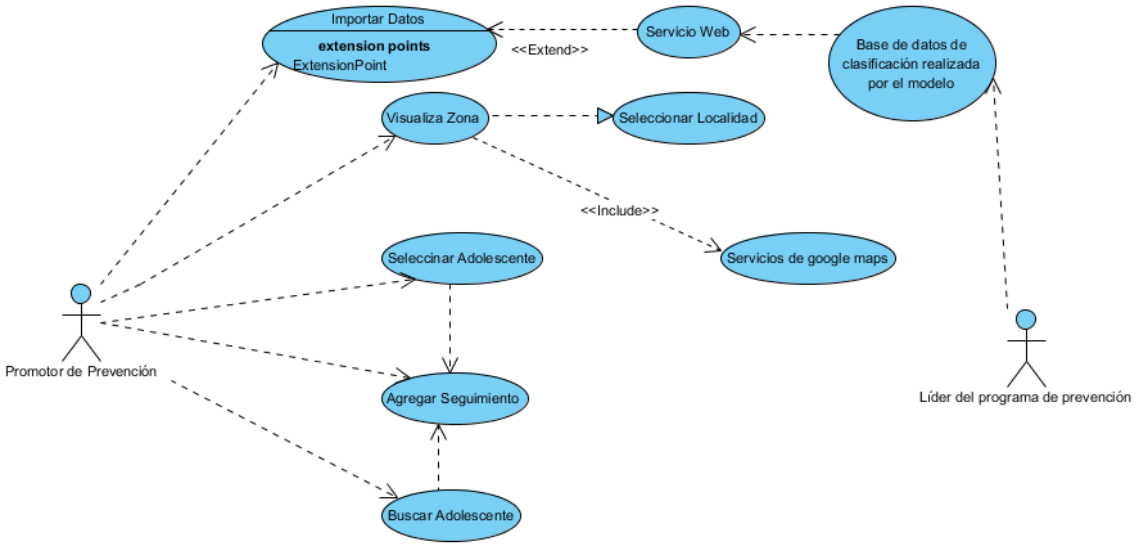


5.2 Arquitectura de la aplicación

En esta sección se hace una breve descripción del funcionamiento de la aplicación desde la arquitectura, para ello se hace uso de algunos elementos del modelo de vistas de arquitectura UML 4+1.

5.2.1 Casos de Uso

Figura 344. Diagrama de casos de uso



5.2.2 Vista Lógica

Diagrama de clases

En esta sección se muestra el diagrama de clases del aplicativo móvil. Se puede diferenciar tres paquetes principales: el paquete del modelo contiene todas las clases referentes a los objetos principales. Existe un paquete cuyas clases se encargan de hacer todas las transacciones con la base de datos SQLite e interactúa con el modelo llamado controlDataBase. Finalmente existe un paquete que contiene l capa encargada del manejo de la visualización de los datos y de los puntos geográficos.

Figura 35. Diagrama de clases

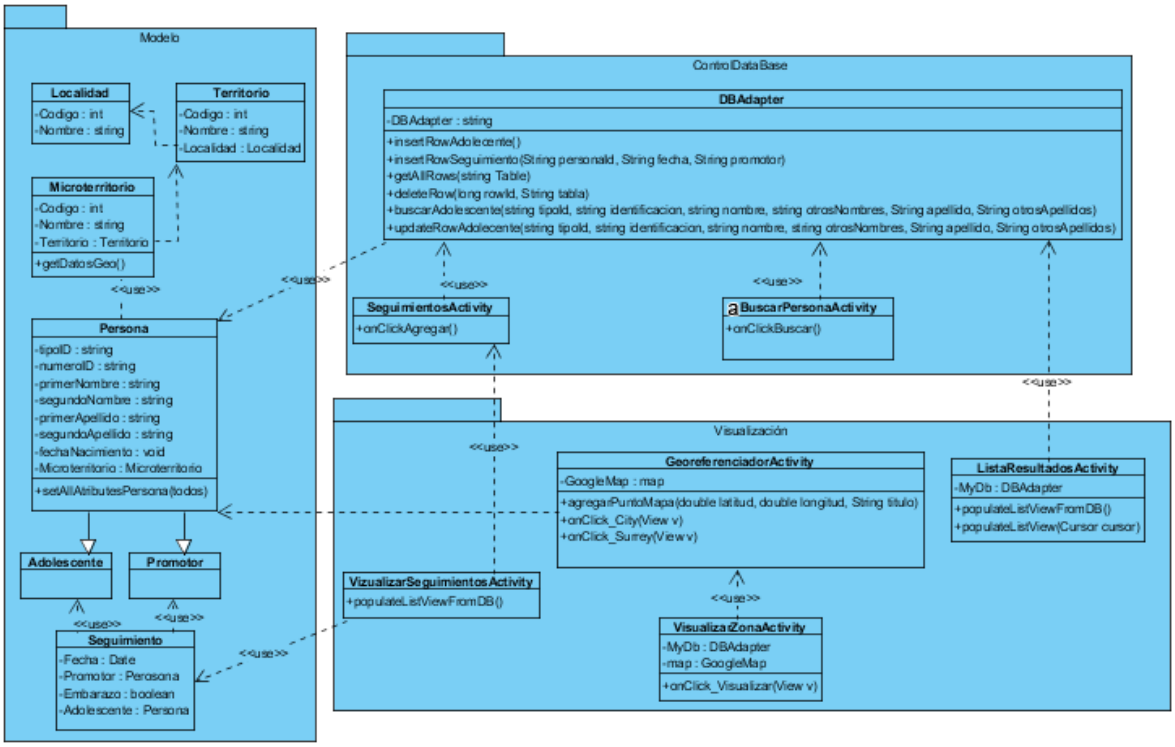
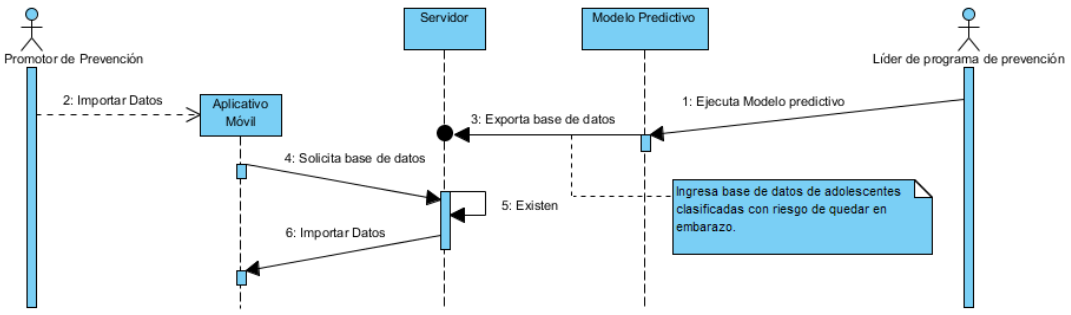


Diagrama de secuencias

El diagrama de secuencia muestra la interacción entre los diferentes partes del sistema con respecto al proceso de la entrega de la base de datos de las adolescentes clasificadas por el modelo con alto riesgo de quedar en embarazo.

Figura 36. Diagrama de secuencias



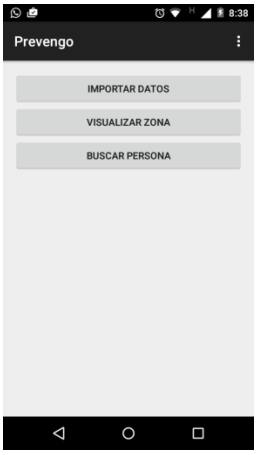
El líder del programa de prevención sería el encargado de ejecutar el clasificador con la base de datos de todas las adolescentes del sistema "Salud a su casa", el modelo predictivo entrega al servidor únicamente el listado de las adolescentes que son clasificadas con mayor riesgo de quedar en embarazo. La otra parte la realiza el promotor de prevención desde la aplicación cliente que para este caso es el aplicativo móvil, en la opción importar datos la cual se sincroniza con el servicio web y descarga la base de datos generada por el modelo al dispositivo móvil.

5.2.3 Vista Física

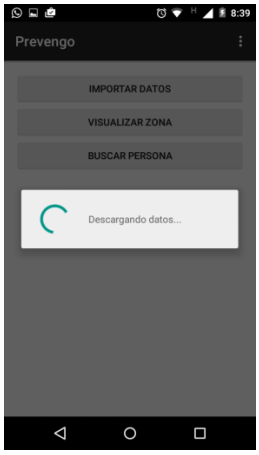
Imágenes del aplicativo

Algunas de imágenes del aplicativo desarrollado se describen a continuación, tal y como se desarrolló durante la etapa de diseño gráfico, se trata de una aplicación básica desarrollada en java para la ubicación de las adolescentes clasificadas por el modelo como aquellas con alto riesgo de quedar en embarazo.

Figura 37. Vistas graficas del aplicativo desarrollado (prevengo.app)



Menú principal: Permite la importación de datos del servidor, la visualización de una zona o hacer una búsqueda de una adolescente específica.

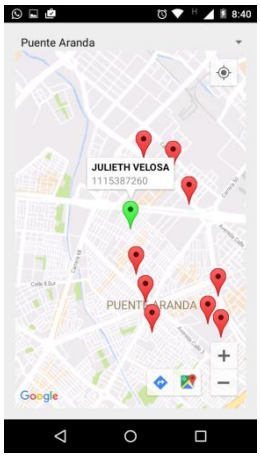


Importar datos: Al seleccionar este botón se realiza la sincronización con el servidor para descargar la base de datos de las adolescentes seleccionadas por el modelo predictivo,

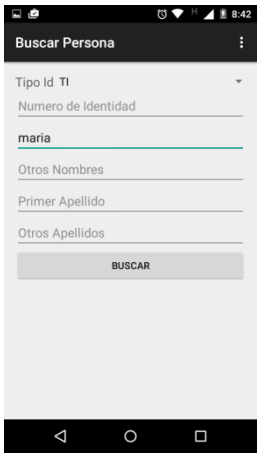
Visualizar Zona: En el mapa se ubican las adolescentes de la localidad seleccionada, Cada adolescente clasificada por el modelo representa un punto en el mapa.

Al seleccionar un punto en el mapa, este muestra el nombre de la adolescente que vive en esa ubicación. Para agregar un seguimiento a la adolescente se debe picar en el nombre.

Para marcar que se le ha realizado el seguimiento a la adolescente se marca agregar seguimiento y se escoge la opción si se encontró un evento nuevo de embarazo.



Sin importar cual fuese el evento, las adolescentes que han recibido seguimiento se identifican con un nuevo color dentro del mapa.



La opción de buscar persona, permite hacer búsquedas por identificación o por nombre.

Una vez realizada la búsqueda se muestra un listado de las adolescentes encontradas en la base de datos según los parámetros ingresados.

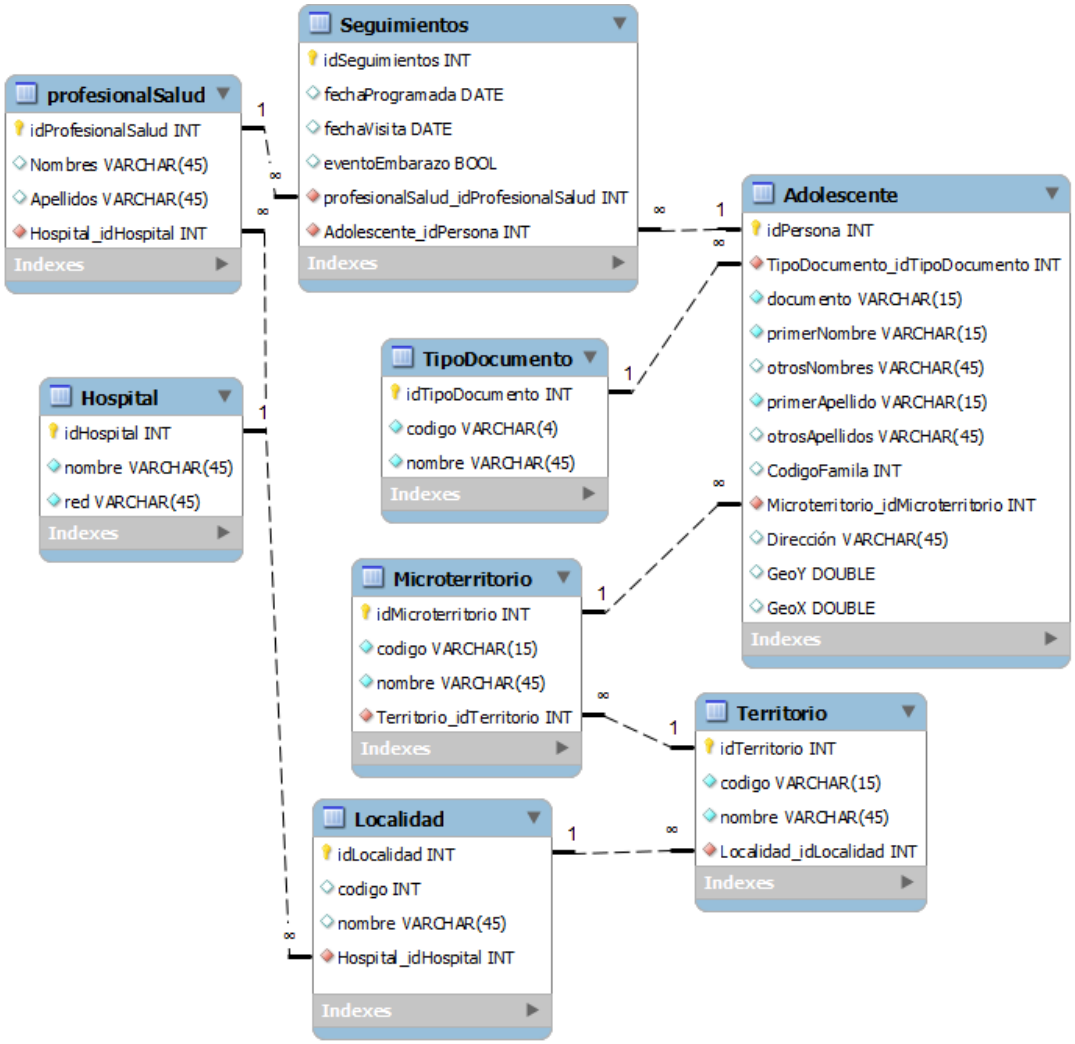
5.2.4 Modelo Entidad Relación

La aplicación móvil cuenta con una base de datos desarrollada en SQLite, la mayoría de los datos ingresados a esta base de datos se lleva a cabo durante la etapa de sincronización con el servidor. Los datos ingresados por la aplicación móvil solo corresponden a las fechas y a los resultados del seguimiento que se realicen a las adolescentes.

El diagrama entidad relación ayuda a modelar el componente general de datos de un sistema. Para la aplicación móvil se desarrollaron ocho entidades. Existe una entidad para los seguimientos que se realizan a una adolescente detectada por el modelo con alto riesgo de quedar en embarazo, esta entidad está relacionada con la adolescente a quien se realiza el seguimiento, guarda la fecha en que se realiza el seguimiento, el profesional de la salud que realiza la visita y el evento de embarazo en caso de que exista durante la etapa de seguimiento. Se crearon otras entidades referentes al funcionamiento del programa “Salud a su casa” tales como Hospital, Localidad, Territorio y Microterritorio.

Los datos de la adolescente incluyen los componentes de identificación y de ubicación con las coordenadas para ser georreferenciadas.

Figura 38. Modelo Entidad Relación (prevengo.app)



6 Conclusiones y recomendaciones

6.1 Conclusiones

El descubrimiento de conocimiento es un proceso que busca encontrar información hasta el momento desconocida, nuevas reglas que permitan determinar características que influyen en alguna situación. Este proceso se realiza a conjuntos de datos como insumo principal y conlleva ciertos pasos para la búsqueda de patrones relacionales ocultos entre las variables. Aplicando este proceso a los conjuntos de datos se pudo establecer nuevo conocimiento respecto a la problemática de embarazo adolescente en la población más vulnerable de Bogotá. Se pudo determinar que existen factores de alto riesgo como la conformación de la familia y la posición que una joven tiene dentro de ella. Se logró verificar que la mayoría de factores de alto riesgo determinados por la OMS influyen en el riesgo de las adolescentes de quedar en embarazo. Las uniones maritales tempranas, el nivel educativo y el ambiente familiar son factores con alta influencia. Se pudo determinar que las adolescentes cuyo jefe de hogar está representado por un hermano, otro pariente o alguien que no es pariente tienen un factor de riesgo más alto que aquellas donde el jefe de hogar son los padres, los abuelos o un padrastro o madrastra.

También se pudo concluir que la jefatura de hogar femenina no representa un factor de riesgo para la población objeto de este estudio. Esta situación en particular es la única que no obedece a lo establecido por los determinantes sociales de la OMS, donde se indica que las jefaturas de hogar representadas por el sexo femenino son un factor de mayor riesgo de embarazos adolescentes.

En cuanto al proceso conocido como minería de datos, se pudo concluir que existen diferentes métodos para la construcción de modelos predictivos los cuales pueden presentar diferente rendimiento conforme al conjunto de datos.

Cuando se evalúa un modelo predictivo se deben considerar varias métricas para medir el desempeño que este tiene, especialmente en conjuntos de datos donde una clase se presenta con mayor frecuencia que otras. Particularmente para el caso de estudio el 92.5% de los casos correspondían a la clase NO embarazo. Los modelos entonces deben ser precisos en la generación de reglas tratando de identificar el mayor número de casos positivos reales, pero esto puede influir en la generación de falsos positivos. El modelo escogido debe tener un punto de equilibrio donde se busque el mejor rendimiento de clasificación basado en las métricas consideradas.

6.2 Recomendaciones

Este tipo de estudios puede ser mejorado con un conjunto de datos que contengan más variables relacionadas con los determinantes sociales definidos por la OMS asociados al embarazo adolescente y que no pudieron ser tenidos en cuenta por ausencia de estas variables en el Sistema de Información “Salud a su Casa”. Nuevas aproximaciones podrían ser desarrolladas para mejorar el rendimiento del modelo si se tienen en cuenta variables de alto impacto relacionadas con la vida sexual de los individuos tales como si ya empezó una vida sexual, el uso de preservativos, entre otros.

Por otra parte, el estudio puede ser mejorado incluyendo información de todos los sectores sociales y económicos de la ciudad, para tener una visión más ampliada de la situación de embarazo adolescente en la ciudad.

Existen diferentes técnicas que no fueron aplicadas en este estudio y que podrían explorarse en búsqueda de algún nuevo descubrimiento o la mejora del modelo predictivo tales como máquinas de soporte vectorial, técnicas de agrupaciones jerárquicas y métodos basados en lógica difusa, entre otros.

A. Anexo: Variables de los conjuntos de datos

Características de las variables de la primera aproximación

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
Embarazo	Categórica - Discreta	**Variable Clase - No Aplica**		No	198428
				Si	11550
CondicionSGSSS	Categórica - Discreta	Acceso a servicios de salud	proximal	Contributivo	96255
				No Asegurado Identificado	26176
				No Asegurado No Identificado	13700
				Régimen Especial	2797
				Subsidiado	71050
TipoAfiliado	Categórica - Discreta	Acceso a servicios de salud	proximal	Adicional	127
				Beneficiario	63806
				Cotizante	2121
				NULL	143924
LocalidadIPS	Categórica - Discreta	Acceso a servicios de salud	proximal	ANTONIO NARI¥O	3758
				BARRIOS UNIDOS	1140
				BOSA	15114
				CANDELARIA	945

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
				CHAPINERO	6512
				CIUDAD BOLIVAR	15572
				ENGATIVA	8657
				FONTIBON	5376
				KENNEDY	18307
				MARTIRES	1748
				PUENTE ARANDA	4018
				RAFAEL URIBE	10376
				SAN CRISTOBAL	11104
				SANTA FE	2484
				SIN DATO	60862
				SUBA	19022
				SUMAPAZ	260
				TEUSAQUILLO	3022
				TUNJUELITO	5874
				USAQUEN	6534
				USME	9293
IPS	Categórica - Discreta	Acceso a servicios de salud	proximal	El nombre de la IPS (En total 901 IPS)	
TiempoDesplaza	Numérica -	Acceso a servicios de	proximal	Tiempo en minutos (Entre 1	

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
mientoIPS	Continua	salud		y 400)	
GastoDesplazamientoIPS	Numérica - Continua	Acceso a servicios de salud	proximal	Valor en pesos (Entre 0 y 30.000)	
AcudioMedicoOdontologo	Categórica - Discreta	Acceso a servicios de salud	proximal	No	116944
				Si	11135
				NULL	81899
AcudioPromotorEnfermero	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119681
				Si	377
				NULL	89920
AcudioFarmaceuta	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119773
				Si	665
				NULL	89540
AsistioTerapiaAlternativa	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119872
				Si	22
				NULL	90084
ConsultoTegua	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119873
				Si	27
				NULL	90078
NotuvoTiempo	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119823
				Si	462
				NULL	89693

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
CentroAtencionLejos	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119852
				Si	285
				NULL	89841
MalServicio	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119856
				Si	297
				NULL	89825
CitaDistante	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119848
				Si	219
				NULL	89911
TramitesCita	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119833
				Si	351
				NULL	89794
NoLoAtendieron	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119827
				Si	254
				NULL	89897
ConsultoAntesSinSolucion	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119868
				Si	57
				NULL	90053
NoConfiaMedicos	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119864
				Si	105
				NULL	90009

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
FaltaDinero	Categórica - Discreta	Acceso a servicios de salud	proximal	No	119803
				Si	945
				NULL	89230
Estudiando	Categórica - Discreta	Acceso a servicios de educación	proximal	No	99286
				Si	110692
En vacaciones escolares	Categórica - Discreta	Acceso a servicios de educación	proximal	No	200332
				Si	9646
Oficios escolares	Categórica - Discreta	Acceso a servicios de educación	proximal	No	206874
				Si	3104
Etnia	Categórica - Discreta	Creencias del individuo	proximal	Afrocolombiano	691
				Indígena	403
				Ninguno	208610
				Rom/Gitano	274
EstadoCivil	Categórica - Discreta	Nupcialidad y/o Uniones tempranas	proximal	Casado	626
				Separado	1576
				Soltero	192925
				Unión Libre	14680
				Viudo	171
NumeroPersonas	Numérica - Continua	Familia	Intermedio	Cantidad de individuos (Entre 1 y 60)	

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
Genero Jefe Hogar	Categórica - Discreta	Hogares con Jefatura femenina	Intermedio	Femenino	74345
				Masculino	135633
IdFamilia (Se utilizó para hacer la búsqueda de embarazo adolescente en la familia)	Categórica - Discreta	Historia de embarazo adolecente en la Familia	Intermedio	Identificador Numérico	
RelacionJefe	Categórica - Discreta	lugar que se ocupa en la familia	Intermedio	Cónyuge	12446
				Hermana	1963
				Hija	160686
				Hijastra	11192
				Jefe Hogar	7781
				Nieta	7660
				No Pariente	1250
				Otro Pariente	5519
				Yerno/Nuera	1481
EdadEncuesta	Numérica - Continua	Edad	Intermedio	Entero correspondiente a la edad (Entre 10 y 19)	
FechaNacimiento	Numérica	Edad	Intermedio	Valores de tipo Fecha.	
Analfabeta	Categórica - Discreta	Nivel educativo	Intermedio	No	206015
				Si	3962
				NULL	1

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
Primaria	Numérica - Continua	Nivel educativo	Intermedio	1	1661
				2	4254
				3	15429
				4	50692
				5	762445
Secundaria	Numérica - Continua	Nivel educativo	Intermedio	1	21952
				2	48214
				3	68508
				4	88296
				5	94670
				6	243798
Técnica	Numérica - Continua	Nivel educativo	Intermedio	1	1335
				2	1940
				3	2157
				4	1852
				5	305
				6	528
Tecnológica	Numérica - Continua	Nivel educativo	Intermedio	1	137
				2	236
				3	255
				4	312

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
				5	160
				6	288
				8	24
Universitaria		Nivel educativo	Intermedio	1	5
				2	13
				3	7
				4	1
				5	0
				6	6
Postgrado	Numérica - Continua	Nivel educativo	Intermedio	1	5
				2	4
				3	3
				4	8
Fuma	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	168117
				Si	25000
				NULL	16861
NotuvoTiempo	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	119823
				Si	462
				NULL	89693
Oficios escolares	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	206874
				Si	3104

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
Oficios del hogar	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	206336
				Si	3642
Otras actividades	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	200112
				Si	9866
OtraActividadPaga	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	137951
				Si	342
				NULL	71685
TrabajooIngresosUltSemana	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	138159
				Si	115
				NULL	71704
TrabajoSinPagaUltSemana	Categórica - Discreta	Manejo del tiempo libre	Intermedio	No	138011
				Si	260
				NULL	71707
Tipo vivienda	Categórica - Discreta	Nivel de pobreza	Distal	Apartamento	32985
				Carpa	349
				Casa	150459
				Casa, Apartamento	16008
				Móvil, Refugio Natural, Carpa	224
				Otro	1730
				Pieza	8079

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
				Refugio Natural	144
EstadoViviendaAdecuado	Categórica - Discreta	Nivel de pobreza	Distal	No	12668
				Si	157707
				NULL	39603
Servicios Sanitarios	Categórica - Discreta	Nivel de pobreza	Distal	Campo Abierto	652
				Letrina	707
				NULL	203415
				Pozo Séptico	1175
				Sin definir	4029
MaterialViviendaApropiado	Categórica - Discreta	Nivel de pobreza	Distal	No	9027
				Si	161347
				NULL	39604
CocinaIndependiente	Categórica - Discreta	Nivel de pobreza	Distal	No	23080
				Si	186898
Gasolina	Categórica - Discreta	Nivel de pobreza	Distal	No	207152
				Si	2826
Gas	Categórica - Discreta	Nivel de pobreza	Distal	No	159564
				Si	50414
Lena	Categórica - Discreta	Nivel de pobreza	Distal	No	208295
				Si	1683
Luz	Categórica -	Nivel de	Distal	No	208161

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
	Discreta	pobreza		Si	1817
EnergiaElectrica	Categórica - Discreta	Nivel de pobreza	Distal	No	2592
				Si	207386
GasNatural	Categórica - Discreta	Nivel de pobreza	Distal	No	32863
				Si	177115
Acueducto	Categórica - Discreta	Nivel de pobreza	Distal	No	6305
				Si	203673
Telefono	Categórica - Discreta	Nivel de pobreza	Distal	No	48457
				Si	161521
RecoleccionBasuras	Categórica - Discreta	Nivel de pobreza	Distal	No	6074
				Si	203904
VectoresVivienda	Categórica - Discreta	Nivel de pobreza	Distal	No	129815
				Si	80163
RiesgosVivienda	Categórica - Discreta	Nivel de pobreza	Distal	No	120098
				Si	50276
				NULL	39604
Desnutrido	Categórica - Discreta	Nivel de pobreza	Distal	No	207875
				Si	2102
				NULL	1
DejoAlimintos	Categórica - Discreta	Nivel de pobreza	Distal	No	183837
				Si	26141

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
RedujoCantidadAlimentos	Categórica - Discreta	Nivel de pobreza	Distal	No	186401
				Si	23577
FaltoDieneroCompraAlimentos	Categórica - Discreta	Nivel de pobreza	Distal	No	165202
				Si	27915
				NULL	16861
PoblacionEspecial	Categórica - Discreta	Inequidad social	Distal	Condición de Desplazamiento	2514
				Habitante de calle	82
				Menor Abandonado	65
				Ninguno	205526
				Reinsertado	246
				Situación de Desplazamiento	1545
Desempleado	Categórica - Discreta	Inequidad social	Distal	No	51986
				Si	157991
				NULL	1
TiempoDesempleado	Categórica - Discreta	Inequidad social	Distal	Cantidad en días (Entre 1 y 120)	
TipoEmpleo	Categórica - Discreta	Inequidad social	Distal	Empleado	4710
				Empleador	43

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
				Miembro de cooperativa de productores	123
				NULL	185716
				Trabajador Familiar Auxiliar	1097
				Trabajador No Formal	4583
				Trabajador por cuenta propia	2758
				Trabajador que no puede clasificarse	10948
Ocupacion	Categórica - Discreta	Inequidad social	Distal	Agricultor	912
				Ama de casa	16704
				Fuerzas Militares	38
				Ninguno	67898
				Profesional titulado	177
				Profesional titulado y especializado	51
				Técnico y tecnólogo	790

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
				Trabajador no calificado	20132
				NULL	103276
PosicionOcupacional	Categórica - Discreta	Inequidad social	Distal	Ayudante sin remuneración	4053
				Empleada Doméstica	989
				Jornalero o peón	610
				Obrero o empleado de empresa particular	5814
				Obrero o empleado del gobierno	156
				Patrón o empleador	79
				Profesional Independiente	210
				Trabajador de su propia finca	84
				Trabajador familiar sin remuneración	20596
				Trabajador independiente o por cuenta	5423

Variable del conjunto de datos	Tipo	Determinante Social	Tipo de determinante Social	valores	Frecuencia del valor
				propia	
				NULL	171964
Trabajando	Categórica - Discreta	Inequidad social	Distal	No	209026
				Si	952
Buscando trabajo	Categórica - Discreta	Inequidad social	Distal	No	209386
				Si	592
TrabajooIngresosUltSemana	Categórica - Discreta	Inequidad social	Distal	No	138159
				Si	115
				NULL	71704
TrabajoSinPagaUltSemana	Categórica - Discreta	Inequidad social	Distal	No	138011
				Si	260
				NULL	71707

B. Anexo: Sentencias SQL para la extracción del conjunto desde el Data Where House

```sql
--- ********************************************************************************************************
--- *************************   Generear tablas de embarazos ***********************************************
--- ********************************************************************************************************
--- dmmoreno: La clausula  YEAR(dbo.FichaAPS.FechaEncuesta) <= 2006 Garantiza que se ha realizado seguimiento durante los ultimos 10 años
  -- Mujeres captadas en estado de embarazo


SELECT b.IdIndividuo , COALESCE(a.Embarazo,'NO') AS Embarazo
INTO SIPETE_Adolescentes
FROM (
        SELECT     dbo.ArchivoIndividuo_NEW.IdIndividuo , 'SI' as Embarazo
        FROM         dbo.ArchivoIndividuo_NEW INNER JOIN
                              dbo.FichaAPS ON dbo.ArchivoIndividuo_NEW.Ficha = dbo.FichaAPS.FichaAPS INNER JOIN
                              dbo.Familia ON dbo.FichaAPS.IDFamilia = dbo.Familia.ID
        WHERE  UPPER(dbo.ArchivoIndividuo_NEW.EmbarazadaActualmente) = 'SI'
                AND round(dbo.ArchivoIndividuo_NEW.EdadEnDiasActual/365,0) BETWEEN 10 AND  19
                --AND YEAR(dbo.FichaAPS.FechaEncuesta) <= 2006
                AND dbo.ArchivoIndividuo_NEW.Genero = 'Femenino'
        UNION  -- Mujeres que en algún momento durante los 10 años de seguimiento reportaron un embarazo
        SELECT    dbo.ArchivoIndividuo_NEW.IdIndividuo, 'SI' as Embarazo
        FROM         dbo.ArchivoIndividuoNovedadAlerta_NEW INNER JOIN
                              dbo.FichaAPS ON dbo.ArchivoIndividuoNovedadAlerta_NEW.FichaAPS = dbo.FichaAPS.FichaAPS COLLATE Latin1_General_CI_AS INNER JOIN
                              dbo.ArchivoIndividuo_NEW ON dbo.ArchivoIndividuoNovedadAlerta_NEW.FichaAPS = dbo.ArchivoIndividuo_NEW.Ficha AND
                              dbo.ArchivoIndividuoNovedadAlerta_NEW.NumeroDocumento = dbo.ArchivoIndividuo_NEW.NumeroDocumento INNER JOIN
                              dbo.Familia ON dbo.FichaAPS.IDFamilia = dbo.Familia.ID
        WHERE     (dbo.ArchivoIndividuoNovedadAlerta_NEW.Nombre = 'EMBARAZO NUEVO')
                AND round(dbo.ArchivoIndividuo_NEW.EdadEnDiasActual/365,0) BETWEEN 10 AND  19
                --AND YEAR(dbo.FichaAPS.FechaEncuesta) <= 2006
                AND dbo.ArchivoIndividuo_NEW.Genero = 'Femenino'
        GROUP BY dbo.ArchivoIndividuo_NEW.IdIndividuo
    )a
    RIGHT JOIN (
        SELECT IdIndividuo
        FROM (
                SELECT     dbo.ArchivoIndividuo_NEW.IdIndividuo
                FROM         dbo.ArchivoIndividuo_NEW INNER JOIN
                                      dbo.FichaAPS ON dbo.ArchivoIndividuo_NEW.Ficha = dbo.FichaAPS.FichaAPS INNER JOIN
                                      dbo.Familia ON dbo.FichaAPS.IDFamilia = dbo.Familia.ID
                WHERE  UPPER(dbo.ArchivoIndividuo_NEW.EmbarazadaActualmente) IN ('NO','SI')
                        AND round(dbo.ArchivoIndividuo_NEW.EdadEnDiasActual/365,0) BETWEEN 10 AND  19
                        --AND YEAR(dbo.FichaAPS.FechaEncuesta) <= 2006
                        AND dbo.ArchivoIndividuo_NEW.Genero = 'Femenino'
                )x
    )b
    ON a.IdIndividuo = b.IdIndividuo
```

```
--- ********************************************************************************************************
--- ************************* DATOS PARA LA INVESTIGACIÓN - SEGUNDO ENFOQUE *********************************
--- ********************************************************************************************************

SELECT        dbo.SIPETE_Adolescentes.IdIndividuo, dbo.SIPETE_Adolescentes.Embarazo,
        -- // DETERMINANTES PROXIMALES
          -- Behavioral Factors
                ArchivoIndividuo_NEW.EstadoCivil,   -- Uniones Tempranas
                ArchivoIndividuo_NEW.Etnia,         -- Creencias*
                ArchivoIndividuo_NEW.CondicionSGSSS, -- Acceso a servicios de salud
                ArchivoIndividuo_NEW.Estudiando, -- Acceso a servicios de educación

        -- // DETERMINANTES INTERMEDIOS Y/O DE ESTRUCTURA
            -- Factores Interpersonales
             ArchivoIndividuosJefesDeHogar.Genero AS GeneroJefeHogar, -- Jefatura de Hogar Femenina
             CASE   WHEN ArchivoIndividuo_NEW.RelacionJefe = 'Abuelo(a)' THEN 'Nieto(a)'
                    WHEN ArchivoIndividuo_NEW.RelacionJefe = 'Suegro(a)' THEN 'Yerno-Nuera'
                    WHEN ArchivoIndividuo_NEW.RelacionJefe = 'Padre o madre' THEN 'Hijo(a)'
                    ELSE  ArchivoIndividuo_NEW.RelacionJefe END
                AS RelacionConJefeDeHogar,  -- Posición que ocupa dentro de la famila
             CASE WHEN ArchivoIndividuosJefesDeHogar.EdadEnDiasActual/365.4 BETWEEN 30 AND 40 THEN 'SI' ELSE 'NO' END
                AS HistoriaEmbarazoAdolescenteMadre, -- Historial de embarazo adolescente en la familia
            ArchivoFamilia.ExitenPersonasSinCuidador, -- Abandono/Sin monitoreo
            ArchivoFamilia.MujeresSinCuidadorMenores, -- Abandono/Sin monitoreo
            -- Factores No Interpersonales
                CONVERT(NUMERIC(5,3),( 6 + ArchivoIndividuo_NEW.Primaria + ArchivoIndividuo_NEW.Secundaria +
                  0.5*ArchivoIndividuo_NEW.Técnica + 0.5*ArchivoIndividuo_NEW.Tecnológica + 0.5*ArchivoIndividuo_NEW.Universitaria)
                  /
                 (ArchivoIndividuo_NEW.EdadEnDiasActual/365.4) ,2)
             as NivelEducativo, -- Nivel Educativo
             ArchivoIndividuo_NEW.[Oficios escolares], -- Manejo del tiempo libre
             ArchivoIndividuo_NEW.[Oficios del hogar],
             ArchivoIndividuo_NEW.[Otras actividades],
        -- // DETERMINANTES DISTALES
            -- Factores Estructurales - Nivel de pobreza e inequidad social
            ArchivoFamilia.FaltoDineroCompraAlimentos,
            ArchivoFamilia.TipoVivienda,
            ArchivoFamilia.RiesgosVivienda,
                      -- Hacinamiento
           ArchivoFamilia.NumeroPersonas,
           ArchivoFamilia.NumeroDormitorios,
           COALESCE(CONVERT(NUMERIC(5,1),convert(float,ArchivoFamilia.NumeroPersonas)
                  /convert(float,ArchivoFamilia.NumeroDormitorios)),NULL)
               AS IndiceHacinamiento,
           CASE WHEN COALESCE(CONVERT(NUMERIC(5,1),convert(float,ArchivoFamilia.NumeroPersonas)
                  /convert(float,ArchivoFamilia.NumeroDormitorios)),NULL) <= 2.4 THEN 'Sin Hacinamiento'
                   WHEN COALESCE(CONVERT(NUMERIC(5,1),convert(float,ArchivoFamilia.NumeroPersonas)
                                /convert(float,ArchivoFamilia.NumeroDormitorios)),NULL) between 2.5 and 4.9 THEN 'Hacinamiento Medio'
                   WHEN COALESCE(CONVERT(NUMERIC(5,1),convert(float,ArchivoFamilia.NumeroPersonas)
                                /convert(float,ArchivoFamilia.NumeroDormitorios)),NULL) >= 5 THEN 'Hacinamiento Critico'
                  ELSE null END
               As Hacinamiento,
           CASE WHEN COALESCE(CONVERT(NUMERIC(5,1),convert(float,ArchivoFamilia.NumeroPersonas)
                     /convert(float,ArchivoFamilia.NumeroDormitorios)),NULL) <= 2.4 THEN 'Sin Hacinamiento'
                   WHEN COALESCE(CONVERT(NUMERIC(5,1),convert(float,ArchivoFamilia.NumeroPersonas)
                                /convert(float,ArchivoFamilia.NumeroDormitorios)),NULL) between 2.5 and 2.9 THEN 'Hacinamiento Medio'
                   WHEN COALESCE(CONVERT(NUMERIC(5,1),convert(float,ArchivoFamilia.NumeroPersonas)
                                /convert(float,ArchivoFamilia.NumeroDormitorios)),NULL) >= 3 THEN 'Hacinamiento Critico'
                  ELSE null END
               As Hacinamiento_2
                /*         mientras que su división de estadística utiliza la medida de personas por habitación (ONU, 2005)
                Se entiende por índice de hacinamiento a la relación:
                   ihacinam. = (personas habitando una vivienda) / (número de dormitorios en la vivienda)
                   Generalmente se aceptan los valores:
                   hasta 2.4 - sin hacinamiento;
                   de 2.5 a 4.9 - hacinamiento medio;
                   más de 5.0 - hacinamiento crítico.

           Hacunamiento 1. Dr. Adonis Arias L. Enfermedades Infecciosas , Medicina Preventiva y Salud Publica. Publicado: 27/05/2009 Portales médicos
           Hacinamiento 2. REPORTE DE INDICADORES ONU-HABITAT en las CIUDADES VERACRUZANAS PERIODO 2000-2010
                */
FROM           dbo.ArchivoIndividuosJefesDeHogar INNER JOIN
                      dbo.SIPETE_Adolescentes INNER JOIN
                      dbo.ArchivoIndividuo_NEW ON dbo.SIPETE_Adolescentes.IdIndividuo = dbo.ArchivoIndividuo_NEW.IdIndividuo INNER JOIN
                      dbo.ArchivoFamilia ON dbo.ArchivoIndividuo_NEW.Ficha = dbo.ArchivoFamilia.FichaAPS ON
                      dbo.ArchivoIndividuosJefesDeHogar.Ficha = dbo.ArchivoIndividuo_NEW.Ficha
```

Bibliografía

Ball, N. M., & Brunner, R. J. (2010). Data mining and machine learning in astronomy. *International Journal of Modern Physics D, 19*(7), 1049–1106.

Bennett, K. P. (1992). Decision tree construction via linear programming. *Proceedings 4th Midwest Artificial Intelligence and Cognitive Science Society Conference*, 97–101.

Berthold, M. R., & Diamond, J. (21 April 1998). Constructive training of probabilistic neural networks. *Neurocomputing, 19, Issues 1–3*, 167–183.

Carse, B. T., & A., M. (1996). Evolving fuzzy rule based controllers using genetic algorithms. *Fuzzy Sets and Systems, 80*(3), 273–293.

Chai, J., Liu, J. N., & Ngai, E. W. (2013). Application of decision-making techniques in supplier selection: A systematic review of literature. *Expert Systems with Applications*.

Chen, Z., & Zhu, Q. (Aug. de 1998). Query construction for user-guided knowledge discovery in databases. *Information Sciences, 109*(1–4), 49–64.

Cismondi, F., Fialho, A. S., & Vieira, S. M. (2011). Predicting laboratory testing in intensive care using fuzzy and neural modeling. *IEEE International Conference on Fuzzy Systems*, 2096–2103.

DNP-DSS.SS, Ministerio de Protección Social. (s.f.). *Revisión bibliográfica estudios de Determinantes de la Salud de la OMS.* Documentos Compes.

Ester , M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD 96)*, (págs. 226-231). Portland, USA.

Ester, M., Frommelt, A., Kriegel, H., & Sander, J. (2000). Spatial data mining: Database primitives, algorithms and efficient DBMS support. *4*(2-3), 193–216.

Fayyad, U. (1996). Data mining and knowledge discovery: making sense out of data. *IEEE Computer Society, 11*(5), 20–25.

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems, 13*(2–3), 99–115.

Fayyad, U., & Uthurusamy, R. (1996). Data mining and knowledge discovery in databases. (39, Ed.) *Communications of the ACM*.

Feyyad, U. (1996). Data mining and knowledge discovery: making sense out of data. *IEEE Computer Society, 11*(5), 20–25.

Friedlin, J., Mahoui, M., Jones, J., & Jamieson, P. (2011). Knowledge discovery and data mining of free text radiology reports. *in Proceedings - 2011 1st IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology*, 89–96.

Gent, I., & Walsh , T. (1992). *The Enigma of SAT Hillclimbing.* Edinburgh: Technical Report 605, Dept. of Artificial Inteligence, University of Edinburgh.

Goodwin, L., Vandyne, M., Lin, S., & Talbert, S. (2003). Data mining issues and opportunities for building nursing knowledge. *Journal of Biomedical Informatics, 36*(4–5), 379–388.

Han, J., & Fu, Y. (1999). Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering, 11*(5), pp. 798–805.

Han, J., Nishio, S., Kawano, H., & Wang, W. (1998). Generalization-based data mining in object-oriented databases using an object cube model. *Data & Knowledge Engineering, 25*(1-2), 55–97.

Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery, 2*(1), 9–37.

Herrera, F. (2008). Genetic fuzzy systems: Taxonomy, current research trends and prospects. *Evolutionary Intelligence, 1*(1), 27–46.

Hilderman, R. J., Hamilton, H. J., & Cercone, N. (1999). Data mining in large databases using domain generalization graphs. *Journal of Intelligent Information Systems, 13*(3), 195–234.

Holzinger, A. (2012). On knowledge discovery and interactive intelligent visualization of biomedical data: Challenges in human-computer interaction & biomedical informatics. *DATA 2012 - Proceedings of the International Conference on Data Technologies and Applications*, SI5–SI16.

Huang, S.-M., Hsu, P.-Y., & Wang, W.-C. (2012). A Study on the Modified Attribute Oriented Induction Algorithm of Mining the Multi-value Attribute Data. *Intelligent Information and Database Systems*, 348-358.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666.

Kavurucu, Y., Senkul, P., & Toroslu, I. H. (Dec. de 2010). Concept discovery on relational databases: New techniques for search space pruning and rule quality improvement. *Knowledge-Based Systems, 23*(8), 743–756.

Köksal, G., Batmaz, I., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications, 38*(10), 13448–13467.

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine, 23*(1), 89–109.

Körting, T. S., Garcia, L. M., & Camara, G. (August de 2013). GeoDMA—Geographic Data Mining Analyst. *Computers & Geosciences, 57*, 133–145.

Li, L., Wang, J., & Leung, H. S. (2012). A Bayesian Method to Mine Spatial Data Sets to Evaluate the Vulnerability of Human Beings to Catastrophic Risk. *Risk Analysis, 32*(6), 1072–1092.

Li, Y. (2012). An improved time-series data mining approach for medical data mining. *Journal of Convergence Information Technology, 7*(11), 141–149.

Magnisalis, I., Demetriadis, S., & Karakostas, A. (2011). Adaptive and intelligent systems for collaborative learning support: A review of the field. *IEEE Transactions on Learning Technologies, 4*(1), 5–20.

Matheus, C. J., Chan, P. K., & Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases. *IEEE Transactions on Knowledge and Data Engineering, 5*(6), 903–913.

Minnie, D., & Srinivasan, S. (2011). Application of Knowledge Discovery in Database to blood cell counter data to improve quality control in clinical pathology. *Proceedings - 2011 6th International Conference on Bio-Inspired Computing: Theories and Applications, BIC-TA 2011*, 338–342.

Padhy, N. P. (2005). *Systems, Artificial Intelligence and Intelligent.* Oxford: Oxford University Press,.

Pal, D., Mandana, K. M., & Pal, S. (2012). Fuzzy expert system approach for coronary artery disease screening using clinical parameters. *Knowledge-Based Systems, 36*(0), 162–174.

Pascual, D., & Pla, F. S. (2007). Algoritmos de agrupamiento.

Riedmiller, M., & Braun, H. (1993). Direct adaptive method for faster backpropagation learning: The RPROP algorithm. *IEEE International Conference on Neural Networks,*, (págs. 586–591). Piscataway, NJ, United States.

Sebastian, Y., & Then, P. H. (2011). Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses. *Knowledge-Based Systems, 24*(5), 609–620.

Secretaria Distrital de Salud. (2007). Boletin de Estadisticas" . 120.

Skulimowski, A. M. (2011). Future trends of intelligent decision support systems and models. *Communications in Computer and Information Science, 184 CCIS*(1), 11–20.

Spetch, D. F. (1990). Probabilistic Neural Networks. *Elsevier*, 109-118.

Srimani, P. K., & Koti, M. S. (2011). A Comparison of different learning models used in data mining for medical data. *AIP Conference Proceedings, 1414*, 51–55.

Srinivasa, K. G., Jagadish, M., Venugopal, K. R., & Patnaik, L. M. (2007). Data mining based query processing using rough sets and genetic algorithms. *in Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2007*, 275–282.

Suematsu, N., Nakayasu, T., & Hayashi, A. (2002). A classifier learning method through data summaries. *Transactions of the Japanese Society for Artificial Intelligence, 17*(5), 565–575.

Tan, P.-N., Steinbach, M., & Kumar, V. (s.f.). *Introduction To Data Mining.* Pearson.

Tesauro, G. J., & Kephart, J. O. (2000). Foresight-based pricing algorithms in agent economies. *Decision Support Systems, 28*(1), 49–60.

Thakur, R., & Mahajan, A. (2015). Preprocessing and Classification of Data Analysis in Institutional System using Weka. *International Journal of Computer Applications, Vol.112, No. 6*, 9-11.

Then, P. H. (2011). Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses. *Knowledge-Based Systems, 24*(5), 609–620.

Tom Fawcett. (2004). ROC graphs: Notes and practical considerations for researchers. *HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto, CA 94304.*

Tsoukalas, L. H. (1998). Neurofuzzy approaches to anticipation: A new paradigm for intelligent systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 28*(4), 573–582.

Vashishtha, J., Kumar, D., & Ratnoo, S. (2011). Revisiting interestingness measures for knowledge discovery in databases. *in Proceedings - 2012 2nd International Conference on Advanced Computing and Communication Technologies, ACCT 2012,*, 72–78.

Verma, I. S. (2015). Knowledge Data Discovery and Its Issues. *Expansion, Impact and Challenges of IT & CS*, 88.

Wang, S.-L., Maskey, R., Jafari, A., & Hong, T.-P. (2008). Efficient sanitization of informative association rules. *Expert Systems with Applications, 35*(1–2), 442–450.

Xu, Y. (2010). A Study for Important Criteria of Feature Selection in Text Categorization. *2nd International Workshop on Intelligent Systems and Applications*, 1-4.

Yang, J. (2001). Intelligence optimization algorithms: A survey. *International Journal of Advancements in Computing Technology, 3*(4), 144–152.

Yang, Y., Slattery, S., & Ghani, R. (2002). A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems, 18*(2–3), 219–241.

Zamudio, L. W., & Rubiano, N. (1999). *El aborto inducido en Colombia: Características demográficas y socioculturales* (ISBN/ISSN: 0122-7815 ed.). Bogotá: Universidad Externado de Colombia.

Zhang, X., Hu, Y., & Xie, K. (May de 2015). An evolutionary trend reversion model for stock trading rule discovery. *Knowledge-Based Systems, 79*, 27–35.